Ciencia amazónica (Iquitos) 6 (1), 73-84
http://dx.doi.org/10.22386/ca.v6i1.110

Artículo Original

Estudio de variables que influyen en la deserción de estudiantes universitarios de primer año, mediante minería de datos

[Study of variables that influence defection in freshman university student, through data mining]

Christian Zarria Torres*, Christian Arce Ramos, Jaime Lam Moraga

Facultad de Ingeniería y Arquitectura, Campus Playa Brava, Universidad Arturo Prat, Av. Arturo Prat 2120, Iquique, Chile.
*e-mail de contacto: czarria@unap.cl

Resumen:

El objetivo de esta investigación fue analizar la deserción de los estudiantes mediante técnicas de minería de datos y obtener un modelo que fuese capaz de clasificar estudiantes desertores a partir de los datos socioeconómicos y académicos de los estudiantes de carreras de pregrado en la Universidad Arturo Prat, Chile. Para el desarrollo de este proyecto se utilizó CRISP-DM como metodología para guiar las etapas del proyecto y se analizaron tres diferentes modelos de clasificación: árboles de decisión, métodos bayesianos y redes neuronales, con el fin de evaluar su comportamiento, encontrándose que Random Forest es el algoritmo de mejor desempeño general, con un 88,9% de exactitud, mientras que el algoritmo Naive Bayes resulto ser el más adecuado para dar respuesta a los objetivos del negocio, dados los niveles de sensibilidad alcanzados. Mediante los experimentos realizados se determinó que las variables académicas de ingreso de los estudiantes no resultan significativas para explicar la deserción de primer año. Con estos resultados, la Universidad podrá generar mejoras en los procesos críticos y en las variables que pudiesen intervenirse, haciendo más eficiente su gestión y mejorando el bienestar del estudiante, y por ende, de la comunidad en la cual se encuentra inmersa.

Palabras clave: Deserción; minería de datos; variables académicas

Abstract:

The aim of this research was to analyze the defection of students through techniques of data mining and obtain a model that was able to classify defectors students through socioeconomic and academic data of undergraduate students at the Arturo Prat University, Chile. In the development of this project CRISP-DM was used as a methodology to guide the project phases and three different classification models were analyzed: decision trees, Bayesian methods and neural networks, to assess their behavior, finding out that Random Forest is the algorithm with the best overall performance, with a 88.9% accuracy, while the Naive Bayes algorithm proved to be the most appropriate to meet business objectives, given the sensitivity levels achieved. Through the experiments performed it was determined that the student's academic entry variables are not significant to explain the defection of first year. With these results, the University may generate improvements in critical processes and in variables that could be intervene, making management more efficient and improving the welfare of the student, and therefore, the community in which it is immersed.

Keywords: Defection; data mining; academic variables

***Recibido**: 12 de abril de 2016*
***Aceptado para publicación**: 30 de mayo de 2016*

INTRODUCCIÓN

En la actualidad, la tasa de deserción de estudiantes en educación superior es uno de los indicadores más utilizados a nivel internacional para evaluar la eficiencia interna de los procesos de enseñanza aprendizaje de las instituciones terciarias. En especial la deserción de primer año, considerando que la mayor fuga de estudiantes se da en ese período. Respecto de las características de la deserción, el trabajo de Himmel (2002) la define como el "abandono prematuro de un programa de estudios antes de alcanzar el título o grado, y considera un tiempo suficientemente largo como para descartar la posibilidad de que el estudiante se reincorpore", siendo esta la definición que se utiliza en este estudio.

La recopilación de información referente a las causales de la Deserción permite identificar una gran diversidad de resultados, demostrándose la influencia de aspectos tales como: sociales y económicos, cómo la como la ocupación y nivel educacional de los padres, la valoración y expectativas educativas de los jóvenes y el compromiso con la meta y los objetivos futuros según lo destacado por Tinto (1993); situación financiera del estudiante y su familia, el trabajo remunerado del estudiante y las responsabilidades familiares, factores motivacionales, que provocan déficit cognitivos según Prieto (2002), valores del arancel, acceso a préstamos y becas, ubicación geográfica, clase y raza de los estudiantes según relata St. John et al.(2004); Himmel (2002). Junto con lo anterior, Medrano et al. (2010) indica que se puede evidenciar la existencia de otras investigaciones que atribuyen una gran relevancia a la explicación de la deserción a factores psicológicos, identificando asociación entre los conceptos de deserción y las creencias irracionales, dado que esto puede interferir con las metas que los estudiantes han establecido y afectar la dirección de los esfuerzos invertidos, y que estudiantes que utilizan estrategias de aprendizaje más complejas, presentan un mejor rendimiento académico y por tanto, menores niveles de deserción, Fernández et al. (2009). A pesar de lo anterior, parece no existir consenso si las causas sociales o académicas son las más determinantes al momento de explicar la deserción, Himmel (2002).

