Identification of Factors Associated with Academic Performance in Mathematics in the Saber 11th Tests Applying Educational Data Mining

Ricardo Timarán Pereira, Ph.D.[1], Javier Caicedo Zambrano, Ph.D.[1], Arsenio Hidalgo Troya, Mg.[1]
[1]Universidad de Nariño, Colombia, ritimar@udenar.edu, jacaza1@gmail.com, arsenio.hidalgo@gmail.com

***Abstract* -- This paper presents one of results obtained in the research project that aimed to apply educational data mining to discover factors associated with the academic performance of Colombian High School students who presented the Saber 11th test between the years 2015 and 2016. Socio-economic, academic and institutional information of those students was selected from the ICFES databases. CRISP-DM was used as methodology. A data repository for data mining was built, cleaned and transformed. Patterns associated with the good or poor academic performance of students were discovered in the mathematics test, using a classification model based on decision trees. The knowledge discovered will be incorporated into the existing one and it can be integrated into the decision-making processes of the MEN, ICFES and others governmental and educational institutions that ensure the quality of education in Colombia.**

***Keywords*– Educational Data Mining, Decision Trees, Academic Performance, High School, Saber 11th Test.**

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2019.1.1.297
ISBN: 978-0-9993443-6-1 ISSN: 2414-6390

17th LACCEI International Multi-Conference for Engineering, Education, and Technology: "Industry, Innovation, And Infrastructure for Sustainable Cities and Communities", 24-26 July 2019, Jamaica.

Identificación de Factores Asociados al Desempeño Académico en Matemáticas en las Pruebas Saber 11° Aplicando Minería de Datos Educativa

Ricardo Timarán Pereira, Ph.D.[1], Javier Caicedo Zambrano, Ph.D.[1], Arsenio Hidalgo Troya, Mg.[1]
[1]Universidad de Nariño, Colombia, ritimar@udenar.edu, jacaza1@gmail.com, arsenio.hidalgo@gmail.com

Abstract*– This paper presents one of results obtained in the research project that aimed to apply educational data mining to discover factors associated with the academic performance of Colombian High School students who presented the Saber 11th test between the years 2015 and 2016. Socio-economic, academic and institutional information of those students was selected from the ICFES databases. CRISP-DM was used as methodology. A data repository for data mining was built, cleaned and transformed. Patterns associated with the good or poor academic performance of students were discovered in the mathematics test, using a classification model based on decision trees. The knowledge discovered will be incorporated into the existing one and it can be integrated into the decision-making processes of the MEN, ICFES and others governmental and educational institutions that ensure the quality of education in Colombia.

Keywords*—Educational Data Mining, Decision Trees, Academic Performance, High School, Saber 11th Test.

Resumen*– En este artículo se presenta uno de los resultados obtenidos en el proyecto de investigación cuyo objetivo fue aplicar la minería de datos educativa para descubrir factores asociados al desempeño académico de los estudiantes colombianos de educación media que presentaron las pruebas Saber 11° entre los años 2015 y 2016. 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 en la prueba de matemáticas, utilizando un modelo de clasificación basado en árboles de decisión. El conocimiento descubierto se incorporará al existente y se podrá integrar a los procesos de toma de decisiones del MEN, ICFES y de otras instituciones gubernamentales y educativas que velan por la calidad de la educación en Colombia.

Palabras Clave*— Minería de Datos Educativa, Árboles de Decisión, Desempeño Académico, Educación Media, Pruebas Saber 11°.

I. Introducción

La evaluación educativa, como rama de la ciencia educativa, contribuye de manera importante a mejorar la calidad de la educación. Se utiliza en cualquier actividad educativa y en todo tipo de actividades para transferir, motivar y adquirir conocimientos y habilidades. La evaluación educativa se aplica para conocer los logros de los alumnos y diagnosticar los resultados educativos como parte de un papel vital en la mejora de la calidad de la educación [1]. En efecto, una adecuada evaluación, que tome en consideración los avances de las ciencias de la cognición, de la pedagogía y de la administración, aporta elementos para una acertada toma de decisiones en los distintos ámbitos educativos tales como: los procesos de enseñanza-aprendizaje, la formulación de políticas, programas y proyectos, la asignación de recursos y el perfeccionamiento de los procesos curriculares, pedagógicos y de gestión [2].

En Colombia, el Instituto Colombiano para Evaluación de la Educación (ICFES) es el encargado de evaluar, mediante exámenes externos estandarizados, la formación que se ofrece en el servicio educativo en los distintos niveles [3]. Actualmente el ICFES diseña y aplica las pruebas Saber 3°, Saber 5°, Saber 9°, Saber 11°, con las cuales evalúa la Educación Básica y Media; y Saber Pro, con esta última se evalúa la Educación Superior.

El Examen de Estado de la educación media, Saber 11°, deben presentarlo estudiantes que se encuentren finalizando el grado undécimo, con el fin de obtener resultados oficiales para efectos de ingreso a la educación superior. También pueden presentarlo quienes ya hayan obtenido el título de bachiller o hayan superado el examen de validación del bachillerato, de conformidad con las disposiciones vigentes. En esta investigación únicamente se tuvo en cuenta los primeros y para este artículo, los resultados de la prueba de matemáticas. Según el Decreto 869 de 2010, los objetivos de esta prueba son: seleccionar estudiantes para la educación superior; monitorear la calidad de la formación que ofrecen los establecimientos de educación media; y producir información para la estimación del valor agregado de la educación superior [3].

