INVESTIGACIÓN e INNOVACIÓN en INGENERÍAS
ISSN2344-8652

Factores asociados al desempeño académico en Lectura Crítica en las pruebas Saber 11o con árboles de decisión

Factors associated with academic performance in Critical Reading on Saber 11th tests with decision trees

Ricardo Timarán Pereira **Arsenio Hidalgo Troya** **Javier Caicedo Zambrano**
Universidad de Nariño, Colombia

OPEN ACCESS

Recibido: 22/09/2020
Aceptado: 22/10/2020
Publicado: 23/11/2020

Correspondencia de autores:
ritimar@udenar.edu.co

Resumen

Objetivo: Descubrir patrones relacionados con el desempeño en la prueba de Lectura Crítica del examen Saber 11o, de los estudiantes de grado once de bachillerato que presentaron este examen en los años 2015 y 2016, utilizando la técnica de árboles de decisión. **Metodología**: Se utilizó CRISP-DM como metodología. A partir de los datos disponibles en el servicio ftp del ICFES sobre las pruebas Saber 11º, se construyó un conjunto de datos, el cual fue preprocesado y transformado. Se seleccionó como atributo clase, el resultado obtenido por los estudiantes en la prueba de lectura crítica. Se obtuvo un modelo de clasificación basado en árboles de decisión, utilizando la herramienta de minería de datos Weka. **Resultados:** Entre los factores relacionados con el rendimiento en Lectura Crítica y que son parte integrante de los patrones descubiertos están: la edad, el estrato, la jornada de estudios y las condiciones TICs del estudiante. **Conclusiones:** Un buen desempeño académico en la prueba de lectura crítica está asociado a estudiantes de estratos socioeconómicos altos, a estudiantes que asisten a colegios con jornada completa o única, así como también a estudiantes con condiciones TIC buenas.

Palabras clave: Árboles de decisión, desempeño académico, factores asociados, Lectura Crítica, examen Saber 11o.

Abstract

Objective: To discover patterns related with the performance of high school students who presented the Critical Reading test on Saber 11 exam between 2015 and 2016, using the classification technique based on decision trees. **Methodology:** The CRISP-DM was used as methodology. From the data available in the ICFES ftp services about the results of the Saber 11º tests, a data set was built. This data set was preprocessed and transformed. The result obtained by the students in the Critical Reading test was selected as target attribute. A classification model based on decision trees was obtained, using the Weka data mining tool. **Results:** Among the factors related with performance in Critical Reading and that are an integral part of the patterns discovered are: the age, the stratum, study day and ICT conditions of the student. **Conclusions:** A good academic performance on the critical reading test is associated with students from high socioeconomic strata, students who attend full-time or single-day schools, as well as students with good ICT conditions.

Keywords: Decision trees, academic performance, associated factors, Critical Reading, Saber 11 exam.

Como citar (IEEE): R. Timarán-Pereira., A. Hidalgo-Troya., J. Calcedo-Zambrano. "Factores asociados al desempeño académico en Lectura Crítica en las pruebas Saber 11o con árboles de decisión" vol. 8, n°3, pp. 29-37, 2020. DOI: https://doi.org/10.17081/invinno.8.3.4701

*Artículos resultados del evento XV Jornadas iberoamericanas de Ingeniería de Software e Ingeniería del Conocimiento - JIISIC 2020

Introducción

La misión del Instituto Colombiano para Evaluación de la Educación-ICFES es la de "evaluar, mediante exámenes externos estandarizados, la formación que se ofrece en el servicio educativo en los distintos niveles" [1]. Actualmente el ICFES evalúa la educación básica y media mediante las pruebas Saber 3°, Saber 5°, Saber 9° y Saber 11° [2].

El examen de estado Saber 11º lo presentan los estudiantes que están en el último grado de educación media, para cumplir con una de las condiciones que se exigen para ingresar a las universidades colombianas [2]. Además, el Ministerio de Educación Nacional - MEN los utiliza para hacer un seguimiento y evaluar la calidad de la educación que se imparte en los colegios de bachillerato del país [2].

