MINERÍA DE DATOS EDUCATIVOS: ANÁLISIS DEL DESEMPEÑO DE ESTUDIANTES DE INGENIERÍA EN LAS PRUEBAS SABER-PRO

Ana Isabel Oviedo Carrascal [1], Jovanny Jiménez Giraldo [2]

[1]Doctora en Ingeniería Electrónica – énfasis en Descubrimiento de Conocimiento, Profesora titular en la Facultad de Ingeniería en Tecnologías de la Información y la Comunicación, Grupo de Investigación GIDATIC, Universidad Pontificia Bolivariana, Medellín, Colombia. Correo electrónico: ana.oviedo@upb.edu.co.
[2] Ingeniero Electrónico, Estudiante de Maestría en Tecnologías de la Información y la Comunicación en Universidad Pontificia Bolivariana, Docente de tiempo completo en la Facultad de Ingeniería Electrónica y del Grupo de Investigación GIMU de la Universidad Católica de Oriente, Rionegro, Colombia. Correo electrónico: joijmenez@uco.edu.co

RESUMEN

En Colombia, las pruebas de Estado Saber-Pro han sido diseñadas para apoyar la evaluación y el mejoramiento de la educación superior en el país. Aplicando la metodología de minería de datos CRISP-DM, se realiza un estudio de los resultados obtenidos en las pruebas Saber-Pro de estudiantes de ingeniería en Antioquia (Colombia). A partir de 108 variables académicas, económicas y socio demográficas se realizan 3 modelos analíticos: 1) agrupación de los tipos de estudiantes, 2) selección de los factores que más influyen en el desempeño de las pruebas, y 3) predicción del desempeño en las pruebas a partir de las variables seleccionadas. Como resultado se encuentra que algunas de las variables más influyentes sobre el resultado de las pruebas son: el número de personas a cargo, método de enseñanza, si el hogar es permanente, el carácter académico de la institución y facilidades económicas como tener horno micro gas y motocicleta.

Palabras clave: Minería de datos educativos; analítica del aprendizaje, aprendizaje de máquinas.

Recibido: 18 de febrero de 2019. Aceptado: 25 de Mayo de 2019
Received: February 18, 2019. Accepted: May 25, 2019

EDUCATIONAL DATA MINING: ANALYSIS OF THE ENGINEERING STUDENTS PERFORMANCE IN SABER-PRO TEST

ABSTRACT

In Colombia, the Saber-Pro test has been created to support the evaluation and improvement of higher education in the country. This article, applies the CRISP-DM data mining methodology to perform a study of the results obtained in the Saber-Pro tests of engineering students in Antioquia (Colombia). Three analytical models are developed from 108 academic, economic and socio-demographic variables: 1) clustering about student types, 2) selection of the most influential factors in the results of the tests, and 3) prediction of performance in the tests from the selected factors. As a result, the most influential variables on the test result are: the number of dependents, teaching method, if the home is permanent, the academic character of the institution and economic facilities such as micro-gas oven and motorcycle.

Keywords: *Educational data mining; learning analytics; machine learning*

Cómo citar este artículo: A. Oviedo, J. jimenez. "Minería de datos educativos: análisis del desempeño de estudiantes de ingeniería en las pruebas SABER-PRO", Revista Politécnica, vol. 15, no.29 pp.128-140, 2019. DOI: 10.33571/rpolitec.v15n29a10

1. INTRODUCCIÓN

La Organización para la Cooperación y el Desarrollo Económico (OECD) publicó en el informe del año 2012 que Colombia ocupó los últimos lugares de las pruebas PISA [1]. Desde entonces se han generado políticas de gobierno como el programa "ser pilo paga" y "la atención a la primera infancia" para amortiguar las dificultades económicas y sociales de la población vulnerable. Posteriormente, en la evaluación del 2015 de la OECD, se afirma que esas políticas han tenido un impacto positivo en Colombia con una mejora significativa de los resultados de las pruebas PISA en el lapso del 2012 al 2014 [2].

Colombia aún no evalúa los procesos de enseñanza – aprendizaje en la educación superior por medio de un examen estandarizado internacionalmente, pero si lo hace localmente por medio de las pruebas Saber-Pro. En este artículo se presenta un estudio sobre las pruebas Saber-Pro y los diferentes factores que influyen en el desempeño de estudiantes de ingeniería en Antioquia por medio de minería de datos, la cual estudia técnicas que permiten la extracción de conocimiento a partir de fuentes masivas de datos.
Las técnicas usadas en minería de datos pueden ser clasificadas como supervisadas y no supervisadas [3]. Las técnicas supervisadas permiten realizar tareas de predicción de datos futuros (clasificación y regresión) mediante el estudio de un histórico de datos y las técnicas no supervisadas permiten describir los datos actuales (clustering, asociación y selección de factores).

El término *Educacional Data Mining* (EDM), representa la minería de datos aplicada a procesos educativos [4][5]. En el mundo se han realizado diversas investigaciones sobre sobre los resultados de las pruebas PISA, en [6] se demostró que existe relación entre los factores socio-económicos y el desempeño en lectura de las pruebas PISA. En [7] se presenta un análisis predictivo sobre los resultados de las pruebas PISA en España, donde se encontró que variables como disponibilidad de computador, género y estado de inmigración son importantes para los resultados en matemáticas. En general, se puede encontrar que las técnicas más comunes en EDM son árboles de decisión, redes bayesianas, regresión, correlación y análisis por clústeres [8][9][10][11], siendo el clustering la técnica más usada.