Cabe señalar que Braxton (2000) y Díaz (2008) indican que existen varios enfoques metodológicos adicionales, los cuales permiten el estudio de la deserción desde distintas perspectivas, los cuales se pueden agrupar en cinco tipos, que son:

- Psicológicos: a través de la Teoría de la Acción Razonada, se analiza el comportamiento como actitudes en respuesta a objetos específicos, considerando normas subjetivas que guían el comportamiento hacia esos objetivos, Fishbein y ajsen (1975).
- Económicos: se distinguen dos modelos; 1) Costo/Beneficio: consiste en que los beneficios sociales y económicos asociados a los estudiantes son percibidos como mayores que los derivados por actividades alternas, y 2) Focalización de Subsidio: consiste en la entrega de subsidios dirigidos a los grupos que presentan limitaciones reales para costear sus estudios y que pueden aumentar el riesgo de abandono, Cabrera, Nora y Asker (1999).
- Sociológicos: considera que elementos como el medio familiar, la congruencia normativa, el desarrollo intelectual y el apoyo de pares, pueden condicionar la integración de los estudiantes en el entorno de la educación superior y por tanto, su riesgo de abandono, Spady (1970).
- Organizacionales: enfocan la deserción desde las características de la institución, en cuanto a los servicios que ésta ofrece a sus estudiantes, Braxton *et al*. (2000). Este modelo sostiene que la deserción depende de las condiciones que otorga la organización hacia la integración social de los estudiantes Berger *y* Milem (2000).
- De Interacciones: Los estudiantes actúan de acuerdo a la teoría del intercambio en la construcción de su integración social y académica, Tinto (1975). Explica el proceso de permanencia en la educación superior como una función del grado de ajuste entre el estudiante y la institución, adquirido a partir de las experiencias académicas y sociales (integración).

Según el modelo de interacciones (Figura 1), los estudiantes actúan de acuerdo a la teoría del intercambio en la construcción de su integración social y académica, explicando el

proceso de permanencia en la educación superior como una función del grado de ajuste entre el estudiante y la institución, adquirido a partir de las experiencias académicas y sociales (integración).

Según datos del SIES (2014) (Servicio de Información de Educación Superior) para la cohorte 2013, la deserción de primer año para Universidades del Consejo de Rectores a nivel nacional alcanza el 30%, mientras que para las universidad de la Región de Tarapacá llega a casi un 40%.

Según cifras de la Red de Unidades de Análisis Institucional (REDUAI) del Consorcio de Universidades del Estado de Chile (CUECH), la Universidad Arturo Prat (UNAP), universidad estatal, ubicada en Iquique, posee una tasa de deserción cercana al 30%, situación que se viene arrastrando desde el año 2008 a la fecha como se puede apreciar en la Tabla 1.

