Cómo citar este artículo: Timarán-Pereira, R., Caicedo-Zambrano, J., & Hidalgo-Troya, A. (2019). Árboles de decisiones para predecir factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas saber 11°. Rev.investig.desarro.innov., 9 (2), 363-378. doi: 10.19053/20278306.v9.n2.2019.9184

Árboles de decisión para predecir factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas Saber 11°

Decision trees for predicting factors associated with academic performance of high school students in Saber 11 tests

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

Recibido: agosto 06 de 2018
Aceptado: diciembre 21 de 2018

Resumen

En este artículo se presentan los resultados obtenidos al aplicar el modelo de clasificación basado en árboles de decisión, con el fin de detectar factores asociados al desempeño académico de los estudiantes colombianos de grado undécimo de educación media, que presentaron las pruebas Saber 11° en los años 2015 y 2016. La investigación fue de tipo descriptivo bajo el enfoque cuantitativo, aplicando un diseño no experimental. Siguiendo 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 y utilizando la herramienta de minería de datos WEKA, se generaron árboles de decisión que permitieron identificar patrones asociados al buen o mal desempeño académico de los estudiantes en las pruebas Saber 11°. Los patrones descubiertos ayudarán en los procesos de toma de decisiones del Ministerio de Educación Nacional, junto con las instituciones que velan por la calidad de la educación en Colombia.

Palabras clave: minería de datos, patrones asociados, desempeño académico, pruebas Saber 11°, árboles de decisión.

Abstract

This article is obtained by applying the classification model based on decision trees in order to detect factors associated with the academic performance of Colombian eleven grade students, who presented the Saber 11 tests in 2015 and 2016. The research was of a descriptive type under the quantitative approach, applying a non-experimental design. Following the CRISP-DM methodology, the socio-economic, academic and institutional information of these students was selected from the ICFES databases. A data repository was built, cleaned and transformed and, using the WEKA data mining tool, decision trees were generated that allowed the identification of patterns associated with the good or poor academic performance of the students in the Saber 11 tests. The patterns discovered will help in the decision-making processes of the Ministerio de Educación Nacional, together with institutions that ensure the quality of education in Colombia.

Keywords: data mining, associated patterns, academic performance, Saber 11 tests, decision trees.

1 Ingeniero de Sistemas, Doctor en Ingeniería énfasis Ciencias de la Computación, Universidad de Nariño, Pasto, Colombia. E-mail: ritimar@udenar.edu.co
2 Licenciado en Matemáticas, Doctor en Educación, Universidad de Nariño, Pasto, Colombia. E-mail: jacaza@udenar.edu.co
3 Licenciado en Matemáticas, Magíster en Estadística, Universidad de Nariño, Pasto, Colombia. E-mail: arsenio.hidalgo@udenar.edu.co

1. Introducción

Es función principal de la evaluación en la educación, orientar y apoyar las acciones de mejoramiento de la calidad mediante la obtención, análisis e interpretación de información válida y confiable. 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 (Fernández, 2005).

La Ley 1324 confiere al Instituto Colombiano para Evaluación de la Educación, ICFES, la misión de evaluar, mediante exámenes externos estandarizados, la formación que se ofrece en el servicio educativo en los distintos niveles. También establece que el Ministerio de Educación Nacional, 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, 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 a los primeros. 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 (ICFES, 2014).

El examen evalúa cinco componentes basados en las aptitudes que deben desarrollar los educandos según los estándares básicos de competencias (MEN, 2006): lectura crítica, matemáticas, sociales y ciudadanas, ciencias naturales e inglés. 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 (ICFES, 2016). La prueba de Matemática evalúa las competencias de los estudiantes para enfrentar situaciones que pueden resolverse con el uso de algunas herramientas matemáticas (ICFES, 2016). 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 relaciones entre distintos eventos y la capacidad de reflexionar y emitir juicios críticos sobre estos (ICFES, 2016).

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 (ICFES, 2016). Finalmente, 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 (ICFES, 2016).

Los resultados de pruebas nacionales e internacionales muestran que Colombia posee un sistema educativo con bajos logros académicos de sus

estudiantes, en cada uno de los niveles de estudio (Posada-Ramos & Mendoza-Martínez, 2014). Según un estudio realizado por la Procuraduría General de la Nación (2006), el gasto público destinado a promover la educación preescolar, básica y media revela una tendencia creciente; en el periodo 1995- 2005, el crecimiento promedio fue del 16,5%, lo cual en términos del PIB llegó a representar 3,1% de este. Ahora bien, en términos de calidad, según el mismo estudio de la Procuraduría (2006), el cual se planteó desde un enfoque de derecho, se evidencia que más de la mitad de los estudiantes de grado 11°, en las pruebas ICFES -ahora llamadas Saber 11°-, de 2004, se ubicaron en los niveles medio bajo y bajo en la adquisición y dominio de las competencias de las áreas de historia, filosofía, física, química, matemáticas y geografía, respectivamente. 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° (Gaviria & Barrientos, 2001; Barrientos-Marín, 2008; Correa, 2004; Chica-Gómez, Galvis-Gutiérrez & Ramírez-Hassan, 2010; Gómez, 2014; Hernández-Angulo, 2015), en su mayoría 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. La aplicación de las técnicas y herramientas de la minería de datos en los diferentes contextos educativos se conoce como minería de datos en educación, del inglés educational data mining.