Saber 11° evalúa cinco componentes basados en las aptitudes que deben desarrollar los educandos según los estándares básicos de competencias [4]: lectura crítica, sociales y ciudadanas, ciencias naturales, inglés y matemáticas. 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 [5]. La prueba de Sociales y Ciudadanas evalúa los conocimientos y competencias del estudiante que lo habilitan para analizar y comprender el mundo social desde la perspectiva propia de las ciencias sociales. Evalúa también su habilidad para establecer

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2019.1.1.297
ISBN: 978-0-9993443-6-1 ISSN: 2414-6390

relaciones entre distintos eventos y la capacidad de reflexionar y emitir juicios críticos sobre estos [5]. La prueba de Ciencias Naturales establece que la formación de niños, niñas y jóvenes debe propiciar el desarrollo de ciudadanos capaces de comprender que la ciencia tiene una dimensión universal, que es cambiante, y que permite explicar y predecir y además, que la ciencia es, ante todo, una construcción humana dinámica de tipo teórico y práctico y entender que, en la medida en que la sociedad y la ciencia se desarrollan, se establecen nuevas y diferentes relaciones entre la ciencia, la tecnología y la sociedad [5]. La prueba de inglés pretende dar cuenta de los niveles de desempeño propuestos por el Marco Común Europeo de Referencia para las Lenguas (aprendizaje, enseñanza y evaluación) del Consejo de Europa. Este marco contempla seis (6) niveles: A1, A2, B1, B2, C1, C2, entre los cuales el MEN propuso como meta para el año 2019 alcanzar el nivel B1 en la población de educación media [5]. Finalmente, la prueba de Matemática evalúa las competencias de los estudiantes para enfrentar situaciones que pueden resolverse con el uso de algunas herramientas matemáticas [5]. Tanto las competencias definidas para la prueba como los conocimientos matemáticos que el estudiante requiere para resolver las situaciones planteadas se contemplan en las definiciones de los Estándares Básicos de Competencias de Matemáticas del Ministerio de Educación Nacional (MEN). En esta prueba, se integran competencias y contenidos en distintas situaciones o contextos, en las cuales las herramientas matemáticas cobran sentido y son un importante recurso para la comprensión, la transformación, la justificación y la solución de los problemas involucrados [5].

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 [6]. 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.

Los estudios que se han realizado en Colombia hasta el momento, para determinar el rendimiento académico en las pruebas Saber 11° [7],[8],[9],[10],[11],[12] 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 se pueden descubrir utilizando un tratamiento más complejo de los datos, que es posible con la minería de datos. En este contexto, la minería de datos emerge como el siguiente paso evolutivo en el proceso de análisis de datos.

La minería de datos en la educación (del término en inglés EDM) 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 [13], [14], [15]. 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 [16]. 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 [17]. 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 [18], [19]. 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 [19].

En este artículo se presentan los resultados de aplicar la minería de datos educativa para descubrir factores asociados al desempeño académico, en la competencia de matemáticas, de los estudiantes colombianos de educación media que presentaron las pruebas Saber 11° entre los años 2015 y 2016.

II. 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 11°. Los datos más actualizados eran de los años 2015 y 2016. Para el descubrimiento de patrones asociados al desempeño académico en las pruebas Saber 11°, se construyó un modelo de clasificación basado en árboles de decisión, utilizando el algoritmo J48 de la herramienta Weka [20]. Se escogió este algoritmo por su simplicidad y facilidad para interpretar los patrones y por ser el más utilizado para este tipo de problemas [21], [22].

Para el descubrimiento de patrones, se aplicó la metodología CRISP-DM (*Cross Industry Standard Process for Data Mining*). Azevedo y Santos [23] 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 [23]. CRISP-DM es la guía de referencia más ampliamente utilizada en el desarrollo de proyectos de minería de datos [24] 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 [25]. En la fase de análisis 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 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 [25]. 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 [25].

III. Resultados

A. *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, en la competencia de matemáticas, de los estudiantes colombianos que encontrándose finalizando el grado undécimo de educación media, presentaron las pruebas Saber 11°, se convirtió en un problema a resolver con minería de datos.

B. *Comprensión de los datos*

En esta fase, se identificó, recopiló y familiarizó con la información socioeconómica, académica e institucional, disponible en las bases de datos del ICFES, correspondiente a los resultados de los estudiantes de educación media que presentaron las pruebas Saber 11° entre los años 2015 y 2016. Se construyó un repositorio inicial donde se integraron los repositorios de cada año, dando como resultado un repositorio compuesto por 1.361.495 registros y 49 atributos al cual se lo denominó T1361495A49, el cual sirvió de base para las subsiguientes fases.

C. *Preparación de los datos*

En esta fase se realizó inicialmente un análisis de la calidad de los datos del repositorio T1361495A49, con el fin de conocer por cada atributo el número de valores distintos, el número de valores nulos, el valor máximo, valor mínimo, moda, media y un histograma para determinar cuáles técnicas de limpieza de datos se debían aplicar.

Los 49 atributos del repositorio base, considerados por el ICFES como los más importantes para capturar la información de las pruebas Saber 11°, fueron depurados, teniendo en cuenta la calidad de los datos y las técnicas de minería de datos a aplicar; se limpiaron (eliminación de datos nulos y valores constantes) e integraron los datos, se generaron atributos adicionales a partir de los existentes por ganancia de información, se realizaron transformaciones o cambios de formato a los valores de los atributos que se consideraron necesarios, se eliminaron los atributos reemplazados, así como los registros de estudiantes que presentaron más de una vez las pruebas Saber 11°. Con el fin de facilitar la detección de patrones de rendimiento académico se discretizaron los valores numéricos de ciertos atributos teniendo en cuenta un rango de valores o que las frecuencias por cada valor sean proporcionales, para evitar sesgos, al construir los modelos de minería de datos.

Como resultado de esta fase se obtuvo un repositorio de datos limpio y transformado, con 1.061.680 registros y 16 atributos, listo para aplicarle las técnicas de minería de datos y al cual se le denominó T1061680A16. En la tabla I se muestra el diccionario de datos de este repositorio.

