Formación Universitaria
Vol. 10(3), 61-68 (2017)
doi: 10.4067/S0718-50062017000300007

Análisis de la Deserción de Estudiantes Universitarios usando Técnicas de Minería de Datos

Mauricio A. Miranda[(1)] y Jheser Guzmán[(2)]
(1) Universidad Católica del Norte, Programa de Magíster en Gestión de Información y Tecnologías Antofagasta, Chile.
(2) Facultad de Economía y Administración, Universidad Católica del Norte, Antofagasta, Chile.
(e-mail: mamiranda@ucn.cl; jguzman02@ucn.cl)

Recibido Sep. 12, 2016; Aceptado Nov. 3, 2016; Versión final Feb. 13, 2017, Publicado Jun. 2017

Resumen

Se realiza un estudio para determinar cuales son y cual es la importancia de las variables que llevan a un estudiante a abandonar sus estudios universitarios, usando técnicas de minería de datos. La deserción de los estudiantes de educación superior genera una serie de inconvenientes que afectan a los estudiantes y las universidades. Los resultados obtenidos a partir de los datos proporcionados por las carreras de Ingeniería de la Universidad Católica del Norte en Antofagasta y Coquimbo (Chile) determinan que las variables que mejor explican la deserción de un estudiante son, las razones socioeconómicas y el puntaje de ingreso a la universidad (PSU). Según el árbol de decisión construido se concluye que la retención se sitúa en un 78,3%. La calidad de los clasificadores permite asegurar que sus predicciones son correctas, con niveles estadísticos de curva ROC de 76%, 75% y 83% de acierto para los clasificadores de red bayesiana, árbol de decisión y red neuronal respectivamente.

Palabras clave: retención universitaria; deserción universitaria; minería de datos; educación; ingeniería

Analysis of Dropouts of University Students using Data Mining Techniques

Abstract

The research discussed in this paper determines the reasons and variables that determine student's decision to abandon their university studies. Student dropout becomes a major problem for educational institutions, as the loss of students can disrupt short and long-term academic and financial strategies. To evaluate these factors, data provided by the School of Engineering of the Catholic University of the North (UCN) in Antofagasta and Coquimbo (Chile) were used. The results are obtained using a decision tree to predict the retention of students within 78,3% of accuracy. The models built in this analysis show a statistical ROC Curve of 76%, 75%, and 83% success rate for the Bayesian network classifier, decision tree, and neural network respectively.

Keywords: university student retention; student dropout; data mining; education; engineering

INTRODUCCIÓN

La deserción, definida como el abandono de un programa de estudios antes de obtener el título o grado correspondiente, considerando un tiempo lo suficientemente largo como para descartar la posibilidad de reincorporación (Delen, 2010). En Chile, el Sistema de Información de la Educación Superior, SIES (2014), ha cifrado la deserción en un 25,40% para el año 2012, en tanto que para los periodos entre 2008 y 2012 la deserción promedio del sistema universitario es de un 24,7%. En relación a la retención según los niveles de puntaje en la PSU, prueba de selección que se debe rendir de manera obligatoria para postular al sistema universitario chileno tradiciona, estudiantes que tienen puntajes superiores a 800 puntos tienen una retención promedio, en 2012, de 96,5%. Por otro lado quienes lograron puntajes inferiores a 450 puntos tienen una tasa de retención promedio de 56,8% en el año 2012 (SIES, 2014). Esto muestra que existe una alta correlación entre un alto puntaje de ingreso a la Universidad y un alto nivel de retención.

De acuerdo a lo indicado por SIES (2014), si bien hay un número considerable de estudiantes que desertan en primer año, es importante advertir que un número importante de ellos no lo hace definitivamente, sino que reingresa en los años siguientes a otras carreras o instituciones o bien reingresan a la misma carrera. Por otro lado, se constata que, al analizar la cohorte 2008; el 13,4% de los alumnos que desertaron reingresan al sistema en los 3 años siguientes, y sólo el 17,2% de los jóvenes pueden considerarse desertores definitivos. Lo anterior implica que los estudiantes buscan mantenerse en el sistema a pesar de poder sufrir de problemas económicos, vocacionales o de conocimientos, especialmente en sus primeros años de estudios. Los trabajos mencionados utilizan diferentes variables o dimensiones para determinar cuáles son las causas de la deserción, estas variables pueden ser de tipo académicas, psicosociales o vocacionales.