La minería de datos en la educación no es un tema nuevo, su estudio y aplicación ha sido muy relevante en los últimos años, pues se pueden utilizar sus técnicas para explicar y/o predecir cualquier fenómeno dentro del campo educativo (Timarán- Pereira et al., 2013a; 2013b). Por ejemplo, utilizando las técnicas de minería de datos, se puede predecir, con un porcentaje muy alto de confiabilidad, la probabilidad de deserción de cualquier estudiante (Valero, 2009; Valero et al., 2010). Blanco (2015), en su tesis de maestría aplica la minería de datos en la educación con el fin de analizar el desempeño académico de los estudiantes del departamento del Cesar que presentaron las Pruebas Saber 11° en el año 2012 -2, para el ingreso a la Educación Superior utilizando la técnica de clustering. Las instituciones educativas pueden usar la minería de datos en la educación para hacer análisis comprensivos de las características de sus estudiantes, métodos evaluativos, develando procesos exitosos o por el contrario, detectando fraudes o inconsistencias (Valero et al., 2010).

En este artículo se presentan los resultados obtenidos al aplicar el modelo de clasificación basado en árboles de decisión para predecir patrones asociados al desempeño académico de los estudiantes colombianos que, encontrándose finalizando el grado undécimo de educación media, presentaron las pruebas Saber 11° entre los años 2015 y 2016, a partir de la información socioeconómica, académica e institucional, almacenada en las bases de datos del ICFES, utilizando la metodología CRISP-DM, Cross Industry Standard Process for Data Mining, y la herramienta de minería de datos WEKA, Waikato Environment for Knowledge Analysis.

2. Materiales y métodos

La investigación fue de tipo descriptivo bajo el enfoque cuantitativo, aplicando un diseño no experimental. Como fuentes de información se utilizaron los datos que se encontraban disponibles, al momento de la investigación, en las bases de datos del ICFES de los resultados de los estudiantes que

presentaron las pruebas Saber 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 (Witten, Frank & Hall, 2011). Se escogió este modelo porque según la experiencia de algunos autores (Han & Kamber, 2001; Sattler & Dunemann, 2001; Timarán & Millán, 2006), para este tipo de proyectos, es probablemente el más utilizado y popular por su simplicidad y facilidad para entender los resultados obtenidos. Además, la importancia de los árboles de decisión se debe a su capacidad de construir modelos interpretables, siendo este un factor decisivo para su aplicación. Por otra parte, se escogió WEKA por ser una herramienta de minería de datos de software libre, distribuida bajo licencia GPL, que contiene una colección de algoritmos para realizar análisis de datos y modelado predictivo, tiene herramientas para la visualización de estos datos y provee una interfaz gráfica que unifica las herramientas para acceder fácilmente a sus funcionalidades (Calleja, 2010; García-Gutiérrez, 2016).

Para el descubrimiento de patrones, se aplicó la metodología CRISP-DM (Chapman et al., 2000; Villena-Román, 2016). En cuanto a las metodologías para desarrollar análisis de minería de datos y en un intento de normalización del proceso, de forma similar a como se hace en ingeniería para normalizar el proceso de desarrollo de software, surgieron a finales de los 90 dos metodologías principales: CRISP-DM (Chapman et al., 2000; Villena-Román, 2016), y SEMMA (Sample, Explore, Modify, Model, and Assess) (Azevedo & Santos, 2008). Ambas especifican las tareas a realizar en cada fase descrita por el proceso, asignando tareas concretas y definiendo lo que es deseable obtener tras cada fase. Azevedo y Santos (2008), comparan ambas implementaciones y llegan a la conclusión que, aunque se puede establecer un paralelismo claro entre ellas, CRISP-DM es más completo porque tiene en cuenta la aplicación al entorno de negocio de los resultados, y por ello es la que se adoptó popularmente.

En encuestas realizadas en KDNuggets en 2002, 2004, 2007 y 2014 se comprobó que CRISP-DM era la principal metodología utilizada, cuatro veces más que SEMMA. La metodología CRISP-DM para proyectos de minería de datos no es la "más actual" o "la mejor", pero es muy útil para comprender esta tecnología o extraer ideas para diseñar o revisar métodos de trabajo para proyectos de similares características (Azevedo & Santos, 2008). CRISP-DM es la guía de referencia más ampliamente utilizada en el desarrollo de proyectos de minería de datos (Hernández et al., 2005) y contempla seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación.

En la fase de análisis del problema se identificó con exactitud la problemática que se solucionaría utilizando la minería de datos, esto permitió recolectar la información necesaria para interpretar con asertividad los resultados encontrados (Villena-Román, 2016). 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 seleccionaron los datos a los cuales se les aplicaría una determinada técnica de modelado, limpieza, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato (Villena-Román, 2016). En la fase de modelado se seleccionaron las técnicas de minería de datos más apropiadas para el proyecto. En la fase de evaluación se verificó si el modelo se ajusta a las necesidades establecidas en el proyecto. Se evaluaron los patrones encontrados con el fin de determinar su validez, remover los redundantes o irrelevantes y traducir los patrones útiles en términos que sean entendibles para el usuario. Finalmente, en la fase de implementación, se trató de explotar la potencialidad de los modelos, integrarlos en los procesos de toma de decisión

del MEN, ICFES y de las instituciones gubernamentales y educativas que velan por la calidad de la educación en Colombia y difundir informes sobre el conocimiento extraído (Villena-Román, 2016).

3. Resultados y discusión

Teniendo en cuenta las fases de la metodología CRISP-DM, los siguientes son los resultados de cada una de las etapas.

