UNIVERSIDAD DE GRANADA

E.T.S. Ingenieros de Caminos, Canales y Puertos
Departamento de Ingeniería Civil
Área de Ingeniería e Infraestructuras de los Transportes

TESIS DOCTORAL

ANÁLISIS DE LA SEVERIDAD DE LOS ACCIDENTES DE TRÁFICO
UTILIZANDO TÉCNICAS DE MINERÍA DE DATOS

ANALYSIS OF TRAFFIC CRASHES' SEVERITY
USING DATA MINING TECHNIQUES

Para la obtención del

GRADO DE DOCTOR POR LA UNIVERSIDAD DE GRANADA CON MENCIÓN DE DOCTORADO INTERNACIONAL

AUTOR:

GRISELDA LÓPEZ MALDONADO

DIRECTORES:

D. JUAN DE OÑA LÓPEZ. Universidad de Granada

D. JOAQUÍN ABELLÁN MULERO. Universidad de Granada

D. ALFONSO MOTELLA. Università degli Studi di Napoli Federico II

GRANADA, 2013

Editor: Editorial de la Universidad de Granada
Autor: Griselda López Maldonado
D.L.: GR 195-2014
ISBN: 978-84-9028-715-6

UNIVERSIDAD DE GRANADA

E.T.S. Ingenieros de Caminos, Canales y Puertos
Departamento de Ingeniería Civil
Área de Ingeniería e Infraestructuras de los Transportes

TESIS DOCTORAL

ANÁLISIS DE LA SEVERIDAD DE LOS ACCIDENTES DE TRÁFICO
UTILIZANDO TÉCNICAS DE MINERÍA DE DATOS

ANALYSIS OF TRAFFIC CRASHES' SEVERITY
USING DATA MINING TECHNIQUES

GRISELDA LÓPEZ MALDONADO

DIRECTORES:

D. JUAN DE OÑA LÓPEZ

Doctor Ingeniero de Caminos, Canales y Puertos

Universidad de Granada

D. JOAQUÍN ABELLÁN MULERO

Doctor en Ciencias Matemáticas

Universidad de Granada

D. ALFONSO MOTELLA.

Dottore di ricerca in Ingegneria dei Trasporti

Università di Roma "La Sapienza"

TESIS DOCTORAL - 2013

Memoria presentada por Dª Griselda López Maldonado para aspirar al grado de Doctor por la Universidad de Granada con mención de Doctorado Internacional.

La doctoranda Griselda López Maldonado y los directores de la tesis "Análisis de los accidentes de tráfico utilizando técnicas de Minería de Datos" garantizamos, al firmar esta tesis doctoral, que el trabajo ha sido realizado por la doctoranda bajo la dirección de los directores de la tesis. Y hasta donde nuestro conocimiento alcanza, en la realización del trabajo, se han respetado los derechos de otros autores a ser citados, cuando se han utilizado sus resultados o publicaciones.

Granada, junio 2013

Los Directores de la Tesis:

Fdo.: D. Juan de Oña López

Fdo.: D. Joaquín Abellán Mulero

Fdo.: D. Alfonso Montella

La Doctoranda:

Fdo: Griselda López Maldonado

AGRADECIMIENTOS

Largo ha sido el camino recorrido hasta llegar a la meta. Pero finalmente, con esfuerzo y constancia, hasta las metas más lejanas pueden alcanzarse.

Quiero agradecer a mis Directores de Tesis, Juan de Oña, Joaquín Abellán y Alfonso Montella, por el tiempo invertido y el interés constante en el avance y el desarrollo de esta tesis doctoral. Muchas gracias por darme la oportunidad de trabajar con vosotros, por guiarme y por asesorarme en todas las etapas de esta investigación.

A lo largo de estos años, muchas han sido las personas que me han mostrado su apoyo constante, que me han dado ánimos y que han estado ahí en los momentos más difíciles. Quiero expresar a todas ellas mi más sincero agradecimiento.

En especial, quiero agradecer a mis padres, a mi hermano y demás familiares, que me han ayudado y apoyado en todos los momentos difíciles por los que se pasa durante la elaboración de una tesis.

A mi marido Alberto, por su apoyo constante, su paciencia, su comprensión, su ayuda, y por darme la fuerza que a veces me faltaba, creyendo en mí, más que yo misma. Muchas gracias de todo corazón.

A mis compañeros del grupo de investigación TRYSE, así como a mis compañeras del resto de departamentos, por todos los momentos de risas y lágrimas que hemos compartido en estos años. Y a todos los amigos que de un modo u otro, me han ayudado, gracias a todos ellos.

A la Consejería de Economía, Innovación, Ciencia y Empleo de la Junta de Andalucía, responsable de la beca de formación de personal docente e investigador en Áreas Deficitarias de la que fui adjudicataria en el año 2009.

Y a la Dirección General de Tráfico, por habernos proporcionado los datos de accidentes necesarios para realizar esta tesis doctoral, y sin los cuales, no hubiera sido posible desarrollar este trabajo.

RESUMEN

Reducir el impacto socio-económico de los accidentes de tráfico sigue siendo una de las prioridades estratégicas planteadas en los planes de seguridad vial, y concretamente una de las claves para conseguir este objetivo, es mejorar de la seguridad vial en las carreteras convencionales. Por ello en esta tesis doctoral se propone un análisis en profundidad de la accidentalidad de estas carreteras.

Existen diferentes enfoques para llevar a cabo el estudio de los accidentes de tráfico, y en esta investigación se realiza en términos de la gravedad de sus consecuencias.

Hasta el momento, las técnicas más utilizadas para analizar la gravedad de los accidentes han sido los Discrete Outcome Models (DOM), sin embargo, en la actualidad numerosos investigadores han comenzado a utilizar técnicas que se encuentran dentro del campo de la Minería de Datos (MD). Estas técnicas permiten extraer conocimiento de los datos, previamente desconocido e indistinguible, y normalmente, no parten de hipótesis ni requieren un previo conocimiento probabilístico del problema objeto de estudio.

Particularmente, los Árboles de Decisión (ADDs) son una técnica de MD muy apropiada para el estudio de los accidentes de tráfico, por diferentes razones: son fácilmente interpretables, pueden trabajar con grandes bases de datos, y descubrir fácilmente complejas interacciones entre los datos. Un aspecto a destacar, es que además, permiten la extracción de Reglas de Decisión (RDs) del tipo "SI-ENTONCES", que pueden ser usadas para descubrir determinados patrones de comportamiento que ocurren dentro de un conjunto de datos. Estos patrones pueden ayudar a la comprensión del suceso de un accidente, así como a la identificación de las principales variables que determinan su gravedad.

Los ADDs pueden ser construidos con diferentes algoritmos. El algoritmo CART (Breiman et al., 1984) ha sido hasta el momento el más ampliamente utilizado en las investigaciones de seguridad vial, sin embargo este método solo permite la construcción de árboles binarios, y su interpretación, para el estudio de los accidentes, puede ser ineficiente en determinadas ocasiones. Dado que existen otros algoritmos, muy utilizados en la literatura de MD, que permiten la construcción de ADDs sin esta restricción binaria, y que pueden ser más explicativos de cara al estudio de la gravedad del accidente, en esta tesis doctoral se propone su aplicación.

La principal limitación de las RDs que se extraen de los ADDs es que son dependientes de la estructura del árbol, de modo que pueden existir ciertos patrones de accidentes que no sean detectados. Por lo que a priori, no se estaría obteniendo todo el conocimiento posible de la base de datos de accidentes analizada. Por ello, en esta tesis doctoral se propone utilizar además, un nuevo método de extracción de RDs, *Information Root Node Variation*, que resuelve esta limitación, y permite extraer todo el conocimiento existente de la base de datos analizada.

Con los resultados obtenidos en esta investigación se demuestra que los ADDs son una herramienta adecuada para analizar los accidentes de tráfico de un modo sencillo y fácilmente comprensible para los analistas de la seguridad vial. Que permiten obtener la importancia de las variables en el modelo, y por tanto las variables con mayor influencia en la gravedad del accidente. Y que la extracción de RDs resulta de vital interés para los analistas y gestores de seguridad vial.

Con las RDs particularmente obtenidas, se han identificado problemáticas concretas de seguridad vial de las carreteras analizadas, sobre las que las Administraciones competentes podrían realizar actuaciones concretas. En una primera fase las actuaciones pueden centrarse en los accidentes graves o mortales, y posteriormente intervenir en los accidentes leves. Dentro de cada uno de estos grupos, el planteamiento propuesto en esta investigación permitiría priorizar actuaciones basándose en parámetros que expresan la importancia de los patrones que se obtendrán. Estos parámetros cuantificarán distintas características de las reglas que surgirán de nuestros procedimientos cuando se aplican a datos concretos.

ABSTRACT

Reducing socio-economic impact of traffic accidents remains one of the strategic priorities raised in road safety plans, and specifically one of the keys for achieving this goal is to improve road safety on the two-lane rural highways. Therefore, in this Ph. D. thesis we propose an in-depth analysis of these road accidents.

There are different approaches to carrying out the study of traffic accidents, and in this research it will be studied in terms of severity of their consequences.

Discrete Outcome Moldels (DOM) have been the techniques the most commonly used to study traffic accidnets severity. Recently, many researchers have begun to use techniques of Data Mining (DM). These techniques allow extracting knowledge from data, previously unknown and indistinguishable. Usually they are not based on assumptions, and do not require prior probabilistic knowledge on the study phenomena.

Decision Trees (DTs) are a DM technique. DTs are particularly appropriate for studing crashes for different reasons: they are easy to interpret, can work with large databases, and easily discover complex interactions between data. In addition, they pertmit the extraction of the Decision Rules (DRs) of the "IF-THEN" type. These rules can be used to discover certain patterns of behavior that occur within a specified set of data. And these patterns canhelp to understand the events leading up to a crahs and identify the variables that determine how serious an accident will be.

DTs can be constructed with different algorithms. The CART algorithm (Breiman et al., 1984) has been the one most commonly used by road safety researchers, however this method always yields binary trees, and their interpretation for study accident may be inefficient in some cases. In this Ph.d Thesis other algorithms widdely used in the literature of DM, and do not involve the binary restriction, and more apprpiated for study accidents, will be used.

The main limitation of the RDs extracted from the DTs is that they are dependent on the tree's structure, so, there could be other patterns of accidents that are not detected. Thus, a priori, we would not be getting that all possible knowledge of the database could be extracted. Therefore, in this thesis, it is used a new method of extracting RDs, Information Variation Root Node; which resolves this limitation, and permit to extract all the knowledge from a particular dataset.

With the results of this research we show that ADDs are a suitable tool for analyzing traffic accidents in a simple and easily understandable manner for road safety analysts. Also it is shown that DTs permit obtain the importance of the variables in the model, and therefore the variables with the most influence on the severity of the accident. And that the extraction of DRs is of vital interest to analysts and managers of road safety.

With the particular DRs obtained, specific patterns on the roads analyzed, have been identified. These patterns can help to carry out specific actions by the Authorities. In the first phase actions can focus on severe and fatal crashes and subsequently intervene in minor accidents. The approach proposed in this paper within each group will be enable actions to be prioritized based on parameter values that express the importance of the patterns to be obtained. These parameters quantify different characteristics of the rules that emerge from our procedures when they are applied to specific data.

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

CAPÍTULO 1.
INTRODUCCIÓN

CAPÍTULO 1. INTRODUCCIÓN

1.1. Visión general de problema.

Los accidentes de tráfico constituyen uno de los principales problemas de la sociedad actual. Según el informe publicado por la Organización Mundial de la Salud (WHO, 2013), el número de muertes por accidentes de tráfico se mantiene inaceptablemente alto (1,24 millones de muertes al año, según los datos de 2010).

En la mayoría de los países desarrollados la tasa de mortalidad por accidentes de tráfico está disminuyendo, mientras que los países con ingresos bajos y medios han sufrido un incremento global del número de muertes y lesiones por accidentes de tráfico. De hecho, si las tendencias actuales se mantienen, en 2030 las muertes por accidente de tráfico se convertirán en la quinta causa de muerte (WHO, 2009), con gran disparidad entre los diferentes países según los ingresos.

Los accidentes de tráfico suponen además, un coste social y económico muy relevante para los gobiernos de los países desarrollados; los cuales se ven obligados a destinar grandes cantidades de recursos para intentar paliar o disminuir este problema.

Sin embargo, la evidencia de muchos países demuestra que se pueden lograr éxitos espectaculares en la prevención de accidentes, mediante la adopción de leyes integrales sobre los factores de riesgo fundamentales, como son: exceso de velocidad, conducción bajo los efectos del alcohol, o la no utilización del casco de motociclista, del cinturón de seguridad y de sistemas de retención para niños (WHO, 2013).

Particularmente en España, la siniestralidad en las carreteras ha mejorado mucho en los últimos años. Es destacable que desde el año 2004 se observa una tendencia decreciente en el número de víctimas mortales producidas por accidentes de tráfico. Este hecho se debe principalmente a diferentes iniciativas relacionadas tanto con las infraestructuras (que han mejorado sustancialmente a lo largo de todo el período de estudio), como con los vehículos (actualmente se posee vehículos más seguros), con el propio conductor (introducción del carnet por puntos desde el año 2006), así como con una legislación cada vez más estricta, que penaliza algunos de los factores que más frecuentemente se encuentran presentes en la ocurrencia de un accidente (el alcohol o la velocidad). De modo que estas medidas han contribuido en el desarrollo de esta tendencia decreciente de la accidentalidad (Ministerio del Interior, 2005).

Sin embargo, reducir el impacto socio-económico de los accidentes de tráfico, sigue siendo una de las prioridades estratégicas planteadas en los planes de Seguridad Vial (Estrategia de Seguridad Vial 2011-2020). Y concretamente se establece como una de

las claves para conseguir este objetivo, la mejora de la seguridad vial en las carreteras convencionales.

Las carreteras convencionales representan una proporción sustancial en la mayoría de los países, y a su vez tienen mayor incidencia de accidentes y de gravedad de las lesiones que otros tipos de carreteras (Muellman y Mueller, 1996; Peek-Asa et al, 2004; Zwerling et al, 2005). Además, está estadísticamente demostrado que la tasa de accidentes mortales en zonas no urbanas es más de dos veces mayor que en zona urbana (Wang et al. 2008).

La justificación del estudio en particular de los accidentes ocurridos en carreteras convencionales, objeto de esta investigación, también se deriva fundamentalmente de dos hechos observados en las estadísticas de siniestralidad de España:

Aunque actualmente la siniestralidad en las zonas urbanas sigue siendo superior que en carreteras (un 54% de los accidentes ocurridos en el año 2010 fueron producidos en zona urbana); si los accidentes son analizados en términos de gravedad, se observa que la gravedad de los ocurridos en carretera es 3,5 veces mayor. Según las cifras del Ministerio del Interior, en el año 2010 se produjeron un 78% de fallecidos en carreteras frente a un 22% en zona urbana.

Dentro de la totalidad de los accidentes ocurridos en carreteras, el 75 % de los mismos se producen en carreteras convencionales, el 20 % en autovías, y el 5% restante en autopistas (Ministerio del Interior, 2010).

Esta situación, junto con el hecho de que los factores concurrentes en los accidentes suelen ser más complejos y complicados de analizar en las carreteras convencionales; hace que sea prioritario estudiar en profundidad este tipo de vías.

Tras esta visión general, se puede afirmar que es necesario seguir buscando medidas que permitan mejorar la seguridad vial de las carreteras. Y por ello se deben seguir dirigiendo grandes esfuerzos orientados a estudiar, analizar y comprender la complejidad de los accidentes de tráfico, con el objeto de así descubrir patrones de comportamiento que ayuden tanto a su prevención, como a la disminución de las consecuencias que se producen como resultado de los mismos.

Existen diferentes enfoques para llevar a cabo el estudio de los accidentes de tráfico. En esta tesis doctoral se realiza su estudio en términos de la gravedad de las lesiones resultantes como consecuencia de su ocurrencia de los mismos.

En general, la gravedad de la lesión se representa por medio de una serie de categorías discretas. Existen diferentes escalas que permiten clasificarla, siendo la escala KABCO la más comúnmente utilizada. Esta escala está basada en el uso de 5 niveles de gravedad: muerte o con lesiones mortales, lesiones incapacitantes, lesiones no incapacitantes, lesiones leves y solo daños materiales. Sin embargo, en España, se utilizan 3 niveles para definir la gravedad de la lesión: muerte, herido grave y herido leve.

Conocer la naturaleza y las particularidades de los datos de accidentes de tráfico es fundamental para el desarrollo y la aplicación de un modelo apropiado.

Hasta el momento, los Discrete Outcome Models (DOM) han sido los modelos más utilizados por los investigadores para analizar la gravedad de los accidentes, como se verá posteriormente en el capítulo del estado del arte. En particular, los más empleados han sido el Modelo Logit, el Modelo Logit Multinomial y el Modelo Probit Ordenado, según las consideraciones realizadas sobre la variable dependiente (la gravedad del accidente).

Dadas las particularidades de los datos de accidentes tales como la infra-detección (que se produce sobre todo en los accidentes con menor gravedad), la heterogeneidad de los datos (que puede producirse por ejemplo por las diferentes conductas entre los conductores, factores fisiológicos, etc), la omisión de variables relevantes de cara a su estudio (limitación debida a la información disponible en los partes de accidentes), los investigadores han ido introduciendo diversas modificaciones de los modelos originales. De modo que la elección del modelo ha estado basada en los problemas detectados a la hora de analizar los accidentes, así como el objetivo perseguido en cada estudio particular.

El principal inconveniente de estos modelos es que parten de hipótesis fijas y predefinen relaciones entre las variables dependientes e independientes, de modo que si estas hipótesis no se cumplen los modelos pueden producir estimaciones erróneas en la probabilidad de la gravedad de la lesión (Chang and Wang, 2006). Esto ha provocado que numerosos investigadores hayan comenzado a utilizar técnicas de Minería de Datos (MD) para estudiar la severidad de los accidentes de tráfico.

Las técnicas de MD permiten extraer conocimiento de los datos previamente desconocido e indistinguible, y normalmente no parten de hipótesis ni requieren un previo conocimiento probabilístico del problema objeto de estudio.

Además, el uso de técnicas de MD en el estudio de los accidentes resulta muy apropiado por la propia naturaleza de estos fenómenos. Un accidente puede definirse como un evento raro, aleatorio y de múltiples factores siempre precedido por una situación en la que uno o más conductores no pueden hacer frente al entorno de la carretera (ROSPA, 2002). Así, cada accidente es el resultado de una cadena de eventos que es, en su totalidad único, pero algunos factores son comunes a varias circunstancias del accidente, y la identificación de estos factores y sus interdependencias puede llevarse a cabo mediante el uso de técnicas de MD (Montella et al., 2012b).

Dentro de los modelos de MD existen diferentes tipos de técnicas. No existe una técnica universal que pueda ser aplicada para la resolución de cualquier tipo de problema (Hernández et al., 2004). Cada técnica presenta sus ventajas e inconvenientes y la elección de la misma dependerá del objetivo perseguido por el analista; siendo las Redes Neuronales Artificiales, las Redes Bayesianas, las Reglas de

Asociación y los Árboles de Decisión, las más utilizadas en el campo de la seguridad vial como se detalla en el capítulo del estado del arte.

Particularmente, los ADDs son una técnica de MD muy apropiada para el estudio de los accidentes de tráfico, ya que, teniendo en cuenta sus ventajas y limitaciones, constituyen uno de los modelos más utilizados en aprendizaje supervisado y en aplicaciones de Minería de Datos (Gehrke et al., 1999).

Entre las ventajas de los ADDs de cara a su utilización para el estudio de los accidentes se puede destacar que permiten la extracción de reglas de decisión del tipo "SI-ENTONCES". Estas reglas son fácilmente comprensibles por los gestores de seguridad vial y pueden ser usadas para descubrir determinados patrones de comportamiento que ocurren dentro de un conjunto de datos. Estos patrones pueden ayudar a la comprensión del suceso de un accidente, a la identificación de las principales variables que determinan su gravedad, así como al establecimiento de actuaciones concretas por parte de las Administraciones competentes con el fin de mejorar la seguridad vial de las carreteras analizadas.

1.2. Objetivos.

El principal objetivo de esta tesis doctoral es identificar patrones de accidentes que ocurren dentro de las carreteras convencionales mediante el uso de Reglas de Decisión extraídas de Árboles de Decisión. Para alcanzar este objetivo principal se han desarrollado una serie de objetivos específicos, tales como: la validación del uso de diferentes algoritmos para la construcción de ADDs para el estudio de la gravedad de los accidentes de tráfico; la identificación de las variables clave que afectan a la gravedad de los accidentes; la validación de las Reglas de Decisión para la identificación de patrones de accidentes; el uso de una metodología específica (*Information Root Node Variation*) que permite la extracción completa del conocimiento existente en la base de datos objeto de estudio, en forma de Reglas de Decisión; la verificación de los patrones obtenidos con esta metodología; y dados los patrones verificados, detección de las principales problemáticas de seguridad vial de las carreteras analizadas.

Los patrones identificados deben ser fácilmente comprensibles por los gestores de seguridad vial, para que las Administraciones competentes puedan realizar actuaciones concretas (ya sea en forma de actuaciones específicas en tramos de carreteras o con diseño de programas de educación y campañas de concienciación en materia de seguridad vial), con el fin de mejorar la seguridad vial de estas carreteras.

1.3. Estructura de la tesis doctoral.

En este epígrafe se describe de forma breve y concisa el contenido de esta tesis doctoral, la cual se desarrolla en los siguientes 7 capítulos:

En el **Capítulo 1** se incluye una introducción de la tesis doctoral, una breve descripción de los objetivos propuestos, la estructura del documento y las principales aportaciones de la investigación.

En el **Capítulo 2** se detalla el concepto de severidad o gravedad de las lesiones resultantes de un accidente. A continuación se realiza una revisión de estado del arte sobre cómo los investigadores han modelizado la gravedad de los accidentes; describiéndose las principales particularidades de los datos de accidentes, los modelos estadísticos más comúnmente empleados en el análisis de severidad; así como nuevas técnicas que se encuentran dentro del campo de la Minería de Datos. Y, finalmente, se describen las principales características de la herramienta particular utilizada en esta tesis doctoral, los Árboles de Decisión.

El **Capítulo 3** recoge los objetivos de esta investigación, distinguiéndose un objetivo principal y una serie de objetivos específicos.

En el **Capítulo 4** se presenta la metodología utilizada. Se describen los diferentes algoritmos utilizados para la construcción de ADDs, a continuación se explica cómo se validan los modelos construidos, cómo se obtiene la importancia de las variables y el proceso para la obtención de las Reglas de Decisión. Posteriormente se describe un método particular, *Information Root Node Variation*, y cómo se extrae el conjunto global de RDs utilizando este método. Y finalmente se describen los datos objeto de esta investigación.

En el **Capítulo 5** se presentan los resultados obtenidos en esta investigación, siguiendo la misma estructura detallada en la metodología, y se realiza una discusión de los mismos.

En el **Capítulo 6** se recogen las conclusiones obtenidas en el desarrollo de esta tesis doctoral.

En el **Capítulo 7** se detallan las futuras líneas de investigación.

Y por último, en el **Capítulo 8** se muestran las referencias bibliográficas consultadas para la elaboración de esta tesis doctoral.

1.4. Principales contribuciones.

A continuación, se detallan las diferentes publicaciones que se han llevado a cabo como resultado de la presente tesis doctoral. Se distingue entre aquellas que están directamente relacionadas con el trabajo de investigación, en las que se describen metodologías o resultados descritos en la tesis, y aquellas que, a pesar de no tener como objetivo la presentación de resultados conseguidos durante el desarrollo de la tesis, se encuentra en relación con el objeto de este trabajo. Particularmente son publicaciones relacionadas con otras metodologías para el análisis de la gravedad de los accidentes de tráfico.

Las publicaciones completas se adjuntan en el anexo III.

1.4.1. Publicaciones directamente relacionadas.

- **Revistas.**

De Oña, J., López, G., Abellán, J., 2013. Extracting decision rules from police accident reports through decision trees. Accident Analysis and Prevention, 50, 1151–1160.

Abellán, J., López, G., De Oña, J., 2013. Analysis of traffic accident severity using Decision Rules via Decision Trees. Expert Systems with Applications. 40, 6047-6054.

López, G., De Oña, J., Abellán, J., 2012. Using decision trees to extract decision rules from police reports on road accidents. PROCEDIA-SOCIAL AND BEHAVIORAL SCIENCES, 53, 106-114.

- **Capítulos de libros.**

G. López, J. de Oña, J. Abellán, 2012. Priorización de actuaciones sobre accidentes de tráfico mediante reglas de decisión. Cuaderno Tecnológico de la PTC, nº 8. Editorial: Plataforma Tecnológica de la Carretera.

- **Congresos.**

López Maldonado, Griselda, Garach Morcillo, Laura, de Oña López, Rocío, de Oña Juan, Calvo Poyo, Francisco Javier. Classification and Regression trees to explore contributory factors by type of accidents. MAMERN' 2013, International conference on approximation methods and numerical modelling in environment and natural resources. Granada (Spain), 22-25 April, 2013.

López Maldonado, Griselda; De Oña López, Juan, Abellán Mulero, Joaquín, 2012. Using decision trees to extract decision rules from police reports on road accidents. SIIV ROMA MMXII: SUSTAINABILITY OF ROAD INFRAESTRUCTURES. Roma (Italy), 29-31 October, 2012.

López Maldonado, Griselda; Garach Morcillo, Laura; De Oña López, Juan José. Análisis de accidentes de tráfico mediante técnicas de Minería de Datos. VI CONGRESO NACIONAL DE LA INGENIERÍA CIVIL. Valencia (España), 23-24 febrero, 2012.

1.4.2. Otras publicaciones y congresos relacionados.

- **Revistas**

De Oña, J., López, G., Mujalli, R.O., Calvo, F.J., 2013. Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks, 51, 1-10.

➢ Congresos

López Maldonado, Griselda; De Oña López, Juan José; De Oña Lopez, Rocio; Mujalli, Randa Oqab; Garach-Morcillo, Laura. Análisis de la severidad de los accidentes de tráfico mediante técnicas paramétricas y no paramétricas. X Congreso de Ingeniería del Transporte (CIT 2012). Granada (España) 20-22, Junio 2012.

Mujalli, Randa; De Oña-López, Juan José; López-Maldonado, Griselda; Garach Morcillo, Laura; Calvo Poyo, Francisco Javier. Analysis of traffic accident injury severity on two-lane highways using simplified Bayesian Networks. X Congreso de Ingeniería del Transporte (CIT 2012). Granada (España) 20-22, Junio 2012.

CAPÍTULO 2.

ESTADO DEL ARTE

CAPÍTULO 2.
ESTADO DEL ARTE

2.1. Introducción.

Los problemas relacionados con la seguridad vial han sido abordados a lo largo de los años desde diferentes enfoques, siendo los más comunes el análisis de los accidentes en términos de frecuencia o en términos de la severidad de las lesiones resultantes.

En los estudios de frecuencia de accidentes, los investigadores han tratado de analizar los factores que afectaban a la probabilidad de ocurrencia de un accidente. Sin embargo, la falta de detalle en los datos relativos al modo de conducción en el momento del accidente (como la aceleración, información relativa al frenado, la respuesta de los estímulos del conductor, etc.) y que podrían ayudar en la identificación de las causas y efectos del accidente, hacen que la mayoría de los investigadores direccionaran el encuadre en términos de la comprensión de los factores que afectan a la frecuencia de los accidentes, que ocurren en un espacio geográfico (segmento de carretera, intersección, etc.) en un período de tiempo específico (Lord and Mannering, 2010). Los modelos estadísticos utilizados por los investigadores para lograr estos objetivos se conocen como modelos de frecuencia (*crash count*).

Como indican Savolainen et al. (2011), reducir la frecuencia o la severidad de los accidentes requiere de planteamientos estratégicos diferentes. Por ejemplo, medidas encaminadas a reducir el radio de curvatura de una determinada irían principalmente encaminadas a reducir la frecuencia de los accidentes, mientras que la colocación de una barrera metálica doble sería una medida fundamentalmente encaminada a reducir la severidad del accidente.

Por otro lado, en el estudio de frecuencias los datos de análisis no suelen ser específicos del accidente (geometría de la carretera, volumen de tráfico, etc.) mientras que los estudios de severidad se realizan utilizando, entre otros, datos posteriores al accidente (número de vehículos involucrados, edad de los ocupantes, condiciones meteorológicas, etc.).

Dado que el objetivo de muchos estudios previos ha sido la compresión e identificación de los principales factores que afectan a la severidad de los accidentes (Donelson et al., 1999; Krull et al., 2000; Abdelwahab and Abdel-Aty, 2001; Abdel-Aty, 2003; Dissanayake, 2004; Wang and Kockelman, 2005; Chang and Wang, 2006; Savolainen y Mannering, 2007; Depaire et al., 2008; Garder, 2009; Paleti et al., 2010; De Oña et al., 2011; Chang and Chien, 2013), e incluso algunos investigadores (Chang and Wang, 2006) han afirmado que reducir la severidad de los mismos es una de las

medidas más efectivas para conseguir mejoras en la seguridad vial; en esta tesis doctoral se llevará a cabo un análisis de los accidentes siguiendo este enfoque.

En este capítulo se detalla el concepto de severidad o gravedad de las lesiones. A continuación se realiza un estudio sobre cómo los investigadores han modelizado la severidad de los accidentes; describiéndose las particularidades de los datos de accidentes que son necesarias tener en cuenta para su modelización; los modelos estadísticos más comúnmente empleados en el análisis de severidad (*crash severity models*); así como nuevas técnicas que se encuentran dentro del campo de la Minería de Datos. Finalmente, se describen las principales características de la herramienta particular utilizada en esta tesis doctoral, los Árboles de Decisión.

Para llevar a cabo la revisión de los modelos, sólo se han considerado aquellos estudios que se centran en el análisis de la gravedad de accidentes en los que hay al menos un automóvil involucrado. Dada la gran cantidad de estudios que existen en la literatura, quedan fuera del ámbito de esta revisión los estudios que únicamente se centran en analizar la gravedad de un determinado tipo de usuario vulnerable (peatones, bicicletas y/o motocicletas), que requerirían de una revisión específica.

2.2. La severidad o gravedad de las lesiones.

La gravedad de una lesión puede representase a través de una serie de categorías discretas. Existen diferentes escalas que permiten clasificarla, y generalmente se utiliza la escala KABCO (Savolainen et al., 2011). Esta escala está basada en el uso de 5 categorías de gravedad: muerte o con lesiones mortales (*fatal injury or killed, K*), lesiones incapacitantes (*incapacitating injury, A*), lesiones no incapacitantes (*non-incapacitating, B*), lesiones leves (*possible injury, C*) o solo daños materiales (*property damage only, O*).

Existen otras escalas que combinan la información de variables discretas y continuas, ya que recogen la información sobre la localización de las lesiones y/o sobre la extensión de las mismas. Este es el caso de la escala AIS (*Abbreviated Injury Scale*) que fue desarrollada por la Asociación Americana de medicina automotriz (*American Association for Automotive Medicine*) (American Medical Association, 1971). Esta escala utiliza un código de 7 dígitos para clasificar la herida, siendo el último dígito el que hace referencia a la gravedad. Para medir la gravedad de la herida utiliza la *Maximum Abbreviated Injury Scale* (MAIS) que consta de 6 niveles de lesiones (de 1-lesiones leves a 6-lesiones mortales).

Otras escalas que también se utilizan son la OIS (*Organ Injury Scales*), propuesta por la Asociación Americana para la cirugía de trauma (*American Association for the Surgery of Trauma*) y cuya codificación consta de 5 categorías: de 1 (lesiones leves) a 5 (lesiones graves) (Moore et al., 1992); y la escala ISS (*Injury Severity Score*), utilizada en hospitales, y que combina la escala AIS y la localización de la lesión según 6 regiones corporales (Baker et al., 1974).

Se puede decir que no existe una definición universalmente aceptada para definir la gravedad de la lesión, sino que los criterios para clasificarla varían de un país a otro, pudiendo incluir entre ellos, el tiempo de permanencia en el hospital, el tipo de lesión y el nivel de gravedad (Montella et al., 2012a).

Las lesiones mortales son definidas como aquellas que producen la muerte de la persona en el momento del accidente o dentro un número determinado de días posteriores al mismo. La OMS recomienda considerar un período de 30 días[1] (OMS, 2013), y en los países del grupo IRTAD (International Road Traffic Accident Database) ya se utiliza esta definición (Austroads, 2009; European Commission, 2004; IRTAD, 2012; NHTSA, 2008). Sin embargo, sobre las lesiones no mortales existen diferentes definiciones. Según IRTAD (2012), en la mayoría de países de la UE un herido grave es una persona que precisa una hospitalización superior a veinticuatro horas. En Irlanda, Rumania y Reino Unido un herido grave es una persona hospitalizada más de veinticuatro horas o que sufren alguna de las siguientes lesiones: fracturas, contusiones, lesiones internas, aplastamientos, quemaduras (excluyendo quemaduras por fricción), cortes y laceraciones graves, shock severo que requiere tratamiento médico y las lesiones que causan la muerte 30 o más días después del accidente. En Polonia y en Suecia (IRTAD, 2012b) la gravedad de las lesiones se clasifican en función de las lesiones registradas (además, en Suecia un herido es grave si las lesiones conducen a la hospitalización). En los Países Bajos un herido grave es una persona con lesiones evaluadas en el nivel 2 o superior según MAIS (su codificación sería: MAIS2+).

En España se utilizan 2 niveles de gravedad de lesiones, siendo las definiciones (según la OM de 18 de febrero de 1993) las siguientes:

- Herido grave: toda persona herida en una accidente de circulación y cuyo estado precisa una hospitalización superior a veinticuatro horas.
- Herido leve: toda persona herida en una accidente de circulación a la que no puede aplicarse la definición de herido grave.

Actualmente se busca el uso de una clasificación internacional y con este objetivo, el grupo IRTAD propone clasificar la severidad de las lesiones basándose en MAIS. Siendo un herido grave aquella persona con lesiones evaluadas en el nivel 3 o más en la escala MAIS (su codificación sería: MAIS3+). Sin embargo el uso de la escala MAIS es todavía muy limitado y por ello CADaS (*Common Accident Data Set*) propone una simple definición basándose en las horas de hospitalización, de modo que herido grave sería toda aquella persona que requiere una hospitalización superior a 24 horas (De Meester, 2011).

[1] La elección de 30 días está basada en investigaciones que muestran que la mayoría de las personas que mueren como consecuencia de las heridas de un accidente de tráfico mueren dentro de este período. Un período mayor, proporcionaría un aumento marginal del número de muertes pero requeriría un aumento desproporcionado en el esfuerzo de vigilancia.

El principal problema en la definición de herido grave, es que el hecho de estar hospitalizado 24 horas no aporta información sobre la verdadera gravedad del accidentado, ni sobre las lesiones que ha sufrido, de forma que sería a través de los diagnósticos médicos como se deberían definir diferentes categorías de heridos.

En el Programa de seguridad vial 2011-2020 de la Comisión Europea (EC, 2010), se pretende alcanzar una definición común de las lesiones graves y leves para fijar objetivos con vistas al establecimiento de un objetivo común de la UE que formará parte de las orientaciones de seguridad vial 2011-2020. De este modo todos los países de la UE tendrían que informar de acuerdo a esas definiciones comunes sobre las lesiones a nivel nacional. Este tipo de acciones posibilita que a nivel nacional se puedan impulsar las tareas necesarias para conocer la información que es requerida a nivel europeo, dado que para ello es necesario poner de acuerdo a diferentes estamentos de la Administración (en este caso el sector salud y el de seguridad), regular los procesos para la puesta en común de la información y realizar las oportunas inversiones para llevarlo a cabo.

2.3. Modelización de la severidad de los accidentes.

Los accidentes de tráfico son sucesos eventuales y su análisis requiere el conocimiento de las particularidades que los definen. En general, los accidentes vienen definidos a través de una serie de variables explicativas que matemáticamente se definen como variables discretas de valor entero (no negativo). Conocer su naturaleza y las particularidades que los definen es fundamental para el desarrollo y la aplicación de un modelo apropiado.

2.3.1. Particularidades de los datos de accidentes.

Los datos de accidentes de tráfico generalmente son recogidos por los oficiales de los cuerpos de seguridad a través de los partes de accidente (o cuestionarios estadísticos). Estos partes garantizan una relativa homogeneidad de la información recogida por los diferentes agentes, permitiendo así una explotación de los datos obtenidos.

Sin embargo, a la hora de analizar los datos de accidentes de tráfico, y al elegir el modelo adecuado para ello, es necesario tener en cuenta una serie de particularidades que los caracterizan (una mayor extensión de las mismas puede consultarse en Savolainen et al., 2011):

- La infra-detección de los accidentes.

Aunque legalmente se establece la obligatoriedad de rellenar e informar de cualquier accidente que suceda, la realidad es que en el caso de los accidentes más leves (solo daños materiales o con heridos leves) los agentes de tráfico no siempre tienen conocimiento de su existencia y, por tanto, no quedan registrados, produciéndose una falta de notificación de esos accidentes. El nivel de infra-detección es mayor cuanto

menor es la gravedad del accidente. También hay determinados tipos de accidente en los que se produce mayor infra-detección, como es el caso de los accidentes con un solo vehículo.

Un informe técnico realizado por la *National Highway Traffic Safety Administration* (2009) estima que el 25% de los accidentes leves y la mitad de accidentes sin lesiones no son registrados; un agudo contraste con los accidentes mortales, donde la tasa de notificación es casi del 100 % (Blincoe et al., 2002).

El problema del infra-registro puede producir sesgos significativos cuando se trata de predecir la probabilidad de la gravedad del accidente (Ye and Lord, 2011). Puesto que los modelos estadísticos se desarrollan bajo el supuesto de que los datos de la muestra se extraen aleatoriamente de la población, y cada accidente tiene igual probabilidad de ser seleccionado. Ben-Akiva and Lerman (1985) indican que hay que considerar que la proporción de la gravedad de lesiones en la muestra registrada por la policía, no es la misma que en la totalidad de los accidentes que se producen.

- Naturaleza ordinal de la gravedad de las lesiones.

El nivel de gravedad de las lesiones tiene naturaleza de carácter ordinal. Dado que las categorías de gravedad están ordenadas y en ocasiones estrechamente relacionadas (por ejemplo, ninguna lesión y posible lesión), pueden ser compartidos los efectos no observados entre las categorías adyacentes de lesiones. Esto puede ser problemático en la estimación de algunas tipologías de modelos, que al no tener en cuenta tal correlación pueden dar lugar a estimaciones sesgadas de los parámetros e inferencias incorrectas (Savolainen and Mannering, 2007; Paleti et al., 2010).

- Parámetros fijos.

La mayoría de los modelos constan de parámetros fijos que restringen el efecto de las variables explicativas a través de las observaciones individuales de la variable dependiente (la gravedad de las lesiones). Sin embargo, dada la heterogeneidad no observada, el efecto de determinadas variables puede variar en los distintos niveles de gravedad y, por tanto, estos parámetros pueden variar a través de observaciones de la gravedad. De modo que, si tales efectos no se tienen en cuenta, pueden resultar un sesgo potencial y dar lugar a inferencias estadísticas erróneas (McFadden and Train, 2000; Train, 2009).

- Omisión de variables relevantes.

Usualmente los datos utilizados para estudiar la severidad de los accidentes son limitados, debido principalmente a la información disponible en los partes de accidente. La omisión de variables explicativas que sean relevantes puede provocar inconsistencia en los parámetros estimados, si tales variables están correlacionadas con otras incluidas en el modelo, o si la variable omitida está correlacionada o tiene diferente varianza entre los niveles de severidad (Washington et al., 2011).

- Bajo tamaño muestral.

El volumen de datos disponibles es a menudo un factor clave en la selección del modelo más adecuado. Cuando el tamaño de la muestra de estudio es pequeño, se prefieren los modelos más simples (tales como aquellos que asumen parámetros fijos), ya que requieren menos datos para llegar a resultados razonables de estimación.

- Heterogeneidad.

La heterogeneidad no observada es otro de los problemas presentes en los datos de accidentes. Si no es tenida en cuenta, cuando se analizan los accidentes, algunas relaciones pueden quedar ocultas (De Oña et al., 2013). Por ejemplo, la heterogeneidad no observada puede existir entre la población de conductores; un accidente puede producirse por las diferencias en las conductas de riesgo entre los conductores, por factores fisiológicos, etc. Algunos autores (Depaire et al., 2008; De Oña et al., 2013) proponen realizar antes de su modelización una segmentación previa de los accidentes para obtener grupos más homogéneos.

- Endogeneidad.

Cuando las variables explicativas están influenciadas por la gravedad de la lesión resultante pueden surgir problemas en la estimación del modelo. Por ejemplo, la variable airbarg tiene influencia en la gravedad de la lesión resultante y además podría hacer que los conductores tengan una percepción diferente del riesgo (Winston et al. 2006). Una definición más general es que un modelo presenta problemas de endogeneidad cuando existe correlación entre las variables explicativas y el término de error (el cual trata de recoger en el modelo la variabilidad existente en un conjunto de características observadas, a través de las variables explicativas).

- Correlación dentro de un mismo accidente.

Muchos estudios identifican el nivel de severidad del accidente con el nivel de gravedad de la persona que resulta más gravemente herida. Sin embargo, cuando se trata de crear un modelo en el que hay más de un vehículo involucrado, o que considere el nivel de gravedad de todos los individuos implicados en el accidente, se debe tener en cuenta la probabilidad de correlación entre las lesiones. Esto se debe a que los elementos no observados en relación con un accidente específico (por ejemplo, las características del impacto, etc.) estarán correlacionados con las lesiones observadas en el accidente.

Por ejemplo, en accidentes múltiples las velocidades de los vehículos implicados (que probablemente no son conocidas) son una posible fuente de correlación con las de lesiones observadas en el accidente. Considerar esta correlación se puede conseguir teniendo en cuenta el conjunto de niveles de gravedad, o mediante el desarrollo de estructuras de modelos más complejos (ver Eluru et al., 2010).

Los modelos estadísticos que no tienen en cuenta la correlación entre las lesiones resultantes que se producen en el mismo accidente probablemente dan lugar a estimaciones sesgadas de los parámetros (Helai et al., 2008).

- Correlaciones espaciales y temporales.

Se producen sobre accidentes que ocurren en lugares próximos (por ejemplo en el mismo segmento de una intersección) o que suceden próximos en el tiempo (mismo día o semana). Estos accidentes tienen probabilidad de compartir efectos no observados, factores que no son tenidos en cuenta con las variables explicativas medibles. De modo que puede ser significativamente complicada la estimación de la estructura del modelo, haciendo las estimaciones de los parámetros más difíciles. Si tales correlaciones no son tenidas en cuenta, habrá una pérdida de eficiencia y los parámetros se estimarán con menor precisión, lo que hace más difícil obtener inferencias estadísticamente defendibles (ver Anselin et al., 2005).

2.3.2. Técnicas de modelización para analizar la severidad de los accidentes.

El desarrollo de un modelo de severidad o gravedad (*crash severity model*) se basa en una condición de partida, que es la ocurrencia del accidente. Por lo tanto, los modelos de gravedad no estiman la probabilidad de ocurrencia de un accidente sino el nivel de gravedad esperado de los mismos.

Las principales técnicas de modelización empleadas para el estudio de la severidad de los accidentes de tráfico han sido recogidas en Savolainen et al. (2011) y Mujalli and De Oña (2012).

A lo largo de los años, han sido utilizadas una gran variedad de técnicas estadísticas, siendo los Discrete Outcome Models (DOM) la técnica analítica prevaleciente. Sin embargo, recientemente han comenzado a utilizarse en este campo técnicas de Minería de Datos (MD), estas técnicas serán explicadas en el apartado 2.4 de este capítulo.

2.3.2.1. Discrete Outcome Models.

En el ámbito del transporte los Discrete Outcome Models (DOM) han sido utilizados para modelar datos discretos que envuelven un comportamiento de elección (por ejemplo la elección de un modo de transporte), y para modelar datos discretos que no hacen referencia a un comportamiento de elección sino a un evento físico (como la gravedad del accidente). El enfoque metodológico para modelar ambos conceptos es estadísticamente igual, sin embargo la teoría que los sustenta es bastante diferente. Los primeros se basan en la teoría económica (concretamente en la maximización de la teoría de la utilidad) y los segundos derivan exclusivamente de la teoría de la probabilidad (Washington et al., 2003).

Teniendo en cuenta los elementos que influyen en el proceso de especificación de los DOM, se puede establecer la clasificación general recogida en la Tabla 1.

<table>
<tr><th colspan="2">ALTERNATIVAS</th><th rowspan="2">FUNCIÓN</th><th colspan="2">REGRESOR</th></tr>
<tr><th>Número</th><th>Tipo</th><th>Características[a]</th><th>Atributos[b]</th></tr>
<tr><td rowspan="3">Modelos de respuesta dicotómica
(2 alternativas)</td><td rowspan="3">Complementarias</td><td>Lineal</td><td colspan="2">Modelo de Probabilidad Lineal Truncado</td></tr>
<tr><td>Logísitica</td><td colspan="2">Modelo Logit</td></tr>
<tr><td>Normal tipificada</td><td colspan="2">Modelo Probit</td></tr>
<tr><td rowspan="4">Modelos de respuesta múltiple
(más de 2 alternativas)</td><td rowspan="2">No ordenadas</td><td>Logísitica</td><td>Logit Multinomial
- Logit Anidado
- Logit Mixto</td><td>Logit Condicional
- Logit Anidado
- Logit Mixto</td></tr>
<tr><td>Normal tipificada</td><td>Probit Multinomial
Probit Multivariante</td><td>Probit Condicional
Probit Multivariante</td></tr>
<tr><td rowspan="2">Ordenadas</td><td>Logísitica</td><td colspan="2">Logit Ordenado</td></tr>
<tr><td>Normal tipificada</td><td colspan="2">Probit Ordenado</td></tr>
</table>

[a] cuando el regresor hace referencia a variables que contienen aspectos específicos del individuo;
[b] cuando el regresor contiene aspectos específicos de las alternativas.

Tabla 1.- Clasificación de los Discrete Outcome Models.

En función del número de categorías de la variable dependiente (o variable endógena) se establecen los dos grupos principales, modelos de respuesta dicotómica y modelos de respuesta múltiple.

2.3.2.1.1. Modelos de respuesta dicotómica.

Son utilizados cuando la variable dependiente presenta 2 categorías (accidente con heridos vs. accidente sin heridos; accidente mortal vs. accidente no mortal). En el análisis de la severidad los modelos de respuesta discreta dicotómica más utilizados han sido los modelos logit (ML) y probit (MP).

Los ML y MP surgen para superar los problemas que presentan los modelos de regresión lineal, que sólo se pueden modelar variables dependientes puramente cuantitativas. Sin embargo, en gran cantidad de estudios la variable dependiente es cualitativa o categórica (como sucede con la severidad del accidente), y para su estudio se pueden utilizar las técnicas de regresión logística, basadas en el Modelo Lineal de Probabilidad (MLP), pero adaptadas a variables categóricas.

El MLP es una variación del modelo de regresión lineal enfocada a la explicación de una variable dependiente cualitativa. Sea Y_i una variable dependiente dicotómica, la expresión del MLP es la siguiente:

$$Y_i = \beta_k x_{ki} + \varepsilon_i \,, \qquad (1)$$

notando con Y_i=1, la ocurrencia de un suceso; Y_i=0, la no ocurrencia del suceso; x_{ki}, un conjunto de variables explicativas; ε_i, una variable aleatoria que se distribuye según una normal $N(0,\sigma^2)$.

La distribución de la muestra en este tipo de modelos se caracteriza por configurar una nube de puntos de tal manera que las observaciones muestrales se dividen en dos subgrupos (ver Figura 1); uno está formado por las observaciones en las que ocurrió el suceso (Y_i=1), y el otro, por los que no ocurrió (Yi=0).

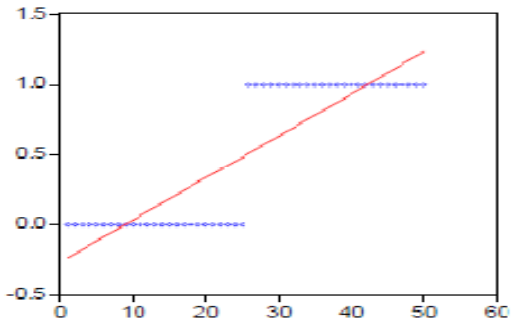


Figura 1.- Inconsistencia en la estimación de los Modelo Lineal de Probabilidad.

La estimación del MLP requiere el ajuste de esa nube de puntos a una función lineal (recta de regresión) capaz de explicar de la mejor manera el comportamiento de la muestra. Sin embargo, la estimación del MLP (por el método de mínimos cuadrados) plantea una serie de limitaciones:

- Inconsistencia en las probabilidades predichas (ya que no se puede garantizar que estén acotadas entre 0 y 1).
- No normalidad de la perturbación aleatoria (al tomar únicamente dos valores, la hipótesis de normalidad del término de perturbación no es aceptable).
- Heterocedasticidad de la perturbación (no se cumple la hipótesis de varianza constante y la perturbación no es homocedástica).
- El coeficiente de determinación no es apropiado (R^2 está subestimado).

Dadas estas limitaciones, son necesarios otros modelos alternativos que permitan estimaciones más fiables de las variables dicotómicas. Para evitar que la variable endógena estimada pueda encontrarse fuera del rango (0, 1), se pueden utilizar modelos de probabilidad no lineales, donde la función de especificación utilizada garantice un resultado en la estimación en el rango 0-1. Las funciones de distribución cumplen este requisito, ya que son funciones continuas que toman valores comprendidos en este rango. Siendo las funciones de distribución más utilizadas, la logística, que ha dado lugar al ML, y la normal tipificada, que ha dado lugar al MP.

Tanto los ML como los MP relacionan por tanto, la variable endógena Y_i con las variables explicativas X_{ki} a través de una función de distribución. En el ML, se utiliza la función logística, por lo que la especificación del modelo sería la siguiente:

$$Y_i = \frac{1}{1+e^{-\alpha-\beta_k X_{ki}}} + \varepsilon_i = \frac{e^{\alpha+\beta_k X_{ki}}}{1+e^{\alpha+\beta_k X_{ki}}} + \varepsilon_i \quad (2)$$

En el MP se utiliza la función normal tipificada, por lo que la especificación del modelo sería la siguiente:

$$Y_i = \int_{-\infty}^{\alpha+\beta X_i} \frac{1}{(2\pi)^{1/2}} e^{-\frac{s^2}{2}} ds + \varepsilon_i \quad (3)$$

Siendo s una variable "muda" de integración con media cero y varianza uno.

Estos modelos pueden ser interpretados en términos probabilísticos, de modo que sirven para medir la probabilidad de que ocurra el suceso de estudio dado (Y_i=1). En cuanto a la interpretación de los parámetros estimados en un ML y un MP, el signo indica la dirección en que se mueve la probabilidad cuando aumenta la variable explicativa correspondiente.

- **Aplicaciones de los modelos logit y probit en el estudio de la gravedad del accidente.**

Los ML han sido ampliamente utilizados por los investigadores en el análisis de severidad de los accidentes con diferentes objetivos. Algunos estudios se han centrado en estudiar la severidad producida por un tipo particular de accidente. Donelson et al. (1999) analizaron mediante ML multivariantes los accidentes producidos por vuelco de camiones ligeros (en los que solo había un vehículo involucrado). Siendo el principal objetivo de su estudio predecir el riesgo de muerte de los ocupantes de estos vehículos. Krull et al. (2000) estudiaron también los accidentes producidos por vuelco (con un solo vehículo involucrado) y su efecto en la gravedad del conductor. Y además estudiaron la gravedad de los mismos en función de la secuencia de eventos que habían sucedido posteriormente al vuelco. Para ello utilizaron ML considerando 2 categorías de severidad: accidentes mortales o severos y accidentes no severos. La contribución de cada variable fue examinada mediante el indicador del efecto marginal, que es el cambio en la probabilidad de tener un accidente severo para cada categoría en relación con la categoría base. La principal limitación del estudio se producía por la calidad de los datos registrados por los agentes de policía (la cantidad de datos perdidos, variables mal codificadas, etc.) y los propios supuestos inherentes al modelo de regresión logística utilizado. Bedard et al. (2002) analizaron las colisiones con objetos fijos. Utilizaron ML para identificar la contribución independiente de las variables relativas al conductor, al accidente y al vehículo en el riesgo de muerte del conductor. Sus resultados muestran la importancia de la edad y el sexo en la severidad del conductor, y concluyen que deben realizarse estudios específicos para conductores mayores y mujeres.

Otros estudios que han utilizado ML se han centrado en estudiar un grupo particular de conductores. Zhang et al. (2000) analizaron los factores que afectan a la gravedad de los accidentes con conductores mayores de 65 años. Consideraron sólo accidentes con 1 o 2 vehículos involucrados producidos por coches, camiones o camiones ligeros. Estimaron 3 ML cambiando el nivel de gravedad en cada uno de ellos: accidentes mortales vs. accidentes muy leves; accidentes graves vs. accidentes muy leves;

accidentes leves vs. accidentes muy leves. Concluyeron que el tratamiento médico, las condiciones físicas, las distracciones, y ciertas condiciones del entrono incrementan el riesgo de muerte de los conductores mayores. Dissanayake (2004) realizó una comparación de los factores que afectan a la gravedad de los conductores jóvenes vs. conductores mayores en accidentes con un solo vehículo involucrado. La metodología utilizada proporcionó un razonable logro de los objetivos perseguidos en su estudio, a pesar de los inconvenientes propios de los datos (exactitud y fiabilidad) y las relaciones internas que pudieran existir entre las variable independientes consideradas. Peek-Asa et al. (2010) también se centraron en el estudio de los factores que influyen en la gravedad de los accidentes con conductores jóvenes. Utilizaron estos modelos para identificar las características propias del conductor (edad, sexo, uso del cinturón, alcohol y distracción) y del accidente que afectan a la gravedad de los accidentes producidos por conductores jóvenes. Realizaron el análisis tanto en zonas urbanas como en zonas no urbanas. Sus resultados demostraron que los riesgos son mayores en zona no urbana, y que los factores que afectan a la gravedad son diferentes según al ámbito analizado.

Al-Ghamdi (2002) utilizó estos modelos para explorar la contribución de ciertas variables en la estimación de la gravedad de las víctimas de accidentes. Estudió accidentes ocurridos en área urbana. Entre sus resultados destacan que los accidentes en intersecciones son menos graves que los ocurridos en otras secciones. Además destacó que los ML son una herramienta que proporciona significativas interpretaciones de cara a ser usadas en futuras mejoras de seguridad. Posteriormente Toy and Hammit (2003) utilizaron los ML para evaluar el riesgo de muerte o de herido grave del conductor. El objetivo era comparar la resistencia al choque y la peligrosidad (definida como el riesgo hacia otros) de diferentes tipos de vehículos (automóviles, vehículos deportivos, furgonetas y camionetas) en colisiones con 2 vehículos involucrados. Para ello desarrollaron diferentes ML; en el primer modelo solo evaluaron el riesgo del conductor respecto al tipo de vehículo y en el resto de modelos introdujeron de forma secuencial las características referentes al conductor y al accidente. Entre sus resultados destacaron que los conductores de camionetas y furgonetas tienen menos riesgo de resultar seriamente heridos en un accidente que los conductores de coches. Un estudio similar, utilizando también ML, fue realizado por Fredette et al. (2008). Recientemente Kononen (2011) utilizó estos modelos para predecir la probabilidad de que un vehículo implicado en un accidente tuviese uno o más ocupantes con heridas graves. Las variables dependientes fueron la dirección del choque (delante, izquierda, derecha y posterior), el cambio en la velocidad, el uso del cinturón, la presencia de al menos un ocupante de más edad ($\geq$ 55 años), la presencia de al menos una mujer en el vehículo y el tipo de vehículo (coche, camión, camioneta y vehículos deportivos utilitarios).

Los ML se caracterizan por su simplicidad en el proceso de estimación. Sin embargo imponen una serie de restricciones (McFadden, 1973; Train, 2009):

- Los coeficientes de las variables son los mismos para toda la población. Esto implicara que diferentes personas con las mismas características atribuyen el

mismo valor a cada una de las variables que entran en el modelo. Sin embargo, pueden existir variaciones aleatorias en los gustos (por ejemplo, en el ámbito del transporte, dos individuos con igual nivel socioeconómico podrían realizar una elección distinta en el modo de transporte, ya que esta elección depende de sus gustos).

- Cumplen la propiedad de independencia de alternativas irrelevantes (por la que se supone que cuando existan dos alternativas con probabilidad no nula de ser elegidas, el cociente de una sobre la otra (odds) no se ve afectado por la presencia o ausencia entre todas las alternativas posibles). Y debido a esta característica, los modelos predicen que una modificación en los atributos de una alternativa modifica las probabilidades de elección de las otras alternativas proporcionalmente, y este patrón de sustitución puede resultar irrealista en muchas ocasiones.

- Consideran que los factores no observados que influyen en la elección son independientes en el tiempo para cada individuo. Por lo que no pueden usarse con datos longitudinales (paneles).

Los MP resuelven las 3 limitaciones que presentan los ML (Train, 2009). Sin embargo, son menos utilizados debido a la complejidad de su proceso de estimación (Daganzo, 1979). Y particularmente, para el estudio de la severidad del accidente también se encuentran menos aplicaciones en la literatura.

Algunos investigadores han utilizado extensiones de los ML y MP para resolver algunos de los inconvenientes que se plantean cuando se trabajan con los datos de accidentes. Por ejemplo, para considerar los diferentes niveles de gravedad resultantes de un mismo accidente (como sucede en las colisiones múltiples), se deben tener en cuenta las correlaciones dentro de un accidente, y los ML y MP pueden producir errores en los parámetros estimados; ya que son modelos que se expresan mediante una sola ecuación y, por las propias especificaciones del modelo, sólo pueden modelar un resultado de gravedad al mismo tiempo (Ouyang et al., 2002). Para tener en cuenta estas correlaciones, Ouyang et al. (2002) utilizaron un modelo logit binario simultáneo que consiste en un sistema de ecuaciones interconectadas que permite modelar la severidad en colisiones múltiples. Helai et al. (2008) utilizaron un modelo logit jerárquico bayesiano para tener en cuenta las correlaciones que existen entre los conductores y vehículos envueltos en un mismo accidente. Los modelos jerárquicos, son modelos de respuesta múltiple (explicados en el siguiente epígrafe) y el enfoque bayesiano hace referencia al modo de estimar el modelo. Hay diferentes métodos para estimar los modelos jerárquicos (Goldstein, 2003), tales como la estimación mediante máxima-verosimilitud, o la estimación mediante inferencia Bayesiana (enfoque Bayesiano).

Para tener en cuenta la presencia de variables explicativas endógenas se pueden utilizar modelos bivariantes (Winston et al., 2006; Lee and Abdel-Aty 2008). Winston et al. (2006) consideraron que la decisión de un conductor de tener un vehículo con airbag y sistema de frenos antibloqueo (ABS) puede estar relacionada con la

probabilidad de verse envuelto en un accidente y con la severidad resultante. Para resolver el problema de variables endógenas, utilizaron un modelo probit bivariante que permitía modelar simultáneamente 4 variables binarias: la elección del airbag, y ABS, la probabilidad de estar involucrado en un accidente (con las variables dependientes de los dos primeros modelos, airbag y ABS, incluidas ahora como variables explicativas), y la probabilidad de la gravedad resultante en una lesión (incluyendo de nuevo las variables como explicativas).

2.3.2.1.2. Modelos de respuesta múltiple.

Estos modelos se utilizan cuando la variable endógena a modelar es una variable discreta con varias alternativas posibles de respuesta o categorías. Se clasifican en dos grandes grupos según si las alternativas de la variable endógena se pueden ordenar (modelos ordenados) o no se pueden ordenar (modelos no ordenados).

A. Modelos no ordenados.

La especificación general de los modelos de respuesta múltiple con datos no ordenados queda recogida a través de la siguiente expresión:

$$Prob(Y_i = j) = \frac{e^{\beta_i X_{in}}}{\sum_i e^{\beta_i X_{in}}} \quad (4)$$

donde X_{in} representa la matriz de los regresores del modelo.

A partir de esta especificación general, y teniendo en cuenta que las variables explicativas hagan referencia a características o atributos (ver Tabla 1), se tiene un modelo logit multinomial en el primer caso, o un modelo logit condicional en el segundo. Dado que las variables explicativas que se utilizan para modelar la gravedad del accidente hacen referencia a características, se utilizaran modelos logit multinomial (MLM).

- **Aplicaciones del modelo logit multinomial en el estudio de la gravedad del accidente.**

La principal ventaja de estos modelos es que son bastante flexibles y, por ello, en el análisis de la severidad han sido ampliamente utilizados.

Abdel-Aty (2003) utilizó MLM, entre otros modelos (probit ordenados y logit anidados), para analizar los factores que afectan a la gravedad de las víctimas de los accidentes ocurridos en segmentos de carretera, en intersecciones señalizadas y en las cercanías de un peaje. Estimó modelos de gravedad para cada una de las ubicaciones y encontró factores que afectan en cada uno de los modelos, independientemente de la ubicación, tales como la edad y el género del conductor, el uso del cinturón, el tipo de impacto, la velocidad y el tipo de vehículo; y factores específicos para cada ubicación, como el alcohol, la iluminación y la presencia de curvas en los segmentos. Para las cercanías de peaje, el autor señala que si el vehículo disponía de un telepeaje (*teletac*) es más probable que el conductor sufra una lesión. Encontró que los MLM daban peores resultados de ajuste que los modelos probit ordenados, y que los modelos logit

anidados eran los que proporcionaban los mejores resultados. Sin embargo, encontró dificultades en su especificación, por lo que recomendó el uso de los modelos probit ordenados.

Posteriormente otros estudios utilizaron estos modelos para analizar la severidad de los conductores. Ulfarsson and Mannering (2004) estudiaron las diferencias en la severidad de los conductores según el género, considerando diferentes tipologías de vehículos (automóviles, furgonetas, vehículos deportivos y camiones ligeros). Analizaron accidentes con 1 y 2 vehículos involucrados y 4 niveles de gravedad de lesiones. Los resultados de los modelos mostraron que existen importantes diferencias en el resultado de la severidad entre los conductores hombres y mujeres; diferencias que pueden ser atribuidas a comportamientos diferenciados en la conducción. Khorashadi et al. (2005) analizaron la severidad de los conductores envueltos en accidentes con camiones pesados tanto en zonas urbanas como en zonas no urbanas. Para su estudio consideraron 4 niveles de gravedad de lesiones: no herido, solo dolor, herido leve y herido severo o mortal. Sus resultados muestran diferencias significativas en la severidad según el ámbito analizado, e indican que estas diferencias pueden, en parte, ser debidas a las diferencias cognitivas, de percepción y respuesta propias de cada ámbito (urbano vs. no urbano). Islam and Manering (2006) también analizaron las diferencias en la severidad de los conductores según el género y diferentes grupos de edad, en accidentes con un solo vehículo y con pasajeros. Construyeron MLM separados para conductores hombres y mujeres, y considerando 4 grupos de edad diferentes. Consideraron 3 niveles de gravedad: accidente sin heridos, accidente leve y accidente grave. Los resultados de su estudio mostraban diferencias estadísticamente significativas en la severidad del accidente según el género del conductor para los diferentes grupos de edad analizados. Savolainen and Ghosh (2008) analizaron la severidad de los conductores envueltos en colisiones con animales (ciervos) con el objetivo de identificar medidas para mitigar esta tipología de accidentes.

Aunque los MLM presentan ventajas en términos de la flexibilidad del modelo, también presentan ciertas limitaciones debido al supuesto de la independencia de las alternativas irrelevantes, originada de la hipótesis de independencia e igual distribución del término de error en cada función de severidad (Xie et al., 2012).

Para superar las limitaciones de los MLM los investigadores han utilizado diferentes extensiones de estos modelos tales como el modelo logit anidado (también llamado modelo logit jerárquico) o el modelo logit mixto (o modelo de parámetros aleatorios). El modelo logit anidado es apropiado cuando el conjunto de alternativas (diferentes gravedades asociadas al conjunto conductor-vehículo) puede descomponerse en subconjuntos jerárquicos denominados nidos (por ello también se conocen como logit jerárquicos). La ventaja es que permite relajar el supuesto de la independencia de las alternativas irrelevantes. En la literatura pueden encontrarse numerosas aplicaciones en el análisis de la gravedad del accidente (Chang and Mannering, 1998; Chang and Mannering, 1999; Lee and Mannering, 2002; Holdridge et al., 2005, Haleem and Abdel-Aty, 2010). El modelo logit mixto es una alternativa intermedia entre el modelo logit y

probit que permite modelizar las preferencias individuales e incluir variables subjetivas; considerando varios términos de error, permite captar el efecto de la heterogeneidad y la correlación de los factores no observados (Train, 2009). Estos modelos también han sido ampliamente utilizados en el estudio de la severidad de los accidentes (ver Savolainen et al., 2011).

B. Modelos con datos ordenados.

Se utilizan cuando las categorías de la variable endógena representan un orden entre ellas. La gravedad de las lesiones de las víctimas de accidentes de tráfico se puede representar a través de una variable categórica ordenada, donde el orden entre las categorías refleja los diferentes niveles de severidad considerados.

Los modelos ordenados están basados en la existencia de una variable continua latente y^* usada como base para modelar el ranking de los datos. Esta variable no observada puede ser especificada como una función lineal dada por:

$$y^* = \boldsymbol{\beta X} + \varepsilon \tag{5}$$

Donde y^* mide la gravedad de la lesión; $\boldsymbol{\beta}$ es el vector de parámetros desconocidos; $\mathbf{X}$ es el vector de los valores observados para cada individuo; ε es el término de perturbación aleatorio (con valor esperado nulo y varianza constante entre los individuos, es homocedástica).

La variable latente y^* no es directamente observable. Sin embargo la variable observable y se relaciona mediante respuestas observadas discretas con la variable latente del siguiente modo:

$$\begin{cases} y=1 & \text{si } y^* \leq \mu_0 \\ y=2 & \text{si } \mu_0 < y^* \leq \mu_1 \\ y=3 & \text{si } \mu_1 < y^* \leq \mu_2 \\ \ldots & \\ y=i & \text{si } y^* \geq \mu_{I-2}, \end{cases} \tag{6}$$

siendo μ, parámetros estimables que representan los valores de los umbrales que definen y, que se corresponde con el orden interno; e I es la alternativa o categoría de mayor valor. Para que todas las probabilidades sean estos umbrales deben cumplir que $0<\mu_1<\ldots<\mu_{I-1}$. Adoptando este enfoque, el modelo probabilístico ordenado se define del siguiente modo:

$$\begin{gathered} P(y=1)=F(-\boldsymbol{\beta X}) \\ P(y=2)=F(\mu_1-\boldsymbol{\beta X})- F(-\boldsymbol{\beta X}) \\ \ldots \\ P(y=I)=1-F(\mu_{I-2}-\boldsymbol{\beta X}), \end{gathered} \tag{7}$$

siendo F la función de distribución de la variable aleatoria ε. La distribución del término de error determina el tipo de modelo. De nuevo, las dos distribuciones más comúnmente usadas son la logística y la normal, dando lugar a los modelos logit ordenado (MLO) y probit ordenado (MPO), respectivamente.

- **Aplicaciones de los modelos logit ordenado y probit ordenado en el estudio de la gravedad del accidente.**

En el análisis de severidad de los accidentes se han encontrado diversas aplicaciones de los MLO. Abdelwahab and Abdel-Aty (2001) utilizaron MLO (junto con técnicas de Minería de Datos) para predecir la gravedad de las lesiones en intersecciones señalizadas en accidentes con dos vehículos. Examinaron la relación entre la severidad de la lesión con el conductor, el vehículo, la carretera y las características del entorno. Entre sus resultados destacaron que los MLO proporcionaban peores ajustes que las técnicas de Minería de Datos utilizadas. Khattak and Rocha (2003) utilizaron también MLO para evaluar la severidad de las lesiones resultantes en los conductores de vehículos deportivos. Entre sus resultados destacaron que estos vehículos tienen mayor probabilidad de sufrir un vuelco y, por tanto, de dañar gravemente a los conductores. Sin embargo, dada su mayor resistencia al choque, proporcionan mayor protección a los conductores cuando el accidente es una colisión.

Respecto a las aplicaciones que utilizan los MPO destaca el trabajo de Kockelman and Kweon (2002), en el que estudiaron la gravedad de las lesiones de las víctimas de accidentes en función de las características del conductor, del vehículo y del accidente. Crearon modelos separados para accidentes con un solo vehículo involucrado y para accidentes con dos vehículos involucrados, considerando 4 niveles de gravedad. Entre sus principales resultados obtuvieron que las mujeres tienen mayor probabilidad sufrir una lesión grave, y que los accidentes más peligrosos son los producidos por colisión frontal o vuelco del vehículo. Gray et al. (2008) los utilizaron para analizar la gravedad de los accidentes con conductores varones jóvenes en carreteras de Gran Bretaña. Consideraron 3 niveles de gravedad de accidente: leve, grave y mortal. Entre sus resultados destacan que la probabilidad de accidente grave o mortal aumenta cuando se conduce en la oscuridad, en carreteras de un carril, con los adelantamientos o cuando se golpea un objeto existente en la calzada. Posteriormente, Garder (2009) los utilizó para analizar la gravedad de las colisiones frontales en carreteras convencionales de 2 carriles. En el estudio realizado por Christoforou et al. (2010), para analizar la gravedad de los ocupantes en accidentes ocurridos en autovías, también se utilizaron los MPO. La particularidad de su estudio fue el uso de variables que hacen referencia a las características del tráfico recogidas a tiempo real (en el momento del accidente). En un estudio más reciente, Obeng (2011) los utilizó para analizar las diferencias según el género en la gravedad del accidente en intersecciones señalizadas. Estimaron modelos diferentes según el género del conductor y encontraron que las condiciones en la conducción, el tipo de accidente, el tipo y las características del vehículo influyen según el género, de un modo diferente en el riesgo de sufrir lesiones graves.

Otros autores han desarrollado estudios en los que utilizan los MPO junto con otras tipologías de modelos. Abdel-Aty (2003) los utilizó junto con modelos logit multinomiales y logit anidados, como se ha descrito anteriormente. Una aplicación más reciente se encuentra en el trabajo desarrollado por Haleem and Abdel-Aty (2010), en el que compararon los resultados de los MPO con los modelos probit binarios y modelos logit anidados, para analizar la severidad de los accidentes en las intersecciones no señalizadas de 3 y 4 brazos. De los modelos probit utilizados, los binarios proporcionaban mejores resultados que los ordenados, y los modelos logit anidados no mostraban ninguna mejora sobre los modelos probit utilizados.

La infra-detección entre los diferentes niveles de gravedad de los accidentes, es uno de los principales inconvenientes de los modelos ordenados, pudiendo producir resultados erróneos, en los que se sobreestima la probabilidad de los accidentes de mayor gravedad y se subestiman los niveles más bajos y, específicamente, los accidentes con daños materiales (Ye and Lord, 2011). Por tanto, las estimaciones de los parámetros pueden ser erróneas e inconsistentes (Islam and Manering, 2006; Ye and Lord, 2011). Cuando surge este problema, una alternativa es utilizar MLM que, en presencia de infra-detección en los datos, son consistentes excepto para el término constante (Washintong et al., 2011).

Otro inconveniente de los modelos ordenados es que los coeficientes de las variables explicativas (que indican el efecto de la variable en el resultado de la severidad) están restringidos a ser iguales en todos los niveles de severidad considerados en el modelo (Xie et al., 2012). Islam and Manering (2006) mostraron este inconveniente utilizando como ejemplo el efecto de la variable apertura del airbag; la cual puede tener diferentes impactos (negativos o positivos) según el nivel de gravedad del accidente. Y propusieron como alternativa utilizar los modelos logit multinomial, en los que para cada nivel de gravedad se tiene una función de gravedad separada, y pueden incluir diferentes conjuntos de variables explicativas. Por tanto, la estructura del modelo es más flexible y puede considerar diferentes efectos de las mismas variables según el nivel de gravedad. Otros autores (Jung et al., 2010; Quddus et al., 2010) utilizaron como alternativa una versión generalizada de los MLO, que permite relajar las restricciones sobre los coeficientes de las variables explicativas.

Los modelos ordenados asumen que la varianza en el término de perturbación es constante (homocedasticidad en el término de perturbación), sin embargo existen diferentes extensiones al supuesto de homocedasticidad en el término de perturbación como son los modelos logit/probit heterocedásticos y los modelos logit ordenados mixtos.

En la modelización de la severidad de los accidentes, los modelos logit/probit heterocedásticos fueron utilizados por O'Donnell and Connor (1996) sobre un conjunto de datos *cross-section,* para explicar la gravedad de las lesiones de accidentes de tráfico con víctimas ocurridos en carreteras australianas. La varianza del término de perturbación dependía de la edad de la víctima, de la velocidad y de la hora del accidente. Como variables explicativas se usaron la edad y sexo del herido, las

características del accidente y el número de vehículos involucrados. Y para la variable dependiente consideraron 4 niveles de gravedad: sin lesión, asistencia médica, hospitalización y muerte. Sus resultados mostraron que la probabilidad de sufrir un accidente con lesiones graves aumentaba con la edad del conductor, la velocidad de circulación, así como con el nivel de alcohol, el tipo de vehículo y el tipo de accidente. Estos modelos también fueron utilizados por Wang and Kockelman (2005) para analizar la gravedad de las lesiones de los ocupantes de un vehículo envueltos en un accidente, haciendo depender la varianza del error de la velocidad a la que circulaba el vehículo, el tipo de vehículo y su peso. Entre otros resultados, mostraron que los ocupantes de mayor edad, así como las mujeres, tienen mayor probabilidad de sufrir un accidente grave. En un estudio más reciente, Lemp et al. (2011) los utilizaron para analizar la severidad de los accidentes con camiones pesados involucrados.

Los modelos logit ordenados mixtos fueron aplicados en el análisis de gravedad por Srinivasan (2002). El objetivo de su estudio fue estimar la gravedad de las víctimas de los accidentes permitiendo a los umbrales de gravedad variar y estar correlacionados para un individuo dado. Esta especificación puede dar cuenta de los efectos no observados (como el comportamiento ante el riesgo de los conductores envueltos en un accidente). Desde el punto de vista teórico, demuestran que el supuesto de umbrales constantes puede llevar a errores e inconsistencias cuando se estima la gravedad, y que la variabilidad de los umbrales depende del individuo, así como de las características del vehículo, del tráfico o del tipo de accidentes, entre otras.

