UNIVERSIDAD DE GRANADA

Departamento de Ciencias de la Computación e Inteligencia Artificial

Tesis Doctoral

Análisis del Sistema Educativo de la Universidad Técnica Estatal de Quevedo Mediante Ciencia de Datos

Memoria presentada por:

Jorge Humberto Guanin Fajardo

Director:

Jorge Casillas

Catedrático

Para optar por el título de Doctor por la Universidad de Granada dentro del *Programa de Doctorado en Tecnologías de la Información y la Comunicación*

SEPTIEMBRE, 2024

Editor: Universidad de Granada. Tesis Doctorales
Autor: Jorge Humberto Guanin Fajardo
ISBN: 978-84-1195-652-9
URI: https://hdl.handle.net/10481/102011

"Porque Jehova es bueno; para siempre es su misericordia. Y su verdad por todas las generaciones."
Salmos 100:5, RV60

A mi Esposa e hijos.
Magdiel Aleshka y Jorge Josías.

A mis Padres.
Jorge y Matilde.

Declaración

La memoria titulada **"Análisis del Sistema Educativo de la Universidad Técnica Estatal de Quevedo Mediante Ciencia de Datos"** que presenta Don Jorge Humberto Guanin Fajardo para optar al grado de doctor en Informática, ha sido realizada en el Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada, bajo la dirección del Doctor D.Jorge Casillas Barranquero, miembro del mismo Departamento.
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, 10 de septiembre de 2024.

Fdo. D. Jorge H. Guanin Fajardo, MsC. Fdo. D. Jorge Casillas Barranquero, PhD.

Agradecimientos

El trabajo de investigación tiende a tener momentos de frustración y dificultad. Son muchas las personas que han sido partícipes de este proyecto, la confianza que me depositó desde su inicio el Ing. Manuel Agustín Haz Álvarez(+) Rector de la Universidad Técnica Estatal de Quevedo así como demás autoridades y colegas.

Ante todo quiero agradecer a mi director por la aceptación y conducción del desarrollo de ésta tesis doctoral.

A mi Esposa *Jenny* por acompañarme en la trayectoria profesional y sobre todo familiar. A mis hijos que dentro de los plazos y planificación de esta tesis no fueron considerados y, que gustosamente estas preciosas criaturas nos acompañan, *Magdiel y Josías*. Sus mentes inquietas siempre me han reclamado tiempo y atención por su tierna edad.

—¡Papi!, ya deja esa computadora, vamos hacer un paseito.

—¡Papi!, vamos a jugar al parque.

A mis Padres, *Jorge y Matilde* agradecerles por apostar y motivarme en mi formación universitaria. De igual manera a mis hermanos y sobrinos por apoyarme y creer siempre que lo conseguiría.

En particular, extiendo mi agradecimiento y de manera muy especial a Rafael Garzón, Miguel Ángel, Oriol, Jaime, Alberto, Carlos, D. Francisco y D. Javier grandes amigos y precursores de mi bienestar familiar. ¡Gracias por secundarnos!.

En la vida de todo ser humano desde la guarde hasta la universidad siempre están formando parte de nuestras vidas. Conocen mucho y poco o mucho de ti, pero al final están allí para estrechar tu mano o darte un abrazo, sin descartar los enfados, el sacarte de apuros o sustos. Son varios amigos y conocidos emergidos desde el CITIC-UGR, aunque me olvidaré nombres y lo siento en verdad, pero no significa que sean menos importantes. Quiero mencionar una breve lista: Sergio, Julín, Pavel, Amilkar, María, Andrés (Cuba); Carlos, Marvin (Nicaragüa); Virgilio, Raúl (México); Luz Marina, Jorge y Angélica (Colombia); Víctor (Bolivia); Gleiston, Byron, Pablo, Efraín, Adolfo, Jesús P., Richard, Jhonny, Wacho, Stephanie (Ecuador); Paco, Sara, David, Juanjo (España); Sara (Italia), entre otros. Gracias por vuestro ánimo.

Otras personas, fuera del núcleo de estudio han contribuido de manera indirecta o directa, por ello, mis agradecimiento de hazañas y desafíos a Puri, Catia y Trini. Al mismo tiempo, expreso mi profunda gratitud a Susana Barrera y Familia, además de la Familia Villegas-Barranco por todo el apoyo y confianza. También a Pepita y Aránzazu por respaldarme. Salvador V., Miguel Ángel M. y Juan M., tres grandes personas que me permitieron conocerlas y quererlas a pesar de su actual ausencia. Finalmente, a mis amigos de carrera que han estado presente en ésta travesía y que no son menos importante, al contrario, han colaborado de todas las formas posibles: Alex, Sara y Stalin.

Muchísimas gracias a todos. Éste, también es vuestro trabajo.

Resumen

La educación desempeña un papel fundamental en el progreso de las sociedades y economías modernas. Con este planteamiento, las instituciones de enseñanza superior se han convertido en depositarias de confianza de una educación de calidad. Su compromiso no sólo radica en la excelencia educativa, sino también en satisfacer la creciente demanda de formación significativa y oportuna. Los investigadores educativos han desempeñado un papel crucial en el desarrollo de herramientas tecnológicas, planes pedagógicos y programas de estudios innovadores. Entre estas herramientas destaca el uso de la inteligencia artificial (IA), que proporciona a los responsables académicos información valiosa para diseñar estrategias que fomenten el éxito universitario. La IA puede analizar datos, identificar patrones de aprendizaje y personalizar la experiencia educativa de cada estudiante de forma personalizada o colectiva en función de sus necesidades. La Inteligencia Artificial y los diversos enfoques marcan una nueva era en la educación moderna. Sin dejar de lado la calidad educativa que busca promover Naciones Unidas a través de los Objetivos de Desarrollo Sostenible (ODS), que busca garantizar una educación inclusiva, equitativa y de alta calidad que fomente oportunidades de aprendizaje para todos. Enmarcado en este contexto, el análisis del sistema educativo de la Universidad Técnica del Estado de Quevedo se desarrolló a través de la ciencia de datos y, para los fines de esta tesis, se aplicó la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) con sus respectivas fases para descubrir patrones o, a su vez, conocimiento oculto en los datos. Previo al análisis, se extrajo información de diferentes departamentos con el fin de limpiar y fusionar sus datos, es decir, aplicar ETL (Extracción, Transformación y Carga) para lograr una base de datos robusta y única. Así, como parte inicial de esta tesis, se realizó un estudio exploratorio de los datos para comprender su comportamiento y facilitar la identificación de información relevante, de ahí que las técnicas de visualización proporcionaran las pautas necesarias en cuanto a proyección, trayectoria y patrón en los datos. Posteriormente, se examinaron sus atributos mediante diversos algoritmos de selección de características e instancias que dieron paso al uso de algoritmos de balanceo de datos. A continuación, la modelización de los datos dio lugar a una base de conocimientos robusta y a otra más simplificada, que reveló el perfil esencial del

profesorado experimentado y maduro, tanto en la enseñanza como en los grupos de edad. Además, la pandemia COVID-19 tuvo un gran impacto en todos los aspectos de la vida universitaria, lo que puso de manifiesto la necesidad de planificar y formar al profesorado en herramientas educativas que mitiguen los riesgos y superen las dificultades de la docencia virtual. Por último, la utilidad de estos resultados a través de la ciencia de datos proporcionó sugerencias para la conveniencia de las partes interesadas de proporcionar una enseñanza significativa y orientar el éxito académico de los estudiantes.

Abstract

Education plays a fundamental role in the progress of modern societies and economies. With this approach, institutions of higher education have become trusted repositories of quality education. Their commitment lies not only in educational excellence, but also in meeting the growing demand for meaningful and timely training. Educational researchers have played a crucial role in the development of innovative technological tools, pedagogical plans and curricula. Chief among these tools is the use of artificial intelligence (AI), which provides academic leaders with valuable information to design strategies that foster college success. AI can analyze data, identify learning patterns and customize each student's educational experience on a personalized or collective basis according to their needs. Artificial Intelligence and diverse approaches mark a new era in modern education. Not to mention the educational quality that the United Nations seeks to promote through the Sustainable Development Goals (SDGs), which seeks to ensure inclusive, equitable and high quality education that promotes learning opportunities for all. Framed in this context, the analysis of the educational system of the Technical University of the State of Quevedo was developed through data science and, for the purposes of this thesis, the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology was applied with its respective phases to discover patterns or, in turn, hidden knowledge in the data. Prior to the analysis, information was extracted from different departments in order to clean and merge their data, i.e. apply ETL (Extraction, Transformation and Loading) to achieve a robust and unique database. Thus, as an initial part of this thesis, an exploratory study of the data was conducted to understand its behavior and facilitate the identification of relevant information, hence visualization techniques provided the necessary guidelines in terms of projection, trajectory and pattern in the data. Subsequently, their attributes were examined using various feature and instance selection algorithms that gave way to the use of data balancing algorithms. Then, modeling of the data resulted in a robust knowledge base and a more simplified one, revealing the essential profile of experienced and mature faculty, both in teaching and age groups. In addition, the COVID-19 pandemic had a major impact on all aspects of university life, highlighting the need to plan and train faculty in educational tools that mitigate the risks and overcome the

difficulties of virtual teaching. Finally, the usefulness of these results through data science provided suggestions for stakeholder convenience to provide meaningful instruction and to guide studentsácademic success.

Índice general

Índice de figuras

Parte I

Tesis Doctoral

1. Introducción

"La educación no es la acumulación de aprendizaje, información, datos, hechos o habilidades -eso es formación o instrucción- sino que es hacer visible lo que está oculto como una semilla."

Thomas More.

Todos los esfuerzos de la investigación en Inteligencia Artificial (IA) se han centrado en construir inteligencias artificiales especializadas y los éxitos alcanzados son muy impresionantes, en particular durante el último decenio gracias sobre todo a la sinergia de dos elementos: la disponibilidad de ingentes cantidades de datos y, el acceso a la computación de altas prestaciones. Dentro de este marco, el éxito de sistemas, como por ejemplo AlphaGo, Watson y los avances que destacan en vehículos autónomos o en diagnóstico médico basado en imágenes, hacen posible el análisis y detección de patrones de forma eficaz (Kraft y Moloney, 2020; Mántaras, 2020). Es por ello, que la IA es el término más amplio aplicado a cualquier técnica que permite a los ordenadores imitar la inteligencia humana utilizando para este propósito la lógica (boleana, predicados, difusa), las reglas "si-entonces" $(X \rightarrow Y)$, reglas de predicados, los árboles de decisión y el aprendizaje automático (incluido el aprendizaje profundo). Por esta razón, el KDD es el proceso organizado de identificación de patrones válidos, novedosos, útiles y comprensibles a partir de grandes y complejos conjuntos de datos. El enfoque holístico y su capacidad para abordar problemas complejos de manera integral está radicado en la ciencia de datos (CD) que es una rama de la inteligencia artificial que implica la inferencia de algoritmos donde se exploran datos, se desarrolla el modelo y descubre patrones previamente desconocidos.

Por ello, el progresivo y creciente interés de la comunidad científica por la ciencia de datos como herramienta para descubrir el conocimiento oculto en los datos ha logrado resolver problemas de manera eficaz. En síntesis, el modelo que se extrae a partir de los datos aplicando técnicas de ciencia de datos comúnmente se utiliza para entender, analizar y predecir en cierta medida fenómenos que han ocurrido y que se necesita preveer o encontrar un patrón de comportamiento (Ghatak, 2017; Maimon y Rokach, 2010).

1.1. Marco de trabajo y motivación

A día de hoy, la ciencia de datos como una rama de la IA es una técnica fiable para solucionar problemas complejos, donde la intervención de un experto humano es mínima pero fundamental para dar validez a los hallazgos. Así, Fayyad et al. (1996) planteó una metodología de trabajo usada para la solución de problemas en ciencia de datos mostrada en la Figura 1.1. De manera general, existen aplicaciones informáticas que están disponibles en la web para llevar a cabo cada una de las fases del KDD, dichas aplicaciones pueden ser de uso comercial o libre (opensource) entre ellas, citamos por ejemplo: R [1],Python [2], Orange [3], Weka [4], KNIME [5], etc. convergiendo con el objetivo de analizar y extraer el conocimiento oculto en los datos, ya sea en forma de patrones o reglas. En la práctica, existen dos tipos de modelos para llevar a cabo ésta práctica: *descriptivo* y *predictivo*.

Modelo descriptivo El modelo descriptivo versa su búsqueda del conocimiento desconociendo la variable objetivo del problema, es decir, se desconoce lo que se desea encontrar apriori. La agrupación de datos (clustering) y las reglas de asociación son tareas comunes vinculadas al modelado descriptivo de los datos.

- **Agrupamiento.** El agrupamiento de datos consiste en la búsqueda o formación de grupo de datos similares dentro de los datos cuando no hay información de la clase que se intenta predecir. Dichos grupos incluyen subgrupos de ejemplos que comparten propiedades o similitudes. Esta técnica en la actualidad cuenta con un sinnúmero de métodos que se emplean para la búsqueda de datos que se aproximen a un mismo grupo. La pertenencia a un mismo grupo de datos debe ser muy similar, mientras que cualquier grupo distinto deben ser diferenciado al máximo posible (Gironés Roig et al., 2017; Manzanares et al., 2019).
- **Reglas de asociación.** En ocasiones el tipo de problema que se desea resolver no siempre está acompañado de una variable que determine el estado de un registro. Es decir, se desconoce la salida de un registro o transacción, o bien, que el problema no cuenta con una variable dependiente. En el trabajo de Agrawal y Srikant (1994); Shi y Zhu (2022), fue presentado el famoso algoritmo de reglas de asociación *Apriori* y, desde entonces, muchas investigaciones se han dedicado al estudio profundo del algoritmo.

[1] https://cran.r-project.org/
[2] https://www.python.org/
[3] https://orangedatamining.com/
[4] https://ml.cms.waikato.ac.nz/weka/
[5] https://www.knime.com/knime-analytics-platform

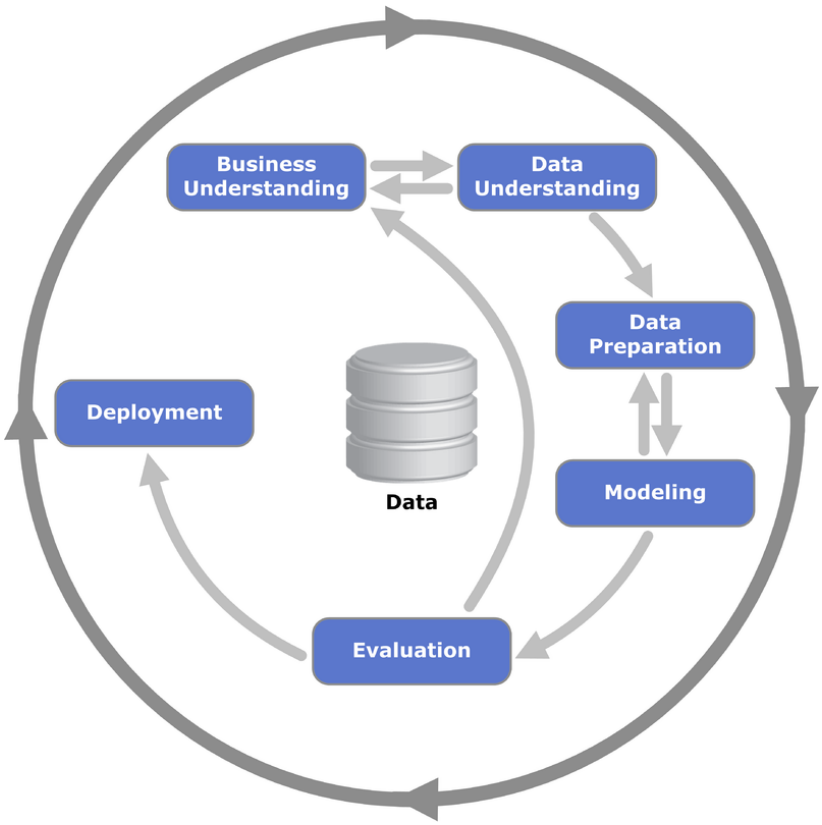


Figura 1.1 Ciclo de vida del modelo CRISP-DM.
Fuente: Wikimedia Commons

Modelo predictivo A diferencia de los modelos descriptivos, en este tipo de modelado la predicción es la estimación de un determinado valor o, a su vez, la distribución de valores de un conjunto de atributos de relevancia para predecir uno similar previamente seleccionado. La evolución del modelado predictivo es diversa y existe una cantidad ingente de investigaciones y mejoras a los algoritmos propuestos para predecir valores.

- **Regresión:** La regresión es uno de los modelos predictivos de mayor uso. Dado que muchos problemas son abarcados a través de esta técnica. El efecto de los avances es especialmente profundo en temas de modelado según el objetivo que se persiga mapeando el espacio de entrada en un dominio de valores reales. Por ejemplo, un regresor o variable objetivo que puede predecir la demanda de un determinado producto o bien dadas sus características (Oded y Rokach, 2005).
- **Clasificación supervisada:** Los métodos de clasificación supervisada son métodos que intentan descubrir la relación entre los atributos de entrada (a veces llamados variables independientes) y un atributo objetivo (a veces denominado variable dependiente).
- **Clasificación flexible.** Estos métodos explotan la capacidad de los ordenadores para buscar en grandes cantidades de datos de forma rápida y eficaz. Sin embargo, los datos a analizar son imprecisos y están afectados por la incertidumbre. En el caso de fuentes de datos heterogéneas, como el texto y el vídeo, los datos pueden ser además ambiguos y parcialmente contradictorios. Además, los patrones y las relaciones de interés suelen ser vagos y aproximadas. Por eso, para que el proceso de extracción de información sea más robusto o, digamos, métodos de búsqueda y aprendizaje similares a los humanos, se requiere tolerancia hacia la imprecisión, la incertidumbre y las excepciones. Por lo tanto, tienen capacidades de razonamiento aproximado y son capaces de manejar la verdad parcial. Las propiedades de este tipo son típicas de la computación blanda o también conocida como computación flexible (Martínez, 2012).

Dentro de este marco, el centro universitario se ha propuesto buscar modelos predictivos tanto del profesorado como del alumnado para que la parte interesada construya alternativas o estrategias que encaminen al alumnado hacia el éxito académico.

La estructura de ésta memoria se divide en dos partes. La primera, enfocada al planteamiento del problema y discusión de los resultados. La segunda, referida a las publicaciones asociadas al estudio. En la Parte I de la memoria comenzamos con una sección dedicada al planteamiento del problema. En la sección dos, la justificación de la tesis. En la sección tres, Los objetivos propuestos. En la sección cuatro, se trata sobre las discusiones de los artículos

publicados. En la sección cinco, las conclusiones finales y, por último, en la sección seis se trata sobre trabajos futuros.

1.2. Preliminares

Como parte del desarrollo de ésta memoria se usó dos áreas importantes de estudio. La primera, referida al análisis exploratorio de datos que permitió conocer un poco más a fondo los datos y de hecho el contexto universitario. La segunda, enfocada propiamente al estudio de ciencia de datos, en concreto, la aplicación de técnicas de predicción supervisada.

1.2.1. Análisis exploratorio de datos

El análisis exploratorio de datos busca de primera mano ofrecer una respuesta previa al análisis exhaustivo de los datos (estadística aplicada, aprendizaje profundo, aprendizaje automático, etc), para ello, es necesario que lo recibido (datos) goce de integridad. A menudo, gran parte de problemas reales lejos de su complejidad propia para resolverlos traen consigo imperfecciones, es decir, carecen de integridad. Los científicos de datos denominan a esta información como "datos crudos" (rawdata). Es importante tener presente que para dar una respuesta clara y precisa a cuestiones de esta tesis doctoral se realizan una serie de procesos para obtener datos idóneos. A modo de reflexión, se puede deducir que: la entrada de datos imprecisos y carente de calidad, tendrá una solución (salida) errónea o deficiente. Esto dará paso a conjeturas fuera de contexto o interpretación desafinada del resultado final. De manera general, el estudio exploratorio convencional sino es bien proyectado solo expresa la representación formal de los datos. En cambio, una apreciación destacada y significativa en la información revela datos claves para el estudio. A la luz de los resultados que se consigan en esta tesis doctoral, se plantea suministrar una descripción gráfica relevante y consolidada de características, patrones, tendencias y relación de los datos. En otras palabras, una buena visualización nos proporciona una instantánea concisa de los datos. Como suele decirse, "una imagen vale más que mil palabras" (Pathak, 2014). En cualquier caso, es importante considerar que cuando la etapa exploratoria es omitida no se tendrá un claro conocimiento del problema. Aplicar técnicas de aprendizaje de forma directa sólo conseguirá resultados relativamente precisos (en ciertos casos). Es decir, la sobre-estimación de variables por parte del algoritmo (supervisado/no supervisado) ocasionalmente provoca sesgos en los resultados. Por ésta razón, el estudio exploratorio, de una forma u otra carece de una respuesta rigurosa sobre la solución del problema, pero es una perspectiva de comprensión global y posible aproximación hacia los resultados esperados, con independencia del tipo

de técnica de aprendizaje aplicada. De ahí que, el estudio exploratorio resulte atractivo y útil cuando se combinan con métricas comúnmente aplicadas en la ciencia de datos para encontrar potenciales relaciones entre los datos.

1.2.2. Ciencia de datos en la educación superior

Estudios multidisciplinarios o dirigidos hacia la educación primaria, secundaria y superior son investigados desde diferentes perspectivas y que convergen en la aplicación de técnicas estadísticas para comprobar hipótesis de una nueva técnica de enseñanza o incursión tecnológica y otros métodos de investigación, pero, con especial atención hacia la ciencia de datos para examinar patrones de comportamiento, aplicar toma de decisiones y otros temas que ocupa cada área. De hecho, lo concerniente hacia la educación superior que es el enfoque que se da a ésta tesis doctoral, sobresalen trabajos relacionados al uso de nuevas tecnologías aplicadas al salón de clases, métodos de aprendizaje a través de las TIC's o sistemas de recomendación para el proceso de matrícula y sugerencia de créditos basados en el historial o logro académico del alumnado con características convergentes (Fernández-García et al., 2020; Guabassi et al., 2021; Masruroh et al., 2021; Sakr et al., 2021; Samin y Azim, 2019; Tavakoli et al., 2022; Verma y Anika, 2018), predicción del rendimiento académico (Albreiki et al., 2021; Alturki et al., 2022; Alyahyan y Düştegör, 2020; Chaka, 2022; de la Cruz-Campos et al., 2023; Dol y Jawandhiya, 2023; Gutierrez-Bucheli et al., 2022; Ijaz et al., 2020; López-Zambrano et al., 2021; Namoun y Alshanqiti, 2021; Nawang et al., 2021; Rahul y Katarya, 2024; Romero y Ventura, 2013, 2020; Saa et al., 2019; Soegoto et al., 2022; Xu et al., 2021; Zaffar et al., 2021), de manera análoga también se examinan sistemas de alerta temprana para detectar el abandono, deserción o fracaso de estudios (Aina et al., 2022; Burgos et al., 2018; Csalódi y Abonyi, 2021; de Oliveira et al., 2021; Jiménez-Macias et al., 2022; Kabathova y Drlik, 2021; Karalar et al., 2021; Pradeep y Thomas, 2015; Siri, 2015), predicción (Cui et al., 2019; Ho et al., 2021; Sghir et al., 2022; Sun et al., 2022), así como temáticas relacionadas a la irrupción de ChatGPT (Baidoo-Anu y Ansah, 2023; Wang et al., 2023), ChatBots (Adamopoulou y Moussiades, 2020; Okonkwo y Ade-Ibijola, 2021; Pérez et al., 2020) y otra perspectiva no trivial pero de importancia sobre educación en temas de habilidades y competencias educativas de la IA, Cloud Computing y la Agenda 2030 para la educación (Askari et al., 2021; Batanero et al., 2019; Costan et al., 2021; Du et al., 2021; Masruroh et al., 2021; Rahayu et al., 2022; Razia, 2023; Santana y Díaz-Fernández, 2023; Zlatic et al., 2021). Con este incentivo, se pretende distinguir sobre temas orientados al desarrollo de esta tesis doctoral, así como el planteamiento y consecución de los objetivos propuestos.

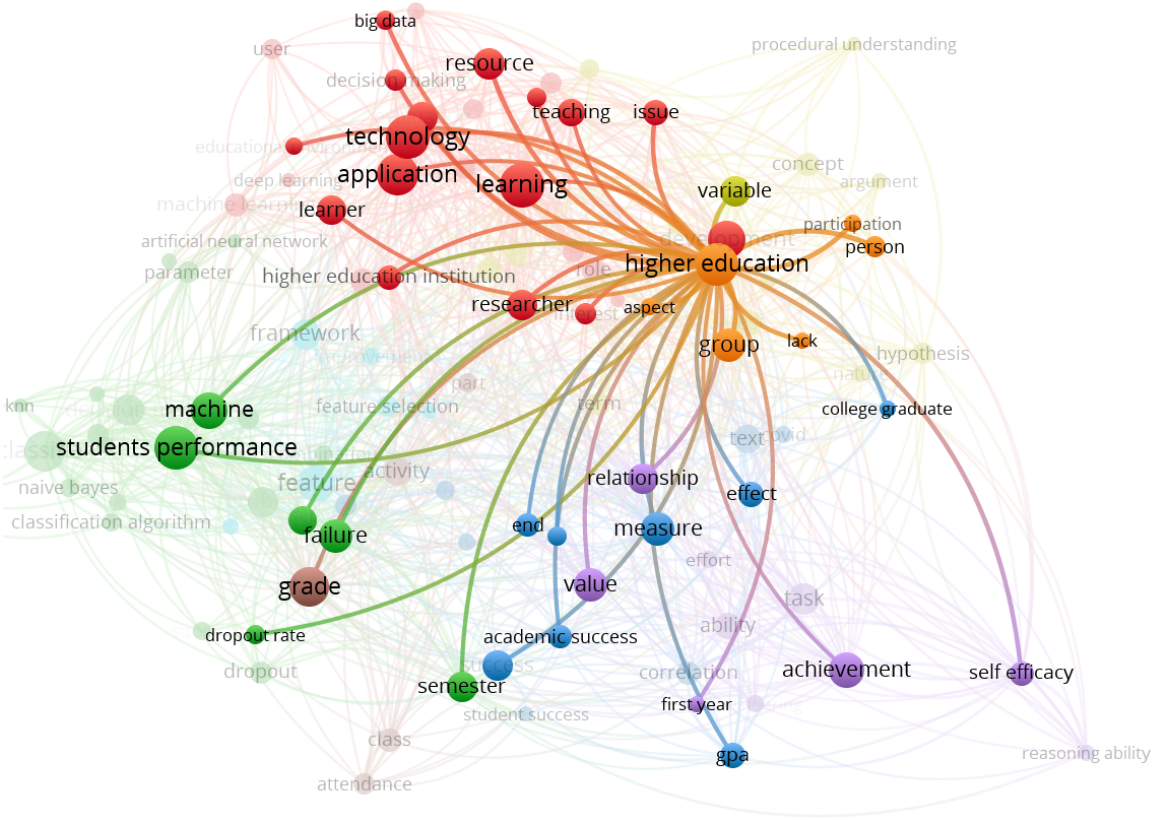


Figura 1.2 Nube de términos vinculados al tópico de educación superior y ciencia de datos, la figura proyectada posee 41 aristas y cinco grupos de datos (color del nodo). Los términos se encuentran según su grado de asociación directamente vinculado el nodo central de educación superior.

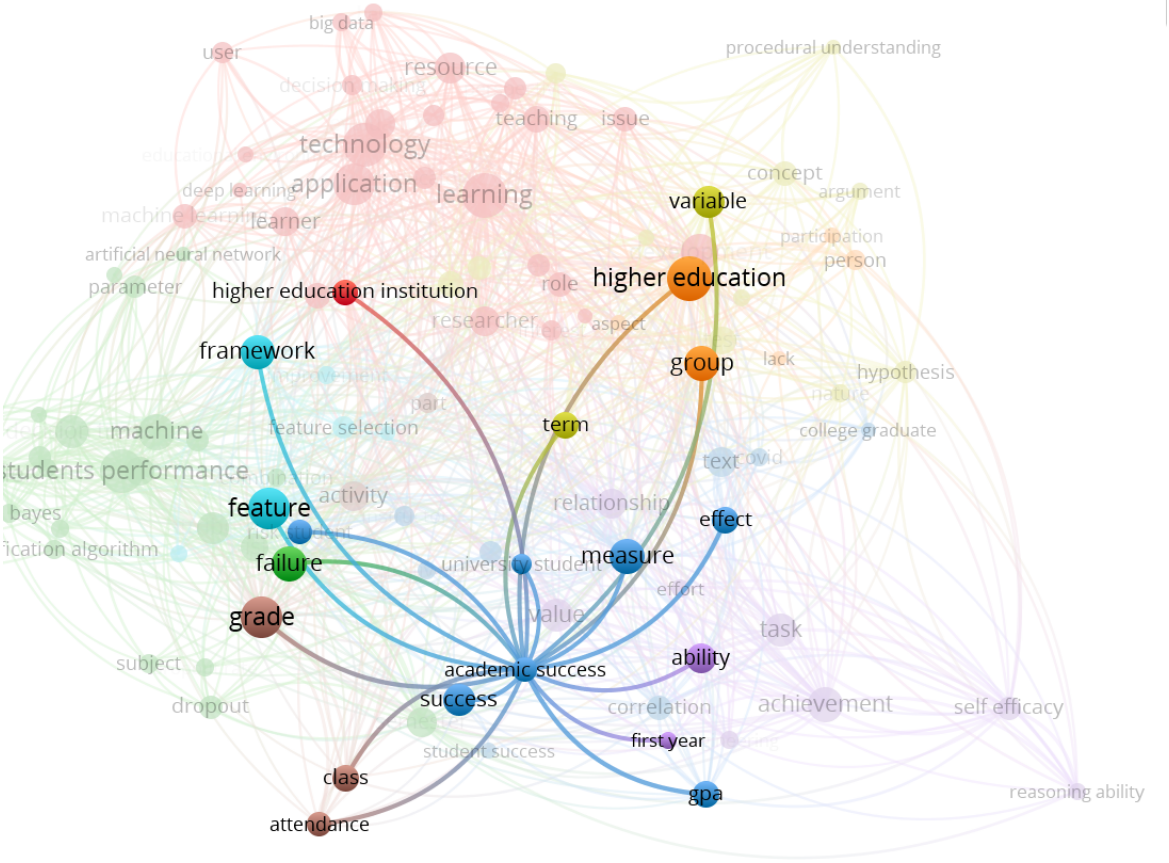


Figura 1.3 Nube de términos vinculados al tópico del éxito académico en la educación superior. En ésta figura se aprecian 19 nodos concentrados en siete grupo de datos (color del nodo).

Actualmente, la ingente cantidad de trabajo científico, divulgativo, ensayos, libros, entre otros tipos de documentos es extenso y, de manera especial todo lo relacionado a la IA. Es así, que la ciencia de datos como disciplina integra múltiples herramientas de inteligencia artificial, así como operaciones de analítica avanzada. Involucra áreas de conocimiento como matemáticas, estadística e incluso ciencias sociales y del comportamiento. Su objetivo general es analizar conjuntos de datos para descubrir conocimiento útil, como tendencias sociales o riesgos. Esta tendencia genera o mejora nuevas tecnologías según su enfoque y de manera especial en lo concerniente al encauzamiento del alumnado hacia el éxito académico. Existe, sin embargo, diversos trabajos en educación superior dirigidos en áreas de estudios tales como: psicología, minería de datos (Data Mining - DM), aprendizaje automático (Machine Learning - ML), Big Data, entre otros y, que se sintetizan en la red de nodos proyectados en la Figura 1.2 y Figura 1.3. La proyección de ambas figuras está basada sobre términos (palabras) de documentos electrónicos en formato "pdf" (Portable Document File) recolectados desde la base de datos Scopus y que posteriormente fueron mapeados (depuración de documentos). La Figura 1.2, se subdivide en cuatro grupos que se distinguen por colores rojo, azul, verde y naranja. Es así que la literatura relacionada con la educación superior "higher education" toma importancia ya que se resuelven problemas complejos que requieren eficacia. Si bien es cierto, el análisis de datos pueden dejar inquietudes por resolver, por ésta razón, existe una cantidad de términos importante para proyectar la red de nodos. En este mismo sentido, se proyecta la Figura 1.3 donde precisan términos asociados con el éxito académico del alumnado y que es el tópico de interés de esta tesis doctoral. De igual manera, existen cuatro grupos diferenciados, pero con menor números de términos. Aquí, la temática central es el éxito académico "academic success" con términos que se asocian al nodo principal.

La presente tesis aborda temas inherentes sobre: análisis y exploración de datos, preprocesado de datos, técnicas de clasificación supervisada y técnicas de visualización sobre datos reales. El denominador común en el que giran dichos temas son las técnicas de ciencia de datos empleadas para extraer información interesante y útil de los datos. La segunda parte del documento está constituida de publicaciones asociadas a esta tesis y organizadas según los objetivos propuestos.

2. Justificación

"Siempre parece imposible, hasta que se hace."
Nelson Mandela.

La toma de decisiones no siempre resulta oportuna cuando se carece de información relevante o experiencia suficiente por parte de los expertos o parte interesada (stakeholder). El centro universitario centra esfuerzos para conseguir resultados que garanticen la toma de decisiones. En este sentido, la justificación proviene del hecho que el centro universitario tiene necesidad de información relevante que refuerce la toma de decisiones a sus gestores académicos. Por esta razón, y, basados en estudios de ciencia de datos en entornos educativos, se busca información relevante que intervengan o condicionen el éxito académico de los alumnos aplicando la ciencia de datos ya que utiliza técnicas de aprendizaje automático para extraer conocimiento favorable a los objetivos estratégicos del centro universitario. Precisamos con ello, que lo significativo de todo esto, es incrementar la eficacia del sistema educativo de la Universidad Técnica Estatal de Quevedo, con el fin de sugerir información válida y contrastada que mejore la calidad de la gestión educativa.

Lo anteriormente expuesto estriba en el propósito de examinar la base de datos del sistema informático universitario. Para este propósito, la Unidad de Planeamiento Académico y la Oficina de Recursos Humanos disponen de datos tanto del alumnado y profesorado respectivamente. La consolidación de estos datos sirvieron como suministro de entrada para analizar la información del estudiantado que al término del primer año consiga el éxito académico en la titulación universitaria matriculado inicialmente y, por esta razón, se propuso el uso de la ciencia de datos para alcanzar los objetivos planteados en esta tesis doctoral.

En este sentido, es conveniente situar primero que la abstracción del éxito académico se perfila en dos partes: (i) baja intención de abandono y, (ii) alto rendimiento académico. Ambas debidamente sincronizadas responden positivamente al éxito académico del alumnado (Respondek et al., 2017). De ahí que, la armonización del rendimiento académico y la motivación del alumnado descansen significativamente en el profesorado. Esta relación, trasciende por la claridad percibida en la enseñanza del profesor, capacidad y apoyo, así como en la implicación del nivel de satisfacción y la experiencia universitaria (Livengood, 1992;

Pascarella et al., 1996). A este respecto, las Unidades Académicas (Facultades) del centro universitario han destacado que la población estudiantil matriculada en los grados universitarios alcanzan el éxito académico con un porcentaje inferior a dos quintiles desde su matrícula inicial. Esta inquietud ha generado en los expertos universitarios un amplio interés de buscar soluciones inmediatas y fiables. Por ello, se han formulado las siguientes preguntas que faciliten la búsqueda de soluciones:

- ¿Es aplicable el uso de las TIC's para determinar los factores que desvelan a estudiantes exitosos?
- ¿Cuál es el perfil apropiado del estudiantado para la titulación universitaria?
- ¿Cómo influye el profesorado en la titulación universitaria para que el estudiantado consiga el éxito académico?
- ¿Qué modelo de predicción se puede aplicar para conocer de manera anticipada el éxito académico del alumnado?

