Formación Universitaria
E-ISSN: 0718-5006
citrevistas@gmail.com
Centro de Información Tecnológica
Chile

Eckert, Karina B.; Suénaga, Roberto
Análisis de Deserción-Permanencia de Estudiantes Universitarios Utilizando Técnica de Clasificación en Minería de Datos
Formación Universitaria, vol. 8, núm. 5, 2015, pp. 3-12
Centro de Información Tecnológica
La Serena, Chile

Formación Universitaria
Vol. 8(5), 3-12 (2015)
doi: 10.4067/S0718-50062015000500002

Análisis de Deserción-Permanencia de Estudiantes Universitarios Utilizando Técnica de Clasificación en Minería de Datos

Karina B. Eckert y Roberto Suénaga
Universidad Gastón Dachary, Departamento de Ingeniería y Ciencias de la Producción,
Av. López y Planes 6511, Posadas, Misiones-Argentina
(e-mail: karinaeck@gmail.com, rsuenaga@ugd.edu.ar)

Recibido Ene. 14, 2015; Aceptado Mar. 20, 2015; Versión final May. 19, 2015, Publicado Oct. 2015

Resumen

Se analiza información académica con el objetivo de identificar factores que influyen sobre la deserción de los estudiantes de la carrera de Ingeniería en Informática de la Universidad Gastón Dachary en Argentina, mediante la aplicación de una técnica de minería de datos. La fuente de datos contiene información proporcionada al ingreso (personales y antecedentes educativos) y la que se genera durante el periodo de estudios. Se realiza la selección y depuración de datos, utilizando diferentes criterios de representación y aplicación de algoritmos de clasificación como árboles de decisión, redes bayesianas y reglas. Se identifica como variables influyentes en la deserción, asignaturas aprobadas, cantidad y resultado de asignaturas cursadas, procedencia y edad de ingreso del estudiante. Mediante este proceso fue posible identificar los atributos que caracterizan a los casos de deserción y su relación con el desempeño académico, especialmente en el primer año de la carrera.

Palabras clave: deserción universitaria, minería de datos, algoritmos de clasificación, estudios universitarios

Analysis of Attrition-Retention of College Students Using Classification Technique in Data Mining

Abstract

Academic information is analyzed to identify the factors that have more impact on desertion of students of Computer Science Engineering of the University Gastón Dachary in Argentina, by applying data mining techniques. The data source comes from the information provided by the student when they entered the university (personal and educational background) and information generated during the studies. Data are selected and analyzed using different criteria for the representation and application of classification algorithms such as decision trees, bayesian networks and rules. Influential variables on desertion are identified: passed courses, number and grades of courses, origin and age of student when he/she entered the university. Through this process it was possible to identify several variables that characterize the cases of desertion and its relation with academic achievement, especially during the first year of study.

Keywords: college retention, data mining, classification algorithms, university studies

INTRODUCCIÓN

La deserción es un fenómeno presente en todo sistema educativo, relacionado con los procesos de selección, rendimiento académico y de la propia eficiencia del sistema en general, es decir, el resultado de la combinación y efecto de distintas variables (Diaz Peralta, 2008). En este sentido, la deserción de estudiantes universitarios vinculado al desempeño académico de los mismos, es un tema que preocupa desde hace varios años. Se han realizado estudios con el objeto de aportar información que contribuya a determinar cuáles son las causas. (Martínez Padilla y Pérez González, 2008) identificaron que las variables relacionadas con la trayectoria académica que mayor efecto tiene en la estimación del desempeño corresponden al promedio general alcanzado en la enseñanza media, el rendimiento académico y la cantidad de materias que fueron reprobadas durante su permanencia en la universidad; determinando así el grado de éxito y fracaso de los estudiantes mexicanos para el examen nacional de egreso de la licenciatura en ingeniería.

Otros estudios (Soria Barreto y Zúñiga Jara, 2014), determinaron que las principales variables que resultaron estadísticamente determinantes en el éxito de los estudiantes fueron, en el orden de importancia, las calificaciones obtenidas en la enseñanza media, el puntaje obtenido en la prueba de aptitud académica de matemáticas, y el número de años de desfase entre el año de egreso de la enseñanza media y el año de ingreso a la universidad. Además, (Díaz, 2009) concluye que los estudiantes de ingeniería, presentan altos riesgos de deserción entre el primer y tercer semestre, siendo máximo en este último semestre, para luego descender y permanecer a tasas más estables.

Muchos de los trabajos relacionados a la problemática (Araque, 2009) se enfocan en la identificación de los factores que más afectan a la deserción, el fracaso y el rendimiento de los estudiantes; toda esa información permanece en las base de datos de las instituciones educativas. Esta gran cantidad de datos almacenados representa una oportunidad para obtener información valiosa sobre los estudiantes (Márquez Vera et al., 2012). (Mercano Aular y Talavera Pereira, 2007) afirman que la capacidad para almacenar datos ha crecido exponencialmente en los últimos años, mientras que la capacidad de procesarlos no ha sido así, y además afirman, que la toma de decisiones efectivas depende de la rapidez con que se identifica y analiza información importante. Esta última afirmación resulta difícil de cumplir si se utilizan métodos clásicos de procesamiento y por tanto resulta necesario aplicar nuevas técnicas adaptadas específicamente para cada caso, que permitan identificar y encontrar información útil, oculta en grandes bases de datos (Quadril y Kalyankar, 2010). Las herramientas de minería de datos (MD), basadas en técnicas inteligentes, facilitan el procesamiento avanzado de datos y permiten realizar un análisis en profundidad de los mismos de manera automática (Britos et al., 2005; Pérez López y Santín González, 2007). Su fortaleza se debe a que forma parte del proceso denominado Descubrimiento de Conocimiento en Base de Datos (KDD), cuyo objetivo es la búsqueda de patrones de datos que sean válidos, novedosos, potencialmente útiles y comprensibles (Fayyad et al., 1996).