3.1 Análisis del problema

En esta fase, se realizaron las actividades que permitieron profundizar y apropiar de una manera completa el problema objeto de estudio, los objetivos y los requisitos de esta investigación, que posibilitaron la recolección de los datos correctos para interpretar adecuadamente los resultados. En esta fase, descubrir factores asociados al desempeño académico de los estudiantes 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.

3.2 Análisis 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° en 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.

3.3 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. Un ejemplo de estos procesos se muestra en la tabla 1 con la generalización del atributo departamento en zonas geográficas.

Tabla 1. Generalización del atributo departamento en zonas geográficas.

ZONAS	DEPARTAMENTOS
ATLÁNTICA	Atlántico, Bolívar, Cesar, Córdoba, Guajira, Magdalena, San Andrés, Providencia y Sucre.
AMAZONAS	Amazonas, Guainía, Guaviare, Putumayo, Vaupés y Caquetá
ANDINA	Boyacá, Cundinamarca, Norte de Santander, Santander, Caldas, Risaralda, Huila, Tolima y Quindío.
ANTIOQUIA	Antioquia
PACÍFICO	Cauca, Chocó, Nariño y Valle del Cauca.
BOGOTÁ	Distrito Capital de Bogotá

Como resultado de esta fase, se obtuvo un repositorio de datos limpio y transformado, con 1.061.680 registros y 17 atributos, listo para aplicarle las técnicas de minería de datos y al cual se le denominó T1061680A17. En la tabla 2 se muestra el diccionario de datos de este repositorio.

Con el fin de tener una comprensión preliminar de los datos del repositorio final T1061680A17, se realizó una caracterización de las variables socioeconómicas de los estudiantes que presentaron las pruebas Saber 11°. Los resultados se muestran en la tabla 3. Además, con el fin de establecer cómo se asocian linealmente las diferentes competencias que evalúa las pruebas Saber 11°, se realizó una correlación a través del coeficiente de correlación de Pearson. Los resultados obtenidos se muestran en la tabla 4.

Tabla 2. Diccionario de datos del repositorio final T1061680A17.

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, TECNOLÓGICO, PROFESIONAL, POSTGRADO, NINGUNO
7	fami_educa_padre	Máximo nivel educativo del padre	PRIMARIA, SECUNDARIA,TÉCNICO, TECNOLÓGICO, PROFESIONAL, POSTGRADO, NINGUNO

NO	ATRIBUTO	DESCRIPCIÓN	VALORES
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	fami_automovil	Si en la familia tienen automóvil	Si o No
11	econ_condicion_vivienda	Condición de la vivienda del estudiante	BUENA, MALA,REGULAR
12	eco_condicion_tic	Condición de uso de TIC en el hogar del estudiante	BUENA, REGULAR, MALA
13	eco_condicion_vive	Condición de vida del estudiante	SIN HACINAMIENTO, HACINAMIENTO MEDIO, HACINAMIENTO CRÍTICO
Académicos			
14	punt_global_cuali	Puntaje global obtenido por el estudiante en las pruebas Saber 11	POR ENCIMA DE LA MEDIA NACIONAL POR DEBAJO DE LA MEDIA NACIONAL
Institucionales			
15	Tipo_cole	Tipo de institución educativa	PÚBLICA,PRIVADA
16	Cole_jornada	Jornada de estudio del estudiante	MAÑANA, TARDE, NOCHE, ÚNICA, SABATINA-DOMINICAL
17	cole_zonageo	Zona geográfica donde se encuentra la institución educativa	Ver TABLA 1.

Tabla 3. Características socioeconómicas de los estudiantes del repositorio final T1061680A17.

VARIABLE SOCIOECONÓMICA		N	%
Género	Femenino	575.659	54,2%
	Masculino	483.168	45,5%
	Sin dato	2.853	0,3%
Grupos de edad	Menor que 18 años	749.256	70,6%
	Entre 18 y 22 años	284.108	26,8%
	Mayor que 22 años	28.316	2,7%
Estrato social	Alto	27.408	2,6%
	Medio	221.438	20,9%
	Bajo	812.834	76,5%

Ingreso familiar	Menos de 1 SM	292.226	27,4%
	Entre 1 y menos de 2 SM	468.384	44,1%
	Entre 2 y menos de 3 SM	167.740	15,8%
	Entre 3 y menos de 5 SM	71.994	6,8%
	Entre 5 y menos de 7 SM	24.629	2,3%
	Entre 7 y menos de 10 SM	14.138	1,3%
	10 o más SM	18.374	1,7%
	Sin dato	4.195	0,4%
Tipo de Colegio	Privado	243.037	22,9%
	Público	818.643	77,1%
Jornada	Completa u Ordinaria	233.321	22,0%
	Única	2.570	0,2%
	Mañana	584.212	55,9%
	Tarde	164.430	15,5%
	Noche	77.147	7,3%
Zona Geográfica	Amazonas	18.605	1,8%
	Andina	286.890	27,0%
	Antioquia	132.851	12,5%
	Atlántica	235.799	22,2%
	Bogotá	189.418	17,8%
	Orinoquia	37.046	3,5%
	Pacífica	161.071	15,2%
Total		1.061.680	100.0%

Caracterizando a los estudiantes que presentaron las pruebas Saber 11 en los años 2015 y 2016: por género la mayoría de estudiantes son mujeres, con un 54,2%; por edad, el mayor porcentaje es menor que 18 años, con un 70,6%. En un alto porcentaje los estudiantes pertenecen a estratos sociales bajos (1,2), con un porcentaje de 76,5%. Igualmente, el 71.5% (27,4% y 44,1%), de las familias a las que pertenecen los estudiantes no alcanzan un ingreso mensual de 2 salarios mínimos, lo que es consistente con el hecho de que el 77.1%de los estudiantes asisten a colegios públicos, en su gran mayoría en jornada de la mañana (55.9%). Por zona geográfica, el 27% se ubica en la zona Andina, el 22.2% en la zona Atlántica, un 17.8% en Bogotá y el 33% restante en las demás zonas del país.