Según el MEN [3], los cinco módulos de evalúa el examen Saber 11° ("lectura crítica, matemáticas, sociales y ciudadanas, inglés y ciencias naturales"), se basan en las habilidades que deben tener los estudiantes de bachillerato, de acuerdo a los estándares básicos de competencias EBC.

En la "Guía de orientación Saber 11º" [1], la prueba de lectura crítica "evalúa las competencias necesarias para comprender, interpretar y evaluar textos que pueden encontrarse en la vida cotidiana y en ámbitos académicos no especializados. Se espera que los estudiantes que culminan la educación media cuenten con una comprensión lectora que les permita tomar posturas críticas frente a diferentes tipos de texto".

Los estudios de factores asociados tienen como objetivo identificar las variables que más influencian el rendimiento escolar de los estudiantes, con el fin de avanzar en la construcción y entendimiento de los aspectos que inciden en la calidad de la educación.

En este sentido, la mayoría de estudios que se han realizado en Colombia [4,5,6,7,8,9,10], para determinar el rendimiento académico en las pruebas Saber 11° han utilizado técnicas estadísticas tradicionales, donde se utilizan conjuntos de datos pequeños o muestras representativas, por la imposibilidad de manejar grandes volúmenes de datos. El problema de estas técnicas es que pueden dejar sin considerar información valiosa y que solo es posible analizar con técnicas de minería de datos. De acuerdo a Timarán et al. [11], "la estadística plantea hipótesis que deben ser validadas a partir de los datos disponibles y la minería de datos descubre patrones no previstos desde la estadística. La minería de datos no se puede utilizar para confirmar o rechazar hipótesis, su objetivo es explorar datos, darles sentido, convertir en conocimiento un volumen de datos que por sí solos no aportan a la toma de decisiones". En este contexto, la minería de datos emerge como el siguiente paso evolutivo en el proceso de análisis de datos.

Para los autores Pérez y Santin [12], "la minería de datos es el proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias utilizando diferentes tareas a partir de grandes volúmenes de datos". Estas tareas tienen como objetivo descubrir patrones, perfiles y tendencias a través del análisis de los datos utilizando técnicas como clasificación, agrupamiento, patrones secuenciales y asociaciones, entre otras.

Metodología

Para la detección de patrones de desempeño académico en la prueba de Lectura Crítica del examen Saber 11º, se utilizó la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) [13,14], por ser según Azevedo & Santos [15] un método probado para proyectos de minería de datos, y según Hernández et al. [16], la metodología más completa y ampliamente utilizada para este tipo de proyectos. Según Chapman et al [13], "esta metodología se compone de seis etapas: análisis del problema, análisis de los datos, preparación de los datos, modelado, evaluación e implementación"

En la etapa de análisis del problema se apropió el conocimiento acerca del examen Saber 11º, y específicamente sobre la prueba de Lectura Crítica. Se identificaron los datos socioeconómicos, académicos e institucionales que a la fecha de esta investigación se encontraban disponibles en las bases de datos del ICFES, lo que permitió seleccionar la información adecuada para obtener los patrones encontrados.

En la etapa de análisis de los datos se construyó un conjunto de datos inicial, denominado T11361495A49 conformado por 1.361.495 filas y 49 columnas Se analizó la calidad de datos de este conjunto, con el fin de determinar por cada columna el número de valores nulos, distintos, valores máximo y mínimo, valor promedio, moda y desviación estándar para tener un primer acercamiento con los datos. Además, por medio del coeficiente de correlación de Pearson, se estableció la dependencia lineal entre la competencia de Lectura Crítica con el resto de competencias y con el puntaje global del examen Saber 11º. Estas correlaciones se muestran en la Tabla 1.

Tabla 1. Correlación de lectura crítica con el resto de competencias del Saber 11°.

Competencias	Lectura Crítica	Inglés	Ciencias Naturales	Matemáticas	Ciudadanas	Global
Lectura Crítica	1	0,629	0,756	0,733	0,772	0,885

Fuente: Elaboración propia

De acuerdo a la Tabla 1, Lectura Crítica presenta correlaciones altas con ciencias naturales, matemáticas y competencias ciudadanas y muy altas con el puntaje Global.

