Modelo Clasificador para Predecir el Desempeño Escolar Terminal de un Estudiante

Mario R. MORENO SABIDO
Depto. de Sistemas y Computación, Instituto Tecnológico de Mérida
Mérida, Yucatán 97118, México

y

Roberto MARTÍNEZ MALDONADO
Depto. de Sistemas y Computación, Instituto Tecnológico de Mérida
Mérida, Yucatán 97118, México

RESUMEN

En este artículo se describe el desarrollo de un modelo clasificador, el cual permite predecir el desempeño escolar terminal de un estudiante con base en la información que se pueda obtener sobre sus características socioeconómicas y académicas en los primeros ciclos escolares de su carrera universitaria. Esto es, conociendo sus antecedentes (socioeconómicos y académicos) el modelo puede predecir las posibilidades que tiene un estudiante que está iniciando sus estudios para terminar satisfactoriamente su carrera, y así se puedan tomar las medidas necesarias que coadyuven en su desarrollo académico. Para la construcción de los modelos se utilizaron los algoritmos Naïve Bayes y arboles de decisión C4.5. Los resultados obtenidos son capaces de predecir correctamente el desempeño terminal de los estudiantes en alrededor del 83% de los casos.

Este trabajo es de gran interés para cualquier institución que desee saber si sus estudiantes podrán concluir sus estudios, y lo más importante es que podrán hacerlo en etapas tempranas. También en caso de no obtener un resultado favorable para cierto estudiante, las instituciones tendrán bases acerca de lo que necesitan trabajar con él para que pueda concluir sus estudios satisfactoriamente. Otro de los usos del modelo clasificador es que establece el perfil que se necesita de un estudiante para cursar una determinada carrera.

Palabras Claves: Modelo Clasificador, Minería de Datos, Inteligencia de Negocios, Educación.

1. INTRODUCCIÓN

Uno de los principales problemas de investigación en el área del aprendizaje ha consistido en buscar los caminos que permitan motivar a los estudiantes y brindarles todas las herramientas para que ellos puedan, exitosamente, desarrollar su proceso de aprendizaje. Idealmente, si se conocieran los factores que determinan el desempeño escolar, se podrían establecer programas pedagógicos que apoyaran a los estudiantes que presenten dificultades observables en éstas áreas.

Un elemento fundamental para poder establecer un mecanismo de observación que permita identificar si un alumno tiene o tendrá dificultades académicas, con base en los factores socioeconómicos, psicológicos, personales, académicos y, en general, todos los factores ambientales que forman parte del entorno, resulta en comprobar que efectivamente estos factores afectan el desempeño escolar de una manera u otra. Es decir, es necesario hallar la existencia de tendencias entre el grupo de estudiantes, para establecer en qué medida son determinantes los factores ambientales en el rendimiento escolar.

Encontrar los factores que afectan el aprendizaje, el cual por sí mismo es difícil de medir, ha sido el objetivo de una gran cantidad de estudios realizados al respecto, los cuáles han sido abordados a partir de un abanico de diferentes puntos de vista [1]. Dentro del objetivo de esta investigación estuvo el acercar al tema de estudio un nuevo enfoque de análisis: obtener un modelo clasificador con un alto nivel de fiabilidad que permita predecir el desempeño escolar terminal de un estudiante, con base en la información que se pueda obtener sobre sus características socioeconómicas y académicas en los primeros ciclos escolares de su carrera universitaria, con apoyo de la minería de datos.

Para desarrollar este trabajo se utilizó la mayor cantidad de datos históricos posibles (socioeconómicos, académicos y datos personales) de los estudiantes del Instituto Tecnológico de Mérida, para buscar y establecer tendencias, estudiando la población de alumnos universitarios de catorce generaciones, aplicando métodos estadísticos, y estableciendo posibles predicciones para determinar las posibilidades que tiene un estudiante que está iniciando sus estudios, de terminar satisfactoriamente su carrera.

2. MARCO TEÓRICO

A lo largo del tiempo se han propuesto modelos conceptuales para estimar la importancia de las consecuencias que generan los factores ambientales en el alumno. Por ejemplo, Salagre y Serrano [2] proponen un modelo conceptual como la relación entre antecedentes familiares, influencia de amigos, características del centro educativo y habilidades personales.

