Solís, Dahyana, Alegría, Diego, Gutiérrez, Édgar, Zapata, Víctor, Vidal, Fredy and Timarán, Ricardo (2019). Identificación de Patrones de Rendimiento Académico en las Pruebas Saber Pro entre 2012-2014, en las Competencias Lectura Crítica y Comunicación Escrita con Técnicas Predictivas de Minería de Datos Cuaderno Activa, 11, 51-64.

Cuaderno Activa
•REVISTA CIENTÍFICA DE LA FACULTAD DE INGENIERÍA•

Identificación de Patrones de Rendimiento Académico en las Pruebas Saber Pro entre 2012-2014, en las Competencias Lectura Crítica y Comunicación Escrita con Técnicas Predictivas de Minería de Datos[1]

Identification of Academic Performance Patterns in the Projects Saber Pro between 2012-2014, in Competences Critical Reading and Written Communication with Data Mining Predictive Techniques

Dahyana Andrea Solís Flórez[2], Diego Fernando Alegría Castrillón[3], Édgar Armando Gutiérrez Vidal[4], Víctor Alfonso Zapata Bedoya[5], Fredy Alonso Vidal Alegría[6], Ricardo Timarán Pereira[7]

Recibido: 8 marzo de 2018 **Aprobado:** 2 de abril de 2019

1 Este artículo es el resultado del proyecto "Identificación de Patrones de Rendimiento Académico en las Pruebas Saber Pro en el Período Comprendido entre 2012-2014, en las Competencias Lectura Crítica y Comunicación Escrita con Técnicas Predictivas de Minería de Datos", desarrollado como trabajo final en el programa de Especialización en Administración de la Información y Bases de Datos de la Institución Universitaria Colegio Mayor del Cauca.

2 Ingeniero de Sistemas. Egresado de la Especialización en Administración de la Información y Bases de Datos. Correo electrónico: asolis@unimayor.edu.co

3 Ingeniero de Sistemas. Egresado de la Especialización en Administración de la Información y Bases de Datos. Correo electrónico: diegofer0419@gmail.com

4 Ingeniero de Sistemas. Egresado de la Especialización en Administración de la Información y Bases de Datos. Correo electrónico: egutierrez@unimayor.edu.co

5 Ingeniero de Sistemas. Egresado de la Especialización en Administración de la Información y Bases de Datos. Correo electrónico: victor_zapata@unimayor.edu.co

6 Magíster en Educación para la Diversidad. Decano de la Facultad de Ingeniería de la Institución Universitaria Colegio Mayor del Cauca. Correo electrónico: decing@unimayor.edu.co

7 Doctor en Informática. Docente de planta de la Universidad de Nariño. Correo electrónico: ritimar@udenar.edu.co

Resumen: La Ley 30 del 28 de diciembre de 1992 (Fundamentos de la Educación Superior), menciona en el artículo 31 la necesidad de: "Propender por la creación de mecanismos de evaluación de la calidad de los programas académicos de las instituciones de Educación Superior". Los exámenes de calidad de la educación superior (ECAES) son una tendencia mundial, de la cual Colombia no es ajena; por el contrario, el Ministerio de Educación Nacional busca garantizar por medio de estas prácticas la calidad en el nivel educativo. En este artículo se identificaron patrones de rendimiento académico en las competencias genéricas de Lectura Crítica y Comunicación Escrita a partir de las bases de datos de las pruebas Saber Pro que presentaron los estudiantes colombianos entre los años 2012 al 2014 utilizando técnicas de minería de datos. Para cumplir este objetivo, y siguiendo las fases de la metodología CRISP-DM, se hizo un análisis de las pruebas Saber Pro y de las bases de datos para tener un conocimiento del negocio y de la información de las pruebas, luego se construyó un repositorio inicial el cual, sirvió sirve de base para la aplicación de un proceso de ETL para construir un repositorio final (limpio y transformado) que contiene los factores socioeconómicos, académicos e institucionales de los estudiantes que presentaron estas pruebas. A este repositorio se le aplicaron técnicas de minería de datos para descubrir patrones de rendimiento académico en estas pruebas. Finalmente, se evaluaron e interpretaron los resultados obtenidos. El conocimiento obtenido sirve como base para realizar recomendaciones que ayuden a los entes gubernamentales e instituciones de educación superior a la toma de decisiones con el fin de mejorar la calidad de la educación superior en Colombia.

Palabras clave: Pruebas Saber Pro, minería de datos, rendimiento académico.

Abstract: The Law 30 of 1992 December 28th (Fundamentals of Higher Education), mentions in Article 31 the need to: "Strive for the creation of mechanisms to evaluate the quality of the academic programs of higher education institutions". The quality exams of Higher Education (ECAES) are a global trend of which Colombia is no stranger; on the contrary, the Ministry of National Education seeks to guarantee through these practices quality at the educational level. In this paper, they were identified patterns of academic performance in the generic competences of Critical Reading and Written Communication from the databases of the Saber Pro tests that Colombian students presented in the period from 2012 to 2014 using mining techniques. data. In order to fulfill this objective, and following the phases of the CRISP-DM methodology, an analysis of the Saber Pro tests and of the databases was made in order to have a knowledge of the business and the information of the tests, then a repository was built initial, which served as the basis for the application of an ETL process to build a final repository (clean and transformed) that contains the socioeconomic, academic and institutional factors of the students who submitted these tests. Data mining techniques were applied to this repository to discover patterns of academic performance in these tests. Finally, the results obtained were evaluated and interpreted. The knowledge obtained serves as a basis for making recommendations that help governmental entities and institutions of higher education to make decisions in order to improve the quality of higher education in Colombia.

Keywords: Saber Pro examination, data mining, academic performance.

Introducción

Al analizar la calidad educativa se debe considerar la existencia de un sistema de evaluación que brinde información sobre los aprendizajes y determine qué se aprende y en qué condiciones se aprende, no únicamente sobre la proporción de asistencia a clase, para determinar niveles aceptables de adquisición de conocimientos (Toranzos, 2017).

Los factores sociales, económicos, académicos e institucionales afectan positiva o negativamente la calidad de la educación de un individuo. Algunos estudios realizados para la ciudad de Medellín en el año 2008 (Tobón, Valencia, Ríos y Bedoya, 2008) complementan la información del ICFES con una encuesta aplicada a profesores y directores, y profundizan el análisis de las características de

ISSN: 2027-8101 - e-ISSN: 2619-5232. Enero-diciembre 2019 • página 53 de 154

la educación asociada al logro. En general, se identificaron factores que afectan positivamente el desempeño como el buen ambiente institucional y la relación profesor/estudiante. Se identificaron, además, factores negativos como las carencias afectivas, nutricionales y cognitivas del estudiante. De igual manera, se destaca que las instituciones públicas obtienen en promedio puntajes más bajos respecto a los planteles privados, están mejor dotadas de capital humano, hecho relacionado con los incentivos salariales, aunque este aspecto no redunda en el mejoramiento de la calidad entendida como el puntaje obtenido en el examen (Tobón *et al.*, 2008).

Igualmente, se observa que existen diferencias en la calidad de la educación lo que ocasiona que los individuos tengan brechas en la calidad de vida, en las posibilidades de acceso a bienes/servicios y en los ingresos. La educación, diferencial en calidad, en vez de ayudar a cerrar las brechas y reducir las diferencias, las profundiza y las perpetúa (Sarmiento, Becerra y González, 2000; Duarte, Bos y Moreno, 2009).

