Aplicación de las Técnicas Árboles de Decisión y Redes Neuronales para la Generación del Modelo para la Proyección de la Deserción Escolar

Humberto Cerón Otero [1] y Martín Verduzco Rodríguez [2]

Resumen

En el estado de Hidalgo, en México, la educación es una preocupación constante ante la deserción de los alumnos, así como su aprovechamiento académico, y uno de los principales intereses es determinar los múltiples factores que pueden influir en él. En el presente trabajo se hace la evaluación de las técnicas de árboles de decisión y redes neuronales para poder generar el modelo de minería de datos que nos permita identificar patrones de comportamiento, con el fin de predecir la deserción escolar. Las técnicas de minería de datos permiten obtener conocimiento oculto en grandes cantidades de datos con información valiosa que, al explotarse, ofrece ventajas competitivas a las organizaciones e Instituciones. En el caso del Instituto Hidalguense de

Acerca de los autores...

[1]Departamento de Redes y Telecomunicaciones, Instituto Hidalguense de Educación para Adultos.
humber.10@hotmail.com

[2]División de Ingeniería en Sistemas Computacionales, Tecnológico de Estudios Superiores de Ecatepec
martinverduzco@yahoo.com.mx

Educación para Adultos, existen muchos datos respecto a los estudiantes, que no se explotan como debería, ya que en la actualidad no se realiza ninguna estrategia para predecir dicha deserción, información que es útil para tomar decisiones estratégicas en pro de los mismos. Este artículo pretende, con base en los resultados obtenidos a través de las herramientas de minería de datos RapidMiner, generar el modelo para poder realizar la proyección de la deserción escolar, y contribuir a una mejor planeación en el área administrativa, docente y psicopedagógica, para evitar el rezago estudiantil y apoyar en todo momento al alumnado.

Palabras Clave:
Instituciones, Deserción, Minería de Datos, Técnicas, Metodología CRISP_DM, herramientas.

Introducción

El Instituto Hidalguense de Educación para Adultos (IHEA) tiene como objeto prestar los servicios de educación básica, la cual comprende a la educación inicial (alfabetización), intermedia (primaria) y avanzada (secundaria), así como la formación para el trabajo, con los contenidos particulares para atender las necesidades educativas específicas de este sector de la población.

En los últimos años, ha surgido una preocupación ante el problema de la deserción escolar y un creciente interés por determinar los múltiples factores que pueden influir en él. La mayoría de los trabajos que intentan resolver este problema están enfocados a determinar cuáles son los factores que más afectan al rendimiento de los estudiantes (abandono y fracaso) en los diferentes niveles educativos, mediante la utilización de gran cantidad de información que los actuales equipos informáticos permiten almacenar en bases de datos, mismos que constituyen una auténtica mina de oro, por la valiosa información sobre los estudiantes. El problema es identificar y encontrar la información útil que está oculta en esas grandes bases de datos.

Una solución muy prometedora para alcanzar este objetivo, es el uso de técnicas de extracción de conocimiento y/o minería de datos en educación, lo que ha dado lugar a la denominada "minería de datos educativa" (Educational Data Mining, EDM). Esta nueva área de investigación, se ocupa del desarrollo de métodos para explorar los datos que se dan en el ámbito educativo, así como de la utilización de estos métodos para entender mejor a los estudiantes y los contextos donde aprenden.

Para el desarrollo de este artículo, se evaluarán las técnicas de árboles de decisión y redes neuronales con la herramienta RapidMiner. Para generar el modelo de minería de datos, se utilizará la metodología CRISP-DM, que estructura el proceso de minería de datos en seis fases, que son: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación, e implantación; algunas de las cuales son bidireccionales, es decir, que de una fase en concreto se puede volver a la anterior, pero en este caso, nos enfocaremos solo en la fase de modelado, en la parte de evaluación de las técnicas, para ver si nos dan un alto porcentaje de confiabilidad y así poder realizar la proyección de la deserción escolar.

1. DESARROLLO DEL PROYECTO

Aquí entramos a la parte más práctica del proyecto, donde iremos aplicando cada una de las etapas de la fase de modelado de la metodología CRISP-DM (Galán, 2015). Antes de construir el modelo, se necesita generar un mecanismo para probar la calidad y validez de las técnicas de árboles de decisión y redes neuronales. En los modelos con funciones supervisadas, es necesario separar los datos en dos conjuntos: uno para entrenamiento y otro para construir el modelo, ello con el fin de analizar el porcentaje de error, así como el porcentaje de confiabilidad; sin embargo, para el uso de funciones no supervisadas, no es necesario realizar tal entrenamiento, puesto que no hay una clase objetivo a buscar (Cortina, 2015).

