Fecha de presentación: julio, 2019, Fecha de Aceptación: septiembre, 2019, Fecha de publicación: octubre, 2019

38

APRENDIZAJE SEMI-SUPERVISADO PARA DESCUBRIR LA ESCALA DE TIEMPO PROMEDIO DE GRADUACIÓN DE ESTUDIANTES UNIVERSITARIOS

SEMI-SUPERVISED LEARNING TO DISCOVER THE AVERAGE SCALE OF GRADUATION OF UNIVERSITY STUDENTS

Jorge Guanin-Fajardo[1]
E-mail: jorgeguanin@uteq.edu.ec
ORCID: http://orcid.org/0000-0001-9150-4009
Jorge Casillas[2]
E-mail: casillas@decsai.ugr.es
ORCID: http://orcid.org/0000-0002-5887-3977
Washington Chiriboga-Casanova[1]
E-mail: chiriboga@uteq.edu.ec
ORCID: https://orcid.org/0000-0001-5166-5350
[1] Universidad Técnica Estatal de Quevedo. Ecuador.
[2] Universidad de Granada. España.

Cita sugerida (APA, sexta edición)

Guanin-Fajardo, J., Casillas, J., & Chiriboga-Casanova, W. (2019). Aprendizaje semi-supervisado para descubrir la escala de tiempo promedio de graduación de estudiantes universitarios. *Revista Conrado*, 15(70), 291-299. Recuperado de http://conrado.ucf.edu.cu/index.php/conrado

RESUMEN

Las instituciones de educación superior omiten hasta cierto punto los factores que retrasan las tasas de promoción de los estudiantes universitarios. El retraso no siempre puede ser revelado debido a la diversidad de los programas de estudio, desde el comienzo de la carrera hasta la finalización del programa y la graduación. En este trabajo se utilizó el conjunto de datos de estudiantes correspondiente a 5 cursos académicos completos (primero-quinto curso), 53 variables y 849 observaciones de las diferentes carreras universitarias. Así, se exploraron variables y se utilizó la minería de datos con técnicas de aprendizaje semi-supervisado para descubrir asociaciones que detectan categorías de graduación de estudiantes. Por lo tanto, las reglas de interés fueron descubiertas usando las métricas de support, confidence, lift y conviction de las reglas de la asociación. Los hallazgos sugieren que las edades del grupo de profesores entre segundo y tercer año, así como la categoría de nota media entre cursos y la empleabilidad de los estudiantes, son los principales factores que influyen en las tasas de graduación de los estudiantes universitarios.

Palabras clave:

Aprendizaje semi-supervisado, análisis educativo, minería de datos, educación superior, tiempo de graduación.

ABSTRACT

Institutions of higher education omit to a certain extent the factors that delay the rates of promotion of university students. The delay cannot always be disclosed due to the diversity of study programs, from the beginning of the career to the completion of the program and graduation. This paper used the student data set for 5 full academic years (grades 1-5), 53 variables, and 849 observations of different university careers. Thus, variables were explored and data mining with semi-supervised learning techniques was used to discover associations that detect graduation categories of students. Therefore, the rules of interest were discovered using the metrics of support, confidence and elevation of the rules of association. The findings suggest that the ages of the group of teachers between second and third year, as well as the grade point average between courses and the employability of students, are the main factors influencing the graduation rates of university students.

Keywords:

Semi-supervised learning, educational analysis, data mining, higher education, graduation time.

INTRODUCCIÓN

Las instituciones de educación superior además del personal relacionado con la enseñanza, la investigación, la administración y los servicios son considerados como una parte importante del ciclo de formación profesional de los estudiantes universitarios. Además, la administración de la gestión de la calidad y la toma de decisiones tiene lugar como resultado del procesamiento de datos y pruebas, en lugar de opiniones o suposiciones de expertos. Esto implica la implementación de diversos procedimientos que garanticen la calidad educativa (González López, 2006). Esta situación se halla enmarcada en que predecir el éxito académico del estudiantado es un desafío para las instituciones de educación superior, esto, debido a la heterogeneidad del alumnado, programas de estudio, docentes, infraestructura, acceso a recursos, etc. Así, la brecha entre la tasa de matrícula y graduación de las universidades se expande en la medida en que la universidad aumenta su oferta académica. En consecuencia, es necesario promover estudios académicos sobre el modelamiento y aplicación del aprendizaje semi-supervisado de minería de datos en la Educación Superior.

En el ámbito educativo, es posible encontrar trabajos previos con diversos métodos para predecir o encontrar patrones de comportamiento asociados con el desempeño del estudiantado. Más específicamente, en este trabajo usamos reglas de asociación. Así mismo Sawant & Shah (2016), estudió y comparó varios tipos de algoritmos de clasificación supervisada y no supervisada en minería de datos para seleccionar la mejor técnica. Posteriormente, la técnica no supervisada fue aplicada para identificar los factores del éxito académico. Por otro lado, Yang (2008), estudió el efecto de la influencia social en la predicción del rendimiento académico. Ya que, el mayor desafío viene de la dificultad de recoger una lista de amigos idónea para los estudiantes. Después los autores construyen la relación social de los estudiantes de acuerdo a su comportamiento en el plantel y predicen el rendimiento académico.