En Colombia se han realizado algunos estudios sobre las pruebas Saber-Pro, analizando el desempeño de algunos programas académicos como ingeniería civil y medicina. En [12] se estudian los factores de éxito en las pruebas Saber-Pro del año 2009 en las facultades de ingeniería civil en Colombia. De forma similar, [13] estudia las pruebas Saber-Pro del 2009 en la facultad de medicina. En la revisión bibliográfica, también se encuentran trabajos en Colombia donde se ha analizado el desempeño en módulos específicos, como lectura crítica e inglés. En [14] se presenta un estudio sobre factores que afectan el resultado del módulo de lectura crítica de las pruebas. En un trabajo similar, [15] presenta un estudio con datos sobre el ambiente social, demográfico y académico de estudiantes de programas profesionales en el módulo de inglés de Saber-Pro del año 2011. En estos trabajos se encontró que el género, la acreditación de la institución educativa, el nivel educativo de los padres y altos ingresos familiares son factores determinantes para el desempeño de las pruebas Saber Pro en los módulos y los programas académicos evaluados. Otro factor relevante se analiza en [16], donde se presenta una comparación entre el desempeño en las pruebas saber pro de programas presenciales y virtuales, encontrando que los presenciales obtienen mejores resultados.

Con el objetivo de aportar en el estudio de las pruebas Saber-Pro en Colombia, en este artículo se analizan variables académicas, económicas, socio demográficas y el resultado de todos los módulos de las pruebas Saber-Pro del año 2016 de estudiantes de ingeniería en Antioquia, por medio de tres modelos analíticos: 1) clustering para encontrar los tipos de estudiantes; 2) selección de los factores que más influyen en el desempeño de las pruebas analizando cada módulo por separado, ya que en la revisión literaria se encontró que los factores relevantes depende de cada módulo, para este análisis se aplica un método avanzado de minería de datos basado en un sistema de votación de varias técnicas; y 3) predicción del desempeño en las pruebas a partir de las variables seleccionadas. La organización del artículo es el siguiente. En la sección 2 se describe la metodología, los datos analizados y los modelos analíticos propuestos. En la sección 3 se presentan los resultados obtenidos en los modelos. En la sección 4 se presenta el análisis de resultados. Finalmente, en la sección 5 se presentan las conclusiones y trabajos futuros.

2. MATERIALES Y METODOS

En esta sección se describe la metodología del proceso de minería de datos, los datos analizados y los modelos analíticos propuestos para el análisis de las pruebas Saber-Pro de estudiantes de ingeniería en Antioquia del año 2016.

Metodología CRISP-DM

El proceso de minería de datos educativos presentado en este artículo es guiado por la metodología CRISP-DM, definida por 6 fases: entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelado, evaluación e implementación [17].

En la primera fase, llamada "Entendimiento del negocio", se analiza la información que suministra el Ministerio de Educación, estudiando e interpretando los resultados de las pruebas Saber-Pro. En la segunda fase, llamada "Entendimiento de los datos", se hace un análisis estadístico descriptivo para observar distribuciones que permitan realizar una descripción cuantitativa y cualitativa de los estudiantes que participaron en la prueba. En la tercera fase, llamada "Preparación de los datos", se perfilan las variables, se eliminan registros duplicados, se gestionan los valores nulos y se eliminan valores atípicos. En la cuarta fase, llamada "Modelado", se diseñan y aplican 3 modelos analíticos para el estudio de los datos correspondientes a un análisis de clustering, una selección de factores y una predicción. En la quinta fase, llamada "Evaluación", se mide el grado de confiabilidad y certeza de los modelos. Y la última fase, llamada "Implementación", se concluye sobre los resultados obtenidos con los modelos analíticos, validando los resultados obtenidos como apoyo a la toma de decisiones.

Datos Analizados

En este estudio se analizaron 49.021 registros correspondientes a estudiantes de ingeniería en Antioquia que presentaron las pruebas Saber-Pro en el año 2016. En total se tienen 108 variables con información económica, social y demográfica del estudiante, además de los resultados de los cinco módulos de las pruebas correspondientes a: 1) lectura crítica, 2) comunicación escrita, 3) inglés, 4) razonamiento cuantitativo, y 5) competencias ciudadanas. En la Tabla 1 se resumen los promedios de cada módulo para todos los estudiantes del país y para los estudiantes de ingeniería en Antioquia, la escala de calificación es de 0-300.

Tabla 1. Resultados pruebas Saber-Pro 2016

Módulo	Promedio Nacional	Promedio Est. Ingeniería en Antioquia
Lectura Crítica	150	154
Comunicación Escrita	150	149
Inglés	150	155
Razonamiento Cuantitativo	150	167
Competencias Ciudadanas	150	152

En los resultados que se entregan a los estudiantes, la calificación se categoriza en 4 niveles: nivel 1 no supera las preguntas de menor complejidad, nivel 2 supera las preguntas de menor complejidad, nivel 3 es el desempeño esperado, y nivel 4 cuando tiene un desempeño sobresaliente.