1.1 Modelado

En esta fase de la metodología, se evaluará y escogerá la técnica (o técnicas) más apropiadas para el objetivo marcado en la minería de datos. A continuación, y una vez realizado un plan de prueba para los modelos, se procederá a aplicar dichas técnicas sobre los datos para generar el modelo y por último se tendrá que evaluar si dicho modelo ha cumplido con los criterios de éxito o no (Pereyra, 2013).

2. ESCOGER LA TÉCNICA DE MODELADO

Se utilizará el software RapidMiner para realizar las pruebas de las técnicas de Minería de Datos; las técnicas de modelado que se utilizarán, son redes neuronales y árboles de decisión. Entre los modelos que nos ofrece RapidMiner, los que mejor se adaptan a nuestro objetivo, son los modelos de regresión, como son los árboles de decisión y redes neuronales, puesto que los problemas que queremos resolver son de predicción y los campos que se quiere predecir contienen valores continuos (Pereyra, 2013).

2.1 Generar el Plan de Prueba

El procedimiento que se empleará para probar la calidad y validez del modelo, será utilizar las medidas del error cuadrático medio (root mean squared error) y el error cuadrado (squared error) para la técnica de redes neuronales y la exactitud (accuracy), así como la precisión para la técnica de árboles de decisión. Estas medidas de error las calcula automáticamente RapidMiner al ejecutar los modelos de minería de datos (Galán, 2015).

2.2 Construir el Modelo

A continuación, se procederá a ejecutar el modelo elegido sobre los 385 registros de la base de datos muestra. En este apartado, se describirán los ajustes de parámetros en el modelo elegido en la herramienta de minería de datos, así como la salida de dicho modelo y su descripción.

Objetivo. Predecir la deserción escolar de los educandos del IHEA.

En este caso, el campo objetivo, es decir aquel sobre el cual queremos hacer la predicción es “Deserto” y el case id será el “id alumno”. En cuanto a los parámetros empleados para el algoritmo de árboles de decisión y redes neuronales, se utilizan los parámetros que vienen por defecto en RapidMiner, sólo seleccionaremos el atributo que queremos predecir y el case id (Fernanda, 2013), como se puede ver en la Figura 1 y 2.

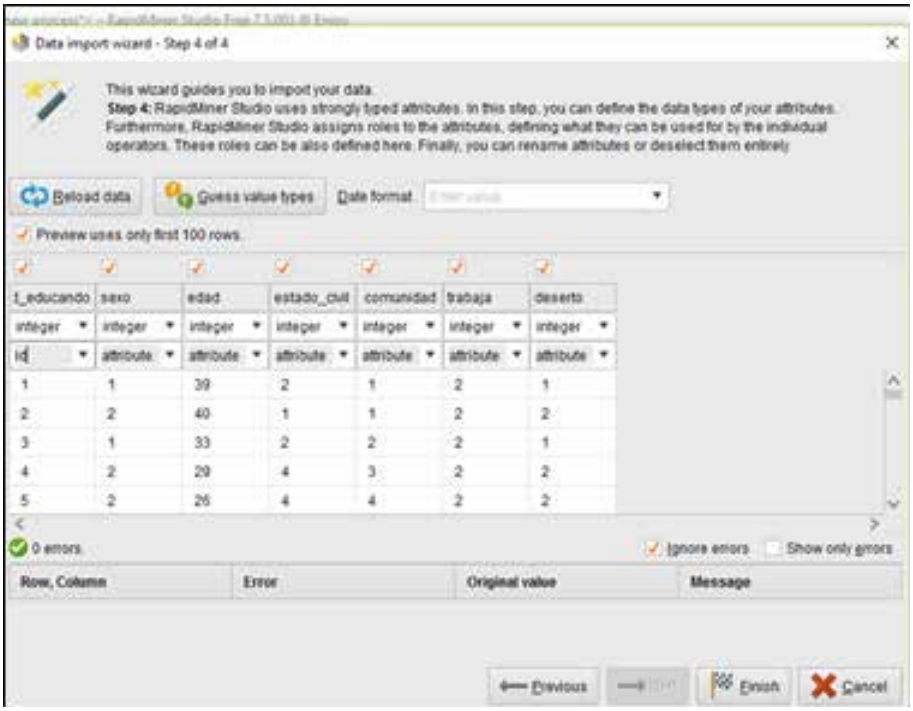

Figura 1

Parámetros case id de los algoritmos árboles de decisión y redes neuronales.