Los estudios más relevantes en esta disciplina son los de Tinto (1975) y Bean (1980). Tinto (1975) formula un modelo teórico que explica los procesos de interacción entre el individuo y la institución, las cuales, determinan las razones para abandonar la universidad y distingue aquellos procesos que dan lugar a diferentes formas de comportamiento de deserción. Las dimensiones más importantes que considera este modelo son: atributos previos al ingreso, metas y compromisos, experiencias institucionales, integración personal y normativa. Bean (1980) aplica un modelo de rotación laboral al problema de la deserción estudiantil. El autor señala que un estudiante deja la Universidad no necesariamente por tener rendimiento malo, puede tener pros y contras, sino por otros factores externos a los académicos. Las variables utilizadas son: factores académicos, factores psicosociales, factores ambientales y factores de socialización. Se determina cuáles son las variables más importantes que inciden en que un estudiante universitario decida abandonar sus estudios superiores antes de su graduación.

Estudios recopilados determinó que las variables que mejor explican este fenómeno son el puntaje de la prueba de ingreso a la Universidad (Prueba de Selección Universitaria, PSU para el caso chileno) y sus calificaciones de enseñanza secundaria. (Sánchez et al. (2005); SIES (2014); Universidad de Chile (2008); Timaran et al. (2013); Rolando et al. (2010)). En tanto que en Chacon et al. (2012), se señala que los determinantes de la retención universitaria son: reclutamiento y admisión, servicios académicos, currículo e instrucción, servicios estudiantiles y ayudas financieras. Estos estudios utilizan técnicas de estadística clásica para determinar sus resultados, sin indagar más en posibles patrones ocultos en los datos, aportando con una perspectiva diferente al problema de la deserción.

En este trabajo se determina la importancia de las variables que influyen a que un estudiante decida abandonar sus estudios universitarios mediante técnicas de minería de datos, comparándolas con las variables determinadas en la revisión de la literatura. Mostramos correlación entre los resultados obtenidos en este trabajo y la literatura, situando la retención en un 77,93%, frente al 75,3% que el SIES (2014) presenta para el periodo 2008-2012 en Chile. En Eckert y Suénaga (2015), se puede encontrar un estudio con una aplicación metodológica similar para un caso en Argentina, en él se encuentran porcentajes de retención de 81,3% y 81% para los algoritmos J48 y BayesNet respectivamente.

METODOLOGÍA

El presente trabajo es de tipo descriptivo, el cual comprobará las siguientes hipótesis mediante el análisis de datos cuantitativos: (i) Las condiciones académicas de entrada (notas de enseñanza media, conocimientos, puntaje PSU) del estudiante determinan su deserción en la carrera de Ingeniería de Universidad Católica del Norte; y (ii) Las condiciones socioeconómicas del estudiante determinan su decisión de abandonar sus estudios universitarios en la carrera de Ingeniería de la Universidad Católica del Norte.

En esta investigación se considera deserción a un periodo suficientemente largo como para que un estudiante decida retomar sus estudios superiores, equivalente a 3 años, tomando como referencia el

tiempo que el SIES del Ministerio de Educación de Chile utiliza como estándar en sus investigaciones. Este estudio se aplica a los estudiantes de Ingeniería de la Universidad Católica del Norte de Chile (UCN), para los años 2000 a 2013 con una muestra total de 9.195 individuos, a través de técnicas de minería de datos. La UCN está ubicada en las ciudades de Antofagasta y Coquimbo, en Chile.