Otras investigaciones han buscado los factores socioeconómicos influyentes en el desempeño escolar, tales como el género, la edad, el estado civil, becas, familia, o datos académicos como calificaciones, profesores, aulas donde estudia, nivel escolar, entre otros [3]. Dekker y Pechenizkiy [4] diseñaron un modelo basado en diferentes algoritmos clasificadores para predecir si un estudiante de ingeniería concluiría un semestre escolar. Dicho modelo se basa en el análisis de 13 atributos que definen el perfil preuniversitario del estudiante, incluyendo año escolar, edad, el número de cursos que ha estudiado y promedio de calificaciones en ciclos anteriores.

3. MÉTODO

Descripción General Del Método

El método a través del cual se realizó el acercamiento al problema de estudio forma parte de la inteligencia de negocios. *Inteligencia de negocios* se refiere al conjunto de tecnologías, aplicaciones, prácticas, conocimiento y habilidades usadas para obtener un mejor conocimiento del contexto de un negocio. En este proyecto las técnicas y tecnologías que se usaron, y que están relacionadas con la Inteligencia de negocios, forman parte de la Minería de Datos.

La *Minería de Datos* (DM) por las siglas en inglés Data Mining, es el proceso de extraer conocimiento útil y comprensible, previamente desconocido, a partir de grandes cantidades de datos almacenados, incluso, en distintos formatos [5].

Existen términos que se utilizan frecuentemente como sinónimos de la minería de datos. Uno de ellos se conoce como extracción o "descubrimiento de conocimiento en bases de datos" (Knowledge Discovery in Databases o KDD, según sus siglas en inglés).

Aunque algunos autores usan los términos Minería de Datos y KDD indistintamente, existen claras diferencias entre los dos. Así, la mayoría de los autores coinciden en referirse al KDD como un proceso que consta de un conjunto de fases, una de las cuales es la minería de datos. De acuerdo con esto, el proceso de minería de datos consiste únicamente en la aplicación de un algoritmo para extraer patrones de datos y se llamará KDD al proceso completo que incluye pre-procesamiento, minería y post-procesamiento de los datos. El KDD es la extracción automatizada de conocimiento o patrones interesantes, no triviales, implícitos, previamente desconocidos, potencialmente útiles y predictivos de la información de grandes Bases de Datos [6].

Proceso Metodológico

El proceso metodológico utilizado consta de 2 etapas principales: la minería de datos en sí, y la interpretación de los resultados (ver Figura 1).

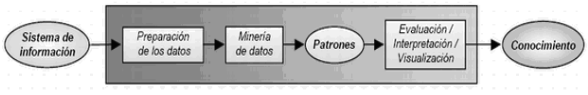

Figura 1. KDD.

Muestra

Por definición, la minería de datos pretende encontrar patrones con base en el análisis de grandes cantidades de datos. En este trabajo se hizo el análisis de toda la población correspondiente a alumnos de catorce generaciones del Instituto Tecnológico de Mérida. Este análisis se encontró limitado por la fecha inicial en la que se comenzaron a almacenar datos sistematizadamente en la institución. Sin embargo, en la fase de preparación de datos se discriminaron todas las instancias que no cumplieron con las restricciones mínimas de aceptabilidad para formar parte del historial de datos básicos para analizar.

Se utilizó la información correspondiente a catorce generaciones, sumando en total 10,520 alumnos. Los alumnos cuya información estuviera incompleta no fueron considerados como casos de estudio

Preparación De Los Datos

El sistema de información de la universidad anfitriona contiene la información socioeconómica, datos personales y académicos de todos los alumnos de 9 carreras de la institución. La mayor parte de estas carreras corresponden a ramas de ingeniería.