Tabla 4. Matriz de correlaciones de las competencias de las pruebas Saber 11°.

COMPETENCIAS	CIENCIAS NATURALES	INGLÉS	LECTURA CRÍTICA	MATEMÁTICAS	CIUDADANAS	GLOBAL
Ciencias Naturales	1	0,715	0,790	0,825	0,814	0,929
Inglés		1	0,691	0,694	0,691	0,795
Lectura Crítica			1	0,761	0,809	0,905
Matemáticas				1	0,782	0,919
Ciudadanas					1	0,923
Global						1

Por cantidad de datos (1.361.495), todas las correlaciones resultan altamente significativas (p valor <0,01). Siguiendo la clasificación de Cohen (1998), para la interpretación del coeficiente de Pearson, se observó que todas las competencias presentan correlaciones altas (r>0,5) y positivas. Sin embargo, se destaca que el puntaje global de la prueba presenta una correlación muy alta (r>0,9) con las demás pruebas, a excepción de la prueba de inglés. Igualmente, es importante la correlación de Ciencias Naturales con Matemáticas y con competencias Ciudadanas (r>0,8) e igualmente de esta última con Lectura Crítica.

3.4 Modelado

Se seleccionó la tarea de clasificación con árboles de decisión, como la técnica predictiva de minería de datos más adecuada para descubrir patrones asociados al desempeño académico de los estudiantes colombianos de grado undécimo de educación media en las pruebas Saber 11°. La clasificación con árboles de decisión considera clases disjuntas, de forma que el árbol conducirá a una y sólo una hoja, asignando una única clase a la predicción (Hernández & Lorente, 2009).

Con esta técnica se pretende obtener un modelo que permita predecir para los nuevos casos de estudiantes, cuales son los factores socioeconómicos, académicos e institucionales asociados al buen (por encima de la media) o mal (por debajo de la media) desempeño académico en las pruebas Saber 11°, teniendo en cuenta el puntaje global obtenido por el estudiante en las pruebas Saber 11°, como atributo clase.

Para la construcción del modelo de clasificación con árboles de decisión se utilizó la herramienta WEKA (Hall et al, 2011) y su algoritmo J48, el cual implementa al algoritmo C.45. El algoritmo J48 se basa en la utilización del criterio de ganancia de información (information gain). De esta manera se consigue evitar que las variables con mayor número de posibles valores salgan beneficiadas en la selección. Además, el algoritmo incorpora una poda del árbol de clasificación una vez que éste ha sido inducido (Hernández & Lorente, 2009). El parámetro más importante que se tuvo en cuenta para la poda fue el factor de confianza C (confidence level), que influye en el tamaño y capacidad de predicción del árbol construido. Cuanto más baja se haga esa probabilidad, se exigirá que la diferencia en los errores de predicción antes y después de podar sea más significativa para no podar. El valor por defecto de este factor es del 25% y conforme va bajando este valor, se permiten más operaciones de poda y por lo tanto llegar a árboles cada vez más pequeños (García & Álvarez, 2010). Otro parámetro utilizado para variar el tamaño del árbol fue a través del factor M que especifica el mínimo número de instancias o registros por nodo del árbol (Hall et al., 2011).

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

estimar la variabilidad del método de aprendizaje con respecto al conjunto de datos. Para esta investigación, se utilizaron 10 particiones (10-fold cross validation) teniendo en cuenta lo recomendado por Hernández et al. (2005).

Además, se evaluó o estimó el coste del clasificador para el repositorio T1061680A17 a través de la matriz de confusión. La matriz de confusión 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) (Fernández, 2009).

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 global de cada estudiante obtenido en las pruebas Saber 11°, el cual fue discretizado en los valores "por encima de la media nacional ", y "por debajo de la media nacional", siendo la media nacional 258 sobre 500. Con el fin de obtener diferentes modelos de árboles y escoger el de mejores resultados, se establecieron 2 porcentajes de prepoda del árbol para el factor M igual a 1% y 2% del total de registros del repositorio de datos, y 2 porcentajes para el factor confianza C igual a 25% y 5% y se construyeron los diferentes modelos combinando estos factores. Se escogió el árbol construido con los parámetros M=20000 (2%) 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 2% 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=2000 y C=2% se muestra en la figura 2.

3.5 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 interpretación y discusión de resultados.