3. Objetivos

Después de presentar la introducción y justificación, ahora se presentan los principales objetivos que fueron precursores para el desarrollo de esta tesis. Entre ellos se incluye el análisis de técnicas visualización, preprocesado de datos y técnicas de ciencia de datos. El objetivo principal de esta tesis fue descubrir el conocimiento oculto de los datos educativos empleando técnicas de ciencia de datos, con el propósito de reforzar la gestión académica y brindar soporte a la Unidad de Planeamiento Académico como parte interesada en la toma de decisiones. A continuación enumeramos los objetivos individuales.

1. Identificar las variables/atributos que proporcionen información significativa sobre la actividad académica y el contexto universitario.

2. Detectar grupos homogéneos de información que tengan un patrón de comportamiento con la finalidad de proporcionar el mejor escenario posible para ofrecer alternativas de mejoras académicas institucionales.

3. Descubrir el conocimiento oculto en los datos aplicando técnicas de ciencia de datos extrayendo información de la base de datos del centro universitario, con el propósito de mejorar la gestión académica.

4. Examinar las técnicas de ciencia de datos utilizadas y propuestas actualmente, que permitan mejorar la adaptación en el entorno de enseñanza/aprendizaje entre docentes y estudiantes.

5. Discriminar los métodos existentes en ciencia de datos para sugerir patrones o reglas que brinden los mejores escenarios académicos y/o metodologías de estudios entre docentes y estudiantes.

El primer objetivo, se encuentran en la sección 4.1 donde se estudia variables relacionadas al profesorado y el rendimiento académico. El segundo objetivo se encuentra descrito en la sección 4.2. El tercer, cuarto y quinto objetivo, está ampliado en la sección 4.3 donde se estudian diversas variables del alumnado por medio de técnicas de preprocesado de datos y aplicación de las técnicas de aprendizaje supervisado.

4. Discusión de resultados

En este capítulo se presenta de forma general los trabajos que fueron presentados en revistas relacionadas con la temática, describiendo sus principales contenidos y una breve discusión de los resultados conseguidos.

4.1. Contexto universitario, profesores y estudiantes: vínculos y éxito académico

4.1.1. Antecedentes

La carencia de soporte para la toma de decisiones acrecienta el esfuerzo y desventajas para optimizar recursos y eficacia. Es decir, no le resulta favorable al centro universitario privarse de herramientas que sugieran alternativas para maximizar la tasa de éxito académico o en el mejor de los casos la tasa de graduados. Es así que, la literatura científica enfatiza que para este propósito existen trabajos sobre la obtención de modelos de tipo regresión logística, lineal, múltiple, modelos mixtos, entre otros para conseguir predecir o localizar las variables que pueden o son importante en el problema. En cambio, en esta propuesta previo al uso de las técnicas de ciencia de datos que se aplican en la sección 4.3 se realiza el análisis exploratorio de datos, donde su propósito era detectar y conocer la incidencia de variables relevantes del contexto universitario, profesorado y alumnado sobre el éxito académico. El estudio exploratorio que forma parte de los objetivos propuestos en ésta tesis también es una "herramienta" habitual por parte del centro universitario para la toma de decisiones basadas en reportes estadísticos según sus propias demandas.

4.1.2. Desarrollo

Sin lugar a dudas, la heterogeneidad de las variables en problemas reales es común, valorar la importancia se vuelve un completo desafío por la propia naturaleza del problema. En este sentido, el conjunto de datos usado para el análisis fue necesario aplicarle métodos de

corrección de valores ausentes, empleando métodos de imputación de datos que típicamente se usan con el algoritmo de bosque aleatorio (RandomForest, R-Package) en lugar del tradicional 3M (media, moda y mediana). Luego, con la completitud de datos perdidos fue necesario conocer la importancia o relevancia de variables que puedan aportar peso a los resultados, es decir, que posean transcendencia. Es así que, existen un sinnúmero de métodos para conocer que variable destaca o es importante en un determinado grupo de datos, para esto se usaron técnicas estadísticas como la asimetría (Skew), curtósis (Kurtosis), ganancia de información (InforGain), tasa de ganancia (GainRatio) y la incertidumbre o entropía (Entropy), de esta forma, se filtró y seleccionó las tres primeras variables que a efectos de la investigación eran relevantes para luego examinarlas de forma individual con la clase. Después del filtrado de variables con menor incertidumbre se generaron tablas de contingencias. Aunque no exista nada novedoso en el uso de las tablas de contingencia el aporte al análisis exploratorio fue agregar a éstas tablas la métrica *lift* usada de manera general en las reglas de asociación y conocer de esta manera que categoría de la variable era relevante en la clase sin necesidad profundizar en otros métodos de estudios. En contraste a estos resultados, se empleó el método mixto generalizado (glm - generalize linear methods) donde el modelo consigue diez variables representativas, entre ellas, las variables que se han estudiado con el método propuesto, destacando que en ambos casos sus categorías fueron examinadas de forma detallada.

4.1.3. Conclusiones

Por sugestivo y estimulante que resulte ser el análisis exploratorio. Sus resultados representan una visión general de las variables. Desde luego, el análisis exploratorio permite comprender y delimitar los datos, sobre todo, para dar respuesta a los objetivos planteados en esta tesis doctoral que consigue identificar variables que proporcionen información significativa sobre la actividad y el contexto universitario. Teniendo en cuenta esto se comprende mejor que existan tres factores o componentes que ponen de relieve el éxito académico de los estudiantes universitarios a la hora de aprobar el grado universitario.

El primer factor, estuvo asociado al profesorado que influyó un 34 %, siendo el factor más importante que impulsó el éxito académico del alumnado, porque cuando al menos dos tercios del total de profesores que dictaron clases en el primer año fueron profesores con edad superior a 65 años o maestros con madurez en la enseñanza (experimentados) el estudiantado tuvo éxito académico. El artículo de revista asociados a esta parte es:

- Guanin-Fajardo, J. H., y Casillas Barranquero, J. (2022). Contexto universitario, profesores y estudiantes: vínculos y éxito académico. Revista Iberoamericana De Educación, 88(1), 127–146. https://doi.org/10.35362/rie8814733

4.2. Adopción del aprendizaje combinado en la educación superior: percepción y evaluación del profesorado

4.2.1. Antecedentes

A la luz de los resultados donde se examinó de forma exploratoria la información recopilada al profesorado, resulta difícil excluir el papel que tuvieron los centros universitarios en el planeta respecto a su continuidad y enseñanza en la pandemia COVID-19. Esta situación condujo a que se aplique un cuestionario de preguntas tanto al profesorado de la Universidad Técnica Estatal de Quevedo como de otras universidades en distintas partes del mundo con profesores hispanohablantes que dieron respuestas sobre todo a la trascendencia del profesorado en tiempo de pandemia. Es así que dentro del pacto establecido las Naciones Unidas se enfatiza que la Educación de Calidad es un pilar fundamental para la consecución de los Objetivos de Desarrollo Sostenible (ODS), especialmente el ODS cuatro[1], que busca garantizar una educación inclusiva, equitativa y de calidad y que promueva oportunidades de aprendizaje para todos. Dentro de este marco, tanto la calidad educativa como el contexto universitario estriban en el éxito académico del alumnado y, siendo en este caso las competencias digitales del profesorado necesarias para alcanzar dicho éxito. De hecho, la formación docente en competencias digitales no solo mejora la metodología de enseñanza, sino que también prepara al alumnado para enfrentarse a desafíos como la globalización o integración laboral. Ya que la educación de calidad en la universidad debe incluir no solo el desarrollo de competencias digitales en el profesorado sino que también el alumnado debe adquirirlas para ampliar sus conocimientos.

4.2.2. Desarrollo

La conducción del estudio sobre las capacidades y habilidades digitales del profesorado de cara al aprendizaje del alumnado bajo situación de pandemia pudo revelar la carencia de competencias digitales del profesorado, también se encontraron dos características comunes entre ellos. La primera, que a pesar de pertenecer a universidades externas el profesorado fue hispanohablante y, la segunda, que su experiencia como docente fue superior a 15 años. En el trabajo presentado, precisó el uso de la entropía, chi-cuadrado y razón de verosimilitudes para valorar las respuestas del grupo de preguntas propuestas al profesorado. La tabla cruzada usada junto con las tres métricas de filtrado fueron empleadas para conseguir el nivel de

[1] 4.c De aquí a 2030, aumentar considerablemente la oferta de docentes calificados, incluso mediante la cooperación internacional para la formación de docentes en los países en desarrollo, especialmente los países menos adelantados y los pequeños Estados insulares en desarrollo

importancia de sus respuestas, en donde seis de las veinticuatro preguntas fueron relevantes. Esto, para deducir la influencia positiva, negativa o de riesgo de las competencias examinadas.

4.2.3. Conclusiones

Combinar la vida ordinaria del profesorado junto con la vida académica en la misma residencia habitual, se transformó en el principal riesgo debido al nivel de estrés generado en la organización de actividades académicas en el plano virtual aún cuando el centro universitario disponía de recursos y medios para aplicar el aprendizaje mixto. De hecho, las habilidades tanto pedagógicas como tecnológicas fueron una herramienta que en algunos casos el profesorado supo gestionar. La repercusión negativa fue vinculada al nivel de afinidad de las asignaturas con la metodología ya que no eran total o parcialmente compatibles. No obstante, el lado positivo fue la disposición y el grado de satisfacción del profesorado para enfrentar el desafío de dar continuidad a la enseñanza.

4.3. Predicting academic success of college students using machine learning techniques

4.3.1. Antecedentes

En este trabajo se presenta el uso de técnicas de ciencia de datos aplicadas para la etapa de preprocesado de datos, tanto en reducción de características, transformación de datos, normalización / estandarización, selección de instancias, entre otras. Al mismo tiempo, se analizaron técnicas de balanceo de datos que dan paso a la aplicación de algoritmos de clasificación supervisada en ciencia de datos. En cierta medida, la asimetría de datos es un problema típico en cualquier área de estudio es así que la duplicidad, la ambigüedad, los datos que faltan y los que se solapan son frecuentes y especialmente en problemas auténticos. De hecho, en las técnicas de clasificación de ciencia de datos, los problemas se presentan como una distribución desigual de ejemplos entre clases (variable objetivo), donde una o más clases (clase minoritaria) están infravaloradas en comparación con las demás (clase mayoritaria) (Haixiang et al., 2017). Comúnmente, en este tipo de problemas se utiliza el método de equilibrado de datos definido por (Chawla et al., 2002) que pretende cubrir el vacío existente en el equilibrado de datos utilizando diferentes métodos de equilibrado para problemas multiclase sobre los datos educativos. La puesta en marcha de los algoritmos de aprendizaje automático consiguen el modelo de predicción en primera instancia del modelo robusto, y en segunda instancia con el modelo simplificado usando el árbol de decisión C4.5.

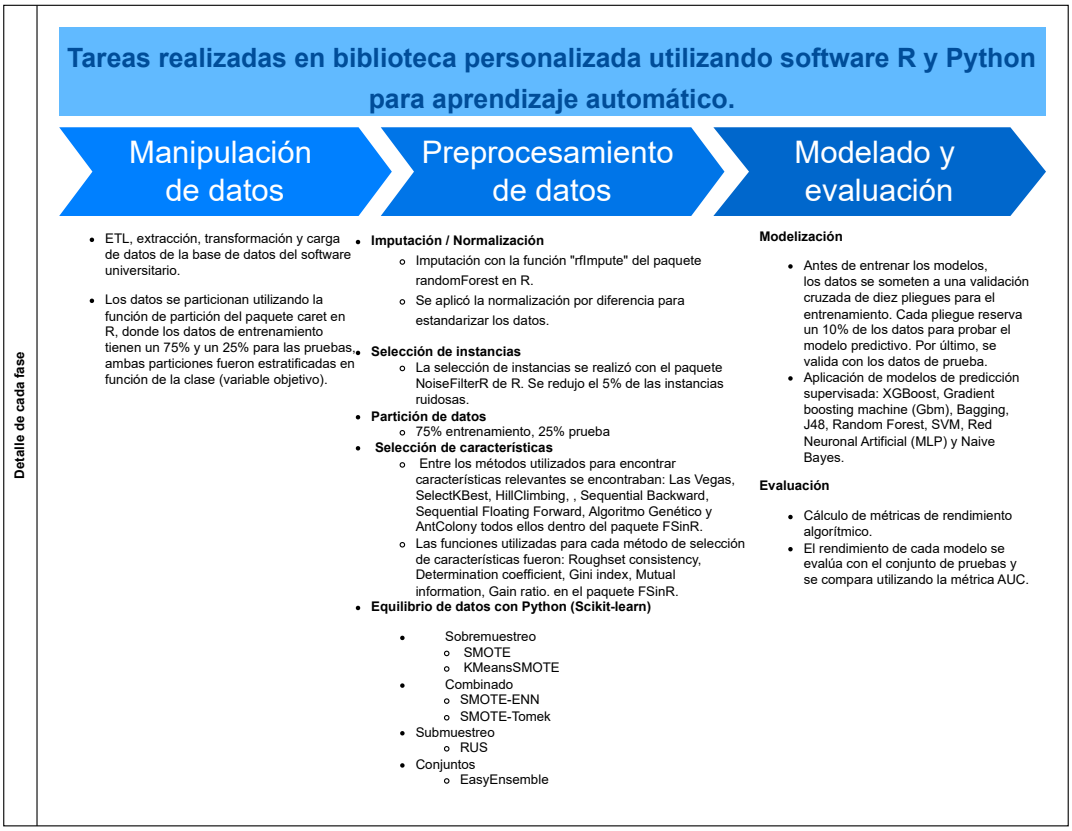


Figura 4.1 Tareas realizadas en el proceso de aprendizaje automático. En cada fase se explica lo efectuado para extraer conocimiento a partir de los datos. Tanto R como Python fueron programas usados para este propósito.

4.3.2. Desarrollo

En esta sección se implementó el grupo de métodos para conseguir el modelo predictivo del conjunto de datos educativo. Para tener una visión más clara del proceso realizado en la Figura 4.1 se muestran dichas tareas, todo bajo la metodología de trabajo plateada por Fayyad. Esta propuesta está inspirada en la aplicación profunda de la etapa de preprocesado de datos, donde por un lado se realiza el remplazo de datos ausentes, selección de características y, por otro, el equilibrado de datos. En ambos casos, se usan diferentes métodos que consiguen resultados adecuados que permitieron como entrada al grupo de algoritmos de clasificación un conjunto de datos expedito. Las tareas desarrolladas se encuentran sintetizadas en la Figura 4.1.

4.3.3. Remplazo de datos

Para el remplazo de valores fue necesario estudiar entre varias alternativas el procedimiento que consiga eficaz, para ello, en primera instancias se uso 3M (media, moda y mediana). Luego, se usó la función interna del algoritmo RandomForest que permite realizar remplazo de valores en función de cada categoría de la clase. En ambos casos, se creó un conjunto de datos donde el método de imputación de RandomForest logró tener resultados con mayor precisión.

4.3.4. Selección de características

Con el conjunto de datos carente de datos ausentes el siguiente paso fue filtrar las variables para conseguir mayor eficacia con el menor número de variables. En este paso, se comprobaron ocho métodos de filtrado de características, donde el algoritmo ReliefF consiguió reducir características que fueron relevantes para los algoritmos de predicción. Los distintos métodos par reducir características están: Las Vegas, SelectKBest, HillClimbing, Sequential Backward, Sequential Floating Forward, GeneticAlgorithm, AntColony.

4.3.5. Equilibrado de datos

Realizados los pasos anteriores de remplazo de valores ausentes y características lo siguiente era equilibrar las categorías de las clases, dado que su razón original era 7 % la clase minoritaria (cambia) y 93 % la mayoritaria (superado y abandono). En este paso se usó el Software Python y su librería scikit-learn, aquí se aplicó seis métodos de equilibrado, relacionados con sobre-muestreo (oversampling), submuestreo (undersamplig) y, equilibrado combinado (Ensemble) donde EasyEnsemble fue el que equilibró mejor los ejemplos del conjunto de entrenamiento.

4.3.6. Aprendizaje supervisado

En principio, el algoritmo XGBoost tuvo mejor desempeño ya que pudo clasificar mejor los ejemplos del conjunto de pruebas, donde ocho de cada diez estudiantes fueron clasificados de forma correcta, en cambio, usando el algoritmo de árbol de decisión C4.5 se logró predecir correctamente siete de cada diez casos. Ahora bien, otro método de predicción probado fuera del trabajo presentado en la revista y, para efectos de contraste de esta tesis fue el modelo mixto generalizado que obtuvo un 35,64 % de precisión.

4.3.7. Conclusiones

Dando respuesta a los objetivos planteados, en este trabajo se consideró dos grupos de variables importantes: el profesorado y el socio-económico del alumnado. La eficacia del modelo de predicción residió en las buenas prácticas realizadas en la fase de preprocesado de datos. Por esta razón, el modelo predictivo conseguido fue de dos tipos: el primero de tipo robusto y el segundo un modelo simplificado. De manera general, ambos fueron eficaces y su diferencia de rendimiento fue mínima. El propósito fue lograr una mayor comprensión con el modelo simplificado de cara a usarse por la parte interesada para la toma de decisiones.

El artículo de revista asociado a esta parte es:

- Guanin-Fajardo, J. H., Guaña-Moya, J., y Casillas, J. (2024). Predicting Academic Success of College Students Using Machine Learning Techniques. Data, 9(4), Artículo 60. https://doi.org/10.3390/data9040060

5. Conclusiones finales

La trascendencia de esta tesis reside sobre la base de los objetivos alcanzados, donde su fecundidad derivó en los principales resultados que fueron presentados en revistas científicas y que posteriormente fueron aceptados y publicados. De manera sustancial, se trató de arrojar luz sobre las posibles alternativas que se pueden considerar con respecto al aprovechamiento de los datos académicos del centro universitario.

- En el primer objetivo, se propuso que mediante técnicas exploratoria de datos combinar la métrica "lift" usada de manera habitual en reglas de asociación junto con la tabla cruzada (CrossTable) usada en estadística descriptiva. Los resultados lograron identificar de forma emergente las categorías significativas de cara a la aplicación de técnicas de ciencia de datos, demostrando eficiencia respecto a otros métodos clásicos.

- En el segundo objetivo, propuesto se logró centrando esfuerzos en la segunda y tercera fase del KDD. Siendo parte de la información significativa, variables asociadas al profesorado que es una de las partes esenciales para la consecución del éxito académico.

- En el tercer objetivo, se propuso la exploración y análisis de datos para localizar la distribución de datos de acuerdo con la variable dependiente.

- En el cuarto objetivo, se aplicaron técnicas de preprocesado de datos para reducir características y datos atípicos de los datos ya que debido al sesgo que se crea cuando la variable clase se encuentra desequilibrada, por ello también se aplicó técnicas de equilibrado para: sobre-muestrear, sub-muestrear y muestreo de conjunto. Por otra parte para conseguir el modelo predictivo fue necesario usar nueve técnicas de clasificación supervisada en ciencia de datos.

- En el quinto objetivo, se centra con la última fase del KDD se consiguió discriminar los modelos predictivos mediante el uso de la curva ROC o área bajo la curva AUC. Con independencia de esta evaluación, también se usó el test no paramétrico de Friedman y Wilcoxon para realizar comparaciones estadísticas entre los modelos predictivos.

6. Trabajos futuros

A partir de las conclusiones extraídas de esta tesis, se pueden proponer nuevas y prometedoras líneas de investigación. Con ellas se pretende mejorar los modelos predictivos existentes, y abordar nuevos problemas que están surgiendo en el escenario evolutivo de la ciencia de datos con especial atención en el campo de la educación.

XAI en Educación Estudiar con detenimiento el sesgo de clasificación de las técnicas de aprendizaje automático con el fin de evitar la discriminación de los algoritmos, sobre todo en el ámbito socio-económico del alumnado para que los modelos resultantes sean más equitativos.

Estudios de IA-Generativa Contrastar el uso de la incursión de la nueva tecnología de la IA-Generativa para construir modelos que permitan sugerir la explotación de esta nueva tecnología con enseñanza del profesorado. Como por ejemplo, los chatbots basados en inteligencia artificial que brindan respuestas rápidas y precisas a las consultas de los estudiantes, mejorando la experiencia de aprendizaje.

Sistema de Alertas Tempranas Proponer el desarrollo de una herramienta de seguimiento continuo y automático para los gestores académicos, profesorado y alumnado que permita anticipar posibles casos de dificultades académicas que trunquen el éxito académico del alumnado. Ya que estos sistemas monitorean y analizan datos relevantes, alertando a los usuarios ante posibles situaciones críticas. El sistema puede aprender a partir de datos históricos (calificaciones, asistencia, etc.). Al aplicar estos sistemas se debe estudiar de forma cuidadosa el sesgo y la interpretación que consigan los modelos predictivos.

Aprendizaje por Refuerzo Investigar sobre cómo los agentes pueden aprender a tomar decisiones óptimas en entornos cambiantes. Es decir, tanto la flexibilidad curricular, formación docente, evaluación del rendimiento académico, entre otras. son temáticas tratadas en el aprendizaje por refuerzo para moldear comportamientos y personalizar el logro o éxito académico.

Parte II

Publicaciones

Trabajos presentados y aceptados

"No tiene ningún sentido ser preciso cuando ni siquiera sabes de lo que estás hablando."

John von Neumann.

La fecundidad de esta tesis se ha reflejado en los trabajos presentados en revistas cualificadas de acuerdo con la normativa de la Escuela Internacional de Posgrado. En el primer trabajo, se presentó el estudio del conjunto de datos con información del rendimiento académico del alumnado así como del profesorado que impartió clases en el primer curso. Ha destacado en primera instancia el papel crucial del profesorado para encauzar al alumnado hacia el éxito académico. En el segundo trabajo, se presentó el estudio del profesorado y su adaptación al estudio combinado que fue implementado para dar continuidad en la enseñanza universiraria debido a la aparición de la pandemia COVID19. Se desveló que el profesorado universitario y, otros profesores hispanohablantes de universidades externas carecieron de recursos para ser frente a la pandemia, no obstante, su predisposición para saltear inconvenientes fue considerada para que se implemente en los programas curriculares y que sirva como una herramienta más de trabajo. En el tercer trabajo, se presentó el estudio profundo sobre la aplicación de ciencia de datos en el contexto universitario, donde se consideró la información socio-económica y del rendimiento académico del alumnado.

7. Contexto universitario, estudiantes y profesores: vínculos y éxito académico

- Jorge Guanin-Fajardo, Jorge Casillas
  - Status: Publicado (Revista Iberoamericana de Educación)
  - WoS “Education & Educational Research”, JCI 2022: 0.6, 605/756 (Q3).
  - WoS “Education & Educational Research”, JIF 2023: 0.6, 520/756 (Q3).

Contexto universitario, profesores y estudiantes: vínculos y éxito académico

Jorge Humberto Guanin-Fajardo jorgeguanin@uteq.edu.ec

Facultad de Ciencias de la Ingeniería, Universidad Técnica Estatal de Quevedo, Ecuador, Los Ríos, Quevedo
CP:EC120509

Jorge Casillas casillas@decsai.ugr.es

Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Granda, España, Granada, CP:18014

DOI: https://doi.org/10.35362/rie8814733

Resumen: La promoción de una educación de calidad en las instituciones de enseñanza superior promueve la autoeficacia. La utilidad del trabajo se ha dirigido al análisis de las características del profesorado y el éxito académico de los estudiantes al final del primer año en el contexto universitario. La población estudiada fue de 6690 estudiantes y 256 profesores, el conjunto de datos tenía 15 variables entre numéricas y categóricas. Se utilizó estadística descriptiva, métricas diseñadas para evaluar datos significativos y técnicas avanzadas de visualización. Los resultados revelaron el perfil esencial de los profesores experimentados y maduros, tanto en la enseñanza como en los grupos de edad. Los profesores experimentados que participaron en la enseñanza en un porcentaje superior al 66%, influyeron con un 72% de certeza en el éxito académico del alumnado. A corto plazo, los profesores noveles cuya tasa de participación fue del 33% mostraron un efecto positivo. A largo plazo, los estudiantes cambiaron (8%) o abandonaron (59%) la carrera universitaria. La utilidad de estos resultados proporciona sugerencias para una enseñanza significativa y oportuna, siempre que la distribución del profesorado experimentado y maduro corresponda a dos o tres tercios del total de profesores del primer año de la titulación universitaria.

Palabras Claves: Éxito académico, contexto universitario, análisis educativo, técnicas de visualización.

University context, teachers and students: links and academic success

Abstract: The promotion of quality education in higher education institutions promotes self-efficacy. The usefulness of the work has been directed to the analysis of faculty characteristics and the academic success of students at the end of the first year in the university context. The population studied was 6690 students and 256 professors, the data set had 15 variables between numerical and categorical. Descriptive statistics, metrics designed to evaluate meaningful data and advanced visualization techniques were used. The results revealed the essential profile of experienced and mature teachers, both in teaching and age groups. Experienced teachers who participated in teaching at a rate of more than 66%, influenced with 72% certainty the academic success of the student body. In the short term, novice teachers whose participation rate was 33% showed a positive effect. In the long term, students either changed (8%) or dropped out (59%) of their university career. The usefulness of these results provides suggestions for meaningful and timely teaching, provided that the distribution of experienced and mature faculty corresponds to two to three-thirds of the total number of first-year faculty in the university degree program.

Keywords: Academic success, university context, educational analysis, visualization techniques.

1 Introducción

La educación, como parte fundamental del progreso de las sociedades y economías modernas impulsadas por la innovación y el desarrollo científico, nunca ha sido tan omnipresente como ahora (Marginson, 2014). Por ello, las instituciones de educación superior se han dedicado a promover la buena educación por diversas razones. En primer lugar, porque les interesa demostrar que son proveedores fiables de una educación de buena calidad, al tiempo que sirven a múltiples partes interesadas con diferentes expectativas (grados universitarios). En segundo lugar, porque deben responder a la creciente demanda de una educación significativa y oportuna. Por último, porque los resultados de la investigación son insuficientes para mantener la reputación de las instituciones de enseñanza superior, por lo que es esencial equilibrar los resultados de la enseñanza y el aprendizaje con los de la investigación (Nasser-Abu Alhija, 2017). En este sentido, los investigadores educativos han logrado crear herramientas tecnológicas, planes pedagógicos y/o curriculares, modelos predictivos, etc., para de esta manera suministrar a los responsables académicos recursos para usar estrategias controladas y proporcionadas para conseguir retener al alumnado en el grado universitario inicial (Araque, Roldán, & Salguero, 2009; B. K. Mishra & Sahoo, 2016; Van Den Berg & Hofman, 2005). Desde de este punto de vista, el contexto universitario gestionado por las instituciones de educación superior es diverso, por ello, la aplicación de políticas y normas que regulan su actividad facilitan el progreso de las instituciones. No obstante, su diversidad más extendida y compleja se localiza en el recurso humano, de servicio, de infraestructura, económico, tecnológico, entre otros. Esta complejidad, trasciende en el grado de impacto y aceptación en la sociedad moderna, de hecho, el prestigio se puede alcanzar ajustando los recursos del contexto universitario. Para lograr el perfil de aceptación institucional, hemos proyectado estudiar el contexto universitario basado en dos recursos humanos esenciales, hablamos del profesorado y alumnado.

Partiendo de este punto, hemos encontrado cinco tipos de investigación en el ámbito educativo que han permitido descubrir información reveladora en datos académicos. En primer lugar, está el análisis de redes sociales que estudia diferentes iteraciones e implicaciones generales (S. Mishra, 2020; Trolian, Jach, & Archibald, 2020). En segundo lugar, los estudios longitudinales destinados a mejorar los resultados de los estudiantes, por ejemplo, (Amida, Algarni, & Stupnisky, 2020; Souchon, Kermarec, Trouilloud, & Bardin, 2020). En tercer lugar, el estudio del análisis factorial para investigar los factores ocultos en las interacciones entre alumnos y profesores (Le, Bolt, Camburn, Goff, & Rohe, 2017). En cuarto lugar, el meta-análisis examina características relacionadas con la implementación de estrategias para el aprendizaje del rendimiento académico de los estudiantes (de Boer, Donker, & van der Werf, 2014), y. Por último, la minería de datos que explora mediante dos técnicas el descubrimiento de conocimiento: (i) técnicas no supervisadas, divididas en dos sub-técnicas. a) agrupamiento basado en estudios de distancia o similitud de vectores (Vo, Nguyen, & Vo, 2016). b) reglas de asociación, para descubrir los hechos que ocurren dentro de los datos (Aleksandrova & Parusheva, 2019; Alyahyan & Düştegör, 2020; Autor, 2019). (ii)

técnicas supervisadas, que predicen los datos mediante una variable dependiente (Shetu, Saifuzzaman, Moon, Sultana, & Yousuf, 2021). La convergencia de las investigaciones ha coincidido en la flexibilidad de la mejora de los resultados académicos, la calidad de las relaciones de la comunidad universitaria, la mejora de los canales de comunicación, la buena enseñanza, la proyección de los objetivos, entre otros.

Objetivo del estudio.

El objetivo principal del trabajo está centrado en descubrir anticipadamente el éxito académico del alumnado dentro del contexto universitario. Por consiguiente, examinaremos el vínculo existente entre estudiantes y profesores, de modo que, nos hemos planteado las siguientes preguntas de investigación:

- ¿Cuáles son los factores del profesorado que han influido en el éxito académico del estudiante?
- En el contexto univsersitario, ¿qué tipo de compatibilidad del profesorado se corresponde con el éxito académico de los estudiantes?

Para ello, examinamos los datos en profundidad para extraer información útil y relevante sobre las características del docente. Dividimos el estudio en tres etapas: (i) recuperación de datos del sistema informático; (ii) análisis y aplicación de procedimientos para extraer datos significativos mediante las métricas propuestas, y; (iii) presentación de los principales resultados mediante técnicas de visualización. El análisis de datos propuesto pretende obtener información significativa sobre los factores del profesorado y el impacto en el alumnado para completar la titulación universitaria al finalizar el primer año. Nuestro trabajo, motiva la toma de decisiones y es precursor de futuros estudios de análisis de datos exhaustivos para comprobar posibles teorías. Para ello, hemos creado una biblioteca personalizada de análisis de datos utilizando el programa estadístico R, que es un lenguaje de libre acceso para la computación estadística y que proporciona una amplia variedad de técnicas estadísticas y gráficas: modelización lineal y no lineal, pruebas estadísticas, clasificación, agrupación, entre otras. (R CoreTeam, 2019).

2 Trabajos relacionados

Las instituciones de educación superior centran sus esfuerzos en el desarrollo de habilidades o atributos curriculares para que los estudiantes tengan una alta probabilidad de éxito académico (Leal Filho, Shiel, & Paço, 2016). Partiendo de este punto, la oferta académica y el alcance de los servicios de las instituciones de educación superior son cruciales para el éxito académico. A vista del trabajo de Respondek (2017), la conceptualización del éxito académico se perfila en dos partes: (i) baja intención de abandono y, (ii) alto rendimiento académico. Ambas debidamente sincronizadas responden positivamente al éxito académico del alumnado. De ahí que, la armonización del rendimiento académico y la motivación del alumnado descansen significativamente en el profesorado. Esta relación, trasciende por la claridad percibida en la enseñanza del profesor, capacidad y apoyo, así como en la implicación del nivel de satisfacción y la experiencia

universitaria (Livengood, 1992; Pascarella, Edison, Hagedorn, Nora, & Terenzini, 1996).

2.1 Contexto universitario

El contexto universitario asocia múltiples factores para fortalecer e influir en el éxito académico de los estudiantes (Struyven, Dochy, & Janssens, 2003). Siguiendo el trabajo de Winterer (2020), los autores sugieren prácticas y normas que han facilitado el éxito académico: (i) la mejora del clima estudiantil; (ii) la calidad del acceso, el conocimiento de los estudiantes y el servicio de orientación (Korobova & Starobin, 2015); y, (iii) el aumento y la mejora de la calidad de los programas y servicios de asistencia académica (Kara, Çubukçuoğlu, & Elçi, 2020). Estas prácticas estimulan la calidad en las relaciones de la comunidad universitaria y promueven espacios socialmente aceptables (Clelia et al., 2014). La incorporación del aprendizaje basado en proyectos dentro del plan de estudios también se ha considerado un enfoque exitoso (Konrad, Wiek, & Barth, 2021; Leal Filho et al., 2016). De hecho, este aprendizaje expone numerosos conocimientos útiles, además de la contribución al crecimiento profesional y al aprendizaje de meta-habilidades (Salminen-Tuomaala & Koskela, 2020). Por otro lado, es difícil no valorar la tutoría como un factor vinculado al éxito académico y al liderazgo de los estudiantes durante la fase universitaria, dado que ha estado empíricamente relacionado con el desarrollo profesional (C. M. Campbell, Smith, Dugan, & Komives, 2012; Cunha, Miller, & Weisburst, 2018; Jacobi, 1991). En líneas generales, en cualquier contexto universitario el profesorado tuvo un alto porcentaje de participación en la formación del estudiantado, es decir, todo el proceso de enseñanza y desarrollo profesional del estudiantado estriba en el profesorado.

2.2 El éxito académico y el profesorado

En cuanto a las expectativas del éxito académico del estudiantado subyace el desánimo como consecuencia del tipo y el contexto de las iteraciones con el profesorado. La importancia de un profesorado de buena calidad y su influencia está bien vista por el alumnado (Cho, Kim, Svinicki, & Decker, 2011). Esto fue reforzado por Lizzio, Wilson y Simons (2002), que dentro de sus hallazgos, afirmaron que la buena enseñanza tenía un efecto positivo en los resultados académicos y que estaba fuertemente asociada con el éxito académico. De esta manera, han surgido muchas investigaciones para conocer los factores influyentes del profesorado y el éxito académico del estudiantado (Chickering, Arthur W.;Gamson, 1987; Crisp, Taggart, & Nora, 2015; Tinto, 1975; Walder, 2016). No obstante, también se han encontrado trabajos que destacan también la influencia negativa (Glogowska, Young, & Lockyer, 2007; Young, Glogowska, & Lockyer, 2007). En cierto modo, el sistema educativo busca constantemente fomentar canales de comunicación profesor-alumno y alumno-alumno que es el eslabón clave para lograr el éxito académico. De alguna manera, la colaboración y participación mutua entre ellos genera un ambiente de confianza y cooperación para lograr los objetivos proyectados (Sarahí Abarca, Mireya; Gómez Pérez, 2015). De ahí que la consolidación de la fluidez de los distintos canales de comunicación hayan promovido iteraciones que han impulsado al alumnado a

alcanzar el éxito académico (Chickering, Arthur W.;Gamson, 1987; S. Mishra, 2020; Trolian et al., 2020; Winterer et al., 2020). Al mismo tiempo, este hecho tiene una consecuencia importante, ya que pueden beneficiarse de oportunidades fuera del aula que enfatizan el valor del trabajo intelectual y el apoyo académico (Nagda, Gregerman, Jonides, Von Hippel, & Lerner, 1998). De todo lo que se ha dicho sobre la eficacia de la comunicación e iteración entre el profesorado y alumnado, otro punto a tener en cuenta está relacionado con los factores y la calidad del profesorado. De ahí que, por ejemplo han surgido estudios sobre: la edad (H. E. Campbell, Steiner, & Gerdes, 2005), titulación académica y experiencia docente (Angervall, 2018; Darling-Hammond, 2000; Jepsen, 2005; Korhonen & Törmä, 2016) que son tres factores que han influido en el éxito académico del estudiantado.