TABLA I
DICCIONARIO DE DATOS REPOSITORIO FINAL T1061680A16

No	Atributo	Descripción	Valores
		Socioeconómicos	
1	estu_genero	Sexo del estudiante	M ,F
2	estu_edad_intervalo	Rango de edad del estudiante en el momento de presentar la prueba	<18 Entre 18 y 22 >22
3	fami_estrato	Estrato socioeconómico del estudiante	BAJO, MEDIO, ALTO
4	fami_nivel_sisben	Nivel de clasificación en el SISBEN al que pertenece el estudiante	NIVELES 1, 2, 3, OTRO NIVEL, NO ESTA EN SISBEN
5	fami_ingreso_fmiliar_mensual	Ingresos mensuales familiares en salarios mínimos	Hasta 1,2,3,4,5,6,7,8,9,10 o mas 10 salarios
6	fami_educa_madre	Máximo nivel educativo de la madre	PRIMARIA, SECUNDARIA,TÉCNICO, TECNOLOGICO,

			PROFESIONAL, POSTGRADO, NINGUNO
7	fami_educa_padre	Máximo nivel educativo del padre	PRIMARIA, SECUNDARIA,TÉCNICO, TECNOLOGICO, PROFESIONAL, POSTGRADO, NINGUNO
8	fami_ocup_padre	Ocupación del padre	DIRECTIVO,EMPLEADO, EMPRESARIO, HOGAR, INDEPENDIENTE,OTRA PENSIONADO,PROFESIONAL
9	fami_ocup_madre	Ocupación de la madre	LOS MISMOS VALORES DE LA OCUPACIÓN DEL PADRE
10	econ_condicion_vivienda	Condición de la vivienda del estudiante	BUENA, MALA,REGULAR
11	eco_condicion_tic	Condición de uso de TIC en el hogar del estudiante	BUENA, REGULAR, MALA
12	eco_condicion_vive	Condición de vida del estudiante	SIN HACINAMIENTO, HACINAMIENTO MEDIO, HACINAMIENTO CRITICO
Académicos			
13	punt_matel_cuali	Puntaje en matemáticas obtenido por el estudiante en las pruebas Saber 11	POR ENCIMA DE LA MEDIA NACIONAL POR DEBAJO DE LA MEDIA NACIONAL
Institucionales			
14	Tipo_cole	Tipo de institución educativa	PÚBLICA,PRIVADA
15	Cole_jornada	Jornada de estudio del estudiante	MAÑANA, TARDE, NOCHE, ÚNICA, SABATINA-DOMINICAL
16	cole_zonageo	Zona geográfica donde se encuentra la institución educativa	ATLANTICA, AMAZONAS, ANDINA, ANTIOQUIA, PACIFICO, BOGOTA

D. Modelado

En esta fase se seleccionó la tarea de clasificación con árboles de decisión como la técnica de minería de datos más adecuada para solucionar el problema objeto de la investigación.

Con clasificación se pretende obtener un modelo que permita predecir para los nuevos casos de estudiantes, los factores socioeconómicos, académicos e institucionales asociados a un probable buen o mal desempeño académico en la competencia de matemáticas evaluada en las pruebas Saber 11°. El modelo de clasificación basado en árboles de decisión, es probablemente el más utilizado y popular por su simplicidad y facilidad para entender [21], [22], [26]. 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 [24]. Por estas razones en esta investigación se escogió este modelo.

Para la construcción del modelo de clasificación con árboles de decisión se utilizó la herramienta Weka [20] y su algoritmo J48, el cual implementa al algoritmo C.45 [27]. El algoritmo J48 se basa en la utilización del criterio de ganancia de información. 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 [24]. 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 [28]. 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 [28].

Antes de construir un 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 [28], se utilizó el método de validación cruzada (*Cross validation*) por ser la opción por defecto y la más comúnmente utilizada. Este mecanismo permite reducir la dependencia del resultado del experimento en el modo en el cual se realiza la partición [24]. 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. Comúnmente, se suelen utilizar 10 particiones (*10-fold cross validation*) [24].

Por otra parte, se evaluó o estimó el coste del clasificador para el repositorio T1061680A16 a través de la matriz de confusión. La matriz de confusión (*Confusion Matrix*) representa de forma detallada el número de instancias que son predichas por clase. La suma de los registros que se representan en cada fila i, i =1…n constituyen el número de instancias que realmente pertenecen a la clase i. Similarmente la sumatoria de los ejemplos o registros en cada columna j, j = 1…n son las instancias que ha predicho el algoritmo al valor j de la clase. Los valores en la diagonal son los aciertos y el

resto son los errores de clasificación (ejemplos que pertenecían a la clase i de la fila i y fueron clasificados incorrectamente en otra) [24].

Teniendo en cuenta los parámetros de evaluación anteriores, se procedió a construir los diferentes árboles de decisión con el algoritmo J48. Se escogió como clase el puntaje en matemáticas de cada estudiante obtenido en las pruebas Saber11, el cual fue discretizado en los valores "por encima de la media nacional" y "por debajo de la media nacional".

Con el fin de obtener diferentes modelos de árboles por competencia y reglas de clasificación generalizadas hasta reglas más detalladas, se establecieron 3 porcentajes de prepoda del árbol para el factor M igual a 0,05 %, 1% y 2% del total de registros del repositorio de datos, y 3 porcentajes para el factor confianza C igual a 5%, 10% y 25% y se construyeron los diferentes modelos combinando estos factores. Se escogió el árbol construido con los parámetros M=5000 (0,05%) y C=5% por los mejores resultados obtenidos y por la facilidad de análisis de los patrones. Una vez construido los árboles 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 0.05% y una confianza del 65%. En la figura 1 se muestra la precisión del árbol y su matriz de confusión. El árbol construido con los parámetros M=5000 y C=5% se muestra en la figura 2.

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

Correctly Classified Instances      709421               66.8206 %
Incorrectly Classified Instances    352259               33.1794 %
Kappa statistic                     0.3331
Mean absolute error                 0.4269
Root mean squared error             0.462
Relative absolute error            85.5402 %
Root relative squared error        92.4953 %
Total Number of Instances         1061680

=== Confusion Matrix ===

     a      b   <-- classified as
 313231 194138 |      a = SOBRE LA MEDIA
 158121 396190 |      b = BAJO LA MEDIA
```

Fig. 1. Precisión del modelo y su matriz de confusión

E. Evaluación

En esta fase se evaluaron los patrones descubiertos con el fin de determinar su validez, remover los patrones redundantes o irrelevantes y traducir los patrones útiles en términos que sean entendibles para el usuario.

La evaluación e interpretación de los patrones descubiertos se describe en la sección de discusión e interpretación de resultados.