2.3.3. Resumen

Los DOM han sido los modelos más utilizados hasta el momento para analizar la severidad de los accidentes. Particularmente, los modelos más frecuentemente empleados por los investigadores han sido el ML, cuando la variable severidad es definida mediante 2 categorías; el MLM, cuando la variable severidad viene dada por más de 2 categorías; y los MLO y MPO, cuando se tiene en cuenta el orden implícito entre las categorías de las lesiones.

Dadas las particularidades analizadas de los datos de accidentes, tales como la infra-detección, la presencia de variables endógenas, correlaciones, etc; los investigadores han ido introduciendo diversas modificaciones de los modelos originales. En general la elección del modelo ha estado basada en los problemas detectados a la hora de analizar los accidentes, así como en el objetivo perseguido en cada estudio particular.

Así, por ejemplo, para tener en cuenta las correlaciones cuando se consideran los diferentes niveles de gravedad resultantes de un mismo accidente se ha utilizado el modelo logit binario simultáneo (Ouyang et al., 2002) y el modelo logit jerárquico bayesiano (Helai et al., 2008). Los modelos probit binario bivariante o multivariante (Wiston et al., 2006; Lee and Abdel-Aty, 2008) fueron empleados cuando existen variables explicativas endógenas con respecto a la gravedad de la lesión.

En presencia de infra-detección en los datos, pueden utilizarse los MLM (Washintong et al., 2011). Los modelos MLM también son una alternativa a los modelos ordenados cuando se buscan modelos más flexibles y que puedan considerar diferentes efectos de las mismas variables según el nivel de gravedad dado (Islam and Manering, 2006).

El modelo logit anidado o el modelo logit mixto son extensiones del MLM. El modelo logit anidado permite relajar el supuesto de la independencia de las alternativas irrelevantes, y se ha utilizado en numerosas aplicaciones para analizar la gravedad del accidente (Chang and Mannering, 1998; Chang and Mannering, 1999; Lee and Mannering, 2002; Holdridge et al., 2005, Haleem and Abdel-Aty, 2010). El modelo logit mixto, permite captar el efecto de la heterogeneidad y la correlación de los factores no observados (Train, 2009). De estos modelos también existen numerosas aplicaciones en la literatura para el estudio de la gravedad del accidentes (un resumen más detallado se puede consultar en Savolainen et al., 2011).

Los modelos ordenados asumen que la varianza en el término de perturbación es constante (homocedasticidad en el término de perturbación). Sin embargo existen diferentes extensiones al supuesto de homocedasticidad en el término de perturbación, como son los modelos logit/probit heterocedásticos y los modelos logit ordenados mixtos. Estas extensiones han sido aplicadas en la modelización de la severidad de los accidentes por diferentes autores (O'Donnell and Connor, 1996; Srinivasan, 2002; Wang and Kockelman, 2005; Lemp et al., 2011).

Por tanto, existe una gran variedad de técnicas que permiten modelar la severidad de los accidentes, y cada técnica presenta sus ventajas e inconvenientes particulares, que deben tenerse en cuenta a la hora de su elección. Sin embargo, el principal inconveniente de todos estos modelos es que parten de hipótesis fijas y predefinen relaciones entre las variables dependientes e independientes, de modo que si estas hipótesis no se cumplen los modelos pueden producir estimaciones erróneas en la probabilidad de la gravedad de la lesión (Chang and Wang, 2006). Esto ha provocado que numerosos investigadores hayan comenzado a utilizar técnicas de Minería de Datos (MD) para estudiar la severidad de los accidentes de tráfico.

Una de las principales diferencias entre al análisis de datos tradicional (estadístico) y la MD es que el análisis de datos tradicional supone que las hipótesis ya están construidas y validadas contra los datos, mientras que la MD normalmente supone que los patrones e hipótesis son automáticamente extraídos de los datos (Hernández, 2004).

Las técnicas de MD permiten extraer conocimiento de los datos previamente desconocido e indistinguible, no parten de hipótesis ni requieren un previo conocimiento probabilístico del problema objeto de estudio.

Además, el uso de técnicas de MD en el estudio de los accidentes resulta muy apropiado por la propia naturaleza de estos fenómenos. Un accidente puede definirse como un evento raro, aleatorio y de múltiples factores siempre precedido por una situación en la que uno o más conductores no pueden hacer frente al entorno de la

carretera (ROSPA, 2002). Así, cada accidente es el resultado de una cadena de eventos, que es, en su totalidad, único, pero algunos factores son comunes a varias circunstancias del accidente, y la identificación de estos factores y sus interdependencias puede llevarse a cabo mediante el uso de técnicas de MD (Montella et al., 2012b).

2.4. Minería de Datos.

Los volúmenes de información que se manejan en la sociedad actual se multiplican día a día, generándose enormes bases de datos, cuyo verdadero valor reside en la información que pueda extraerse de éstas. Por ejemplo, los accidentes de tráfico son almacenados en las bases oficiales de datos de la DGT. De estas bases de datos puede extraerse y consultarse gran cantidad de información. Sin embargo, los datos por sí solos pueden tener un valor relativo, y lo realmente interesante es el conocimiento que puede inferirse a partir de ellos, y más aún, la capacidad de poder usar este conocimiento (Hernández et al., 2004).

A lo largo de los años, se han desarrollado un gran número de métodos de análisis de datos basados en la estadística (ver Michalski et al., 1982). Sin embargo, en la medida en que se ha incrementado la cantidad de información almacenada en las bases de datos, estos métodos han comenzado a enfrentarse a problemas de eficiencia y escalabilidad.

Los problemas y limitaciones de las aproximaciones clásicas han hecho surgir la necesidad de herramientas y técnicas que permitan la extracción de conocimiento útil, a partir de la información disponible (Hernández et al., 2004), las cuales se engloban bajo el nombre de Minería de Datos (MD).

La MD puede definirse como un conjunto de técnicas y herramientas aplicadas al proceso no trivial de extraer y presentar conocimiento implícito (previamente desconocido), potencialmente útil y humanamente comprensible, a partir de grandes conjuntos de datos, con objeto de predecir de forma automatizada tendencias y comportamientos y/o describir de forma automatizada modelos previamente desconocidos (Piatetsky-Shapiro and Frawley, 1991). Por tanto, los retos de la MD son:

- Trabajar con grandes bases de datos, con los problemas que ello conlleva (ruido, datos perdidos o en blanco, intratabilidad, etc).
- Usar técnicas adecuadas para el análisis de los datos, y extraer conocimiento novedoso y útil. La utilidad del conocimiento minado puede estar relacionada con la comprensibilidad del modelo inferido, ya que el usuario final no tiene por qué ser experto en estas técnicas, ni debe invertir mucho tiempo interpretando los resultados (Hernández et al., 2004). Por ello es importante que la información descubierta sea lo más comprensible posible (usando representaciones gráficas,

convirtiendo los patrones a lenguaje natural o utilizando técnicas de visualización de datos).

La MD es una etapa del proceso "Knowledge Discovery in Databases" (KDD), aunque muchos autores lo han identificado como si fuesen el mismo proceso (Fayyad et al. 1996b). En Fayyad et al. (1996a) se define el KDD como un proceso no trivial para la identificación de patrones a partir de los datos, válidos, novedosos, potencialmente útiles, y en última instancia comprensibles.

El desarrollo del proceso KDD es un proceso complejo que involucra la consecución de 5 pasos, esquematizados en la Figura 2 tomada de Fayyad et al. (1996b).

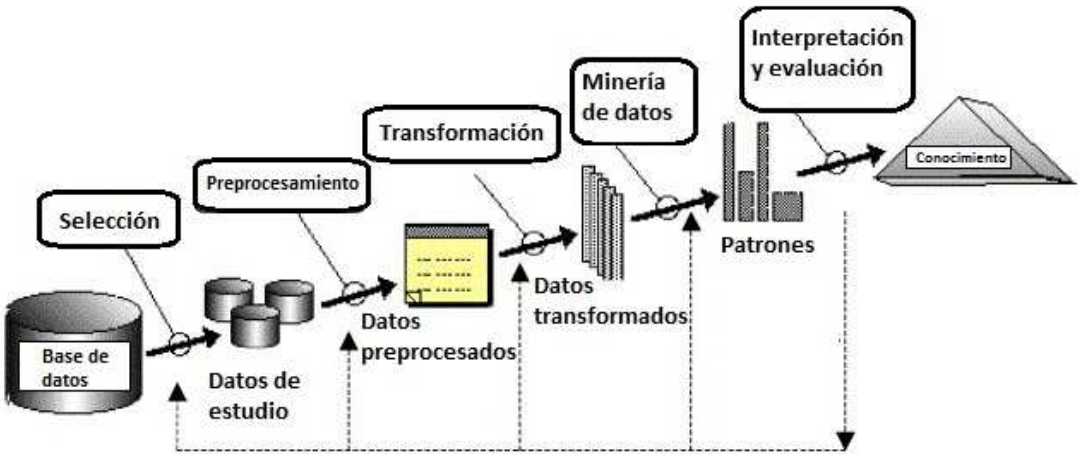


Figura 2.- Esquema del proceso Knowledge Discovery in Databases (Fayyad et al., 1996b).

- **Selección de datos.** En esta etapa se determinan las fuentes de datos y el tipo de información a utilizar. Es la etapa donde los datos relevantes para el análisis son extraídos desde las fuentes originales de datos.

- **Preprocesamiento.** Esta etapa consiste en la preparación y limpieza de los datos extraídos desde las distintas fuentes de datos en una forma manejable, necesaria para las fases posteriores. En esta etapa se utilizan diversas estrategias para manejar datos perdidos o datos en blanco, datos inconsistentes o que están fuera de rango, obteniéndose al final una estructura de datos adecuada para su posterior transformación.

- **Transformación.** Consiste en el tratamiento preliminar de los datos, transformación y generación de nuevas variables a partir de las ya existentes con una estructura de datos apropiada. Aquí se realizan operaciones de agregación o normalización, consolidando los datos de una forma necesaria para la fase siguiente.

- **Minería de Datos.** Es la fase en la que se construye el modelo, en la que se aplican métodos inteligentes con el objetivo de extraer patrones previamente

desconocidos, válidos, nuevos, potencialmente útiles y comprensibles, y que están contenidos u "ocultos" en los datos.

- **Interpretación y Evaluación.** Se identifican los patrones obtenidos y que son realmente interesantes, basándose en algunas medidas y además se realiza una evaluación de los resultados obtenidos.

2.4.1. Técnicas de Minería de Datos.

De un modo general se puede decir que el objetivo de la MD es analizar los datos para extraer conocimiento. El conocimiento puede ser expresado en forma de relaciones, patrones o reglas inferidos de los datos y previamente desconocido, o bien en forma de una descripción más concisa (como un resumen de los mismos). Estas relaciones o resúmenes constituyen el modelo de los datos analizados, distinguiéndose dos tipos de modelos (ver Hernández et al., 2004):

- **Predictivos:** pretenden estimar valores futuros o desconocidos de las variables de interés, usando otras variables o campos de la base de datos.

- **Descriptivos:** identifican patrones que explican o resumen los datos, de modo que se utilizan para explorar las propiedades de los datos examinados, y no para predecir nuevos datos.

Dentro de MD, existen diferentes técnicas que permiten la extracción del conocimiento que reside en los datos. En general, una técnica es el enfoque conceptual que permite dicha extracción y es implementada por un determinado algoritmo.

Las técnicas de MD se pueden clasificar en dos grandes grupos en función del proceso utilizado para la adquisición del conocimiento: técnicas supervisadas y técnicas no supervisadas (ver Weiss and Indurkaya, 1998).

- **Técnicas supervisadas.** Se caracterizan porque el aprendizaje se realiza a partir de ejemplos. Los ejemplos de entrada van acompañados de una clase o salida correcta. De modo que existe un atributo especial, normalmente denominado clase, presente en todos los casos, que especifica si un caso pertenece o no a un cierto concepto, que será el objetivo del aprendizaje.

- **Técnicas no supervisadas.** Se caracterizan porque el aprendizaje se realiza por observación, no existe ningún atributo especial que guíe el proceso de aprendizaje. Se construyen descripciones, hipótesis o teorías a partir de un conjunto de datos u observaciones sin que exista una clasificación previa de los ejemplos

En la Figura 3 (adaptada de Rokach and Maimon, 2008) se muestra un esquema dónde se recogen las técnicas más utilizadas en Minería de Datos:

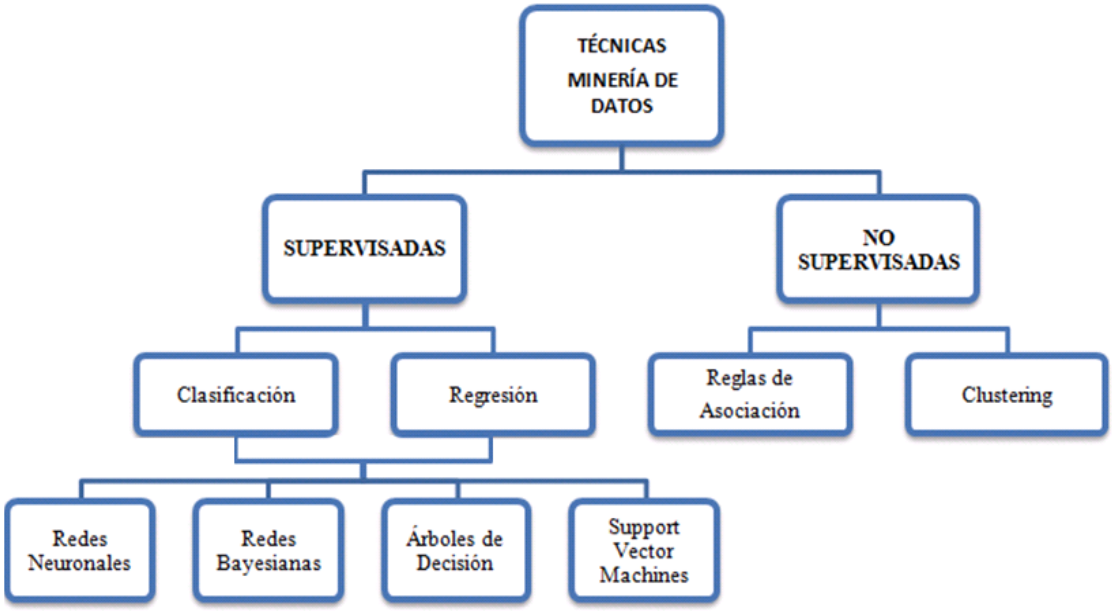


Figura 3.- Esquema de las principales técnicas de Minería de Datos (adaptada de Rokach and Maimon, 2008.

Las técnicas de MD han sido utilizadas en campos muy diversos, tales como las finanzas, el marketing, el comercio, las telecomunicaciones, la manufactura e incluso en la industria de la salud (seguros médicos, diagnóstico y tratamiento de enfermedades), es decir, en sectores que han requerido de técnicas avanzadas para la extracción de información útil y rentable dado los grandes volúmenes de datos que manejan habitualmente.

2.4.2. Técnicas de Minería de datos para analizar la gravedad de los accidentes.

Las técnicas de MD también han comenzado a aplicarse en el campo de la seguridad vial. Particularmente en el análisis de la severidad de los accidentes, las técnicas más utilizadas has sido las técnicas de clasificación supervisadas tales como las Redes Neuronales, las Redes Bayesianas, las Reglas de Asociación y los Arboles de Decisión; y dentro de las técnicas no supervisadas las Reglas de Asociación.

A continuación se explican las principales características de cada una de estas técnicas junto con los estudios que las han utilizado para analizar la gravedad de los accidentes.

2.4.2.1. Redes Neuronales Artificiales (RNA).

Son generalizaciones de modelos estadísticos clásicos cuya estructura y operación está inspirada en las redes neuronales biológicas. Una RNA puede verse como un grafo dirigido formado por un conjunto interconectado de elementos simples de procesamiento, unidades o nodos. La capacidad de procesamiento de la red se almacena en las fuerzas de conexión entre las unidades, o pesos, obtenidos por un proceso de aprendizaje a partir de un conjunto de patrones de entrenamiento

(Gurney, 1997). El objetivo de las RNA es conseguir que la red aprenda automáticamente las propiedades deseadas a partir de un conjunto de datos de entrada (suficientemente significativo).

Se componen de unidades simples llamadas neuronas. Cada neurona tiene asociada una función matemática (función de transferencia) que genera la salida de la neurona a partir de las señales de entrada. La entrada de la función es la suma de todas las señales de entrada por el peso asociado a la conexión de entrada de la señal. De este modo, la función de transferencia es la relación entre la señal de salida y de entrada.

En la Figura 4 se muestra la estructura de una RNA. La cual está formada por una serie de variables descriptoras que forman la capa de entrada; éstas se multiplican por unos pesos (capa oculta) que van variando durante la fase de aprendizaje hasta llegar a la solución buscada, que es dada en la capa de salida. El tipo de RNA más utilizada es el perceptrón multicapa (MLP).

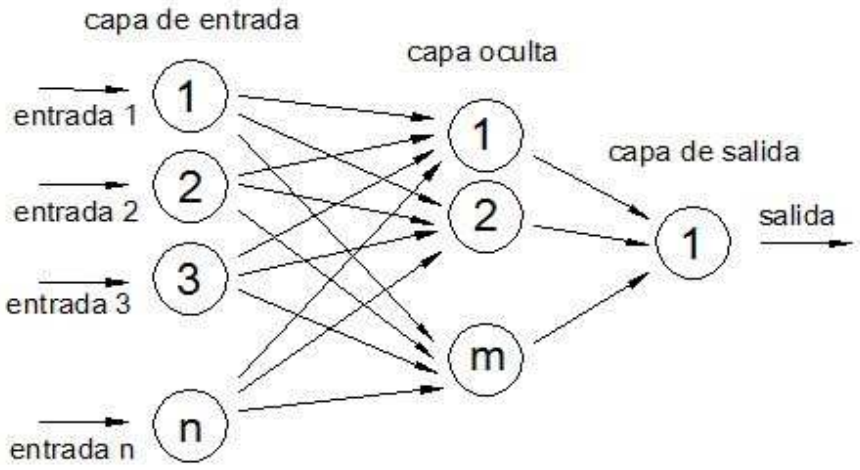


Figura 4.- Ejemplo de estructura de una Red Neuronal Artifical del perceptrón multicapa.

Las ventajas de las RNA es que permiten modelar problemas complejos en los que puede haber interacciones no lineales entre variables. El principal inconveniente es que tiene el efecto de "caja negra". Los datos entran en la "caja negra" y se obtienen las predicciones, pero no se revela normalmente la naturaleza de las relaciones entre las variables independientes y dependientes. En una RNA el conjunto de los pesos determina el conocimiento de esa red y permiten resolver el problema para el que ha sido entrenada. Sin embargo, el conocimiento de la red expresado de este modo (en forma de pesos), impide la inteligibilidad de las asignaciones de clase que se realizan. Estos pesos son ocultos y no pueden ser modificados por el operador (ver Tullis and Jensen, 2003). Es decir, de las variables descriptoras no se extrae nuevo conocimiento para el usuario, sino que esa extracción de conocimiento es interna de la red y no revierte en el usuario salvo en la asignación de clases realizada. Otro de los inconvenientes de las RNA es que el modelo aprendido es difícilmente comprensible, y requieren gran cantidad de datos para su entrenamiento.

- **Aplicaciones de RNA en el estudio de la gravedad del accidente.**

Las primeras aplicaciones de RNA para el estudio de la severidad de los accidentes fueron desarrolladas por Abdelwahab y Abdel-Aty (2001; 2004). En el primer estudio investigaron el uso de las RNA para predecir la gravedad de las lesiones de accidentes con 2 vehículos involucrados en intersecciones señalizadas. El objetivo era encontrar las relaciones entre la severidad del conductor con los factores relativos al conductor, al vehículo, a la carretera y al entorno. Utilizaron un total de 13 variables, pero sólo 6 resultaron significativas (género, culpable, tipo de vehículo, uso del cinturón de seguridad, punto del impacto y tipo de zona). Compararon los resultados de las RNA (creadas con el enfoque del MLP) con modelos probit ordenados, obteniendo que la precisión del modelo creado con RNA era superior (65.6 y 60.4% frente a un 58.9 y 57.1%). Posteriormente (Abdelwahab and Abdel-Aty, 2002) aplicaron la teoría difusa para el desarrollo de una RNA. El objetivo de su investigación era analizar la efectividad de la técnica para predecir la gravedad de los conductores implicados en un accidente. Analizaron para ello los accidentes ocurridos en 3 localizaciones diferentes. En Abdel-Aty and Abdelwahab (2004) compararon dos tipos de RNA (enfoque del MLP y las creadas con la teoría difusa) con modelos probit ordenados para predecir el nivel de severidad resultante en un accidente de tráfico. Obteniendo que los modelos con RNA proporcionaban mejores ajustes.

En el estudio realizado por Delen et al. (2006) también se utilizaron las RNA para modelar la severidad de los accidentes. Desarrollaron diferentes RNA (creadas con el enfoque del MLP) para los 5 niveles de severidad que habían considerado. Entre sus resultados destacaron que todos los modelos presentaban alto poder predictivo y ayudaban a identificar las variables más importantes para cada nivel de gravedad considerado. En un estudio más reciente (Moghaddam et al., 2011) las RNA fueron aplicadas para analizar la severidad de los accidentes en carreteras de ámbito urbano. En este estudio se subrayó como una de sus ventajas la capacidad de predecir y mostrar resultados adecuados incluso teniendo en cuenta las limitaciones de los datos de accidentes.

2.4.2.2. Redes Bayesianas (RBs).

Son modelos gráficos acíclicos dirigidos en el que cada nodo representa una variable y cada arco representa una dependencia probabilística. La variable a la que apunta el arco es dependiente (causa-efecto) de la que está en el origen de éste. De este modo, la topología o estructura de la red es una forma compacta de representar el conocimiento que aporta información sobre las dependencias probabilísticas entre las variables y sobre las independencias condicionales de una variable (o conjunto de variables) dada otra variable(s).

Las RBs pueden codificarse a partir del conocimiento de un experto o pueden ser inferidas a partir de los datos. La obtención de la red a partir de datos, es un proceso de aprendizaje que se divide en dos etapas: el aprendizaje estructural y el aprendizaje paramétrico (Pearl, 1988). La primera consiste en obtener la estructura de la red

bayesiana, es decir, las relaciones de dependencia e independencia entre las variables involucradas. Y la segunda etapa tiene como finalidad obtener las probabilidades a priori y condicionales requeridas a partir de una estructura dada. Por tanto, las RBs permiten establecer relaciones causales y efectuar predicciones.

En la Figura 5 se muestra un ejemplo de una RB con dos eventos (visibilidad de la carretera y condiciones atmosféricas de lluvia fuerte) que pueden causar la ocurrencia de otro (un accidente grave). Si las 3 variables consideradas pueden tomar dos posibles valores, T (para verdadero) y F (para falso) y la lluvia tiene un efecto sobre la visibilidad (cuando llueve hay menor visibilidad), la RB que modela el problema puede ser la que se muestra en la Figura 5.

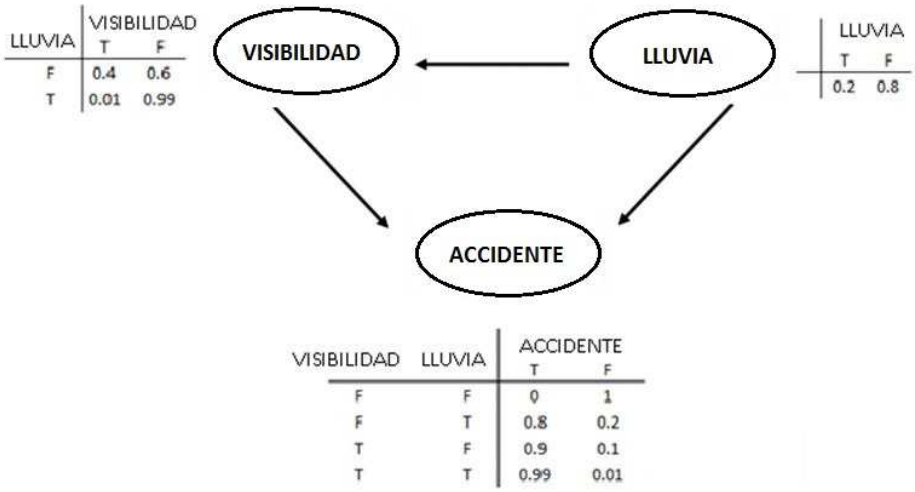


Figura 5.- Esquema de una Red Bayesiana.

Entre las principales ventajas de las RBs destacan las siguientes: permiten modelar sistemas complejos entendiéndose las relaciones causales, visualizándolas por medio del grafo; permiten hacer inferencia en ambos sentidos, es decir, las variables de entrada pueden ser usadas para predecir las variables de salida y viceversa; la salida de una RB es una probabilidad de distribución en lugar de valores únicos (por ejemplo, en una variable con estados bajo, medio, y alto, las RBs estiman la probabilidad de cada uno de los estados); permiten combinar conocimiento con datos (Heckerman et al., 1995); evitan el sobre-ajuste de los datos; y pueden manejar bases de datos incompletas (Heckerman, 1995; Ramoni and Sebastiani, 1996).

Sin embargo también presentan ciertos inconvenientes, entre los que se pueden citar los siguientes: puede ser difícil describir una estructura compleja incluso para expertos del dominio, especialmente si el dominio es nuevo, creando desigualdades entre el problema del dominio y el modelo construido; si dos variables están relacionadas de forma poco evidente, entonces habrá una dependencia entre ellas y puede que el modelo no tenga esta relación en cuenta; en un sistema con un gran número de variables puede ser difícil asegurar su consistencia. La inferencia Bayesiana es útil sólo si se puede confiar en los parámetros ya que una mala estimación de los parámetros distorsionaría toda la red e invalidaría los resultados.

- **Aplicaciones de RBs en el estudio de la gravedad del accidente.**

Las aplicaciones de RBs que se encuentran en el estudio de severidad son recientes. Simoncic (2005) utilizó RBs para modelar la severidad de los accidentes de tráfico con 2 vehículos involucrados. Mostraron la posibilidad de utilizar esta técnica en el estudio de los accidentes, indicando que en comparación con otros conocidos métodos estadísticos, la principal ventaja de las RBs es su complejo enfoque en el que las variables del sistema son interdependientes, de modo que no son necesarias variables dependientes e independientes.

Además, las RBs han sido utilizadas en 3 estudios que tienen como objetivo identificar los factores claves que afectan a la severidad de los accidentes ocurridos en carreteras convencionales de dos carriles. En De Oña et al. (2011) se muestra el uso de esta técnicas para alcanzar el objetivo anteriormente comentado. Sus resultados indicaron que las RBs proporcionan modelos con unos valores de ajuste razonables para el estudio del problema. Sin embargo, el principal inconveniente que encontraron es la gran cantidad de datos que se necesitan para trabajar con las RBs, y su uso en muestras no-balanceadas. En Mujalli and De Oña (2011) el objetivo era crear RBs más simples y que ayudaran a una mejor comprensión del problema. Para reducir el número de variables a modelar utilizaron diversas técnicas de selección de variables. Concluyeron que es posible modelar el problema con un menor número de variables, sin disminuir la precisión de la RB creada. Finalmente, De Oña et al. (2013) combinaron la técnica de Análisis de Clases Latentes junto con RBs. El objetivo era realizar una segmentación previa de los datos y reducir su heterogeneidad para posteriormente modelar el problema con RBs. Sus resultados mostraron ciertas relaciones entre las variables que forman las RBs que no podrían haber sido detectadas sin realizar una previa segmentación de los datos.

2.4.2.3. Reglas de Asociación (RA).

Las RA constituyen un mecanismo de representación del conocimiento muy simple y útil para caracterizar las regularidades que se pueden encontrar en grandes bases de datos. Son modelos de aprendizaje no supervisado, que se utilizan cuando el resultado de interés no es conocido y el sistema debe aprender directamente de los datos disponibles.

La extracción de RA se ha aplicado tradicionalmente a bases de datos transaccionales (bases de datos que tienen como objetivo la recepción y envío de datos a gran velocidad). En estas bases, una transacción T es un conjunto de artículos o ítems, junto a un identificador único; y una transacción contiene un conjunto de ítems I, si I está incluido en T. Un ejemplo típico de su uso es en el análisis de la cesta de la compra (*market basket analysis*), para identificar productos que se compran juntos frecuentemente.

Así una regla de asociación es una implicación del tipo "X→Z", donde X y Z son ítems, y se conocen como: "X" - antecedente y "Z" - consecuente. En estas reglas, tanto el antecedente como el consecuente, pueden estar formados por conjuntos de valores (o

conjuntos de variables). Así, el significado intuitivo de una regla de asociación "X→Z" es que las transacciones en la base de datos que contienen a X también tienden a contener a Z.

Existen diferentes algoritmos eficaces que permiten extraer RA. No obstante, el algoritmo clásico para generar RA es el algoritmo Apriori de Agrawal et al. (1993). La idea básica de este algoritmo es generar de forma progresiva y recursiva conjuntos de ítems frecuentes que aparecen juntos en la base de datos con un porcentaje mínimo de ocasiones.

Para medir el interés de una RA normalmente se utilizan 3 parámetros:

- Support: mide la cantidad de veces que aparecen los ítems de dicha regla en la base de datos (se calcula dividiendo el número de casos que contienen el antecedente y el consecuente entre el número total de casos).
- Confidence: es el porcentaje de veces que, apareciendo en una instancia los ítems del antecedente, aparecen también los ítems del consecuente (se computa como el número de veces que aparecen juntos el antecedente y el consecuente dividido entre el número de veces que aparece solo el antecedente).
- Lift: indica qué probabilidad existe de encontrar el consecuente limitando la búsqueda a aquellos conjuntos de ítems donde el antecedente está presente (se obtiene como la proporción de la confidence de la regla entre el support del consecuente).

Un ejemplo de cómo se calculan estas RA se muestra a continuación:

Dadas las 4 transacciones de la base de datos de Figura 6, se pretenden encontrar las reglas de asociación que tienen un mínimo de 50% de support y confidence.

Num. Transacción	Elementos Comprados					
	A	B	C	D	E	F
1	1	1	1	0	0	0
2	1	0	1	0	0	0
3	1	0	0	1	0	0
4	0	1	0	0	1	1

Figura 6.- Ejemplo de Transacciones.

Las reglas con un support mínimo de 50% y una confidence mínima de 50% son:

$$A \rightarrow C: \text{Support} = 2/4 = 0{,}5 \rightarrow 50\%; \text{Confidence} = 2/3 = 0{,}66 \rightarrow 66{,}6\%$$

$$C \rightarrow A: \text{Support} = 2/4 = 0{,}5 \rightarrow 50\%; \text{Confidence} = 2/2 = 1 \rightarrow 100\%$$

La principal característica de las RA es que permiten generar patrones que son muy fácilmente comprensibles. Y como principal limitación de los modelos que se utilizan para obtenerlas, se destaca que cuando éstos son capaces de generar muchas reglas, suelen aparecer una gran cantidad que no son "interesantes" o son redundantes. La

aparición de patrones debidos al azar es lo que se conoce como riesgo de error Tipo I (Webb, 2007).

- **Aplicaciones de RA en el estudio de la gravedad del accidente.**

Hasta dónde se conoce no se han encontrado estudios que apliquen esta técnica con el objetivo de esta revisión bibliográfica (estudios sobre el análisis de la gravedad de accidentes en los que hay al menos un automóvil involucrado, quedando fuera de este ámbito aquellos que únicamente se centran en analizar la gravedad de un determinado tipo de usuario vulnerable (peatones, bicicletas y/o motocicletas)). Sin embargo, si se han encontrado diversas aplicaciones en el campo de la seguridad vial:

Geurts et al. (2005) utilizaron RA para identificar y diferenciar patrones de accidentes que ocurren dentro y fueran de las denominadas áreas negras (o segmentos de carretera peligrosos). Y entre sus resultados destacaron que la metodología puede identificar potenciales relaciones que no son conocidas en la literatura de la seguridad vial. Pande and Abdel-Aty (2009) las utilizaron para detectar patrones en los accidentes ocurridos en carreteras sin intersecciones. Y concluyeron que los resultados eran consistentes con la comprensión de las características del accidente y subrayaron el potencial de esta metodología para ser utilizada como una herramienta de decisión por las administraciones de seguridad. Montella (2011) también utilizó RA para detectar patrones en los accidentes con el objetivo de identificar los factores que contribuyen a los accidentes en glorietas urbanas. Posteriormente las RA fueron utilizadas con el objetivo de analizar las características que influyen en los accidentes con peatones (Montella et al., 2011) y para analizar los accidentes con motocicletas (Montella et al., 2012b).

2.4.2.3. Árboles de Decisión (ADDs).

Los ADDs son grafos direccionados formados por nodos (variables de entrada), ramas (asociadas a los valores de la variable que forma el nodo) y hojas, nodos hoja o nodos terminales (valores de la variable de salida).

Destacan por su sencillez y transparencia, y se explican por sí mismos, de modo que no es necesario ser un experto en Minería de Datos para ser capaz de seguir una determinada decisión del árbol. Además, su representación gráfica como una estructura jerárquica hace que sean fácilmente comprensibles, y por tanto, más fáciles de interpretar que otras técnicas.

Dado que los ADDs son la herramienta utilizada para el desarrollo de esta tesis doctoral, son explicados de modo más detallado en el epígrafe 2.5 de este capítulo.

- **Aplicaciones de ADDs en el estudio de la gravedad del accidente.**

En la literatura de seguridad vial se encuentran diversas aplicaciones que utilizan ADDs para analizar la gravedad de los accidentes.

Existen numerosos algoritmos que permiten la construcción de ADDs, sin embargo el algoritmo CART (Classification and Regresion Trees) desarrollado por Breiman et al., (1984) es el que comúnmente más se ha utilizado para analizar la severidad de los accidentes.

Autores como Kuhnert et al. (2000) compararon los resultados obtenidos con CART, con los Splines de Regresión Adaptativa Multivariante (MARS) y con un modelo de regresion logística para analizar un caso de control epidemiológico de las heridas resultantes de accidentes tráfico. Los resultados indicaron que las técnicas no paramétricas como CART y MARS proporcionaban modelos más informativos y atractivos cuyos componentes individuales podían visualizarse gráficamente. Chang and Wang (2006) analizaron las relaciones entre la severidad de los accidentes con las características relacionadas con el conductor y el vehículo, así como con variables relacionadas con la carretera, el accidente y el entorno. Mostraron la técnica CART como una herramienta útil en problemas de predicción y clasificación. En un estudio posterior, Pakgohar et al. (2010) analizaron el rol que juegan las características del conductor en la severidad resultante del accidente utilizando la técnica CART y una regresión logísitica multinomial. Obtuvieron que el método CART proporcionaba resultados con mayores precisiones, además de ser éstos más simples y fáciles de interpretar. Kashani et al. (2011) analizaron los principales factores que afectan a la gravedad de los conductores que se ven envueltos en un accidente en carreteras convencionales de 2 carriles. Posteriormente, Kashani and Mohaymany (2011) utilizaron CART para identificar los principales factores que afectan a la gravedad de los ocupantes de un vehículo que sufren un accidente en ese tipo de carreteras. En una aplicación reciente (Chang and Chien, 2013) los utilizaron para identificar los factores que influyen en la severidad de los conductores envueltos en accidentes con camiones.

2.5. Árboles de decisión: conceptos generales.

Los ADDs son uno de los modelos más utilizados en aprendizaje supervisado y en aplicaciones de Minería de Datos (Gehrke, 1999). En general, el objetivo de cualquier algoritmo de aprendizaje supervisado es construir un modelo de clasificación a partir de un conjunto de datos de entrada, normalmente denominado conjunto de entrenamiento (training set). El conjunto de entrenamiento está formado por datos (casos u ejemplos) con cada una de las clases que se pretenden modelar, así como por una serie de atributos o características que se utilizarán para construir un modelo de clasificación. La variable que se pretende modelar o predecir es llamada, normalmente, variable clase, y el resto de variables del conjunto de datos son las variables atributo.

El objetivo del proceso de aprendizaje es la obtención de un modelo que puede ser utilizado para clasificar nuevos ejemplos (casos cuyas clases se desconozcan a priori), para detectar patrones en los datos o para comprender mejor el fenómeno que se está analizando.

Los ADDs son una técnica muy versátil que puede utilizarse en áreas de muy diversa índole, desde aplicaciones de diagnóstico médico hasta sistemas de predicción meteorológica o, recientemente, en el campo de la seguridad vial.

En MD un árbol de decisión es un modelo predictivo que puede ser utilizado para tareas de clasificación o para tareas de regresión, según sea la naturaleza de la variable clase: discreta (árbol de clasificación) o continua (árbol de regresión).

Esta investigación se centra en los árboles de clasificación, ya que la variable en estudio será de naturaleza discreta.

Según se describe en Murthy (1998), un ADD puede utilizarse para el análisis de datos con uno o varios de los siguientes fines:

- Descripción: reducir una gran cantidad de datos transformándolos en una forma más compacta que preserva las características esenciales y proporciona un resumen preciso.
- Clasificación: descubrir si los datos contienen clases de objetos bien diferenciadas que puedan ser interpretadas de manera significativa en el contexto de una teoría sustantiva.
- Generalización: descubrir una relación entre variables independientes y dependientes que sea útil para predecir el valor de la variable dependiente en el futuro.

Un ADD es una forma de representar el conocimiento obtenido en el proceso de aprendizaje inductivo. Puede verse como un conjunto de condiciones organizadas en una estructura jerárquica en forma de árbol, formada por diferentes nodos que se conectan con arcos (o ramas) dirigidos, diferenciándose:

- Nodo raíz: Es el nodo inicial, sólo tiene ramas salientes y en él queda recogida la totalidad de la población (o datos de estudio).
- Nodos intermedios (u hijos): Poseen ramas entrantes (que proviene de los nodos padre) y ramas salientes (que apuntan a los nodos hijo). Cada uno de estos nodos contiene una pregunta sobre un atributo concreto, que es una unidad que evalúa una función de decisión para determinar cuál es el próximo nodo hijo por el cual la rama se divide (existe un nodo hijo por cada posible respuesta).
- Nodo terminal (u hoja): Representan la partición final, sólo tienen ramas entrantes (no tienen nodos hijos) y se asocia con una etiqueta o valor que caracteriza a los datos que llegan al nodo. De modo que cada nodo hoja se refiere a una decisión (etiquetada con una de las clases del problema).

Así, dentro del árbol, cada nodo representa una variable atributo (X) y, cada rama representa un estado de esa variable. Normalmente cada nodo terminal representa el valor esperado de la variable clase o variable en estudio (C) según la información contenida en el conjunto de datos utilizado para construir el modelo. La clasificación

de una nueva instancia se realiza en base a una serie de preguntas sobre los valores de sus atributos, empezado por el nodo raíz y siguiendo el camino determinado por las respuestas a las preguntas de los nodos intermedios, hasta llegar a un nodo hoja. La etiqueta asignada a esta hoja es la que se asignará a la instancia a clasificar.

La Figura 7 muestra un ejemplo de un problema de clasificación resuelto con un ADD. Dada una muestra de datos de accidentes, se pretende conocer la gravedad de los mismos en función de dos variables atributo: *edad* y *velocidad*. La estructura del árbol muestra cómo son clasificados los accidentes según las 2 posibles categorías de la variable clase: *accidente leve vs. accidente mortal*. Además, en el gráfico de la derecha se representan el número de casos que recogen en cada nodo terminal u hoja (nodos sombreados en el árbol), distinguiéndose los casos que se predicen correctamente en cada uno de los mismos.

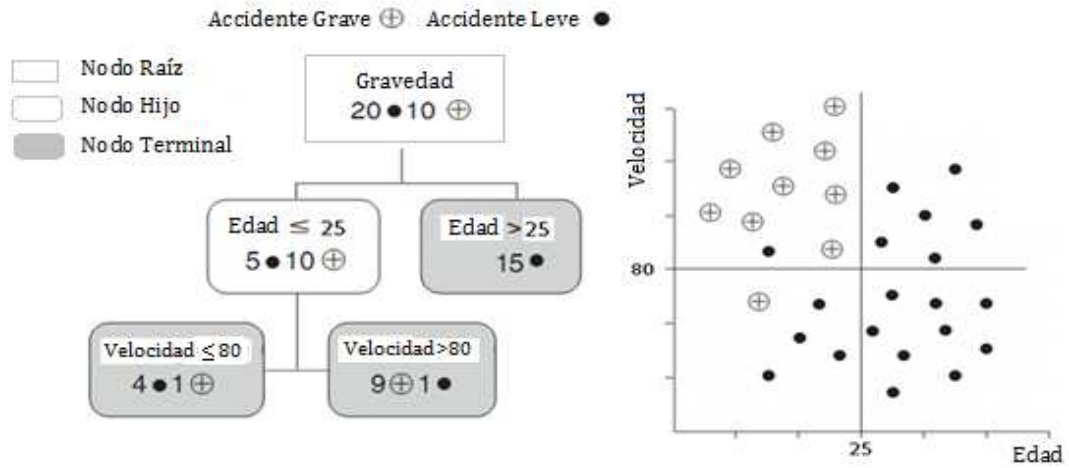


Figura 7.- Estructura de un Árbol de Decisión y del proceso de clasificación.

En la parte derecha de la Figura 7 se muestra además la representación gráfica de los casos agrupados en cada nodo. De esta forma puede entenderse el funcionamiento de un método de clasificación, y la idea de precisión. En cada área del gráfico se observan los casos clasificados en cada nodo según una categoría de la variable clase, y estos casos pueden estar bien clasificados (cuando coinciden con el valor de la variable clase) o mal clasificados (cuando no coinciden con el valor de la variable clase). En este ejemplo el porcentaje de los casos mal clasificados (que nos da la precisión del método) es pequeño (gráfico de la Figura 7).

2.5.1. Construcción de ADDs.

Los ADDs se construyen recursivamente siguiendo una estrategia descendente, desde los conceptos generales hasta los ejemplos particulares. Por ello, a la familia de algoritmos de construcción de ADDs se la conoce como "*Top-Down Induction of Decision Tress, TDIDT*".

La generación de la estructura de un árbol se fundamenta en el principio de "divide y vencerás". A partir del conjunto completo de datos (que forma el nodo raíz) y dado un

criterio de partición determinado se divide el conjunto en subconjuntos cada vez más pequeños (dando lugar a los nodos intermedios del árbol). Recursivamente se va dividiendo cada uno de los subconjuntos creados hasta que todos ellos son puros (cuando los casos del nodo son de una misma clase) o hasta que su "pureza" no puede incrementarse; se forman así los nodos hoja del árbol. Y finalmente, a los nodos hoja se le asigna una etiqueta o valor determinado de la variable clase.

Si no se establece ningún límite, la construcción del árbol se detiene cuando el nodo es "puro". Sin embargo, este criterio puede dar lugar a un sobreajuste de los datos (overfitting), que reduce la aplicabilidad del modelo de clasificación aprendido. Hace que el modelo construido sea muy específico, poco general, y por tanto malo para otros conjuntos de datos. Pero además, si los datos contienen ruido (errores en los atributos o en las clases), el modelo intentará ajustarse a los errores, perjudicando el comportamiento global del modelo aprendido. Para evitar el sobreajuste existen numerosas estrategias (que en ocasiones son complementarias), tales como reglas de parada y los métodos de poda (ver Murthy, 1998).

Normalmente, el objetivo cuando se construye un ADD es alcanzar el máximo grado de pureza en los nodos usando el menor número de particiones posibles, así el árbol resultante debe ser pequeño y el número de instancias en cada subconjunto grande.

Lo deseable es que la complejidad del árbol, que puede ser medida según diferentes métricas (número de nodos, número de hojas, cota de profundidad del árbol y número de atributos usados en su construcción), sea la menor posible para lograr la máxima comprensión del modelo. Para controlar la complejidad se pueden utilizar determinadas reglas de parada y/o un determinado método de poda. Además hay que tener en cuenta que las la complejidad del árbol tiene un efecto crucial en la precisión del método (Breiman et al., 1984).

- **Criterios de partición.**

Los criterios de partición o ramificación se basan generalmente en medidas de la impureza de nodo. Y la bondad de la partición se mide como el decrecimiento de la impureza que se consigue con ella. Por tanto, la mayoría de los criterios de partición tratan de maximizar la bondad de la partición (lo que equivale a minimizar la bondad de la impureza del árbol generado por la partición).

Los criterios de partición más utilizados (y que serán explicados en el capítulo de materiales y métodos) son: el índice de Gini usado en el algoritmo CART (Breiman et al., 1984); la ganancia de información, usada en el algoritmo ID3 (Quinlan, 1986); o el ratio de ganancia de información, usado en el algoritmo C4.5 (Quinlan, 1993).

- **Reglas de parada.**

Las reglas de parada tratan de predecir si conviene seguir construyendo el árbol por una determinada rama o no. En general, la fase de crecimiento del árbol continúa hasta que un criterio de parada se activa. Las condiciones más comúnmente utilizadas como criterios de parada son las siguientes:

- Pureza del nodo. Cuando un nodo solamente contiene casos de una misma clase, el proceso de construcción del árbol finaliza, ya que el nodo es "puro". Sin embargo, también puede utilizarse un umbral de pureza para detener la ramificación, cuando ésta no suponga una disminución significativa de la pureza del nodo (según alguna medida estadística de pureza). El grado de pureza está relacionado con el de información sobre la variable en estudio. Así mismo es posible usar medidas de información, como en Quinlan (1986, 1993), donde se sustituye el concepto de pureza por el grado de información que una variable atributo presenta sobre la variable clase.

- Cota de profundidad. Se puede establecer de antemano una determinada cota de profundidad para no construir árboles excesivamente complejos, es decir, excesivamente grandes. Así, cuando un nodo se halle a más de cierta profundidad, se detiene el proceso de generación del árbol.

- Umbral de soporte. Cuando hay un nodo con menos de N casos, también se puede detener el proceso de construcción del árbol, ya que no se considera fiable una clasificación avalada por menos de N casos (menos de N casos se consideran insuficientes para estimar probabilidades adecuadamente).

Las reglas de parada también se conocen como reglas de pre-poda porque reducen la complejidad del árbol durante su construcción. Son establecidas a priori por el propio investigador en función de investigaciones anteriores, análisis previos o incluso su propia experiencia (Pérez, 2007).

➢ Métodos de poda.

En general, el método recursivo de construcción de árboles divide el conjunto de casos hasta que se encuentra un nodo puro o no se puede seguir ramificando el árbol, pudiendo producirse un sobreajuste de los datos. La poda permite realizar una simplificación del árbol (que permite además mejorar tanto la precisión del método como la capacidad predictiva) y evita el sobreajuste de los datos. Los métodos de poda dependen del algoritmo empleado en la construcción de árbol, siendo los métodos más comunes:

- Poda por estimación del error. Utilizada en Quinlan (1987).

- Poda por coste-complejidad. Utilizada en Breiman et al. (1984).

- Poda pesimista. Utilizada en Quinlan (1993). Este tipo de poda no describirá, ya que no es el motivo fundamental de este trabajo de investigación.

Los métodos de poda por estimación del error y por coste-complejidad serán explicados en el capítulo de materiales y métodos. La poda pesimista no se describirá, ya que no es motivo fundamental de este trabajo de investigación.

2.5.2. Algoritmos de construcción de ADDs.

Existen numerosos algoritmos que permiten la construcción de ADDs. Dentro de la familia de los algoritmos TDIDT, los más conocidos son: el algoritmo CHAID (Chi-squared Automatic Interaction Detection) implementado por Kass (1980); el método CART (*Classification and Regression Trees*) desarrollado por Breiman et al. (1984); el algoritmo ID3 (Quinlan, 1986) y sus posteriores evoluciones C4.5 (Quinlan, 1993) y C5.0 (Quinlan, 1997). Las principales diferencias entre los mismos radican en el criterio adoptado para realizar las particiones, en el tipo de variables que pueden manejar, así como en las restricciones que se imponen en el número de ramas en las que se puede dividir cada nodo, o el criterio de poda adoptado, entre otras.

CHAID: Es un método exploratorio de análisis de datos útil para identificar variables importantes y sus interacciones con fines de segmentación, análisis descriptivo o como paso previo a otros análisis posteriores (ver Pérez, 2007). Se caracteriza porque permite la construcción de árboles no binarios (con más de 2 ramas por cada nodo). Se puede utilizar en tareas de clasificación (cuando la variable clase es categórica) y en este caso el criterio de partición está basado en test de Chi-cuadrado; y para tareas de regresión (cuando la variable clase es continua), y en este caso el criterio de partición está basado en el test de la F de Snedecor. No presenta post-poda.

CART: Constituye una alternativa al CHAID y, de hecho, fue desarrollado por Breiman et al. (1984) para superar algunas de las limitaciones de este algoritmo. CART sólo permite la construcción de árboles binarios. Se puede utilizar en tareas de clasificación el criterio de partición en este caso está basado en el Índice de Gini; y en tareas de regresión, siendo el criterio de partición la reducción del error cuadrático o la desviación media absoluta de la mediana. Permite realizar post-poda por coste-complejidad.

ID3: El algoritmo ID3 es un algoritmo simple pero potente, que realiza la construcción del árbol de manera similar al método CART, con la diferencia de que desde un nodo surgen tantas ramas como valores posibles de dicho atributo (no existe restricción binaria). Sólo puede trabajar con atributos categóricos, por lo que solo puede utilizarse en tareas de clasificación, siendo el criterio de partición la Ganancia de Información. No presenta post-poda.

C4.5: El algoritmo C4.5 es una extensión del ID3 que surge para resolver algunas de sus limitaciones. En este sentido permite trabajar con atributos continuos; el criterio de partición está basado en la razón de Ganancia de Información; y permite realizar post-poda (poda pesimista).

En el capítulo 3 de metodología se explicarán con detalle las características específicas de los algoritmos CART, ID3 y C4.5, utilizados como herramienta para el desarrollo de esta tesis doctoral. El algoritmo CART se ha seleccionado puesto que hasta el momento ha sido el utilizado en el campo de la seguridad vial. Sin embargo, presenta una restricción binaria que puede ser una limitación cuando se pretende analizar la influencia de una categoría específica de una variable en la gravedad del accidente. Se

han seleccionado además los algoritmos ID3 y C4.5 dado que no presentan esta restricción y en la literatura de MD son los más utilizados.

2.5.3. Reglas de Decisión obtenidas de ADDs.

Una de las principales características de los ADDs es que su estructura puede transformarse en conocimiento "directo" en forma de Reglas de Decisión (RDs). Las RDs que se obtienen de los ADDs se caracterizan porque el consecuente es un caso concreto de la variable de estudio; por tanto son del tipo "SÍ (condición)→ ENTONCES (clase)". Las RDs así obtenidas equivalen completamente al árbol original. Este hecho facilita la compresión del modelo, y es aún más importante cuando los ADDs son grandes, ya que al aumentar el tamaño disminuye su inteligibilidad.

En un árbol, cada regla se obtiene siguiendo el camino desde el nodo raíz a cada nodo hoja del árbol. De modo que desde el nodo raíz se deriva un conjunto de reglas cuyo antecedente es una conjunción de literales relativos a los valores de los atributos situados en los nodos intermedios del árbol, y cuyo consecuente es la decisión a la que hace referencia la hoja del árbol (la clasificación realizada).

Las RDs son potencialmente útiles para los gestores de seguridad vial ya que permiten identificar determinados patrones de comportamiento de un modo muy fácilmente comprensible.

2.5.4. Ventajas y desventajas de los ADDs.

Las principales ventajas de los ADDs son las siguientes:

- Son auto-explicativos, y cuando su tamaño no es muy grande, son muy fáciles de interpretar (incluso para no expertos en la materia).
- Permiten la extracción del conocimiento de un modo comprensible, en forma de Reglas de Decisión del tipo "SÍ-ENTONCES". Y estas reglas pueden ser usadas para descubrir determinados patrones de comportamiento que ocurren dentro de un conjunto de datos para una variable objeto de estudio.
- Los ADDs son normalmente un método no paramétrico por lo que no establecen hipótesis sobre su distribución espacial ni de la estructura del clasificador.
- Pueden trabajar con un gran número de variables predictivas sin problemas de multicolinealidad.
- Son flexibles para trabajar con atributos tanto numéricos como nominales.
- Adaptabilidad en el pre-procesamiento de la base de datos, la cual puede contener errores o valores perdidos.
- Alto rendimiento predictivo con relativo esfuerzo computacional.

- Pueden trabajar con grandes bases de datos, y descubrir fácilmente complejas interacciones entre los datos.

Y las principales desventajas son:

- El exceso de sensibilidad sobre el conjunto de entrenamiento.
- La presencia de atributos irrelevantes y el ruido (Quinlan, 1993), que pueden provocar que el modelo aprendido sea especialmente inestable, de modo que un pequeño cambio en los datos puede cambiar toda la estructura del árbol.
- Los ADDs no permiten realizar análisis de elasticidades (o de sensibilidades), los cuales permiten examinar los efectos marginales de las variables dependientes, sobre la variable objeto de estudio (Chang and Wang, 2006).
- Las RDs que se pueden obtener de un sólo ADD vienen muy condicionadas por la variable que se usa como raíz, dando lugar a pérdida de otra/s informaciones importantes sobre la variable en estudio.

2.6. Conclusiones.

La principal ventaja de la MD con respecto a los DOM es que permite la extracción conocimiento de los datos (previamente desconocido e indistinguible), y que no requiere un previo conocimiento probabilístico del problema objeto de estudio.

Además, los MDE parten de hipótesis fijas y predefinen relaciones entre las variables dependientes e independientes, de modo que si estas hipótesis no se cumplen los modelos pueden producir estimaciones erróneas en la probabilidad de la gravedad de la lesión (Chang and Wang, 2006). Por ello, recientemente, muchos investigadores han comenzado a utilizar las técnicas de MD para analizar los accidentes de tráfico, y en particular, la gravedad de los mismos.

Dentro de los modelos de MD se distinguen modelos predictivos y descriptivos. Los predictivos, se caracterizan porque realizan el aprendizaje mediante un procedimiento supervisado. De este modo, la técnica supervisa en el modelo en construcción el grado de ajuste a la realidad conocida. Y en este sentido, estos modelos pretenden estimar valores futuros o desconocidos de una variable respuesta. Dentro de las técnicas no supervisadas, las RNA, las RB y los ADDs son las más utilizadas en el campo de la seguridad vial.

Mientras que los modelos descriptivos se caracterizan porque realizan el aprendizaje mediante un procedimiento no supervisado. Su objetivo es identificar patrones en los datos sin identificadores externos que guíen al algoritmo, es decir, sin conocimiento previo de la realidad. Y en este sentido, los modelos descriptivos sirven para explorar las propiedades de los datos a examinar. Dentro de las técnicas supervisadas, las RA son la técnica más utilizada en el campo de la seguridad vial.

Las principales características de las RNA radican en el aprendizaje secuencial, en el hecho de utilizar transformaciones de las variables originales para la predicción y que pueden existir interacciones no lineales entre variables. Así, permiten aprender en contextos difíciles, sin precisar la formulación de un modelo concreto. Sin embargo, su principal inconveniente es que para el usuario son una "caja negra" (no se revela normalmente la naturaleza de las relaciones entre las variables independientes y dependientes del modelo). Además el modelo aprendido es difícilmente comprensible, y requieren gran cantidad de datos para su entrenamiento.

Respecto de las RBs se puede destacar su buen rendimiento en clasificación, su estructura en forma de grafo y la capacidad para hacer predicción. Sin embargo presentan problemas con las variables ocultas y puede haber problemas de inconsistencia en las probabilidades. Además requieren un gran número de datos para desarrollar el modelo (De Oña et al., 2012).

Respecto a las RA, su principal característica es que permiten generar patrones que son muy fácilmente comprensibles. Y como limitación destaca la posible aparición de patrones debidos al azar, y es lo que conoce como error Tipo I (Web, 2007).

Para realizar el análisis de los accidentes de tráfico recogidos en este estudio se utilizará la técnica de ADDs, ya que, teniendo en cuenta sus ventajas y limitaciones, constituyen uno de los modelos más utilizados en aprendizaje supervisado y en aplicaciones de Minería de Datos (Gehrke et al., 1999). Además resulta ser un método muy versátil, que muy frecuentemente es utilizado en investigaciones de muy diversa índole, desde aplicaciones de diagnóstico médico, sistemas de predicción meteorológica, análisis de riesgo, hasta en el propio campo de la seguridad vial.

Los ADDs resultan muy apropiados para el estudio de los accidentes ya que normalmente no utiliza parámetros y no suelen suponer ninguna relación previa entre las variables. Son muy fácilmente interpretables. Permiten la extracción de reglas de decisión, que pueden ser usadas para descubrir determinados patrones de comportamiento que ocurren dentro de un conjunto de datos. Y estos patrones pueden ayudar a la comprensión del suceso de un accidente.

Al igual que los ADDs, las RNA también pueden ser utilizadas para la búsqueda de patrones presentes en los datos (Bayam et al., 2005). Sin embargo, los ADDs proporcionan decisiones más comprensibles y explicables que las RNA, por lo que resultan modelos más prácticos para ser utilizados por los agentes policiales (Pande and Abdel-Aty, 2009), así como para su posible implementación como herramienta de análisis.

Del mismo modo que las RA, los ADDs pueden interpretarse mediante un conjunto de reglas que proporcionan resultados específicos y fáciles de comprender, que describen las relaciones entre los atributos del accidente (Pande and Abdel-Aty, 2009). Sin embardo, las reglas obtenidas por un modelo de árbol tienen poder descriptivo y predictivo desde el momento que se evalúa su precisión con los datos del conjunto de

test. Mientras que esta particularidad no queda recogida en los patrones obtenidos con RA.

Los ADDs pueden ser construidos con diferentes algoritmos. Sin embargo, el método CART ha sido el más ampliamente utilizado en las investigaciones de seguridad vial (Chang and Wang, 2006; Kashani and Mohaymany, 2011; Kashani et al., 2011; Montella et al, 2011; Montella et al., 2012b). Dado que CART solo permite la construcción de árboles binarios, su interpretación puede ser ineficiente en determinadas ocasiones (Breiman et al., 1984).

Sin embargo, existen en la literatura otros algoritmos que permiten la construcción de ADDs y no presentan esta restricción binaria (ID3 y C4.5). En el caso de los accidentes de tráfico, estos métodos pueden resultar muy apropiados cuando se pretende analizar la influencia de una categoría específica de una variable en la gravedad del accidente.

Las RDs (que se extraen de los ADDs) permiten identificar determinados patrones de comportamiento de un modo fácilmente comprensible; lo cual permite a los gestores de seguridad vial, y a las Administraciones competentes que puedan establecer determinadas contramedidas.

La principal limitación de las RDs que se extraen de los ADDs es que son dependientes de la estructura del árbol, de modo que pueden existir ciertos patrones de accidentes que, a priori, no sean detectados. Por lo que, a priori, no se estaría obteniendo todo el conocimiento posible de la base de datos de accidentes analizada, y no se estaría actuando sobre la totalidad de la problemática analizada. Por ello, en esta tesis doctoral se propone utilizar un nuevo método de extracción de RDs que ayudaría a resolver esta limitación.

CAPÍTULO 3.
OBJETIVOS

CAPÍTULO 3. OBEJTIVOS

En esta tesis doctoral se aplica la herramienta de Árboles de Decisión (ADDs) en el campo de la seguridad vial, dada su capacidad para el estudio de problemas complejos, tal y como son los accidentes de tráfico. Teniendo en cuenta que muchos estudios previos han estudiado la gravedad de los accidentes mediante el uso de diferentes técnicas estadísticas, la aplicación de ADDs en este ámbito resulta todavía novedosa y proporciona una nueva visión para resolver algunas de las limitaciones de los métodos estadísticos más utilizados.

3.1. Objetivo principal.

El principal objetivo de esta tesis doctoral es identificar patrones de accidentes que ocurren dentro de las carreteras convencionales, que sean fácilmente comprensibles por los gestores de seguridad vial, y sobre los que las Administraciones competentes puedan realizar actuaciones concretas (en forma de actuaciones específicas en tramos de carreteras o con diseño de programas de educación y campañas de concienciación en materia de seguridad vial); con el fin de mejorar la seguridad vial de estas carreteras. La identificación de estos patrones se hace mediante el uso de Reglas de Decisión extraídas de Árboles de Decisión.

3.2. Objetivos específicos.

De este objetivo principal se derivan los objetivos específicos que se presentan a continuación

- En el campo de la seguridad vial, la metodología CART ha sido la aplicada hasta el momento para el estudio de los accidentes de tráfico, sin embargo, esta metodología sólo permite la construcción de árboles binarios. Por ello, el **primer objetivo específico** es validar el uso de diferentes algoritmos para la construcción de ADDs destindos al estudio de la gravedad de los accidentes de tráfico.

- Los ADDs permiten identificar las variables más importantes del modelo. Por ello, como **segundo objetivo específico** se propone la identificación de las variables clave que afectan a la gravedad de los accidentes.

- Los ADDs permiten la extracción de Reglas de Decisión, que son una herramienta clave de cara a la interpretación de los resultados y potencialmente útil de cara a la identificación de patrones de comportamiento. De este modo, el **tercer objetivo**

específico es la validación de las Reglas de Decisión para la identificación de patrones de accidentes.

- La principal limitación de las Reglas de Decisión que se extraen de los ADDs es que son dependientes de la estructura del árbol, de modo que pueden existir ciertos patrones específicos que no sean detectados. Como **cuarto objetivo específico** se propone el uso de un nuevo método, propuesto en esta tesis doctoral, que resuelve esta limitación y permite la extracción completa del conocimiento existente en la base de datos objeto de estudio. La principal característica de esta metodología es que se extraen reglas de ADDs construidos variando el nodo raíz que genera su construcción.

- La mayoría de las reglas que se obtienen cuando se aplica el nuevo método propuesto en esta investigación provienen de ADDs dónde el nodo raíz se impone, por lo que el **quinto objetivo específico** es la validación del nodo raíz en los patrones obtenidos con esta metodología.

- Dado que la aplicación de contramedidas conlleva un coste, y los recursos disponibles suelen ser limitados, los patrones obtenidos deben representar fielmente las problemáticas de la vía. Por tanto como **sexto objetivo específico se plantea una** validación completa de los patrones obtenidos.

CAPÍTULO 4.

MATERIALES Y MÉTODOS

CAPÍTULO 4.
MATERIALES Y MÉTODOS

4.1. Introducción.

En este capítulo se exponen las diferentes fases del trabajo de investigación llevado a cabo en esta tesis doctoral. Posteriormente, se realiza una descripción de la metodología utilizada, así como una descripción tanto del tratamiento realizado sobre los datos de estudio como de las principales características que los definen.

4.2. Fases del trabajo de investigación.

En el presente trabajo de investigación se han desarrollado dos experimentaciones diferentes para alcanzar los objetivos propuestos.

- **Fase preliminar**

- Creación de una base de datos de estudio.
- Tratamiento de datos para la aplicación de modelos.
- Las características del los datos de estudio así como del proceso de creación de la base de datos, y el tratamiento de los mismos son explicados en el epígrafe 4.4.

- **Extracción de Reglas de Decisión mediante el uso de ADDs.**

- Construcción de ADDs mediante 3 algoritmos diferentes: CART, ID3 y C4.5 (epígrafe 4.3.1).
- Evaluación y comparación de los 3 modelos generados utilizando los indicadores descritos en el epígrafe 4.3.2 y 4.3.3.
- De los modelos que mejores resultados proporcionan, se identificaron las variables con mayor efecto clave en la severidad del accidente (epígrafe 4.3.4).
- Se extraen y se evalúan las Reglas de Decisión (RDs) con los criterios definidos en el epígrafe 4.3.5.

- **Exposición del método propuesto, *Information Root Node Variation* (IRNV), y extracción de RDs con esta metodo.**

Las RDs extraídas de un ADD dependen de la propia estructura del árbol. De modo que la extracción de conocimiento (en forma de RDs) sólo se realiza en el sentido

dictado desde el nodo raíz hasta cada uno de los nodos terminales del árbol. Sin embargo, es posible que existan otras reglas importantes que no sean detectadas por la configuración del árbol, que depende del nodo raíz que inicia su construcción.

Por ello en la siguiente fase de la investigación se expone el nuevo método propuesto en esta tesis doctoral. Es un método que permite la extracción de reglas a partir de diferentes ADDs, obtenidos mediante variación del nodo raíz que genera su construcción. El nombre utilizado para hacer referencia a este método es *Information Root Node Variation method* (IRNV). Otra de sus principales características, es que permite utilizar diferentes criterios de partición. Las fases en las que se divide esta experimentación serán las siguientes:

- Aplicación de IRNV con el criterio de partición basado en el Índice de Gini (GI).
- Aplicación de IRNV con el criterio de partición basado en el Ratio de Ganancia de Información (RGI) utilizado en C4.5.
- Creación de un conjunto global de Reglas de Decisión.
- Validación de las reglas.
- Extracción del conjunto final de Reglas de Decisión.

4.3. Metodología de la investigación.

En este punto se describe con detalle la metodología utilizada en esta tesis doctoral. Se describen los diferentes algoritmos utilizados para la construcción de ADDs (epígrafe 4.3.1). A continuación se explica el proceso utilizado para validar los modelos construidos (epígrafe 4.3.2) y los criterios utilizados para evaluar los modelos (epígrafe 4.3.3). Una vez que se tienen los modelos, en el epígrafe 4.3.4, se explica cómo se obtiene la importancia de las variables en los modelos; y el proceso para la obtención de las RDs (epígrafe 4.3.5). En el epígrafe 4.3.6 se describe el método IRNV y en el epígrafe 4.3.7 cómo se obtiene el conjunto final de RDs, del cual se analizarán los patrones en el capítulo de resultados.

4.3.1. Métodos para la construcción de ADDs.

Existe mucha información en la literatura sobre los procedimientos de construcción de un ADD, sin embargo, todos tienen en común las siguientes particularidades:

- Los criterios utilizados para la selección del atributo que se coloca en un nodo y que produce la ramificación. Este criterio es conocido como criterio de partición.
- Los criterios para detener la ramificación del árbol. Este criterio es conocido como criterio de parada.

- El método para asignar una clase o una distribución de probabilidad en los nodos hoja.
- El proceso de poda (pre-poda o post-poda), que simplifica la estructura del árbol y evita el sobreajuste de los mismos.

A continuación se describen con detalle las características específicas de los algoritmos CART, ID3 y C4.5.

4.3.1.1. CART.

El método CART fue desarrollado por Breiman et al. (1984). Es un método no-paramétrico que genera árboles binarios. En función de la naturaleza de la variable dependiente o variable clase se desarrolla un árbol de clasificación o un árbol de regresión, si la variable clase es discreta o si la variable clase es continua respectivamente.

Dado que la variable clase objeto de este estudio es una variable discreta (la gravedad del accidente) solo serán explicados en este epígrafe los árboles de clasificación.

El desarrollo de un ADD mediante el método CART consiste generalmente en tres pasos:

- **Construcción del árbol máximo.**

La construcción del árbol se realiza de modo recursivo, siguiendo una estrategia descendente, con el objetivo de maximizar la pureza de los nodos (que se alcanza cuando todos los casos que contiene el nodo son de la misma clase). El proceso se inicia en el nodo raíz, el cual se obtiene a partir del conjunto completo de datos. Utilizando el criterio de partición basado en el Índice de Gini (ecuación 8), el conjunto total de datos se divide en 2 subconjuntos mutuamente excluyentes, formando así 2 nodos intermedios del árbol. La variable que se utiliza para realizar la partición es aquella que crea la mejor homogeneidad en los dos nodos secundarios creados (también llamados nodos hijos). De hecho, los datos en cada nodo hijo son más homogéneos que los del nodo padre. Recursivamente se va dividiendo cada uno de los subconjuntos creados hasta que todos ellos son puros o hasta que su "pureza" no puede incrementarse; formándose así los nodos terminales del árbol o nodos hoja.

El método CART crea árboles binarios. Así, cuando la variable que realiza la partición es de naturaleza categórica, sus diferentes categorías se combinan en las dos asociaciones que crean la mayor pureza en los dos nodos secundarios. Y cuando la variable es de naturaleza continua, los datos se dividen de acuerdo a un valor umbral, que genera también la mejor homogeneidad en los dos nodos secundarios creados.

Las reglas de partición de un nodo dependen exclusivamente de las variables (o atributos). De modo que, durante el crecimiento del árbol, se crea un conjunto de reglas de partición candidatas, que consta de todas las divisiones posibles para todas

las variables incluidas en el análisis, y se obtiene de forma diferente, dependiendo de la naturaleza de los atributos (discreta o continua):

- Para el caso de atributos discretos, suponiendo que el atributo tiene N categorías (C1, C2, ..., CN), el conjunto de las posibles divisiones será: $2^{N-1}-1$
- Para el caso de atributos continuos, el número de divisiones posibles en un nodo dado es uno menos que el número de sus valores distintos observados: N-1.

En definitiva, en ambos casos las reglas son una cantidad finita y están perfectamente determinadas. Estas divisiones son evaluadas y clasificadas usando un criterio diferente para un árbol de clasificación o para un árbol de regresión.

El criterio de partición en los árboles de clasificación está basado en el Índice de diversidad de Gini, mientras que de los árboles de regresión se basa en la reducción del error cuadrático.

El índice de diversidad de Gini es una medida de la diversidad de clases en un nodo del árbol, que trata de minimizar la impureza en los subconjuntos de casos generados al ramificar el árbol. Para una variable C, el Índice de Gini es definido como:

$$gini(C) = 1 - \sum_j p^2\,(C = c_j) \tag{8}$$

De este modo el criterio de partición basado en el Índice de Gini se define como:

$$\mathrm{GI}x(C, X) = gini(C|X) - gini(X), \tag{9}$$

donde $gini\,(C|X) = \sum_t p(\mathrm{X} = x_t) gini\,(C|X = x_t)$ y X otra variable conocida. Siendo la mejor variable para usar en la partición aquella que minimiza $\mathrm{GI}x(C, X)$.

Siguiendo este procedimiento se crea el árbol de mayor tamaño, que es un modelo sobreajustado a los datos. El problema de un modelo sobreajustado es que no ayuda a clasificar con precisión otro conjunto de datos, y más aún si los datos utilizados para crear el modelo contienen ruido (errores en los atributos o en las clases), puesto que el modelo creado intentará ajustarse a los errores, perjudicando el comportamiento global del modelo aprendido.

Para desarrollar un modelo CART, normalmente, los datos se suelen dividir en dos subconjuntos, uno de aprendizaje o entrenamientro (*training*) y otro de validación (*test*). El conjunto del *training* se utiliza para crear el modelo, mientras que el *test* se utiliza para chequear la bondad del modelo. De este modo, a partir de los datos de aprendizaje se construye el árbol máximo, que es un modelo sobreajustado a los datos.

➢ Poda del árbol.

Con el método CART lo primero es construir el árbol máximo con la única condición de no permitir nodos con muy pocos elementos (Breiman et al., 1984), para posteriormente aplicar un proceso de poda. El objetivo de la poda es reducir la

complejidad del árbol máximo y evitar un excesivo sobreajuste a los datos de aprendizaje

De modo general, a partir del árbol máximo se podan (o eliminan) aquellas ramas o sub-árboles que determinen beneficios muy pequeños, en lo que respecta a la disminución de la impureza. Con este procedimiento se obtiene un sub-árbol que permite, para determinados nodos, que una de sus ramas permanezca y la otra se pode. La construcción del árbol óptimo, se realiza a partir de un proceso de selección de sub-árboles, en el que interviene de manera fundamental el error asociado a cada uno de ellos.

Para los árboles de clasificación, el método CART utiliza un método de poda basado en el algoritmo de coste-complejidad. La poda por coste-complejidad se basa en una medida que combina los criterios de precisión frente a la complejidad en el número de nodos y velocidad de procesamiento, buscando el árbol que obtiene el menor valor para este parámetro.

La complejidad del árbol viene dada por el número de nodos terminales (u hojas) que posee. Si T es el ADD usado para clasificar N casos del conjunto entrenamiento y se clasifican mal M casos, la medida de coste-complejidad del T para un parámetro de complejidad α es:

$$R_\alpha(T) = R(T) + \alpha l(T), \tag{10}$$

dónde $l(T)$ es la complejidad del árbol, el número de hojas del árbol T; α es un parámetro de complejidad; y $R(T) = M/N$ es un estimador del error de clasificación en el árbol T, también llamada tasa de mala clasificación.

Por tanto, $R_\alpha(T)$ es una combinación lineal del error o coste del árbol y de su complejidad. El parámetro α es un número real mayor o igual a 0, cuyo valor óptimo a priori es desconocido. A la hora de implementar la técnica de poda, a partir del árbol máximo T se genera una secuencia finita de árboles podados, decrecientes en cantidad de hojas $\{T_i, T_{i-1}, \ldots, T_1 = T\}$. Valores altos de α penalizan árboles con muchas hojas. Si por ejemplo $\alpha \to \infty$, el tamaño del árbol tiende a ser una sola hoja (asociado con el mayor coste R(T)). Y para $\alpha = 0$ no se tiene en cuenta el tamaño del árbol, y se obtiene el de menor coste R(T), es decir, el árbol máximo T.

Hay que señalar que conforme el parámetro de complejidad α crece, el tamaño del árbol que minimiza $R_\alpha(T)$ decrece; de modo que el objetivo es encontrar el óptimo entre el coste y la complejidad. Por tanto el árbol podado es aquel que hace mínima la medida de coste-complejidad $R_\alpha(T)$.

➢ Selección del árbol óptimo.

En el último paso se selecciona el árbol óptimo. En la Figura 8 se muestra gráficamente la elección de este árbol. De toda la secuencia de árboles creados se escoge aquel que tiene asociado el menor error, utilizando para estimar el error el conjunto de *Test* o el procedimiento de validación cruzada, como proponen Breiman et al. (1984).

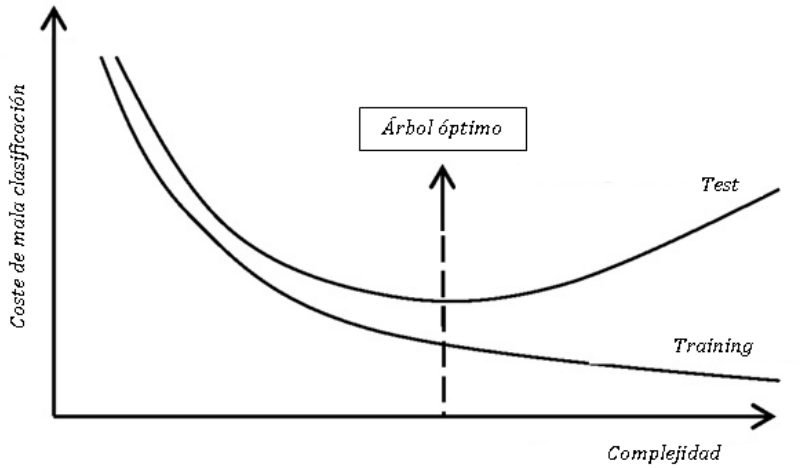


Figura 8.-Selección del árbol óptimo.