La minería de datos en la educación (MDE) no es un concepto nuevo, su estudio y aplicación ha tomado mayor relevancia en los últimos años. La utilización de las técnicas de MD permite deducir fenómenos dentro del ámbito educativo; de esta forma, es posible determinar la probabilidad de desertar o continuar con sus estudios de los estudiantes, así como el desempeño de los mismos durante el cursado. El producto final de los modelos beneficia a estudiantes, docentes, padres y gestores de la educación, no sólo para informar sobre la situación de los estudiantes cuyo desempeño podría estar asociado con una característica particular (positiva o negativa), sino también como asesoramiento para la toma de decisiones. Dicho de otra manera, se pretende que estos modelos finales faciliten la reflexión y la autorregulación durante los estudios (Anand Kumar y Uma, 2009; Ramaswani y Bhaskaran, 2009, Romero y Ventura, 2010).

En otras investigaciones, se han empleado con éxito las técnicas de MDE para la creación de modelos predictivos del rendimiento de los estudiantes (Kotsiantis et al., 2010). Un caso de aplicación (Más Estellés et al., 2010), demuestra los resultados de un estudio sobre rendimiento académico de los estudios de Informática (ingenierías técnicas y superiores) de ocho universidades españolas públicas. En éste, se estudiaron de manera separada los alumnos de nuevo ingreso y el total de alumnos de una titulación, desagregados por sexo, edad, nota de ingreso y procedencia. Globalmente, los alumnos de nuevo ingreso en Informática presentaron perfiles diferenciados según estudien la titulación técnica o la superior. La edad resultó ser un factor determinante para explicar el abandono el primer año de estudios y la calificación en el ingreso en la titulación refleja diferentes duraciones medias de los estudiantes en la carrera (aquellos alumnos que acceden a una titulación con nota más alta, tardan menos años en titularse, y lo opuesto).

En (Superby et al., 2006) han analizado los factores de deserción en el primer año de la carrera en base a datos recabados mediante un cuestionario específico para el estudio, donde se consideraron factores internos y externos a la universidad; han utilizado cuatro técnicas y concluyen que el nivel de aproximación

en la estimación de los factores de deserción dependen en gran medida de la propia universidad ya que obtuvieron resultados dispares entre las tres instituciones consideradas en el estudio. Otro trabajo (Herzog, 2006) afirma que en base a un estudio de la retención y tasa de graduación en carreras de posgrado, los algoritmos de minería de datos, trabajando con grandes bases de datos, logran mejores resultados en cuanto a la determinación del tiempo de graduación que identificando factores de retención de alumnos.

Considerando afirmaciones como las de (Planck Barahona, 2014) que indican que diversos estudios han logrado determinar que las calificaciones obtenidas al inicio de la universidad guarda una estrecha relación con el rendimiento académico posterior; se incluyó como parte del proceso el análisis del rendimiento académico de los alumnos en el primer año académico y calendario. Como lo afirma (García Ortiz et al., 2014), los factores o variables que inciden en el rendimiento académico son diversos y corresponden a múltiples interacciones de los estudiantes y su entorno.

Este trabajo se centra en el análisis de variables relacionadas directamente con los resultados académicos del estudiante y su interacción con la universidad, basado en los datos que se obtienen del trayecto del estudiante en la carrera. Se propone la utilización de la técnica de clasificación en MD para detectar, cuáles son las características y los factores de mayor incidencia en los estudiantes de la carrera Ingeniería en Informática de la Universidad Gastón Dachary (UGD), con relación a la suspensión o abandono de sus estudios. Para ello, se propone la utilización de tres algoritmos de clasificación para una mayor confiabilidad de los resultados. El problema abordado es complejo, debido a que los datos pueden presentar una alta dimensionalidad (muchas variables o características que pueden influir) y suelen estar desbalanceados (muchos estudiantes suelen aprobar y sólo algunos pocos desertan). El objetivo es detectar con anterioridad cuales son los estudiantes que presentan características relacionadas con la posibilidad de abandono y así proveer contención o ayuda especial y así evitar y/o disminuir los casos de deserción estudiantil; la elección de la carrera es debido a que es la que presenta mayor cantidad de casos de deserción, en relación a las demás carreras (relacionadas a la administración, turismo, derecho, nutrición, entre otras) dictadas en la universidad.

METODOLOGÍA

Esta investigación analiza la situación académica de alumnos universitarios en base a los datos de su trayectoria en la UGD. La muestra sobre la cual se trabajó corresponde a los estudiantes de la carrera de Ingeniería en Informática, modalidad presencial, 5 años de duración y la tesis de grado. El período seleccionado para el estudio corresponde a los estudiantes ingresados desde el año 2000 al 2009, totalizando 855 casos analizados.