```
fami_estrato = BAJO
|  estu_edad_intervalo = Entre 18 y 22 años: BAJO LA MEDIA
(230240.0/62546.0)
|  estu_edad_intervalo = Menor que 18 años
|  |  eco_condicion_tic = MALA
|  |  |  estu_genero = F: BAJO LA MEDIA (174650.24/64402.34)
|  |  |  estu_genero = M
|  |  |  |  cole_zonageo = ANDINA
|  |  |  |  |  cole_jornada = Completa u Ordinaria: SOBRE LA MEDIA
(12506.22/5145.72)
|  |  |  |  |  cole_jornada = Mañana: SOBRE LA MEDIA (21259.56/8833.03)
|  |  |  |  cole_zonageo = PACIFICA
|  |  |  |  |  fami_ingreso_fmiliar_mensual = Entre 1 y menos de 2 SM: SOBRE
LA MEDIA (7406.31/3489.47)
|  |  |  |  |  fami_ingreso_fmiliar_mensual = Menos de 1 SM: BAJO LA MEDIA
(10332.16/4839.66)
|  |  |  |  cole_zonageo = BOGOTA: SOBRE LA MEDIA (5217.25/2064.63)
|  |  |  |  cole_zonageo = ORINOQUIA: SOBRE LA MEDIA (5735.13/2216.66)
|  |  |  |  cole_zonageo = ANTIOQUIA: BAJO LA MEDIA (8262.31/3733.84)
|  |  |  |  cole_zonageo = ATLANTICA: BAJO LA MEDIA (39287.29/15354.75)
|  |  eco_condicion_tic = REGULAR
|  |  |  estu_genero = F
|  |  |  |  cole_jornada = Completa u Ordinaria: SOBRE LA MEDIA
(27679.27/10710.55)
|  |  |  |  cole_jornada = Mañana
|  |  |  |  |  fami_ingreso_fmiliar_mensual = Entre 1 y menos de 2 SM
|  |  |  |  |  |  fami_educa_madre = Primaria completa: BAJO LA MEDIA
(6463.93/2839.56)
|  |  |  |  |  |  fami_educa_madre = Secundaria completa
|  |  |  |  |  |  |  eco_condicion_vivienda = MALA: BAJO LA MEDIA
(8414.3/3774.12)
|  |  |  |  |  |  |  eco_condicion_vivienda = BUENA: SOBRE LA MEDIA
(11080.67/5456.61)
|  |  |  |  |  |  |  eco_condicion_vivienda = REGULAR: BAJO LA MEDIA
(766.12/363.12)
|  |  |  |  |  |  fami_educa_madre = Primaria incompleta: BAJO LA MEDIA
(7087.3/3425.37)
|  |  |  |  |  |  fami_educa_madre = Secundaria incompleta: BAJO LA MEDIA
(9422.05/4153.68)
|  |  |  |  |  |  fami_educa_madre = Educación técnica o tecnológica completa:
SOBRE LA MEDIA (5328.68/2132.0)
|  |  |  |  |  fami_ingreso_fmiliar_mensual = Menos de 1 SM: BAJO LA MEDIA
(12301.74/5090.68)
|  |  |  |  |  fami_ingreso_fmiliar_mensual = Entre 2 y menos de 3 SM: SOBRE
LA MEDIA (18660.41/8372.3)
|  |  |  |  cole_jornada = Tarde
|  |  |  |  |  fami_educa_madre = Secundaria completa: BAJO LA MEDIA
(10830.41/4802.74)
|  |  |  |  |  fami_educa_madre = Secundaria incompleta: BAJO LA MEDIA
(6128.3/2491.93)
|  |  |  estu_genero = M
|  |  |  |  cole_jornada = Completa u Ordinaria: SOBRE LA MEDIA
(22351.73/6292.45)
|  |  |  |  cole_jornada = Mañana: SOBRE LA MEDIA (67319.62/24772.38)
|  |  |  |  cole_jornada = Tarde: SOBRE LA MEDIA (22519.19/9126.28)
|  estu_edad_intervalo = Mayor que 22 años: BAJO LA MEDIA (24862.0/2552.0)
fami_estrato = MEDIO
|  estu_edad_intervalo = Entre 18 y 22 años
|  |  cole_jornada = Completa u Ordinaria: SOBRE LA MEDIA (12523.0/3803.0)
|  |  cole_jornada = Mañana
|  |  |  estu_genero = F: BAJO LA MEDIA (8755.7/3323.47)
|  |  |  estu_genero = M: SOBRE LA MEDIA (11121.3/4896.76)
|  |  cole_jornada = Tarde: BAJO LA MEDIA (7491.0/3392.0)
|  |  cole_jornada = Noche: BAJO LA MEDIA (5531.0/1612.0)
|  estu_edad_intervalo = Menor que 18 años
|  |  cole_jornada = Completa u Ordinaria: SOBRE LA MEDIA (70260.0/12602.0)
|  |  cole_jornada = Mañana: SOBRE LA MEDIA (77835.0/25356.0)
|  |  cole_jornada = Tarde: SOBRE LA MEDIA (21613.0/8444.0)
|  |  cole_jornada = Noche: BAJO LA MEDIA (2654.0/966.0)
fami_estrato = ALTO
|  cole_jornada = Completa u Ordinaria: SOBRE LA MEDIA (19920.0/1781.0)
|  cole_jornada = Mañana: SOBRE LA MEDIA (6091.0/1832.0)

