Proyección de Estudiantes en Riesgo de Desertar Mediante Técnicas de Minería de Datos

Proyection of Students in Desertion Risk through Data Mining Techniques

Jhon Jaime Méndez Alandete[1]

Recibido: diciembre 17 de 2014
Aceptado: julio de 2015

Resumen

Uno de los problemas que se presentan en la universidad es el alto porcentaje de deserción y la dificultad para identificar las causas y prever los estudiantes en riego de desertar. El objetivo de esta investigación fue la predicción de estudiantes desertores aplicando técnicas de minería de datos, utilizando algoritmos de clasificación de caja blanca con reglas de inducción y árboles de decisión sobre diferentes conjuntos de datos. Se obtuvo, como resultado, un conjunto de reglas para predecir el estudiante desertor de la Corporación Universitaria del Caribe – CECAR.

Palabras clave: deserción, inteligencia de negocios, minería de datos.

Abstract

One of the problems presented in the university is the high percentage of desertion and the difficulty to identify its causes and foreseeing those students in desertion risk. The purpose of this research was to predict students desertion by applying data mining techniques using white box classification algorithms with induction rules and decision trees about different data sets. It was obtained as a result a set of rules to predict deserting students at Corporación Universitaria del Caribe – CECAR.

Keywords: desertion, business intelligence , data mining.

[1]Docente de la facultad de Ingeniería, Ciencias Básicas y arquitectura de la Corporación Universitaria del Caribe, CECAR
jhon.mendez@cecar.edu.co

1. Introducción

La deserción estudiantil es una problemática de las instituciones de educación superior de muchos países. En Colombia, las cifras de abandono en carreras universitarias llega al 45.3% [1]. En relación a esto, la inteligencia de negocio y en especial la minería de datos, la cual juega hoy por hoy un papel fundamental en las ventajas competitivas de una empresa, apoyando los procesos de toma de decisiones a través de la descripción y proyección de información útil y oculta en grandes bases de datos [2], sería una solución prometedora en pro de mitigar las tasas de deserción estudiantil en las instituciones de educación superior.

El uso de técnicas de minería de datos en el contexto educativo ha conllevado a la creación de una disciplina especial denominada minería de datos educativa [3], cuyo principal propósito es el desarrollo de métodos para la explotación de datos, provenientes de las instituciones educativas, que ayuden a entender mejor a los estudiantes y su entorno de aprendizaje [4]. La minería de datos educativa se ha aplicado con éxito en la deserción y retención de estudiantes [3][5][6], debido a la adquisición de conocimientos a través de la generación de modelos descriptivos que caractericen al estudiante en riesgo de desertar. Así mismo, se generan modelos predictivos que permiten determinar que estudiantes se encuentran en riesgo de abandono académico y con esto definir estrategias que apunten a disminuir los índices de deserción.

Instituciones de educación básica, media y superior han logrado disminuir los índices de deserción [7][8] y en algunos casos mejorar el rendimiento académico estudiantil [9] a través de la utilización de técnicas de minería de datos que les han permitido la generación de estrategias encaminadas a la retención y éxito académico de la población estudiantil.

En este artículo se propone la utilización de técnicas de minería de datos para detectar cuales son los factores que afectan en la deserción de estudiantes en riesgo de desertar. Además se utilizan diferentes datos teniendo en cuenta la alta dimensionalidad de estos y el desbalanceo que presentan. Lo anterior busca la detección temprana de estudiantes en riego de desertar y de esta forma definir estrategias que disminuyan el fenómeno de la deserción.

2. Metodología

La metodología usada para el desarrollo de esta investigación se basó en el proceso de extracción de conocimiento utilizando herramientas de minería de datos para la proyección de estudiantes en riesgo de desertar. El proceso en mención se compone de las siguientes etapas [10]:

- Selección de datos: se recopila toda la información académica y personal de los estudiantes y se integra en un solo conjunto de datos almacenados en un repositorio de datos distinto a las fuentes de donde provienen los datos.
- Pre procesamiento: se realiza la preparación y limpieza de los datos con el objetivo de adaptarlos y

posteriormente aplicar alguna técnica de minería de datos. Se excluyeron datos faltantes, redundantes e inconsistente derivados de las entrevistas realizadas a los estudiantes en su primer ingreso [11].