3 Metodología

3.1 Contexto

Para realizar el estudio hemos considerado la Institución de Educación Superior (IES) como parte del estudio; la política universitaria y el cumplimiento de requisitos mínimos permiten al estudiantado superar cada año académico. La modalidad de estudio de la IES es presencial y el ciclo académico consta de dos semestres, el estudiantado aprobará ambos semestres para pasar al curso inmediato superior, debiendo alcanzar la nota mínima exigida en cada asignatura (7, en una escala de 0 a 10). La IES está localizada geográficamente en el cantón Quevedo, Los Ríos, Ecuador.

3.2 Recolección de datos

Los datos para el estudio lo hemos recuperado del sistema informático de la Universidad, donde fue almacenada la actividad académica entre profesores y alumnos. Primero, hemos recuperado la información sobre las actividades evaluativas desarrolladas a lo largo del grado académico entre ambos. Después, el Departamento de Recursos Humanos ha proporcionado información cualitativa del profesorado que hemos unido y relacionado para formar el conjunto de datos final. Por último, hemos filtrado los datos en seis periodos académicos entre el primer y el quinto año de todas las titulaciones universitarias. Además, hemos aplicado las respectivas políticas de protección de datos proporcionadas por la Dirección de Planificación Académica de la Universidad que aprobó la recolección de datos.

3.3 Datos

El conjunto de datos original tiene 6.690 registros y 15 variables categóricas y numéricas (véase el anexo 1). En este estudio, utilizamos como población a los estudiantes que han estado matriculados en el primer año y que han completado el grado académico. Además, la población total de profesores que analizamos fue de 286, incluyendo profesores titulares, agregados y ocasionales.

3.4 Preparación de los datos

Esta es una etapa importante del estudio, ya que es fundamental contar con datos claros y de buena calidad. Para ello, hemos aplicado correctivos para los valores que faltan, ya que es común en problemas reales la omisión involuntaria de transcripción o recuperación automática de datos que se quedan sin valores, después, hemos dado uniformidad a los datos según la ecuación 3.

3.5 Métricas

En este estudio hemos usado seis tipos de métricas para transformar y evaluar la calidad de los datos. La ecuación 1 ha ponderado las titulaciones del profesorado, donde x = cantidad de profesores con grado académico (licenciatura, ingeniería, biología, etc.); y = cantidad de profesores con títulos de máster; z = cantidad de profesores con títulos de doctorado; y, por último, n = total de profesores que han impartido clases en el curso académico.

$$w = \left[\left(\frac{x}{n} * 0.10 \right) + \left(\frac{y}{n} * 0.30 \right) + \left(\frac{z}{n} * 0.60 \right) \right] \quad \text{(Ecuación 1)}$$

La ecuación 2, es la métrica *lift*, comúnmente utilizada en minería de datos, donde se obtiene la mejora de la confianza de las reglas de asociación, donde tanto *x* como *y* son elementos del conjunto de datos (Brin, Motwani, Ullman, & Tsur, 1997). La ecuación se define como:

$$Lift(X \rightarrow Y) = \frac{Confidence(X \rightarrow Y)}{Support(Y)} \quad \text{(Ecuación 2)}$$

Donde $confidence = \frac{Support(X \cup Y)}{Support(Y)}$, y, $Support(Y)$ es definido como la proporción de transacciones en el conjunto de datos que contiene Y. Por otra parte, en la ecuación 3 hemos propuesto la fórmula que ha servido para obtener datos uniformes. Donde Z_i es la variable normalizada [0-1], siendo $x_{\min}$ y $x_{\max}$ el valor mínimo y máximo de la variable respectivamente.

$$Z_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \quad \text{(Ecuación 3)}$$

Las ecuaciones 4, 5 y 6 fueron planteadas para un análisis profundo de la información, donde se analizó el desorden de los datos (entropía de la información), la ganancia de información y la tasa de ganancia (Romanski & Kotthoff, 2016). Con estas ecuaciones se han obtenido cuantitativamente el comportamiento de las variables.

$$Symmetrical.uncertainty = 2 \frac{H(Class) + H(Attribute) - H(Class, Attribute)}{H(Attribute) + H(Class)} \quad \text{(Ecuación 4)}$$

$$Information.gain = H(Class) + H(Attribute) - H(Class, Attribute) \quad \text{(Ecuación 5)}$$

$$Gain.ratio = \frac{H(Class) + H(Attribute) - H(Class, Attribute)}{H(Attribute)}$$ (Ecuación 6)

3.6 Procedimiento

Para lograr los objetivos propuestos en este trabajo, hemos creado una librería personalizada con el programa estadístico R; esta librería tiene varias funciones para el procesamiento, análisis de datos y visualización de resultados. Para ello, hemos considerado seis pasos: Primero, hemos sustituido los datos ausentes en el conjunto de datos por datos aproximados utilizando medidas estadísticas de posicionamiento central (media, mediana y moda) (Breiman, 2001), según el tipo de variable. Segundo, hemos normalizado las variables para obtener datos homogéneos [0-1]. Tercero, hemos estudiado inicialmente las variables a través del cluster jerárquico, que ha servido para concentrar las variables según el grado de similitud utilizando para ello la distancia euclidiana. Cuarto, hemos analizado el comportamiento de las variables con las métricas: curtosis, asimetría, incertidumbre, ratio de ganancia y ganancia de información, esto, para el filtrado de las tres variables principales que servirán para el análisis. Quinto, hemos categorizado las variables para calcular la tabla de contingencia, con el fin de obtener las proporciones de las categorías frente a la variable dependiente (Abandono, Cambio, Superación), además hemos calculado la métrica lift (ecuación 2) para obtener el grado de confianza entre los datos encontrados. Por último, hemos conseguido proyectar los resultados significativos en gráficos para mejorar la comprensión de los hallazgos obtenidos.

4 Resultados

En esta sección hemos presentado los principales resultados obtenidos a través del análisis profundo de los datos. En respuesta a las preguntas iniciales de este trabajo, hemos empleado las métricas propuestas para conocer el comportamiento de las variables. Asimismo, hemos utilizado técnicas estadísticas que han estudiado, por un lado, la forma y distribución general de los datos y, por otro, la relación que ha existido entre ellos.

4.1 Estudio exploratorio

Como punto de partida para la exploración de los datos, hemos evaluado la información desde dos perspectivas. La primera fue crear el conglomerado jerárquico calculando la similitud entre las variables mediante la distancia Euclidiana que ha ayudado y mucho a comprender los grupos de variables. En consonancia con lo anterior, hemos mostrado en la columna 3 de la Tabla 1, la media de participación del profesorado respecto a sus edades ha sido superior en Edad2 e inferior en Edad3 y Edad1. Sin embargo, la desviación típica (columna 2) en Edad1 fue mayor, lo que indica que también hubo una participación significativa de los profesores en este grupo de edad. Por otra parte, la media de experiencia docente y titulación docente fue de 16 años y 0.26 respectivamente. La razón

fundamental de estos valores fue dada por la participación en mayor cantidad de profesores del grupo Edad2.

Tabla 1. Métricas de centralidad y tendencia de las variables independientes*.

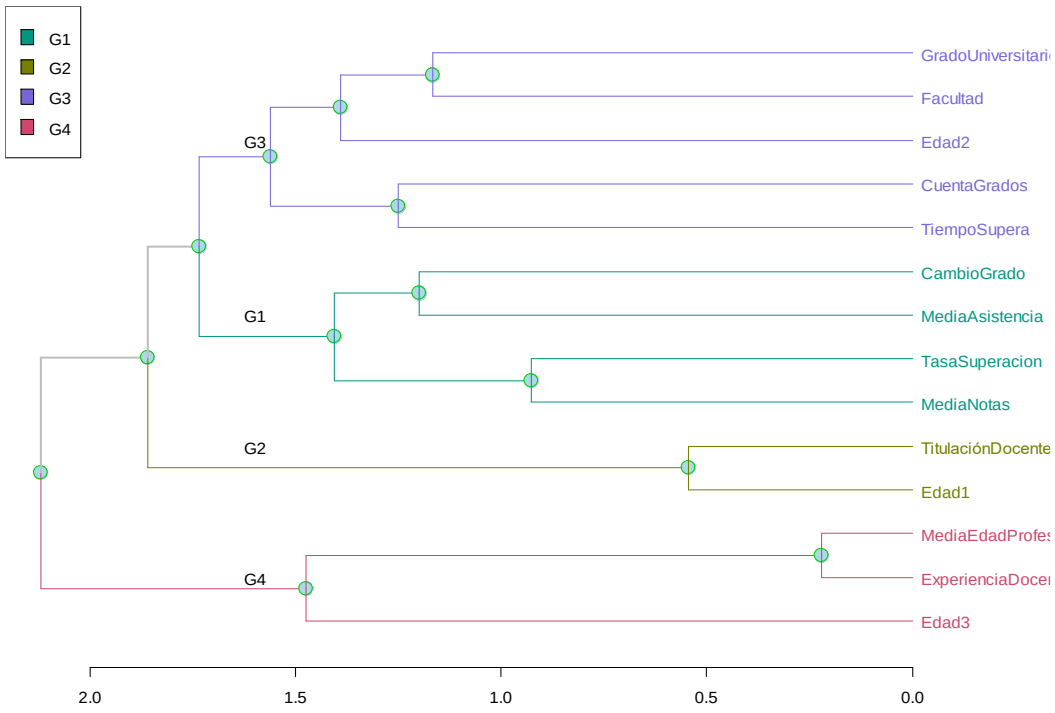
Variables	Desviación típica	Media	Mínimo	Máximo
Facultad			1	5
GradoUniversitario			1	22
Edad1	2.9127	2	0	15
Edad2	2.3867	5	0	11
Edad3	1.8906	2	0	10
ExperienciaDocente	4.2912	16	4	26
MediaEdadProfesor	6.2228	53	35	64
TitulaciónDocente	0.0943	0.26	0.1	0.6
MediaNotas	1.3044	7.33	0.03	10
MediaAsistencia	6.5976	97	21	100
TiempoSupera	0.4911	1	1	5
TasaSuperacion	0.2440	1.05	0.052	2.708
CuentaGrados	0.2490	0	0	2
CambioGrado			1	2

*Las variables categóricas tienen valores vacíos en la columna de desviación típica y media.

Fuente: Elaboración propia

En el Gráfico 1, presentamos el conglomerado jerárquico que fue dividido en cuatro grupos de variables según el grado de similitud. Para ello, hemos calculado y agrupado las variables según la distancia euclidiana. Es decir, medimos la distancia que hay entre una variable y otra, han destacado dos grupos relacionados con el profesorado (G2 y G4). El rendimiento académico del alumnado estuvo agrupado en G1. Por último, G3 estuvo compuesto con variables vinculadas al rendimiento académico (Tiempo superación, cuenta grados), profesorado (Edad2) y titulación académica (Facultad, grado universitario).

Gráfico 1. En el conglomerado jerárquico hemos diferenciado cuatro subgrupos de variables, Los subgrupos G2 y G4 se han vinculado con las características del profesorado, G1 fue asociado con el rendimiento académico, y G3 ha implicado una mezcla de variables entre estudiantado y profesorado.



Fuente: Elaboración propia

4.2 **Análisis de los datos relacionados con los factores del profesorado**

En este apartado, examinamos en profundidad las variables vinculadas al profesorado. Para ello, hemos usado el diagrama de Sankey que visualiza la carga y distribución de los datos entre las variables. Asimismo, en la Tabla 2 presentamos el estudio de las variables con las métricas: asimetría, curtosis, incertidumbre, ganancia de información y tasa de ganancia. De acuerdo con este análisis, hemos filtrado las tres primeras variables para examinar la incidencia del profesorado con los alumnos que han tenido éxito académico.

Tabla 2. Estudio profundo de variables relacionadas al profesorado, ordenadas según el nivel de incertidumbre (Uncertainty).

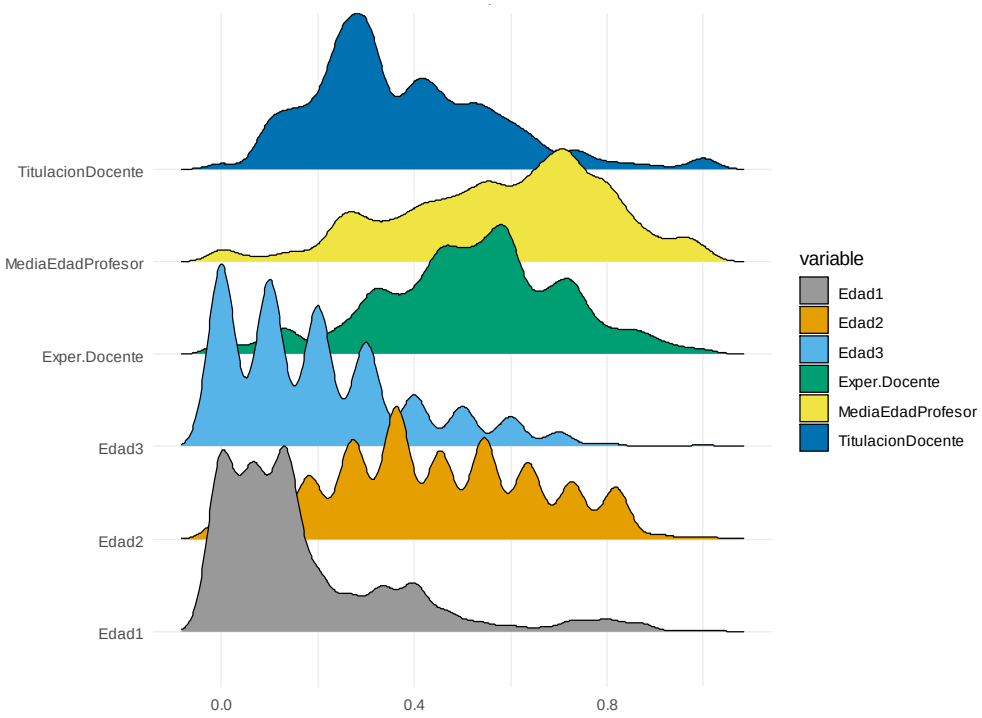
Variables	**Asimetría**	**Curtosis**	***Uncertainty****	**InforGain**	**GainRatio**
Edad3	1,0079	0,6380	0,0229	0,0213	0,0218
Edad2	0,0790	-0,6948	0,0294	0,0333	0,0241
Experiencia Profesorado	0,2699	0,0489	0,0486	0,0557	0,0394
MediaEdadPofesorado	0,4500	-0,2568	0,0609	0,0703	0,0492
TitulacionDocente	0,8466	0,7942	0,0701	0,0813	0,0564
Edad1	1,7087	2,8320	0,0742	0,0662	0,0731

* Orden ascendente

Fuente: Elaboración propia

Las variables asociadas al profesorado, por ejemplo, la asimetría de *Edad3* fue 1.0079, donde fue evidenciado inicialmente que pocos profesores de ese rango de edad participan en las clases del alumnado. Lo contrario ha sucedido con la *Edad1* con asimetría de 1,7087, que evidenció mayor presencia de profesores de este rango de edad. Por otra parte, *Edad2* y *ExperienciaProfesorado* han evidenciado también pocos profesores de esas edades y experiencia docente respectivamente.

Gráfico 2. Densidad de variables relacionadas con las características del profesorado. La asimetría de las variables Exper.Docente y MediaEdadProfesor son mostradas hacia la izquierda, mientras que el resto de variables hacia la derecha.

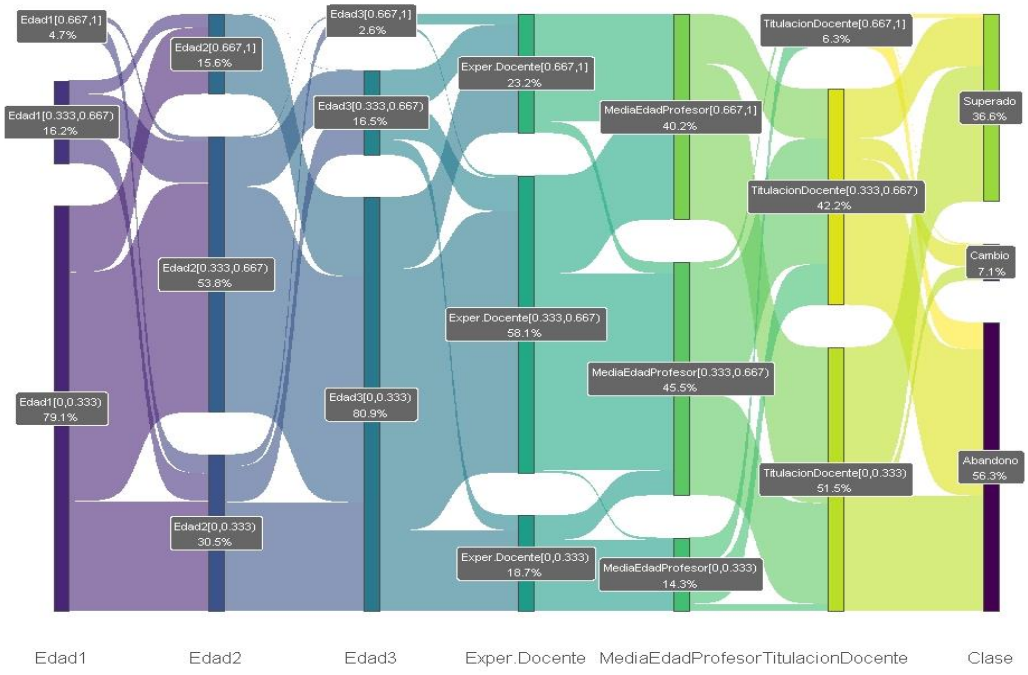


Fuente: Elaboración propia

En el Gráfico 2, hemos condensado los resultados de la Tabla 2, en concreto lo relacionado con la asimetría y curtósis, donde gráficamente valoramos la distribución y tendencia que los datos han tenido. Por otra parte, el diagrama de Sankey (Gráfico 3) nos ha llamado la atención seis grupos de datos: primero, *Edad1[0-0,33)* ha mostrado 79,1% del profesorado menor de 45 años; segundo, *Edad2[0,33-0,667]* ha concentrado del 53,8% del profesorado entre 45 y 60 años; tercero, *Edad3 [0,00-0,333]* ha presentado el 80,9% del profesorado con edad superior a 60 años; cuarto, *MediaEdadProfesorado[0,33-0,667)* ha concentrado el 45,52%; quinto, *Exper.Docente[0,33-0,667)* ha agrupado el 58,1% de datos relacionados con la variable experiencia docente; por último, la categoría *TitulaciónDocente[0-0,333)* ha concentrado 51,53% de datos relacionados con la titulación académica del docente. La iteración del flujo de datos ha evidenciado la tendencia en la distribución del profesorado a lo largo de los grados académicos

del alumnado, por ende, la variable objetivo “Clase” agrupa el porcentaje del alumnado que ha superado, cambiado o abandonado el grado académico.

Gráfico 3. Flujo y carga de datos con características del profesorado. Las variables fueron normalizadas en un rango entre 0 y 1, posteriormente fueron las separamos en tres categorías, después, mostramos la carga (%) de las categorías que facilita la comprensión de lo que ocurre entre las categorías. Al final del Gráfico, mostramos la variable denominada “Clase” que representa el estado académico final del alumnado.



Fuente: Elaboración propia

Tabla 3. Resultados de la relación entre las categorías de la variable *Edad3* frente a la situación académica del alumno (*Superado*). La métrica *Lift* ha destacado los menos significativos cuyo valor es inferior a 1 y los más significativos mayores o iguales a 1.

Edad3		**Estado académico**			
		Abandona	**Cambio**	***Superado***	**Total**
[0,00-0,33)	Número de casos	3188	431	1791	5410
	Frecuencia sobre el total [a]	58,90%	8,00%	33,10%	*80,90 %*
	Distribución sobre la categoría	84,60%	90,90%	73,20%	
	Relación categoría/total* (Lift)	1.046	1.127	0.904	
[0,33-0,667)	Número de casos	536	38	529	1103
	Frecuencia sobre el total	48,60%	3,40%	48,00%	16,50 %
	Distribución sobre la categoría	14,20%	8,00%	21,60%	
	Relación categoría/total (Lift)	0,863	0,479	1,311	
[0,667-1,00]	Número de casos	44	5	128	177
	Frecuencia sobre el total	24,90%	2,80%	72,30%	2,60%
	Distribución sobre la categoría	1,20%	1,10%	5,20%	
	Relación categoría/total (Lift)	0.442	0.394	*1.975*	
Total		3768	474	2448	6690
Porcentaje [b]		56,30%	7,10%	36,60%	

*Relación categoría/total = (a / b)
Fuente: Elaboración propia

En la Tabla 3, hemos mostrado las categorías de la variable “Edad3” que representa la cantidad de profesores con edad superior a 60 años, donde el 80,90% de los datos se han concentrado en la categoría “[0-0,33)”, es decir, el primer tercio del total profesores. Al mismo tiempo, hemos evidenciado dos categorías relevantes: (i) la categoría “[0,33-0,667)” que del total 16,50%, 48% han superado el grado académico y la relación categoría/total fue 1,311; (ii) la categoría “[0,667-1,00]” que del total 2,60%, 72,30% han superado el grado académico, y la relación categoría/total fue 1,975. En líneas generales, hemos detectado que la participación del tercer tercio del profesorado en el proceso de enseñanza ha influenciado en positivo al estudiantado para superar el grado universitario.

Tabla 4. Resultados de la relación de datos entre la categoría de la variable *Edad2* frente al estado académico del alumnado (*Superado*). Utilizamos la métrica *Lift* para destacar los menos significativos cuyo valor es inferior a 1 y los más significativos mayores o iguales a 1.

Edad2		Estado académico			
		Abandona	**Cambio**	***Superado***	**Total**
[0,00-0,33)	Número de casos	1446	124	473	2043
	Frecuencia sobre el total [a]	70,80%	6,10%	23,20%	30,50%
	Distribución sobre la categoría	38,40%	26,20%	19,30%	
	Relación categoría/total* (Lift)	1,258	0,859	0,634	
[0,33-0,667)	Número de casos	1761	256	1585	3602
	Frecuencia sobre el total	48,90%	7,10%	44,00%	*53,80%*
	Distribución sobre la categoría	46,70%	54,00%	64,70%	
	Relación categoría/total (Lift)	0.869	1,000	1,202	
[0,667-1,00]	Número de casos	561	94	390	1045
	Frecuencia sobre el total	53,70%	9,00%	37,30%	15,60%
	Distribución sobre la categoría	14,90%	19,80%	15,90%	
	Relación categoría/total (Lift)	0,954	1,268	1,019	
Total		3768	474	2448	6690
Porcentaje [b]		56,30%	7,10%	36,60%	

*Relación categoría/total = (a / b)

Fuente: Elaboración propia

Como hemos visto, en la Tabla 4 las categorías de “Edad2”, siendo “(0,33-0,667]” la que tuvo una densidad del 53,80% de los datos. Como consecuencia de haber explorado esta variable, encontramos dos categorías relevantes: (i) la categoría “[0,33-0,667)” que del total del 53,8%, el 44% han superado el grado universitario, mientras que, la relación categoría/total fue 1,203. (ii) La categoría “[0,667-1,00]” del total 15,6%, 37,30% han superado el grado académico, y la relación categoría/total de 1,02. En otras palabras, el segundo y tercer tercio de los profesores entre 45 y 60 años influyeron positivamente en el alumnado para superar el grado universitario. Si bien observamos, en la Tabla 5, hemos examinado las categorías de la variable experiencia del profesorado, la densidad de los datos fue 58,10% en la categoría “[0,33-0,667)”. Teniendo en cuenta los datos de la tabla, hemos encontrado que dos categorías fueron relevante: (i) la categoría “[0,33-0,667)” que del total 58,10%, el 37,20% ha superado el grado universitario, y la relación categoría/total fue 1,016; (ii) la categoría “[0,667-1,00]” del total del 23,20%, el 54,30% ha superado el grado universitario, y la relación categoría/total fue 1,484. Una importante distinción que hacer en la experiencia del profesorado fue que la segunda y tercera categoría ha influido en los estudiantes para superar el grado universitario.

Tabla 5. Resultados de la relación de datos entre la categoría de la variable experiencia del profesor frente a la situación académica del alumno (*Superado*). Utilizamos la métrica Lift para destacar los menos significativos cuyo valor es inferior a 1 y los más significativos mayores o iguales a 1.

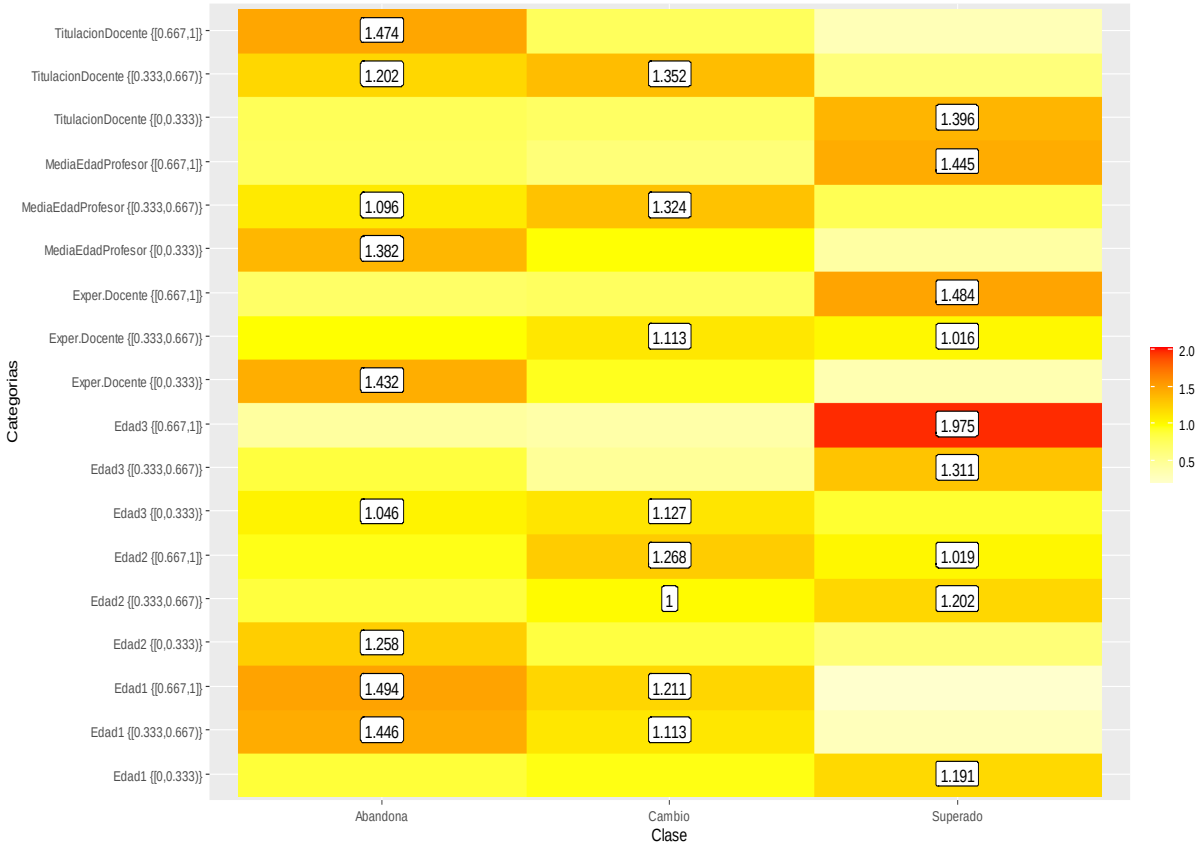
Experiencia profesorado		**Estado académico**			
		Abandona	**Cambio**	***Superado***	**Total**
[0,00-0,33)	Número de casos	1007	84	159	1250
	Frecuencia sobre el total [a]	80,60%	6,70%	12,70%	18,70%
	Distribución sobre la categoría	26,70%	17,70%	6,50%	
	Relación categoría/total* (Lift)	1,432	0,944	0,347	
[0,33-0,667)	Número de casos	2134	307	1445	3886
	Frecuencia sobre el total	54,90%	7,90%	37,20%	*58,10%*
	Distribución sobre la categorí	56,60%	64,80%	59,00%	
	Relación categoría/total (Lift)	0,975	1,113	1,016	
[0,667-1,00]	Número de casos	627	83	844	1554
	Frecuencia sobre el total	40,30%	5,30%	54,30%	23,20%
	Distribución sobre la categoría	16,60%	17,50%	34,50%	
	Relación categoría/total (Lift)	0,716	0,746	*1,484*	
Total		3768	474	2448	6690
Porcentaje [b]		56,30%	7,10%	36,60%	

*Relación categoría/total (Lift) = (a / b).

Fuente: Elaboración propia

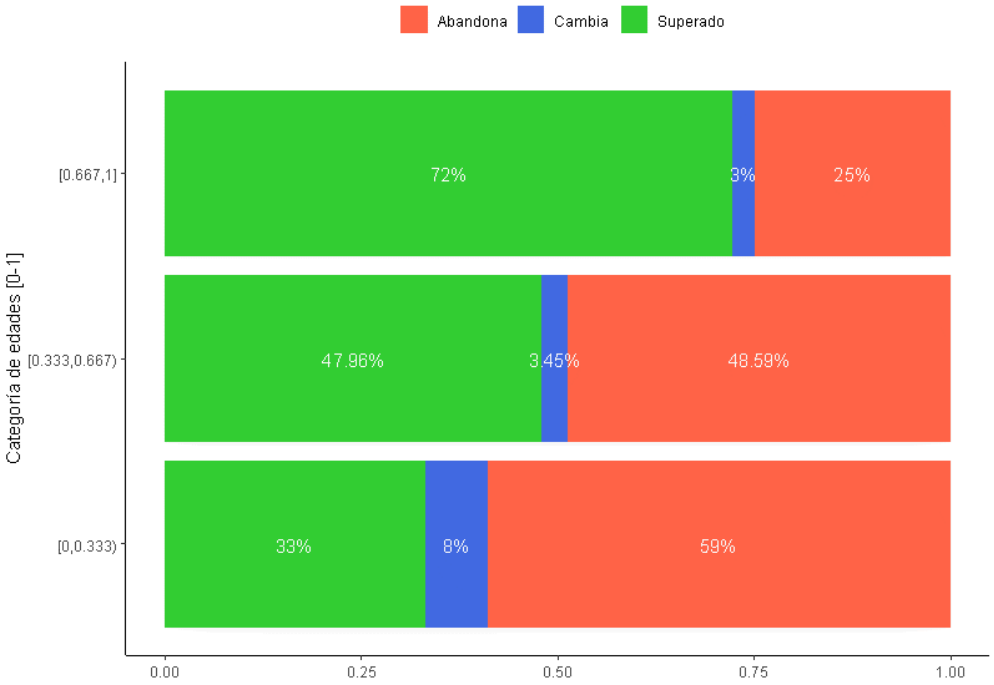
Basados en lo evidenciado en las Tablas 3,4 y 5. Si bien se observa, la "relación categoría/total (Lift)" ha conseguido captar categorías significativas en las variables del profesorado, por tanto, la transmisión consolidada y comprensible de los resultados lo hemos mostrado en la Gráfico 4. De ahí que, han destacado las etiquetas con valor superior a 1 y mostradas con colores más oscuros (rojizos). En concreto, la etiqueta Edad3 [0,667-1,00] con 1,976 en el estado académico *Superado* ha sido altamente significativa para los datos analizados. De manera general, tanto el segundo y tercer tercio de las variables experiencia docente, media de edad, Edad2 y Edad3 han influido en la superación o cambio del grado universitario. Es decir, el encauzamiento del éxito académico del estudiantado se potencia cuando la participación de este grupo de variables es superior o igual a dos tercios. Por otra parte, en la Gráfico 5, mostramos la incidencia que tuvo el profesorado experimentado (Edad3). De acuerdo con los resultados evidenciamos que cuanto mayor es la participación de ellos en el proceso de enseñanza, el alumnado tuvo mayores casos de éxito académico.

Gráfico 4. Mapa de calor con las categorías variables y el estado académico del alumnado, el color más oscuro (rojizo) indica una alta asociación entre la categoría y la superación del grado universitario. La mayor concentración de asociaciones se ha dado en el segundo y tercer tercio de las variables categorizadas. La categoría Edad3 [0,667-1,00] muestra una alta asociación.



Fuente: Elaboración propia

Gráfico 5. Proporción de participación del profesorado experimentado en el proceso de enseñanza del alumnado y el estado académico del alumnado, cada categoría contiene tres barras horizontales que van de izquierda a derecha. La primera relacionada con los que han superado el grado académico, la segunda de los que han cambiado de grado y la tercera de los que han abandonado. La mayor proporción del alumnado que ha superado el grado académico estuvo centrado en docentes con participación superior al 66%, concretamente la categoría [0.66-1].



Fuente: Elaboración propia

5 Discusión y conclusiones

En respuesta a las preguntas de investigación sobre la compatibilidad y qué factores del profesorado eran influyentes para que el alumnado consiguiera el éxito académico. En cierta medida, existen estudios análogos tal como se muestra la literatura citada, sin embargo, nos hemos centrado en evidenciar y facilitar la comprensión de resultados relevantes sobre los factores y vínculos significativos que llevan al alumnado a superar el grado universitario. En este sentido, el proceso de enseñanza en las titulaciones universitarias es efectivo cuando la participación del profesor con experiencia y madurez académica y los grupos de edades del profesorado se distribuyen proporcionalmente entre el segundo y tercer tercio del total de profesores que imparten docencia en el primer curso, esto promueve un ambiente de confianza, además, consolida la positividad y motiva al alumnado a un mayor compromiso para alcanzar el éxito académico.

Respecto a la experiencia del profesorado, nuestros resultados coinciden con Pascarella et.al (1996) que entre los hallazgos encontrados han sugerido que

la práctica docente eficaz influye positivamente en el aprendizaje, además que también incrementa el número de alumnos con éxito académico. De hecho, Roksa y Whitley (2017) afirman que la madurez y el tipo de enseñanza del docente, a través de la iteración alumno-profesor han contribuido como precursor en el alumnado para superar el grado universitario. Por otra parte, Boluda y López (2012) en su investigación expresan que la "calidad" del profesorado es un poderoso predictor que se relaciona directamente con el rendimiento del alumnado y es posiblemente uno de los componentes más decisivos de cualquier proceso formativo. De igual forma, el éxito académico no solo está ligado a las actividades y cualidades del profesorado, sino también a la calidad del esfuerzo realizado por el alumno (Valadas, Almeida, & Araújo, 2017). A pesar de estos hallazgos, los resultados de este estudio deben interpretarse con cierta cautela, ya que los datos solo representan a una institución y los resultados de este estudio pueden no ser generalizables a otras universidades.