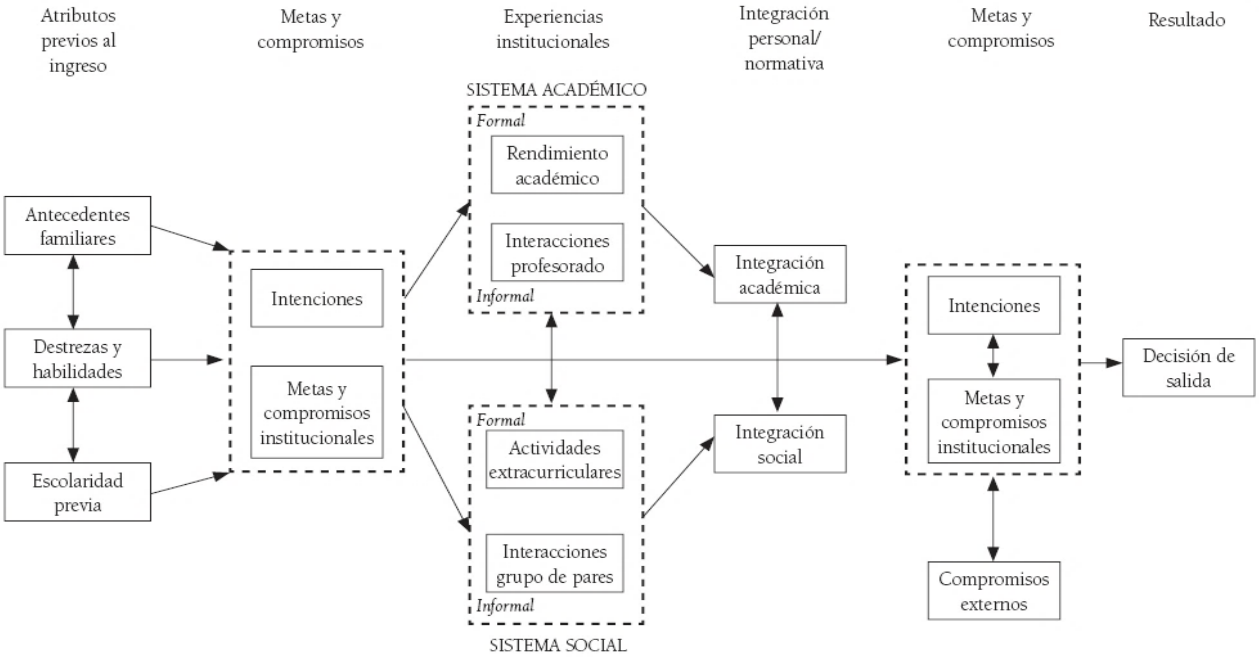

Figura 1. Modelo longitudinal de la salida institucional, (Tinto, 1975)

Tabla 1. Deserción Institucional, Fuente: REDUAI CUECH

Universidad	2008	2009	2010	2011	2012
UBIOBIO	18,80%	16,70%	13,60%	17,70%	9,40%
UTALCA	11,30%	8,40%	10,00%	12,50%	10,90%
UMAG	13,20%	13,40%	17,50%	16,70%	11,80%
ULS	13,90%	13,60%	15,60%	16,20%	12,70%
ULL	13,40%	16,00%	13,30%	20,00%	12,90%
USACH	12,81%	10,68%	11,10%	14,94%	14,42%
UANTOF	21,00%	20,40%	20,20%	25,70%	15,10%
UFRO	12,70%	14,50%	15,40%	17,40%	15,50%
Promedio CUECH	17,29%	16,72%	16,45%	19,36%	15,99%
UV	20,80%	20,40%	16,10%	17,90%	16,60%
UMCE	16,90%	15,90%	17,60%	15,50%	17,60%
UPLA	15,50%	15,20%	18,20%	24,10%	18,00%
UDA	24,80%	23,30%	22,80%	30,50%	18,70%
UNAP	25,60%	29,90%	28,00%	36,80%	26,50%
UTEM	23,40%	23,90%	25,60%	24,80%	29,50%

Pese a la implementación de diferentes planes remediales, aún existe una dificultad para responder a esta problemática en forma eficiente, lo cual da luces de que probablemente los recursos están siendo concentrados en las variables incorrectas o en los conjuntos de estudiantes inadecuados.

Mejorar la gestión de la deserción en la institución resulta de vital importancia, dado que esto repercutirá en una mejora de la gestión financiera, ya que si se considera que cada estudiante que deserta paga alrededor de $2 000 000 (pesos chilenos) al año, la universidad tiene una pérdida por concepto de arancel anual de $510 000 000, la cual representa un 29% de los ingresos totales, y esto es solo en el nivel de pregrado, donde se enfocará esta investigación. Por lo tanto, se puede decir que es de una gran importancia la resolución de este problema.

Teniendo en consideración lo anteriormente expuesto, la Universidad Arturo Prat, plantea el objetivo de tomar como herramienta principal la minería de datos para poder estudiar las variables (input) que inciden en la deserción de primer año y generar un modelo que permita clasificar a los estudiantes con mayor riesgo de deserción, para de esta forma enfocar de mejor manera los esfuerzos y recursos asociados a la disminución de esta.

Es por tanto importante indicar que se pretende inferir es si la Deserción de primer año de la Universidad Arturo Prat depende de las variables académicas de ingreso de los estudiantes mediante la minería de datos.

Para llevar a cabo este trabajo se propone la aplicación de la metodología CRISP-DM, ya que se hace necesario contar con un marco metodológico que guie las etapas del proyecto. Se ha decidido utilizar esta metodología debido a que es una de las mayormente utilizadas en proyectos de este tipo según Piatetsky (2015). Esta metodología será adecuada al contexto y características del problema planteado.