- Transformación: se modifican y particionan los datos. Además se aplican técnicas de selección de atributos con el fin de disminuir la alta dimensionalidad que presentan los datos. Por último, se hizo un rebalanceo de datos debido al desbalanceo del conjunto de datos.
- Minería de datos: se aplican técnicas de minería de datos tipo clasificación, previamente seleccionadas en base al criterio del mayor porcentaje de instancias clasificadas correctamente. Lo anterior con el objetivo de proyectar los estudiantes en riesgo de desertar.
- Interpretación y evaluación: en esta última etapa se analizan los resultados obtenidos por los algoritmos seleccionados, estudiando los factores que muestran las reglas y árboles de decisión, sus valores y las relaciones entre cada uno de ellos.

3. Resultados y discusión

A continuación se describe en detalle el caso de estudio aplicado a los estudiantes de pregrado presencial de la Corporación Universitaria del Caribe CECAR:

3.1. Selección de datos

Para el Ministerio de Educación existen tres factores fundamentales que alimentan los altos niveles de deserción en Colombia: razones de carácter económico, bajas competencias académicas y fallas en la orientación vocacional [2]. Para el caso de estudio hemos utilizado los datos de los estudiantes de pregrado presencial de la Corporación Universitaria del Caribe – CECAR. Tomando como base los estudiantes de nuevo ingreso de los períodos comprendidos entre 2010 y 2013. Se recopilaron datos desde dos fuentes distintas:

- Entrevista realizada a los estudiantes de nuevo ingreso de cada programa a través de un aplicativo desarrollado por el departamento de sistemas de CECAR.
- Datos académicos provenientes de las bases de datos del sistema de información de gestión académica desarrollado por el departamento de sistemas de CECAR.

En la Tabla 1 se muestran todas las variables agrupadas desde las dos fuentes de datos utilizadas como base.

3.2. Pre procesamiento

Normalmente antes de ejecutar alguna técnica de minería de datos se deben realizar algunas tareas de pre procesado, adaptando los datos originales al formato que exige el algoritmo de minería de datos seleccionado. Por ende se realizaron sobre los datos originales las tareas de integración y limpieza.

La integración consistió en agrupar todos los datos de los estudiantes desde las dos fuentes en una tabla sobre un repositorio de datos creado en el sistema manejador de base de datos postgres, utilizando la herramienta ETL libre talend.

Tabla 1. Variables utilizadas y fuente de datos

Fuente	Variables
Entrevista	Estado civil, estrato socioeconómico, sexo, lugar de residencia, si trabaja o no, edad de ingreso, grupo étnico al que pertenece, grupo familiar, personas a cargo, si posee alguna discapacidad y/o capacidad excepcional, número de salarios mínimos que recibe el núcleo familiar, ingreso familiar, si posee vivienda propia o no, si la vivienda posee alguna deuda, categoría de sisben en caso que lo posea, grado de escolaridad del padre y madre, sector en que labora el padre y la madre en caso que se encuentren laborando, número de hermanos que posee, si requiere ayuda financiera, si recibe algún apoyo financiero, entidad financiera que recibe apoyo, si requiere ayuda académica, si pertenece algún grupo vulnerable, si pertenece o no a una población con necesidades educativa especiales, si es víctima del conflicto armado, cuáles son las metas que posee, qué personas influenciaron en la carrera, si recibió alguna orientación vocacional, qué tipo de orientación vocacional en caso de haberla recibido, tiempo promedio que dedica a los estudios, qué cualidades y defectos posee (solo se registran tres de cada una a partir de una lista predefinida).
Bases de datos del Sistema de información de gestión académica	Código estudiantil, nombre completo, programa al que se encuentra matriculado, promedio estudiantil, número de asignaturas reprobadas a los largo de su estadía en la Corporación Universitaria, si alguna vez ha suspendido o desertado de algún programa académico al cual se encontraba matriculado en la Corporación.

