ISSN: 1692-7257 - Volumen 2 – Número 38 - 2021

Revista Colombiana de Tecnologías de Avanzada

Recibido: 8 de agosto de 2021
Aceptado: 6 de septiembre de 2021

EDUCATIONAL DATA MINING TO DISCOVER PATTERNS ASSOCIATED WITH ACADEMIC PERFORMANCE IN GENERIC SKILLS

MINERÍA DE DATOS EDUCATIVA PARA DESCUBRIR PATRONES ASOCIADOS AL DESEMPEÑO ACADÉMICO EN COMPETENCIAS GENÉRICAS

Ing. Andrea Timarán Buchely[*], PhD. Ricardo Timarán Pereira[**]

[*] **Pontificia Universidad Javeriana,** Facultad de Ingeniería, Programa de Ingeniería de Sistemas y Computación.
Calle 18, Cali, Valle del Cauca, Colombia.
(57) (2) 3218200.
andreaestefania12@javerianacali.edu.co.
[**] **Universidad de Nariño,** Facultad de Ingeniería, Departamento de Sistemas.
Calle 18 Carrera 50, Pasto, Nariño, Colombia.
(57) (27)244309.
ritimar@udenar.edu.co.

Resumen: El objetivo de este estudio fue descubrir patrones asociados al desempeño académico en las competencias genéricas del examen estatal Saber Pro que presentaron los estudiantes de la Universidad Javeriana Cali (Colombia) en los años 2017 y 2018, aplicando técnicas de minería de datos educativa. Se utilizó la metodología CRISP-DM. Se seleccionó, de las bases de datos del ICFES, la información socioeconómica, académica e institucional de estos estudiantes. Se construyó, limpió y transformó un repositorio de datos para la minería de datos. Se descubrieron patrones asociados al buen o mal desempeño académico de los estudiantes con respecto al puntaje global obtenido en las competencias genéricas del examen Saber Pro. Se utilizó un modelo de clasificación basado en árboles de decisión. El conocimiento generado permitirá soportar la toma de decisiones de las directivas universitarias, con el fin de mejorar la calidad de la educación en la Universidad Javeriana.

Palabras clave: Minería de Datos Educativa, Árboles de Decisión, Desempeño Académico, Competencias Genéricas, Pruebas Saber Pro.

Abstract: The goal of this study was to discover patterns associated with academic performance in the generic skills of the Saber Pro state exam presented by students of the Universidad Javeriana Cali (Colombia) in the years 2017 and 2018, applying educational data mining techniques. The CRISP-DM methodology was used. The socioeconomic, academic and institutional information of these students was selected from the ICFES databases. Built, cleaned, and transformed a data repository for data mining. Patterns associated with the good or poor academic performance of the students were discovered with respect to the overall score obtained in the generic skills of the Saber Pro exam. A classification model based on decision trees was used. The knowledge generated will support the decision-making of university directors, in order to improve the quality of education at the Javeriana University.

Keywords: Educational Data Mining, Decision Trees, Academic Performance, Generic Skills, SaberPro tests.

1. INTRODUCCIÓN

En toda institución de educación superior-IES, el desempeño académico de los estudiantes es el factor que determina la calidad de educación de dichas instituciones, es por eso, que el desempeño ha adquirido un gran interés, ya que permite mejorar el nivel de educación superior en las mismas. La calidad de la educación superior supone el esfuerzo continuo de las instituciones para cumplir en forma responsable con las exigencias propias de cada una de sus funciones. Estas funciones que, en última instancia pueden reducirse a docencia, investigación e interacción social, reciben diferentes énfasis en una institución u otra, dando lugar a distintos estilos de institución (icfes,2017).

La calidad de la educación superior es una prioridad. Ofrecerla es un deber de las IES. Para lograrlo, el Ministerio de Educación Nacional-MEN y el Instituto Colombiano para la Evaluación de la Educación - ICFES definieron tres programas entrelazados: Estándares Mínimos de Calidad - EMC- para pregrado y posgrado, incentivos a la acreditación de excelencia, y exámenes de calidad (icfes,2012).

En Colombia, el Decreto 3963 de 2009, establece que las pruebas Saber Pro son instrumentos estandarizados para evaluación externa de la calidad de la educación superior; forman parte, con otros procesos y acciones de un conjunto de instrumentos que el Estado dispone para evaluar la calidad del servicio público educativo y ejercer su inspección y vigilancia (MEN, 2006).

El examen está compuesto por pruebas que evalúan competencias genéricas y específicas. De acuerdo a los lineamientos Saber Pro del Instituto Colombiano para la Evaluación de la Educación (ICFES), todos los estudiantes deben presentar los módulos de competencias genéricas sin importar el programa de formación que cursen, que incluye competencias de lectura crítica, razonamiento cuantitativo, escritura, inglés y competencias ciudadanas (icfes,2017).

