Análisis inteligente de datos aplicado al proceso de nivelación en la Universidad Técnica Estatal de Quevedo

H. Escobar. Universidad Técnica Estatal de Quevedo
W. Burbano. Universidad Técnica Estatal de Quevedo
A. Puris. Universidad Técnica Estatal de Quevedo

Resumen: La presente investigación es un primer acercamiento para estudiar los factores que influyen en el alto índice de deserción estudiantil que ocurre en el proceso de nivelación implementado por el Sistema Nacional de Nivelación Estudiantil que lleva a cabo el gobierno de Ecuador. En este escenario, se toma como caso de estudio la Universidad Técnica Estatal de Quevedo, la cual consta con un registro de 5 periodos académicos que sustentan la base para un análisis inteligente de datos. Para este análisis, se emplean algunos algoritmos basados en arboles de decisión para encontrar un modelo matemático que sea capaz de obtener un alto grado de precisión en relación con la problemática. En este proceso se realiza un preprocesamiento de la información para obtener los mejores ajustes del modelo y se concluye que el algoritmo LMT es le mejor representación de problema obtiene, con una exactitud del 83% y una cantidad razonable de árboles y hojas.

Palabras claves: Proceso de nivelación, árboles de decisión, minería de datos

Intelligent data analysis applied to the grading process at the State Technical University of Quevedo

Abstract: This research is a first approach to study factors influencing the high rate of dropout occurs in the leveling process implemented by the National System of Equalization Student holding the government of Ecuador. In this scenario, is taken as a case study State Technical University of Quevedo, which has a record of 5 academic periods that sustain the foundation for intelligent data analysis. For this analysis, some algorithms based on decision trees to find a mathematical model capable of obtaining a high degree of accuracy in relation to the problem used. In this process the information preprocessing is performed to obtain the best fit model and concludes that the LMT algorithm is better representation of problem you get with an accuracy of 83% and a reasonable amount of trees and leaves.

Keywords: Leveling process, decision trees, data mining

1. INTRODUCCIÓN

Las Universidades del Ecuador ambicionan proporcionar una educación de calidad, con la finalidad de que los estudiantes de dichas instituciones sean más competitivos en el ámbito profesional y puedan aportar con mayor eficiencia al desarrollo de nuestro País.

Uno de los grandes desafíos a los que se enfrenta la educación superior hoy en día, es pronosticar las trayectorias individuales de los estudiantes, determinando la importancia que toma este sector tanto en el ámbito gubernamental como privado. Por tal efecto estas instituciones cuentan con una mayor responsabilidad, como la de proponer una mejora continua en sus prácticas y calidad educacional.

Unas de las iniciativas efectuadas por el gobierno ecuatoriano para mejorar el rendimiento de los estudiantes es el curso de "Nivelación" del Sistema Nacional de Nivelación y Admisión (SNNA) orientada a atender las limitaciones por las que transita hoy en día la enseñanza en el perfil de bachiller. El proceso se lleva a cabo con estudiantes que ya tiene asignada un cupo para ingresar a la universidad y son agrupados por carreras a cursar.

El proceso aunque está bien justificado presenta algunas deficiencias que provoca que un porcentaje elevado de los estudiantes que ingresan no culminen la etapa de preparación y otros mucho no pasan del primer año de su carrera.

Por otro lado la minería de datos ha sido una de las herramientas más utilizadas en los procesos educativos para descubrir y conocer patrones que representen las conductas individuales con gran precisión, ayudando en gran medida a mejorar los procesos pedagógicos existentes (Bae, 2015) (Holte, 1993) (García, 2010).

El presente estudio cuenta con un ingente repositorio de datos acerca de los alumnos que formaron parte del proceso de nivelación entre los años 2012 y 2014 en la Universidad Técnica Estatal de Quevedo (UTEQ). Esta información es utilizada para encontrar un modelo inteligente basado en árboles de decisión que ayude a identificar estudiantes que pueden estar en peligro de una posible deserción y de esta manera tomar decisiones a tiempo.