Number of Leaves  :     73
Size of the tree :   92
```

Fig. 2. Mejor árbol obtenido con M=0.05% y C=5%

E. Implementación

En esta fase, a través de la difusión de los informes de esta investigación, el conocimiento descubierto se incorporará al existente y se podrá integrar a los procesos de toma de decisiones del MEN, ICFES y de otras instituciones gubernamentales y educativas que velan por la calidad de la educación media y superior en Colombia.

IV. Interpretación y Discusión de Resultados

Analizando los resultados obtenidos con el árbol de decisión construido con el conjunto de datos T1061680A16, en el cual se almacenan los datos válidos sobre los factores socioeconómicos, académicos e institucionales correspondientes a 1.061.680 estudiantes, quienes presentaron las prueba Saber 11° en los periodos 2015 y 2016 donde se escogió el atributo puntaje en la prueba de matemáticas (puntaje_mate_cuali) como clase, se puede observar que este clasifica correctamente a 709.421 instancias, que corresponde a un porcentaje de precisión del 67% y 352.259 instancias incorrectamente clasificadas, correspondiente a un porcentaje del 33% (ver figura 1).

Teniendo en cuenta la matriz de confusión (figura 1), del total de 1.061.680 estudiantes evaluados, el desempeño académico de 507.369 estudiantes en la prueba de matemáticas, esta sobre la media y el desempeño de 554.311 estudiantes se ubica bajo la media. El modelo clasifica correctamente a 313.231 casos de estudiantes cuyo desempeño en matemáticas está sobre la media, y a 396.190 casos que están bajo la media. Por otra parte, clasifica incorrectamente como bajo la media a 194.138 casos de estudiantes que están sobre la media y como sobre la media a 158.121 casos que están bajo la media. Esto significa que el modelo clasifica correctamente al 61,7 % de los estudiantes que están sobre la media en la prueba de matemáticas y al 71,5% de las estudiantes que están bajo la media.

Para efectos de la discusión de los resultados, se escogieron los patrones más representativos, teniendo en cuenta un mínimo soporte del 0,05% y una confianza mínima de 60%, tanto los que se ubican por encima de la media, como aquellos que se sitúan por debajo de ella. Entre los patrones más importantes están:

Regla 1. Si el estudiante es de estrato socioeconómico bajo y su edad está entre 18 y 22 años entonces su desempeño académico en la prueba de matemáticas de Saber 11° tiene mayor probabilidad de estar bajo la media nacional. El 21,7% del total de 1.061.680 estudiantes evaluados se clasifican de esta manera. El 72,8% de los 230.240 estudiantes que se clasifican así, están correctamente clasificados y el 30,3% de los 554.311 que están bajo la media, cumplen este patrón.

Regla 2. Si el estudiante es de estrato socioeconómico bajo, es menor que 18 años, el índice de condición tic es malo y es de sexo femenino entonces su desempeño académico en la prueba de matemáticas de Saber 11° tiene mayor probabilidad de estar bajo la media nacional. El 16,5% del total de 1.061.680 estudiantes evaluados se clasifican de esta manera. El 63,1% de los 174.650 estudiantes que se clasifican así, están correctamente clasificados y el 19,9% de los 566.068 que están bajo la media, cumplen este patrón.

Regla 3. Si el estudiante es de estrato socioeconómico bajo, es menor que 18 años, el índice de condición tic es malo, es de sexo masculino y el colegio se situa en la zona geográfica de Bogotá entonces su desempeño académico en en la prueba de matemáticas de Saber 11° tiene mayor probabilidad de estar sobre la media nacional. El 0,5% del total de 1.061.680 estudiantes evaluados se clasifican de esta manera. El 60,4% de los 5.217 estudiantes que se clasifican así, están correctamente clasificados y el 0,6% de los 507.369 que están sobre la media, cumplen este patrón.

Regla 4. Si el estudiante es de estrato socioeconómico bajo, es menor que 18 años, el índice de condición tic es malo, es de sexo masculino y el colegio se situa en la zona geografica de la Orinoquia entonces su desempeño académico en en la prueba de matemáticas de Saber 11° tiene mayor probabilidad de estar sobre la media nacional. El 0,5% del total de 1.061.680 estudiantes evaluados se clasifican de esta manera. El 61,4% de los 5.735 estudiantes que se clasifican así, están correctamente clasificados y el 0,7% de los 507.369 que están sobre la media, cumplen este patrón.

Regla 5. Si el estudiante es de estrato socioeconómico bajo, es menor que 18 años, el índice de condición tic es malo, es de sexo masculino y el colegio se sitúa en la zona geografica del Atlántico entonces su desempeño académico en en la prueba de matemáticas de Saber 11° tiene mayor probabilidad de estar bajo la media nacional. El 3,7% del total de 1.061.680 estudiantes evaluados se clasifican de esta manera. El 60,9% de los 39.287 estudiantes que se clasifican así, están correctamente clasificados y el 4,3% de los 566.068 que están bajo la media, cumplen este patrón.

Regla 5. Si el estudiante es de estrato socioeconómico bajo, es menor que 18 años, el índice de condición tic es regular, es de sexo femenino y la jornada de estudio es completa entonces su desempeño académico en la prueba de matemáticas de Saber 11° tiene mayor probabilidad de estar sobre la media nacional. El 2,6% del total de 1.061.680 estudiantes evaluados se clasifican de esta manera. El 61,3% de los 27.679 estudiantes que se clasifican así, están correctamente clasificados y el 3,3% de los 507.369 que están sobre la media, cumplen este patrón.

Regla 6. Si el estudiante es de estrato socioeconómico bajo, es menor que 18 años, el índice de condición tic es regular, es de sexo femenino, la jornada de estudio es en la mañana, los ingresos familiares mensuales está entre 1 y 2 salarios mínimos mensuales vigentes y la educación de la madre es técnica o tecnológica entonces su desempeño académico en la prueba de matemáticas de Saber 11° tiene mayor probabilidad de estar sobre la media nacional. El 1,1% del total de 1.061.680 estudiantes evaluados se clasifican de esta manera. El 60% de los 5.328 estudiantes que se clasifican así, están correctamente clasificados y el 0.6% de los 507.369 que están sobre la media, cumplen este patrón.