Por otro lado, las pruebas Saber Pro surgen con el objetivo de comprobar el grado de desarrollo de las competencias de estudiantes próximos a culminar el plan de estudios de programas profesionales. Estas pruebas evalúan competencias genéricas y específicas. Los estudiantes deben presentar los módulos de competencias genéricas sin importar el programa de formación que cursen, lo cual para las pruebas Saber Pro 2012-2014 incluye competencias de lectura crítica, comunicación escrita, razonamiento cuantitativo e inglés (ICFES, 2011).

Los estudios realizados hasta el momento, con excepción de Timarán, Hernández, Caicedo, Hidalgo y Alvarado (2015), se han centrado en el análisis estadístico de los resultados y no han incluido técnicas de minería de datos que permitan identificar patrones asociados al rendimiento académico. El estudio en mención se centra en los resultados de las pruebas Saber Pro aplicadas en el segundo semestre de 2011, donde falta evaluar la competencia genérica de Competencias Ciudadanas.

Este artículo presenta el desarrollo de un proyecto enfocado al análisis de los resultados obtenidos en las competencias genéricas de lectura crítica y comunicación escrita de las pruebas Saber Pro, presentadas en el periodo comprendido entre los años 2012 y 2014. En la sección 1 se aborda la metodología usada para el desarrollo del trabajo, los antecedentes, el marco teórico y el proceso de ingeniería aplicado. En la sección 2 se presentan los resultados obtenidos en el proceso de análisis; y finalmente en la sección 3 se presentan las conclusiones.

Metodología

En esta sección se describe el proceso de desarrollo del trabajo, donde se presentan los antecedentes y el fundamento teórico que brinda soporte al dominio del problema y la solución propuesta. Luego se presenta el proceso de ingeniería utilizado, el cual se basa en la metodología CRISP-DM, y se incluyen los elementos relevantes del desarrollo del trabajo.

Antecedentes

En el año 2010, Álvaro Jiménez y Hugo Álvarez (Jiménez y Álvarez, 2010) destacaron las ventajas de utilizar técnicas de minería de datos en los entornos educativos en lugar de los paradigmas de investigación tradicional, teniendo en cuenta el incremento del interés por este tipo de estudios. El estudio se centró en diferentes enfoques que se pueden aplicar en el ámbito educativo y resaltó los métodos de inferencia mediante el empleo de modelos y la destilación de datos, como aquellos de particular importancia dentro de la minería de datos educacionales.

Por su parte, Márquez, Romero y Ventura llevaron a cabo en el año 2012 un estudio utilizando técnicas de minería de datos para predecir el fracaso escolar, para lo cual utilizaron datos de 670 escuelas del estado de Zacatecas, México, y mediante reglas de inducción y árboles de decisión construyeron un modelo para predecir el fracaso escolar y el abandono. El estudio concluye que, en efecto, los algoritmos de clasificación pueden utilizarse con éxito para predecir el rendimiento

académico de los estudiantes (Márquez, Romero y Ventura, 2012).

En Colombia, el Instituto Colombiano para la Evaluación de la Educación Superior (ICFES) presenta de forma frecuente diferentes análisis sobre los resultados de las pruebas Saber Pro. Estos informes contienen información de instituciones, de nivel regional, nacional y por año, entre otros datos, y basan su análisis sólo en interpretación estadística de los resultados; sin embargo, les falta realizar una exploración profunda que permita identificar patrones de desempeño a través de técnicas como la minería de datos.

Con base en los resultados de las pruebas Saber Pro 2011-2, Timarán *et al.* llevaron a cabo un estudio para descubrir patrones de desempeño académico en competencias genéricas de los estudiantes de programas profesionales en las pruebas de este periodo (Timarán *et al.,* 2015). Este estudio aplicó la metodología CRISP-DM mediante técnicas predictivas de minería de datos (clasificación por árboles de decisión), y entre los resultados obtenidos se destaca la importancia de la acreditación institucional como un factor que influye en el desempeño de los estudiantes en las diferentes competencias genéricas.

Fundamento teórico

Pruebas Saber Pro.
Las pruebas Saber Pro son un conjunto de exámenes aplicados por el ICFES, que buscan evaluar los resultados de la calidad de la educación superior en Colombia. Actualmente, estas pruebas se aplican cada año a los estudiantes que están próximos a finalizar los estudios de los programas académicos de pregrado en las instituciones de educación superior (IES), como un requisito adicional para obtener el grado académico (ICFES, 2011). Los objetivos principales de las pruebas Saber Pro son:

- Comprobar el desarrollo de las competencias de los estudiantes próximos a culminar sus programas académicos.
- Generar indicadores de valor agregado de la educación superior. Otorgar insumos para realizar comparaciones entre programas e instituciones.
- Proveer información para la construcción de indicadores de evaluación de la calidad de los programas e Instituciones de Educación Superior (ICFES, 2017a).

Aunque las pruebas Saber Pro se reglamentaron mediante el Decreto 1781 de 2003, fue sólo a partir del año 2009 que se fijaron los parámetros y criterios para organizar el sistema de evaluación de los resultados de la calidad de la educación superior a través de la Ley 1324 y el Decreto 3963. De acuerdo con este marco normativo, los exámenes de Estado se convierten en un instrumento para que el Ministerio de Educación Nacional (MEN) obtenga información que permita mejorar la calidad de la educación (ICFES, 2011, 2017b; Congreso de Colombia, 2009; MEN, 2003, 2009).

Las pruebas Saber Pro evalúan dos tipos de competencias: genéricas y específicas. Las competencias genéricas son aquellas que deben desarrollar los estudiantes, independiente de su programa académico, y se evalúan con los siguientes módulos: Comunicación escrita, Razonamiento cuantitativo, Lectura crítica, Competencias ciudadanas e Inglés. Las competencias específicas, por su parte, son aquellas que dependen de los elementos disciplinares propios de cada programa o área de formación y se evalúan en función de la naturaleza del programa de pregrado a partir de un conjunto de 41 módulos específicos entre los que se encuentran: Análisis de problemáticas psicológicas, Análisis económico, Atención en salud medicina, Atención en salud y Comunicación jurídica, entre otras (ICFES, 2017b).

Proceso de Descubrimiento de Conocimiento.
El proceso de descubrimiento de conocimiento (KDD, por sus siglas en inglés —Knowledge Discovery from Databases—) es la extracción no trivial de información implícita, previamente desconocida y potencialmente útil de datos (Frawley, Piatetsky y Matheus, 1992). Al proceso de descubrimiento de conocimiento es común conocerlo con el nombre de minería de datos, aunque en realidad la minería de datos es una de

ISSN: 2027-8101 - e-ISSN: 2619-5232. Enero-diciembre 2019 • página 55 de 154

las cinco fases que componen todo el proceso, las cuales son:

- Integración y recopilación
- Selección, limpieza y transformación
- Minería de datos
- Evaluación e interpretación
- Difusión y uso

En la Figura 1 se presentan las fases del proceso de descubrimiento de conocimiento, el cual es de naturaleza iterativa e interactiva. Iterativa, porque el resultado de una fase puede servir como realimentación de fases previas y ajustar la información útil para la obtención de conocimiento. Interactiva, porque requiere una participación activa de un experto con amplio conocimiento sobre la información en todas las fases del proceso.