En cuanto a las edades del profesorado, establecemos una idea clara, conviene distinguir entre profesores experimentados y noveles. Dado que la edad, como tal, puede ser discriminatoria si no la contextualizamos adecuadamente. Apliquemos esta distinción al estudio de Fogarty, Wang y Creek (1983), donde ellos observaron que el profesorado experimentado tenían en cuenta una mayor variedad de objetivos e instrucciones para la toma de decisiones en el aula; curiosamente, lo contrario ocurría con los profesores noveles. No obstante, los novatos eran más propensos en detectar signos de rendimiento académico del alumnado que los profesores experimentados. Dicho esto, y en consonancia con nuestros resultados, los profesores maduros (Edad2) y experimentados (Edad3) fueron eficaces en tutelar la superación del grado académico del alumnado, siempre que su participación comprendiera el segundo y tercer tercio del total de profesores que impartieron clase. En términos generales, los profesores con madurez y experiencia educativa generan un entorno académico fiable y positivo para los alumnos.

Consideramos que los resultados obtenidos en este estudio tienen implicaciones relevantes. Tenemos razones para creer que las implicaciones asociadas al éxito académico del alumnado deben ajustarse al contexto, política y normas universitarias y que, a pesar de ello, se sugiere la distribución del profesorado asignado a impartir clases en el primer curso. Más concretamente, hemos propuesto dos alternativas basadas en los resultados: La primera, estaba relacionada con la edad del profesorado experimentado y maduro que tuviera una tasa de participación superior al 33%. Todo ello, con el fin de retener al alumno y descartar la deserción universitaria. La segunda, sugerir que profesores con un nivel de madurez en la enseñanza superior al segundo y tercer tercio del total de profesores participe en el proceso de enseñanza.

En cuanto a las limitaciones de este estudio, está la necesidad de incorporar al estudio las características socioeconómicas del alumnado dado que existe mayor posibilidad de éxito académico en estudiantes con mayores recursos (Roksa & Kinsley, 2019). Por otra parte, van Herpen (2017) ha examinado las limitaciones que hemos considerado en el estudio y que tienen que ver con el perfil del alumnado, importando y mucho la autoeficacia dada la correspondencia con las características del profesorado. Como estudio posterior, se debe examinar

el factor familiar y socio económico del alumnado, ya que el refuerzo con programas extracurriculares, acceso a recursos y ayudas impulsan al alumnado a superar el grado universitario, a su vez, permitirá al centro universitario promover estrategias que encauzan al alumnado hacia el éxito académico.

6 Referencias bibliográficas

Aleksandrova, Y., & Parusheva, S. (2019). Social media usage patterns in higher education institutions - An empirical study. *International Journal of Emerging Technologies in Learning*, *14*(5), 108–121. https://doi.org/10.3991/ijet.v14i05.9720

Alyahyan, E., & Düştegör, D. (2020, December 1). Predicting academic success in higher education: literature review and best practices. *International Journal of Educational Technology in Higher Education*, Vol. 17, pp. 1–21. Springer. https://doi.org/10.1186/s41239-020-0177-7

Amida, A., Algarni, S., & Stupnisky, R. (2020). Testing the relationships of motivation, time management and career aspirations on graduate students' academic success. *Journal of Applied Research in Higher Education*. https://doi.org/10.1108/JARHE-04-2020-0106

Angervall, P. (2018). The academic career: a study of subjectivity, gender and movement among women university lecturers. *Gender and Education*, *30*(1), 105–118. https://doi.org/10.1080/09540253.2016.1184234

Araque, F., Roldán, C., & Salguero, A. (2009). Factors influencing university drop out rates. *Computers and Education*, *53*(3), 563–574. https://doi.org/10.1016/j.compedu.2009.03.013

Boluda, I. K., & López, N. V. (2012). El docente universitario y sus efectos en el estudiante. *Estudios Sobre Educacion*, *23*(23), 157–182. Retrieved from https://www.unav.edu/publicaciones/revistas/index.php/estudios-sobre-educacion/article/view/2055

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brin, S., Motwani, R., Ullman, J. D., & Tsur, S. (1997). Dynamic itemset counting and implication rules for market basket data. *ACM SIGMOD Record*, *26*(2), 255–264. https://doi.org/10.1145/253262.253325

Campbell, C. M., Smith, M., Dugan, J. P., & Komives, S. R. (2012). Mentors and college student leadership outcomes: The importance of position and process. *Review of Higher Education*, *35*(4), 595–625. https://doi.org/10.1353/rhe.2012.0037

Campbell, H. E., Steiner, S., & Gerdes, K. (2005). Student Evaluations of Teaching: How You Teach and Who You Are. *Journal of Public Affairs Education*, *11*(3), 211–231. https://doi.org/10.1080/15236803.2005.12001395

Chickering, Arthur W.;Gamson, Z. F. (1987). Seven Principles for Good Practice in Undergraduate Graduation. *AAHE Bulletin;*, *39*(7), 3–7.

Cho, Y., Kim, M., Svinicki, M. D., & Decker, M. L. (2011). Exploring teaching concerns and characteristics of graduate teaching assistants. *Teaching in*

Higher Education, *16*(3), 267–279. https://doi.org/10.1080/13562517.2010.524920
Clelia, P. B., José-Javier, B. A., Ángela, R. B., Natalia, P. G., Rodrigo, S. G., & Fabián, C. B. (2014). Student engagement and academic performance in the colombian university context. *RELIEVE - Revista Electronica de Investigacion y Evaluacion Educativa*, *20*(2), 1–19. https://doi.org/10.7203/relieve.20.2.4238
Crisp, G., Taggart, A., & Nora, A. (2015). Undergraduate Latina/o Students: A Systematic Review of Research Identifying Factors Contributing to Academic Success Outcomes. *Review of Educational Research*, *85*(2), 249–274. https://doi.org/10.3102/0034654314551064
Cunha, J. M., Miller, T., & Weisburst, E. (2018). Information and College Decisions: Evidence From the Texas GO Center Project. *Educational Evaluation and Policy Analysis*, *40*(1), 151–170. https://doi.org/10.3102/0162373717739349
Darling-Hammond, L. (2000). Teacher Quality and Student Achievement. *Education Policy Analysis Archives*, *8*(0), 1. https://doi.org/10.14507/epaa.v8n1.2000
de Boer, H., Donker, A. S., & van der Werf, M. P. C. (2014). Effects of the Attributes of Educational Interventions on Students' Academic Performance: A Meta-Analysis. *Review of Educational Research*, *84*(4), 509–545. https://doi.org/10.3102/0034654314540006
Fogarty, J. L., Wang, M. C., & Creek, R. (1983). A Descriptive Study of Experienced and Novice Teachers' Interactive Instructional Thoughts and Actions. *The Journal of Educational Research*, Vol. 77, pp. 22–32. Taylor & Francis, Ltd. https://doi.org/10.2307/27540012
Glogowska, M., Young, P., & Lockyer, L. (2007). Should I go or should I stay? *Active Learning in Higher Education*, *8*(1), 63–77. https://doi.org/10.1177/1469787407074115
Jacobi, M. (1991). Mentoring and Undergraduate Academic Success: A Literature Review. *Review of Educational Research*, *61*(4), 505–532. https://doi.org/10.3102/00346543061004505
Jepsen, C. (2005). Teacher characteristics and student achievement: Evidence from teacher surveys. *Journal of Urban Economics*, *57*(2), 302–319. https://doi.org/10.1016/j.jue.2004.11.001
Kara, N., Çubukçuoğlu, B., & Elçi, A. (2020). Using social media to support teaching and learning in higher education: An analysis of personal narratives. *Research in Learning Technology*, *28*, 1–16. https://doi.org/10.25304/rlt.v28.2410
Konrad, T., Wiek, A., & Barth, M. (2021). Learning processes for interpersonal competence development in project-based sustainability courses – insights from a comparative international study. *International Journal of Sustainability in Higher Education*, *ahead-of-p*(ahead-of-print). https://doi.org/10.1108/ijshe-07-2020-0231
Korhonen, V., & Törmä, S. (2016). Engagement with a teaching career - how a group of finnish university teachers experience teacher identity and professional growth. *Journal of Further and Higher Education*, *40*(1), 65–82. https://doi.org/10.1080/0309877X.2014.895301
Korobova, N., & Starobin, S. S. (2015). A comparative study of student

engagement, satisfaction, and academic success among international and american students. *Journal of International Students*, *5*(1), 72–85. Retrieved from http://jistudents.org

Le, T., Bolt, D., Camburn, E., Goff, P., & Rohe, K. (2017). Latent Factors in Student–Teacher Interaction Factor Analysis. *Journal of Educational and Behavioral Statistics*, *42*(2), 115–144. https://doi.org/10.3102/1076998616676407

Leal Filho, W., Shiel, C., & Paço, A. (2016). Implementing and operationalising integrative approaches to sustainability in higher education: the role of project-oriented learning. *Journal of Cleaner Production*, *133*, 126–135. https://doi.org/10.1016/j.jclepro.2016.05.079

Livengood, J. M. (1992). Students' motivational goals and beliefs about effort and ability as they relate to college academic success. *Research in Higher Education*, *33*(2), 247–261. https://doi.org/10.1007/BF00973581

Lizzio, A., Wilson, K., & Simons, R. (2002). University students' perceptions of the learning environment and academic outcomes: Implications for theory and practice. *Studies in Higher Education*, *27*(1), 27–52. https://doi.org/10.1080/03075070120099359

Marginson, S. (2014). Higher education and public good. In *Thinking About Higher Education* (Vol. 9783319032, pp. 53–69). Wiley/Blackwell (10.1111). https://doi.org/10.1007/978-3-319-03254-2-5

Mishra, B. K., & Sahoo, A. K. (2016). Evaluation of faculty performance in education system using classification technique in opinion mining based on GPU. In *Advances in Intelligent Systems and Computing* (Vol. 411, pp. 109–119). Springer. https://doi.org/10.1007/978-81-322-2731-1_10

Mishra, S. (2020, February 1). Social networks, social capital, social support and academic success in higher education: A systematic review with a special focus on 'underrepresented' students. *Educational Research Review*, Vol. 29, p. 100307. Elsevier Ltd. https://doi.org/10.1016/j.edurev.2019.100307

Nagda, B. A., Gregerman, S. R., Jonides, J., Von Hippel, W., & Lerner, J. S. (1998, September). Undergraduate student-faculty research partnerships affect student retention. *Review of Higher Education*, Vol. 22, pp. 55–72. Johns Hopkins University Press. https://doi.org/10.1353/rhe.1998.0016

Nasser-Abu Alhija, F. (2017). Teaching in higher education: Good teaching through students' lens. *Studies in Educational Evaluation*, *54*, 4–12. https://doi.org/10.1016/j.stueduc.2016.10.006

Pascarella, E. T., Edison, M., Hagedorn, L. S., Nora, A., & Terenzini, P. T. (1996). Influences on students' internal locus of attribution for academic success in the first year of college. *Research in Higher Education*, *37*(6), 731–756. https://doi.org/10.1007/BF01792954

R CoreTeam, D. C. (2019). A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*, Vol. 739, p. ISBN 3-900051-07-0-ISBN 3-900051-07-0. Vienna, Austria. https://doi.org/10.1007/978-3-540-74686-7

Respondek, L., Seufert, T., Stupnisky, R., & Nett, U. E. (2017). Perceived academic control and academic emotions predict undergraduate university student success: Examining effects on dropout intention and achievement.

Frontiers in Psychology, *8*(MAR), 243. https://doi.org/10.3389/fpsyg.2017.00243
Roksa, J., & Kinsley, P. (2019). The Role of Family Support in Facilitating Academic Success of Low-Income Students. *Research in Higher Education*, *60*(4), 415–436. https://doi.org/10.1007/s11162-018-9517-z
Roksa, J., & Whitley, S. E. (2017). Fostering Academic Success of First-Year Students: Exploring the Roles of Motivation, Race, and Faculty. *Journal of College Student Development*, *58*(3), 333–348. https://doi.org/10.1353/csd.2017.0026
Romanski, P., & Kotthoff, L. (2016). *FSelector: Selecting Attributes*. Retrieved from https://cran.r-project.org/package=FSelector
Salminen-Tuomaala, M., & Koskela, T. (2020). How can simulation help with learning project work skills? Experiences from higher education in Finland. *Educational Research*, *62*(1), 77–94. https://doi.org/10.1080/00131881.2020.1711791
Sanvitha Kasthuriarachchi, K. T., Liyanage, S. R., & Bhatt, C. M. (2018). A data mining approach to identify the factors affecting the academic success of tertiary students in sri lanka. In *Lecture Notes on Data Engineering and Communications Technologies* (Vol. 11, pp. 179–197). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/978-3-319-68318-8_9
Sarahí Abarca, Mireya; Gómez Pérez, M. T. C. V. M. L. (2015). Análisis de los factores que contribuyen al éxito a cadémico en estudianes universitarios. *Revista Internacional de Educación y Aprendizaje*, *3*(2).
Shetu, S. F., Saifuzzaman, M., Moon, N. N., Sultana, S., & Yousuf, R. (2021). Student's performance prediction using data mining technique depending on overall academic status and environmental attributes. *Advances in Intelligent Systems and Computing*, *1166*, 757–769. Springer. https://doi.org/10.1007/978-981-15-5148-2_66
Souchon, N., Kermarec, G., Trouilloud, D., & Bardin, B. (2020). Influence of teachers' political orientation and values on their success prediction toward students from different socioeconomic background. *Revue Europeenne de Psychologie Appliquee*, *70*(5), 100553. https://doi.org/10.1016/j.erap.2020.100553
Struyven, K., Dochy, F., & Janssens, S. (2003). Students' Perceptions about New Modes of Assessment in Higher Education: A Review BT - Optimising new modes of assessment: In search of qualities and standards. *Optimising New Modes of Assessment: In Search of Qualities and Standards*, *1*(Chapter 8), 171–223. https://doi.org/10.1007/0-306-48125-1_8
Tinto, V. (1975). Dropout from Higher Education: A Theoretical Synthesis of Recent Research. *Review of Educational Research*, *45*(1), 89–125. https://doi.org/10.3102/00346543045001089
Trolian, T. L., Jach, E. A., & Archibald, G. C. (2020). Shaping Students' Career Attitudes toward Professional Success: Examining the Role of Student-Faculty Interactions. *Innovative Higher Education*. https://doi.org/10.1007/s10755-020-09529-3
Valadas, S. T., Almeida, L. S., & Araújo, A. M. (2017). The Mediating Effects of

Approaches to Learning on the Academic Success of First-Year College Students. *Scandinavian Journal of Educational Research*, *61*(6), 721–734. https://doi.org/10.1080/00313831.2016.1188146

Van Den Berg, M. N., & Hofman, W. H. A. (2005). Student success in university education: A multi-measurement study of the impact of student and faculty factors on study progress. *Higher Education*, *50*(3), 413–446. https://doi.org/10.1007/s10734-004-6361-1

van Herpen, S. G. A., Meeuwisse, M., Hofman, W. H. A., Severiens, S. E., & Arends, L. R. (2017). Early predictors of first-year academic success at university: pre-university effort, pre-university self-efficacy, and pre-university reasons for attending university. *Educational Research and Evaluation*, *23*(1–2), 52–72. https://doi.org/10.1080/13803611.2017.1301261

Vo, T. N. C., Nguyen, H. P., & Vo, T. N. T. (2016). Making kernel-based vector quantization robust and effective for incomplete educational data clustering. *Vietnam Journal of Computer Science*, *3*(2), 93. https://doi.org/10.1007/s40595-016-0060-6

Walder, A. M. (2016). *Pedagogical Innovation in Canadian higher education: Professors' perspectives on its effects on teaching and learning*. https://doi.org/10.1016/j.stueduc.2016.11.001

Winterer, E. R., Froyd, J. E., Borrego, M., Martin, J. P., & Foster, M. (2020, December 1). Factors influencing the academic success of Latinx students matriculating at 2-year and transferring to 4-year US institutions—implications for STEM majors: a systematic review of the literature. *International Journal of STEM Education*, Vol. 7, p. 34. Springer. https://doi.org/10.1186/s40594-020-00215-6

Young, P., Glogowska, M., & Lockyer, L. (2007). Conceptions of early leaving: A comparison of the views of teaching staff and students. *Active Learning in Higher Education*, *8*(3), 275–287. https://doi.org/10.1177/1469787407081882

Anexo I

Descripción de las variables usadas para el estudio.

Variables	Descripción	Tipo	Rango de datos
Facultad	Nombre de la Facultad	Categórica	5 Facultades
GradoUniversitario	Titulación académica en la que se ha matriculado el estudiante	Categórica	22 titulaciones
Edad1	Cantidad de profesores menores de 45 años	Numérica	0-15
Edad2	Cantidad de profesores entre 46 y 60 años	Numérica	0-15
Edad3	Cantidad de profesores mayores de 61 años	Numérica	0-15
ExperienciaDocente	Media de experiencia docente del profesorado (años)	Numérica	4-35
MediaEdadProfesor	Media de edad del profesorado	Numérica	35-75
TitulaciónDocente	Ponderación del profesorado que dictó clases en primer curso	Numérica	0-1
MediaNotas	Media de notas del primer curso del estudiante	Numérica	0-10
MediaAsistencia	Media de asistencia a clases del estudiante	Numérica	0-100
TiempoSupera	Cantidad de cursos matriculado para superar el primer año	Numérica	0-5
TasaSuperacion	Ponderación de convocatorias a exámenes	Numérica	0-3
CuentaGrados	Cantidad de grados matriculado en primer curso	Numérica	1-5
CambioGrado	Indicador para conocer si el grado académico es el inicial o ha cambiado de grado.	Numérica	0-1
Clase	Estado académico del estudiante al finalizar el grado académico	Categórica	Abandona, Cambio, Superado

8. Adopción del aprendizaje combinado en la educación superior: percepción y evaluación del profesorado

- Jorge Guanin-Fajardo, Jorge Casillas, Adolfo Elizondo-Saltos
  - Status: Publicado (Revista Electrónica Calidad en la Educación Superior)
  - WoS “Education & Educational Research”, JIF 2023: 0.1, 699/756 (Q4)

Adopción del aprendizaje combinado en la educación superior: percepción y evaluación del profesorado

Jorge Humberto Guanin-Fajardo jorgeguanin@uteq.edu.ec

Facultad de Ciencias de la Ingeniería, Universidad Técnica Estatal de Quevedo, Ecuador, Los Ríos, Quevedo
CP:EC120509

Jorge Casillas casillas@decsai.ugr.es

Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Granda, España, Granada, CP:18014

Adolfo-Hernán Elizondo-Saltos aelizondos@uteq.edu.ec

Facultad de Ciencias Económicas y Sociales, Universidad Técnica Estatal de Quevedo, Ecuador, Los Ríos, Quevedo
CP:EC120509

DOI: https://doi.org/10.22458/caes.v14i2.4915

Resumen

La pandemia del COVID-19 ha tenido un gran impacto en todo el mundo, afectando no solo la salud pública sino también la economía y la educación. Este estudio se enfocó en examinar la relación entre los distintos factores del profesorado para atender el aprendizaje del estudiantado bajo situación de pandemia. Valiéndose de las oportunidades, habilidades y capacidades del profesorado en el contexto universitario latinoamericano. Se llevó a cabo un estudio cuantitativo con un diseño descriptivo transversal y campo no experimental, donde se aplicó un cuestionario autoadministrado a profesores con más de 15 años de experiencia en seis universidades nacionales e internacionales. Los resultados del estudio indicaron la presencia de cuatro factores positivos, un factor de riesgo, y un factor negativo. La relevancia de los factores fue obtenida por medio de métricas estadísticas propuestas en el estudio. De ahí que los resultados desvelaran que una planificación cuidadosa y medida en la formación de nuevas herramientas educativas para el profesorado estimula de forma positiva a éstos, además de mitigar el riesgo y superar la fase negativa en la enseñanza virtual.

Palabras clave: Educación superior, metodología mixta, COVID-19, contexto universitario, Análisis educacional.

Adoption of blended learning in higher education: faculty perception and evaluation.

Abstract

The COVID-19 pandemic has had a major impact worldwide, affecting not only public health, but also the economy and education. This study focused on examining the relationship between different teacher factors in addressing student learning under pandemic conditions. Using the opportunities, skills and capabilities of faculty in the Latin American university context. A quantitative study was conducted with a descriptive cross-sectional and non-experimental field design, where a self-administered questionnaire was applied to professors with more than 15 years of experience in six national and international universities. The results of the study indicated the presence of four positive factors, one risk factor and one negative factor. The relevance of the factors was obtained by means of statistical metrics proposed in the study. Thus, the results revealed that careful and measured planning in the training of new educational tools for teachers stimulates them positively, in addition to mitigating the risk and overcoming the negative phase in virtual teaching.

Keywords: Higher education, blended learning, COVID-19, university context, educational analysis.

1 INTRODUCCIÓN

La preferencia del aprendizaje combinado o aprendizaje mixto (Blended Learning) fue urgido por las Instituciones de Educación Superior (IES) a raíz del COVID-19 que generó una multitud de consecuencias considerables desde el ámbito laboral, sanitario, económico, social y académico (Miranda-Chavez et al., 2022; Ferreira et al., 2021; Nicola et al., 2020). Varios países del mundo carecieron de un plan de contingencia efectivo para mitigar sus efectos devastadores. Las IES tampoco estuvieron preparadas para afrontar la sorpresiva pandemia que se derivó en confinamiento obligatorio, suspendiéndose para ello todo tipo de actividad académica. De ahí que, aproximadamente 600 millones de estudiantes que asistían de forma presencial en todo el mundo se vieron afectados por el cierre de los centros educativos (Sanchez-Pujalte et al., 2021).

A la vista de los trabajos de (Obeidat et al., 2020; Rojas et al., 2020; Sá & Serpa, 2020), los autores entre sus resultados manifiestan que durante la finalización del curso académico las IES convergieron en la toma de decisiones que cubrirían en particular tres objetivos: a) Flexibilizar las tareas docentes; b) Sortear dificultades generadas por la no presencialidad; c) Buscar alternativas al área docente y su evaluación. De este modo, su adaptación al aprendizaje combinado fue apresurado y requirió valorar el contexto y metodología a utilizar; ya que es poco recomendable una transposición, sin una reflexión sobre el diseño educativo (Yildiz et al., 2022; Santana-Sardi et al., 2020). A diferencia del estudiantado con modalidad de educación a distancia que resultó menos afectado debido al empleo de sistemas de gestión del aprendizaje (LMS- Learning Management Software) como herramienta de soporte para armonizar la interacción de enseñanza-aprendizaje entre profesores y estudiantes, todo esto estribado por las Tecnologías de la Información y Comunicación (TIC's) (Ortega et al., 2021; Teichgräber et al., 2021; The Chronicle of Higher Education, 2020).

Queda claro, pues, que las nuevas TIC han resultado claves para abordar el proceso de enseñanza-aprendizaje en la universidad. Los efectos de la pandemia sobre el ámbito educativo, entre otras cosas, han servido para concienciar a todos los agentes implicados en la educación superior de la necesidad de avanzar en una auténtica cultura digital para que se produzca una verdadera transformación. Es evidente que las TIC's fueron clave para enfocar la enseñanza-aprendizaje en las IES hacia una reflexión profunda en la necesidad de avanzar en una autentica cultura digital a efectos de que se produzca una verdadera transformación digital.

El acertado uso de los sistemas LMS en la educación a distancia fue fundamental para que la educación tradicional (presencial) ubicara su actividad académica hacia el uso de sistemas LMS y otras herramientas tecnológicas necesarias para implantar el aprendizaje combinado, todo ello debido al obligado confinamiento del alumnado. Esta adopción era impensable dada la falta de políticas reguladoras entre los modelos de enseñanza (tradicional-combinado) y, en base a los hechos, se generaron políticas paliativas para la transición, además de la poca o nula preparación del profesorado que generó frustración y agobio, además de los riesgos y negatividad del profesorado para gestionar los medios o recursos que demandaban los sistemas LMS. No obstante, el profesorado se adaptó a este nuevo modelo de enseñanza combinado, ejerciendo para ello tanto su vida cotidiana y academia en el mismo sitio de residencia, lo cual evidenció que tienen capacidad y competencia para implementar el aprendizaje combinado.

Por otra parte (Obeidat et al., 2020; Rojas et al., 2020; Sá & Serpa, 2020) corroboran que la adaptación del aprendizaje combinado valorado dentro del contexto universitario coincide en que el principal obstáculo para aplicar la modalidad fueron las habilidades y capacidades

del profesorado para desempeñarla de forma positiva y agradable. Aunque muchos estudiantes y profesores siguen mostrándose escépticos ante el aprendizaje combinado, hay otros que están muy satisfechos. Dado que, el entorno combinado ofrece múltiples modalidades de aprendizaje, interactividad significativa y conexiones sostenidas con profesores y estudiantes, existiendo aceptación y preferencia en los usuarios para esta modalidad. (Banerjee, 2022).

En síntesis, la pandemia de COVID-19 planteó retos importantes para los sistemas educativos y sociales de los países hispanoamericanos. Estos retos y estas lecciones brindan hoy la posibilidad de replantearnos el propósito de la educación y su papel en el sostenimiento de la vida y la dignidad humana, esta crisis ofreció una oportunidad sin precedentes para aumentar la capacidad de recuperación del sistema educativo y transformarlo en sistema equitativo e inclusivo que contribuya al cumplimiento del compromiso colectivo asumido por Naciones Unidas en la Agenda 2030 en los Objetivos de Desarrollo Sostenible (Educación de calidad). Tanto la enseñanza docente y la metodología mixta o combinada se refieren al mismo contexto, es decir, el procedimiento que el profesor o grupo de profesores emplean para impartir las clases.

Por esta razón, como objetivo central del estudio se pretende examinar la asociación de factores positivos, de riesgos y negativos; oportunidades de las capacidades y habilidades del profesorado en el contexto de la educación superior (Hispanoamérica). Para ello, se han planteado las siguientes preguntas de investigación: i) ¿Qué nivel de conocimiento sobre la enseñanza mixta tiene el profesorado? ii) ¿Cómo intervienen las capacidades y habilidades del profesorado en la metodología mixta? Para dar respuestas a estas preguntas se examina el contraste de variables, así como una exploración gráfica que acompañe a los resultados y que ayude a interpretarlos. El resto del trabajo está organizado de la siguiente manera: En la sección 2, planteamos la metodología utilizada y las valoraciones estadísticas usadas. La sección 3, se presentan los principales resultados obtenidos. La sección 4 enfatiza la discusión de los resultados. La sección 5 con las principales conclusiones del trabajo.

2 METODOLOGÍA

El enfoque de esta investigación es mixto; es decir cualitativo-cuantitativo, para la metodología se han empleado los métodos analítico, inductivo y deductivo. Entre las técnicas para la recolección de la información se ha aplicado una encuesta, y como instrumento se elaboró un cuestionario de 24 preguntas de carácter cerrado, las mismas que fueron creadas usando el formulario de Google, dichas encuestas se proporcionaron a una muestra seleccionada del profesorado de las diferentes Instituciones de Educación Superior a nivel Hispanoamericana y una de Marruecos (con profesores hispanos). Luego de recolectada la información fue calculada su confiabilidad por medio del coeficiente Alfa de Cronbach dicho valor fue de 0,7952 comprobándose la coherencia y consistencia de los datos. Después, se realizó un estudio exploratorio para filtrar las variables con mayor relevancia, además se realizó el contraste de variables y computo del estadístico chi-cuadrado (χ^2) para comprobar si las frecuencias observadas de una o más categorías se ajustan a las esperadas, la entropía para medir el grado de incertidumbre en los datos; y, la razón de verosimilitudes. Todo esto, facilita crear de gráficos para entender mejor los resultados. La tabulación, cálculo estadístico de métricas y gráficos se realizó con el software estadístico SPSS[1] (versión académica) que aportó las evidencias empíricas de este estudio.

[1] https://www.ibm.com/es-es/products/spss-statistics

2.1 Población y Muestra

Se tomó como población a un grupo determinado de profesores participantes de siete países como son: Colombia, Brasil, Chile, Marruecos, México, España y Ecuador. De los cuales los participantes para esta encuesta fueron, el 17.5% mujeres y el 82.5% hombres, en su mayoría todos desarrollan su labor docente en la educación superior y cuentan con un promedio de 15 años de experiencia en las diferentes titulaciones (grados universitarios) que ofertan las IES.

2.2 Instrumento y proceso de recogida de datos

El presente estudio parte de un diseño de investigación de corte cuantitativo, apoyado en un enfoque exploratorio-descriptivo y de correlación, donde el método usado fue la encuesta en línea debido a la situación excepcional generada por la COVID-19. Para ello, se utilizó la herramienta de administración de encuesta “Formularios de Google”. Al momento de suministrarla a los participantes esta herramienta entre otras cosas ofreció ventajas tales como: ubicuidad, facilidad, eficiencia para recopilar y extraer datos en múltiples formatos. El cuestionario de preguntas estuvo organizado en cuatro dimensiones: la primera, con los datos identificativos del centro de estudio; la segunda, conduce al conocimiento de la enseñanza mixta con preguntas tipo selección múltiple y escala de Likert (0-Nada de acuerdo, 5-Totalmente de acuerdo); la tercera, sobre las capacidades y habilidades de la metodología mixta; por último, la formación docente en la enseñanza mixta. Los profesores participantes tuvieron acceso al formulario de la encuesta por diferentes canales de comunicación correo electrónico, mensajería instantánea y redes sociales. El enlace estuvo operativo durante el curso académico 2020-2021.

3 RESULTADOS

3.1 Estudio descriptivo

La Figura 1, proyecta de forma proporcionada las respuestas de cada pregunta (Sí, No, NC), se evidencia que el “Sí” en las preguntas tiene mayor tendencia, es decir, presenta una instantánea de los datos. En cambio, el estudio exploratorio de las variables relevantes se muestra detalladamente en la Tabla 1. En cuanto a la pregunta 1, sobre el conocimiento de la metodología mixta, se ha podido visualizar que el 80% del profesorado conoce sobre la metodología mixta empleada en la educación superior, el 17.5% no conoce sobre este método y el 2.5% eligió por no contestar. Con lo cual, se puede decir que los docentes cuentan con un nivel intermedio-alto sobre el uso de las TICs.

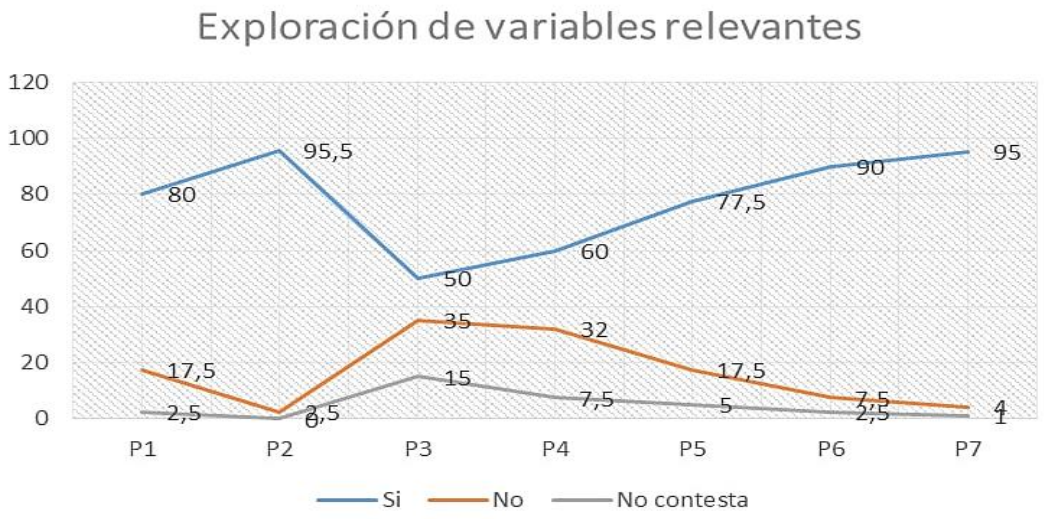
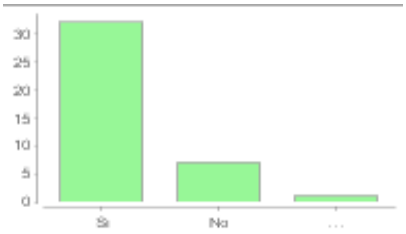
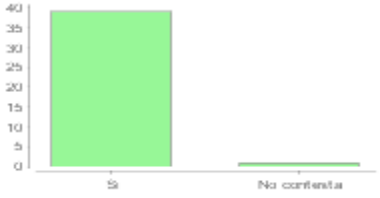
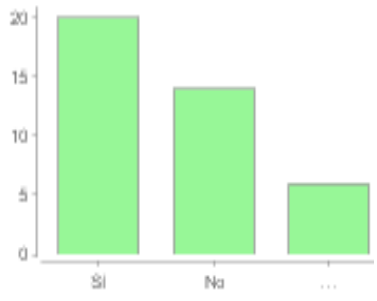
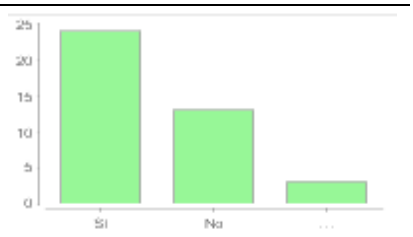
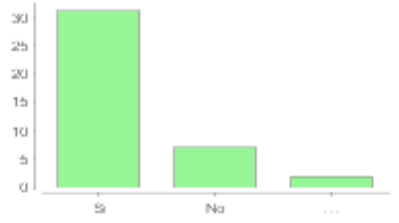
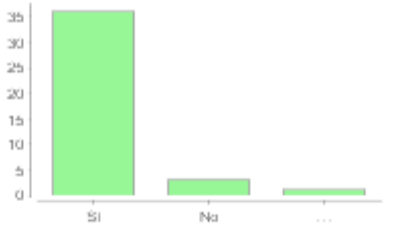
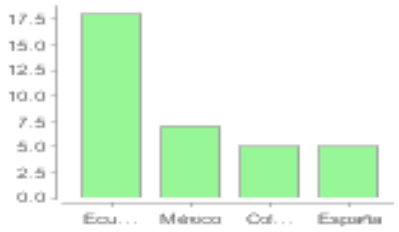


Figura 1. Síntesis del grupo de variables relevantes (abscisa). Las líneas representan respuestas de cada pregunta. El punto de quiebre entre cada línea y el número equivale a la proporción de la respuesta contestada.

Fuente: Elaboración propia.

Tabla 1. Estudio exploratorio de variables relevantes.

