UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Revista Científica

http://revistas.udistrital.edu.co/ojs/index.php/revcie/index

CIDC

Artículo de investigación científica y tecnológica

Detección de patrones de desempeño académico en la competencia de matemáticas en las pruebas Saber 5º

Detection of Patterns Regarding Academic Performance in the Mathematics Component of the Saber 5 Tests

Detecção de padrões de desempenho acadêmico nas habilidades matemáticas nos testes Saber 5

Ricardo Timarán-Pereira[1]
Javier Caicedo-Zambrano[2]
Arsenio Hidalgo-Troya[3]

Recibido: marzo de 2022 **Aceptado:** mayo de 2023

Para citar este artículo: Timarán-Pereira, R., Caicedo-Zambrano, J. y Hidalgo-Troya, A. (2023). Detección de patrones de desempeño académico en la competencia de matemáticas en las pruebas Saber 5º. *Revista Científica*, *47*(2), 127-137. https://doi.org/10.14483/23448350.20908

Resumen

Las pruebas Saber 5° buscan contribuir al mejoramiento de la calidad de la educación en Colombia. Su aplicación es periódica, evalúa las competencias básicas de los estudiantes y analiza los factores que inciden en sus logros. En este artículo se presenta uno de los resultados de una investigación cuyo objetivo fue aplicar técnicas de minería de datos para detectar patrones de desempeño académico en la competencia de matemáticas de las pruebas Saber 5º del año 2017. Esta prueba fue presentada por los estudiantes de grado quinto de las instituciones educativas colombianas de Básica Primaria. Para cumplir con este objetivo, se utilizó la metodología CRISP-DM. Se obtuvo información socioeconómica, académica e institucional de las bases de datos del ICFES. Esta información fue preprocesada utilizando técnicas de limpieza y transformación de datos. Se seleccionó el modelo de clasificación por árboles de decisión por su facilidad para interpretar patrones. Entre los factores más importantes de los patrones asociados al buen o mal desempeño académico en matemáticas están la naturaleza y la zona de ubicación del colegio y si el estudiante reprobó o no grado. El conocimiento generado en esta investigación constituye información de calidad para la toma de decisiones del Ministerio de Educación Nacional, las secretarias de educación y las directivas de las instituciones educativas de Básica Primaria en la definición de planes de mejoramiento que redunden en la calidad de la educación en Colombia.

Palabras clave: árboles de decisión; competencia de matemáticas; desempeño académico; detección de patrones; pruebas Saber 5º.

1. Ph. D. Universidad de Nariño (Pasto-Nariño, Colombia). ritimar@udenar.edu.co.
2. M. Sc. Universidad de Nariño (Pasto-Nariño, Colombia). jacaza@udenar.edu.co.
3. M. Sc. Universidad de Nariño (Pasto-Nariño, Colombia). arch@udenar.edu.co.

Abstract
The Saber 5° tests seek to contribute to improving the quality of education in Colombia. Their application is periodic, they evaluate students' basic skills, and they analyze the factors that affect their achievements. This paper presents one of the results of a research work whose objective was to apply data mining techniques to detect patterns of academic performance regarding the mathematics component of the Saber 5 tests in the year 2017. This test was taken by the fifth-grade students of Colombian Primary Education institutions. To meet this objective, the CRISP-DM methodology was used. Socioeconomic, academic, and institutional information was obtained from the ICFES databases. This information was preprocessed using data cleaning and transformation techniques. The decision tree classification model was selected, as it allows easily interpreting patterns. Among the most important factors in the patterns associated with good or poor academic performance in mathematics are the nature and location of the school and whether or not the student had failed a grade. The knowledge generated in this research constitutes quality information for decision-making by the Ministry of National Education, education secretariats, and the executives of Primary Education institutions with regard to the definition of improvement plans that result in the quality of Primary Education in Colombia.
Keywords: academic performance; decision trees; mathematics skill; pattern detection; Saber 5o tests.