Además, construyen la red social mediante el aprendizaje semi-supervisado y evalúan el algoritmo propuesto. Por otro lado, Romero & Romero (2010), exploró la extracción de reglas de asociación raras recogiendo datos de estudiantes en una plataforma LMS específicamente Moodle. Por lo general, estas reglas se generan cuando los conjuntos de datos están desequilibrados. Este tipo de reglas es más difícil de encontrar cuando se aplican algoritmos tradicionales de minería de datos. Otro trabajo, también relacionado con las reglas de la asociación tiene que ver con He (2011), donde los autores tomaron como fuente de información a 300 estudiantes, ellos dividen la información en cinco categorías para el primer nivel y 14 categorías para el segundo nivel; así que en el primer nivel se consideró como fondo y aplicaron el algoritmo apriori para obtener las reglas que forman parte del conocimiento oculto de los datos.

Así mismo, Liu (2012), realizó el análisis de minería de datos con un conjunto de datos vinculados a las puntuaciones finales del estudiante. Como parte del trabajo, los autores procesan previamente los datos dividiéndolos en tres clases: Pobre, Media y Mejor. Una vez preparada esta información utilizan los algoritmos de clasificación como una herramienta de minería de datos para obtener conocimiento. De todos los trabajos revisados Sawant (2016), está en consonancia con nuestra investigación, ya que usamos técnicas de preprocesamiento de datos y técnicas no supervisadas para la obtención de patrones útiles. Sin embargo, a pesar de su potencial importancia a nivel académico es un área poco investigada, por lo tanto, se requiere nuevas investigaciones que profundicen su aplicabilidad.

En este contexto planteamos como objetivo determinar a través de la minería de datos, cuáles son los factores que se asocian con la escala del tiempo promedio de graduación del alumnado. Para este propósito, adquirimos información relevante usando reglas de asociación estimulando el consecuente con una salida deseada.

MATERIALES Y MÉTODOS

En esta sección abordamos el estudio del aprendizaje semi-supervisado utilizando técnicas de aprendizaje no supervisado. Concretamente usamos reglas de asociación estimulando su consecuente por medio de una variable categórica. Los datos obtenidos para este estudio se recuperaron de la base de datos institucional de un centro de estudio universitario en Ecuador, el conjunto de datos dispone de variables numéricas y categóricas, en concreto, 53 variables y 849 observaciones de estudiantes de las diferentes carreras universitarias.

DESARROLLO

En la publicación de García, Luengo, & Herrera (2016), se ha manifestado que una parte esencial para cualquier algoritmo en minería de datos es disponer de un conjunto de datos íntegro y limpio. Basándonos en este trabajo hemos efectuado el pre-procesado de los datos. En primer lugar, realizamos el filtrado de los datos atípicos, después etiquetamos estos datos como un dato perdido, esto, con el fin de que sean usados en la siguiente fase del preprocesado de datos. En segundo lugar, las observaciones con datos perdidos fueron remplazadas por un valor que es asignado por la función rfimpute del algoritmo

Random Forest (Breiman, 2001), que resultó más eficiente y precisa al tratar anomalías en los datos. En tercer lugar, aplicamos un filtrado de características, para seleccionar las variables más relevantes. En cuarto lugar, obtenidas las características relevantes del conjunto de datos, hemos efectuado la filtración de observaciones usando la función NoiseFiltersR (Morales, et al., 2017). Por último, hemos balanceado los datos con el algoritmo SMOTE (Chawla, Bowyer, Hall & Kegelmeyer, 2002). Es decir, se ha equiparado las observaciones para evitar el sesgo en los datos. Todo esto para filtrar los de mayor relevancia, así el conjunto de datos original se ha reducido en 16 variables descrito en la Tabla 1.

Tabla 1. Descripción de variables del conjunto de datos estudiado.