Los datos fueron inicialmente analizados por medio de la herramienta "DQAnalyzer" para evaluar el perfil de las variables y "Frill" para encontrar registros duplicados.

Modelos Analíticos

Con los datos descritos en la sección anterior, se realizaron 3 modelos analíticos: clustering, selección de factores y predicción de desempeño. Las técnicas empleadas fueron seleccionadas según la revisión bibliográfica sobre minería de datos educativos en el mundo. Adicionalmente, se propone un método avanzado de minería de datos mediante un sistema de votación para realizar la selección de factores.

1) Clustering: el objetivo de este modelo es encontrar los tipos de estudiantes por medio del algoritmo K-means, identificando los registros con comportamiento similar según la distancia euclídea entre los datos, de tal manera que los registros pertenecientes a un mismo cluster son aquellos que tienen una corta distancia entre ellos [3].

2) Selección de factores: el objetivo de este modelo es encontrar las variables que más influyen en el desempeño de cada módulo de las pruebas. La selección de factores se realiza con un sistema de votación de 4 métodos: correlaciones, análisis de componentes principales, árboles de decisión y reglas de asociación. En la Fig. 1 se ilustra el proceso de análisis propuesto para este trabajo, en el que se propone aplicar 4 técnicas cuyo resultado alimenta un sistema de votación.

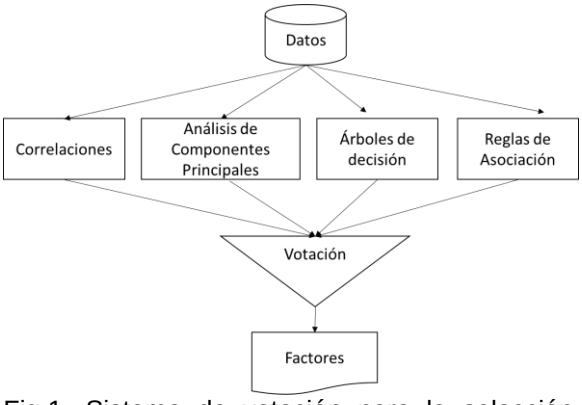


Fig.1. Sistema de votación para la selección de factores

Para aplicar los métodos del sistema de votación se asume como variable objetivo los resultados de cada uno de los módulos de las pruebas saber pro. Mediante el análisis de correlaciones se puede encontrar las variables que presentan una mayor relación con la variable objetivo, seleccionando así las más relevantes. Mediante el análisis de componentes principales (PCA) se identifican las variables con mayor coeficiente en los componentes encontrados por el método. Con los árboles de decisión, se seleccionan las variables que se encuentran en la cabecera del árbol como las más relevantes. Finalmente, con las reglas de asociación se seleccionan las variables que se encuentran en el precedente de las reglas como las más relevantes [3][18].

3) Predicción de desempeño: el objetivo de este modelo es anticipar el resultado en las pruebas por medio del algoritmo k-vecinos más cercanos (KNN) usando los factores elegidos en el modelo anterior. El algoritmo KNN, también conocido como método perezoso, busca los k registros más cercanos y asigna la respuesta de los registros vecinos [3].

3. RESULTADOS

Los experimentos se realizaron con la herramienta WEKA versión 3.8. ("Waikato Environment for Knowledge Analysis"), con la que se obtuvieron los siguientes resultados.

Clustering

Se ejecutó el método K-means con diferente cantidad de clústeres, encontrando el mejor desempeño con 6 grupos, donde la cohesión (distancias de los datos en cada clúster respecto a su centro) fue menor. Posteriormente se realizó el perfilamiento de los clústeres, explicando el centroide encontrado por el método K-means. En la Fig. 2 se presenta el tamaño de los grupos.

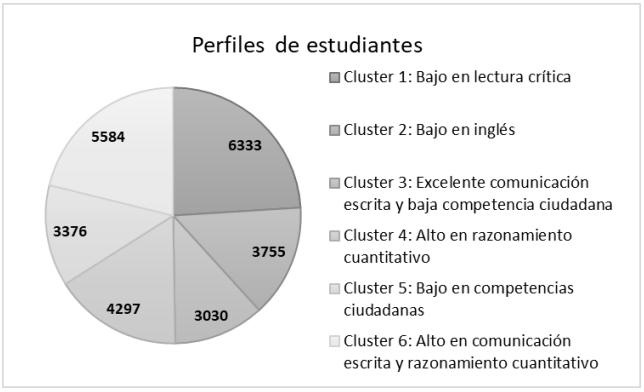


Fig.2. Cantidad de registros en cada clúster.

Clúster 1) Estudiantes con bajo desempeño en lectura crítica. Son en su mayoría hombres solteros de una buena condición socio económica, no son cabeza de familia, sin personas a cargo, viven en cabecera municipal, egresados de un bachiller académico, pagaron una matrícula a crédito entre 4 y 5.5 millones de pesos colombianos, sin beca, dedican a internet entre 1 y 3 horas, cuyos hogares poseen de 26 a 100 libros, cuyos padres no poseen una ocupación bien definida, estrato 3, con hogares de 4 personas y de 3 habitaciones, cuya madre posee formación tecnológica completa y el padre formación tecnológica incompleta, con automóvil, con internet y horno micro-gas. Los puntajes promedio obtenidos en los módulos de las pruebas Saber-Pro se presentan en la Tabla 2.