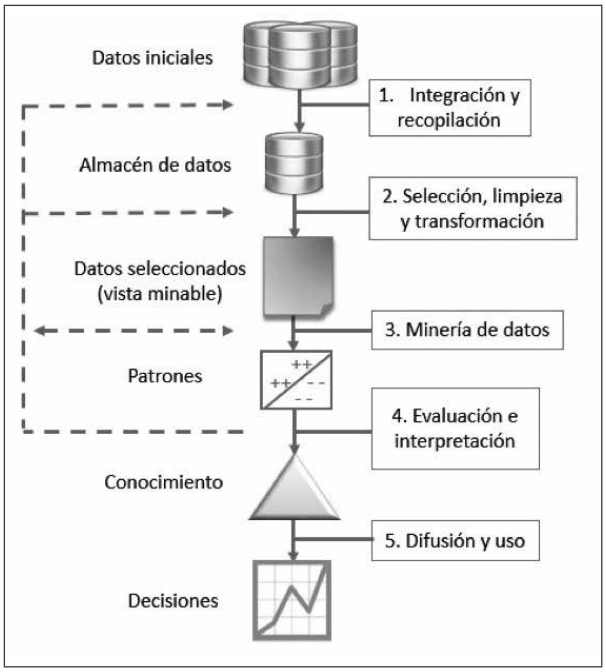

Figura 1. Fases del Proceso de Descubrimiento de Conocimiento (KDD)
Fuente: (Hernández, Ramírez y Ferri, 2010)

En la fase de integración y recopilación se identifican los orígenes de la información que puede ser útil y ésta se recopila en un almacén de datos o repositorio unificado. Esta tarea representa un gran reto, ya que los orígenes de datos pueden no estar en un mismo sistema de base de datos o incluso no tener ningún soporte digital. En la fase de selección, limpieza y transformación se corrigen datos —o se eliminan, según el caso— y se decide la estrategia que se utilizará para el tratamiento de los datos incompletos. Durante la fase de minería de datos se decide qué tareas se utilizarán para la identificación de patrones, pueden ser tareas predictivas o tareas descriptivas. En la fase de evaluación e interpretación se evalúa la calidad de los patrones identificados en la fase anterior. Finalmente, en la fase de difusión se hace uso del nuevo conocimiento mediante la toma de decisiones y la aplicación de las recomendaciones obtenidas.

CRISP-DM.

Para el desarrollo del proyecto se tomó como punto de referencia las fases de la metodología *Cross Industry Standard Process for Data Mining* (CRISP-DM), que constituyeron la base para identificar patrones de rendimiento académico en las pruebas Saber Pro, en el período comprendido entre 2012 y 2014, en las competencias de Lectura crítica y Comunicación escrita (Shearer, 2000). La metodología incluye un modelo de proceso de minería de datos que describe los enfoques comunes que utilizan los expertos en esta técnica. El portal Knowledge Discovery Nuggets (KDNuggets), en los años 2002, 2004, 2007 y 2014, presentó varias encuestas que indican que la metodología CRISP-DM es la más utilizada para los proyectos de minería de datos en comparación con SEMMA[1] u otras metodologías propias (KDNuggets, 2002; 2004; 2007; 2014). Además, en una revisión y crítica de los modelos de minería de datos en 2009 a CRISP-DM se le llamó el "estándar de facto para el desarrollo de la minería de datos y los proyectos de descubrimiento de conocimiento" (Marbán, Mariscal y Segovia, 2017).

En la Figura 2 se presentan las seis fases que componen la metodología CRISP-DM: Comprensión del problema o negocio, Comprensión de datos,

1 SEMMA (por sus siglas en inglés: Sample, Explore, Modify, Model, and Assess) es una metodología de minería de datos propuesta por la empresa SAS Institute Inc.

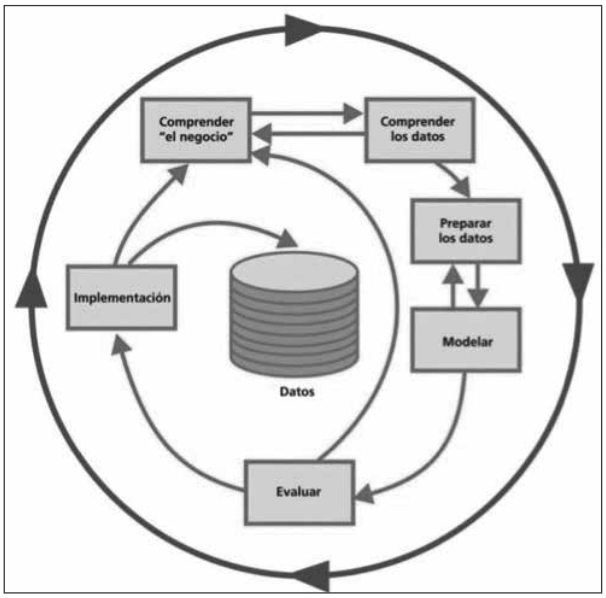

Figura 2. Relación entre las diferentes fases de CRISP-DM Fuente: IBM (2016, p. 1)

Preparación de datos, Modelado, Evaluación e implementación (Harper y Pickett, 2006):

Comprensión del problema o negocio: Esta fase se centró en la comprensión de las pruebas Saber Pro, inicialmente desde un enfoque legal e histórico y luego desde el punto de vista de los componentes de las pruebas, en especial de las competencias genéricas que están dentro del alcance del proyecto: Lectura crítica y Comunicación escrita.

Comprensión de datos: En esta fase se recopiló la información existente sobre los resultados de las pruebas y se analizaron la base de datos y todos los atributos de información. Lo anterior permitió obtener conocimiento de los datos, entender cómo están distribuidos e identificar cuáles son los procesos que se deben aplicar sobre ellos.

Preparación de datos: Durante esta fase se realizó todo el proceso de transformación de los datos (ETL) de forma repetitiva y sin ningún orden diferente a la necesidad de transformación de cada atributo. Además, se genera el repositorio inicial de datos en un repositorio final, limpio y transformado, sobre el cual se construye el modelo predictivo.

Modelado: En esta fase se utilizaron técnicas predictivas de minería de datos para construir un modelo óptimo que brinda información útil sobre los factores socioeconómicos que influyen en el desempeño académico. Además, se consideraron las dos competencias genéricas (Lectura crítica y Comunicación escrita) para construir un modelo para cada caso, ya que los factores que influyen en el desempeño de los estudiantes en una de las competencias no influyen necesariamente en la otra.

Evaluación: En esta fase se obtuvieron los modelos y se interpretaron desde la perspectiva de los patrones identificados en los árboles de clasificación desde un enfoque dimensional, a partir de la construcción de un cubo OLAP[2] para el análisis del desempeño en la dimensión más representativa de cada árbol.

Implementación: En esta fase se hicieron un conjunto de recomendaciones basadas en los patrones de desempeño identificados, las cuales sirven de soporte para que las instituciones de educación superior, el Ministerio de Educación y el Gobierno Nacional tomen las decisiones estratégicas que ayuden a mejorar la calidad de la educación en los niveles local, regional y nacional.

Herramientas tecnológicas empleadas

Para cumplir a cabalidad con los objetivos propuestos en el proyecto, se utilizaron las siguientes herramientas tecnológicas, por la funcionalidad y las características que se mencionan:

PostgreSQL: Es un sistema de base de datos relacional de código abierto que cuenta con más de 15 años de desarrollo activo y una arquitectura sólida y comprobada con la que ha obtenido una reputación de fiabilidad, seguridad, integridad y confiabilidad de los datos. En este motor de bases de datos se realizó la gran mayoría de operaciones de ETL[3] sobre el repositorio de datos de las pruebas Saber Pro porque es una herramienta

2 Cubo OLAP: On Line Analytical Processing o procesamiento Analítico en Línea.

3 ETL es la sigla para el proceso de datos que se realiza durante los procesos de construcción de bodegas de datos.

ISSN: 2027-8101 - e-ISSN: 2619-5232. Enero-diciembre 2019 • página 57 de 154

capaz de realizar sin mayores inconvenientes todas las transacciones necesarias sobre el gran volumen de datos del repositorio.