Variable	Tipo	Valores	Descripción
carrera	numérico	[1-14]	Carrera universitaria.
sostenimiento	categórico	{Público, Privado}	Colegio donde obtuvo el bachillerato.
taprobacion_1	numérico	[0,3]	Cantidad de cursos repetidos en el primer año de la carrera.
pedad2	numérico	[0,15]	Cantidad de docentes con edades menores a 40 años, que dieron clases en primer curso.
taprobacion_2	numérico	[0,3]	Cantidad de cursos repetidos en el segundo año de la carrera.
sedad2	numérico	[0,15]	Cantidad de docentes con edades entre 46 y 60 años, que dieron clases en segundo curso.
taprobacion_3	numérico	[0,3]	Cantidad de cursos repetidos en el tercer año de la carrera.
tedad3	numérico	[0,15]	Cantidad de docentes con edad superior a 60 años, que dieron clases en tercer curso.
habito estudios	categórico	{si, no}	Hábito de estudios.
tamaño familia	numérico	[1-3]	Número de personas constituida por la familia.
jornada trabajo	categórico	{Tiempo completo, medio tiempo, tiempo parcial, eventual}	Jornada laboral del estudiante.
Núm. Matriculas	numérico	[1-15]	Total de matrículas acumuladas.
financiamiento	categórico	{Fondos propios, crédito, beca}	Tipo de financiamiento de estudios.
estratoAprob	categórico	{Muy alta, alta, baja}	Cantidad total de cursos aprobados
estExpDoc	entero	[1-3]	Promedio de años de experiencia del profesorado.
estrato graduación	categórico	{Muy alta, alta, baja}	Medida de tiempo en años para graduarse.

Diseño experimental

El objetivo del trabajo lo hemos centrado en analizar la escala de tiempo promedio de graduación del estudiante. Durante los últimos años ha existido un gran interés en mejorar las tasas de graduación de parte de las instituciones de educación superior. En trabajos similares, se han calculado diversas medidas para evaluar la importancia de los patrones encontrados en el aprendizaje automático y determinar las reglas de asociación de mayor importancia (Berzal & Cubero, 2010; Kumar & Chadha, 2012; Prajapati, Garg & Chauhan, 2017; Varshali & Jitendra, 2012)there is an increasing demand in mining interesting patterns from the big data. The process of analyzing such a huge amount of data is really computationally complex task when using traditional methods. The overall purpose of this paper is in twofold. First, this paper presents a novel approach to identify consistent and inconsistent association rules from sales data located in distributed environment. Secondly, the paper also overcomes the main memory bottleneck and computing time overhead of single computing system by applying computations to multi node cluster. The proposed method initially extracts frequent itemsets for each zone using existing distributed frequent pattern mining algorithms. The paper also compares the time efficiency of Mapreduce based frequent pattern mining algorithm with Count Distributed Algorithm (CDA. Por ello, el proceso de extracción de reglas lo hemos realizado de la siguiente manera. En primer lugar, hemos transformado las variables numéricas a categóricas. En segundo lugar, hicimos un estudio exploratorio de las variables para detectar la magnitud de correspondencia entre variables. En tercer lugar, el conjunto de datos lo hemos convertido en transacciones para facilitar la detección de reglas. En cuarto lugar, se filtró las reglas más relevantes de acuerdo con las métricas planteadas en este trabajo. Por último, las reglas de mayor interés son seleccionadas y mostradas en gráficos para facilitar la comprensión de las reglas descubiertas.

Métricas usadas.

En este trabajo se incluyó métricas de importancia para evaluar las reglas de asociación, hemos usado ***Confidence***, ***Support***, ***Lift*** y ***Conviction.*** Esto, con el fin de reducir el volumen de reglas, y conseguir las reglas con mejor representación (Han, et al., 2017).

$$Confidence = (X \rightarrow Y) = P(X|Y) = \frac{Support(X \cup Y)}{Support(X)} \quad (1)$$

La confianza es el porcentaje de transacciones en el conjunto de datos D con el conjunto de elementos X que también contiene el conjunto de elementos Y.

$$Support = (X \rightarrow Y) = P(X \cup Y) \quad (2)$$

El soporte es el porcentaje de transacciones en el conjunto de datos D que contienen ambos conjuntos de elementos X e Y.

$$Lift(X \rightarrow Y) = \frac{Confidence(X \rightarrow Y)}{Support(Y)} \quad (3)$$

El valor de lift 1 indica que X e Y aparecen tan frecuentemente juntos bajo la suposición de independencia condicional.

$$Conviction(X \rightarrow Y) = \frac{1 - Support(X)}{1 - Confidence(X \rightarrow Y)} \quad (4)$$

Conviction mide la fuerza de implicación de la regla de independencia estadística, donde P (Y) es la probabilidad de que Y no aparezca en una transacción.

Obtención de reglas de asociación.

Hemos obtenido el patrón de reglas transformando el conjunto de datos en transacciones, con el fin de que el algoritmo trabaje con ellas y devuelva reglas de asociación. Para este fin, hemos creado una rutina en algoritmo 1 que permitió extraer las reglas usando todas las posibles combinaciones entre las métricas de soporte y confianza. También, a esta rutina hemos establecido el número de términos que componen el antecedente de la regla y el consecuente (Tabla 2).

Tabla 2. Reglas de Asociación Generada.

