Formación Universitaria
Vol. 12(4), 55-62 (2019)
http://dx.doi.org/10.4067/S0718-50062019000400055

Extracción de Conocimiento para la Predicción y Análisis de los Resultados de la Prueba de Calidad de la Educación Superior en Colombia

José R. García-González[(1)]*, Paola A. Sánchez-Sánchez[(1)], Manuel Orozco[(2)] y Sergio Obredor[(2)]
(1) Facultad de Ingenierías, Universidad Simón Bolívar, Barranquilla – Colombia.
(e-mail: jgarcia122@unisimonbolivar.edu.co; psanchez9@unisimonbolivar.edu.co).
(2) Departamento de Ciencias Básicas, Universidad Simón Bolívar, Barranquilla – Colombia
(e-mail: morozco10@unisimonbolivar.edu.co; sergio.obredor@unisimonbolivar.edu.co).

* Autor a quien debe ser dirigida la correspondencia

Recibido Dic. 4, 2018; Aceptado Feb. 1, 2019; Versión final Mar. 15, 2019, Publicado Ago. 2019

Resumen

Se presentan y analizan los resultados de la prueba de la calidad de la educación superior en Colombia – Saber Pro. Se usó la metodología de extracción de conocimiento en bases de datos KDD sobre la cual se construyó una base de datos del desempeño académico del estudiante en áreas asociadas con los contenidos de la prueba Saber Pro, y se utilizó redes neuronales como técnica para la minería de datos. Las redes neuronales permitieron la predicción de los resultados de la prueba Saber Pro con alta exactitud tanto en rangos cualitativos como cuantitativos. Además, se comprobó una correlación entre el desempeño académico y los resultados de Saber Pro. Los hallazgos sugieren que la metodología usada es una excelente guía para el descubrimiento de patrones ocultos en los datos, y permite establecer estrategias de mejora de los resultados de las pruebas Saber Pro que involucren el desempeño académico del estudiante.

Palabras clave: minería de datos; extracción de conocimiento; bases de datos KDD; Saber Pro; educación superior

Knowledge Capture for the Prediction and Analysis of Results of the Quality Test of Higher Education in Colombia

Abstract

In this paper, the prediction and results analysis of the quality test of higher education in Colombia - Saber Pro is performed. The knowledge extraction in databases methodology KDD was used, on which a database of the student's academic performance was built in areas associated with the contents of the Saber Pro test, and neural networks were used as a technique for data mining. The neural networks allowed the prediction of the results of the Saber Pro test with high exactness in both qualitative and quantitative ranges. A correlation between academic performance and Saber Pro results was also found. The findings suggest that the methodology used is an excellent guide for the discovery of hidden patterns in the data, and allows to establish strategies to improve the results of the Saber Pro tests that involve the student's academic performance.

Keywords: data mining; knowledge extraction; databases methodology KDD; Saber Pro; higher education

INTRODUCCIÓN

Los procesos asociados con la calidad de la educación y formación profesional en las Instituciones de Educación Superior en Colombia (IES), han generado en los últimos años un debate interesante en torno al impacto de sus egresados en los sectores industriales, comerciales y productivos del país. Uno de los apartes interesantes del mismo, se concentra en la dinámica inherente a los exámenes académicos de estado para la medición de la calidad de los procesos de formación profesional en las IES. El ministerio de Educación Nacional MEN, concibe estos exámenes como pruebas académicas de carácter oficial y obligatorio, denominadas genéricamente, Pruebas SABER PRO, las cuales, forman parte de un conjunto de instrumentos que el estado colombiano dispone para evaluar la calidad de la educación formal recibida por quienes culminan los distintos programas académicos profesionales en las Instituciones de Educación Superior en Colombia (Ministerio de Educación Nacional - MEN, 2009).