En la competencia de lectura crítica se evalúan los desempeños asociados a lectura, pensamiento crítico y entendimiento interpersonal (icfes,2017). En la competencia de razonamiento cuantitativo se evalúan los desempeños relacionados con uso de lenguaje cuantitativo y solución de problemas (icfes,2017). En escritura se evalúa la competencia para comunicar ideas por escrito referidas a un tema dado (MEN,2006; icfes,2017). En inglés se evalúa la competencia del estudiante para comunicarse efectivamente en inglés y finalmente en competencias ciudadanas se evalúa los conocimientos y habilidades que posibilitan la construcción de marcos de comprensión del entorno, los cuales promueven el ejercicio de la ciudadanía y la coexistencia inclusiva según la Constitución Política Colombiana.

A pesar que en la prueba Saber Pro no se pretende que los estudiantes de todas las formaciones desarrollen las competencias genéricas a un mismo nivel, ni aún las comunes a grupos de programas, sí es importante determinar cómo influyen los factores socioeconómicos, académicos e institucionales del estudiante para obtener un determinado nivel de desempeño de estas competencias en las pruebas Saber Pro. Los estudios que se han realizado hasta el momento (MEN,2006; icfes,2012; Posada and Mendoza,2014) con respecto al análisis de los resultados de las pruebas Saber Pro se basan en información procesada mediante un análisis estadístico, donde fundamentalmente se consideran variables y relaciones primarias, sin tener en cuenta las verdaderas interrelaciones, que por lo general están ocultas y que únicamente se pueden descubrir utilizando un tratamiento más complejo de los datos, que es posible con la minería de datos.

Los resultados de pruebas nacionales e internacionales muestran que Colombia posee un sistema educativo con bajos logros académicos de sus estudiantes, en cada uno de los niveles de estudio (Posada and Mendoza,2014). Esta situación es crítica, pues de continuar persistiendo esos rendimientos académicos en la mayor parte del estudiantado colombiano, los rendimientos asociados a las economías de escala entre el capital físico y el capital humano seguirán llevando al país por una senda de desarrollo restringido y bajo crecimiento económico.

La minería de datos en la educación EDM (del término en inglés Educational Data Mining) no es un tema nuevo, su estudio y aplicación ha sido muy relevante en los últimos años, se puede utilizar sus técnicas para explicar y/o predecir cualquier fenómeno dentro del campo educativo (Valero *et al*.,2005; Baker,2010; Peña, 2014; Timarán *et al.*,2016; Algarni,2016; Timarán *et al.*,2017). El

EDM se define como el área de la investigación científica centrada en el desarrollo de métodos para realizar descubrimientos dentro de los tipos únicos de datos que provienen de entornos educativos, y usar esos métodos para comprender mejor a los estudiantes y el contexto en el que aprenden (Baker,2010). Además, EDM extrae información interesante, interpretable, útil y novedosa de datos educativos. El EDM es útil en muchas áreas diferentes, como la identificación de estudiantes con alto riesgo académico, las necesidades de aprendizaje prioritarias para diferentes grupos de estudiantes, el aumento de los índices de graduación, la evaluación del desempeño institucional, la maximización de los recursos del campus y la optimización de la renovación del currículo de la asignatura (Algarni,2016). Usando técnicas de extracción de datos, por ejemplo, se puede predecir, con un porcentaje muy alto de confiabilidad, la probabilidad de deserción de cualquier estudiante (Valero, 2009). Las instituciones de educación pueden usar la minería de datos para hacer análisis comprensivos de las características de sus estudiantes, métodos evaluativos, develando procesos exitosos o, por el contrario, detectando fraudes o inconsistencias (Valero *et al*., 2005).

En este artículo se presentan los resultados de aplicar la minería de datos educativa para descubrir factores asociados al desempeño académico, en el puntaje general de las competencias genéricas que obtuvieron los estudiantes de la Universidad Javeriana Cali que presentaron las pruebas Saber Pro en los años 2017 y 2018.

2. MATERIALES Y MÉTODOS

La investigación fue de tipo descriptivo bajo el enfoque cuantitativo, aplicando un diseño no experimental. Como fuentes de información se utilizaron los datos que se encontraban disponibles, al momento de la investigación, en las bases de datos del ICFES de los resultados de los estudiantes que presentaron las pruebas Saber Pro. Los datos más actualizados eran de los años 2017 y 2018. Para el descubrimiento de patrones asociados al desempeño académico en las pruebas Saber Pro, se construyó un modelo de clasificación basado en árboles de decisión, utilizando el algoritmo J48 de la herramienta Weka (Witten, 2011). Se escogió este algoritmo por su simplicidad y facilidad para interpretar los patrones y por ser el más utilizado para este tipo de problemas (Hand and Kamber, 2001; Sattler and Dunemann,2001).