Tabla 2. Puntajes en los módulos de las pruebas Saber-Pro de los estudiantes del clúster 1.

Módulos	Puntaje promedio
Razonamiento cuantitativo	3
Lectura crítica	2
Competencia ciudadana	3
Inglés	B1
Comunicación escrita	3

Clúster 2) Estudiantes con bajo desempeño en Inglés. Son en su mayoría, mujeres que viven en unión libre en cabecera municipal, tienen un bachiller técnico y pagan una matrícula entre 1 y 2.5 millones, no usan crédito para pagarla, dedican menos de una hora al uso de internet y tienen en sus casas menos de 10 libros, poseen un cuarto en el hogar en el que viven dos personas, el padre y la madre trabajan por cuenta propia, la madre posee formación secundaria incompleta y el padre formación profesional incompleta, son de estrato dos, no tienen automóvil pero si motocicleta, no tienen televisión ni horno micro gas, no tienen personas a cargo, con una condición socio económica media – baja. Los puntajes promedio obtenidos en los módulos de las pruebas Saber-Pro se presentan en la tabla 3.

Tabla 3. Puntajes en los módulos de las pruebas Saber-Pro de los estudiantes del clúster 2.

Módulos	Puntaje promedio
Razonamiento cuantitativo	3
Lectura crítica	3
Competencia ciudadana	3
Inglés	A2
Comunicación escrita	3

Clúster 3) Estudiantes con excelente desempeño en comunicación escrita y bajo desempeño en competencia ciudadana. Son en su mayoría hombres con excelentes habilidades en comunicación escrita, buen desempeño en razonamiento cuantitativo y lectura crítica, pero un desempeño bajo en competencias ciudadanas, son casados, viven en cabecera municipal, graduados como bachiller académico, cuya matrícula esta entre 500 mil y un millón de pesos colombianos, no usan crédito para pagar su matrícula, le dedican a internet entre 1 y 3 horas, existen menos de diez libros, la madre era empleada doméstica y tiene primaria completa y su padre patrono o empleador y tiene primaria incompleta, son de estrato 4, tienen televisión y servicio de micro-gas, pero no tienen automóvil ni motocicleta. Son cabeza de familia, pero no tienen personas a cargo. Los puntajes promedio obtenidos en los módulos de las pruebas Saber-Pro se presentan en la Tabla 4.

Tabla 4. Puntajes en los módulos de las pruebas Saber-Pro de los estudiantes del clúster 3.

Módulo	Puntaje promedio
Razonamiento cuantitativo	3
Lectura crítica	2
Competencia ciudadana	1
Inglés	B1
Comunicación escrita	4

Clúster 4) Estudiantes con alto desempeño en razonamiento cuantitativo. Son en su mayoría mujeres, con desempeño alto en razonamiento cuantitativo y bajo desempeño en los otros módulos. Son solteras, viven en área rural, graduadas de bachiller académico, pagan entre un millón y 2.5 millones de pesos colombianos, tienen crédito y cuya deuda la asumen sus padres, le dedican a internet entre 1 y 3 horas y en sus casa existen entre 11 y 25 libros, la madre posee una ocupación no definida y el padre trabaja por cuenta propia, ambos poseen formación secundaria completa, son de estrato 1 con tres cuartos en sus hogares en la que habitan cinco personas, no son cabeza de familia, ni tienen personas a cargo, no tienen automóvil pero si motocicleta, así como servicio de televisión y micro gas, viven en un hogar temporal. Los puntajes promedio obtenidos en los módulos de las pruebas Saber-Pro se presentan en la Tabla 5.

Tabla 5. Puntajes en los módulos de las pruebas Saber-Pro de los estudiantes del clúster 4.

Módulo	Puntaje promedio
Razonamiento cuantitativo	3
Lectura crítica	2
Competencia ciudadana	2
Inglés	A1
Comunicación escrita	2

Clúster 5) Estudiantes con bajo desempeño en competencias ciudadanas. Son en su mayoría mujeres solteras, viven en cabecera municipal, graduadas de bachiller académico, pagan una matrícula entre 2.5 y 4 millones, la pagan con un

crédito, graduadas de bachillerato académico, le dedican menos de una hora a internet, en sus casas existen entre 11 y 25 libros. La madre es empleada doméstica con formación secundaria completa y el padre es obrero empleado de empresa particular, estrato 2 con 3 habitaciones en sus casas donde habitan 3 personas. Tienen televisión más no automóvil, motocicleta ni horno micro gas. Su hogar es habitual o permanente y no poseen personas a cargo. Los puntajes promedio obtenidos en los módulos de las pruebas Saber-Pro se presentan en la Tabla 6.

Tabla 6. Puntajes en los módulos de las pruebas Saber-Pro de los estudiantes del clúster 5.