Algoritmo 1. Pseudocódigo de reglas de asociación generadas
Mientras iterminos entre 3 y 9 inicio Mientras isoporte entre 0.05 y 1 Inicio Mientras iconfianza entre 0.05 y 1 inicio reglas Obtener Reglas (iterminos, isoporte, iconfianza) fin fin fin

En esta sección analizamos los resultados encontrados, en respuesta a la pregunta de investigación sobre cuáles son las variables que se relacionan con el tiempo de graduación de los estudiantes.

Es importante conocer la actividad o comportamiento de los datos, ya que esto ha permitido en primer plano tener una idea más clara de los datos. Para ello, hemos usado la matriz de correlación y el método de correlación de Pearson que presentamos en la Figura 1. Aquí se calculó la magnitud o el grado de asociación entre las variables.

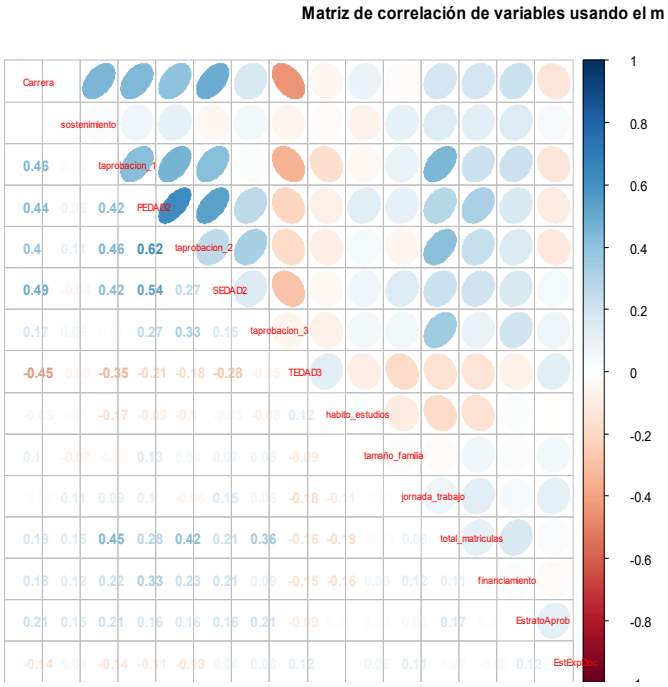

Figura 1. Matriz de correlación entre variables usando el coeficiente de Pearson.

En la figura 1 la correspondencia gradual de variables tanto positivas (azulado) como negativas (rojizo). Las formas de las elipses determinan el grado de correspondencia, cuanto más fina la elipse mayor correspondencia existió entre las variables.

Transacciones del conjunto de datos.

En esta etapa, se consiguió transformar las observaciones en transacciones. Las transacciones simplifican en 364 de las primeras 849 iniciales. Esta reducción es notoria debido a la eliminación de transacciones redundantes y menos significativas. En tal sentido mostramos la Figura 2 que muestra las etiquetas (ítems) de las transacciones a través de un mapa de calor.

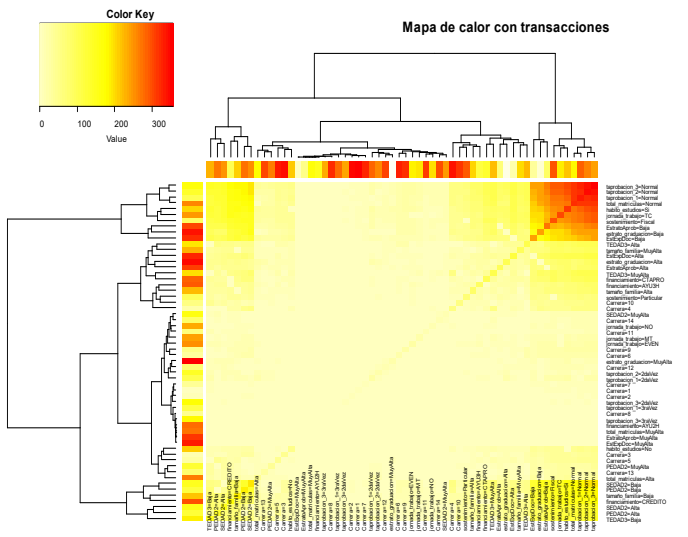

Figura 2. Mapa de calor de transacciones del conjunto de datos, con concentraciones de transacciones más relevantes en la parte superior derecha.

El color rojizo indicó mayor concentración de transacciones. Las transacciones tienen un orden de agrupamiento jerárquico. La cantidad de transacciones se definió por el color, el cuadro superior izquierdo valora el número de transacciones y la proximidad del color**.**

En la Fig. 2, las etiquetas del eje X e Y con mayor concentración de datos se han agrupado de manera jerárquica y aparecen con mayor frecuencia en las transacciones. De modo que las etiquetas frecuentes de acuerdo con la figura son: taprobacion_1, taprobacion_2, taprobacion_3, total-matriculas, habito-estudios, jornada-trabajo, sostenimiento, EstratoAprob, estrato_graduacion y EstratoExpDoc.