En la fase de limpieza se eliminó la duplicidad que existía con algunos estudiantes y se filtraron aquellos que poseían datos inconsistentes o por el contrario no poseían estos, como por ejemplo lugar de residencia, edad de ingreso y colegio donde realizó sus estudios de secundaria.

3.3. Transformación

En la etapa de transformación se realiza una discretización del promedio acumulado de los estudiantes cambiando el formato numérico de 1 a 5 a un formato nominal o categórico teniendo en cuenta los siguientes rangos: si el promedio acumulado se encuentra entre 1 y 3 se considera bajo de 3,1 a 4 medio y de 4,1 a 5 alto. A continuación se genera un archivo con formato CSV desde el repositorio de datos y se modifican algunos datos que contenían caracteres especiales como comillas dobles, remplazando estos por espacios en blanco. Lo anterior, se realiza debido a que la herramienta de minería de datos presenta problemas en la lectura de los anteriores.

Seguidamente desde la herramienta WEKA se crea un archivo con formato ARFF y se modifica el atributo *desertó* de numérico a class, de forma que se tome como base para la aplicación de técnicas de minería de datos.

Después de realizar las anteriores actividades se obtiene un archivo con formato ARFF con 54 atributos y 2084 estudiantes. Con este archivo se crean 10 particiones divididas aleatoriamente con el objetivo de realizar una validación cruzada en las pruebas de clasificación. Las particiones son utilizadas en la etapa de entrenamiento (training) y pruebas (test).

Teniendo en cuenta la gran cantidad de atributos recopilados (54), se realiza un análisis de selección de atributos para determinar cuáles afectan en mayor medida la deserción académica de los estudiantes de CECAR. Para esto, se utilizaron varios algoritmos de selección de atributos disponibles en WEKA, entre los cuales se usaron CfsSubsetEval, Consistency-SubsetEval y FilteredSubsetEval algoritmos de tipo filtro que seleccionan y evalúan los atributos en forma independiente al algoritmo de aprendizaje y ChiSquared-AttributeEval, Filtered-AttributeEval, GainRatio-AttributeEval, InfoGain-AttributeEval y SymmetricalUncert-AttributeEval algoritmos de tipo wrappers que usan el desempeño de algún clasificador (algoritmo de aprendizaje) para determinar las características deseables de un subconjunto [12]. Los resultados obtenidos por cada algoritmo se muestran en la Tabla 2.

Para seleccionar los mejores atributos se contabilizó el número de veces que un atributo fue seleccionado por algún algoritmo y se escogieron aquellos cuya frecuencia de aparición sea igual o mayor a 2. Aplicando lo anterior, solo quedaron 11 atributos los cuales se muestran en la

Tabla 2. Atributos seleccionados por algoritmos de selección de atributos

Algortimo	Método Escogido	Atributos de importancia seleccionados
CfsSubsetEval	BestFirst	Promedio Estudiantil y colegio.
ChiSquared-AttributeEval.	Ranker	Colegio, municipio colegio, promedio estudiantil, número de asignaturas reprobadas, tipo de orientación vocacional, tipo colegio, edad ingreso, sexo, sector laboral del padre, recibió orientación vocacional y vivienda propia.
Consistency SubsetEval	BestFirst	Sexo, promedio estudiantil, edad ingreso, número de asignaturas reprobadas, vivienda propia, sector laboral del padre, colegio, municipio colegio, recibió orientación vocacional y tipo de orientación vocacional.
Filtered-AttributeEval	Ranker	Colegio, municipio colegio, promedio estudiantil, número de asignaturas reprobadas, tipo de orientación vocacional, edad ingreso, sexo, sector laboral del padre, tipo colegio, recibió orientación vocacional y vivienda propia.
FilteredSubsetEval	BestFirst	Promedio estudiantil y colegio.
GainRatio-AttributeEval	Ranker	Promedio estudiantil, colegio, municipio colegio, número de asignaturas reprobadas, tipo colegio, tipo de orientación vocacional, edad ingreso, sector laboral del padre, sexo, vivienda propia y recibió orientación vocacional.
InfoGain-AttributeEval	Ranker	Colegio, municipio colegio, promedio Estudiantil, número de asignaturas reprobadas, tipo de orientación vocacional, edad ingreso, tipo colegio, sexo, sector laboral del padre y vivienda propia.
SymmetricalUncert-AttributeEval	Ranker	Promedio estudiantil, colegio, municipio colegio, número de asignaturas reprobadas, tipo colegio, edad ingreso, sector laboral del padre, sexo, vivienda propia y recibió orientación vocacional.