El uso de la minería de datos aplicada a la educación no es un tema nuevo; existen diversas publicaciones en el ámbito nacional como internacional que dan cuenta del estudio y aplicación de técnicas de minería de datos para intentar resolver el problema de la deserción estudiantil, destacando en Chile principalmente el uso de la metodología CRISP-DM y modelos de árboles de clasificación. A continuación se presentan algunas de estas experiencias desarrolladas nacional e internacionalmente sobre la temática de minería de datos aplicada a la deserción estudiantil.

A nivel latinoamericano se tiene el trabajo realizado por Timaran (2013) en la Universidad de Manizales, Colombia, el cual consistió en la aplicación de técnicas de minería de datos para la extracción de perfiles de los estudiantes desertores. En Argentina se encuentran trabajos como el de Sposito *et al*. (2010), donde se aplicaron técnicas de minería de datos para evaluar el rendimiento académico y la deserción de los estudiantes del Departamento de Ingeniería e Investigaciones Tecnológicas.

A nivel local se han llevado a cabo trabajos donde se ha aplicado la minería de datos principalmente para determinar el riesgo de desertar. Trabajos como el desarrollado por Fisher (2012) en la Universidad de las Américas, el cual consistió en proponer una metodología que permita identificar en forma automática a los estudiantes con mayor riesgo de deserción de las carreras de ingeniería.

En el ámbito internacional es posible encontrar trabajos como el caso de estudio realizado por Zhang *et al*. (2010) en la Universidad de Thames Valley, denominado "Use data mining to improve student retention in higher education – A case study".

En el trabajo desarrollado por Herzog (2006) se utilizaron los datos institucionales y pruebas de ingreso a la universidad (ACT, American College Test), para estimar la retención de los estudiantes y el tiempo de obtención del título.

Todas las experiencias relatadas permiten identificar que consistentemente se ha estado utilizando la minería de datos para comprender e intentar dar explicación al fenómeno de la deserción estudiantil, encontrándose que las principales tendencias en este tema van por el camino de la utilización de una metodología que guíe el proceso de minería, como es CRISP-DM, acompañada por el uso de técnicas de clasificación (dependiendo del tipo de solución buscada), como son los árboles de decisión, las redes neuronales y los métodos bayesianos.

MATERIAL Y MÉTODOS

A continuación se presenta el desarrollo del proceso de minería de datos guiado según la metodología CRISP-DM, la cual está compuesta por los siguientes pasos: Comprensión del negocio, Comprensión de los datos, Muestra, Preparación de los datos, Eliminación de valores nulos, Valores perdidos, Transformaciones de atributos y Selección de variables.

Selección de metodología

Dadas las características del problema planteado y los resultados encontrados en las investigaciones antes descritas, es que la solución de la problemática se define como una tarea de clasificación, en el caso del problema planteado esta variable será la deserción, que identificará el estado de un estudiante como: "Deserta" o "Permanece".

Algunas de las mejores técnicas para el estudio de la deserción de nivel de pregrado en el contexto de los algoritmos de clasificación corresponden a los árboles de decisión, métodos bayesianos y redes neuronales. A continuación se presenta la selección de algoritmos a utilizar, para cada una de estas categorías.

Árboles de decisión

La literatura revisada sugiere que los árboles de decisión entregan buenos resultados para clasificar a los estudiantes desertores, principalmente con el algoritmo J48, Timaran (2013) y Sposito (2010). Aparte de su uso para el modelado, los árboles de decisión pueden ser utilizados para explorar y clarificar datos en cubos dimensionales en ambientes de inteligencia de negocios. La utilidad de los árboles de decisión viene dada porque son una forma de análisis multivariante y como tal permite predecir, explicar, describir o clasificar un objetivo.

Para este trabajo se evaluarán los siguientes algoritmos para árboles de decisión:

- J48
- Random Forest
- Simple CART

Métodos bayesianos

Las Bayesian Networks (2007), pertenecen a la familia de modelos probabilísticos gráficos. Estas estructuras son utilizadas para representar conocimiento de un dominio incierto.

Para este trabajo se evaluarán los métodos bayesianos que poseen buenos resultados como lo demuestran los trabajos de Er (2012) y Zhang *et al*. (2010).

- Naive Bayes
- Bayes Net

Redes neuronales artificiales

Una red neuronal es un sistema de entradas y salidas compuesto por muchos elementos de procesamiento simples y similares; cada uno de estos elementos tiene un número interno de parámetros los cuales se les denomina "pesos". Al cambiar el peso de un elemento, se altera el comportamiento del mismo y por lo tanto, también cambiará el comportamiento de toda la red. La meta es poder seleccionar los pesos de la red que sean capaces de alcanzar una cierta relación de entrada y salida, a este proceso es lo que se conoce como "entrenamiento de la red neuronal" Nguyen y Berbard (1990).