En dichas pruebas, se evalúan competencias genéricas, las cuales aplican para estudiantes de todos los programas de formación, estas incluyen las áreas de lectura crítica, razonamiento cuantitativo, composición escrita, inglés y competencias ciudadanas, además de la medición de las competencias específicas, estas últimas, para el caso de las ingenierías, son las asociadas y tomadas como insumo para este primer ejercicio de predicción con la herramienta que se discute en este artículo. En correspondencia con lo anterior, existe un marcado interés de las Instituciones de Educación Superior (IES) por mejorar el desempeño académico de los estudiantes que presentan estos exámenes, debido a que los resultados obtenidos en las mismos, son tenidos en cuenta como indicadores que inciden directamente en la medición que realiza en Ministerio de Educación Nacional de Colombia, a través el Instituto Colombiano de Fomento de la Educación Superior ICFES y cuyo impacto se evidencia significativamente en todos y cada uno de los procesos académicos y de formación de dichas Instituciones. Así las cosas, puede entenderse la importancia que los resultados de este tipo de pruebas tienen para las Instituciones de Educación Superior, teniendo en cuenta los aspectos que denotan la relación entre la calidad de un programa académico de formación profesional y los resultados que obtengan sus estudiantes en la prueba de estado SABER PRO, además, de asociar a dicha relación, el desempeño académico del estudiante durante todo su proceso de formación (García, 2008). En consecuencia, en el presente artículo se referencian las diferentes etapas y fases que se llevaron a cabo durante todo el proceso, las cuales incluyeron desde luego, los análisis sobre experiencias similares, casos exitosos de implementación de alternativas de trabajo académico tendientes a mejorar los indicadores y resultados de las Pruebas de estado SABER PRO en las diferentes Instituciones de Educación Superior (IES) de la ciudad de Barranquilla inicialmente.

Seguidamente, se analizaron algunas de las diferentes estrategias y programas con los cuales las Instituciones de Educación Superior, promueven la adopción de metodologías orientadoras para mejorar el desempeño y valoración de la prueba en sus estudiantes, así como la incorporación de las TIC en diferentes tipos de acciones de formación y acompañamiento académico a fin de promover el cambio de roles, enriquecer los procesos de enseñanza y de aprendizaje y transformar las prácticas de formación profesional tradicionales, orientando siempre los esfuerzos hacia mejorar los indicadores de calidad propuestos por el Ministerio de Educación Nacional (Lagunes-Domínguez et al., 2017; Hernandez et al., 2013; Rodriguez et al., 2014). Se revisaron diversos análisis y estudios asociados con el impacto de los resultados de la prueba en la calidad educativa y en los procesos de formación en las diferentes IES, los cuales han involucrado el uso de herramientas de análisis estadísticos y algoritmos computacionales, estudios en los cuales se han tenido en cuenta variables de tipo socioeconómico, tales como: genero, estado civil, estrato económico, ocupación del padre y de la madre; de igual forma, estudios en los cuales se han tratado temáticas como: la deserción estudiantil, cierre de programas académicos, etc., pero que no han sido del todo suficientes para encontrar posibles argumentos que denoten una relación directa entre la formación académica impartida por la IES a los estudiantes, reflejado en su historial de notas académicas, con el puntaje obtenido en la prueba de estado SABER PRO.

El conjunto de datos del historial académico, junto con los resultados de la prueba de estado SABER PRO, constituyen un gran volumen de información, con una estructura compleja, lo cual dificulta las tareas de análisis y extracción de conocimiento oculto (Jiménez et al., 2013). Debido a esto, la minería de datos surge como una posible respuesta, a la necesidad de análisis, manipulación y extracción de conocimiento de los datos, toda vez que ésta hace uso de sofisticados algoritmos que permiten encontrar patrones ocultos en un conjunto de datos y predecir comportamientos. Así las cosas, no es novedoso el uso de la minería de datos para el análisis de información de índole académica, sin embargo, estudios como (Ruby y David, 2014; Nghe et al., 2007; Miguéis et al., 2018; Asif et al., 2017; Shahiri et al., 2015; Goga et al., 2015; Costa et al., 2017), entre otros, denotan la pertinencia del uso de dichas técnicas en el ámbito académico. No obstante, las investigaciones analizadas, no evidencian un análisis en profundidad sobre la posible correlación que puedan existir entre los datos arrojados por los resultados de la prueba de estado SABER PRO y el desempeño académico de un estudiante, ni tampoco una investigación cuyo norte sea la predicción, fruto del análisis correlacional entre los dos mencionados factores.

Finalmente, la problemática expuesta orienta al desarrollo de una herramienta para el análisis de los resultados de las Pruebas de Estado de la Educación Superior en Colombia (SABER PRO) con respecto al promedio académico por áreas genéricas mediante el uso de minería de datos.

METODOLOGÍA

Se describen las diferentes fases del proyecto "P*roceso de extracción de conocimiento KDD para la predicción y análisis de resultados del examen Saber- Pro",* cuyo objetivo contempla el desarrollo de una herramienta resultado último y aplicado de una serie de procesos, enmarcados en la metodología para el descubrimiento de conocimiento en bases de datos conocida como KDD, metodología que hace uso de minería de datos, como principal proceso para descubrir conocimiento oculto, resultado de correlacionar la data correspondiente a la información del historial académico con la data correspondiente a la información de los resultados de la prueba de estado SABER PRO, para un grupo de estudiantes.