La primera etapa del estudio, preparación de datos, consistió en determinar las variables o factores con los cuales se entrenaría el modelo clasificador. En este proyecto se utilizaron todos los atributos presentes en la base de datos que contuvieran información personal, socioeconómica y académica de los tres primeros ciclos escolares de los alumnos. Así mismo, se seleccionaron los atributos que tuvieran datos válidos para la mayor parte de la población de alumnos estudiada. Además, en esta etapa se definió el dominio de cada variable, es decir, se establecieron los valores discretos que tomaron las mismas con base en la información contenida en el sistema de información preexistente. Se seleccionaron 21 atributos de los estudiantes. La Figura 2 enumera todas las variables que fueron consideradas. La variable dependiente a predecir puede tomar dos posibles valores: *egresado (EG)*, esto es, si el alumno concluye el grado completo de ingeniería (5,195 de los casos); o *baja (BD)*, si el alumno cursó tres o más ciclos escolares sin concluir sus estudios por completo (5,325 de los casos). La distribución de estos casos (5,195 egresados y 5,325 dados de baja) no toma en cuenta como fueron clasificados (correcta o incorrectamente) por el algoritmo C4.5.

En la segunda etapa se encuentra la minería de datos propiamente dicha. Los pasos a seguir para la realización de un proyecto de minería de datos son siempre los mismos, independientemente de la técnica específica de extracción de conocimiento usada. El proceso de minería de datos utilizado en este proyecto pasó por las siguientes fases:

1. Comprensión del negocio y del problema que se quiere resolver.
2. Filtrado de datos (dejar los datos crudos en un formato adecuado para el algoritmo)
3. Selección de variables.
4. Extracción del conocimiento.
5. Interpretación y evaluación.

Atributo	Tipo	Descripción
Residencia	Nominal	{ciudad, poblado_próximo}
Genero	Nominal	{M,F}
Edo_civil	Nominal	{Soltero, casado, unión, otro}
Edad_ingreso	Numérico	Edad de inicio de la carrera
Estatus_ingreso	Nominal	{Directo, propedéutico}
Carrera	Nominal	{Una de las 9 carreras incluídas en el estudio}
Becado	Nominal	{Si, No}
Promedio_semestre_1	Numérico	Promedio de calificaciones
Reprobadas_semestre_1	Numérico	Número de materias reprobadas en primer semestre.
Cursadas_semestre_2	Numérico	Número de materias cursadas en segundo semestre.
Promedio_semestre_2	Numérico	Promedio de calificaciones
Reprobadas_semestre_2	Numérico	Número de materias reprobadas en segundo semestre.
Cursadas_semestre_3	Numérico	Número de materias cursadas en tercer semestre.
Promedio_semestre_3	Numérico	Promedio de calificaciones
Reprobadas_semestre_3	Numérico	Número de materias reprobadas en tercer semestre.
Promedio_3_ingresos	Numérico	Promedio general de calificaciones
Verano_cursadas	Numérico	Número de materias cursadas en escuela de verano.
Verano_reprobadas	Numérico	Número de materias reprobadas en escuela de verano.
Revalidado	Nominal	{Si, No} Si comenzó en alguna otra escuela revalidando materias
Adelanto_semestre_2	Nominal	{Si, No} Si el alumno cursa materias de semestres posteriores
Adelanto_semestre_3	Nominal	{Si, No} Si el alumno cursa materias de semestres posteriores

Figura 2. Atributos.

Instrumentos

En la etapa de extracción de conocimiento es realmente en donde se aplicaron los instrumentos y las herramientas estadísticas. Para la minería de datos se usó la metodología CRISP-DM [7], la cual consiste en un conjunto de tareas descritas en cuatro niveles de abstracción: fase, tarea genérica, tarea especializada, e instancia de proceso, organizados de forma jerárquica en tareas que van desde el nivel más general hasta los casos más específicos.

Fase: Se le denomina fase al paso dentro del proceso. CRISP-DM consta de 6 fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelación, evaluación y explotación.

Tarea Genérica: Cada fase está formada por tareas genéricas, o sea, la tarea genérica es la descripción de las actividades que se realizan dentro de cada fase. Por ejemplo, la tarea *Limpiar los datos* es una tarea genérica.

Tarea Especializada: La tarea especializada describe cómo se pueden llevar a cabo las tareas genéricas en situaciones específicas. Por ejemplo, la tarea *Limpiar los datos* tiene tareas especializadas, como *limpiar valores numéricos*, y *limpiar valores categóricos*.