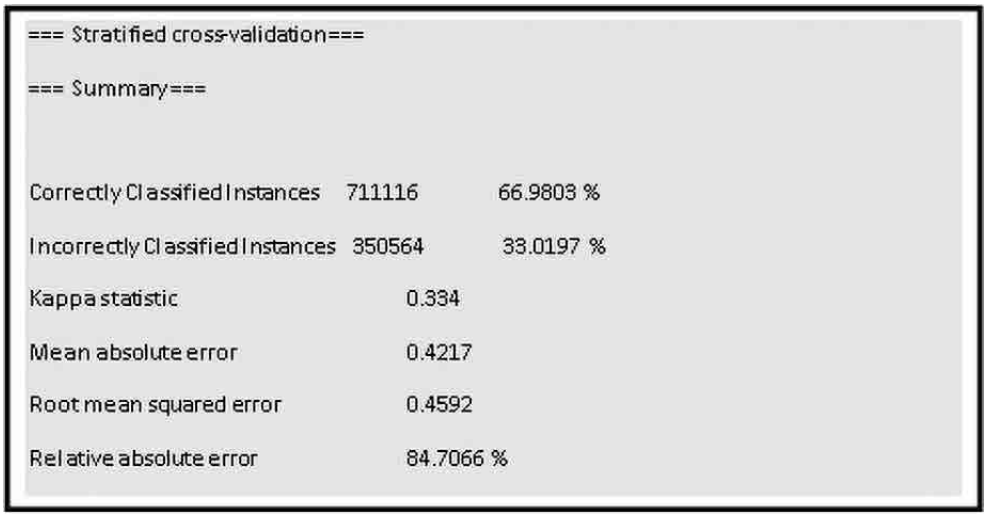

Figura 1. Precisión del modelo y matriz de confusión obtenidos con la herramienta Weka.

3.6 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 las instituciones educativas que velan por la calidad de la educación media y superior en Colombia. Una vez estas instituciones intervengan los factores asociados al desempeño académico en las Pruebas Saber 11°, será posible analizar los resultados y determinar sus efectos.

3.7 Interpretación y discusión de resultados

Analizando los resultados obtenidos con el árbol de decisión construido con el conjunto de datos T1061680A17, 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 global (puntaje_global_cuali) como clase, se puede observar que este clasifica correctamente a 711.116 instancias, que corresponde a un porcentaje de precisión del 67%; mientras que 350.564 instancias fueron incorrectamente clasificadas, correspondiendo a un porcentaje del 33%, ver figura 1.

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

fami_estrato = ALTO: ENCIMA MEDIA (27408.0/3737.0)
fami_estrato = BAJO
|  estu_edad_intervalo = Menor que 18 años
|  |  eco_condicion_tic = MALA
|  |  |  fami_nivel_sisben = Sin dato: DEBAJO MEDIA (549.0/208.0)
|  |  |  fami_nivel_sisben = No esta clasificada:  DEBAJO MEDIA (17094.0/8207.0)
|  |  |  fami_nivel_sisben = Nivel 2
|  |  |  |  estu_genero = M: ENCIMA MEDIA (22319.39/10424.59)
|  |  |  |  estu_genero = F: DEBAJO MEDIA (30233.61/13951.2)
|  |  |  fami_nivel_sisben = Nivel 1: DEBAJO MEDIA (220888.0/78852.0)
|  |  |  fami_nivel_sisben = Nivel 3: ENCIMA MEDIA (1890.0/871.0)
|  |  |  fami_nivel_sisben = Otro nivel del sisben: DEBAJO MEDIA (1192.0/525.0)
|  |  eco_condicion_tic = REGULAR
|  |  |  cole_jornada = Mañana
|  |  |  |  fami_ingreso_fmiliar_mensual  = Sin dato: ENCIMA MEDIA (0.0)
|  |  |  |  fami_ingreso_fmiliar_mensual  = Entre 1 y menos de 2 SM
|  |  |  |  |  estu_genero = M: ENCIMA MEDIA (38155.38/16323.26)
|  |  |  |  |  estu_genero = F
|  |  |  |  |  |  eco_condicion_vivienda  = REGULAR: DEBAJO MEDIA (2028.43/942.72)
|  |  |  |  |  |  eco_condicion_vivienda  = MALA: DEBAJO MEDIA (21736.73/10124.29)
|  |  |  |  |  |  eco_condicion_vivienda  = BUENA: ENCIMA MEDIA (27242.45/12814.58)
|  |  |  |  fami_ingreso_fmiliar_mensual  = Menos de 1 SM: DEBAJO MEDIA (20196.0/9048.0)
|  |  |  |  fami_ingreso_fmiliar_mensual  = Entre 3 y menos de 5 SM: ENCIMA MEDIA (9306.0/2961.0)
|  |  |  |  fami_ingreso_fmiliar_mensual  = Entre 2 y menos de 3 SM: ENCIMA MEDIA (34265.0/13420.0)
|  |  |  |  fami_ingreso_fmiliar_mensual  = 10 o más SM: ENCIMA MEDIA (161.0/64.0)
|  |  |  |  fami_ingreso_fmiliar_mensual  = Entre 7 y menos de 10 SM: ENCIMA MEDIA (352.0/104.0)
|  |  |  |  fami_ingreso_fmiliar_mensual  = Entre 5 y menos de 7 SM: ENCIMA MEDIA (1535.0/472.0)
|  |  |  cole_jornada = Tarde
|  |  |  |  estu_genero = M: ENCIMA MEDIA (22518.87/10158.14)
|  |  |  |  estu_genero = F: DEBAJO MEDIA (29434.13/13356.26)
|  |  |  cole_jornada = Noche: DEBAJO MEDIA (5999.0/1030.0)
|  |  |  cole_jornada = Completa u Ordinaria: ENCIMA MEDIA (50031.0/16981.0)
|  |  |  cole_jornada = Unica: ENCIMA MEDIA (605.0/242.0)
|  estu_edad_intervalo = Entre 18 y 22 años: DEBAJO MEDIA (230240.0/57208.0)
|  estu_edad_intervalo = Mayor que 22 años: DEBAJO MEDIA (24862.0/2338.0)
fami_estrato = MEDIO
|  cole_jornada = Mañana: ENCIMA MEDIA (98648.0/34883.0)
|  cole_jornada = Tarde: ENCIMA MEDIA (29321.0/12288.0)
|  cole_jornada = Noche: DEBAJO MEDIA (10142.0/2794.0)
|  cole_jornada = Completa u Ordinaria: ENCIMA MEDIA (82910.0/14175.0)
|  cole_jornada = Unica: ENCIMA MEDIA (417.0/139.0)