Fases del proceso

Selección de fuentes de datos: Consistió en buscar en los datos atributos apropiados de entrada. Esto quiere decir, saber qué es lo que se quiere obtener y cuáles son los datos que facilitaran este proceso para lograr los resultados.

Pre-procesamiento de datos: En este paso se depuraron los datos, incluyendo los datos incompletos (donde hay atributos o valores de atributos perdidos), el ruido (valores incorrectos o inesperados) y datos inconsistentes (conteniendo valores y atributos con nombres diferentes). Los datos sin coherencia en algunos casos debieron ser eliminados, debido a que pudieren permitir un análisis inadecuado y por ende resultados incorrectos. En resumen, este proceso consistió de tres (3) fases: definir y determinar los tipos de errores, buscar e identificar las instancias que contienen errores y corregir los errores descubiertos.

Normalización de datos: En esta etapa se contempló el uso de los siguientes sub procesos: (i) Integración de Datos: Se combinaron datos de múltiples procedencias incluyendo múltiples bases de datos, que podrían tener diferentes contenidos y formatos. La inconsistencia en el formato puede llevar una redundancia e inconsistencia en los atributos y valores de los datos. Normalmente cuando se trabaja en un problema de proceso de descubrimiento es necesario primero formar un único conjunto con todos los datos que provienen de distintas fuentes; (ii) Transformación de Datos: Las transformaciones consisten principalmente en modificaciones sintácticas llevadas a cabo sobre datos sin que supongan un cambio para la técnica de minería aplicada; y (iii) Reducción de Datos: Se redujo el tamaño de los datos, encontrando las características más significativas para representar los datos dependiendo del objetivo del proceso. Se pueden utilizar métodos de transformación para reducir el número efectivo de variables a ser consideradas, o para encontrar otras representaciones de los datos.

Minería de datos: Este es un proceso que consiste en la búsqueda de los patrones de interés que pueden expresarse como un modelo o simplemente que expresen dependencia de los datos. El modelo encontrado depende de su tarea (por ej. Clasificación, Predicción, Agrupamiento, etc.) y de su forma de representarlo (Por ejemplo. Arboles de decisiones, reglas, redes neuronales, entre otras). Se tiene que especificar un criterio de preferencia para seleccionar un modelo de un conjunto de posibles modelos (véase para referencia de tareas y modelos (Shu-Hsien et al., 2012; Sumathi y Sivanandam, 2006; Klösgen y Zytkow, 2002; Peña-Ayala, 2014; Fernandes et al., 2018; Wang y Liao, 2011). En éste proceso el modelo se desarrolla para la predicción y la técnica de extracción de conocimiento usada son las redes neuronales artificiales, debido a la robustez que posee la técnica para el manejo de datos, adaptabilidad y su reconocida capacidad de generalización (Sánchez y García, 2017).

Evaluación de patrones: Se identificaron los patrones interesantes que representan conocimiento usando diferentes técnicas incluyendo análisis estadísticos y lenguajes de consultas.

Interpretación de resultados: Se asumió como el proceso de entender los resultados del análisis y sus implicaciones y puede llevar a regresar a algunos de los pasos anteriores. Hay técnicas que pueden ser útiles en este paso para facilitar el entendimiento de los patrones descubiertos.

Luego de conocer cada una las etapas que conforman la metodología KDD, se hace énfasis en el proceso de extracción de conocimiento el cual está enmarcado en la fase minería de datos.

Aplicación de la metodología KDD

Se muestra en forma detallada las tareas que fueron realizadas en cada etapa, como módulo de prueba, se llevó a cabo el análisis correlacional entre las notas obtenidas por los estudiantes de la Facultad de Ingenierías

de la Universidad Simón Bolívar, tanto para el módulo de razonamiento cuantitativo, como para las notas obtenidas en cada una de las asignaturas de la malla académica, asociadas al mencionado módulo. En este proceso, se incorporaron los datos recopilados de las distintas fuentes de información en una base de datos creada en el motor Microsoft SQL Server 2012®, mientras que en las subsiguientes etapas del proceso, se empleó la herramienta de administración Microsoft SQL Server Management Studio 2012®.

La base de datos "dataset" de notas tras el preprocesado y la normalización contiene 2040 registros, los cuales corresponden a información de 12 asignaturas de 170 estudiantes. Así mismo, la respectiva base de datos de registros de los resultados de la prueba de estado Saber Pro contiene 170 registros correspondientes a los mismos estudiantes identificados. Seguidamente, con los datasets del sistema de notas de la Universidad Simón Bolívar y prueba SABER PRO, se procedió a entrenar una red neuronal, Implementada en MatLab®, y posteriormente enlazada a una aplicación desarrollada en visual Basic.NET®, la cual proporciona una forma visual para el resultado del análisis llevado a cabo. Con base en lo anterior, a continuación se resumen de manera gráfica las tareas que se realizaron en cada fase de este trabajo teniendo en cuenta el marco establecido por la metodología KDD, tal como se muestra en la Fig. 1.