Se construyen 3 clasificadores que permiten categorizar a los estudiantes entre las clases abandono y no abandono. Para ello utilizamos 3 algoritmos: redes bayesianas, redes neuronales y árbol de decisión. De ellos se mide su calidad como clasificadores y las variables más importantes en cada algoritmo, que permiten determinar si un estudiante desertará o no. El proceso KDD (Proceso de Descubrimiento de Conocimiento en Bases de Datos) permite la selección, limpieza, transformación y proyección de los datos; analizar los datos para extraer patrones y modelos adecuados, evaluar e interpretar los patrones para convertirlos en conocimiento; consolidar el conocimiento resolviendo posibles conflictos con conocimiento previamente extraído y hacer el conocimiento disponible para su uso. (Hernández et al., 2005). El proceso KDD se compone de las siguientes etapas: (a) Fuente de información, (b) Preparación de los datos, (c) Minería de datos, (d) Interpretación y evaluación, y (e) Conocimiento; los cuales son explicados a mayor detalle en lo que sigue.

(a) Fuente de Información. La información utilizada para la presente investigación fue proporcionada por la Universidad Católica del Norte, y se analizaron 12 carreras de Ingeniería de las Sedes Antofagasta y Coquimbo, para los años 2000 a 2013. El detalle de esta información se presenta en la tabla 1.

Tabla 1: Variables utilizadas en esta investigación (Elaboración propia)

Variable	*Método de normalización*
Estado académico (variable clase)	Variables que no requieren normalización
Año de admisión a la carrera	
Año de egreso de enseñanza media del estudiante	
Código carrera a la que pertenece el estudiante	
Orden en la lista de selección	Método de máximos y mínimos
Promedio ponderado PSU (Prueba de Selección Universitaria)	Método de media de la serie
Preferencia de postulación del estudiante	
Nota promedio de los 4 años de Enseñanza Media (NEM)	
Puntaje PSU para la NEM (Puntaje estandarizado equivalente a la nota promedio de enseñanza media del estudiante)	
Beneficios estudiantiles	
Calificaciones por semestre para cada asignatura cursada por el estudiante	

(b) Preparación de los datos. Una vez que los mencionados datos son limpiados y depurados, se procede a la construcción de un modelo Entidad – Relación y su esquema relacional, el cual posteriormente permitió la construcción de la base de datos en SQL Server. Todo el procedimiento explicado se basa en Elmasri et al. (2002). Posteriormente se construye una única vista que incluye los datos de las tablas que componen la base de datos y dicha vista se procesa para su posterior uso en los análisis. Luego se agrupan las vistas incluyendo los datos correspondientes a cada variable para períodos consecutivos de 4 semestres, haciendo un seguimiento de indicadores de forma secuencial por estudiante entre el primero de 2000 hasta el segundo de 2013. La muestra corresponde a 73.958 individuos. Esta muestra se alcanza al realizar un seguimiento uno a uno de cada estudiante por semestre, entre el total de semestres que el estudiante permaneció en la universidad y de esta manera se llega de los 9.195 a los 73.958 registros mencionados en este apartado. Dado que las variables no se encuentran estandarizadas en su tabulación se establece un procedimiento de normalización. Mediante este proceso las variables tomaron valores estándar entre 0 y 1, los cuales se establecieron a través de dos tipos de procedimiento: (1) Método de la media de la serie. Se toma la media de la serie y se resta a cada variable dividiéndolo entre su desviación estándar, a través de la siguiente formula: $x = (x-\mu)/\sigma;$ y (2) Método de máximos y mínimos. Se toma el máximo y mínimo de la serie en análisis, a través de la siguiente formula: $x' = (x-min)/(máx-min).$

Posteriormente y debido a que la cantidad de individuos para ambas clases analizadas (abandono y retención) no están equilibradas, se aplica de manera automática en el software Weka el algoritmo SMOTE, que de acuerdo a lo señalado por Witten, Frank y Hall (2011), este algoritmo se usa para muestrear los datos alterando la distribución de la clase, equiparando la cantidad de ejemplos o casos de cada clase. De esta manera la nueva base de datos posee 89.056 individuos con los cuales se construyen los clasificadores.

(c) Minería de datos. Para cumplir con el objetivo de esta investigación se utilizan las técnicas de minería de datos conocidas como clasificadores. De acuerdo a lo señalado por Pang-Ning et al. (2006), la clasificación consiste en la tarea de asignación de objetos a una de varias categorias predefinidas, en relación a ello, los autores señalan que es un tipo de problema generalizado que abarca aplicaciones diversas.