Waikato Environment for Knowledge Analysis (WEKA): Es un entorno para experimentación de análisis de datos que permite aplicar, analizar y evaluar las técnicas más relevantes de análisis de datos, principalmente las que provienen del aprendizaje automático sobre cualquier conjunto de datos del usuario. Para ello únicamente se requiere que los datos para analizar se almacenen con un cierto formato, conocido como ARFF (*Attribute-Relation File Format*). WEKA se distribuye como software de libre distribución desarrollado en Java y está constituido por una serie de paquetes de código abierto con diferentes técnicas de preprocesado, clasificación, agrupamiento, asociación y visualización, así como facilidades para su aplicación y análisis de prestaciones cuando son aplicadas a los datos de entrada seleccionados. Estos paquetes se pueden integrar en cualquier proyecto de análisis de datos, e incluso pueden extenderse con contribuciones de los usuarios que desarrollen nuevos algoritmos. Con el objeto de facilitar su uso por un mayor número de usuarios, WEKA incluye una interfaz gráfica de usuario para acceder y configurar las diferentes herramientas integradas (García y Molina, 2012).

PDI Pentaho (Data Integration Previous Kettle): Pentaho es una *suite* de herramientas de inteligencia de negocios que tiene dos versiones, la versión comercial y la versión *open source;* la segunda opción es la utilizada en el proyecto. *Pentaho Data Integration* (cuyo nombre clave es Kettle), es una herramienta de la *suite* de Pentaho de las que se denominan ETL (*Extract – Transform – Load*), es decir, una herramienta de extracción de datos de una fuente, transformación de esos datos y su carga en otro sitio. Con el uso de Kettle se pueden evitar grandes cargas de trabajo manual frecuentemente difícil de mantener y de desplegar. PDI es el componente de Pentaho responsable de la extracción, transformación y procesos de carga (ETL). Aunque las herramientas de ETL se utilizan con mayor frecuencia en entornos almacenes de datos, PDI también se puede utilizar para otros fines:

- Migración de datos entre aplicaciones o bases de datos
- Exportación de datos desde bases de datos a archivos planos
- Carga de datos en bases de datos de forma masiva
- Limpieza de datos
- Integración de aplicaciones

PDI es fácil de usar. Cada proceso se crea con una herramienta gráfica donde se especifica qué hacer sin necesidad de escribir código para indicar cómo hacerlo; por eso, se podría decir que la PDI está orientada a metadatos. Además, se puede utilizar como una aplicación independiente, o como parte de la más grande Pentaho Suite. Como una herramienta ETL, es la herramienta de código abierto más popular disponible. PDI es compatible con una amplia gama de formatos de entrada y salida, incluyendo archivos de texto, hojas de datos y motores de bases de datos comerciales y gratuitos. Por otra parte, las capacidades de transformación de PDI le permiten manipular los datos con pocas limitaciones" (Curto Díaz, 2010).

Modelado

La construcción del modelo predictivo mediante árboles de clasificación se realizó sobre las competencias de Lectura crítica y Comunicación escrita. Los resultados en el repositorio final están clasificados como "Sobre la media" y "Bajo la media". Este repositorio cuenta con un total de 711.604 registros. Mediante la aplicación de técnicas predictivas de minería de datos se obtuvieron dos modelos que permiten predecir el buen o mal desempeño académico en las competencias de Lectura crítica y Comunicación escrita, con base en los diferentes factores socioeconómicos de índole demográfico, social, económico, familiar y académico, de acuerdo con los resultados de las pruebas Saber Pro 2012-2014.

La técnica de clasificación que se utilizó para encontrar patrones de desempeño académico fue la de árboles de decisión, que permite pronosticar resultados futuros de acuerdo con la clasificación obtenida durante una fase de entrenamiento, el cual construye un conjunto de reglas que es conocido como "árbol de decisión". Para la aplicación de esta técnica se carga la base de

datos dentro de la herramienta WEKA. Por cada competencia analizada fue necesario ejecutar el proceso de modelado. La primera vez se dejó como única clase el atributo de Lectura crítica para el entrenamiento del árbol. Al finalizar se repitió el proceso de entrenamiento, pero utilizando como única clase el atributo de Comunicación escrita, y así se encontró el segundo modelo.

Durante el entrenamiento de los árboles se empleó la estrategia de validación cruzada y se utilizan 10 iteraciones (o pliegues). Con esta estrategia se reduce la dependencia del resultado entre los conjuntos de entrenamiento y de prueba. De igual manera, se utilizó el algoritmo C4.5 para obtener el modelo de clasificación. En WEKA se encuentra una implementación de este algoritmo que es conocida como J48. "El algoritmo J48 se basa en la utilización del criterio del coeficiente de ganancia de información (*information gain ratio*). De esta manera, se consigue evitar que las variables con mayor número de posibles valores salgan beneficiadas en la selección" (Timarán *et al.*, 2016).

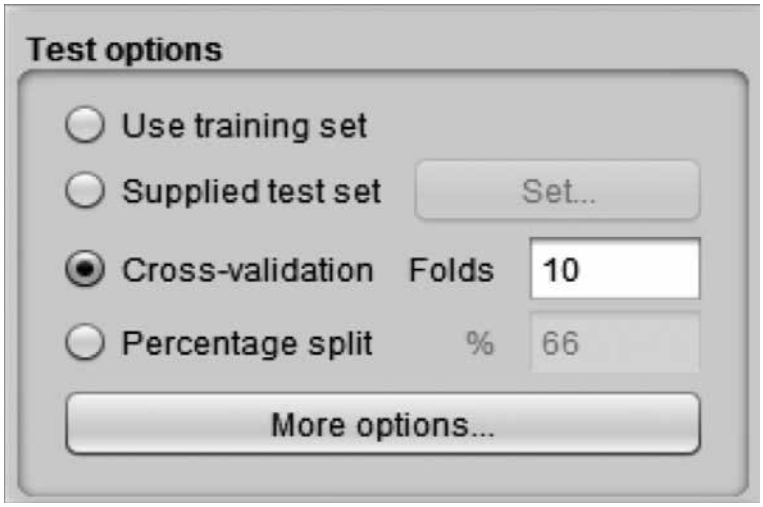

Figura 3. Opciones de prueba para el entrenamiento de los árboles en WEKA.
Fuente: Elaboración propia

Lectura crítica

El primer modelo que se construyó permite identificar los patrones socioeconómicos asociados al desempeño en la competencia genérica de Lectura crítica. De acuerdo con el diccionario de datos del repositorio final, el atributo que almacena esta información es mod_lectura_critica_desemp y sus valores son "SOBRE LA MEDIA" y "BAJO LA MEDIA", así indica el desempeño general del estudiante en esta competencia.

El primer paso es cargar los datos del repositorio final en WEKA y se eliminan los atributos que no serán objeto del presente análisis. El atributo seleccionado como "clase" para la construcción del modelo es Lectura crítica. En la Figura 4 se presenta la distribución de valores sobre la media y bajo la media para la clase mod_lectura_critica_desemp.