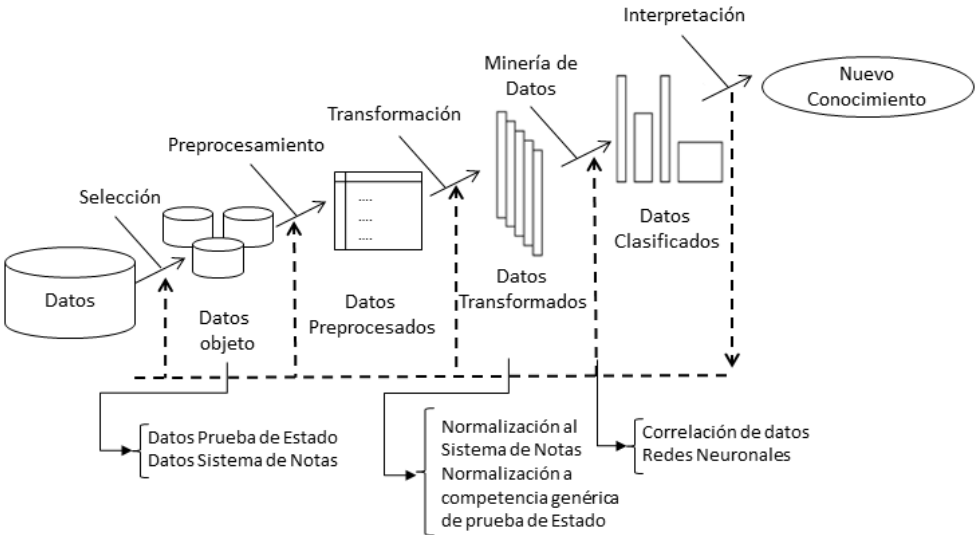

Fig. 1: Diagrama de Flujo del Proceso KDD aplicado a la correlación y predicción de datos, prueba de estado vs datos académicos.

RESULTADOS

El proceso de extracción y selección de los datos correlacionados se obtuvo a partir de la utilización de un algoritmo descriptivo correlacional. Para lo anterior se utilizó una herramienta integrada al gestor de bases de datos Microsoft SQL Server Management Studio, llamada Analysis Services, la cual es capaz por medio del uso de algoritmos de correlación, encontrar las relaciones existentes entre las diferentes fuentes de datos ya procesadas.

El proceso para llevar a cabo la correlación, es el resultado del cruce de los dos conjuntos de datos, producto de los procesos anteriormente descritos, para luego extraer las notas definitivas, llevarlas a un vector el cual permite calcular el promedio ponderado de dichas asignaturas y posteriormente compararlas con el puntaje del módulo de razonamiento cuantitativo. Para reducir el margen de error, se convierten las notas de una escala cuantitativa a una escala cualitativa, determinando 4 rangos continuos en la escala de notas académicas mostradas en la Tabla 1.

Tabla 1: Conversión notas académicas a notas de la prueba de estado SABER PRO.

Promedio Ponderado	
Valor	*Escala Cualitativa*
3,00 - 3,49	Regular
3,50 - 3,99	Bueno
4,00 - 4,59	Muy Bueno
4,60 - 5,00	Excelente

Además de utilizar 6 rangos continuos en la escala de resultados de la prueba de estado SABER PRO. Para poder hacer esta comparación es necesario llevar la nota del conjunto de datos académico a la misma escala del puntaje de la prueba SABER PRO, para este fin se diseñó una escala de conversión, la cual se muestra en la Tabla 2.

Tabla 2: Escala de conversión prueba de estado SABER PRO.

Saber Pro	
Valor	*Escala Cualitativa*
> 9,0	Muy Bajo
9,0 - 9,5	Bajo
9,6 - 10,2	Regular
10,3 - 10,6	Bueno
10,7 - 11	Muy Bueno
< 11,0	Excelente

Correlación de datos

Para la elaboración de este artículo, se parte del supuesto de que existe una correlación positiva entre el promedio ponderado de las notas académicas del estudiante, y el puntaje del módulo de razonamiento cuantitativo de la prueba de estado SABER PRO. Por lo tanto, realizaremos la validación de dicha relación mediante el cálculo del coeficiente de correlación. Un coeficiente de correlación, mide el grado de relación o asociación existente generalmente entre dos variables aleatorias. En la literatura existen varias pruebas orientadas al cálculo del coeficiente de correlación, dos ampliamente difundidas son: el coeficiente de *Spearman* y el de *Pearson*.