El atributo para determinar los casos de deserción, es decir si los estudiantes abandonan o no sus estudios es de tipo dicotómico, deserta (“Des”) y no deserta (“NoDes”). Para ello se tomó como atributo la condición final, de tipo nominal, que puede adoptar uno de los cuatro valores posibles: egresado, en curso, baja temporal y baja definitiva. La baja definitiva indica que el alumno ha desertado de la carrera (“Des”) y las demás condiciones, para su procesamiento, fueron agrupadas como no deserción (“NoDes”); cabe aclarar que la baja temporal hace referencia a la suspensión temporal de la actividad del alumno. Como rendimiento académico se considera el grado de éxito de los estudiantes, relacionado con la obtención de buenas calificaciones, escasos exámenes reprobados, pocos o ninguna materia re-cursada, cursado y aprobación sin retraso respecto al plan de estudios de la carrera.

Tabla 1: Atributos seleccionados y estandarizados

Atributos Seleccionados	Estandarización	Tipo de dato
Condición de Deserción	Deserción	Nominal: Des, NoDes
Total de Finales Aprobados de 1°año	1°Apr	Numérico
Proporción de Materias Cursadas del Año 1 (calendario)	Curs1	Numérico
Cantidad de Fracasos de Cursado del Año 1 (calendario)	FracC1	Numérico
Número de Finales Aprobados en el Año 1 (calendario)	Apro1	Numérico
Promedio General de 1° Año	PromA1°	Numérico
Promedio Materias Aprobadas de 1° Año	PromG1°	Numérico
Edad de Ingreso	EdadI	Numérico
Establecimiento educativo (previo)	Est	Nominal: Bachi, EscEdMed, Cen, Tec, Com, Inst, Col, EdSup, Nor, Otros.
Localización geográfica (de origen)	Loc	Nominal: Pdas, IntProv, Otras

Los atributos seleccionados para el estudio luego de la integración, recopilación y filtrado de los datos (los detalles se describen posteriormente) se presentan en la Tabla 1, allí se incluye la descripción y la

denominación estandarizada de cada atributo, así como el tipo de dato y valores posibles para los nominales. Cabe aclarar que año calendario corresponde a la fecha de la actividad (por ejemplo: actividades realizadas en el año 2005), y por año académico corresponde al ordenamiento anual del plan de estudio de la carrera (por ejemplo: primer año de ingeniería en informática).

Como metodología de análisis se implementa el Proceso KDD, que consta de las fases: *Integración y Recopilación de Datos, Filtrado de Datos, Minería de Datos y Evaluación e Interpretación de Resultados*. Durante el desarrollo del proceso de KDD, como consecuencia de los resultados intermedios, es frecuente interrumpir la secuencia de fases del proceso, para volver a retomar en alguno de los pasos anteriores, siendo así un proceso iterativo e interactivo necesario para lograr una alta calidad del conocimiento a descubrir (Fayyad et al., 1996; Hernández Orallo et al., 2004).

En minería de datos, existe la necesidad de determinar en qué nivel de madurez se encuentran los procesos y modelos, y si son adecuados para resolver el o los problemas planteados, por lo que deben ser revisados, interpretados y evaluados, y finalmente concluir si es posible extraer conocimiento significativo.

Partiendo de los datos operacionales provenientes de la base de datos de la Universidad, se llevó a cabo una interpretación del dominio de la aplicación que se refiere a la información registrada de índole académico (fase de integración y recopilación de datos). Debido a la gran cantidad de atributos disponibles, más de 50 tablas almacenadas en una base de datos relacional, que cuentan con información de los estudiantes de índole personal, asistencias a clases, calificaciones, etc.; se realizó una recopilación de atributos para determinar los de mayor relevancia respecto a la condición de deserción. En la fase de filtrado de datos se realizó un control y depuración exhaustiva de los datos para hallar una coherencia completa de las tablas y subconjuntos de datos a utilizar.

Se utilizaron dos técnicas de selección de atributos disponibles en la herramienta Weka (descripta posteriormente). La primera técnica utiliza algoritmos que se distinguen por su forma de evaluar los atributos, clasificándose en: filtros, donde se seleccionan y evalúan los atributos en forma independiente del algoritmo de aprendizaje, y envoltorios (wrappers), los cuales usan el desempeño de algún clasificador (algoritmo de aprendizaje) para determinar lo deseable de un subconjunto. Otra técnica aplicada es la denominada 'selección de atributos', utilizada para identificar, en base a un atributo en particular, cuales son los que más inciden sobre el atributo objeto (en este caso la condición de deserción). Esto permite a su vez, optimizar posteriores pruebas y resultados a obtener con la técnica de clasificación, sobre todo para evitar clasificaciones muy complejas, como por ejemplo árboles de decisiones extensos y por ende difíciles de interpretar. El método de evaluación aplicado es CfsSubsetEval y el de búsqueda BestFirst, los que ofrecen una selección de subconjuntos de atributos de mayor calidad según (Witten y Eibe, 2005). Se han probado alternativas a los algoritmos para cada método, pero a los efectos prácticos, no se han encontrado variaciones significativas en los resultados finales.