Como parte de la exploración de reglas de asociación, hemos explorado el conjunto de reglas de forma general, en primera instancia obtuvimos 41 019 reglas, luego discriminamos las reglas que tuvieron un valor de ***lift*** inferior a 1 quedando reducido a 34 593 reglas. En segundo lugar, debido a que el alto número de reglas representa una vaguedad para ofrecer un conocimiento válido, hemos eliminado las reglas redundantes quedando un total de 35 064 reglas. La idea principal es reducir las reglas a un punto que sean manejables e interpretables, es por esto, que también filtramos las reglas de acuerdo con la métrica ***conviction*** y hemos conseguido finalmente 18 298 reglas. Aunque, todavía sigue siendo un número de reglas alto, se logró reducir un 45% de reglas que carecían de importancia. De esta manera, nos hemos quedado con reglas relevantes y de mayor utilidad para el análisis.

En esta etapa, usamos el aprendizaje semi-supervisado, que consistió en establecer una salida personalizada de las reglas de asociación. En este sentido, hemos establecido una variable con etiquetas de interés que permitió obtener un patrón útil respecto al tiempo de graduación del estudiantado. Por ello, hemos considerado un nuevo conjunto de reglas con el consecuente inducido, es decir, hemos forzado que el consecuente tuviera solo la etiqueta vinculada con la escala de tiempo de graduación. Esto, para obtener un patrón de regla útil. Las etiquetas se asocian al tiempo que tarda el estudiantado en graduarse: "Muy Alta" (superior a 10 años), "Alta" (entre 7 y 10 años) y "Baja" (menor a 7 años) (Figura 3).

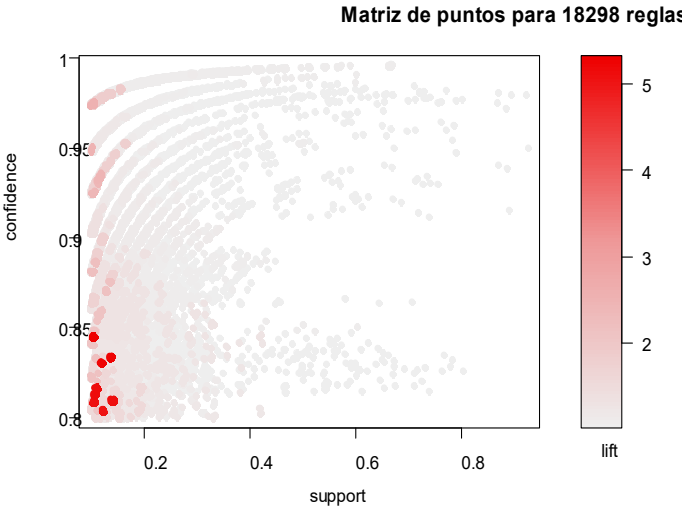

Figura 3. Matriz de puntos de las reglas de asociación filtrada de acuerdo con la métrica lift. Los puntos más rojizos representan a reglas de asociación de alto interés.

Los parámetros usados para obtener un conjunto de reglas adecuado fueron confianza=0.86, Soporte =0.01 y el número de etiquetas del antecedente = 5. En la Tabla 3 mostramos el conjunto de reglas obtenido y la valoración de las métricas. Las reglas encontradas se encuentran ordenadas de forma descendente según la columna conviction (Conv). De acuerdo con estos resultados, llama la atención las primeras cuatro reglas de asociación, donde el consecuente con la variable inducida ha demostrado que el tiempo de espera promedio para la graduación del estudiantado es "muy alta" y "alta".

Tabla 3. Conjunto de reglas de asociación encontradas.

Antecedente		Consecuente	Medidas			
			Sop.	Conf.	Lift.	Conv.
taprobacion_1=2daVez + taprobacion_2=Normal	→	Alta	0.049	0.947	3.448	13.780
sostenimiento=Particular + SEDAD2=Alta + taprobacion_3=Normal + EstratoAprob=Alta	→	Alta	0.030	0.917	3.337	8.703
Carrera=12 + SEDAD2=Alta + TEDAD3=Baja	→	Muy Alta	0.019	0.875	15.925	7.560
Carrera=12 + TEDAD3=Baja + financiamiento=CREDITO	→	Muy Alta	0.019	0.875	15.925	7.560
taprobacion_1=Normal + SEDAD2=Baja + financiamiento=CREDITO	→	Baja	0.203	0.892	1.330	3.040
taprobacion_2=Normal + SEDAD2=Baja + TEDAD3=Muy Alta	→	Baja	0.187	0.883	1.317	2.821
taprobacion_1=Normal + PEDAD2=Baja + SEDAD2=Baja + jornada_trabajo=TC	→	Baja	0.228	0.874	1.303	2.610
SEDAD2=Baja + TEDAD3=Muy Alta	→	Baja	0.187	0.872	1.301	2.571
SEDAD2=Baja + financiamiento=CREDITO	→	Baja	0.203	0.871	1.299	2.547
taprobacion_2=Normal + taprobacion_3=Normal + TEDAD3=MuyAlta + jornada_trabajo=TC	→	Baja	0.184	0.870	1.298	2.538
sostenimiento=Fiscal + taprobacion_1=Normal + PEDAD2=Baja + SEDAD2=Baja	→	Baja	0.236	0.869	1.296	2.511
taprobacion_1=Normal + taprobacion_2=Normal + SEDAD2=Baja + tamaño_familia=Baja	→	Baja	0.225	0.863	1.288	2.409