Instancias De Proceso: Las instancias de proceso son las acciones y resultados de las actividades realizadas dentro de cada fase del proyecto.

Las fases del proyecto de minería, de acuerdo a lo establecido por la metodología CRISP-DM, interactúan entre ellas de forma iterativa durante el desarrollo de la misma. La secuencia de las fases no siempre es ordenada.

Ahora bien, el instrumento que se utilizó es el programa de Sofware Open Source Weka, de la Universidad de Waikato en Nueva Zelanda [8].

Algoritmos

Los algoritmos que se utilizaron en ésta investigación son los siguientes: el primero es el algoritmo clasificador Naïve Bayes que se basa precisamente en el teorema de Bayes asumiendo una fuerte independencia entre todas las variables de entrada. Los modelos generados por este algoritmo asumen la presencia o ausencia de una característica en particular o clase; esta característica es totalmente independiente a la ausencia o presencia de cualquier otra característica o clase del modelo.

Aunque en el mundo real la mayoría de los sistemas son más complicados de lo que el modelo maneja, el clasificador Naïve Bayes resulta ser práctico por su sencillez y resultados aceptables. Este tipo de clasificadores han trabajado de manera eficaz resolviendo situaciones complejas del mundo real, por ejemplo, clasificando grandes cantidades de texto o en filtros anti-spam. Aunque no parece ser razonable que un modelo que no refleja la manera en que se comportan las variables en el mundo real pueda ser eficaz para predecir su comportamiento, ha sido probada la eficacia de los clasificadores bayesianos, tanto teóricamente como en forma práctica [9].

Aún así, otros métodos han demostrado mejor rendimiento que el Naïve Bayes. Primeramente, están los algoritmos de Bayes mejorados. Uno de ellos es el Bayes Tree o Árbol Naïve Bayes, que difiere del algoritmo Naïve Bayes en que sí considera las influencias que las variables pueden ejercer entre sí. El árbol Naïve Bayes realmente utiliza un árbol de decisión como estructura general y despliega el clasificador Naïve Bayes básico como ramas u hojas. La mejora que propone este árbol sobre el método Naïve Bayes básico es que permite una mayor precisión en la clasificación cuando se trabaja con mayores cantidades de datos.

Como tercer algoritmo de prueba se utilizó el C4.5. Este método fue generado por Ross Quinlan. Proviene de otro algoritmo, el ID3. Este algoritmo ha sido previamente utilizado en investigaciones de minería de datos en el área educativa para predecir el desempeño escolar de acuerdo al perfil del alumno [10]. Los árboles generados por este algoritmo son exhaustivos. El algoritmo se basa en el hecho de que cada característica o clase de datos se puede utilizar para tomar una decisión que divida los datos en subconjuntos más pequeños. C4.5 examina la ganancia de la información para elegir una característica que divida los datos, es decir, en qué grado la presencia o ausencia del valor de una variable hace que se dividan todos los registros de entrada. La característica o clase con la ganancia de información más alta es la que es usada para formar el nodo de decisión. El algoritmo entonces se repite en las sublistas más pequeñas, es decir, los nodos del árbol se van formando a partir de las características que dividan la mayor cantidad de los registros de entrada.

4. DESCRIPCIÓN DEL PROYECTO

Filtrado De Datos Y Selección De Variables
Se seleccionaron 21 variables que representan diferentes factores que repercuten en el desarrollo de un estudiante universitario. Se eligieron de esta manera debido a la disponibilidad y limpieza con que se encontraban en la base de datos institucional. Adicionalmente, también se determinó la variable dependiente, la cual fue la variable a observar; en este caso se averiguó si el estudiante en estudio va a egresar o no de la institución.

Preprocesamiento
En esta etapa se extrajeron los datos de la base de datos mediante consultas, con la finalidad de obtener los datos en el formato adecuado para aplicar sobre ellos algoritmos de minería de datos, de acuerdo a los dominios definidos para cada vabriable. De esta manera, se obtuvieron los registros que realmente pasaron por los filtros definidos. En la Figura 3 se muestra como la aplicación de minería de datos ha precargado los datos; además se contabilizaron las apariciones de cada métrica dentro del dominio de cada variable.