Para el descubrimiento de patrones, se aplicó la metodología CRISP-DM (Cross Industry Standard Process for Data Mining). Azevedo y Santos (2008) comparan las metodologías de minería de datos CRISP-DM y SEMMA (Sample, Explore, Modify, Model, and Assess) y llegan a la conclusión de que, aunque se puede establecer un paralelismo claro entre ellas, CRISP-DM es más completo porque tiene en cuenta la aplicación al entorno de negocio de los resultados, y por ello es la que se adoptó popularmente. En encuestas realizadas en KDNuggets en 2002, 2004, 2007 y 2014 se comprobó que CRISP-DM era la principal metodología utilizada, cuatro veces más que SEMMA. La metodología CRISP-DM para proyectos de minería de datos no es la "más actual" o "la mejor", pero es muy útil para comprender esta tecnología o extraer ideas para diseñar o revisar métodos de trabajo para proyectos de similares características (Azevedo and Santos,2008). CRISP-DM es la guía de referencia más ampliamente utilizada en el desarrollo de proyectos de minería de datos (Hernández and Ramírez, 2005) y contempla seis fases: comprensión del problema, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación.

En la fase de comprensión del problema se identificó con exactitud la problemática que se solucionaría utilizando la minería de datos, esto permitió recolectar la información necesaria para interpretar con asertividad los resultados encontrados (Villena,2016). En la fase de comprensión de los datos se realizó la recolección inicial de datos, para establecer un primer contacto con el problema, familiarizarse con ellos, identificando su calidad y establecer las relaciones más evidentes que permitieron definir las primeras hipótesis. En la fase de preparación de los datos se seleccionó los datos a los cuales se les aplicaría una determinada técnica de modelado, limpieza, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato (Villena,2016). En la fase de modelado se seleccionaron las técnicas de minería de datos más apropiadas para el proyecto. En la fase de evaluación se verificó si el modelo se ajusta a las necesidades establecidas en el proyecto. Se evaluaron los patrones encontrados con el fin de determinar su validez, remover los redundantes o irrelevantes y traducir los patrones útiles en términos que sean entendibles para el usuario. Finalmente, en la fase de implementación, se trató de explotar la potencialidad de los modelos, integrarlos en los procesos de toma de decisión del MEN, ICFES y de otras instituciones

gubernamentales y educativas que velan por la calidad de la educación en Colombia y difundir informes sobre el conocimiento extraído (Villena,2016) para su evaluación y publicación.

3. RESULTADOS Y DISCUSIÓN

3.1 Comprensión del Problema

En esta fase, se realizaron las actividades que permitieron profundizar y apropiar de una manera completa el problema objeto de estudio, los objetivos y los requisitos de esta investigación, que posibilitaron la recolección de los datos correctos para interpretar adecuadamente los resultados. En esta fase, descubrir factores asociados al desempeño académico, de los estudiantes de la Universidad Javeriana Cali, que presentaron las pruebas Saber Pro, se convirtió en un problema a resolver con minería de datos.

3.2 Comprensión de los datos

En esta fase, se identificó, recopiló y familiarizó con la información socioeconómica, académica e institucional, disponible, al momento de realizar esta investigación, en las bases de datos en el ICFES, correspondiente a los resultados de las competencias genéricas de los estudiantes de la Universidad Javeriana que presentaron las pruebas Saber Pro entre los años 2017 y 2018. Los repositorios de cada año se integraron en uno solo y como resultado se obtuvo un conjunto de datos inicial al cual se le denominó sbpro2052A101, con 2052 registros y 101 atributos.

3.3 Preparación de los datos

Teniendo en cuenta que la alta dimensionalidad es un problema para el descubrimiento de patrones con minería de datos (Hernández *et al*.,2005), al conjunto sbpro2052A101 se le aplicaron técnicas de limpieza y transformación con el fin de eliminar los datos ruidosos, nulos, atípicos, transformar algunos atributos para que obtengan mayor ganancia de información y eliminar aquellos atributos irrelevantes que no aportaban al proceso de detección de patrones, dando como resultado el conjunto de datos denominado sbpro2052A32 compuesto por 2052 registros y 32 atributos, el cual sirvió de base para la fase de modelado.

3.4 Modelado

En la fase del modelado se seleccionó el modelo de clasificación con árboles de decisión como la técnica de minería de datos más adecuada para solucionar el problema objeto de la investigación. Este modelo, es probablemente el más utilizado y popular por su simplicidad y facilidad para entender (Han and Kamber,2001; Sattler and Dunemann, 2001; Timarán et al, 2014). La importancia de los árboles de decisión se debe a su capacidad de construir modelos interpretables, siendo este un factor decisivo para su aplicación.