De modo que para seleccionar el árbol óptimo, el objetivo es encontrar la proporción óptima entre la tasa de mala clasificación y la complejidad del árbol, según se observa en la Figura 8.

4.3.1.2. ID3.

El algoritmo ID3, desarrollado por Quinlan (1986), es un algoritmo simple pero potente. Realiza la construcción del árbol de manera similar al método CART, pero sin la restricción binaria en el número de ramas, es decir, el conjunto de datos asociado a un nodo puede dividirse en tantas particiones como ramas posibles, y esta división depende de la naturaleza del atributo:

- Atributos numéricos: el número de bifurcaciones que genera un nodo que realiza una comparación con un atributo numérico es igual a dos. La división se realiza en función de un valor umbral: menores que el umbral vs. mayores que el umbral. Para determinar ese valor umbral, es necesario ordenar las instancias por el valor del atributo. El algoritmo ID3 está optimizado de modo que sólo hacen esa ordenación en el nodo raíz, manteniendo una estructura de datos que permite aprovechar esa reordenación a medida que se van construyendo las ramas.

- Atributos nominales: el número de bifurcaciones que genera un nodo que realiza una comparación con un atributo nominal es igual al número de posibles valores que pueda tomar ese atributo.

Al igual que CART, la construcción del árbol se realiza de modo recursivo, siguiendo una estrategia descendente, con el objetivo de maximizar la pureza de los nodos. El algoritmo ID3 utiliza para medir la impureza del nodo, el concepto de entropía de la información o entropía de Shannon (1948).

Utilizando la definición de Shannon (1948), sea P una distribución de probabilidad con $P = \{p_1, p_2, \ldots, p_n\}$, tal que $pi = prob(X = x_i);\ p_i \geq 0\ y\ \sum_{i=1}^{n} p_i = 1$. Se llama entropía b-aria de la distribución P a:

$$H_b(p_1, \dots, p_n) = -\sum_{i=1}^{n} p_i \, log_b p_i, \quad (11)$$

si b=2, la entropía se define como:

$$H_2(p, 1-p) = p log_2 \frac{1}{p} + (1-p) log_2 \frac{1}{1-p}, \quad (12)$$

Y en este caso, se obtiene la función de entropía H(p). La entropía que es una medida utilizada comúnmente en la teoría de la información, que caracteriza la pureza (o impureza) de un conjunto de datos.

El criterio de partición utilizado por el algoritmo ID3, denominado ganancia de información (*Information Gain*) está basado en la entropía. La idea es medir la ganancia de información asociada al elegir particionar por un determinado atributo (o variable); que se define como la diferencia de entropía del nodo actual y la suma ponderada de las entropías correspondientes a bifurcar por ese atributo. Por tanto, el criterio de partición para una variable atributo X dada una variable clase C viene dado por la siguiente expresión:

$$Ganancia\ de\ Información\ (C, X) = IG(C, X) = H(C) - H(C|X), \quad (13)$$

donde $H(C)$ es la entropía de C, dada por:

$$H(C) = -\sum_j p(c_j) \, log \, p(c_j), \quad (14)$$

con $p(c_j) = p(C = c_j)$, la probabilidad de cada valor de la variable clase estimada en el conjunto de *training*.

Del mismo modo, $H(C|X)$ es la entropía de clasificación del conjunto de casos del atributo X, y viene dada por:

$$H(C|X) = -\sum_t \sum_j p(c_j|x_t) \, log p(c_j|x_t), \quad (15)$$

donde x_t, t=1,...,|X|, es cada posible estado de la variable atributo, X y c_j, j=1,...,k es cada posible estado de la variable clase C.

El algoritmo ID3 tiene cierta preferencia implícita a bifurcar los atributos nominales con muchas categorías, produciendo árboles que desprecian de forma prematura el resto de atributos ya que llegan muy rápidamente a ramas con pocos casos. Otra de las características de este algoritmo es que no presenta ningún proceso de post-poda. Una descripción más detallada de la metodología ID3 puede encontrarse en Quinlan (1986).

4.3.1.3. C4.5.

El algoritmo C4.5, desarrollado por Quilan (1993), es una extensión del ID3 que surge para resolver algunas de sus limitaciones:

- Como criterio de partición utiliza la razón de ganancia de información (*Gain Ratio*). La maximización de la ganancia de información, utilizada en el algoritmo ID3,

puede dar lugar a errores en atributos nominales con muchas categorías, ya que se crean ramas para cada una de ellas recogiendo así un menor número de casos (que si por ejemplo, la bifurcación fuese binaria), y por tanto es más fácil que esas ramas tiendan a ser puras. El algoritmo C4.5 introduce una mejora para evitar este efecto, utilizando como criterio la razón de ganancia en lugar de la ganancia de información. La razón de ganancia de información para una variable atributo X y una variable clase C se define como:

$$IGR(C,X) = \frac{IG(C,X)}{H(X)} \quad (16)$$

Cuando la división realizada del conjunto de casos de entrenamiento es trivial, el denominador de $IGR(C,X)$ es cercano a cero. Por tanto, se ha de escoger el atributo que maximice el cociente $IGR(C,X)$, siendo su ganancia al menos, tan grande como la ganancia media de todas las alternativas analizadas.

- Permite trabajar con atributos continuos y con valores perdidos
- Presenta un método de post-poda del tipo pesimista. El algoritmo C4.5 incorpora una poda del árbol de clasificación, una vez que este ha sido inducido, que se basa en la aplicación de un test de hipótesis si expandir o no una determinada rama.

  El algoritmo C4.5 considera dos operaciones de poda (esquematizadas en la Figura 9):

  - Sustituir el sub-árbol por una hoja hija (*Subtree replacement*). En la Figura 9 se observa cómo se realiza esta poda: el sub-árbol que se crea a partir del nodo C del árbol 1 es remplazado en el árbol 2 por el nodo 1'.
  - Elevar el sub-árbol (*Subtree raising*), sustituyéndolo por una rama completa correspondiente a uno de sus nodos hijos. En la Figura 9 se observa cómo se realiza esta poda: el sub-árbol se crea a partir del nodo B del árbol 1, que es remplazado en el árbol 3 por el nodo 1', 2' y 3'.

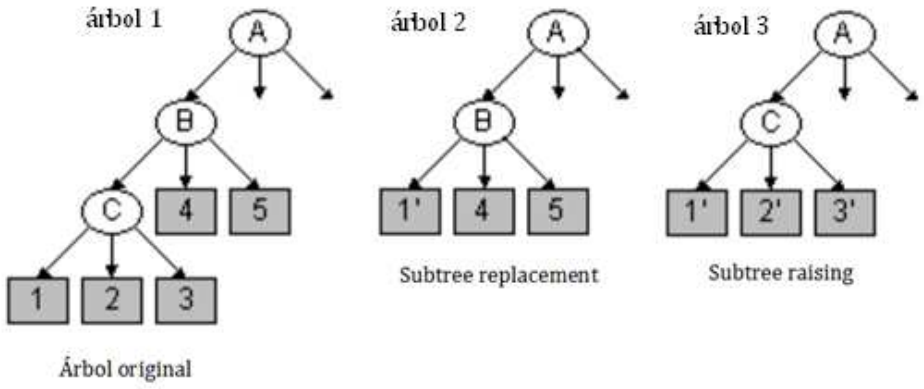


Figura 9.- Tipos de operaciones de poda en C4.5.

El proceso de poda comienza en los nodos hoja y recursivamente continúa hasta llegar al nodo raíz. La cuestión es decidir reemplazar un nodo interno por una hoja (*replacement*) o reemplazar un nodo interno por uno de sus nodos hijo (*raising*). Para ello se compara el error estimado de clasificación del árbol antes y después de una operación de poda, de modo que sólo se lleva a cabo cuando esta diferencia resulta favorable a la poda. Naturalmente, el problema está en que el árbol sin podar tendrá un error de entrenamiento cero o muy próximo a cero. Por tanto, sería muy optimista asumir que el error antes de la poda es el error de entrenamiento sin más. La solución más sencilla consistiría en retirar un pequeño número de instancias del conjunto de entrenamiento antes de construir el árbol, y hacer la estimación de los errores con dicho conjunto. Sin embargo, esta aproximación plantea un problema para aquellos conjuntos de datos que cuenten con pocas instancias de entrenamiento.

El algoritmo C4.5 evita este problema utilizando todas las instancias del conjunto de *training* de un sub-árbol para ver si un nodo se puede eliminar o no. Y compensa el efecto optimista de utilizar el conjunto de *training* haciendo una estimación pesimista del error.

Cuando una hoja del árbol cubre N casos del *training*, de los cuales E casos son clasificados incorrectamente, su error de resustitución es E/N. El estimador del error de resustitución asociado a un subárbol será la suma de los errores estimados para cada una de sus ramas.

La probabilidad real del error cometido en un nodo del árbol no se puede determinar con exactitud, y menos aún a partir del conjunto de *training*, que se emplea para construir el ADD. Sin embargo, se puede considerar que los E errores de un nodo corresponden a E "éxitos" en N experimentos aleatorios, por lo que, de forma heurística, se le asocia al nodo del árbol una distribución de probabilidad binomial. Dado un grado de confianza (CF), se puede establecer un intervalo de confianza para el valor de una distribución binomial y se puede considerar el límite superior de este intervalo U_{CF} (E, N) como una estimación del error en el nodo. Si bien esta estimación carece de una base sólida, se emplea para predecir el número de errores de un nodo del árbol: N x U_{CF} (E, N).

Al utilizar poda pesimista, se poda un subárbol si el intervalo de confianza del error de resustitución (generalmente de amplitud dos veces el error estándar) incluye el error de resustitución del nodo si se trata como hoja. De esta forma se eliminan los subárboles que no mejoran significativamente la precisión del clasificador. El método es tan cuestionable como cualquier otra heurística sin base teórica, pero suele producir resultados aceptables.

4.3.2. Validación del ADD.

Tradicionalmente los métodos clasificación han utilizado una parte del conjunto total de los datos (conjunto de training) para construir el modelo y otra parte (conjunto de test) para medir el ratio de error del modelo construido (o la precisión del modelo).

Normalmente se establece que el *training* debe estar formado por 2/3 de los datos y el *test* por 1/3. Además, los conjuntos de datos (*training* y *test*) deben ser elegidos al azar y ser representativos del conjunto total de datos, de modo que la proporción de datos que componen cada categoría de la variable clase en el conjunto total de datos deben estar representadas con aproximadamente la misma proporción en los subconjuntos creados.

En los estudios de seguridad vial analizados se ha utilizado esta técnica para validar los modelos, estableciendo unos valores del 70% de los datos para formar el conjunto de training y el 30% restante para el conjunto de test (Chang and Wang, 2005; Chang and Wang, 2006; Kashani and Mohaymany, 2011; Montella, 2011; Montella et al., 2011; Chang and Chien, 2013).

Sin embargo existe otra técnica de validación del modelo llamada *k-fold cross validation,* que no requiere dividir la base de datos en dos conjuntos diferentes, sino que permite aprovechar el conjunto total de los mismos sin prescindir de una parte de los registros. Esta técnica ofrece resultados más fiables y asegura que los conjuntos de datos sean representativos (Lewis, 2000).

Este método divide de forma aleatoria la muestra utilizada en la fase de aprendizaje en k conjuntos (*k-fold cross validation*). Secuencialmente, cada uno de estos subconjuntos se reserva para emplearse como conjunto de *test* frente al modelo de árbol generado por los k-1 subconjuntos restantes. Se obtienen así k modelos diferentes, donde se puede evaluar la precisión de las clasificaciones tanto en el conjunto de aprendizaje (k-1) como en los subconjuntos de prueba (k). Normalmente, la precisión del modelo se expresa como la media obtenida en los k conjuntos de prueba.

En la Figura 10 (adaptada de Lewis, 2000) se muestra un esquema del proceso *k-fold cross validation.*

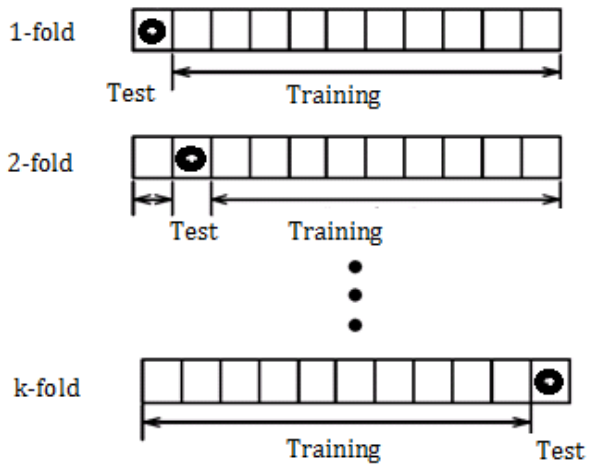


Figura 10.- Procedimiento de k-fold cross validation.

El número k es seleccionado por el investigador. En esta investigación se ha utilizado un valor de *10x10-fold cross validation* (con el objeto de obtener el valor real de funcionamiento de los modelos usados).

4.3.3. Evaluación del método de construcción de ADDs.

En general las medidas que pueden utilizarse para evaluar los resultados obtenidos al aplicar un determinado método de clasificación son: *Accuracy*, Sensitivity, *Specificity, y Receiver Operating Characteristic Curve Area* (De Oña et al., 2011; Mujalli and De Oña, 2011).

Si la variable clase es la severidad del accidente, definida según dos estados posibles (HL-Accidentes con heridos leves y HGM-Accidentes con heridos graves o muertos), la definición de estos indicadores es la siguiente:

- *Accuracy* - Es la precisión del método, definida como el porcentaje de de casos correctamente clasificados, por el clasificador.
- *Sensitivity* – Representa la proporción de casos correctamente clasificados como heridos leves del total de los casos clasificados como heridos leves.
- *Specificity* – Representa la proporción de casos correctamente clasificados como heridos graves o muertos del total de los casos clasificados como heridos graves o muertos.
- *Receiver Operating Characteristic Curve Area (ROC)* – Este indicador representa la curva de casos positivos correctamente clasificados (*sensitivity*) frente a los casos falsos positivos (1-*specificity*), de modo que un valor de 1 describe un perfecto ajuste.

Y las ecuaciones que definen estos indicadores, expresadas en porcentajes son las siguientes:

$$Accuracy = \frac{THL+THGM}{THL+THGM+FHL+FHGM}100\% \quad (17)$$

$$Sensitivity = \frac{THL}{THL+FHGM}100\% \quad (18)$$

$$Specificity = \frac{THGM}{THGM+FHL}100\% \quad (19)$$

Donde, *TrueHL* (THL) – es el número de casos de heridos leves; *TrueHGM* ($THGM$) – es el número de casos de heridos graves o muertos; *FalseHL* (FHL) – es el número de casos falsos de heridos leves (es decir, incorrectamente clasificados como heridos leves); *FalseHGM* ($FHGM$) – es el número de casos falsos de heridos graves o muertos (es decir, incorrectamente clasificados como heridos graves o muertos).

La principal ventaja de estos indicadores es que su interpretación es muy sencilla: cuanto mayor sea el valor que arroja un indicador, mejores son los resultados del modelo obtenido.

Por tanto, para evaluar los modelos construidos en esta investigación, los indicadores utilizados serán: *Accuracy*, *Sensitivity, Specificity* y *ROC*.

4.3.4. Importancia de las variables.

Los ADDs permiten además obtener la importancia de cada una de las variables independientes en el modelo, siendo un resultado fundamental a la hora de identificar las variables con un efecto clave en la severidad del accidente.

Los criterios de partición expuestos anteriormente, representan una medida interesante para calcular la importancia directa de cada una de las variables atributo sobre la variable en estudio. La importancia de las variables que intervienen en el modelo se define del siguiente modo: para una variable X, con los posibles estados $\{x_1...x_n\}$, la siguiente ecuación expresa la importancia o medida informativa de X sobre la variable en estudio C:

$$VIMX = \sum_{i=1}^{h} \frac{nx_i}{n} (I(C/X = x_i) - I(C)), \qquad (20)$$

donde n_{xi} el número de casos de $X=x_i$; n el número de casos totales; I es el índice de Gini del método CART, la ganancia de información en ID3 o el ratio de ganancia de información en C4.5.

4.3.5. Reglas de decisión.

Como se ha indicado anteriormente, la propia estructura de un árbol puede ser transformada en un conjunto de reglas que permiten toda la extracción de la información, a priori potencialmente útil, del conjunto de datos de estudio.

Las RDs son una estructura condicional lógica del tipo "Sí (A) →Entonces (B)". Siendo A, el conjunto de estados de las diferentes variables atributo que conforman una determinada rama del árbol, y B el estado de la variable clase en el nodo hoja determinado.

La parte A de la regla se conoce como ANTECEDENTE, mientras que la parte B es la CONSECUENCIA. Así una RD puede ser expresada, por ejemplo, de la siguiente forma:

SI (ANCHO DE LA CALZADA = estrecho & ILUMINACIÓN = insuficiente) **ENTONCES** (SEVERIDAD=*accidente con heridos graves*).

En un ADD las reglas se configuran desde el nodo raíz, y en ese punto comienza la estructura condicionada (SI). Cada variable que interviene en la división árbol conforma un "SI" de la regla, que termina en los nodos hoja con un valor de "ENTONCES" asociado con la clase resultante en el nodo hijo; siendo la clase del nodo,

el estado de la variable de clase que presenta el mayor número de casos en el nodo hijo analizado.

4.3.5.1. Extracción de RDs.

A priori, se pueden identificar tantas reglas como nodos terminales tiene un árbol. Sin embargo, no todas las reglas que puedan extraerse del árbol resultan fácilmente comprensibles o potencialmente útiles para los analistas o gestores de la seguridad vial. Por ejemplo, si una regla implica la consecución de un gran número de variables, no será fácilmente comprensible y, por tanto, no resultará útil de cara a implantar estrategias de seguridad vial. Por ello, en los métodos usados, que se basan en ADDs, limitaremos la profundidad de éstos, lo que implica la obtención de reglas con pocas variables como antecedente. Esto aumentará notablemente la comprensibilidad de los resultados.

Con objeto de extraer las reglas potencialmente útiles (que se denominarán reglas fuertes) y que permitan a los gestores establecer determinadas estrategias de seguridad vial, se utilizan 3 parámetros: support de la regla completa, population y confidence.

Sea una regla RD, que proviene del nodo hoja t del árbol, RD_t: $A \rightarrow B$ donde A es el antecedente y B el consecuente. Sea $|A_t|$ el número de casos que cumplen el antecedente, sea $|(A \rightarrow B)_t|$ el número de casos que cumplen la regla RD_t, y sea N el número total de casos de la muestra.

Si se define el support del antecedente, como el porcentaje de casos que verifican el antecedente de la regla, los parámetros utilizados se definen como sigue:

- Support de la regla (S): es el porcentaje de casos que cumplen la regla RD. El support es una medida de la frecuencia con la que ocurre una combinación de antecedente y consecuencia en la base de datos.

$$S(A \rightarrow B) = \frac{|(A \rightarrow B)_t|}{N} \tag{21}$$

- Population (Po): es el support del antecedente de la regla RD. Es decir, es el porcentaje de casos en los que el consecuente de la regla se produce, dada el antecedente.

$$Po = S(A) = \frac{|A_t|}{N} \tag{22}$$

- Confidence (o probabilidad) (C): es el porcentaje de casos en los que la regla coincide con la clase.

$$C = \frac{S(A \rightarrow B)}{S(A)} \tag{23}$$

Estos conceptos de support y confidence son centrales para las Reglas de Asociación y han sido usados por diversos autores (Agrawal et al, 1993; Pande and Abdel-Aty,

2009; Montella et al., 2011; Montella et al., 2012b). Se observa que la population (Po), que mide esencialmente la fuerza de una regla, se deduce del S y la C: Po = S/C.

La obtención de las RDs "fuertes", o potencialmente útiles de la base de datos, requiere establecer unos valores mínimos de estos parámetros, de modo que cuanto más grandes sean los valores de los mismos, mayor será la "fuerza" de la regla. Sin embargo, dado que algunos eventos de gran interés en el análisis de accidentes de tráfico suelen ser ocasionales (por ejemplo, "los accidentes mortales"), el support de algunas reglas muy importantes para los analistas de seguridad vial podría ser muy bajo. Así que los valores límite de estos parámetros tienen que ser establecidos teniendo en cuenta la naturaleza de los datos que se estudian. Y dependerán de la homogeneidad de la muestra inicial (número de casos de cada una de las clases que pueda tomar la variable clase), del interés de los analistas en los sucesos ocasionales, así como del tamaño de la muestra (bases de datos pequeñas o grandes).

Pande and Abdel-Aty (2009) establecen 0,90% y 10% como valores de umbral para el support y la probabilidad respectivamente. Eso significa que reglas con support <0,90% y / o probabilidad <10% no serían consideradas. En Montella et al. (2012b) se utilizan umbrales más bajos: de 0,10% para el support y un 1,00% para la confidence.

Teniendo en cuenta el tipo de datos utilizados y los objetivos perseguidos en esta investigación, se ha considerados que, dado que la muestra de accidentes no es muy grande (el tamaño y las características de la misma se explican en el punto 4.4), los valores de umbral utilizados para la extracción de reglas "fuertes" deben ser: 0,60% para el support y 60% para la confidence. Dados estos valores, la population mínima (Po) debe ser mayor o igual a un 1%.

4.3.5.2. Validación de RDs.

Con el objeto de comprobar la validez de las reglas obtenidas y además evitar reglas debidas a la casuística propia del conjunto de datos analizado; y siguiendo el procedimiento establecido por otros analistas de seguridad vial (Montella et al., 2011; Montella et al., 2012b), el conjunto de datos total es dividido en dos partes: un conjunto de *training* (que contiene el 70% de los datos) y un conjunto de *test* (que contiene el 30% restante). Así, con el conjunto de training se crean los árboles de los cuales se extraen las RDs, y posteriormente con el conjunto de *test*, se chequean o validan las reglas obtenidas con el *training*.

4.3.6. Método "*Information Root Node Variation*".

En esta tesis doctoral se presenta un nuevo método que permite la extracción de un conjunto global de reglas de una base de datos, a partir de diferentes ADDs, obtenidos mediante la variación del nodo raíz que genera su construcción. Este método será denominado como *Information Root Node Variation* (IRNV).

La idea se apoya en el buen funcionamiento de un método de clasificación similar (Abellán and Masegosa, 2010), que está basado en probabilidades imprecisas, y en el que también se varía el nodo raíz. Sin embargo, en este método sólo se toman como variables posibles para formar el nodo raíz, aquellas que resultan informativas (porque cuando se trabaja con probabilidades imprecisas, no todas las variables resultan informativas).

En IRNV se utilizan todas las variables existentes en la base de datos como nodo raíz. El método de elección de una variable para insertar en un nodo diferente al nodo raíz, será siempre el mismo para todos esos nodos. De este modo, se impondrá el nodo raíz, pero el resto del procedimiento de construcción de los árboles será un método estándar prefijado según un criterio de partición. Concretamente seguiremos el método de construcción de árboles de decisión de Abellán and Moral (2003).

Cuando se extraen RDs de un ADD, las reglas son dependientes de la estructura del mismo. De este modo, la extracción del conocimiento sólo se realiza en el sentido dictado desde el nodo raíz hasta cada uno de los nodos terminales del árbol. Sin embargo, es posible que existan otras reglas importantes que no sean detectadas por la configuración del árbol, que depende del nodo raíz que inicia su construcción. La principal ventaja del método IRNV, es que se permite extraer todo el conocimiento (expresado en forma de RDs) existente en una base de datos.

El proceso mediante el cual se obtiene el conjunto global de reglas con el método IRNV es esquematizado en la Figura 11. A partir de la base de datos del *training* se generan tantos ADDs cómo atributos se tengan. De cada árbol son extraídos los conjuntos de reglas que posteriormente se verifican en el *test*, dando lugar finalmente al conjunto global de reglas.

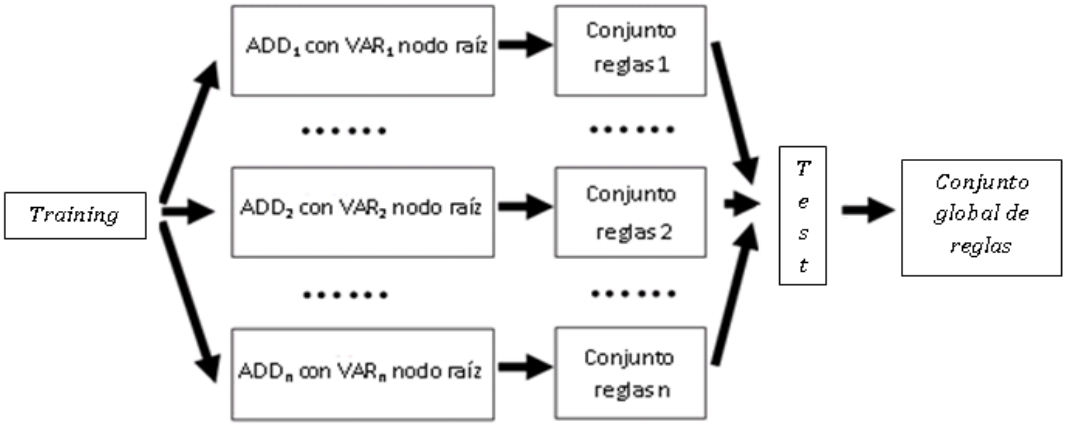


Figura 11.- Esquema del método IRNV.

Desde un punto de vista de la seguridad vial, el conjunto global de reglas obtenidas es de vital interés para los analistas y/o administraciones, ya que se obtienen patrones de comportamiento de los accidentes, que de un modo fácilmente comprensible, permitirán llevar a cabo actuaciones concretas de seguridad vial que ayuden a disminuir su número y/o su gravedad.

4.3.6.1. Procedimiento de construcción de los ADDs en IRNV.

Teniendo en cuenta estas características comunes en los diferentes procedimientos de construcción de ADDs, el procedimiento de construcción de ADDs utilizado por Abellán and Moral (2003) con probabilidades imprecisas y medidas de incertidumbre, ha sido adaptado en esta investigación, para utilizarlo con probabilidad precisas; particularmente para los criterios de partición usados: el basado en el Índice de Gini (GI), y el basado en el Ratio Ganancia de Información (RGI).

El procedimiento recursivo de Abellán and Moral (2003) utilizado para la construcción de un ADD, puede ser expresado según el algoritmo de la Figura 12. Cada nodo N de un ADD produce una partición D de las base de datos (para el nodo raíz, se considera la base de datos completa). Además, cada nodo N tiene asociada una lista "Γ" de características (características que no están presentes en el camino desde el nodo raíz hasta el nodo N considerado).

Procedimiento de construcción de árboles (N, Γ)

1. *Si Γ =Φ, entoces* ***Exit***
2. *Sea D la partición asociada al nodo N*
3. *Calcular el valor de la maxima ganancia de información para el elemento D (usando un criterio de partición: CP)*
   δ= max SC(C,X)
4. *Si δ es menor o igual a 0 entonces* ***Exit***
5. *Si no:*
   6. *Sea X_t la variable para la cual la máxima δ es alcanzada*
   7. *Quitar X_t de Γ*
   8. *Asignar X_t al nodo N*
   9. *Para cada possible valor x_t de X_t*
   10. *Añadir un nodo N_t*
   11. *Hacer N_t un hijo de N*
   12. *Llamar árbol construido (N_t, Γ)*

Figura 12.- Algoritmo para la construcción de un Árbol de Decisión.

Cada resultado "*Exit*" del algoritmo corresponde a la creación de un nodo hoja del árbol, al cual se le asocia el valor más probable de la variable clase (asociado con la correspondiente partición).

Los árboles creados con este procedimiento, son árboles simples, en los que la elección del atributo que conforma cada nodo se realiza en base a los criterios de partición (GI e IRG, en esta investigación); y, en los que de cada nodo surgen tantas ramas como categorías tiene el atributo. De este modo se construye el árbol hasta llegar a un nivel de profundidad determinado (en este caso se han considerado 4 niveles de profundidad). Este criterio también ha sido usado en estudios previos con objetivos

similares (Montella et al., 2011; Montella et al., 2012b)). Otra de sus características es que no tienen implementado el proceso de poda.

4.3.6.2. Obtención del conjunto global de reglas: IRNV

Cuando se obtienen reglas de un solo ADD, éstas vienen determinadas por la variable que se utiliza como nodo raíz, es decir, se usa la información que se tiene en la base de datos según la "dirección" que indica la variable raíz, que es la variable de mayor índice de información sobre la variable clase (en base al criterio de partición utilizado). Con el método IRNV se resuelve esta limitación, ya que permite obtener un conjunto global de reglas a partir de la construcción de ADDs variando el nodo raíz.

Además, tal y como se ha explicado en el epígrafe anterior (4.3.6.1), en esta tesis se propone aplicar este método utilizando dos criterios de partición diferentes, GI e RGI, garantizándose así una mayor extracción de conocimiento de la base de datos analizada. Ya que al utilizar dos criterios, que se complementan el uno al otro (es decir, a partir del mismo conjunto de datos son capaces de obtener modelos diferentes), se pueden extraer un mayor número de reglas.

La construcción del ADD variando el nodo raíz se realiza del siguiente modo: dado un conjunto de m variables independientes o atributos y sea RX_i la variable que ocupa la posición i en importancia con respecto al criterio de partición (CP) utilizado, se construye el árbol ADD_i, con i=1,…, m, utilizando como nodo raíz, RX_i y a continuación se extrae el conjunto de reglas RS_i. El procedimiento utilizado para construir ADD_i es el explicado en el epígrafe 4.3.6.1, con la diferencia de que en cada ADD el nodo raíz es impuesto.

De este modo se obtienen m árboles, de los cuales se extraen m conjuntos de reglas. Para la extracción de cada conjunto de reglas se utilizan los parámetros de support, population y confidence siguiendo el procedimiento explicado en el epígrafe 4.3.5.1. Posteriormente cada conjunto de reglas es validado utilizando el conjunto de *test* (según se ha explicado en el epígrafe 4.3.5.2).

De forma sistemática el proceso de obtención del conjunto global de reglas puede esquematizarse en los siguientes pasos:

1.- Seleccionar GI como CP para la construcción de los ADDs.

2.- Construir el ADD_i utilizando como nodo raíz RX_i y el determinado CP, con i=1, …, m.

3.- Extraer el conjunto de reglas RS_i del ADD_i.

4.- Chequear RS_i en el conjunto de *test* y entonces seleccionar las reglas.

5.- Extraer el subconjunto final de reglas de RS_i obtenidas con el determinado CP.

6.- Seleccionar RGI como CP y volver al paso 2.

7.- Unión de los conjuntos de reglas obtenidos con GI e RGI: Conjunto global de reglas.

4.3.7. Conjunto final de RDs.

Dado que la implantación de cualquier medida correctiva conlleva un coste, y que los recursos disponibles son limitados, es necesario garantizar que los patrones que se obtiene son aquellos que definen las principales problemáticas de las carreteras. Por ello se deben extraer las reglas más "fuertes"; para que de ese modo, las Administraciones competentes puedan centrarse en aquellas actuaciones que le permitan conseguir la mayor efectividad posible en cuanto a la reducción de la accidentalidad y/o su gravedad.

Para conseguir estas reglas, se utiliza un parámetro adicional, el parámetro Lift, que también ha sido usado previamente en otros estudios de seguridad vial con objetivos similares a los de esta investigación (ver Pande and Abdel-Aty, 2009; Montella, 2011; Montella et al., 2011). De hecho, este parámetro es muy utilizado en la literatura de Reglas de Asociación, y es considerado como el más importante para determinar la fuerza de una regla, más incluso, que la confidence o el support (Montella et al., 2012b).

El lift es una mejora de la confidence de una regla, que indica, qué probabilidad existe de encontrar el consecuente de la regla, limitando la búsqueda a aquellos conjuntos de "ítems" dónde el antecedente está presente. Así, el lift puede definirse como la proporción entre el support observado de un conjunto de casos respecto del support teórico de ese conjunto dado el supuesto de independencia:

$$Lift = \frac{S\,(A \rightarrow B)}{S(A)xS(B)} \tag{24}$$

Un valor de lift $= 1$ indica que la regla aparece una cantidad de veces acorde a lo esperado bajo condiciones de independencia. Un valor de lift > 1 indica que aparece una cantidad de veces superior a lo esperado bajo condiciones de independencia. Un valor de lift < 1 indica que ese conjunto aparece una cantidad de veces inferior a lo esperado bajo condiciones de independencia (ver en Brijs et al., 1999).

Con el objeto de extraer las reglas que aparecen un número de veces mayor de lo esperado, y siguiendo el procedimiento de trabajos previos (Pande and Abdel-Aty 2009; Montella et al., 2011; Montella et al., 2012b), se ha establecido un valor mínimo de lift $\geq$1,2; de modo que las reglas que verifican esta última condición son las que conforman el conjunto final de RDs.

El esquema de todos los pasos seguidos para la obtención del conjunto final de RDs puede observarse en la Figura 13.

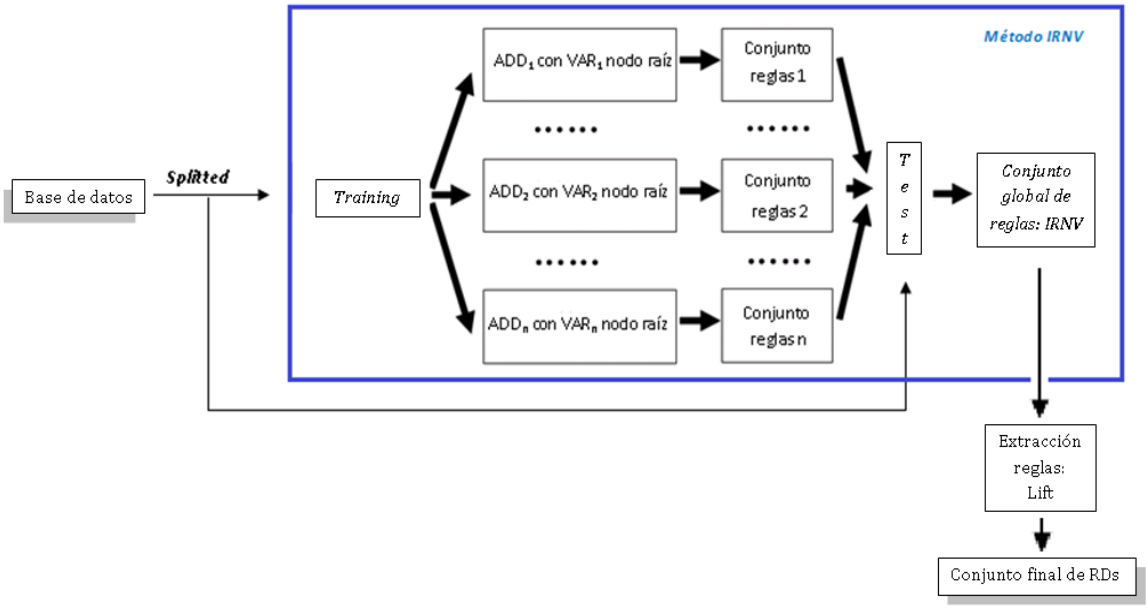


Figura 13.- Esquema de la metodología para obtener el conjunto final de Reglas de Decisión.

Se quiere destacar, que dentro del método IRNV sólo se realiza de forma automática la construcción de los ADDs (a modo de ilustración se incluyen en los anexos I y II de esta tesis doctoral, dos árboles, uno creado con GI y otro creado con RGI). Una vez construidos los árboles, la estructura de los mismos se exporta al programa Excel, y cada nodo terminal de cada árbol se transforma en una posible RDs. Así se obtiene el conjunto inicial de RDs del *Training*. Un paso adicional, antes de la verificación de estas reglas en el *Test*, consiste en eliminar las reglas del *Training* que son iguales (puesto que del total de reglas pueden existir gran cantidad que se repitan); para ello se programan en Excel algoritmos específicos de búsqueda para cada una de las reglas. En el siguiente paso, se eliminan las reglas iguales y, sobre las reglas restantes, se calculan los parámetros de Po, S, C y lift. Aquellas que no verifican los umbrales mínimos, se eliminan. De este modo se obtiene el conjunto de reglas del *Training* (llamados en la Figura 13 conjunto de reglas 1, conjunto de reglas 2, etc.).

A continuación, estas reglas se verifican en el *Test*. Se calculan de nuevo los parámetros de Po, S, C y lift, y se eliminan las reglas que no verifican los umbrales mínimos establecidos. Esta verificación de las reglas en el conjunto de *Test* tampoco se realiza de forma automática, sino que es necesario obtener las reglas y sus parámetros una a una mediante un chequeo en la base de datos del *Test*.

Para realizar el resto de pasos de la metodología utilizada para la obtención del conjunto final de RDs (ver Figura 13), y todo el proceso de verificación de las reglas (que se expone a continuación), también se ha utilizado el programa Excel.

4.3.7.1. Validación de los patrones en el conjunto final de RDs.

La validación de los patrones que conforman el conjunto final de RDs se realizará mediante dos comprobaciones sucesivas; en la primera, se verifica el nodo raíz

impuesto, y en la segunda, se verifican la totalidad de las variables que conforman la regla.

➢ Validación del nodo raíz.

Dado que la mayor parte de las reglas que conforman el conjunto final de RDs han sido obtenidas de ADDs que se han construido imponiendo el nodo raíz; podrían obtenerse reglas en las que la variable impuesta como nodo raíz no sea realmente importante en el patrón que describe la regla. Por ello, en cada regla que provenga de un ADD dónde se impone el nodo raíz será necesario verificar la importancia de esta variable.

Sea RD la regla en la que se debe validar la variable *nodo raíz,* y que denominaremos regla ampliada; y sea RD^{-}, la regla sin la variable *nodo raíz*, y que denominaremos regla simple. Supongamos que el antecedente A de RD está formado por n variables (X'_1, X'_2, …, X'_n) y que el antecedente A^{-} de RD^{-} está formada por n-1 variables del siguiente modo, (X'_2, …; X'_n), y por tanto las reglas se definen como:

$$RD\text{: } A\ (X'_1, X'_2, \ldots, X'_n) \rightarrow B \text{ vs. } RD^{-}\text{: } A^{-}\ (X'_2, \ldots, X'_n) \rightarrow B,$$

con $X'_i=\{X_1, X_2, \ldots, X_n\}$ y B el consecuente de la regla. Siguiendo el criterio establecido en Montella et al. (2012b) en el que una regla con n+1 "ítems" es seleccionada sobre otra de con n "ítems", si experimenta un aumento del ratio del lift mayor que α, se establece la siguiente ecuación:

$$\frac{Soporte\ (A \rightarrow B)}{S(A)xS(B)} > \frac{Soporte\ (A^{-} \rightarrow B)}{S(A^{-})xS(B)} x\alpha, \qquad (25)$$

simplificando ($\frac{Soporte\ (A \rightarrow B)}{S(A)} > \frac{Soporte\ (A^{-} \rightarrow B)}{S(A^{-})} x\alpha$), y teniendo en cuenta que en Montella et al. (2012b) se estable un valor de α=1,05. En esta investigación se ha relajado el umbral a 1,03, obteniéndose así la condición 1:

$$\text{Condición 1: } \frac{C(A \rightarrow B)}{C\ (A^{-} \rightarrow B)} > 1{,}03 \qquad (26)$$

Sin embargo, considerando esta única condición, no estaríamos garantizando que en todas las reglas que se cumpla, el nodo raíz impuesto sea verdaderamente importante en la regla. Esto es debido a que la condición es un ratio, y pueden obtenerse casos extremos en los que la RD no deba ser seleccionada, como es el caso del ejemplo mostrado en la Tabla 2.

REGLA	A	B	A→B	S(A)	S(B)	S(A→B)	C(A→B)	C(A→B)/C(A⁻→B) > 1,03
RD	6	110	5	0,030	0,550	0,025	0,833	1,068 > 1,03
RD⁻	100	110	78	0,500	0,550	0,390	0,780	

A-Casos que cumplen el antecedente de la regla; B- casos que cumplen el consecuente; A→B casos que cumplen antecedente y consecuente; S- support de la regla;C- confidence de la regla.

Tabla 2.- Ejemplo de aplicación de la condición 1.

En la Tabla 2 se observa que aunque se cumple la condición 1, la diferencia de casos que cumple la regla RD con respecto a la RD^{-} es muy elevada, y por tanto introducir una variable más en la regla no compensa, puesto que desde el punto de vista de

seguridad vial, al introducir más variables, se aumenta su complejidad y se disminuye su capacidad informativa. Por ello algunos autores restringen el número de variables en el antecendente (Pande and Abdel-Aty, 2009).

Con el objeto de evitar que las regla ampliada, RD, tengan un support muy bajo con respecto al de la regla simple, RD^{-}, se impone una condición adicional sobre los supports (véase la ecuación 27).

$$\text{Condición 2: } \frac{S\,(A \rightarrow B)}{S\,(A^{-} \rightarrow B)} > 0{,}2 \qquad (27)$$

De este modo, la regla ampliada RD sería seleccionada cuando se cumplen las condiciones 1 y 2 simultáneamente. Y la regla simple RD^{-} sería seleccionada cuando no se cumple una de las 2 condiciones anteriores.

- **Validación de las variables de la regla: condición del incremento del lift (CIL).**

Con el objeto de validar la cadena concreta de las variables que conforman las reglas, a continuación, se mide el incremento del lift al introducir cada variable en la regla. Siguiendo el criterio establecido en Montella et al. (2012b) en el que una regla con n+1 "ítems" es seleccionada sobre otra de con n "ítems" si experimenta un aumento del ratio del lift, se establece la siguiente ecuación:

$$\text{CIL: } \frac{L(A_{n+1} \rightarrow B)}{L(A_n \rightarrow B)} > 1{,}03 \qquad (28)$$

Dónde $\boldsymbol{A_n}$ es el antecedente de la regla formada por i variables; $\boldsymbol{A_{n+1}}$ es el antecedente de la regla formada por n+1 variables; y B es el consecuente.

La regla $\mathbf{A_{n+1} \rightarrow B}$ es seleccionada sobre $\mathbf{A_n \rightarrow B}$ cuando se cumple la condición CIL.

Finalmente, sobre el conjunto de patrones en el que hemos verificado la cadena de variables, se verifican los parámetros de: Po, S, C y lift con los umbrales previamente definidos (1%, 0,6%, 60% y 1,2, respectivamente). Estas serán las reglas que conforman el conjunto final de RDs, y las que serán descritas desde el punto de vista de la seguridad vial.

4.4. Datos de estudio.

Los datos utilizados para llevar a cabo esta investigación han sido suministrados por la Dirección General de Tráfico (DGT).

La DGT es el organismo sobre el cual recaen las competencias de la elaboración de la estadística de los accidentes (según la Orden Ministerial del 24 de febrero de 1993). La recogida e inserción de datos en los cuestionarios es realizada por los agentes encargados de la vigilancia y control del tráfico. En particular, le corresponde a la Dirección General de la Guardia Civil y a las Policías autonómicas y municipales en el ámbito de sus respectivas competencias.

El envío de los cuestionarios a las Jefaturas Provinciales de Tráfico se realiza dentro de los cinco días siguientes al accidente. En el caso de que se trate de un accidente con víctimas, la remisión nunca se realizará antes de haber efectuado el seguimiento del estado de los heridos durante las primeras veinticuatro horas, a fin de poder determinar si, a efectos estadísticos, se trata de un fallecido en accidente de circulación dentro de las veinticuatro horas o de un herido grave o leve.

Una vez que se tienen los cuestionarios de accidentes en las distintas Jefaturas Provinciales de Tráfico, se introducen los datos que contienen los cuestionarios en los ficheros informáticos de los Servicios Centrales de la Dirección General de Tráfico. Particularmente, la información es almacenada en la aplicación ARENA (Accidentes Recogida de Información y Análisis). Un resumen de la información que se recoge en el sistema puede observarse en la Figura 14 (tomada de la página web de la DGT).

La aplicación ARENA trabaja con la información general del accidente, incluyendo datos de su localización, fecha de ocurrencia, circunstancias, características, climatología, etc. Por otra parte, considera la información particular de los vehículos involucrados, a nivel de matrícula, características técnicas, condiciones, mercancías peligrosas, etc. Por último, recopila información de las personas que se ven afectadas por el accidente, distinguiendo entre conductores, pasajeros y peatones. Para cada uno de ellos, se recoge información identificativa, circunstancial, de lesividad, administrativa, etc.

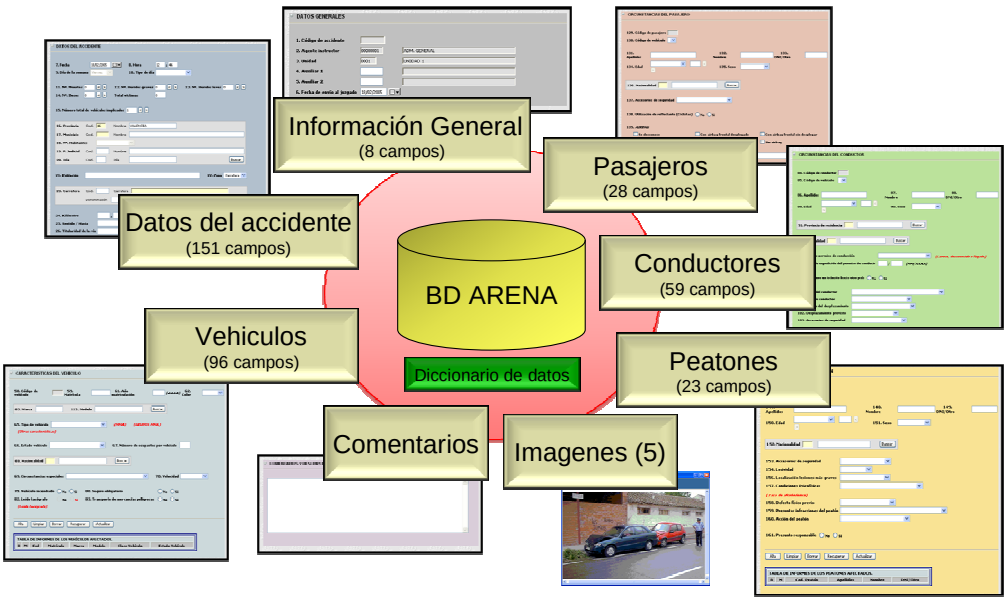


Figura 14.- Información recogida en el sistema ARENA (www.dgt.es).

4.4.1. Severidad del accidente.

La Dirección General de Tráfico (DGT) define accidente con víctimas como aquél en que una o varias personas resultan muertas o heridas, pudiendo ser el resultado:

- Muerto: Toda persona que, como consecuencia del accidente, fallezca en el acto o dentro de los 30 días siguientes.
- Herido grave: Toda persona herida en un accidente de circulación y cuyo estado precisa una hospitalización superior a 24 horas.
- Herido leve: Toda persona herida en un accidente de circulación a la que no pueda aplicarse la definición de herido grave.

En esta investigación, dado que se pretende estudiar la severidad de los accidentes, sólo serán analizados los accidentes con víctimas (quedando fuera del ámbito de este estudio los accidentes con daños materiales).

Siguiendo estudios previos (Chang and Wang, 2006; De Oña et al., 2011; Kashani and Mohaymany, 2011; De Oña et al., 2013), la severidad del accidente se ha definido conforme a las más severas consecuencias resultantes para cualquier víctima implicada en el accidente. De este modo, la severidad del accidente se define de acuerdo a 3 categorías: accidentes con heridos leves, accidentes con heridos graves y accidentes mortales.

4.4.2. Tratamiento de los datos de estudio.

Como se ha indicado en el capítulo 2, antes de aplicar una herramienta concreta de MD, los datos requieren un tratamiento previo, cuyos pasos son esquematizados en Figura 15, y se desarrollan a continuación.

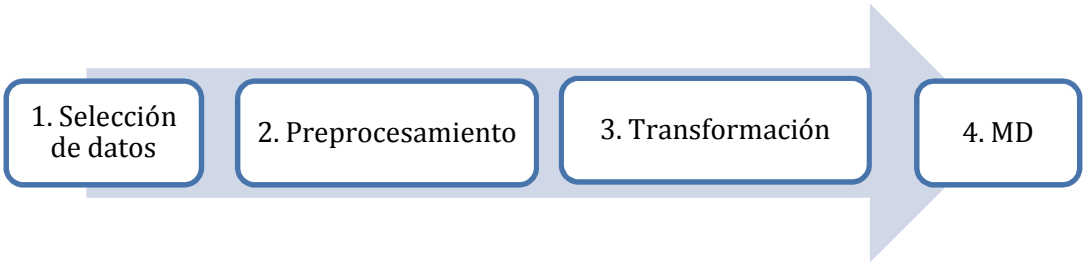


Figura 15.- Proceso de tratamiento de los datos de estudio.

4.4.2.1. Selección de datos.

La información de los accidentes suministrada por la DGT viene recogida en tres bases de datos diferentes: datos del accidente, datos de las personas y datos de los vehículos. El primer paso es unificar toda la información en una sola base de datos.

4.4.2.2. Preprocesamiento.

El proceso de preprocesamiento se realiza tanto sobre los datos de estudio, como sobre las variables seleccionadas para llevar a cabo el análisis.

➢ Filtrado de datos.

En las bases de datos originales suministrada por la DGT se tienen recogidos la totalidad de los accidentes ocurridos en las carreteras de Granada. Dado el objetivo planteado en el Plan Estratégico de Seguridad Vial (Estrategia de Seguridad Vial 2011-2020), que la gravedad de los accidentes en zona no urbana es superior a la de los accidentes en zona urbana, y que en las carreteras convencionales se producen el mayor número de accidentes (puntualizaciones comentadas el capítulo 1), en esta tesis sólo se han analizado los accidentes de tráfico ocurridos en las carreteras convencionales de la provincia.

Además de filtrar los accidentes para reducir la muestra a accidentes ocurridos en carreteras convencionales; se han filtrado los accidentes ocurridos en intersecciones, dado que los factores relativos a los accidentes que ocurren en las intersecciones son diferentes y, por tanto, se recomienda estudiarlas por separado (Moore et al., 2010).

La gravedad del accidente depende de factores de diversa naturaleza (tales como el vehículo, la carretera o el conductor). Sin embargo, los factores que afectan a los accidentes con un solo vehículo involucrado pueden ser diferentes de los que afectan a accidentes con más vehículos involucrados. En los accidentes múltiples la gravedad está altamente relacionada con factores tales como el tipo de colisión, el tamaño y peso de los vehículos involucrados en el accidente, los puntos de contacto, etc. (Krull et al., 2000).

La gravedad de los accidentes de un solo vehículo ha sido ampliamente estudiada por al menos tres razones (Chang and Yeh, 2006). En primer lugar, por lo general, estos accidentes presentan mayor gravedad que los accidentes múltiples, por lo que son un objetivo prioritario de cara al desarrollo de estrategias de mejora de la seguridad vial. En segundo lugar, en estos accidentes el comportamiento del conductor o los factores humanos que contribuyen a la gravedad del accidente pueden ser explorados con mayor eficacia. Y en tercer lugar, estudiar la gravedad de los accidentes con un solo vehículo puede simplificar el diseño de la investigación mediante la exclusión de los efectos relativos a otros vehículos.

Por tanto, en esta investigación sólo se han estudiado accidentes con un vehículo involucrado, siendo filtrados de la muestra aquellos con más de un vehículo involucrado.

Finalmente, dada la disponibilidad de los datos, se han analizado los accidentes ocurridos durante 7 años, desde 2003 hasta 2009. Después del proceso de filtrado de datos se cuenta con una muestra de 1.801 accidentes válidos.

➢ Filtrado de variables.

Una vez que se tienen los accidentes que van a ser objeto de estudio, el siguiente paso es realizar un análisis del total de las variables disponibles para describir un accidente, para posteriormente seleccionar aquellas que se consideren más significativas para el estudio.

En la base de datos unificada se tienen de un total de 106 variables que aportan información general y particular de cada accidente relacionada con las condiciones de la vía en el momento del accidente, los elementos de seguridad de la vía, posibles factores que lo han producido (en opinión del agente que recoge la información en el momento del siniestro), las características propias del vehículo, e información tanto del conductor como de los pasajeros implicados en el accidente. De este modo, el total de estas variables es almacenado en tres grupos de características diferentes:

- **Características relacionadas con el accidente**. En este grupo se incluyen 70 variables que definen las características generales y particulares del accidente, así como, de los elementos de seguridad de la vía y posibles factores que lo han producido.
- **Características relacionadas con las personas**. En este segundo grupo se incluyen 20 variables relacionadas con los datos relativos a los conductores y a los pasajeros de los vehículos implicados en el accidente.
- **Características relacionadas con los vehículos**. En este último grupo se incluyen 16 variables que recogen características propias de los vehículos implicados en el accidente.

Analizando el total de las variables, se puede observar que algunas de ellas se encuentran incluidas en los tres grupos analizados (por ejemplo el número de identificación del accidente o la población en la que se ha producido dicho accidente). Además, existen otras variables que se recogen exclusivamente para su almacenamiento en la base de datos y solo aportan información descriptiva del siniestro (como la titularidad de la carretera en la que se produce el siniestro). Por lo tanto es necesaria una selección de las variables que se van a utilizar en el análisis.

Dado que el objetivo de este estudio es analizar la severidad de los accidentes de tráfico, y teniendo en cuenta las variables que han sido seleccionadas en otros estudios de características similares (Chang and Wang, 2006; De Oña et al., 2011; Kashani and Mohaymany, 2011; Pakgohar et al., 2010); finalmente se han seleccionado un total de 19 variables (ver Tabla 3), relacionadas con:

- **Características del conductor**: edad y sexo.
- **Características del accidente**: causa, día, estacionalidad, hora, número de heridos, número de ocupantes, tipo de accidente.
- **Características relativas a la carretera**: ancho de arcén, ancho de calzada, ancho de carril, arcén pavimentado, barreras de seguridad y marcas viales.
- **Características relativas al entorno**: condiciones atmosféricas, luminosidad y visibilidad.
- **Características relativas al vehículo**: tipo de vehículo.

Cabe destacar que si se usasen otras variables (por disponibilidad de las mismas, por el uso de otra base de datos, etc.) se alcanzarían resultados diferentes en el caso de

que las nuevas variables utilizadas fueran relevantes en el modelo. No obstante, el procedimiento utilizado en esta tesis seguiría siendo igual de válido.

4.4.2.3. Transformación.

Por otra parte, cada una de las variables seleccionadas ofrece un número diferente de respuestas posibles (o categorías), incluso algunas variables recogen inicialmente un número tan elevado que sería imposible trabajar con todas sus categorías iniciales, realizar su análisis y obtener posteriormente una información relevante y comprensible. De modo que el siguiente paso consiste en una transformación de las categorías de las variables seleccionadas.

- **Variables independientes.**

Edad (EDAD): en la base original la variable edad es una variable continua, y para su análisis ha sido re-categorizada en 5 categorías diferentes que hacen referencia a una tipología determinada de conductor: menores, jóvenes, adultos, mayores y de edad desconocida.

Sexo (SEX): esta variable ha sido tomada directamente de la base original.

Causas (CAU): en la base de datos original no existe la variable causas como tal, sino que existen un total de 13 variables que hacen referencia a las posibles causas del accidente. En la Figura 16 se muestra un esquema de cómo se ha creado esta variable, y cómo se han definido las 5 categorías que la forman.

Variables originales		Transformación	Variable nueva
Distracción			
Alcohol			
Cansancio	⟹	Factores del conductor	**CAUSAS (CAU):**
Velocidad inadecuada			
Infracción			
Estado Vía			
Estado Señalización	⟹	Factores de la carretera	Conductor Carretera
Obras			
Estado Vehículo			Vehículo
Avería	⟹	Factores del vehículo	Combinación factores*
Meteorología Adversa			Otra causa
Otros factores	⟹	Otros factores	
Sin opinión			

**El oficial de policía que rellena el parte del accidente puede marcar más de un factor como concurrente en el accidente, en este caso, la categoría de la variable causas del accidente es: combinación de factores.*

Figura 16.- Creación de la variable causas.

Día (DIA): esta variable ha sido tomada directamente de la base original.

Estacionalidad (EST): en la base original se dispone de la variable meses con 12 categorías (correspondientes a cada mes), y para su análisis ha sido re-categorizada en 4 categorías correspondientes con las diferentes estaciones del año y llamada estacionalidad.

NUM	VARIABLES	CATEGORÍAS		%	SEVERIDAD	
	Código: nombre	Código	Descripción	TOTAL	%HL	%HGM
1	EDAD: Edad	≤ 20	≤ 20	12,22	52,73	47,27
		(20-27]	(20-27]	25,65	50,00	50,00
		(27-60]	(27-60]	53,64	51,76	48,24
		>60	>60	6,89	59,68	40,32
		DES	Desconocida	1,61	27,59	72,41
2	SEX: Sexo	M	Mujeres	15,30	62,18	37,82
		H	Hombres	84,50	49,61	50,39
		DES	Desconocido	0,20	75,00	25,00
3	CAU: Causas del accidente	CON	Conductor	82,70	48,99	51,01
		CAR	Carretera	1,40	84,00	16,00
		VEH	Vehículo	1,20	63,64	36,36
		COF	Combination de factores	13,40	61,16	38,84
		OT	Otra	1,20	72,73	27,27
4	DIA: Tipo de día	L	Laborable	45,00	51,05	48,95
		AF	Anterior a festivo	15,90	52,26	47,74
		F	Festivo	30,60	50,36	49,64
		PF	Posterior a festivo	8,40	57,62	42,38
5	EST: Estacionalidad	INV	Invierno	24,00	47,92	52,08
		PRI	Primavera	25,20	53,64	46,36
		VER	Verano	27,30	51,63	48,37
		OTO	Otoño	23,50	53,07	46,93
6	HORA: Hora	[0-6]	[0-6]	20,00	48,06	51,94
		(6-12]	(6-12]	21,00	58,73	41,27
		(12-18]	(12-18]	32,10	52,77	47,23
		(18-24]	(18-24]	26,90	47,22	52,78
7	HER: Número de heridos	[1]	1 herido	69,60	53,43	46,57
		[>1]	Más de 1 herido	30,40	47,35	52,65
8	OCU: Ocupantes	[1]	1 occupante	64,70	51,20	48,80
		[2]	2 occupantes	22,50	51,48	48,52
		[>2]	Más de 2 occupantes	12,70	53,71	46,29
9	TAC: Tipo de accidente	CO	Colisión con obstáculos	0,90	76,47	23,53
		CP	Colisión con peatones	7,70	33,33	66,67
		VUE	Vuelco	6,60	61,86	38,14
		SV	Salida de la vía	82,90	51,77	48,23
		OT	Otra	1,90	68,57	31,43
10	ANARC: Ancho de arcén	EST	< 1,5 m	40,40	52,54	47,46
		MED	[1,5-2,5] m	10,50	50,28	49,72
		NE	Inexistente o impracticable	49,10	50,57	49,43
11	ANCAR: Ancho de carril	EST	< 3,25 m	27,50	46,87	53,13
		MED	[3,25-3,75] m	70,20	53,20	46,80
		ANC	> 3, 75 m	2,30	58,54	41,46
12	ANCAL: Ancho de calzada	EST	< 6 m	14,40	45,56	54,44
		MED	[7-6] m	30,50	53,19	46,81
		ANC	> 7 m	55,10	52,27	47,73
13	APAV: Arcén pavimentado	N	No	17,10	49,35	50,65
		NE	Inexistente o impracticable	31,30	50,89	49,11
		SI	Sí	51,60	52,74	47,26
14	BAR: Barreras de seguridad	N	No	96,90	48,30	54,70
		SI	Sí	3,10	53,60	46,40
15	MVIAL: Marcas viales	NE	No existen o borrosas	9,40	52,35	47,65
		MAR	Separación de márgenes	9,90	48,31	51,69
		CAR	Separación de carriles	4,90	46,59	53,41
		CYM	Separación de carriles y márgenes	75,80	52,23	47,77
16	CAT: Condiciones atmosféricas	BT	Buen tiempo	86,40	50,58	49,42
		LF	Lluvia fuerte	2,10	63,16	36,84
		LL	Lluvia ligera	8,90	58,75	41,25
		O	Otra	2,60	51,06	48,94
17	LUM: Luminosidad	DIA	Dia	53,10	55,49	44,51
		ATA	Atardecer	5,80	54,29	45,71
		INS	Noche: insuficiente iluminación	7,30	51,15	48,85
		SUF	Noche: suficiente iluminación	40,00	59,72	48,28
		SI	Noche: sin iluminación	29,80	43,10	56,90
18	VIS: Visibilidad	SR	Sin restricción	73,10	51,94	48,06
		ATM	Factores atmosféricos	2,20	67,50	32,50
		EDI	Edificios	0,60	36,36	63,64
		TOP	Topografía	22,70	49,39	50,61
		VEG	Vegetación	0,70	50,00	50,00
		OT	Otra	0,70	50,00	50,00
19	VEH: Tipo de vehículo	VL	Vehículo ligero	70,90	47,10	52,90
		VP	Vehículo pesado	4,90	53,80	46,20
		MOT	Motocicleta o ciclomotor	21,70	35,60	64,40
		OT	Otro	2,50	50,60	49,40
20	SEV: Severidad	HL	Accidente con heridos leves	51,58	-	-
		HGM	Accidente con heridos graves o mortal	48,42		

Tabla 3.- Descripción y distribución de variables.

Hora (HORA): en la base original la variable hora varía de 0-24 h. Para no tener un número muy elevado de posibles categorías se ha re-categorizado en 4 categorías correspondientes con las diferentes franjas horarias mostradas en la Tabla 3.

Número de heridos (HER): en la base original no se tiene la variable número de heridos directamente, sino que esta ha sido construida mediante la combinación de 2 variables de la base original (heridos graves y heridos leves), siguiendo un procedimiento similar al mostrado en la Figura 16. Además, esta variable, inicialmente numérica, se ha re-categorizado (según los valores indicados en la Tabla 3) para no tener un número demasiado elevado de posibles categorías.

Número de ocupantes (OCU): en la base original se tiene una variable numérica que indica el número de ocupantes, pero al igual que la variable HER se ha re-categorizado (según los valores indicados en la Tabla 3) para no tener un número demasiado elevado de posibles categorías.

Tipo de accidente (TAC): en la base original la variable TAC presenta 33 categorías diferentes. Sin embargo dentro de estas categorías existen 5 tipologías principales de los accidentes (salidas de la vía, atropellos, vuelco, colisiones y otro tipo), que han sido las utilizadas para re-categorizar esta variable.

Ancho de calzada (ANCAL): esta variable ha sido tomada directamente de la base original.

Ancho de carril (ANCAR): esta variable ha sido tomada directamente de la base original.

Ancho de arcén (ANARC): esta variable ha sido tomada directamente de la base original.

Arcén pavimentado (APAV): esta variable ha sido tomada directamente de la base original.

Barreras de seguridad (BAR): esta variable ha sido tomada directamente de la base original.

Condiciones atmosféricas (CAT): en la base original se dispone de esta variable. Sin embargo, presenta 9 categorías iniciales que han sido re-categorizadas en 4 categorías correspondientes con los factores climatológicos más frecuentes.

Marcas viales (MVIAL): esta variable ha sido tomada directamente de la base original.

Luminosidad (LUM): esta variable ha sido tomada directamente de la base original.

Visibilidad (VIS): en la base original se dispone de esta variable. Sin embargo, presenta 8 categorías iniciales que han sido re-categorizadas en 6 categorías atendiendo a la frecuencia de las categorías iniciales (por ejemplo, la categoría deslumbramiento representa menos del 10% de los accidentes, por tanto se ha codificado dentro de la categoría otros).

Tipo de vehículo (VEH): en la base original se dispone de esta variable. Sin embargo presenta 27 categorías iniciales que han sido re-categorizada en 4 categorías correspondientes con las principales tipologías de vehículos.

En la Tabla 3 se muestra la descripción de las variables seleccionadas con cada una las categorías finalmente establecidas.

- **Variable clase.**

La variable clase es la severidad del accidente (variable 20 en la Tabla 3), inicialmente presenta 3 categorías con la siguiente distribución: 149 – Accidentes mortales, 723 – Accidentes con heridos graves y 929 – Accidentes con heridos leves.

Dado que las diferentes categorías de la variable clase no se encuentran balanceadas, y que este hecho afecta tanto la precisión total del modelo como a la probabilidad en cada clase (Kashani and Mohaymany, 2011); y teniendo en cuenta que los principales factores que influyen en el resultado de un accidente grave o un accidente mortal son muy similares; la variable severidad ha sido re-codificada en 2 categorías:

- HL – Accidentes con heridos leves (929 casos).
- HGM - Accidentes con heridos graves o mortales (872 casos).

La re-categorización de las variables se ha basado en estudios previos y se ha intentado realizar desde un punto de vista ingenieril. Sin embargo, el autor de esta investigación pone de manifiesto que se podría realizar otra re-categorización de las variables objeto de este estudio. Y al igual que se ha indicado al final del epígrafe 4.4.2.2, si se realizase otra re-categorización los resultados alcanzados podrían ser distintos pero la metodología utilizada seguiría siendo igual de válida.

4.4.3. Descripción de los datos de estudio.

En el período de estudio, la distribución de severidad en las carreteras convencionales analizadas (carreteras convencionales de 2 carriles) es de: 6,5% accidentes mortales, 37,9% accidentes graves y 62,1% accidentes leves. Para el mismo periodo, si se tienen en cuenta el total de los accidentes ocurridos en la provincial (incluyendo accidentes en autovías, autopistas, intersecciones, etc.) la distribución es de: 8,3% accidentes mortales, 40,1% accidentes graves y 51,6% accidentes leves.

En la Tabla 4 se muestra la distribución temporal del número de accidentes con víctimas objeto de este estudio, junto con su distribución de severidad.

ACCIDENTES	2003	2004	2005	2006	2007	2008	2009	Total
Leves	122	111	122	106	160	152	156	929
Graves	113	81	125	124	99	103	78	723
Mortales	18	22	35	25	20	16	13	149
Total	**253**	**214**	**282**	**255**	**279**	**271**	**247**	**1801**

Tabla 4.- Distribución temporal de los accidentes con víctimas analizados.

Según se observa en la Tabla 4, la tendencia de los accidentes es decreciente. Los accidentes mortales y graves también presentan esta tendencia. Sin embargo en los accidentes leves se muestra una tendencia creciente.

Como se observa en la Tabla 4, la distribución de severidad en la muestra de estudio es de: 51,58% de accidentes con heridos leves y 48,42% de accidentes con heridos graves o mortales.

Respecto a la distribución de severidad de las variables analizadas, se observa las siguientes particularidades:

- **Características del conductor.**

Edad: El mayor porcentaje de accidentes se produce en los conductores adultos (53,64%). Este resultado es lógico ya que recogen la franja de edades con mayor número de años. La distribución según gravedad es bastante homogénea en todos los grupos de edad, siendo los conductores de (20-27] años los que mayor porcentaje de accidentes graves presentan (50%).

Sexo: Los conductores varones se ven involucrados en un 84,5% de los accidentes, siendo además el porcentaje de accidentes graves superior para estos conductores.

- **Características del accidente.**

Causas: El 82,70% de los accidentes son debidos a causas atribuibles al conductor. Además, destaca que en los accidentes debidos a causas de la carretera, del vehículo u otras causas, tienen un porcentaje de heridos leves muy superior al de heridos graves o mortales (ver Tabla 3).

Día: El mayor porcentaje de accidentes se produce en días laborables, sin embargo el porcentaje de accidentes en días festivos es muy alto, casi de un 31%; destacando además que los días festivos representan una proporción muy baja con respecto a los días laborables. Respecto a la distribución de severidad, también es mayor en este tipo de días.

Estacionalidad: La distribución de accidentes es muy homogénea en las distintas estaciones. Respecto a la gravedad, los accidentes en invierno presentan mayor gravedad con un porcentaje del 52,08%.

Hora: La distribución de accidentes es muy homogénea en las distintas franjas horarias. Atendiendo a la distribución de gravedad, los accidentes ocurridos de (18-24] horas son los de mayor gravedad (52,78%).

Heridos: El mayor porcentaje de accidentes tienen un herido involucrado (69,60%). Sin embargo, la gravedad es superior cuando hay más de 2 heridos involucrados en el accidente (52,65%).

Ocupantes: El mayor porcentaje de accidentes tienen un ocupante (64,70%), este resultando es consistente según el obtenido con la variable heridos. Respecto a la gravedad, es superior en accidentes con 1 y 2 ocupantes involucrados (ver Tabla 3).

Tipo de accidente: Las salidas de la vía representan el 82,90% de los accidentes. Respecto a la gravedad, las colisiones con peatones son los accidentes de mayor gravedad representando un 66,67%.

Ancho de arcén: Las carreteras con arcén inexistente o impracticable son las que recogen mayor porcentaje de los accidentes (49,10%). La distribución de gravedad es muy homogénea (50% vs. 49%) salvo en carreteras con arcén estrecho, dónde los accidentes leves representan casi un 53%.

Ancho de carril: Las carreteras con ancho de carril de 3,25 a 3,75 m. recogen más del 70% de los accidentes. Sin embargo, los accidentes más graves se producen en carreteras de carriles estrechos (con un porcentaje del 53,13%).

Ancho de calzada: Las carreteras con ancho de calzada mayor a 7 m. recogen más de la mitad de los accidentes (55,10%). Sin embargo, los accidentes más graves se producen en carreteras de calzada estrecha (54,44%).

Arcén pavimentado: Los accidentes en carreteras con el arcén pavimentado representan un 51,60%. La distribución de gravedad es muy homogénea en carreteras sin arcén o con el arcén sin pavimentar, mientras que en carreteras con arcén pavimentado, los accidentes leves representan el mayor porcentaje (52,74%).

Barreras de seguridad: El 96,9% de accidentes se producen en carreteras sin barreras, siendo también en estas circunstancias en las que se producen los accidentes más graves.

Marcas viales: Los accidentes en carreteras con marcas viales en separación de carriles y en los márgenes representan un 75,80%. Destaca una mayor gravedad cuando sólo hay marcas viales en la separación de los carriles (53,41% son accidentes con heridos graves o mortales).

- **Condiciones del entorno.**

Condiciones atmosféricas: El buen tiempo es el factor dominante en los accidentes (86,60%), y a su vez es el que representa el mayor porcentaje de accidentes graves. Es de resaltar que cuando las condiciones atmosféricas son adversas (lluvia fuerte o ligera) la gravedad de los accidentes es menor (36% y 41% respectivamente).

Luminosidad: El mayor porcentaje de los accidentes se producen durante el día (53,10%) seguido de noche con suficiente iluminación (40%). Atendiendo a la gravedad, esta es superior en accidentes que se producen en noche sin iluminación (56,90%).

Visibilidad: Los accidentes que se producen cuando la visibilidad no presenta ningún tipo de restricción son los más frecuentes (73,10%); siendo los más graves cuando la visibilidad está restringida por edificios (63,64%) o por la topografía (50,61%).

- **Características del vehículo.**

Tipo de vehículo: Los accidentes con vehículos ligeros son los que representan el mayor porcentaje (70,90%). Atendiendo a la gravedad, estos vehículos, junto con las motocicletas o ciclomotores, son los que recogen los accidentes más graves (52,90% y 64,40%).

CAPÍTULO 5.

RESULTADOS Y DISCUSIÓN

CAPÍTULO 5.
RESULTADOS Y DISCUSIÓN

5.1. Introducción.

En este capítulo se exponen los principales resultados del trabajo de investigación realizado. El software utilizado para llevar a cabo la experimentación ha sido Weka (Witten and Frank, 2005). Se trata de un software de libre distribución y difusión, disponible en: http://www.cs.waikato.ac.nz/ml/weka/

5.2. Preparación de datos.

El primer paso antes de construir los modelos fue dividir aleatoriamente la base de datos en dos conjuntos diferentes: el conjunto de *training*, que contiene el 70% de los datos, y el conjunto de *test,* que contiene el 30% restante. Para ello se ha aplicado un filtro no supervisado de Weka que realiza esta operación de forma automática. De este modo, el conjunto de *training* está formado por 1260 accidentes, en los que la distribución de severidad es la siguiente: HGM – 646 (51,3%) y HL – 614 (48,7%). El conjunto de *test* está formado por los 541 accidentes restantes, siendo su distribución de severidad: HGM – 254 (47%) y HL – 287 (53%).

5.3. Construcción y descripción de los modelos de ADDs.

Una vez que se tienen dividida la muestra, con el conjunto de *training* se construyen los ADDs utilizando los algoritmos CART, C4.5 e ID3.

Con el fin de estudiar el funcionamiento de los métodos anteriores sobre los datos en estudio, se ha realizado una experimentación exhaustiva utilizando la técnica de 10x10-fold CV. Esta técnica de experimentación muestra de forma más exacta, la potencia de cada uno de los criterios de partición para obtener información de los datos. Finalmente se reportarán los valores medios de los 100 experimentos realizados (100 conjuntos de entrenamiento y sus 100 conjuntos de test correspondientes).

Los resultados de los indicadores que se utilizan para evaluar los modelos (*accuracy, sensitivity, specificity y Receiver Operating Characteristic Curve Area* (ROC)) se muestran en la Tabla 5. Los resultados de los árboles generados con los distintos algoritmos se han comparado utilizando el método de la diferencia mínima significativa o método LSD (*Least Significant Difference*), con el que se comprueba si

existen diferencias significativas entre un método u otro, en promedio (en los 10 conjuntos de datos del *training* y *test* generados a partir del conjunto original de datos, son 10 conjuntos porque se está utilizando un 10x10-fold CV). El nivel de significación utilizado para este test ha sido de un 0,05.

Los resultados del test vienen recogidos en la Tabla 5. Se puede observar que se incluyen las 3 comparaciones posibles (CART, C4.5 e ID3) y, en cada una de ellas, se comparan los resultados de un algoritmo (sombreado en la Tabla 5) con los resultados obtenidos por los 2 restantes sobre cada uno de los 4 indicadores.