Las reglas mostradas en la Tabla 3 con el consecuente “Muy alta”, tienen una confianza del 87.5% de aparición en las transacciones, los términos del antecedente SEDAD2 y TEAD3 se han relacionado con edades de profesores que dictaron clases en el segundo y tercer año respectivamente, SEDAD2=Alta, que fue el número de profesores entre 46 y 60 años; y TEDAD3=Baja, que fue el número de profesores con edades superiores a 60 años. Adicionalmente, la etiqueta carrera=12 se refirió a la carrera de ingeniería agropecuaria.

Visualización de los patrones.

De acuerdo con los hallazgos encontrados, en la Tabla 4 hemos resumido las etiquetas comunes del antecedente respecto al consecuente.

Tabla 4. Etiquetas comunes en las reglas de asociación.

Etiquetas comunes	Indecencia en la graduación
Carrera=12 TEDAD3=Baja SEDAD2= Alta	Muy alta
TEDAD3=Baja, taprobacion_1=2daVez, taprobacion_2=2daVez	Alta
EstratoAprob=Baja TEDAD3=Muy Alta taprobacion_1= Normal taprobacion_2=Normal	Baja

En la Tabla 4, la incidencia de graduación “Baja” tuvo etiquetas de taprobacion-1 y taprobacion-2 como “Normal”, TEDAD3=Muy Alta y, EstratoAprob = “Bajo”. El estrato de aprobación indicó tres niveles bajo, alto y muy alto. Es decir, que el curso es superado sin la presentación a exámenes de suspensión.

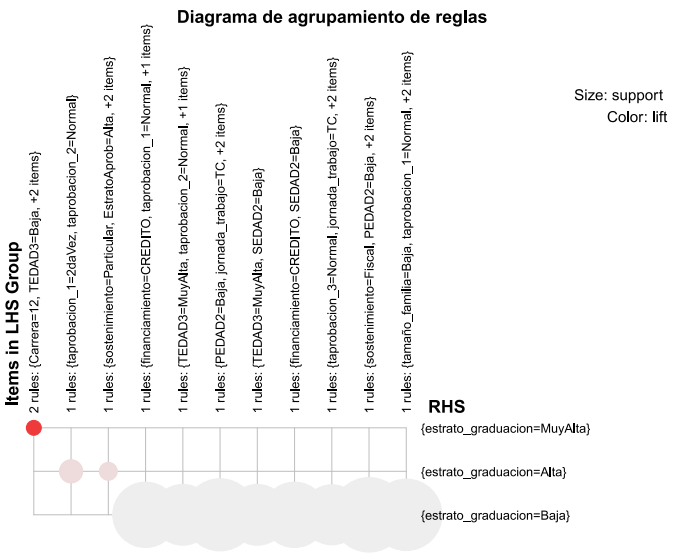

Figura 4. Agrupamiento de reglas de asociación, en esta figura se muestra el consecuente RHS con las tres categorías de la variable que sirve de pivote para filtrar las reglas.

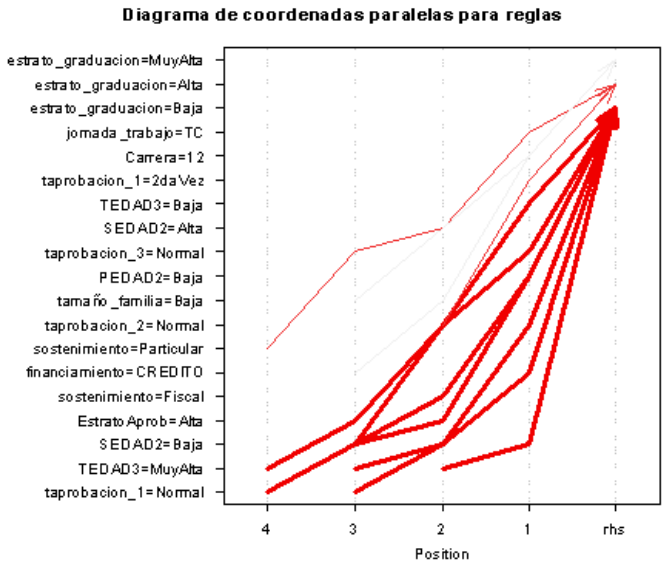

Figura 5. Diagramas de coordenadas paralelas que ha permitido visualizar el recorrido de las etiquetas hasta el consecuente. El color rojizo de las líneas representa.