La clasificación con árboles de decisión considera clases disjuntas, de forma que el árbol conducirá a una y sólo una hoja, asignando una única clase a la predicción (Hernández and Lorente, 2009). Esta técnica presenta varias ventajas. Primero, el proceso de razonamiento detrás del modelo resulta claramente evidente cuando se examina el árbol. Esto contrasta con otras técnicas de modelado de caja negra, en las que la lógica interna puede resultar difícil de averiguar. En segundo lugar, el proceso incluye automáticamente en su regla únicamente los atributos que realmente importan en la toma decisiones. Los atributos que no contribuyan a la precisión del árbol se omiten (Sattler and Dunemann, 2001).

Antes de construir el modelo, se definió el procedimiento para probar la calidad del modelo y su validez. Teniendo en cuenta que, para entrenar y probar un modelo de clasificación, se divide los datos en dos conjuntos: entrenamiento y prueba (Witten *et al*.,2011), se utilizó el método de validación cruzada (Cross validation) porque permite reducir la dependencia del resultado del experimento en el modo en el cual se realiza la partición (Hernández *et al.*,2005). Para este caso particular se utilizó el método de evaluación validación cruzada con n pliegues (n-fold cross validation). Este método consiste en dividir el conjunto de entrenamiento en n subconjuntos disjuntos de similar tamaño llamados pliegues (folds) de forma aleatoria. El número de subconjuntos se puede introducir en el campo Folds. Posteriormente se realizan n iteraciones (igual al número de subconjuntos definido), donde en cada una se reserva un subconjunto diferente para el conjunto de prueba y los restantes n-1 (uniendo todos los datos) para construir el modelo (entrenamiento). En cada iteración se calcula el error de muestra parcial del modelo. Por último, se construye el modelo con todos los datos y se obtiene su error promediando los obtenidos anteriormente en cada una de las iteraciones. Otra ventaja de la validación cruzada es que la varianza de los n errores de muestra parciales, permite estimar la variabilidad del método de aprendizaje

con respecto al conjunto de datos. Para esta investigación, se utilizó 10 particiones (10-fold cross validation) teniendo en cuenta lo recomendado por Hernández *et al*. (2005).

Al seleccionar la técnica de clasificación con árboles de decisión se pretende obtener un modelo que permita predecir para los nuevos casos de estudiantes de la Universidad Javeriana Cali, los factores socioeconómicos, académicos e institucionales asociados al buen (por encima de la media) o mal (por debajo de la media) desempeño académico en las pruebas Saber Pro, teniendo en cuenta, como atributo clase, el puntaje global obtenido en las pruebas Saber Pro.

Se escogió el algoritmo J48 de la herramienta Weka para la construcción de los modelos de clasificación con árbol de decisión. El algoritmo J48 se basa en la utilización del criterio de ganancia de información (information gain). De esta manera se consigue evitar que las variables con mayor número de posibles valores salgan beneficiadas en la selección. Además, el algoritmo incorpora una poda del árbol de clasificación una vez que éste ha sido inducido. El parámetro más importante que se tuvo en cuenta para la poda fue el factor de confianza C (confidence level), que influye en el tamaño y capacidad de predicción del árbol construido. Cuanto más baja se haga esa probabilidad, se exigirá que la diferencia en los errores de predicción antes y después de podar sea más significativa para no podar. El valor por defecto de este factor es del 25% y conforme va bajando este valor, se permiten más operaciones de poda y por lo tanto llegar a árboles cada vez más pequeños (García and Álvarez, 2010). Otro parámetro utilizado para variar el tamaño del árbol fue a través del factor M que especifica el mínimo número de instancias o registros por nodo del árbol.

Se generaron diferentes modelos de árboles de decisión con el fin de escoger el árbol de decisión que mejor clasifique a los estudiantes y con mayor nivel de interpretabilidad de los patrones asociados al desempeño académico. Por esta razón, se configuraron dos valores para el factor de confianza C en 25%, 50%, combinándolos con dos valores para el factor M en 2.5 % (52 ejemplos) % y 5%(104 ejemplos). Además, se aplicó un proceso de pospoda para dejar las ramas y por ende las reglas más representativas, que son aquellas que sobrepasan un mínimo soporte del 2.5% y una confianza del 60%.