El modo de evaluación utilizado en los algoritmos de selección de atributos y de clasificación, es el de validación cruzada, el cual divide *n* veces el mismo conjunto de datos mutuamente excluyente y de igual tamaño; *n-1* conjuntos se utilizan para construir el clasificador y con el conjunto restante se valida (particiones estratificadas); y así las particiones de test no se superponen. El clasificador final se construye con todos los subconjuntos de datos y la precisión se obtiene del promedio total. El número de subconjuntos o pliegues de validación cruzada utilizados en el estudio es de 10, lo que provoca que la evaluación sea lenta, pero precisa. (Witten y Eibe, 2005; Molina López y García Herrero, 2006; García Castellano, 2009).

Como atributo-indicador se consideraron diversos parámetros, entre los más relevantes podemos mencionar: promedio de asignaturas aprobadas; promedio general; condición final de cursado de materias (si el alumno regulariza o no la materia al final del curso); calificaciones obtenidas en exámenes finales; graduación final (obtención del título); abandono (deserción); entre otros, quedando seleccionados los que se presentan en la Tabla 1.

Una vez definido, dispuesto y adecuado el conjunto de datos a procesar (vista minable) se procede a la aplicación de técnicas y algoritmos de MD (fase de minería de datos). La técnica utilizada es la de clasificación, basada en un modelo destinado a predecir la categoría de instancias en función de una serie de atributos de entrada, a partir del cual el clasificador aprende un esquema de clasificación de los datos.
El primer algoritmo utilizado es C4.5, que genera un árbol de decisión a partir de las variables disponibles, mediante particiones realizadas recursivamente, según la estrategia de primero en profundidad, su implementación en WEKA se denomina J48. (Witten y Eibe, 2005; Molina López y García Herrero, 2006).

El segundo algoritmo empleado se denomina Naïve Bayes aumentado a árbol (Tree Augmented Network (TAN)), como todos los clasificadores Bayesianos, se basan en el teorema de Bayes, conocido como la

fórmula de la probabilidad de las causas. Naïve Bayes (NB) es una simplificación que ha demostrado una alta exactitud y velocidad cuando se ha aplicado a grandes volúmenes datos. El modelo TAN de manera general obtiene mejores resultados que NB, manteniendo la simplicidad computacional y la robustez; el conjunto de padres del atributo a clasificar C, es vacio, mientras que el conjunto de variables padres de cada uno de los atributos predictores Xi, contiene necesariamente al atributo a clasificar, y como mucho otro atributo (Mitchell, 1997; Gámez y Puerta Callejón, 1998; Duda et al., 2000; Hernández Orallo et al., 2004; García Castellano, 2009). La implementación del algoritmo TAN en WEKA se denomina BayesNet..

Como último algoritmo a evaluar, se escogió a OneR, el cuál es uno de los algoritmos clasificadores más sencillos y rápidos; dado que simplemente identifica el atributo que mejor explica la clase de salida. Si hay atributos numéricos, busca los umbrales para hacer reglas con mejor tasa de aciertos (Witten y Eibe, 2005). Al utilizar un clasificador, su precisión y fiabilidad depende principalmente de los casos clasificados correctamente a partir del número total de elementos (fase de evaluación e interpretación de resultados).

La herramienta de MD utilizada para la investigación es WEKA, la cual se caracteriza por utilizarse bajo licencia GNU, y además se diseñó específicamente para ser utilizada en investigación y con fines educativos. El paquete WEKA contiene una colección de herramientas de visualización, algoritmos para el análisis de datos, modelado predictivo y descriptivo, unido a una interfaz gráfica de usuario para acceder fácilmente a sus funcionalidades (WEKA; Witten y Eibe, 2005).

RESULTADOS

Algoritmo Clasificador C4.5 - Árbol de Decisión

Los resultados de la herramienta se presentan en forma de esquema y gráficamente. A partir de los atributos de entrada, se obtiene como resultado una serie de condiciones representadas de forma escrita mediante un conjunto de reglas, condiciones del tipo si-sino (if-else) y gráfica mediante un árbol de decisión. En la Figura 1 se puede apreciar el conjunto de reglas generado para clasificar los casos de deserción (“Des”) y permanencia (“NoDes”). El nodo o condición inicial representa la cantidad de exámenes finales aprobados correspondientes al primer año de la carrera (“1°Apro”), donde se dividen en dos sub-clasificaciones, una para cantidades de materias aprobadas menores o iguales a siete y para las mayores a siete.

El segundo criterio de clasificación en ambos casos es la localización geográfica de donde proviene los estudiantes (“Loc”), la cual se encuentran discriminados en Posadas (“Pdas”) (capital de provincia), interior de la provincia (“IntProv”) u otros (otra provincia o país). El tercer criterio de clasificación varía según la combinación de condiciones anteriores, para instancias con un número de asignaturas menores o iguales a siete, para los estudiantes provenientes del interior de la provincia, en este nivel se hacen presentes los atributos que refieren a la edad de ingreso (“Edadl”), en cambio para alumnos de la ciudad de Posadas aparece el número de asignaturas cursadas en el primer año calendario (“Curs1”) y para los provenientes de “otras” localizaciones geográficas, la cantidad de asignaturas que el estudiante no ha podido regularizar o promocionar (“FracC1”); por otro lado, para un número de asignaturas aprobadas en primer año superior a siete, se identificó relevante para estudiantes provenientes de “otras” localizaciones o de Posadas, al número de asignaturas aprobadas en el primer año calendario (“Apro1”), para “otras” localidades el punto de corte es tres, para “Pdas” siete, y para los provenientes del interior de la provincia (“IntProv”) se predijo de forma directa que no desertarán.