Number of Leaves  :         31

Size of the tree :          42
```

Figura 2. Mejor árbol de decisión textual obtenido con la herramienta Weka.

Teniendo en cuenta la matriz de confusión (figura 1), del total de 1.061.680 estudiantes evaluados, el modelo clasifica a 495.612 estudiantes con desempeño académico sobre la media, correspondiente a un 46.7% del total de estudiantes y a 566.068 estudiantes con un desempeño académico bajo la media, que corresponde al 53.7%. Del 46.7% de estudiantes que están sobre la media, el modelo clasifica correctamente a 304.114 estudiantes, que corresponde a un porcentaje de 61.4%. Del 53.7% de estudiantes que están bajo la media, el modelo clasifica correctamente a 407.002 estudiantes, que corresponde a un porcentaje de 71.9%.

Para efectos de la discusión de los resultados, se escogieron los patrones más representativos del mejor árbol obtenido, ver figura 2, teniendo en cuenta un mínimo soporte del 2% 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, es menor que 18 años, el índice de condición tic es malo y el nivel SISBEN es 1 entonces su desempeño académico en las pruebas Saber 11° es posible que este bajo la media nacional. El 20,81% del total de 1.061.680 estudiantes evaluados se clasifican de esta manera. El 64,30% de los 220.888 estudiantes que se clasifican así, están correctamente clasificados, y el 25,1% de los 566.068 que están bajo la media, cumplen este patrón.

Regla 2. Si el estudiante es de estrato socioeconómico bajo y su edad está entre 18 y 22 años entonces su desempeño académico en las pruebas Saber 11° es posible que este bajo la media nacional. El 21,69% del total de 1.061.680 estudiantes evaluados se clasifican de esta manera. El 75,15% de los 230.240 estudiantes que se clasifican así, están correctamente clasificados y el 30,57% de los 566.068 que están bajo la media, cumplen este patrón.

Regla 3. Si el estudiante es de estrato socioeconómico medio y su jornada de estudio es en la mañana, entonces su desempeño académico en las pruebas Saber 11° es posible que esté sobre la media nacional. El 9,29% del total de 1.061.680 estudiantes evaluados se clasifican de esta manera. El 64,64% de los 98.648 estudiantes que se clasifican así, están correctamente clasificados, y el 12.87% de los 495.612 que están sobre la media, cumplen este patrón.

Regla 4. Si el estudiante es de estrato socioeconómico medio y su jornada de estudio es completa, entonces su desempeño académico en las pruebas Saber 11° es posible que esté sobre la media nacional. El 7,81% del total de 1.061.680 estudiantes evaluados se clasifican de esta manera. El 82,9% de los 82.910 estudiantes que se clasifican así, están correctamente clasificados, y el 13.87% de los 495.612 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 regular y su jornada de estudio es completa, entonces su desempeño académico en las pruebas Saber 11° es posible que esté sobre la media nacional. El 4,71% del total de 1.061.680 estudiantes evaluados se clasifican de esta manera. El 66,06% de los 50.031 estudiantes que se clasifican así, están correctamente clasificados y el 6,67% de los 495.612 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, su jornada de estudio es en la mañana y los ingresos de su familia están entre 2 y menos que 3 salarios mínimos, entonces su desempeño académico en las pruebas Saber 11° es posible que este sobre la media nacional. El 3,23% del total de 1.061.680 estudiantes evaluados se clasifican de esta manera. El 60,83% de los 34.265 estudiantes que se clasifican así, están correctamente clasificados y el 4,21% de los 495.612 que están sobre la media, cumplen este patrón.

Regla 7. Si el estudiante es de estrato socioeconómico alto entonces su desempeño académico en las pruebas Saber 11° es posible que esté sobre la media nacional. El 2,58% del total de 1.061.680 estudiantes evaluados se clasifican de esta manera. El 86,37% de los 27.408 estudiantes que se clasifican así, están correctamente clasificados y el 4,78% de los 495.612 que están sobre la media, cumplen este patrón.

Regla 8. Si el estudiante es de estrato socioeconómico bajo y es mayor que 22 años entonces su desempeño académico en las pruebas Saber 11° es posible que esté bajo la media nacional. El 2,34% del total de 1.061.680 estudiantes evaluados se clasifican de esta manera. El 90,6% de los 24.862 estudiantes que se clasifican así, están correctamente clasificados y el 3.98% de los 566.068 que están bajo la media, cumplen este patrón.

De acuerdo con los resultados anteriores, el estrato socioeconómico y los ingresos familiares están asociados al desempeño académico de los estudiantes que presentaron las pruebas 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-Vargas (2007), Seibold (2000), y Montero-Rojas, Villalobos-Palmas y Cubero, (2004), 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. Igualmente, Chica-Gómez, Galvis-Gutiérrez y Ramirez-Hassan (2010), afirman que los resultados obtenidos en su estudio: "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°. En particular, las variables nivel de ingreso y nivel de escolaridad de los progenitores, las cuales presentan un impacto positivo y significativo en el resultado de las pruebas.