El mejor árbol fue construido con los parámetros C=0.25 y M=52 para la prepoda y con soporte mayor o igual al 2.5% para la postpoda. En la figura 1 se muestra el modelo de clasificación obtenido con la herramienta Weka.
Para evaluar o estimar el coste del modelo de clasificación construido, se utilizó la matriz de confusión, también llamada tabla de contingencia. La matriz de confusión es una herramienta que permite visualizar el desempeño de un algoritmo de aprendizaje supervisado. Esta se muestra en la figura 2.

```
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

facultades = Ciencias Economicas y Administrativas
|   estu_grupo_etario = [22]: Bajo la Media (111.0/53.0)
|   estu_grupo_etario = [21]: Sobre la Media (99.0/40.0)
|   estu_grupo_etario = [23]: Bajo la Media (107.0/51.0)
|   estu_grupo_etario = [24]: Bajo la Media (81.0/31.0)
|   estu_grupo_etario = [>=25]: Bajo la Media (217.0/70.0)
facultades = Ingeniería y Ciencias: Sobre la Media (467.0/181.0)
facultades = Humanidades y Ciencias Sociales
|   fami_educacionmadre = Postgrado: Sobre la Media (132.0/51.0)
|   fami_educacionmadre = Educacion profesional incompleta: Sobre la Media (62.0/23.0)
|   fami_educacionmadre = Secundaria (Bachillerato) completa: Bajo la Media (92.0/40.0)
|   fami_educacionmadre = Educacion profesional completa
|   |   fami_numlibros = 26 A 100 LIBROS: Sobre la Media (105.23/51.82)
|   |   fami_numlibros = MAS DE 100 LIBROS: Sobre la Media (74.87/27.58)
|   |   fami_numlibros = 11 A 25 LIBROS: Bajo la Media (52.61/23.2)
|   fami_educacionmadre = Tecnica o tecnologica completa: Bajo la Media (81.0/26.0)
facultades = Ciencias de la Salud: Sobre la Media (194.0/45.0)

Number of Leaves  :     21

Size of the tree :  25

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1214               59.1618 %
Incorrectly Classified Instances       838               40.8382 %

Total Number of Instances             2052
```

Fig. 1. Árbol de Clasificación SaberPro

```
Total Number of Instances          2052

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0,687    0,516    0,600      0,687   0,641      0,175    0,618     0,629     Sobre la Media
                 0,484    0,313    0,578      0,484   0,527      0,175    0,618     0,572     Bajo la Media
Weighted Avg.    0,592    0,421    0,590      0,592   0,587      0,175    0,618     0,602

=== Confusion Matrix ===

   a   b   <-- classified as
 747 341 |   a = Sobre la Media
 497 467 |   b = Bajo la Media
```

Fig. 2. Matriz de Confusión

3.4 Discusión de Resultados

Analizando los resultados sobre el desempeño general de los estudiantes de la Universidad Javeriana Cali en las pruebas Saber Pro 2017-2018, obtenidos en el árbol de decisión que se muestra en la figura 1, se puede observar que este clasifica correctamente a 1214 instancias, que corresponde a una exactitud del 59,1% y 838 instancias incorrectamente clasificadas, correspondiente a un porcentaje del 40.9%.

Evaluando el modelo con la matriz de confusión, obtenido con la herramienta Weka, de la figura 2, este predice correctamente a 747 casos de estudiantes cuyo desempeño general está sobre la media (TP) y a 467 casos que están bajo la media (TN). Por otra parte, 341 casos cuyo desempeño esta sobre la media, el modelo los clasifica incorrectamente como bajo la media (FN) y 497 casos cuyo desempeño está bajo la media, el modelo los clasifica incorrectamente sobre la media (FP).

Para el caso de los estudiantes que están sobre la media en el puntaje global, el modelo tiene una precisión de predicción de 0.60, lo que quiere decir que, del total de casos predichos que están sobre la media, el 60% son correctos. La sensibilidad (TPR) y Recall del modelo, es de 0.687, lo que indica que el modelo clasifica correctamente al 68.7 % de los estudiantes que realmente están sobre la media. Por otra parte, la tasa de Falsos Positivos (FP Rate) del modelo es de 0.516, lo que significa que el 51.6% de estudiantes que estaban bajo la media fueron clasificados como sobre la media. El F-measure es de 0.641 lo que significa que la media armónica entre la precisión y el recall de los que están sobre la media es del 64.1%. En la combinación de estas medidas se aprecia un mejor desempeño del modelo para los que están sobre la media.

Para el caso de los estudiantes que están bajo la media en el puntaje general, el modelo tiene una precisión de predicción de 0.578, lo que quiere decir que del total de casos predichos que están bajo la media, el 57.8 % son correctos. La especificidad (TNR) y Recall del modelo, es de 0.484, lo que indica que el modelo clasifica correctamente al 48.4 % de los estudiantes que realmente están bajo la media. Por otra parte, la tasa de Falsos Negativos (FN Rate) del modelo es de 0.313, lo que significa que el 31.3% de estudiantes que estaban sobre la media fueron clasificados como bajo la media. El F-measure es de 0.527 lo que significa que la media armónica entre la precisión y el recall de los que están bajo la media es del 52.7%. En la combinación de estas medidas se aprecia un desempeño moderado del modelo para los que están bajo la media.