Regla 7. Si el estudiante es de estrato socioeconómico bajo, es menor que 18 años, el índice de condición tic es regular, es de sexo masculino y la jornada de estudio es completa entonces su desempeño académico en la prueba de matemáticas de Saber 11° tiene mayor probabilidad de estar sobre la media nacional. El 2,1% del total de 1.061.680 estudiantes evaluados se clasifican de esta manera. El 71,8% de los 22.351 estudiantes que se clasifican así, están correctamente clasificados y el 3,2% de los 507.369 que están sobre la media, cumplen este patrón.

Regla 8. Si el estudiante es de estrato socioeconómico bajo, es menor que 18 años, el índice de condición tic es regular, es de sexo masculino y la jornada de estudio es en la mañana entonces su desempeño académico en la prueba de matemáticas de Saber 11° tiene mayor probabilidad de estar sobre la media nacional. El 6,3% del total de 1.061.680 estudiantes evaluados se clasifican de esta manera. El 63,2% de los 67.319 estudiantes que se clasifican así, están correctamente clasificados y el 8,4% de los 507.369 que están sobre la media, cumplen este patrón.

Regla 9. Si el estudiante es de estrato socioeconómico bajo y es mayor que 22 años entonces su desempeño académico en la prueba de matemáticas de Saber 11° tiene mayor probabilidad de estar bajo la media nacional. El 2,3% del total de 1.061.680 estudiantes evaluados se clasifican de esta manera. El 89,7% de los 24.862 estudiantes que se clasifican así, están correctamente clasificados y el 4% de los 554.311 que están bajo la media, cumplen este patrón.

Regla 10. Si el estudiante es de estrato socioeconómico medio, su edad está entre 18 y 22 años y la jornada de estudio es completa entonces su desempeño académico en la prueba de matemáticas de Saber 11° tiene mayor probabilidad de estar sobre la media nacional. El 1,2% del total de 1.061.680 estudiantes evaluados se clasifican de esta manera. El 69,6% de los 12.523 estudiantes que se clasifican así, están correctamente clasificados y el 1,7% de los 507.369 que están sobre la media, cumplen este patrón.

Regla 11. Si el estudiante es de estrato socioeconómico medio, su edad está entre 18 y 22 años, la jornada de estudio es en la mañana y es de sexo femenino entonces su desempeño académico en la prueba de matemáticas de Saber 11° tiene mayor probabilidad de estar bajo la media nacional. El 0,8% del total de 1.061.680 estudiantes evaluados se clasifican de esta manera. El 62% de los 8.755 estudiantes que se clasifican así, están correctamente clasificados y el 1% de los 554.311 que están bajo la media, cumplen este patrón.

Regla 12. Si el estudiante es de estrato socioeconómico medio, su edad está entre 18 y 22 años y la jornada de estudio es en la noche entonces su desempeño académico en la prueba de matemáticas de Saber 11° tiene mayor probabilidad de estar bajo la media nacional. El 0,5% del total de 1.061.680 estudiantes evaluados se clasifican de esta manera. El 70,9% de los 5.531 estudiantes que se clasifican así, están correctamente clasificados y el 0,7% de los 554.311 que están bajo la media, cumplen este patrón.

Regla 13. Si el estudiante es de estrato socioeconómico medio, es menor que 18 años y la jornada de estudio es completa entonces su desempeño académico en la prueba de matemáticas de Saber 11° tiene mayor probabilidad de estar sobre la media nacional. El 6,6% del total de 1.061.680 estudiantes evaluados se clasifican de esta manera. El 82,1% de los 70.260 estudiantes que se clasifican así, están correctamente clasificados y el 11,4% de los 507.369 que están sobre la media, cumplen este patrón.

Regla 14. Si el estudiante es de estrato socioeconómico medio, es menor que 18 años y la jornada de estudio es en la mañana entonces su desempeño académico en la prueba de matemáticas de Saber 11° tiene mayor probabilidad de estar sobre la media nacional. El 7,3% del total de 1.061.680 estudiantes evaluados se clasifican de esta manera. El 67,4% de los 77.835 estudiantes que se clasifican así, están correctamente clasificados y el 10,3% de los 507.369 que están sobre la media, cumplen este patrón.

Regla 15. Si el estudiante es de estrato socioeconómico medio, es menor que 18 años y la jornada de estudio es en la tarde entonces su desempeño académico en la prueba de matemáticas de Saber 11° tiene mayor probabilidad de estar sobre la media nacional. El 2% del total de 1.061.680 estudiantes evaluados se clasifican de esta manera. El 60,9% de los 21.613 estudiantes que se clasifican así, están correctamente clasificados y el 2,6% de los 507.369 que están sobre la media, cumplen este patrón.

Regla 16. Si el estudiante es de estrato socioeconómico alto y la jornada de estudio es completa entonces su desempeño académico en la prueba de matemáticas de Saber 11° tiene mayor probabilidad de estar sobre la media nacional. El 1,9% del total de 1.061.680 estudiantes evaluados se clasifican de esta manera. El 91,1% de los 19.920 estudiantes que se clasifican así, están correctamente clasificados y el 3,6% de los 507.369 que están sobre la media, cumplen este patrón.

Regla 17. Si el estudiante es de estrato socioeconómico alto y la jornada de estudio es en la mañana entonces su desempeño académico en la prueba de matemáticas de Saber 11° tiene mayor probabilidad de estar sobre la media nacional. El 0,6% del total de 1.061.680 estudiantes evaluados se clasifican de esta manera. El 69,9% de los 6.091 estudiantes que se clasifican así, están correctamente clasificados y el 0,8% de los 507.369 que están sobre la media, cumplen este patrón.

De acuerdo a los resultados anteriores, el estrato socioeconómico está asociado al desempeño académico de los estudiantes que presentaron la prueba de matemáticas en el Saber 11°, específicamente los estudiantes de estratos bajos tienen un desempeño bajo, lo que no sucede con estudiantes de estratos altos que están sobre la media. Estos resultados coinciden con Garbanzo [29], Seibold [30] y Montero y Villalobos [31], en el sentido de que un resultado generalmente aceptable en el desempeño académico es la existencia de una asociación significativa entre el nivel