Resumo
Os testes Saber 5° buscam contribuir para melhorar a qualidade da educação na Colômbia. Sua aplicação é periódica, avalia as habilidades básicas dos alunos e analisa os fatores que afetam suas realizações. Este artigo apresenta um dos resultados de uma investigação cujo objetivo foi aplicar técnicas de mineração de dados para detectar padrões de desempenho acadêmico na habilidade matemática dos testes Saber 5º do ano 2017. Este teste foi apresentado por alunos de graduação quinto do ensino colombiano instituições da Primária Básica. Para atender a esse objetivo, foi utilizada a metodologia CRISP-DM. Informações socioeconômicas, acadêmicas e institucionais foram obtidas nas bases de dados do ICFES. Essas informações foram pré-processadas usando técnicas de limpeza e transformação de dados. O modelo de classificação da árvore de decisão foi selecionado por sua facilidade de interpretação dos padrões. Entre os fatores mais importantes nos padrões descobertos, associados ao bom ou mau desempenho acadêmico em matemática, estão a natureza e a localização da escola e se o aluno foi reprovado ou não. O conhecimento apurado nesta pesquisa constitui informação de qualidade para a tomada de decisão do Ministério da Educação Nacional, das secretarias de educação e das diretrizes das instituições de ensino fundamental básico, na definição de planos de melhoria que resultem na qualidade do ensino fundamental fundamental em Colômbia.
Palavras-chaves: árvores de decisão; competência matemática; desempenho acadêmico; detecção de padrões; testes Saber 5o.

INTRODUCCIÓN

Los resultados de pruebas nacionales e internacionales muestran que Colombia posee un sistema educativo con bajos logros académicos por parte de sus estudiantes en cada uno de los niveles de estudio (Posada and Mendoza, 2014). Esta situación es crítica pues, si persisten estos desempeños 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 (OECD, 2016).

La Ley 1324 le confiere al Instituto Colombiano para Evaluación de la Educación (ICFES) la misión de evaluar, mediante exámenes externos estandarizados, la formación que se ofrece en los distintos niveles del servicio educativo. Esta ley también establece que el MEN define lo que debe evaluarse en estos exámenes (ICFES, 2014). Actualmente, el ICFES diseña y aplica las pruebas Saber 3°, Saber 5°, Saber 9° y Saber 11°, con las cuales evalúa la Educación Básica y Media; y Saber Pro, para evaluar la Educación Superior.

La prueba Saber 5° está dirigida a estudiantes de quinto grado de Básica Primaria, y su objetivo es contribuir al mejoramiento de la calidad de la educación. Esto, mediante la realización de pruebas periódicas en las que se evalúan las competencias básicas de los estudiantes y se analizan los factores que inciden en sus logros. Los resultados de estas evaluaciones permiten que los establecimientos educativos, las secretarías de educación, el Ministerio de Educación Nacional (MEN) y la sociedad en general conozcan cuáles son las fortalezas y debilidades del sistema educativo y, a partir de estas, puedan definir planes de mejoramiento en sus respectivos ámbitos de actuación. Su carácter periódico posibilita, además, valorar cuáles han sido los avances en un determinado periodo de tiempo, así como determinar el impacto de programas y acciones específicas de mejoramiento (ICFES, 2014). Según la Guía de Orientación de las pruebas Saber 5º del ICFES (2017), en la prueba de matemáticas se evalúa el uso flexible de esta ciencia en diversas situaciones. Las competencias comunicativas en lenguaje se evalúan a través de dos instrumentos: uno enfocado en evaluar la comprensión lectora y otro enfocado en evaluar la competencia escritora. En las competencias ciudadanas se evalúa la capacidad de los estudiantes para participar de manera constructiva y activa como ciudadanos en la sociedad. En la prueba de ciencias naturales se evalúa la capacidad de comprender y usar nociones, conceptos y teorías de las ciencias naturales en la resolución de problemas. La prueba, además, involucra el proceso de indagación, lo que implica observar y relacionar patrones en los datos para derivar conclusiones de fenómenos naturales.

En el caso de esta investigación, se analizaron los resultados en matemáticas y lenguaje. Dichas competencias se evaluaron en el año 2017, y esta es la última información disponible en las bases de datos del ICFES. En el año 2022, el ICFES volvió aplicar estas pruebas, pero, por el momento, no hay información disponible sobre los resultados obtenidos.