Para este trabajo se utilizará el siguiente algoritmo de redes neuronales:

- Multilayer Perceptron

Evaluación

Para la evaluación de los modelos y la verificación de la hipótesis se prepararon una serie de experimentos cuya metodología de validación fue realizada utilizando los parámetros de: exactitud, sensibilidad, área bajo la curva ROC y estadístico de Kappa. Para la evaluación de los algoritmos se ha agregado también la ejecución del algoritmo ZeroR, es conveniente la utilización de este algoritmo debido a su propiedad de clasificar todos los datos en base a la clase mayoritaria, por lo tanto el porcentaje de aciertos obtenidos por este algoritmo será la meta base a superar.

A continuación se definen una serie de experimentos que buscan comprobar la hipótesis planteada y los resultados obtenidos de la evaluación de los algoritmos.

- **1er Experimento:** se utilizan todas las variables identificadas durante el proceso de selección de variables.
- **2do Experimento:** se excluyen las variables académicas de ingreso y se

miden nuevamente los parámetros de evaluación.

- **3er Experimento:** como una forma de validación se excluyen las variables de aprobación y promedio final, para comprobar el efecto que poseen para explicar la deserción.
- **4to Experimento:** se analiza el comportamiento de los modelos utilizando sólo las variables académicas de ingreso.

Metodología de validación

Ya se ha mencionado que el problema planteado corresponde a uno de clasificación, por lo tanto es lógico que las metodologías de validación a utilizar sean las que permitan analizar y evaluar algoritmos de este tipo. Es así que para la evaluación de los modelos se utilizan los siguientes parámetros:

Exactitud: corresponde según Chand y Manoj (2013), al porcentaje de tuplas correctamente clasificadas del total.

$$Exactitud\ (Accuracy) = \frac{TP + TN}{TP + FP + FN + TN}$$

TP: verdaderos positivos, TN: verdaderos negativos, FP: falsos positivos, FN: falsos negativos.

Sensibilidad: También conocida como la tasa de verdaderos positivos (True Positive Rate), mide la proporción de observaciones correctamente clasificadas como positivas, Chand y Manoj (2013).

$$Tasa\ de\ verdaderos\ Positivos\ (TPR) = \frac{TP}{P}$$

P: Positivos totales

Área bajo la curva de ROC: es una de las principales técnicas usadas por Vuk y Tomaz (2006) para evaluar el desempeño de modelos de clasificación y es usualmente utilizada como una medida de calidad de los clasificadores probabilísticos. Una de las características que destaca Fawcett (2003) que la hacen especialmente útil para evaluar clasificadores es el hecho de no ser afectada por sesgos en la distribución de las clases.

Estadístico de Kappa: este coeficiente mide el grado de acuerdo entre la predicción y la clase real ajustando el efecto del azar. Cuando no existe acuerdo, más que el esperado por el azar el valor del estadístico será 0, mientras que cuando exista un total acuerdo el valor del estadístico será 1, según Carletta (1996).

Para la validación de los resultados y la aplicación de los modelos se utilizará la herramienta WEKA (*Waikato Environment for Knowledge Analysis*) (2013).

RESULTADOS

Comprensión de los Datos

Para esta investigación se han utilizado diversas fuentes de datos que posee la Institución, siendo una de las principales, el *data warehouse*. Este almacén de datos posee, en su mayoría, información referente a la progresión del estudiante destacando: proceso de selección estudiantes, matriculas de estudiantes, rendimiento académico y titulación.

Sistema Curricular y Docente (SICDO)

El sistema SICDO es un sistema desarrollado internamente en la Universidad Arturo Prat, basado en Oracle Forms y bases de datos Oracle 11g. Este sistema es el encargado de gestionar la información académica de los estudiantes considerando mantención de los planes de formación, carreras, generación certificados, ingreso de notas parciales, administración de guía académica y asignaturas.

Sistema Financiero (ICON)

Corresponde al ERP financiero de la institución sistema desarrollado en Oracle Forms sobre una base de datos Oracle.

Bases de datos DEMRE

Corresponde a la información de los estudiantes inscritos en el proceso de selección universitaria (PSU) la cual es entregada por el Departamento de Evaluación, Medición y Registro Educacional (DEMRE). Esta información se hace llegar a las universidades que participan en el proceso e incluye información académica y socioeconómica de los estudiantes inscritos.

A partir de las fuentes de información descritas, se seleccionar las variables visualizadas en la Tabla 2.