2. METODOLOGÍA

Para el desarrollo de la presente investigación se utilizó el método analítico el cual se fundamenta a partir del análisis desarrollado entre las diferentes técnicas de minería de datos para seleccionar una de las más utilizadas con descripción detallada de su funcionamiento.

La observación se utilizó para obtener los registros de las variables utilizadas en el proceso de extracción del conocimiento y la técnica cuasi experimental sirvió para realizar las comparaciones pertinentes entre los arboles de decisión que se utilizaron en la investigación.

Minería de datos

La Minería de Datos (Data Mining) por las siglas en inglés Data Mining es el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos (Frank, 2000). Las herramientas de Minería de Datos predicen futuras tendencias y comportamientos, ayudando a la toma de decisiones. El proceso se reduce en 4 etapas fundamentales como se describen a continuación:

- *Determinación de los objetivos***.** Trata de la delimitación de los objetivos que se desean obtener con el análisis de datos.
- *Preprocesamiento de los datos***.** Se refiere a la selección, la limpieza, el enriquecimiento, la reducción y la transformación de las bases de datos. Esta etapa consume generalmente alrededor del setenta por ciento del tiempo total de un proyecto y supone un conocimiento importante del problema.
- *Determinación del modelo***.** Se comienza realizando unos análisis estadísticos de los datos, y después se lleva a cabo una visualización gráfica de los mismos para tener una primera aproximación. Según los objetivos planteados y la tarea que debe llevarse a cabo, pueden utilizarse algoritmos desarrollados en diferentes áreas de la Inteligencia Artificial con diferente naturaleza.
- *Análisis de los resultados*. Verifica si los resultados obtenidos son coherentes y los coteja con los obtenidos por los análisis estadísticos y de visualización gráfica.

El cliente determina si son novedosos y si le aportan un nuevo conocimiento que le permita considerar sus decisiones.

La Minería de Datos constituye una de las etapa del descubrimiento de conocimientos conocido como KDD (Fayyad, 1996) el cual consiste en el uso de algoritmos concretos los cuales generan una enumeración de patrones a partir de los datos anteriormente procesados, apoyándose con algoritmos de Aprendizaje Automático (Frank, 2000).

Para la ejecución del preprocesamiento de datos y la determinación del modelo, existe un grupo de herramientas ya implementadas puestas a disposición de la comunidad científica, para facilitar el análisis de datos, Weka (Waikato Environment for Knowledge Analysis) (Witten & Eibe Frank, 2007) es una de las más utilizadas por su distribución libre y las cantidad de técnicas que tiene incorporada.

Los sistemas de aprendizaje basados en árboles de decisión son quizás el método más fácil de utilizar y de entender. Un árbol de decisión es un conjunto de condiciones organizadas en una estructura jerárquica, de tal manera que la decisión final a tomar se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas.

Una de las grandes ventajas de los árboles de decisión es que, en su forma más general, las opciones posibles a partir de una determinada condición son excluyentes. Esto permite analizar una situación y, siguiendo el árbol de decisión apropiadamente, llegar a una sola acción o decisión a tomar.

Proceso de nivelación

Los estudiantes que concluyen sus estudios de bachillerato se presentan al Examen Nacional de Educación superior (ENES), el mismo que busca igualdad de condiciones de acceso a los estudiantes en igualdad de oportunidades. Con los resultados postulan hasta en cinco carreras en una o varias universidades, en función del puntaje el Sistema Nacional de Nivelación y Admisión les asigna un cupo con el que participan en el curso de nivelación que aborda contenidos de Buen Vivir y Proyecto de Vida, y de conocimientos en función de la carrera elegida, compensando las desigualdades

producidas por la heterogeneidad educativa del bachillerato. El proceso de nivelación tiene una duración de 18 semanas con una carga de 38 horas semanales.

En el curso de Nivelación los docentes inician sus actividades con una prueba diagnóstica que permite conocer las situación del estudiante, en el proceso de evaluación formativa se aplican procesos de retroalimentación y mejora y se busca generar alternativas que propicie que el estudiante alcance los resultados de aprendizaje propuesto, para lo cual se monitorea el trabajo autónomo del mismo. Es destacable el aporte del proceso tutorial, que permite generar estrategias de acompañamiento y seguimiento académico al estudiante.