Se han realizado varios estudios sobre el desempeño académico en las pruebas Saber 5º, tales como los realizados por Torres *et al*. (2014), Martín (2015) y Gutiérrez (2015), quienes buscaban identificar las variables asociadas al rendimiento académico, en especial al desempeño en las pruebas Saber 5°, con base en solo una de las áreas fundamentales, *i.e.*, ciencias naturales, matemáticas y lenguaje respectivamente. En otro estudio (ICFES, 2009) se analizaron los factores asociados de las pruebas de grado 5° y 9°. Una de las conclusiones de dicho estudio fue que, entre más alto sea nivel socioeconómico de los alumnos y sus familias, mayor será el desempeño esperado en ambas áreas y grados evaluados. Además, los estudiantes matriculados en colegios privados tienden a obtener puntajes más altos en las pruebas, y las diferencias frente a quienes asisten a planteles oficiales aumentan en la medida en que mejoran las condiciones socioeconómicas. En el informe realizado sobre factores asociados en las pruebas Saber 5° y 9° del ICFES (2011), se identificaron variables relacionadas con el rendimiento. Se aplicaron técnicas estadísticas que permitieron visualizar los elementos que inciden en el desempeño académico. Por otra parte, para extender sus procesos de evaluación, el ICFES dio paso al estudio de los factores asociados al rendimiento escolar, utilizando modelos teóricos para explicar las relaciones existentes entre los elementos que determinan el aprendizaje y están presentes en tres niveles: instituciones educativas, aulas de clase y estudiantes (ICFES, 2017).

Según Timarán *et al*. (2021a, 2021b) los estudios que se han realizado hasta el momento, con respecto al análisis de los resultados de las pruebas Saber 5º se basan en información procesada mediante análisis estadístico, donde fundamentalmente se consideran variables y relaciones primarias, sin tener en cuenta las verdaderas interrelaciones, que generalmente están ocultas y únicamente se pueden descubrir utilizando un tratamiento más complejo de los datos. Esto es posible con la minería de datos, que descubre patrones no previstos con la estadística en vista de que la estadística plantea hipótesis que deben ser validadas a partir de los datos disponibles.

En este artículo se aplican técnicas de minería de datos para descubrir patrones de desempeño académico en matemáticas en las pruebas Saber 5°. Este estudio se centra en los estudiantes de grado quinto de las instituciones educativas colombianas de básica primaria en el año 2017.

METODOLOGÍA

Esta investigación fue de tipo descriptivo y empleó un enfoque cuantitativo aplicando un diseño no experimental. Se utilizó la metodología CRISP-DM (*cross-industry standard process for data mining*), que involucra la minería de datos. Según Timarán *et al*. (2013) y Valero *et al*. (2005), CRISP-DM es uno de los modelos utilizados, principalmente, en los ambientes académico e industrial y la guía de referencia más ampliamente utilizada en el desarrollo de este tipo de proyectos. CRISP-DM está compuesta por seis fases: análisis del problema, análisis de los datos, preparación de los datos, modelado, evaluación e implementación.

En la fase de análisis del problema, se recopiló y seleccionó el material bibliográfico necesario para que los investigadores pudieran conocer y apropiar el conocimiento acerca de las pruebas Saber 5º y de las competencias que evalúa, haciendo énfasis en el área de matemáticas. Este proceso posibilitó la recolección de los datos correctos para obtener resultados adecuados.

En la fase de análisis de datos, los investigadores identificaron, recopilaron y se familiarizaron con la información socioeconómica, académica e institucional que, al momento de realizar el estudio, estaba disponible en las bases de datos del ICFES y correspondía a los resultados en matemáticas obtenidos por los estudiantes colombianos que presentaron las pruebas Saber 5º en el año 2017. Como resultado, se obtuvo un conjunto de datos inicial, denominado *sbr5_776436A56*, con 776 436 registros y 56 atributos.

En la fase de preparación de los datos, y teniendo en cuenta que la alta dimensionalidad es un problema para descubrir patrones en minería de datos, al conjunto *sbr5_776436A56* se le aplicaron técnicas de limpieza y transformación, con el fin de eliminar los datos ruidosos, nulos y atípicos, así como transformar algunos atributos para obtener mayor ganancia de información y eliminar los atributos irrelevantes que no aportaban al proceso de detección de patrones. Esto resultó en el conjunto de datos denominado *sbr5_776436A15*, compuesto por 776436 registros y 15 atributos, el cual sirvió de base para la fase de modelado. En la Tabla 1 se muestra el diccionario de datos del conjunto *sbr5_776436A15*.