Ítem	Variable	Valores	Gráfico
1	**Pregunta 1 de la dimensión 2, ¿Conoce usted sobre la metodología de docencia mixta?**	Si=80% No=17,5% No contesta=2,5%	
2	**Pregunta 2 dimensión 2, ¿Es interesante, combinar ambas?**	Si=97,5% No contesta=2,5%	
3	**Pregunta 3 de la dimensión 2, ¿La Universidad ha tenido un plan de mejora para esta nueva modalidad de enseñanza?**	Si=50% No=35% No contesta=15%	

4	**Pregunta 1, dimensión 4¿La Universidad lo ha preparado para aplicar la enseñanza mixta?**	Si=60% No=32% No contesta=7,5%	
5	**Pregunta 3, dimensión 2. ¿Cree usted que la puede usar en cualquier ámbito educativo y hacer mejoras de ser el caso?**	Si=77,5% No=17,5% No contesta=5%	
6	**Pregunta 4, dimensión 2. ¿De volver a la normalidad, considera importante incorporar esta metodología al plan de estudio?**	Si=90% No=7,5% No contesta=2,5%	
7	**Pregunta 1, dimensión 1. País de los profesores participantes en las encuestas**	Ecuador=45% México=17.5% Colombia=12.5% España=12.5% Marruecos=4% Brasil=4.5% Chile=4%	

Otra de pregunta importante dentro del estudio fue referida a que si los docentes están de acuerdo en combinar ambas metodologías en esta nueva modalidad de enseñanza (Pregunta 2). Las respuestas indicaron que el 97% de docentes estuvo de acuerdo en aplicar ambas metodologías, el otro 2,5% no contestaron, esto reveló que el profesorado estuvo interesado en aplicar ambas metodologías en la enseñanza superior como una herramienta adicional en el marco pedagógico de enseñanza.

Por otra parte, en la pregunta 3, el profesorado considera que, si la Universidad ha tenido un plan de mejoras para esta nueva modalidad de enseñanza, se obtuvo como resultados que el 50% de los docentes consideran que si han contado con un plan de mejoras en la enseñanza por parte de la universidad; el 35% considera que la universidad carece de un plan de mejoras para la nueva modalidad y, el 15% no contestaron. En general, la mayoría de las IES sí han contado con un plan de mejoras en la nueva modalidad.

En lo que corresponde a la pregunta 4, se consultó si la Universidad ha preparado al profesorado para aplicar el aprendizaje combinado, se obtuvo como resultado que el 60% del profesorado considera que si han recibido preparación de la Universidad; el 32.5% consideran que falta capacitar al profesorado de las IES y, el 7.5% no contestaron. Por lo que cual sí existe una preparación aceptable del profesorado para aplicar el aprendizaje combinado en la educación superior directamente en el proceso de enseñanza aprendizaje. De la misma forma, la pregunta 5, respecto a la creencia sobre la predisposición de usar el aprendizaje combinado en cualquier ámbito

educativo y hacer las mejoras del caso, se confirmó que el 77.5% del profesorado opinó que si se puede usar en cualquier ámbito educativo; el 17.5% consideró que no se la puede usar y, el 5% no contestaron. Con los resultados obtenidos se mostró que los docentes estuvieron de acuerdo en usar estas metodologías en cualquier ámbito educativo.

Uno de los principales desafíos en tiempo de pandemia era volver a la normalidad (clases presenciales). Por esta razón se consultó al profesorado que, si de volver a la normalidad debería ser importante incorporar esta metodología al plan de estudios, donde el 90% de los docentes decidió que si se debe incorporar esta metodología en los nuevos planes de estudios; el 7,5% opinan que no están de acuerdo con aplicar esta metodología y, el 2,5% no contestaron. Es así que los docentes admiten conciliar la metodología en los nuevos planes de estudios en tiempos normalidad. Los países participantes en el cuestionario según las encuestas registradas son docentes hispanoamericanos. Donde Ecuador, tuvo mayor participación; México en el segundo lugar, España y Colombia en tercer lugar, Marruecos, Brasil y Chile en quinto lugar.

3.1.1 Estudio de contraste de variables.

En el contexto de la formación docente en cuanto al aprendizaje combinado, es importante examinar las capacidades y habilidades del profesorado y valorar su competencia digital. Para ello, se han utilizado preguntas planteadas en la cuarta dimensión y se ha hecho referencia a dos documentos clave: el Plan de Acción de Educación Digital (2021-2027) de la Unión Europea (European Union, n.d.); y, el Marco de Competencias de los Docentes en Materia de TIC de la UNESCO (UNESCO, 2019). Ambos documentos enfatizan la importancia del uso de nuevas tecnologías en la educación, y la necesidad de que los docentes estén preparados para desempeñar nuevas funciones y utilizar nuevas pedagogías y métodos.

Para medir el nivel de conocimiento de la cuarta dimensión, se realizó un cruce de preguntas y se obtuvieron resultados estadísticos, que se muestran en la Tabla 2. Para graduar las observaciones en cuanto a las competencias digitales del docente, se establecieron tres categorías: positivo, donde el docente domina ampliamente las competencias digitales; riesgo, donde puede existir insuficiencia del dominio de las competencias digitales; y negativo, donde el docente necesita capacitación total o parcial en cuanto al dominio de las competencias digitales. Estas categorías están fundamentadas en los criterios establecidos por el Plan de Acción de la Educación Digital y el Marco de Competencias de los Docentes en Materia de TIC. Por lo que, la evaluación de la competencia digital del profesorado es fundamental para adaptarse a los nuevos desafíos educativos y mejorar la calidad de la enseñanza.

Tabla 2. Tabla cruzada entre las preguntas del cuestionario. Se calculó el valor de chi-cuadrado, la razón de verosimilitudes y el desorden de los datos (entropía) entre la pregunta central y pivote.

Ítem	Pregunta central	Pregunta pivote			Observación
		Chi-cuadrado	Razón de verosimilitudes	Entropía	
1	Si me capacitan, estoy dispuesto a usar la metodología.	Me da fatiga al organizar las actividades y tareas del entorno virtual			*Positivo*, ya que se cuenta con la predisposición del docente para recibir capacitaciones periódicas e interactuar con la metodología mixta.
		0,987	0,971	0,096	

2	Si me capacitan, estoy dispuesto a usar la metodología.	En esta metodología pierdo el dominio como profesor			*Positivo*, dada la inclinación abierta del docente para recibir capacitaciones periódicas e interactuar con la metodología mixta, aunque perciba la pérdida de dominio como profesor.
		0,970	0,900	0,121	
3	Me da fatiga al organizar las actividades y tareas del entorno virtual.	Tengo las bases pedagógicas para llevar paralelamente ambas modalidades			*Positivo*, el manejo pedagógico del profesor en ambas modalidades frente a la fatiga, a pesar de ello y al poco uso o desconocimiento de herramientas digitales que faciliten la organización de tareas en asignaturas del entorno virtual.
		0,933	0,767	0,088	
4	Me da fatiga al organizar las actividades y tareas del entorno virtual.	Tengo suficiente conocimiento en el manejo de dispositivos digitales (WiFi, móvil/celular, tabletas, etc.) y puedo aplicar esta metodología			*Positivo*, aunque el cambio abrupto de metodología de enseñanza generó agotamiento al docente, y que su conocimiento fue sólido en conectividad de dispositivos digitales supo trascender en aplicar el aprendizaje combinado.
		0,902	0,846	0,086	
5	Me da fatiga al organizar las actividades y tareas del entorno virtual	Los recursos e infraestructura que tiene la Universidad facilitan la aplicación de la metodología			*Riesgo*, es posible que el docente no esté muy de acuerdo con el manejo y uso de las TICS para impartir la cátedra. Mostrándose reacio a todo tipo de cambio por falta de capacitación.
		0,728	0,536	0,139	
6	Tengo las bases pedagógicas para llevar paralelamente ambas modalidades.	Las asignaturas asignadas, entorpece el uso de la metodología			*Negativo*, a pesar de tener asignado en el plan curricular asignaturas que carecen de afinidad para llevar la metodología, el profesor tiene bases pedagógicas para aplicar ambas modalidades de enseñanza.
		0,525	0,258	0,164	

Fuente: Elaboración propia

Se considera que los datos recolectados se puedan disponer de información relevante ya que se necesita responder a las preguntas del estudio, por ello, en la Tabla 2, se muestra el valor de las métricas propuestas (chi-cuadrado, verosimilitud y entropía). Conjuntamente el valor de chi-cuadrado superior o igual a 0,05; y, la razón de verosimilitud con valor superior o igual a 0,75 son valores explícitos para estimar el tipo de factor entre las variables de contraste, que son: positivo, mientras que valores de riesgo entre 0,75 y 0,50, y por último, valores inferiores a 0,5 se consideran factores negativo.

Hechas las consideraciones anteriores se observa que en el ítem 1, el valor chi-cuadrado fue 0,987 y 0,97 de verosimilitud, siendo estos resultados de factor positivo respecto a la organización y clasificación de actividades causa fatiga en el docente, a pesar de ello, se siente predispuesto a recibir capacitación para agilizar su acceso y comprensión. El ítem 2, en cambio refleja valores de 0,97 en chi-cuadrado y 0,90 en la razón de verosimilitud. En el contraste de estas preguntas se establece un factor positivo, en este caso el docente percibe una pérdida del dominio sobre el estudiante y a pesar de ello está abierto a capacitarse y utilizar el aprendizaje combinado como método de enseñanza.

Por otra parte, el ítem 3 hace referencia a la fatiga causada por las actividades y la virtualidad de la metodología de enseñanza frente al dominio pedagógico que el profesor lo toma como un riesgo, dado que el profesorado, a pesar de tener bases pedagógicas, puede encontrarse limitado por el desconocimiento de herramientas que faciliten elaborar el material de clase en el modo virtual, el valor chi-cuadrado de este ítem fue 0,933 y 0,767 de razón de verosimilitud. Dando pie a que esta vinculación sea positiva a pesar de carecer de asesoría especializada en organizar tareas virtuales. De igual manera, el ítem 4 hace énfasis a la fatiga de actividades y tareas del entorno virtual, aunque el profesorado tenga conocimiento suficiente de conectividad en dispositivos digitales, el valor chi-cuadrado fue 0,902 y la razón de verosimilitud 0,846 derivándose en un factor positivo en el análisis de ambas preguntas.

Otro punto de interés que salta a la vista es el ítem 5 que fue catalogado como un factor de riesgo, ya que el contraste ha reflejado inexperiencia en el profesorado para armonizar las tecnologías digitales para fines académicos. Si bien la Universidad dispone de recursos tecnológicos que facilitan la aplicación del aprendizaje combinado y, algunos profesores pueden usarlas con destrezas en su vida ordinaria, muchos no combinan estos recursos con las asignaturas y actividades del aula. El profesor percibe entonces fatiga para organizarlas, dentro de los resultados el valor chi-cuadrado al cruce de estas preguntas ha sido de 0,728 y la razón de verosimilitudes 0,536. En este ítem se debe considerar que el confinamiento y la privación de una asesoría dirigida limita y produce frustración al docente en la organización de su clase.

Por otra parte, se encontró una relación excluyente (efecto negativo) en el ítem 6, donde el docente con bases pedagógicas para llevar adelante ambas modalidades se enfrenta a un reto mayor, que son las asignaturas incompatibles con el aprendizaje combinado, por ejemplo, se puede citar “pruebas de laboratorio” o “prácticas de campo agrícola” asignaturas que, en el modo virtual, no pueden desarrollarse de buena manera dado que para el estudiante es importante mantenerse en el laboratorio o campo agrícola y ser parte de la experiencia que conllevan las asignaturas prácticas. El resultado de este ítem tuvo un valor de chi-cuadrado 0,525 y la razón de verosimilitudes 0,258. El profesorado de forma general se somete a todo tipo de desafío dentro del ámbito de la educación superior, aplicar el cuestionario de estudio logró detectar el discernimiento que tiene el profesorado sobre el uso del aprendizaje combinado. La Figura 2 se proyecta el mapa de calor que transmite el extracto de la Tabla 2, donde valores significativos se establecieron de color rojizo y lo contrario con valores claros.

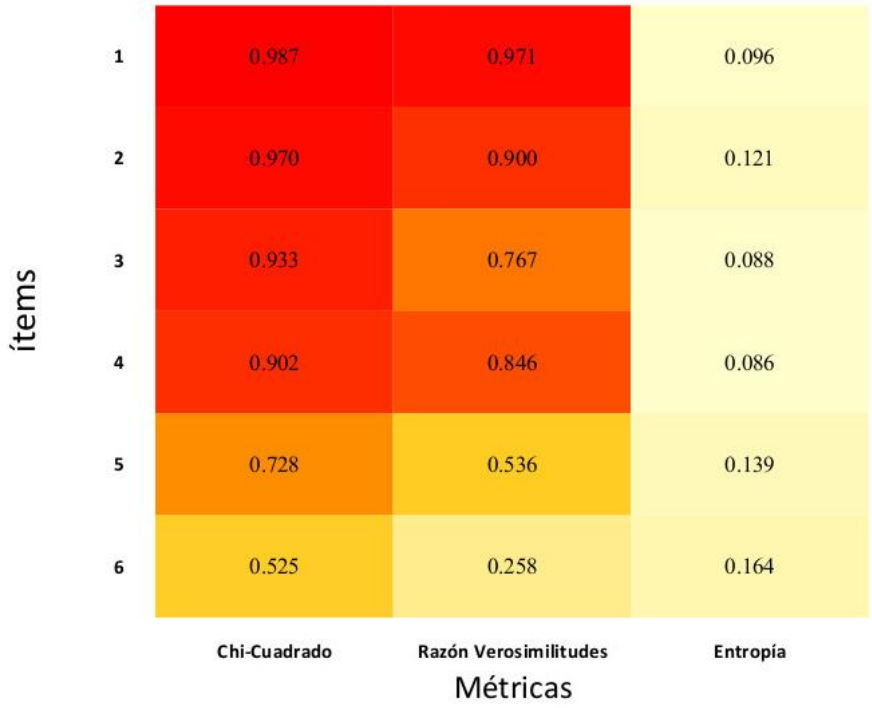


Figura 2. Se muestran por un lado el valor de las métricas (abscisas) y por otro, el contraste de las preguntas (ordenadas), los valores relevantes y altos en colores rojizos, mientras que, colores bajos (pálidos) fueron menos significativos. El ítem 5 y 6, tuvieron menos relevancia, es decir, existe factor de riesgo y negativo respectivamente. Se observa que la entropía es menor a 0,2, esto quiere decir que está dentro del umbral aceptable del estudio de los datos.

Fuente: Elaboración propia.

4 DISCUSIÓN

Examinados los resultados obtenidos, podemos indicar que, bajo la situación de pandemia fue crucial aprender de la experiencia de otros países al implementar estrategias y políticas de salud pública. También, resultó oportuno detectar problemas asociados con el aprendizaje combinado y la comunidad universitaria. Dadas las consideraciones que anteceden, el presente estudio destaca la importancia de tres dimensiones del cuestionario que identificaron factores positivos, de riesgo y negativos para el profesorado que acogió el aprendizaje combinado. Esta información fue esencial para que los gestores educativos esgriman sus criterios sobre la toma de decisiones, informadas y fundamentadas que beneficien al profesorado y la comunidad universitaria.

En este sentido Rappoport et al (2020), en su trabajo destacó que el apoyo recibido de forma inicial al profesorado fue crucial para realizar tareas educativas, sin embargo, planteó incorporar dichas tareas al quehacer cotidiano educativo. Así, el ítem 1, se observa que la inclinación del profesorado de educación superior sobre la capacitación y disposición para utilizar el aprendizaje combinado, a pesar de los desafíos que esto implica en términos de organización de actividades y tareas en el entorno virtual, es acogida de forma positiva. Esto deja entrever que la capacitación continua al profesorado es necesaria. Esta actitud sugiere que la nueva realidad ha puesto en una situación inesperada al profesorado, quienes deben enfrentar no solo su situación personal, familiar y profesional (en el caso de actividades esenciales), sino también la responsabilidad de garantizar la continuidad de las actividades de enseñanza para sus estudiantes. Este desafío generó una serie de situaciones complejas que deben abordarse a corto, mediano y largo plazo, con el fin de minimizar el impacto en el aprendizaje y la formación profesional del estudiantado universitarios (Sanz, 2020).

Otro panorama habitual en pandemia que toma desprevenida a las IES fue la planificación insuficiente para enfrentar la pandemia y la falta de formación y recursos adecuados tanto para

profesores como estudiantes, que generó una alta probabilidad de consecuencias negativas en la docencia telemática (Hu, Santuzzi, & Barber, 2019). Sin embargo, los resultados del ítem 2 sugiere que el profesorado está dispuesto en acoger el aprendizaje combinado aún cuando la fatiga de organizar actividades y tareas en el entorno virtual sean tediosas. Esta adaptación es parte del proceso de acoplamiento a nuevos procesos, lo cual puede generar estabilidad y eficiencia en el desempeño de las funciones (Guzmán, 2018). Por lo tanto, es positivo que los profesores estén dispuestos a recibir capacitaciones abiertamente y utilizar la metodología mixta (Gil-Galván, 2018).

El aprendizaje combinado presenta retos en el proceso de enseñanza, incluyendo el temor de los profesores a perder el control sobre los alumnos. Sin embargo, el ítem 2 del estudio, que evalúa la disposición del profesor a utilizar esta metodología indicó una actitud positiva hacia su uso, aunque deba o tenga que perder el dominio como profesor, por motivos de la virtualidad. Este resultado es consistente con los hallazgos de Oliva et al., (2020), donde manifiesta que la pérdida tácita de control del profesor sobre el alumno se asocia a un incremento de la carga lectiva y exceso de actividades o tareas fuera del horario habitual. Además, el paso de la presencialidad a la virtualidad se ha efectuado de modo precipitado y brusco, lo que ha generado problemas emocionales en el alumnado. A pesar de los esfuerzos de las IES por implementar medidas de vigilancia y control frente a conductas disruptivas y al fraude académico, el profesorado sigue percibiendo la pérdida del dominio con el alumnado como el peor elemento de la docencia virtual (Oliva, Ponce, Fernández, & Rivero, 2022).

El ítem 3, denotó la fatiga del profesorado para organizar las actividades y tareas del entorno virtual, en paralelo con las bases pedagógicas de ambas modalidades, se consideró positivo. Aunque, ésta situación puede desperdiciar el potencial pedagógico de los docentes, ya que la falta de conocimiento en el manejo de herramientas que faciliten la organización de tareas en el entorno virtual puede obstaculizar su capacidad para ofrecer una enseñanza de calidad. Por tanto, es importante que se brinden oportunidades de capacitación y apoyo técnico para que los docentes puedan adquirir las habilidades necesarias para optimizar el uso de herramientas tecnológicas y así evitar la fatiga y el estrés en la organización de las actividades en el entorno virtual. De esta manera, se puede asegurar que los docentes puedan desempeñarse con eficacia en la enseñanza mixta y aprovechar al máximo las posibilidades pedagógicas usando la tecnología (Melo-Hernandez, 2018).

No se puede desconocer el nivel de estrés y agotamiento a la que fueron sometidos tanto estudiantes como docentes para efectuar de manera exitosa la enseñanza a través del aprendizaje combinado durante el proceso de la virtualización de la educación (Zhang, 2020). En efecto, el ítem 4 revela la fatiga de organizar actividades y tareas del entorno virtual frente al suficiente conocimiento del profesorado en el manejo de dispositivos digitales (Wifi, móvil/celular, tabletas, etcétera.), se estimó este resultado como factor positivo. Un caso muy similar fue detectado por Sánchez et al. (2020) donde se enfatizó que el profesorado con destrezas digitales tuvo problemas en la aplicación de estas habilidades. Siendo una causa posible la falta de motivación que recibe el profesorado respecto a la organización de actividades curriculares en la asignatura, aun cuando su destreza tecnológica es altamente comprobable, no logra aprovecharla por la desmotivación trasladándose como una demanda pedagógica mayormente reflexiva, afectiva y emocional (Barros & Da Costa, 2020).

Varios fueron los factores positivos que se han encontrado en este estudio destacan la importancia de que el profesorado se capacite para integrar docentes capaces de promover el desarrollo de los estudiantes y educar para la adopción del aprendizaje combinado. Aunque no existe una solución universal que garantice la resolución de los problemas encontrados en el ítem 6, frente a estos destaca la capacidad de gestionarlos desde el escenario de la virtualidad (Olivia, Ponce , Hernández & Rivero, 2020). Por ello, es posible realizar una planificación cuidadosa para reducir al mínimo la consecuencia negativa y el factor de riesgo existente. La planificación debe tener en cuenta las posibles dificultades del alumnado y ofrecer pautas de actuación dentro de la estrategia institucional para afrontarlas de manera efectiva y mejorar la calidad de la enseñanza y el aprendizaje en el entorno mixto.

Tal como se ha visto, aunque la metodología mixta implica una mayor carga de trabajo y esfuerzo por parte del profesorado, se aprecia la disposición y capacidad de los docentes para adaptarse a esta nueva realidad. Esto es crucial para mantener la calidad de la educación superior en tiempos de crisis. Es importante destacar también, que los recursos tecnológicos (TIC) por sí solos no garantizó el éxito en la situación de pandemia. No obstante, son herramientas potentes que puede brindar oportunidades increíbles, pero la planificación, organización y flexibilidad son los elementos fundamentales que permitirán aprovechar los desarrollos tecnológicos y afrontar este y otros retos similares.

Cabe mencionar que el presente estudio se ha limitado a cuestionarios de Google Form que fue suministrado al profesorado. Por esta razón, se consideran posibles líneas de investigación a efectos de la adopción del aprendizaje combinado el uso de realidad virtual o realidad aumentada para corresponder a entornos prácticos de enseñanza en la educación superior, lo que permitiría un análisis y comparaciones con respecto a dicho tema entre profesores y estudiantes.

5 CONCLUSION

Las conclusiones de este estudio presentan los factores identificados que influyeron de forma positiva sobre la adopción del aprendizaje combinado/mixto, además de la gestión del profesorado sobre distintos escenarios presentados en la pandemia. En resumen, las capacidades y habilidades del profesorado se afianzaron a su destreza y pedagogía en la enseñanza, así como en sus competencias digitales. Los resultados sugieren que la decisión institucional de adoptar o rechazar el aprendizaje combinado está influida por el nivel del deseo institucional y el nivel de integración y capacitación en nuevas herramientas dirigidas al profesorado.

AGRADECIMIENTOS

Este trabajo ha sido apoyado por el proyecto FOCICYT-2023-IX-29 aprobado por el Honorable Consejo Universitario en comunicación UTEQ-SECGEN-2023-0222-M de la Universidad Técnica Estatal de Quevedo.

6 REFERENCIAS

Bartolic, S. K.; Boud, D., Agapito, J.; Verpoorten, D.; Williams, S.; Lutze-Mann, L.; Matzat, U.; Moreno M.; Polly P.; Tai J.; Marsh H.; Lin L.; Jamie-Lee B.; Habtu S.; Rodrigo M.; Roth M.; Heap T & Guppy, N. (2022). A multi-institutional assessment of changes in higher education teaching and learning in the face of COVID-19. *Educational Review*, *74*(3), 517-533.

Banerjee, G. (2011). Blended Environments: Learning Effectiveness and Student Satisfaction at a Small College in Transition. Online Learning, 15(1), 8–19. https://doi.org/10.24059/OLJ.V15I1.190

Barros, P., & Da Costa, J. (2021). Pedagogía en tiempos de pandemia: afectos y memorias de la enseñanza-aprendizaje. *593 Digital Publisher CEIT*, 6(2-1), 229-241. https://doi.org/10.33386/593dp.2021.2-1.505

Cabrera, L. (2020). Efectos del coronavirus en el sistema de enseñanza: aumenta la desigualdad de oportunidades educativas en España. *Revista de Sociología de la Educación-RASE*, *13*(2), 114-139.

Ferreira, P. D. C.; Barros, A.; Pereira, N.; Marques Pinto, A. & Veiga Simão, A. M. (2021). How Presenteeism Shaped Teacher Burnout in Cyberbullying Among Students During the COVID-19 Pandemic. *Frontiers in Psychology*, *12*, 745252.

Gil-Galván, R. (2018). El uso del aprendizaje basado en problemas en la enseñanza universitaria. Análisis de las competencias adquiridas y su impacto. *Revista mexicana de investigación educativa*, *23*(76), 73-93.

Guzmán, J. C. (2018). Las buenas prácticas de enseñanza de los profesores de educación superior. REICE. Revista iberoamericana sobre calidad, eficacia y cambio en educación, 16(2), 133-149.

Irugalbandara, A. I. (2021). The potential of Zoom technology for enabling creativity in the drama classroom through peer-assisted learning and group collaboration in pre service teacher education. *NJ*, *45*(2), 144-159.

Kor, P. P. K.; Leung, A. Y. M.; Parial, L. L.; Wong, E. M. L.; Dadaczynski, K.; Okan, O.; Adusei P.; Wang S.S.; Deng R.; Chi T.C. & Molassiotis, A. (2021). Are People With Chronic Diseases Satisfied With the Online Health Information Related to COVID-19 During the Pandemic?, *Journal of Nursing Scholarship*, *53*(1), 75-86.

Mendiola, M. S.; Hernández, A. M. D. P. M.; Torres, R.; Carrasco, M. D. A. S.; Romo, A.; Mario, A. & Cazales, V. (2020). Retos educativos durante la pandemia de COVID-19: una encuesta a profesores de la UNAM. *Revista digital universitaria*, *21*(3), 1-24. https://doi.org/10.22201/codeic.16076079e.2020.v21n3.a12

Miranda-Chavez, B.; Copaja-Corzo, C.; Rivarola-Hidalgo, M. & Taype-Rondan, Á. (2022). Fear of Death in Medical Students from a Peruvian University during the COVID-19 Pandemic. *Behavioral Sciences*, *12*(5), 142.

Mishra, P. & Koehler, M. J. (2008, March). Introducing technological pedagogical content knowledge. In *annual meeting of the American Educational Research Association* (Vol. 1, p. 16).

Nicola, M.; Alsafi, Z.; Sohrabi, C.; Kerwan, A.; Al-Jabir, A.; Iosifidis, C.; Agha M. & Agha, R. (2020). The socio-economic implications of the coronavirus pandemic (COVID-19): A review. *International journal of surgery*, *78*, 185-193.

Obeidat, A.; Obeidat, R. & Al-Shalabi, M. (2020). The effectiveness of adopting e-learning during COVID-19 at Hashemite University. *International Journal of Advanced Computer Science and Applications*, *11*(12).

Oliva, M. F. R.; Ponce, H. H.; Fernández, R. J. & Rivero, A. G. (2022). Retos en educación superior ante nuevos escenarios docentes durante la pandemia de la COVID-19. *Educação e Pesquisa*, *48*, 2022. https://doi.org/10.1590/S1678-4634202248258278ES

Ortega Ortigoza, D.; Rodríguez Rodríguez, J. & Mateos Inchaurrondo, A. (2021). Educación superior y la COVID-19: adaptación metodológica y evaluación online en dos universidades de Barcelona. *Revista Digital de Investigación en Docencia Universitaria*, *15*(1).

Rappoport, S., Rodríguez Tablado, M. S., & Bressanello, M. (2020). Enseñar en tiempos de COVID-19: una guía teórico-práctica para docentes. Unesco, 37. https://unesdoc.unesco.org/ark:/48223/pf0000373868

Rojas, L. V.; Huamán, C. J. V. & Salazar, F. M. (2020). Pandemia COVID-19: repercusiones en la educación universitaria. *Odontología sanmarquina*, *23*(2), 203-205.

Sá, M. J. & Serpa, S. (2020). The COVID-19 pandemic as an opportunity to foster the sustainable development of teaching in higher education. *Sustainability*, *12*(20), 8525.

Santana-Sardi, G. A.; Gutiérrez-Santana, J. A.; Zambrano-Palacios, V. C. & Castro-Coello, R. L. (2020). La Educación Superior ecuatoriana en tiempo de la pandemia del Covid-19. *Dominio de las Ciencias*, *6*(3), 757-775.

Sánchez-Pujalte, L. ., Gómez Yepes, T. ., Albalá Genol, M. ., & Etchezahar, E. . (2021). Percepción del profesorado y del alumnado universitario argentino sobre la adaptación a la educación virtual durante la pandemia por COVID-19. *Calidad De Vida Y Salud*, 14(2), 2-14. Recuperado a partir de http://revistacdvs.uflo.edu.ar/index.php/CdVUFLO/article/view/353

Tejedor, S.; Cervi, L.; Tusa, F. & Parola, A. (2020). Education in times of pandemic: Reflections of students and teachers on virtual university education in Spain, Italy and Ecuador. *Revista Latina de Comunicacion Social*, *2020*(78), 1–21. https://doi.org/10.4185/RLCS-2020-1466

Teichgräber, U.; Mensel, B.; Franiel, T.; Herzog, A.; Cho-Nöth, C. H.; Mentzel, H. J.; Ingwersen, M. & Aschenbach, R. (2021). Virtual inverted classroom to replace in-person radiology lectures at the time of the COVID-19 pandemic-a prospective evaluation and historic comparison. *BMC Medical Education*, *21*(1), 1-10.

UNESCO (2020). Education in a post-COVID world: Nine ideas for public action. International Commission on the Futures of Education. https://unesdoc.unesco.org/ark:/48223/pf0000373717/PDF/373717eng.pdf.multi

Yıldız, G.; Şahin, F.; Doğan, E. & Okur, M. R. (2022). Influential factors on e-learning adoption of university students with disability: Effects of type of disability. *British Journal of Educational Technology*.

Zhang, C. (2020). From Face-to-Face to Screen-to-Screen: CFL Teachers' Beliefs about Digital Teaching Competence during the Pandemic. *International Journal of Chinese Language Teaching*, *1*(1) 35-52. https://doi.org/10.46451/ijclt.2020.06.03

9. Predicting academic success of college students using machine learning techniques

- Jorge Guanin-Fajardo, Javier Guaña-Moya, Jorge Casillas
  - Status: Publicado (Data)
  - WoS “Computer Science, Information Systems”, JIF 2023: 2.2, 138/249 (Q3).
  - WoS “Multidisciplinary Sciences”, JIF 2023: 2.2, 49/134 (Q2).
  - SJR 2023: h-index 38, “Information Systems” (Q2).

PREDICTING ACADEMIC SUCCESS OF COLLEGE STUDENTS USING MACHINE LEARNING TECHNIQUES

Jorge Humberto Guanin-Fajardo
Facultad de Ciencias de la Ingeniería , Universidad Técnica Estatal de Quevedo, Quevedo, Ecuador;
jorgeguanin@uteq.edu.ec

Javier Guaña-Moya
Facultad de Ingeniería; Pontificia Universidad Católica del Ecuador, Quito, Ecuador;
eguana953@puce.edu.ec

Jorge Casillas
Department of Computer Science and Artificial Intelligence; University of Granada, Granada, Spain;
casillas@decsai.ugr.es

DOI: https://doi.org/10.3390/data9040060

Abstract: College context and academic performance are important determinants of academic success; using students' prior experience with machine learning techniques to predict academic success before the end of the first year reinforces college self-efficacy. Dropout prediction is related to student retention and has been studied extensively in recent work; however, there is little literature on predicting academic success using educational machine learning. For this reason, CRISP-DM methodology was applied to extract relevant knowledge and features from the data. The dataset examined consists of 6690 records and 21 variables with academic and socioeconomic information. Preprocessing techniques and classification algorithms were analyzed. The area under the curve was used to measure the effectiveness of the algorithm; XGBoost had an AUC = 87.75% and correctly classified eight out of ten cases, while the decision tree improved interpretation with ten rules in seven out of ten cases. Recognizing the gaps in the study and that on-time completion of college consolidates college self-efficacy, creating intervention and support strategies to retain students is a priority for decision makers. Assessing the fairness and discrimination of the algorithms was the main limitation of this work. In the future, we intend to apply the extracted knowledge and learn about its influence of on university management.

Keywords: educational data mining; machine learning; educational analysis; higher education; academic success

1. Introduction

Higher education has developed a fundamental role due to the versatility and complexity of today's world, which has led to the rapid growth of scientific literature dedicated to predicting academic success or the risk of student dropout [1–7]. Higher education institutions and their traditional role of knowledge dissemination have changed; innovation in new knowledge especially with the irruption of artificial intelligence [8] and the training of qualified professionals make many of them interact in different areas of society. In fact, their missions through teaching, research, and the ability to share and transfer this knowledge constitute central functions of their academic and cultural activity, with the aim of improving the level of knowledge in society. They have the important role of transmitting knowledge, skills, and values to students to create competitive professionals in society. Therefore, channeling students towards academic success is transcendental, as HEIs must continue the work undertaken and further deepen their involvement, significance, and service capacity in relation to the social, cultural, and economic framework [9]. Thus, the prediction of academic success with past information of students who have successfully completed their university studies has become a tool of interest for educational managers since it allows them to strengthen decisions and build improvement alternatives or educational policies. ICT is one of the most widely used alternatives today, especially machine learning.

Hence, advances in machine learning techniques, along with other areas of study, are precursors to educational data mining. In higher education, the academic success of students is statistically measured by the graduation rate, which is defined as the total number of students graduating among the total number of entering students. In fact, ref. [10] states that it is possible to think about student success more broadly by studying endogenous and exogenous factors in the student environment. Thus, the constant need to be effective in the academic success of students has led to the customization of machine learning, this to achieve specific predictive models that provide useful information.

In the last decade, many studies have focused on investigative works that address the problems of performance, dropout, and academic success in university students. As detailed in [11–14], the authors emphasize that university dropout or failure converges with students from disadvantaged social strata who project university dropout behavior. To sustain university permanence among their findings, the authors are inclined to consider that extra-university activities that guarantee retention should be strengthened. Therefore, early detection has become a tool for solving these problems. Academic history, university context (tangible and intangible resources), and other data were used as the input elements to predict the results [4]. For this purpose, qualitative and quantitative research methods have been used to solve these problems. More recently, multiple studies have been derived that employed data mining or machine learning techniques that, among other things, use algorithms and two well-known techniques to extract useful knowledge from data. The first technique, supervised classification, evaluates the data and predicts the target variable (class). The work of [6,15–17] has shown results related to supervised classification.

Similarly, in [18,19], using another approach based on supervised classification, they used a set of pre-selected algorithms that classify the data by applying the voting technique. Both approaches attempt to predict students' academic success or performance effectively. The second technique, unsupervised classification, is one in which the target variable is unknown and that focuses on finding hidden patterns among the data. In general, association rules are used to discover facts occurring within the data and are composed of two parts: antecedent and consequent; for example, the rule $\{A, B\} \Rightarrow \{C\}$ means that, when A and B occur, then C occurs. In [20–22], they look for the occurrence of data by focusing on the association rules and evaluating the rules with metrics such as support, confidence, and lift, among others.

In the studies of [23–25], related to machine learning, the convergence of objectives and techniques applied for the data preprocessing stage was observed, both in feature reduction, data transformation, normalization, and instance selection, among others. At the same time, data balancing techniques and "black box" classification algorithms were analyzed. The synergy of the studies lies in the simplification of the predictive models obtained given the high degree of complexity of the extracted knowledge, for which they used decision trees, since this technique simplifies the knowledge by

means of the representation of rules of type (X⇒Y). To some extent, the methods applied are part of the KDD process proposed in [26]. However, data asymmetry is a typical problem in any area of study. Duplicity, ambiguity, and missing and overlapping data are frequent, especially in authentic problems. Indeed, in data mining classification techniques, problems are presented as an unequal distribution of examples among classes (target variable), where one or more classes (minority class) are underrepresented compared to the others (majority class) [27]. Commonly, the data balancing method defined by Chawla [28] is used in this type of problem. However, it is intended to fill the existing gap of data balancing with educational data by using different balancing methods for multiclass problems.