Muestra

La muestra está conformada por 5.547 registros de estudiantes (cohorte 2009-2015), los cuales han sido clasificados como

estudiantes que Desertan o Permanecen en primer año, según si se mantienen o no matriculados en el año siguiente en la misma carrera. La Tabla 3 muestra la distribución de esta información, además se analiza la distribución porcentual de la deserción.

Eliminación de valores nulos

Se encontró que los datos correspondientes a estudiantes de ingresos por vías especiales (ingreso no PSU) no contaban con muchos de los valiosos datos que provienen desde las bases de datos de DEMRE, y ante la imposibilidad de generarlos u obtenerlos desde otras bases de datos externas, se decidió no considerar a estos estudiantes en la muestra, los que corresponden a un 19% del total.

Valores perdidos

En casos excepcionales se encontraron valores perdidos para puntajes promedio PSU, puntaje de prueba de lenguaje y prueba de matemáticas. Para corregir estos datos se utilizó la herramienta de pre-procesamiento de WEKA y la técnica "*Replace missing values*" la cual reemplaza los valores perdidos por el valor medio encontrado para el atributo.

Transformaciones de atributos

En la Tabla 4 se presentan las principales parametrizaciones realizadas para las trasformaciones.

Selección de variables

Para llevar a cabo este trabajo, se utilizó la herramienta WEKA (2015) mediante los siguientes métodos: ChiSquaredAttributeEval, GainRatioAttributeEval, InfoGainAttributeEval y ReliefFAttributeEval.

Cada uno de los métodos anteriormente descritos fueron aplicados utilizando la opción de entrenamiento de validación cruzada con 10 carpetas (*cross-validation 10 folds*).

De la ejecución de los algoritmos de evaluación de variables de entrada (input) se encontró que todos generaban un ranking similar de variables. En todas las evaluaciones se encontró que las variables de aprobación y promedio final fueron las más significativas para explicar la variable objetivo (deserción), mientras que el RUT y código de carrera (id_carrera) tienen la menor significancia para explicarla.

A continuación se presenta Tabla 5 con los resultados de la evaluación en base a los parámetros establecidos para las metodologías seleccionadas:

A. Exactitud:

Los resultados demostraron que el algoritmo Random Forest en general tiene una mejor exactitud en cada uno de los experimentos. Los datos del experimento 2 muestran que, en general, el excluir las variables académicas de ingreso no tiene mayores efectos en el desempeño de los algoritmos, no así como ocurre en el experimento 3, donde se observa una baja considerable de rendimiento, disminuyendo en un 30% la exactitud de los algoritmos. Este comportamiento se puede apreciar para todos los algoritmos exceptuando a ZeroR, algo esperable considerando que la clase mayoritaria es "Permanece", con una distribución del 73,41% de la muestra. Finalmente el experimento 4, donde solo se analizan las variables académicas de ingreso tiene un desempeño igualmente pobre que el experimento 3, demostrando que las variables académicas de ingreso no son significativas a la hora de clasificar correctamente las instancias.

B. Sensibilidad: para el análisis de sensibilidad los resultados son evaluados y analizados para la clase desertor, ya que el interés de la institución es identificar correctamente a los estudiantes desertores.

De acuerdo a lo observado, se puede apreciar que el algoritmo Naive Bayes logra el mejor rendimiento en este parámetro para el experimento 1 (todas las variables). En los cuatro experimentos realizados se puede observar que los métodos bayesianos son los que obtienen la mayor sensibilidad, específicamente NaiveBayes, lo cual implica que es el algoritmo más adecuado a utilizar para identificar a los estudiantes desertores.

Según los experimentos realizados (2, 4) se demuestra que bajo la métrica de sensibilidad, las variables académicas de ingreso no son significativas a la hora de predecir la deserción de estudios.

C. Área bajo la curva ROC:

En general el algoritmo Random Forest ofrece mayor capacidad para discriminar estudiantes desertores. El experimento 2 muestra que el desempeño de los modelos se ve afectado solo marginalmente al excluir las variables académicas de ingreso de los estudiantes. Al

excluir las variables de aprobación y promedio final en el experimento 3 se puede observar una disminución sustancial en la capacidad para discriminar de todos los algoritmos viéndose especialmente afectados los algoritmos Multilayer Perceptron y J48, llegando a presentar resultados equivalentes al azar. Al realizar el experimento 4 utilizando solo las variables académicas se puede observar un bajo desempeño de los algoritmos en general, excepto solo por el Algoritmo Multilayer Perceptron que logra un leve aumento en su capacidad para discernir entre las clases "Desertor" y "Permanece". Bajo el parámetro del área bajo la curva de ROC, es posible decir que las variables académicas de ingreso no tienen un efecto significativo a la hora de determinar la deserción de estudios.