Algoritmo Clasificador Naïve Bayes Aumentado a Árbol (TAN)

En la Figura 2 se puede observar el grafo obtenido para la predicción de la condición de deserción de los estudiantes, donde todos los atributos involucrados se relacionan al nodo padre atributo/nodo objeto a clasificar. Por cada uno de los atributos se pueden visualizar la relación probabilística que posee en relación al atributo objeto (condición de deserción). Para algunos atributos el algoritmo no se detectó relación probabilística con la condición de deserción, esto ocurrió para en número de fracasos en el cursado (“FracC1”), el promedio general (“PromG1°”) y de materias aprobadas (“PromA1°”) en el primer año y la edad de ingreso del estudiante.

De los atributos probabilísticamente significativos, en la parte inferior de la figura se muestra la relación que tiene el atributo correspondiente a la cantidad de exámenes finales aprobados en el primer año calendario en relación a la condición de deserción; a la izquierda se observa el atributo padre con sus posibles valores, a su derecha los intervalos respecto al número de asignaturas aprobadas y sus respectivas probabilidad, donde se destacan los casos de deserción en los que aprueban un máximo de tres asignaturas en el primer año académico con una probabilidad del 0.612 para esa condición (valor ,“Des”); para las instancias que

refieren a una continuidad de sus estudios, se observa que el 0.719 de este valor (“NoDes”) corresponde a estudiantes que aprueban entre cuatro y ocho finales el primer año dentro de la universidad. Respecto al número de materias cursadas del primer año (“1°Curs”), se obtuvo una probabilidad de deserción del 0.58 para el rango de 0 a 5 materias y casos de no deserción con el 0.48 entre 6 y 8, y para más de 8 materias un 0.32 de probabilidad. De las materias aprobadas del primer año (“1°Apro”), entre 0 y 3 materias se predicen casos de deserción en un 0.54 y al contrario, con más de 7 materias la probabilidad es de 0.7 para los casos que no desertan. Las materias cursadas en el año que ingresa a la universidad (“Curs1”), predice que desertan en un 0.70 casos si cursan entre 0 a 8 y una probabilidad de 0.57 que no desertan con más de 8 materias cursadas. El atributo localización geográfica predice permanencia en una probabilidad de 0.42 para Posadas (“Pdas”) y para el resto de la provincia (“IntProv”) de 0.27, y paro los casos no incluidos en los anteriores (valor, “otras”) una probabilidad de 0.74 de los casos de deserción. En relación a los establecimientos previos, la distribución de probabilidades es muy homogénea.

```
J48
------------------
1°Apro <= 7
|  Loc = Otras
|  |  FracC1 <= 5
|  |  |  PromA1° <= 5.25
|  |  |  |  1°Curs <= 6
|  |  |  |  |  Curs1 <= 8: Des
|  |  |  |  |  Curs1 > 8
|  |  |  |  |  |  PromA1° <= 5.14: Des
|  |  |  |  |  |  PromA1° > 5.14: NoDes
|  |  |  |  1°Curs > 6: NoDes
|  |  |  PromA1° > 5.25: Des
|  |  FracC1 > 5
|  |  |  1°Apro <= 5
|  |  |  |  PromG1° <= 2.9: NoDes
|  |  |  |  PromG1° > 2.9: Des
|  |  |  1°Apro > 5: NoDes
|  Loc = Pdas
|  |  Curs1 <= 5: Des
|  |  Curs1 > 5
|  |  |  Est = Com
|  |  |  |  Apro1 <= 4: Des
|  |  |  |  Apro1 > 4: NoDes
|  |  |  Est = Inst
|  |  |  |  EdadI <= 22
|  |  |  |  |  1°Apro <= 5
|  |  |  |  |  |  Curs1 <= 11
|  |  |  |  |  |  |  EdadI <= 17: NoDes
|  |  |  |  |  |  |  EdadI > 17: Des
|  |  |  |  |  |  Curs1 > 11: Des
|  |  |  |  |  1°Apro > 5: NoDes
|  |  |  |  EdadI > 22: NoDes
|  |  |  Est = Otros
|  |  |  |  1°Apro <= 5
|  |  |  |  |  PromA1° <= 6.9: Des
|  |  |  |  |  PromA1° > 6.9: NoDes
|  |  |  |  1°Apro > 5: NoDes
|  |  |  Est = Col
|  |  |  |  EdadI <= 19: Des
|  |  |  |  EdadI > 19: NoDes
|  |  |  Est = EdSup: Des
|  |  |  Est = Nor
|  |  |  |  PromG1° <= 3.57: NoDes
|  |  |  |  PromG1° > 3.57: Des
|  |  |  Est = Tec
|  |  |  |  Curs1 <= 10: Des
|  |  |  |  Curs1 > 10: NoDes
|  |  |  Est = Bachi: NoDes
|  |  |  Est = EscEdMed: NoDes
|  |  |  Est = Cen: Des
|  Loc = IntProv
|  |  EdadI <= 22
|  |  |  PromG1° <= 3.18: NoDes
|  |  |  PromG1° > 3.18
|  |  |  |  Curs1 <= 10: Des
|  |  |  |  Curs1 > 10: NoDes
|  |  EdadI > 22
|  |  |  PromG1° <= 7.63: NoDes
|  |  |  PromG1° > 7.63
|  |  |  |  1°Curs <= 2: NoDes
|  |  |  |  1°Curs > 2: Des
1°Apro > 7
|  Loc = Otras
|  |  Apro1 <= 3
|  |  |  1°Apro <= 9: Des
|  |  |  1°Apro > 9: NoDes
|  |  Apro1 > 3
|  |  |  1°Apro <= 9
|  |  |  |  Curs1 <= 7: NoDes
|  |  |  |  Curs1 > 7
|  |  |  |  |  Apro1 <= 4: NoDes
|  |  |  |  |  Apro1 > 4
|  |  |  |  |  |  1°Apro <= 8: Des
|  |  |  |  |  |  1°Apro > 8
|  |  |  |  |  |  |  1°Curs <= 9: NoDes
|  |  |  |  |  |  |  1°Curs > 9: Des
|  |  |  1°Apro > 9: NoDes
|  Loc = Pdas: NoDes
|  |  Apro1 <= 7
|  |  |  1°Apro <= 9: Des
|  |  |  1°Apro > 9: NoDes
|  |  Apro1 > 7
|  |  |  PromG1° <= 7.5: NoDes
|  |  |  PromG1° > 7.5: Des
|  Loc = IntProv: NoDes
```