ISSN: 2027-8101 - e-ISSN: 2619-5232. Enero-diciembre 2019 • página 59 de 154

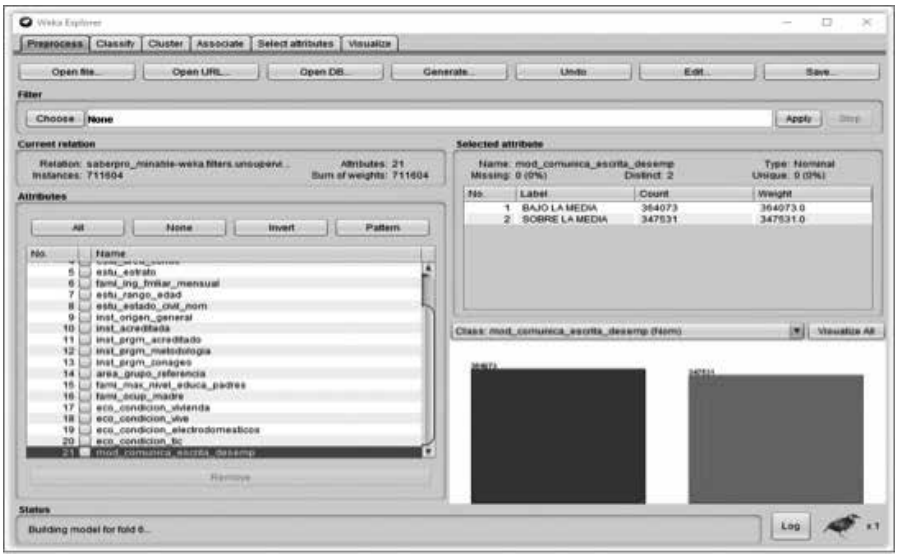

Figura 4. Información de preprocesamiento para la competencia Lectura crítica.
Fuente: Elaboración propia

Durante la etapa de entrenamiento es necesario múltiples ajustes para calibrar el modelo realizando operaciones de prepoda para el factor M y el factor de confidencia. Los diferentes valores utilizados durante la poda se presentan en la Tabla 1. En cada iteración se indica el porcentaje de instancias clasificadas correcta e incorrectamente, así como el tamaño del árbol y el número de hojas obtenido. También se encontró que el más óptimo es aquel que cuenta con un factor de confidencia de 0,25 y un factor M = 4 %. Esto significa que, de 711.604 registros, el valor de M es 28.464 y se obtiene un árbol con un total de 21 hojas y con un tamaño de 28. Una vez construido el árbol, se aplicó un proceso de postpoda para dejar las ramas que brindan mayor información (ramas más representativas) mediante la selección de las reglas que sobrepasan un soporte mínimo del 0,5 % y una confianza del 60 %.

Tabla 1. Ajuste de valores del factor M y factor de confidencia para el entrenamiento del modelo (Lectura crítica)

Porcentaje	Factor Confidencia	M	Instancias correctamente clasificadas	%	Instancias incorrectamente clasificadas	%	#Hojas	Tamaño Árbol	A (Sobre la media)	B (Bajo la media)
0,5%	0,25	3558	457290	64,2619%	254314	35,7381%	75	93	210328	145548
	0,4		457068	64,2307%	254536	35,7693%	82	62	213840	142036
	0,5		457048	64,2279%	254556	35,7721	116	145	214435	141441
	0,6		457049	64,2280%	254555	35,7720	140	174	214396	141480
1%	0,25	7116	453379	63,7123%	258225	36,2877	57	71	208503	147373
	0,4		453528	63,7332%	258076	36,2668	66	83	208808	147068
	0,5		453550	63,7363%	258054	36,2637	70	89	299004	146872
	0,6		453546	63,7357%	258058	36,2643	70	89	209005	244541
	0,19		453410	63,7120%	258227	36,2880	51	64	207805	148071

Porcentaje	Factor Confidencia	M	Instancias correctamente clasificadas	%	Instancias incorrectamente clasificadas	%	#Hojas	Tamaño Árbol	A (Sobre la media)	B (Bajo la media)
1,5%	0,25	10674	452146	63,5390%	259458	36,4610	56	70	200821	155055
	0,4		452168	63,5421%	259436	36,4579	61	77	201274	154602
	0,5		452221	63,5495%	259383	36,4505	61	77	201667	154209
	0,6		452215	63,5487%	259389	36,4513	61	77	201665	154211
2%	0,25	14232	449286	63,1317%	262318	36,8629	40	50	207878	147998
	0,4		449360	63,1475%	262244	36,8525	42	53	209116	146760
	0,5		449375	63,1496%	262229	36,8504	47	59	209285	146591
	0,6		449375	63,1496%	262229	36,8504	47	59	20285	146591
	0,19		449284	63,1368%	262320	36,8632	34	43	207180	148696
2,5%	0,25	17790	444422	62,4536%	267182	37,5664	33	41	208920	267182
	0,4		444500	62,4645%	267104	37,5664	35	44	208390	147486
	0,5		444480	62,4617%	267124	37,5383	37	47	207173	148703
	0,6		444480	62,4617%	247124	37,5383	37	47	207173	148703
	0,19		444419	62,4531%	267185	37,5496	27	34	208415	177461
3%	0,25	21348	444085	62,4062%	267519	37,5938	35	45	211087	144789
3,5%	0,25	24906	444059	62,4025%	267545	37,5975	25	33	209797	146079
4%	0,25	28464	439591	61,7747%	272013	38,2253	21	28	210211	145665
5%	0,25	35580	483600	61,6354%	273004	383646	13	17	210373	145503
7%	0,25	49812	434224	61,0205%	277380	38,9795	9	12	187279	168597
10%	0,25	71160	434224	61,0205%	277380	38,9795	9	12	187279	168597

Fuente: Elaboración propia

Modelo obtenido

La Figura 5 presenta la precisión del modelo obtenido que indica que se tiene una precisión del 61,77 %. También, se incluye la matriz de confusión, la cual indica que el modelo clasifica el desempeño en la competencia de Lectura crítica de forma correcta para 210.211 sobre la media y 229.380 casos bajo la media. Además, clasifica incorrectamente 145.665 casos de estudiantes sobre la media y 126.348 casos bajo la media. Esto significa que el modelo clasifica correctamente al 59,1 % de los estudiantes que están sobre la media en la competencia lectura crítica y al 64,5 % que están bajo la media en esta competencia. En la Figura 6 se presenta el árbol que se obtiene, donde se identifica que el atributo que más información brinda es la acreditación institucional, razón por la cual se convierte en la raíz del árbol según el algoritmo de clasificación utilizado.

```
Time taken to build model: 10.42 seconds

=== Stratified cross-validation ===

Correctly Classified Instances      439591
61.7747 %
Incorrectly Classified Instances         7
38.2253 %
Kappa statistic                          0.2355
Mean absolute error                      0.4645
Root mean squared error                  0.4819
Relative absolute error                 92.8902 %
Root relative squared error             96.3819 %
Total Number of Instances               711604
```

Figura 5. Información sobre la precisión y matriz de confusión del árbol de Lectura crítica.