socioeconómico del estudiante y su desempeño académico. De igual manera Chica, Galvis y Ramírez [32] afirman que los resultados obtenidos en su estudio denominado "Determinantes del rendimiento académico en Colombia: pruebas Saber 11°", enseñan la relevancia que tienen las variables socioeconómicas en el desempeño académico en las pruebas Saber 11°.

La jornada académica es otro factor asociado al rendimiento académico de los estudiantes en la prueba de matemáticas en el Saber 11°, especialmente los de jornada completa que están sobre la media nacional. Hecho que coincide con los resultados del estudio realizado por Chica, Galvis y Ramirez [32] utilizando un modelo Logit Ordenado Generalizado en el cual los bachilleres de jornada completa obtienen puntajes más altos comparados con los estudiantes pertenecientes a otras jornadas. De igual manera en los resultados obtenidos en el estudio denominado "La jornada escolar y el rendimiento de los alumnos" de Ridao y Gil [33], se registran mejores calificaciones en los centros con jornada completa, con relación a la jornada continua.

El índice de condición TIC de los estudiantes que mide la posibilidad que tienen los estudiantes de utilizar internet, el computador, la telefonía en su casa es otro factor asociado al desempeño académico de los estudiantes que presentaron la prueba de matemáticas en el Saber 11°, específicamente si este índice es MALO su desempeño estará por debajo de la media. Hecho que se corrobora en otras investigaciones como el de Botello y Guerrero [34] que estudian el impacto que tienen las tecnologías de la información y comunicación sobre el desempeño académico de los estudiantes de América latina utilizando la prueba PISA del 2012. Los resultados muestran 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 niños, incrementando el puntaje promedio en cada una de las áreas de estudio entre un 5% y un 6%.

La edad del estudiante es otro factor asociado al rendimiento académico de los estudiantes en la prueba de matemáticas en el Saber 11°. Específicamente los menores que 22 años tienen un mejor desempeño que el resto de estudiantes. Este hecho valida lo expuesto por Murillo [35] en que los jóvenes, poseen un gran potencial de imaginación y talento creativo, que debe aprovecharse para el trabajo matemático en la Educación Básica y Media.

El sexo es un factor que asociado con otros factores como el estrato socioeconómico, la edad y la jornada de estudio determinan el buen o bajo desempeño académico de los estudiantes que presentaron la prueba de matemáticas en el Saber 11°. En este contexto, están los hallazgos de Gómez y Soares[36], quienes en su análisis sobre la diferencia de sexo en relación con el desempeño académico, expresan que tales diferencias no son significativas; que no se puede afirmar de modo definitivo que exista una relación directa entre el rendimiento académico y el sexo. Sin embargo de acuerdo a los resultados obtenidos los hombres si tienen un ligero mejor desempeño académico que las mujeres en la prueba de matemáticas en el Saber 11°.

V. Conclusiones y Trabajos futuros

Los resultados obtenidos con el modelo de clasificación por árboles de decisión para descubrir factores asociados al desempeño académico de los estudiantes colombianos que encontrándose finalizando el grado undécimo de educación media, presentaron la prueba de matemáticas dentro de las pruebas Saber 11° entre los años 2015 y 2016, indican que este es capaz de generar modelos consistentes con la realidad observada y el respaldo teórico, basándose únicamente en los datos que se encuentran almacenados en las bases de datos del ICFES.

Entre los atributos con mayor ganancia de información que forman parte de los patrones descubiertos, se destacan el estrato socioeconómico, la jornada de estudio, el índice tic, la edad y el sexo de los estudiantes como factores importantes asociados al buen o bajo desempeño académico de los estudiantes en la prueba de matemáticas.

Entre las dificultades presentadas en el desarrollo de la investigación están la mala calidad de los datos de las bases de datos del ICFES, que se tuvieron que descartar ciertos atributos por la imposibilidad de obtener sus valores en otras fuentes, y que de alguna manera, podrían influir en el descubrimiento de los patrones objeto de este estudio, además del gran consumo de recursos que implicó el proceso de limpieza y transformación de datos.

Se plantea como trabajos futuros complementar este estudio con el desempeño académico en el resto de competencias que evalúa las pruebas Saber 11°. Además, utilizar otras técnicas de minería de datos que permitan relacionar cuales atributos se presentan juntos asociados al desempeño académico en las pruebas Saber 11° y cómo se agrupan los estudiantes de acuerdo a su rendimiento en dichas pruebas.

Además, sería recomendable realizar estudios sobre la relación entre el desempeño académico de los estudiantes en las pruebas Saber 11°, el desempeño académico en las Instituciones de Educación Superior en su formación profesional y en las pruebas Saber Pro que presentan los estudiantes próximos a terminar una carrera profesional en Colombia.

Agradecimientos

Este proyecto de investigación se financió con recursos del sistema de investigaciones de la Universidad de Nariño (Colombia).

Referencias

[1] R. Jahanian. Educational Evaluation: Functions and Applications in Educational Contexts. International Journal of Academic Research in Economics and Management Sciences. Vol. 1, No. 2 ISSN: 2226-3624. 2012.

[2] H. Fernández. Como interpretar la evaluación pruebas Saber. Subdirección de Estándares y Evaluación. Ministerio de Educación Nacional. Bogotá, Colombia.2005.

[3] Instituto Colombiano para la Evaluación de la Educación (ICFES). Alineación del examen SABER 11° Lineamientos generales 2014 – 2 Sistema Nacional de Evaluación Estandarizada de la Educación.. ISBN: 978-958-11-0630-1. Bogotá, Colombia. 2014.