ALGORITMOS	INDICADORES							
	Accuracy	p-value	Sensitivity	p-value	Specificity	p-value	ROC	p-value
C4.5	54,16*	0,004	54,83^{v}	0,015	53,54*	0,000	54,39*	0,000
ID3	52,71*	0,000	53,17	0,527	52,28*	0,000	52,81*	0,000
(1) CART	55,57		52,58		58,38		56,99	
CART	55,57^{v}	0,004	52,58*	0,015	58,38^{v}	0,000	56,99^{v}	0,000
ID3	52,71*	0,003	53,17	0,073	52,28	0,138	52,81*	0,003
(2) C4.5	54,16		54,83		53,54		54,39	
CART	55,57^{v}	0,000	52,58	0,527	58,38v	0,000	56,99^{v}	0,000
C4.5	54,16^{v}	0,003	54,83	0,073	53,54	0,138	54,39^{v}	0,003
(3) ID3	52,71		53,17		52,28		52,81	

*v los resultados mejoran significativamente; * los resultados empeoran significativamente.*

Tabla 5.- Comparación de algoritmos utilizados para la construcción de Árboles de Decisión.

Según el parámetro de *accuracy*, se observa que C4.5 y CART presentan valores similares, siendo significativamente mejor en CART, comparado tanto con C4.5 como con ID3. El valor de *accuracy obtenido* con C4.5 también es significativamente mejor que el obtenido con ID3.

Cabe destacar que los valores de *accuracy* se encuentran dentro del rango de los valores obtenidos en otros estudios en los que se han aplicado métodos de clasificación con objetivos similares. Abdel Wahab and Abdel-Aty (2001) obtuvieron precisiones del 61% al aplicar redes bayesianas y 58,1% con redes neuronales. De Oña et al. (2011) obtuvieron precisiones del 58%, 59% y 61% aplicando redes Bayesianas con diferentes algoritmos (AIC, MDL y BDeu respectivamente). En un estudio más reciente, De Oña et al. (2013), aplicando análisis cluster y redes bayesianas, obtuvieron precisiones que variaban de un 55,1% a un 64%.

Si se analiza el parámetro *sensitivity,* se observa que C4.5 presenta mayores valores que CART e ID3, siendo esta mejora significativa cuando se compara con CART. Para el parámetro *specificity*, CART arroja un valor significativamente superior a C4.5 e ID3. Una medida global de estos dos parámetros viene dada por el indicador *ROC*, observándose que CART alcanza los mayores resultados (57%), seguido de C4.5 con un 54%, mientras que ID3 obtiene nuevamente los valores más bajos (con un 53%), produciendo un empeoramiento significativo de los mismos.

Teniendo en cuenta estos resultados se puede decir que el algoritmo ID3 es el método que peores resultados arroja. Sin embargo, entre CART y C4.5 las diferencias analizadas no son tan significativas, y aunque CART obtiene valores algo superiores en los parámetros de *accuracy*, *specificity* y *ROC* analizados, no podemos decir, a priori,

que un método sea mejor que el otro. Así que se ha considerado adecuado analizar los modelos obtenidos con ambos algoritmos: C4.5 y CART.

5.3.1. CART.

El árbol creado con el método CART (Figura 17), tiene un total de 18 nodos, de los cuales 10 son nodos terminales.

La variable raíz que genera la construcción del árbol es el sexo del conductor (SEX), y divide el árbol en dos ramas dando lugar a los nodos 1 y 2. Para los conductores mujeres, y en función de la variable luminosidad de la vía (LUM), se llega a los nodos terminales 5 y 6, que son clasificados con diferente nivel de severidad, según las condiciones de luminosidad: los accidentes son mortales o graves (HGM) cuando la luminosidad es insuficiente o la vía carece de iluminación, con una probabilidad del 61% (nodo 5), mientras que si la luminosidad es suficiente, o es igual a atardecer o pleno día, los accidentes resultan leves, con una probabilidad del 69% (nodo 6).

El resto del árbol se genera cuando los conductores son hombres (nodo 1). Este resultado es coherente con los datos de estudio dado que en el 84,5% de los accidentes analizados los conductores son hombres (ver Tabla 3). A continuación el árbol se divide según la variable tipo de accidente (TAC). Cuando los accidentes son colisiones con obstáculos, vuelcos u otra tipología, se producen accidentes HL con un 64% de probabilidad (nodo 4). Mientras que cuando los accidentes son salidas de la vía o colisiones con peatones, la probabilidad de accidente HGM es mayor (nodo 3 en la Figura 17).

El nodo 3 se divide nuevamente por medio de la variable condiciones atmosféricas (CAT). Cuando CAT son iguales a lluvia ligera, los accidentes producidos son HL con un 63% de probabilidad. Este resultado nuevamente resulta coherente con los patrones de comportamiento detectados en los conductores: con condiciones meteorológicas adversas se extreman las precauciones. Para el resto de CAT el árbol continúa creciendo por medio de la variable día (DIA).

Cuando DIA es fin de semana, festivo o posterior a festivo los accidentes resultan HGM con un probabilidad de un 65% (nodo 10). Este resultado es coherente con la tendencia observada en España, donde la mayoría de víctimas mortales en accidentes de tráfico se producen los fines de semana (el 31,4% de los accidentes de tráfico ocurrido en el año 2009 se produjeron en los fines de semana y días festivos, y en ellos se registraron 818 muertes, es decir, el 38,4% del total número de víctimas mortales de ese año, según los datos del Ministerio del Interior, 2009).

Cuando el día es un día laborable o es anterior a festivo, el árbol se divide nuevamente por medio de la variable hora. A partir de éste nivel de profundidad, la interpretación del árbol comienza a ser más compleja. Y desde un punto de vista de la seguridad vial no puede obtenerse una interpretación directa de los patrones de comportamiento,

porque intervienen un número muy elevado de variables, es decir, están presentes muchos condicionantes. A partir de este punto los resultados son los siguientes:

En la franja horaria de 6 a 12 am, y dependiendo de la pavimentación del arcén, se tienen: accidentes HL cuando el arcén no existe o está pavimentado (nodo 15) y accidentes HGM cuando el arcén no está pavimentado (nodo 16).

Para el resto de franjas horarias, la variable que participa en la construcción del árbol es la estacionalidad (EST). Llegando directamente al nodo 13 cuando los accidentes se producen en verano; y dividiendo nuevamente por la variable luminosidad cuando los accidentes ocurren en cualquier otra estación del año y dando lugar a los dos últimos nodos terminales del árbol: el nodo 17, donde los accidentes se clasifican como leves cuando las condiciones de iluminación de la vía son suficientes; y el nodo 18, dónde los accidentes son graves o mortales en cualquier otra circunstancia de iluminación.

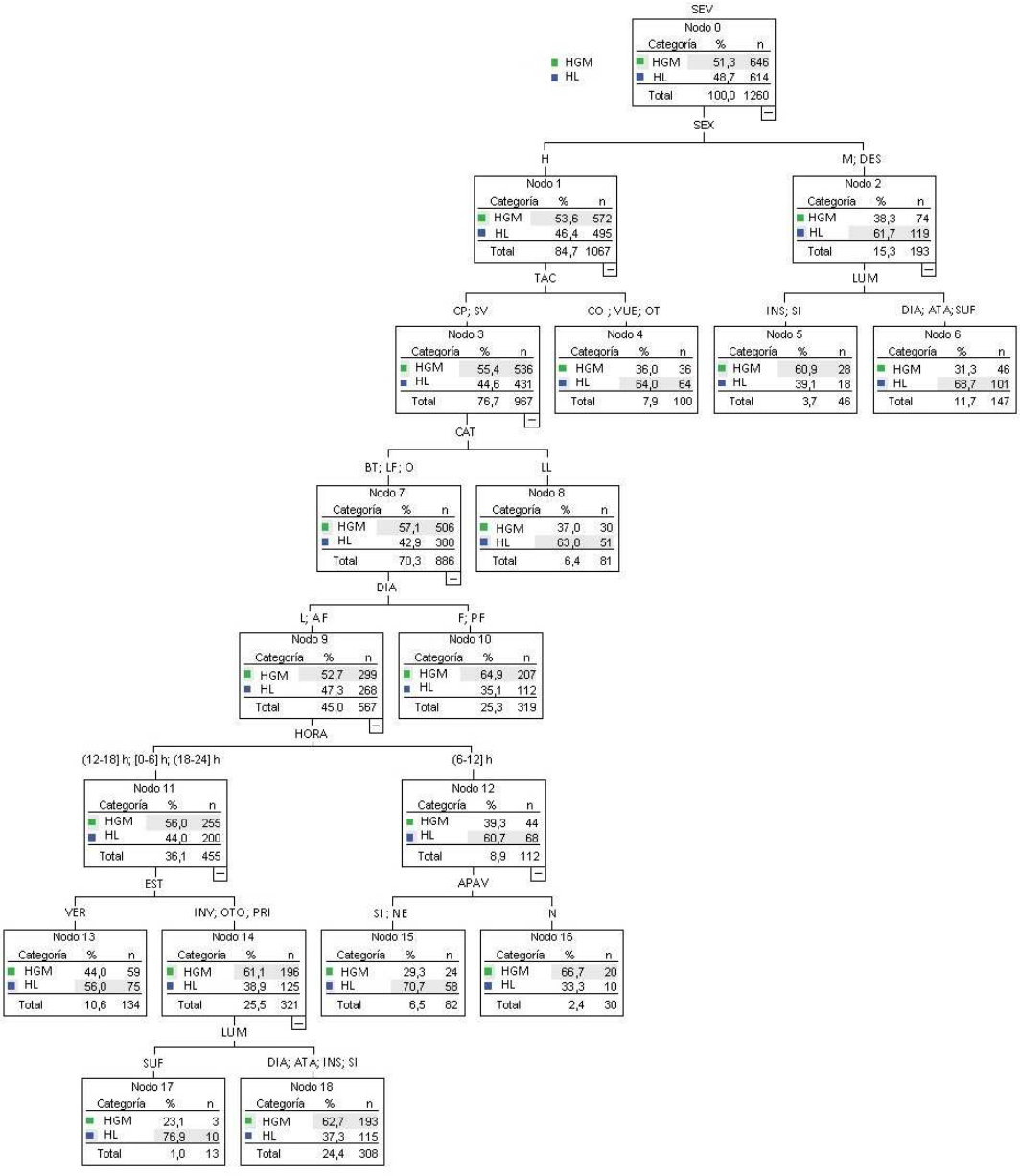


Figura 17.- Árbol de Decisión construido con el método CART.

➢ Importancia de las variables.

En base al criterio de partición usado en el método CART, se puede obtener la importancia de las variables que participan en el modelo. Aplicando la ecuación 20, se obtienen los valores de importancia normalizada que se observan en la Tabla 6. De las 19 variables analizadas, 12 de ellas tienen importancia con valores que varían del 100% al 9,9% (se ha normalizado por el valor de mayor medida informativa). Las variables HER, OCU, ANARC, ANCAL, BAR, MVIAL y VEH tienen importancia en el modelo con valores inferiores al 10%.

VARIABLES	IMPORTANCIA
LUM	100%
CAT	83,6%
HORA	77,1%
TAC	76,0%
SEX	72,0%
APAV	55,9%
DIA	54,9%
EST	49,9%
CAU	32,8%
EDAD	30,6%
VIS	28,4%
ANCAR	9,9%

Tabla 6.- Importancia normalizada de las variables.

La variable luminosidad (LUM) es la de mayor importancia. Este resultado coincide con muchos estudios previos que también la destacan como una de las más significantes en la severidad del accidente: Gray et al. (2008) identificaron que los accidentes más graves se producían por la noche; Abdel-Aty (2003) y Helai et al. (2008) obtuvieron los mismos resultados; en el estudio de Pande and Abdel-Aty (2009) se obtuvo una importante correlación entre la falta de iluminación de la vía y la severidad del accidente; y de la misma forma, en De Oña et al. (2011) se observó que los accidentes graves o mortales estaban asociados a carreteras sin iluminación.

Las condiciones atmosféricas (CAT) son la segunda variable, con una importancia del 83,6%. Este resultado también coincide con estudios previos, tales como Xie et al. (2009), Mujalli and De Oña (2011) o Kashani and Mohyamany (2011), quienes la destacan como una de las variables claves en el análisis de la severidad de los accidentes. La tercera variable es la hora del accidente (HORA), con una importancia del 77,1%. Este resultado también es coherente porque la hora suele estar relacionada con la variable luminosidad. Con un 76% participa la variable tipo de accidente (TAC). Al-Ghamdi (2002), Kcoleman and Kweon (2002) y De Oña et al. (2011) también encontraron la tipología del accidente como una de las variables más importantes en la severidad del mismo. Y por último, cabe destacar la variable sexo (SEX) con un 72% de importancia. Muchos estudios previos han identificado que el género tienen un efecto importante en la severidad del accidente (Evans, 2001; Abdel-Aty, 2003; Ulfarsson and Mannering, 2004; Obeng, 2011). El resto de variables tienen un valor más bajo de importancia (menores del 55,9%).

➢ Extracción y verificación de RDs.

Todos los nodos del árbol pueden ser transformados en reglas de decisión de la forma: "Sí (X) → Entonces (Y)". En la Tabla 7 se muestra las 10 posibles RDs (coincidentes con los nodos terminales del árbol) junto con los valores de los parámetros Po, S y C. Cabe destacar que el número de variables que conforman las reglas varía desde 2 variables (reglas 5 y 6) hasta 7 variables (reglas 17 y 18).

NODO	REGLAS DEL ÁRBOL: SÍ…	ENTONCES	Po(%)	S(%)	C(%)
4	SI [(SEX=H) y (TAC=VUE ó TAC=CO ó TAC=OT)]	HL	7,94	5,08	64,00
5	SI [(SEX= M ó SEX= DES) y (LUM=INS ó LUM=SI)]	HGM	3,65	2,22	60,87
6	SI [(SEX= M ó SEX= DES) y (LUM=DIA ó LUM=ATA ó LUM=SUF)]	HL	11,67	8,02	68,71
8	SI [(SEX=H) y (TAC=SV ó TAC=CP) y (CAT=LL)]	HL	6,43	4,05	62,96
10	SI [(SEX=H) y (TAC=SV ó TAC=CP) y (CAT≠LR) y (DIA=F ó DIA =PF)]	HGM	25,32	16,43	64,89
13	SI [(SEX=H) y (TAC=SV ó TAC=CP) y (CAT≠LR) y (DIA=L ó DIA =AF) y (HORA≠(6-12)) y (EST=VER)]	HL	10,63	5,95	55,97
15	SI [(SEX=H) y (TAC=SV ó TAC=CP) y (CAT≠LR) y (DIA=L ó DIA =AF) y (HORA=(6-12)) y (APAV≠N)]	HL	6,51	4,60	70,73
16	SI [(SEX=H) y (TAC=SV ó TAC=CP) y (CAT≠LR) y (DIA=L ó DIA =AF) y (HORA= (6-12)) y (APAV=N)]	HGM	2,38	1,59	66,67
17	SI [(SEX=H) y (TAC=SV ó TAC=CP) y (CAT≠LR) y (DIA=L ó DIA =AF) y (HORA≠(6-12)) y (EST≠VER) y (LUM=SUF)]	HL	1,03	0,79	76,92
18	SI [(SEX=H) y (TAC=SV ó TAC=CP) y (CAT≠LR) y (DIA=L ó DIA =AF) y (HORA≠(6-12)) y (EST≠VER) y (LUM ≠SUF)]	HGM	24,44	15,32	62,66

Tabla 7.- Extracción de Reglas Decisión con CART.

A continuación se deben verificar las reglas de decisión extraídas del árbol. El objeto de la verificación es eliminar aquellos patrones de comportamiento que puedan ser debidos a la aleatoriedad de los datos (es decir, fruto de la probabilidad), y filtrar aquellos que verdaderamente tengan solidez de cara a fijar determinadas actuaciones desde un punto de vista de la Seguridad Vial.

Para la verificación se filtran las reglas del *training* que cumplen los parámetros mínimos fijados (Po≥1%, S≥ 0,6% y C≥60%). Todas las reglas recogidas en la Tabla 7, salvo la regla del nodo 13, cumplen los parámetros. Y estas reglas son validadas utilizando la muestra de datos que forma el *test*. Para ello se calculan los mismos parámetros de Po, S y C de estas reglas en el *test* y se eliminan aquellas reglas que no alcanzan estos valores mínimos. En la Tabla 8 se muestran los valores de reglas calculados en el *test*.

Teniendo en cuenta los valores de las reglas en el test (Po, S y C), se observa que 6 de las 9 reglas analizadas cumplen el proceso de verificación.

NODO	CLASE	Po(%)	S(%)	C(%)	CHEKEO
4	HL	9,06	5,73	63,27	OK
5	HGM	6,65	5,18	77,78	OK
6	HL	13,12	9,61	73,24	OK
8	HL	7,02	5,73	81,58	OK
10	HGM	26,06	13,49	51,77	NO
15	HL	5,73	3,88	67,74	OK
16	HGM	1,85	1,11	60,00	OK
17	HGM	1,29	0,74	57,14	NO
18	HGM	22,37	13,31	59,50	NO

Tabla 8.- Verificación de las Reglas de Decisión obtenidas con CART en el *test*.

En la Tabla 9 se muestra un resumen de las reglas que han sido verificadas, ordenadas según nivel de severidad y valor de confidence. Se puede observar que los valores de support varían desde el 1,6% (regla 16) a 8,0% (regla 6), que todas las reglas incluyen al menos 1% de la population, y que todos los valores de confidence son superiores a 60,9%.

Num.	REGLAS: SÍ…	ENTONCES	Po(%)	S(%)	C(%)
16	SI [(SEX=H) y (TAC=SV ó TAC=CP) y (CAT≠LR) y (DIA=L ó DIA =AF) y (HORA= (6-12)) y (APAV=N)]	HGM	2,38	1,59	66,67
5	SI [(SEX= M ó SEX= DES) y (LUM=INS ó LUM=SI)]	HGM	3,65	2,22	60,87
15	SI [(SEX=H) y (TAC=SV ó TAC=CP) y (CAT≠LL) y (DIA=L ó DIA =AF) y (HORA=(6-12)) y (APAV≠N)]	HL	6,51	4,60	70,73
6	SI [(SEX= M ó SEX= DES) y (LUM=DIA ó LUM=ATA ó LUM=SUF)]	HL	11,67	8,02	68,71
4	SI [(SEX=H) y (TAC=VUE ó TAC=CO ó TAC=OT)]	HL	7,94	5,08	64,00
8	SI [(SEX=H) y (TAC=SV ó TAC=CP) y (CAT=LL)]	HL	6,43	4,05	62,96

Tabla 9.- Resumen las de Reglas de Decisión con CART.

De las 6 reglas extraídas, 2 hacen referencia a accidentes HGM y 4 hacen referencia a accidentes HL. Los patrones obtenidos pueden analizarse teniendo en cuenta el sexo del conductor implicado en el accidente. Para conductores mujeres, y en función de la iluminación de la vía se distinguen dos patrones particulares:

- Regla 5: en carreteras con insuficiente iluminación o sin iluminación la probabilidad de accidente HGM es del 61%.
- Regla 6: cuando la iluminación de la carretera es suficiente, o es igual a atardecer o pleno día los accidentes son HL, con una probabilidad del 69%.

El resto de patrones se obtienen para conductores varones y, en función del tipo de accidente, se tienen:

- Regla 4: cuando los accidentes son colisiones con obstáculos, vuelcos u otra tipología, se producen accidentes HL, con un 64% de probabilidad.

- Regla 8: cuando los accidentes son salidas de la vía o colisiones con peatones y las condiciones atmosféricas son iguales a lluvia ligera, los accidentes producidos son HL con un 63% de probabilidad.

Para accidentes producidos por salidas de la vía o colisiones con peatones, si las condiciones atmosféricas no son de lluvia ligera, el tipo de día es laborable ó es anterior a festivo, la franja horaria es de 6 a 12 am, y dependiendo de la pavimentación del arcén, se tienen los dos patrones finales:

- Regla 15: accidentes HL cuando el arcén no está pavimentado, con una probabilidad del 71%.
- Regla 16: accidentes HGM cuando el arcén no existe o está pavimentado, con una probabilidad del 67%.

5.3.2. C4.5.

El árbol creado con el algoritmo C4.5 (Figura 18) tiene un total de 52 nodos, de los cuales 39 son nodos terminales.

El aumento en el número de nodos se justifica debido a que este algoritmo crea una rama por cada categoría de las variables que intervienen en el análisis. Sin embargo, en este caso sólo se obtienen 9 reglas que cumplen con los parámetros mínimos fijados para support, population y confidence (ver Tabla 10). Dado el tamaño del árbol generado con este método, sólo será descrita la estructura del árbol que genera las reglas finalmente validas (reglas de la Tabla 10).

Al igual que CART, la variable raíz que genera la construcción del árbol es el sexo del conductor (SEX). Para conductores mujeres y cuando la luminosidad es igual a día, los accidentes son HL (regla 8). Es la regla de mayor population (9,9%) y mayor support (6,8%). Se muestra así un patrón que ya se había obtenido en CART (regla 6), confirmándose que las mujeres parecen verse bastante afectadas por las condiciones de iluminación de la vía.

Al igual que CART, la mayor parte del árbol se genera para conductores hombres y en función del tipo de accidente (TAC) (ver Figura 18). La Figura 18 y la Tabla 10 muestran los siguientes patrones: si el tipo de accidente es vuelco la severidad es HL, con un 61% de probabilidad (regla 12); mientras que si el accidente es una colisión con peatones el patrón obtenido depende del arcén (APAV). Así, en carreteras con arcén pavimentado se tienen accidentes HGM (regla 16), siendo esta regla la de mayor probabilidad (78%).

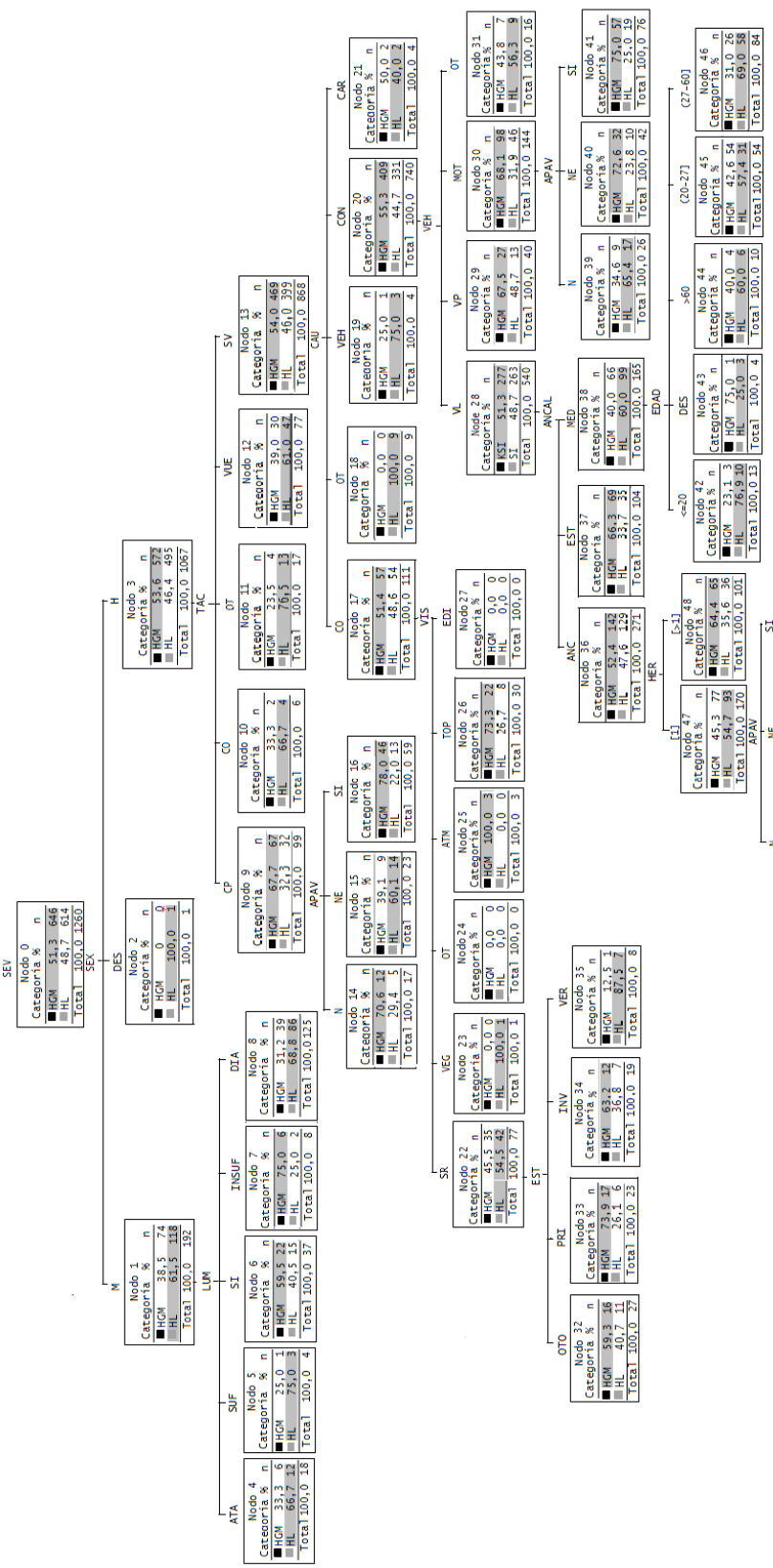


Figura 18.- Árbol de Decisión construido con el método C4.5.

El resto de reglas son obtenidas para accidentes producidos por salidas de la vía y en función de la variable causas del accidente (CAU) se tienen 6 reglas: cuando las causas son una combinación de factores, la visibilidad de la carretera (VIS) no presenta restricciones y los accidentes se producen en primavera, se tiene un patrón de accidentes HL con una probabilidad de casi un 74% (regla 33). Para causas debidas al conductor, y dependiendo del tipo de vehículo (VEH), se observa que los accidentes son HGM cuando los vehículos implicados son vehículos pesados, con un 67,5% de probabilidad (regla 29). Los accidentes son también HGM cuando el vehículo implicado es una motocicleta en carreteras sin arcén o con arcén impracticable (regla 40). Y cuando la tipología de vehículo implicada en el accidente es un vehículo ligero el patrón depende del ancho de calzada.

En carreteras con ancho de calzada (ANCAL) entre 6 y 7 m, y cuando los conductores tienen edades comprendidas entre 28 y 60 años, los accidentes son HL con una probabilidad de casi un 69% (regla 46). En carreteras con ancho mayor de 7 m, y en función del número de heridos (HER), se tienen los dos últimos patrones: si hay más de un herido, los accidentes son HGM en el 64,4% de los casos (regla 48); y si hay un herido y la carretera no tiene arcén o es impracticable, los accidentes son HL con una probabilidad de casi un 71% (regla 50).

Estas tres últimas reglas (46, 48 y 50) son menos útiles para los gestores de seguridad vial porque el antecedente implica la combinación de muchas variables (6, 6 y 7 variables respectivamente), lo que dificulta la interpretación de los resultados y la posibilidad de tomar medidas de actuación directas. En este sentido, Pande and Abdel-Aty (2009) limitan el número de variables en el antecedente a tres.

Num.	REGLAS: SÍ…	ENTONCES	Po(%)	S(%)	C(%)
16	SI (SEX=H) y (TAC=CP) y (APAV=Y)	HGM	4,68	3,65	77,97
40	SI (SEX=H) y (TAC=SV) y (CAU=CON) y (VEH=MOT) y (APAV=NE)	HGM	3,33	2,54	76,19
29	SI (SEX=H) y (TAC=SV) y (CAU=CON) y (VEH=VP)	HGM	3,17	2,14	67,50
48	SI (SEX=H) y (TAC=SV) y (CAU=CON) y (VEH=VL) y (ANARC=ANC) y (HER=[>1])	HGM	8,02	5,16	64,36
33	SI (SEX=H) y (TAC=SV) y (CAU=COF) y (VIS=SR) y (EST=PRI)	HL	1,83	1,35	73,91
50	SI (SEX=H) y (TAC=SV) y (CAU=CON) y (VEH=VL) y (ANARC=ANC) y (HER=[1]) y (APAV=NE)	HL	1,90	1,35	70,83
46	SI (SEX=H) y (TAC=SV) y (CAU=CON) y (VEH=VL) y (ANARC=MED) y (EDAD=(28-60])	HL	6,67	4,60	69,05
8	SI (SEX=M) y (LUM=DIA)	HL	9,92	6,83	68,80
12	SI (SEX=H) y (TAC=VUE)	HL	6,11	3,73	61,04

Tabla 10.- Resumen las de Reglas de Decisión con C4.5.

➢ Importancia de las variables.

Con el criterio de partición del algoritmo C4.5 también es posible obtener la importancia de las variables en el modelo generado. Aplicando la ecuación 20 se obtienen los valores de importancia normalizada que se observan en la Tabla 11. De las 19 variables analizadas, 14 de ellas tienen valores de importancia entre el 100% y el 11,2% (y como en el caso anterior, son valores normalizados según el mayor valor).

Las variables BAR, APAV, ANCAL, OCU y MVIAL aparecen con valores de importancia menores al 5%.

VARIABLES	IMPORTANCIA
TAC	100,%
CAU	80,4%
SEX	69,1%
LUM	67,5%
VEH	65,7%
CAT	59,8%
ANARC	42,8%
EDAD	41,2%
HORA	39,7%
VIS	36,3%
HER	32,1%
DIA	25,7%
ANCAR	20,2%
EST	11,2%

Tabla 11.- Importancia normalizada de las variables.

Once de las variables identificadas con C4.5 han sido también identificadas con CART. Las variables vehículo (VEH) y número de heridos (HER) son además identificadas con C4.5, y otros estudios existentes en la literatura también las han identificado como variables que tiene un efecto importante en la gravedad del accidente (Chang and Wang, 2006; Mujalli and de Oña, 2011). La variable ancho de arcén (ANARC) también se identifica en C4.5, mientras que en CART se identifica la variable arcén pavimentado (APAV). Sin embargo, este resultado es consistente ya que entre estas variables existe correlación. Por tanto, ambos métodos identifican que el arcén tiene influencia en la gravedad del accidente.

La variable tipo de accidente (TAC) es la más importante en el modelo, seguida de la variable causas (CAU) y el sexo del conductor (SEX). Los resultados obtenidos con C4.5 son consistentes con los obtenidos con CART, y como se ha visto anteriormente, los resultados son también consistentes con los obtenidos en la literatura.

5.3.3. Conclusiones

Los ADDs son una herramienta que permite analizar los accidentes de tráfico de un modo sencillo y fácilmente comprensible para los analistas de la seguridad vial. Además, permiten realizar una clasificación de los accidentes en base a su severidad. De este modo, los ADDs se presentan como un método alternativo a los modelos paramétricos debido a su capacidad para identificar los patrones en los datos, sin la necesidad de establecer una relación funcional entre las variables. Por otra parte, estos modelos de clasificación se pueden utilizar para determinar las interacciones entre los variables que serían imposibles de establecer directamente, utilizando las técnicas de modelización tradicionales.

Sobre los métodos particulares para construir los modelos de ADDs utilizados en esta tesis doctoral, se destacan las siguientes conclusiones:

- CART construye árboles binarios y, por tanto, determinadas categorías de las variables divisoras son agrupadas en una rama, aumentando el support de un nodo, pero imposibilitando analizar la influencia de una categoría concreta en la severidad del accidente; a diferencia de C4.5 que crea una rama para cada categoría y permite así analizar la influencia de todas las categorías de las variables que participan en la construcción del árbol. Por tanto, se puede decir que las reglas obtenidas con CART pueden ser menos informativas.

- C4.5 genera árboles con mayor número de ramas que CART y, por tanto, produce mayor número de reglas. Sin embargo, no todas cumplen con los parámetros mínimos establecidos de support, population y confidence, y por ello estas reglas pueden no ser de mucha utilidad de cara a implantar futuras estrategias de seguridad vial.

- Ambos algoritmos permiten obtener la importancia de las variables en el modelo.

- Atendiendo a las características del ADD, se puede destacar que ambos algoritmos presentan semejanzas en cuanto a la estructura del árbol generado. Para ambos la variable raíz es el sexo del conductor separando desde el inicio los patrones de comportamiento en función del mismo:

    - Para las mujeres el árbol únicamente crece por medio de la variable luminosidad, llegando directamente a los nodos hijos, en función de las condiciones de iluminación en el momento del accidente.
    - Los conductores hombres conforman el grueso del árbol, y los diferentes patrones son obtenidos en función de la tipología del accidente.

Los ADDs permiten identificar determinadas reglas, potencialmente útiles, y que pueden ser usadas por los analistas y gestores de seguridad vial. En una primera fase los gestores pueden centrarse en los accidentes graves o mortales (HGM) y posteriormente intervenir en los accidentes leves (HL). Dentro de cada uno de estos grupos, el planteamiento propuesto en esta investigación permitiría priorizar actuaciones en base a los valores de support, population y confidence.

Desde una perspectiva de la gestión de la seguridad vial, se deben destacar algunas conclusiones generales sobre los resultados particulares obtenidos:

- Los accidentes HGM son producidos fundamentalmente por conductores hombres.

- La probabilidad de accidente HGM aumenta en el caso de que haya involucrados peatones (nodo 3 en la Figura 17 y regla 16 en la Tabla 10). Desde un punto de vista de la seguridad vial, se podría actuar sobre estos accidentes colocando barreras de seguridad en los tramos de carretera dónde puedan existir circulación peatonal (recordar que se han analizado carreteras convencionales, luego estos accidentes son localizados en carreteras que conectan 2 núcleos de población cercanos).

- Cuando una conductora está involucrada en el accidente, ambos métodos predicen accidentes HL si existe iluminación en la vía (pleno día, suficiente iluminación o atardecer), (regla 6 en la Tabla 9 y nodos 4, 5 y 8 en la Figura 18). Sin embargo, ambos predicen HGM cuando no existe iluminación o esta es insuficiente (regla 5 en la Tabla 9 y nodos 6 y 7 en la Figura 18). Estas reglas no se observan para conductores varones y podrían indicar que las mujeres aumentan su riesgo de severidad en el accidente en condiciones de menor iluminación de la vía.

Desde un punto de vista de la seguridad vial, la mayor parte de las reglas obtenidas coinciden con los problemas tradicionales de las carreteras convencionales en los países desarrollados, como muchos estudios previos han señalado (Evans, 2001; Abdel-Aty, 2003; Gray et al., 2008; Helai et al., 2008; Pande and Abdel-Aty, 2009; Xie et al., 2009; Kashani and Mohyamany, 2011; Montella, 2011; Mujalli and De Oña, 2011; De Oña et al., 2011; De Oña et al., 2013).

Por último, hay que destacar que cada método presenta ventajas e inconvenientes y revela información diferente, por lo que ambos métodos pueden resultar complementarios y, para un análisis completo, se recomienda usar ambos de forma conjunta.

5.4. Extracción de reglas de decisión mediante el uso del método *Information Root Node Variation method* (IRNV).

Las RDs extraídas de un ADD dependen de la propia estructura del árbol. De modo que la extracción de conocimiento (en forma de RDs) sólo se realiza en el sentido dictado desde el nodo raíz hasta cada uno de los nodos terminales del árbol. Sin embargo, es posible que existan otras reglas importantes que no sean detectadas por la configuración del árbol, que depende del nodo raíz que inicia su construcción.

Por ello, en la siguiente fase de la investigación se ha aplicado un método que permite la extracción de reglas a partir de diferentes ADDs obtenidos mediante variación del nodo raíz que genera su construcción. El nombre utilizado para hacer referencia a este método es *Information Root Node Variation* (IRNV) *method.* Hemos desarrollado este método a partir de un método similar basado en probabilidad imprecisas, y presentado en Abellán and Masegosa (2010).

El fundamento principal del método es que no siempre la variable que expresa mayor importancia (grado de información) sobre la variable clase, y que se inserta como nodo raíz en un ADD, da lugar a mejores resultados. En muchas bases de datos reales podemos encontrar relaciones de 2 o más variables que conjuntamente explican mejor el funcionamiento de la variable en estudio. Además es posible que ninguna de esas variables sea la de mayor importancia, tomada normalmente como nodo raíz.

Como se ha comentado con anterioridad el software utilizado para llevar a cabo la experimentación ha sido Weka. Sin embargo ha sido necesaria la implementación del

procedimiento de construcción de ADDs variando el nodo raíz (epígrafe 4.3.6), así como de los dos criterios específicos de partición utilizados para la construcción de los ADDs (el criterio de ganancia de información - GI y el criterio del ratio de ganancia de información – RGI), utilizando como base el método propuesto en Abellán and Masegosa (2010).

El primer paso antes de aplicar esta metodología fue dividir aleatoriamente la base de datos en los conjuntos de *training* y de *test* (estos conjuntos son los mismos que los indicados en el epígrafe 5.2 de este capítulo). De modo que el conjunto de *training* está formado por 1260 accidentes, en los que la distribución de severidad es la siguiente: HGM – 646 (51,3%) y HL-614 (48,7%).

A continuación, utilizando las variables definidas en la Tabla 3 y utilizando el conjunto de *training* se construyen los 38 ADDs posibles del siguiente modo: 19 ADDs utilizando GI y 19 ADDs utilizando RGI.

Además, como se ha explicado anteriormente, no nos interesa obtener las reglas que surgen de niveles muy bajos de los árboles porque dan lugar a combinaciones complejas de casos de variables atributo que son poco informativas de cara a posibles campañas de actuación por organismos oficiales, tales como la DGT. Por ello, los ADDs son construidos limitando la profundidad del árbol a 4 niveles. Este criterio también ha sido utilizado en estudios previos con objetivos similares (Montella et al., 2011; Montella et al., 2012b). Como se ha visto en el epígrafe 5.3, las reglas que combinan la consecución de muchas variables en el antecedente dificultan la interpretación de los resultados y la posibilidad de tomar medidas de actuación directas. En este sentido, Pande and Abdel-Aty (2009) limitan el número de variables en el antecedente a tres.

Una vez que se han construido los ADDs, se procede a la extracción de reglas. A priori, se pueden identificar tantas reglas como nodos terminales tiene el árbol, sin embargo, con el objetivo de extraer las "reglas fuertes", se utilizan 3 parámetros (S, Po y C) y unos valores umbrales mínimos: $S \geq 0{,}6\%$, $Po \geq 1\%$ y $C \geq 60\%$. Los valores de estos umbrales son los mismos que los utilizados en la experimentación realizada en el epígrafe 5.3. De modo que, sólo serán extraídas las reglas del *training* que cumplen con los parámetros mínimos establecidos. Posteriormente, estas reglas se comprueban en el conjunto de *test*.

Finalmente, con el objeto de extraer las reglas más fuertes y relevantes, se calcula el parámetro lift. Como se ha indicado en el epígrafe 4.3.7, en esta investigación se ha utilizado un valor de $lift > 1{,}2$ para la extracción de las reglas que forman el conjunto final. Cabe destacar que si lo parámetros se suavizaran, es decir, se tomaran valores más bajos, se obtendrían mayor número de reglas. De forma contraria, si los parámetros fuesen más altos, se obtendrían menos reglas (y desde el punto de vista de seguridad vial se perdería información potencialmente útil).

5.4.1. Análisis de reglas obtenidas con GI.

En la Tabla 12 se muestra el número de reglas obtenidas en los diferentes pasos del método IRNV, indicándose la variable raíz que inicia la construcción de cada árbol, así como el número de reglas que se obtiene en el *training* y en el *test* en cada uno de los árboles creados. Además, se recoge el número de reglas que forman el conjunto final de RDs, formado por aquellas reglas que tienen un valor de lift≥1,2.

Si no se aplicase el método IRNV sólo se obtendría un ADD en el que el nodo raíz es la variable TAC (ADD_1 en la Tabla 12). De este árbol se pueden extraer 14 reglas del *training*, de las que sólo 5 verifican el *test* y tienen un lift≥1,2. Por lo que el conocimiento extraído de la base de datos de accidentes analizada resultaría muy reducido.

ADDs	IRNV: REGLAS CON GI			LIFT Conj. Final de RDs
	NODO RAÍZ	TRAINING	TEST	
ADD_1	TAC	14	5	5
ADD_2	CAU	16	4	4
ADD_3	SEX	8	4	4
ADD_4	LUM	17	2	2
ADD_5	VEH	12	4	1
ADD_6	CAT	14	6	6
ADD_7	ANCAL	10	3	2
ADD_8	EDAD	7	3	3
ADD_9	HORA	12	3	2
ADD_{10}	VIS	8	4	4
ADD_{11}	HER	9	3	3
ADD_{12}	DIA	12	5	4
ADD_{13}	ANCAR	14	8	7
ADD_{14}	EST	16	3	2
ADD_{15}	MVIAL	8	5	5
ADD_{16}	OCU	12	7	7
ADD_{17}	ANARC	14	5	5
ADD_{18}	BAR	11	3	3
ADD_{19}	APAV	13	1	1
TOTAL		**227**	**78**	**70**

Tabla 12.- Número de reglas obtenidas con GI.

Sin embargo, cuando se aplica el método IRNV se construyen 19 ADDs, de los que pueden extraerse un total de 227 reglas del *training*. De estas reglas, 78 verifican el test y 70 tienen un valor de lift≥1,2. Con este método se obtiene un conocimiento mucho mayor: 70 reglas vs. 5 reglas.

Se observa que el nodo raíz LUM es el que mayor número de reglas genera en el *training* (17 reglas), más que el ADD_1 con 14 reglas. Respecto a las reglas validadas en el *test,* es el árbol ADD_{13} con el que se validan el mayor número de reglas (8 reglas).

Finalmente, el árbol que más patrones validados produce en el conjunto final de RDs es el ADD_{13}, en el que el nodo raíz es la variable ANCAR (7 reglas), y el ADD_{16} (y nodo raíz OCU) con 7 reglas, seguido del árbol ADD_6 (y nodo raíz CAT) con 6 reglas.

Atendiendo a la gravedad de los patrones obtenidos, cabe destacar que se obtienen un 54,3% de patrones que hacen referencia a accidentes graves o mortales y un 45,7% que hacen referencia a accidentes leves (38 HGM frente a 32 HL).

- **Validación del nodo raíz.**

Dado que la mayor parte de las reglas obtenidas (65 de las 70 totales) proviene de ADDs que se han construido imponiendo el nodo raíz; podrían obtenerse reglas en las que la variable impuesta como nodo raíz no sea realmente "importante" en el patrón que describe la regla. Por ello, es necesario validar el nodo raíz en cada una de estas reglas.

Si RD es la regla ampliada, considerando la variable impuesta como nodo raíz (que coincide con la primera variable de cada regla), por ejemplo: RD = (**SEX=H**; TAC=SV; VEH=OT; ANCAR=MED); y RD^- es la regla simple, sin la variable impuesta como nodo raíz. Para el ejemplo considerado RD^- sería: RD^- = (TAC=SV; VEH=OT; ANCAR=MED). El objetivo es saber qué patrón es el correcto (RD ó RD^-). Se elegirá RD cuando se cumplen la condición 1 (sobre el ratio de confidence) y la condición 2 (sobre el ratio de supports), ecuaciones 26 y 27 respectivamente.

En la Tabla 13 se muestran los resultados de la validación del nodo raíz. Se han verificado todas las reglas que provienen de los árboles ADD_2 - ADD_{19} (y se han codificado como GI06-GI70); las reglas que provienen del ADD_1 (codificadas como GI01-GI05) no necesitan verificación del nodo raíz, dado que en este caso dicho nodo no ha sido impuesto.

Como puede observase en la Tabla 13, de las 65 reglas evaluadas, 48 reglas cumplen la condición 1 del ratio de confidences (ecuación 26) y 17 reglas no la cumplen. Respecto a la condición 2 del ratio de supports (ecuación 27), se observa que la cumplen un mayor número de reglas, 62 de las 65 totales.

Como resultado final se tiene que 45 reglas cumplen simultáneamente las 2 condiciones. Por lo tanto, en 45 reglas se debe mantener el nodo raíz impuesto (y en estas reglas se verifica la regla ampliada RD), mientras que en las 20 reglas restantes (sombreadas en la Tabla 13) no se cumple al menos una de las dos condiciones impuestas, por lo que se extrae la regla simple RD^- (que son los patrones RD pero eliminando la primera variable que los forma, que es el nodo raíz impuesto, marcado en cursiva en las reglas de la Tabla 13).

NUM.	REGLAS: RD	CONDICIÓN 1		CONDICIÓN 2		PATRÓN FINAL
		C(A→B)/C(A⁻→B)	≥ 1,03	S(A→B)/S(A⁻→B)	≥0,2	
GI06	CAU=COF;HORA=(18-24];APAV=NE;ANCAL=MED	1,68	SI	0,37	SI	RD
GI07	*CAU=COF*;HORA=(6-12];VEH=VL;CAT=BT	1,36	SI	0,18	NO	RD-
GI08	*CAU=CON*;VEH=MOT;CAT=BT;TAC=SV	1,02	NO	0,88	SI	RD-
GI09	*CAU=CON*;VEH=VL;EDAD=(20-27];ANCAL=EST	1,01	NO	0,92	SI	RD-
GI10	SEX=H;TAC=CP;APAV=N;VEH=VL	1,15	SI	1,00	SI	RD
GI11	SEX=H;TAC=SV;VEH=MOT;APAV=NE	1,95	SI	1,30	SI	RD
GI12	*SEX=H*;TAC=SV;VEH=OT;ANCAR=MED	0,98	NO	0,91	SI	RD-
GI13	SEX=M ;LUM=DIA;MVIAL=CYM;OCU=[1]	3,03	SI	0,43	SI	RD
GI14	LUM=DIA;SEX=H;VEH=OT;ANCAR=MED	1,10	SI	0,75	SI	RD
GI15	*LUM=SI*;CAT=BT;EST=VER;DIA=F	1,01	NO	0,31	SI	RD-
GI16	*VEH=VL*;TAC=SV;CAU=COF;ANARC=EST	1,01	NO	0,83	SI	RD-
GI17	CAT=BT;LUM=DIA;SEX=H;VEH=OT	2,09	SI	1,36	SI	RD
GI18	CAT=BT;LUM=DIA;SEX=M;OCU=[>2]	1,08	SI	1,00	SI	RD
GI19	CAT=BT;LUM=DIA;SEX=M;OCU=[1]	2,20	SI	1,26	SI	RD
GI20	*CAT=BT*;LUM=SI;EST=INV;MVIAL=CYM	1,02	NO	0,85	SI	RD-
GI21	*CAT=BT*;LUM=SI;EST=OTO;EDAD=(20-27]	1,02	NO	0,93	SI	RD-
GI22	CAT=LL;DIA=F;ANARC=NE;MVIAL=CYM	1,66	SI	0,32	SI	RD
GI23	ANCAL=ANC;CAU=CON;VEH=MOT;TAC=SV	1,45	SI	0,30	SI	RD
GI24	ANCAL=MED;CAU=COF;LUM=DIA;OCU=[1]	1,07	SI	0,40	SI	RD
GI25	EDAD=(20-27];CAU=CON;ANCAL=EST;LUM=SI	1,54	SI	0,44	SI	RD
GI26	EDAD=(27-60];VEH=MOT;CAU=CON;APAV=NE	2,05	SI	0,65	SI	RD
GI27	*EDAD=(27-60]*;VEH=MOT;CAU=CON;APAV=SI	1,02	NO	0,58	SI	RD-
GI28	HORA=[0-6];APAV=SI;EDAD=(20-27];LUM=SI	1,33	SI	0,56	SI	RD
GI29	HORA=[0-6];APAV=SI;EDAD=<=20;OCU=[1]	1,17	SI	0,40	SI	RD
GI30	VIS=SR;CAU=CON;LUM=DIA;HORA=(6-12]	1,13	SI	0,79	SI	RD
GI31	*VIS=SR*;CAU=CON;LUM=INS;EST=INV	0,93	NO	0,80	SI	RD-
GI32	VIS=SR;CAU=CON;LUM=SI;ANCAR=EST	1,11	SI	0,74	SI	RD
GI33	VIS=TOP;EDAD=(20-27];EST=INV;LUM=SI	1,94	SI	0,68	SI	RD
GI34	HER=[>1];EDAD=(27-60];CAU=CON;VIS=TOP	1,16	SI	0,42	SI	RD
GI35	*HER=[1]*;OCU=[1];VEH=MOT;TAC=SV	1,00	NO	1,00	SI	RD-
GI36	HER=[1];OCU=[1];VEH=VL;TAC=CP	2,39	SI	6,08	SI	RD
GI37	DIA=AF;HORA=(6-12];OCU=[1];MVIAL=CYM	1,73	SI	0,25	SI	RD
GI38	DIA=F;TAC=CP;ANCAR=MED	1,51	SI	0,33	SI	RD
GI39	DIA=F;TAC=SV;VIS=TOP;EDAD=<=20	1,49	SI	0,54	SI	RD
GI40	DIA=L;CAU=CON;SEX=M;VEH=VL	2,46	SI	0,78	SI	RD
GI41	ANCAR=EST;LUM=DIA;EDAD=(20-27];VEH=VL	1,51	SI	0,41	SI	RD
GI42	ANCAR=EST;LUM=DIA;EDAD=<=20;SEX=H	1,42	SI	0,47	SI	RD
GI43	ANCAR=EST;LUM=SI;CAT=BT;DIA=F	1,17	SI	0,45	SI	RD
GI44	ANCAR=EST;LUM=SI;CAT=BT;DIA=L	1,14	SI	0,33	SI	RD
GI45	ANCAR=MED;VEH=MOT;CAU=CON;TAC=VUE	1,20	SI	0,87	SI	RD
GI46	ANCAR=MED;VEH=VL;TAC=CP;EST=INV	1,14	SI	0,80	SI	RD
GI47	ANCAR=MED;VEH=VL;TAC=SV;HER=[1]	1,08	SI	0,77	SI	RD
GI48	EST=OTO;TAC=SV;CAU=CON;HORA=(6-12]	1,15	SI	3,18	SI	RD
GI49	EST=PRI;CAT=BT;LUM=DIA;ANCAL=MED	1,15	SI	3,10	SI	RD
GI50	*MVIAL=CYM*;CAU=COF;HORA=[0-6];ANCAL=ANC	1,00	NO	1,00	SI	RD-
GI51	MVIAL=CYM;CAU=COF;HORA=[18-24];LUM=SI	1,70	SI	1,07	SI	RD
GI52	MVIAL=CYM;CAU=COF;HORA=[6-12];VEH=VL	1,92	SI	1,17	SI	RD
GI53	MVIAL=CYM;CAU=CON;VEH=VL;TAC=CP	1,09	SI	0,81	SI	RD
GI54	*MVIAL=NE*;SEX=H;TAC=SV;LUM=SI	1,30	SI	0,14	NO	RD-
GI55	*OCU=[1]*;TAC=CP;EST=INV;EDAD=[27-60]	0,99	NO	0,92	SI	RD-
GI56	*OCU=[1]*;TAC=CP;EST=OTO;VEH=VL	1,00	NO	1,00	SI	RD-
GI57	*OCU=[1]*;TAC=SV;VEH=MOT;APAV=NE	0,96	NO	0,81	SI	RD-
GI58	OCU=[1];TAC=SV;VEH=MOT;APAV=SI	1,06	SI	0,81	SI	RD
GI59	OCU=[1];TAC=SV;VEH=VL;ANCAL=MED	1,09	SI	0,59	SI	RD
GI60	*OCU=[1]*;TAC=VUE;LUM=DIA;HORA=(12-18]	0,98	NO	0,76	SI	RD-
GI61	OCU=[2];CAT=BT;LUM=SI;DIA=L	2,54	SI	0,41	SI	RD
GI62	ANARC=EST;TAC=SV;VEH=MOT;LUM=SI	1,20	SI	0,59	SI	RD
GI63	ANARC=EST;TAC=SV;VEH=VL;HER=[1]	1,07	SI	0,47	SI	RD
GI64	*ANARC=MED*;VIS=SR;LUM=DIA;TAC=SV	1,31	SI	0,16	NO	RD-
GI65	*ANARC=NE*;SEX=H;LUM=DIA;VEH=OT	0,92	NO	0,64	SI	RD-
GI66	ANARC=NE;SEX=M;CAU=CON;VEH=VL	1,05	SI	0,48	SI	RD
GI67	BAR=N;TAC=SV;VEH=MOT;APAV=NE	1,90	SI	1,38	SI	RD
GI68	*BAR=N*;TAC=SV;VEH=OT;ANCAR=MED	1,00	NO	1,00	SI	RD-
GI69	BAR=N;TAC=VUE;CAU=CON;HORA=[12-18]	1,06	SI	1,00	SI	RD
GI70	APAV=SI;TAC=SV;VEH=VL;CAU=COF	2,58	SI	0,86	SI	RD
NÚMERO DE REGLAS: RD		**CONDICIÓN 1**	**48**	**CONDICIÓN 2**	**62**	**45**
NÚMERO DE REGLAS: RD-			**17**		**3**	**20**

Tabla 13.- Evaluación de nodo raíz en las reglas GI

➢ Validación de las variables de la regla: condición del incremento del lift (CIL).

A continuación se verifican las variables que conforman los patrones finales, tanto los obtenidos para el ADD_1 como los obtenidos después de la verificación del nodo raíz. Cabe destacar que se verifican 69 reglas en lugar de las 70 iniciales, porque tras la verificación del nodo raíz, las reglas GI12 y GI68 (del tipo RD^-) son iguales. De este modo se tienen 50 reglas ampliadas RD y 19 reglas simples RD^-.

En las Tabla 14 y 15 se muestran los resultados de la CIL (ecuación 28) para los accidentes HGM y HL respectivamente. Se señalan en sombreado los patrones seleccionados (que cumple el CIL≥1,03), y en negrita su valor de lift y de incremento.

NUM.	COMPARACIÓN - REGLAS BASE		Lift	Incr. Lift
	TAC=SV	HGM	1,00	
	TAC=SV;VEH=MOT	HGM	1,33	32,5%
	TAC=SV;VEH=MOT;APAV=SI	HGM	1,44	8,7%
GI01	TAC=SV;VEH=MOT;APAV=SI;LUM=SI	HGM	**1,75**	**21,3%**
	TAC=SV	HGM	1,00	
	TAC=SV;VEH=MOT	HGM	1,33	32,5%
	TAC=SV;VEH=MOT;APAV=NE	HGM	1,42	6,7%
GI02	TAC=SV;VEH=MOT;APAV=NE;EST=VER	HGM	**1,64**	**16,1%**
	TAC=SV	HGM	1,00	
	TAC=SV;VEH=MOT	HGM	1,33	32,5%
	TAC=SV;VEH=MOT;APAV=NE	HGM	**1,42**	**6,7%**
GI03	TAC=SV;VEH=MOT;APAV=NE;EST=PRI	HGM	1,20	-15,2%
	TAC=SV	HGM	**1,00**	
	TAC=SV;VEH=VL	HGM	0,92	-8,0%
	TAC=SV;VEH=VL;ANCAL=EST	HGM	1,20	30,0%
GI04	TAC=SV;VEH=VL;ANCAL=EST;LUM=SI	HGM	1,48	23,4%
	CAU=COF	HGM	0,94	
	CAU=COF;HORA=(18-24]	HGM	1,20	28,0%
	CAU=COF;HORA=(18-24];APAV=NE	HGM	**1,56**	**30,0%**
GI06	CAU=COF;HORA=(18-24];APAV=NE;ANCAL=MED	HGM	1,50	-3,8%
	VEH=MOT	HGM	**1,16**	
	VEH=MOT;CAT=BT	HGM	1,19	2,5%
GI08	VEH=MOT;CAT=BT;TAC=SV	HGM	1,34	12,4%
	VEH=VL	HGM	**0,95**	
	VEH=VL;EDAD=(20-27]	HGM	0,98	2,6%
GI09	VEH=VL;EDAD=(20-27];ANCAL=EST	HGM	1,46	49,4%
	SEX=H	HGM	1,05	
	SEX=H;TAC=CP	HGM	1,32	26,2%
	SEX=H;TAC=CP;APAV=N	HGM	1,38	4,3%
GI10	SEX=H;TAC=CP;APAV=N;VEH=VL	HGM	**1,65**	**19,9%**
	SEX=H	HGM	**1,05**	
	SEX=H;TAC=SV	HGM	1,05	0,8%
	SEX=H;TAC=SV;VEH=MOT	HGM	1,32	25,6%
GI11	SEX=H;TAC=SV;VEH=MOT;APAV=NE	HGM	1,46	10,5%
	CAT=BT	HGM	**1,04**	
	CAT=BT;EST=VER	HGM	0,96	-7,1%
GI15	CAT=BT;EST=VER;DIA=F	HGM	1,24	28,5%
	LUM=SI	HGM	1,15	
	LUM=SI;EST=INV	HGM	**1,31**	**13,8%**
GI20	LUM=SI;EST=INV;MVIAL=CYM	HGM	1,20	-8,3%
	LUM=SI	HGM	**1,15**	
	LUM=SI;EST=OTO	HGM	1,17	2,4%
GI21	LUM=SI;EST=OTO;EDAD=(20-27]	HGM	1,22	3,8%
	ANCAL=ANC	HGM	1,01	
	ANCAL=ANC;CAU=CON	HGM	1,06	4,8%
	ANCAL=ANC;CAU=CON;VEH=MOT	HGM	1,21	13,6%
GI23	ANCAL=ANC;CAU=CON;VEH=MOT;TAC=SV	HGM	**1,50**	**24,3%**
	EDAD=(20-27]	HGM	1,00	
	EDAD=(20-27];CAU=CON	HGM	1,05	5,6%
	EDAD=(20-27];CAU=CON;ANCAL=EST	HGM	1,39	32,4%
GI25	EDAD=(20-27];CAU=CON;ANCAL=EST;LUM=SI	HGM	**1,76**	**26,7%**
	EDAD=(27-60]	HGM	0,98	
	EDAD=(27-60];VEH=MOT	HGM	1,19	22,1%
	EDAD=(27-60];VEH=MOT;CAU=CON	HGM	1,23	3,3%
GI26	EDAD=(27-60];VEH=MOT;CAU=CON;APAV=NE	HGM	**1,39**	**13,3%**
	VEH=MOT	HGM	1,16	
	VEH=MOT;CAU=CON	HGM	1,21	3,9%
GI27	VEH=MOT;CAU=CON;APAV=SI	HGM	**1,26**	**3,9%**

Tabla 14.- Aplicación de la Condición del Incremento del Lift en las reglas HGM obtenidas con GI.

	HORA=[0-6]	HGM	**1,06**	
	HORA=[0-6];APAV=SI	HGM	0,89	-16,1%
	HORA=[0-6];APAV=SI;EDAD=<=20	HGM	0,98	9,3%
GI29	HORA=[0-6];APAV=SI;EDAD=<=20;OCU=[1]	HGM	1,30	33,3%
	CAU=CON	HGM	**1,03**	
	CAU=CON;LUM=INS	HGM	1,02	-1,5%
GI31	CAU=CON;LUM=INS;EST=INV	HGM	1,33	31,0%
	VIS=SR	HGM	0,98	
	VIS=SR;CAU=CON	HGM	1,02	4,3%
	VIS=SR;CAU=CON;LUM=SI	HGM	1,21	18,9%
GI32	VIS=SR;CAU=CON;LUM=SI;ANCAR=EST	HGM	**1,46**	**20,7%**
	VIS=TOP	HGM	1,09	
	VIS=TOP;EDAD=(20-27]	HGM	1,25	14,9%
	VIS=TOP;EDAD=(20-27];EST=INV	HGM	1,63	30,1%
GI33	VIS=TOP;EDAD=(20-27];EST=INV;LUM=SI	HGM	**1,80**	**10,8%**
	OCU=[1]	HGM	1,00	
	OCU=[1];VEH=MOT	HGM	1,20	19,8%
GI35	OCU=[1];VEH=MOT;TAC=SV	HGM	**1,38**	**14,6%**
	HER=[1]	HGM	0,95	
	HER=[1];OCU=[1]	HGM	**0,99**	**4,5%**
	HER=[1];OCU=[1];VEH=VL	HGM	0,92	-7,4%
GI36	HER=[1];OCU=[1];VEH=VL;TAC=CP	HGM	1,36	48,3%
	DIA=F	HGM	1,10	
	DIA=F;TAC=CP	HGM	1,76	59,7%
GI38	DIA=F;TAC=CP;ANCAR=MED	HGM	**1,95**	**10,5%**
	DIA=F	HGM	**1,10**	
	DIA=F;TAC=SV	HGM	1,11	0,8%
	DIA=F;TAC=SV;VIS=TOP	HGM	1,36	22,2%
GI39	DIA=F;TAC=SV;VIS=TOP;EDAD=<=20	HGM	1,80	32,3%
	ANCAR=EST	HGM	**1,08**	
	ANCAR=EST;LUM=DIA	HGM	0,93	-14,4%
	ANCAR=EST;LUM=DIA;EDAD=<=20	HGM	1,30	40,4%
GI42	ANCAR=EST;LUM=DIA;EDAD=<=20;SEX=H	HGM	1,66	27,5%
	ANCAR=EST	HGM	1,08	
	ANCAR=EST;LUM=SI	HGM	1,32	22,0%
	ANCAR=EST;LUM=SI;CAT=BT	HGM	**1,36**	**2,8%**
GI43	ANCAR=EST;LUM=SI;CAT=BT;DIA=F	HGM	1,58	16,5%
	ANCAR=EST	HGM	1,08	
	ANCAR=EST;LUM=SI	HGM	1,32	22,0%
	ANCAR=EST;LUM=SI;CAT=BT	HGM	**1,36**	**2,8%**
GI44	ANCAR=EST;LUM=SI;CAT=BT;DIA=L	HGM	1,23	-9,1%
	ANCAR=MED	HGM	**0,97**	
	ANCAR=MED;VEH=VL	HGM	0,92	-5,4%
	ANCAR=MED;VEH=VL;TAC=CP	HGM	1,49	61,2%
GI46	ANCAR=MED;VEH=VL;TAC=CP;EST=INV	HGM	1,64	10,5%
	MVIAL=CYM	HGM	**0,99**	
	MVIAL=CYM;CAU=COF	HGM	0,93	-5,7%
	MVIAL=CYM;CAU=COF;HORA=[18-24]	HGM	1,20	29,0%
GI51	MVIAL=CYM;CAU=COF;HORA=[18-24];LUM=SI	HGM	1,56	30,0%
	MVIAL=CYM	HGM	0,99	
	MVIAL=CYM;CAU=CON	HGM	**1,02**	**3,2%**
	MVIAL=CYM;CAU=CON;VEH=VL	HGM	0,95	-7,0%
GI53	MVIAL=CYM;CAU=CON;VEH=VL;TAC=CP	HGM	1,45	53,5%
	SEX=H	HGM	**1,05**	
	SEX=H;TAC=SV	HGM	1,05	0,8%
GI54	SEX=H;TAC=SV;LUM=SI	HGM	1,15	8,8%
	TAC=CP	HGM	**1,29**	
	TAC=CP;EST=INV	HGM	1,19	-7,5%
GI55	TAC=CP;EST=INV;EDAD=[27-60]	HGM	1,69	41,8%
	TAC=CP	HGM	1,29	
	TAC=CP;EST=OTO	HGM	1,79	38,8%
GI56	TAC=CP;EST=OTO;VEH=VL	HGM	**1,95**	**9,1%**
	TAC=SV	HGM	1,00	
	TAC=SV;VEH=MOT	HGM	1,33	32,5%
GI57	TAC=SV;VEH=MOT;APAV=NE	HGM	**1,42**	**6,7%**
	OCU=[1]	HGM	**1,00**	
	OCU=[1];TAC=SV	HGM	1,00	-0,2%
	OCU=[1];TAC=SV;VEH=MOT	HGM	1,38	37,5%
GI58	OCU=[1];TAC=SV;VEH=MOT;APAV=SI	HGM	1,53	11,3%
	OCU=[2]	HGM	1,03	
	OCU=[2];CAT=BT	HGM	1,07	3,7%
	OCU=[2];CAT=BT;LUM=SI	HGM	1,27	18,7%
GI61	OCU=[2];CAT=BT;LUM=SI;DIA=L	HGM	**1,39**	**9,2%**
	ANARC=EST	HGM	0,62	
	ANARC=EST; TAC=SV	HGM	0,96	54,1%
	ANARC=EST; TAC=SV; VEH=MOT	HGM	1,27	32,2%
GI62	ANARC=EST;TAC=SV;VEH=MOT;LUM=SI	HGM	**1,696**	**33,5%**
	BAR=N	HGM	**1,00**	
	BAR=N;TAC=SV	HGM	1,00	0,3%
	BAR=N;TAC=SV;VEH=MOT	HGM	1,31	31,3%
GI67	BAR=N;TAC=SV;VEH=MOT;APAV=NE	HGM	1,42	7,9%

Continuación de Tabla 14.

Tras la verificación de la CIL en las 38 reglas HGM (Tabla 14) se comprueba que seis patrones se repiten (GI11=GI54; GI03=GI57; GI43=GI44). Por tanto, la distribución según el número de variables en los 35 patrones HGM es la siguiente: 15 reglas quedan reducidas a una sola variable, 3 reglas están formadas por 2 variables, 7 reglas por 3 variables, y las 10 reglas restantes están formadas por las 4 variables.

NUM.	COMPARACIÓN - REGLAS BASE		Lift	Incr. Lift
	TAC=VUE	HL	**1,20**	
	TAC=VUE;CAU=CON	HL	1,11	-7,3%
	TAC=VUE;CAU=CON;HORA=(12-18]	HL	1,62	45,6%
GI05	TAC=VUE;CAU=CON;HORA=(12-18];OCU=[1]	HL	1,64	1,3%
	HORA=(6-12]	HL	1,16	
	HORA=(6-12];VEH=VL	HL	1,26	8,3%
GI07	HORA=(6-12];VEH=VL;CAT=BT	HL	**1,31**	**4,2%**
	TAC=SV	HL	1,00	
	TAC=SV;VEH=OT	HL	1,07	7,2%
GI12	TAC=SV;VEH=OT;ANCAR=MED	HL	**1,61**	**50,6%**
	SEX=M	HL	1,26	
	SEX=M ;LUM=DIA	HL	**1,41**	**11,9%**
	SEX=M ;LUM=DIA;MVIAL=CYM	HL	1,35	-4,1%
GI13	SEX=M ;LUM=DIA;MVIAL=CYM;OCU=[1]	HL	1,36	0,4%
	LUM=DIA	HL	**1,04**	
	LUM=DIA;SEX=H	HL	0,96	-8,0%
	LUM=DIA;SEX=H;VEH=OT	HL	1,44	49,6%
GI14	LUM=DIA;SEX=H;VEH=OT;ANCAR=MED	HL	1,89	31,9%
	TAC=SV	HL	1,00	
	TAC=SV;CAU=COF	HL	1,08	8,7%
GI16	TAC=SV;CAU=COF;ANARC=EST	HL	**1,30**	**19,5%**
	CAT=BT	HL	0,96	
	CAT=BT;LUM=DIA	HL	**1,00**	**4,0%**
	CAT=BT;LUM=DIA;SEX=H	HL	0,90	-9,7%
GI17	CAT=BT;LUM=DIA;SEX=H;VEH=OT	HL	1,40	55,5%
	CAT=BT	HL	0,96	
	CAT=BT;LUM=DIA	HL	1,00	4,0%
	CAT=BT;LUM=DIA;SEX=M	HL	1,43	43,3%
GI18	CAT=BT;LUM=DIA;SEX=M;OCU=[>2]	HL	**2,05**	**43,2%**
	CAT=BT	HL	0,96	
	CAT=BT;LUM=DIA	HL	1,00	4,0%
	CAT=BT;LUM=DIA;SEX=M	HL	**1,43**	**43,3%**
GI19	CAT=BT;LUM=DIA;SEX=M;OCU=[1]	HL	1,38	-3,4%
	CAT=LL	HL	1,34	
	CAT=LL;DIA=F	HL	1,47	10,1%
	CAT=LL;DIA=F;ANARC=NE	HL	1,87	27,3%
GI22	CAT=LL;DIA=F;ANARC=NE;MVIAL=CYM	HL	**1,94**	**3,8%**
	ANCAL=MED	HL	**1,12**	
	ANCAL=MED;CAU=COF	HL	0,92	-18,2%
	ANCAL=MED;CAU=COF;LUM=DIA	HL	1,19	29,7%
GI24	ANCAL=MED;CAU=COF;LUM=DIA;OCU=[1]	HL	1,31	9,9%
	HORA=[0-6]	HL	0,93	
	HORA=[0-6];APAV=SI	HL	1,11	19,3%
	HORA=[0-6];APAV=SI;EDAD=(20-27]	HL	**1,54**	**38,3%**
GI28	HORA=[0-6];APAV=SI;EDAD=(20-27];LUM=SI	HL	1,54	0,0%
	VIS=SR	HL	**1,02**	
	VIS=SR;CAU=CON	HL	0,98	-4,4%
	VIS=SR;CAU=CON;LUM=DIA	HL	1,04	6,4%
GI30	VIS=SR;CAU=CON;LUM=DIA;HORA=(6-12]	HL	1,27	21,8%
	HER=[>1]	HL	0,88	
	HER=[>1];EDAD=(27-60]	HL	**0,97**	**10,5%**
	HER=[>1];EDAD=(27-60];CAU=CON	HL	0,98	1,1%
GI34	HER=[>1];EDAD=(27-60];CAU=CON;VIS=TOP	HL	1,28	30,1%
	DIA=AF	HL	1,06	
	DIA=AF;HORA=(6-12]	HL	1,69	60,2%
	DIA=AF;HORA=(6-12];OCU=[1]	HL	1,97	16,2%
GI37	DIA=AF;HORA=(6-12];OCU=[1];MVIAL=CYM	HL	**2,05**	**4,3%**
	DIA=L	HL	**1,04**	
	DIA=L;CAU=CON	HL	1,01	-3,3%
	DIA=L;CAU=CON;SEX=M	HL	1,33	32,0%
GI40	DIA=L;CAU=CON;SEX=M;VEH=VL	HL	1,37	2,9%
	ANCAR=EST	HL	0,91	
	ANCAR=EST;LUM=DIA	HL	1,08	18,0%
	ANCAR=EST;LUM=DIA;EDAD=(20-27]	HL	1,60	48,1%
GI41	ANCAR=EST;LUM=DIA;EDAD=(20-27];VEH=VL	HL	**1,78**	**11,4%**
	ANCAR=MED	HL	**1,03**	
	ANCAR=MED;VEH=MOT	HL	0,86	-16,8%
	ANCAR=MED;VEH=MOT;CAU=CON	HL	0,80	-5,9%
GI45	ANCAR=MED;VEH=MOT;CAU=CON;TAC=VUE	HL	1,28	59,4%

Tabla 15.- Aplicación de la Condición del Incremento del Lift en las reglas HL obtenidas con GI.

	ANCAR=MED	HL	1,03	
	ANCAR=MED;VEH=VL	HL	1,08	5,4%
	ANCAR=MED;VEH=VL;TAC=SV	HL	1,13	4,9%
GI47	ANCAR=MED;VEH=VL;TAC=SV;HER=[1]	HL	**1,29**	**13,4%**
	EST=OTO	HL	**1,01**	
	EST=OTO;TAC=SV	HL	1,02	1,7%
	EST=OTO;TAC=SV;CAU=CON	HL	1,02	-0,6%
GI48	EST=OTO;TAC=SV;CAU=CON;HORA=(6-12]	HL	1,34	32,1%
	EST=PRI	HL	**1,01**	
	EST=PRI;CAT=BT	HL	0,96	-4,8%
	EST=PRI;CAT=BT;LUM=DIA	HL	0,99	2,2%
GI49	EST=PRI;CAT=BT;LUM=DIA;ANCAL=MED	HL	1,30	31,7%
	CAU=COF	HL	1,07	
	CAU=COF;HORA=[0-6]	HL	1,23	15,6%
GI50	CAU=COF;HORA=[0-6];ANCAL=ANC	HL	**1,54**	**25,0%**
	MVIAL=CYM	HL	1,01	
	MVIAL=CYM;CAU=COF	HL	1,07	5,8%
	MVIAL=CYM;CAU=COF;HORA=[6-12]	HL	1,39	29,4%
GI52	MVIAL=CYM;CAU=COF;HORA=[6-12];VEH=VL	HL	**1,44**	**4,0%**
	OCU=[1]	HL	**1,00**	
	OCU=[1];TAC=SV	HL	1,00	0,2%
	OCU=[1];TAC=SV;VEH=VL	HL	1,14	13,7%
GI59	OCU=[1];TAC=SV;VEH=VL;ANCAL=MED	HL	1,36	19,5%
	TAC=VUE	HL	1,20	
	TAC=VUE;LUM=DIA	HL	1,24	3,3%
GI60	TAC=VUE;LUM=DIA;HORA=(12-18]	HL	**1,52**	**22,2%**
	ANARC=EST	HL	1,01	
	ANARC=EST; TAC=SV	HL	1,04	3,3%
	ANARC=EST; TAC=SV; VEH=VL	HL	1,12	7,9%
GI63	ANARC=EST;TAC=SV;VEH=VL;HER=[1]	HL	1,27	**13,0%**
	VIS=SR	HL	1,02	
	VIS=SR;LUM=DIA	HL	**1,10**	**7,5%**
GI64	VIS=SR;LUM=DIA;TAC=SV	HL	1,11	0,6%
	SEX=H	HL	**0,95**	
	SEX=H;LUM=DIA	HL	0,96	0,9%
GI65	SEX=H;LUM=DIA;VEH=OT	HL	1,44	49,6%
	ANARC=NE	HL	0,98	
	ANARC=NE; SEX=M	HL	**1,34**	**36,5%**
	ANARC=NE; SEX=M; CAU=CON	HL	1,27	-5,2%
GI66	ANARC=NE;SEX=M;CAU=CON;VEH=VL	HL	1,31	3,5%
	BAR=N	HL	1,00	
	BAR=N;TAC=VUE	HL	**1,25**	**24,3%**
	BAR=N;TAC=VUE;CAU=CON	HL	1,13	-9,2%
GI69	BAR=N;TAC=VUE;CAU=CON;HORA=[12-18]	HL	1,71	51,0%
	APAV=SI	HL	**1,01**	
	APAV=SI;TAC=SV	HL	1,03	2,2%
	APAV=SI;TAC=SV;VEH=VL	HL	1,18	14,0%
GI70	APAV=SI;TAC=SV;VEH=VL;CAU=COF	HL	1,42	20,3%

Continuación de Tabla 15.

Una vez realizada la verificación de la CIL en las 31 reglas HL (ver Tabla 15), se tiene la siguiente distribución (según el número de variables que forman los patrones): 11 reglas quedan reducidas a una sola variable, 6 reglas están formadas por 2 variables, 7 reglas por 3 variables, y 7 reglas por 4 variables.

- **Conjunto final de patrones.**

Sobre los patrones finales obtenidos al aplicar la CIL se calculan los parámetros Po, S, C y lift, y se comprueban de nuevo que se cumplen los umbrales mínimos previamente fijados (1%, 0,6%, 60% y 1,2 respectivamente). Las reglas que no cumplen estas condiciones se eliminan del conjunto final.