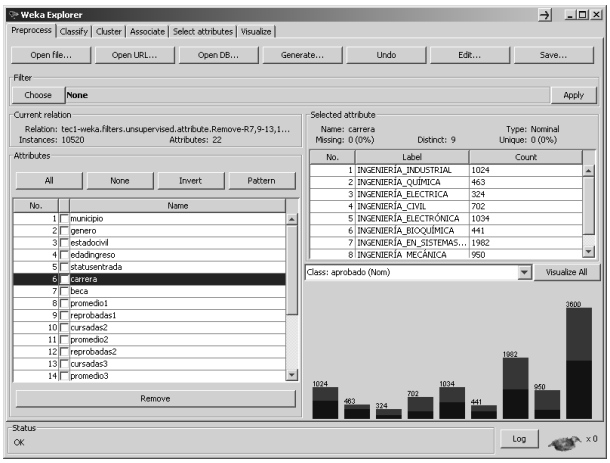

Figura 3. Preprocesamiento de Datos Históricos.

En la Figura 4 se puede observar la incidencia del promedio de calificaciones en el primer semestre.

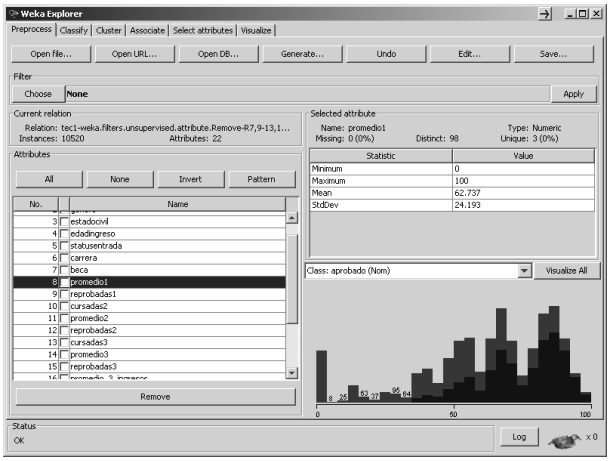

Figura 4. Incidencia de las Variables.

5. RESULTADOS

Después de analizar los resultados y comparar los algoritmos, estos fueron modificados hasta lograr obtener un modelo clasificador que permitiera predecir con alta fiabilidad, el desempeño escolar terminal de un estudiante.

Como se mencionó anteriormente, los algoritmos que se utilizaron fueron los siguientes: Naïve Bayes, Naïve Bayes Tree y el C4.5 (o J48). En resumen, el procedimiento que se aplicó para cada algoritmo fue el siguiente: la aplicación de minería de datos tomó el banco de datos conformado por una cantidad de 10,520 instancias provenientes de los datos históricos de varias generaciones de estudiantes del Instituto Tecnológico de Mérida, es decir, datos de 10,520 estudiantes, y posteriormente se generó un modelo por cada algoritmo. El modelo obtenido es un modelo matemático para el algoritmo Naïve Bayes y un árbol para los dos últimos. Luego, con base en este modelo, se clasificaron las 10,520 instancias con la finalidad de poder analizar la confiabilidad y el grado de acierto que tiene cada uno de los modelos generados. Con base en estos resultados se pudo notar si el modelo tiene utilidad como clasificador para esta investigación. Si se desea clasificar un nuevo caso de estudio se introduce como una nueva instancia y el mismo modelo lo clasificará.

Para evaluar el desempeño de los algoritmos clasificadores se utilizó el método de validación cruzada con 10 particiones. Este método consiste en tomar el 90% de los datos para entrenar el modelo y utilizar el restante 10% para validar el mismo, repitiendo este proceso diez veces para cubrir todos los datos. Los resultados que se presentan a continuación son los promedios de dichas particiones.

Después de realizar varias pruebas y ajustar varias veces los algoritmos, se pudo comprobar que el modelo generado por el algoritmo Naïve Bayes tuvo una exactitud (*accuracy*) del 79.9%. El modelo del Naïve Bayes Tree presentó una fiabilidad del 82.5%. Es importante mencionar que se utilizó el algoritmo C4.5 (o J48 en Weka) en 2 variantes: la primera prueba fue hecha podando el algoritmo, es decir, se optimizó para que genere pocas ramas, y la segunda fue dejándolo tal cual es (en

este caso, sin podar). De esta manera se obtuvieron 2 modelos clasificadores para este algoritmo. El algoritmo podado obtuvo una exactitud del 82.4%, en comparación cuando utilizó sin podar, en donde se obtuvo un modelo clasificador con una exactitud del 79.3%.