Por otro lado, en las Figuras 4, 5 y 6 hemos presentado las reglas de asociación obtenidas de forma visual. Basados en el trabajo de Yang (2008)one on\\neach parallel coordinate, with polynomial curves. In the presence\\nof item taxonomy, an item taxonomy tree is displayed as coordinate\\nand can be expanded or shrunk by user interaction. This interaction\\nintroduces a border in the generalized itemset lattice, which separates\\ndisplayable itemsets from non-displayable ones. Only those frequent\\nitemsets on the border are displayed. This approach can be generalized\\nto the visualization of general monotone Boolean functions on lattice\\nstructure. Its usefulness is demonstrated through examples.","author":[{"dropping-particle":"","family":"Yang","given":"Li","non-dropping-particle":"","parse-names":false,"suffix":""}],"container-title":"Visual Data Mining","id":"ITEM-1","issued":{"date-parts":[["2008"]]},"page":"60-75","title":"Visual Exploration of Frequent Itemsets and Association Rules", "type":"chapter"},"suppress-author":1,"uris":["http://www.mendeley.com/documents/?uuid=be7b4eb5-06f3-3282-9eb7-387322abd37d"]}],"mendeley":{"formattedCitation":"(2008, donde se describió la importancia del enfoque de la exploración visual y las reglas de asociación mediante el uso del grafico de coordenadas paralelas, ya que es aquí donde se puede visualizar la correspondencia entre los conjunto de ítems. Por ello, hemos presentado la Figura 5 con el recorrido de cada etiqueta desde su inicio hasta el consecuente.

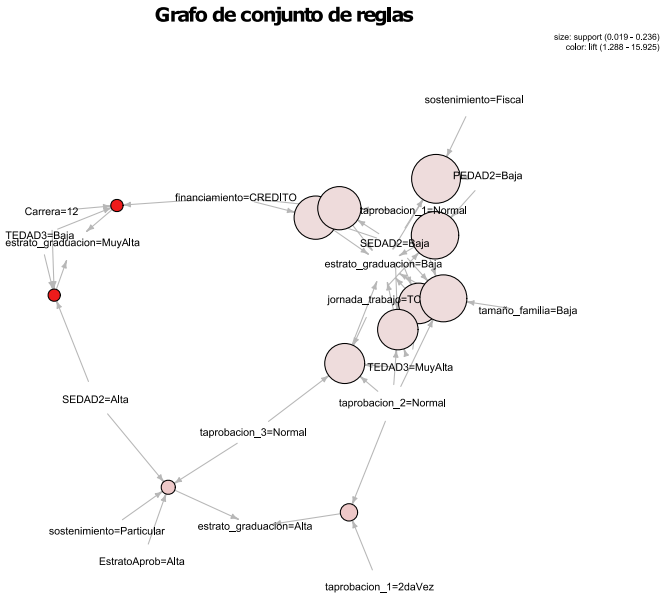

Figura 6. Grafo dirigido con la concurrecia entre etiquetas del conjunto de reglas asociación obtenidas. El color rojizo lo hemos asociado con la métrica lift y el tamaño de los nodos con la métrica support.

De acuerdo con los hallazgos encontrados, las etiquetas del consecuente indican una incidencia de graduación "Baja" que corresponden a factores donde el profesorado del tercer curso con edad superior a los 60 años tuvo una participación "Muy Alta". Además, los tiempos de

aprobación en primer y segundo curso fueron "Normal", es decir, que el estudiantado con bajo tiempo de graduación no suspendió asignaturas. Otras etiquetas menos frecuentes fueron jornada_trabajo=TC (TC=tiempo completo), financiamiento=CREDITO y tamaño_familia=Baja. Por otra parte, la incidencia de graduación "Alta" se debe a que tanto en primer y segundo curso las tasas de aprobación fueron "2daVez" o "3raVez", es decir, se evidenció que el estudiantado ha suspendido asignaturas y tuvo matriculas adicionales para aprobar un año; también, que el profesorado del tercer curso con edad superior a 60 años tuvo una baja participación en la impartición de clases.

Por último, la incidencia de graduación "Muy Alta" se relacionó con las categorías de edades del profesorado en segundo y tercer curso. Para el segundo curso el profesorado con edades entre 46 y 60 años tuvo una "Alta" participación en las clases; mientras que, para el tercer curso el profesorado con edad superior a 60 años tuvo una "Baja" participación en las clases. Por otro lado, se encontró otra etiqueta conexa con el tiempo de graduación "Muy Alta" y fue la etiqueta carrera= 12 que correspondió al estudiantado de la carrera de Agropecuaria.