En la fase del modelado se seleccionó un modelo de clasificación con árboles de decisión como la técnica de minería de datos más adecuada para solucionar el problema del estudio, dada su facilidad y simplicidad para interpretar los patrones obtenidos (Azevedo and Santos, 2008; Hernández *et al*., 2005; Timarán *et al*., 2017). Esta técnica tiene varias ventajas. En primer lugar, 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, de manera automática, el proceso incluye 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 (Han *et al*., 2011, Sattler and Dunemann, 2001; Timarán *et al*., 2006).

Tabla 1. *Diccionario de datos sbr5_776436A15*

No	Atributo	Descripción
1	estu_genero	Sexo del estudiante
2	estu_edad	Intervalo de edad al cual pertenece el estudiante al momento de presentar la prueba
3	estu_reprobogrado	Si el estudiante reprobó o no grado
4	fami_educacion_madre	Máximo nivel educativo de la madre
5	fami_educacion_padre	Máximo nivel educativo del padre
6	fami_hacinamiento	Índice de hacinamiento de la familia
7	fami_tics	Condición de uso de TIC en el hogar del estudiante
8	fami_electrodomesticos	Condición relacionada con los electrodomésticos en el hogar del estudiante
9	cole_genero	Si el colegio es mixto, femenino o masculino
10	cole_naturaleza	Si el colegio es oficial o privado
11	cole_caracter	Si el colegio es técnico, académico o técnico/académico
12	cole_area_ubicación	Si el colegio está en el área urbana o rural
13	cole_jornada	Jornada de estudio del colegio
14	cole_región_ubicación	Región del país en la que se ubica el colegio
15	clase_desemp_matematicas	Desempeño en matemáticas (sobre la media, bajo la media)

De acuerdo con las recomendaciones de Hernández *et al*. (2005), para probar la calidad y validez del modelo, se utilizó el método de validación cruzada con 10 particiones (*10-fold cross validation*), con el fin de reducir la dependencia del resultado del modelo con respecto al modo en que se realiza la partición. Se tomó como atributo clase o variable objetivo el puntaje obtenido en la prueba de matemáticas, el cual se discretizó en dos valores: "sobre la media" y "bajo la media" nacional. En la fase de evaluación, se estimó el coste del clasificador para los repositorios de entrenamiento y prueba a través de una matriz de confusión (Sattler and Dunemann, 2001). Por otra parte, se evaluaron los patrones descubiertos con el fin de determinar su validez, remover los patrones redundantes o irrelevantes e interpretar los patrones útiles en términos que sean entendibles para el usuario, teniendo en cuenta el soporte y la confianza del patrón.

En la fase de implementación se documentaron los patrones descubiertos, los cuales constituyen información de calidad para la toma de decisiones del MEN, las secretarías de educación y las directivas de las instituciones educativas de Básica Primaria en la definición de planes de mejoramiento que redunden en la calidad de la educación en Colombia.

RESULTADOS

Análisis exploratorio de datos

Con el fin de entender los datos, se analizaron las variables socioeconómicas de los estudiantes de las instituciones educativas de Colombia que presentaron las pruebas Saber 5º en el año 2017. En la Tabla 2 se muestran los resultados.

La Tabla 2 muestra que, por género, la mayoría de los estudiantes son hombres, con un 52.3%; por edad, la mayoría está entre los 10 (39.9 %) y los 11 años (35.2 %). Casi la totalidad de los estudiantes (89.3%) no presentaba hacinamiento en su vivienda. Por otra parte, los padres de los estudiantes que presentaron las pruebas son mayoritariamente bachilleres (en las madres el 48. 2% y en los padres el 47 %). La mayoría de los estudiantes tienen buenos índices de TIC y electrodomésticos en sus hogares (43.6 y

47.7 % respectivamente). Finalmente, el 75 % de los estudiantes que presentaron las pruebas Saber 5º en el año 2017 no habían reprobado ningún grado.

Definición del modelo

Se evaluaron diferentes algoritmos de árboles de decisión con la herramienta Weka versión 3.9.4, un *software* de minería de datos desarrollado en la Universidad de Waikato de Nueva Zelanda (Hall *et al.*,2011), con el fin de seleccionar la técnica de árboles de decisión que mejor clasificara el conjunto de datos *sbr5_776436A15*. Los resultados se muestran en la Tabla 3.