Con base en la etapa anterior, en la etapa de preparación de los datos se aplicaron técnicas de limpieza y transformación de datos con el fin de eliminar columnas con un alto porcentaje de valores nulos o valores constantes, reemplazar otros con valores como la media, la mediana o la moda, crear nuevas columnas que tengan mayor ganancia de información que las que las originan, discretizar valores continuos para disminuir el número de valores distintos y generalizar otros. Al final, resultó el conjunto de datos T11061680A16, con 1.061.680 filas y 16 columnas, limpio y transformado para continuar con la etapa de modelado.

Teniendo en cuenta las recomendaciones de [17,18,19,20], en la fase de modelado se construyó un clasificador basado en árboles de decisión utilizando la herramienta Weka ver 3.9.4 [21], donde se escogió el algoritmo J48 para la obtención del árbol, el cual es una versión del algoritmo C 4.5 de Quinlan [22]. Weka es una herramienta de Machine Learning de código abierto desarrollada en la Universidad de Waikato (Nueva Zelandia) muy utilizada en proyectos de minería de datos [23]. Además, se utilizó el método de validación cruzada con 10 pliegues para el proceso de entrenamiento y prueba del árbol, por ser este el método que mejor resultados da según [17]. Por otra parte, se establecieron los parámetros de prepoda con el nivel de confianza C y el nivel de soporte M. El nivel de confianza permite modificar el tamaño y capacidad de predicción del árbol. Su valor por defecto es del 25%. El nivel de soporte determina el número de instancias mínimas que deben llegar a cada hoja del árbol para ser considerada.

En la etapa de evaluación se constató que el modelo se ajuste a los parámetros de exactitud establecidos en el proyecto. Se determinó que la exactitud del modelo, que es el porcentaje de predicciones correctas, esté por encima del 65% de instancias correctamente clasificadas. Para ello se utilizó la matriz de confusión (Confusion Matrix) que es una tabla que permite observar fácilmente que tipos de aciertos y errores tiene el modelo que se está entrenando. Los valores en la diagonal son los aciertos y se los conoce, en el caso de que el atributo clase tenga dos valores, como verdaderos positivos (VP) y verdaderos negativos (VN) y el resto son los errores de clasificación conocidos como falsos positivos (FP) y falsos negativos (FN) (instancias que pertenecen a una clase y fueron clasificados incorrectamente en otra) [24].

Se calcularon los parámetros de sensibilidad y especificidad del modelo. La sensibilidad es la fracción de casos positivos clasificados correctamente, en cambio la especificidad, es la fracción de casos negativos clasificados correctamente.

Una vez cumplido la exactitud del modelo, se procedió a evaluar las reglas obtenidas en el árbol teniendo en cuenta soporte y una confianza mínima como parámetros de pospoda. Se eliminaron las reglas que no cumplían estos parámetros y se procedió a interpretar los reglas en términos que sean entendibles para el usuario.

Finalmente, en la etapa de implementación, se socializaron y documentaron los resultados obtenidos en esta investigación con el fin de que el conocimiento obtenido pueda integrarse al existente y sirva de soporte a la toma de decisiones del MEN, para mejorar la calidad de la educación en Colombia.

Resultados

Para efectos de escoger el mejor árbol que permita identificar claramente los patrones de rendimiento académico en lectura crítica, se construyeron diferentes modelos de árboles variando el nivel de confianza desde un 25% hasta un 5% y combinándolos con diferentes valores de niveles de soporte de 5000, 10.000 y 20.000 ejemplos por nodo, correspondientes al 0,05%, 1% y 2% respectivamente con respecto al total de instancias clasificadas.