The approach of this study is like previous work described in [6,29–31], where similar tasks were performed with predictions in binary and multiclass classes. However, the main difference with our approach focuses on the in-depth analysis of data balancing and feature selection techniques to avoid biases in predictions. Using 53% fewer variables and improving its accuracy by 10% over the preliminary results with the raw data, we not only built classification models to identify the relevant factors of college students' academic success, but also obtained a general model from the decision tree to obtain a higher readability of the predictive model. In this way, it is intended to provide additional guidance to academic decision makers in decision making. The open license software used for this work was R [32] through a customized library to visualize, preprocess and classify the data. The Python library scikit-learn [33] was used for data balancing.

The core of the work focuses on the study of machine learning techniques that predict academic success. This has allowed us to establish the objective of the work, which is to know in advance the factors that explain the academic success of students at the end of their first year of university. To do this, it has been necessary to pose the research questions since we intend to identify the factors that contribute to the academic success of students during their first year of college. This will allow us to examine the preprocessing techniques, the predictive model, the determinants of academic success and, of course, the visualization techniques to improve its interpretation before and after obtaining the predictive model. In this sense, the following research questions were posed:

RQ1: Which balancing and feature selection technique is relevant for supervised classification algorithms?

RQ2: Which predictive model best discriminates students' academic success?

RQ3: Which factors are determinants of students' academic success?

Most studies on predicting academic success by machine learning have focused solely on finding a predictive model, which is, to some extent, highly effective. In contrast, the work presented, in line with RQ1, seeks the group of features that are most significant for the model and, on the other hand, also seeks a balanced training dataset, using different data balancing techniques and avoiding biases in the prediction. RQ2, on the other hand, aims to find the effective predictive model using different supervised learning algorithms. Finally, RQ3 examines which variables were relevant in the predictive model achieved by the machine learning algorithms to then obtain another model with a better interpretation for the decision maker.

The presented work differs, among other things, by the following contributions: (i) we unveil the effectiveness of educational data mining techniques, to identify academically successful students early enough to act and reduce the failure rate; (ii) the impact of data preprocessing is analyzed; (iii) the important variables underlying the predictive model of better performance are unveiled. Thus, an approach to the presented work is associated with the works of [23,29,34], where the authors have examined the characteristics and impact of the best-performing algorithm. The rest of the paper is organized as follows: in Section 2, a literature review is carried out; in Section 3, the methodology used in this work is explained; in Section 4, the main results obtained by applying machine learning are presented; in Section 5, the discussion is presented; in Section 6, the relevant conclusions, in Section 7, limitations; and finally, in Section 8 future work are described.

2. Literature Review

In the cited literature, there are works related to the study of machine learning in higher education and its impact on the prediction of academic performance or success. In prediction, the purpose is to predict the target variable (class) of a dataset. The works cited in Table 1 employ supervised classification algorithms that focus on obtaining the predictive model.

Table 1. Summary of papers related to the prediction of academic performance or success of university students.

Objective	Inst. [1]	Feat. [2]	Class	DPM [3]	Accuracy	Citation	Scope
Performance	6948	55	2	Data preprocessing methods	82%	[35]	Higher Education
Performance	3830	27	2	Data transformation, Discretization	83%	[36]	Higher Education
Prediction	1854	4	2		75%	[37]	
Academic Success Assessment	731	12	2	Extraction Feature, Imbalanced Dataset	78%	[6]	Higher Education
Achievement	339	15	3	Extraction Feature	69.3%	[23]	Higher Education
Performance	32,593	31	4	Extraction Feature, Imbalanced Dataset	72.73%	[38]	Higher Education
Prediction	9652	68	2	Extraction Feature, Imbalanced Dataset	75.43%	[24]	Higher Education
Prediction	3225	57	2	Extraction Feature, Imbalanced Dataset	79.5%	[28]	Higher Education
Prediction	300	18	2	Extraction Feature	63.33%	[34]	Higher Education
Prediction	1491	13	2	Extraction Feature, Imbalanced Dataset	75.78%	[5]	Higher Education
Prediction	7936	29	2	Extraction Feature	69.3%	[30]	Higher Education
Prediction	4413		2			[18]	Higher Education
Prediction	6690	21	3	Selection Feature, Selection Instance, Data imbalanced	81%	Our proposal	Higher Education

[1] Number of instances. [2] Number of features. [3] Data preprocessing methods.

Among other works, the use of machine learning techniques to predict the success or failure of university courses or degrees stands out. The use of the recommender system proposed by [35] suggests to computer science students the subjects they can take, in addition to the prediction of success or failure based on the previous experience of other university students. In the work, data preprocessing and example balancing techniques were applied. Then, the preprocessed data were used as input for the classification algorithms to learn and obtain the prediction model from the test data. The results achieved provide guidelines for university administrators to enhance educational quality. In this sense, the early provision of useful information to predict a given event in the student body is valuable. Hence, the study of academic performance is a relevant contribution in higher education. Helal [36], in his work, predicted the academic performance of the student body; the data used in his work were divided into groups, and each subgroup of data was evaluated with different classification algorithms to predict academic performance. Their results suggest that external students and female students performed well in the prediction.

The work of Bertolini [29] set out to examine different classification algorithms to predict final exam grades with reasonable accuracy, considering midterm grades. Similarly, Alyahyan [23] proposed the use of decision trees to predict students' academic performance and generate an early warning when low performance is detected. Different decision tree approaches as well as relevant feature extraction were employed to obtain a simpler model for decision making by academic experts. In line with this, refs. [29,34] also examined high-impact features in the data to fit representative

variables with respect to college retention and dropout, to develop interventions to help improve student academic success.

Similarly, in Beaulac [39], the prediction of the academic success of university students has been studied by applying the random forest and decision tree algorithms, the latter being very intuitive for decision making; the authors propose the use of these techniques to know if at the end of the first two semesters the student would achieve the university degree. Their results have indicated that there is a strong relationship between underperforming grades and the likelihood of succeeding in a degree program, although this did not necessarily indicate a causal connection.

Several of the related articles reveal the variety of work linked to improving the educational system. The approach of Guerrero-Higueras [7], which proposes the use of the GIT version control system as an evaluation methodology to observe the frequency and use of the tool to help predict the student's academic success, stands out. The variables studied describe the student's ability with tasks related to the development of the computer science subject. This methodology as introduced differs from the rest given the adaptation of the GIT version control platform and the issues specific to the computer science area.

The literature cited above emphasizes gradualism to achieve features that achieve high accuracy in the algorithms and obtain a simple and readable model. The lack of salient features prevents obtaining an effective prediction model. This is because of the ambiguity or irrelevance of the variables [40]. On the other hand, of significant importance is the reduction of outliers in the data due to duplicate observations or overlapping data [41–43]. It is understood, of course, that all of this leads to the application of each stage suggested in the CRISP-DM [26], methodology that allows obtaining a reliable model at the end. The validity of the model obtained is checked by the performance metrics of the classification algorithms. Based on what has been presented in this section, it was observed in the literature that the work focuses mainly on two fronts: identifying significant attributes to predict student performance, success, or failure in higher education, and finding the best prediction method to improve the accuracy of the predictive model achieved.

3. Materials and Methods

3.1. Context

The Institution of Higher Education (IES) is geographically located in the Municipality of Quevedo, Province of Los Ríos, Ecuador. Its coordinates are set at: 1°00′46″ S 79°28′09″ W/−1.012778, −79.469167. According to the policies of the IES and its minimum requirements, each university course is taught in face-to-face mode, and in addition, each academic year of the university course must be passed. In this case, each academic year consists of two academic cycles (semesters). Students must enroll in the university degree program and obtain grades in each subject, with a minimum grade of seven on a scale of zero to ten. As a result of the academic activities performed and their permanence in the university degree, the academic status of the student body is determined (dependent variable/class). Academic statuses are established in three categories. The first is "Passed", when the student has completed and passed all academic courses. The second is "Change", when the student passes courses other than the initial degree. And finally, third is "Dropout", when the student leaves the university completely.

3.2. Data Collection

Data collection was performed using SQL server scripts. The data were extracted from the university's information system database server. The dataset used in this work consisted of two parts: student body and faculty, which were subsequently merged. It should be noted that the criterion for the merger was the classes taught in the first year by the faculty in the teaching process for the university degree. Thus, the first part of the information referring to the students dealt with academic and socioeconomic data, while that relating to the teaching staff referred to degrees obtained, age, and academic experience, among others. Among the diversity of professors in charge of university teaching of first-year students, there were full, associate, and occasional professors, totaling 286 professors selected for this study.

On the other hand, the number of regular students was 6690. Although the number of professors and students does not coincide, it is necessary to clarify that a professor can teach different subjects. The students selected were those who were enrolled and had completed the first year of all university courses. In short, all of the above was framed within a retrospective of six complete academic years of each university degree, that is, ten calendar years. It should also be noted that any identifying reference to both faculty and students was eliminated to obtain an anonymous dataset. Among other things, the information extracted for this work had the endorsement and permission of the competent authority of the higher education institution detailed in the Institutional Review Board Statement section[M1][JG2]. The database with the raw data had 21 variables and 6690 records (see Appendix A, Table A1 for a description of the variables used).

So far, one of the main differences in algorithms between machine learning (ML) and traditional statistical methods lies in their purpose, as the former is still focused on the ability to capture complex relationships between features and make predictions as accurate as possible, while the latter, especially linear regression (LR), logistic regression (LOR), generalized mixed models and relevance-based prediction and others, aim at inferring relationships between variables. However, the key difference between traditional statistical approaches and ML is that, in ML, a model learns from examples rather than being programmed with rules. For a given task, examples are provided in the form of inputs (called features or attributes) and outputs (called labels or classes) [44,45].

In this work, we used the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology proposed by [26], which comprises seven phases: understanding the problem, understanding the data, data preparation, modeling, evaluation and implementation; the data preparation or data preprocessing is a stage that gained importance and became a key stage, since its function is related to data preparation. In other words, the objective is to reduce the complexity of the original dataset to obtain a readable predictive model with useful variables. Therefore, the work is based on the best practice for data preprocessing suggested in [46–48]. For this reason, Appendixes B and C detail the results of the various methods used for data preprocessing using feature filtering, instance selection, and class balancing. The main advantage of efficient data preprocessing was the transfer of suitable data to classification algorithms for simple and accurate learning. First, the compacted data were cleaned and transformed and then analyzed with visualization techniques that allowed, among other things, the location of trajectories, overlaps and data behavior. Second, the data were stratified into two subsets of data: training and test. Then, the training set was filtered for relevant instances and features to balance the data using different methods. The already balanced dataset was used as input data for the classification algorithms, together with the test data that were used to obtain the predictive model. Finally, this model was evaluated with the metrics proposed in this work. Figure 1 shows the activities that were performed.

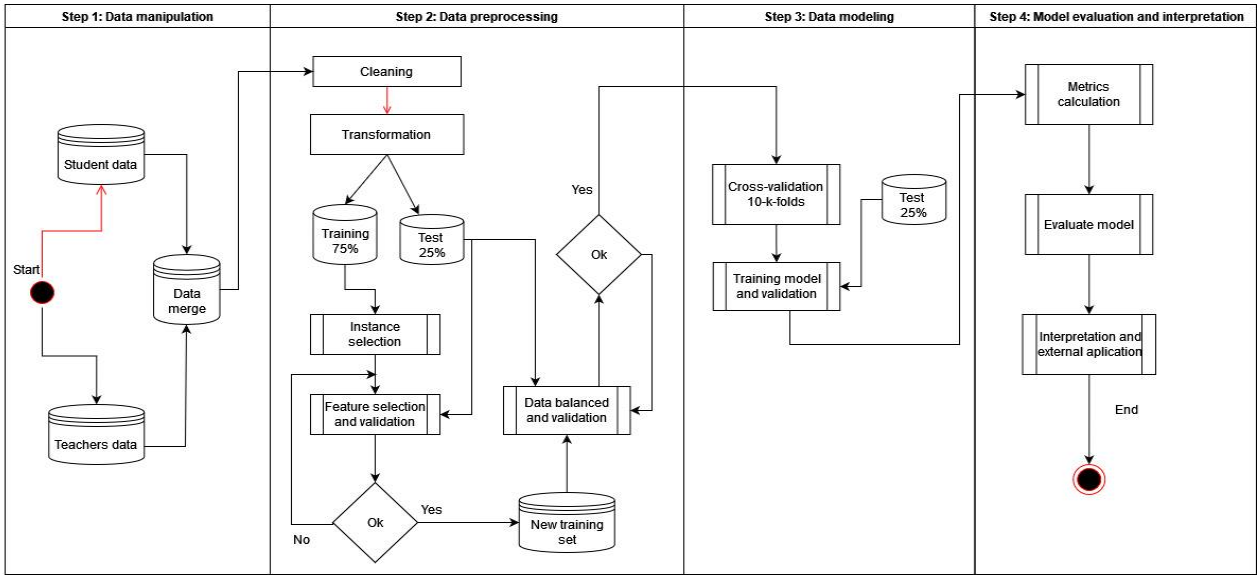


Figure 1. Diagram of activities performed. The processes conducted are described in four stages.

3.3. Metric Assessment

The metrics referred to in this section are used to evaluate the performance of the set of algorithms used to obtain predictive models. In Equation (4), the term α represents P(Tp) = Sensitivity, and (1 − β) represents P(Tn) = Specificity [49].

$$\text{Accuracy}=\frac{TP+TN}{TP+FN+FP+TN} \tag{1}$$

$$\text{Sensitivity}=\text{Recall}=\frac{TP}{TP+FN} \tag{2}$$

$$\text{Specificity}=\frac{TN}{TN+FP} \tag{3}$$

$$\text{AUC}=\sum\left\{\left(1-\beta_i.\Delta\alpha\right)+\frac{1}{2}\left[\left(1-\beta\right).\Delta\alpha\right]\right\} \tag{4}$$

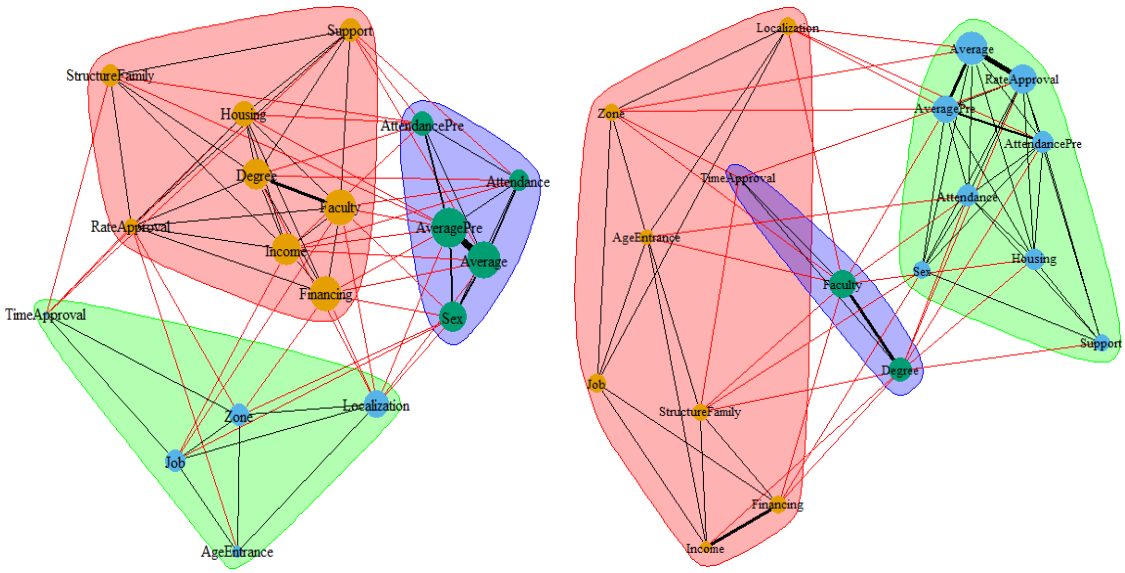
$$\text{Precision}=\frac{TP}{TP+FP} \tag{5}$$

$$\text{Cohen's Kappa}=\frac{\Pr(a)-\Pr(e)}{1-\Pr(e)} \tag{6}$$

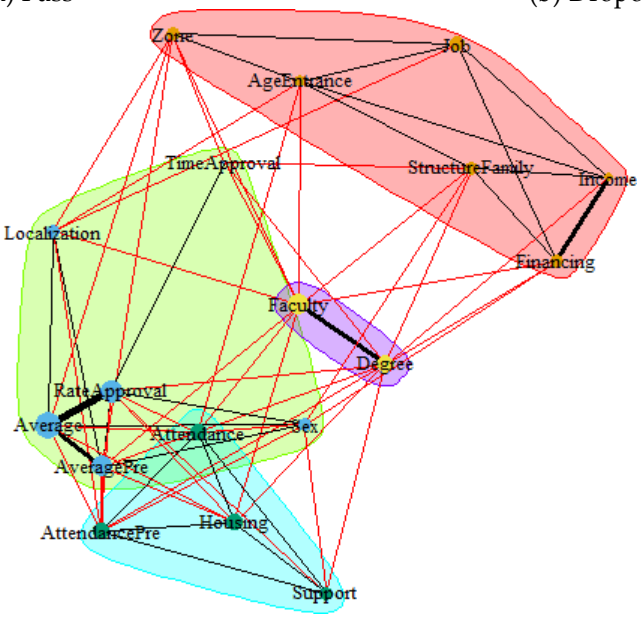
$$LogLoss=-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}y_{i,j}\log\left(p_{i,j}\right) \tag{7}$$

3.4. Data Exploratory

The importance of data exploration is that it serves to understand the activity and behavior of the data. Visualization techniques have been used that detected significant information in the data; specifically, variables were examined according to each category of the class using graphs (Figure 2).



(a) Pass (b) Dropout



(c) Change

Figure 2. Undirected graph calculated from the correlation matrix (Pearson's method). Both the arcs and the adjacency matrix were filtered with cut-off points obtained from the weighted mean of the nodes (Pass = 0.0007804694, Dropout = 0.0061971, Change = 0.01684287). The graphs had weights associated with each of the arcs, and this weight fixed their density. Three groups of subfigures were separated according to the target variable (pass, dropout, change). Subfigure (a) showed three subgroups of variables (8, 5, 5) where a common variable overlaps. Cluster (b) showed three subgroups of variables (8, 3, 8); this subfigure lacks overlap. Group (c) showed four subgroups of variables (6, 7, 4, 2) overlapped by three common variables. On the other hand, red lines indicate a lower degree of association, while black lines and thickness indicate their strength of association.

3.5. Data Preprocessing

The importance of data preprocessing is to synthesize and achieve expeditious data. This fact has an important consequence for classification algorithms since the integrity of the data is gradually assessed by the hit rate, i.e., the number of true positives that the prediction algorithm can detect. Within this context, the aim is to obtain the set of features and instances that are close to a reasonable hit rate. The problem around which the data preprocessing revolves is the

different search strategies such as sequential, random, and complete that are proposed for this task. The evaluation criterion is set with filtering (distance, information, dependency, and consistency), hybrid and wrapper methods [50–54].

The data preprocessing was divided into four phases. First, missing values in the data were replaced using the k-nearest neighbor's algorithm KNN_MV [55]. Second, unrepresentative instances were excluded using the "NoiseFiltersR" algorithm. Third, feature selection was studied with different algorithms and functions that have evaluated feature quality. Finally, data balancing was applied to avoid bias in the prediction model due to the small amount of minority class data.

3.6. Missing Values

Data in their original form contain inconsistent data and often have missing values. That is, when the value of a variable is not stored, it is considered missing data. Multiple techniques have been developed to replace missing values. In general, statistical techniques of central tendency are usually used; for numerical values, the mean or median is used, while for nominal values, the "mode" is usually used. Another common technique is to remove the entire record from the dataset. Deletion can cause significant loss of information. Frequent techniques are easy to use and solve the problem of missing values, although, in data mining practice, there is a tendency to implement algorithms that solve this problem by examining the entire dataset. Specifically, in this work, we have used the "rfImpute" function, which replaced missing values by the nearest neighbor technique that takes the class (target variable) as reference.

3.7. Instance Selection

Instance selection was also key in the data preprocessing, since poor-quality examples were eliminated by using the NoiseFiltersR algorithm [41], which filtered out the 5% of examples that were not within the data standard. In other words, when a value is at an unusual distance from the rest of the values in the dataset, it is considered an outlier or noise.

3.8. Feature Selection

There is an important distinction to be made in this section since the generality and accuracy of the predictive model will depend on the quality of the variables. Therefore, it is crucial to decide which variables are relevant to include in the study. For this, we used nine feature selection algorithms among them: "LasVegas-LVF", "Relief" [56], "selectKBest", "hillClimbing", "sequentialBackward", "sequentialFloatingForward", "deepFirst", "geneticAlgorithm", and "antColony". On the other hand, the algorithms used distinct functions to value the attributes. Among the functions, we had "mutualInformation" [57], "MDLC" [58], "determinationCoefficient" [59], "GainRatio" [60], "Gini Index" [61], and "roughsetConsistency" [62,63]. The group of algorithms used for the study of significant characteristics obtained subgroups of variables that have been evaluated and are shown in Table 2 and Appendix C Table A3.

Table 2. Feature filtering by the "Relief" algorithm using different k and bestk filters. The lowest feature selection and the highest accuracy achieved by the C4.5 classification algorithm were established with the "bestk" filtering (10 variables).

Filter	Variable	Value	Accuracy	Kappa	Sensitivity	Specificity	Precision	Recall	F1
k = 9	11	−0.002	0.75	0.56	0.83	0.85	0.79	0.83	0.80
k = 7	11	−0.001	0.76	0.5	0.82	0.85	0.78	0.82	0.80
k = 5	11	−0.003	0.74	0.52	0.80	0.83	0.77	0.80	0.78
k = 3	14	−0.001	0.76	0.56	0.82	0.85	0.79	0.82	0.80
bestk	10	0.062	0.79	0.62	0.85	0.87	0.81	0.85	0.83

3.9. Data Balancing

Sample balancing is another important step in data preprocessing. Currently, there are several techniques for data balancing or resampling using Python software 3.9 and its scikit-learn library [33]. In this work, the following techniques have been studied: oversampling, combined, undersampling and ensemble. The first used the methods "Smote" [28]

and "KMeansSMOTE" (oversampling with SMOTE, followed by undersampling with edited nearest neighbors) [64]. The second used both "Smote-ENN" and "Smote-Tomek" (oversampling with SMOTE) [65]. The third technique used was subsampling with the "RUS" method [66]. Finally, the ensemble technique used "EasyEnsemble" [67] and "Bagging". Specifically, new balanced training datasets were generated. All of this was from the initial training set, in which the different techniques and methods were used to balance the data (See Table 3).

Table 3. The table displays the distribution of data per class using different data balancing techniques, along with the corresponding imbalance ratio (IR) between the majority and minority classes. A higher IR indicates a more severe class imbalance problem.

	Classes				
Algorithms Used	Dropout	Change	Pass	Overall	IR
Origin data (not use algorithm)	3.346	466	2.080	5.892	7.180
Over (SMOTE)	2.826	5.652	8.478	16.956	3
Over (KMeansSMOTE	5.655	8.481	2.829	16.965	2.997
Combined (SMOTE-ENN)	5.365	2.822	4.164	12.351	1.901
Combined (SMOTE-Tomek)	5.360	2.826	7.894	16.080	1.472
Under (RUS)	355	1.065	710	2.130	3
Under (Tomelinks)	2.439	4.229	3.874	10.542	1.733
Ensembles (EasyEnsemble)	2.826	5.017	4.662	12.505	1.775
Ensembles (Bagging)	2.826	5.017	4.662	12.505	1.775

3.10. Classification Algorithms

The use of supervised classification techniques aims to achieve a prediction model that is highly accurate. Hence, several algorithms have been created that use different mathematical models to achieve the model. In this section, we detail the types of algorithms and a provide a brief description of how each works.

Decision Trees: Consists of building a tree structure in which each branch represents a question about an attribute. New branches are created according to the answers to the question until reaching the leaves of the tree (where the structure ends). The leaf nodes indicate the predicted class; see [35].

Support Vector Machine (SVM): A relatively simple supervised machine learning algorithm used in regression or classification related problems. In many cases, it is used for classification, although it is preferably useful for regression. Basically, SVM creates a hyperplane with boundaries between data types in a two-dimensional space; this hyperplane is nothing more than a line. In SVM, each datum in the dataset is plotted in an N-dimensional space, where N is the number of features/attributes of the data; see [68].

Neural Network: Multilayer perceptrons (MLP) are the best known and most widely used type of neural network. They consist of neuron-like units, multiple inputs, and an output. Each of these units forms a weighted sum of its inputs, to which a constant term is added. This sum is then passed through a nonlinearity, usually called an activation function. Most of the time, the units are interconnected in such a way that they form no loop; see [69].

Random Forest: A combination of tree predictors, where each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The use of random feature selection to split each node produces error rates that compare favorably with "Adaboost" but are more robust with respect to noise. The internal estimates control for error, strength, and correlation, and are used to show the response to increasing the number of features used in the split. Internal estimates are also used to measure the importance of variables; see [70].

Gradient Boosting Machine: Gradient boosting is a machine learning technique used to solve regression or classification problems, which builds a predictive model in the form of decision trees. It develops a general gradient descent "boosting" paradigm for additive expansions based on any fitting criteria. Gradient boosting of regression trees produces competitive, very robust, and interpretable regression and classification procedures, especially suitable for the extraction of not-so-clean data; see [71].

XGBoost: XGBoost is a distributed and optimized gradient boosting library designed to be highly efficient, flexible, and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides parallel tree boosting (also known as GBDT, GBM) that solves many data science problems in a fast and accurate manner; see [72].

Bagging: Predictor bagging is a method of generating multiple versions of a predictor and using them to obtain an aggregate predictor. Bagging averages the versions when predicting a numerical outcome and performs plural voting when predicting a class. Multiple versions are formed by making bootstrap replicas of the learning set and using them as new learning sets. Tests on real and simulated datasets show that bagging can provide a substantial increase in accuracy; see [73].

Naïve Bayes: A probabilistic machine learning model used for classification tasks. The core of the classifier is based on Bayes' theorem: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$, which is the probability of A occurring, given that B has occurred. Here, B is the evidence, and A is the hypothesis. The assumption made here is that the predictors/features are independent. That is, the presence of a particular feature does not affect the other; see [74].

4. Results

In response to the research questions posed, different data preprocessing algorithms have been employed to reduce the dimensionality of the dataset, so that the classification algorithms obtain a simple and accurate predictive model. In the following sections, we study data preprocessing for feature selection first. Second, we study data balancing using different data balancing algorithms and, finally, the results using the metrics calculated from the confusion matrix where the performance of the algorithms was evaluated.

4.1. Data Preprocessing

4.1.1. Feature Selection

Prior to preprocessing, the dataset was separated into two parts: 75% of the total was selected for training data, and the other 25% for testing. The latter were used to evaluate the predictive model achieved by the classification algorithms, while the training set was subjected to preprocessing techniques to reduce dimensionality and obtain adequate data. In this sense, the work has focused on achieving simplicity and improving the accuracy of the predictive model, for which different feature and filter selection methods have been configured. Table 2 shows the results of the algorithm that obtained the lowest features; the rest of the runs of other algorithms and their results can be found in Appendix C.

In view of the cited works, in the studies of [15,28], relevant features in the data were examined to improve the predictive model, in line with these. Table 2 presents the results for the pre-selected feature set, where each evaluative filter and method rated the variables according to the performance metric. Specifically, the Relief method together with the "bestk" evaluative filter achieved better efficiency, i.e., higher accuracy with fewer variables. Based on these results, a new dataset with the new characteristics was established and used as input data for the data balancing phase described in the next section.

4.1.2. Data Balancing

The importance of data balancing is fundamental to classification algorithms since the disparity of examples between one class and another can lead to bias in the prediction model. There are two common techniques for data balancing. The first is the oversampling of examples technique, in which the data are balanced to the same number of examples in the majority class. The second is to reduce the other classes to the same number of examples in the minority class. Both techniques, although not very efficient, are useful for obtaining primary results since the redistribution of the data is achieved with the judgment and experience of the data analyst. To some extent, this personalized judgment is avoided by the intervention of algorithms that perform data balancing. The algorithms augment, reduce or equalize the examples depending on the technique applied. From the above, Table 3 shows the data imbalance index according to the

algorithms used. Thus, each algorithm generated a new balanced dataset that was used to train the classification algorithms.

4.2. Classification Algorithms

In this section, we examine the effectiveness of the set of classification algorithms proposed for this work, which is related as a multiclass problem, that is, a dependent variable (class) with three types of outputs: Dropout, Change and Passed. For this reason, and as is common in supervised classification problems, two datasets have been used: the first, for the algorithms to learn and obtain a prediction model; and the second, to evaluate the effectiveness of the model obtained. Hence, we worked with two types of analysis: the first with the original data (without data preprocessing) and the second with the different datasets generated from the preprocessing techniques used.

It is difficult not to appreciate the importance of data preprocessing, as it provides classification algorithms with balanced and clean datasets. Obtaining the predictive model requires the algorithm to learn from the provided data (training set), as the effectiveness of the model will depend on it. Therefore, for the algorithm to achieve adequate learning, the cross-validation technique k-fold cross-validation (CV) was applied; this approach randomly subdivided the training set into 10 folds with approximately equal size, and each fold, in turn, was fragmented into two sections: training and test. This was done so that at the end of training, the mean prediction was obtained from among the folds. On the other hand, to check what was learned by the algorithms, the metrics proposed in the section of methodology were used, which helped to discriminate the most effective predictive models. While it is true that effectiveness is fundamental to evaluate the predictive model, the comprehensibility of the model obtained is also important, since the experts evaluate the simplicity of the model.

Here, we present the best result of the classification algorithms that were achieved using the dataset balanced by the "EasyEnsemble" algorithm and the performance assessment of the classifiers using the ROC curve presented in Figure 3. [M3][JG4]The rest of the results with different datasets derived from the application of the data balancing algorithms are presented in Appendix B, Table A2.

In view of the results, Table 4 (raw data) and Table 5 (preprocessed data) show differences in the performance of the algorithms. Negative values −0.0214 and −0.0222 for precision and AUC, respectively, are evident. This negative effect between raw data and preprocessed data is a consequence of preprocessing, so data preprocessing should be interpreted not as a contradictory process but as an improvement of the predictive model by using fewer variables from the original set. Therefore, the advantage of applying data preprocessing has been observed.

Table 4. Preliminary results for the original dataset, omitting data preprocessing.

Algorithms	Accuracy	Kappa	Sensitivity	Specificity	Precision	Recall	F1	AUC	LogLoss
XGBoost	0.8133	0.6617	0.8492	0.8861	0.8456	0.8492	0.8462	0.8997	0.3736
RandomForest	0.8163	0.6664	0.8523	0.8873	0.8428	0.8523	0.8468	0.8978	NA
Gbm	0.8062	0.6473	0.8460	0.8800	0.8352	0.8460	0.8401	0.8930	0.3925
Bagging	0.8008	0.6379	0.8423	0.8769	0.8291	0.8423	0.8351	0.8781	NA
C4.5	0.7822	0.6039	0.8378	0.8642	0.8033	0.8378	0.8193	0.8308	NA
NaiveBayes	0.6549	0.3847	0.5215	0.8025	0.7622	0.5215	0.5059	0.8168	NA
SvmRadial	0.7284	0.4934	0.7781	0.8218	0.7673	0.7781	0.7709	0.7973	NA
SvmPoly	0.7165	0.4687	0.7571	0.8132	0.7685	0.7571	0.7616	0.7754	0.5484
MLP	0.6895	0.4501	0.7673	0.8143	0.7471	0.7673	0.7511	0.7621	0.5378

Table 5. Evaluation results of the predictive models obtained by the classification algorithms. The training set was balanced with the "EasyEnsemble" technique. Model validation was performed on the test dataset. The data were sorted according to the AUC column.

Algorithms	Accuracy	Kappa	Sensitivity	Specificity	Precision	Recall	F1	AUC	LogLoss
XGBoost	0.7949	0.6299	0.8425	0.8753	0.8214	0.8425	0.8306	0.8775	6.3430
RandomForest	0.7925	0.6269	0.8444	0.8747	0.8205	0.8444	0.8305	0.8744	NA
Gbm	0.7752	0.5923	0.8318	0.8605	0.8043	0.8318	0.8171	0.8606	5.6340
Bagging	0.7752	0.5933	0.8268	0.8617	0.8088	0.8268	0.8168	0.8591	NA
C4.5	0.7644	0.5803	0.8334	0.8594	0.7964	0.8334	0.8110	0.8249	NA
SvmPoly	0.6861	0.4094	0.7347	0.7919	0.7466	0.7347	0.7384	0.7679	4.1072
SvmRadial	0.6814	0.4073	0.7460	0.7920	0.7321	0.7460	0.7377	0.7676	NA
MLP	0.6539	0.4059	0.7620	0.8013	0.7462	0.7620	0.7360	0.7446	3.2832
NaiveBayes	0.6389	0.3850	0.6348	0.8022	0.7879	0.6348	0.6442	0.8018	6.3015

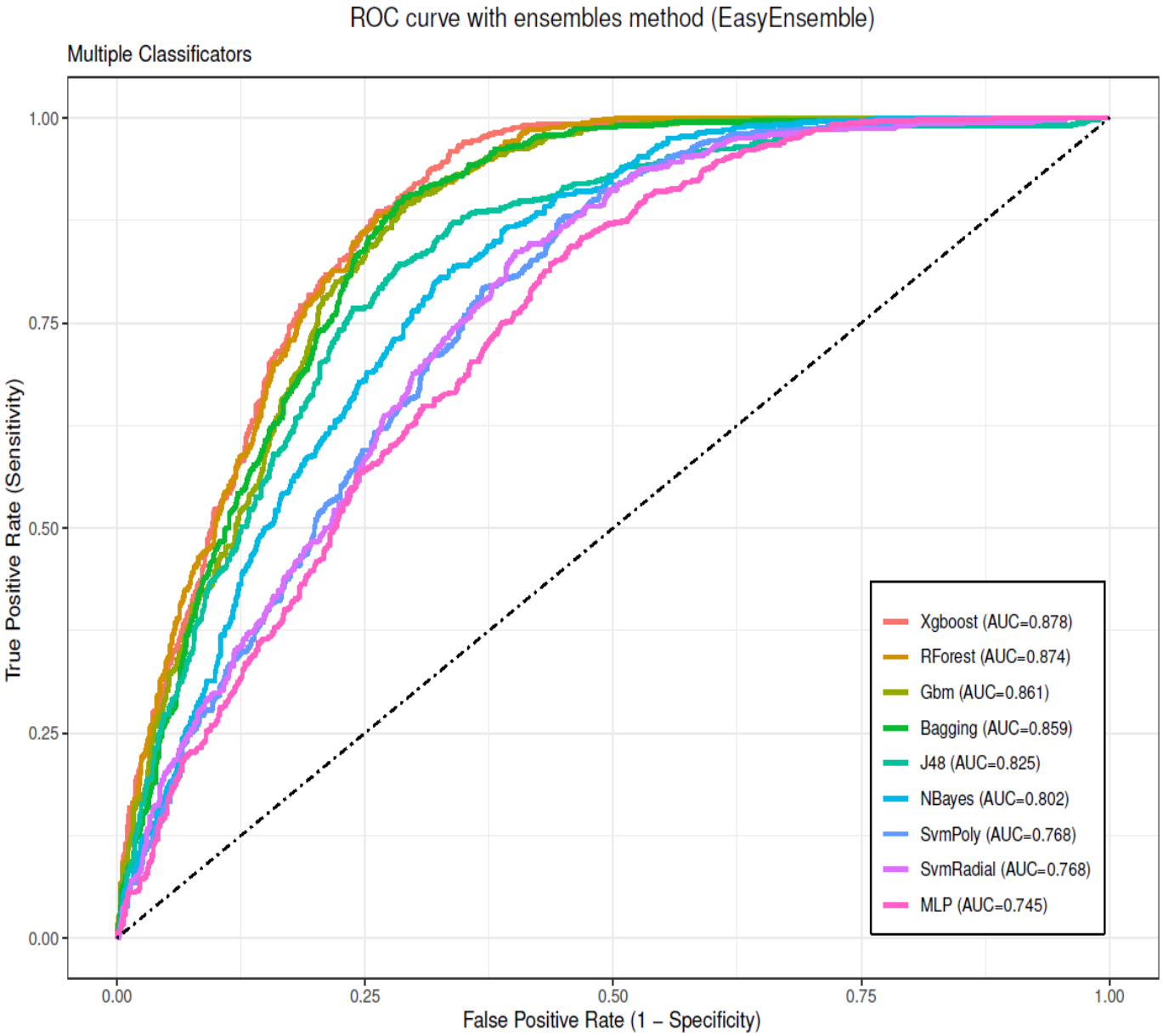


Figure 3[M5][JG6]. Performance of the group of algorithms by plotting the area under the AUC curve. On the ordinate axis is the true positive rate, and on the abscissa axis the false positive rate. The classifier lines above the diagonal (dashed line) represent good classification results (better than random), while those below represent bad results (worse than random). The best performance in classifying the test data examples was obtained by the XGBoost algorithm; two algorithms had an AUC above 0.87, the rest performed below 0.86. This performance clearly indicates the effectiveness of the predictive model against the test set.