De acuerdo a lo expuesto por Hernández et al. (2005), el objetivo de la clasificación es aprender una función $\lambda = E \rightarrow S$, denominada clasificador, que represente la correspondencia existente, es decir, para cada valor de E tenemos un único valor para S. Además S es nominal, o sea, puede tomar un conjunto de valores C_1, C_2,…,C_n, denominados clases. La función aprendida será capaz de determinar la clase para cada nuevo ejemplo sin etiquetar, es decir, dará un valor de S para cada valor de E. Por ejemplo, para el caso particular que analiza esta investigación, se busca poder determinar si un estudiante abandonará su carrera universitaria antes de graduarse, entonces, para cada nuevo estudiante sin etiquetar, el clasificador ya entrenado deberá ser capaz de determinar si este nuevo estudiante se graduará o no a partir de las variables que determinan esta decisión. Se desarrolló la metodología explicada, pues lo que se buscaba era identificar las variables que afectan al problema en ambos casos (los que permanecen y los que abandonan), para ello se ha construido un clasificador que indica que variable es la que influye más en la clasificación.

En la presente investigación se construyen 3 tipos de clasificadores: (i) clasificador de redes bayesianas; (ii) clasificación basada en árboles de decisión; y (iii) clasificación basada en redes neuronales.

Clasificador de redes bayesianas. Este clasificador se basa en la teoría de probabilidad bayesiana, propuesta por el matemático Thomas Bayes, cuyos principios fundamentales se encuentran en el texto "*An Essay towards Solving a Problem in the Doctrine of Chances*", de Bayes y Price (1763). De acuerdo a Hernández, et al.(2005), las redes bayesianas representan el conocimiento cualitativo del modelo mediante un grafo dirigido acíclico. Los autores señalan que este conocimiento se articula en la definición de relaciones de independencia/dependencia entre las variables que componen el modelo. Estas relaciones abarcan desde una independencia completa hasta una dependencia funcional entre variables del modelo.

Clasificación basada en árboles de decisión. De acuerdo a lo expuesto por Hernández, et al. (2005), la tarea de aprendizaje para la cual los árboles de decisión se adecuan mejor es la clasificación. Los autores indican que de hecho, clasificar es determinar de entre varias clases a qué clase pertenece un objeto; la estructura de condición y ramificación de un árbol de decisión es idónea para este problema. Una de las ventajas de este tipo de modelos, en relación, por ejemplo, a las redes neuronales o los vectores de soporte, es que el resultado que se obtiene es inteligible para los seres humanos (y también para sistemas semi-automáticos que procesen reglas). (Hernández et al., 2005)

Clasificación basada en redes neuronales. Las redes neuronales nacen a partir de los intentos de los investigadores por establecer un sistema que logrará representar las características de funcionamiento del sistema nervioso de las personas. Para el desarrollo de la red neuronal utiliza un modelo de optimización no lineal.

El análisis de redes neuronales suele ser considerado una "caja negra", dado que a diferencia de otros análisis es difícil poder interpretar con claridad a diferencia de los resultados y parámetros entregados por otros análisis. Para la construcción del clasificador usado en esta investigación se utilizó la red neuronal del perceptrón multicapa, el cual, de acuerdo a lo señalado por Hernández et al. (2005), construye una red neuronal en forma de cascada, que tiene una o más capas ocultas.

(d) Interpretación y evaluación. Se analizaron los resultados arrojados por el árbol de decisión y los parámetros de los clasificadores Naive Bayes y de red neuronal.

Los parámetros de evaluación que se utilizan para cada clasificador fueron los siguientes (i) Precisión: Determina la fracción de registros que en realidad resulta ser positiva y que ha sido efectivamente clasificada como positiva. Cuanto mayor sea la precisión, menor será el número de casos erróneamente clasificados como falsos negativos. (Pang-Ning et al., 2006); (ii) Alcance o sensibilidad: Determina la fracción de casos positivos correctamente clasificados. (Pang-Ning et al., 2006); (iii) Ratio falso positivo: Es el porcentaje de casos negativos erróneamente clasificados como positivos. (Pang-Ning et al., 2006); (iv) Ratio falso negativo: Es la proporción de casos positivos erróneamente clasificados como negativos. (Pang-Ning et al., 2006); (v) Curva ROC: Es una representación gráfica de la relación entre los ratios verdaderos positivos y falsos positivos del clasificador. De acuerdo a lo señalado por (Pang-Ning, Steinbach y Kumar, 2006), el área bajo la curva ROC (AUC) proporciona un enfoque para evaluar qué modelo es mejor en promedio. Si el modelo es perfecto, entonces su AUC sería igual a 1. Si el modelo se realiza al azar su AUC sería igual a 0,5. (Pang-Ning et al., 2006); y (vi) F-Measure: Es una métrica que unifica la precisión y el recall. (Pang-Ning et al., 2006)