Los resultados de esta comprobación se muestran en las Tablas 16 (accidentes HGM) y 17 (accidentes HL). En sombreado se indican las reglas que cumplen la verificación de los parámetros.

NUM.	REGLAS: SI…	ENTONCES	Po%	S%	C%	Lift	CHEKEO
GI01	TAC=SV;VEH=MOT;APAV=SI;LUM=SI	HGM	2,30	2,06	89,66	1,75	OK
GI02	TAC=SV;VEH=MOT;APAV=NE;EST=VER	HGM	1,51	1,27	84,21	1,64	OK
GI03	TAC=SV;VEH=MOT;APAV=NE	HGM	4,05	2,94	72,55	1,42	OK
GI04	TAC=SV	HGM	82,22	42,22	51,35	1,00	NO
GI06	CAU=COF;HORA=(18-24];APAV=NE	HGM	1,98	1,59	80,00	1,56	OK
GI08	VEH=MOT	HGM	21,03	12,50	59,62	1,16	NO
GI09	VEH=VL	HGM	72,14	35,32	48,95	0,95	NO
GI10	SEX=H;TAC=CP;APAV=N;VEH=VL	HGM	1,03	0,87	84,62	1,65	OK
GI11	SEX=H	HGM	84,68	45,40	53,61	1,05	NO
GI15	CAT=BT	HGM	85,71	45,56	53,15	1,04	NO
GI20	LUM=SI;EST=INV	HGM	9,60	6,43	66,94	1,31	OK
GI21	LUM=SI	HGM	30,08	17,70	58,84	1,15	NO
GI23	ANCAL=ANC;CAU=CON;VEH=MOT;TAC=SV	HGM	6,19	4,76	76,92	1,50	OK
GI25	EDAD=(20-27];CAU=CON;ANCAL=EST;LUM=SI	HGM	1,67	1,51	90,48	1,76	OK
GI26	EDAD=(27-60];VEH=MOT;CAU=CON;APAV=NE	HGM	2,22	1,59	71,43	1,39	OK
GI27	VEH=MOT;CAU=CON;APAV=SI	HGM	9,37	6,03	64,41	1,26	OK
GI29	HORA=[0-6]	HGM	18,33	10,00	54,55	1,06	NO
GI31	CAU=CON	HGM	81,51	43,10	52,87	1,03	NO
GI32	VIS=SR;CAU=CON;LUM=SI;ANCAR=EST	HGM	5,71	4,29	75,00	1,46	OK
GI33	VIS=TOP;EDAD=(20-27];EST=INV;LUM=SI	HGM	1,03	0,95	92,31	1,80	OK
GI35	OCU=[1];VEH=MOT;TAC=SV	HGM	10,79	7,62	70,59	1,38	OK
GI36	HER=[1];OCU=[1]	HGM	64,60	32,86	50,86	0,99	NO
GI38	DIA=F;TAC=CP;ANCAR=MED	HGM	1,43	1,43	100,00	1,95	OK
GI39	DIA=F	HGM	29,29	16,59	56,64	1,10	NO
GI42	ANCAR=EST	HGM	28,17	15,63	55,49	1,08	NO
GI43	ANCAR=EST;LUM=SI;CAT=BT	HGM	8,10	5,63	69,61	1,36	OK
GI46	ANCAR=MED	HGM	69,60	34,76	49,94	0,97	NO
GI51	MVIAL=CYM	HGM	75,63	38,25	50,58	0,99	NO
GI53	MVIAL=CYM;CAU=CON	HGM	60,24	31,43	52,17	1,02	NO
GI55	TAC=CP	HGM	8,65	5,71	66,06	1,29	OK
GI56	TAC=CP;EST=OTO;VEH=VL	HGM	1,59	1,59	100,00	1,95	OK
GI58	OCU=[1]	HGM	65,87	33,89	51,45	1,00	NO
GI61	OCU=[2];CAT=BT;LUM=SI;DIA=L	HGM	3,02	2,14	71,05	1,39	OK
GI62	ANARC=EST;TAC=SV;VEH=MOT;LUM=SI	HGM	1,83	1,59	86,96	1,70	OK
GI67	BAR=N	HGM	97,06	49,60	51,10	1,00	NO

Tabla 16.- Comprobación de parámetros en reglas HGM obtenidas con GI.

Según se observa en la Tabla 16, los umbrales mínimos de los parámetros de Po, S, C y lift son verificados por un total de 19 patrones. Su distribución según el número de variables que los forman es la siguiente: 1 regla está formada por 1 variable, 1 regla están formada por 2 variables, 7 reglas están formadas por 3 variables y 10 reglas por 4 variables.

De las 31 reglas de accidentes HL (recogidas en la Tabla 17), 17 verifican los umbrales mínimos de los parámetros (Po, S, C y lift). Y su distribución según el número de variables es: 3 reglas están formadas por 2 variables, 7 reglas están formadas por 3 variables y 7 reglas por 4 variables.

Una vez que se tienen identificados los patrones finales se procede a su descripción desde el punto de vista de la seguridad vial. Para ello se describirán por separado aquellos patrones que hacen referencia a accidentes graves o mortales de aquellos que hacen referencia a accidentes leves.

NUM.	REGLAS: SI…	ENTONCES	Po%	S%	C%	Lift	CHEKEO
GI05	TAC=VUE	HL	6,51	3,81	58,54	1,20	NO
GI07	HORA=(6-12];VEH=VL;CAT=BT	HL	13,89	8,89	64,00	1,31	OK
GI12	TAC=SV;VEH=OT;ANCAR=MED	HL	1,11	0,87	78,57	1,61	OK
GI13	SEX=M ;LUM=DIA	HL	9,92	6,83	68,80	1,41	OK
GI14	LUM=DIA	HL	53,33	27,14	50,89	1,04	NO
GI16	TAC=SV;CAU=COF;ANARC=EST	HL	4,52	2,86	63,16	1,30	OK
GI17	CAT=BT;LUM=DIA	HL	46,11	22,46	48,71	1,00	NO
GI18	CAT=BT;LUM=DIA;SEX=M;OCU=[>2]	HL	1,03	1,03	100,00	2,05	OK
GI19	CAT=BT;LUM=DIA;SEX=M	HL	8,41	5,87	69,81	1,43	OK
GI22	CAT=LL;DIA=F;ANARC=NE;MVIAL=CYM	HL	1,51	1,43	94,74	1,94	OK
GI24	ANCAL=MED	HL	28,81	15,71	54,55	1,12	NO
GI28	HORA=[0-6];APAV=SI;EDAD=(20-27]	HL	2,22	1,67	75,00	1,54	OK
GI30	VIS=SR	HL	73,65	36,75	49,89	1,02	NO
GI34	HER=[>1];EDAD=(27-60]	HL	17,62	8,33	47,30	0,97	NO
GI37	DIA=AF;HORA=(6-12];OCU=[1];MVIAL=CYM	HL	1,67	1,67	100,00	2,05	OK
GI40	DIA=L	HL	48,02	24,37	50,74	1,04	NO
GI41	ANCAR=EST;LUM=DIA;EDAD=(20-27];VEH=VL	HL	2,38	2,06	86,67	1,78	OK
GI45	ANCAR=MED	HL	69,60	34,84	50,06	1,03	NO
GI47	ANCAR=MED;VEH=VL;TAC=SV;HER=[1]	HL	28,10	17,62	62,71	1,29	OK
GI48	EST=OTO	HL	23,65	11,60	49,00	1,01	NO
GI49	EST=PRI	HL	25,08	12,38	49,37	1,01	NO
GI50	CAU=COF;HORA=[0-6];ANCAL=ANC	HL	1,27	0,95	75,00	1,54	OK
GI52	MVIAL=CYM;CAU=COF;HORA=[6-12];VEH=VL	HL	2,14	1,51	70,37	1,44	OK
GI59	OCU=[1]	HL	65,87	31,98	48,55	1,00	NO
GI60	TAC=VUE;LUM=DIA;HORA=(12-18]	HL	1,83	1,35	73,91	1,52	OK
GI63	ANARC=EST;TAC=SV;VEH=VL;HER=[1]	HL	17,46	10,79	61,82	1,27	OK
GI64	VIS=SR;LUM=DIA	HL	38,12	20,5	53,60	1,10	NO
GI65	SEX=H	HL	84,68	39,29	46,39	0,95	NO
GI66	ANARC=NE; SEX=M	HL	7,06	4,60	65,17	1,34	OK
GI69	BAR=N;TAC=VUE	HL	6,27	3,81	60,76	1,25	OK
GI70	APAV=SI	HL	51,349	25,32	49,30	1,01	NO

Tabla 17.- Comprobación de parámetros en reglas HL obtenidas con GI.

➢ Patrones de accidentes HGM.

En la Tabla 18 se describen los 19 patrones de accidentes HGM, ordenados según el valor de la confidence. Los valores de confidence de las reglas varían desde un 100% (en las reglas GI38 y GI56) hasta un valor de 64,41% (regla GI27). Respecto al support de las reglas, todas tienen valores superiores al mínimo establecido (0,6%), variando de 0,87% (regla GI10) a 7,62% (regla GI35). El valor de population varía de un 10,8% en la regla GI35, hasta casi el mínimo establecido, un 1,03%, en las reglas GI10 y GI33. Y respecto a los valores del lift, varían desde 1,26 (regla GI27) hasta 1,95 (reglas GI38 y GI56).

En la "Estrategia de Seguridad Vial 2011-2020" (DGT, 2011) se señalan los accidentes con peatones en zona urbana como una de las principales políticas de seguridad vial en las que se debe continuar trabajando; dado que los accidentes con peatones en zona urbana son más frecuentes que en zona no urbana. Sin embargo, con este estudio también se muestra la problemática de los accidentes con peatones existente en las carreteras convencionales analizadas. Además, se muestra que las consecuencias de estos accidentes son las más severas: todos los patrones obtenidos para accidentes con peatones son HGM (reglas GI38, GI56, GI10 y GI55). En la regla GI55 se relacionan directamente las colisiones con peatones con accidentes graves o mortales con una

probabilidad del 66%. El resto de patrones más específicos se describen a continuación:

- Regla GI38: identifica estos accidentes, en días festivos y en carreteras con ancho de carril comprendido entre 3,25 y 3,75 m, con una probabilidad de un 100% de que el accidente en esta situación sea HGM.
- Regla GI56: identifica también estos accidentes con una probabilidad del 100% y un support mayor (1,6%). Se producen en otoño cuando el vehículo involucrado es un vehículo ligero.
- Regla GI10: describe un patrón para estos accidentes en los que el conductor involucrado es un hombre, en carreteras sin arcén pavimentado y con un vehículo ligero como vehículo implicado en el accidente. La probabilidad de esta regla es de un 85% y el support de casi un 0,9%.

NUM.	REGLAS (SI...)	ENTONCES	Po%	S%	C%	Lift
GI38	DIA=F;TAC=CP;ANCAR=MED	HGM	1,43	1,43	100,00	1,95
GI56	TAC=CP;EST=OTO;VEH=VL	HGM	1,59	1,59	100,00	1,95
GI33	VIS=TOP;EDAD=(20-27];EST=INV;LUM=SI	HGM	1,03	0,95	92,31	1,80
GI25	EDAD=(20-27];CAU=CON;ANCAL=EST;LUM=SI	HGM	1,67	1,51	90,48	1,76
GI01	TAC=SV;VEH=MOT;APAV=SI;LUM=SI	HGM	2,30	2,06	89,66	1,75
GI62	ANARC=EST;TAC=SV;VEH=MOT;LUM=SI	HGM	1,83	1,59	86,96	1,70
GI10	SEX=H;TAC=CP;APAV=N;VEH=VL	HGM	1,03	0,87	84,62	1,65
GI02	TAC=SV;VEH=MOT;APAV=NE;EST=VER	HGM	1,51	1,27	84,21	1,64
GI06	CAU=COF;HORA=(18-24];APAV=NE	HGM	1,98	1,59	80,00	1,56
GI23	ANCAL=ANC;CAU=CON;VEH=MOT;TAC=SV	HGM	6,19	4,76	76,92	1,50
GI32	VIS=SR;CAU=CON;LUM=SI;ANCAR=EST	HGM	5,71	4,29	75,00	1,46
GI03	TAC=SV;VEH=MOT;APAV=NE	HGM	4,05	2,94	72,55	1,42
GI26	EDAD=(27-60];VEH=MOT;CAU=CON;APAV=NE	HGM	2,22	1,59	71,43	1,39
GI61	OCU=[2];CAT=BT;LUM=SI;DIA=L	HGM	3,02	2,14	71,05	1,39
GI35	OCU=[1];VEH=MOT;TAC=SV	HGM	10,79	7,62	70,59	1,38
GI43	ANCAR=EST;LUM=SI;CAT=BT	HGM	8,10	5,63	69,61	1,36
GI20	LUM=SI;EST=INV	HGM	9,60	6,43	66,94	1,31
GI55	TAC=CP	HGM	8,65	5,71	66,06	1,29
GI27	VEH=MOT;CAU=CON;APAV=SI	HGM	9,37	6,03	64,41	1,26

Tabla 18.- Reglas para accidentes HGM con GI.

Con los patrones obtenidos cabe destacar que el tipo de vehículo involucrado en estos accidentes suele ser un vehículo ligero (GI56 y GI10), y que los días festivos aumentan la probabilidad de estos accidentes (GI38). Este resultado es coherente con la tendencia de la accidentalidad observada en España, dónde la mayoría de los accidentes mortales en carretera ocurren durante el fin de semana (el 31,4% de los accidentes de 2009 ocurrieron en fin de semana, en ellos se produjeron un total de 814 muertes, lo que representa el 38,4% de las muertes por accidente de tráfico del año 2009 (Ministerio del Interior, 2009)).

También es de destacar que en la regla GI10 aparece un factor que hace referencia a los márgenes de la carretera, por lo que medidas específicas en las secciones de

carretera con tránsito peatonal podrían ayudar a mejorar la seguridad vial de las mismas.

Otra de las políticas señaladas en la "Estrategia de Seguridad Vial 2011-2020" (DGT, 2011) se centra en el estudio de los accidentes producidos por salida de la vía. Con esta investigación se obtienen 6 patrones de accidentes graves para esta tipología de accidentes (reglas GI01, GI62, GI02, GI23, GI03 y GI35):

- Regla GI01: muestra un patrón para accidentes con motocicletas en carreteras con arcén pavimentado y sin iluminación. La probabilidad de esta regla es de casi un 90% y el support es de un 2%.

- Regla GI62: describe un patrón muy similar a la regla GI01. Sin embargo, en éste caso son accidentes que suceden en carreteras con arcén estrecho (menor de 1,5 m), en lugar de hacer referencia al arcén pavimentado (como en la regla GI01). Los valores de probabilidad y support son un 87% y un 1,6% respectivamente.

- Regla GI03: muestra un patrón para accidentes con motocicletas en carreteras en las que el arcén pavimentado es inexistente o impracticable. La probabilidad de esta regla es de casi un 73% y el support es de un 2,9%.

- Regla GI02: describe el mismo patrón que la regla GI03, pero puntualizando que estos accidentes ocurren en verano, y con un valor de probabilidad algo superior (84%).

- Regla GI35: describe accidentes por salida de la vía de motocicletas con un ocupante (que hace referencia al propio conductor). Presenta unos valores de probabilidad y support de casi un 71% y 8% respectivamente. Se puede observar que no se ha obtenido ningún patrón de estos accidentes con número de ocupantes igual a 2. Por lo que los resultados indican que hay mayor gravedad en estos accidentes cuando en la motocicleta circula el conductor sin acompañante.

- Regla GI23: identifica salidas de la vía con motocicletas por causas debidas al conductor en carreteras con ancho de calzada mayor de 7 m. La probabilidad de esta regla es de casi un 77% siendo el support también muy elevado (4,8%).

La principal característica de los patrones obtenidos para accidentes por salida de la vía es que todos se relacionan con las motocicletas, y resultan ser accidentes HGM. Estudios previos (Daniello and Gabler, 2011; Montella and Pernetti, 2010; Perandones et al., 2008, Montella et al., 2012b) también han indicado que estos accidentes están relacionados con las consecuencias más severas. En este sentido es importante que las Administraciones competentes realicen esfuerzos orientados a disminuir este tipo de accidentes.

También se ha identificado la gravedad de estos accidentes con condiciones de iluminación insuficientes; este resultado también ha sido mostrado en otros estudios que han analizado la gravedad de los accidentes con motocicletas (Quddus et al., 2002; Rifaat et al., 2011; Savolainen and Mannering, 2007).

Los accidentes en cuyos factores concurrentes aparecen causas asociadas al estado de la vía son otra de las prioridades marcadas como línea de trabajo en "Estrategia de Seguridad Vial 2011-2020" (DGT, 2011). En 4 de los 7 patrones obtenidos para estos accidentes (reglas GI01, GI62, GI02, GI03), aparece al menos un factor que hace referencia a los márgenes de la carretera. Por tanto, actuaciones específicas en los márgenes de estas carreteras podría ayudar a mejorar la seguridad vial, así como, a proteger a usuarios vulnerables, tales como los conductores de las motocicletas.

En la regla GI23 también se identifican factores relativos a la carretera: accidentes que se producen en carreteras con ancho de calzada mayor de 7 m, y que son debidos al conductor. Este patrón puede estar relacionado con una mayor velocidad de circulación de los conductores de motocicletas, ya que las características de la vía lo permiten. De cara a mitigar estos patrones, la nueva mediad de seguridad vial propuesta en España sobre la reducción de la velocidad máxima autorizada para circular en las carreteras convencionales de un carril por sentido, puede resultar efectiva (se espera que entre en vigor con el nuevo código de circulación en verano de 2013).

Atendiendo al tipo de vehículo, se han obtenido dos patrones adicionales para motocicletas (en los que no hacen referencia a la tipología particular del accidente) y que se relacionan de nuevo con los márgenes de la carretera:

- Regla GI27: identifica accidentes con motocicletas por causas debidas al conductor en carreteras con arcén pavimentado. La probabilidad de esta regla es un 64% y el support es muy elevado (6%).

- Regla GI26: muestra un patrón similar a la regla GI27, en el que el arcén pavimentado es inexistente o impracticable y la edad del conductor está comprendida entre 27 y 60 años. Presenta unos valores de probabilidad y support de un 71% y 1,6% respectivamente.

Atendiendo a la edad de los conductores se han obtenido 3 patrones particulares (reglas GI33, GI25 y GI26). Salvo la regla GI26 (que ha sido descrita anteriormente), todas hacen referencia a conductores jóvenes, con edad comprendida entre 21 y 27 años:

- Regla GI33: identifica accidentes HGM para estos conductores cuando la visibilidad está restringida por la topografía, se producen en invierno, en carreteras sin iluminación; con una probabilidad del 92%. La probabilidad de que los accidentes HGM aumenta en invierno en carreteras sin iluminación también es identificada en la regla GI20.

- Regla GI25: muestra accidentes HGM para conductores jóvenes por causas debidas al conductor cuando el ancho de calzada estrecho (menor de 6 m), en carreteras sin iluminación con una probabilidad del 90%.

Estos patrones muestran un problema particular que puede estar relacionado con la inexperiencia de los conductores jóvenes en la conducción en carreteras

convencionales. Particularmente cuando las condiciones de la carretera no son las más favorables, tienen visibilidad restringida (regla GI33) o son carreteras estrechas (reglas GI25).

Los conductores jóvenes son potencialmente vulnerables a sufrir accidentes en carreteras convencionales dada su inexperiencia y falta de madurez en la conducción, comparada con conductores mayores y más experimentados (Mayhew, et al., 2003; McCartt et al., 2009; Peek-Asa et al., 2010). De hecho, está demostrado que la inexperiencia se relaciona con el riesgo del accidente, particularmente en los años siguientes a la obtención del permiso de circulación (Young Drivers: The Road to Safety, 2006).

De los 4 patrones restantes, 3 de ellos describen accidentes que se producen en noches sin iluminación (reglas GI32, GI61, GI43):

- Regla GI32: identifica también accidentes HGM en carreteras sin iluminación con arcén estrecho (menor de 1,5 m) cuando la visibilidad no presenta restricciones y son producidos por causas debidas al conductor. La probabilidad de esta regla es de un 75% y el support es de casi un 4,3%.

- Regla GI61: muestra accidentes HGM en carreteras sin iluminación y condiciones atmosféricas de buen tiempo, cuando los ocupantes del vehículo involucrado son 2 (conductor más un ocupante), en días laborables, con una probabilidad de casi un 71%.

- Regla GI43: muestra accidentes HGM en carreteras sin iluminación con arcén estrecho (menor de 1,5 m) y condiciones atmosféricas de buen tiempo, con una probabilidad de casi un 70%.

Con estos patrones se muestra una correlación entre la falta de iluminación de la carretera y la gravedad del accidente. Resultados similares han sido mostrados en otros estudios. Gray et al. (2008) identificaron que los accidentes más graves se producen durante la noche. El mismo resultado también fue mostrado en el estudio de Abdel-Aty (2003) y en Helai et al. (2008). De Oña et al. (2011) y De Oña et al. (2013) también subrayaron que la probabilidad de accidente grave en carreteras convencionales aumenta en carreteras sin iluminación.

En el último patrón de accidentes HGM (regla GI06) se identifican accidentes por causas debidas a una combinación de factores, durante la franja horaria de 18 a 24 horas, en carreteras con arcén pavimentado inexistente o impracticable. La probabilidad de esta regla es de 80% y el support de 1,6%.

➢ Patrones de accidentes HL.

En la Tabla 19 se describen los 17 patrones de accidentes HL, ordenados según el valor de la confidence. Los valores de confidence de las reglas HL varían desde un 100% (en las reglas GI18 y GI37) hasta un valor de 60,7% (regla GI69). Respecto al support de las reglas, se observa que varía de 17,6% (regla GI47) a 0,87% (regla GI12). El valor

de population varía de 28,1% en la regla GI47, hasta casi el mínimo establecido (un 1,1%) en la regla GI12. Y respecto a los valores del lift, estos varían desde un 1,25 (regla GI69) hasta 2,05 en las reglas GI18 y GI37.

NUM.	REGLAS (SI…)	ENTONCES	Po%	S%	C%	Lift
GI18	CAT=BT;LUM=DIA;SEX=M;OCU=[>2]	HL	1,03	1,03	100,00	2,05
GI37	DIA=AF;HORA=(6-12];OCU=[1];MVIAL=CYM	HL	1,67	1,67	100,00	2,05
GI22	CAT=LL;DIA=F;ANARC=NE;MVIAL=CYM	HL	1,51	1,43	94,74	1,94
GI41	ANCAR=EST;LUM=DIA;EDAD=(20-27];VEH=VL	HL	2,38	2,06	86,67	1,78
GI12	TAC=SV;VEH=OT;ANCAR=MED	HL	1,11	0,87	78,57	1,61
GI28	HORA=[0-6];APAV=SI;EDAD=(20-27]	HL	2,22	1,67	75,00	1,54
GI50	CAU=COF;HORA=[0-6];ANCAL=ANC	HL	1,27	0,95	75,00	1,54
GI60	TAC=VUE;LUM=DIA;HORA=(12-18]	HL	1,83	1,35	73,91	1,52
GI52	MVIAL=CYM;CAU=COF;HORA=[6-12];VEH=VL	HL	2,14	1,51	70,37	1,44
GI19	CAT=BT;LUM=DIA;SEX=M	HL	8,41	5,87	69,81	1,43
GI13	SEX=M ;LUM=DIA	HL	9,92	6,83	68,80	1,41
GI66	ANARC=NE; SEX=M	HL	7,06	4,60	65,17	1,34
GI07	HORA=(6-12];VEH=VL;CAT=BT	HL	13,89	8,89	64,00	1,31
GI16	TAC=SV;CAU=COF;ANARC=EST	HL	4,52	2,86	63,16	1,30
GI47	ANCAR=MED;VEH=VL;TAC=SV;HER=[1]	HL	28,10	17,62	62,71	1,29
GI63	ANARC=EST;TAC=SV;VEH=VL;HER=[1]	HL	17,46	10,79	61,82	1,27
GI69	BAR=N;TAC=VUE	HL	6,27	3,81	60,76	1,25

Tabla 19.- Reglas para accidentes HL con GI.

Atendiendo al tipo de accidente, 4 reglas describen patrones para salidas de la vía (reglas GI12, GI16, GI47 y GI63) y 2 reglas describen patrones para los vuelcos (reglas GI60 y GI69). Los patrones particulares se describen a continuación:

- Regla GI12: describe accidentes por salida de la vía cuando el vehículo involucrado es otro, en carreteras con ancho de carril comprendido entre 3,25 y 3,75 m. Tiene una probabilidad de casi un 79% y un support de 0,87%.

- Regla GI16: describe accidentes por salida de la vía por causas debidas a una combinación de factores en carreteras con arcén estrecho (menor de 1,5 m). Esta regla tiene un support de casi un 2,9%, y una probabilidad de un 63%.

- Regla GI47: muestra estos accidentes cuando el vehículo involucrado es un vehículo ligero, con un herido, y que se producen en carreteras con ancho de carril entre 3,25 y 3,75 m. El support de esta regla es uno de los más elevados (17,6%), siendo la probabilidad también elevada (63%).

- Regla GI63: muestra el mismo patrón que la regla GI47, pero haciendo referencia al ancho de arcén (en lugar del ancho de carril). El support de esta regla es también muy elevado (10,8%), y la probabilidad es de un 62%.

En la regla GI12, se muestran accidentes por salida de la vía en los que el vehículo involucrado es otra tipología, lo cual incluye camiones de PM>3500, autobuses, etc. Por tanto, se muestra que los accidentes por salida de la vía con estos vehículos tienen consecuencias leves para el conductor y/o ocupantes de dicho vehículo (regla GI12). Desde la perspectiva de la seguridad vial, es destacado que los vehículos más pesados proporcionan mayor protección a sus conductores (Bedard et al., 2002).

También se han obtenido patrones para accidentes leves por salida de la vía cuando el vehículo involucrado es un turismo (reglas GI47 y GI63); en contraste con los patrones obtenidos para accidentes graves por salida de la vía, en los que el vehículo implicado era una motocicleta (reglas GI01, GI62, GI02, GI23, GI03 y GI35).

Respecto a los patrones identificados para los accidentes por vuelco, se observan las siguientes características:

- Regla GI69: identifica estos accidentes en carretas sin barreras de seguridad con una probabilidad de un 61%.
- Regla GI60: también identifica los accidentes producidos por vuelco cuando la luminosidad es igual a día en la franja horaria de 12 a 18 horas. La probabilidad es de un 74% y el support de un 1,3%.

La gravedad de los accidentes producidos por vuelco, en las carreteras analizadas, es HL, en la franja horaria coincidente con horas laborales (de 12 a 18 h). Este resultado es coherente con la tendencia observada en España, en la que la mayoría de los accidentes se producen durante la franja diurna. Sin embargo, la gravedad de los mismos es menor (en el año 2009 el índice de gravedad en el tramo horario nocturno fue de 6,5 muertos por cada 100 accidentes frente a los 4,6 por cada 100 accidentes para el resto del día (Ministerio del Interior, 2009)).

Atendiendo a la iluminación, cuando es igual a pleno día, se tienen 4 patrones particulares (reglas GI18, GI41, GI19 y GI13):

- Regla GI13: muestra un patrón para accidentes HL que se producen a pleno día cuando el conductor es una mujer, con una probabilidad del 69%. La regla GI19 describe el mismo patrón, añadiendo una nueva variable (condiciones atmosféricas de buen tiempo). La probabilidad de esta regla es algo superior: 70%.
- Regla GI18: muestra el mismo patrón que la regla GI19, añadiendo una nueva variable, número de ocupantes (igual a 2). La probabilidad de esta regla es de un 100%.
- Regla GI41: identifica accidentes cuando la iluminación es igual a pleno día, los conductores son jóvenes (de 20 a 27 años), el vehículo implicado es un vehículo ligero y que se producen en carreteras con ancho de carril estrecho (< 3,25 m). La probabilidad y support de esta regla son de un 87% y un 2% respectivamente.

Desde el punto de vista de la seguridad vial los accidentes analizados (en carreteras convencionales con un solo vehículo involucrado) que ocurren durante el día tienen consecuencias menos graves; particularmente cuando el conductor es una mujer (regla GI13, GI18 y GI19), o cuando el vehículo es un vehículo ligero y el conductor tiene edad de 20 a 27 años (regla GI41).

En la regla GI66 se muestra otro accidente leve para mujeres, particularmente, en carreteras con arcén inexistente o impracticable. La probabilidad de esta regla es de un 65% y el support de un 4,6%.

Atendiendo a las marcas viales de la carretera, cuando están presentes en la separación de carriles y márgenes, se muestran 2 patrones (reglas GI37, GI22 y GI52):

- Regla GI37: muestra accidentes HL en estas carreteras, con una probabilidad de un 100% y un support de 1,67%, cuando ocurren en la franja horaria de 6 a 12 h, hay un solo ocupante involucrado y se producen en días anteriores a festivos.

- Regla GI22: muestra accidentes HL en estas carreteras, cuando las condiciones atmosféricas son de lluvia ligera, el tipo de día es festivo y se producen en carreteras con arcén inexistente o impracticable. La probabilidad de esta regla es de un 95% y el support también es de 1,4%.

- Regla GI52: identifica accidentes en estas carreteras, en la franja horaria de 6 a 12, por causas debidas a una combinación de factores, cuando el vehículo involucrado en el accidente es un vehículo ligero. La probabilidad es de un 70% y el support es de 1,5%.

De nuevo se pone de manifiesto que los accidentes que ocurren durante el día tienen consecuencias menos graves (regla GI37); que los días festivos se relaciona con la probabilidad de ocurrencia de un accidente (reglas GI37 y GI22); y que la gravedad de los accidentes que ocurren en horas laborales es inferior (reglas GI37 y GI52).

Atendiendo particularmente a la franja horaria en la que se produce el accidente se observan los siguientes patrones:

- Regla GI28: muestra accidentes HL que se producen en la franja horaria de 0 a 6 horas, con conductores jóvenes (de 21 a 27 años) en carretas con arcén pavimentado. La probabilidad es de un 75% y el support es de 1,7%.

- Regla GI50: muestra también accidentes HL que se producen de 0 a 6 horas, por causas debidas a una combinación de factores, en carreteras con ancho de calzada mayor de 7 m. La probabilidad es también de un 75% y el support algo inferior 0,95%.

- Regla GI07: identifica los accidentes HL en la franja horaria de 6 a 12, en los que el vehículo involucrado es un vehículo ligero, en condiciones atmosféricas de buen tiempo. Siendo la probabilidad de un 64% y el support de un 8,9%.

5.4.2. Análisis de reglas obtenidas con RGI.

En la Tabla 20 se muestra el número de reglas obtenidas en los diferentes pasos del método IRNV, indicándose la variable raíz que inicia la construcción de cada árbol, así como el número de reglas que se obtiene en el *training* y en el *test* en cada uno de los árboles creados. Además, se recoge el número de reglas que forman el conjunto final de RDs formado por aquellas reglas que tienen un valor de lift≥1,2.

Si no se aplicase el método IRNV sólo se obtendría un ADD en el que el nodo raíz es la variable SEX (ADD_1 en la Tabla 20). De este árbol se pueden extraer 8 reglas del training, de las que sólo 6 verifican el test, y 5 tienen un lift≥1,2.

Al aplicar el método IRNV con el criterio de partición RGI de nuevo se construyen 19 ADDs, de los que pueden extraerse un total de 174 reglas del *training*. De estas reglas, 81 verifican el *test* y 76 tienen un valor de lift≥1,2. Se observa que con este método se obtiene un conocimiento mucho mayor: 76 reglas vs. 5 reglas.

ADDs	IRNV: REGLAS CON RGI			LIFT Conj. Final de RDs
	NODO RAÍZ	TRAINING	TEST	
ADD_1	SEX	8	6	5
ADD_2	TAC	8	2	2
ADD_3	CAU	12	5	5
ADD_4	CAT	15	7	7
ADD_5	VEH	6	1	1
ADD_6	LUM	16	7	6
ADD_7	HER	5	2	2
ADD_8	VIS	10	3	3
ADD_9	ANCAL	7	3	3
ADD_{10}	EDAD	7	3	3
ADD_{11}	ANCAR	4	3	3
ADD_{12}	HORA	11	6	6
ADD_{13}	DIA	11	5	5
ADD_{14}	BAR	6	4	4
ADD_{15}	EST	10	4	3
ADD_{16}	MVIAL	7	3	3
ADD_{17}	OCU	5	1	1
ADD_{18}	ANARC	13	10	10
ADD_{19}	ARPAV	13	6	4
TOTAL		**174**	**81**	**76**

Tabla 20.- Número de reglas obtenidas con RGI.

El nodo raíz LUM es el que mayor número de reglas genera en el *training* (16 reglas), más que el ADD_1 con 8 reglas. Respecto a las reglas validadas en el *test* con el árbol ADD_{18} se validan el mayor número de reglas (10 reglas).

Finalmente, el árbol que más patrones validados produce en el conjunto final de RDs es el ADD_{18}, en el que el nodo raíz es la variable ANARC (10 reglas), seguido del árbol ADD_4 (y nodo raíz CAT) con 7 reglas.

Atendiendo a la gravedad de los patrones obtenidos, cabe destacar que se obtienen un 40% de patrones que hacen referencia a accidentes graves o mortales y un 60% a accidentes leves (30 HGM frente a 46 HL).

➢ Validación del nodo raíz.

A continuación se realiza la verificación del nodo raíz impuesto en cada regla (reglas que provienen de los árboles ADD_2 - ADD_{19}, codificadas RGI06-RGI76). Al igual que en el caso de GI, las 5 reglas que provienen del ADD_1 (codificadas como RGI01-RGI05) no necesitan verificación del nodo raíz.

Como puede observase en la Tabla 21, de las 71 reglas evaluadas 52 reglas cumplen la condición 1 del ratio de confidences (ecuación 26) y 19 reglas no la cumplen. Respecto a la condición 2 del ratio de supports (ecuación 27), se observa que la cumplen un mayor número de reglas, 63 de las 71 totales.

NUM	REGLAS RD	CONDICIÓN 1		CONDICIÓN 2		PATRÓN FINAL
		C(A→B)/C(A⁻→B)	≥1,03	S(A→B)/S(A⁻→B)	≥0,2	
RGI06	*TAC=CP*;CAT=BT;APAV=SI;ANCAR=MED	1,57	SI	0,17	NO	RD-
RGI07	CAT=BT;SEX=M;LUM=DIA;OCU=[>2]	1,08	SI	1,00	SI	RD
RGI08	*CAT=O*;OCU=[1];ANCAR=MED;TAC=SV	1,43	SI	0,05	NO	RD-
RGI09	VIS=SR;CAU=CON;VEH=OT;EDAD=(27-60]	1,13	SI	1,00	SI	RD
RGI10	ANCAL=ANC;CAU=CON;ANCAR=EST;ANARC=EST	1,26	SI	0,52	SI	RD
RGI11	ANCAL=MED;BAR=N;CAU=CON;TAC=VUE	1,39	SI	0,31	SI	RD
RGI12	EDAD=(20-27];CAU=CON;SEX=M;DIA=L	1,37	SI	0,35	SI	RD
RGI13	ANCAR=MED;CAT=O;HER=[1];TAC=SV	1,44	SI	0,92	SI	RD
RGI14	HORA=(12-18];CAT=LL;VEH=VL;TAC=SV	1,13	SI	0,30	SI	RD
RGI15	DIA=PF;CAT=BT;HORA=[6-12];LUM=DIA	1,53	SI	0,25	SI	RD
RGI16	DIA=F;TAC=CP;ANCAR=MED	1,51	SI	0,33	SI	RD
RGI17	EST=OTO;TAC=CP;VEH=VL	1,15	SI	0,32	SI	RD
RGI18	MVIAL=CYM;CAU=CON;CAT=BT;VEH=OT	1,18	SI	0,79	SI	RD
RGI19	ANARC=MED;ANCAL=ANC;CAT=BT;VIS=TOP	1,29	SI	0,29	SI	RD
RGI20	*APAV=N*;BAR=N;SEX=M;VEH=VL	1,19	SI	0,18	NO	RD-
RGI21	TAC=SV;CAU=CON;VEH=MOT;CAT=BT	1,10	SI	0,77	SI	RD
RGI22	CAU=CON;SEX=H;CAT=BT;VEH=OT	1,07	SI	0,76	SI	RD
RGI23	*CAU=CON*;SEX=H;CAT=LF;MVIAL=CYM	1,00	NO	1,00	SI	RD-
RGI24	CAU=CON;SEX=H;CAT=LL;VEH=VL	1,76	SI	0,37	SI	RD
RGI25	*CAU=CON*;SEX=H;CAT=LL;TAC=SV	0,97	NO	0,76	SI	RD-
RGI26	*CAU=CON*;SEX=M;VEH=VL;LUM=DIA	0,95	NO	0,74	SI	RD-
RGI27	*CAT=BT*;SEX=H;TAC=SV;VEH=MOT	1,01	NO	0,99	SI	RD-
RGI28	*CAT=BT*;SEX=H;TAC=OT;ANCAR=MED	0,96	NO	0,83	SI	RD-
RGI29	*CAT=BT*;SEX=M;LUM=DIA;OCU=[1]	1,00	NO	0,85	SI	RD-
RGI30	CAT=BT;SEX=M;LUM=SI;VEH=VL	1,05	SI	1,36	SI	RD
RGI31	CAT=LL;TAC=SV;DIA=F;VEH=VL	1,53	SI	0,26	SI	RD
RGI32	VEH=VL;TAC=CP;HER=[1];CAT=BT	1,12	SI	0,96	SI	RD
RGI33	LUM=DIA;SEX=M;MVIAL=CYM;OCU=[1]	1,11	SI	0,79	SI	RD
RGI34	*LUM=INS*;BAR=N;EST=INV;SEX=H	1,11	SI	0,10	NO	RD-
RGI35	LUM=SI;CAT=BT;SEX=H;ANCAR=EST	1,20	SI	0,47	SI	RD
RGI36	*LUM=SI*;CAT=LL;VEH=VL;SEX=H	0,95	NO	0,33	SI	RD-
RGI37	*LUM=ATA*;BAR=N;CAT=BT;ANARC=EST	1,32	SI	0,08	NO	RD-
RGI38	*LUM=ATA*;BAR=N;CAT=BT;ANARC=NE	1,16	SI	0,07	NO	RD-
RGI39	*HER=[1]*;OCU=[1];BAR=N;SEX=M	1,00	NO	0,99	SI	RD-

Tabla 21.- Evaluación de nodo raíz en las reglas RGI.

RGI40	HER=[>1];TAC=SV;CAU=CON;SEX=M	1,04	SI	0,29	SI	RD
RGI41	*VIS=SR*;CAU=CON;VEH=VL;TAC=CP	0,97	NO	0,81	SI	RD-
RGI42	VIS=TOP;BAR=N;TAC=SV;EDAD=<=20	1,28	SI	0,33	SI	RD
RGI43	ANCAL=MED;BAR=N;CAU=CON;TAC=CP	1,20	SI	0,33	SI	RD
RGI44	EDAD=(27-60];BAR=N;CAT=LL;TAC=SV	1,07	SI	0,65	SI	RD
RGI45	EDAD=(20-27];CAU=CON;SEX=H;CAT=LL	1,07	SI	0,29	SI	RD
RGI46	ANCAR=MED;CAT=BT;VEH=VL;TAC=CP	1,07	SI	0,78	SI	RD
RGI47	*ANCAR=MED*;CAT=LL;TAC=SV;MVIAL=CYM	1,01	NO	0,78	SI	RD-
RGI48	*HORA=(12-18]*;CAT=BT;LUM=DIA;SEX=M	0,89	NO	0,51	SI	RD-
RGI49	HORA=[0-6];CAU=CON;OCU=[2];TAC=SV	1,32	SI	0,35	SI	RD
RGI50	HORA=(18-24];CAT=BT;LUM=INS;OCU=[1]	1,06	SI	0,55	SI	RD
RGI51	*HORA=(18-24]*;CAT=BT;LUM=SI;BAR=N	1,02	NO	0,52	SI	RD-
RGI52	HORA=(6-12];VEH=VL;CAU=COF;CAT=BT	1,64	SI	0,34	SI	RD
RGI53	DIA=L;CAU=CON;SEX=M;VEH=VL	1,09	SI	0,52	SI	RD
RGI54	*DIA=PF*;CAT=BT;HORA=[0-6];ANCAR=MED	1,19	SI	0,19	NO	RD-
RGI55	DIA=F;TAC=SV;CAT=BT;VIS=TOP	1,30	SI	0,45	SI	RD
RGI56	*BAR=N*;SEX=H;TAC=CP;CAT=BT	1,01	NO	0,98	SI	RD-
RGI57	*BAR=N*;SEX=H;TAC=OT;ANCAR=MED	1,00	NO	1,00	SI	RD-
RGI58	*BAR=N*;SEX=M;VEH=VL;CAU=CON	1,00	NO	0,95	SI	RD-
RGI59	*BAR=N*;SEX=M;VEH=VL;CAU=COF	1,00	NO	1,00	SI	RD-
RGI60	EST=INV;CAT=BT;LUM=SI;SEX=H	1,12	SI	0,34	SI	RD
RGI61	EST=OTO;TAC=SV;CAU=CON;SEX=M	1,15	SI	0,32	SI	RD
RGI62	MVIAL=CYM;CAU=CON;CAT=LL;TAC=SV	1,03	SI	0,88	SI	RD
RGI63	MVIAL=NE;ANARC=NE;SEX=H;TAC=SV	1,11	SI	0,22	SI	RD
RGI64	*OCU=[2]*;CAT=BT;HER=[1];LUM=DIA	1,56	SI	0,07	NO	RD-
RGI65	ANARC=EST;TAC=SV;CAU=COF;ANCAR=MED	1,18	SI	0,59	SI	RD
RGI66	*ANARC=EST*;TAC=VUE;ANCAR=MED;BAR=N	1,00	NO	0,38	SI	RD-
RGI67	ANARC=EST;TAC=CP;ANCAR=MED;CAT=BT	1,22	SI	0,71	SI	RD
RGI68	ANARC=NE;SEX=H;CAT=BT;LUM=SI	1,04	SI	0,50	SI	RD
RGI69	ANARC=NE;SEX=H;CAT=BT;LUM=SUF	1,17	SI	0,71	SI	RD
RGI70	ANARC=NE;SEX=H;CAT=BT;LUM=ATA	1,36	SI	0,64	SI	RD
RGI71	*ANARC=NE*;SEX=H;CAT=LL;TAC=SV	0,99	NO	0,51	SI	RD-
RGI72	*ANARC=NE*;SEX=M;TAC=SV;CAT=BT	0,98	NO	0,46	SI	RD-
RGI73	ANARC=MED;ANCAL=ANC;CAT=LL;SEX=H	1,42	SI	0,42	SI	RD
RGI74	APAV=SI;TAC=SV;VEH=VL;CAU=COF	1,23	SI	0,59	SI	RD
RGI75	APAV=SI;TAC=VUE;BAR=N;SEX=H	1,07	SI	0,64	SI	RD
RGI76	APAV=NE;CAU=CON;HORA=6-12];BAR=N	1,25	SI	0,39	SI	RD
NÚMERO DE REGLAS: RD		**CONDICIÓN 1**	52	**CONDICIÓN 2**	63	44
NÚMERO DE REGLAS: RD-			19		8	27

Continuación de Tabla 21.

Como resultado final se tiene que 44 reglas cumplen simultáneamente las 2 condiciones. Por lo tanto, en 44 reglas se debe mantener el nodo raíz impuesto (y en estas reglas se verifica la regla ampliada *RD*), mientras que en las 27 reglas restantes (sombreadas en la Tabla 21) no se cumple al menos una de las dos condiciones impuestas, por lo que se extrae la regla simple RD^- (que son los patrones RD pero eliminando la primera variable que los forma, que es el nodo raíz impuesto, marcado en cursiva en las reglas de la Tabla 21).

➢ Validación de las variables de la regla: condición del incremento del lift (CIL).

A continuación se verifican las variables que conforman los patrones finales. Esta verificación se realiza tanto en las 5 reglas que provienen del ADD_1 como en las obtenidas después de la verificación del nodo raíz. Por tanto, se verifican 76 reglas, de las que 49 son *RD* y 27 son RD^-.

En las Tablas 22 y 23 se muestran los resultados de la CIL de las reglas HGM y HL, respectivamente. El procedimiento de cálculo es el mismo que se indica en las Tablas 14 y 15. La regla final (con CIL≥1,03) se muestra en sombreado en negrita.

NUM	COMPARACIÓN - REGLAS BASE		Lift	Incr. Lift
	SEX=H	HGM	**1,05**	
	SEX=H;TAC=SV	HGM	1,05	0,79%
	SEX=H;TAC=SV;CAU=CON	HGM	1,08	2,29%
RGI01	SEX=H;TAC=SV;CAU=CON;VEH=MOT	HGM	1,33	23,13%
	SEX=H	HGM	1,05	
	SEX=H;TAC=CP	HGM	**1,32**	**26,24%**
	SEX=H;TAC=CP;CAT=BT	HGM	1,31	-0,46%
RGI02	SEX=H;TAC=CP;CAT=BT;APAV=SI	HGM	1,54	17,19%
	SEX=H	HGM	1,05	
	SEX=H;TAC=CP	HGM	**1,32**	**26,24%**
	SEX=H;TAC=CP;CAT=BT	HGM	1,31	-0,46%
RGI03	SEX=H;TAC=CP;CAT=BT;APAV=N	HGM	1,30	-1,04%
	CAT=BT	HGM	**1,04**	
	CAT=BT;APAV=SI	HGM	1,03	-0,53%
RGI06	CAT=BT;APAV=SI;ANCAR=MED	HGM	1,00	-3,38%
	ANCAL=ANC	HGM	1,01	
	ANCAL=ANC;CAU=CON	HGM	1,06	4,77%
	ANCAL=ANC;CAU=CON;ANCAR=EST	HGM	1,58	49,09%
RGI10	ANCAL=ANC;CAU=CON;ANCAR=EST;ANARC=EST	HGM	**1,81**	**14,29%**
	DIA=F	HGM	1,10	
	DIA=F;TAC=CP	HGM	1,76	59,74%
RGI16	DIA=F;TAC=CP;ANCAR=MED	HGM	**1,95**	**10,53%**
	MON=OTO	HGM	0,99	
	MON=OTO;TAC=CP	HGM	**1,79**	**79,71%**
RGI17	MON=OTO;TAC=CP;VEH=VL	HGM	1,70	-5,14%
	ANARC=MED	HGM	0,94	
	ANARC=MED;ANCAL=ANC	HGM	0,98	3,85%
	ANARC=MED;ANCAL=ANC;CAT=BT	HGM	1,10	12,73%
RGI19	ANARC=MED;ANCAL=ANC;CAT=BT;VIS=TOP	HGM	1,65	**50,12%**
	TAC=SV	HGM	1,00	
	TAC=SV;CAU=CON	HGM	1,03	3,08%
	TAC=SV;CAU=CON;VEH=MOT	HGM	**1,34**	**30,20%**
RGI21	TAC=SV;CAU=CON;VEH=MOT;CAT=BT	HGM	1,36	1,37%
	CAU=CON	HGM	1,03	
	CAU=CON;SEX=H	HGM	**1,07**	**4,02%**
	CAU=CON;SEX=H;CAT=LL	HGM	0,74	-31,13%
RGI24	CAU=CON;SEX=H;CAT=LL;VEH=VL	HGM	1,22	65,00%
	SEX=H	HGM	**1,05**	
	SEX=H;TAC=SV	HGM	1,05	0,79%
RGI27	SEX=H;TAC=SV;VEH=MOT	HGM	1,32	25,63%
	CAT=BT	HGM	**1,04**	
	CAT=BT;SEX=M	HGM	0,76	-26,43%
	CAT=BT;SEX=M;LUM=SI	HGM	1,44	89,44%
RGI30	CAT=BT;SEX=M;LUM=SI;VEH=VL	HGM	1,33	-7,95%
	VEH=VL	HGM	0,95	
	VEH=VL;TAC=CP	HGM	**1,41**	**47,26%**
	VEH=VL;TAC=CP;HER=[1]	HGM	1,34	-4,52%
RGI32	VEH=VL;TAC=CP;HER=[1];CAT=BT	HGM	1,34	-0,49%
	BAR=N	HGM	1,00	
	BAR=N;MON=INV	HGM	**1,07**	**7,02%**
RGI34	BAR=N;MON=INV;SEX=H	HGM	1,10	2,93%
	LUM=SI	HGM	1,15	
	LUM=SI;CAT=BT	HGM	**1,18**	**3,22%**
	LUM=SI;CAT=BT;SEX=H	HGM	1,17	-1,65%
RGI35	LUM=SI;CAT=BT;SEX=H;ANCAR=EST	HGM	1,41	21,11%
	BAR=N	HGM	1,00	
	BAR=N;CAT=BT	HGM	**1,03**	**3,83%**
RGI38	BAR=N;CAT=BT;ANARC=NE	HGM	1,05	1,79%
	CAU=CON	HGM	**1,03**	
	CAU=CON;VEH=VL	HGM	0,98	-5,05%
RGI41	CAU=CON;VEH=VL;TAC=CP	HGM	1,33	35,96%
	VIS=TOP	HGM	**1,09**	
	VIS=TOP;BAR=N	HGM	1,08	-0,81%
	VIS=TOP;BAR=N;TAC=SV	HGM	1,10	1,82%
RGI42	VIS=TOP;BAR=N;TAC=SV;EDAD=<=20	HGM	1,42	29,19%
	ANCAL=MED	HGM	**0,89**	
	ANCAL=MED;BAR=N	HGM	0,88	-0,33%
	ANCAL=MED;BAR=N;CAU=CON	HGM	0,87	-1,49%
RGI43	ANCAL=MED;BAR=N;CAU=CON;TAC=CP	HGM	1,44	65,61%
	ANCAR=MED	HGM	0,97	
	ANCAR=MED;CAT=BT	HGM	**1,02**	**5,13%**
	ANCAR=MED;CAT=BT;VEH=VL	HGM	0,98	-4,59%
RGI46	ANCAR=MED;CAT=BT;VEH=VL;TAC=CP	HGM	1,49	52,25%
	HORA=[0-6]	HGM	1,06	
	HORA=[0-6];CAU=CON	HGM	1,10	3,47%
	HORA=[0-6];CAU=CON;OCU=[2]	HGM	1,38	25,65%
RGI49	HORA=[0-6];CAU=CON;OCU=[2];TAC=SV	HGM	**1,49**	**7,84%**
	CAT=BT	HGM	1,04	
	CAT=BT;LUM=SI	HGM	**1,18**	**14,28%**
RGI51	CAT=BT;LUM=SI;BAR=N	HGM	1,18	0,02%

Tabla 22.- Aplicación de la Condición del Incremento del Lift en las reglas HGM obtenidas con RGI.

	CAT=BT	HGM	1,04	
	CAT=BT;HORA=[0-6]	HGM	**1,09**	**5,37%**
RGI54	CAT=BT;HORA=[0-6];ANCAR=MED	HGM	1,02	-6,72%
	DIA=F	HGM	**1,10**	
	DIA=F;TAC=SV	HGM	1,11	0,81%
	DIA=F;TAC=SV;CAT=BT	HGM	1,19	7,26%
RGI55	DIA=F;TAC=SV;CAT=BT;VIS=TOP	HGM	1,50	25,78%
	SEX=H	HGM	1,05	
	SEX=H;TAC=CP	HGM	**1,32**	**26,24%**
RGI56	SEX=H;TAC=CP;CAT=BT	HGM	1,31	-0,46%
	MON=INV	HGM	1,07	
	MON=INV;CAT=BT	HGM	1,12	4,92%
	MON=INV;CAT=BT;LUM=SI	HGM	**1,35**	**19,82%**
RGI60	MON=INV;CAT=BT;LUM=SI;SEX=H	HGM	1,31	-2,85%
	MVIAL=NE	HGM	**1,01**	
	MVIAL=NE;ANARC=NE	HGM	1,03	1,65%
	MVIAL=NE;ANARC=NE;SEX=H	HGM	1,17	13,07%
RGI63	MVIAL=NE;ANARC=NE;SEX=H;TAC=SV	HGM	1,23	5,21%
	ANARC=EST	HGM	0,99	
	ANARC=EST;TAC=CP	HGM	**1,50**	**51,16%**
	ANARC=EST;TAC=CP;ANCAR=MED	HGM	1,54	2,92%
RGI67	ANARC=EST;TAC=CP;ANCAR=MED;CAT=BT	HGM	1,57	1,60%
	ANARC=NE	HGM	1,02	
	ANARC=NE;SEX=H	HGM	**1,08**	**5,81%**
	ANARC=NE;SEX=H;CAT=BT	HGM	1,11	2,70%
RGI68	ANARC=NE;SEX=H;CAT=BT;LUM=SI	HGM	1,21	9,58%
	ANARC=NE	HGM	1,02	
	ANARC=NE;SEX=H	HGM	**1,08**	**5,81%**
	ANARC=NE;SEX=H;CAT=BT	HGM	1,11	2,70%
RGI70	ANARC=NE;SEX=H;CAT=BT;LUM=ATA	HGM	1,30	17,37%

Continuación de Tabla 22.

Tras la verificación de la CIL en las 30 reglas HGM (ver Tabla 20) y la selección del patrón (con su correspondiente número de variables), se comprueba que 5 patrones son iguales (RGI02=RGI03=RGI56; RGI01=RGI27; RGI06=RGI30; RGI68=RGI70). Por tanto, el número de reglas HGM se reduce a 25, siendo su distribución según el número de variables, la siguiente: 7 reglas quedan reducidas a una sola variable, 12 reglas están formadas por 2 variables, 3 reglas por 3 variables, y 3 reglas están formadas por las 4 variables.

NUM	COMPARACIÓN - REGLAS BASE		Lift	Incr. Lift
	SEX=M	HL	1,26	
	SEX=M;VEH=VL	HL	1,30	**3,11%**
	SEX=M;VEH=VL;TAC=SV	HL	1,30	0,24%
RGI04	SEX=M;VEH=VL;TAC=SV;CAU=CON	HL	1,27	-2,73%
	SEX=M	HL	1,26	
	SEX=M;VEH=VL	HL	**1,30**	**3,11%**
	SEX=M;VEH=VL;TAC=SV	HL	1,30	0,24%
RGI05	SEX=M;VEH=VL;TAC=SV;CAU=COF	HL	1,47	12,45%
	CAT=BT	HL	0,96	
	CAT=BT;SEX=M	HL	1,25	29,98%
	CAT=BT;SEX=M;LUM=DIA	HL	1,43	14,64%
RGI07	CAT=BT;SEX=M;LUM=DIA;OCU=[>2]	HL	**2,05**	**43,24%**
	OCU=[1]	HL	1,00	
	OCU=[1];ANCAR=MED	HL	**1,05**	**4,89%**
RGI08	OCU=[1];ANCAR=MED;TAC=SV	HL	1,05	0,82%
	VIS=SR	HL	**1,02**	
	VIS=SR;CAU=CON	HL	0,98	-4,35%
	VIS=SR;CAU=CON;VEH=OT	HL	1,59	61,93%
RGI09	VIS=SR;CAU=CON;VEH=OT;EDAD=(27-60]	HL	1,80	13,24%
	ANCAL=MED	HL	**1,12**	
	ANCAL=MED;BAR=N	HL	0,95	14,92%
	ANCAL=MED;BAR=N;CAU=CON	HL	1,14	19,31%
RGI11	ANCAL=MED;BAR=N;CAU=CON;TAC=VUE	HL	1,58	38,93%
	EDAD=(20-27]	HL	**1,00**	
	EDAD=(20-27];CAU=CON	HL	0,95	-5,82%
	EDAD=(20-27];CAU=CON;SEX=M	HL	1,37	44,72%
RGI12	EDAD=(20-27];CAU=CON;SEX=M;DIA=L	HL	1,82	33,33%
	ANCAR=MED	HL	1,03	
	ANCAR=MED;CAT=O	HL	**1,34**	**30,29%**
	ANCAR=MED;CAT=O;HER=[1]	HL	1,34	0,00%
RGI13	ANCAR=MED;CAT=O;HER=[1];TAC=SV	HL	1,54	15,00%

Tabla 23.- Aplicación de la Condición del Incremento del Lift en las reglas HL obtenidas con RGI.

	HORA=(12-18]	HL	0,98	
	HORA=(12-18];CAT=LL	HL	1,39	42,15%
	HORA=(12-18];CAT=LL;VEH=VL	HL	1,44	3,88%
RGI14	HORA=(12-18];CAT=LL;VEH=VL;TAC=SV	HL	**1,50**	**3,85%**
	DIA=AF	HL	**1,06**	
	DIA=AF;CAT=BT	HL	1,00	-5,09%
	DIA=AF;CAT=BT;HORA=[6-12]	HL	1,74	73,60%
RGI15	DIA=AF;CAT=BT;HORA=[6-12];LUM=DIA	HL	1,84	5,67%
	MVIAL=CYM	HL	**1,01**	
	MVIAL=CYM;CAU=CON	HL	0,98	-3,23%
	MVIAL=CYM;CAU=CON;CAT=BT	HL	0,92	-6,10%
RGI18	MVIAL=CYM;CAU=CON;CAT=BT;VEH=OT	HL	1,61	74,96%
	BAR=N	HL	1,00	
	BAR=N;SEX=M	HL	**1,27**	**26,71%**
RGI20	BAR=N;SEX=M;VEH=VL	HL	1,31	2,71%
	CAU=CON	HL	**0,97**	
	CAU=CON;SEX=H	HL	0,92	-4,51%
	CAU=CON;SEX=H;CAT=BT	HL	0,89	-4,06%
RGI22	CAU=CON;SEX=H;CAT=BT;VEH=OT	HL	1,33	50,56%
	SEX=H	HL	0,95	
	SEX=H;CAT=LF	HL	1,11	16,76%
RGI23	SEX=H;CAT=LF;MVIAL=CYM	HL	**1,48**	**33,33%**
	SEX=H	HL	0,95	
	SEX=H;CAT=LL	HL	**1,32**	**38,57%**
RGI25	SEX=H;CAT=LL;TAC=SV	HL	1,31	-0,83%
	SEX=M	HL	1,26	
	SEX=M;VEH=VL	HL	1,30	3,11%
RGI26	SEX=M;VEH=VL;LUM=DIA	HL	**1,43**	**9,77%**
	SEX=H	HL	0,95	
	SEX=H;TAC=OT	HL	1,57	64,84%
RGI28	SEX=H;TAC=OT;ANCAR=MED	HL	**1,64**	**4,62%**
	SEX=M	HL	1,26	
	SEX=M;LUM=DIA	HL	**1,41**	**11,95%**
RGI29	SEX=M;LUM=DIA;OCU=[1]	HL	1,38	-2,11%
	CAT=LL	HL	**1,34**	
	CAT=LL;TAC=SV	HL	1,30	-2,63%
	CAT=LL;TAC=SV;DIA=F	HL	1,43	9,94%
RGI31	CAT=LL;TAC=SV;DIA=F;VEH=VL	HL	1,47	2,38%
	LUM=DIA	HL	1,04	
	LUM=DIA;SEX=M	HL	**1,41**	**35,19%**
	LUM=DIA;SEX=M;MVIAL=CYM	HL	1,35	-4,13%
RGI33	LUM=DIA;SEX=M;MVIAL=CYM;OCU=[1]	HL	1,36	0,39%
	CAT=LL	HL	**1,34**	
	CAT=LL;VEH=VL	HL	1,32	-1,26%
RGI36	CAT=LL;VEH=VL;SEX=H	HL	1,32	0,18%
	BAR=N	HL	**1,00**	
	BAR=N;CAT=BT	HL	0,96	-4,00%
RGI37	BAR=N;CAT=BT;ANARC=EST	HL	0,99	2,97%
	OCU=[1]	HL	1,00	
	OCU=[1];BAR=N	HL	1,06	6,02%
RGI39	OCU=[1];BAR=N;SEX=M	HL	**1,27**	**20,52%**
	HER=[>1]	HL	**0,88**	
	HER=[>1];TAC=SV	HL	0,90	2,04%
	HER=[>1];TAC=SV;CAU=CON	HL	0,90	0,47%
RGI40	HER=[>1];TAC=SV;CAU=CON;SEX=M	HL	1,29	43,17%
	EDAD=(27-60]	HL	**1,03**	
	EDAD=(27-60];BAR=N	HL	1,04	1,21%
	EDAD=(27-60];BAR=N;CAT=LL	HL	1,42	36,80%
RGI44	EDAD=(27-60];BAR=N;CAT=LL;TAC=SV	HL	1,41	-0,55%
	EDAD=(20-27]	HL	**1,00**	
	EDAD=(20-27];CAU=CON	HL	0,95	-5,82%
	EDAD=(20-27];CAU=CON;SEX=H	HL	0,89	-5,66%
RGI45	EDAD=(20-27];CAU=CON;SEX=H;CAT=LL	HL	1,37	53,40%
	CAT=LL	HL	**1,34**	
	CAT=LL;TAC=SV	HL	1,30	-2,63%
RGI47	CAT=LL;TAC=SV;MVIAL=CYM	HL	1,32	1,55%
	CAT=BT	HL	0,96	
	CAT=BT;LUM=DIA	HL	1,00	3,96%
RGI48	CAT=BT;LUM=DIA;SEX=M	HL	**1,43**	**43,32%**
	HORA=(18-24]	HL	**0,94**	
	HORA=(18-24];CAT=BT	HL	0,92	-1,79%
	HORA=(18-24];CAT=BT;LUM=INS	HL	1,06	14,48%
RGI50	HORA=(18-24];CAT=BT;LUM=INS;OCU=[1]	HL	1,31	24,24%
	HORA=(6-12]	HL	1,16	
	HORA=(6-12];VEH=VL	HL	1,26	8,32%
	HORA=(6-12];VEH=VL;CAU=COF	HL	1,52	20,85%
RGI52	HORA=(6-12];VEH=VL;CAU=COF;CAT=BT	HL	**1,78**	**17,20%**
	DIA=L	HL	**1,04**	
	DIA=L;CAU=CON	HL	1,01	-3,31%
	DIA=L;CAU=CON;SEX=M	HL	1,33	32,05%
RGI53	DIA=L;CAU=CON;SEX=M;VEH=VL	HL	1,37	2,90%

Continuación de Tabla 23

	SEX=H	HL	0,95	
	SEX=H;TAC=OT	HL	1,57	64,84%
RGI57	SEX=H;TAC=OT;ANCAR=MED	HL	**1,64**	**4,62%**
	SEX=M	HL	1,26	
	SEX=M;VEH=VL	HL	**1,30**	**3,11%**
RGI58	SEX=M;VEH=VL;CAU=CON	HL	1,25	-3,90%
	SEX=M	HL	1,26	
	SEX=M;VEH=VL	HL	1,30	3,11%
RGI59	SEX=M;VEH=VL;CAU=COF	HL	**1,47**	**12,71%**
	MON=OTO	HL	**1,01**	
	MON=OTO;TAC=SV	HL	1,02	1,65%
	MON=OTO;TAC=SV;CAU=CON	HL	1,02	-0,61%
RGI61	MON=OTO;TAC=SV;CAU=CON;SEX=M	HL	1,43	40,29%
	MVIAL=CYM	HL	**1,01**	
	MVIAL=CYM;CAU=CON	HL	0,98	-3,23%
	MVIAL=CYM;CAU=CON;CAT=LL	HL	1,32	34,62%
RGI62	MVIAL=CYM;CAU=CON;CAT=LL;TAC=SV	HL	1,30	-1,56%
	CAT=BT	HL	0,96	
	CAT=BT;HER=[1]	HL	**1,02**	**5,59%**
RGI64	CAT=BT;HER=[1];LUM=DIA	HL	1,02	0,59%
	ANARC=EST	HL	1,01	
	ANARC=EST;TAC=SV	HL	1,04	3,30%
	ANARC=EST;TAC=SV;CAU=COF	HL	1,30	24,49%
RGI65	ANARC=EST;TAC=SV;CAU=COF;ANCAR=MED	HL	**1,34**	**3,64%**
	TAC=VUE	HL	1,20	
	TAC=VUE;ANCAR=MED	HL	1,31	8,71%
RGI66	TAC=VUE;ANCAR=MED;BAR=N	HL	**1,37**	**4,76%**
	ANARC=NE	HL	**0,98**	
	ANARC=NE;SEX=H	HL	0,92	-6,36%
	ANARC=NE;SEX=H;CAT=BT	HL	0,89	-3,35%
RGI69	ANARC=NE;SEX=H;CAT=BT;LUM=SUF	HL	1,62	82,74%
	SEX=H	HL	0,95	
	SEX=H;CAT=LL	HL	**1,32**	**38,57%**
RGI71	SEX=H;CAT=LL;TAC=SV	HL	1,31	-0,83%
	SEX=M	HL	**1,26**	
	SEX=M;TAC=SV	HL	1,28	1,33%
RGI72	SEX=M;TAC=SV;CAT=BT	HL	1,28	0,36%
	ANARC=MED	HL	**1,06**	
	ANARC=MED;ANCAL=ANC	HL	1,03	-3,57%
	ANARC=MED;ANCAL=ANC;CAT=LL	HL	1,82	77,78%
RGI73	ANARC=MED;ANCAL=ANC;CAT=LL;SEX=H	HL	2,05	12,50%
	APAV=SI	HL	**1,01**	
	APAV=SI;TAC=SV	HL	1,03	2,20%
	APAV=SI;TAC=SV;VEH=VL	HL	1,18	14,00%
RGI74	APAV=SI;TAC=SV;VEH=VL;CAU=COF	HL	1,42	20,28%
	APAV=SI	HL	1,01	
	APAV=SI;TAC=VUE	HL	1,26	24,18%
	APAV=SI;TAC=VUE;BAR=N	HL	1,34	6,52%
RGI75	APAV=SI;TAC=VUE;BAR=N;SEX=H	HL	**1,40**	**4,55%**
	APAV=NE	HL	**0,99**	
	APAV=NE;CAU=CON	HL	0,99	0,74%
	APAV=NE;CAU=CON;HORA=6-12]	HL	1,32	32,69%
RGI76	APAV=NE;CAU=CON;HORA=6-12];BAR=N	HL	1,38	4,48%

Continuación de Tabla 23

Tras la verificación de la CIL en las 46 reglas HL y la selección del patrón correspondiente (ver Tabla 23), se comprueba que 9 reglas se repiten (RGI04=RGI05=RGI58; RGI31=RGI36=RGI47; RGI28=RGI57; RGI12=RGI45; RGI18=RGI62; RGI25=RGI71; RGI29=RGI33). De este modo se tienen un total de 37 patrones con la siguiente distribución: 18 reglas con 1 variable, 7 reglas con 2 variables, 7 reglas con 3 variables y 5 reglas con 4 variables.

➢ Conjunto final de patrones.

Sobre los patrones finales obtenidos al aplicar la CIL se calculan los parámetros Po, S, C y lift, y se comprueban de nuevo que se cumplen los umbrales mínimos previamente fijados (1%, 0,6%, 60%, 1,2, respectivamente). Las reglas que no cumplen estas condiciones se eliminan del conjunto final.

Los resultados de esta comprobación se muestran en las Tablas 24 (accidentes HGM) y 25 (accidentes HL). En sombreado se indican las reglas que cumplen la verificación de los parámetros.

NUM	REGLAS: SI...	ENTONCES	Po%	S%	C%	Lift	CHEKEO
RGI01	SEX=H	HGM	84,68	45,40	53,61	1,05	NO
RGI02	SEX=H;TAC=CP	HGM	7,86	5,32	67,68	1,32	OK
RGI06	CAT=BT	HGM	85,71	45,56	53,15	1,04	NO
RGI10	ANCAL=ANC;CAU=CON;ANCAR=EST;ANARC=EST	HGM	1,11	1,03	92,86	1,81	OK
RGI16	DIA=F;TAC=CP;ANCAR=MED	HGM	1,43	1,43	100,00	1,95	OK
RGI17	MON=OTO;TAC=CP	HGM	1,90	1,75	91,67	1,79	OK
RGI19	ANARC=MED;ANCAL=ANC;CAT=BT;VIS=TOP	HGM	2,06	1,75	84,62	1,65	OK
RGI21	TAC=SV;CAU=CON;VEH=MOT	HGM	11,75	8,10	68,92	1,34	OK
RGI24	CAU=CON;SEX=H	HGM	69,84	38,41	55,00	1,07	NO
RGI32	VEH=VL;TAC=CP	HGM	6,83	4,92	72,09	1,41	OK
RGI34	BAR=N;MON=INV	HGM	24,52	13,41	54,69	1,07	NO
RGI35	LUM=SI;CAT=BT	HGM	25,87	15,71	60,74	1,18	NO
RGI38	BAR=N;CAT=BT	HGM	83,02	44,05	53,06	1,03	NO
RGI41	CAU=CON	HGM	81,51	43,10	52,87	1,03	NO
RGI42	VIS=TOP	HGM	23,49	13,10	55,74	1,09	NO
RGI43	ANCAL=MED	HGM	28,81	13,10	45,45	0,89	NO
RGI46	ANCAR=MED;CAT=BT	HGM	60,16	31,59	52,51	1,02	NO
RGI49	HORA=[0-6];CAU=CON;OCU=[2];TAC=SV	HGM	4,05	3,10	76,47	1,49	OK
RGI51	CAT=BT;LUM=SI	HGM	25,87	15,71	60,74	1,18	NO
RGI54	CAT=BT;HORA=[0-6]	HGM	15,87	8,89	56,00	1,09	NO
RGI55	DIA=F	HGM	29,29	16,59	56,64	1,10	NO
RGI60	MON=INV;CAT=BT;LUM=SI	HGM	7,94	5,48	69,00	1,35	OK
RGI63	MVIAL=NE	HGM	9,76	5,08	52,03	1,01	NO
RGI67	ANARC=EST;TAC=CP	HGM	4,13	3,17	76,92	1,50	OK
RGI68	ANARC=NE;SEX=H	HGM	41,90	23,17	55,30	1,08	NO

Tabla 24.- Comprobación de parámetros en reglas HGM obtenidas con RGI.

Según se observa en la Tabla 24, los umbrales mínimos de los parámetros de Po, S, C y lift son verificados por un total de 10 patrones; siendo su distribución según el número de variables que los forman la siguiente: 4 reglas formadas por 2 variables, 3 reglas formadas por 3 variables y 3 reglas formadas por 4 variables.

De las 37 reglas de accidentes HL (recogidas en la Tabla 25), 19 verifican los umbrales mínimos de los parámetros (Po, S, C y lift), siendo su distribución según el número de variables: 2 reglas están formadas por 1 variable, 5 reglas por 2 variables, 7 reglas por 3 variables y 5 por 4 variables.

Una vez que se tienen identificados los patrones se procede a su descripción desde el punto de vista de la seguridad vial. Para ello se describirán por separado aquellos patrones que hacen referencia a accidentes graves o mortales de aquellos que hacen referencia a accidentes leves.

NUM	REGLAS: SI	ENTONCES	Po%	S%	C%	Lift	CHEKEO
RGI04	SEX=M;VEH=VL	HL	13,65	8,65	63,37	1,30	OK
RGI07	CAT=BT;SEX=M;LUM=DIA;OCU=[>2]	HL	1,03	1,03	100,00	2,05	OK
RGI08	OCU=[1];ANCAR=MED	HL	47,06	23,97	50,93	1,05	NO
RGI09	VIS=SR	HL	73,65	36,75	49,89	1,02	NO
RGI11	ANCAL=MED	HL	28,81	15,71	54,55	1,12	NO
RGI12	EDAD=(20-27]	HL	25,63	12,54	48,92	1,00	NO
RGI13	ANCAR=MED;CAT=O	HL	1,83	1,19	65,22	1,34	OK
RGI14	HORA=(12-18];CAT=LL;VEH=VL;TAC=SV	HL	2,06	1,51	73,08	1,50	OK
RGI15	DIA=AF	HL	15,87	8,17	51,50	1,06	NO
RGI18	MVIAL=CYM	HL	75,63	37,38	49,42	1,01	NO
RGI20	BAR=N;SEX=M	HL	14,60	9,05	61,96	1,27	OK
RGI22	CAU=CON	HL	81,51	38,41	47,13	0,97	NO
RGI23	SEX=H;CAT=LF;MVIAL=CYM	HL	1,43	1,03	72,22	1,48	OK
RGI25	SEX=H;CAT=LL	HL	6,67	4,29	64,29	1,32	OK
RGI26	SEX=M;VEH=VL;LUM=DIA	HL	9,13	6,35	69,57	1,43	OK
RGI28	SEX=H;TAC=OT;ANCAR=MED	HL	1,19	0,95	80,00	1,64	OK
RGI29	SEX=M;LUM=DIA	HL	9,92	6,83	68,80	1,41	OK
RGI31	CAT=LL	HL	8,89	5,79	65,18	1,34	OK
RGI37	BAR=N	HL	97,06	47,46	48,90	1,00	NO
RGI39	OCU=[1];BAR=N;SEX=M	HL	10,87	6,75	62,04	1,27	OK
RGI40	HER=[>1]	HL	30,40	13,02	42,82	0,88	NO
RGI44	EDAD=(27-60]	HL	54,13	27,06	50,00	1,03	NO
RGI48	CAT=BT;LUM=DIA;SEX=M	HL	8,41	5,87	69,81	1,43	OK
RGI50	HORA=(18-24]	HL	27,54	12,62	45,82	0,94	NO
RGI52	HORA=(6-12];VEH=VL;CAU=COF;CAT=BT	HL	1,83	1,59	86,96	1,78	OK
RGI53	DIA=L	HL	48,02	24,37	50,74	1,04	NO
RGI59	SEX=M;VEH=VL;CAU=COF	HL	2,22	1,59	71,43	1,47	OK
RGI61	MON=OTO	HL	23,65	11,59	48,99	1,01	NO
RGI64	CAT=BT;HER=[1]	HL	59,84	29,60	49,47	1,02	NO
RGI65	ANARC=EST;TAC=SV;CAU=COF;ANCAR=MED	HL	4,37	2,86	65,45	1,34	OK
RGI66	TAC=VUE;ANCAR=MED;BAR=N	HL	5,00	3,33	66,67	1,37	OK
RGI69	ANARC=NE	HL	49,05	23,41	47,73	0,98	NO
RGI72	SEX=M	HL	15,24	9,37	61,46	1,26	OK
RGI73	ANARC=MED	HL	10,71	5,56	51,85	1,06	NO
RGI74	APAV=SI	HL	51,35	25,32	49,30	1,01	NO
RGI75	APAV=SI;TAC=VUE;BAR=N;SEX=H	HL	3,49	2,38	68,18	1,40	OK
RGI76	APAV=NE	HL	31,19	15,00	48,09	0,99	NO

Tabla 25.- Comprobación de parámetros en reglas HL obtenidas con RGI.