Fig. 1: Clasificación del Algoritmo J48 (C4.5) para predecir casos de deserción y permanencia

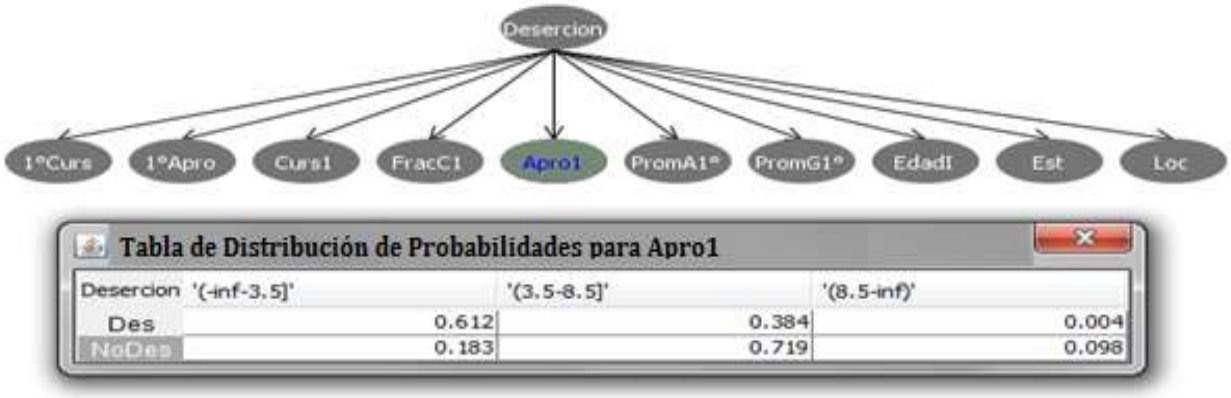

Fig. 2: Clasificación del Algoritmo BayesNet (TAN) para predecir casos de deserción y permanencia

Algoritmo Clasificador Reglas OneR

Para la predicción de la condición de deserción, se identifica como atributo condicionante a la cantidad de materias aprobadas del primer año de la carrera ("1°Apro"), con punto de corte en el valor 7, indicando que por debajo de éste número, se clasifica como casos de deserción, así como ocurre con los casos donde no aprobaron ninguna materia ("?": indica que el campo está vacío) y para las instancias con 7 o más materias aprobadas, como casos de permanencia (esto ocurrió en el 76,6 % de los casos (655 de 855)); lo mencionado se encuentra reflejado en las reglas obtenidas y representadas en la Figura 3.

```
=== OneR ===

1°Apro:
        < 7.5    -> Des
        >= 7.5   -> NoDes
        ?        -> Des
```

Fig. 3: Clasificación Algoritmo OneR para predecir casos de deserción y permanencia

En la Tabla 2 se representa el detalle de las evaluaciones de resultados obtenidas por los tres algoritmos de clasificación (J48, TAN y OneR). Para cada uno los algoritmos, se puede apreciar en la segunda columna el porcentaje de instancias clasificadas correctamente (ICC). Podemos observar que el algoritmo J48 obtuvo mayor porcentaje de ICC (80,23%) y en contrapartida el algoritmo que menor porcentaje de ICC es para el algoritmo OneR (76,61%), estos valores se deben en gran medida a la robustez de los algoritmos.