Módulo	Puntaje promedio
Razonamiento cuantitativo	3
Lectura crítica	3
Competencia ciudadana	2
Inglés	B1
Comunicación escrita	3

Clúster 6) Estudiantes con buen desempeño en comunicación escrita y razonamiento cuantitativo. Son en su mayoría hombres solteros, viven en cabecera municipal, provienen de un bachiller a académico y pagan por valor de matrícula entre 2.5 y 4 millones, cuya matrícula se la pagan sus padres, sin endeudarse y sin beca. Dedican a internet entre 1 y 3 horas, tienen en sus casas entre 11 y 25 libros, la madre trabaja por cuenta propia y el padre es obrero o empleado de una empresa particular, ambos poseen formación secundaria completa, viven en estrato 3 con hogares que poseen 3 habitaciones, en el que habitan 4 personas. Tienen servicio de televisión y micro-gas pero no tienen automóvil ni motocicleta, no son cabeza de familia pero tienen una persona a cargo. Su hogar es habitual o permanente. Los puntajes promedio obtenidos en los módulos de las pruebas Saber-Pro se presentan en la Tabla 7.

Tabla 7. Puntajes en los módulos de las pruebas Saber-Pro de los estudiantes del clúster 6.

Módulo	Puntaje promedio
Razonamiento cuantitativo	3
Lectura crítica	2
Competencia ciudadana	2
Inglés	A2
Comunicación escrita	3

Selección de Factores

Los resultados del modelo de selección de factores por medio de un sistema de votación, se presentan organizados en dos subsecciones correspondientes a los resultados individuales de cada método y la votación final de los resultados.

1) Resultados Individuales de los Métodos
Los métodos aplicados son análisis de correlaciones, análisis de componentes principales (PCA), árboles de decisión y reglas de asociación. El análisis de correlaciones fue realizado con el método de Pearson, seleccionando las 20 variables más correlacionadas con cada módulo. En la Tabla 8 se presenta la variable con mayor correlación en cada módulo, se puede observar que la variable más común en la mayoría de módulos es el carácter académico de la institución.

Tabla 8. Variable con mayor correlación en cada módulo.

Módulo	Variable seleccionada con mayor correlación
Comunicación escrita	Carácter académico de la institución
Inglés	Personas a cargo
Competencias ciudadanas	Carácter académico de la institución
Lectura crítica	Carácter académico de la institución
Razonamiento cuantitativo	Carácter académico de la institución

El análisis de componentes principales (PCA) permite determinar las variables de las que más dependen linealmente los resultados de los módulos. Las variables con la dependencia más alta son los que más influyen en los resultados de cada una de las pruebas. En la Tabla 9 se presenta la variable con la dependencia más alta en cada módulo, se puede observar que la variable más común en la mayoría de módulos es el nivel de educación de la madre.

Tabla 9. Variable principal seleccionada con PCA en cada módulo.

Módulo	Variable seleccionada con PCA
Comunicación escrita	Nivel socio económico
Inglés	Educación de la madre
Competencias ciudadanas	Educación de la madre

Lectura crítica	Educación de la madre
Razonamiento cuantitativo	Educación de la madre

El método de árbol de decisión permite predecir el desempeño de los módulos de las pruebas Saber-Pro a partir de las variables académicas, económicas y socio demográficas. El algoritmo usado fue C5.0 donde las variables que se encuentran en la cabecera del árbol (ramas más altas) son aquellas que tienen más influencia en la predicción. Se realizó el mismo procedimiento, seleccionando las 20 variables más relevantes de cada módulo. En la Tabla 10 se presenta la variable con mayor relevancia en cada módulo, se puede observar que la variable más común en la mayoría de módulos es el carácter académico de la institución.

Tabla 10. Variable principal seleccionada con un árbol de decisión en cada módulo.

Módulo	Variable seleccionada con árbol de decisión
Comunicación escrita	Tener beca
Inglés	Nivel socio económico
Competencias ciudadanas	Carácter académico de la institución
Lectura crítica	Carácter académico de la institución
Razonamiento cuantitativo	Fecha de nacimiento

Finalmente, la creación de reglas de asociación por medio del algoritmo apriori permitió determinar aquellas variables que ocurren de forma conjunta repetidas veces en cada módulo. Se realizó el mismo procedimiento, seleccionando las 20 variables más relevantes de cada módulo. En la Tabla 11 se presenta la variable con mayor relevancia en cada módulo, se puede observar que la variable más común en la mayoría de módulos es el estado civil.

Tabla 11. Variable principal seleccionada con apriori en cada módulo.

Módulo	Variable seleccionada con a priori
Comunicación escrita	Género
Inglés	Estado civil
Competencias ciudadanas	Crédito para la matrícula
Lectura crítica	Estado civil
Razonamiento cuantitativo	Estado civil

2) Votación Final de los Métodos

En los experimentos anteriores se presentaron resultados parciales donde se evidencian algunas diferencias en las variables seleccionadas en cada módulo. A continuación, se procede a agrupar por módulo los resultados de cada uno de los métodos. En la Tabla 12 se presenta el sistema de votación para el módulo de competencias ciudadanas, se listan las variables con mayor votación.

Tabla 12. Sistema de votación para el módulo de competencias ciudadanas. Abreviaciones: A priori (Ap), Análisis de Componentes Principales (PCA), Correlación (Cor), Árboles de decisión (Arb) y Frecuencia (Frec).