Fuente: Elaboración propia

ISSN: 2027-8101 - e-ISSN: 2619-5232. Enero-diciembre 2019 • página 61 de 154

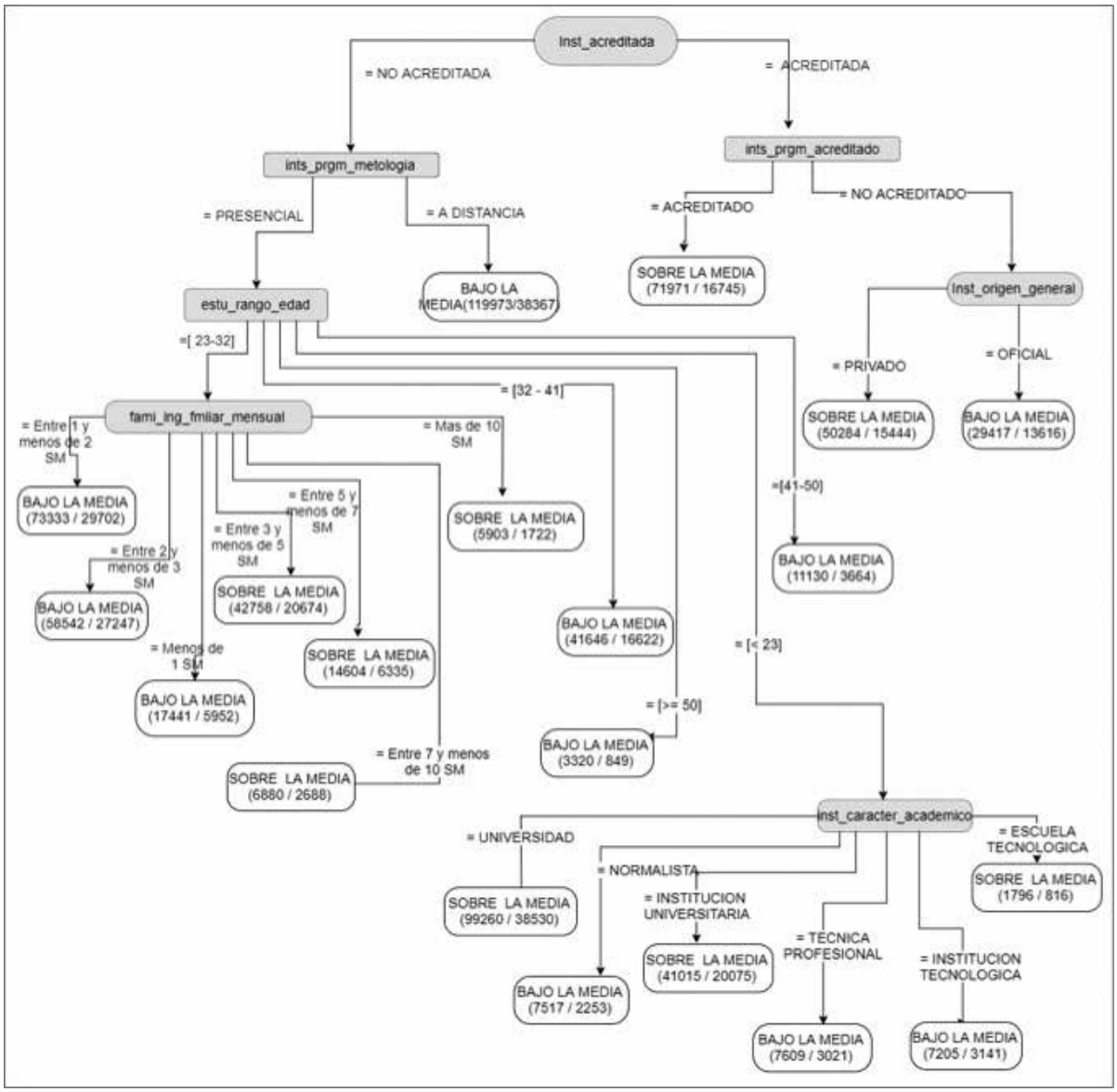

Figura 6. Árbol obtenido para la competencia de Lectura crítica.
Fuente: Elaboración propia

Análisis dimensional

Una forma de comprender el comportamiento de los datos es mediante las diferentes dimensiones que lo componen, consiste en realizar un análisis dimensional de los registros. De esta forma se puede obtener una vista del desempeño global de los estudiantes en las pruebas Saber Pro 2012-2014 a través de los diferentes factores socioeconómicos.

Modelado dimensional

Por medio de la estrategia de modelado de datos dimensionales se construyó una bodega de datos compuesta por una tabla de hechos principal, relacionada con ocho dimensiones que representan los factores socioeconómicos de los estudiantes que presentaron las pruebas Saber Pro en los años 2012-2014. La Figura 7 contiene el modelo dimensional correspondiente.

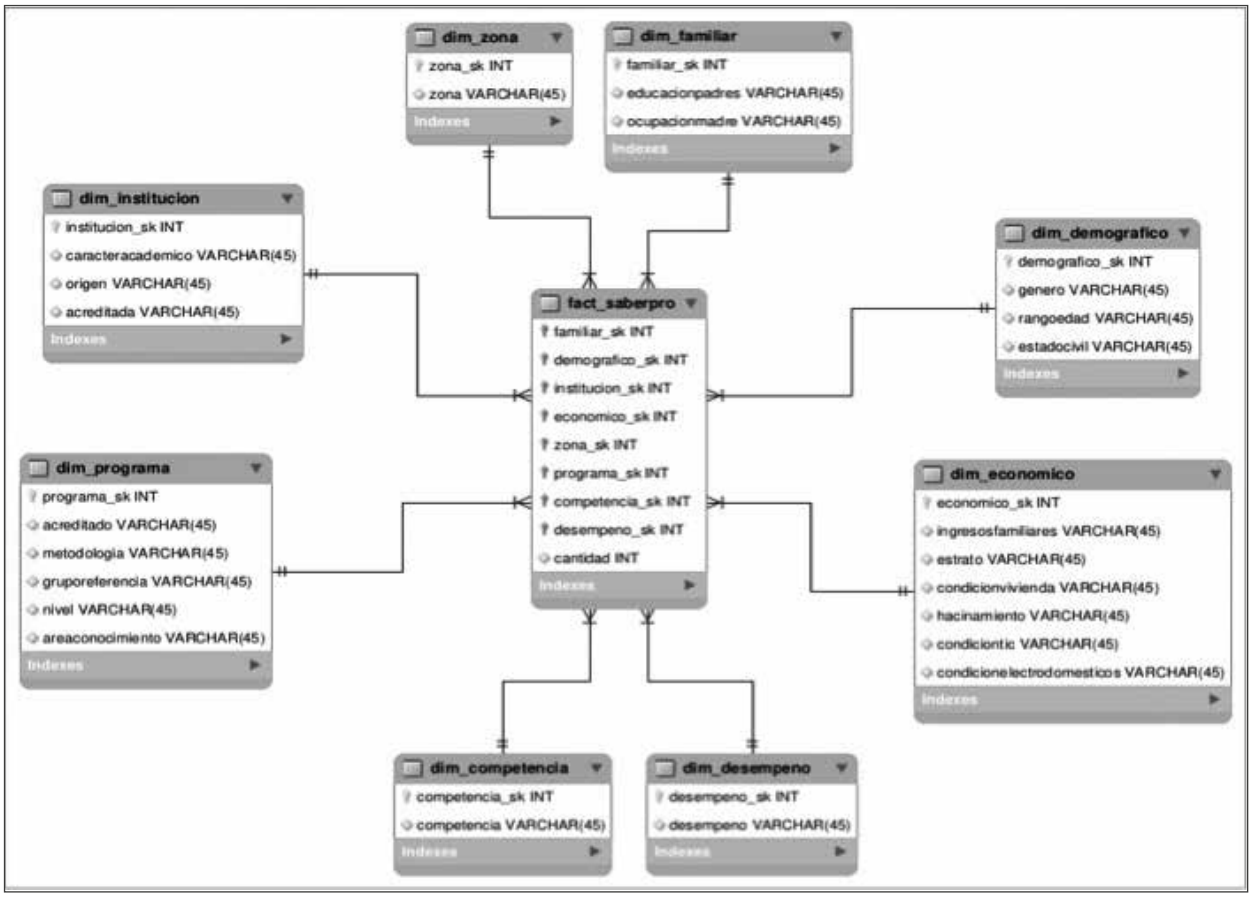

Figura 7. Modelo dimensional correspondiente a la bodega de datos construida para el análisis de datos.
Fuente: Elaboración propia

Proceso ETL

A partir del modelo construido fue necesario realizar una transformación para procesar los datos desde el repositorio final construido durante la fase de Preparación de los datos hasta la bodega de datos. En la Figura 8 se presentan los pasos ejecutados durante el proceso de transformación con la herramienta PDI (*Pentaho Data Integration*).