Tabla 3 con su frecuencia de aparición. A partir de éstos se crea un fichero con formato ARTFF con 11 atributos y 2084 estudiantes, al que nuevamente se particiona en 10 archivos de datos para entrenamiento y prueba.

Por último, al revisar los datos de los ficheros no particionados se evidencia un desbalanceo entre estudiantes desertores y no desertores, 1752 contra 337 respectivamente. La problemática de utilizar datos no balanceados en algoritmos de clasificación es que los clasificadores tienen en la etapa entrenamiento una clase mayoritaria que conlleva a clasificar en la etapa de prueba elementos de la clase minoritaria con una baja probabilidad.

Para solucionar el problema descrito anteriormente se realiza un balanceo de distribución de clases utilizando el algoritmo SMOTE (SyntheticMinority Oversampling Technique) [13] disponible en Weka como un filtro de datos.

Para el caso de estudio se rebalanceo los 10 ficheros particionados con los 11 mejores atributos de forma que se tuviera un 50% de instancias de estudiantes no desertados y 50% de estudiantes desertados.

Después de realizar las tareas de selección de datos, pre-procesado y transformación se cuenta con:

- 10 archivos de entrenamiento y prueba con todos los atributos
- 10 archivos de entrenamiento y prueba con los mejores atributos
- 10 archivos de entrenamiento y prueba con los mejores atributos rebalanceados.

3.4. Minería de datos

En esta etapa se describen las actividades de minería de datos utilizadas para la obtención de modelos de predicción de estudiantes en riesgo de desertar. Con el fin de obtener la máxima exactitud se han realizado varios experimentos sobre

Tabla 3. Atributos de mayor frecuencia de aparición en los algoritmos de selección de atributos escogidos.

Atributo	Frecuencia de aparición
Promedio estudiantil	8
Colegio	8
Municipio colegio	6
Número de asignaturas reprobadas	6
Edad Ingreso	6
Sector laboral del padre	6
Sexo	6
Vivienda propia	6
Recibió orientación vocacional	5
Tipo colegio	5
Tipo de orientación vocacional	5

diferentes archivos de entrenamiento y prueba con distintos algoritmos de clasificación, seleccionado 4 de inducción de reglas de clasificación (JRip, NNge, OneR y Ridor) y 5 de árboles de decisiòn (J48, SimpleCart, ADTree, RandomTree y REPTree) disponibles en la herramienta de minería de datos WEKA. Los algoritmos se seleccionaron en base a la facilidad de comprensión de los modelos de salida resultantes para usuarios no expertos, obteniendo reglas de clasificación "Si - entonces" que representan el conocimiento de una manera simple y sencilla comprensible por todo tipo de usuarios finales o árboles de decisión que muestran condiciones organizadas en una estructura jerárquica que contienen nodos de tipo interno o hojas.

En el primer experimento se ejecutaron los 9 algoritmos seleccionados realizando validaciones cruzadas sobre las 10 particiones con todos los atributos. Con cada algoritmo se obtiene la media de las ejecuciones sobre cada partición (10 ejecuciones), los resultados se muestran en la Tabla 4, indicando el porcentaje de instancias clasificadas correctamente e incorrectamente, coeficiente kappa y el porcentaje de error absoluto.

De la Tabla 4 se observa que los algoritmos que tuvieron mayor porcentaje de instancias clasificadas son el SimpleCart, REPtree, JRip y ADTree.

En el segundo experimento se ejecutaron los 9 algoritmos seleccionados realizando validaciones cruzadas sobre las 10 particiones con los mejores atributos. Al final se comprueba cómo afectan los porcentajes en relación al primer experimento. La Tabla 5 muestra los resultados del segundo experimento.