Tabla 2. *Características socioeconómicas de los estudiantes de grado quinto que presentaron las pruebas Saber 5º en el año 2017-*

Variables Socioeconómicas		N	%
Genero	Masculino	406 117	52.3
	Femenino	370 319	47.7
Edad	9 años o menos	27 032	3.5
	10 años	310 129	39.9
	11 años	273 647	35.2
	12 años o mas	165 628	21.3
Educación del padre	Primaria	216 688	27.9
	Bachillerato	365 220	47.0
	Técnico o Tecnólogo	49 224	6.3
	Universitario – Pregrado	61 981	8.0
	Universitario – Posgrado	83 323	10.7
Educación de la madre	Primaria	186 504	24.0
	Bachillerato	374 350	48.2
	Técnico o Tecnólogo	50 385	6.5
	Universitario – Pregrado	74 520	9.6
	Universitario – Posgrado	90 677	11.7
Hacinamiento	Hacinamiento crítico	8 400	1.1
	Hacinamiento medio	75 064	9.7
	Sin hacinamiento	692 972	89.3
TICS	Malo	324 183	41.7
	Regular	113 957	14.7
	Bueno	338 296	43.6
Electrodomésticos	Malo	188 962	24.3
	Regular	217 084	28.0
	Bueno	370 390	47.7
Reprobado	No	582 240	75.0
	Sí	194 196	25.0
Total		**776 436**	**100.0**

Tabla 3. *Evaluación de diferentes algoritmos de árboles de decisión.*

Algoritmo	Exactitud
Decision Stump (árbol de decisión de un nivel)	62.09 %
J48	63.91 %
LMT (Logistic Model Tree)	64.86 %
Random Forest	63.61 %
Random Tree	62.04 %
RepTree	63.20 %
AdTree	63.83 %

Según la Tabla 2, el algoritmo con mayor exactitud fue LMT, pero, dada la dificultad para interpretar los patrones, no fue posible escogerlo. Por esa razón se escogió el algoritmo J48 para la construcción de los modelos de clasificación con árbol de decisión, además del hecho de que facilita el entendimiento de los patrones.

Una vez seleccionado el algoritmo y el método para el entrenamiento y prueba de los modelos, se procedió a construir los diferentes árboles de decisión con la herramienta Weka 3.9.4 y su algoritmo J48, el cual implementa el algoritmo C4.5 (Quinlan, 1993, 1996).

Para la poda del árbol se tuvo en cuenta el nivel de confianza C (*confidence level*), que influye en el tamaño y capacidad de predicción del árbol construido. El valor por defecto de este factor es del 25 % y, conforme va bajando este valor, se permiten más operaciones de poda y, por tanto, árboles cada vez más pequeños (García and Álvarez,2010). Se utilizó también el parámetro M, que determina el mínimo número de registros por nodo del árbol. Se escogió como clase el atributo del puntaje obtenido en matemáticas, cuyos valores fueron discretizados en las categorías "sobre la medial" y "bajo la media".

Para la competencia de matemáticas, el mejor árbol se obtuvo con los parámetros C 0.5 % y M 1 %, como se muestra en la Figura 1. La matriz de confusión se muestra en la Figura 2.

```
J48 pruned tree
------------------
cole_naturaleza = OFICIAL
|   estu_reprobogrado = No
|   |   cole_region_ubicacion = ATLANTICA: BAJO MEDIA (112240.0/37025.0)
|   |   cole_region_ubicacion = CENTRAL
|   |   |   fami_electrodomesticos = MALO: BAJO MEDIA (14188.0/6126.0)
|   |   |   fami_electrodomesticos = BUENO: SOBRE MEDIA (22185.0/10597.0)
|   |   |   fami_electrodomesticos = REGULAR: BAJO MEDIA (13926.0/6917.0)
|   |   cole_region_ubicacion = ANTIOQUIA: BAJO MEDIA (58152.0/23135.0)
|   |   cole_region_ubicacion = PACIFICO: BAJO MEDIA (66062.0/27002.0)
|   |   cole_region_ubicacion = ORIENTAL
|   |   |   estu_edad = 10 anos: SOBRE MEDIA (40700.0/17552.0)
|   |   |   estu_edad = 12 anos o mas: BAJO MEDIA (5221.0/2059.0)
|   |   |   estu_edad = 11 anos: SOBRE MEDIA (29114.0/12684.0)
|   |   |   estu_edad = 9 anos o menos: SOBRE MEDIA (2888.0/1432.0)
|   |   cole_region_ubicacion = ORINOQUIA/AMAZONIA: BAJO MEDIA (33873.0/15678.0)
|   |   cole_region_ubicacion = BOGOTA D.E.
|   |   |   estu_edad = 10 anos
|   |   |   |   estu_genero = Hombre: SOBRE MEDIA (10805.0/4993.0)
|   |   |   |   estu_genero = Mujer: BAJO MEDIA (11792.0/5723.0)
|   |   |   estu_edad = 12 anos o mas: BAJO MEDIA (2716.0/901.0)
|   |   |   estu_edad = 11 anos
|   |   |   |   estu_genero = Hombre: SOBRE MEDIA (10091.0/4611.0)
|   |   |   |   estu_genero = Mujer: BAJO MEDIA (9800.0/4769.0)
|   |   |   estu_edad = 9 anos o menos: BAJO MEDIA (1853.0/727.0)
|   estu_reprobogrado = Si: BAJO MEDIA (174073.0/52116.0)
cole_naturaleza = NO OFICIAL: SOBRE MEDIA (156757.0/48134.0)

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

Correctly Classified Instances      494169               63.6458 %
Incorrectly Classified Instances    282267               36.3542 %
Total Number of Instances           776436
```