El coeficiente de correlación de *Spearman*, ρ (rho) es una medida de la correlación (la asociación o interdependencia) entre dos variables aleatorias continuas. En otras palabras, permite identificar si dos variables se relacionan en una función monótona (es decir, cuando un número aumenta, el otro también o viceversa). Para calcular ρ, los datos son ordenados y reemplazados por su respectivo orden.
El estadístico ρ viene dado por la ecuación (1).

$$\rho = 1 - \frac{6 \sum D^2}{N(N^2-1)} \quad (1)$$

Donde D es la diferencia entre los correspondientes estadísticos de orden de x – y. N es el número de parejas. El coeficiente de correlación de *Pearson* es una medida de la relación lineal entre dos variables aleatorias cuantitativas. A diferencia de la covarianza, la correlación de *Pearson* es independiente de la escala de medida de las variables.

La interpretación del coeficiente rho de *Spearman* es igual que la del coeficiente de correlación de *Pearson*, concuerda en valores próximos a 1; indican una correlación fuerte y positiva. Valores próximos a –1 indican una correlación fuerte y negativa. Valores próximos a cero indican que no hay correlación lineal. Puede que exista otro tipo de correlación, pero no lineal. Los signos positivos o negativos solo indican la dirección de la relación; un signo negativo indica que una variable aumenta a medida que la otra disminuye o viceversa, y uno positivo que una variable aumenta conforme la otra también lo haga disminuye, si la otra también lo hace. En la experimentación se calcularon los coeficientes de correlación de *Spearman* y *Pearson* para los datos del promedio ponderado de las notas relacionadas con el modulo y el puntaje del módulo genérico de razonamiento cuantitativo de 170 estudiantes, obteniéndose como resultados 0.8439 para el coeficiente de *Spearman* (rho) y 0.9396 para el coeficiente de *Pearson* (r). La r de *Pearson* y la rho de *Spearman*, reflejan una asociación positiva entre el promedio ponderado y la prueba de estado SABER PRO: los promedios ponderados más altos tienden a tener mejor desempeño en la prueba SABER PRO, esto es, valores altos.

Predicción con redes neuronales

Para llevar a cabo la predicción se implementó una red neuronal Perceptron Multicapa con 12 entradas correspondientes a cada una de las asignaturas asociadas al módulo de razonamiento cuantitativo, 1 capa oculta de 7 neuronas, y una salida, correspondiente al resultado predicho para el mencionado modulo (Fig. 2). Posteriormente se integró la red neuronal desarrollada a la herramienta construida en el lenguaje visual basic.NET. Esto se logró gracias al consumo de las librerías de MATLAB desde Basic. La red neuronal fue desarrollada en la toolbox de redes neuronales de MATLAB® versión R2014a.

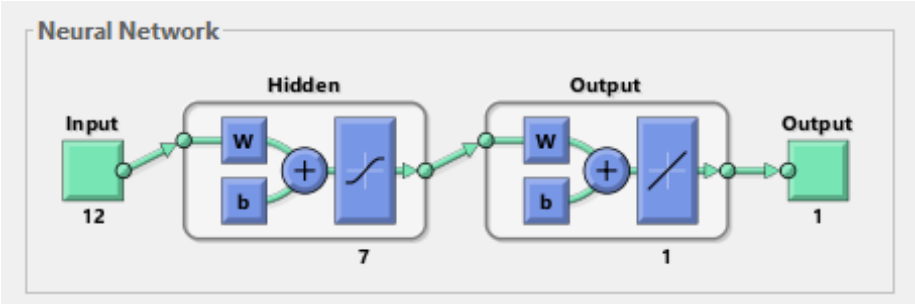

Fig. 2: Arquitectura de la red neuronal implementada

El proceso de entrenamiento y las características de la red neuronal utilizada para llevar a cabo la predicción, además de la integración con la herramienta desarrollada se describen brevemente a continuación, teniendo en cuenta que la data destinada para el entrenamiento de la red neuronal fueron en total 2040 registros, provenientes del conjunto de datos de notas académicas. Así mismo, para llevar a cabo esta tarea se utilizó un algoritmo de división de datos aleatorio, el cual divide en tres (3) partes los datos, asignando de manera aleatoria los datos a alguno de los 3 siguientes rangos: Rango para entrenamiento: 70% = 1.428 registros (119 estudiantes); Rango para validación: 20% = 408 registros (34 estudiantes); Rango para pruebas: 10% = 204 registros (17 estudiantes).