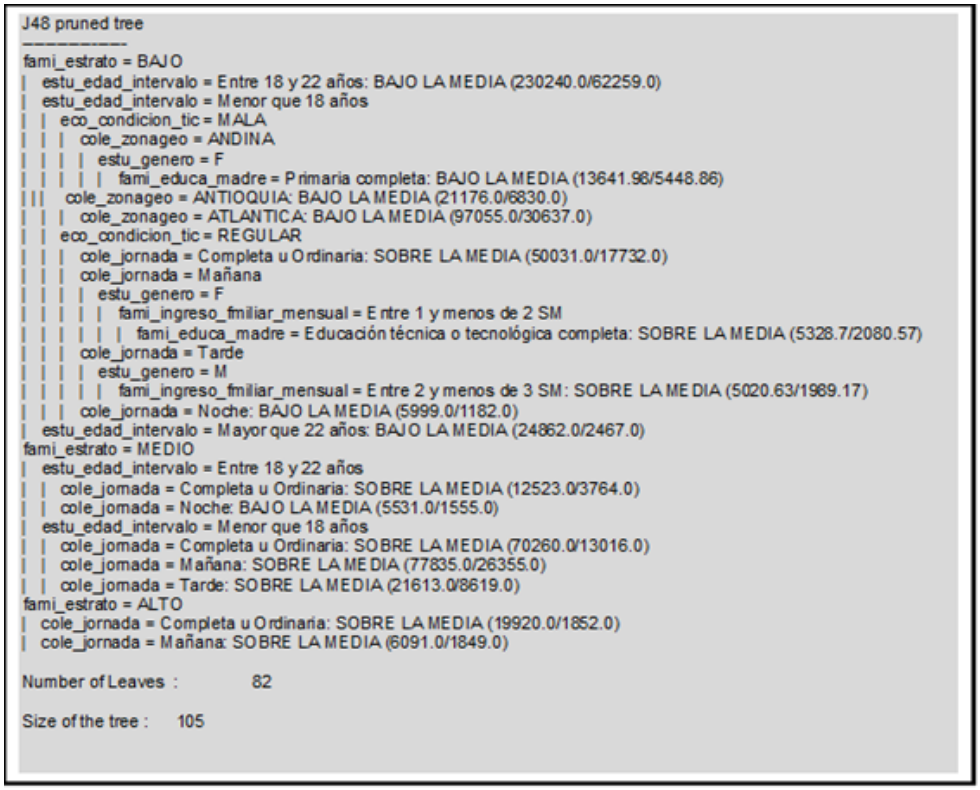
Analizando todos los modelos de árboles obtenidos, se escogió el árbol construido con los niveles C=5% y M=5000 por su mayor precisión y por la facilidad de interpretar los patrones. A este árbol se le aplicó un proceso de pospoda con el fin de dejar las ramas más representativas que cumplan un mínimo soporte del 0,05% con respecto al total de ejemplos y una confianza mínima del 65% con respecto a las instancias correctamente clasificadas en cada nodo. La evaluación del mejor árbol a través de la matriz de confusión se muestra en la Figura 1. En la figura 2 se muestra el mejor modelo de árbol obtenido con la herramienta Weka, con los parámetros antes mencionados.

Figura 1. Evaluación del mejor modelo de árbol

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

Correctly Classified Instances      703241               66.2385 %
Incorrectly Classified Instances    358439               33.7615 %
Kappa statistic                          0.3209
Mean absolute error                      0.4284
Root mean squared error                  0.4629
Relative absolute error                 86.0331 %
Root relative squared error             92.7592 %
Total Number of Instances          1061680

=== Confusion Matrix ===

     a      b   <-- classified as
 392460 171984 |     a = BAJO LA MEDIA
 186455 310781 |     b = SOBRE LA MEDIA
```

Fuente: Elaboración propia

Figura 2. Mejor árbol podado