CONCLUSIONES

El presente trabajo, centra su atención en el aprendizaje semi-supervisado usado para descubrir la escala de tiempo promedio de graduación del estudiantado. De manera más concreta, se ha investigado y descubierto las variables relevantes para la obtención del patrón de reglas de asociación. Por un lado, hemos encontrado variables relacionadas con edades del profesorado y el aprovechamiento del estudiante en los tres primeros años. Por otro lado, variables socioeconómicas tales como el financiamiento de estudios, tamaño de la familia y jornada de trabajo del estudiante. El aprendizaje semi-supervisado permitió conseguir reglas de asociación con alto nivel de confianza y además permitió desechar reglas espurias del análisis. También fue clave para lograr una mejor interpretación de las reglas usando gráficos.

Proponemos estudiar mediante otras técnicas de machine learning como: K-vecinos cercanos (KNN), Maquina de soporte vectorial (SVM) o Agrupamiento difuso (Clustering fuzzy), esto con el fin de detectar otros patrones que ayuden a los administradores académicos tener un nuevo conocimiento.

REFERENCIAS BIBLIOGRÁFICAS

Berzal, F., & Cubero, J. C. (2010). Interestingness Measures for Association Rules. Ipmu, 298–307. Recuperado de https://www.researchgate.net/profile/Fernando_Berzal/publication/221453256_Interestingness_Measures_for_Association_Rules_within_Groups/links/0deec52f764e881c87000000/Interestingness-Measures-for-Association-Rules-within-Groups.pdf

Breiman, L. (2001). Random Forests. ***Machine Learning***, ***45***(1), 5–32. Recuperado de https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. ***Journal of Artificial Intelligence Research***, ***16(1)***, 321–357. Recuperado de https://jair.org/index.php/jair/article/view/10302

García, S., Luengo, J., & Herrera, F. (2016). Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. ***Knowledge-Based Systems***, ***98***, 1–29. Recuperado de https://dl.acm.org/citation.cfm?id=2905525

González López, I. (2006). Dimensiones de evaluación de la calidad universitaria en el Espacio Europeo de Educación Superior. ***Revista Electrónica de Investigación Psicoeducativa***, 4(3), 445–468. Recuperado de http://lafacultadinvisible.com/wp-content/uploads/2015/03/González-López-2006.pdf

Han, W., et al. (2017). Interestingness Classification of Association Rules for Master Data. ***17th Industrial Conference on Data Mining***. New York.

Kumar, V., & Chadha, A. (2012). Mining association rules in student's assessment data. ***International Journal of Computer Science Issues***, ***9***(5), 211–216. Recuperado de http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.402.3422&rep=rep1&type=pdf

Liu, Y. (2012). AISC 159 - Applying Data Mining in Score Analysis. ***Advances in FCCS***, ***1***, 175–180.

Morales, P., et al. (2017). The NoiseFiltersR package. ***The R Journal***, ***9***(1), 1–8. Recuperado de https://journal.r-project.org/archive/2017/RJ-2017-027/index.html

Prajapati, D. J., Garg, S., & Chauhan, N. C. (2017). Interesting Association Rule Mining with Consistent and Inconsistent Rule Detection from Big Sales Data in Distributed Environment. ***Future Computing and Informatics Journal***, *2*(1), 19–30. Recuperado de https://www.researchgate.net/profile/Sanjay_Garg7/publication/316894904_Interesting_Association_Rule_Mining_with_Consistent_and_Inconsistent_Rule_Detection_from_Big_Sales_Data_in_Distributed_Environment/links/5949feb34585158b8fd5c5b3/Interesting-Association-Rule-Mining-with-Consistent-and-Inconsistent-Rule-Detection-from-Big-Sales-Data-in-Distributed-Environment.pdf

Romero, C., & Romero, J. (2010). Mining rare association rules from e-learning data. ***Educational Data Mining,*** 171–180. Recuperado de http://sci2s.ugr.es/keel/pdf/keel/congreso/RareAssociationRuleMining.pdf

Sawant, V., & Shah, K. (2016). Performance Evaluation of Distributed Association Rule Mining Algorithms. ***Procedia Computer Science***, *79*, 127–134. Recuperado de https://core.ac.uk/download/pdf/82536091.pdf

Varshali, J., & Jitendra, A. (2012). The Evolution of the Association Rules. ***International Journal of Modeling and Optimization***, *2*(6), 726–729. Recuperado de http://www.ijmo.org/papers/220-S20042.pdf

Yang, L. (2008). Visual Exploration of Frequent Itemsets and Association Rules. ***Visual Data Mining***, 60–75. Recuperado de https://link.springer.com/chapter/10.1007/978-3-540-71080-6_5