A partir de la tercera columna se evalúan la precisión detallada por clases para los dos valores posibles que puede tomar la variable objetivo, "Deserción" ("Des", "NoDes"),; en la cuarta columna, para cada uno de los valores ("Des", "NoDes"), los verdaderos positivos (VP) representan la proporción de instancias que están clasificados dentro de una clase, de entre todos los elementos que realmente son de la clase, por ejemplo se clasifica una instancia como un caso de deserción ("Des") y efectivamente corresponde como tal; el caso contrario a los VP, son los FP ubicados en la quinta columna, representan la proporción de casos que han sido clasificados dentro de una clase, pero pertenecen a una clase diferente (FP: falsos positivos), siguiendo con el ejemplo anterior, son los casos donde se clasifican como no desertores y corresponden a casos de deserción.

Finalmente, la última columna indica la precisión, es decir, la proporción de casos que realmente son de una clase de entre todos los elementos que han sido clasificados dentro de la misma. Con relación a la precisión obtenida para los casos que desertan ("Des") o no desertan ("NoDes"), podemos ver de qué manera general los 3 algoritmos clasifican con más exactitud los casos que no desertan, donde el algoritmo OneR obtuvo el porcentaje superior (82,3%) en relación a los demás; y para los casos de deserción, el algoritmo que obtuvo mayor precisión es J48 (79,7%).

Tabla 2: Evaluaciones de Resultados de los Algoritmos de Clasificación

Algoritmos	ICC	Deserción	VP	FP	Precisión
J48 (C4.5)	80.234%	Des	79.1%	18.7%	79.7%
		NoDes	81.3%	20.9%	80.8%
BayesNet (TAN)	78.129%	Des	81.0%	24.5%	75.3%
		NoDes	75.5%	19.0%	81.1%
OneR	76.608%	Des	83.7%	30.0%	72.1%
		NoDes	70.0%	16.3%	82.3%

CONCLUSIONES

Las herramientas de MD brindan resultados que deben ser interpretados y traducidos a diagnósticos y consecuencias del ámbito real (en este caso la universidad). Esto implica que los resultados de la aplicación de las técnicas se han utilizado para explicar parte del comportamiento de la situación en cuanto a la permanencia y su determinación a partir del desempeño académico de los estudiantes. Las posibles consecuencias y acciones tendientes a la toma de decisiones específicas, está sujeta a consideraciones de otros integrantes del cuerpo académico de la institución educativa.

En el proceso KDD, la preparación y acondicionamiento de los datos es la etapa más extensa y a la vez fundamental porque en gran medida los resultados posteriores dependen de ésta. En las etapas intermedias, es crítico llevar a cabo análisis e interpretaciones de resultados parciales, ya que a partir de éstos se retoma el proceso y continúa la depuración y refinamiento del conocimiento extraído.

Mediante la aplicación de algoritmos de minería de datos llevada a cabo en el presente trabajo se pudo identificar que durante el primer año de la carrera es donde adquieren mayor importancia las acciones de contención, apoyo, tutoría y todas aquellas actividades que mejoren la situación académica del alumno al ingreso en la universidad.

Se detectaron atributos que al procesarlos y asociarlos a criterios específicos, se relacionan fuertemente con la deserción y permanencia, el principal de ellos es cantidad de asignaturas aprobadas del primer año, debido a que marca una tendencia notable sobre el resto de la carrera; otros que se destacan son: el número de asignaturas cursadas, los casos donde el estudiante no regulariza la materia al cursarlas, la edad de ingreso, la procedencia; la combinación de estos criterios obtuvo porcentajes de aciertos, de entre un 76% y un 80% de los casos clasificados correctamente.

En el análisis de los resultados obtenidos de los algoritmos de clasificación C4.5 (J48), TAN (BayesNet) y OneR, se pudo observar porcentajes de aciertos similares, sin embargo, no identifican exactamente los mismos atributos. Incluso, dentro de un mismo modelo (por ejemplo, el árbol de decisión C4.5), no todos los atributos tienen la misma importancia, existiendo algunos que el método no lo considera significativo y que podría suponerse importantes (por ejemplo, el establecimiento educativo previo de los estudiantes).

REFERENCIAS

Anand Kumar N. V. y Uma G. V., "*Improving Academic Performance of Students by Applying Data Mining Technique"*, ISSN: 1450-216X, European Journal of Scientific Research, 34 (4), 526-534, (2009)

Araque F. y otros dos autores, *"Factors Influencing University Drop Out Rates",doi:10.1016/j.compedu.2009.03.013,* Computers & Education, 53, 563–574, (2009)

Britos P. V. y otros tres autores, "*Minería de Datos basada en Sistemas Inteligentes"*, 1ª ed., Nueva Librería, Buenos Aires, Argentina, (2005)

Díaz C. J., "*Factores de Deserción Estudiantil en Ingeniería: Una Aplicación de Modelos de Duración"*, doi: 10.1612/inf.tecnol.4095it.08, Información Tecnológica, 20 (5), 129-145, (2009)

Díaz Peralta C., "*Modelo Conceptual para Deserción Estudiantil Universitaria Chilena"*, doi: 10.4067/S0718-07052008000200004, Estudios pedagógicos, 34 (2), 65-86, (2008)

Duda R. O. y otros dos autores, "Pattern Classification", 2ª ed., Wiley - Interscience, USA y Canadá, (2012)