(e) Conocimiento. Una vez que se obtienen los resultados de esta investigación se espera que éstos puedan ser utilizados por la Universidad analizada y por otras de similares características para el diseño de políticas de retención estudiantil.

Para el desarrollo de la presente investigación se utilizan las siguientes herramientas de software: SQL Server, para el almacenamiento de la base de datos, SPSS Statistics, para el clasificador de redes neuronales y el árbol de decisión, y Weka para el clasificador de redes bayesianas. Para la repetibilidad del experimento es necesario contar con al menos las mismas variables utilizadas para este análisis, utilizando para la vista final: carrera, puntajes PSU por prueba y ponderado, NEM (Nota promedio del estudiante de sus 4 años de enseñanza media), puntaje NEM (corresponde al puntaje estandarizado equivalente a la nota promedio de enseñanza media del estudiante), beneficios asignados por año, promedio semestral de calificaciones y estado académico.

RESULTADOS

En este apartado se presentan los principales resultados de la construcción de los clasificadores: (1) construcción del clasificador de redes bayesianas; (2) construcción del clasificador de árbol de decisión; y (3) construcción del clasificador mediante redes neuronales.

Construcción del clasificador de redes bayesianas

Se desarrolla el clasificador mediante el algoritmo BayesNet en el software Weka (Durrant et al., 2011). Se obtiene como resultado un grafo mediante el cual es posible caracterizar a los estudiantes desertores. Entre los resultados obtenidos a través de la red bayesianas se puede mencionar que el abandono encontrado por este clasificador es de un 33,9% y la retención de un 66,1%, estos porcentajes corresponden al resultado de la clasificación para la variable clase (abandono y no abandono). En tanto que un 94% de aquellos estudiantes que abandonan sus estudios poseen un puntaje dentro del promedio en la PSU, en tanto que quienes no abandonan se encuentran en esta media en un 87%.

Otro resultado encontrado, entre las variables más importantes para explicar la deserción, es el orden de postulación a las carreras. Esto consiste en que cuando un estudiante postula a las universidades chilenas tiene 10 opciones de postulación en orden de descarte, es decir si es seleccionado en su primera opción (carrera, universidad), las siguientes se descartan automáticamente. Esta preferencia no incide de manera importante en la decisión de desertar pues sólo un 9% se encuentra en las preferencias promedio, de acuerdo a lo detectado por el algoritmo. Esto puede estar mostrando que estudiantes que no necesariamente colocaron las Ingenierías de la UCN en los primeros lugares abandonan sus estudios.

Aquellos estudiantes desertores tienen una probabilidad condicionada de 89% de tener una nota de enseñanza media promedio dentro de la muestra analizada para este estudio. Respecto al puntaje de la PSU de matemáticas. Los estudiantes que desertan tienen en un 98% un puntaje cercano o en la media de la muestra en dicha PSU. En cambio en el caso de la prueba de historia un 77% de quienes abandonan pertenecen a la cota superior, esto puede obedecer a un problema vocacional o de conocimientos. Quienes se encuentran en la cota inferior de puntaje para la prueba de ciencias tienen un 69% de probabilidad de desertar.

En cuanto a los beneficios socioeconómicos, estos fueron analizados en 4 niveles y se puede afirmar que tanto quienes desertan como quienes permanecen se encuentran en el nivel promedio de beneficios. Con respecto a la calidad del clasificador desarrollado, se obtiene una baja tasa de casos positivos mal clasificados, 24% y un AUC para su curva ROC cercano al umbral considerado aceptable con un 76%. En tanto que el recall indica que un 76% de los casos positivos, efectivamente fueron correctamente clasificados. Un 76% de los casos fueron correctamente clasificados en la etapa de prueba, en la cual un 24% de los casos fueron erróneamente etiquetados (tabla 3).