La función de entrenamiento que se utilizó fue la Levenberg-Marquardt backpropagation, el cual es un algoritmo de aprendizaje supervisado que actualiza los pesos de las entradas de acuerdo con la optimización de Levenberg-Marquardt. Para el cálculo de errores se utilizó la función que implementa el algoritmo de Error cuadrático medio (MSE por sus siglas en ingles). El cual es una función de riesgo, que a partir de un estimador y un estimado, mide el promedio de los errores al cuadrado, es decir, la diferencia entre el estimador y lo que se estima. Esta función permitió minimizar la diferencia que existe entre los dos factores que procesa, realizándolo en cada iteración del proceso de aprendizaje de la red neuronal.

La Tabla 3 presenta los resultados de 17 estudiantes del rango de pruebas donde se contrastan el promedio ponderado – PP, la escala cualitativa de la calificación promedio calculada según la Tabla 1, el valor real obtenido en la prueba de estado, el rango en escala cualitativa del valor de la prueba de estado, el valor predicho, el rango cualitativo del valor predicho, la diferencia entre el valor real y el predicho, y el error cuadrático medio – MSE obtenido con la muestra del rango de pruebas.

Tabla 3: Valores reales y predichos de la prueba de estado de Colombia

Promedio Ponderado	*Escala Cualitativa*	*Valor Real obtenido*	*Rango*	*Valor Predicho*	*Rango Predicho*	*Diferencia*
3,11	Regular	9,8	Regular	9,6	Regular	-0,2
4,34	Muy bueno	10,5	Bueno	10,8	Muy bueno	0,4
4,29	Muy bueno	11,4	Excelente	11,7	Excelente	0,3
3,45	Regular	10,0	Regular	10,1	Regular	0,1
3,22	Regular	9,4	Bajo	9,6	Regular	0,2
3,56	Bueno	9,9	Regular	9,6	Regular	-0,3
4,06	Muy bueno	11	Excelente	11,8	Excelente	0,8
4,31	Muy bueno	11,2	Excelente	11,8	Excelente	0,6
3,78	Bueno	10	Regular	9,9	Regular	-0,1
3,54	Bueno	9,8	Regular	10,2	Bueno	0,4
3,87	Bueno	10,4	Bueno	10,5	Bueno	0,1
4,01	Muy bueno	10,6	Muy bueno	10,4	Bueno	-0,2
3,06	Regular	9,2	Bajo	9,1	Bajo	-0,1
3,37	Regular	9,5	Regular	10,1	Regular	0,6
3,93	Bueno	10,4	Bueno	10,3	Bueno	-0,1
4,19	Muy bueno	11,1	Excelente	11,3	Excelente	0,2
3,15	Regular	9,1	Bajo	9,2	Bajo	0,1
					MSE	0,116

El MSE obtenido por el rango de pruebas es de 0.116, lo cual indica que el rango de variación en la respuesta (error), es menor a 0.2, dando lugar a valores muy acertados. Lo anterior se acentúa al observar los rangos cualitativos predichos que no coinciden con el rango real obtenido en la prueba, donde puede observarse que 3 de los 17 valores no corresponden (18%). Ahora bien, si comparamos igualmente con la escala cualitativa obtenido según el promedio ponderado de las notas los rangos marcados corresponden en todos los casos, lo cual es razonable, toda vez, que se comprobó una correlación entre el promedio ponderado y los resultados de la prueba de estado.

DISCUSIÓN DE RESULTADOS

Para el desarrollo de esta herramienta, se utilizaron técnicas de minería de datos de las dos categorías (técnicas supervisadas y no supervisadas). Debido a que la primera mencionada nos ayuda para predicción de los resultados y la segunda para el análisis de la prueba. Para analizar los resultados de la prueba Saber Pro con respecto al promedio académico de estudiantes, se utilizó la técnica de minería llamada correlación de datos, esta pertenece a la categoría de aprendizaje no supervisado y para la predicción se usó una red neuronal tipo perceptrón multicapa, con una función de entrenamiento Levenberg-Marquardt backpropagation está técnica pertenece a la categoría de aprendizaje supervisado. Finalmente, la selección de las técnicas adecuadas fue un tanto difícil, debido a que la minería de datos ofrece un abanico de técnicas útiles para varios objetivos, por lo tanto se valoraron las ventajas de las diferentes técnicas dando mejor valoración en nuestro estudio a la correlación y las redes neuronales, como ya se mencionó. El algoritmo permitió: El análisis estadístico y correlacional de la prueba y la predicción de resultados. El lenguaje de programación utilizado para la codificación del algoritmo diseñado fue Visual Basic .NET ®. La programación de la herramienta se realizó por componentes lo que facilito la complejidad de este. Además para el almacenamiento de los datos ultimamos el motor Microsoft SQL Server 2012®. Para llevar a cabo las gestión de estos, se empleó la herramienta de administración Microsoft SQL Server Management Studio 2012®.