En la Figura 5 se presenta un resumen de los resultados obtenidos por los algoritmos. Los datos mostrados indican que los cuatro algoritmos ofrecen resultados similares, no obstante, el árbol de decisión Naïve Bayes y el algoritmo C4.5 ofrecen los más altos índices de exactitud (*accuracy*) pero al mismo tiempo, los más estables índices de precisión. El algoritmo Naïve Bayes presenta una gran precisión para clasificar la condición de baja definitiva y una alta tasa de positivos verdaderos para la condición de egreso. Sin embargo, en promedio, los resultados son significativamente inferiores a los dos algoritmos mencionados primeramente.

	Naïve Bayes	Naïve Bayes Tree	C4.5	C4.5 (sin podar)
Accuracy	79.9%	**82.5%**	**82.4%**	79.3%
Precision EG	73.4%	**80.2%**	**81.0%**	78.9%
Precision BD	**91.1%**	**85.0%**	**83.9%**	79.6%
Recall EG	**93.3%**	**85.6%**	**84.1%**	79.2%
Recall BD	67.0%	**79.4%**	**80.7%**	79.3%

Figura 5. Resumen de Resultados.

En la Figura 6 se presentan otros valores importantes arrojados por el modelo clasificador del algoritmo C4.5, el cual presentó un mayor número de aciertos (junto con el algoritmo Naïve Bayes Tree) y un razonablemente alto índice de precisión (alrededor del 83% en promedio). Las columnas *TP Rate* o *Recall* ilustran la taza de positivos verdaderos, la cual es la proporción, en este caso, de casos egresados (EG) que fueron correctamente identificados (84.1%) o de dados de baja (BD) dependiendo de qué resultados se tome como "positivo" (80.7% para BD). En pocas palabras, alrededor del 84% de los egresados fueron correctamente clasificados bajo este modelo, y por otro lado, alrededor del 81% de los dados de baja fueron correctamente clasificados. La columna *FP Rate* o de falsos positivos es la proporción de casos negativos incorrectamente clasificados como positivos, lo cual sería en consecuencia, el complemento de la taza *TP Rate*, es decir, el 82% y el 18% respectivamente (si se toman en cuenta los promedios).

La columna *Precision* es la tasa de aciertos con base en el total de las clasificaciones, es decir, el 81% de los clasificados como egresados fueron correctamente clasificados. Al mismo tiempo, alrededor del 84% de los clasificados como bajas, realmente fueron dados de baja. *F-1* combina *Recall (R)* y *Precision (P)* dándoles un peso combinado. Se calcula por medio de la fórmula $F\text{-}1 = 2RP/R+P$ [11]. Este parámetro puede también ser considerado para evaluar la efectividad de las predicciones de los modelos clasificadores.

TP Rate	FP rate	Precision	Recall	F-1	Clase
0.841	0.193	0.81	0.841	0.825	EG
0.807	0.159	0.839	0.807	0.823	BD
0.824	0.176	0.824	0.824	0.824	Promedio

Figura 6. Resultados de Precisión del C4.5.

El producto principal de modelo es la matriz de confusión, la cual contiene la información actual y las clasificaciones realizadas por el modelo clasificador. Se puede evaluar el rendimiento un modelo clasificador con base en ésta. En la Figura 7 se presenta la matriz de confusión generada por el C4.5.

EG	BD	< Clasificaciones Valores reales V
4369	826	EG
1027	4298	BD

Figura 7. Matriz de Confusión Generada por el C4.5.

En este caso, la matriz de confusión es de dos dimensiones dado que la variable de estudio EGRESADO puede tomar dos valores, Egresado (EG) o Baja Definitiva (BD), para definir si un alumno egresará satisfactoriamente o será dado de baja.