Construcción del clasificador de árbol de decisión

Se desarrolla un clasificador mediante el algoritmo de árbol de decisión. De entre los resultados obtenidos se puede mencionar los siguientes: para la variable clase (abandono y retención), un 21,7% se clasifica como abandonado y un 78,3% como retenido. En tanto que aquellos estudiantes que hoy tienen beneficios estudiantiles (créditos, becas), en el promedio o menos que este valor tienen un 89,3% de probabilidad de permanecer en su carrera universitaria y un 10,7% de abandonar. Por otro lado, quienes tienen mayores o menores beneficios a la media (extremos) desertan con una probabilidad de un 28,1% de probabilidad de abandonar sus estudios frente a un 71,9% que no lo haría. Este grupo de estudiantes corresponde al 63,1% de los estudiantes analizados y demuestra que quienes poseen un beneficio económico tienen una mayor posibilidad de permanecer y terminar sus carreras.

Quienes poseen un promedio en su prueba de selección PSU, justo en la media de la muestra, tienen un 76,7% de probabilidad de permanecer en sus carreras y 23,3% de abandonar. En tanto que aquellos puntajes que se encuentran en los extremos de la muestra (máximos y mínimos), tienen un 25,7% de probabilidad de permanecer en sus carreras y un 74,3% de abandonar, el cual se podría suponer que está en el extremo inferior de esta categoría.

Otra variable, que está al mismo nivel del promedio de la PSU como buen predictor en el árbol de decisión, es el puntaje de la prueba de historia. Quienes se encuentran en el extremo superior de la muestra tienen un 91,1% de probabilidad de permanecer en la carrera que estudian. En tanto que quienes se encuentran bajo la media del promedio en la prueba de historia tienen un 88,8% de permanecer en su carrera y un 12% de probabilidad de abandonarla. En tanto que en el tercer nivel de predictores de abandono del árbol de decisión, aparecen aquellos estudiantes que tienen beneficios desde hace 3 periodos hacía atrás, donde destacan quienes se encuentran en los extremos de esta muestra, es decir aquellos estudiantes que en este periodo tienen beneficios muy altos o de montos muy pequeños, tienen un 80% de probabilidad de abandonar sus estudios superiores.

El clasificador además arrojó el conjunto de variables normalizadas más importantes para el problema analizado. Estas variables son, en orden de importancia, las siguientes: beneficios económicos en el nivel 0, beneficios desde hace 3 periodos, beneficios desde hace 1 periodo, puntaje en la PSU de Lenguaje, promedio de puntajes en la PSU y puntaje en la PSU de Historia.

Para desarrollar el entrenamiento del modelo se utiliza un 70% de la muestra, de los cuales fue clasificado correctamente un 79% de la muestra, esto implica que este porcentaje de casos será predicho de manera correcta por el modelo. Esto implica un bajo nivel de riesgo (19,8%). El clasificador tiene los siguientes parámetros de evaluación. El clasificador desarrollado clasifica correctamente un 81,3% de los casos en el entrenamiento y un 82,2% en el contraste. En promedio se puede afirmar que el modelo es bueno, con un nivel de AUC de 75% para su curva ROC (tabla Nº3).

Construcción del clasificador mediante redes neuronales.

Se desarrolla una red neuronal mediante el algoritmo Perceptrón Multicapa con un error admisible de 0,0001. En cuanto a la arquitectura del modelo, se utilizó una metodología personalizada para la activación de las capas de entrada y salida, a través de la función de tangente hiperbólica. La red neuronal es una caja negra a nivel de interpretación de predicciones, sin embargo, se analizan los resultados que el clasificador entrega respecto a la importancia normalizada de las variables analizadas como clasificadores efectivos del abandono y/o retención estudiantil. En cuanto a la importancia normalizada de las variables entregadas por este análisis mencionamos las más importantes: promedio prueba PSU, beneficio últimos 3 periodos, beneficio nivel 0, promedio ponderado de postulación, puntaje de la nota de enseñanza media del estudiante y beneficios económicos en los últimos 2 periodos. Si bien este resultado no coincide con el obtenido en el clasificador Naive Bayes, se debe mencionar que ambos utilizan diferentes técnicas de análisis y deben ser evaluados en ese contexto. (Tabla 2). Es destacable que el quinto lugar lo ocupa el puntaje EM, el cual representa la nota obtenida por los estudiantes en sus 4 años de enseñanza secundaria, transformado en puntaje estandarizado de la prueba PSU, este puntaje sirve al estudiante para postular a la Universidad. Para evaluar la calidad del clasificador construido, en la tabla 3 se presentan los parámetros de evaluación de la red neuronal.