```
J48 pruned tree
------------------
fami_estrato = BAJO
|   estu_edad_intervalo = Entre 18 y 22 años: BAJO LA MEDIA (230240.0/62259.0)
|   estu_edad_intervalo = Menor que 18 años
|   |   eco_condicion_tic = MALA
|   |   |   cole_zonageo = ANDINA
|   |   |   |   estu_genero = F
|   |   |   |   |   fami_educa_madre = Primaria completa: BAJO LA MEDIA (13641.98/5448.86)
|   |   cole_zonageo = ANTIOQUIA: BAJO LA MEDIA (21176.0/6830.0)
|   |   |   cole_zonageo = ATLANTICA: BAJO LA MEDIA (97055.0/30637.0)
|   |   eco_condicion_tic = REGULAR
|   |   |   cole_jornada = Completa u Ordinaria: SOBRE LA MEDIA (50031.0/17732.0)
|   |   |   cole_jornada = Mañana
|   |   |   |   estu_genero = F
|   |   |   |   |   fami_ingreso_fmiliar_mensual = Entre 1 y menos de 2 SM
|   |   |   |   |   |   fami_educa_madre = Educación técnica o tecnológica completa: SOBRE LA MEDIA (5328.7/2080.57)
|   |   |   cole_jornada = Tarde
|   |   |   |   estu_genero = M
|   |   |   |   |   fami_ingreso_fmiliar_mensual = Entre 2 y menos de 3 SM: SOBRE LA MEDIA (5020.63/1989.17)
|   |   |   cole_jornada = Noche: BAJO LA MEDIA (5999.0/1182.0)
|   estu_edad_intervalo = Mayor que 22 años: BAJO LA MEDIA (24862.0/2467.0)
fami_estrato = MEDIO
|   estu_edad_intervalo = Entre 18 y 22 años
|   |   cole_jornada = Completa u Ordinaria: SOBRE LA MEDIA (12523.0/3764.0)
|   |   cole_jornada = Noche: BAJO LA MEDIA (5531.0/1555.0)
|   estu_edad_intervalo = Menor que 18 años
|   |   cole_jornada = Completa u Ordinaria: SOBRE LA MEDIA (70260.0/13016.0)
|   |   cole_jornada = Mañana: SOBRE LA MEDIA (77835.0/26355.0)
|   |   cole_jornada = Tarde: SOBRE LA MEDIA (21613.0/8619.0)
fami_estrato = ALTO
|   cole_jornada = Completa u Ordinaria: SOBRE LA MEDIA (19920.0/1852.0)
|   cole_jornada = Mañana: SOBRE LA MEDIA (6091.0/1849.0)