Al comparar los resultados de las tablas 4 y 5 se puede evidenciar que todos los algoritmos han mejorado el porcentaje de instancias clasificadas correctamente. Se puede observar que los algoritmos con mejores porcentajes son: REPTree, SimpleCart, Ridor y J48.

Tabla 4. Validación cruzada con los 54 atributos que afectan la deserción académica en CECAR.

Algortimo	% instancias clasificadas correctamente	% instancias clasificadas incorrectamente	Coeficiente Kappa	Error absoluto
SimpleCart	83,19076	17,01924	0	0,28163
REPTree	82,88461	17,11539	-0,00188	0,28116
JRip	82,50001	17,49999	0,25061	0,24953
ADTree	82,06732	17,93268	0,23983	0,27153
J48	81,97116	18,02884	0,14889	0,25892
Ridor	81,68268	18,31732	0,16853	0,18316
OneR	80,57693	19,42307	0,00198	0,19422
Random Tree	79,03846	20,96154	0,03525	0,26558
NNge	70,62501	29,37499	0,02405	0,29375

Fuente: cálculos del estudio.

Tabla 5. Validación cruzada con los 11 mejores atributos que afectan la deserción académica en CECAR.

Algortimo	% instancias clasificadas correctamente	% instancias clasificadas incorrectamente	Coeficiente Kappa	Error absoluto
REPTree	84,18269	15,81731	-0,00034	0,26183
SimpleCart	84,13461	15,86539	-0,00372	0,25839
Ridor	82,40386	17,59614	0,05771	0,17597
J48	82,35578	17,64422	0,06182	0,25133
OneR	82,01922	17,98078	-0,02836	0,17981
JRip	81,49039	18,50961	0,11652	0,25037
ADTree	81,20194	18,79806	0,1416	0,26843
Random Tree	79,27884	20,72116	0,03799	0,25203
NNge	72,88463	27,11537	-0,02701	0,27115

Fuente: cálculos del estudio.

La Tabla 6 muestra los resultados del tercer experimento realizando el mismo procedimiento anterior, pero sobre las particiones con los mejores atributos balanceados.

Al comparar los resultados de las tablas 5 y 6 se observa que la mayoría de algoritmos mejoraron sus porcentajes, pero en el caso de REPTree, SimpleCart, J48 y NNge disminuyeron. Pero, en general la tendencia es de mejora.

3 Interpretación y evaluación

A continuación se analizan los modelos de reglas y árboles de clasificación generados por los algoritmos con los mejores porcentajes de clasificación obtenidos en el la etapa de experimentación con los porcentajes más altos: OneR, ADTree, JRip y Ridor.

La Tabla 7 muestra las reglas de clasificación obtenidas por el algoritmo OneR, el cual solo descubre como regla los colegios de donde provienen los estudiantes de CECAR y tienen un riego de desertar.

Tabla 6. Validación cruzada con los 11 mejores atributos rebalanceados que afectan la deserción académica en CECAR.

Algortimo	% instancias clasificadas correctamente	% instancias clasificadas incorrectamente	Coeficiente Kappa	Error absoluto
OneR	88,350618	17,2077	0,76700909	0,116481
ADTree	87,701782	12,298218	0,75402727	0,248336
JRip	85,051345	14,948655	0,70102727	0,215190
Ridor	83,60925	16,39075	0,67219	0,16393
REPTree	83,482745	16,517255	0,66964545	0,271063
NNge	83,057227	16,942773	0,66115455	0,169436
Random Tree	81,965491	14,1877083	0,63443	0,22477
J48	78,0527	21,9473	0,56106364	0,284945
SimpleCart	66,776536	33,223464	0,33554545	0,343409

Tabla 7. Reglas obtenidas usando ONER con los 11 mejores atributos que afectan la deserción académica en CECAR.