El modelo construido para detectar patrones de rendimiento general en las pruebas Saber Pro de los estudiantes de la Universidad Javeriana Cali no está excesivamente desbalanceado ya que hay una diferencia pequeña de casos entre los que están sobre la media (53%) y los que están bajo la media (47%) y es de 124 casos (6%). Por esta razón, dentro de las métricas de evaluación calculadas anteriormente, se puede decir que el modelo tiene una exactitud del 59.1% y que predice mejor a los estudiantes que están sobre la media que a los que están bajo la media. Esto también lo muestra en la relación entre el Recall y la Precisión dada en el PRC área, donde para los estudiantes que están sobre la media es de 0,629 y los que están bajo la media es de 0.572. El coeficiente de correlación de Mathews MCC del modelo es de 0.175, lo que indica que hay una relación débil entre lo predicho y lo observado, es decir una baja calidad en la predicción. Finalmente, en cuanto a las áreas, el área ROC del modelo por ser mayor que 0.5 indica que el modelo tiene un desempeño bueno en la clasificación de los estudiantes de la Universidad Javeriana Cali con respecto al puntaje global obtenido en las pruebas Saber Pro 2017-2018.

Con respecto a los patrones de desempeño global en las pruebas Saber Pro que presentaron los estudiantes de la Universidad Javeriana Cali en los años 2017 y 2018 y que se pueden observar en la figura 1, las siguientes reglas son la interpretación de estos patrones. Para escoger los patrones más representativos, se tuvo en cuenta aquellos que superen un soporte mínimo del 2.5% y una confianza mínima del 60%:

Regla 1. Si el estudiante es de la facultad de Ciencias Económicas y Administrativas y pertenece al grupo etario de los 21 años entonces su desempeño general en las pruebas Saber Pro tiene mayor probabilidad de estar sobre la media. El 4.8% del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 59.6% de los 99 estudiantes que tienen estas características, son correctamente clasificados y el 5.4 % del total de estudiantes observados que están sobre la media cumplen este patrón.

Regla 2. Si el estudiante es de la facultad de Ciencias Económicas y Administrativas y pertenece al grupo etario de los 24 años entonces su desempeño general en las pruebas Saber Pro

tiene mayor probabilidad de estar bajo la media. El 3.9% del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 61.7% de los 81 estudiantes que tienen estas características, son correctamente clasificados y el 5.2 % del total de estudiantes observados que están bajo la media cumplen este patrón.

Regla 3. Si el estudiante es de la facultad de Ciencias Económicas y Administrativas y pertenece al grupo etario de los mayores o iguales que 25 años entonces su desempeño general en las pruebas Saber Pro tiene mayor probabilidad de estar bajo la media. El 10.6% del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 67.7% de los 217 estudiantes que tienen estas características, son correctamente clasificados y el 15.2 % del total de estudiantes observados que están bajo la media cumplen este patrón.

Regla 4. Si el estudiante es de la facultad de Ingeniería y Ciencias entonces su desempeño general en las pruebas Saber Pro tiene mayor probabilidad de estar sobre la media. El 22.8% del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 61.2% de los 467 estudiantes que tienen estas características, son correctamente clasificados y el 26.3 % del total de estudiantes observados que están sobre la media cumplen este patrón.

Regla 5. Si el estudiante es de la facultad de Humanidades y Ciencias Sociales y el nivel de educación de la madre es postgrado entonces su desempeño general en las pruebas Saber Pro tiene mayor probabilidad de estar sobre la media. El 6.4% del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 61.4% de los 132 estudiantes que tienen estas características, son correctamente clasificados y el 7.4 % del total de estudiantes observados que están sobre la media cumplen este patrón.

Regla 6. Si el estudiante es de la facultad de Humanidades y Ciencias Sociales y el nivel de educación de la madre es profesional incompleta entonces su desempeño general en las pruebas Saber Pro tiene mayor probabilidad de estar sobre la media. El 3.0% del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 62.9% de los 62 estudiantes que tienen estas características, son correctamente clasificados y el 3.6 % del total de estudiantes observados que están sobre la media cumplen este patrón.

Regla 7. Si el estudiante es de la facultad de Humanidades y Ciencias Sociales, el nivel de educación de la madre es profesional completa y la familia tiene más de 100 libros entonces su desempeño general en las pruebas Saber Pro tiene mayor probabilidad de estar sobre la media. El 3.6% del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 63.5% de los 74 estudiantes que tienen estas características, son correctamente clasificados y el 4.3 % del total de estudiantes observados que están sobre la media cumplen este patrón.