➢ Patrones de accidentes HGM.

En la Tabla 26 se describen los 10 patrones de accidentes HGM, ordenados según el valor de la confidence. Los valores de confidence de las reglas varían desde un 100% (regla RGI16) hasta un valor de 67,7% (regla RGI02). Respecto al support de las reglas, todas tienen valores superiores al 1%, siendo el máximo de 8,1% (en la regla RGI21). El valor de population varía de 11,8% (regla RGI21), hasta 1,1% (regla RGI10). Y respecto a los valores del lift, varían desde un mínimo de 1,32 (regla RGI02) hasta un 1,95 (regla RGI16).

Con los patrones obtenidos con este método, de nuevo se pone de manifiesto la importancia de las colisiones con peatones en zona no urbana, así como de su gravedad; todos los patrones obtenidos para las colisiones con peatones son accidentes HGM (reglas RGI67, RGI16, RGI17, RGI02 y RGI32). Los patrones

particulares se describen a continuación, salvo la regla RGI16 que ya ha sido identificada y descrita con el método GI (regla GI38 en la Tabla 18).

- Regla RGI32: es una regla más general que regla GI56 (obtenida con GI) en la que se utilizaba una variable más (EST) para describir este mismo patrón: colisiones con peatones en los que el vehículo involucrado es un vehículo ligero. En esta regla la probabilidad de HGM es de un 72% (frente al 100% de regla GI56).

- Regla RGI17: es también un patrón más general que regla GI56 (en ese caso la variable adicional era VEH), que identifica colisiones con peatones en otoño. La probabilidad de esta regla es de casi un 92%.

- Regla RGI02: también es un patrón más general de una regla identificada con GI (regla GI10) en la que se identificaban 2 variables adicionales (VEH y APAV). De este modo, la regla RGI02 muestra colisiones con peatones cuando el conductor involucrado es un hombre con una probabilidad de HGM de casi un 68%.

- Regla RGI67: identifica colisiones con peatones en carreteras con arcén estrecho (menor de 1,5 m). La probabilidad de HGM con esta regla es de un 77%.

NUM.	REGLAS (SI…)	ENTONCES	Po%	S%	C%	LIFT
RGI16	DIA=F;TAC=CP;ANCAR=MED	HGM	1,43	1,43	100,00	1,95
RGI10	ANCAL=ANC;CAU=CON;ANCAR=EST;ANARC=EST	HGM	1,11	1,03	92,86	1,81
RGI17	EST=OTO;TAC=CP	HGM	1,90	1,75	91,67	1,79
RGI19	ANARC=MED;ANCAL=ANC;CAT=BT;VIS=TOP	HGM	2,06	1,75	84,62	1,65
RGI67	ANARC=EST;TAC=CP	HGM	4,13	3,17	76,92	1,50
RGI49	HORA=[0-6];CAU=CON;OCU=[2];TAC=SV	HGM	4,05	3,10	76,47	1,49
RGI32	VEH=VL;TAC=CP	HGM	6,83	4,92	72,09	1,41
RGI60	EST=INV;CAT=BT;LUM=SI	HGM	7,94	5,48	69,00	1,35
RGI21	TAC=SV;CAU=CON;VEH=MOT	HGM	11,75	8,10	68,92	1,34
RGI02	SEX=H;TAC=CP	HGM	7,86	5,32	67,68	1,32

Tabla 26.- Reglas para accidentes HGM con RGI.

Con los patrones obtenidos con RGI para las colisiones con peatones, se reafirman los resultados que ya se señalaban con las reglas GI. De nuevo se relaciona la estación de otoño con la gravedad de estas colisiones (RGI17); que el tipo de vehículo involucrado en estos accidentes suele ser un vehículo ligero (RGI32) y el conductor un hombre (RGI02); o se muestra alguna variable relacionada con los márgenes de la carretera (regla RGI67). Por lo que de nuevo se confirma que los usuarios vulnerables en carreteras convencionales tienen mayor riesgo de sufrir un accidente grave. Este resultado también fue identificado en el estudio realizado por Giacomo et al. (2013) sobre accidentes en esta tipología de carreteras.

Para los accidentes producidos por salida de la vía se identifican 2 patrones (RGI49 y RGI21) con gravedad HGM:

- Regla RGI49: muestra estos accidentes por causas debidas al conductor, en la franja horaria de 0 a 6 horas, con 2 ocupantes involucrados. La probabilidad de esta regla es de un 76% y el support de un 3%.

- Regla RGI21: identifica salidas de la vía de motocicletas por causas debidas al conductor. Es la regla de mayor support (8,1%), siendo la probabilidad también elevada (69%). Comparada con las reglas GI, se observa que este patrón es más general que el obtenido en la regla GI23 (que identificaba además la variable ANCAL).

Respecto a los 3 patrones restantes, 2 de ellos (reglas RGI19 y RGI10) involucran variables relacionadas con la carretera:

- Regla RGI10: identifica los accidentes en carreteras con calzada mayor de 7 m, con arcén estrecho (menor de 1,5 m), y carriles también estrechos (menores de 3,25 m), por causas debidas al conductor. El valor de probabilidad de esta regla es de casi un 93% y el support de un 1%. Con este patrón se relaciona condiciones poco favorables de la vía, tales como ancho de carril y arcén estrecho, con patrones de accidentes graves.

- Regla RGI19: identifica los accidentes en carreteras con ancho de calzada mayor de 7 m, con ancho de arcén entre 1,5 y 2,5 m, cuando las condiciones atmosféricas son de buen tiempo y la visibilidad de la carretera está restringida por la topografía. Los valores de probabilidad y support son de 85% y 1,8% respectivamente. En este patrón las condiciones de la vía son adecuadas pero las del entorno no (visibilidad restringida por la topografía).

Finalmente, la regla RGI60 identifica accidentes graves en invierno, en condiciones de buen tiempo, durante noches sin iluminación. Los valores de probabilidad y support de esta regla son de 69% y 5,5% respectivamente. Este patrón es más específico que el descrito en la regla GI20 que identificaba accidentes graves en invierno durante noches sin iluminación. Por lo que de nuevo se pone de manifiesto la relación entre la falta de iluminación y la gravedad del accidente, en las carreteras analizadas.

➢ Patrones de accidentes HL.

En la Tabla 27 se muestran los 19 patrones de accidentes HL ordenados según el valor de la confidence. Los valores de confidence de las reglas HL varían desde un 100% (en la regla RGI07) hasta un valor de 61,5% (regla RGI72). Respecto al support de estas reglas, se observa que varía de 9,4% (regla RGI72) a 0,95% (regla RGI28). El valor de population varía de 15,2% en la regla RGI72, hasta un 1,03%, en la regla RGI07. Y los valores del lift, varían desde 1,26 (regla RGI72) hasta 2,05 (regla RGI07).

De los 19 patrones obtenidos, 3 de ellos (reglas RGI07, RGI48 y RGI29) fueron también identificados con GI (reglas GI18, GI19 y GI13) en la Tabla 19. Y al igual que sucedía con las reglas para accidentes HGM, muchos de los patrones obtenidos con RGI son patrones más generales (o específicos) que los obtenidos previamente con GI. Por ejemplo, la regla RGI26 es un patrón más específico de las reglas RGI29 o GI13.

NUM.	REGLAS (SI…)	ENTONCES	Po	S%	C%	LIFT
RGI07	CAT=BT;SEX=M;LUM=DIA;OCU=[>2]	HL	1,03	1,03	100,00	2,05
RGI52	HORA=(6-12];VEH=VL;CAU=COF;CAT=BT	HL	1,83	1,59	86,96	1,78
RGI28	SEX=H;TAC=OT;ANCAR=MED	HL	1,19	0,95	80,00	1,64
RGI14	HORA=(12-18];CAT=LL;VEH=VL;TAC=SV	HL	2,06	1,51	73,08	1,50
RGI23	SEX=H;CAT=LF;MVIAL=CYM	HL	1,43	1,03	72,22	1,48
RGI59	SEX=M;VEH=VL;CAU=COF	HL	2,22	1,59	71,43	1,47
RGI48	CAT=BT;LUM=DIA;SEX=M	HL	8,41	5,87	69,81	1,43
RGI26	SEX=M;VEH=VL;LUM=DIA	HL	9,13	6,35	69,57	1,43
RGI29	SEX=M;LUM=DIA	HL	9,92	6,83	68,80	1,41
RGI75	APAV=SI;TAC=VUE;BAR=N;SEX=H	HL	3,49	2,38	68,18	1,40
RGI66	TAC=VUE;ANCAR=MED;BAR=N	HL	5,00	3,33	66,67	1,37
RGI65	ANARC=EST;TAC=SV;CAU=COF;ANCAR=MED	HL	4,37	2,86	65,45	1,34
RGI13	ANCAR=MED;CAT=O	HL	1,83	1,19	65,22	1,34
RGI47	CAT=LL	HL	8,89	5,79	65,18	1,34
RGI71	SEX=H;CAT=LL	HL	6,67	4,29	64,29	1,32
RGI58	SEX=M;VEH=VL	HL	13,65	8,65	63,37	1,30
RGI39	OCU=[1];BAR=N;SEX=M	HL	10,87	6,75	62,04	1,27
RGI20	BAR=N;SEX=M	HL	14,60	9,05	61,96	1,27
RGI72	SEX=M	HL	15,24	9,37	61,46	1,26

Tabla 27.- Reglas para accidentes HL con RGI.

Según el tipo de accidente se identifican 5 patrones, 2 para accidentes por salida de la vía (reglas RGI14 y RGI65), 2 para accidentes por vuelco (reglas RGI75 y RGI66) y 1 para accidentes clasificados como otra tipología (regla RGI28):

- Regla RGI14: describe accidentes por salida de la vía cuando el vehículo involucrado es un vehículo ligero y las condiciones atmosféricas son de lluvia ligera, en la franja horaria de 12-18 h. Tiene una probabilidad del 73% y un support de 1,5%.

- Regla RGI65: identifica accidentes por salida de la vía producidos en carreteras con ancho de carril comprendido entre 3,25 y 3,75 m, y ancho de arcén menor de 1,5 m, por causas debidas a una combinación de factores. La probabilidad de esta regla es de un 65% y el support de 2,9%.

- Regla RGI75: describe un patrón para accidentes producidos por vuelco, cuando el conductor es un hombre, en carreteras sin barreras de seguridad y con arcén pavimentado. La probabilidad de esta regla es de un 68% y el support de 2,4%.

- Regla RGI66: muestra accidentes por vuelco, en carreteras sin barreras de seguridad, con ancho de carril comprendido entre 3,25 y 3,75 m, con una probabilidad de 66,7% y un soporte de 3,3%.

- Regla RGI28: identifica accidentes clasificados como otra tipología (que incluyen colisiones con animales en rebaño, sueltos, etc.), con conductores hombres en carreteras con ancho de carril medio (entre 3,25 y 3,75 m). La probabilidad de esta regla es de un 80% y el support de 0,9%.

Los patrones obtenidos para accidentes por vuelco son patrones más específicos que los obtenidos con el método GI (regla GI69) en el que se identificaban accidentes leves producidos por vuelco en carreteras sin barreras. Esto pone de manifiesto la influencia de las barreras en los accidentes por vuelco.

Al igual que con GI, se han obtenido patrones para accidentes leves por salida de la vía cuando el vehículo involucrado es un vehículo ligero (reglas RGI14); en contraste con los patrones obtenidos para accidentes graves por salida de la vía, en los cuales el vehículo implicado era siempre una motocicleta. Con RGI también se obtiene un patrón grave de accidentes por salida de la vía y motocicletas (regla RGI21).

Cabe destacar que en la mayoría de estos los patrones (RGI14, RGI66 y RGI65) existen condiciones específicas de la vía que son favorables (ancho de carril comprendido entre 3,25 y 3,75 m), lo que puede estar relacionado con una mayor velocidad de circulación. En este sentido la medida de reducción de velocidad propuesta por la DGT puede ser una contramedida efectiva en este tipo de carreteras.

Atendiendo a las condiciones atmosféricas se distinguen 5 nuevos patrones:

- Regla RGI47: identifica accidentes HL con condiciones atmosféricas de lluvia ligera. La regla RGI71 muestra un patrón más específico de estos accidentes en el que el conductor es un hombre (siendo las probabilidades de ambas reglas muy similares 65% vs. 64%).

- Regla RGI13: que muestran accidentes HL en carreteras con ancho de carril comprendido entre 3,25 y 3,75 m, con condiciones atmosféricas clasificadas como otras (nieve, granizo, etc). La probabilidad de esta regla es de un 65% y el support de 1,2%.

- Regla RGI23: identifica también accidentes HL cuando las condiciones atmosféricas son de lluvia fuerte, el conductor es un hombre, y existen marcas viales que separan carriles y márgenes de la carretera. Esta regla tiene una probabilidad del 72% y un support de un 1%.

- Regla RGI52: identifica un patrón de accidentes HL cuando las condiciones atmosféricas son de buen tiempo, que se producen de 6 a 12 h, el vehículo involucrado es un vehículo ligero, y las causas son una combinación de factores. La probabilidad de esta es regla de un 87% y el support de 1,6%.

En contraste con los accidentes graves que suelen producirse con condiciones atmosféricas de buen tiempo (reglas RGI19 y RGI60), los resultados apuntan a que cuando las condiciones atmosféricas son adversas la gravedad del accidente es menor (RGI47, RGI14, RGI23, RGI13 y RGI71). Este resultado también fue indicado por Xie et al. (2013), quienes subrayaron que bajo condiciones meteorológicas adversas los conductores extreman su precaución y disminuyen su velocidad.

El resto de patrones que se obtienen con este método hacen referencia a conductores mujeres. La regla RGI72 identifica accidentes HL cuando los conductores son mujeres con una probabilidad del 61%. A partir de este patrón surgen otros más específicos:

- Regla RGI20: cuando la carretera no tiene barreras de seguridad, los accidentes son HL con probabilidad del 62% y un support de 9%. Y si además sólo hay un ocupante involucrado se obtiene la regla RGI39 (con aproximadamente el mismo valor de probabilidad, 62%).
- Regla RGI58: muestra también accidentes HL para mujeres, cuando el vehículo involucrado en el accidente es un vehículo ligero. Y en la regla RGI59 se muestran estos accidentes cuando las causas son debidas a una combinación de factores (con una probabilidad superior 71% vs. 63%).
- Regla RGI26: muestra accidentes HL para mujeres cuando la iluminación es igual a día y el vehículo implicado es un vehículo ligero. La probabilidad de esta regla es de un 69,5% y el support de 6,4%.

5.4.3. Comparación de métodos.

Respecto de los 2 métodos utilizados para la obtención de patrones (GI y RGI), se pueden destacar las siguientes conclusiones:

Del total de patrones obtenidos con ambos métodos, se han identificado 4 patrones iguales: 1 para accidentes HGM (reglas GI38 y RGI16), y 3 para accidentes HL (reglas GI18, GI19 y GI13 con las reglas RGI07, RGI48 y RGI29, respectivamente), aunque la mayoría de los patrones obtenidos presentan características similares. Por ejemplo, la relación entre los accidentes graves producidos por colisiones con peatones es obtenida con ambos criterios (reglas GI38, GI56, GI10 y GI55 y reglas RGI67, RGI16, RGI17, RGI02 y RGI32). En este caso la mayoría de los patrones obtenidos con RGI son patrones más generales que los obtenidos con GI. Además con ambos criterios se identifica una relación entre estos accidentes y los vehículos ligeros (GI56, GI10 y RGI32). Particularmente, la regla RGI32 es un patrón más general que la GI56 (en la que se utiliza una variable más, EST) para describir colisiones con peatones en los que el vehículo involucrado es un vehículo ligero. La regla RGI17 también es un patrón más general que la regla GI56 (en este caso la variable adicional es VEH), que identifica colisiones con peatones HGM en otoño. Y la regla RGI02 es un patrón más general de la regla GI10 en la que se identifican 2 variables adicionales (VEH y APAV), para describir colisiones con peatones cuando el conductor involucrado es un hombre.

Patrones particulares para los accidentes graves producidos por salida de la vía son también obtenidos con ambos métodos (reglas GI01, GI62, GI02, GI23, GI03 y GI35 y reglas RGI49 y RGI21). Todas las reglas obtenidas con GI muestran accidentes HGM por salida de la vía en las que el vehículo implicado es una motocicleta. Con el método RGI, se identifica también este patrón en la regla RGI21 (que es una más general de la regla GI23).

Finalmente, la regla RGI60 identifica accidentes graves en invierno, en condiciones de buen tiempo, durante noches sin iluminación, siendo un patrón más específico que la regla GI20.

Por tanto, de un modo global, las problemáticas de seguridad vial detectadas por ambos métodos son similares. Sin embargo, cada método identifica, sobre cada problemática general, determinados aspectos concretos, a través de las variables particulares que son identificadas en las reglas que se obtienen con uno u otro método.

Además, con cada método se ha obtenido algún patrón particular, como por ejemplo los obtenidos para accidentes HGM y conductores jóvenes, con las reglas GI33 y GI25. Estos resultados son coherentes, porque los 2 métodos utilizan criterios diferentes para dividir la base de datos y, por tanto, pueden identificar variables específicas en determinados patrones, y que no son detectadas con el otro método.

Para los accidentes HL también se observan estos hechos. Por ejemplo, los patrones obtenidos para accidentes por vuelco con el método RGI (reglas RGI66 y RGI75) son patrones más específicos que los obtenidos con el método GI (regla GI69), en los que se pone de manifiesto la influencia de las barreras en estos accidentes. Accidentes leves por salida de la vía con vehículos ligeros son también identificados con ambos métodos (reglas GI47 y RGI14).

Con ambos métodos se identifican patrones de accidentes leves con mujeres (reglas GI18, GI19, GI13 y GI66, y reglas RGI07, RGI59, RGI48, RGI26, RGI29, RGI58, RGI39, RGI20 y RGI72). Incluso con el método RGI se muestra un patrón directo en el que no influye ninguna otra variable (regla RGI72); y a partir de este patrón se muestran otros más específicos con ambos métodos. Por ejemplo, las reglas RGI26 y GI19 son patrones más específicos de la regla RGI29 (o GI13), en las que se muestran patrones de accidentes leves para mujeres con condiciones de iluminación igual a día, y particularmente, cuando el vehículo involucrado es un vehículo ligero (RGI26) o cuando las condiciones atmosféricas son de buen tiempo (GI19, RGI07 y GI18).

Con cada método también se han obtenido algunos patrones particulares. Por ejemplo, con RGI se han identificados accidentes leves para conductores hombres (RGI28, RGI23, RGI75 y RGI71). Y en algunos de estos patrones se relacionan los accidentes leves, en los que conductor es un hombre, con condiciones atmosféricas adversas (RGI71 y RGI23). La relación entre accidentes leves y condiciones atmosféricas adversas también se muestra con GI (regla GI22).

Por tanto, de nuevo se pone de manifiesto que de un modo global las problemáticas de seguridad vial detectadas con ambos métodos son similares. Y que cada método, sobre cada problemática general, identifica determinados aspectos concretos.

5.4.4. Conclusiones.

Como se ha visto en los resultados obtenidos, cuando se utiliza un solo ADD para extraer el conocimiento de una base de datos (ADD_1 en las Tablas 12 y 20), el número

de patrones que se obtienen es mucho menor que cuando se aplica el método propuesto en esta tesis (IRNV). Por tanto, para el estudio completo de una base de datos, se recomienda utilizar el método IRNV.

La principal limitación de este método, es que la mayoría de los patrones provienen de ADDs en los que se impone el nodo raíz que genera su construcción. Por tanto, se tienen reglas en las que la variable impuesta como nodo raíz no tiene porque ser realmente importante en el patrón que describe la regla. Por ello es necesario una previa verificación de la importancia de esta variable con el objeto de saber si el patrón original (RD) debe ser simplificado (RD^{-}). Para realizar esta verificación se ha impuesto dos condiciones (ecuaciones 26 y 27).

Los resultados ponen de manifiesto que este paso es muy importante, ya que en el caso de las reglas obtenidas con GI se tienen 45 RD y 20 reglas RD^{-}; y en el caso de las reglas obtenidas con RGI se tienen 44 RD y 27 reglas RD^{-}. El número de reglas ampliadas (RD) y número de reglas reducidas (RD^{-}) varía según los umbrales establecidos en las condiciones 1 y 2 (ecuaciones 26 y 27).

Dado que la implantación de cualquier medida correctiva conlleva un coste, y que los recursos disponibles son limitados, se deben extraer los patrones más fuertes que describen las problemáticas de seguridad vial de las carreteras analizadas. Las variables que forman cada patrón han sido también verificadas utilizando el criterio del incremento del lift (CIL). De este modo, sólo los patrones que sufren una mejora en el incremento del lift se mantienen de cara al posterior análisis de seguridad vial. Cabe destacar, que este paso también es muy importante, porque muchos patrones quedan reducidos a una, dos o tres variables (teniendo en cuenta que las reglas originalmente extraídas de los ADDs tenía 4 variables como máximo).

Atendiendo a los resultados particulares que se obtienen con cada método, se puede observar que ambos métodos identifican algunos patrones que son exactamente iguales, y que la mayoría de ellos resultan similares. Es decir, de un modo global, las problemáticas de seguridad vial detectadas son las mismas con ambos métodos. Por ejemplo: accidentes graves y colisiones con peatones, accidentes graves por salida de la vía y motocicletas, y accidentes leves con conductores mujeres, etc. Sin embargo, cada método identifica, a partir de cada problemática general, determinados aspectos específicos, a través de variables particulares que son identificadas en las reglas que se obtienen con uno u otro método.

Estos resultados son coherentes, ya que la base de accidentes analizada es la misma para ambos métodos, y las problemáticas generales de seguridad vial deben de ser iguales. Sin embargo, los métodos utilizados para la construcción de los ADDs y la posterior extracción de reglas son diferentes. Están basados en criterios de partición diferentes (la medida de pureza con la que se dividen los datos en el caso de GI se basa en el índice de diversidad, mientras que en RGI se basa en la entropía). Por ello, cada método puede realizar diferentes particiones de la base de datos y, por tanto, pueden identificar variables específicas para un determinado patrón general, e incluso identificar los patrones adicionales que se obtienen con cada método.

Por lo tanto, para una extracción completa del conocimiento existente en la base de datos analizada, y para una gestión global de las problemáticas de las carreteras analizadas, se recomienda el uso del método IRNV con ambos criterios (GI e RGI).

CAPÍTULO 6.
CONCLUSIONES

CAPÍTULO 6. CONCLUSIONES

6.1. Conclusiones.

En este capítulo se presentan las principales conclusiones obtenidas con el trabajo de investigación desarrollado en esta tesis doctoral.

La compresión e identificación de los principales factores que afectan a la gravedad de los accidentes de tráfico ha sido el objetivo de muchos analistas de la seguridad vial, siendo los Discrete Outcome Models los más comúnmente empleados para llevar a cabo este análisis.

Dadas las particularidades de los datos de esta tipología de accidentes, tales como la infra-detección, la presencia de variables endógenas, correlaciones, etc; los investigadores han ido introduciendo diversas modificaciones en los modelos originales. En general, la elección del modelo ha estado basada en los problemas detectados a la hora de analizar los accidentes, así como en el objetivo perseguido en cada estudio particular.

Sin embargo, el principal inconveniente de estos modelos es que parten de hipótesis fijas y predefinen relaciones entre las variables dependientes e independientes, de modo que si estas hipótesis no se cumplen los modelos pueden producir estimaciones erróneas en la probabilidad de la gravedad de la lesión (Chang and Wang, 2006). Esto ha provocado que numerosos investigadores hayan comenzado a utilizar técnicas de Minería de Datos (MD) para estudiar la severidad de los accidentes de tráfico.

Las técnicas de MD permiten extraer conocimiento de los datos previamente desconocido e indistinguible, y normalmente no parten de hipótesis, ni requieren un previo conocimiento probabilístico del problema objeto de estudio.

Además, el uso de técnicas de MD en el estudio de los accidentes resulta muy apropiado por la propia naturaleza de estos fenómenos. Un accidente es un evento raro, aleatorio y en el que influyen múltiples factores, siempre precedido por una situación en la que uno o más conductores no pueden hacer frente al entorno de la carretera (ROSPA, 2002). De modo que cada accidente es el resultado de una cadena de eventos, que es en su totalidad único pero algunos factores son comunes a varias circunstancias del accidente. La identificación de estos factores y sus interdependencias puede llevarse a cabo mediante el uso de técnicas de MD (Montella et al., 2012b).

Dentro de los modelos de MD existen diferentes tipos de técnicas. No existe una técnica universal que pueda ser aplicada para la resolución de cualquier tipo de problema (Hernández et al., 2004). Sino que cada técnica presenta sus ventajas e inconvenientes, y la elección de la misma dependerá del objetivo perseguido por el analista. Las Redes Neuronales Artificiales, las Redes Bayesianas, las Reglas de Asociación y los Árboles de Decisión son las más utilizadas en el campo de la seguridad vial.

Para realizar el análisis de los accidentes de tráfico recogidos en esta investigación se han utilizado ADDs, ya que, teniendo en cuenta sus ventajas y limitaciones, constituyen uno de los modelos más utilizados en aprendizaje supervisado y en aplicaciones de Minería de Datos (Gehrke et al., 1999). Entre sus ventajas, de cara a su utilización para el estudio de los accidentes, se pueden destacar que: son una técnica que normalmente no utiliza parámetros, y no suele suponer ninguna relación previa entre las variables; son fácilmente interpretables; pueden trabajar con grandes bases de datos, y descubrir fácilmente complejas interacciones entre los mismos. Además, permiten la extracción de reglas de decisión del tipo "SI-ENTONCES", que pueden ser usadas para descubrir determinados patrones de comportamiento que ocurren dentro de un conjunto de datos.

Los datos de accidentes utilizados para llevar a cabo esta investigación han sido suministrados por la Dirección General de Tráfico. Sobre la muestra total se ha realizado un pre-tratamiento previo con el objeto de analizar los accidentes ocurridos en carreteras convencionales. Se han eliminado los accidentes ocurridos en intersecciones, dado que los factores relativos a estos accidentes son diferentes se recomienda estudiarlas por separado (Moore et al., 2010).

La gravedad del accidente depende de factores de diversa naturaleza, tales como el vehículo, la carretera o el conductor. Además, los factores que afectan a los accidentes con un solo vehículo involucrado pueden ser diferentes de los que afectan a las colisiones múltiples (con más de un vehículo involucrado). En los accidentes múltiples la gravedad está altamente relacionada con factores tales como el tipo de colisión, el tamaño y peso de los vehículos involucrados en el accidente, los puntos de contacto, etc. (Krull et al., 2000). Por ello, en esta investigación solo se han analizado accidentes con un vehículo involucrado.

Finalmente, dada la disponibilidad de los datos, se han analizado los accidentes ocurridos durante 7 años, desde 2003 hasta 2009. El entorno geográfico en el que se sitúan estos accidentes son las carreteras convencionales de la provincia de Granada. Después del proceso de filtrado de datos se cuenta con una muestra de 1.801 accidentes.

Dada esta muestra de estudio, en una primera fase de la investigación el objetivo fue validar el uso de ADDs, construidos con diferentes algoritmos, para el estudio de la gravedad de los accidentes de tráfico; así como la identificación de las variables que tiene un efecto clave en la gravedad de los mismos. De esta primera fase de la investigación se pueden extraer las siguientes conclusiones:

- Los ADDs son una herramienta que permite analizar los accidentes de tráfico de un modo sencillo y fácilmente comprensible para los analistas de la seguridad vial. Además permiten realizar una clasificación de los accidentes en base a su gravedad. De este modo, los ADDs se presentan como un método alternativo a los modelos paramétricos, debido a su capacidad para identificar los patrones en los datos sin la necesidad de establecer una relación funcional entre las variables. Por otra parte, estos modelos de clasificación se pueden utilizar para determinar las interacciones entre los variables que serían imposibles de establecer directamente utilizando las técnicas de modelización tradicionales.

- Los ADDs pueden ser construidos con diferentes algoritmos. El método CART ha sido el más utilizado en las investigaciones de seguridad vial (Chang and Wang, 2006; Kashani and Mohaymany, 2011; Kashani et al., 2011; Montella et al., 2011; Montella et al., 2012b). Sin embargo, CART sólo permite la construcción de árboles binarios. Existen otros algoritmos que permiten la construcción de ADDs, tales como ID3 y C4.5, que no presentan esta restricción binaria. En el caso de los accidentes de tráfico, estos métodos resultan muy apropiados cuando se pretende analizar la influencia de una categoría específica de una variable en la gravedad del accidente.

- Los ADDs construidos con algoritmos CART, ID3 y C4.5 han sido evaluados con los indicadores *accuracy, sensitivity, specificity y ROC*. Los resultados muestran que el algoritmo ID3 es el método que peores resultados arroja. Sin embargo entre CART y C4.5 las diferencias no son muy significativas y, aunque CART obtiene valores algo superiores en los parámetros de *accuracy, specificity y ROC*, no podemos decir, a priori que un modelo sea mejor que otro. Así que resulta adecuado analizar los modelos obtenidos con los algoritmos C4.5 y CART.

- CART construye árboles binarios y por tanto determinadas categorías de las variables divisoras son agrupadas en una rama, aumentando el support de un nodo, pero imposibilitando analizar la influencia de una categoría concreta en la severidad del accidente. Por el contrario, C4.5 crea una rama para cada categoría y permite analizar la influencia de todas las categorías de las variables que participan en la construcción del árbol. Por tanto, se puede decir que las reglas obtenidas con CART son menos informativas.

- C4.5 genera árboles con mayor número de ramas que CART y, por tanto, produce mayor número de reglas. Sin embargo, no todas cumplen con los parámetros mínimos establecidos de support, population y confidence. Por ello, estas reglas pueden resultar de menos utilidad de cara a implantar futuras estrategias de seguridad vial.

- Ambos algoritmos permiten obtener la importancia de las variables en el modelo. Con CART se han identificado 12 variables con importancias superiores al 9,9%. La variable luminosidad, las condiciones atmosféricas, la hora, el tipo de accidente y el sexo del conductor son las variables de mayor importancia. Estos resultados coinciden con los obtenidos en estudios previos (Evans, 2001; Al-Ghamdi, 2002;

Kcoleman and Kweon, 2002; Abdel-Aty, 2003; Ulfarsson and Mannering, 2004; Gray et al., 2008; Helai et al,. 2008; De Oña et al., 2011; Pande and Abdel-Aty, 2009; Xie et al., 2009; Kashani and Mohyamany, 2011; Mujalli and De Oña, 2011; Obeng, 2011). Con C4.5 se identifican 14 variables, siendo 11 de estas variables también identificadas en CART. Por orden de importancia, la primera es el tipo de accidente, seguida de las causas, el sexo, la iluminación y el tipo de vehículo.

- Desde una perspectiva de la gestión de la seguridad vial, se pueden destacar algunas conclusiones:

  - Los accidentes HGM son producidos fundamentalmente por conductores hombres.
  - La probabilidad de accidente HGM aumenta en el caso de que haya involucrados peatones. Desde un punto de vista de la seguridad vial, se podría actuar sobre estos accidentes colocando barreras de seguridad en los tramos de carretera convencional dónde pueda existir circulación peatonal (por ejemplo en carreteras que conectan 2 núcleos de población cercanos).
  - Cuando una conductora está involucrada en el accidente, ambos métodos predicen accidentes HL si existe iluminación en la vía (pleno día, suficiente iluminación o atardecer). Sin embargo, ambos predicen HGM cuando no existe iluminación o esta es insuficiente. Estas reglas no se observan para conductores varones y podrían indicar que las mujeres aumentan su riesgo de severidad en el accidente en condiciones de menor iluminación de la vía.

- Por último, hay que destacar que cada método presenta ventajas e inconvenientes y revela información diferente, por lo que ambos métodos pueden resultar complementarios y, para un análisis completo, se recomienda usar ambos de forma conjunta.

Dado que las RDs que se pueden obtener de un sólo ADD vienen muy condicionadas por la variable que se usa como raíz, se puede dar lugar a pérdida de otra/s informaciones importantes sobre la variable en estudio. Por ello, en una segunda fase de la investigación se ha propuesto una metodología específica (*Information Root Node Variation*) que resuelve esta limitación; y que permite la extracción completa del conocimiento existente en la base de datos objeto de estudio. Las conclusiones específicas obtenidas son las siguientes:

- Cuando se utiliza un solo ADD para extraer el conocimiento de una base de datos el número de patrones que se obtienen es mucho menor que cuando se aplica el método propuesto (IRNV). Por tanto, para el estudio completo de una base de datos, se recomienda utilizar el método IRNV.

- La principal limitación del método IRNV, es que la mayoría de los patrones provienen de ADDs en los que se impone el nodo raíz que genera su construcción. Por tanto, se tienen reglas en las que la variable impuesta como nodo raíz puede no ser realmente importante en el patrón que describe la misma. Por ello, es necesario una verificación previa de la importancia de esta variable, con el objeto

de saber si el patrón original (RD) puede ser simplificado (RD^{-}). Los resultados ponen de manifiesto que este paso es muy importante.

- Dado que la implantación de cualquier medida de seguridad vial conlleva un coste, y que los recursos disponibles son limitados, se deben extraer los patrones más fuertes que describen las problemáticas de seguridad vial de las carreteras analizadas. Por ello, las variables que forman cada patrón han sido verificadas utilizando el criterio del incremento del lift. De este modo, sólo los patrones que sufren una mejora en el incremento del lift se mantienen de cara al posterior análisis de seguridad vial. Cabe destacar que este paso también es muy importante porque muchos patrones quedan reducidos a una, dos o tres variables en el antecedente.

- Atendiendo a los resultados particulares que se obtienen con cada método, se puede decir que ambos métodos identifican algunos patrones que son exactamente iguales, y la mayoría de ellos resultan muy similares. Así, de un modo global las problemáticas de seguridad vial detectadas son similares con ambos métodos. Por ejemplo: accidentes graves y colisiones con peatones, accidentes graves por salida de la vía y motocicletas, accidentes leves con conductores mujeres, etc. Sin embargo, cada método identifica sobre cada problemática general, determinados aspectos particulares. Esto es lógico, ya que la base de accidentes analizada es la misma para ambos métodos, y las problemáticas generales de seguridad vial deben de ser iguales. Sin embargo, los métodos utilizados para la construcción de los ADDs y la posterior extracción de reglas son diferentes, están basados en criterios de partición diferentes (la medida de pureza de GI se basa en el índice de diversidad, mientras que en RGI se basa en la entropía). Por ello, cada método puede realizar diferentes particiones la base de datos y pueden identificar variables específicas para un determinado patrón general, e incluso identificar otros patrones adicionales.

- Respecto a los resultados particulares obtenidos, desde el punto de vista de la seguridad vial, se pueden destacar las siguientes conclusiones particulares:

  - La relación entre los accidentes graves producidos por colisiones con peatones es obtenida con ambos criterios (GI y RGI). Esto confirma que los usuarios vulnerables en carreteras convencionales tienen mayor riesgo de sufrir un accidente grave. Con lo patrones obtenidos cabe destacar que el tipo de vehículo involucrado en estos accidentes suele ser un vehículo ligero; que los días festivos aumentan la probabilidad de estos accidentes; y que en estos accidentes aparecen variables relacionadas con los márgenes de la carretera, por lo que medidas específicas en las secciones de carretera con tránsito peatonal pueden ayudar a mejorar la seguridad vial de las mismas.
  - Patrones particulares para los accidentes graves producidos por salida de la vía son obtenidos también con ambos métodos. Todas las reglas obtenidas con GI muestran accidentes HGM por salida de la vía en los que el vehículo implicado es una motocicleta. Con el método RGI se identifica también este patrón en una

regla. También se muestra la influencia del arcén en estos accidentes. En este sentido, las Administraciones competentes pueden realizar esfuerzos orientados a disminuir este tipo de accidentes.
- Con ambos métodos se identifican bastantes patrones de accidentes leves con mujeres. Incluso con el método RGI se muestra un patrón directo en el que no influye ninguna otra variable; y a partir de este patrón se muestran otros más específicos con ambos métodos.
- Y con los dos métodos se identifica también patrones de accidentes leves cuando las condiciones meteorológicas son adversas.

- Por lo tanto, para un una extracción completa del conocimiento existente en la base de datos analizada, y para una gestión global de las problemáticas de las carreteras, se recomienda el uso conjunto de ambos criterios (GI y RGI).

- Finalmente, se puede destacar que en esta tesis doctoral se han analizado las problemáticas concretas de las carretas convencionales de la provincia de Granada. Sin embargo, la metodología presentada resulta de interés para el estudio de las problemáticas existentes en otras vías, o en un ámbito más general.

Desde un punto de vista de la seguridad vial, la mayor parte de las reglas obtenidas coinciden con los problemas tradicionales de las carreteras convencionales en los países desarrollados, como muchos estudios previos han señalado (Evans, 2001; Abdel-Aty, 2003; Gray et al., 2008; Helai et al., 2008; Pande and Abdel-Aty, 2009; Xie et al., 2009; Kashani and Mohyamany, 2011; Montella, 2011; Mujalli and De Oña, 2011; De Oña, 2011; De Oña et al., 2013). Este resultado permite validar la metodología propuesta para el estudio de los accidentes.

Las RDs (que se extraen de los ADDs) permiten identificar determinados patrones de comportamiento de un modo fácilmente comprensible, que permite a los gestores de seguridad vial, y a las Administraciones competentes, que puedan establecer determinadas contramedidas y/o acciones de carácter preventivo.

Con el enfoque establecido en los resultados, las acciones deben ser dirigidas en un primer lugar sobre los accidentes mortales y graves, para posteriormente actuar sobre los accidentes leves. En el enfoque de esta investigación, permite además, dentro de cada grupo de accidentes (HGM o HL) priorizar las acciones en base a los valores del support, population, confidence y lift.

6.2. Conclusions.

This chapter presents the mains conclusions from the research work developed in this Ph.D Tesis.

Tha main goal of many safety research has been to understand and identify the factors that have an impact on traffic crashes'severity. Discrete Outcome Models have been the techniques most commonly use to perform this analysis.

Accident data have some particularities such as underreporting, the presence of endogenous variables, correlations, and so on; and for this reason, researchers have been introducing various modifications of the original models. Thus, the choice of the model has been based on detected problems in analyzing accidents and in the objective of each study.

However, most of these models have their own model assumptions and pre-defined underlying relationships between dependent and independent variables. If these assumptions are violated, the model could lead to erroneous estimations of the likelihood of injury severity (Chang and Wang, 2006).

For this reason, researchers have begun to use Data Mining Techniques (DM) to analyze the severity of traffic accidents.

DT techniques aimed at extracting knowledge from large amounts of data previously unknown and indistinguishable, and usually they are not based on assumptions.

DT techniques are particularly appropriate for studying crashes; the reason is related also to the nature of the crash phenomenon. A crash can be defined as a rare, random, multi-factor event always preceded by a situation in which one or more road users fail to cope with the road environment (ROSPA, 2002). Each crash is the result of a chain of events which is, in its entirety, unique, but some factors are common to several crash circumstances and the identification of these factors and their interdependences can be done using DM techniques (Montella et al., 2012b).

Within DM models there are different types of techniques. There is no universal technique that can be applied to solve any problem (Hernández et al., 2004). But each technique has its advantages and disadvantages, and the choice of it depends on the objective of the analyst; being the Artificial Neural Network, Bayesian Network, Association Rules and Decision Trees (DTs), the most widely used in the field of road safety.

In this research, DTs have been used in order to analyze traffic accidents. Considering its advantages and limitations, they are one of the most used models in supervised learning and DM applications (Gehrke et al., 1999). Among its advantages for the study of accidents can be highlighted the following: they usually a non-parametric technique, and do not establish prior relationship between the variables. They are easy to interpret, can work with large databases, and easily discover complex interactions between data. In addition, they permit the extraction of the Decision Rules (DRs) of

the "IF-THEN" type. These rules can be used to discover certain patterns of behavior that occur within a specified set of data.

The accident data used to conduct this research have been provided by the Spanish General Traffic Accident Directorate (DGT). In order to analyze road accidents on rural road, a pretreatment of the data was carried out. Accidents at intersections were deleted, since the factors related to these accidents are different, and therefore it is recommended to study them separately (Moore et al., 2010).

The severity of the accident depends on factors of different nature (such as the vehicle, the road or the driver). In addition, the factors that affect to single vehicle accidents may be different from those to multiple collisions (more than one vehicle involved). In multiple accidents, the severity is highly correlated with factors such as the type of collision, the size and weight of the vehicles involved, contact points and so on (Krull et al., 2000). Therefore, in this research only single vehicle accidents were analyzed.

Finally, given the availability of data, we have analyzed the accidents occurred during seven years, from 2003-2009. Accident data were for two-lane rural highways in the province of Granada (South of Spain). After the filtering process data, we have a sample of 1,801 accidents.

In a first phase of the research, the goal was to validate the use to DTs built with different algorithms, for the study of the severity of accidents. Also, the identification of the variables that a key effect on the severity. And, the following conclusions were obtained:

- The DTs are a tool for analyzing traffic accidents in a simple and easily manner, understandable for road safety analysts. Also they allow accident classification based on crash severity. DTs provide an alternative to parametric models due to their ability to identify patterns based on data, without the need to establish a functional relationship between variables. Moreover, classification models can be used to determine interactions between variables that would be impossible to establish directly, using ordinary statistical modelling techniques.

- There are many algorithms that can be used to build DTs, but CART is the once most commonly used in road safety researches (Chang and Wang, 2006; Kashani and Mohaymany, 2011; Kashani et al., 2011; Montella et al., 2011; Montella et al., 2012b). However, CART always yields binary trees. There are others algorithms to built DTs, such as ID3 y C4.5, and they do not have this restriction.

- In the case of road accidents, these methods may be very practical when it comes to analysing the impact of a specific category of variable in crash severity.

- DTs built with algorithms CART, ID3 and C4.5 were evaluated using the indicators: accuracy, sensitivity, specificity and ROC. The results show that ID3 algorithm is the method with worse results. However, between CART and C4.5 there are not very significant differences, and although CART showed higher values in the parameters of accuracy, specificity and ROC, a priori we cannot say that one model

is better than another. So it is appropriate to analyze the models obtained with C4.5 and CART algorithms.

- CART builds binary DTs and therefore certain categories of splitting variables are grouped in some branches, increasing node support, but making it impossible to analyze the influence of a specific category on severity. C4.5 creates a branch for each category, thereby permitting an analysis of the influence of all the categories of variables used to build the DT. Consequently, it could be said that the rules obtained with CART are less informative.

- C4.5 generates DTs with more branches than CART, and therefore it produces more rules. However, not all the rules meet the established minimal number of support, population and probability parameters, and therefore the rules may not be very useful for implementing future road safety strategies.

- The importance of the variables in the model can be obtained using either algorithm. With CART were identified 12 variables, with values of importance bigger than 9.9%. The variable lighting, atmospheric condition, time, type of accident, and gender were the most important variables in the model. These results agree with those obtained in previous studies (Evans, 2001; Al-Ghamdi, 2002; Kcoleman and Kweon, 2002; Abdel-Aty, 2003; Ulfarsson and Mannering, 2004; Gray et al., 2008; Helai et al., 2008; De Oña et al., 2011; Pande and Abdel-Aty, 2009; Xie et al., 2009; Kashani and Mohyamany, 2011; Mujalli and De Oña, 2011; Obeng, 2011). C4.5 identify 14 variables, being 11 of them also identified with CART In order of importance, the first is the type of accident, followed by causes, sex, lighting and the type of vehicle.

- DTs permit certain potentially useful rules to be determined that can be used by road safety analysts and managers.

- From a management perspective of road safety, it should highlight some general conclusions about the particular results obtained:

  - Male drivers are the main cause of crashes with killed or seriously injured (KSI).
  - The probability of KSI increases if pedestrians are involved. From the point of view of road safety, it could act on these accidents by placing barriers at road sections where pedestrian traffic may exist (remember that rural roads have been analyzed, then these accidents are located on roads that connect two nearby towns).
  - When women drivers are involved in an accident, both methods predict SI when lighting exists (full daylight, sufficient lighting and dusk). However, both methods predict KSI when the lighting is non-existent or insufficient. These rules are not observed for men and may indicate that women increase their risk of severity under conditions of less lighting on the road.

- Finally, it should be stressed that each method has advantages and drawbacks, and reveals different information. Therefore, the two methods complement each other and the recommendation is to use both of them for a full analysis.

Since DRs obtained from a single DT are strongly influenced by the variable used as root node, it can lead to loss of other important information about the variable under study. Therefore, in a second phase of this research the aim was to use a specific methodology (Information Variation Root Node) that solved this limitation, and allow complete extraction of the existing knowledge in the database under study. And, the following conclusions were obtained:

- When a single DT is used to extract knowledge from a database, the number of patterns obtained is much lower than when applying the proposed method (IRNV). Therefore, for the complete study of a database, is recommended use IRNV method.

- The main issue of this method is that most rules are extracted from DTs in which the root node has been imposed, and this node (or this variable) would not be essential for the pattern that describes the rule. Therefore, the rule should be simplified. And so, it is necessary a prior verification of the importance of this variable, in order to know if the original pattern (DR) can be simplified (DR^{-}). The results show that this step is very important.

- The implementation of any corrective road safety measure has a cost, and the available resources are limited, so, we must extract the strongest patterns that describe the problems of road safety. Therefore, the variables of each pattern were verified using the Increase Lift Criterion. Only patterns with an improvement in this criterion were keeping for the road safety analysis. Note that this step is also very important because many patterns are reduced to one, two or three variables in the antecedent.

- In response to the particular results obtained with each method, both methods identify some patterns that are exactly the same, and most of them are very similar. Thus, on a global road safety issues identified are the same with both methods, for example, accidents and collisions with pedestrians, serious accidents for run off the road and motorcycles, and minor accidents with women drivers, and so on. However, each method on each general problem identified certain particular variables. It is very consistent, because the database of accidents analyzed is the same for both methods, and the general road safety issues must be equal. However, the methods used for the construction of the DTs and subsequent extraction of the rules are different. They are based on different criteria partition (purity measure in GI is based on diversity index, while in RGI is entropy). Therefore, each method can perform various database partitions, and thus may identify specific variables for a given overall pattern, and even identify others additional patterns.

- Regarding the particular results obtained, from the point of view of road safety, we can highlight the following specific findings:

- The relationship between serious accidents caused by collisions with pedestrians is obtained with both criteria. So this confirms that vulnerable road users are more risk of suffer a serious accident. With the patterns obtained, it is showed that the type of vehicle involved in these accidents is usually a light vehicle, that in holidays increase the likelihood of these accidents, and that these accidents are variables related to road margins, so specific measures of road sections with pedestrian traffic can help improve road safety thereof.
- Particular serious patterns by run off the road are also obtained with both methods. All rules obtained by GI show KSI accidents off the road on which the vehicle involved is a motorcycle. With the RGI method, this pattern is also identified in one rule. It also shows the influence of the shoulder in these accidents. In this regard it is important that competent authorities make efforts to reduce such accidents.
- Both methods identify patterns of minor accidents with women. Even the RGI method shows a direct pattern in which does not influence any other variable, and from this pattern other more specific with both methods are obtained.
- And with the two methods are also identified patterns of minor accidents when weather conditions are adverse.

- So for a complete extraction of existing knowledge in the database analyzed, and overall management of the problems of roads, we recommend the combined use of both criteria (GI and RGI).

- Finally, we may note that in this thesis, we have analyzed the specific problems of rural roads of the province of Granada. However, the methodology presented is of interest for the study of existing problems in other roads, or in a more general.

From the standpoint of road safety, most of the rules extracted coincide with the conventional problems found on rural highways in developed countries, as most previous studies point out. This validates the method proposed in this paper, and therefore it is positive. However, the primary importance of this proposal is that other data bases not used here (i.e. other infrastructure, roads and countries) could be used to identify unconventional problems in a manner easy for road safety managers to understand, as decision rules (Evans, 2001; Abdel-Aty, 2003; Gray et al., 2008; Helai et al., 2008; Pande and Abdel-Aty, 2009; Xie et al., 2009; Kashani and Mohyamany, 2011; Montella, 2011; Mujalli and De Oña, 2011; De Oña, 2011; De Oña et al., 2013). This validates the method proposed for study accidents.

DTs permit certain potentially useful rules to be determined that can be used by road safety analysts and managers. Initially, they should focus on severe crashes and subsequently intervene in minor accidents. The approach proposed in this research permits within each group will enable actions to be prioritized on the basis of support, population, confidence and lift.

CAPÍTULO 7. FUTURAS LÍNEAS DE INVESTIGACIÓN

CAPÍTULO 7. FUTURAS LÍNEAS DE INVESTIGACIÓN

7.1. Futuras líneas de investigación.

La metodología desarrollada en la presente tesis doctoral, así como los resultados obtenidos de su aplicación pueden convertirse en una herramienta interesante para el estudio de la seguridad vial de cara a la prevención de accidentes y a la reducción de su gravedad.

Sin embargo, como se ha citado a lo largo del presente trabajo de investigación, hay ámbitos tratados en el mismo que precisan investigaciones complementarias:

- En esta investigación sólo accidentes con un vehículo involucrado han sido analizados. En el futuro se pretende ampliar el estudio a accidentes múltiples.
- Esta investigación se ha centrado exclusivamente en el estudio de las carreteras convencionales de la provincia de Granada. En las futuras investigaciones se pretende ampliar el ámbito de estudio a las carreteras convencionales de toda Andalucía (ya comenzado), y nivel nacional.
- También como futura línea de investigación se plantea el uso de la metodología aplicada en esta tesis para analizar los patrones de accidentes en otras tipologías de carretera (autovías), en localizaciones particulares (intersecciones), así como extenderla al estudio de accidentes en ámbito urbano.

Como se ha mostrado a lo largo del desarrollo de esta tesis doctoral, analizar la gravedad de los accidentes de tráfico no es una tarea fácil de resolver y, por ello, los investigadores están comenzando a utilizar las nuevas técnicas de análisis de datos de las que disponemos hoy día. Respecto a las técnicas de análisis, se proponen las siguientes líneas de investigación:

- Dada la naturaleza ordinal de la gravedad del accidente (accidentes mortales, graves, leves y solo daños materiales), se pueden utilizar modelos de clasificación específicos del campo de la Minería de Datos que tienen en cuenta la estructura jerárquica de los datos, y a partir de estos modelos, se pueden obtener los patrones de comportamiento particulares.
- Cuando la variable clase es no balanceada, es decir, no existe homogeneidad entre sus categorías, como es el caso de la gravedad del accidente cuando se tienen en cuenta todos sus niveles: accidentes leves, accidentes graves y accidentes mortales,

los ADDs presentan problemas en la predicción de las categorías infra-representadas. En futuros trabajos se pretende realizar el análisis específico de cada una de las categorías de gravedad, y por ello se requerirán de técnicas que puedan trabajar con este tipo de datos. Algunas de las soluciones que se proponen en la literatura se centran en el proceso de pre-tratamiento de los datos: duplicar los casos que corresponden a la clase infra-representada en el conjunto de *training* (Ling and Li, 1998), eliminar casos de la muestra sobre-representada (Kubat and Matwin, 1997).

- La heterogeneidad de los datos es un problema que puede producir que determinadas relaciones existentes en los datos permanezcan ocultas, si no es reducida previamente (Depaire et al., 2008). Por ello se propone como futura línea de investigación, el uso de técnicas cluster que permiten segmentar los datos y crear mayor homogeneidad en los mismos, para posteriormente aplicar ADDs.

- En el método IRNV se pueden implementar nuevos métodos de construcción de ADDs. Por ejemplo, métodos de construcción de ADDs que utilizan criterios de partición basados en probabilidades imprecisas, como los que se exponen en Abellán and Moral (2003). Estos criterios de partición generan árboles muy diferentes a los obtenidos con los criterios de partición clásicos y pueden proporcionar información adicional sobre la base de datos analizada.

CAPÍTULO 8. REFERENCIAS BIBLIOGRÁFICAS

CAPÍTULO 8.
REFERENCIAS BIBLIOGRÁFICAS

Abdel-Aty, M. (2003). Analysis of driver injury severity levels at multiple locations using ordered probit models. Journal of Safety Research 34 (5), pp. 597-603.

Abdel-Aty, M., Abdelwahab, H. (2004). Modeling rear-end collisions including the role of driver's visibility and light truck vehicles using a nested logit structure. Accident Analysis and Prevention 36 (3), pp. 447–456.

Abdelwahab H. and Abdel-Aty M. (2002). Investigating Driver Injury Severity in Traffic Accidents Using Fuzzy ARTMAP, Journal of Computer-Aided Civil and Infrastructure Engineering 17, pp. 396-408.

Abdelwahab, H.T., Abdel-Aty, M.A. (2001). Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized inter-sections. Transportation Research Record 1746, pp. 6-13.

Abellán, J., Masegosa, A. (2010). An ensemble method using credal decision trees. European Journal of Operational Research 205 (1), pp. 218–226.

Abellán, J., Moral, S. (2003). Building classification trees using the total uncertainty criterion. International Journal of Intelligent Systems 18 (12), pp. 1215–1225.

Abellán, J., Masegosa, A. (2010). An ensemble method using credal decision trees. European Journal of Operational Research 205 (1), pp. 218–226.

Agrawal, R., Imielinski, T., Swami, A. (1993). Mining association rules between sets of items in large databases. In: Proceedings of ACM SIGMOD Conference on Management of Data, pp. 207–216

Al-Ghamdi, A. (2002). Using logistic regression to estimate the influence of accident factors on accident severity. Accident Analysis and Prevention 34 (6), pp. 729–741.

American Medical Association, Committee on Medical Aspects of Automotive Safety (1971). Rating the Severity of Tissue Damage. I The Abbreviated Scale. Journal of the American Medical Association 215, pp. 277-280.

Anselin, L., Florax, R., Rey, S. (2005). Advances in Spatial Econometrics: Methodology, Tools and Application. Springer-Verlag, Berlin, Germany.

Baker SP, O'Neill B, Haddon W, Long WB. (1974). The injury severity score: a method for describing patients with multiple injuries and evaluating emergency care. Journal of Trauma 14, pp. 187-196.

Bayam, E., Liebowitz, J., Agresti, W. (2005). Older drivers and accidents: a meta analysis and data mining application on traffic accident data. Expert Systems with Applications 29 (3), pp. 598–629.

Bedard, M., Guyatt, G., Stones, M., Hirdes, J. (2002). The independent contribution of driver, crash, and vehicle characteristics to driver fatalities. Accident Analysis and Prevention 34 (6), pp. 717–727.

Ben-Akiva, M., Lerman, S. (1985). Discrete Choice Analysis: Theory and Application to Travel Demand. MIT Press, Cambridge, MA.

Blincoe, L., Seay, A., Zaloshnja, E., Miller, T., Romano, E., Luchter, S., Spicer, R. (2002). The economic impact of motor vehicle crashes, NHTSA Technical Washington, DC.

Breiman, L., Friedman, J., Olshen, R., Stone, C. (1984). Classification and Regression Trees. Chapman & Hall, Belmont, CA.

Brijs, T., Swinnen, G., Vanhoof, K., Wets.G. (1999). The use of association rules for product assortment decisions: a case study". In: Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining, San Diego (USA), pp. 254-260.

Chang, H., Yeh, T. (2006). Risk factors to driver fatalities in single-vehicle crashes: comparisons between non-motorcycle drivers and motorcyclists. Journal of Transportation Engineering 132 (3), pp. 227–236.

Chang L.-Y., Chien, J.T. (2013). Analysis of driver injury severity in truck involved accidents using a non-parametric classification tree model. Safety Science 51, pp. 17–22.

Chang, L.-Y., Mannering, F. (1999). Analysis of injury severity and vehicle occupancy in truck- and non-truck-involved accidents. Accident Analysis and Prevention 31 (5), pp. 579–592.

Chang, L.-Y., Mannering, F. (1998). Predicting vehicle occupancies from accident data: an accident severity approach. Transportation Research Record 1635, pp. 93–104.

Chang, L.-Y., Wang, H.W. (2006). Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accident Analysis and Prevention 38, pp. 1019-1027.

Chang, H., Yeh, T. (2006). Risk factors to driver fatalities in single-vehicle crashes: comparisons between non-motorcycle drivers and motorcyclists. Journal of Transportation Engineering 132 (3), pp. 227–236.

Chang, L.-Y.,Chen, W.-C. (2005). Data mining of tree-based models to analyze freeway accident frequency. Journal of Safety Research 36 (4), pp. 365-375.

Christoforou, Z., Cohen, S., Karlaftis M.G. (2010). Vehicle occupant injury severity on highways: an empirical investigation. Accident Analysis and Prevention, 42, pp. 1606–1620.

Daganzo, C.F. (1979) Multinomial Probit: The Theory and its Applications to Demand Forecasting. Academic Press, Nueva York.

Daniello, A., Gabler, H.C. (2011). Fatality risk in motorcycle collisions with roadside objects in the United States. Accident Analysis and Prevention 43, pp. 1167–1170.

Delen, D., Sharda, R., Bessonov, M. (2006). Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. Accident Analysis and Prevention 38 (3), pp. 434–444.

De Meester, D. (2011). Recommendation for a Common Accident Data Set, Reference Guide, Version 3.11.

Daniello, A., Gabler, H.C. (2011). Fatality risk in motorcycle collisions with roadside objects in the United States. Accident Analysis and Prevention 43, pp. 1167–1170.

De Oña, J., Mujalli, R.O., Calvo, F.J. (2011). Analysis of traffic accident injury on Spanish rural highways using Bayesian networks. Accident Analysis and Prevention 43, pp. 402–411.

De Oña, J., López, G., Mujalli, R.O., Calvo, F.J. (2013). Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks. Accident Analysis and Prevention 51, pp. 1–10.

Depaire, B., Wets, G., Vanhoof, K. (2008). Traffic accident segmentation by means of latent class clustering. Accident Analysis and Prevention 40 (4), pp. 1257–1266.

Dissanayake, S. (2004). Comparison of severity affecting factors between young and older drivers in single vehicle crashes. IATSS Research 28 (2), pp. 48-54.

Donelson, A., Ramachandran, K., Zhao, K. & Kalinowski, A. (1999). Rates of occupant deaths in vehicle rollover: Importance of fatality-risk factors. Transportation Research Record 1665 (1), pp. 109-117.

EC, European Commission (2004). CADaS - The Common Accident Data Set. D.1.14 CADaS.

EC, European Comission (2010). Towards a European road safety area: policy orientations on road safety 2011-2020. COM (2010) 389 final.

Eluru, N., Paleti, R., Pendyala, R., Bhat, C. (2010). Modeling multiple vehicle occupant injury severity: a copula-based multivariate approach. Transportation Research Record 2165, pp. 1–11.

Evans, L. (2001). Female compared with male fatality risk from similar physical impacts. The Journal of Trauma: Injury, Infection and Critical Care 50, pp. 281–288.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., (1996a). From Data Mining to Knowledge Discovery: An Overview. In: Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P.,

Uthurusamy, R. (Eds.), Advances in Knowledge Discovery and Data Mining. AAAI Press/The MIT Press, pp. 1–34.

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. and Uthursamy, R. (1996b). Advances in Knowledge Discovery and Data Mining. California: AAAI/MIT Press.

Fredette, M. Mambu, L., Chouinard, A., Bellavance, F. (2008). Safety impacts due to the incompatibility of SUVs, minivans, and pickup trucks in two-vehicle collisions. Accident Analysis and Prevention 40, pp. 1987–1995.

Garder, P. (2006). Segment characteristics and severity of head-on crashes on two-lane rural highways in Maine. Accident Analysis and Prevention 38 (4), pp. 652–661.

Gehrke, J., Ganti, V., Ramakrishnan, R., Loh, W-Y. (1999). BOAT-Optimistic Decision Tree Construction. SIGMOD Conference, pp. 169-180.

Geurts, K., Thomas, I., Wets, G. (2005). Understanding spatial concentrations of road accidents using frequent item sets. Accident Analysis and Prevention 37 (4), pp. 787–799.

Giacomo, C., Rasmussen, T., Kaplan, S. (2013). Risk Factors Associated with Crash Severity on Low-Volume Rural Roads in Denmark. Journal of Transportation Safety & Security. DOI: 10.1080/19439962.2013.796027

Goldstein, H. (2003). Multilevel Statistical Models, 3nd ed. Edward Arnold, London.

Gray, R.C., Quddus, M.A., Evans, A. (2008). Injury severity analysis of accidents involving young male drivers in Great Britain. Journal of Safety Research 39 (5), pp. 483-495.

Gurney K., 1997. An introduction to neural networks. London (UK): UCL Press.

Haleem, K., Abdel-Aty, M. (2010). Examining traffic crash injury severity at unsignalized intersections. Journal of Safety 41 (4), pp. 347-357

Heckerman, D. Geiger, D. Chickering, D. (1995). Learning Bayesian networks: The Combination of Knowledge and Statistical Data. Machine Learning 20, pp. 197-243

Helai, H., Chor, C., Haque, M. (2008). Severity of driver injury and vehicle damage in traffic crashes at intersections: a Bayesian hierarchical analysis. Accident Analysis and Prevention 40 (1), pp.45–54.

Hernández Orallo, J., Ramírez Quintana, M. J. y Ferri Ramírez, C. (2004). Introducción a la minería de datos. Madrid: Pearson Educación S.A.

Holdridge, J., Shankar, V., Ulfarsson, G. (2005). The crash severity impacts of fixed roadside objects. Journal of Safety Research 36 (2), pp. 139–147.

IRTAD - International Traffic Safety Data and Analysis Group (2012). Road Safety Annual Report 2011. Paris, France.

IRTAD - International Traffic Safety Data and Analysis Group (2012b). Reporting on Serious Road Traffic Casualties: combining and using different data sources to improve understanding of non-fatal road traffic crashes. Paris, France.

Islam, S., Mannering, F. (2006). Driver aging and its effect on male and female single-vehicle accident injuries: some additional evidence. Journal of Safety Research 37 (3), pp. 267–276.

Jung, S., Qin, X. & Noyce, D.A. (2010). Rainfall effect on single-vehicle crash severities using polychotomous response models. Accident Analysis and Prevention 42 (1), pp. 213-224.

Kashani, A., Mohaymany, A. (2011). Analysis of the traffic injury severity on twolane, two-way rural roads based on classification tree models. Safety Science 49, pp. 1314–1320.

Kashani, A., Mohaymany, A., Ranjbari, A. (2011). A data mining approach to identify key factors of traffic injury severity. Promet-Traffic and Transportation 23 (1), pp. 11–17.

Kass, G.V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. Applied Statistics 29 (2), pp.119–127.

Khattak, A., Rocha, M. (2003). Are SUVs 'Supremely Unsafe Vehicles'? Analysis of rollovers and injuries with sport utility vehicles. Transportation Research Record 1840, pp. 167–177.

Khorashadi, A., Niemeier, D., Shankar, V., Mannering, F. (2005). Differences in rural and urban driver-injury severities in accidents involving large-trucks: an exploratory analysis. Accident Analysis and Prevention 37 (5), pp. 910–921.

Kockelman, K. M., Kweon, Y. J. (2002). Driver injury severity: an application of ordered probit models. Accident; Analysis and Prevention 34, pp. 313–321.

Kononen, D.W., Flannagan, C.A.C., Wang, S.C. (2011). Identification and validation of a logistic regression model for predicting serious injuries associated. Accident Analysis and Prevention 43 (1), pp. 112-122

Krull, K., Khattak, A., Council, F. (2000). Injury effects of rollovers and events sequence in single-vehicle crashes. Transportation Research Record 1717, pp. 46–54.

Kubat M, Matwin S. (1997). Addressing the curse of imbalanced training sets: one-sided selection. Proc. 14th Intl.Conf. on Machine Learning, pp. 179–186.

Kuhnert, P.M., Do, K.A., McClure, R. (2000). Combining non-parametric models with logistic regression: an application to motor vehicle injury data. Computational Statistics & Data Analysis 34 (3), pp. 371–386.

Lee, C., Abdel-Aty, M. (2008). Presence of passengers: does it increase or reduce driver's crash potential? Accident Analysis and Prevention 40 (5), pp. 1703–1712.

Lee, J., Mannering, F. (2002). Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis. Accident Analysis and Prevention 34 (2), pp. 149–161.

Lemp, J.D., Kockelman, K.M., Unnikrishnan, A. (2011). Analysis of large truck crash severity using heteroskedastic ordered probit models. Accident Analysis and Prevention 43 (1), pp. 370-380.

Lewis RJ. (2000). An Introduction to Classification and Regression Trees (CART) Analysis. 2000 Annual Meeting of the Society for Academic Emergency Medicine In San Francisco, California.

Ling CX, Li C. (1998). Data mining for direct marketing: problems and solutions. In: Proc. 4th Intl. Conf. on Knowledge Discovery and Data Mining, pp. 73–79.

Lord, D., Mannering, F. (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. Transportation Research Part A, 44, pp. 291-305.

Mayhew, D. R., Simpson, H. M., Pak, A. (2003). Changes in collision rates among novice drivers during the first months of driving. Accident Analysis and Prevention 25, pp. 683−691.

McCartt, A. T., Mayhew, D. R., Braitman, K. A., Ferguson, S. A., Simpson, H. M. (2009). Effects of age and experience on young driver crashes: review of recent literature. Traffic Injury Prevention 10(3), pp. 209−219.

McFadden, D. (1973). Conditional Logit Analysis of Qualitative Choice Models Frontiers in Econometrics, Zarembka (ed,), New York: Academic Press.

McFadden, D., Train, K. (2000). Mixed MNL models for discrete response. Journal of Applied Econometrics 15 (5), pp. 447–470.

Michalski, R.S., Baskin, A.B., Spackman, K.A. (1982). A Logic-Based Approach to Conceptual Database Analysis. Sixth Annual Symposium on Computer Applications on Medical Care, George Washington University, Medical Center, Washington, DC, EE.UU.

Ministerio de Relaciones con las Cortes y la Secretaría del Gobierno. (1993). Orden Ministerial por la que se modifica la estadística de accidentes de circulación. Boletín Oficial del Estado, 24 de febrero de 1993, 47, pp. 6016-620.

Ministerio del Interior (2005). Plan Estratégico de Seguridad Vial 2005-2008 Disponible en:

http://www.dgt.es/portal/es/seguridad_vial/planes_seg_vial/plan_estrategico/

Ministerio del Interior (2010). Las principales de la Siniestralidad Vial. España 2010. General de Tráfico, Madrid, Madrid. Disponible en:

http://www.dgt.es/was6/portal/contenidos/es/seguridad_vial/estadistica/publicaciones/princip_cifras_siniestral/cifras_siniestralidadl011.pdf

Ministerio del Interior (2011). Estrategia de Seguridad Vial 2011-2020. Dirección General de Tráfico, Madrid. Disponible en:

http://www.dgt.es/was6/portal/contenidos/documentos/seguridad_vial/planes_seg_vial/estrategico_seg_vial/estrategico_2020_004.pdf

Moghaddam, F.R., Afandizadeh, S., Ziyadi, M. (2011). Prediction of accident severity using artificial neural networks. International Journal of Civil Engineering 9 (1), pp. 41-49

Montella A., Andreassen D., Tarko A., Turner S., Mauriello F., Imbriani L., Singh R. (2012a). Critical Review of the International Crash Databases and Proposals for Improvement of the Italian National Databases. Procedia - Social and Behavioral 53, pp. 49-61.

Montella A., Aria M., D'Ambrosio A., Mauriello F. (2011). Data Mining Techniques for Exploratory Analysis of Pedestrian Crashes. Transportation Research Record 2237, pp. 107–116.

Montella, A., Pernetti, M. (2010). In-depth investigation of run-off-the-road crashes on the Motorway Naples-Candela. In: Presented at the 4th International Symposium on Highway Geometric Design of the Transportation Research Board, Valencia.

Montella, A. (2011). Identifying crash contributory factors at urban roundabouts and using association rules to explore their relationships to different crash types. Accident Analysis and Prevention 43, pp. 1451–1463.

Montella A., Aria M., D'Ambrosio A., Mauriello F. (2012b). Analysis of powered two-wheeler crashes in Italy by classification trees and rules discovery. Accident Analysis and Prevention 49, pp. 58–72.

Moore, D.N., Schneider IV, W.H., Savolainenb, P.T., Farzaneh, M. (2010). Mixed logit analysis of bicyclist injury severity resulting from motor vehicle crashes at intersection and non-intersection locations. Accident Analysis and Prevention, Accident Analysis and Prevention 43 (3), 621-630.

Moore, E.E., Cogbill, T.H., Jurkovich, G.J., McAninch J.W., Champion H.R., Gennarelli T.A., Malangoni M.A., Shackford S.R., Trafton P.G. (1992). Organ injury scaling III: chest wall, abdominal vascular, ureter, bladder and urethra. J Trauma 33, pp. 337-339.

Mujalli, R.O., de Oña, J. (2011). A method for simplifying the analysis of traffic accidentsinjury severity on two-lane highways using Bayesian networks. Journal of Safety Research 42, pp. 317–326.

Murthy, S.K. (1998). Automatic construction of decision trees from data: A multi-disciplinary survey. Data Mining and Knowledge Discovery 2 (4), pp. 345-389.

NHTSA - National Highway Traffic Safety Administration (2008). "MMUCC Guideline, Model Minimum Uniform Crash Criteria". Third Edition. Report DOT HS 810 957, US Department of Transportation, Washington, D.C., USA.

Obeng, K. (2011). Gender differences in injury severity risks in crashes at signalized intersections. Accident Analysis and Prevention 43 (4), pp. 1521–1531.

O'Donnell, C., Connor, D. (1996). Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice. Accident Analysis and Prevention 28 (6), pp. 739–753.

Ouyang, Y., Shankar, V., Yamamoto, T. (2002). Modeling the simultaneity in injury causation in multi-vehicle collisions. Transportation Research Record 1784, pp. 143–152.

Pakgohar, A., Tabrizi, R.S., Khalilli, M., Esmaeili, A. (2010). The role of human factor in incidence and severity of road crashes based on the CART and LR regression: a data mining approach. Procedia Computer Science 3, pp. 764–769.

Paleti, R., Eluru, N., Bhat, C. (2010). Examining the influence of aggressive driving behavior on driver injury severity in traffic crashes. Accident Analysis and Prevention 42 (6), pp. 1839–1854.

Pande, A., Abdel-Aty, M. (2009). Market basket analysis of crash data from large jurisdictions and its potential as a decision supporting tool. Safety Science, 47, pp. 145–154.

Pearl, J. (1988). Probabilistic Reasoning in Intelligence Systems. Morgan Kaufmann, San Mateo, CA.

Peek-Asa, C., Britton, C., Young, T., Pawlovich, M., Falb, S. (2010). Teenage driver crash incidence and factors influencing crash injury by rurality. Journal of Safety Research 41 (6), pp. 487–492.

Perandones, J.M., Molinero, A., Martin, C., Mansilla, A., Pedrero, D. (2008). Recommendations for location of motorcyclist protection devices in Spanish regional road network of Castilla y Leon. In: Presented at the 87th Annual Meeting of the Transportation Research Board , Washington, DC.

Pérez López, C. (2007). Minería de datos. Técnicas y herramientas. Madrid. Thomson.

Piatetski-Shapiro, G., Frawley, W.J., Matheus, C.J. (1991). Knowledge discovery in databases: an overview. AAAI-MIT Press, Menlo Park, California.

Quddus, M.A., Noland, R.B., Chin, H.C. (2002). An analysis of motorcycle injury and vehicle damage severity using ordered probit models. Journal of Safety Research 33, pp. 445–462.

Quddus, M.A., Wang, C., Ison, S.G. (2010). Road traffic congestion and crash severity: Econometric analysis using ordered response models. Journal of Transportation Engineering 136 (5), pp. 424-435.

Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), pp. 81–106.

Quinlan, J. R. (1987). Generating production rules from decision trees. In McDermott, J. (Ed.), Proc of the 10th Int Joint Conf on Artificial Intelligence. Milan, Italy: Morgan Kaufmann, Los Altos, CA.

Quinlan, J.R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, California.

Quinlan, J. R., Cameron-Jones, R. M. (1995). Oversearching and layered search in empirical learning. In Proc of the 14th Int Joint Conf on Artificial Intelligence (pp. 1019–1024). Montreal, Canada: Morgan Kaufmann, San Francisco, CA.

Ramoni, M., Sebastiani, P. (1996). Robust Parameter Learning in Bayesian Networks with Missing Data. Technical Report KMI-TR-29, Knowledge Media Institute, The Open University, October.

Rifaat, S.M., Tay, R., de Barros, A. (2011). Severity of motorcycle crashes in Calgary. Accident Analysis and Prevention 49, pp. 44-49.

Rokach, L., Maimon, O. (2008). Data Mining with Decision Trees: Theory and Applications (Series in Machine Perception and Artifical Intelligence). Singapore: World Sicientific Publishing Co. Pte. Ltd.

ROSPA – The Royal Society for Prevention of Accidents (2002). Road Safety Engineering Manual. Birmingham.

Savolainen, P., Ghosh, I. (2008). Examination of factors affecting driver injury sever-ity in Michigan's single-vehicle-deer crashes. Transportation Research Record 2078, pp. 17–25.

Savolainen, P., Mannering, F. (2007). Probabilistic models of motorcyclists' injury severities in single- and multi-vehicle crashes. Accident Analysis and Prevention 39 (5), pp. 955–963.

Savolainen, P., Mannering, F., Lord, D., Quddus, M. (2011). The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. Accident Analysis and Prevention 43, pp. 1666-1676.

Shannon, C.E. (1948). A mathematical theory of communication. Bell System Technical Journal 27, pp. 379–423, 623–656.

Simoncic, M. (2005). A Bayesian network model of two-car accidents. Journal of transportation and Statistics 7 (2-3), pp. 13–25.

Srinivasan, K.K. (2002). Injury severity analysis with variable and correlated thresholds: ordered mixed logit formulation. Transportation Research Record 1784 (1), pp. 132-142.

Toy, E., Hammitt, J. (2003). Safety impacts of SUVs, vans, and pickup trucks in two vehicle crashes. Risk Analysis 23 (4), pp. 641–650.

Train, K., 2009. Discrete Choice Methods with Simulation. Cambridge University Press, New York, NY.

Tullis, J.A., Jensen, J.R. (2003). Expert System House Detection in High Spatial Resolution Imagery Using Size, Shape, and Context. Geocarto International 18(1), pp. 5-15.

Ulfarsson, G., Mannering, F. (2004). Differences in male and female injury severities in sport-utility vehicle, minivan, pickup and passenger car accidents. Accident Analysis and Prevention 36 (2), pp. 135–147.