Figura 1. *Modelo de árbol de decisión generado por el algoritmo J48 de Weka.*

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0,467    0,221    0,640      0,467   0,540      0,260    0,662     0,607     SOBRE MEDIA
                 0,779    0,533    0,635      0,779   0,699      0,260    0,662     0,663     BAJO MEDIA
Weighted Avg.    0,636    0,390    0,637      0,636   0,627      0,260    0,662     0,637

=== Confusion Matrix ===

     a      b   <-- classified as
 165799 188916 |     a = SOBRE MEDIA
  93351 328370 |     b = BAJO MEDIA
```

Figura 2. *Matriz de confusión del modelo de árbol de decisión.*

DISCUSIÓN

Analizando los resultados del desempeño en matemáticas en las pruebas Saber 5º de 2017, el modelo que se muestra en la Figura 1 clasifica 494 169 instancias correctamente, lo que corresponde a una exactitud del 63.6 %, y 282 267 instancias incorrectamente (36.4 % de exactitud).

Evaluando el modelo con la matriz de confusión (Figura 2), obtenido con la herramienta Weka, este predice correctamente 165 799 casos de estudiantes cuyo desempeño en matemáticas está sobre la media (*true positives*, TP) y 328 370 casos que están bajo la media (true negatives, TN). Por otra parte, en el caso de 93 351 estudiantes cuyo desempeño está bajo la media, el modelo los clasifica incorrectamente sobre la media (*false positives*, FP). Asimismo, hay 188 916 casos cuyo desempeño está sobre la media y el modelo los clasifica incorrectamente (*false negatives*, FN).

Para el caso de los estudiantes que están sobre la media en el puntaje de matemáticas, el modelo tiene una precisión de predicción de 0.64, lo que quiere decir que, del total de casos predichos que están sobre la media, el 64 % son correctos. La sensibilidad (TPR) y el *recall* del modelo muestran un valor de 0.47, lo que indica que el modelo clasifica correctamente al 47 % de los estudiantes que realmente están sobre la media. Por otra parte, la tasa de falsos positivos (*FP rate*) del modelo es de 0.22, lo que significa que el 22 % de estudiantes que estaban bajo la media fueron clasificados incorrectamente. El *F-measure* es de 0.54. Esto significa que la media armónica entre la precisión y el *recall* de los que están sobre la media es de 54. En la combinación de estas medidas se aprecia un peor desempeño del modelo para los que están sobre la media.

Para el caso de los estudiantes que están bajo la media en el puntaje de matemáticas, el modelo tiene una precisión de predicción de 0.64, lo que quiere decir que, del total de casos predichos que están bajo la media, el 64 % son correctos (igual que para lenguaje). La sensibilidad (TPR) y el *recall* del modelo reportan un valor de 0.78. Esto indica que el modelo clasifica correctamente al 78 % de los estudiantes que realmente están bajo la media. Por otra parte, la tasa de falsos negativos (*FP rate*) del modelo es de 0.54, lo que significa que el 54 % de estudiantes que estaban sobre la media fueron clasificados bajo la media. El *F-measure* es de 0.70, *i.e.*, la media armónica entre la precisión y el *recall* de los que están bajo la media es de 70. En la combinación de estas medidas se aprecia un mejor desempeño del modelo para los estudiantes que están bajo la media.

Dentro de las métricas de evaluación calculadas anteriormente para la competencia de matemáticas, se puede decir que el modelo tiene una exactitud del 63.6 % y que predice mejor a los estudiantes que están bajo la media. Esto también lo muestra en la relación entre el *recall* y la precisión dada en el PRC *area*. Aquí, para los estudiantes que están bajo la media, el valor es de 0.66, y, para los que están sobre la