Regla 8. Si el estudiante es de la facultad de Humanidades y Ciencias Sociales y el nivel de educación de la madre es técnica o tecnológica entonces su desempeño general en las pruebas Saber Pro tiene mayor probabilidad de estar bajo la media. El 3.9% del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 67.9% de los 81 estudiantes que tienen estas características, son correctamente clasificados y el 5.7 % del total de estudiantes observados que están bajo la media cumplen este patrón.

Regla 9. Si el estudiante es de la facultad de Ciencias de la Salud entonces su desempeño general en las pruebas Saber Pro tiene mayor probabilidad de estar sobre la media. El 9.5% del total de estudiantes de la Universidad Javeriana Cali analizados se clasifican de esta manera. El 76.8% de los 194 estudiantes que tienen estas características, son correctamente clasificados y el 13.7 % del total de estudiantes observados que están sobre la media cumplen este patrón.

4. CONCLUSIONES

En esta investigación se escogió el modelo predictivo de clasificación con árboles de decisión basado en el algoritmo J48 de la herramienta Weka, para detectar patrones de desempeño general en las pruebas Saber Pro que presentaron los estudiantes de la Universidad Javeriana Cali en los años 2017 y 2018. Para la preparación de los datos, la construcción y evaluación del modelo se siguieron las diferentes etapas de la metodología CRISP-DM.

Entre las variables predictoras de los patrones descubiertos teniendo en cuenta el puntaje global obtenido por los estudiantes de la Universidad

Javeriana Cali en las pruebas Saber Pro 2017-2018, están la facultad, el grupo etario, la educación de la madre y el número de libros disponibles en la casa del estudiante, como cuatro variables importantes asociadas al buen o bajo desempeño académico en estas pruebas. Particularmente, entre los patrones descubiertos se destacan el buen desempeño de los estudiantes de las facultades de Ingeniería y Ciencias y la de Ciencias de la Salud, con una participación de un buen número de estudiantes con relación a otras facultades.

De acuerdo a las métricas de calidad del modelo analizadas, el modelo tiene un mejor desempeño para los que están sobre la media (positivos) que los que están bajo la media (negativos). Esto significa que el modelo es más específico que sensible, lo que indica que el modelo trata de evitar los falsos positivos.

Se plantean como trabajos futuros construir modelos de árboles de decisión para predecir el desempeño de los estudiantes de la Universidad Javeriana Cali en cada una de las competencias genéricas de las pruebas Saber Pro. Aplicar técnicas descriptivas de minería de datos con el fin de analizar las relaciones de asociación existentes entre los atributos socioeconómicos, académicos e institucionales de los estudiantes teniendo en cuenta el desempeño en las competencias genéricas de las pruebas Saber Pro y analizar la forma como se pueden agrupar estos estudiantes de acuerdo a su similitud en el rendimiento en estas pruebas.

REFERENCIAS

Algarni, A., 2016. Data Mining in Education. In (IJACSA) International Journal of Advanced Computer Science and Applications. Vol. 7, No. 6. 2.

Azevedo,A. and Santos,M., 2008. KDD, SEMMA and CRISP-DM: a parallel overview. *Proceedings of IADIS European Conference on Data Mining*. (pp. 182-185). Amsterdan, Netherlands. ISBN: 978-972-8924-63-8.: https://www.researchgate.net/publication/220969845_KDD_semma_and_CRISP-DM_A_parallel_overview.

Baker, R.,2010. Data Mining for Education. In McGaw, B., Peterson, P., Baker, E. (Eds.) International. Encyclopedia of Education (3rd edition), vol. 7, pp. 112-118. Oxford, UK: Elsevier.

Escobar, H., Alcívar, M., Márquez, C. and Escobar, C., 2017. Implementación de Minería de Datos en la Gestión Académica de las Instituciones de Educación Superior. *Didasc@lia: Didáctica y Educación*. ISSN 2224-2643, 8(3), 203-212. [En Línea]. Disponible en: http://revistas.ult.edu.cu/index.php/didascalia/article/view/637.

García, M. and Álvarez, A., 2010. Análisis de datos en WEKA–pruebas de selectividad. [En línea]. Disponible en: *http://www.it.uc3m.es/~jvillena/irc/practicas/06-07/28.pdf.*

Hall, M., Frank, E. and Witten, I., 2011. Practical Data Mining: Tutorials. University of Waikato. [En línea]. Disponible en: *http://www.micai.org/2012/tutorials/Weka%20tutorials%20Spanish.pdf.*

Han, J. and Kamber, M., 2001. *Data Mining: Concepts and Techniques*, Third Edition (3 edition.). Burlington, MA: Morgan Kaufmann.

Hernández, J., Ramírez, M. and Ferri, C., 2005. *Introducción a la Minería de Datos*. Editorial Pearson Prentice Hall. Madrid, España. ISBN: 84-205-4091-9.