Wang, X., Kockelman, K.M. (2005). Use of heteroscedastic ordered logit model to study severity of occupant injury: distinguishing effects of vehicle weight and type. Transportation Research Record 1908, pp.195–204.

Washington, S., Karlaftis, M.G., Mannering, F. (2011). Statistical and Econometric Methods for Transportation Data Analysis, 2nd ed. Chapman and Hall/CRC, Boca Raton, FL.

Webb, G.I. (2007). Discovering significant patterns. Machine Learning 68, pp. 1–33.

Weiss, S. M., and Indurkhya, N. (1998). Predictive Data Mining. Morgan Kaufmann Publishers.

WHO - World Health Organization (2009). Informe Global sobre el estado de la Seguridad Vial: Tiempo para la Acción. Disponible en:

www.who.int/violence_injury_prevention/road_safety_status/2009

WHO - World Health Organization (2013). Global status report on road safety 2013: supporting a decade of action. Disponible en:

http://www.who.int/violence_injury_prevention/road_safety_status/2013/report/en/index.html

Winston, C., Maheshri, V., Mannering, F. (2006). An exploration of the offset hypothesis using disaggregate data: the case of airbags and antilock brakes. of Risk and Uncertainty 32 (2), pp. 83–99.

Witten, I.H., Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed. Morgan Kaufmann, San Francisco, CA.

Xie Y., Zhao, K., Huynh, N. (2012). Analysis of driver injury severity in rural single-vehicle crashes. Accident Analysis and Prevention 47, pp. 36-44.

Xie, Y., Zhang, Y. & Liang, F. (2009). Crash injury severity analysis using Bayesian ordered probit models. Journal of Transportation Engineering 135 (1), pp. 18-25.

Ye, F., Lord, D. (2011). Investigating the effects of underreporting of crash data on three commonly used traffic crash severity models: multinomial logit, ordered probit and mixed logit models. Transportation Research Record 2241, pp. 51-58.

Young Drivers: The Road to Safety, 2006. Organization for Economic Co-Operation and Development, Transportation Research Center. OECD Publishing, France.

Zhang, J., Lindsay, J., Clarke, K., Robbins, G., Mao, Y. (2000). Factors affecting the severity of motor vehicle traffic crashes involving elderly drivers in Ontario. Accident Analysis and Prevention 32 (1), pp. 117–125.

ANEXOS

ANEXOS

ANEXO I. ÁRBOL DE DECISIÓN 1 CREADO CON GI.

```
=== Run information ===
Scheme:       weka.classifiers.trees.IPTree -LeafEstimation MLE -
AlphOTOINSitSIMIP 0.0 -S 1.0 -StopLevel 4 -SM GiniIndex -KTH 1
Relation:     TR_pr2
Instances:    1260
Attributes:   20
              VEH
              BAR
              EDAD
              SEX
              MON
              HORA
              DIA
              HER
              OCU
              ANCAL
              MVIAL
              ANCAR
              ANARC
              APAV
              LUM
              CAT
              VISIB
              TAC
              CAU
              SEV
Test mode:    evaluate on training data
=== Classifier model (full training set) ===
IPTree

TAC = SV
|  VEH = OT
|  |  ANCAR = MED
|  |  |  EDAD = (27-60]: 0.0 (6.0/6.0) - [1.0,0.0,
|  |  |  EDAD = (20-27]: 1.0 (3.0/4.0) - [0.25,0.75,
|  |  |  EDAD = >60: 0.0 (1.0/1.0) - [1.0,0.0,
|  |  |  EDAD = DES: 0.0 (0/0) - [0,0
|  |  |  EDAD = <=20: 0.0 (3.0/3.0) - [1.0,0.0,
|  |  ANCAR = EST
|  |  |  DIA = L: 1.0 (5.0/5.0) - [0.0,1.0,
|  |  |  DIA = PF: 1.0 (2.0/2.0) - [0.0,1.0,
|  |  |  DIA = AF: 0.0 (1.0/2.0) - [0.5,0.5,
|  |  |  DIA = F: 0.0 (0/0) - [0,0,
|  |  ANCAR = ANC: 0.0 (12.0/23.0) -
[0.5217391304347826,0.4782608695652174,
|  VEH = MOT
|  |  APAV = SI
|  |  |  LUM = DIA: 1.0 (37.0/49.0) -
[0.24489795918367346,0.7551020408163265,
```

```
|   |   |  LUM = INS: 0.0 (2.0/4.0) - [0.5,0.5,
|   |   |  LUM = SI: 1.0 (26.0/29.0) -
[0.10344827586206896,0.896551724137931,
|   |   |  LUM = SL: 0.0 (2.0/4.0) - [0.5,0.5,
|   |   |  LUM = ATA: 0.0 (5.0/6.0) -
[0.8333333333333334,0.16666666666666666,
|   |  APAV = NE
|   |   |  EST = VER: 1.0 (16.0/19.0) -
[0.15789473684210525,0.8421052631578947,
|   |   |  EST = INV: 0.0 (2.0/2.0) - [1.0,0.0,
|   |   |  EST = PRI: 1.0 (8.0/13.0) -
[0.38461538461538464,0.6153846153846154,
|   |   |  EST = OTO: 1.0 (13.0/17.0) -
[0.23529411764705882,0.7647058823529411,
|   |  APAV = N
|   |   |  EST = VER: 0.0 (6.0/10.0) - [0.6,0.4,
|   |   |  EST = INV: 0.0 (5.0/5.0) - [1.0,0.0,
|   |   |  EST = PRI: 0.0 (6.0/11.0) -
[0.5454545454545454,0.45454545454545453,
|   |   |  EST = OTO: 1.0 (3.0/3.0) - [0.0,1.0,
|  VEH = VL
|   |  ANCAL = ANC
|   |   |  HER = [1]: 0.0 (164.0/274.0) -
[0.5985401459854015,0.40145985401459855,
|   |   |  HER = [>1]: 1.0 (85.0/140.0) -
[0.39285714285714285,0.6071428571428571,
|   |  ANCAL = MED
|   |   |  MVIAL = CYM: 0.0 (115.0/179.0) -
[0.6424581005586593,0.3575418994413408,
|   |   |  MVIAL = NE: 0.0 (13.0/18.0) -
[0.7222222222222222,0.2777777777777778,
|   |   |  MVIAL = CAR: 1.0 (20.0/26.0) -
[0.23076923076923078,0.7692307692307693,
|   |   |  MVIAL = MAR: 0.0 (14.0/21.0) -
[0.6666666666666666,0.3333333333333333,
|   |  ANCAL = EST
|   |   |  LUM = DIA: 0.0 (29.0/58.0) - [0.5,0.5,
|   |   |  LUM = INS: 0.0 (7.0/14.0) - [0.5,0.5,
|   |   |  LUM = SI: 1.0 (47.0/62.0) -
[0.24193548387096775,0.7580645161290323,
|   |   |  LUM = SL: 0.0 (2.0/2.0) - [1.0,0.0,
|   |   |  LUM = ATA: 1.0 (3.0/4.0) - [0.25,0.75,
|  VEH = VP
|   |  MVIAL = CYM
|   |   |  DIA = L: 1.0 (16.0/23.0) -
[0.30434782608695654,0.6956521739130435,
|   |   |  DIA = PF: 1.0 (6.0/6.0) - [0.0,1.0,
|   |   |  DIA = AF: 0.0 (2.0/2.0) - [1.0,0.0,
|   |   |  DIA = F: 1.0 (4.0/4.0) - [0.0,1.0,
|   |  MVIAL = NE: 0.0 (1.0/1.0) - [1.0,0.0,
|   |  MVIAL = CAR
|   |   |  EST = VER: 0.0 (1.0/1.0) - [1.0,0.0,
|   |   |  EST = INV: 0.0 (1.0/1.0) - [1.0,0.0,
|   |   |  EST = PRI: 1.0 (1.0/1.0) - [0.0,1.0,
|   |   |  EST = OTO: 0.0 (0/0) - [0,0
|   |  MVIAL = MAR: 0.0 (4.0/4.0) - [1.0,0.0,
TAC = VUE
|  CAU = CON
```

```
|   |  HORA = (12-18]
|   |  |  OCU = [1]: 0.0 (12.0/15.0) - [0.8,0.2,
|   |  |  OCU = [>2]: 1.0 (1.0/1.0) - [0.0,1.0,
|   |  |  OCU = [2]: 0.0 (3.0/3.0) - [1.0,0.0,
|   |  HORA = [0-6]
|   |  |  HER = [1]: 1.0 (4.0/5.0) - [0.2,0.8,
|   |  |  HER = [>1]: 0.0 (4.0/4.0) - [1.0,0.0,
|   |  HORA = (18-24]
|   |  |  ANCAL = ANC: 0.0 (7.0/14.0) - [0.5,0.5,
|   |  |  ANCAL = MED: 0.0 (4.0/4.0) - [1.0,0.0,
|   |  |  ANCAL = EST: 1.0 (6.0/6.0) - [0.0,1.0,
|   |  HORA = (6-12]
|   |  |  ANARC = EST: 1.0 (6.0/6.0) - [0.0,1.0,
|   |  |  ANARC = NE: 0.0 (0/0) - [0,0,
|   |  |  ANARC = MED: 0.0 (1.0/1.0) - [1.0,0.0,
|   CAU = COF
|   |  APAV = SI: 0.0 (3.0/3.0) - [1.0,0.0,
|   |  APAV = NE
|   |  |  EDAD = (27-60]: 1.0 (2.0/3.0) -
[0.3333333333333333,0.6666666666666666,
|   |  |  EDAD = (20-27]: 0.0 (1.0/1.0) - [1.0,0.0,
|   |  |  EDAD = >60: 0.0 (0/0) - [0,0,
|   |  |  EDAD = DES: 1.0 (4.0/6.0) -
[0.3333333333333333,0.6666666666666666,
|   |  |  EDAD = <=20: 1.0 (2.0/2.0) - [0.0,1.0,
|   |  APAV = N: 0.0 (4.0/4.0) - [1.0,0.0,
|   CAU = VEH: 1.0 (3.0/3.0) - [0.0,1.0,
|   CAU = OT: 0.0 (0/0) - [0,0,
|   CAU = CAR: 0.0 (7.0/7.0) - [1.0,0.0,
TAC = CP
|   APAV = SI
|   |  DIA = L
|   |  |  HORA = (12-18]: 1.0 (10.0/10.0) - [0.0,1.0,
|   |  |  HORA = [0-6]: 1.0 (1.0/1.0) - [0.0,1.0,
|   |  |  HORA = (18-24]: 1.0 (15.0/18.0) -
[0.16666666666666666,0.8333333333333334,
|   |  |  HORA = (6-12]: 1.0 (4.0/7.0) -
[0.42857142857142855,0.5714285714285714,
|   |  DIA = PF
|   |  |  EST = VER: 0.0 (1.0/1.0) - [1.0,0.0,
|   |  |  EST = INV: 0.0 (1.0/1.0) - [1.0,0.0,
|   |  |  EST = PRI: 1.0 (1.0/1.0) - [0.0,1.0,
|   |  |  EST = OTO: 1.0 (1.0/1.0) - [0.0,1.0,
|   |  DIA = AF
|   |  |  HORA = (12-18]: 0.0 (2.0/2.0) - [1.0,0.0,
|   |  |  HORA = [0-6]: 1.0 (4.0/5.0) - [0.2,0.8,
|   |  |  HORA = (18-24]: 1.0 (2.0/3.0) -
[0.3333333333333333,0.6666666666666666,
|   |  |  HORA = (6-12]: 0.0 (3.0/3.0) - [1.0,0.0,
|   |  DIA = F: 1.0 (12.0/12.0) - [0.0,1.0,
|   APAV = NE
|   |  EST = VER
|   |  |  DIA = L: 0.0 (4.0/4.0) - [1.0,0.0,
|   |  |  DIA = PF: 1.0 (1.0/1.0) - [0.0,1.0,
|   |  |  DIA = AFH: 1.0 (1.0/1.0) - [0.0,1.0,
|   |  |  DIA = F: 1.0 (1.0/1.0) - [0.0,1.0,
|   |  EST = INV
|   |  |  EDAD = (27-60]: 1.0 (1.0/1.0) - [0.0,1.0,
```

```
|   |   |  EDAD = (20-27]: 0.0 (4.0/5.0) - [0.8,0.2,
|   |   |  EDAD = >60: 0.0 (0/0) - [0,0,
|   |   |  EDAD = DES: 1.0 (1.0/1.0) - [0.0,1.0,
|   |   |  EDAD = <=20: 0.0 (4.0/4.0) - [1.0,0.0,
|   |  EST = PRI: 0.0 (3.0/3.0) - [1.0,0.0,
|   |  EST = OTO: 1.0 (4.0/4.0) - [0.0,1.0,
|  APAV = N
|   |  ANCAL = ANC
|   |   |  DIA = L: 0.0 (5.0/5.0) - [1.0,0.0,
|   |   |  DIA = PF: 0.0 (0/0) - [0,0,
|   |   |  DIA = AF: : 0.0 (0/0) - [0,0,
|   |   |  DIA = F: 1.0 (1.0/1.0) - [0.0,1.0,
|   |  ANCAL = MED
|   |   |  HORA = (12-18]: 1.0 (2.0/2.0) - [0.0,1.0,
|   |   |  HORA = [0-6]: 0.0 (1.0/1.0) - [1.0,0.0,
|   |   |  HORA = (18-24]: 1.0 (1.0/1.0) - [0.0,1.0,
|   |   |  HORA = (6-12]: 1.0 (6.0/6.0) - [0.0,1.0,
|   |  ANCAL = EST
|   |   |  EST = VER: 0.0 (1.0/1.0) - [1.0,0.0,
|   |   |  EST = INV: 1.0 (1.0/1.0) - [0.0,1.0,
|   |   |  EST = PRI: 1.0 (1.0/1.0) - [0.0,1.0,
|   |   |  EST = OTO: : 0.0 (0/0) - [0,0,
TAC = OT
|  EST = VER: 0.0 (4.0/4.0) - [1.0,0.0,
|  EST = INV
|   |  HORA = (12-18]: 0.0 (0/0) - [0,0,
|   |  HORA = [0-6]: 1.0 (1.0/1.0) - [0.0,1.0,
|   |  HORA = (18-24]: 1.0 (2.0/2.0) - [0.0,1.0,
|   |  HORA = (6-12]: 0.0 (2.0/2.0) - [1.0,0.0,
|  EST = PRI
|   |  HORA = (12-18]: 0.0 (0/0) - [0,0,
|   |  HORA = [0-6]: 0.0 (2.0/2.0) - [1.0,0.0,
|   |  HORA = (18-24]: 1.0 (1.0/1.0) - [0.0,1.0,
|   |  HORA = (6-12]
|   |   |  VEH = OT: 0.0 (0/0) - [0,0,
|   |   |  VEH = MOT: 0.0 (1.0/1.0) - [1.0,0.0,
|   |   |  VEH = VL: 1.0 (1.0/1.0) - [0.0,1.0,
|   |   |  VEH = VP: 0.0 (0/0) - [0,0,
|  EST = OTO
|   |  VEH = OT: 0.0 (0/0) - [0,0,
|   |  VEH = MOT
|   |   |  EDAD = (27-60]: 0.0 (1.0/1.0) - [1.0,0.0,
|   |   |  EDAD = (20-27]: 1.0 (1.0/1.0) - [0.0,1.0,
|   |   |  EDAD = >60: 0.0 (0/0) - [0,0,
|   |   |  EDAD = DES: 0.0 (0/0) - [0,0,
|   |   |  EDAD = <=20: : 0.0 (0/0) - [0,0,
|   |  VEH = VL: 0.0 (5.0/5.0) - [1.0,0.0,
|   |  VEH = VP: 0.0 (0/0) - [0,0,
TAC = CO
|  DIA = L
|   |  HER = [1]: 0.0 (7.0/7.0) - [1.0,0.0,
|   |  HER = [>1]: 1.0 (1.0/1.0) - [0.0,1.0,
|  DIA = PF: 0.0 (0/0) - [0,0,
|  DIA = AF: 1.0 (1.0/1.0) - [0.0,1.0,
|  DIA = F: 0.0 (3.0/3.0) - [1.0,0.0,

Time taken to buINSd model: 0.05 seconds
```

```
=== Evaluation on training set ===
=== VERmarSI ===

CorrectlSI Classified Instances         890               70.6349 %
IncorrectlSI Classified Instances       370               29.3651 %
Kappa statistic                          0.4153
Mean absolute error                      0.3678
Root mean squared error                  0.4288
Relative absolute error                 73.6044 %
Root relative squared error             85.793  %
Total Number of Instances             1260

=== DetaINSed TACuracSI BSI Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure
ROC Area  Class
                 0.801     0.384      0.665     0.801     0.727
0.784    SI
                 0.616     0.199      0.765     0.616     0.683
0.784    KSI
Weighted Avg.    0.706     0.289      0.716     0.706     0.704
0.784

=== Confusion Matrix ===

   a   b   <-- classified as
 492 122 |   a = SI
 248 398 |   b = KSI
```

ANEXO II. ÁRBOL DE DECISIÓN 1 CREADO CON RGI.

```
=== Run information ===

Scheme:       weka.classifiers.trees.IPTree -LeafEstimation MLE -
AlphaUtilityMIP 0.0 -S 1.0 -StopLevel 4 -SM J48InfoGainRatio -KTH 1
Relation:     TR_pr2
Instances:    1260
Attributes:   20
              VEH
              BAR
              AGE
              SEX
              MON
              HORA
              DIA
              HER
              OCU
              ANCAL
              MVIAL
              ANCAR
              ANARC
              APAV
              LUM
              CAT
              VISIB
              ACC
              CAU
              SEV
Test mode:    evaluate on training data

=== Classifier model (full training set) ===

IPTree


SEX = H
|  TAC = SV
|  |  CAU = CON
|  |  |  VEH = OT: 0.0 (9.0/16.0) - [0.5625,0.4375,
|  |  |  VEH = MOT: 1.0 (98.0/144.0) -
[0.3194444444444444,0.6805555555555556,
|  |  |  VEH = VL: 1.0 (277.0/540.0) -
[0.48703703703703705,0.512962962962963,
|  |  |  VEH = VP: 1.0 (27.0/40.0) - [0.325,0.675,
|  |  CAU = COF
|  |  |  BAR = N: 1.0 (57.0/110.0) -
[0.4818181818181818,0.5181818181818182,
|  |  |  BAR = SI: 0.0 (1.0/1.0) - [1.0,0.0,
|  |  CAU = VEH
|  |  |  VIS = SR: 0.0 (3.0/3.0) - [1.0,0.0,
|  |  |  VIS = TOP: 1.0 (1.0/1.0) - [0.0,1.0,
|  |  |  VIS = ATM: 0.0 (0/0) - [0,0,
|  |  |  VIS = OT: 0.0 (0/0) - [0,0,
|  |  |  VIS = VEG: 0.0 (0/0) - [0,0,
|  |  |  VIS = EDI: 0.0 (0/0) - [0,0,
|  |  CAU = OT: 0.0 (9.0/9.0) - [1.0,0.0,
|  |  CAU = CAR
```

```
|   |   |  EST = VER: 0.0 (2.0/4.0) - [0.5,0.5,
|   |   |  EST = INV: 0.0 (2.0/2.0) - [1.0,0.0,
|   |   |  EST = PRI: 0.0 (0/0) - [0,0,
|   |   |  EST = OTO: 0.0 (0/0) - [0,0,
|   TAC = VUE
|   |   BAR = N
|   |   |  CAT = BT: 0.0 (41.0/68.0) -
[0.6029411764705882,0.39705882352941174,
|   |   |  CAT = LF: 0.0 (0/0) - [0,0,
|   |   |  CAT = O: 0.0 (3.0/3.0) - [1.0,0.0,
|   |   |  CAT = LL: 0.0 (3.0/3.0) - [1.0,0.0,
|   |   BAR = SI: 1.0 (3.0/3.0) - [0.0,1.0,
|   TAC = CP
|   |   CAT = BT
|   |   |  APAV = SI: 1.0 (45.0/57.0) -
[0.21052631578947367,0.7894736842105263,
|   |   |  APAV = NE: 0.0 (14.0/23.0) -
[0.6086956521739131,0.391304347826087,
|   |   |  APAV = N: 1.0 (10.0/15.0) -
[0.3333333333333333,0.6666666666666666,
|   |   CAT = LF: 0.0 (1.0/1.0) - [1.0,0.0,
|   |   CAT = O: 1.0 (2.0/2.0) - [0.0,1.0,
|   |   CAT = LL: 1.0 (1.0/1.0) - [0.0,1.0,
|   TAC = OT
|   |   ANCAR = MED
|   |   |  ANARC = EST: 0.0 (10.0/11.0) -
[0.9090909090909091,0.09090909090909091,
|   |   |  ANARC = NE: 0.0 (2.0/3.0) -
[0.6666666666666666,0.3333333333333333,
|   |   |  ANARC = MED: 1.0 (1.0/1.0) - [0.0,1.0,
|   |   ANCAR = EST: 1.0 (1.0/1.0) - [0.0,1.0,
|   |   ANCAR = ANC: 0.0 (1.0/1.0) - [1.0,0.0,
|   ACC = CO
|   |   APAV = SI: 0.0 (3.0/3.0) - [1.0,0.0,
|   |   APAV = NE
|   |   |  HORA = (12-18]: 1.0 (2.0/2.0) - [0.0,1.0,
|   |   |  HORA = [0-6]: 0.0 (0/0) - [0,0,
|   |   |  HORA = (18-24]: 0.0 (1.0/1.0) - [1.0,0.0,
|   |   |  HORA = (6-12]: 0.0 (0/0) - [0,0,
|   |   APAV = N: 0.0 (0/0) - [0,0,
SEX = M
|   VEH = OT: 0.0 (4.0/4.0) - [1.0,0.0,
|   VEH = MOT
|   |   HORA = (12-18]
|   |   |  OCU = [1]: 0.0 (5.0/8.0) - [0.625,0.375,
|   |   |  OCU = [>2]: 0.0 (0/0) - [0,0,
|   |   |  OCU = [2]: 1.0 (1.0/1.0) - [0.0,1.0,
|   |   HORA = [0-6]: 0.0 (0/0) - [0,0,
|   |   HORA = (18-24]: 1.0 (7.0/7.0) - [0.0,1.0,
|   |   HORA = (6-12]: 0.0 (0/0) - [0,0,
|   VEH = CAR
|   |   TAC = SV
|   |   |  CAU = CON: 0.0 (76.0/123.0) -
[0.6178861788617886,0.3821138211382114,
|   |   |  CAU = COF: 0.0 (20.0/28.0) -
[0.7142857142857143,0.2857142857142857,
|   |   |  CAU = VEH: 1.0 (3.0/4.0) - [0.25,0.75,
|   |   |  CAU = OT: 0.0 (4.0/4.0) - [1.0,0.0,
```

```
|   |   |  CAU = CAR: 0.0 (0/0) - [0,0,
|   |   TAC = VUE
|   |   |  EDAD = (27-60]: 1.0 (1.0/1.0) - [0.0,1.0,
|   |   |  EDAD = (20-27]: 0.0 (1.0/1.0) - [1.0,0.0,
|   |   |  EDAD = >60: 0.0 (0/0) - [0,0,
|   |   |  EDAD = DES: 0.0 (0/0) - [0,0,
|   |   |  EDAD = <=20: 0.0 (0/0) - [0,0,
|   |   TAC = CP
|   |   |  ANARC = EST: 1.0 (3.0/3.0) - [0.0,1.0,
|   |   |  ANARC = NE: 0.0 (3.0/4.0) - [0.75,0.25,
|   |   |  ANARC = MED: 10.0 (0/0) - [0,0,
|   |   TAC = OT: 0.0 (1.0/1.0) - [1.0,0.0,
|   |   TAC = CO: 0.0 (3.0/3.0) - [1.0,0.0,
|   VEH = VP: 0.0 (0/0) - [0,0,
SEX = DES: 0.0 (1.0/1.0) - [1.0,0.0,

Time taken to build model: 0.06 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         763               60.5556 %
Incorrectly Classified Instances       497               39.4444 %
Kappa statistic                          0.2013
Mean absolute error                      0.4455
Root mean squared error                  0.472
Relative absolute error                 89.1563 %
Root relative squared error             94.4226 %
Total Number of Instances             1260

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure
ROC Area  Class
                 0.36      0.161      0.68      0.36      0.471
0.663    SI
                 0.839     0.64       0.58      0.839     0.686
0.663    KSI
Weighted Avg.    0.606     0.407      0.629     0.606     0.581
0.663

=== Confusion Matrix ===

   a   b   <-- classified as
 221 393 |   a = SI
 104 542 |   b = KSI
```

ANEXO III. Publicaciones.

En este anexo se incluyen las principales publicaciones relacionadas con la presente tesis doctoral:

- De Oña, J., López, G., Abellán, J., 2013. Extracting decision rules from police accident reports through decision trees. Accident Analysis and Prevention, 50, 1151–1160.
- Abellán, J., López, G., De Oña, J., 2013. Analysis of traffic accident severity using Decision Rules via Decision Trees. Expert Systems with Applications. 40, 6047-6054.
- López, G., De Oña, J., Abellán, J., 2012. Using decision trees to extract decision rules from police reports on road accidents. PROCEDIA-SOCIAL AND BEHAVIORAL SCIENCES, 53, 106-114.

Accident Analysis and Prevention 50 (2013) 1151–1160

Contents lists available at SciVerse ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Extracting decision rules from police accident reports through decision trees

Juan de Oña [a,*], Griselda López [a], Joaquín Abellán [b]

[a] TRYSE Research Group. Department of Civil Engineering, University of Granada, ETSI Caminos, Canales y Puertos, c/ Severo Ochoa, s/n, 18071 Granada, Spain
[b] Department of Computer Science & Artificial Intelligence, ETSI Informática, c/Periodista Daniel Saucedo Aranda, s/n, 18071 Granada, Spain

ARTICLE INFO

Article history:
Received 6 March 2012
Received in revised form 17 August 2012
Accepted 5 September 2012

Keywords:
Traffic accident
Severity
Decision trees
CART
C4.5
Decision rules

ABSTRACT

Given the current number of road accidents, the aim of many road safety analysts is to identify the main factors that contribute to crash severity. To pinpoint those factors, this paper shows an application that applies some of the methods most commonly used to build decision trees (DTs), which have not been applied to the road safety field before. An analysis of accidents on rural highways in the province of Granada (Spain) between 2003 and 2009 (both inclusive) showed that the methods used to build DTs serve our purpose and may even be complementary. Applying these methods has enabled potentially useful decision rules to be extracted that could be used by road safety analysts. For instance, some of the rules may indicate that women, contrary to men, increase their risk of severity under bad lighting conditions. The rules could be used in road safety campaigns to mitigate specific problems. This would enable managers to implement priority actions based on a classification of accidents by types (depending on their severity). However, the primary importance of this proposal is that other databases not used here (i.e. other infrastructure, roads and countries) could be used to identify unconventional problems in a manner easy for road safety managers to understand, as decision rules.

1. Introduction

Traffic accidents are considered a major public health problem worldwide, claiming 1.27 million annual deaths and between 20 and 50 million injuries (WHO, 2009). Therefore, the aim of many studies to date has been to understand and identify the main factors that have an impact on road accident severity. Regression-type generalized linear models, Logit models and Probit models have been the techniques most commonly used to conduct such analyses (Kashani and Mohaymany, 2011; Savolainen et al., 2011; Mujalli and de Oña, in press). However most of them have their own model assumptions and pre-defined underlying relationships between dependent and independent variables (Chang and Wang, 2006).

Recently, data mining (DM) techniques have been used to study crash-injury severities by different researchers (Kuhnert et al., 2000; Sohn and Shin, 2001; Chang and Wang, 2006; Kashani and Mohaymany, 2011; Kashani et al., 2011; Pakgohar et al., 2010). The term decision trees (DTs) encompasses a series of techniques for extracting processable knowledge, implicit in databases, which is based on artificial intelligence and statistical analysis. One part of the DM could be defined as the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable pattern in data (Fayyad et al., 1996). These techniques are aimed at extracting knowledge from large amounts of data previously unknown and indistinguishable. DT techniques are particularly appropriate for studying crashes because they are non-parametric techniques that do not require prior probabilistic knowledge on the study phenomena. Furthermore, they consider conditional interactions among input data (Montella et al., in press). Other advantages of DTs compared to other methods with similar aims include the extraction of decision rules of the "if-then" type (Kashani et al., 2011), and that they can be used to discover behaviours that occur within a specified set of data. Moreover, conclusions on behaviour can be drawn from the structure of DTs to understand the events leading up to a crash and identify the variables that determine how serious an accident will be.

There are many algorithms that can be used to build DTs, but CART (Classification and Regression Trees) developed by Breiman et al. in 1984 is the once most commonly used to analyse crash severity. Authors such as Kuhnert et al. (2000) compared the results obtained with CART, multivariate adaptive regression splines (MARS) and logistic regression in the analysis of an epidemiological case–control study of injuries resulting from motor vehicle accidents. The findings indicated that non-parametric techniques such as CART and MARS can provide more informative and attractive models whose individual components can be displayed graphically. Chang and Wang (2006) studied the relationships between crash severity with characteristics related to drivers and vehicles, as well as variables related to roads, road accidents and the environment characteristics. Pakgohar et al. (2010) used CART and Multinomial

* Corresponding author. Tel.: +34 958 24 99 79.
E-mail address: jdona@ugr.es (J. de Oña).

0001-4575/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.aap.2012.09.006

Logistic Regression to study the role played by drivers' characteristics in the resulting crash severity. They found that the CART method provided more precise results, which were also simpler and easier to interpret. Kashani et al. (2011) studied the key factors that affect the injury severity of drivers involved in crashes on two-lane two-way rural roads. Subsequently, Kashani and Mohaymany (2011) used CART to identify the main factors that affect the injury severity of vehicle occupants involved in crashes on those roads.

However, CART always yields binary trees, which sometimes cannot be summarized as efficiently for interpretation and/or presentation (Breiman et al., 1984). In the case of road accidents, they may not be very practical when it comes to analysing the impact of a specific category of variable in crash severity. Liu (2009) mentions the existence of other popular algorithms for building DTs, such as C4.5. He does not apply it, however, because a binary DT is sufficient to develop his work.

Other simple algorithms, such as ID3 (Quinlan, 1986) and C4.5 (Quinlan, 1993), have been widely used in the literature of DM for building DTs, and do not involve the binary restriction. Therefore, this study proposes to make a comparison between the various methods, and to use CART and other methods to identify the main factors that affect crash severity. Then we extract certain decision or association rules (Agrawal et al., 1993) from the methods that give the best results. We show that the methods used complement our objective. Finally, our results could be used for the predictive purposes pursued by road safety analysts.

This paper is organized as follows: Section 2 gives an introduction to procedures for building DTs, focusing on the ones used in this study. It also describes the parameters used to assess the various methods, the procedure for extracting rules and the main features of the study data. Section 3 presents the results and a discussion on them. Finally, the last section presents the conclusions.

2. Materials and methods

2.1. Decision trees

A DT is a predictive model which can be used to represent both classifiers and regression models. DTs are popular due to their simplicity and transparency; moreover, they are usually presented graphically as hierarchical structures, which make them easy to interpret.

A DT is a simple structure that can be used as a classifier. Within a DT, each node represents an attribute variable[1] X and, each branch represents one of the states of this variable. Normally, a terminal node, or leaf, specifies the expected value of the class variable or variable in study C, depending on the information contained in the training data set, i.e. the set used to build the model. The set of data used to check the model is called test set. When we obtain a new instance or case of the test data set, we can make a decision or prediction about the state of the variable class following the path to the tree from the root node to a leaf node, using the sample values and the tree structure. Subsequently, the model obtained can be used to classify new examples (cases whose classes are not known a priori), to detect patterns, or simply to gain a better understanding of the phenomenon being analysed.

DTs are built recursively, following a descending strategy, starting with the full data set (made by the root node). Using specific split criteria, the full set of data is then split into even smaller subsets. Each subset is split recursively until all of them are pure (when the cases in each subset are all of the same class) or their "purity" cannot be increased. That is how the tree's terminal nodes are formed, which are obtained according to the answer values of the class variable.

The main difference between DTs building procedures lies in the splitting criteria. The most commonly applied splitting criteria in simple algorithms are the Gini Index (which measures the degree of purity), used in the CART system (Breiman et al., 1984); Information Gain, used in the ID3 algorithm (Quinlan, 1986); and the Information Gain Ratio, used in the C4.5 algorithm (Quinlan, 1993). ID3 and C4.5 are based on the entropy, which measures the degree of confusion (the greater the confusion, less information). The procedures also differ in the strategies they use after building a tree, in the process known as pruning. This is when the model obtained is simplified and adjusted more closely to the data set used to build it.

2.2. Methods for building decision trees

2.2.1. CART

Depending on the nature of the dependent variable, a classification tree (case discrete) or a regression tree (case continuous) will be built. The CART model generates binary trees by using impurity as a measure to split the Gini Index of diversity (which is a measure of the diversity of classes in a tree node being used). For a variable C, it is defined as:

$$gini(C) = 1 - \sum_j p^2(C = c_j). \quad (1)$$

In this way, we can define the split criterion based on the Gini Index as:

$$GIx(C, X) = gini(C|X) - gini(X), \quad (2)$$

where $gini(C|X) = \sum_t p(x_t) gini(C|X = x_t)$ and X another known variable.

Thus, the best split is the one that minimizes $GIx(C,X)$. With this procedure, the maximal tree that overfits the data is created. To decrease its complexity, the tree is pruned using a cost-complexity measure that combines the precision criteria as opposed to complexity in the number of nodes and processing speed, searching for the tree that obtains the lowest value for this parameter. A more detailed description of the CART method can be found in Breiman et al. (1984).

2.2.2. ID3

Builds a tree in a manner similar to the CART method but without the binary restriction. It can only be used with discrete variables, does not allow pruning and the function used to measure impurity is the Shannon's entropy (Shannon, 1948), which is an information-based uncertainty measure.

The ID3 algorithm uses the Information Gain criterion to choose which attribute goes into a decision node. Information Gain could be defined as a difference of entropies in the current node, considering the information that an attribute variable gives us about the class variable. This split criterion can therefore be defined on an attribute variable X, given the class variable C, as follow:

$$Information\ Gain\,(C, X) = IG(C, X) = H(C) - H(C|X) \quad (3)$$

where $H(C)$ is the entropy of C, $H(C) = \sum_j p(c_j) \log p(c_j)$, with $p(c_j) = p(C = c_j)$, the probability of each value of the variable class estimated in the training data set. In the same way, $H(C|X) = -\sum_t \sum_j p(c_j|x_t) \log p(c_j|x_t)$, where x_t, $t = 1, \ldots, |X|$, is each possible state of X and c_j, $j = 1, \ldots, k$ each possible state of C.

Notice that the Information Gain criterion has implicit preference for splitting nominal attributes with lots of values. Therefore, it produces trees that discard the remaining attributes prematurely because they soon come to branches that have only a few cases.

[1] Also called *feature* or *predictor variable*.

A more detailed description of the ID3 algorithm can be found in Quinlan (1986).

2.2.3. C4.5

In order to improve the ID3 algorithm, Quinlan (1993) introduces the C4.5 algorithm, where the Information Gain split criterion is replaced by an Information Gain Ratio criterion which penalizes variables with many states. Moreover, this model makes it possible to deal with continuous attributes and missing values, and to carry out a post-pruning process. The algorithm incorporates classification tree pruning once a tree has been induced, by applying a hypothesis test on whether or not to expand a branch.

The Information Gain Ratio of an attribute variable X on a variable class C can be expressed as:

$$IGR(C,X) = \frac{IG(C,X)}{H(X)} \quad (4)$$

2.3. Method assessment

Taking into consideration the indicators used to evaluate the goodness of a classification method in De Oña et al. (2011) and Mujalli and de Oña (2011), and that the variable class used shows 2 possible response categories (state A and state B), the parameters that can be defined are described below:

- *Accuracy* – the method's precision, defined as the percentage of cases correctly classified by the classifier.
- *Sensitivity* – the proportion of cases correctly classified as state A among all the observed as state A.
- *Specificity* – the proportion of cases correctly classified as state B among all the observed as state B.
- *Receiver operating characteristic curve* (*ROC*) *area* – this indicator represents the curve of positive cases correctly classified (sensitivity), as opposed to the cases of false positives (1-specificity), in such a way that a value 1 describes a perfect adjustment.

If the variable class is accident severity and its potential states are accidents with slightly injured (SI) (state A) and accidents with killed or seriously injured (KSI) (state B), the equations that define these indicators are:

$$\text{Accuracy} = \frac{\text{TSI} + \text{TKSI}}{\text{TSI} + \text{TKSI} + \text{FSI} + \text{FKSI}} 100\% \quad (5)$$

$$\text{Sensitivity} = \frac{\text{TSI}}{\text{TSI} + \text{FKSI}} 100\% \quad (6)$$

$$\text{Specificity} = \frac{\text{TKSI}}{\text{TKSI} + \text{FSI}} 100\% \quad (7)$$

where, TSI – number of cases of SI; TKSI – number of cases of KSI; FSI – number of false cases of SI (i.e. incorrectly classified as SI); FKSI – number of false cases of KSI (i.e. incorrectly classified as KSI).

The software used to build the DTs was Weka (Witten and Frank, 2005), which is an open source freeware, available at: http://www.cs.waikato.ac.nz/ml/weka/.

Moreover, in order to obtain a more reliable result for each method (CART, ID3 and C4.5) in classification, a repeated Cross Validation procedure (CV) was used. In our case, we use a 10×10-fold CV. In general, a k-fold CV uses the whole data set, and randomly divides the sample used in the training phase into k sets: Sequentially, each subset is kept to be used as a testing set against the tree model generated by the remaining $k-1$ subsets. Thus, different k models are obtained, in which the accuracy of the classifications in the training set $(k-1)$ and in the testing subsets (k) can be evaluated and the optimal tree can be selected.

Finally, a corrected paired t-test implemented in Weka, which is a corrected version of the standard paired t-test, was used to compare the results of the trees generated with the different algorithms. This test checks whether a method is better or worse than another, on average, in all the training and testing data sets based on an initial data set. In our case, we used the classification results from the 100 test set for this test, i.e. the sets obtained from a 10×10-fold CV procedure. The level of significance used for this paired t-test was 0.1.

It should be pointed out that the ID3 algorithm implemented in Weka allows instances without classification. To compare results, we implemented a similar procedure but classified all the instances of the test set as in Abellán and Masegosa (2010). In the case of no classification, we took into account the decision in the parent node. For sake of simplicity, we call this procedure ID3 too.

2.4. Rules extraction and validation

The DT's structure was transformed into rules in order to extract its potentially useful information. The rules make a logic conditional structure of the type "$X \rightarrow Y$", where in our case, X is a set of statuses of several attribute variables; and Y is only one state of the class variable:

IF (a set of statuses of several attribute variables) – THEN (status of the class variable).

For example:

IF (**accident type** = *rollover* & **atmospheric condition** = *light rain*) THEN (**severity** = *slightly injured accident*).

The part X of the rule is called the antecedent and the part Y is called the consequent.

In a DT, rules are configured from the root node, which is where the conditioned structure (IF) begins. Each variable that intervenes in tree division makes an IF of the rule, which ends in child nodes with a value of THEN, which is associated with the state resulting from the child node. The resulting state is the status of the class variable that shows the highest number of cases in the child node analysed.

A priori, as same number of rules can be identified as the number of terminal nodes on the tree. However, 3 parameters were used on each possible rule "$X \rightarrow Y$", in order to extract significant rules that could provide useful information for the implementation of road safety strategies in the future.

It is known as support of X, as the percentage of the data set where X appears. In the same vein, we can talk about the support of the entire rule, as the percentage of the data set where X & Y appear. For each rule, the 3 parameters that we use are the following: *support* (S), which will be the support of the entire rule; *population* (Po), which is the support of the antecedent of the rule; and *probability* (P), which is the percentage of cases in which the rule is accurate (i.e. P = S/Po expressed as percentage).

The concepts of support (S) and probability (P) are central to association rules and have been used by several authors (Agrawal et al., 1993; Pande and Abdel-Aty, 2009; Montella et al., in press). Population (Po) is deduced from S and P (Po = S/P). Support is a measure of how frequently any given combination of antecedent and consequent occurs in a database. Probability[2] is defined by the percentage of cases in which a consequent appears, given that the antecedent has occurred. It essentially measures the strength of an association rule. For further clarification of these parameters see Pande and Abdel-Aty (2009).

Association rule discovery is the process of finding strong associations with a minimum support and probability. It is desirable

[2] Pande and Abdel-Aty (2009) and Montella et al. (in press) call this parameter confidence.

Table 1
Variables used from the police accident reports.

Num	Variables			%Total	Severity	
	Description	Code	Values		%SI	%KSI
1	ACT: accident type	CO	Fixed objects collision	0.90	76.47	23.53
		CP	Collision with pedestrian	7.70	33.33	66.67
		OT	Other (collision with animals, etc.)	1.90	68.57	31.43
		RO	Rollover (in carriage without any collision)	6.60	61.86	38.14
		ROR	Run off road (with or without collision)	82.90	51.77	48.23
2	AGE: age	≤20	≤20	12.22	52.73	47.27
		[21–27]	[21–27]	25.65	50.00	50.00
		[28–60]	[28–60]	53.64	51.76	48.24
		≥ 61	≥61	6.89	59.68	40.32
		UN	Unknown	1.61	27.59	72.41
3	ATF: atmospheric factors	GW	Good weather	86.40	50.58	49.42
		HR	Heavy rain	2.10	63.16	36.84
		LR	Light rain	8.90	58.75	41.25
		O	Other	2.60	51.06	48.94
4	BAR: safety barriers	N	No	96.90	48.30	54.70
		Y	Yes	3.10	53.60	46.40
5	CAU: cause	DC	Driver characteristics	82.70	48.99	51.01
		CO	Combination of factors	13.40	61.16	38.84
		OT	Other	1.20	72.73	27.27
		RC	Road characteristics	1.40	84.00	16.00
		VC	Vehicle characteristics	1.20	63.64	36.36
6	DAY: day	APH	Working day after the weekend or public holiday (Monday or day after public holiday)	8.40	57.62	42.38
		BPH	Working day before the weekend or public holiday (Friday or day before public holiday)	15.90	52.26	47.74
		PH	On a weekend (Saturday or Sunday) or public holiday	30.60	50.36	49.64
		WD	Regular working day (Tuesday, Wednesday or Thursday nor before neither after public holiday)	45.00	51.05	48.95
7	LAW: lane width	THI	<3.25 m	27.50	46.87	53.13
		MED	[3.25–3.75] m	70.20	53.20	46.80
		WID	>3.75 m	2.30	58.54	41.46
8	LIG: lighting	DAY	Daylight	53.10	55.49	44.51
		DU	Dusk	5.80	54.29	45.71
		IL	Insufficient (night-time)	7.30	51.15	48.85
		SL	Sufficient (night-time)	40.00	59.72	48.28
		WL	Without lighting (night-time)	29.80	43.10	56.90
9	MON: month	AUT	Autumn	23.50	53.07	46.93
		SPR	Spring	25.20	53.64	46.36
		SUM	Summer	27.30	51.63	48.37
		WIN	Winter	24.00	47.92	52.08
10	NOI: number of injuries	[1]	1 injury	69.60	53.43	46.57
		[>1]	>1 injury	30.40	47.35	52.65
11	OI: occupants involved	[1]	1 occupant	64.70	51.20	48.80
		[2]	2 occupants	22.50	51.48	48.52
		[>2]	>2 occupants	12.70	53.71	46.29
12	PAS: paved shoulder	N	No	17.10	49.35	50.65
		NE	Non-existent or impassable	31.30	50.89	49.11
		Y	Yes	51.60	52.74	47.26
13	PAW: pavement width	MED	[6–7] m	30.50	53.19	46.81
		THI	<6 m	14.40	45.56	54.44
		WID	>7 m	55.10	52.27	47.73
14	ROM: pavement markings	DME	Does not exist or was deleted	9.40	52.35	47.65
		DMR	Separate margins of roadway	9.90	48.31	51.69
		SLD	Separate lanes and define road margins	75.80	52.23	47.77
		SLO	Separate lanes only	4.90	46.59	53.41
15	SEX: gender	F	Female	15.30	62.18	37.82
		M	Male	84.50	49.61	50.39
		UN	Unknown	0.20	75.00	25.00
16	SHT: shoulder type	THI	<1.5 m	40.40	52.54	47.46
		MED	[1.5–2.5] m	10.50	50.28	49.72
		NE	Non-existent or impassable	49.10	50.57	49.43
17	SID: sight distance	ATM	Atmospheric	2.20	67.50	32.50
		BU	Building	0.60	36.36	63.64
		OT	Other	0.70	50.00	50.00
		TOP	Topography	22.70	49.39	50.61
		VEG	Vegetation	0.70	50.00	50.00
		WR	Without restriction	73.10	51.94	48.06
18	TIM: time	[0–6)	[00:00–05:59]	20.00	48.06	51.94
		[6–12)	[06:00–11:59]	21.00	58.73	41.27
		[12–18)	[12:00–17:59]	32.10	52.77	47.23
		[18–24)	[18:00–23:59]	26.90	47.22	52.78
19	VEH: vehicle type	CAR	Cars	70.90	47.10	52.90
		TRU	Trucks	4.90	53.80	46.20
		MOT	Mortorbikes and motorcycles	21.70	35.60	64.40
		OT	Other	2.50	50.60	49.40

for the rules to have a large probability factor and a high level of support. However, since some events of interest in traffic safety analysis are very rare (e.g., "crashes with fatal injury"), the support for some rules of interest could be quite low. The threshold values for these parameters depend on the nature of the data (balanced or unbalanced), significant interest in fatal crashes (rare events) and sample size (small or large databases). Pande and Abdel-Aty (2009) set 0.90% and 10% as threshold values for support and probability respectively. It means that no rules with support <0.90% and/or probability <10% would be considered. Montella et al. (in press) used lower thresholds for their analysis (0.10% and 1.00% for support and probability respectively). In this paper, as the sample size is not very large and the sample is balanced, the threshold values used are 0.60% for support and 60% for probability. With these thresholds the minimum population (Po) will be 1%. It is worth highlighting that if other lower threshold values were established, more rules could be obtained.

In order to test that spurious rules, and due to the large number of patterns considered, DTs could suffer from an extreme risk of type-1 error, that is, of finding patterns that appear due to chance alone to satisfy constraints on the sample data (Webb, 2007). To reduce this error and following other authors (Montella et al., in press; Kashani and Mohaymany, 2011), the dataset was split randomly in two parts: a training set (70% of the data) and a testing set (remaining 30%).

The training set was used to build a DT and obtain the significant rules that satisfied the three parameters defined (S, Po and P). Next, the rules were validated in the testing set to prevent spurious rules (checking that they still met minimum values S, Po and P).

We also used a binomial test to check if the rule support measure deviates significantly (at 0.05 level) from the theoretically expected value (values from the training set) when the antecedent and the consequent items are independent.

2.5. Importance of the variables

The importance of the variables that intervene in the model is defined for a variable X with possible states $\{x_1 \ldots x\}$ by the following equation:

$$VIM\,X = \sum_{i=1}^{h} \frac{nx_i}{n}(I(C|X = x_i) - (C)) \quad (8)$$

where C is class variable (severity), nx_i the number of cases that $X = x_i$, n the number of total cases. I is Gini Index in CART, Information Gain in ID3 and Information Gain Ratio in C4.5

2.6. Description of the data

Accident data were obtained from the Spanish General Traffic Accident Directorate (DGT) for two-lane rural highways in the province of Granada (South of Spain) over a period of 7 years (2003–2009). In this study, rural highways with only two lanes (one for each direction) were used. The horizontal curves radius of these roads ranged from 16 m to 2824 m. And the AADT ranged from 210 to 8681 veh/day. The accidents analysed involved 1 vehicle and they did not occur on intersections. The total number of 1801 accidents met these conditions

In the period of study, the severity distribution for two-lane rural highways was: 6.1% fatal, 35.6% severe injury and 58.3% slight injury. For the same period, the severity distribution of all accidents (including accidents in freeways, multilane highways, two-lane highways, intersections, etc.) in the province of Granada was: 8.3% fatal, 40.1% severe injury and 51.6% slight injury. This study uses the DGT definition for injuries: severe injury is any person injured in a traffic accident and whose condition requires hospitalization for more than 24 h; slight injury is any person that does not meet the severe injury definition; and fatal injury is any person that dies on the spot or within the subsequent 30 days as a result of a traffic accident.

Table 2
Comparison of the parameters produced by the various algorithms.

	CART	C4.5	ID3
Accuracy	55.87	54.16	52.72[a]
Sensitivity	54.00	55.00	53.00
Specificity	58.00	54.00	52.00
ROC area	57.00	54.00	53.00[a]

[a] The results worsen significantly.

Following previous studies (Chang and Wang, 2006; De Oña et al., 2011; Kashani and Mohaymany, 2011), severity of accident was defined according to the worst injured occupant, and two level of severity were identified: accident with slightly injured (SI) and accidents with killed or seriously injured (KSI).

To identify the main factors that affect accident severity, 19 variables were analysed (see Table 1). The variables chosen were based on:

- Variables available in the original dataset (from DGT).
- Variables selected in others studies with similar objectives (Chang and Wang, 2006; De Oña et al., 2011; Kashani and Mohaymany, 2011; Pakgohar et al., 2010).

The variables describe characteristics related to the driver (age and gender); accident (month, time, day, number of injuries, occupants involved, accident type and cause); road (safety barriers, pavement width, lane width, shoulder type, paved shoulder, pavements markings and sight distance[3]); vehicle (vehicle type); and context (atmospherics factors and lighting). Some variables were re-coded in a reduced number of categories to be able to work with them. For instance, in the original dataset MON had 12 categories (12 months), and it was recoded into four periods (see Table 1). Other variables, such as CAU, DAY, LAW, LIG, PAS, PAW, ROM, SEX, SHT, SID, were used as they were in the original dataset. Table 1 gives a description of the variables used for the analysis, together with the frequency distribution.

3. Results and discussion

The first step was to build DTs using the three algorithms (CART, C4.5 and ID3) with the aim of classification using 10 × 10-fold CV procedure. In order to compare the results, corrected paired t-tests were conducted. The results of the tests, comparing the methods to each other on the indicators *accuracy*, *sensitivity*, *specificity* and *ROC_area* are shown in Table 2.

C4.5 and CART show similar values for accuracy. ID3 shows significantly worse values than the other two algorithms. The accuracy values are within the range of values obtained in other studies in which classification methods with similar objectives were applied: Abdel Wahab and Abdel-Aty (2001) obtained 61% accuracy when they applied Bayesian networks and 58.1% accuracy on neural networks. De Oña et al. (2011) obtained 58%, 59% and 61% accuracy applying Bayesian networks with different algorithms (AIC, MDL and BDeu, respectively).

[3] The sight distance refers only to the horizontal visibility limitation at the site of the accident (i.e. the 'without restrictions' category means that there were no visibility limitations at the point of the accident; the 'building' category means that the visibility limitation at the point of the accident was a building; the same applies for topography, vegetation, atmospheric factors and others.

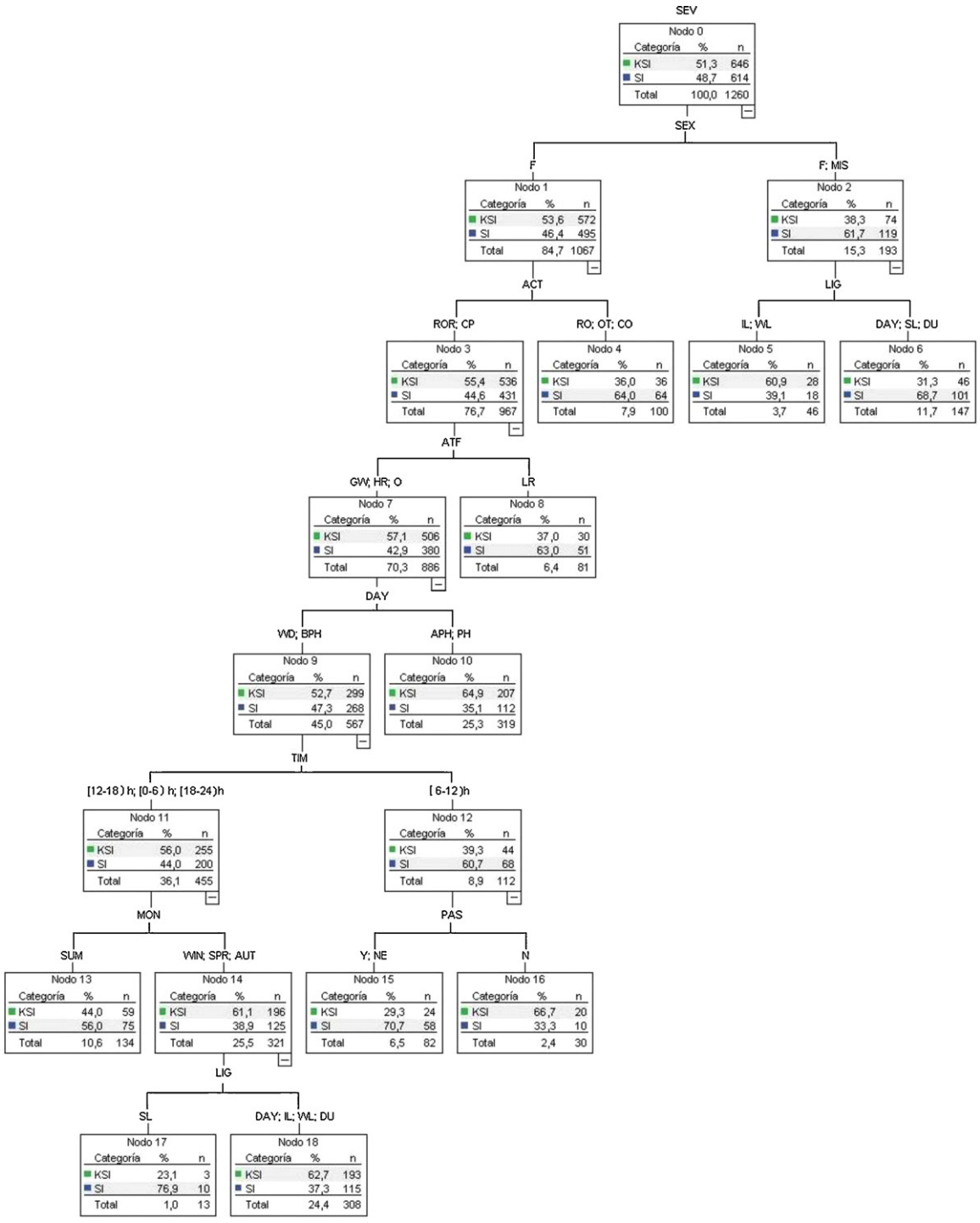


Fig. 1. Decision tree built with CART.

The C4.5 algorithm gives a higher value than CART (55% vs. 54%) in the sensitivity parameter analysis. The improvement is not significant, however. CART gives a higher value than C4.5 for the specificity parameter, although the improvement is not significant either. For ID3, both sensitivity and specificity are poorer, in comparison to the values of the other two algorithms. A global measure given by the ROC area indicator shows that CART gives the best results (57%) whereas ID3 obtains the lowest values again (53%).

The computational time it took each algorithm to build the DT was another indicator analysed. It was obtained that the CART method requires the most time to build a tree, being 55 times slower than for the C4.5 algorithm and 42 times slower than ID3. C4.5 is the algorithm that takes the less time, needing only 0.03 s to build a DT with 19 variables and 1801 data. This result is logical because the CART algorithm is more complex, and in turn, C4.5 is more complex than ID3, since it has more optimization parameters in order to improve the results. The implementation of the C4.5 algorithm is optimized in Weka, and therefore the computational time is lower than for ID3.

Taking the above results in consideration, it can be seen that the ID3 algorithm is the method that gives the worst results. The difference in improvement using CART and C4.5 is not significant,

Table 3
Description of the rules according to the CART.

Node/rule	Rules CART: IF,...	THEN	S (%)	Po (%)	P (%)
16	IF (SEX = M) AND (ACT = ROR OR ACT = CP) AND (ATF ≠ LR) AND (DAY = WD OR DAY = BPH) AND (TIM = [6–12)) AND (PAS = N)	KSI	1.59	2.38	66.67
5	IF (SEX ≠ M) AND (LIG = IL OR LIG = WL)	KSI	2.22	3.65	60.87
15	IF (SEX = M) AND (ACT = ROR OR (ACT = CP) AND (ATF ≠ LR) AND (DAY = WD OR DAY = BPH) AND (TIM = [6–12)) AND (PAS ≠ N)	SI	4.60	6.51	70.73
6	IF (SEX ≠ M) AND (LIG ≠ IL OR LIG ≠ WL)	SI	8.02	11.67	68.71
4	IF (SEX = M) AND (ACT = RO OR ACT = CO OR ACT = OT)	SI	5.08	7.94	64.00
8	IF (SEX = M) AND (ACT = ROR OR ACT = CP) AND (ATF = LR)	SI	4.05	6.43	62.96

however. Although CART obtains slightly higher values in the precision and specificity parameters analysed, the improvement is not significant, and therefore, we cannot assert a priori that one method is better than the other. It would be worthwhile to analyse the decision rules obtained with the algorithms that attained the best results: C4.5 and CART.

3.1. CART

Fig. 1 shows the DT built using the CART method with 70% of the data for training and the remaining data (30%) for testing, as used by Montella et al. (in press). The CART method creates a tree with 19 nodes and 10 terminal nodes.

Table 3 shows a description of the six rules identified in the DT that verify the minimum values of the parameters S, Po and P in the training and in the test sets. Support varies from 1.6% (rule 16) to 8.0% (rule 6). All the rules include at least 1% of the population, and probability values are higher than 60.9%, with 70.7% being the highest value (rule 15).

With regards to the binomial test that was performed, all the rules obtained from the training set with the minimum threshold have a grade of lift (see Montella et al., in press) different than 1. Hence the antecedent and consequent are independent. These results were not included in the paper because they are not important for our aims. The binomial test showed that all the rules given in Table 3 have no significant differences (at 0.05 level), based on support when they are applied on the test set. Only the rule 5 (see Table 3) has a high level of support in the test set compared to the support in the training set. This difference is significant at 0.05 level of significance.

The root variable that generates the tree is SEX (see Fig. 1) which splits into two branches (nodes 1 and 2). For female drivers, and depending on LIG, nodes 5 and 6 are obtained, with different degrees of severity (see Fig. 1): accidents are KSI if LIG is insufficient or without lighting, with a probability of 61% (rule 5); while if LIG is sufficient, dusk or day light the severity is SI, with a probability of 69% (rule 6). This result shows a direct relationship between KSI accidents and female drivers on rural highways with insufficient or without lighting.

The rest of the rules are attributable to male drivers (node 1). This result is coherent with the study data, given that in 84.5% of the accidents analysed the drivers were men (see Table 1). After this node, the tree splits according to ACT. The accident type has been identified in several previous studies (Al-Ghamdi, 2002; De Oña et al., 2011; Kashani and Mohaymany, 2011) as one of the key variables in analyses of accident severity. This study shows that if the accident type is rollover, collision with obstacles or other accidents types the probability of SI is 64.0% (rule 4 in Table 3). However, in the case of run off road or collision with pedestrian the probability of KSI is higher than the probability of SI (node 3 in Fig. 1). So, in this kind of facilities road safety managers should pay attention to this type of accidents (run off road and collision with pedestrian).

Node 3 (Fig. 1) splits by the variable ATF: if ATF is light rain the accident is SI, with a probability of 63% (rule 8 in Table 3). This result proves that drivers try to be very careful under bad atmospheric conditions. In other cases, the tree continues to grow according to DAY. If DAY is on a weekend or public holiday (PH) or a working day after the weekend or public holiday (APH) the accident is KSI, with a probability of 65% (node 10 in Fig. 1). This result is coherent with the trend observed in Spain, where most of fatalities in road accidents occur on weekends (31.4% of the car accidents in 2009 occurred at the weekends, in which 818 deaths were recorded, that is 38.4% of the total number of fatalities in the year 2009).

When DAY is a working day before the weekend or public holiday (BPH) or a regular working day (WD) the tree is divided according to TIM. From this point of DT's structure, the rule interpretation is difficult because many variables are involved in the accident. However, the following results are highlighted: from [6–12) h, accidents with SI are obtained when PAS is paved or non-existent (rule 15, which is the one that represents the highest probability: almost 71%) whereas when it is not paved the severity is KSI (rule 16); and from [12–18) h, tree is divided by MON and LIG (see Fig. 1), however neither of the obtained nodes are rules because they do not meet the threshold limits for S, Po or P.

Following Eq. (8), it is possible to obtain the importance of the variables in the model. Table 4 shows the normalized importance of these variables. 12 variables were detected as having the greatest influence on accident severity, with percent which varying from 100% to 9.9%.

LIG is the most important variable, coinciding with previous studies. Gray et al. (2008) identified that more severe injuries are predicted during darkness. Abel-Aty (2003) and Helai et al. (2008) found the same results. Pande and Abdel-Aty (2009) concluded that there is a significant correlation between lack of illumination and high severity of crashes. De Oña et al. (2011) also pointed that KSI accidents are associated with roadways without lighting.

ATF is the second variable with 83.6% importance in the model. This result matches with other previous studies, such as Xie et al. (2009) and Mujalli and de Oña (2011). TIM has 77.1% importance in the model which is coherent because there is already a degree of relationship between the time and lighting variables. Next comes

Table 4
Importance of the variables with CART.

Variables	Importance normalized
LIG	100%
ATF	83.6%
TIM	77.1%
ACT	76.0%
SEX	72.0%
PAS	55.9%
DAY	54.9%
MON	49.9%
CAU	32.8%
AGE	30.6%
SID	28.4%
LAW	9.9%

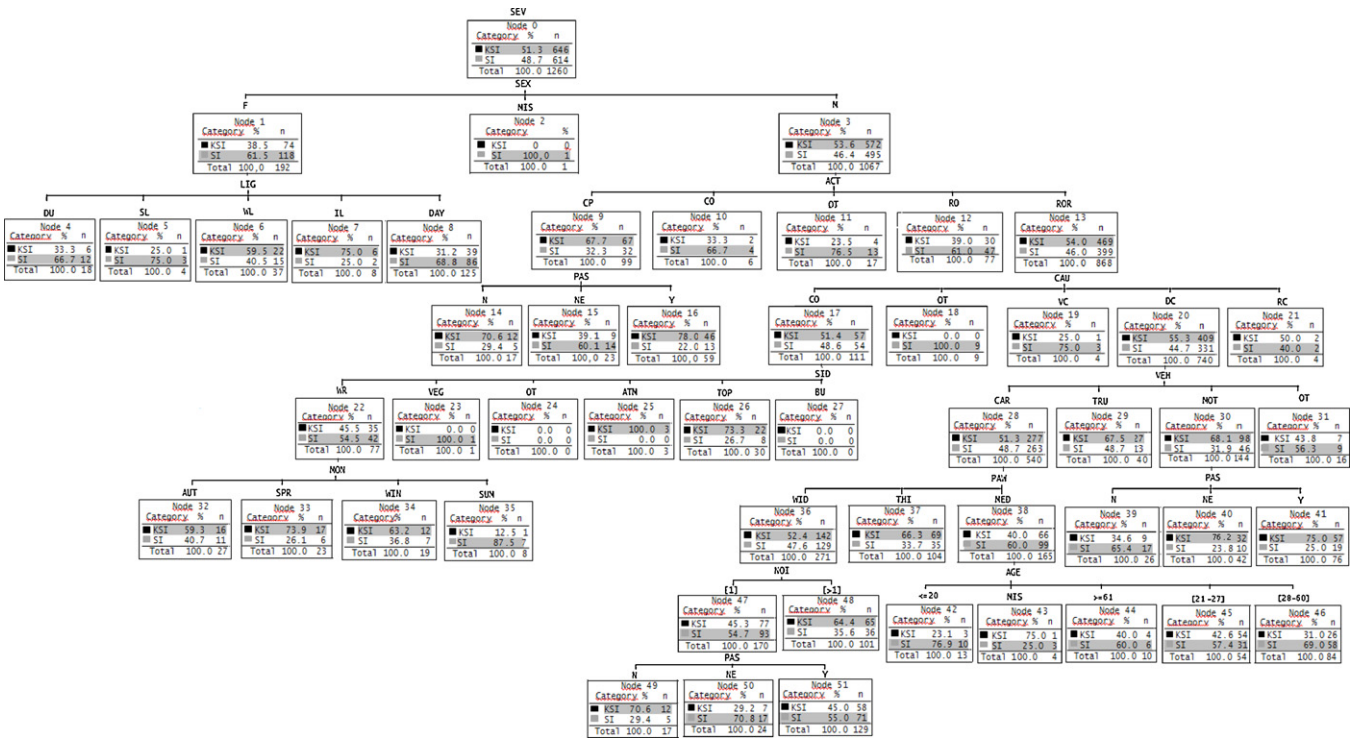


Fig. 2. Decision tree built with C4.5.

ACT with 76.0%. Kockelman and Kweon (2002) and De Oña et al. (2011) also found this variable as one of the most important in the study of severity. SEX represented 72% of the variables' importance. The other variables (see Table 4) in the model are less important, with percentages between 55.9% and 9.9%.

3.2. C4.5

Fig. 2 represents a DT built using the C4.5 algorithm based on the training set. It shows 52 nodes, with 39 terminal nodes. The increase in the number of nodes is justified be the fact that this algorithm creates a branch for each category of variable used in the analysis. In this case, however, only 9 rules that meet the minimal values for S, Po and P were obtained (see Table 5).

Since the tree generated with C4.5 is larger, only the rules extracted in Table 5 are used to describe the following tree structure. In this case, the rules in Table 5 also verify the threshold values for S, Po and P in both training and testing sets. For C4.5, the binomial test showed that all the rules obtained from the training set with the minimum threshold have a grade of lift (see Montella et al., in press) different than 1. And none of the rules given in Table 5 have significant differences (at 0.05 level) on support when they are applied to the test set.

As in CART, the root variable is the variable SEX. For female drivers when LIG is daylight, the rule with the highest population (9.9%) and support (6.8%) gives a severity result of SI (rule 8 in Table 5). This result agrees with the previous CART's results: female drivers seem to be highly affected by lighting conditions.

Most of the tree is generated by male drivers (see Fig. 2) and according to ACT, the same as CART. Fig. 2 and Table 5 show the following patterns: if ACT is rollover the severity is SI, with a probability of 61% (rule 12); whereas, if ACT is collision with pedestrian, it depends on PAS. This result is very important because if PAS is paved the severity is KSI and we obtained the rule with the highest probability (78%) (rule 16). Thus, from the perspective of road safety, precautions against accidents could be taken by placing safety barriers on stretches of road where pedestrians walk on the shoulder (roads that link two towns that are close to each other).

The rest of the rules are obtained for run off road accidents (ROR represents of 82.9% of the accident analysed) and depending on CAU. When CAU is a combination of factors, SID is without restriction and MON is spring the severity of accident is SI with almost 74% of the probability (rule 33). For CAU attributable to driver and depending on VEH the following patterns are shown: when VEH is a truck the accident is KSI (67.5%), rule 29; when is a motorbike or motorcycle and PAS is non-existent or impassable, the same severity (KSI) is obtained (rule 40); and for car two more rules are obtained depending on PAW. This result indicates the need to raise male drivers' awareness of vehicles of this type.

When PAW is between [6–7] m and driver's age is 28–60 (rule 46) the severity is SI with a probability of almost 69%. When PAW is >7 m, the tree splits according to NOI, and when it is higher than 1, accidents are SI in 64.4% of cases (rule 48); but when NOI is 1 and PAS is non-existent or impassable (rule 50) accidents are also SI in 70.8% of cases. These last three rules (rules 46, 48 and 50) are less useful to policy makers because they imply a combination of many more variables than in the preceding rules (6, 6 and 7 variables respectively), which makes it difficult to interpret the results and impossible to take direct preventive measures. That is why Pande and Abdel-Aty (2009) restricted the number of variables in the antecedent to three.

Following Eq. (8), it is possible to obtain the importance of the variables in the C4.5 model (see Table 6).

Fourteen variables were detected as having the greatest influence on accident severity, with percent which varying from 100% to 11.2%. ACT is the most important variable in the C4.5 model, followed by CAU. These results are in accordance with Al-Ghamdi (2002) and Kashani and Mohaymany (2011), who situate crash cause among the top variables influencing severity. The CART algorithm identified eleven of the previous fourteen variables. Moreover C4.5 identified VEH, PAW, and NOI.

Table 5
Description of the rules according to the C4.5.

Node/rule	Rules C4.5: IF,...	THEN	S (%)	Po (%)	P (%)
16	IF (SEX = M) AND (ACT = CP) AND (PAS = Y)	KSI	3.65	4.68	77.97
40	IF (SEX = M) AND (ACT = ROR) AND (CAU = DC) AND (VEH = MOT) AND (PAS = NE)	KSI	2.54	3.33	76.19
29	IF (SEX = M) AND (ACT = ROR) AND (CAU = DC) AND (VEH = TRU)	KSI	2.14	3.17	67.50
48	IF (SEX = M) AND (ACT = ROR) AND (CAU = DC) AND (VEH = CAR) AND (PAW = WID) AND (NOI = [>1])	KSI	5.16	8.02	64.36
33	IF (SEX = M) AND (ACT = ROR) AND (CAU = CO) AND (SID = WR) AND (MON = SPR)	SI	1.35	1.83	73.91
50	IF (SEX = M) AND (ACT = ROR) AND (CAU = DC) AND (VEH = CAR) AND (PAW = WID) AND (NOI = [1]) AND (PAS = NE)	SI	1.35	1.90	70.83
46	IF (SEX = M) AND (ACT = ROR) AND (CAU = DC) AND (VEH = CAR) AND (PAW = MED) AND (AGE = [28–60])	SI	4.60	6.67	69.05
8	IF (SEX = F) AND (LIG = DAY)	SI	6.83	9.92	68.80
12	IF (SEX = M) AND (ACT = RO)	SI	3.73	6.11	61.04

Table 6
Importance of the variables with C4.5.

Variables	Importance normalized
ACT	100.0%
CAU	80.4%
SEX	69.1%
LIG	67.5%
VEH	65.7%
ATF	59.8%
PAW	42.8%
AGE	41.2%
TIM	39.7%
SID	36.3%
NOI	32.1%
DAY	25.7%
LAW	20.2%
MON	11.2%

4. Conclusions

DTs allow accident classification based on crash severity. They provide an alternative to parametric models due to their ability to identify patterns based on data, without the need to establish a functional relationship between variables. Moreover, such classification models can be used to determine interactions between variables that would be impossible to establish directly, using ordinary statistical modelling techniques.

The main conclusions regarding the methods used in this paper to build DTs are the following:

- CART builds binary DTs and therefore certain categories of splitting variables are grouped in some branches, increasing node support, but making it impossible to analyse the influence of a specific category on severity. C4.5 creates a branch for each category, thereby permitting an analysis of the influence of all the categories of variables used to build the DT. Consequently, it could be said that the rules obtained with CART are less informative.
- C4.5 generates DTs with more branches than CART, and therefore it produces more rules. However, not all the rules meet the established minimal number of support, population and probability parameters, and therefore the rules may not be very useful for implementing future road safety strategies.
- The importance of the variables in the model can be obtained using either algorithm.
- The two algorithms have certain similarities with regards to the structure of the tree generated. For example, the root variable for both is SEX and tree density is obtained by the branch male drivers, and the value that continues to split the tree is ACT.

DTs permit certain potentially useful rules to be determined that can be used by road safety analysts and managers. Initially, they should focus on severe crashes and subsequently intervene in minor accidents. The approach proposed in this paper within each group will enable actions to be prioritized on the basis of support, population and probability. It is worth highlighting certain overall conclusions from a road safety perspective.

The rules drawn from the two methods are coincidental in that:

- Male drivers are the main causes of KSI crashes.
- The probability of KSI increases if pedestrians are involved (node 3 Fig. 1 and rule 16 in Table 5).
- When women drivers are involved in an accident, both methods predict SI when lighting exists (full daylight, sufficient lighting and dusk) (rule 6 in Table 3; rule 8 in Table 5, and nodes 4 and 5 in Fig. 2). However, both methods predict KSI when the lighting is non-existent or insufficient (rule 5 in Table 3 and nodes 6 and 7 in Fig. 2). These rules are not observed for men and may indicate that women increase their risk of severity under conditions of less lighting on the road.

From a road safety point of view, most of the rules extracted coincide with the conventional problems found on rural highways in developed countries, as most previous studies point out. This validates the method proposed in this paper, and therefore it is positive. However, the primary importance of this proposal is that other data bases not used here (i.e. other infrastructure, roads and countries) could be used to identify unconventional problems in a manner easy for road safety managers to understand, as decision rules.

However, using these two types of DTs permitted the identification of a specific problem worthy of further study: Although less women than men are involved in accidents (15.3% vs. 84.5%, see Table 1), and accident severity is SI in 62.2% of cases, the two methods indicate that women increase their risk of severity under conditions of non-existent or insufficient lighting. The efforts of multidisciplinary teams with experts on psychology, physiology, road safety and illumination should focus on a search of the reason why women, contrary to men, present higher risk of severity under conditions of less lighting on the road.

Finally, it should be stressed that each method has advantages and drawbacks, and reveals different information. Therefore, the two methods complement each other and the recommendation is to use both of them for a full analysis.

5. Future work

When we use a DT to obtain decision rules, such rules are highly dependent on the variable entered in the root node, which permits

knowledge to be extracted only in the sense dictated by said root variable. For future research, it is worth studying the possibility of generating DTs by varying the root node and analysing all the rules that may be obtained from a single set of data.

For the same purpose, we would like to apply new split criteria based on new mathematical models for representing information, as well as the new procedures used in classification to date. These criteria and procedures can be seen in Abellán and Masegosa (2010) and Abellán et al. (2011).

Acknowledgements

The authors express their gratitude to the Spanish General Directorate of Traffic (DGT) for supporting this research and offering all the resources that are available to them. Griselda López wishes to express her acknowledgement to the regional ministry of Economy, Innovation and Science of the regional government of Andalusia (Spain) for their scholarship to train teachers and researchers in Deficit Areas, which has made this work possible.

The authors appreciate the reviewer's comments and effort in order to improve the paper.

References

Abdel Wahab, H.T., Abdel-Aty, M.A., 2001. Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. Transportation Research Record 1746, 6–13.

Abdel-Aty, M., 2003. Analysis of driver injury severity levels at multiple locations using ordered probit models. Journal of Safety Research 34, 597–603.