De acuerdo a la matriz, 4,369 instancias de entrada fueron correctamente clasificadas como egresados, 4,298 alumnos fueron correctamente clasificados como dados de baja. Por otro lado, 826 egresados fueron erróneamente clasificados como dados de baja y 1,027 bajas fueron clasificados erróneamente como egresados. De esta manera se pudo comprobar que el modelo generado es altamente aceptable. Los mejores indicadores para ello son los términos *Accuracy y Precision*, que es la tasa de predicciones correctas y el grado de desviación de los resultados respectivamente. En este caso se obtuvieron 8,667 instancias correctamente clasificadas que corresponden al 82.40% del total de las 10,520 instancias evaluadas. Adicionalmente, la *Precision* resultó altamente aceptable.

6. CONCLUSIONES

Con base en los resultados obtenidos y en los planteamientos iniciales de la investigación, se puede concluir que se obtuvo un modelo clasificador, el cual permite predecir con alto porcentaje de fiabilidad, si un estudiante podrá concluir sus estudios universitarios con base en las características socioeconómicas y académicas que le afectan.

Los algoritmos C4.5 y el Naïve Bayes Tree probaron ser los que presentan modelos clasificadores con mayor número de aciertos. Alrededor del 83% de las instancias estudiadas por dichos modelos son perfectamente clasificadas, de manera que el poder de predicción es bastante confiable. Además, se obtuvieron altos índices de precisión, arriba del 80% para ambas condiciones de egreso (EG) y baja definitiva (BD).

Otro punto importante que se tiene que mencionar es relativo a las variables independientes de entrada. La información con que se cuenta en el Instituto Tecnológico de Mérida, principalmente la relacionada con aspectos socioeconómicos de los estudiantes, es bastante limitada. La primera recomendación para futuras investigaciones, consiste en planear la búsqueda de patrones a largo plazo, establecer la información que podría resultar del interés de las investigaciones, y proyectar su recolección en los formatos correspondientes.

Para el desarrollo de este trabajo también se recurrió a información académica de los estudiantes. Ésta información sí ha sido recopilada de una manera más precisa, de tal manera que se trabajó con la información de los tres primeros semestres de estudio de un estudiante dentro del instituto. Por lo tanto, el poder predictivo del modelo y la utilidad de los resultados, requiere que un estudiante esté en tercer grado para que sea un nuevo caso de estudio.

7. REFERENCIAS

1. Chance, B. and J. Garfield, *New Approaches to Gathering Data on Student Learning for Research in Statistics Education.* Statistics Education Research Journal, 2001. **53**.
2. Salagre, D.J. and S.O. Serrano, *Determinacion de los factores que afectan al rendimiento académico en la educación superior*, in *XII Jornadas de la Asociación de Economía de la Educación*. 2003.
3. Valdivieso, M., K. Monar, and M.L. Granda, *Análisis de los determinantes del rendimiento de los estudiantes de ESPOL - 2002*. Revista Tecnológica, 2004. **17**(1).
4. Dekker, G.W., M. Pechenizkiy, and J.M. Vleeshouwers, *Predicting Students Drop Out: A Case Study*, in *2nd International Conference On Educational Data Mining*. 2009: Cordoba, Spain. p. 41-50.
5. Witten, I.H. and E. Frank, *Data Mining: Practical machine learning tools and techniques*. 2005: Morgan Kaufmann Pub.
6. Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth, *The KDD process for extracting useful knowledge from volumes of data*. Communications of the ACM, 1996. **39**(11): p. 27-34.
7. Chapman, P., et al., *CRISP-DM 1.0: Step-by-step data mining guide*. 2000: SPSS.
8. Witten, I.H. and U.o.W.D.o.C. Science, *Weka: Practical machine learning tools and techniques with Java implementations*. 1999: Citeseer.
9. Zhang, H., *The optimality of Naïve Bayes*, in *FLAIRS*. 2004. p. 3.
10. Vialardi, C., et al., *A Case Study: Data Mining Applied to Student Enrollment*, in *International Conference on Educational Data Mining*. 2010: Pittsburgh, USA. p. 333-334.
11. Yang, Y. and X. Liu, *A re-examination of text categorization methods*, in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. 1999, ACM: Berkeley, California, United States. p. 42-49.