D. Estadístico de Kappa: se encuentra que para los experimentos 1 y 2 el algoritmo Random Forest es el que presenta el mayor nivel de concordancia con las clases. Para la evaluación del nivel de aceptación de este índice se puede utilizar una tabla como la propuesta por López y Fernández (1999), la cual se refleja en la Tabla 6.

Según lo indicado por la Tabla 6 se puede decir que para los experimentos 1 y 2, en general, los algoritmos poseen una buena concordancia para predecir. Los experimentos 3 y 4 nuevamente vuelven a presentar los peores resultados, generando valores pobres de concordancia.

Los resultados del estadístico de Kappa indican que al excluir las variables académicas de ingreso no existe una pérdida importante de concordancia entre la clase a predecir y la clase real; por otro lado al generar los modelos utilizando solo estas variables disminuye significativamente la capacidad de los modelos para lograr concordancia entre la clase predicha y la clase real de la variable objetivo.

Tabla 2. Variables de estudio

Variables	**Variables**
RUT	Periodo de Egreso Enseñanza Media
Deserción	Tramo de Ingreso Bruto Familiar
Facultad	Preferencia de Postulación
Carrera	Código de Carrera
Educación de los Padres	Tipo de Colegio
Sede	Promedio PSU
Sexo	Nota de Enseñanza Media (NEM)
Lejanía del Hogar	Comuna de Procedencia
Prueba Mejor Desempeño	Puntaje de la Prueba de Matemáticas
Puntaje Prueba Lenguaje	Posee Beneficios
Porcentaje de Asignaturas Aprobadas al término del año	Año de Ingreso

Tabla 3. Distribución de datos en cantidad y porcentual de la deserción UNAP, para estudiantes ingreso PSU

	2009		2010		2011		2012		2013		2014		2015		Total General
Estado Alumno	cant	%	cant	%	cant	%	cant	%	cant	%	cant	%	cant	%	
Deserta	267	28	212	26	232	34	164	24	163	26	203	23	255	29	1496
Permanece	696	72	600	74	454	66	525	76	465	74	687	77	624	71	4051
Total General	**963**		**812**		**686**		**689**		**628**		**890**		**879**		**5547**

Tabla 4. Parametrización de atributos

Variable	Descripción	Parametrización
Educación de los Padres	El atributo de educación de los padres viene dado desde la base de datos DEMRE. Este atributo ha sido transformado para ser expresado de forma dicotómica.	• Algún padre profesional • Ningún padre profesional
Periodo de Egreso	El atributo de periodo de egreso indica el año en el que el estudiante egreso de la enseñanza media. Permite identificar si el estudiante se encuentra recién egresado de enseñanza medio (último año) o es un estudiante egresado de años anteriores.	• Último año • Años anteriores
Ingreso Bruto Familiar	Ingreso bruto familiar: proveniente de las bases de datos de DEMRE, este atributo tiene 15 clasificaciones según documento oficial y se transforma a 3 clasificaciones.	• 0 – 288 000 • 288 001 – 576 000 • 576 001 – montos mayores
Distancia del Hogar	Variable creada en base a la comuna de procedencia del estudiante y la ubicación de la sede en donde estudia, se categoriza como "cerca" a aquel estudiante cuya procedencia es igual a la sede de estudios y "lejos" para aquellos estudiantes que provienen de una comuna distinta a la sede de estudios.	• Cerca • Lejos
Preferencia de Postulación	Esta variable indica el lugar en el que el estudiante postulo a la carrera estudiada.	• Primera preferencia • Otras preferencias
Deserción	Esta es la variable objetivo e indica si el estudiante continua sus estudios al año N+1 en la misma carrera.	• Deserta • Permanece
Prueba de Mejor Desempeño	Como forma de reconocer las habilidades particulares del estudiante, se genera esta variable basada en las variables de las pruebas de matemáticas y lenguaje, la cual obtendrá el valor según la prueba donde el estudiante obtuvo su mejor desempeño.	• Matemáticas • Lenguaje
Posee Beneficios	Variable que permite determinar si un estudiante posee beneficios de algún tipo, ya sea becas, crédito, ambos o ninguno de los casos anteriores.	• Solo Beca • Solo Crédito • Beca y Crédito • Sin Beneficios
Aprobación	Corresponde al porcentaje de aprobación anual de primer año del estudiante y se define con la siguiente formula: $\frac{Nro\ de\ asignaturas\ aprobadas}{Nro\ de\ asignaturas\ cursadas}$	• No aplica

Tabla 5. Cuadro resumen de análisis para los experimentos en relación a los parametros evaluados.