It should be noted that the logloss was lower with the original data than with the preprocessed data. The increase with the latter was due to the smaller imbalance between classes. That is, the smaller the imbalance between classes, the greater the logloss, due to the smaller proportion of observations in the minority class. Table 3 shows the imbalance index between the original set and the dataset preprocessed with "EasyEnsemble" (column IR: 7.18 and 1.775 respectively).

In Table 6, the confusion matrix of the best-scoring algorithm (XGBoost) aimed to explain the predicted values of the test dataset, and the prediction model obtained by the algorithm was established. First, the type II error or β type error was analyzed, where (a) the "Dropout" class had predicted values of 868 cases, of which 741 were correct, and 127 cases

were classified as "Pass"; (b) the "Change" class had 126 cases, of which 115 were correct and 11 were classified as "Pass"; (c) the "Pass" class of the 679 predicted cases had 474 that were correct, four cases were classified as "Change", and 201 were classified as "Dropout". Secondly, the type I error or type α error was analyzed, where (a) the class "Dropout" had 942 cases, of which 741 were correct and 201 "Pass"; (b) the class "Change" had 119 cases, of which 115 were correct and four were classified as "Pass"; (c) the class "Pass" had 612 cases, of which 474 were correct, 11 were classified as "Change", and 127 were classified as "Dropout".

Table 6. Confusion matrix of the XGBoost algorithm. Here, the actual values (rows) are shown versus the values predicted by the classifier (columns).

Prediction / Actual	Dropout	Change	Pass	Total	Error Type II (β)
Dropout	741	0	127	868	0.8536
Change	0	115	11	126	0.9126
Pass	201	4	474	679	0.6980
Total	942	119	612	1673	μ = 0.8214
Error Type I (α)	0.7866	0.9663	0.7745	μ = 0.8431	

Overall, a more efficient predictive model was obtained with the XGBoost classification algorithm. In the work of [75], they highlight that the random forest algorithm obtained a better result in accuracy (ACC: 0.81) using only 10 features of the original dataset, pointing out the importance of improving academic performance and increasing the graduation rate of the students of the educational center. Consequently, it is necessary to consider that the accuracy of the model increases, and its complexity needs to be explainable as well. In this context, we looked for a way to apply a simple and readable method. The decision tree provides a simple rule-based model that improves comprehensibility. The use of the decision tree, although less efficient, is very easy to interpret. Figure 4 shows the decision tree generated from the training data and Figure 5 shows the important variables.

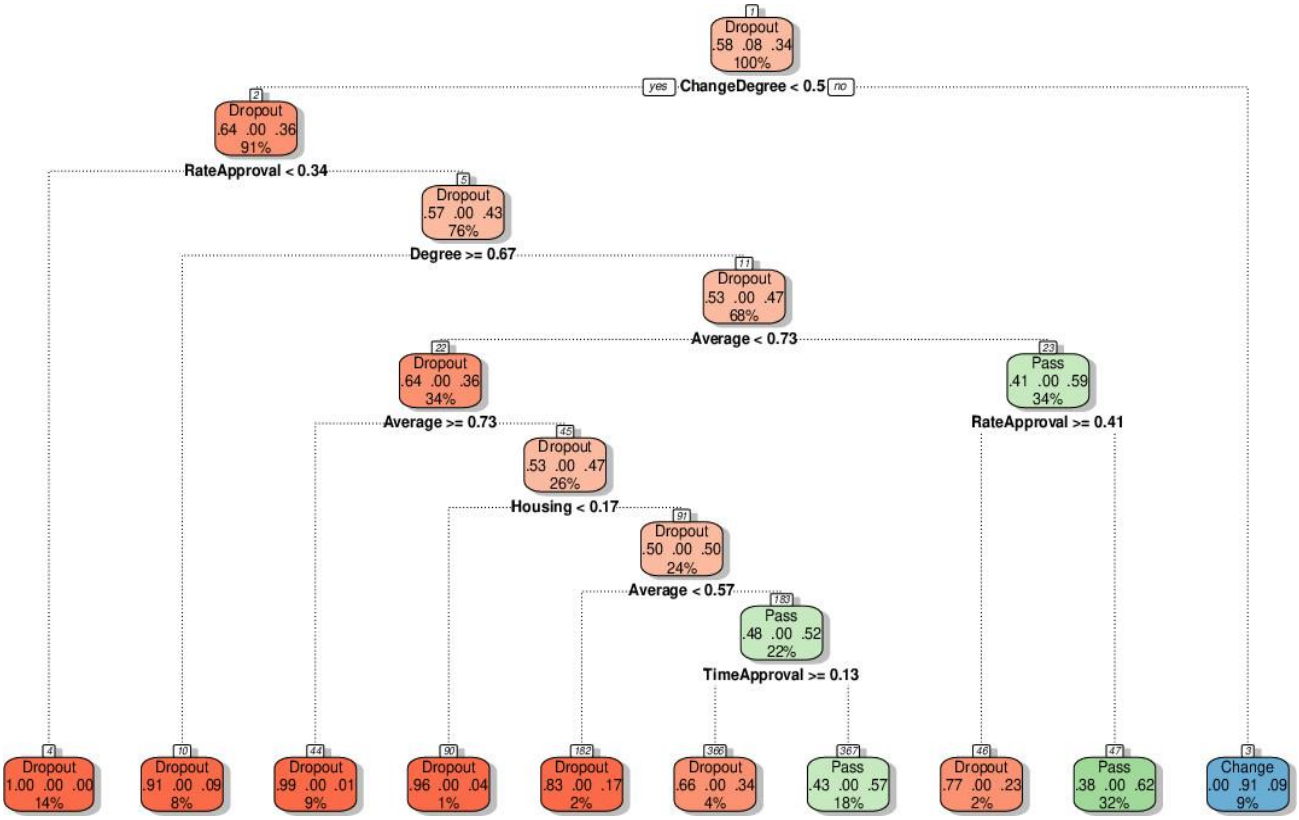


Figure 4. The decision tree drawn is based on the rules obtained. The nodes represent the class. The three decimal values within the node represent the probability of each class with respect to the evaluation of the rule. In turn, the total percentage of cases for the rule (cover) is shown. Below the node, the condition of the rule is displayed.

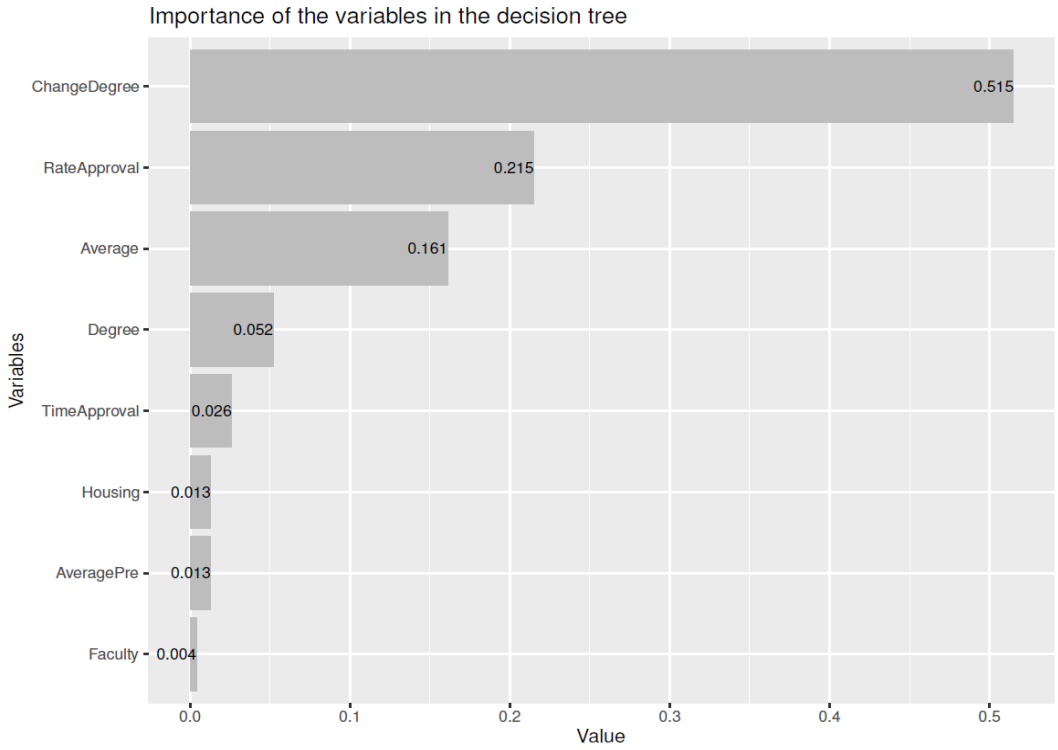


Figure 5. The importance of the variable is calculated by summing the decrease in error when divided by a variable. Thus, the higher the value, the more the variable contributes to improve the model, so the values are bounded between 0 and 1.

4.3. Static Comparison of Several Classifiers

Formally, statistical significance is defined as a probability measure to assess experiments or studies. Ronald Fisher promoted the use of the null hypothesis [76], establishing a significance threshold of 0.05 (1/20) to determine the validity of the results obtained in empirical tests. In this way, it is guaranteed that the provenance of their results is not due to chance coincidences. In the work of Demšar [77], the statistical significance of different classification algorithms and real-world datasets was validated by different empirical tests. In this context, the nonparametric Friedman and Wilcoxon tests were used, which are suitable for this type of analysis because they both do not skimp on the normal distribution of the data or on the homogeneity of variances, making them suitable for studies with data of a real or unmanipulated nature.

Prior to the calculation of the nonparametric tests, the results matrix of the group of algorithms and the datasets was organized, using the area under the curve (AUC, see Appendix D Table A6) as the metric. The significance threshold was set at 0.05 for the Friedman and Wilcoxon tests to determine if there were significant differences between more than two dependent groups. To perform the empirical tests, we used the null hypothesis H0: there are no significant differences between the groups of algorithms, and the alternative hypothesis, Ha: there is at least one significant difference between the groups of algorithms. The results of the Friedman test yielded a chi-square ($\chi 2$) of 52.305 with 8 degrees of freedom and a p-value of $1.47 \times 10{-}8$ (See Appendix D, Table A4). Since the p-value was below the threshold, the null hypothesis was rejected, and the alternative hypothesis was accepted, confirming the existence of significant differences. Next, a pairwise comparison of algorithms will be performed using the Wilcoxon test to assess the significance of these differences.

The above analysis established that there were significant differences, so a test was performed for each pair of algorithms using the Wilcoxon test, which is a Friedman post hoc test and is presented in Table 7, where the p-values obtained are shown.

Table 7. Wilcoxon signed rank test.

	XGBoost	RF	Gbm	Bagging	C4.5	NaiveBayes	SvmRadial	SvmPoly
RF	0.018	-						
Gbm	0.018	0.063*	-					
Bagging	0.018	0.018	0.018	-				
C4.5	0.018	0.018	0.018	0.018	-			
NaiveBayes	0.018	0.018	0.018	0.018	0.612*	-		
SvmRadial	0.018	0.018	0.018	0.018	0.028	0.018	-	
SvmPoly	0.018	0.018	0.018	0.018	0.028	0.018	0.398*	-
MLP	0.018	0.018	0.018	0.018	0.043	0.018	0.091*	0.128*

*Reject the null hypothesis.

According to the results, significant differences were found in RF vs. Gbm (0.063); C4.5 vs. NaiveBayes (0.612); SVMRadial vs. SVMPoly (0.398); SVMRadial vs. MLP (0.091); and SVMPoly vs. MLP (0.128) (See Appendix D, Table A5 for detailed results). In [78–80], opinions on statistics and significance tests have been discussed, because they are often misused, either by misinterpretation or by overemphasizing their results. It should be stated that statistical tests provide some assurance of the validity and non-randomness of the results [77].

5. Discussion

This paper explores and discusses three research questions related to machine learning techniques that are applied to achieve a predictive model with greater accuracy and readability, in addition to the study of factors that lead to the academic success of university students when they finish the first course. The answers to the questions posed are detailed.

RQ1: Which balancing and feature selection technique is relevant for supervised classification algorithms? In general, it is evident that with the increase in variables, the accuracy of the model increases, and so does its complexity, since the classification algorithms improve performance, although the readability of the model decreases. Against this in the work of Alwarthan [24], they apply recursive feature elimination (RFE) with Pearson correlation coefficient, RFE with mutual information and GA to find relevant features, in addition to class balancing using SMOTE-TomekLink to build the final prediction model. The relevant variables were related to English courses and GPA, as well as students' social variables. Alwarthan [24] used 68 features and achieved 93% accuracy with the initial results, while feature filtering detected 44 relevant variables and 90% accuracy. On the other hand, they analyzed eight relevant characteristics that achieved 77% accuracy; the variables were directly related to the academic performance of the student body.

In [6], the filtering of characteristics using the Gini index was proposed, from which seven characteristics were selected, achieving 79% accuracy using the random-forest algorithm. These results were very similar to ours, but far from being explainable, due to the bias derived from the imbalance of the data. In the proposal made in this study, different data processing techniques were used to obtain an expeditious dataset. On the one hand, the instance filtering method was considered to reduce duplicate or noisy observations by 5%. On the other hand, for feature group filtering, six methods were used, and five filters were applied, with which an accuracy between 58% and 78% was achieved. On the other hand, when applying the "ReliefF" method, 10 features were obtained with an accuracy of 79% (algorithm C4.5). In contrast, with the literature presented, the analyzed datasets had accuracy values below 84% and 32 features on average. The difference with what is proposed in this work is greater than 5% in accuracy, initially attractive. However, the handling of 22 additional features generates a robust and poorly explainable model for decision support.

Consequently, data balancing as part of data preprocessing was crucial to achieve a robust predictive model. The literature reviewed generically posits data balancing as a step prior to feature filtering. The approach taken so far is to obtain a filtered dataset (instances and features) and then apply data balancing. Among the best classification accuracies achieved by the data balancing methods, a range between 73% and 79% was obtained. The "EasyEnsemble" method obtained the best accuracy, AUC and logloss. The latter was far from the original data, as the imbalance rate was high. For example, the imbalance rates (IR) of the original data (7.35 IR) for undergraduate academic statuses (dropout, change and pass) were 57%, 7% and 36%, while for the balanced data (1.75 IR), they were 23%, 40% and 37% with synthetic observations. The accuracy of the XGBoost model with balanced data was approximately 80%. In summary, the proposed data preprocessing made the dataset unbiased and the predictive model simple and explainable.

RQ2: Which predictive model best discriminates students' academic success? Currently, there are several supervised algorithms used in higher education to predict different educational contexts in higher education. Specifically, the best discrimination was performed by the XGBoost algorithm. This criterion was based first on the values collected with the predictive model, where the accuracy value was 79.49% and the AUC was 87.75%. Sensitivity = 84.25%, which indicated the rate of positive examples that the algorithm was able to classify, while specificity = 87.53% for negative examples. Next, the logloss metric measuring computational cost had 0.3736 and an imbalance rate of 7.18 with the original dataset. However, the logloss value went to 6.34 with the preprocessed dataset and an imbalance rate of 1.775, i.e., lower computational cost and a higher data imbalance rate were inversely proportional to the performance of the predictive model. Although the predictive model obtained using XGBoost is poorly explainable due to its high complexity, it performed better by classifying examples from the test set. Explainability of the predictive model was obtained when the decision tree was applied to the training set to obtain a predictive model based on rules (If, Then) and readable for decision makers.

Similarly, [6,16–19,24,75] converge in their predictions on higher education data using classifiers such as Random Forest (RF), SVM, Neural Networks and decision trees. Likewise, linear regression or logistic regression was used to obtain predictive models that detect failure, success, or academic performance early enough [1,81], or in turn, semi-supervised

learning to obtain patterns in students who managed to pass the courses for a university degree [22]. Being the main objective to achieve very attractive and reliable accuracies, undoubtedly, accuracy always comes hand in hand with the quantity and quality of the data. For example, Gil [38] obtained accuracy rates with "random forest" of 77%, 91% and 94% with features of 30, 44 and 68, respectively, where the positive correlation between number of features and accuracy was evidenced. That said, in our results, accuracies very close to 80% were achieved with only 10 features and a completely readable model (10 rules).

RQ3: Which factors are determinants of students' academic success? As part of the development of this study, variables that play a significant role in the academic success of students were found. Specifically, the variables ChangeDegree, RateApproval, Average, and Degree were determinants for the prediction model obtained. These findings are close to the results obtained by Alturki [34], where individual results from the third and fourth semester were examined, both with accuracies of 63.33% (six variables) and 92.6% (nine variables), respectively. The influential variables were grade point average, number of credits taken and academic assessment performance, applying the selection of characteristics for each academic semester. Similarly, Alyahyan [23] identified variables related to GPA and key subjects that detect student performance early enough. As detailed by Beaulac [39] in their study, they identified variables associated with undergraduate degree completion as a first group of variables, whereas the second group of variables was related to the type of major. In summary, the first-year students opt for computer and English related subjects to reach their academic achievement, i.e., characteristics related to academic performance.

Specifically, data preprocessing provides as input an expedited dataset for classification algorithms to achieve an adequate predictive model. Although the results in the reviewed literature resemble ours, and these can be improved by inducing endogenous or exogenous variables for the model to achieve more optimal results, the results can also be improved by over-fitting parameters in the algorithms. It is also worth mentioning that, for example, Ismanto [82] obtained an RF prediction model with an accuracy higher than 90% without preprocessing the data, which resulted in a complex predictive model due to its explainability. Therefore, even if the model obtains the highest accuracy, the prediction bias can also be extended if the parameters are over-fitted or the data preprocessing phase is omitted.

Kaushik [83] has defined feature selection as increasing the quality in the data to facilitate better results, all according to the proposed method set of techniques for feature selection in educational data. What is applied in this paper fits with Kaushik's perspective. It is important to anticipate early enough and with general quality characteristics to take effective countermeasures, providing timely warnings to students to achieve academic success. In this way, the percentage of underachieving students can be reduced, and appropriate counseling and intervention can be provided to them by the college.

The results provide conclusive support for the anticipation of college completion [84–86], which is essential to assist students in the learning process and ensure their academic success. Thus, taking advantage of the fact that predictions made early enough by machine learning manage to reveal possible difficulties or improvements from students' historical data, its effective use requires building specific strategies [84]. Consequently, the application of the knowledge obtained from the data is leveraged, for example, in constant monitoring or continuous tracking that acts as a tool to assess progress in academic performance, class attendance, extracurricular activities and other key indicators [87]. Other strategies include personalized tutorial support or intervention plans, remediation and other resources for students who have demonstrated compelling needs [88,89]. Machine learning, along with other data analysis techniques, offers valuable suggestions for targeted interventions for the benefit of students, with the goal of helping them achieve academic success in the shortest possible time. The results presented support the authenticity of the analyses performed, as the information is not based on mere coincidences, but on real data. In this context, significant tests were performed using statistical methods such as the nonparametric Friedman and Wilcoxon test, which are widely recognized for comparing the performance of machine learning algorithms [77,90,91]. Although these tests are not recommended for a

comprehensive study, due to the need to conform to other assumptions, some authors have deepened their analysis and proposed alternatives to the tests [92,93]. In summary, significant tests are essential for a solid and objective interpretation of the results obtained.

6. Conclusions

In response to the research questions, the effectiveness of the prediction model lies in the good practice conducted in the data preprocessing phase. Hence, the importance of obtaining an expeditious dataset is crucial. Unlike the methodologies reviewed in the literature, our applied methodology avoided bias in the accuracy rates of the predictive model, as well as in the academic status (class). In fact, both the robust predictive model achieved by means of XGBoost as well as the simplified decision tree model proved to be effective. The simplified predictive model was able to detect students with high potential for academic success in seven out of ten cases, while the robust model detected them in eight out of ten cases. The simplification and explainability of the model were based on a set of rules obtained from the decision tree used, to make them understandable and provide them to academic experts as suggestions for decision making. Overall, this study provides valuable information on the factors underlying college students' academic success expectations and highlights the importance of effective data preprocessing and model simplification techniques for making accurate, meaningful, and understandable predictions about college students' academic success.

7. Limitations

The main limitation of this work was the absence of variables that help to have consistent measurements in the classification algorithms in terms of gender, scholarships, and financial aid, since it is important to analyze the evaluation of equity and discrimination aspects in the decisions made by the algorithms to build the predictive model.

8. Future Work

Looking ahead, we intend to explore how the knowledge extracted in this work and the university practices applied with this knowledge can influence classroom management, with the aim of improving students' academic outcomes and reducing the disparity in educational opportunities. To this end, we propose studies related to (i) examining how the personalization of predictive models can be adapted to the phenotype (characteristics) of the student body, where the objective is to examine the use of fuzzy logic to make uncertainty flexible and how the fuzzy model can manage the university context; (ii) designing early warning systems to intervene early and prevent failure or dropout; and (iii) other approaches, such as longitudinal studies, that aid evaluation and effectiveness over time to adjust the models as needed.

Author Contributions: Individual contributions: J.H.G.-F. and J.C.: conceptualization, methodology, software, validation, formal analysis, research, writing—writing of the original draft, writing—revising and editing, visualization, supervision; J.G.-M.: resources, writing—revising and editing, visualization, support, project administration. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The work is supported as part of the UTEQ-FOCICYT IX-2023/29 project "Factors that affect the completion of time to degree and affect the retention of UTEQ students" approved by the twentieth resolution made by the Honorable University Council dated 14 February 2023. The study keeps the respective confidentiality of the information stipulated in the Organic Law for the Protection of Personal Data of the Republic of Ecuador, in addition to the application of the respective "Code of Ethics for Officials and Servants Designated or Contracted by the Universidad Tecnica Estatal de Quevedo" approved by the Honorable University Council on 6 September 2011. Therefore, the research group has declared this research approved for publication in any journal with the document CERT-ETHIC-001-2023.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset is not available but can be obtained from the corresponding author upon reasonable request.

Acknowledgments: We would like to express our deep appreciation to the authorities of the IES for authorizing and allowing access, exploration, and analysis of the information, especially for the support provided by Javier Guaña-Moya, and Efraín Díaz Macías as Director of the project "Factors that influence the completion of the time to degree and affect the retention of students at UTEQ". The work is supported as part of the UTEQ-FOCICYT IX-2023/29 project.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

This section presents the information used in the work. The dataset used consists of data such as career, class attendance, students' academic performance and socioeconomic information. Numerical and categorical data are according to each variable.

Table A1. Description of the dataset used for the study.

Variable Names	Values	Description	Type
Faculty	1–5	Names of the faculties.	Categorical
Degree	1–27	Names of the university degrees.	Categorical
Sex	1. Male, 2. Female	Sex of students.	Categorical
Age Entrance	16–50	Age at entrance to university.	Numeric
Support	1. Public 2. Private	Type of financial support from the high school where the student completed high school.	Categorical
Localization	1. Local, 2. Outside of Quevedo, 3. Other Province	The geographical area of the school where the student finished high school.	Categorical
AveragePre	0–10	Average of the grades of the university leveling program (Pre-university/Admission/Selectivity).	Numeric
AttendancePre	0–100	Pre-university attendance percentage.	Numeric
Average	0–10	Average of the subjects taught in the first year.	Numeric
Attendance	0–100	Average of the student's attendance percentage in all subjects enrolled. Must meet the minimum attendance percentage of 70%.	Numeric
TimeApproval	1–3	Number of enrollments used by the student to pass the first course.	Numeric
RateApproval	0–3	Weighting of the effort in the exams to pass the subjects; the first exam (recovery) has a value of 0.25, while the second one has a value of 0.75.	Numeric
CounterDegree	0–2	The number of college courses in which the student was enrolled.	Numeric
StructureFamily	1. I am independent 2. Only with mom, 3. Only with dad, 4. Both parents, 5. Couple, 6. Other relative.	Variable associated with the student's family structure.	Categorical
Job	1. Does not work, 2. Full time,	This variable is linked to the student's work or occupational situation.	Categorical

	3. Part-time, 4. Part-time by the hour, 5. Occasionally.		
Financing	1. Family support, with 1 or 2 children studying. 2. Self-employed (own account). 3. Family support, with more than three children studying. 4. Loan, scholarship, or current credit.	This variable is related to the student's economic disposition to pay for the academic year.	Categorical
Zone	1. Outside of Quevedo, 2. Urban, 3. Slum, 4. Rural.	Describes the geographic district where the student lives.	Categorical
Income	1. More than $400, 2. Between $399 and $200, 3. Between $199 and $100, 4. Less than or equal to $99.	Monthly cash income (approximate) of the family nucleus.	Categorical
Housing	1. Own housing, 2. Rental, 3. Mortgaged, 4. Borrowed.	This variable is related to the usufruct of the housing where the student and his family live.	Categorical
ChangeDegree	1. Yes, 2. No.	This variable describes whether the student has changed degrees when repeating the first year.	Categorical
Class	1. Dropout, 2. Change, 3. Pass.	Variable with the student's academic status at the end of the university degree.	Categorical

Appendix B

Table A2 presents various results from the calculation of the metrics applied to the group of classification algorithms. The results presented in this appendix are complementary trainings, as six different balancing techniques were used to generate new datasets that were contributed to train and achieve effective predictive models. Each technique applied balancing methods related to oversampling, undersampling and combined balancing based on the SMOTE algorithm. The "EasyEnsemble" data balancing was the best performing of the algorithms and has been presented in the Results section as part of the data input supply for the group of classification algorithms to obtain the predictive model.

Table A2. Performance results of the classification algorithms that trained and tested the predictive models using new datasets constructed using the data balancing algorithms.

Bal.	Algorithms	Acc.	Kappa	Sensi.	Speci.	Preci.	Recall	F1	AUC	LogLoss
SMOTE	XGBoost	0.7878	0.6214	0.8472	0.8740	0.8237	0.8472	0.8318	0.8743	6.7890
	RF	0.7812	0.6118	0.8418	0.8720	0.8143	0.8418	0.8229	0.8671	
	Gbm	0.7723	0.5984	0.8446	0.8679	0.8105	0.8446	0.8205	0.8575	5.0183
	Bagging	0.7687	0.5887	0.8315	0.8633	0.8045	0.8315	0.8135	0.8546	
	C4.5	0.7639	0.5771	0.8248	0.8577	0.8008	0.8248	0.8101	0.7936	
	SvmPoly	0.6999	0.4740	0.7819	0.8242	0.7618	0.7819	0.7631	0.7640	5.8172
	SvmRadial	0.6970	0.4681	0.7835	0.8215	0.7587	0.7835	0.7630	0.7649	
	MLP	0.6545	0.4190	0.7779	0.8087	0.7552	0.7779	0.7350	0.7512	5.0928
	NaiveBayes	0.6198	0.3802	0.7478	0.7990	0.7817	0.7478	0.6957	0.8040	5.0538
KMeans.SMOTE	XGBoost	0.7956	0.6366	0.8565	0.8802	0.8262	0.8565	0.8367	0.8702	6.3079
	RF	0.7794	0.6080	0.8420	0.8702	0.8125	0.8420	0.8226	0.8600	
	Gbm	0.7693	0.5929	0.8396	0.8660	0.8060	0.8396	0.8160	0.8515	5.1901
	Bagging	0.7681	0.5860	0.8228	0.8620	0.8049	0.8228	0.8103	0.8467	
	C4.5	0.7663	0.5828	0.8259	0.8605	0.8036	0.8259	0.8113	0.7979	
	SvmPoly	0.6946	0.4613	0.7751	0.8187	0.7548	0.7751	0.7584	0.7616	5.8277
	SvmRadial	0.6892	0.4499	0.7717	0.8139	0.7492	0.7717	0.7551	0.7591	
	MLP	0.6712	0.4229	0.7703	0.8045	0.7353	0.7703	0.7452	0.7424	4.9865
	NaiveBayes	0.6067	0.3644	0.7505	0.7933	0.7804	0.7505	0.6862	0.7970	5.0905
SMOTE.Tomek	XGBoost	0.7914	0.6278	0.8474	0.8766	0.8241	0.8474	0.8320	0.8665	6.5445
	Bagging	0.7747	0.5970	0.8269	0.8656	0.8090	0.8269	0.8148	0.8468	
	Gbm	0.7741	0.6029	0.8430	0.8705	0.8137	0.8430	0.8199	0.8577	4.9046
	RF	0.7717	0.5922	0.8295	0.8639	0.8050	0.8295	0.8139	0.8562	
	C4.5	0.7579	0.5639	0.8088	0.8526	0.7947	0.8088	0.8001	0.7623	
	SvmPoly	0.6975	0.4722	0.7822	0.8242	0.7627	0.7822	0.7619	0.7634	5.7663
	SvmRadial	0.6910	0.4579	0.7749	0.8182	0.7550	0.7749	0.7565	0.7633	
	MLP	0.6724	0.4459	0.7885	0.8182	0.7631	0.7885	0.7483	0.7592	4.8305
	NaiveBayes	0.6372	0.4053	0.7622	0.8079	0.7805	0.7622	0.7104	0.7959	
SMOTE.ENN	XGBoost	0.7478	0.5573	0.8216	0.8542	0.7933	0.8216	0.7988	0.8335	5.9690
	Gbm	0.7406	0.5481	0.8239	0.8519	0.7923	0.8239	0.7965	0.8230	5.2251
	RF	0.7394	0.5492	0.8285	0.8534	0.7962	0.8285	0.7972	0.8192	
	Bagging	0.7352	0.5387	0.8179	0.8485	0.7908	0.8179	0.7926	0.8165	
	C4.5	0.7310	0.5274	0.8109	0.8429	0.7828	0.8109	0.7888	0.7548	
	SvmRadial	0.6880	0.4590	0.7846	0.8198	0.7594	0.7846	0.7589	0.7511	
	SvmPoly	0.6880	0.4598	0.7809	0.8206	0.7608	0.7809	0.7568	0.7490	5.5081
	MLP	0.6665	0.4398	0.7875	0.8169	0.7661	0.7875	0.7436	0.7650	4.7577
	NaiveBayes	0.6186	0.3810	0.7615	0.7989	0.7428	0.7615	0.6858	0.7764	4.2434
RUS	Gbm	0.7346	0.5346	0.8188	0.8455	0.7861	0.8188	0.7938	0.8102	5.1165
	XGBoost	0.7328	0.5344	0.8205	0.8466	0.7886	0.8205	0.7931	0.8173	4.4343

	RF	0.7304	0.5303	0.8187	0.8450	0.7868	0.8187	0.7914	0.8153	
	Bagging	0.7197	0.5062	0.8034	0.8346	0.7731	0.8034	0.7813	0.7962	
	C4.5	0.6987	0.4954	0.8131	0.8387	0.7957	0.8131	0.7667	0.7764	
	SvmRadial	0.6629	0.4081	0.7634	0.7992	0.7249	0.7634	0.7368	0.7360	
	SvmPoly	0.6605	0.4040	0.7622	0.7976	0.7275	0.7622	0.7374	0.7348	4.3219
	MLP	0.6402	0.3871	0.7610	0.7950	0.7320	0.7610	0.7251	0.7304	2.9982
	NaiveBayes	0.6031	0.3575	0.7428	0.7907	0.7773	0.7428	0.6827	0.7841	

Appendix C

This table presents the results of the filtering of characteristics using the different methods proposed in the study. Each method, according to its nature, filtered the group of variables that best represented the data. Then, the group of variables was evaluated with the C4.5 classification algorithm.

Table A3. Selection of characteristics used for the evaluation of the best group of variables. The best group of variables was selected by the RelefFbestK algorithm.

Filter	Var.	Method	Value	Acc.	Kappa	Sensi.	Speci.	Preci.	Recall	F1
Roughset consistency	11	Las Vegas	1.00	0.68	0.39	0.53	0.79	0.65	0.53	0.56
	9	SelectKBest	0.02	0.67	0.36	0.47	0.78		0.47	
	8	HillClimbing	1.00	0.62	0.21	0.41	0.73		0.41	
	9	Sequential Backward	1.00	0.67	0.36	0.47	0.78		0.47	
	9	Sequential Floating Forward	1.00	0.67	0.36	0.47	0.78		0.47	
	10	Genetic Algorithm	1.00	0.67	0.34	0.47	0.78		0.47	
	20	AntColony	1.00	0.78	0.60	0.83	0.86	0.81	0.83	0.82
Determination coefficient	13	Las Vegas	0.48	0.72	0.46	0.60	0.82	0.70	0.60	0.62
	9	SelectKBest	0.06	0.67	0.36	0.47	0.78		0.47	
	20	HillClimbing	0.48	0.78	0.60	0.83	0.86	0.81	0.83	0.82
	20	Sequential Backward	0.48	0.78	0.60	0.83	0.86	0.81	0.83	0.82
	20	Sequential Floating Forward	0.48	0.78	0.60	0.83	0.86	0.81	0.83	0.82
	20	Genetic Algorithm	0.48	0.78	0.60	0.83	0.86	0.81	0.83	0.82
	20	AntColony	0.48	0.78	0.60	0.83	0.86	0.81	0.83	0.82
Gini index	11	Las Vegas	1.00	0.68	0.39	0.53	0.79	0.65	0.53	0.56
	9	SelectKBest	0.51	0.67	0.36	0.47	0.78		0.47	
	8	HillClimbing	1.00	0.62	0.21	0.41	0.73		0.41	
	9	sequentialBackward	1.00	0.67	0.36	0.47	0.78		0.47	
	9	sequential Floating Forward	1.00	0.67	0.36	0.47	0.78		0.47	
	11	Genetic Algorithm	1.00	0.62	0.21	0.41	0.73		0.41	
	20	AntColony	1.00	0.78	0.60	0.83	0.86	0.81	0.83	0.82
Mutual information	12	Las Vegas	1.27	0.72	0.46	0.56	0.82	0.69	0.56	0.58
	9	SelectKBest	0.16	0.67	0.36	0.47	0.78		0.47	
	6	HillClimbing	1.27	0.58	0.09	0.37	0.69		0.37	
	8	Sequential Backward	1.27	0.62	0.21	0.41	0.73		0.41	
	8	Sequential Floating Forward	1.27	0.62	0.21	0.41	0.73	0.41		
	4	GeneticAlgorithm	1.27	0.67	0.34	0.47	0.78		0.47	
	20	AntColony	1.27	0.78	0.60	0.83	0.86	0.81	0.83	0.82
Gain ratio	7	Las Vegas	0.10	0.59	0.15	0.39	0.71		0.39	
	9	SelectKBest	0.13	0.67	0.36	0.47	0.78		0.47	
	7	HillClimbing	0.10	0.59	0.15	0.39	0.71		0.39	
	11	SequentialBackward	0.10	0.68	0.39	0.53	0.79	0.65	0.53	0.56
	11	Sequential Floating Forward	0.10	0.68	0.39	0.53	0.79	0.65	0.53	0.56
	1	GeneticAlgorithm	0.10	0.59	0.15	0.39	0.71		0.39	
	19	AntColony	0.10	0.72	0.48	0.60	0.82	0.71	0.60	0.62

Appendix D

This section presents the results of the nonparametric Friedman and Wilcoxon tests performed. For this purpose, the value of the AUC metric was used. The calculation was performed using the R statistical program. Table A4 presents the values obtained from the calculation of the Friedman test. Table A5 presents the matrix of the Wilcoxon test results, both the Z-value on the left and the p-value on the right. Table A6 is the matrix used for the calculation of the tests.