3. RESULTADOS

Preparación de los datos

La preparación de datos en general tiene como objetivo principal organizar y representar las vistas minables a las que se les pueda aplicar las herramientas concretas de Minería de Datos. Esta organización de los datos debe ir acompañada de una limpieza e integración de los mismos para que estén en condiciones para su análisis.

En el caso específico de este trabajo, la información utilizada proviene de dos años de recopilación de información donde se observaron un conjunto de variables como se describe a continuación, la mayoría de estas se obtuvieron a partir de encuestas.

Preprocesamiento de datos

Debido a la enorme cantidad de datos que no eran relevantes, la presencia de valores faltantes, los datos ruidosos y las inconsistencias que presentaban fue necesario realizar una limpieza de datos.

En primer lugar se realizó un análisis de las 42 variables para poder identificar cuáles eran las más relevantes que permitieran conseguir una confiable fuente de conocimientos, identificamos 13 atributos relevantes los cuales son:

IDENTIFICACIÓN DE VARIABLES	
UBV	Universidad y Buen Vivir equivale al 10% de la nota Final

ICA	Introducción al Conocimiento Analítico equivale al 10% de la nota Final
M1	Corresponde a la Evaluación del Módulo 1 equivale al 20% de la nota Final
M2	Corresponde a la Evaluación del Módulo 2 equivale al 20% de la nota Final
M3	Corresponde a la Evaluación del Módulo 3 equivale al 20% de la nota Final
PIS	Proyecto de Integración de Saberes equivale al 10% de la nota Final
EX_FIN	Corresponde a la nota obtenida en el examen final que equivale al 10% de la nota Final
NOT_FIN	La calificación final obtenida en los diferentes módulos de aprendizaje
ASISTENCIA	Es el porcentaje obtenido en las asistencias totales de los estudiantes
CARRERA	Nombre de la carrera en la que se matriculó en la UTEQ
MODALIDAD	Indica si el estudiante estudia de forma presencial o semipresencial
ESTILO_AP	Es el estilo de aprendizaje que fue aplicado al estudiante en el proceso de nivelación.
APROBACION	Detalla si el estudiante aprobó o no el curso

donde APROBACION representa la variable de decisión y la variable ESTILO_AP puede tener 4 posibles valores *read/write, Kinesthetic, AK, aural, dudoso*, los 4 primeros definen una forma de preparación de los estudiantes y el último cuando el estudiantes no lo tiene bien definido.

Para eliminar los valores perdidos se realizó una imputación de datos utilizando la función ReplaceMissingValues, la cual realiza la imputación de los datos utilizando la media en el caso de los valores continuos y la moda en el caso de los valores nominales.

Seguido eliminamos los outliers y los valores extremos, comenzamos utilizando la función InterquartileRange presente en weka, la cual se encarga de evaluar que los valores se encuentren dentro del rango esperado, si algún valor no cumple con el rango lo detecta

como outlier o como extremo y almacenarlos como otro atributo y la función RemoveWithValues la cual elimina los datos seleccionados, en este caso los outliers

Con el fin de mantener balanceados los datos de cada una de las clases determinadas, se utilizó la función Smote, ya que el caso representativo de la clase aprobado era superior en un 74% de los que no aprueban.

Obtención del modelo

Para el modelado se utilizaron 4 clasificaciones por árboles de decisión, los cuales permitieron obtener resultados diversos, los algoritmos aplicados son: J48, DecisionStump, LMT y REPTree.

El algoritmo J48 de Weka es una implementación del algoritmo C4.5, Se trata de un refinamiento del modelo generado con OneR y supone una mejora moderada en las prestaciones, con su aplicación se obtuvieron los siguientes resultados:

- Precisión = 83.77%
- Error = 16.23%
- N° de hojas = 108
- N° de árboles = 231

El algoritmo DecisionStump es un modelo de aprendizaje que consiste en un árbol de decisión de 1 nivel, se trata de un árbol de decisión con una raíz, que está inmediatamente conectado a los nodos terminales. El algoritmo hace una predicción basada en el valor de una sola entidad de entrada.