Number of Leaves  :         82

Size of the tree :      105
```

Fuente: Elaboración propia

De acuerdo a la Figura 1, el modelo clasifica correctamente al 66% del total de estudiantes del conjunto de datos T1061680A16 que presentaron la prueba de lectura crítica. Esto significa que 703.241 estudiantes, están correctamente clasificados. La precisión del modelo por lo tanto es de 0,66. Por otra parte, el modelo clasifica incorrectamente al 34% del total de registros del conjunto de datos T1061680A16, lo que implica que 358.439 estudiantes están incorrectamente clasificadas.

Según la matriz de confusión (Figura 1), los verdaderos positivos corresponden en el modelo a 392.460 estudiantes que, en la prueba de lectura crítica, su desempeño académico está bajo la media, lo que significa que el 70% del total de estudiantes que están bajo la media (564.444) están correctamente clasificados. Esta es la precisión del modelo, que mide el porcentaje de casos positivos detectados. Así mismo, los verdaderos negativos son 310.781, que corresponden a estudiantes que en la prueba de lectura crítica están sobre la media. Esto significa que el 63% del total de estudiantes que están sobre la media (497.236) están correctamente clasificados. De igual manera, los falsos positivos son 171.984 que el modelo los clasifica incorrectamente sobre la media, pero que realmente son estudiantes que están bajo la media. A este grupo pertenece el 30% del total de estudiantes que están bajo la media. Igualmente, los falsos negativos son 186.455 que el modelo los clasifica incorrectamente bajo la media, pero que realmente son estudiantes que están sobre la media. A este grupo pertenece el 37% del total de estudiantes que están sobre la media.

En cuanto a la sensibilidad de modelo (recall), este identifica correctamente al 67% del total de los verdaderos positivos que el modelo clasifica como por debajo de la media. En cuanto a la especificidad del modelo, este identifica correctamente al 35% del total de verdaderos negativos que el modelo clasifica como por encima de la media.

En la Tabla 2 se interpretan los patrones en forma de reglas que el modelo de clasificación basado en árboles de decisión generó con la herramienta Weka (Figura 2) y que cumplen con un soporte mínimo del 0,05% y una confianza mínima de 65%.

Tabla 2. Reglas que cumplen con el soporte y la confianza mínima en lectura crítica

No. Regla	Antecedente	Consecuente	Soporte Mínimo	Confianza Mínima
1	"Estrato= bajo & 18 <=edad <=22 años"	"Bajo la media nacional"	21.7%	73%
2	"Estrato=bajo & edad< 18 años & índice TIC=malo & zona colegio = ATLANTICA"	"Bajo la media nacional"	9,1%	68.4%
3	"Estrato= bajo & edad <18 años & índice TIC=regular & jornada=nocturna"	"Bajo la media nacional"	0,6%	80.3%
4	"Estrato=medio & 18 <=edad <=22 años & jornada=completa"	"Sobre la media nacional"	1,2%	69,9%
5	"Estrato=medio & 18 <=edad <=22 años & jornada =nocturna"	"Bajo la media nacional"	0,5%	71,9%
6	"Estrato=medio & edad< 18 años & jornada=completa"	"Sobre la media nacional"	6.6%	81.5%
7	"Estrato=medio & edad< 18 años & jornada=mañana"	"Sobre la media nacional"	7.3%	66,1%
8	"Estrato= alto & jornada=completa"	"Sobre la media nacional"	1,9%	90,7%
9	"Estrato=alto & jornada=mañana"	"Sobre la media nacional"	0,6%	69,6%

Fuente: Elaboración propia

De acuerdo con la Tabla 2, entre los factores asociados al desempeño estudiantil en lectura crítica y que se encuentran entre las reglas encontradas están: el estrato socioeconómico, la edad, la jornada de estudio y el índice TIC del estudiante.

En cuanto al estrato socioeconómico del estudiante, los resultados de esta investigación en el desempeño académico en la prueba de lectura crítica muestran que los estratos altos están asociados al buen desempeño en esta prueba y en cambio los estratos bajos se asocian con un bajo desempeño. Estos datos coinciden con algunos resultados obtenidos en la investigación realizada por [25] sobre el desempeño académico de los estudiantes con respecto al puntaje general obtenido en el examen Saber 11º. Además, estos resultados son similares con los trabajos de [26,27,28], en los cuales se afirman que" existe una asociación importante entre el nivel socioeconómico del estudiante y su desempeño académico".