De acuerdo a la curva ROC, se puede decir que la clasificación fue realizada correctamente en un 83% de los casos y con una alta exactitud para los casos positivos clasificados con un 73% de precisión y un 88% para el ratio de casos negativos clasificados de forma exacta. Tanto en la prueba como el entrenamiento de este clasificador, se etiquetaron correctamente un 80% de los individuos, dado que en ambos test se obtiene un nivel similar de clasificación correcta. Esto implica un buen nivel de exactitud en los resultados arrojados (ver tabla 3). A nivel general se puede señalar que, en cuanto a la calidad de los clasificadores construidos, existe una baja probabilidad de haber obtenido valores clasificados incorrectamente de acuerdo a la curva ROC de los clasificadores de red neuronal, árbol de decisión y red bayesiana (83%, 74% y 76% respectivamente). (Tabla 3)

El análisis efectuado arroja una tasa de deserción de un 33,9% (red bayesiana) y 21,7% (árbol de decisión) para las carreras de Ingeniería analizadas. Al comparar estas probabilidades con los obtenidos por el SIES (2014) del Ministerio de Educación de Chile se puede decir que las carreras de Ingeniería de la UCN se encuentran en el promedio de lo obtenido por el estudio del SIES, estudio que cifra la deserción en un 25,4% a nivel nacional para el año 2012. En tanto que en el área tecnológica (donde se clasifica a las Ingenierías) la tasa fue de un 27,9% y en las regiones en las que se encuentra la UCN, Antofagasta (21,43%) y Coquimbo (32,14%). (SIES, 2014).

Tabla 2: Importancia normalizada variables red neuronal. (Elaboración propia en base a resultados obtenidos desde el software SPSS)

Importancia de las variables independientes		
Variables	Importancia	*Importancia normalizada*
Orden_lista_1	,021	16,8%
Promedio_Ponderado_1	,088	68,9%
Preferencia_1	,053	41,5%
NEM_1	,078	61,0%
PuntajeEM_1	,084	65,9%
Matemáticas_1	,031	24,2%
Lenguaje_1	,058	45,7%
Historia_1	,063	49,4%
Ciencias_1	,064	50,4%
Promedio_prueba_1	,128	100,0%
Beneficio_n3_1	,110	85,8%
Beneficio_n2_1	,080	62,5%
Beneficio_n1_1	,051	40,2%
Beneficio_n0_1	,089	70,0%

Tabla 3: Comparación de los parámetros de evaluación de los clasificadores

	Precisión	*Recall*	*TPR*	*TNR*	*FPR*	*FNR*	*ROC Curve*	*F-Measure*	*clasificados correcta*	*clasificados incorrecto*
Red neuronal	73%	65%	65%	88%	12%	35%	83%	69%	80%	20%
Árbol de decisión	72%	64%	64%	87%	12%	36%	74%	68%	82%	18%
Red Bayesiana	76%	76%	76%	70%	30%	24%	76%	76%	76%	24%

No es posible afirmar que un clasificador es mejor que otro, inclusive se puede señalar que los 3 son útiles para analizar el problema teniendo en consideración sus ventajas y desventajas. El experimento es replicable bajo condiciones similares y replicando el uso y tipo de datos utilizados en esta investigación.

CONCLUSIONES

Se construyeron tres clasificadores (redes bayesianas, árbol de decisión y redes neuronales). Cada uno de estos resultados tiene sus particularidades y ventajas por lo cual no son totalmente comparables. Sin embargo, existe coincidencia respecto a que las becas y créditos que los estudiantes tienen como beneficio son determinantes a la hora determinar la deserción. Por el lado académico la variable que mejor explicaría la deserción es el puntaje de la PSU (Tabla Nº4).