media, es de 0.61. El coeficiente de correlación de Mathews (MCC) del modelo es de 0.26, lo que indica que hay una concordancia media entre lo predicho y lo observado, es decir, una calidad regular en la predicción. Finalmente, en cuanto a las áreas, el área ROC de 0.66 del modelo, al ser mayor que 0.5, indica que el modelo tiene un desempeño bueno en la clasificación de los estudiantes colombianos con respecto al puntaje en matemáticas obtenido en las pruebas Saber 5º.

Para escoger los patrones más representativos del desempeño en matemáticas en las pruebas Saber 5º que presentaron los estudiantes colombianos de las instituciones educativas de educación Básica Primaria en el año 2017 (Figura 1), se tuvieron en cuenta aquellos que superaran un soporte mínimo del 1 % y una confianza mínima del 60 %. Las siguientes reglas son la interpretación de los patrones que cumplen estas métricas.

Regla 1. Si el estudiante es de un colegio oficial, no reprobó grados y el colegio está ubicado en la región Atlántica, entonces su desempeño en matemáticas probablemente está bajo la media nacional, con un soporte del 14.5 % y una confianza del 67 %. El 17.8 % del número total de estudiantes analizados que están bajo la media cumple este patrón.

Regla 2. Si el estudiante es de un colegio oficial, no reprobó grados y el colegio está ubicado en la región de Antioquia, entonces su desempeño en matemáticas probablemente está bajo la media nacional, con un soporte del 7.5 % y una confianza del 60.2 %. El 8.3 % del número total de estudiantes analizados que están bajo la media cumple este patrón.

Regla 3. Si el estudiante es de un colegio oficial y reprobó grados, entonces su desempeño en matemáticas probablemente está bajo la media nacional, con un soporte del 22.4 % y una confianza del 70.1 %. El 28.9 % del número total de estudiantes analizados que están bajo la media cumple este patrón.

Regla 4. Si el estudiante es de un colegio no oficial, entonces su desempeño en matemáticas probablemente está sobre la media nacional, con un soporte del 20.2 % y una confianza del 69.3 %. El 30.6 % del número total de estudiantes analizados que están sobre la media cumple este patrón.

CONCLUSIONES

Este estudio presenta los resultados obtenidos al aplicar la técnica de minería de datos y clasificación por árboles de decisión con el fin de detectar patrones de desempeño académico en la competencia de matemáticas de los estudiantes de Básica Primaria que presentaron las pruebas Saber 5º en el año 2017.

Es importante destacar que la naturaleza del colegio es un factor determinante para el desempeño académico en la competencia analizada. Para los estudiantes que son de colegios privados o no oficiales se muestra un buen desempeño, con altos porcentajes tanto de soporte como de confianza. Por otra parte, para los estudiantes de colegios oficiales, el desempeño no es bueno, si bien se tienen en cuenta otros factores, *e.g.*, si el estudiante reprobó grados o no y la zona de ubicación del colegio.

Estos hechos deben tenerse en cuenta para que las instituciones gubernamentales relacionadas con la educación Básica Primaria, como el MEN, las secretarías de educación departamentales y municipales y las instituciones educativas, tomen medidas para mejorar la calidad de la educación, especialmente en los colegios oficiales.

Se plantean como trabajos futuros 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 de lenguaje y matemáticas de las pruebas Saber 5º; y analizar la manera en la que se pueden agrupar estos estudiantes de acuerdo con su similitud en el rendimiento en estas pruebas.

AGRADECIMIENTOS

Al sistema de investigaciones de la Universidad de Nariño por financiar esta investigación.

CONTRIBUCIÓN DE AUTORÍA

Ricardo Timarán-Pereira: investigación, curación de datos, análisis formal, escritura-edición y revisión.
Javier Caicedo-Zambrano: investigación, conceptualización, escritura-edición y revisión.