Seguidamente, para poder llevar a cabo el análisis, se utilizó una herramienta integrada a al gestor de bases de datos Microsoft SQL Server Management Studio, llamada Analysis Services, la que permitió por medio del uso de algoritmos de correlación, encontrar las relaciones existentes entre las diferentes fuentes de datos ya procesadas y para poder predecir los resultados de la prueba utilizamos una herramienta de software matemático que nos ofrece un entorno de desarrollo integrado llamado MATLAB ®. Lo anterior se realizó soportados en la metodología de extracción de conocimiento KDD, pasando por cada una de sus fases, de las cuales, el resultado fue el análisis y a través del análisis logramos la predicción. La validación del algoritmo permitió determinar la efectividad del análisis y predicción de resultados de la prueba de estado en la educación superior, para esto se contrastaron los resultados obtenidos con datos reales y simulados los cuales fueron presentados con un análisis de las diferentes pruebas realizadas y una tabla de resultados globales, Mostrando que el algoritmo cumplió su objetivo sin fallas consiguiendo el análisis y la predicción de los resultados de la prueba Saber Pro. Las características esenciales que definen la correcta funcionalidad del algoritmo desarrollado en el objetivo anterior son las siguientes: el correcto análisis de las notas de los estudiantes con respecto a los puntajes de la prueba saber Pro y la predicción de los resultados de la prueba; por lo tanto, la validación de la herramienta se orienta a su cumplimiento.

Se requirió el desarrollo, implementación y validación de un algoritmo que fue construido basado en el proceso KDD. Dicho proceso permitió hacer el descubrimiento de conocimiento en las diferentes fuentes de datos tratadas, aplicando paso a paso cada etapa. Las cuales permitieron la selección de los datos, en esta etapa se determinó las fuentes de datos y el tipo de información a utilizar. Luego pasamos al pre procesamiento en esta etapa consistió en la preparación y limpieza de los datos extraídos desde las distintas fuentes de datos en una forma manejable, necesaria para las fases posteriores. Posterior a esto se realizó la etapa de normalización que consistió en el tratamiento preliminar de los datos, transformación y generación de nuevas variables a partir de las ya existentes con una estructura de datos apropiada para luego aplicar la minería de datos el cual fue la fase de modelado propiamente, en donde métodos inteligentes fueron aplicados para la extracción de conocimiento, como fueron los algoritmos de correlación y predicción con el objetivo de extraer patrones previamente desconocidos y por último la fase de Interpretación que nos permitió mostrar el conocimiento extraídos de los datos.

CONCLUSIONES

Tras la construcción de la herramienta y la realización de las pruebas en diferentes escenarios (reales, artificiales y uso de datos masivo) se resaltan los siguientes hechos: (1) En todos los casos analizados la herramienta pudo relacionar las asignaturas y pudo calcular el promedio ponderado de cada estudiante; (2) En todos los casos analizados el algoritmo aplico la correlación y pudo predecir de manera satisfactoria los puntajes de la prueba de razonamiento cuantitativo, comparándolo con el resultado real del estudiante; (3) En algunos casos la predicción no fue exacta, pero los resultados son acordes con el desempeño del estudiante.

Para poder realizar estos análisis, la correlación de los datos tuvo un papel importante debido a que permitió crear relación entre dos fuentes de datos de estructura diferente, cabe resaltar que este proceso fue muy complejo, debido al gran volumen de los datos y a su estructura. Si bien, se realizaron aportes metodológicos y prácticos entorno a la aplicación de la metodología KDD para el análisis y predicción de resultados de la prueba de estado Saber Pro, existen aún, espacios de investigación inexplorados que pueden dar lugar a trabajo posterior.