Hernández, E. and Lorente, R., 2009. Mineria de datos aplicada a la detección de Cáncer de Mama. Universidad Carlos III de Madrid. Disponible en: *http://tps5to-utn-frre.googlecode.com/svn/trunk/BI/Cancer%20de%20Mama/14.pdf.*

Icfes, 2012. *Saber Pro: Principales resultados en Competencias Genéricas.* Santa Marta, Colombia. [En línea]. Disponible en: www.icfes.gov.co/examenes/.../151-saber-pro-modulos-de-competencias.

Icfes, 2017. *Saber Pro: Módulos de Competencias Genéricas 2017*. Instituto Colombiano para la Evaluación de la Educación Superior. [En línea]. Disponible en: https://www.icfes.gov.co/documents/20143/495161/Guia%20de%20orientacion%20modulos%20de%20competencias%20genericas-saber-pro-2017.pdf.

MEN, 2006. Ministerio de Educación Nacional (MEN). Estándares Básicos de Competencias en Lenguaje, Matemáticas, Ciencias y Ciudadanas: Guía sobre lo que los estudiantes deben saber y saber hacer con lo que aprenden. ISBN: 958-691-290-6. Bogotá D.C., Colombia.

Peña, 2014. Educational data mining: A survey and a data mining-based analysis of recent works. Expert Systems with Applications, 1432–1462.

Posada, J. and Mendoza, F., 2014. Determinantes del logro académico de los estudiantes de grado 11 en el periodo 2008-2010. Una perspectiva de género y región. Estudios sobre calidad de la educación en Colombia, ICFES, Ministerio de Educación Nacional. Bogotá, Colombia.

Quinlan, J., 1993. *C4. 5: programs for machine learning (Vol. 1)*. Morgan kaufmann. Disponible en: http://books.google.com.co/books?hl=es&lr=&id=HExncpjbYroC&oi=fnd&pg=PR7&dq=Programs+for+Machine+Leraning&ots=nLkbbRq2Yj&sig=Y5h5CQUdtbZjs1Fjd8ilbJfyRLE.

Sattler, K. and Dunemann, O., 2001. SQL database primitives for decision tree classifiers. *Proceedings*

of the tenth international conference on Information and knowledge management (pp. 379–386). ACM. Disponible en: http://dl.acm.org/citation.cfm?id=502650.

Timarán, R., Hernández, I., Caicedo, J., Hidalgo, A. and Alvarado, J., 2016. *Descubrimiento de patrones de desempeño académico*. Ediciones Universidad Cooperativa de Colombia, Bogotá, abril de 2016. ISBN (digital): 978-958-760-050-6. DOI: http://dx.doi.org/10.16925/9789587600490.

Timarán, R., Jiménez, J. and Calderón, A., 2017. *Detección de patrones de deserción estudiantil con minería de datos*. Editorial Universidad de Nariño, Pasto, Colombia. ISBN:978-958-8958-38-5. Disponible en: https://editorial.udenar.edu.co/?p=2383.

Unal, 2012. *Análisis de los resultados obtenidos por la Universidad Nacional de Colombia sede Bogotá en las pruebas Saber Pro 2011–2*. Universidad Nacional de Colombia. Bogotá: Universidad Nacional de Colombia. [En linea]. Disponible en: ww.unal.edu.co/diracad/evaluacion/SaberPro_2012/analisis_de_resultados.pdf.

Valero, S., Vargas, A. and García, M., 2005. Minería de datos: Predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos. Ene, 779(73), 33. [En línea]. Disponible en: *http://fcaenlinea.unam.mx/anexos/1566/1566_u6_act1b.pdf*.

Valero, S., 2009. Aplicación de técnicas de minería de datos para predecir deserción. Puebla, México: Universidad Tecnológica de Izúcar de Matamoros. Disponible en: http://www.utim.edu.mx/~svalero/docs/MineriaDesercion.pdf.

Villena, J., 2016. *CRISP-DM: La metodología para poner orden en los proyectos de Data Science*. Disponible en: https://data.sngular.team/es/art/25/crisp-dm-la-metodologia-para-poner-orden-en-los-proyectos-de-data-science.

Witten, I., Frank, E. and Hall, M., 2011. *Data Mining: Practical Machine Learning Tools and Techniques (Third Edition)*. Morgan Kaufmann ISBN:978-0-12-374856-0.

Zapata, L., 2011. *Factores académicos asociados al bajo rendimiento en inglés en las pruebas ECAES presentadas por los estudiantes de la Facultad de Educación en el año 2009*. (Trabajo de grado de pregrado). Fundación Universitaria Luis Amigó, Facultad de Educación, Licenciatura en Educación Básica con Énfasis en inglés. Medellín, Colombia.