Table A4. Average Rankings of the algorithms.

Algorithm	Ranking
XGBoost	0.9999999999999998
RandomForest	2.2857142857142856
Gbm	2.714285714285714
Bagging	3.999999999999999
C4.5	5.999999999999999
NaiveBayes	5.428571428571429
SvmRadial	7.428571428571429
SvmPoly	7.571428571428571
MLP	8.571428571428571

Friedman statistic considering reduction performance (distributed according to chi-square with 8 degrees of freedom: 52.3047619047619 p-value computed by Friedman test: 1.474479383034577 × 10−8.

Table A5. Z Score and significance on Wilcoxon test (Z/p-value, within table).

Algorithms	XGBoost	RF c	Gbm	Bagging	C4.5	NaiveBayes	SvmRadial	SvmPoly
RF	−2.366 a/0.018	-						
Gbm	−2.366 a/0.018	−1.859 a/0.063 *	-					
Bagging	−2.371 a/0.018	−2.366 a/0.018	−2.366 a/0.018	-				
C4.5	−2.366 a/0.018	−2.366 a/0.018	−2.366 a/0.018	−2.366 a/0.018	-			
NaiveBayes	−2.366 a/0.018	−2.366 a/0.018	−2.366 a/0.018	−2.366 a/0.018	−0.507 b/0.612 *	-		
SvmRadial	−2.366 a/0.018	−2.366 a/0.018	−2.366 a/0.018	−2.366 a/0.018	−2.197 a/0.028	−2.366 a/0.018	-	
SvmPoly	−2.366 a/0.018	−2.366 a/0.018	−2.366 a/0.018	−2.366 a/0.018	−2.197 a/0.028	−2.366 a/0.018	−0.845 a/0.398 *	-
MLP	−2.366 a/0.018	−2.366 a/0.018	−2.366 a/0.018	−2.366 a/0.018	−2.028 a/0.043	−2.366 a/0.018	−1.690 a/0.091 *	−1.521 a/0.128 *

a Based on positive rankings. . b Based on negative rankings. c Random Forest. *Reject the null hypothesis.

Table A6. AUC value metrics with different classifiers and dataset.

DataSet	Algorithms								
	XGBoost	RF	Gbm	Bagging	C45	NaiveBayes	SvmRadial	SvmPoly	MLP
RawData	0.8997	0.8978	0.8930	0.8781	0.8308	0.8168	0.7973	0.7754	0.7621
EasyEnsemble	0.8775	0.8744	0.8606	0.8591	0.8249	0.8018	0.7676	0.7679	0.7446
SMOTE	0.8743	0.8671	0.8575	0.8546	0.7936	0.8040	0.7649	0.7640	0.7512
KmeansSMOTE	0.8702	0.8600	0.8515	0.8467	0.7979	0.7970	0.7591	0.7616	0.7424
SMOTETomek	0.8665	0.8562	0.8577	0.8468	0.7623	0.7959	0.7633	0.7634	0.7592
SMOTEENN	0.8335	0.8192	0.8230	0.8165	0.7548	0.7764	0.7511	0.7490	0.7650
RUS	0.8173	0.8153	0.8102	0.7962	0.7764	0.7841	0.7360	0.7348	0.7304

References

1. Realinho, V.; Machado, J.; Baptista, L.; Martins, M.V. Predicting Student Dropout and Academic Success. Data 2022, 7, 146.
2. Ortiz-Lozano, J.M.; Rua-Vieites, A.; Bilbao-Calabuig, P.; Casadesús-Fa, M. University student retention: Best time and data to identify undergraduate students at risk of dropout. Innov. Educ. Teach. Int. 2018, 57, 74–85.
3. Urbina-Nájera, A.B.; Téllez-Velázquez, A.; Barbosa, R.C. Patterns to Identify Dropout University Students with Educational Data Mining. Rev. Electron. De Investig. Educ. 2021, 23, e1507.
4. Lopes Filho JA, B.; Silveira, I.F. Early detection of students at dropout risk using administrative data and machine learning. RISTI—Rev. Iber. De Sist. E Tecnol. De Inf. 2021, 40, 480–495.
5. Guanin-Fajardo, J.H.; Barranquero, J.C. Contexto universitario, profesores y estudiantes: Vínculos y éxito académico. Rev. Iberoam. De Educ. 2022, 88, 127–146.
6. Zeineddine, H.; Braendle, U.; Farah, A. Enhancing prediction of student success: Automated machine learning approach. Comput. Electr. Eng. 2020, 89, 106903.
7. Guerrero-Higueras, M.; Llamas, C.F.; González, L.S.; Fernández, A.G.; Costales, G.E.; González, M.C. Academic Success Assessment through Version Control Systems. Appl. Sci. 2020, 10, 1492.
8. Rafik, M. Artificial Intelligence and the Changing Roles in the Field of Higher Education and Scientific Research. In Artificial Intelligence in Higher Education and Scientific Research. Bridging Human and Machine: Future Education with Intelligence; Springer: Singapore, 2023; pp. 35–46.
9. BOE. BOE-A-2023-7500 Ley Orgánica 2/2023, de 22 de marzo, del Sistema Universitario. 2023. Available online: https://www.boe.es/buscar/act.php?id=BOE-A-2023-7500 (accessed on 23 March 2024).
10. Guney, Y. Exogenous and endogenous factors influencing students' performance in undergraduate accounting modules. Account. Educ. 2009, 18, 51–73.
11. Tamada, M.M.; Giusti, R.; Netto, J.F.d.M. Predicting Students at Risk of Dropout in Technical Course Using LMS Logs. Electronics 2022, 11, 468.
12. Contini, D.; Cugnata, F.; Scagni, A. Social selection in higher education. Enrolment, dropout and timely degree attainment in Italy. High. Educ. 2017, 75, 785–808.
13. Costa, E.B.; Fonseca, B.; Santana, M.A.; De Araújo, F.F.; Rego, J. Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. Comput. Hum. Behav. 2017, 73, 247–256.
14. Márquez-Vera, C.; Cano, A.; Romero, C.; Noaman, A.Y.M.; Fardoun, H.M.; Ventura, S. Early dropout prediction using data mining: A case study with high school students. Expert Syst. 2015, 33, 107–124.
15. Fernández, A.; del Río, S.; Chawla, N.V.; Herrera, F. An insight into imbalanced Big Data classification: Outcomes and challenges. Complex Intell. Syst. 2017, 3, 105–120.
16. Rodríguez-Hernández, C.F.; Musso, M.; Kyndt, E.; Cascallar, E. Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation. Comput. Educ. Artif. Intell. 2021, 2, 100018.
17. Contreras, L.E.; Fuentes, H.J.; Rodríguez, J.I. Academic performance prediction by machine learning as a success/failure indicator for engineering students. Form. Univ. 2020, 13, 233–246.
18. Hassan, H.; Anuar, S.; Ahmad, N.B.; Selamat, A. Improve student performance prediction using ensemble model for higher education. In Frontiers in Artificial Intelligence and Applications; IOS Press: Amsterdam, The Netherlands, 2019; Volume 318, pp. 217–230.
19. Bolón-Canedo, V.; Alonso-Betanzos, A. Ensembles for feature selection: A review and future trends. Inf. Fusion 2018, 52, 1–12.

20. Meghji, A.F.; Mahoto, N.A.; Unar, M.A.; Shaikh, M.A. The role of knowledge management and data mining in improving educational practices and the learning infrastructure. Mehran Univ. Res. J. Eng. Technol. 2020, 39, 310–323.
21. Crivei, L.; Czibula, G.; Ciubotariu, G.; Dindelegan, M. Unsupervised learning based mining of academic data sets for students' performance analysis. In Proceedings of the SACI 2020—IEEE 14th International Symposium on Applied Computational In-telligence and Informatics, Proceedings, Timisoara, Romania, 21–23 May 2020; Volume 17, pp. 11–16.
22. Guanin-Fajardo, J.; Casillas, J.; Chiriboga-Casanova, W. Semisupervised learning to discover the average scale of graduation of university students. Rev. Conrado 2019, 15, 291–299.
23. Alyahyan, E.; Düşteargör, D. Decision trees for very early prediction of student's achievement. In Proceedings of the 2020 2nd International Conference on Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia, 3–15 October 2020; pp. 1–7.
24. Alwarthan, S.; Aslam, N.; Khan, I.U. An Explainable Model for Identifying At-Risk Student at Higher Education. IEEE Access 2022, 10, 107649–107668.
25. Adekitan, A.I.; Noma-Osaghae, E. Data mining approach to predicting the performance of first year student in a university using the admission requirements. Educ. Inf. Technol. 2018, 24, 1527–1543.
26. Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. Knowledge Discovery and Data Mining: Towards a Unifying Framework. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, 2–4 August 1996; pp. 82–88.
27. Haixiang, G.; Yijing, L.; Shang, J.; Mingyun, G.; Yuanyue, H.; Bing, G. Learning from class-imbalanced data: Review of methods and applications. Expert Syst. Appl. 2017, 73, 220–239.
28. Chawla, N.; Bowyer, K. SMOTE: Synthetic Minority Over-sampling Technique Nitesh. J. Artif. Intell. Res. 2002, 16, 321–357.
29. Bertolini, R.; Finch, S.J.; Nehm, R.H. Enhancing data pipelines for forecasting student performance: Integrating feature selection with crossvalidation. Int. J. Educ. Technol. High. Educ. 2021, 18, 44. [PubMed]
30. Febro, J.D. Utilizing Feature Selection in Identifying Predicting Factors of Student Retention. Int. J. Adv. Comput. Sci. Appl. 2019, 10, 269–274.
31. Ghaemi, M.; Feizi-Derakhshi, M.-R. Feature selection using Forest Optimization Algorithm. Pattern Recognit. 2016, 60, 121–129.
32. R Development Core Team. R: A Language and Environment for Statistical Computing; R Foundation for Statistical Computing: Vienna, Austria, 2020.
33. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 2011, 12, 2825–2830.
34. Alturki, S.; Alturki, N.; Stuckenschmidt, H. Using Educational Data Mining to Predict Students' Academic Performance for Applying Early Interventions. J. Inf. Technol. Educ. JITE. Innov. Pract. IIP 2021, 20, 121–137.
35. Fernández-García, A.J.; Rodríguez-Echeverría, R.; Preciado, J.C.; Manzano, J.M.C.; Sánchez-Figueroa, F. Creating a recommender system to support higher education students in the subject enrollment decisión. IEEE Access 2020, 8, 189069–189088.
36. Helal, S.; Li, J.; Liu, L.; Ebrahimie, E.; Dawson, S.; Murray, D.J.; Long, Q. Predicting academic performance by considering student heterogeneity. Knowl.-Based Syst. 2018, 161, 11.
37. Yağci, M. Educational data mining: Prediction of students' academic performance using machine learning algorithms. Smart Learn. Environ. 2022, 9, 1–19.
38. Gil, P.D.; Martins, S.d.C.; Moro, S.; Costa, J.M. A data-driven approach to predict first-year students' academic success in higher education institutions. Educ. Inf. Technol. 2020, 26, 2165–2190.

39. Beaulac, C.; Rosenthal, J.S. Predicting University Students' Academic Success and Major Using Random Forests. Res. High. Educ. 2019, 60, 1048–1064.
40. Fernandes, E.R.; de Carvalho, A.C. Evolutionary inversion of class distribution in overlapping areas for multiclass imbalanced learning. Inf. Sci. 2019, 494, 141–154.
41. Morales, P.; Luengo, J.; García, L.P.F.; Lorena, A.C.; de Carvalho, A.C.P.L.F.; Herrera, F.; Ciencias, I.D.; Paulo, U.D.S.; Av, T.S.-C.; Carlos, S.; et al. Noisefiltersr the noise-filtersr package. R J. 2017, 9, 219–228.
42. Zeng, X.; Martinez, T. A noise filtering method using neural networks. In Proceedings of the IEEE International Workshop on Soft Computing Techniques in Instrumentation and Measurement and Related Applications (SCIMA2003), Provo, UT, USA, 17 May 2003; pp. 26–31.
43. Verbaeten, S.; Assche, A. Ensemble methods for noise elimination in classification problems. In Multiple Classifier Systems. MCS 2003; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2003; pp. 317–325.
44. Ali, A.; Jayaraman, R.; Azar, E.; Maalouf, M. A comparative analysis of machine learning and statistical methods for evaluating building performance: A systematic review and future benchmarking framework. J. Affect. Disord. 2024, 252, 111268.
45. Rajula, H.S.R.; Verlato, G.; Manchia, M.; Antonucci, N.; Fanos, V. Comparison of Conventional Statistical Methods with Machine Learning in Medicine: Diagnosis, Drug Development, and Treatment. Medicina 2020, 56, 455. [PubMed]
46. García, S.; Luengo, J.; Herrera, F. Tutorial on practical tips of the most influential data preprocessing algo-rithms in data mining. Knowl.-Based Syst. 2016, 98, 1–29.
47. Cruz RM, O.; Sabourin, R.; Cavalcanti GD, C. Dynamic classifier selection: Recent advances and perspectives. Inf. Fusion 2018, 41, 195–216.
48. Yadav, S.K.; Pal, S. Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification. arXiv 2012, arXiv:1203.3832.
49. Bradley, A.P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit. 1997, 30, 1145–1159.
50. Nájera, A.B.U.; de la Calleja, J.; Medina, M.A. Associating students and teachers for tutoring in higher education using clustering and data mining. Comput. Appl. Eng. Educ. 2017, 25, 823–832.
51. Kononenko, I. Estimating Attributes: Analysis and Extensions of RELIEF. In European Conference on Machine Learning; Springer: Berlin/Heidelberg, Germany, 1994; pp. 171–182.
52. Liu, H.; Setiono, R. Feature selection and classification: A probabilistic wrapper approach. In Proceedings of the 9th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEAAIE´96), Fukuoka, Japan, 4–7 June 1996; pp. 419–424.
53. Zhu, Z.; Ong, Y.-S.; Dash, M. Wrapper–Filter Feature Selection Algorithm Using a Memetic Framework. IEEE Trans. Syst. Man Cybern. Part B 2007, 37, 70–76.
54. Liu, H.; Yu, L. Toward integrating feature selection algorithms for classification and clustering. IEEE Trans. Knowl. Data Eng. 2005, 17, 491–502.
55. Batista, G.E.A.P.A.; Monard, M.C. An analysis of four missing data treatment methods for supervised learning. Appl. Artif. Intell. 2003, 17, 519–533.
56. Kira, K.; Rendell, L. The feature selection problem: Traditional methods and a new algorithm. In Proceedings of the AAAI'92: Proceedings of the Tenth National Conference on Artificial Intelligence, San Jose, CA, USA, 12–16 July 1992; pp. 129–134.
57. Qian, W.; Shu, W. Mutual information criterion for feature selection from incomplete data. Neurocomputing 2015, 168, 210–220.

58. Sheinvald, J.; Dom, B.; Niblack, W. A modeling approach to feature selection. In Proceedings of the 10th International Conference on Pattern Recognition, Atlantic City, NJ, USA, 16–21 June 1990; Volume I, pp. 535–539.
59. Coefficient of Determination. In The Concise Encyclopedia of Statistics; Springer: New York, NY, USA, 2008; pp. 88–91.
60. Quinlan, J. Induction of decision trees. Mach. Learn. 1986, 1, 81–106.
61. Ceriani, L.; Verme, P. The origins of the Gini index: Extracts from Variabilità e Mutabilità (1912) by Corrado Gini. J. Econ. Inequal. 2011, 10, 421–443.
62. Pawlak, Z. Imprecise Categories, Approximations and Rough Sets; Springer: Dordrecht, The Netherlands, 1991; Volume 19, pp. 9–32.
63. Wang, D.; Zhang, Z.; Bai, R.; Mao, Y. A hybrid system with filter approach and multiple population genetic algorithm for feature selection in credit scoring. J. Comput. Appl. Math. 2018, 329, 307–321.
64. Douzas, G.; Bacao, F.; Last, F. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. Inf. Sci. 2018, 465, 1–20.
65. Batista, G.E.; Bazzan, A.L.; Monard, M.C. Balancing training data for automated annotation of keywords: A case study. WOB 2003, 3, 10–18.
66. Ivan, T. Two modifications of cnn. IEEE Trans. Syst. Man Commun. SMC 1976, 6, 769–772.
67. Liu, X.-Y.; Wu, J.; Zhou, Z.-H. Exploratory Undersampling for Class-Imbalance Learning. IEEE Trans. Syst. Man Cybern. Part B 2008, 39, 539–550.
68. Hearst, M.A. Support vector machines. IEEE Intell. Syst. 1998, 13, 18–28.
69. Almeida, L.B. C1. 2 multilayer perceptrons. In Handbook of Neural Computation; Oxford University Press: New York, NY, USA, 1997; pp. 1–30.
70. Breiman, L. Random forests. Ensemble Mach. Learn. 2001, 45, 5–32.
71. Friedman, J.H. Greedy function approximation: A gradient boosting machine. Ann. Stat. 2001, 29, 1189–1232.
72. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
73. Breiman, L. Bagging predictors. Mach. Learn. 1996, 24, 123–140.
74. Webb, G.I. Naïve Bayes. Encycl. Mach. Learn. 2010, 15, 713–714.
75. Shetu, S.F.; Saifuzzaman, M.; Moon, N.N.; Sultana, S.; Yousuf, R. Student's performance prediction using data mining technique depending on overall academic status and environmental attributes. In Advances in Intelligent Systems and Computing; Springer: Berlin/Heidelberg, Germany, 2021; Volume 1166, pp. 757–769.
76. Fisher, R.A. The Design of Experiments; Oliver & Boyd: Thomas Oliver, NY, USA, 1935.
77. Demšar, J. Statistical Comparisons of Classifiers over Multiple Data Sets. J. Mach. Learn. Res. 2006, 7, 1–30. Available online: http://jmlr.org/papers/v7/demsar06a.html (accessed on 9 April 2024).
78. Cohen, J. The eart is round ($p < 0.05$). Am. Psychol. 1994, 49, 997–1003.
79. Schmidt, F.L. Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. Psychol. Methods 1996, 1, 115–129.
80. Harlow, L.L.; Mulaik, S.A.; Steiger, J.H. (Eds.) Multivariate Applications Book Series. In What If There Were No Significance Tests? Lawrence Erlbaum Associates Publishers: Mahwah, NJ, USA, 1997.
81. Al-Fairouz, E.I.; Al-Hagery, M.A. Students Performance: From Detection of Failures and Anomaly Cases to the Solutions-Based Mining Algorithms. Int. J. Eng. Res. Technol. 2020, 13, 2895–2908.
82. Ismanto, E.; Ghani, H.A.; Saleh, N.I.B.M. A comparative study of machine learning algorithms for virtual learning environment performance prediction. IAES Int. J. Artif. Intell. 2023, 12, 1677–1686.

83. Kaushik, Y.; Dixit, M.; Sharma, N.; Garg, M. Feature Selection Using Ensemble Techniques. In Futuristic Trends in Network and Communication Technologies; FTNCT 2020. Communications in Computer and Information Science; Springer: Singapore, 2021; Volume 1395, pp. 288–298.
84. Mayer, A.-K.; Krampen, G. Information literacy as a key to academic success: Results from a longitudinal study. Commun. Comput. Inf. Sci. 2016, 676, 598–607.
85. Harackiewicz, J.M.; Barron, K.E.; Tauer, J.M.; Elliot, A.J. Predicting success in college: A longitudinal study of achievement goals and ability measures as predictors of interest and performance from freshman year through graduation. J. Educ. Psychol. 2002, 94, 562–575.
86. Meier, Y.; Xu, J.; Atan, O.; van der Schaar, M. Predicting Grades. IEEE Trans. Signal Process. 2015, 64, 959–972.
87. Lord, S.M.; Ohland, M.W.; Orr, M.K.; Layton, R.A.; Long, R.A.; Brawner, C.E.; Ebrahiminejad, H.; Martin, B.A.; Ricco, G.D.; Zahedi, L. MIDFIELD: A Resource for Longitudinal Student Record Research. IEEE Trans. Educ. 2022, 65, 245–256.
88. Tompsett, J.; Knoester, C. Family socioeconomic status and college attendance: A consideration of individual-level and school-level pathways. PLoS ONE 2023, 18, e0284188.
89. Ma, Y.; Cui, C.; Nie, X.; Yang, G.; Shaheed, K.; Yin, Y. Pre-course student performance prediction with multi-instance multi-label learning. Sci. China Inf. Sci. 2018, 62, 29101.
90. Berrar, D. Confidence curves: An alternative to null hypothesis significance testing for the comparison of classifiers. Mach. Learn. 2017, 106, 911–949.
91. Berrar, D.; Lozano, J.A. Significance tests or confidence intervals: Which are preferable for the comparison of classifiers? J. Exp. Theor. Artif. Intell. 2013, 25, 189–206.
92. García, S.; Herrera, F. An Extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all Pairwise Comparisons. J. Mach. Learn. Res. 2008, 9, 2677–2694.
93. Biju, V.G.; Prashanth, C. Friedman and Wilcoxon Evaluations Comparing SVM, Bagging, Boosting, K-NN and Decision Tree Classifiers. J. Appl. Comput. Sci. Methods 2017, 9, 23–47.

Bibliografía

Adamopoulou, E. y Moussiades, L. (2020). Chatbots: History, technology, and applications. *Machine Learning with Applications*, 2:100006.

Agrawal, R. y Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB_94, page 487–499, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Aina, C., Baici, E., Casalone, G., y Pastore, F. (2022). The determinants of university dropout: A review of the socio-economic literature. *Socio-Economic Planning Sciences*, 79.

Albreiki, B., Zaki, N., y Alashwal, H. (2021). A systematic literature review of student' performance prediction using machine learning techniques. *Education Sciences*, 11.

Alturki, S., Hulpuș, I., y Stuckenschmidt, H. (2022). Predicting academic outcomes: A survey from 2007 till 2018. *Technology, Knowledge and Learning*, 27:275–307.

Alyahyan, E. y Düştegör, D. (2020). Predicting academic success in higher education: literature review and best practices. *International Journal of Educational Technology in Higher Education*, 17(1):1–21.

Askari, S. H., Ahmad, F., Umair, S., y Khan, S. A. (2021). Cloud computing education strategies: A review. *Research Anthology on Architectures, Frameworks, and Integration Strategies for Distributed and Cloud Computing*, pages 2519–2530.

Baidoo-Anu, D. y Ansah, L. O. (2023). Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning. *SSRN Electronic Journal*.

Batanero, J. M. F., Rebollo, M. M. R., y Rueda, M. M. (2019). Impact of ict on students with high abilities. bibliographic review (2008–2018). *Computers and Education*, 137:48–58. REsultados encontrados factor importante la familia.

Burgos, C., Campanario, M. L., de la Peña, D., Lara, J. A., Lizcano, D., y Martínez, M. A. (2018). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers and Electrical Engineering*, 66:541–556.

Chaka, C. (2022). Educational data mining, student academic performance prediction, prediction methods, algorithms and tools: an overview of reviews. *Journal of E-Learning and Knowledge Society*, 18:58–69.

Chawla, N. V., Bowyer, K. W., Hall, L. O., y Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Costan, E., Gonzales, G., Gonzales, R., Enriquez, L., Costan, F., Suladay, D., Atibing, N. M., Aro, J. L., Evangelista, S. S., Maturan, F., Selerio, E., y Ocampo, L. (2021). Education 4.0 in developing economies: A systematic literature review of implementation barriers and future research agenda. *Sustainability (Switzerland)*, 13.

Csalódi, R. y Abonyi, J. (2021). Integrated survival analysis and frequent pattern mining for course failure-based prediction of student dropout. *Mathematics*, 9:1–17.

Cui, Y., Chen, F., Shiri, A., y Fan, Y. (2019). Predictive analytic models of student success in higher education: A review of methodology. *Information and Learning Science*, 120:208–227.

de la Cruz-Campos, J. C., Victoria-Maldonado, J. J., Martínez-Domingo, J. A., y Campos-Soto, M. N. (2023). Causes of academic dropout in higher education in andalusia and proposals for its prevention at university: A systematic review. *Frontiers in Education*, 8:1130952.

de Oliveira, C., Sobral, S., Ferreira, M., y Moreira, F. (2021). How does learning analytics contribute to prevent students' dropout in higher education: A systematic literature review. *Big Data and Cognitive Computing*, 5.

Dol, S. M. y Jawandhiya, P. M. (2023). Classification technique and its combination with clustering and association rule mining in educational data mining — a survey. *Engineering Applications of Artificial Intelligence*, 122.

Du, X., Yang, J., Shelton, B. E., Hung, J. L., y Zhang, M. (2021). A systematic meta-review and analysis of learning analytics research. *Behaviour and Information Technology*, 40:49–62.

Fayyad, U., Piatetsky-Shapiro, G., y Smyth, P. (1996). Knowledge Discovery and Data Mining: Towards a Unifying Framework. *Int Conf on Knowledge Discovery and Data Mining*, pages 82–88.

Fernández-García, A. J., Rodríguez-Echeverría, R., Preciado, J. C., Manzano, J. M. C., y Sánchez-Figueroa, F. (2020). Creating a recommender system to support higher education students in the subject enrollment decision. *IEEE Access*, 8:189069–189088.

Ghatak, A. (2017). *Machine Learning with R*. Springer Singapore.

Gironés Roig, J., Casas Roma, J., Minguillón Alfonso, J., y Caihuelas Quiles, R. (2017). *Minería de datos Modelos y Algoritmos*. Editorial UOC, Madrid, Barcelona.

Guabassi, I. E., Bousalem, Z., Marah, R., y Qazdar, A. (2021). A recommender system for predicting studentsádmission to a graduate program using machine learning algorithms. *International journal of online and biomedical engineering*, 17:135–147.

Gutierrez-Bucheli, L., Kidman, G., y Reid, A. (2022). Sustainability in engineering education: A review of learning outcomes. *Journal of Cleaner Production*, 330.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., y Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications.

Ho, I. M. K., Cheong, K. Y., y Weldon, A. (2021). Predicting student satisfaction of emergency remote learning in higher education during covid-19 using machine learning techniques. *PLoS ONE*, 16.

Ijaz, S., Safdar, T., y Sanaullah, M. (2020). Educational data mining: A review and analysis of student's academic performance. *Communications in Computer and Information Science*, 1198:510–523.

Jiménez-Macias, A., Moreno-Marcos, M., Muñoz-Merino, P., Ortiz-Rojas, M., y Kloos, C. (2022). Analyzing feature importance for a predictive undergraduate student dropout model. *Computer Science and Information Systems*, pages 50–50.

Kabathova, J. y Drlik, M. (2021). Towards predicting student's dropout in university courses using different machine learning techniques. *Applied Sciences (Switzerland)*, 11.

Karalar, H., Kapucu, C., y Gürüler, H. (2021). Predicting students at risk of academic failure using ensemble model during pandemic in a distance learning system. *International Journal of Educational Technology in Higher Education*, 18.

Kraft, D. y Moloney, C. (2020). *Introduction to Artificial Intelligence*, volume 163. Springer Science and Business Media Deutschland GmbH.

Livengood, J. M. (1992). Students'motivational goals and beliefs about effort and ability as they relate to college academic success. *Research in Higher Education*, 33(2):247–261.

López-Zambrano, J., Torralbo, J., y Romero, C. (2021). Early prediction of student learning performance through data mining: A systematic review | predicción temprana del rendimiento académico con minería de datos: una revisión sistemática. *Psicothema*, 33:456–465.

Maimon, O. y Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook*. Springer US.

Mántaras, R. L. D. (2020). ¿ Hacia una nueva Ilustración ? Una década trascendente El futuro de la IA : hacia inteligencias artificiales realmente inteligentes. *Instituto de Investigación en Inteligencia Artificial (IIIA)*, page 1 15.

Manzanares, M. C. S., del Camino Escolar Llamazares, M., y Medina, J. R. (2019). Investigación cualitativa: aplicación de métodos mixtos y de técnicas de minería de datos.

Martínez, F. J. M.-L. &. J. C. &. F. J. (2012). *SISTEMAS INTELIGENTES DE MARKETING PARA MODELADO CAUSAL*. Delta Publicaciones.

Masruroh, S. U., Rosyada, D., Zulkifli, Sururin, y Vitalaya, N. A. R. (2021). Adaptive recommendation system in education data mining using knowledge discovery for academic predictive analysis: Systematic literature review. *2021 9th International Conference on Cyber and IT Service Management, CITSM 2021*.

Namoun, A. y Alshanqiti, A. (2021). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences (Switzerland)*, 11:1–28.

Nawang, H., Makhtar, M., y Hamzah, W. M. A. F. W. (2021). A systematic literature review on student performance predictions. *International Journal of Advanced Technology and Engineering Exploration*, 8:1441–1453.

Oded, M. y Rokach, L. (2005). *Data Mining and Knowledge Discovery Handbook*. Springer-Verlag.

Okonkwo, C. W. y Ade-Ibijola, A. (2021). Chatbots applications in education: A systematic review. *Computers and Education: Artificial Intelligence*, 2.

Pascarella, E. T., Edison, M., Hagedorn, L. S., Nora, A., y Terenzini, P. T. (1996). Influences on studentsínternal locus of attribution for academic success in the first year of college. *Research in Higher Education*, 37(6):731–756.

Pathak, M. A. (2014). *Data Visualization*, pages 31–60. Springer International Publishing, Cham.

Pradeep, A. y Thomas, J. (2015). Predicting college students dropout using edm techniques. *International Journal of Computer Applications*, 123:26–34.

Pérez, J. Q., Daradoumis, T., y Puig, J. M. M. (2020). Rediscovering the use of chatbots in education: A systematic literature review. *Computer Applications in Engineering Education*, 28:1549–1565.

Rahayu, N. W., Ferdiana, R., y Kusumawardani, S. S. (2022). A systematic review of learning path recommender systems. *Education and Information Technologies*.

Rahul y Katarya, R. (2024). A systematic review on predicting the performance of students in higher education in offline mode using machine learning techniques. *Wireless Personal Communications 2024*, pages 1–32.

Razia, B. (2023). A systematic review of the use of blockchain in higher education. *Lecture Notes in Networks and Systems*, 485:631–648.

Respondek, L., Seufert, T., Stupnisky, R., y Nett, U. E. (2017). Perceived academic control and academic emotions predict undergraduate university student success: Examining effects on dropout intention and achievement. *Frontiers in Psychology*, 8(MAR):243.

Romero, C. y Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3:12–27.

Romero, C. y Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10.

Saa, A. A., Al-Emran, M., y Shaalan, K. (2019). Factors affecting students' performance in higher education: A systematic review of predictive data mining techniques. *Technology, Knowledge and Learning*, 24:567–598.

Sakr, N., Salama, A., Tameesh, N., y Osman, G. (2021). Edupal leaves no professor behind: Supporting faculty via a peer-powered recommender system. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12749 LNAI:302–307.

Samin, H. y Azim, T. (2019). Knowledge based recommender system for academia using machine learning: A case study on higher education landscape of pakistan. *IEEE Access*, 7:67081–67093.

Santana, M. y Díaz-Fernández, M. (2023). Competencies for the artificial intelligence age: visualisation of the state of the art and future perspectives. *Review of Managerial Science*, 17:1971–2004.

Sghir, N., Adadi, A., y Lahmer, M. (2022). Recent advances in predictive learning analytics: A decade systematic review (2012–2022). *Education and Information Technologies*.

Shi, L. y Zhu, Q. (2022). Association rule analysis of influencing factors of literature curriculum interest based on data mining. *Computational Intelligence and Neuroscience*, 2022.

Siri, A. (2015). Predicting students'dropout at university using artificial neural networks.

Soegoto, E. S., Soegoto, H., Soegoto, D. S., Soegoto, S. W., Rafdhi, A. A., Saputra, H., y Oktafiani, D. (2022). A systematic literature review of internet of things for higher education: Architecture and implementation. *Indonesian Journal of Science and Technology*, 7:511–528.

Sun, X., Fu, Y., Zheng, W., Huang, Y., y Li, Y. (2022). Big educational data analytics, prediction and recommendation: A survey. *Journal of Circuits, Systems and Computers*, 31.

Tavakoli, M., Faraji, A., Vrolijk, J., Molavi, M., Mol, S. T., y Kismihók, G. (2022). An ai-based open recommender system for personalized labor market driven education. *Advanced Engineering Informatics*, 52.

Verma, R. y Anika (2018). Applying predictive analytics in elective course recommender system while preserving student course preferences. *Proceedings of the 6th IEEE International Conference on MOOCS Innovation and Technology In Education, MITE 2018*, pages 52–59.

Wang, T., Lund, B. D., Marengo, A., Pagano, A., Mannuru, N. R., Teel, Z. A., y Pange, J. (2023). Exploring the potential impact of artificial intelligence (ai) on international students in higher education: Generative ai, chatbots, analytics, and international student success. *Applied Sciences (Switzerland)*, 13.

Xu, F., Li, Z., Yue, J., y Qu, S. (2021). *A Systematic Review of Educational Data Mining*, pages 764–780. Springer Nature.

Zaffar, M., Hashmani, M. A., Savita, K., y Khan, S. A. (2021). A review on feature selection methods for improving the performance of classification in educational data mining. *International Journal of Information Technology and Management*, 20:110.

Zlatic, L., Slibar, B., y Redep, N. B. (2021). Decision making styles in higher education institutions: Systematic literature review. *2021 44th International Convention on Information, Communication and Electronic Technology, MIPRO 2021 - Proceedings*, pages 826–832.