Liceo Moderno Del Litoral -> 1
Intitucion Educativa San Jorge -> 1
Institución Educativa Nuestra Señora De Fàtima-> 1
Institución Educativa Miguel De Servantes Saavedra -> 1
Cedi -> 1
Institución Educativa Luis Carlos Galán -> 1
Institución Educativa Santiago Apostol -> 1
Colegio Tercer Milenio -> 1
Institución Educativa De Macajan -> 1
Instituto Luis Carlos Galan -> 1
Institución Educativa Gabriel Taboada Santodomingo-> 1
Instituto Educ. San Marcos -> 1
Institución Educativa Rural Pto Claver -> 1
Colpec -> 1
Inst. Educ. Antonio Prieto -> 1
Institución Educativa Liceo 20 De Julio -> 1
Institución Educativa Liceo Politecnico Del Sinu -> 1
Institución Educativa Indigena Tecnico Agropecuaria De Escobar Arriba -> 1
Liceo Bartolomé De Las Casas -> 1
Liceo De Cervantes Saavedra -> 1
Institución Educativa Para Poblaciones Especiales Inpes -> 1
Instituto Sur Oriental San Jorge -> 1
Institución Educativa Tecnico Agropecuario Artesanal -> 1
Inst. Educativa Normal Superior Sincelejo -> 1
Institución Educativa Comfasucre -> 1
Institución Educativa Juan Jacobo Rosean -> 1
Institución Educativa Comfasucre -> 1
Institución Educativa Juan Jacobo Rosean -> 1
Institución Educativa San Martin Inesam -> 1
Institución Educativa Liceo Moderno Del Litoral -> 1
Institución Educativa Tecnica Comercial Maria Inmaculada -> 1
Institución Educativa Avanzado Icsa-> 1
Institución Educativa Nuevo Milenio -> 1
Liceo Moderno El Litoral -> 1
Institución Educativa Tecnica Alvaro Ulcue Chocue -> 1
Institución Educativa Limasor -> 1
Institución Educativa Cucuta -> 1
Colegio San Rafael -> 1
Instituto Técnico Juan Mejia Gomez-> 1
Institución Educativa Los Palmitos. -> 1
Institución Educativa Igmnasio Del Rosario -> 1
Institución Educativa San Jose DE MAJAGUAL -> 1

En el árbol de decisión mostrado en la Tabla 8 generado por el algoritmo ADTree se evidencian las siguientes reglas: primera, los estudiantes con promedio acumulado bajo[1], número de asignaturas reprobadas entre 1 y 7 están en riesgo de desertar. Segunda, los estudiantes que provengan del liceo Bartolomé de las Casas se encuentran en riesgo de abandono académico. Por último, estudiantes con promedio acumulado medio[1] o alto[1], número de asignaturas reprobadas menor a 4 y no recibieron orientación vocacional o edad de ingreso superior a los 18 años corren la misma suerte.

Los resultados obtenidos por el algoritmo JRip, indicados en la Tabla 9, evidencian una regla donde aquellos estudiantes con riesgo de desertar presentan un promedio bajo[1], un número de asignaturas reprobadas entre 1 y 7 y recibieron ayuda vocacional de tipo charla o test.

Por su parte, la Tabla 10 muestra varias reglas generadas por el algoritmo Ridor para estudiantes con riesgo de desertar. La primera indica que aquellos estudiantes con promedio estudiantil bajo, edad de ingreso entre 17 y 18 años, tipo de colegio público y tipo de orientación