Variable	Ap	PCA	Cor	Arb	Frec
Tiene internet	x	x	x		3
Hogar actual	x	x		x	3
Núm. de personas a cargo	x	x	x	x	4
Metodología del programa	x		x	x	3
Carácter académico de la institución	x		x	x	3
Tiene moto		x	x	x	3
Estrato vivienda		x	x	x	3
Tiene horno micro-gas		x	x	x	3

De forma similar, se seleccionan las variables más relevantes de cada módulo con un sistema de votación, seleccionando las variables con al menos 2 votos. Para el módulo de competencias ciudadanas, las variables más relevantes según el sistema de votación son: crédito para la matrícula, estado civil, tiene lavadora, tiene computador, tiene internet, hogar actual, número de personas a cargo, metodología del programa académico, carácter académico de la institución, cabeza familia, tiene motocicleta, cantidad de cuartos del hogar, estrato vivienda, tiene horno micro-gas, tiene automóvil, numero libros, valor matricula, pago matricula recursos propios, índice socio económico, padres pagan matricula y fecha de nacimiento.

Para el módulo de comunicación escrita, las variables más relevantes según el sistema de

votación son: género, estado civil, dedicación a internet, crédito para la matrícula, tiene computador, tiene internet, hogar actual, número de personas a cargo, carácter académico de la institución, cabeza familia, tiene motocicleta, estrato vivienda, tiene horno micro-gas, número libros y el índice socio económico.

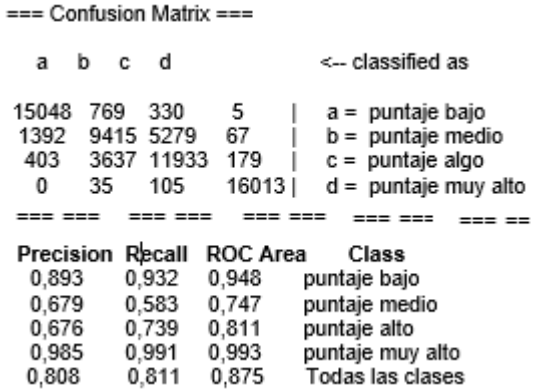
Para el módulo de inglés, las variables más relevantes según el sistema de votación son: tiene lavadora, tiene computador, tiene internet, hogar actual, número de personas a cargo, metodología del programa académico, carácter académico de la institución, cabeza familia, tiene servicio tv, tiene motocicleta, estrato vivienda, tiene horno micro-gas, valor matricula universidad, pago matricula con recursos propios y el índice socio económico.

Para el módulo de lectura crítica, las variables más relevantes según el sistema de votación son: crédito para la matrícula, estado civil, tiene lavadora, tiene computador, tiene internet, hogar actual, número de personas a cargo, metodología del programa académico, cabeza familia, tiene servicio tv, tiene motocicleta, estrato vivienda, número de personas que viven en el hogar, tiene horno micro-gas, tiene automóvil, número libros, valor matricula de la universidad, índice socio económico, fecha nacimiento.

Finalmente, para el módulo de razonamiento cuantitativo, las variables más relevantes según el sistema de votación son: género, crédito para la matrícula, estado civil, beca para la matrícula, tiene lavadora, tiene computador, tiene internet, hogar actual, número de personas a cargo, metodología del programa académico, carácter académico de la institución, cabeza familia, tiene motocicleta, estrato vivienda, tiene horno micro-gas, tiene automóvil, número libros, valor matricula de la universidad, índice socio económico, padres pagan matrícula y fecha de nacimiento.

Predicción de Desempeño

En el experimento anterior se seleccionaron las variables más relevantes para cada módulo. Para realizar la predicción de desempeño se analizarán el conjunto de todas las variables seleccionadas en el experimento anterior. Las variables utilizadas en la predicción son: índice socio económico, tiene computador, tiene internet, hogar actual, número de personas a cargo, carácter académico de la institución, cabeza familia, tiene motocicleta, estrato vivienda, tiene horno micro-gas, crédito para matrícula, estado civil, tiene lavadora, metodología del programa académico, número libros, valor matricula de la universidad, tiene automóvil, fecha nacimiento, si pagó la matrícula con recursos propios, si pagó la matrícula con recursos de los padres, género, tiene servicio de TV, cantidad de cuartos en el hogar, dedicación a internet, cantidad de personas en el hogar y si tenía beca. La variable que se desea predecir (variable objetivo) corresponde al resultado promedio de los cinco módulos, para ello fue necesario crear una nueva variable “PROMEDIO”, con base en el puntaje obtenido en los módulos (puntaje entre cero y 300 para todos los casos). La variable creada fue discretizada en cuatro categorías según los percentiles: bajo desempeño (0 - 121.25), puntaje medio (121.26- 165.7), puntaje alto (165.8 - 210.15) y la cuarta categoría que corresponde a los puntajes más altos (210.16 – 300). Se aplicó un balanceo de los datos para que cada categoría posea aproximadamente la misma cantidad de muestras. Se aplicó entonces un método de predicción perezoso con el algoritmo k-vecinos más cercanos (Knn) y una validación cruzada de 10 grupos, obteniendo los resultados de la Fig. 3.