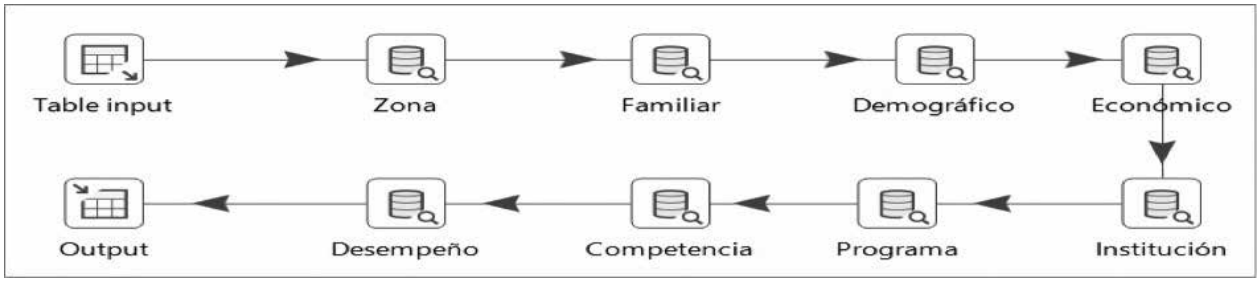

Figura 8. Proceso ETL automatizado en la herramienta Pentaho Data Integration (PDI) utilizado en la creación de la bodega de datos para análisis.
Fuente: Elaboración propia

ISSN: 2027-8101 - e-ISSN: 2619-5232. Enero-diciembre 2019 • página 63 de 154

Cubo OLAP

Los cubos de datos son estructuras que proporcionan herramientas para el análisis de gran cantidad de datos a partir de una estructura de datos sencilla, de este modo superan las limitaciones propias de las bases de datos relacionales. Para realizar el análisis dimensional de los resultados de las pruebas Saber Pro, se construye un cubo de datos que representa la estructura propuesta en el modelo dimensional que se presenta en la Figura 7. El cubo se construyó utilizando la herramienta *Schema Workbench* (de Pentaho). En la Figura 9 se presenta una captura de pantalla con la forma parcial de la configuración del cubo y sus dimensiones.

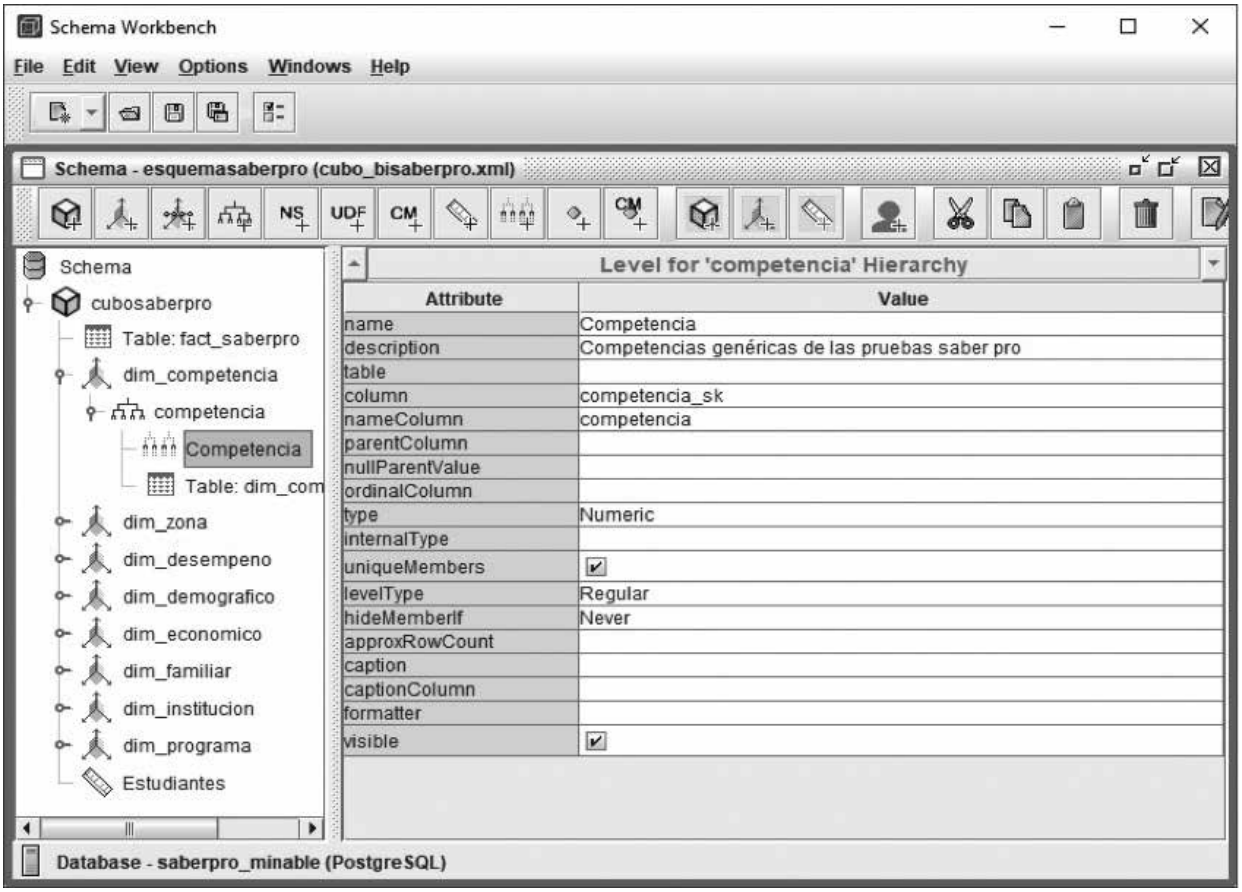

Figura 9. Configuración del cubo para consultas OLAP.
Fuente: Captura de pantalla de Schema Workbench

Resultados

Una vez construido el cubo y publicado en el servidor de Pentaho (*Pentaho BI Server*) es posible realizar diferentes operaciones de análisis descriptivo de los datos. A continuación se describen los resultados obtenidos.

Desempeño general. El desempeño de los estudiantes en las competencias genéricas de comunicación escrita y lectura crítica se mide cualitativamente como "Sobre la media" y "Bajo la media". En la Tabla 2 se indica la cantidad de estudiantes que se encuentran sobre la media y bajo la media en las dos competencias genéricas analizadas.

Tabla 2. Desempeño en las competencias genéricas de comunicación escrita y lectura crítica

Desempeño	Comunicación Escrita	Lectura Crítica
Sobre la media	347.531	355.876
Bajo la media	364.073	355.728
Total	711.604	711.604

Fuente: Elaboración propia

En la Figura 10 se presenta el desempeño obtenido por los estudiantes, el cual, como es de esperarse, tiene una distribución más o menos equilibrada, ya que el análisis se realiza con base en el promedio general.