[4] 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, Colombia. 2006.
[5] Instituto Colombiano para la Evaluación de la Educación (ICFES). Sistema Nacional de Evaluación Estandarizada de la Educación: Lineamientos generales para la presentación del examen de Estado Saber 11°. ISBN: 978-958-11-0680-6. Bogotá, Colombia. 2016.
[6] J. Posada y F. Mendoza. 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.2014.
[7] A. Gaviria y J. Barrientos. Calidad de la educación y rendimiento académico en Bogotá. Revista Coyuntura Social, Núm. 24. ISSN: 0121-2532. Bogotá D.C., Colombia. 2001.
[8] 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. ISSN: 0120-2596. Universidad de Antioquia. Medellín, Colombia.2008
[9] J. Correa. Determinantes del Rendimiento Educativo de los Estudiantes de Secundaria en Cali: un análisis multinivel. En: Revista Sociedad y Economía. No. 6. pp. 81-105. 2004.
[10] S. Chica, D. Galvis y A. Ramirez. Determinantes del rendimiento académico en Colombia: pruebas ICFES Saber 11º. Revista Universidad EAFIT, Vol. 46, Núm. 160. ISSN: 0120-341X. Medellín, Colombia.2010.
[11] 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. Disponible en: https://repository.icesi.edu.co/biblioteca_digital/bitstream/10906/77946/1/gomez_analisis_competencias_2014.pdf.2014.
[12] 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. Disponible en: http://repositorio.escuelaing.edu.co/bitstream/001/349/1/AA-Econom%C3%ADa-1077087614.pdf.2015.
[13] R.Timarán, A. Calderón y J. Jiménez. Aplicación de la minería de datos en la extracción de perfiles de deserción estudiantil. Revista Ventana Informática, No 28(ene.-jun.). Manizales, Colombia: Facultad de Ciencias e Ingeniería, Universidad de Manizales. p. 31-47.ISSN:0123-9678.2013.
[14] R. Timarán, A. Calderón y J. Jiménez. La minería de datos como un método innovador para la detección de patrones de deserción estudiantil en programas de pregrado en Instituciones de Educación Superior. En Memorias Foro Mundial de Educación en Ingeniería, WEEF 2013. Cartagena, Colombia: ACOFI & IFEES. 2013.
[15] A. Peña. Educational data mining: A survey and a data mining-based analysis of recent works. Expert Systems with Applications, 1432–1462. 2014.
[16] R. Baker. 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. 2010.
[17] A. Algarni. Data Mining in Education. In (IJACSA) International Journal of Advanced Computer Science and Applications. Vol. 7, No. 6. 2016.
[18] S. Valero, S. 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. 2009.
[19] S.Valero, A. Salvador y M. García. Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos. Puebla, México: Universidad Tecnológica de Izúcar de Matamoros. Disponible en: www.utim.edu.mx/~svalero/docs/e1.pdf. 2010.
[20] University of Waikato. Weka 3: Data Mining Software in Java. Nueva Zelanda: Machine Learning Group at the University of Waikato. Disponible en: http://www.cs.waikato.ac.nz/ml/weka/
[21] J. Han and M. Kamber. Data Mining: Concepts and Techniques. San Francisco, USA: Morgan Kaufmann Publishers; 2001. 550 p. 2001.
[22] K, Sattler and O. Dunemann. SQL Database Primitives for Decision Tree Classifiers. In: Paques H, Liu L, Grossman D, editors. The 10th ACM International Conference on Information and Knowledge Management. Atlanta, USA: ACM New York. p. 379-86. 2001.
[23] A. Azevedo and M. Santos. KDD, SEMMA and CRISP-DM: a parallel overview. In: Proceedings of IADIS European Conference on Data Mining. Amsterdam, Netherlands. p. 182-185. ISBN: 978-972-8924-63-8. 2008.
[24] J. Hernández, M. Ramírez, and C. Ferri. Introducción a la Minería de Datos. *Editorial Pearson Educación SA, Madrid*. Recuperado a partir de http://dspace.ucbscz.edu.bo/dspace/handle/123456789/526. 2005.
[25] J. Villena. 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. 2016.
[26] R. Timarán and M. Millán. New algebraic operators and SQL primitives for mining classification rules. En Computational Intelligence (pp. 61–65). Disponible en: http://www.actapress.com/PaperInfo.aspx?PaperID=29048&reason=500. 2006.
[27] J.R. Quinlan. C 4. 5: Programs for Machine Learning. San Francisco (CA, USA): Morgan Kaufmann Publishers. 299 p. ISBN: 1-55860-238-0. 1993.
[28] M. García y A. Álvarez A. Análisis de Datos en WEKA -Pruebas de Selectividad. Disponible en: http://www.it.uc3m.es/jvillena/irc/practicas/06-07/28.pdf. 2010.
[29] G. Garbanzo. Factores asociados al rendimiento académico en estudiantes universitarios, una reflexión desde calidad de la educación superior pública. Revista Educación, 31(1):43-63. 2007.
[30] 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, 23. Disponible en: http://www.rieoei.org/rie23a07.htm. 2000.
[31] E. Montero y J. Villalobos. Factores institucionales, pedagógicos, psicosociales y sociodemográficos asociados al rendimiento académico y a la repetición estudiantil en la Universidad de Costa Rica. San José, Costa Rica: Universidad de Costa Rica.pdf. 2004.
[32] S. Chica, D. Galvis y A. Ramirez. Determinantes del rendimiento académico en Colombia: pruebas ICFES Saber 11º. Revista Universidad EAFIT, Vol. 46, Núm. 160. ISSN: 0120-341X. Medellín, Colombia. 2010.
[33] I. Ridao y J. Gil. La jornada escolar y el rendimiento de los alumnos. En revista de Educación, No. 327(2002), pp.141-156.2002.
[34] H. Botello y A. Rincón. La influencia de las TIC en el desempeño académico de los estudiantes en América Latina: Evidencia de la prueba PISA 2012. Repositorio Digital Reposital. Universidad Nacional Autónoma de México. 2015. Disponible en: http://reposital.cuaed.unam.mx:8080/xmlui/handle/123456789/4050.
[35] E. Murillo. Factores que inciden en el Rendimiento Académico en el area de Matemáticas de los estudiantes de noveno grado en los Centros de Educación Básica de la ciudad de Tela, Atlántida Tesis de maestria. Universidad Pedagógica Nacional Francisco Morazán.San Pedro Sula, Honduras.2013.
[36] G.Gómez y Soares. Diferencias de género con relación al desempeño académico en estudiantes de nivel básico. Alternativas en Psicología,XVII(28),106-118. 2013.