Fayyad U. M. y otros tres autores, *"Advances in Knowledge Discovery & Data Mining"*, 1ª ed., MIT Press Cumberland, Rhode Island, EE.UU, (1996)

Gámez Martín J. A. y Puerta Callejón J.M., "*Sistemas Expertos Probabilísticos"*, Cuenca: Universidad de Castilla-La Mancha, (1998)

García Castellano F. J., "*Modelos Bayesianos para la clasificación supervisada. Aplicaciones al análisis de datos de expresión genética"*, Tesis Doctoral, Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada, –Granada, España, (2009)

García Ortiz Y. y otros dos autores, *"Estudiantes universitarios con bajo rendimiento académico, ¿qué hacer?"*, http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2077-28742014000200018&lng=es&nrm=iso, ISSN: 2077-2874, EDUMECENTRO, 6 (2), 272-278, (2014)

Hernández Orallo J. y otros dos autores, *"Introducción a la Minería de Datos"*, 1ª ed., Pearson, Madrid, España, (2004)

Herzog S., *"Estimating Student Retention and Degree-Completion Time: Decision Trees and Neural Networks Vis-à-Vis Regression"*, doi: 10.1002/ir, New Directions for Institutional Research, 131(1), 17-33, (2006)

Kotsiantis S., Patriarcheas K. y Xenos M., *"A Combinational Incremental Ensemble of Classifiers as a Technique for Predicting Students´ Performance in Distance Education"*, doi: 10.1016 / j.knosys.2010.03.010, Knowledge Based System, 23(6), 529-535, (2010)

Márquez Vera C. y otros dos autores, *"Predicción del Fracaso Escolar mediante Técnicas de Minería de Datos"*, http://rita.det.uvigo.es/201208/uploads/IEEE-RITA.2012.V7.N3.A1.pdf, ISSN 1932-8540 ,IEEE-RITA, 3 (7), 109-117, (2012)

Martínez Padilla J. H. y Pérez González J. A., *"Efecto de la Trayectoria Académica en el Desempeño de Estudiantes de Ingeniería en Evaluaciones Nacionales"*, doi: 10.4067/S0718-50062008000100002, Formación Universitaria, 1 (1), 3-12, (2008)

Más Estellés J. y otros cuatro autores, *"Rendimiento académico de los estudios de Informática en algunos centros españoles",* http://dialnet.unirioja.es/servlet/articulo?codigo=3417223, ISSN: 0211-2124, Novática: Revista de la Asociación de Técnicos de Informática. España, 204, 55-61, (2010)

Mercano Aular Y. J. y Talavera Pereira R., *"Minería de Datos como soporte a la toma de decisiones empresariales"*, http://www.scielo.org.ve/scielo.php?script=sci_arttext&pid=S1012-15872007000100008&lng=es&nrm=iso, ISSN: 1012-1587, Opción, 23 (52), 104-118, (2007)

Mitchell T.M., "Machine Learning", 1ª ed., McGraw Hill, 154-199, (1997)

Molina López J. M. y García Herrero J. *"Técnicas de Análisis de Datos – Aplicaciones prácticas utilizando Microsoft Excel y Weka"* (en línea), 12 mayo de 2014, http://www.giaa.inf.uc3m.es/docencia/II/ADatos/apuntesAD.pdf, (2006)

Pérez López C. y Santín González D., "*Minería de Datos - Técnicas y Herramientas*", 1ª ed., Thomson, Madrid, España,(2007)

Planck Barahona U., "*Factores determinantes del rendimiento académico de los estudiantes de la Universidad de Atacama"*, doi: 10.4067/S0718-07052014000100002, Estudios pedagógicos, 40 (1), 25-39, (2014)

Quadril M. N. y Kalyankar N. V., *"Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques"*, Global Journal of Computer Science and Technology, 10 (2), 2-5, (2010)

Ramaswani M. y Bhaskaran R., *"A Study on Feature Selection Techniques in Educational Data Mining"*, http://arxiv.org/ftp/arxiv/papers/0912/0912.3924.pdf, ISSN: 2151-9617, Journal of Computing, 1 (1), (2009)

Romero C. y Ventura S., *"Educational Data mining: A Review of the State of the Art"*, doi: 10.1109 / TSMCC.2010.2053532, IEEE Transactions on Systems, Man, and Cybernetics, 40 (6), 601-618, (2010)

Soria Barreto K. y Zúñiga Jara S., *"Aspectos Determinantes del Éxito Académico de Estudiantes",* doi: 10.4067/S0718-50062014000500006, Formación Universitaria, 7 (5), 41-50, (2014)

Superby J. y otros dos autores, "*Determination of factors influencing the achievement of the first-year university students using data mining methods*", Workshop on Educational Data Mining at the 8th International Conference on Intelligent Tutoring Systems, 37-44, Jhongli, Taiwan, (2006)

WEKA Waikato Envioronment for Knowledge Analysis (en línea), Nueva Zelanda, http://www.cs.waikato.ac.nz/ml/weka/, fecha consulta 1 de Abril de 2014

Witten I. H. y Eibe F., *"Data Mining: Practical Machine Learning Tools and Techniques*", 2ª ed., Morgan Kaufmann, San Francisco, EEUU, (2005)