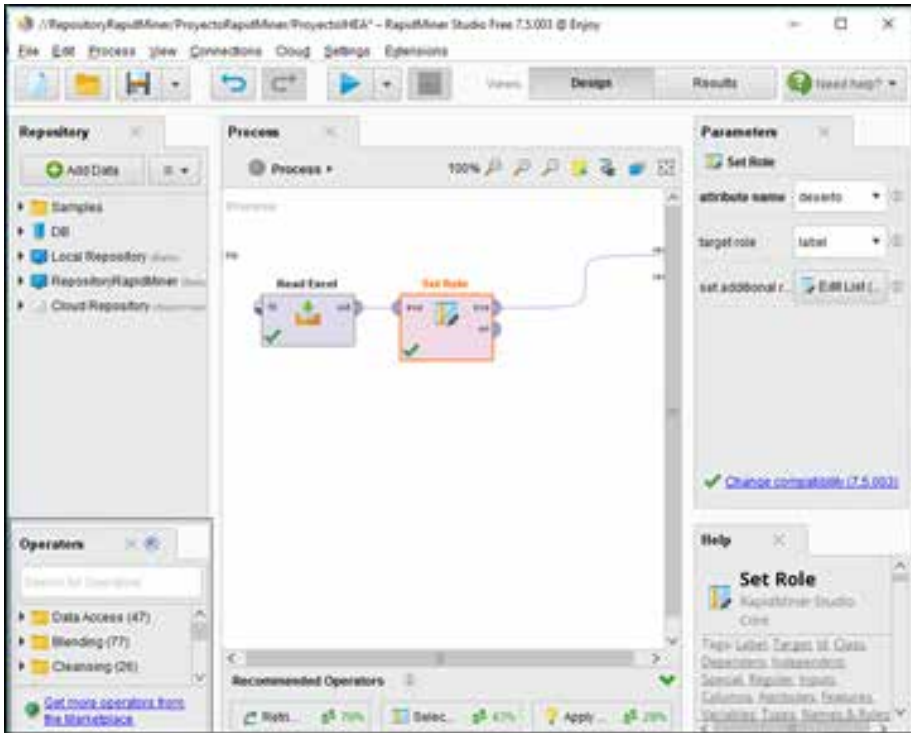

Figura 2

Parámetros deserto de los algoritmos árboles de decisión y redes neuronales.

MODELOS

Se ejecutan los dos modelos, uno por cada técnica de la minería de datos, sobre un conjunto de datos de entrenamiento del 60%, con lo cual se deja el 40% de datos para el conjunto de prueba. Los detalles de la ejecución de cada modelo se pueden ver a continuación en las siguientes figuras (rapidminer, s.f.).

En la Figura 3 se muestra el error cuadrático medio de la ejecución del algoritmo redes neuronales.

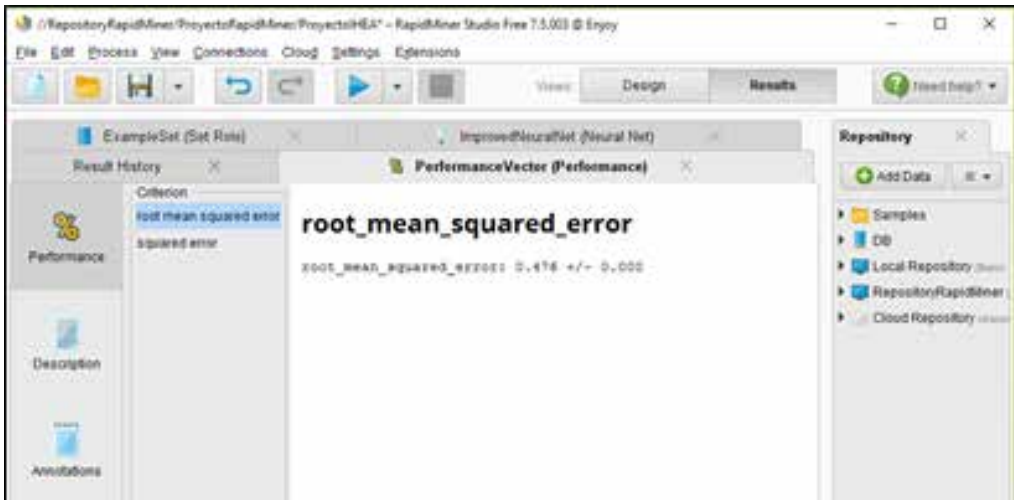

Figura 3

Resultados del algoritmo redes neuronales para el error cuadrático medio.

En la Figura 4 se muestra el error cuadrado de la ejecución del algoritmo redes neuronales.