Tabla 8. Reglas obtenidas usando ADTree con los 11 mejores atributos que afectan la deserción académica en CECAR.

```
: -0.823
| (1)promedioestudiantil < 1.5: 0.672
| | (5)numeroasignaturasreprobadas < 7.5: 0.097
| | (5)numeroasignaturasreprobadas >= 7.5: -0.429
| (1)promedioestudiantil >= 1.5: -0.279
| | (2)edadingreso < 18.5: -0.169
| | (2)edadingreso >= 18.5: 0.233
| | (3)numeroasignaturasreprobadas < 4.5: 0.035
| | | (4)recibioorientacionvocacional < 1.5: 0.125
| | | | (7)colegio = INSTITUCION EDUCATIVA GABRIEL TABOADA SANTODOMINGO: 1.244
| | | | (7)colegio != INSTITUCION EDUCATIVA GABRIEL TABOADA SANTODOMINGO: -0.024
| | | (4)recibioorientacionvocacional >= 1.5: -0.137
| | (3)numeroasignaturasreprobadas >= 4.5: -0.602
| | | (9)edadingreso < 17.5: -1.137
| | | (9)edadingreso >= 17.5: 0.166
| | (6)colegio = INSTITUCION EDUCATIVA NORMAL SUPERIOR DE SINCELEJO: -1.28
| | (6)colegio != INSTITUCION EDUCATIVA NORMAL SUPERIOR DE SINCELEJO: 0.011
| (8)colegio = INSTITUCION EDUCATIVA SIMON ARAUJO: -0.717
| (8)colegio != INSTITUCION EDUCATIVA SIMON ARAUJO: 0.011
| (10)colegio = LICEO BARTOLOME DE LAS CASAS: 0.944
| (10)colegio != LICEO BARTOLOME DE LAS CASAS: -0.009
Legend: -ve = 0, +ve = 1
Tree size (total number of nodes): 31
Leaves (number of predictor nodes): 21
```

[1]Promedio discretizado en la fase de transformación.

Tabla 9. Reglas obtenidas usando JRip con los 11 mejores atributos que afectan la deserción académica en CECAR.

promedioestudiantil <= 1) and (numeroasignaturasreprobadas <= 7) and (tipoorientacionvocacional >= 5) => deserto=1 (38.0/15.0)

Tabla 10. Reglas obtenidas usando Ridor con los 11 mejores atributos que afectan la deserción académica en CECAR.

deserto = 0 (2089.0/337.0) Except (promedioestudiantil <= 1.5) and (numeroasignaturasreprobadas <= 7.5) and (edadingreso <= 18.5) and (edadingreso > 17.5) and (tipocolegio > 1.5) and (tipoorientacionvocacional > 2.5) => deserto = 1 (7.0/0.0) [3.0/1.0] Except (promedioestudiantil <= 1.5) and (numeroasignaturasreprobadas <= 7.5) and (edadingreso <= 18.5) and (edadingreso <= 16.5) and (tipocolegio <= 1.5) and (recibioorientacionvocacional > 1.5) => deserto = 1 (12.0/1.0) [9.0/4.0] Except (promedioestudiantil <= 1.5) and (viviendapropia <= 0.5) and (situacionlaboralpadre > 1.5) => deserto = 1 (22.0/4.0) [12.0/6.0] Total number of rules (incl. the default rule): 4 Time taken to build model: 0.08 seconds

vocacional de tipo test están en riego de abandono académico. Otra regla indica que estudiantes con promedio estudiantil bajo, edad de ingreso menor o igual a 18 años, tipo de colegio público y no recibió orientación vocacional corren riesgo de desertar. Por último, se tiene que estudiantes con promedio bajo, carecen de vivienda propia y sus padres no poseen trabajo se encuentren en igual medida propensos a abonar los programas académicos que se encuentran cursando.

Finalmente, es importante reseñar que después analizar los resultados obtenidos por los algoritmos seleccionados de clasificación, se ha detectado dos factores comunes presentes en todos los resultados que influyen en la deserción de los estudiantes de la Corporación Universitaria de Caribe - CECAR: el promedio estudiantil y el número de asignaturas reprobadas.

De igual forma, se pueden señalar otros factores, que si bien no están presentes en los resultados de algunos algoritmos vale la pena su estudio y análisis de influencia en el fenómeno de la deserción estudiantil de CECAR, como son: edad de ingreso, si recibió orientación vocacional, tipo de orientación vocacional que recibió, tipo y colegio donde realizó los estudios de secundaria, si posee vivienda propia y si el padre se encuentra laborando.