La jornada académica es otro factor asociado al rendimiento académico de los estudiantes en las pruebas Saber 11, especialmente los de jornada completa que están sobre la media nacional. Hecho que coincide con los resultados del estudio de Chica-Gómez, Galvis-Gutiérrez y Ramirez-Hassan (2010), 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. Igualmente, en el estudio: "La jornada escolar y el rendimiento de los alumnos", de Ridao-García y Gil-Flórez (2002), se registran mejores calificaciones en los centros con jornada partida, con relación a la jornada continua.

El índice de condición TIC, que mide la posibilidad que tienen los estudiantes de utilizar internet, el computador y la telefonía en su casa, es otro factor asociado al desempeño académico de los estudiantes que presentaron las pruebas Saber 11: específicamente, si este índice es MALO, su desempeño estará por debajo de la media. Este hecho se corrobora en investigaciones como la de Alberto-Botello y Guerrero-Rincón (2014), que estudian el impacto que tienen las Tecnologías de la Información y Comunicación, TIC, 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 de 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%.

4. Conclusiones

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 las pruebas Saber 11° entre los años 2015 y 2016, indican que estos son capaces 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.

Considerando el buen desempeño académico en las pruebas Saber 11° como aquellos puntajes globales por encima de la media y un bajo desempeño en estas pruebas como aquellos puntajes globales por debajo de la media, es mayor el porcentaje de estudiantes colombianos que tienen un desempeño académico bajo, comparado con el porcentaje de estudiantes que tienen un buen desempeño.

Por otra parte, entre los atributos con mayor ganancia de información que forman parte de los patrones descubiertos, asociados al buen desempeño académico en las pruebas Saber 11°, están: el estrato socioeconómico medio o alto, la jornada de estudio en la mañana o completa, el índice TIC regular y la edad menor que 18 años.

Asimismo, entre los atributos con mayor ganancia de información que forman parte de los patrones descubiertos, asociados a un bajo desempeño académico en las pruebas Saber 11°, están: el estrato socioeconómico bajo, el índice TIC bajo y el nivel SISBEN 1.

Entre las dificultades presentadas en el desarrollo de la investigación están la mala calidad de los datos de las bases de datos del ICFES, ya 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 trabajo futuro complementar este estudio utilizando otras técnicas predictivas y algoritmos, con el fin de comparar los resultados obtenidos con árboles de decisión con el algoritmo J48. Igualmente, aplicar otras tareas 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 individuos de acuerdo a su rendimiento en dichas pruebas. Además, sería recomendable realizar estudios sobre la relación entre el rendimiento académico de los estudiantes en las pruebas Saber 11°, el desempeño académico en las Instituciones de Educación Superior en su formación profesional y las pruebas Saber Pro que presentan los estudiantes próximos a terminar una carrera profesional en Colombia.

Agradecimientos

Este proyecto se financió con recursos del sistema de investigaciones de la Universidad de Nariño, en Pasto, Colombia.

Referencias

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

Barrientos-Marín, J. (2008). Calidad de la educación pública y logro académico en Medellín 2004-2006: Una aproximación por regresión intercuartil. *Revista Lecturas de Economía*, 68.

Blanco, V. (2015). *Análisis del Desempeño Académico del Examen de Estado para el Ingreso a la Educación Superior Aplicando Minería de Datos* (Tesis de Maestría). Universidad Nacional de Colombia. Valledupar, Colombia.

Alberto-Botello, L. H., & Guerrero-Rincón, A. (2014). La influencia de las TIC en el desempeño académico de los estudiantes en América Latina: Evidencia de la prueba PISA 2012. Memorias Virtual Educa. Lima, Perú.

Calleja, A. (2010). *Minería de Datos con Weka para la Predicción del Precio de Automóviles de Segunda Mano* (Trabajo de pregrado). Universidad Politécnica de Valencia. Recuperado de: https://riunet.upv.es/bitstream/handle/10251/10097/PFC_DSIC-80_Agust%C3%ADnCalleja.pdf

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. CRISP-DM consortium: NCR Systems Engineering Copenhagen (USA and Denmark), DaimlerChrysler AG (Germany), SPSS Inc. (USA), and OHRA Verzekeringen en Bank Groep B.V. (The Netherlands).

Chica-Gómez, S. M., Galvis-Gutiérrez, D. M., & Ramirez-Hassan, A. (2010). Determinantes del rendimiento académico en Colombia: pruebas ICFES Saber 11º. *Revista Universidad EAFIT*, 46 (160), 48-72. Recuperado de: http://publicaciones.eafit.edu.co/index.php/revista-universidad-eafit/article/view/754

Cohen, J. (1988). *Análisis de poder estadístico para las Ciencias del comportamiento*. Segunda ed. Nueva Jersey: Lawrence Erlbaum.

Correa, J. J. (2004). Determinantes del Rendimiento Educativo de los Estudiantes de Secundaria en Cali: un análisis multinivel. *Revista Sociedad y Economía*, 6, 81-105. Recuperado de: https://www.redalyc.org/pdf/996/99617648003.pdf