Los resultados hallados en este análisis permiten demostrar que las dos hipótesis planteadas se comprueban positivamente, puesto que tanto los resultados académicos como la situación socioeconómica influyen en la decisión de permanencia de un estudiante en su respectiva carrera. Dada la alta correlación que los resultados obtenidos tienen con la bibliografía consultada y la buena calidad de los modelos, es posible afirmar que gestionando estas variables es posible reducir las tasas de deserción del sistema universitario.

Tabla 4: Variables que mejor clasifican a los estudiantes desertores según cada clasificador.

Árbol de decisión	*Redes Bayesianas*
Beneficio Nivel 0	Promedio prueba PSU
Beneficio Nivel 3	Beneficio nivel 3
Beneficio Nivel 1	Beneficio nivel 1

El desarrollo de esta investigación tuvo como limitación el acceso a la información psicosocial del estudiante, la cual aparece como una variable importante en el marco referencial, esta podría aportar a estudiar la deserción desde la perspectiva vocacional, sumado a las dimensiones económicas y académicas estudiadas en este artículo. Esta podría ser una interesante corriente a seguir por una futura investigación.

AGRADECIMIENTOS

Los autores agradecen a la Universidad Católica del Norte por permitir el uso de sus datos académicos.

REFERENCIAS

Bayes, M. y M. Price, An essay towards solving a problem in the doctrine of chances. By the late rev. Mr. Bayes, frs communicated by Mr. price, in a letter to john canton, amfrs. pp. 370–418, Philosophical Transactions (1763)

Bean, J. P., Dropouts and turnover: The synthesis and test of a causal model of student attrition, Research in higher education, 12 (2), 155–187 (1980)

Bean, J., Interaction effects based on class level in an explanatory model of college student dropout syndrome, American educational Research journal, 22 (1), 35–64 (1985)

Chacon, F., Spicer, D. y A. Valbuena, Analytics in support of student retention and success, Research Bulletin 3. Louisville, CO: Educause, Center for Applied Research (2012)

Delen, D., A comparative analysis of machine learning techniques for student retention management, Decision Support Systems, 49 (4), 498–506 (2010)

Durrant, B., Frank, E., Hunt, L., Holmes, G., Mayo, M., Pfahringer, B., Smith, T. y Witten, I., Weka (Versión 3.6.6) [Software] Recuperado de http://www.cs.waikato.ac.nz/ml/weka/index.html (2011)

Elmasri, R., S.B. Navathe, V.C. Castillo, B.G. Espiga y G.Z. Pérez, Fundamentos de sistemas de bases de datos. Addison-Wesley (2002)

Eckert, K. B. y R. Suénaga, Análisis de deserción-permanencia de estudiantes universitarios utilizando técnica de clasificación en minería de datos, Formación Universitaria, 8 (5) (2015)

Hernández, J., M. Ramírez y C. Ferri, Introducción a la minería de datos (2005)

Pang-Ning, T., M. Steinbach, V. Kumar, Introduction to data mining, in Library of Congress, p. 74 (2006)

Rolando, R., J. Salamanca y A. Lara, Retención de primer año en el pregrado: Descripción y análisis de la cohorte de ingreso 2007, Ministerio de Educación de Chile (2010)

Sánchez, A., M.A. Gutiérrez, C. Meneses, Using data mining to support the university decision process: A case in a chilean university, AMCIS 2005 Proceedings, p. 350 (2005)

SIES, Panorama de la educación superior en Chile en 2014, en línea: https://goo.gl/A55ckP, acceso: 2 Septiembre 2016, (2014)

Timaran, R., A. Calderón, y J. Jiménez, Application of data mining in the detection of student dropout patterns, Revista Vinculos,10 (1) (2013)

Tinto, V., Dropout from higher education: A theoretical synthesis of recent research, Review of educational research, pp. 89–125 (1975)

Universidad de Chile, Estudio sobre causas de la deserción universitaria, Centro de Microdatos, (en línea: https://goo.gl/TNnrrj, acceso: 2 Septiembre 2016) (2008)