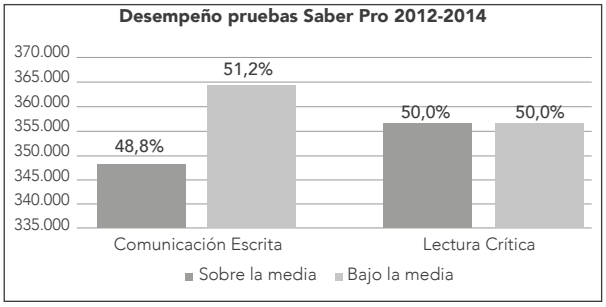

Figura 10. Distribución del desempeño en las pruebas Saber Pro para las competencias genéricas de Comunicación escrita y Lectura crítica.
Fuente: Elaboración propia

Análisis dimensional de los árboles de clasificación.

Los modelos obtenidos para las competencias de Lectura crítica y Comunicación escrita indican que el atributo que aporta mayor información sobre el desempeño de los estudiantes es la institución educativa de donde provienen, en particular cuando la institución es o no acreditada, lo que representa un indicio del impacto de la acreditación institucional en la calidad de la educación. Tomando como base la acreditación institucional, a través del modelo dimensional se puede confirmar que cuando la institución está acreditada, el 68,4 % de los estudiantes obtienen una calificación por encima de la media en la competencia Lectura crítica y un 58,8 % por encima de la media en Comunicación escrita. Mientras que cuando la institución no está acreditada hay un mayor porcentaje de estudiantes que se encuentran por debajo de la media, más específicamente de 55 % para la competencia Lectura crítica y de 53,9 % en la Comunicación escrita.

Conclusiones

- La aplicación correcta de los procesos de ETL ayuda a transformar los datos de forma tal que se faciliten los procesos de análisis e interpretación de la información.
- El uso de las técnicas predictivas de minería de datos es útil para identificar información que en ocasiones es desconocida y funciona mejor con grandes volúmenes de datos, como el de las pruebas Saber Pro.
- Los resultados de este proyecto pueden ser de interés estratégico para el Ministerio de Educación Nacional, ya que realiza de forma continua diferentes estudios e informes sobre los resultados en las pruebas Saber Pro.
- En los resultados obtenidos se puede observar que los factores socioeconómicos influyen en el desempeño de los estudiantes en los resultados de las competencias genéricas de Lectura crítica y Comunicación escrita de las pruebas Saber Pro.
- Existe una gran variedad de herramientas para el análisis de datos que sirven de soporte para la toma de decisiones estratégicas. Entre ellas encontramos las bodegas de datos, la inteligencia de negocios y, por supuesto, la minería de datos.
- El seguimiento de una metodología para el proceso de minería de datos brinda un marco de trabajo organizado que facilita el desarrollo del proyecto y los procesos de descubrimiento de información.

Referencias

Congreso de la República de Colombia (2009). Ley 1324 de 2009, "Por la cual se fijan parámetros y criterios para organizar el sistema de evaluación de resultados de la calidad de la educación, se dictan normas para el fomento de una cultura de la evaluación, en procura de

ISSN: 2027-8101 - e-ISSN: 2619-5232. Enero-diciembre 2019 • página 65 de 154

facilitar la inspección y vigilancia del Estado y se transforma el ICFES". Diario Oficial, 47.409. Santafé de Bogotá. D.C.

Curto, J. (2010). Introducción al Business Intelligence. Barcelona: Editorial UOC.

Duarte, J., Bos, S. y Moreno, M. (2009). Inequidad en los Aprendizajes Escolares en Latinoamérica. Banco Interamericano de Desarrollo. Nota Técnica #4.

Frawley, W., Piatetsky-Shapiro, G. y Matheus, C. (1992). Knowledge Discovery in Databases: An Overview. AI Magazine, 13(3), 58.

García, J. y Molina, J. (2012). Técnicas de análisis de datos. Aplicaciones prácticas utilizando Microsoft Excel y Weka. Madrid: Universidad Carlos III de Madrid.

Harper, G. y Pickett, S. (2006). Methods for mining HTS data. Drug Discovery Today, 11(15-16), 694-699.

Hernández O., Ramírez Q. y Ferri R. (2010). Introducción a la minería de datos. Madrid: Pearson.

IBM (2016). Guía de CRISP-DM de IBM SPSS Modeler. España: IBM Corporation.

Instituto Colombiano para la Evaluación de la Educación (ICFES) (diciembre de 2011). Exámenes de Estado de calidad de la educación superior SABER PRO. Resultados del periodo 2005-2009. Bogotá.

Instituto Colombiano para la Evaluación de la Educación (ICFES) (2017a). Informe nacional de resultados Examen Saber Pro 2016. Bogotá.

Instituto Colombiano para la Evaluación de la Educación (ICFES) (2017b). Módulos Saber Pro 2016-2. Recuperado de http://www.icfes.gov.co/terminos-de-uso/item/1982-modulos-saber-pro-2016-2 [Consultado el 04 de noviembre de 2017].

Jiménez, G. y Álvarez, H. (2010). Minería de Datos en la Educación. Universidad Carlos III de Madrid. Recuperado de http://www.it.uc3m.es/jvillena/irc/practicas/10-11/08mem.pdf

KDNuggets (2002). Poll: What main methodology are you using for data mining? Recuperado de https://www.kdnuggets.com/polls/2002/methodology.htm [Consultado el 3 de noviembre de 2017].

KDNuggets (2004). Poll: Data Mining Methodology. Recuperado de https://www.kdnuggets.com/polls/2004/data_mining_methodology.htm [Consultado el 3 de noviembre de 2017].

KDNuggets (2007). "Poll: Data Mining Methodology. Recuperado de https://www.kdnuggets.com/polls/2007/data_mining_methodology.htm [Consultado el 3 de noviembre de 2017].

KDNuggets (2014). Poll: What main methodology are you using for your analytics, data mining, or data science projects? Recuperado de https://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html [Consultado el 3 de noviembre de 2017].

Marbán, O., Mariscal, G. y Segovia, J. (2017). A Data Mining and Knowledge Discovery Process Model. En J. Ponce & A. Karahoca (Eds.), Data Mining and Knowledge Discovery in Real Life Applications (pág. 442). Viena, Austria: I-Tech.

Márquez V., Romero M. y Ventura S. (2012). Predicción del Fracaso Escolar mediante Técnicas de Minería de Datos. Revista Iberoamericana de Tecnologías del Aprendizaje, 7(3), 109-117.

Ministerio de Educación Nacional de Colombia (MEN) (2003). Decreto 1781, "Por el cual se reglamentan los Exámenes de Estado de Calidad de la Educación Superior, ECAES, de los estudiantes de los programas académicos de pregrado". Bogotá.

Ministerio de Educación Nacional de Colombia (MEN) (2009). Decreto 3963, "Por el cual se reglamenta el Examen de Estado de Calidad de la Educación Superior". Bogotá.

Sarmiento, A., Becerra, L. y González, J. (2000). La incidencia del plantel en el logro educativo del alumno y su relación con el nivel socioeconómico. Coyuntura Social (22).

Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing, 5(4), 13-22.

Timarán, S., Hernández, I., Caicedo, S. J., Hidalgo, A. y Alvarado, J. (2016). Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional. Bogotá: Ediciones Universidad Cooperativa de Colombia. doi: http://dx.doi.org/10.16925/9789587600490

Tobón, D., Valencia, G., Ríos, P. y Bedoya, J. (2008). Organización jerárquica y logro escolar en Medellín. Un análisis a partir de la función de producción educativa. Lecturas de economía, (68), 145-173.

Toranzos, L. (2017). Evaluación y Calidad. Revista Iberoamericana de Educación, (10), 65.