=== Confusion Matrix ===

a	b	c	d		<-- classified as
15048	769	330	5	|	a = puntaje bajo
1392	9415	5279	67	|	b = puntaje medio
403	3637	11933	179	|	c = puntaje algo
0	35	105	16013	|	d = puntaje muy alto

Precision	Recall	ROC Area	Class
0,893	0,932	0,948	puntaje bajo
0,679	0,583	0,747	puntaje medio
0,676	0,739	0,811	puntaje alto
0,985	0,991	0,993	puntaje muy alto
0,808	0,811	0,875	Todas las clases

Fig.3.Evaluación del método de predicción Knn.

A partir del área ROC puede identificarse que las mejores predicciones ocurren en los puntajes más bajos y más altos, con valores de 0.948 y 0.993 respectivamente. Respecto a la medida *recall* (cobertura), la categoría en la que más se dificulta el aprendizaje es con los estudiantes de puntaje medio. El porcentaje de instancias correctamente clasificadas es del 81.12% y el área ROC superior a 0.87 indican que el modelo es confiable y que las variables seleccionadas representan de una

manera significativa las condiciones académicas, sociales y económicas de los estudiantes.

4. DISCUSIÓN

Teniendo en cuenta los resultados presentados en la sección anterior, se presenta un análisis de los 3 experimentos analíticos realizados.

En el análisis de clustering se encontraron 6 tipos de estudiantes. Los dos primeros grupos (clúster 1 y 2) corresponden a estudiantes que obtienen buenos resultados en razonamiento cuantitativo, competencias ciudadanas y comunicación escrita; sin embargo, poseen grandes diferencias en el resultado de inglés y lectura crítica. Al comparar ambos grupos se encuentra una diferencia alta en la condición económica de ambos, de tal manera que una condición económica alta corresponde a un buen desempeño de inglés, y una baja a un buen resultado de competencia lectora (pero no en inglés). El clúster 3 de estudiantes con excelente desempeño en comunicación escrita, tiene un desempeño muy bajo en competencias ciudadanas. Los clústeres 4 y 6 son los de más bajo rendimiento y tienen en común varias características: dedican a internet entre 1 y 3 horas, tienen en sus casas entre 11 y 25 libros, la madre y el padre poseen formación secundaria.

En la selección de factores se logró reducir 108 variables iniciales a 26 variables. Para cada módulo se seleccionaron las variables más relevantes por medio de un sistema de votación, encontrando que:

- El factor género fue determinante en Razonamiento cuantitativo y Comunicación escrita. Sin embargo, su participación fue prácticamente nula en el resto de módulos.
- El número de libros en casa es determinante para la comunicación escrita.
- La condición económica es relevante para el módulo de inglés.
- Es un hecho que tener personas a cargo o ser cabeza de familia se constituye en un impedimento para obtener buenos resultados en las pruebas Saber-Pro.
- Generalmente, se obtienen buenos resultados cuando la metodología de enseñanza es presencial, este factor también ha sido identificado en otros estudios.
- Los resultados indican que las instituciones de educación superior cuyo carácter académico es UNIVERSIDAD, ofrecen los estudiantes con mejor resultados. Esto excluye a las instituciones con un carácter académico tal como: ESCUELA NORMAL SUPERIOR, INSTITUCIÓN TECNOLÓGICA o INSTITUCIÓN UNIVERSITARIA, entre otras. Es posible que el carácter de Universidad acerque a los estudiantes a los problemas de la sociedad por medio de la investigación y la extensión.

En la predicción de desempeño, con las 26 variables seleccionadas se creó un modelo analítico para predecir el desempeño general en las pruebas Saber-Pro. Dicho modelo alcanzó mejores resultados en las clases en las que el desempeño del estudiante era o muy malo o muy bueno, lo cual significa que las variables seleccionadas explican con fidelidad estos dos casos extremos.

5. CONCLUSIONES

En este trabajo se presentó un estudio sobre los factores económicos, sociales y demográficos que influyen en el desempeño de las pruebas Saber Pro en estudiantes de ingeniería en Antioquia. Mediante la revisión bibliográfica se identificaron las técnicas analíticas más usadas en la minería de datos educativos en el mundo, resaltando el clustering entre ellas. Sobre los estudios realizados en Colombia para analizar las pruebas saber-pro, se resalta que han sido aplicados para los módulos de inglés y lectura crítica encontrando la influencia de variables diferentes. Los programas analizados han sido ingeniería química y medicina. En este trabajo se realiza un estudio completo de todos los módulos para estudiantes de ingeniería en Antioquia, aplicando diferentes modelos analíticos como clustering, selección de factores y predicción. En especial se resalta el modelo de selección de factores, en el cual se aplica un sistema de votación de varios métodos, este tipo de modelos recibe el nombre de ensambles y constituyen modelos avanzados de minería de datos.

En el experimento del Clustering usando K-means se observó una separación clara entre los estudiantes que obtienen buenos resultados en inglés, quienes generalmente poseen una buena condición económica. Los estudiantes que obtienen buen resultado en lectura crítica generalmente son de estratos bajos y pagan matriculas de bajo valor. En el experimento de selección de factores se encontró que las variables más relevantes son: el

número de personas a cargo, método de enseñanza, si el hogar es permanente, el carácter académico de la institución y facilidades económicas como tener horno micro gas y motocicleta. Finalmente, a partir de 26 variables seleccionadas en el experimento anterior, se logró predecir el desempeño en las pruebas con una exactitud del 81%.