En cuanto la edad del estudiante, los resultados de esta investigación en el desempeño académico en la prueba de lectura crítica muestran que los estudiantes cuya edad es menor que 18 años generalmente tienen mejor desempeño que los mayores que esta edad. Estos resultados coinciden con lo que Tejedor [29] señala que la edad es un factor explicativo del rendimiento académico, mencionando que dentro de un mismo curso aquellos alumnos que son más jóvenes obtienen un mejor promedio.

En cuanto a la jornada de estudio del colegio al cual asiste el estudiante, los resultados de esta investigación en el desempeño académico en la prueba de lectura crítica muestran que los estudiantes que asisten a colegios con jornada única o completa tienen mejor desempeño que aquellos estudiantes que asisten a colegios con otras jornadas. Esta tendencia se mantiene también en la investigación de [25] e igualmente en el estudio realizado por [8] y por [30].

En cuanto al índice de condición TIC, que determina la existencia del servicio de internet, computador y telefonía en casa del estudiante, los resultados de esta investigación en el desempeño académico en la prueba de lectura crítica muestran que los estudiantes cuyo índice es malo o regular tienen un desempeño bajo la media nacional.

Hecho que se corrobora con el estudios, en el que se concluye que "la tenencia de tecnologías y el uso de éstas en el aprendizaje escolar mediante actividades contenido digital, afectan positivamente el desempeño académico de los estudiantes de educación básica y media".

Conclusiones

En este estudio se abordó el problema de determinar cuáles son los factores relacionados con el desempeño académico de los estudiantes que presentaron la prueba de Lectura Crítica dentro del examen Saber 11 con un enfoque diferente a los trabajos que se han realizado hasta el momento sobre estas pruebas. Se aplicó la técnica de minería de datos clasificación con árboles de decisión y se descubrieron patrones que predicen cuales son los factores que están relacionados con el buen o mal rendimiento académico de los estudiantes que presentaron la prueba de lectura crítica del Saber 11º.

El estrato al cual pertenece el estudiante, la jornada de estudio del colegio en el que el estudiante está matriculado, los elementos tecnológicos (tics) con los que se apoya el estudiante en casa y finalmente la edad del estudiante al momento de presentar el examen, son los factores que el modelo asocia, de manera automática, con el desempeño académico en la prueba de Lectura Crítica, a partir del conjunto de datos T1061680A16. Como trabajo futuro se plantea utilizar los conjuntos de datos construidos en esta investigación con otras técnicas predictivas de minería de datos con el fin de comparar la exactitud de los modelos y aplicar la mejor. Por otra parte, aplicar técnicas descriptivas de minería de datos como Reglas de Asociación y Clustering y comparar los resultados obtenidos.

Referencias bibliográficas

1. Icfes, "Guía de orientación Saber 11º", Bogotá, Colombia: Publicación del Instituto Colombiano para la Evaluación de la Educación (icfes), 2019.
2. Icfes, "Lineamientos generales para la presentación del examen de Estado Saber 11°", Bogotá, Colombia: Publicación del Instituto Colombiano para la Evaluación de la Educación, 2018.ISBN: 978-958-11-0680-6.
3. 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", Bogotá, Colombia: Ministerio de Educación Nacional, 2006.
4. F. Rodríguez, H. Benavides and A. Riascos. Predicción del desempeño académico usando técnicas de aprendizaje de máquinas. 2019. Disponible en: https://www.icfes.gov.co/documents/20143/234129/Prediccion+desempeno+academico+usando+un+enfoque+de+mineria+de+datos.pdf/0e5d0f1d-20ac-dffc-f3f1-88ccfde6b0bc.
5. A. Gaviria and J. Barrientos, Calidad de la educación y rendimiento académico en Bogotá, Revista Coyuntura Social, núm. 24, jun, 2001.

6. J. Barrientos, Calidad de la educación pública y logro académico en Medellín 2004-2006: Una aproximación por regresión intercuartil. Revista Lecturas de Economía, núm. 68, pp. 121-144, ene-jun,2008. DOI: https://doi.org/10.17533/udea.le.n68.
7. J.J. Correa, Determinantes del Rendimiento Educativo de los Estudiantes de Secundaria en Cali: un análisis multinivel, Revista Sociedad y Economía, núm. 6. pp. 81-105, abr,2004.
8. S. Chica, D. Galvis and A. Ramírez,. Determinantes del rendimiento académico en Colombia: pruebas ICFES Saber 11º. Revista Universidad EAFIT, vol. 46, núm. 160, 2010.