- Precisión = 33.58%
- Error = 66.42%

El algoritmo LMT (Logistic Model Tree) es un modelo de clasificación con un algoritmo de entrenamiento supervisado asociado que combina regresión logística y el aprendizaje árbol de decisión, con su aplicación se obtuvieron los siguientes resultados:

- Precisión = 83.02%

- Error = 16.98%
- N° de hojas = 29
- N° de árboles = 43

El algoritmo REPTree es un árbol de aprendizaje de decisión rápida, Construye un árbol de decisión utilizando la información de varianza, sólo ordena los valores para los atributos numéricos de una vez, con su aplicación se obtuvieron los siguientes resultados:

- Precisión = 58.30%
- Error = 41.70%
- Tamaño del Árbol = 41

Como podemos observar, el árbol con mayor precisión es el presente en el algoritmo J48, con una precisión de 83.77%, seguido del algoritmo LMT. Sin embargo se puede observar que la diferencia en precisión es mínima (0.75), pero si tomamos en cuenta el número de hojas y la cantidad de árboles encontrados, podemos determinar que el algoritmo LMT es el que mejor modelo presenta para este problema.

Interpretación

Los resultados demuestran claramente que existe una mayoría de estudiantes en kinesthetic, Dudo, Aural, Read/Write, ARK y Visual en comparación con las metodologías de estudio restantes, por lo cual procederemos a explicar y detallar los resultados

Class Kinesthetic

Class Kinesthetic tiene un total de 141 estudiantes en el repositorio de datos obtenido, de los cuales 85 de ellos han reprobado, lo que equivale al 60.28%, los factores más influyentes se encuentra en la asistencia y las calificaciones obtenidas en los módulos, los cuales no fueron los necesarios para aprobar, por el contrario un total de 56 estudiantes correspondientes al 39.72% si aprobaron la nivelación, esto nos permite destacar que, a pesar de ser la modalidad con mayor cantidad de estudiantes, muestra una mayoría de estudiantes reprobados.

Class Dudo

Class Dudo con un total de 101 estudiantes es la segunda modalidad más influyente, esta destaca sobre la kinesthetic por tener un menor porcentaje de estudiantes reprobados, pues 28 estudiantes han reprobado, es decir, el 27.72% de los estudiantes no lograron aprobar el curso, en su mayoría por factores de promedio de notas en los módulos de estudio, además debemos destacar que su porcentaje de aprobados es 72.28%, es decir un total de 73 estudiantes han aprobado.

Class Aural

Esta modalidad de estudio es un caso especial, pues presenta partes iguales de aprobados y reprobados, con un total de 82 estudiantes, solo 41 han aprobado, demostrando ser estadísticamente más efectiva que kinesthetic pero no en comparación con la modalidad dudo, el análisis mostró que se debe en su mayoría a los promedios obtenidos en los módulos a lo largo del curso, siendo influenciados muy poco por la asistencia.