REFERENCIAS

Asif, R., A. Merceron, S. Abbas y N. Ghani. Analyzing Undergraduate Students' Performance Using Educational Data Mining, doi: 10.1016/j.compedu.2017.05.007, Computers & Education, 113, 177-194 (2017)

Costa, E., B. Fonseca, M. Almeida, F. Ferreira y J. Rego, Evaluating the Effectiveness of Educational Data Mining Techniques for Early Prediction of Students' Academic Failure in Introductory Programming Courses, doi: 10.1016/j.chb.2017.01.047, Computers in Human Behavior, 73, 247-256 (2017)

Fernandes, E., M. Holanda y otros cuatro autores, Educational Data Mining: Predictive Analysis of Academic Performance of Public School Students in the Capital of Brazil, doi: 10.1016/j.jbusres.2018.02.012, Journal of Business Research (2018)

Garcia, J., La Aplicabilidad del Enfoque de Sistemas como Método para la Transposición Didáctica de Situaciones Profesionales Enmarcadas en los Procesos Administrativos, Dimensión Empresarial, ISSN: 2322-956X, 6(1), 52-57 (2008)

Goga, M., S. Kuyoro y N. Goga, A Recommender for Improving the Student Academic Performance, doi: 10.1016/j.sbspro.2015.02.296, Procedia - Social and Behavioral Sciences, 180, 1481-1488 (2015)

Hernández, U., J. Moreno y otros cuatro autores, Evaluación y Aprendizajes de una Experiencia Colombiana de Formación Docente en TIC, Revista Virtual Universidad Católica del Norte, 40, 31-52 (2013)

Jiménez, S., L. Reyes y M. Cañón, Enfrentando Resultados Programa de Ingeniería de Sistemas de la Universidad Simón Bolívar con las Pruebas Saber Pro, doi: 10.17081/invinno.1.1.2072, Investigación e Innovación en Ingenierías, 1(1), 50-56 (2013)

Klösgen, W. y J. Zytkow, Handbook of Data Mining and Knowledge Discovery, Oxford University Press, Inc., New York, USA (2012)

Lagunes-Domínguez, A., C.A. Torres-Gastelú, J. Angulo-Armenta y M.Á. Martínez-Olea, Prospectiva hacia el Aprendizaje Móvil en Estudiantes Universitarios, doi: 10.4067/S0718-50062017000100011, Formación Universitaria, 10 (1), 101-108 (2017)

Miguéis, V.L., A. Freitas, P. Garcia y A. Silva, Early Segmentation of Students According to their Academic Performance: A Predictive Modelling Approach, doi: 10.1016/j.dss.2018.09.001, Decision Support Systems, 115, 36-51 (2018)

Ministerio de Educación Nacional - MEN. Ley 1324, Decreto Numero 3963: Examen de Estado de Calidad de la Educación Superior, República de Colombia (2009)

Nghe N., P. Janecek y P. Haddawy, A Comparative Analysis of Techniques for Predicting Academic Performance, doi: 10.1109/FIE.2007.4417993, 37 ASEE/IEEE Conference Frontiers in Education Conference, T2G7-T2G11, Milwaukee - USA 10-13 de Octubre (2007)

Peña-Ayala, A., Educational Data Mining: A Survey and a Data Mining-Based Analysis of Recent Works, doi: 10.1016/j.eswa.2013.08.042, Expert Systems with Applications, 41(4-1), 1432-1462 (2014)

Rodríguez, G., M. Ariza y J.L. Ramos, Calidad Institucional y Rendimiento Académico: El Caso de las Universidades del Caribe Colombiano, doi: 10.1016/S0185-2698(14)70607-5, Perfiles Educativos, 36(143), 10-29 (2014)

Ruby, J. y K. David, Predicting the Performance of Students in Higher Education Using Data Mining Classification Algorithms - A Case Study, International Journal for Research in Applied Science & Engineering Technology, 2(9), 173-180 (2014)

Sánchez-Sánchez, P. y J.R. García-González, A New Methodology for Neural Network Training Ensures Error Reduction in Time Series Forecasting, doi: 10.3844/jcssp.2017.211.217, Journal of Computer Science, 13(7), 211-217 (2017)

Shahiri, A., W. Husain y N. Rashid, A Review on Predicting Student's Performance Using Data Mining Techniques, doi: 10.1016/j.procs.2015.12.157, Procedia Computer Science, 72, 414-422 (2015)

Shu-Hsien, L., C. Pei-Hui y H. Pei-Yuan, Data Mining Techniques and Applications – A Decade Review from 2000 to 2011, Expert Systems with Applications, 39(12), 11303-11311 (2012)

Sumathi, S. y S. Sivanandam. Introduction to Data Mining and its Applications.Studies in Computational Intelligence, Springer-Verlag, Heidelberg, Alemania (2006)

Wang, Y.H. y H-C. Liao, Data Mining for Adaptive Learning in a TESL-Based E-Learning System, doi: 10.1016/j.eswa.2010.11.098, Expert Systems with Applications, 38(6), 6480-6485 (2011)