4. Conclusiones

En el proyecto se realizaron varios experimentos sobre distintas particiones con diferentes algoritmos con el fin de obtener el mayor grado de exactitud en la predicción de estudiantes con riesgo de desertar. En relación a la metodología aplicada y los resultados obtenidos, se concluye lo siguiente:

- La importancia de utilizar algoritmos de selección de atributos cuando se tiene una gran cantidad de características debido a que mejoran los porcentajes de instancias clasificadas correctamente.
- En el caso del aplicar técnicas de clasificación es importante determinar si los datos se encuentran balanceados. En caso que no se encuentren se recomienda balancearlos utilizando algoritmos como el SMOTE (Synthetic Minority Oversampling Technique) debido a que junto con la selección de atributos mejoran los porcentajes de instancias clasificadas correctamente.
- Los algoritmos seleccionados de tipo "caja-blanca" generan resultados que son fáciles de entender por usuarios no expertos en minería de datos, facilitándoles la comprensión de las reglas y árboles de decisión generados por éstos y de esta forma apoyar el proceso de toma de decisiones en relación al fenómeno de deserción.
- Los factores que más influyen en el fenómeno de la deserción académica de la Corporación Universitaria del Caribe – CECAR son el promedio estudiantil y el número de asignaturas cursadas. Otros factores que influyen en menor medida son la edad de ingreso, si recibió orientación vocacional, tipo de orientación vocacional que recibió, tipo y colegio donde realizó los estudios de secundaria, si posee vivienda propia y si el padre se encuentra laborando.

A partir de los modelos de reglas y árboles de decisión generados por los algoritmos seleccionados de clasificación, se creó un sistema de alertas que envía a través de correo electrónico una alerta al terminar cada corte académico al personal encargado de definir estrategias para mitigar el fenómeno de la deserción estudiantil.

Agradecimientos

El autor agradece la ayuda prestada por el departamento de sistemas de CECAR en la entrega y suministro de información relacionada a los sistemas de información.

Referencias

[1] Cardozo. Duque. (2014). Deserción, lunar en educación superior [Online]. Disponible en: http://www.elcolombiano.com/BancoConocimiento/D/desercion_lunar_en_educacion_superior/desercion_lunar_en_educacion_superior.asp

[2] Pinzón, Liza. "Aplicando minería de datos al marketing educativo", Notas D Marketing 1, vol. 1, pp 45- 61, 2011.

[3] A. Hall and G. Holmes. "Benchmarking Attribute Selection Techniques for Data Mining", Technical Report 00/10, University of Waikato, Department of Computer Science, Hamilton,

New Zealand, Julio 2002. Available: http://www.cs.waikato.ac.nz/~ml/publications/2000/00MH-GH-Benchmarking.pdf. M

[4] R. Baker y K. Yacef. "The State of Educational Data Mining in 2009: A Review and Future Visions", Journal of Educational Data Mining, Vol. 1, pp 3-16, 2009.

[5] A. Salazar, J. Gosalbez, I. Bosch, R. Miralles. "A case study of knowledge discovery on academic achievement, student desertion and student retention" en Information Technology: Research and Education, 2004. ITRE 2004. 2nd International Conference on, 2004, p 150-154.

[6] S. Kotsiantis, K. Patriarcheas and M. Xenos. "A Combinational Incremental Ensemble of Classifiers as a Technique for Predicting Students´ Performance in Distance Education", Knowledge Based System, vol. 23, no. 6, pp. 529-535, 2010.

[7] F. Díaz, M. Osorio, A. Amadeo y D. Romero. "Aplicando estrategias y tecnologías de Inteligencia de Negocio en sistemas de gestión académica" en XV Workshop de investigadores en ciencias de la computación, 2013.

[8] (2012) Govloop Website. [En línea]. Disponible en http://www.govloop.com/profiles/blogs/using-business-intelligence-tools-to-improve-school-districts.

[9] (2010) Educause website. [En línea]. Disponible en http://www.educause.edu/ero/article/signals-applying-academic-analytics.

[10] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence. (1996).

[11] E. Espíndola, A. León. "La Deserción Escolar en América Latina un Tema Prioritario Para la Agenda Regional", Revista Iberoamericana de Educación, no. 30, pp. 1-17, 2002.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, W.P. Kegelmeyer. "Synthetic Minority Over-sampling Technique", Journal of Artificial Intelligence Research, 2002, 16:321-357.

[13] C. Márquez, C. Romero y S. Ventura. "Predicción del Fracaso Escolar mediante Técnicas de Minería de Datos", IEEE-RITA, Vol. 7, pp 109-117, Nov. 2012.