Arsenio Hidalgo-Troya: investigación, análisis formal, escritura-borrador original.

REFERENCIAS

Azevedo, A., Santos, M. (2008). KDD, SEMMA and CRISP-DM: A parallel overview. En *Proceedings of IADIS European Conference on Data Mining* (pp. 182-185).

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

Gutiérrez, Y. (2015). *Relación entre la estructura familiar y el rendimiento académico en el área de matemáticas*. Instituto Latinoamericano de Altos Estudios.

Hall, M., Frank E., Witten, I. (2011). *Practical data mining: Tutorials*. University of Waikato. https://www.micai.org/2012/tutorials/Weka%20tutorials%20Spanish.pdf

Han, J., Kamber, M., Pei, J. (2011). *Data mining: Concepts and techniques* (3ra ed.). Morgan Kaufmann.

Hernández, J., Ramírez, M., Ferri, C. (2005). *Introducción a la minería de datos*. Editorial Pearson Prentice Hall.

ICFES. (2009). *Lineamientos generales Saber 5° y 9°*. Instituto Colombiano para la Evaluación de la Educación.

ICFES. (2011). *Informe técnico Saber 5° y 9°*. Instituto Colombiano para la Evaluación de la Educación.

ICFES. (2014). *Pruebas Saber 3°, 5° y 9°: lineamientos para las aplicaciones muestral y censal*. Instituto Colombiano para la Evaluación de la Educación.

ICFES. (2017). *Saber 5°: guía de orientación*. Instituto Colombiano para la Evaluación de la Educación.

Martín, S. (2015). *Pruebas Saber de lenguaje 3° y 5°: posibilidades y retos desde la perspectiva de la evaluación formativa*. Universidad Pedagógica Nacional.

OECD. (2016). *Education in Colombia*. http://www.oecd.org/edu/school/Education-in-Colombia-Highlights.pdf

Posada, J., 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.

Quinlan, J. R. (1993). *C4. 5: Programs for machine learning* (vol. 1). Morgan Kaufmann.

Quinlan, J. R. (1996). Improved use of continuous attributes in C 4.5. *Journal of Artificial Intelligence Research, 4*, 77-90.

Sattler, K., Dunemann, O. (2011). SQL database primitives for decision tree classifiers. En *Proceedings of the Tenth International Conference on Information and Knowledge Management* (pp. 379-386). http://dl.acm.org/citation.cfm?id=502650

Timarán, R., Millán, M. (2006). New algebraic operators and SQL primitives for mining classification rules. En *Computational Intelligence* (pp. 61–65). http://www.actapress.com/PaperInfo.aspx?PaperID=29048&reason=500

Timarán, R., Calderón, A., Jiménez, J. (2013). Aplicación de la minería de datos en la extracción de perfiles de deserción estudiantil. *Ventana Informática, 28*, 31-47. https://doi.org/10.30554/ventanainform.28.181.2013

Timarán, R., Jiménez, J., Calderón, A. (2017). *Detección de patrones de deserción estudiantil con minería de datos*. San Juan de Pasto, Colombia. Editorial Universidad de Nariño. https://editorial.udenar.edu.co/?p=2383.

Timarán, R., Caicedo, J., Hidalgo, A. (2021a). *Aplicación de la minería de datos en la detección de patrones de desempeño académico de las pruebas Saber Pro*. Editorial Universidad de Nariño.

Timarán, R., Caicedo, J., Hidalgo, A. (2021b). *Minería de datos educativa para el descubrimiento de factores asociados al desempeño académico en las Pruebas Saber 11º*. Editorial Universidad de Nariño.

Torres, J., Pachajoa, L., Pantoja, R. (2014). Resultados de las Pruebas Saber en el grado quinto del área de las ciencias naturales en tres instituciones educativas oficiales del municipio de Pasto. *Revista Fedumar Pedagogía y Educación, 1*(1), 55-69.

Valero, S., Vargas, A. S., 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*. http://fcaenlinea.unam.mx/anexos/1566/1566_u6_act1b.pdf

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

Witten, I., Frank, E., Hall, M. (2011). *Data mining: Practical machine learning tools and techniques* (3ra ed.). Morgan Kaufmann.