9. J. Gómez, Análisis de las competencias en matemáticas y lenguaje de los bachilleres colombianos. Trabajo de grado, Facultad de Ciencias Administrativas y Económicas Economía y Negocios Internacionales. Universidad ICESI. Cali, Colombia, 2014. https://repository.icesi.edu.co/biblioteca_digital/bitstream/10906/77946/1/gomez_analisis_competencias_2014.pdf.
10. O. Hernández, Determinantes del Rendimiento Académico en la Educación Media de Cundinamarca. Trabajo de grado, Facultad de Economía, Escuela Colombiana de Ingeniería Julio Garavito. Bogotá D.C., Colombia, 2015. http://repositorio.escuelaing.edu.co/handle/001/349.
11. S.R. Timarán, I. Hernández, S.J. Caicedo, A. Hidalgo and J.C. Alvarado, Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional. Bogotá, Colombia: Ediciones Universidad Cooperativa de Colombia, 2016. DOI: https://dx.doi.org/10.16925/9789587600490.
12. C. Pérez, & D. Santín. *Data Mining: Soluciones con Enterprise Miner*. México: Editorial Alfaomega, 2007.
13. Chapman, P., Clinton, J., Kerber, R, Khabaza, T., Reinartz, T.,Shearer, C.and Wirth, R.(2000). *CRISP-DM 1.0: Step-by-step data mining guide* [en línea]. Disponible en: https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf.
14. Villena J. CRISP-DM: La metodología para poner orden en los proyectos de Data Science. (2016). Disponible en: https://data.sngular.team/es/art/25/crisp-dm-la-metodologia-para-poner-orden-en-los-proyectos-de-data-science.
15. A. Azevedo and M. Santos, "KDD, SEMMA and CRISP-DM: a parallel overview", presentado en IADIS European Conference on Data Mining, Amsterdam, Netherlands. pp. 182-185, 2008.
16. J. Hernández, M. Ramírez and C. Ferri, *Introducción a la Minería de Datos*. Madrid, España: Editorial Pearson Educación SA, 2005. http://dspace.ucbscz.edu.bo/dspace/handle/123456789/526.
17. J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. San Francisco, USA: Morgan Kaufmann Publishers; 2001. 550 p.
18. K. Sattler and O. Dunemann, "SQL Database Primitives for Decision Tree Classifiers", presentado en la 10th ACM International Conference on Information and Knowledge Management. Atlanta, USA: ACM New York. p. 379-86, 2001.
19. S.R. Timarán, I. Hernández, S.J. Caicedo, A. Hidalgo and J.C. Alvarado, Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional. Bogotá, Colombia: Ediciones Universidad Cooperativa de Colombia, 2016. DOI: https://dx.doi.org/10.16925/9789587600490.
20. I. Witten, I., E. Frank and M. Hall, Data Mining: Practical Machine Learning Tools and Techniques (Third Edition). New York,USA:Morgan Kaufmann, 2011. ISBN:978-0-12-374856-0. DOI: https://doi.org/10.1016/C2009-0-19715-5.
21. J.R. Quinlan, J.R. (1993). C 4. 5: Programs for Machine Learning. San Francisco, USA: Morgan Kaufmann Publishers. 299 p., 1993.
22. M.García and A. Álvarez, Análisis de Datos en WEKA:Pruebas de Selectividad. 2010. http://www.it.uc3m.es/jvillena/irc/practicas/06-07/28.pdf.

23. G. Fernández, Extracción de Información de la Web usando Técnicas de Minería de Datos, 2009. http://www.tdg-seville.info/Download.ashx?id=48

24. R. Timarán., R., J. Caicedo and A. Hidalgo, A. (2019). Árboles de decisión para predecir factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas saber 11°. Revista de Investigación, Desarrollo e Innovación, 9 (2), pp. 363-378, ene-jun,2019. DOI: https://doi.org/10.19053/20278306.v9.n2.2019.9184.

25. G.M. Garbanzo, Factores asociados al rendimiento académico en estudiantes universitarios, una reflexión desde calidad de la educación superior pública. Revista Educación, vol. 31(1), pp. 43-63, 2007.

26. J. Seibold, La calidad integral en educación. Reflexiones sobre un nuevo concepto de calidad educativa que integre valores y equidad educativa. Revista Iberoamericana de Educación, vol. 23, may,2000. DOI: https://doi.org/10.35362/rie2301012.

27. E. Montero, J. Villalobos and A. Valverde, Factores institucionales, pedagógicos, psicosociales y sociodemográficos asociados al rendimiento académico en la Universidad de Costa Rica: un análisis multinivel, Revista RELIEVE, vol. 3(2), pp. 215-234, 2007. www.uv.es/RELIEVE/v13n2/RELIEVEv13n2_5.htm.

28. J. Tejedor. Poder explicativo de algunos determinantes del rendimiento eI. Ridao and J. Gil, La jornada escolar y el rendimiento de los alumnos, Revista de Educación, núm 327, pp.141-156, 2002. https://hdl.handle.net/11441/77859.

29. H.Botello and A. Guerrero, La influencia de las TIC en el desempeño académico de los estudiantes en América Latina: Evidencia de la prueba PISA. Memorias Virtual Educa, Lima, Perú, 2014. http://hdl.handle.net/20.500.12579/4050.