Fernández, H. (2005). *Cómo interpretar la evaluación pruebas Saber.* Subdirección de Estándares y Evaluación. Bogotá, Colombia: Ministerio de Educación Nacional.

Garbanzo-Vargas, G. M. (2007). Factores asociados al rendimiento académico en estudiantes universitarios, una reflexión desde calidad de la educación superior pública. *Revista Educación, 31*(1), 43-63. Recuperado de: https://www.redalyc.org/articulo.oa?id=44031103

García-Gutiérrez, J. A. (2016). Comenzando con Weka: Filtrado y selección de subconjuntos de atributos basada en su relevancia descriptiva para la clase. *Technical report.* Recuperado de: https://www.researchgate.net/publication/308141950.

Gaviria, A., & Barrientos, J. (2001). Calidad de la educación y rendimiento académico en Bogotá. *Revista Coyuntura Social,* 24, 112-127. Recuperado de: https://www.repository.fedesarrollo.org.co/handle/11445/1759

Gómez, J. (2014). *Análisis de las competencias en matemáticas y lenguaje de los bachilleres Colombianos* (Tesis de pregrado). Universidad ICESI. Cali, Colombia. Recuperado de: https://repository.icesi.edu.co/biblioteca_digital/bitstream/10906/77946/1/gomez_analisis_competencias_2014.pdf.

Han, J., & Kamber, M. (2001). *Data Mining: Concepts and Techniques*. San Francisco, USA: Morgan Kaufmann Publishers.

Hernández, J., Ramírez, M., & Ferri, C. (2005). *Introducción a la Minería de Datos.* Madrid, España: Pearson Prentice Hall.

Hernández-Angulo, O. E. (2015). *Determinantes del Rendimiento Académico en la Educación Media de Cundinamarca* (Tesis de pregrado). Escuela Colombiana de Ingeniería Julio Garavito. Bogotá, Colombia. Recuperado de: https://repositorio.escuelaing.edu.co/bitstream/001/349/1/Hern%C3%A1ndez%20Angulo%2C%20Oscar%20Eduardo-2015.pdf

Hernández-Martínez, E., & Lorente-Sanjurjo, R. (2009). *Minería de datos aplicada a la detección de Cáncer de Mama.* Madrid, España: Universidad Carlos III de Madrid. Recuperado de: https://www.researchgate.net/publication/265891193_Minera_de_datos_aplicada_a_la_deteccion_de_Cancer_de_Mama

Instituto Colombiano para la Evaluación de la Educación, ICFES. (2014). *Alineación del examen SABER 11° Lineamientos generales 2014 – 2 Sistema Nacional de Evaluación Estandarizada de la Educación.* Bogotá, Colombia.

Instituto Colombiano para la Evaluación de la Educación, ICFES. (2016). *Sistema Nacional de Evaluación Estandarizada de la Educación: Lineamientos generales para la presentación del examen de Estado Saber 11°.* Bogotá, Colombia.

Ministerio de Educación Nacional, MEN. (2006). *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.

Montero-Rojas, E., Villalobos-Palmas, J., & Cubero, Z. R. (2004). 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.

Posada-Ramos, J. M., & Mendoza-Martínez, 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. Bogotá, Colombia: Ministerio de Educación Nacional. Recuperado de: http://webcache.googleusercontent.com/search?q=cache:dkK95ExkHmAJ:www2.icfes.gov.co/docman/investigadores-y-estudiantes-de-posgrado/resultados-de-investigaciones/factores-asociados/educacion-superior/1011-determinantes-del-logro-academico-de-los-estudiantes-de-grado-11-en-el-periodo-2008-2010-una-perspectiva-de-genero-y-region+&cd=1&hl=es-419&ct=clnk&gl=co

Procuraduría General de la Nación (2006). El derecho a la educación: la educación en la perspectiva de los Derechos Humanos. Bogotá, Colombia.

Ridao-García, I., & Gil-Flórez, J. (2002). La jornada escolar y el rendimiento de los alumnos. *Revista de Educación,* 327, 141-156.

Sattler, K., & Dunemann, O. (2001). SQL Database Primitives for Decision Tree Classifiers. En: Paques H, Liu L, Grossman D, editors. *The 10th ACM International Conference on Information and Knowledge Management*. 379-86. Atlanta, USA: ACM New York.

Seibold, J. R. (2000). La calidad integral en educación. Reflexiones sobre un nuevo concepto de calidad educativa que integre valores y equidad educativa. *Revista Iberoamericana de Educación,* 23, 215-231. Recuperado de: https://rieoei.org/RIE/article/view/1012

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

Timarán-Pereira, R., Calderón-Romero, A., & Jiménez-Toledo, J. (2013a). Aplicación de la minería de datos en la extracción de perfiles de deserción estudiantil. *Revista Ventana Informática,* 28, 31-47. Recuperado de: http://webcache.googleusercontent.com/search?q=cache:5ZShtZGF8WQJ:revistasum.umanizales.edu.co/ojs/index.php/ventanainformatica/article/download/181/228+&cd=1&hl=es-419&ct=clnk&gl=co

Timarán-Pereira, R., Calderón-Romero, A., & Jiménez-Toledo, J. (2013b). 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. *Foro Mundial de Educación en Ingeniería, WEEF 2013.* Cartagena, Colombia: ACOFI & IFEES.

Valero, S. (2009). *Aplicación de técnicas de minería de datos para predecir deserción*. Puebla, México: Universidad Tecnológica de Izúcar de Matamoros. Recuperado de: http://www.utim.edu.mx/~svalero/docs/Mineria-Desercion.pdf.

Valero, S., Salvador, A., & García, M. (2010). *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. Recuperado de: www.utim.edu.mx/~svalero/docs/e1.pdf.

Villena-Román, J. (2016). *CRISP-DM: La metodología para poner orden en los proyectos de Data Science.* Recuperado de: 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.* Third Edition. Morgan Kaufmann.