Abellán, J., Baker, R.M., Coolen, F.P.A., 2011. Maximising entropy on the nonparametric predictive inference model for multinomial data. European Journal of Operational Research 212 (1), 112–122.

Abellán, J., Masegosa, A., 2010. An ensemble method using credal decision trees. European Journal of Operational Research 205 (1), 218–226.

Agrawal, R., Imielinski, T., Swami, A., 1993. Mining association rules between sets of items in large databases. In: Proceedings of ACM SIGMOD Conference on Management of Data (SIGMOD 1993), pp. 207–216.

Al-Ghamdi, A., 2002. Using logistic regression to estimate the influence of accident factors on accident severity. Accident Analysis and Prevention 34 (6), 729–741.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Chapman & Hall, Belmont, CA.

Chang, L.Y., Wang, H.W., 2006. Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accident Analysis and Prevention 38, 1019–1027.

De Oña, J., Mujalli, R.O., Calvo, F.J., 2011. Analysis of traffic accident injury on Spanish rural highways using Bayesian networks. Accident Analysis and Prevention 43, 402–411.

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., 1996. From data mining to knowledge discovery: an overview. In: Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (Eds.), Advances in Knowledge Discovery and Data Mining. AAAI Press/The MIT Press, pp. 1–34.

Gray, R.C., Quddus, M.A., Evans, A., 2008. Injury severity analysis of accidents involving young male drivers in Great Britain. Journal of Safety Research 39, 483–495.

Helai, H., Chor, C.H., Haque, M.M., 2008. Severity of driver injury and vehicle damage in traffic crashes at intersections: a Bayesian hierarchical analysis. Accident Analysis and Prevention 40, 45–54.

Kashani, A., Mohaymany, A., 2011. Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models. Safety Science 49, 1314–1320.

Kashani, A., Mohaymany, A., Ranjbari, A., 2011. A data mining approach to identify key factors of traffic injury severity. Promet-Traffic & Transportation 23 (1), 11–17.

Kockelman, K.M., Kweon, Y.J., 2002. Driver injury severity: an application of ordered probit models. Accident Analysis and Prevention 34, 313–321.

Kuhnert, P.M., Do, K.A., McClure, R., 2000. Combining non-parametric models with logistic regression: an application to motor vehicle injury data. Computational Statistics & Data Analysis 34 (3), 371–386.

Liu, P., 2009. A self-organizing feature maps and data mining based decision support system for liability authentications of traffic crashes. Neurocomputing 72, 2902–2908.

Montella, A., Aria, M., D'Ambrosio, A., Mauriello, F. Analysis of powered two-wheeler crashes in Italy by classification trees and rules discovery. Accident Analysis and Prevention, http://dx.doi.org/10.1016/j.aap.2011.04.025, in press.

Mujalli, R.O., de Oña, J. Injury severity models for motorized vehicle accidents: a review, Proceedings of the Institution of Civil Engineering – Transport, http://dx.doi.org/10.1680/tran.11.00026, in press.

Mujalli, R.O., de Oña, J., 2011. A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks. Journal of Safety Research 42, 317–326.

Pakgohar, A., Tabrizi, R.S., Khalilli, M., Esmaeili, A., 2010. The role of human factor in incidence and severity of road crashes based on the CART and LR regression: a data mining approach. Procedia Computer Science 3, 764–769.

Pande, A., Abdel-Aty, M., 2009. Market basket analysis of crash data from large jurisdictions and its potential as a decision supporting tool. Safety Science 47, 145–154.

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, California.

Quinlan, J.R., 1986. Induction of decision trees. Machine Learning 1 (1), 81–106.

Savolainen, P., Mannering, F., Lord, D., Quddus, M., 2011. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accident Analysis and Prevention 43, 1666–1676.

Shannon, C.E., 1948. A mathematical theory of communication. Bell System Technical Journal 27, 379–423 and 623–656.

Sohn, S.Y., Shin, H.W., 2001. Data mining for road traffic accident type classification. Ergonomics 44, 107–117.

WHO, World Health Organisation, 2009. Informe Global sobre el estado de la Seguridad Vial: Tiempo para la Acción. Available at: www.who.int/violence_injury_prevention/road_safety_status/2009

Webb, G.I., 2007. Discovering significant patterns. Machine Learning 68, 1–33.

Witten, I.H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed. Morgan Kaufmann, San Francisco, CA.

Xie, Y., Zhang, Y., Liang, F., 2009. Crash injury severity analysis using bayesian ordered Probit models. Journal of Transportation Engineering ASCE 135 (1), 18–25.

Expert Systems with Applications 40 (2013) 6047–6054

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Analysis of traffic accident severity using Decision Rules via Decision Trees

Joaquín Abellán [a,*], Griselda López [b], Juan de Oña [b]

[a] Department of Computer Science & Artificial Intelligence, University of Granada, ETSI Informática, c/Periodista Daniel Saucedo Aranda s/n, 18071 Granada, Spain
[b] TRYSE Research Group, Department of Civil Engineering, University of Granada, ETSI Caminos, Canales y Puertos, c/Severo Ochoa s/n, 18071 Granada, Spain

ARTICLE INFO

Keywords:
Traffic accident
Severity
Road safety
Decision Trees
Decision Rules

ABSTRACT

A Decision Tree (DT) is a potential method for studying traffic accident severity. One of its main advantages is that Decision Rules (DRs) can be extracted from its structure. And these DRs can be used to identify safety problems and establish certain measures of performance. However, when only one DT is used, rule extraction is limited to the structure of that DT and some important relationships between variables cannot be extracted. This paper presents a more effective method for extracting rules from DTs. The method's effectiveness when applied to a particular traffic accident dataset is shown. Specifically, our study focuses on traffic accident data from rural roads in Granada (Spain) from 2003 to 2009 (both included). The results show that we can obtain more than 70 relevant rules from our data using the new method, whereas with only one DT we would have extracted only five relevant rules from the same dataset.

1. Introduction

The current large number of road accidents implies an unacceptable burden on the community in terms of human injury and economic cost. Therefore, one of the main tasks of safety analysts is to make a comprehensive assessment of traffic accidents to determine what caused them, so measures can be taken to mitigate the severity of their consequences.

Usually, an accident severity analysis is carried out to study a particular dataset of traffic accidents with the aim of obtaining useful knowledge to tackle this problem. In most countries, traffic accidents are recorded in accident reports by police officers, and subsequently the information is stored in a dataset. A huge amount of information can be obtained from such datasets. It could be said that their true potential consists in the knowledge that can be extracted from them.

Traditionally, regression techniques such as Logit and Porbit have been used to analyze traffic accident severity (Kashani & Mohaymany, 2011; Mujalli & de Oña, 2013; Savolainen, Mannering, Lord, & Quddus, 2011). However, these techniques establish their own model assumptions and pre-defined underlying relationships between dependent and independent variables. If the assumptions are violated, the model can lead to erroneous estimations of injury likelihood (Chang & Wang, 2006).

Data Mining (DM) techniques are one of the solutions used to analyze huge amounts of data and turn it into useful information and knowledge (Han & Kamber, 2006). DM has been widely used in crash severity analysis with satisfactory results. Abdel Wahab and Abdel-Aty (2001) investigated the use of Artificial Neural Network models for predicting injury severity in two-vehicle crashes at signalized intersections. Recently, Bayesian Networks have been used to analyze traffic accident severity (De Oña, López, Mujalli, & Calvo, 2013b, 2011; Mujalli & de Oña, 2011). Decision Trees (DT) is another DM technique used to study crash severity (Chang & Chien, 2013; Chang & Wang, 2006; De Oña, López, & Abellán, 2013a; Montella, Aria, D'Ambrosio, & Mauriello, 2011, 2012).

DTs, in particular, represent a set of useful methods for analyzing traffic accident severity because, normally, they are non-parametric methods that do not depend on any functional form and require no prior probabilistic knowledge on the phenomena under study. Moreover, the structure of a DT permits the extraction of Decision Rules (DR) that can be used to discover behaviors that occur within a specific dataset. Safety analysts could use these rules to understand the events leading up to a crash and identify the variables that determine how serious an accident will be (De Oña et al., 2013a).

DTs have been largely reported in road safety literature. Specifically, the most widely used method in the literature on traffic accident severity is the CART method (Chang & Chien, 2013; Chang & Wang, 2006; De Oña et al., 2013a; Kashani & Mohaymany, 2011; Kashani, Mohaymany, & Ranjbari, 2011; Kuhnert, Do, & McClure, 2000; Montella et al., 2011, 2012; Pakgohar, Tabrizi, Khalilli, & Esmaeili, 2010). However, CART always yields binary trees, which sometimes cannot be summarized as efficiently for interpretation and/or presentation (Breiman, Friedman, Olshen, & Stone, 1984).

* Corresponding author. Tel.: +34 958242376; fax: +34 958243371.
E-mail address: jabellan@decsai.ugr.es (J. Abellán).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.05.027

In the case of road accidents, they may not be very practical when it comes to analyzing the impact of a specific category of variable on crash severity. The C4.5 algorithm (Quinlan, 1993) is another method that is frequently used in several fields because it does not present the binary restriction when tree building. It has been used before to analyze traffic accident severity (De Oña et al., 2013a). An important difference between the two methods (CART vs. C4.5) is the split criterion: the CART method uses the Gini Index, based on a measure of diversity; and the C4.5 algorithm uses the Info Gain Ratio (IGR), based on the entropy measure on probabilities (Shannon, 1948).

However, using DRs from DTs to extract knowledge from a specific dataset also poses certain limitations. The extraction of knowledge is constrained by the tree's structure, for instance, and the DRs are dependent on a DT's structure. The DRs are extracted from each tree branch from the root node to the terminal node, and therefore knowledge is extracted only in that direction. However, there could be other important rules that depend on the root node from which the tree is built, and that are not detected by the tree's structure.

In this paper, a particular method for extracting DRs from DTs is used to extract all the knowledge from a particular dataset. The main characteristic of this method is that different DTs are built by varying the root node. Thus, every possible set of DRs is obtained from each tree. The resulting useful rules could be used by road safety analysts to establish specific measures of performance.

To conduct a full analysis of the dataset, in our method for extracting DRs, we use different DTs built using two different split criteria, both each with a different meaning. In fact, the two criteria complement each other, and even a previous study recommends using the both criteria for a full analysis (De Oña et al., 2013a). By doing so, a broader range of rules can be obtained from a single dataset.

The paper is structured as follows: Section 2 shows the main features of the traffic accident data used to validate the methodology. The necessary prior knowledge on decisions trees and the procedure to build them is presented. It also describes the method used to obtain Decision Rules, and how to obtain the importance of each of the variables considered in the model. Section 3 presents the main results obtained and the discussion. Finally, the last section presents the conclusions.

2. Materials and methods

2.1. Traffic accident data

Traffic accidents where only 1 vehicle was involved, for two-lane rural highways in Granada (Spain), were collected from the Spanish General Traffic Accident Directorate (DGT). The study period was 7 years (2003–2009) and accidents at intersections were not considered. Thus, the total number of accidents was 1801.

In order to identify the main factors that had an impact on accident severity and taking into account the available variables in the original dataset, 19 variables were used (see Table 1). The variables described characteristics related to the driver (age and gender); accident (month, time, day, number of injuries, occupants involved, accident type and cause); road (safety barriers, pavement width, lane width, shoulder width, shoulder type, road markings and sight distance); vehicle (vehicle type); and environment (atmospherics factors and lighting conditions).

The class variable was accident severity (SEV in Table 1). Following previous studies (Chang & Wang, 2006; De Oña et al., 2011; Kashani & Mohaymany, 2011), accident severity was defined according to the worst injured occupant, and two levels of severity were identified: accident with slightly injured (SI) and accidents with killed or seriously injured (KSI).

2.2. Classification and Decisions Trees

In the general domain of DM, a supervised classification problem is normally defined as follows: given a dataset of observations, called a *training set*, we want to obtain a set of rules that can be used to assign a value of the variable to be predicted to each new observation. To verify the quality of this set of rules, a different set of observations is used; this set is called the *test set*. The variable to be predicted (classified) is called *class variable* and the rest of variables in the dataset are called *predictive attributes* or *features*. There are important applications of classification in fields such as medicine, bioinformatics, physics, pattern recognition, economics, civil engineering, etc.

A DT is a structure that can be used in classification and regression tasks. If the class variable (i.e., the variable under study) has a finite set of possible states or values, the task is called a classification; otherwise, it is called a regression.

Within a DT, each node represents a feature and each branch represents one of the states of this variable. A tree leaf (or terminal node) specifies the expected value of the class variable depending on the information contained in the training dataset. Associated to each node is the most informative variable which has not already been selected in the path from the root to the node (as long as this variable provides more information than if it had not been included). In the latter case, a leaf node is created with the most probable class value for the partition of the dataset defined with the configuration given by the path from the root node to that leaf node.

When a new sample or instance of the test dataset is obtained, a decision or prediction about the state of the class variable can be made by following the path in the tree from the root to a leaf, using the sample values and the tree's structure.

A DT allows us to extract DRs directly. A DR is a logic conditional structure of the type "IF A THEN B". Where A is the antecedent of the rules (in our case, a set of statuses of several attribute variable); and B is the consequent (in our case, it is only one state of the class variable). Thus, each rule starts at the root node, and each variable that intervenes in tree division makes an IF of the rule, which ends in leaf nodes with a value of THEN (which is associated with the state resulting from the leaf node). The resulting state is the status of the class variable that shows the highest number of cases in the leaf node analyzed. Thus, a priori, the number of rules can be identified with the number of terminal nodes in the tree.

Fig. 1 shows an example of a DT built using a dataset of accidents. The DT is formed by two attribute variables, and the class variable is the *severity* (two states) of the accidents. This example shows how accidents are classified by each status of the class variable (slight accidents vs. severe accidents). In addition, the chart gives the number of cases shown in each leaf or terminal node (shaded nodes in the tree), distinguishing the cases that are predicted correctly in each terminal node. One example of DRs is the following: IF (*age* $\leqslant$25 yrs AND *speed* $\leqslant$80 km/h) THEN (*severity* = slight accident).

There is a wealth of information in the literature about different procedures to build DT, but normally they have the following characteristics in common:

- The criterion used for selecting the attribute to be placed in a node and branching. This criterion is known as the split criterion.
- The criterion used to stop the branching of the tree.
- The method for assigning a class label or a probability distribution at the leaf nodes.

Table 1
Variable description.

Num	Variables	Description: code	Severity		
			Count	%SI	%KSI
1	ACT: accident type	Fixed objects collision: **CO**	19	76.47	23.53
		Collision with pedestrian: **CP**	152	33.33	66.67
		Other (collision with animals, etc.): **OT**	32	68.57	31.43
		Rollover (in carriage without any collision): **RO**	118	61.86	38.14
		Run off road (with or without collision): **ROR**	1480	51.77	48.23
2	AGE: age	$\leqslant$20: $\leqslant$20	219	52.73	47.27
		[21–27]: [21–27]	492	50	50
		[28–60]: [28–60]	948	51.76	48.24
		$\geqslant$61: $\geqslant$61	110	59.68	40.32
		Unknown: **UN**	32	27.59	72.41
3	ATF: atmospheric factors	Good weather: **GW**	1540	50.58	49.42
		Heavy rain: **HR**	43	63.16	36.84
		Light rain: **LR**	161	58.75	41.25
		Other: **O**	57	51.06	48.94
4	BAR: safety barriers	No: **N**	1740	48.3	54.7
		Yes: **Y**	61	53.6	46.4
5	CAU: cause	Driver characteristics: **DC**	1471	48.99	51.01
		Combination of factors: **CO**	262	61.16	38.84
		Other: **OT**	29	72.73	27.27
		Road characteristics: **RC**	24	84	16
		Vehicle characteristics: **VC**	15	63.64	36.36
6	DAY: day	Working day after the weekend or public holiday: **APH**	131	57.62	42.38
		Working day before the weekend or public holiday: **BPH**	286	52.26	47.74
		On a weekend or public holiday: **PH**	532	50.36	49.64
		Regular working day: **WD**	852	51.05	48.95
7	LAW: lane width	<3,25 m: **THI**	503	46.87	53.13
		[3,25-3,75] m: **MED**	1264	53.2	46.8
		>3,75 m: **WID**	34	58.54	41.46
8	LIG: lighting	Daylight: **DAY**	958	55.49	44.51
		Dusk: **DU**	103	54.29	45.71
		Insufficient (night-time): **IL**	131	51.15	48.85
		Sufficient (night-time): **SL**	66	59.72	48.28
		Without lighting (night-time): **WL**	543	43.1	56.9
9	MON: month	Autumn: **AUT**	412	53.07	46.93
		Spring: **SPR**	440	53.64	46.36
		Summer: **SUM**	479	51.63	48.37
		Winter: **WIN**	470	47.92	52.08
10	NOI: number of injuries	1 injury: **[1]**	1233	53.43	46.57
		>1 injury: **[>1]**	568	47.35	52.65
11	OI: occupants involved	1 occupant: **[1]**	1171	51.2	48.8
		2 occupants: **[2]**	374	51.48	48.52
		>2 occupants: **[>2]**	256	53.71	46.29
12	SHT: shoulder type	No: **N**	309	49.35	50.65
		Non existent or impassable: **NE**	580	50.89	49.11
		Yes: **Y**	912	52.74	47.26
13	PAW: pavement width	[6–7] m: **MED**	530	53.19	46.81
		<6 m: **THI**	282	45.56	54.44
		>7 m: **WID**	989	52.27	47.73
14	ROM: pavement markings	Does not exist or was deleted: **DME**	168	52.35	47.65
		Separate margins of roadway: **DMR**	180	48.31	51.69
		Separate lanes and define road margins: **SLD**	1368	52.23	47.77
		Separate lanes only: **SLO**	85	46.59	53.41
15	SEX: gender	Female: **F**	286	62.18	37.82
		Male: **M**	1513	49.61	50.39
		Unknown: **UN**	2	75	25
16	SHW: shoulder width	<1.5 m: **THI**	699	52.54	47.46
		[1.5–2.5] m: **MED**	898	50.28	49.72
		Non existent or impassable: **NE**	204	50.57	49.43
17	SID: sight distance	Atmospheric: **ATM**	30	67.5	32.5
		Building: **BU**	6	36.36	63.64
		Other: **OT**	12	50	50
		Topography: **TOP**	420	49.39	50.61
		Vegetation: **VEG**	13	50	50
		Without restriction: **WR**	1320	51.94	48.06
18	TIM: time	[00:00–05:59]: **[0–6)**	340	48.06	51.94
		[06:00–11:59]: **[6–12]**	380	58.73	41.27
		[12:00–17:59]: **[12–18]**	591	52.77	47.23
		[18:00–23:59]: **[18–24]**	490	47.22	52.78
19	VEH: vehicle type	Cars: **CAR**	1287	47.1	52.9
		Trucks: **TRU**	78	53.8	46.2
		Motorbikes and motorcycles: **MOT**	385	35.6	64.4
		Other: **OT**	51	50.6	49.4
20	SEV: severity	Accident with slightly injured: **SI**	929	–	–
		Accidents with killed or seriously injured: **KSI**	872		

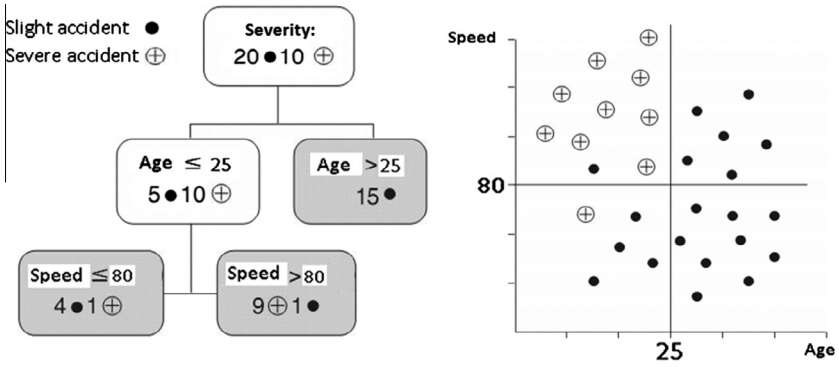


Fig. 1. Example of a DT's structure and classification.

– The pruning process (pre or post building process), which simplifies the structure of the tree and prevents over-fitting (i.e., the dependence of the data used to build the model).

DTs started to play an important role in machine learning following publication of the CART method (Breiman et al., 1984) and Quinlan's ID3 algorithm (Quinlan, 1986). The former uses a split criterion based on the Gini Index. The Quinlan's method uses a split criterion, called the Information Gain (IG), based on the entropy measure on probabilities (Shannon, 1948). Subsequently, Quinlan (1993) also presented the algorithm C4.5, which is an advanced version of ID3 with a split criterion, called the Information Gain Ratio (IGR), which is similar to the one used in the ID3 procedure penalizing the variables with many states. Since then, C4.5 has been considered as a standard model in supervised classification. It has also been widely applied to very different fields as a data analysis tool.

The Gini Index is a measure of diversity, and for a variable C (for example, the class variable in a classification problem), it can be expressed as follows:

$$gini(C) = 1 - \sum_j p^2(C = c_j) \quad (1)$$

In the same line, Shannon's entropy is a measure of information based on uncertainty that can be expressed as:

$$H(C) = -\sum_j p(C = c_j) log(p(C = c_j)) \quad (2)$$

The split criterion used in CART, which we call GInf, is based on the Gini Index and can be expressed as follows:

$$\mathrm{GInf}(C,X) = gini(C|X) - gini(X), \quad (3)$$

where $gini(C|X) = \sum_t p(x_t) gini(C|X = x_t)$ and X is another known variable (for example, a feature variable in a classification problem).In the C4.5 procedure, the split criterion is called the info gain ratio and it is a measure based on Shannon's entropy. It is defined as:

$$\mathrm{IGR}(C,X) = \frac{IG(C,X)}{H(X)}, \quad (4)$$

where $IG(C,X) = H(C) - H(C|H)$, IG is the Info Gain measure defined by Quinlan (1986) and $H(C)$ is the entropy of C. The probability of each value of the class variable is estimated in the training dataset. In the same way, $H(C|X) = -\sum_t \sum_j p(c_j|x_t) log(p(c_j|x_t))$, where x_t, $t = 1,\ldots,|X|$, is each possible state of X; and c_j, $j = 1,\ldots,k$, each possible state of C.

2.3. Procedure for building Decision Trees

In this section, we explain how to build a simple DT using the above mentioned split criteria. The procedure proposed by Abellán and Moral (2003) to build DTs using imprecise probabilities and uncertainty measures is used. The method can easily be adapted to be used with precise probabilities; for example, via the GInf or IGR split criteria.

Each node N in a DT produces a partition D of the dataset (for the root node the entire dataset is considered). Also, each node N has associated a list "Γ" of labels of features (features that are not in the path from the root node to N). The recursive and simple procedure formulated by Abellán and Moral (2003) for building DTs can be expressed in the algorithm shown in Fig. 2.

Each ***Exit*** state in the above procedure corresponds to a leaf node. Here, the most probable value of the class variable, associated with the corresponding partition, is selected.

2.4. Method to obtain Decision Rules: Information Root Node Variation method

When rules are obtained from a single DT, they are determined by the variable that is used as a root node. In other words, the information we select from our dataset depends on the direction indicated by the variable in that root node. This is the most informative variable about the class variable using a split criterion.

The method that we propose here for obtaining rules, which we call the *Information Root Node Variation* (IRNV) method, is based on using different trees obtained by varying the root node. In this method, if there are m features, and RX_i is the feature that occupies position i in importance with regards to the split criterion, RX_i is used as the root to build DT_i ($i = 1,\ldots,m$). We use the simple method for building trees explained in Section 2.2, nonetheless now the root node is selected directly for each tree (the rest of the building procedure remains the same). Thus, we obtain m trees and m rule sets, DT_i and RS_i ($i = 1,\ldots,m$), respectively. Each RS_i is checked in the test set to obtain the final rule set. The entire procedure is carried out using GInf and IGR criteria.

The following chart gives a more systematic explanation of the entire process:

(1) Select Ginf as the split criterion (SC) for building trees.
(2) Build DT_i using RX_i as the root node and SC; for $i = 1,\ldots,m$.
(3) Extract RS_i from each DT_i.
(4) Check RS_i in the corresponding TEST set → Selection of rules from RS_i.
(5) Extract the final rule set obtained by using the SC.
(6) Use the IGR as SC and go back to step 2. Skip if IGR was used before.
(7) Join the final rule sets obtained using GInf and IGR.

Fig. 3 gives a graphic explanation of the procedure for each split criterion. In other words, the method shown in Fig. 3 must be applied as many times as split criteria that we apply.

Procedure BuildTree (N, Γ)

1. *If Γ =Φ, then* ***Exit***
2. *Let D be the partition associated with node N*
3. *Compute the value of the maximum gain of information for a feature on D (using a split criterion: SC)*
   δ= max SC(C,X)
4. *If δ is lower than or equal to 0 then* ***Exit***
5. *Else*
   6. *Let X_t be the variable for which the maximum δ is attained*
   7. *Remove X_t from Γ*
   8. *Assing X_t to the node N*
   9. *For each possible value x_t of X_t*
   10. *Add a node N_t*
   11. *Make N_t a child of N*
   12. *Call BuilTree (N_t, Γ)*

Fig. 2. Algorithm to build a DT.

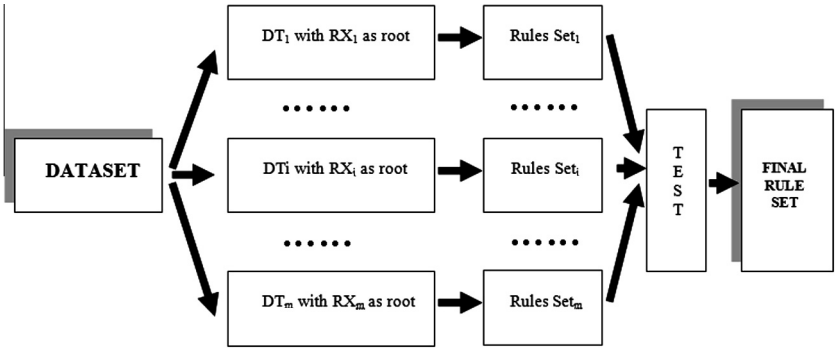


Fig. 3. Information Root Node Variation method for each split criterion.

In our case, when we use the IRNV method on our dataset, 19 DTs are obtained, i.e., one DT for each feature (see Table 1), for each one of the split criteria (GInf and IGR). All the DRs are extracted for each of the DTs built. Finally, each RS_i obtained from each DT_i is verified on the corresponding test set.

It is important to point out that we use two very different split criteria that can be used to build different trees, despite the fact that they begin with the same root node.

2.5. Significant Decision Rules

In order to extract significant and useful rules (i.e., rules that could provide useful information for the implementation of road safety strategies in the future), of the type IF A THEN B ("A→B"), the parameters and the minimum threshold used by Montella et al. (2011) and De Oña et al. (2013a) are used:

- *Support*
  (S) is the percentage of the dataset where "A and B" appear. The minimum threshold selected is $S \geqslant 0.6\%$.
- *Population*
  (Po) is the percentage of the dataset where "A" appears. The minimum threshold selected is $Po \geqslant 1\%$.
- *Probability*
  (P) is the percentage of cases in which the rule is accurate (i.e., $P = S/Po$ expressed as percentage). The minimum threshold selected is $P \geqslant 60\%$.

The threshold values for parameters (P, S and Po) normally are selected depending on the following characteristics: nature of the data (balanced or unbalanced); significant interest in fatal crashes (rare events); and sample size (small or large datasets). Montella et al. (2011) used a large amount of data (crash data referred to the period 2003–2008) of an unbalanced type, with the aim of analyzing rare events. Therefore, they use a low value for S and Po. However, in De Oña et al. (2013a) the sample size was smaller, the sample was balanced, and their aims were different; so they established higher thresholds for S and Po (the threshold for P is obviously determined by the ones of S and Po). Our data here have the same characteristics as the data used in De Oña et al. (2013a). Hence, we use the same set of thresholds for the parameters.

Due to the large number of patterns considered, DTs can suffer from an extreme risk of type-1 error, that is, of finding patterns that appear due to chance alone to satisfy constraints on the sample data (Webb, 2007). To reduce this risk error and following the suggestions of other authors (Chang & Chien, 2013; De Oña et al.,

2013a; Kashani & Mohaymany, 2011; Montella et al., 2011) the rules extracted on the training set (with the minimum value of the parameters) are validate using the testing set.

2.6. Importance of the variables

The importance of a variable in the model is defined following Eq. (5):

$$\mathrm{VIM}(X) = \sum_{i=1}^{h} \frac{nxi}{n} I(C, X = x_i)' \quad (5)$$

where X is the variable with possible states $\{x_i \dots x_n\}$, C is the class variable (SEV in our case), nxi is the number of cases that $X = xi$, and n is the number of total cases; and I is the GInf or the IGR split criterion, i.e., an information gain measure.

This measure expresses the gain in information that we obtain on the class variable C, when we use the information expressed on C via a feature X. The values of the *VIM* measure on a feature X can be different if we use different split criteria. If we divide by the largest value obtained for a feature, we will obtain the *normalized importance* of each variable with respect to the class variable.

3. Results and discussion

The software used to build the DTs was Weka (Witten & Frank, 2005). The procedures for building the DTs based on each split criterion and the root node variation procedure were implemented using the method proposed by Abellán and Masegosa (2010).

In order to obtain DRs that would be useful and easy to understand by the analysts, we built DTs with only four levels. Previous studies (Montella et al., 2011, 2012) used the same number of levels.

Following the method exposed in Section 2.2 to obtain DRs we used only one DT (DT_1 in Table 2). Following the IRNV method, by varying the root node, 19 DTs, can be used to obtain DRs, (DT_1 to DT_{19} in Table 2) for each of the split criteria (GInf and IGR). Thus, the total number of DTs generated is 38 (19 for GInf and 19 for IGR).

DT_1 presents a different root node depending on the split criteria: ACT is selected as the root node when GInf is used, whereas SEX is selected when using IGR (Table 2). For this DT, 22 rules were extracted from the training set (14 with GInf and 8 with IGR) but only 11 rules (5 with GInf and 6 with IGR) were validated with the testing set.

Table 2 shows the number of the DRs obtained from each DT for each root node. Both criteria (GInf and IGR) generate more than 170 rules validated on the training set, (i.e., verify the minimum threshold fixed for the parameters *S*, Po and *P*). LIG is the variable that generates the highest number of rules when it is used as a root node. Depending on the criteria, the number of rules is: 17 rules when GInf is used, and 16 rules when using IGR.

When the rules are validated using the testing set, the number of rules decreases considerably (78 rules with GInf and 81 rules with IGR). We would like to remark that all DTs generate valid DRs. When GInf is used, the root node that generates the highest number of valid rules is LAW (8 rules). When IGR is used, the root node that generates the highest number of valid rules is SHW (10 rules). In both cases, the number of valid rules obtained from a single tree, using both criteria, is lower (5 with GInf and 6 with IGR).

Table 3 shows the normalized importance of the variables in the model. Six variables were detected as having the greatest impact on accident severity with GInf, with percentages that vary from 100% to 61.21%. Five variables were detected with IGR, with percentages ranging from 100% to 51.96%. Both split criteria identify

Table 3
Normalized importance of the variables.

Variable	GInf (%)	Variable	IGR (%)
ACT	100.00	SEX	100.00
CAU	77.89	ACT	94.51
LIG	69.56	CAU	81.72
SEX	69.53	ATF	69.17
VEH	67.43	VEH	51.96
ATF	61.21	LIG	36.30
PAW	44.20	NOI	33.37
TIM	41.09	SID	32.64
AGE	40.72	PAW	27.35
SID	35.47	AGE	21.81
NOI	33.78	LAW	18.29
DAY	26.60	TIM	18.25
LAW	20.91	DAY	13.59
MON	11.58	BAR	9.24
ROM	4.64	MON	5.06
OI	4.54	ROM	3.46
SHT	3.52	OI	3.15
BAR	2.02	SHT	2.23
SHW	0.77	SHW	0.46

Table 2
Number of rules obtained in the different steps of the IRNV method.

DTS	GInf			IGR		
	Root node	Rules training	Validated rules	Root node	Rules training	Validated rules
DT_1	ACT	14	5	SEX	8	6
DT_2	CAU	16	4	ACT	8	2
DT_3	SEX	8	4	CAU	12	5
DT_4	LIG	17	2	CAT	15	7
DT_5	VEH	12	4	VEH	6	1
DT_6	CAT	14	6	LIG	16	7
DT_7	PAW	10	3	NOI	5	2
DT_8	AGE	7	3	SID	10	3
DT_9	TIM	12	3	PAW	7	3
DT_{10}	SID	8	4	AGE	7	3
DT_{11}	NOI	9	3	LAW	4	3
DT_{12}	DAY	12	5	TIM	11	6
DT_{13}	LAW	14	8	DAY	11	5
DT_{14}	MON	16	3	BAR	6	4
DT_{15}	ROM	8	5	MON	10	4
DT_{16}	OI	12	7	ROM	7	3
DT_{17}	SHW	14	5	OI	5	1
DT_{18}	BAR	11	3	SHW	13	10
DT_{19}	SHT	13	1	SHT	13	6
Total		227	78		174	81

Table 4
DRs from the IRNV method.

Num.	rules (IF...)	THEN	*S*%	Po%	*P*%
1	NOI = [1];OCU = [1];VEH = MOT;ACT = ROR	KSI	7.62	10.79	70.59
2	**CAU = DC;VEH = MOT;ATF = GW;ACT = ROR**	**KSI**	**8.10**	**11.59**	**69.86**
3	SEX = M;ACT = ROR;CAU = DC;VEH = MOT	KSI	7.78	11.43	68.06
4	**ACT = ROR;CAU = DC;VEH = MOT;ATF = GW**	**KSI**	**8.10**	**11.59**	**69.86**
5	ATF = GW;SEX = M;ACT = ROR;VEH = MOT	KSI	8.81	12.86	68.52
6	LIG = WL;ATF = GW;SEX = M;LAW = THI	KSI	5.40	7.46	72.34
7	SID = WR;CAU = DC;VEH = MOT;BAR = N	KSI	7.06	11.67	60.54
8	TIM = [18-24];ATF = GW;LIG = WL;BAR = N	KSI	8.10	13.02	62.20
9	BAR = N;SEX = M;ACT = CP;ATF = GW	KSI	5.00	7.38	67.74
10	SHW = NE;SEX = M;ATF = GW;LIG = WL	KSI	7.06	11.35	62.24

Note: rules 1 and 2 have been obtained from the GInf criterion and rules 3–10 from the IGR criterion. In bold are the rules that are repeated in both methods.

the same variables, although with different orders of importance: ACT, CAU, SEX, VEH, and ATF. The variable LIG is also detected with GInf (with a percentage higher than 50%), whereas IGR's percentage in the model is slightly lower (36.3%). However, it occupies sixth place in the importance ranking.

From the point of view of safety, these results are consistent with previous studies. Several authors (Kockelman & Kweon, 2002; De Oña et al., 2011, 2013a, 2013b) have pointed out that *accident type* is a key variable in severity. Chang and Wang (2006) stressed that the most important variable associated with crash severity was *vehicle type*. *Causes of the accident* also match previous studies (Al-Ghamdi, 2002; Kashani & Mohaymany, 2011). Xie, Zhang, and Liang (2009) and Mujalli and de Oña (2013) found that *atmospheric factors* have an important effect on severity. Many studies have also indicated gender differences in injury severity (Abdel-Aty, 2003; Evans, 2001; Obeng, 2011; Ulfarsson & Mannering, 2004). *Lighting conditions* have been also identified as a variable with effects on severity. In fact, Gray, Quddus, and Evans (2008), Abdel-Aty (2003) and Helai, Chor, and Haque (2008) found that more severe injuries are predicted during darkness. Pande & Abdel-Aty, 2009 concluded that there is a significant correlation between lack of illumination and high crash severity. De Oña et al. (2011) and De Oña et al. (2013a) also pointed out that KSI accidents are associated with roadways with no lighting.

In order to describe the pattern showed in the rules, only rules with the most severe consequences (accidents with killed or seriously injured, KSI) are extracted in Table 4. The IRNV method generates 4 KSI rules with GInf and 3 KSI rules with IGR (DT_1) and 36 KSI rules for GInf and 28 for IGR (DT_{2-19}). Due to the large number of rules obtained with each method, only rules with $S > 5\%$ are extracted on Table 4. Support is a parameter that combines confidence and population. Therefore, a support higher than 5% implies that the rule is met by at least 63 accidents in the sample under study.

Table 4 shows the following patterns. Using the IRNV method, we identified two rules (rules 1 and 2) with GInf (and neither of them was obtained from DT_1); and seven rules (rules 3 to 10) with IGR (rule 3 was obtained from DT_1).

Rules 1 to 5 allow the identification of one of the most important concerns for road safety in Spain: run-off-road for motorcycles in two-lane rural highways (DGT, 2011). Precisely, one of the priorities of the Spanish Road Safety Strategy 2011–2020 (DGT, 2011) is to diminish this type of accidents, as well as their severity.

- Rule 1 identifies this kind of accident when only one occupant is involved (therefore, there is also only one injury). The probability of KSI in these cases is one of the highest (70.6%).
- Rules 2 and 4 are the same. Motorcyclists' run-off-road accidents under good weather conditions when the cause of the accident is due to the driver. The probability of KSI in these cases is 69.9%.
- Rule 3 identifies motorcyclists' run-off-road accidents for male drivers and due to driver characteristics. The probability of KSI is 68%.
- Rule 5 shows a similar pattern: motorcyclists' run-off-road accidents under good weather conditions when the driver is a male. The probability of KSI in these cases is 68.5%.

In this sense, the DGT is making an important effort to lower the number of accidents of this type (e.g., advertising campaigns that target motorcyclists; more stringent monitoring on two-lane rural highways; lowering the maximum speed limit on two-lane rural highways; etc.). The DGT also tries to lower motorcycle crash severity (e.g., with projects that target improvements on the shoulders of two-lane rural highways that have no safety barriers). On the other hand, one of the priorities in the DGT's 2013–2016 Research Plan (DGT, 2011) is to identify the main factors that lead to accidents of this type (run-off-road for motorcycles on two-lane rural highways).

Table 4 shows that three rules (rules 7–9) identify KSI accidents on two-lane rural highways with no safety barriers:

- Rule 7 identifies motorcyclists' accidents with no-restrained sight distance due to the driver. Even if this rule does not present a very high probability (only 60.5%), it represents 11.7% of the population.
- Rule 8 identifies accidents in the evening (18–24 h) under good weather conditions on roads with no lighting. This rule presents the highest population (13.0%).
- Rule 9 identifies collision with pedestrian accidents under good weather conditions when the driver is a male.

These rules show that safety barriers play a fundamental role in crash severity on two-lane rural highways.

Finally, rules 6 and 10 share 3 variables: ACT, LIG and SEX. Thus, the pattern described for these rules refers to an accident on roads with no lighting, when atmospheric factors are good and the driver is male. If the road has a lane width of <3.25 m, rule 6 is obtained, whereas rule 10 is for roads where the shoulder is non-existent or impassable. Thus, from the point of view of road safety, bad lighting conditions and bad road features increase accident severity.

4. Conclusions

If we use a single DT to extract knowledge based on a dataset, in the form of DRs, we are constrained by the DT's structure. However, the method proposed in this paper uses one DT for each variable under study (variables that describe the data), which allows us to extract much more knowledge. If we add that our model uses two split criteria, the extraction is even more extensive.

More than 70 significant validated rules were obtained from the practical study conducted on traffic accident data from rural roads

in Granada (Spain). For the KSI rules, only one rule was repeated in both methods (rule 2 with rule 4); however some patterns were similar in both methods (rules 1–5). Although the criterion based on IGR detected a higher number of rules (with the minimum parameters established), it could be said that the two criteria complement each other when searching for the key factors that have an impact on accident severity, because each criterion detects different patterns within the same dataset.

With regards to the special patterns detected for the KSI accidents analyzed, we could highlight the high number of rules for the motorcyclists' run-off-road accidents (rules 1 to 5). These results are in line with current concerns for road safety on two-lane rural highways. The Spanish Road Safety Strategy 2011–2020 (DGT, 2011) promotes specific studies on the factors associated with the highest levels of severity in run-off-road accidents on two-lane rural highways (i.e., KSI) when motorcyclists are involved.

Our study also highlights the need for studying the conditions in the environment of two-lane rural highways (i.e., safety barriers, shoulders, visibility, lighting, etc.), because they have a substantial impact on crash severity.

Finally, it should be pointed out that the proposed method can be extrapolated for specific studies on other datasets (i.e., other infrastructure, roads and countries). This method can also provide DRs that would be useful and easy for road safety analysts and managers to use to identify problems. Also, other split criteria can be applied in the IRNV method, as the one of Abellán, Baker, Coolen, Crossman, and Masegosa (2013), based on the tools of Abellán, Baker, and Coolen (2011).

Acknowledgements

The authors express their gratitude to the Spanish General Directorate of Traffic (DGT) for supporting this research and offering all the resources that are available to them. Griselda López wishes to express her acknowledgement to the Regional Ministry of Economy, Innovation and Science of the regional government of Andalusia (Spain) for their scholarship to train teachers and researchers in Deficit Areas, which has made this work possible. The authors appreciate the reviewer's comments and effort in order to improve the paper.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.eswa.2013.05.027.

References

Abdel Wahab, H. T., & Abdel-Aty, M. A. (2001). Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. *Transportation Research Record, 1746*, 6–13.

Abdel-Aty, M. (2003). Analysis of driver injury severity levels at multiple locations using ordered probit models. *Journal of Safety Research, 34*, 597–603.

Abellán, J., & Moral, S. (2003). Building classification trees using the total uncertainty criterion. *International Journal of Intelligent Systems, 18*(12), 1215–1225.

Abellán, J., & Masegosa, A. (2010). An ensemble method using credal Decision Trees. *European Journal of Operational Research, 205*(1), 218–226.

Abellán, J., Baker, R. M., & Coolen, F. P. A. (2011). Maximising entropy on the nonparametric predictive inference model for multinomial data. *European Journal of Operational Research, 212*(1), 112–122.

Abellán, J., Baker, R. M., Coolen, F. P. A., Crossman, R., & Masegosa, A. (2013). Classification with Decision Trees from a nonparametric predictive inference perspective. *Computational Statistics and Data Analysis*. http://dx.doi.org/10.1016/j.csda.2013.02.009.

Al-Ghamdi, A. (2002). Using logistic regression to estimate the influence of accident factors on accident severity. *Accident Analysis and Prevention, 34*, 729–741.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*. Belmont, CA: Chapman and Hall.

Chang, L. Y., & Chien, J. T. (2013). Analysis of driver injury severity in truck involved accidents using a non-parametric classification tree model. *Safety Science, 51*, 17–22.

Chang, L. Y., & Wang, H. W. (2006). Analysis of traffic injury severity: An application of non-parametric classification tree techniques. *Accident Analysis and Prevention, 38*, 1019–1027.

De Oña, J., López, G., & Abellán, J. (2013a). Extracting Decision Rules from police accident reports through Decision Trees. *Accident Analysis and Prevention, 50*, 1151–1160.

De Oña, J., López, G., Mujalli, R. O., & Calvo, F. J. (2013b). Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks. *Accident Analysis and Prevention, 51*, 1–10.

De Oña, J., Mujalli, R. O., & Calvo, F. J. (2011). Analysis of traffic accident injury on Spanish rural highways using Bayesian networks. *Accident Analysis and Prevention, 43*, 402–411.

DGT (2011). *Spanish road safety strategy 2011–2020*. Madrid: Traffic General Directorate (pp. 222).

Evans, L. (2001). Female compared with male fatality risk from similar physical impacts. *The Journal of Trauma: Injury, Infection and Critical Care, 50*, 281–288.

Gray, R. C., Quddus, M. A., & Evans, A. (2008). Injury severity analysis of accidents involving young male drivers in Great Britain. *Journal of Safety Research, 39*, 483–495.

Han, J., & Kamber, M. (2006). *Data mining: concepts and techniques*. San Fransisco: Morgan kufman Publishers.

Helai, H., Chor, C. H., & Haque, M. M. (2008). Severity of driver injury and vehicle damage in traffic crashes at intersections: a Bayesian hierarchical analysis. *Accident Analysis and Prevention, 40*, 45–54.

Kashani, A., & Mohaymany, A. (2011). Analysis of the traffic injury severity on twolane, two-way rural roads based on classification tree models. *Safety Science, 49*, 1314–1320.

Kashani, A., Mohaymany, A., & Ranjbari, A. (2011). A data mining approach to identify key factors of traffic injury severity. *Promet-Traffic and Transportation, 23*(1), 11–17.

Kockelman, K. M., & Kweon, Y. J. (2002). Driver injury severity: an application of ordered probit models. *Accident Analysis and Prevention, 34*, 313–321.

Kuhnert, P. M., Do, K. A., & McClure, R. (2000). Combining non-parametric models with logistic regression: An application to motor vehicle injury data. *Computational Statistics & Data Analysis, 34*, 371–386.

Montella, A., Aria, M., D'Ambrosio, A., & Mauriello, F. (2011). Data mining techniques for exploratory analysis of pedestrian crashes. *Transportation Research Record, 2237*, 107–116.

Montella, A., Aria, M., D'Ambrosio, A., & Mauriello, F. (2012). Analysis of powered two-wheeler crashes in Italy by classification trees and rules discovery. *Accident Analysis and Prevention, 49*, 58–72.

Mujalli, R. O., & de Oña, J. (2013). Injury severity models for motorized vehicle accidents: A review. *Proceedings of the Institution of Civil Engineering – Transport*. . http://dx.doi.org/10.1680/tran.11.00026.

Mujalli, R. O., & de Oña, J. (2011). A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks. *Journal of Safety Research, 42*, 317–326.

Obeng, K. (2011). Gender differences in injury severity risks in crashes at signalized intersections. *Accident Analysis and Prevention, 43*(4), 1521–1531.

Pakgohar, A., Tabrizi, R. S., Khalilli, M., & Esmaeili, A. (2010). The role of human factor in incidence and severity of road crashes based on the CART and LR regression: A data mining approach. *Procedia Computer Science, 3*, 764–769.

Pande, A., & Abdel-Aty, M. (2009). Market basket analysis of crash data from large jurisdictions and its potential as a decision supporting tool. *Safety Science, 47*, 145–154.

Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning, 1*(1), 81–106.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Mateo, California: Morgan Kaufmann Publishers.

Savolainen, P., Mannering, F., Lord, D., & Quddus, M. (2011). The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. *Accident Analysis and Prevention, 43*, 1666–1676.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379–423. and 623–656.

Ulfarsson, G. F., & Mannering, F. L. (2004). Differences in male and female injury severities in sport-utility vehicle, minivan, pickup and passenger car accidents. *Accident Analysis and Prevention, 36*, 135–147.

Webb, G. I. (2007). Discovering significant patterns. *Machine Learning, 68*, 1–33.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (2nd ed.). San Francisco, CA: Morgan Kaufmann.

Xie, Y., Zhang, Y., & Liang, F. (2009). Crash injury severity analysis using Bayesian ordered probit models. *Journal of Transportation Engineering ASCE, 135*(1), 18–25.

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 53 (2012) 106 – 114

Procedia
Social and Behavioral Sciences

SIIV - 5th International Congress - Sustainability of Road Infrastructures

Using Decision Trees to extract Decision Rules from Police Reports on Road Accidents

López Griselda[a]*, de Oña Juan[b] and Abellán Joaquín[c]

[a] *Ph. D. Student. TRYSE Research Group. Department of Civil Engineering, University of Granada, ETSI Caminos, Canales y Puertos, c/Severo Ochoa s/n, 18071, Granada, Spain*

[a] *Ph. D. TRYSE Research Group. Department of Civil Engineering, University of Granada, ETSI Caminos, Canales y Puertos, c/ Severo Ochoa, s/n, 18071 Granada (Spain)*

[c] *Ph. D. Department of Computer Science & Artificial Intelligence, University of Granada, ETSI Informática, c/Periodista Daniel Saucedo Aranda s/n, 18071, Granada, Spain*

Abstract

The World Health Organization (WHO) considers that traffic accidents are major public health problem worldwide, for this reason safety managers try to identify the main factors affecting the severity as consequence of road accidents. In order to identify these factors, in this paper, Data Mining (DM) techniques such as Decision Trees (DTs), have been used. A dataset of traffic accidents on rural roads in the province of Granada (Spain) have been analyzed.
DTs allow certain decision rules to be extracted. These rules could be used in future road safety campaigns and would enable managers to implement certain priority actions.

Keywords: Type your keywords here, separated by semicolons ;

1. Introduction

The World Health Organization (WHO) considers that traffic accidents are major public health problem worldwide that every year claiming 1.27 million annual deaths and between 20 and 50 million injuries [1]. For this reason, many safety researchers have attempted to identify affecting the severity as consequence of road accidents. Different techniques, such as: Regression-type generalized linear models, Logit/Probit models, ordered Logit/Probit models have been used to achieve these objectives [2, 3, 4]. However most of these models have their own model assumptions and pre-defined underlying relationships between dependent and independent

* Corresponding author. Tel.: +34-958-249-450; fax: +34-958-246-138.
E-mail address: griselda@ugr.es

1877-0428 © 2012 The Authors. Published by Elsevier Ltd. Selection and/or peer-review under responsibility of SIIV2012 Scientific Committee
doi:10.1016/j.sbspro.2012.09.864

variables [5]; if these assumptions are violated, the model could lead to erroneous estimations, for example of the likelihood of severity accident.

To solve these limitations other method such as, Classification and Regression Trees (CART), have been used in the field of Road Safety. Kuhnert et al. [6] compared the results obtained with logistic regression, CART, multivariate adaptive regression splines (MARS) in the analysis of study of injuries resulting from motor vehicle accidents. The findings indicated that non-parametric techniques (CART and MARS) could provide more informative and attractive models whose individual components can be displayed graphically. Chang and Wang [5] employed CART to study the relationships between crash severity with characteristics related to drivers and vehicles, as well as variables related to roads, road accidents and the environment characteristics. They obtained that vehicle type was the most important affecting the severity of the accident. Recently, Pakgohar et al. [7] used CART and Multinomial Logistic Regression to study the role played by drivers' characteristics in the resulting crash severity. They found that the CART method provided more precise results, which are also simpler and easier to interpret. Kashani et al. [8] studied the most important factors that affect the injury severity of drivers involved in crashes on two-lane two-way rural roads. Subsequently, Kashani and Mohaymany [2] used CART to identify the main factors that affect the injury severity of vehicle occupants involved in crashes on those roads. And the results indicated that improper overtaking and not using a seatbelt was the most important factors associated with crash severity.

CART is particularly appropriate for studying traffic accident because is non-parametric techniques that do not require a priori probabilistic knowledge about the phenomena under studying and consider conditional interactions among input data [9].

Moreover, CART method allow certain decision rules of the "if-then" type to be extracted [8], and these rules can be used to discover behaviours that occur within a particular set of data. So, the aim of this work is to use CART method to identify the main factors that affect of the traffic injury severity and to extract certain decision rules which could be used in future road safety campaigns.

The paper is organized in four mayor sections. Section 2 presents an introduction to the main concepts of CART method, Decision Rules and the database used in the analysis. Section 3 presents the results and discussion. And, finally, section 4 presents the main conclusions of the study.

2. Materials and Methods

2.1. CART

A decision tree (DT) could be defined as a predictive model which can be used to represent both classifiers and regression models (depending on the nature of the variable class). When the value of the target variable is discrete, a regression trees is developed, whereas a regression trees is developed for the continuous target variable. CART method is a particularly type of DTs which allow developed either type of tree. In this wok a classification tree is developed because target variable (injurity severity) is discrete (slight injured -SI; killed or seriously injured –KSI).

A DT is a simple structure formed by number finite of “nodes” (which represent an attribute variable) connected by “branches” (which represents one of the states of the one variable) and finally, “terminal nodes or leafs” which specify the expected value of the variable class or target variable. The principle behind tree growing is to recursively partition the target variable to maximize “purity” in the child node. DTs are built recursively, following a descending strategy. The root node (which contained all of the data), is divide by two branchs (because the CART model generates binary trees) on the basis of an independent variable (splitter) that creates the best homogeneity. Each branch connected with a child node, the data in each child node are more homogenous than those in the upper parent node. Then, each child node is split recursively until all of them are pure (when all the cases are of the same class) or their "purity" cannot be increased. That is how the tree's terminal nodes are formed, which are obtained according to the answer values of the variable class.

There are different splitting criteria, however in the CART system the most commonly applied splitting criteria is the Gini index (GI); it could be defined for node c, as:

$$\text{gini}(c) = 1 - \sum_{j} p^2(j|c) \quad (1)$$

With: $p(j|c) = p(j,c) \Big/ p(c)$, $p(j,c) = \pi(j) N_j(c) \Big/ N_j$ and $p(c) = \sum_j p(j,c)$. Where: j – number of target variable or classes; $\pi(j)$ - prior probability for class j; $p(j|c)$ - conditional probability of a case being in class j provided that is in node m, $N_j(c)$ - number of cases of class j of node m, N_j - number of cases of class j in the roof node.

GI is one measure the degree of purity of the node, so when GI is equal to cero, the node is pure (all the cases in the node have the same class). When CART is development the aim is to achieve the maximum purity in the nodes, so the best split is the one that minimizes GI. Following this procedure the maximal tree that overfits the data is created. To decrease its complexity, the tree is pruned using a cost-complexity measure that combines the precision criteria as opposed to complexity in the number of nodes and processing speed, searching for the tree that obtains the lowest value for this parameter. At great length description of the CART method could be found in Breiman [10].

Following de Oña et al., [11], the goodness of a classification method is evaluated by accuracy. Accuracy is the percentage of cases correctly classified by the classifier of the method, and it is defined by following equation:

$$\text{accuracy} = \frac{TSI + TKSI}{TSI + TKSI + FSI + FKSI}.100\% \quad (2)$$

Where, TSI- Number of cases of SI; TKSI- Number of cases of KSI; FSI- Number of false cases of SI (i.e. incorrectly classified as SI); FKSI- Number of false cases of KSI (i.e. incorrectly classified as KSI).

On the other hand, one of the most valuable outcome provided by CART analysis is the value of the importance of independent variables that intervene in the model, which shows the impact of such predictor variables on the model.

2.2. Decision Rules

Decision Rules (DRs) could be obtained from the DT's structure. DRs are important because could be used to extract the potentially useful information from the data. The rules have the form of logic conditional: if "A" then "B", where "A" is the antecedent (a state or a set of statuses of one or several variables) and "B" is the consequent (one status of the variable class).

So, the conditioned structure (IF) of DR, begins in root node. Each variable that intervenes in tree division makes an IF of the rule, which ends in child nodes with a value of THEN, which is associated with the class resulting (the status of the variable class that shows the highest number of cases in the terminal node) from the child node. A priori, as same number of rules can be identified as the number of terminal nodes on the tree.

However, 2 parameters (population -Po; class probability -P) were used in order to extract important rules that could provide useful information for the implementation of road safety strategies in the future. The parameters that have been used could be defined as: population (Po), is the percentage of cases of a node in relation to the total number of cases analysed; and class probability (P), is the percentage of cases for the resulting class. The minimum values used so the selected rules will be representative are: Po≥1% and P≥60%.

2.3. Data

In this work, traffic accident data for rural highways for the province of Granada (South of Spain) have been used. These data have been obtained from Spanish General Traffic Accident Directorate (DGT). The period of the study is 5 years (2004-2008), and only data for 1 vehicle involved were used for this analysis. The total number of accident's records used is 1,801.

Considering that the main objective of this study is to identify the principal factors that affect the severity of traffic accidents, 17 explanatory variables were used based on De Oña et al. [11], and as a class variable, the injury severity level was considered with two classes (SI or KSI).

The data included variables describing the conditions that contributed to the accident and injury severity (see Table 1): characteristics of the accidents (month, time, day type, number of injuries, number of occupants, accident type and cause); weather information (atmospherics factors and lighting); driver characteristics (age and gender); and road characteristics (pavement width, lane width, shoulder width, paved shoulder, road markings and sight distance).

Table 1. Explanatory variables description

VARIABLE (CODE)	DESCRIPTION (CODE)	KSI	SI cases
Accident type (ACT)	Fixed objets collision (CO)	4	13
	Collision with pedrastian (CP)	92	46
	Other (OT)	11	24
	Rollover (RO)	45	73
	Run off road (ROR)	720	773
Age (AGE)	≤ 20	104	116
	(20-27]	231	231
	(27-60]	466	500
	>60	50	74
	Missing (MIS)	21	8
Atmospheric factors (ATF)	Good weather (GW)	769	787
	Heavy rain (HR)	66	94
	Light rain (LR)	14	24
	Other (O)	23	24
Cause (CAU)	Driver characteristics (DC)	760	730
	Combination of factors (COF)	94	148
	Other (OT)	6	16
	Road characteristics (RC)	4	21
	Vehicle characteristics (VC)	8	14
Day (DAY)	After holiday (AH)	137	150
	Before holiday (BH)	64	87
	Holiday (H)	274	278
	Working day (W)	397	414
Lane width (LAW)	< 3,25 m (THI)	263	232
	[3,25-3,75] m (MED)	592	673
	> 3, 75 m (WID)	17	24
Lighting (LIG)	Daylight (DAY)	426	531
	Dusk (DU)	48	57
	Insufficient (IL)	64	67
	Sufficient (SL)	29	43
	Without lighting (WL)	305	231
Month (MON)	Autumn (AUT)	199	225
	Spring (SPR)	210	243
	Summer (SUM)	238	254
	Winter (WIN)	225	207
Number of injuries	1 injury	584	670

(NOI)	> 1 injury	288	259
Occupants involved (OI)	1 occupant	569	597
	2 occupants	197	209
	> 2 occupants	106	123
Paved shoulder (PAS)	No (N)	156	152
	Non existent or impassable (NE)	277	287
	Yes (Y)	439	490
Pavement width (PAW)	[7-6] m (MED)	257	292
	< 6 m (THI)	141	118
	> 7 m (WID)	474	519
Road markings (ROM)	Does not exist or was deleted (DME)	81	89
	Separate margins of roadway (DMR)	92	86
	Separate lanes and define road margins	652	713
	Separate lanes only (SLO)	47	41
Gender (SEX)	Female (F)	104	171
	Male (M)	767	755
	Missing (MIS)	1	3
Shoulder type (SHT)	< 1,5 m (THI)	345	382
	[1,5-2,5] m (MED)	90	100
	Non existent or impassable (NE)	437	447
Sight distance (SID)	Atmosferic (ATM)	13	27
	Building (BU)	7	4
	Other (OT)	6	6
	Topological (TOP)	207	202
	Vegetation (VEG)	6	6
	Without restriction (WR)	633	684
Time (TIM)	[0-6] h	187	173
	(6-12] h	156	222
	(12-18] h	256	229
	(18-24] h	273	305

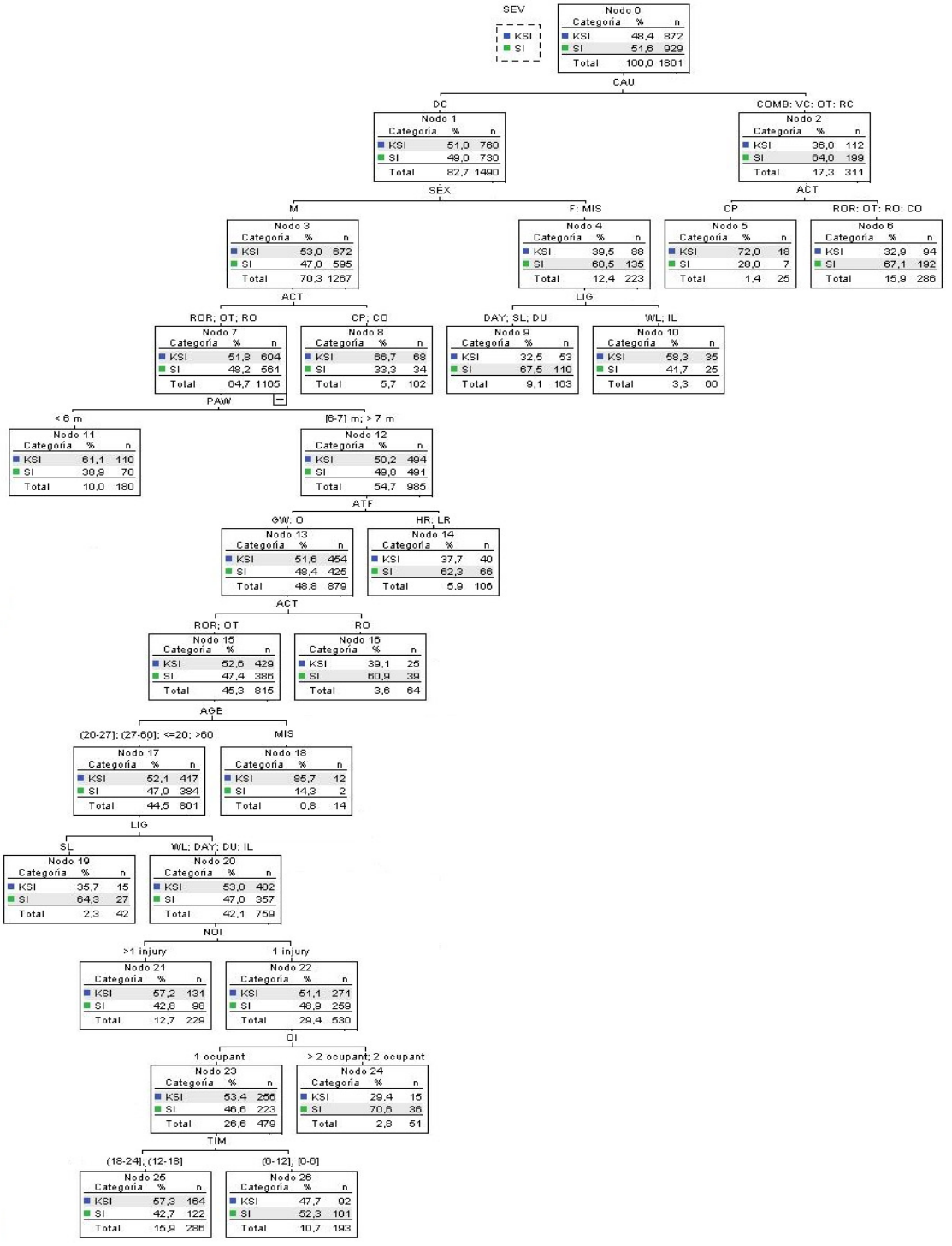
3. Results

The accuracy obtained for CART method was 54.43%. This value is within the range of values obtained in other studies in which classification methods with similar objectives [12; 11].

DT obtained contains 27 nodes (14 of them are terminal nodes). The identifier number, total number of accidents present in that node, and node classification based on the 2 categories (SI and KSI) are indicated for each node. Figure 1 shows the DT built and the interpretation is given below.

A root node is variable CAU, which is divided into two child nodes (node 1 and 2, see Figure 1). Node 2 shows accidents which not due to the driver, and depending on type, nodes 5 and 6 are obtained, with varying degrees of severity in collision with pedestrians the resulting severity is KSI with a probability of 72%, while for other accident types the severity is SI with a probability of 67.1%. Node 1 shows data related to accidents which are due to driver. This node is divided by gender variable. So, if the driver is a woman, nodes 10 and 9 are obtained, depending on road lighting: if the lighting is insufficient or without lighting, the accident is KSI with a probability of 58.33%; whereas if the road is sufficiently lit, it is broad daylight or dusk, the severity is SI in 67.48%.

However, most of the tree is generated by male driver (node 3). This node is splitted according to the ACT variable. For accidents involving pedestrians (with or without obstacles), accident severity is KSI. Whereas, for all other accident types, the tree splits by the variable lane width. In lanes narrower than 6m (narrow lanes), accidents are KSI in 61.11% of cases. In lanes wider than 6 m, severity with light or heavy rain is SI with a probability of 62.26%, depending on ATF variable.



Figue 1. Classification tree.

When atmospheric factors of good weather and others, the tree continues to grow according to ACT variable. When ACT is rollover, the accident is SI with a probability of 60.94%, and if it is run off the road or another type, the tree is divided according to the age of the driver involved in the accident. For most of the age groups (4 of the 5 analysed), the variable lighting causes the tree to grow, so accidents are SI if the road lighting is sufficient, whereas severity in the other cases is related to the number of injured.

If there is more than one injured person involved, the accident has a 57% probability of being KSI (node 21), while if there is only one injured person, the severity will also depend on the number of occupants, so that if the number is equal or more than two, the accident will be SI. If there is only one occupant, depending on the time of day, the 2 last nodes of the decision tree are obtained: from and from 12-24 h accident severity is KSI (node 25) and from 0-12 h accidents with SI (node 26).

3.1. Variable importance

The CART modelling process has an important phase in which the variables that are of key importance in the prediction of the dependent variable are identified. This is achieved by using the importance index [2].

Using this index, thirteen variables were detected as having the greatest influence on accident severity (see Table 2). Accident type is the most important variable, coinciding with previous studies [13, 11, 9]. The next important variable has been causes of the accident with a 57,6% importance, result that is coherent with other studies [14, 2], who situate crash cause among the top variables influencing severity. The variable lighting has 42.3% importance in the model. Lighting conditions were also highlighted in studies by Abel-Aty [15], Gray et al. [16], Heali et al [17], De Oña et al. [11] and Montella et al [18]. And gender variable has 34.9%. The other variables in the model are less important, with percentages of 26.9% to 4.5%.

Table 2. Importance of the variables.

ACT	CAU	LIG	SEX	OI	TIM	ATF	PAW	AGE
100%	57.6%	42.3%	34.9%	26.9%	23.5%	20.5%	18.2%	17.8%

3.2. Decision Rules

DT obtained has 14 terminals nodes, 9 of them has been identified as DRs. Table 3 shows a description of the DRs obtained which have been ordered by number of the node . As it could be observed, most of the rules are SI rules (6 of 9), however 3 important KSI rules have been identified.

About parameter analyzed, all the rules include at least 1% of the population, having rule 6 a percent of 16% of the population. Probability parameter, it could de remark that probability values are higher than 60%, with 70.59% being the highest value (rule 24).

About the length of one rule it could be said that less numbers of variables involved in the rule to imply the higher its predictive capacity of the rule. In DRs analyzed, rule length varies from 2 variables (as in node 5) to a maximum of 11 variables (as in node 26), so DRs obtained are enough informative

Seeing Table 3, it could be remark that:

- Two of three KSI rules have a male drivers involved.
- When there are pedestrians involved in accident, the probability of KSI increases: two out of three accidents involving pedestrians and male drivers will be KSI (rule 8).
- In general, accidents due to causes not attributable to the driver tend to have minor consequences, SI accident (rule 6). With a probability of 67.1% , this rule representing almost 16% of the total population.
- Also, a higher probability of KSI for accidents that were not collisions caused by male drivers on roads with a pavement width of less than 6m are identified.
- About DRs obtained with variable LIG, when women drivers cause a crash, CART methods predict SI accident if lighting exists (full daylight, sufficient lighting and dusk) (node 9). However, when the lighting is

non-existent or insufficient (node 10) the rule obtained is KSI accident. This rule is not observed for men and may indicate that women increase their risk of severity under conditions of less lighting on the road.

Table 3. DRs obtained.

NODE	VARIABLES OF THE RULES: [IF (AND … AND)]	THEN	Po (%)	P (%)
5	IF [(CAU≠DC) AND (ACT=CP)]	KSI	1.39	72.00
6	IF (CAU ≠ DC) AND (ACT=ROR OR ACT=OT OR ACT = ROOR ACT = CO).	SI	15.88	67.13
8	IF (CAU=DC) AND (SEX=M) AND (ACT=CP OR ACT = CO).	KSI	5.66	66.67
9	IF (CAU =DC) AND (SEX ≠ M) AND (LIG ≠ WL AND LIG ≠ IL).	SI	9.05	67.48
11	IF (CAU =DC) AND (SEX = M) AND (ACT ≠ CP AND ACT ≠ CO) AND (PAW = THI).	KSI	9.99	61.11
14	IF (CAU =DC) AND (SEX = M) AND (ACT ≠ CP AND ACT≠ CO) AND (PAW ≠ THI) AND (ATF = LR OR ATF = HR).	SI	5.89	62.26
16	IF (CAU=DC) AND (SEX=M) AND (ACT≠CP AND ACT≠CO) AND (PAW≠THI) AND (ACT=RO).	SI	3.55	60.94
19	IF (CAU=DC) AND (SEX=M) AND (ACT≠CP AND ACT≠CO) AND (PAW ≠ THI) AND (ATF ≠ LR AND ATF ≠ HR) AND (ACT =ROR OR ACT=OT) AND (AGE ≠ UN) AND (LIG = SL).	SI	2.33	64.29
24	IF (CAU =DC) AND (SEX = M) AND (ACT ≠ CP AND ACT ≠ CO) AND (PAW ≠ THI) AND (ATF ≠ LR AND ATF ≠ HR) AND (ACT =ROR OR ACT=OT) AND (AGE ≠ UN) AND (LIG ≠ SL) AND (NOI ≠ [>1]) AND (OI = [>2] OR OI = [2]).	SI	2.83	70.59

4. Conclusions

CART method allows classification based on crash severity and provides an alternative to parametric models because of their ability to identify patterns based on data, without the need to establish a functional relationship between variables. In fact, CART analysis does not need to specify a functional form as ordinary statistical modelling techniques, such as regression models. In regression analysis if the model is misspecified, the estimated relationship between dependent variable and independent variables as well as model predictions will be erroneous. So, CART model has a number of benefits compared to other widely used parametric models.

One of the most important advantages of the CART model is that the outcomes of the analysis are easy to understand and perform due to the graphical nature of its results. Also, the CART analysis allows a great many explanatory variables and it can easily find the important variables of the model.

Moreover, CART has permitted certain potentially useful rules to be determined that can be used by road safety analysts and managers. DRs obtained have been classified based on their severity, so, firstly the safety analysts should focus on severe or mortal crashes and subsequently intervene in accidents whose results are slight injuries. The approach proposed in this work within each group will enable the actions to give priority on the basis of population and probability.

Analyzing DRs, certain overall conclusions from a road safety perspective could be remarked:

- Due to length of DRs obtained, it could be said that they are enough informative.
- The CART method enables to obtain the importance of the variables in the model. In this case, the most important variables are: accident type, cause of the accident and lighting.
- The structure of the tree is generated by variable “cause of accident”.
- Male drivers are the main cause of KSI crashes.
- Women drivers have more probability than men drivers from suffering KSI accident when the lighting is non-existent or insufficient.

However, DTs models are often unstable. They could suffer variations if different strategies such as stratified random sampling (with injury severity as the stratification variable) are applied for creating learning and testing datasets [5].

Finally, the main problem observed with CART is that only binary trees can be built. For this reason certain categories of splitting variables are grouped in some branches, increasing node support, but making impossible to

analyse the influence of a specific category on crash severity. For this reason it could be suitable to use other methods to built DTs which allow trees without binary restriction in the branches.

Acknowledgements

The authors are grateful to the Spanish General Directorate of Traffic (DGT) for providing the data necessary for this research. Griselda López wishes to express her acknowledgement to the regional ministry of Economy, Innovation and Science of the regional government of Andalusia (Spain) for their scholarship to train teachers and researchers in Deficit Areas, which has made this work possible.

References

[1] WHO, World Health Organization, (2009). Informe Global sobre el estado de la Seguridad Vial: Tiempo para la Acción. Available at: www.who.int/violence_injury_prevention/road_safety_status/2009
[2] Kashani, A. and Mohaymany, A. (2011). "Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models". Safety Science 49, pp.1314-1320.
[3] Savolainen, P., Mannering, F., Lord, D., Quddus, M., (2011). The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. Accident Analysis and Prevention. In press
[4] Mujalli, R.O., De Oña, J. (in press). Injury Severity Models for Motorized Vehicle Accidents: A review, Proceedings of the Institution of Civil Engineering - Transport. In Press.
[5] Chang, L.Y. and Wang, H.W. (2006). "Analysis of traffic injury severity: an application of non-parametric classification tree techniques". Accident Analysis and Prevention 38, pp. 1019–1027.
[6] Kuhnert, P.M., Do, K.A., McClure, R., (2000). Combining non-parametric models with logistic regression: an application to motor vehicle injury data. Computational Statistics & Data Analysis, Vol 34(3), 371-386.
[7] Pakgohar, A., Tabrizi, R.S., Khalilli, M., Esmaeili, A., (2010). The role of human factor in incidence and severity of road crashes based on the CART and LR regression: a data mining approach. Procedia Computer Science 3, 764-769.
[8] Kashani, A., Mohaymany, A., Ranjbari, A., (2011). A Data Mining Approach to Identify Key Factors of Traffic Injury Severity. Promet-Traffic & Transportation, 23 (1), 11-17.
[9] Montella, A., Aria, M., D´Ambrosio, A.,Mauriello, F., 2011. Analysis of powered two-wheeler crashes in Italy by classification trees and rules discovery. Accident Analysis and Prevention, in press.
[10] Breiman, L., Friedman, J.,Olshen, R., and Stone, C., (1984). "Classification and Regression Trees". Belmont, CA: Chapman & Hall.
[11] De Oña, J., Mujalli, R.O., Calvo, F.J., (2011). "Analysis of traffic accident injury on Spanish rural highways using Bayesian networks". Accident Analysis and Prevention 43, pp. 402–411.
[12] Abdel Wahab, H.T., Abdel-Aty, M.A., (2001). Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. Transportation Research Record 1746, 6–13.
[13] Kockelman, K.M., Kweon, Y.J., (2002). Driver injury severity: an application of ordered probit models. Accident Analysis and Prevention 34, 313–321.
[14] Al-Ghamdi, A., (2002). Using logistic regression to estimate the influence of accident factors on accident severity. Accident Analysis and Prevention 34 (6), 729–741.
[15] Abdel-Aty, M., (2003). "Analysis of driver injury severity levels at multiple locations using ordered probit models". Journal of Safety Research 34, 597–603.
[16] Gray, R.C., Quddus, M.A., Evans, A., (2008). "Injury severity analysis of accidents involving young male drivers in Great Britain". Journal of Safety Research 39, 483–495.
[17] Helai, H., Chor, C.H., Haque, M.M., (2008). "Severity of driver injury and vehicle damage in traffic crashes at intersections: a Bayesian hierarchical analysis". Accident Analysis and Prevention 40, 45–54.
[18] Montella, A., Aria, M., D´Ambrosio, A., Mauriello, F., (2011). Data Mining Techniques for Exploratory Analysis of Pedestrian Crashes. Transportation Research Record: Journal of Transportation Research Board, No. 2237, Transportation Research Board of the National Academies, Washington, D.C., 107-116.