Class Read/Write

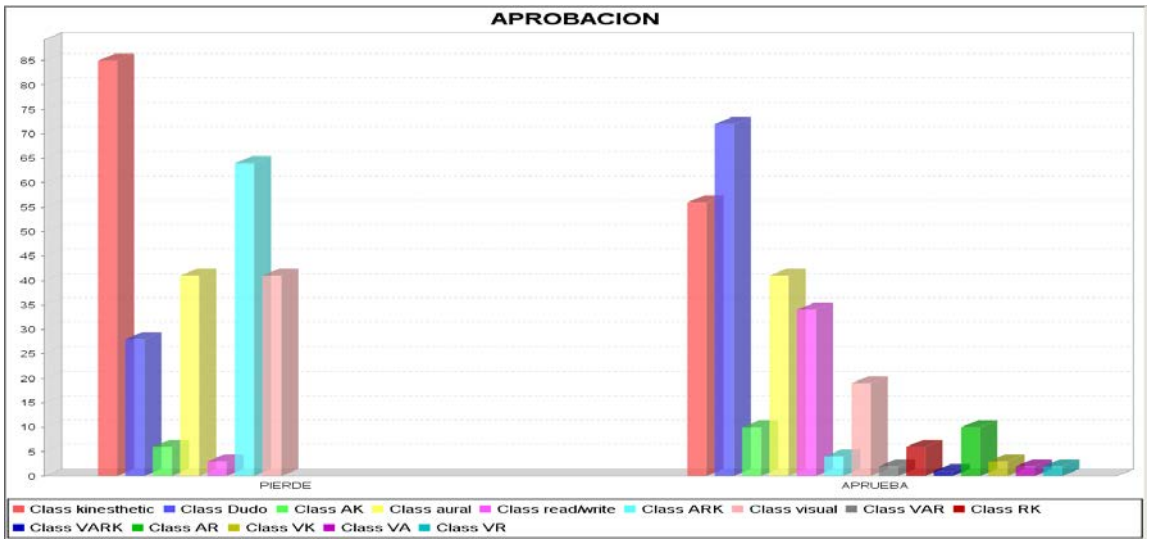
La modalidad Read/Write es aquella que ha mostrado un mayor porcentaje de aprobados de todas las metodologías aplicadas, con un total de 35 estudiantes, solamente el 5.71% han reprobado, es decir solamente 2 estudiantes no han conseguido aprobar, teniendo un total de aprobados de 33 estudiantes, es decir el 94.29% presentando el mayor índice de éxito entre todas las modalidades de estudio aplicadas.

Class ARK

Al contrario de Read/Write que presentaba el mayor porcentaje de aprobados, la modalidad ARK presenta el mayor porcentaje de reprobados, pues de un total de 66 estudiantes, 63 han reprobado, esto corresponde a tan solo el 4.55% de aprobados y el 95.45% de reprobados, lo que indica que la modalidad de ARK es la más ineficaz de las modalidades presentes en esta investigación, esto se debe a múltiples factores, casi en igualdad de condiciones reprobaron debido a asistencia y calificaciones de los múltiples módulos, otro factor que destaca es la baja calificación en el examen final.

Class Visual

Visual es una modalidad en la que se presentaron 60 estudiantes, con un total de 19 estudiantes aprobados, lo que equivale al 31.67% y un total de 41 reprobados, que equivale al 68.33%, esta cantidad de reprobados se debe a las bajas calificaciones



obtenidas durante los módulos del curso de nivelación

4. CONCLUSIONES

En el presente trabajo se construyó un modelo inteligente basado en arboles de decisión para identificar estudiantes que pueden fracasar en su formación de nivelación, a partir de un grupo de variables relacionadas con las notas en la primera etapa del proceso de nivelación, así como información de estilo de aprendizaje que utilizan los estudiantes. El modelo obtenido con el algoritmo LMT fue el que mejores resultados obtuvo, el mismo identifica 83 casos correctos de 100 alternativas, lo que supone una taza de precisión del 83%. Además solo presenta un conjunto de 43 árboles con un promedio de 19 hojas, muy superior en este indicador a los resultados obtenido por el algoritmo j48. Para obtener estos valores, fue necesario realizar un proceso de preprocesamiento donde se seleccionaron las variables más relacionadas en el proceso y otro conjunto de técnicas para eliminar datos ruidosos, así como evitar el sobre entrenamiento del modelo.

5. REFERENCIAS

Bae, B. P. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Syst. Appl, 42*(6), 2928–2934.

Fayyad. (1996). *The KDD Process for Extracting Useful Knowledge from Volumes of Data.*

Frank, I. H. (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* Morgan Kaufmann.

García, G. Á. (2010). Minería de Datos en la Educación," , 2010. *Intel. en Redes Comun.*, 12--21.

Holte, R. C. (1993). Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. *Mach. Learn, 11*(1), 63–90.

Witten, I. H., & Eibe Frank, L. T. (2007). Weka: Practical Machine Learning Tools and Techniques with Java Implementations. *ICONIP/ANZIIS/ANNES'99 Workshop on Emerging Knowledge Engineering and Connectionist-Based Information Systems.*, (págs. 192–196).