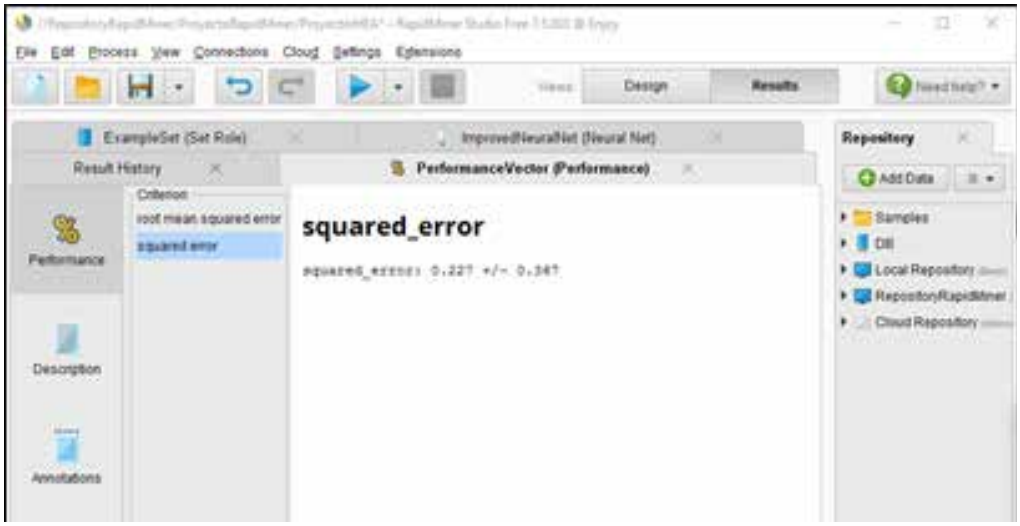

Figura 4

Resultados del algoritmo redes neuronales para el error cuadrado.

En la Figura 5 se muestra la exactitud de la ejecución del algoritmo árboles de decisión.

Figura 5

Resultados del algoritmo árboles de decisión para la exactitud.

Figura 6

Resultados del algoritmo árboles de decisión para la precisión.

En la Figura 6 se muestra la precisión de la ejecución del algoritmo árboles de decisión.

2.3 Descripción de la evaluación de los modelos

A continuación, se describe el resultado de la ejecución de cada uno de los modelos para cada técnica, los cuales muestran si el modelo es factible para usar.

En la siguiente tabla se pueden observar los valores para las técnicas de redes neuronales y árboles de decisión.

Tabla 1.
Resultados de la Evaluación de las Técnicas.

	Error Cuadrático Medio	Error Cuadrado	Exactitud	Precisión
Modelo 1 (Redes neuronales)	0.47%	0.22%		
Modelo 2 (Árboles de Decisión)			76.52%	75.00%

El primer modelo tiene un valor de 0.476 de error cuadrático medio y un valor de 0.227 de error cuadrado para la técnica redes neuronales, por lo que sí es factible emplear este modelo para resolver el objetivo, ya que tiene un bajo porcentaje de error, el cual disminuiría si aumentamos los registros.

El segundo modelo tiene un valor de 76.52% de exactitud y 75.00% de precisión para la técnica árboles de decisión, por lo que igualmente es factible emplear este modelo para resolver el objetivo, ya que tiene un alto porcentaje de exactitud y precisión, tomando en cuenta que es una muestra de 385 registros, el cual puede mejorar aumentando datos.

Conclusiones y trabajos futuros

La aplicación de una adecuada técnica de minería de datos es de gran importancia, ya que ésta permitirá predecir de manera adecuada la deserción de los educandos. La evaluación de estas dos técnicas para su correcto funcionamiento requiere de las herramientas más adecuadas, que con base en sus aplicaciones permita el desarrollo correcto de la técnica.

Después de una buena selección de herramientas, métodos y técnicas de minería de datos para predecir la deserción escolar, se llevó a cabo la evaluación de los modelos, obteniendo buenos resultados, tomando en cuenta que se utilizó una muestra de toda la base de datos, obtenida con el teorema del límite central, que calcula el valor más probable para nuestro universo, y ya finalizando este modelo, podemos utilizarlo con nuestros datos para poder predecir la variable "Deserto"; así también fue posible saber cómo evaluar un modelo en RapidMiner, desde la etapa de configuración de datos hasta la de evaluación del modelo.

En la siguiente fase del desarrollo del proyecto, se llevarán a cabo las siguientes etapas de la metodología CRISP-DM, aplicando los modelos y usando la misma base de datos, pero con nuevos registros.

Referencias

Cortina, V. G. (2015). *Aplicación de la metodología crisp-dm a un proyecto de minería de datos en el entorno universitario*. En V. G. Cortina, Madrid, (págs. 9-11).

Fernanda, O. B. (2013). *Aplicación de técnicas de Minería de datos para predecir la deserción de los estudiantes de primer ciclo de la UTPL*. Loja, Ecuador.

Galán, V. C. (2015). *Aplicación de la metodología crisp-dm a un proyecto de minería de datos*. Madrid, (págs. 9-11).

Pereyra, R. T. (2013). *Aplicación de la minería de datos en la extracción de perfiles de deserción estudiantil.* Colombia.

Rapidminer. (s.f.). Obtenido de RapidMiner: https://rapidminer.com/

Fernanda, O. B. (2013). A*plicación de técnicas de Minería de Datos para predecir la deserción de los estudiantes de primer ciclo de la UTPL*. Loja, Ecuador.