El paso siguiente de este trabajo es identificar si las estrategias que intentan mitigar las causas del bajo rendimiento escolar que implementa el Ministerio de Educación, han tenido el efecto deseado en la población estudiantil. Para ello se propone como trabajo futuro realizar un estudio histórico con base a los datos de años anteriores para observar la evolución de los resultados.

6. AGRADECIMIENTOS

Los autores expresan su agradecimiento a los grupos de investigación GIDATIC de la Universidad Pontificia Bolivariana y GIMU de la Universidad Católica de Oriente. Este documento es resultado de la tesis de Maestría en TIC de la Universidad Pontifica Bolivariana titulada “Minería de datos educativos: análisis de los factores económicos, sociales y demográficos que influyen en el desempeño de las pruebas saber-pro en estudiantes de ingeniería en Antioquia”.

7. REFERENCIAS BIBLIOGRÁFICAS

[1] OCDE. Resultados de las pruebas PISA 2012 en foco. Organización para la Cooperación y el Desarrollo Económico, 2012. Obtenido de https://www.oecd.org/pisa/keyfindings/PISA2012_Overview_ESP-FINAL.pdf

[2] OCDE. PISA 2015, Results in focus. Organización para la Cooperación y el Desarrollo Económico, 2015. Obtenido de https://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf

[3] Oviedo, A. Velez, G., y Oviedo, E. Minería de datos: aportes y tendencias en el servicio de salud de ciudades inteligentes. Revista Politecnica, 11, 111-120, 2015. DOI: https://doi.org/10.22395/rium.v16n31a6

[4] Chadha, A. Efficient Clustering Algorithms in Educational Data Mining. En: Handbook of Research on Knowledge Management for Contemporary Business Environments, IGI Global, 279-312, 2018.

[5] Oviedo, A., y Jiménez, G. Estudio sobre Estilos de Aprendizaje mediante Minería de Datos como apoyo a la Gestión Académica en Instituciones Educativas. RISTI-Revista Ibérica de Sistemas e Tecnologias de Informação, (29), 1-13, 2018.

[6] Jehangir, K., Glas, C. A. W., y van den Berg, S. Exploring the relation between socio-economic status and reading achievement in PISA 2009 through an intercepts-and-slopes-as-outcomes paradigm. International Journal of Educational Research, 71, 1–15, 2015.

[7] Gorostiaga, A., y Rojo-Álvarez, J. L. On the use of conventional and statistical-learning techniques for the analysis of PISA results in Spain. Neurocomputing, 171, 625–637, 2016.

[8] Ganesh, S. H., y Christy, A. J. Applications of educational data mining: a survey. Documento presentado en 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 1-6, 2015.

[9] Peña-Ayala, A. Educational data mining: A survey and a data mining-based analysis of recent works. Expert Systems with Applications, 41(4 PART 1), 1432–1462, 2014.

[10] Kaur, R., & Singh, S. (2016). A survey of data mining and social network analysis based anomaly detection techniques. Egyptian Informatics Journal, 17(2), 199–216, 2016.

[11] Mohamad, S. K., y Tasir, Z. Educational Data Mining: A Review. Procedia - Social and Behavioral Sciences, 97, 320–324, 2013. http://doi.org/10.1016/j.sbspro.2013.10.240

[12] Cantillo, V., y García, L. Gender and Other Factors Influencing the Outcome of a Test to Assess Quality of Education in Civil Engineering in Colombia. Journal of Professional Issues in Engineering Education Practise, 1–7, 2014.

[13] Gil, F. A., Rodríguez, V. A., Sepúlveda, L. A., Rondón, M. A., y Gómez-Restrepo, C. Impacto de las facultades de medicina y de los estudiantes

sobre los resultados en la prueba nacional SABER PRO). Revista Colombiana de Anestesiología, 41(3), 196–204, 2013,

[14] Timarán, R., Hidalgo, A, Caicedo, J., Hernández, I. y Alvarado, J. Descubrimiento de Patrones de Desempeño Académico en la Competencia de Lectura Crítica. Documento presentado en 13th LACCEI Annual International Conference: "Engineering Education Facing the Grand Challenges, What Are We Doing?", Santo Domingo, Dominican Republic, julio 29-31, 2015.

[15] Timarán, R., Hidalgo, A. y Caicedo J. Proceso de Descubrimiento de Patrones de Desempeño Académico en la Competencia de Inglés con CRISP-DM. Documento presentado en Décima Quinta Conferencia Iberoamericana en Sistemas, Cibernética e Informática: CISCI 2016,: Orlando-Florida-USA, Volume: I, 2016.

[16] Rodríguez, G., Gómez, V., y Ariza , M. Calidad de la educación superior a distancia y virtual: un análisis de desempeño académico en Colombia. Investigación Y Desarrollo, 22(1), 79–119, 2014.

[17] Sharma, S., Osei-Bryson, K.-M., y Kasper, G. M. Evaluation of an integrated Knowledge Discovery and Data Mining process model. Expert Systems with Applications, 39(13), 11335–11348, 2012. http://doi.org/10.1016/j.eswa.2012.02.044

[18] Oviedo, A. I., y Sánchez, S. Minería de datos de la salud: Sistema de votación de técnicas analíticas para identificar los factores que influyen en la realización de cirugías estéticas. Revista Politécnica, 13(25), 43-52, 2017.