Algoritmo/ Parámetro		ZeroR	J48	SimpleCART	RandomForest	NaiveBayes	BayesNet	MultilayerPerceptron
Exp 1	**A**	73,03%	89,49%	89,09%	89,87%	87,29%	87,72%	87,18%
	B	0	0,722	0,729	0,726	0,767	0,755	0,595
	C	0,499	0,856	0,878	0,913	0,902	0,902	0,868
	D	0	0,7184	0,7106	0,728	0,678	0,685	0,6362
Exp 2	**A**	73,03%	89,18%	89,13%	89,23%	87,89%	88,12%	88,44%
	B	0	0,714	0,729	0,734	0,767	0,753	0,699
	C	0,499	0,856	0,875	0,907	0,905	0,904	0,857
	D	0	0,7097	0,7113	0,7307	0,6908	0,6933	0,6897
Exp 3	**A**	73,03%	73,03%	73,03%	69,57%	71,77%	71,82%	68,40%
	B	0	0	0	0,173	0,202	0,203	0,1
	C	0,499	0,499	0,499	0,587	0,623	0,62	0,506
	D	0	0	0	0,0741	0,1331	0,1345	-0,0003
Exp 4	**A**	73,03%	72,65%	73,03%	71,41%	70,74%	70,81%	68,84%
	B	0	0,031	0	0,164	0,33	0,326	0,199
	C	0,499	0,527	0,499	0,625	0,659	0,657	0,506
	D	0	0,0201	0	0,1009	0,1924	0,191	-0,0003

A: Análisis de exactitud de los algoritmos
B: Análisis de sensibilidad de los algoritmos (clase desertor)
C: Análisis área bajo la curva de ROC de los algoritmos
D: Análisis del estadístico de Kappa de los algoritmos

Tabla 6. Valoración del índice de Kappa

Valor de K	Fuerza de la concordancia
<0,20	Pobre
0,21 – 0,40	Débil
0,41 – 0,60	Moderada
0,61 – 0,80	Buena
0,81 – 1,00	Muy Buena

DISCUSIÓN

En base a los resultados obtenidos por este trabajo se puede establecer que: "La Deserción de primer año de la Universidad Arturo Prat no depende de las variables académicas de ingreso de los estudiantes". Esto ya que los experimentos realizados demostraron que las variables académicas de ingreso no afectan significativamente la deserción de primer año. Por otra parte, los experimentos realizados encontraron que las variables de desempeño académico de primer año (aprobación y promedio final) tienen una alta significancia a la hora de explicar la deserción.

El algoritmo Random Forest demostró tener el mejor desempeño en todos los parámetros por sobre el resto de los modelos al utilizar todas las variables predictivas, a excepción del análisis de sensibilidad, donde el algoritmo Naive Bayes resultó ser el de mejor desempeño y, por lo tanto, el escogido para la explotación del modelo, dado que cumple de mejor manera con el objetivo del negocio (detectar la mayor cantidad de estudiantes desertores) y no presenta grandes diferencias de desempeño para el resto de los parámetros analizados.

Con los resultados generados es posible comenzar a buscar en base al uso de información, como guía del proceso, las verdaderas causas tras la deserción de estudios en la universidad.

Producto de este trabajo de investigación, la Universidad Arturo Prat cuenta con dos productos que le permitirán focalizar de mejor manera esfuerzos y recursos; por un lado, el conocimiento que no se tenía previamente

acerca del efecto de las variables de ingreso académico y por otro, se encuentra el modelo generado que entrega a la institución un insumo de información clave para la gestión de la deserción de primer año, con esto se podrá optimizar la ejecución de programas como tutorías, apoyo financiero, reforzamientos y otros que actualmente no tienen los resultados esperados para las autoridades.

Otro de los aportes de este trabajo desarrollado, además de los resultados expuestos, corresponde a la propuesta de un plan de explotación de los resultados, monitoreo y mantención del modelo de minería, etapa que según la literatura estudiada no suele ser profundizada, donde la definición de métodos de monitoreo y métricas de evaluación de uso y resultados resultan claves para el éxito y ejecución sistemática de este tipo de proyectos al interior de las instituciones.

La metodología CRISP-DM representa un aporte significativo para el desarrollo del proceso de minería de datos, entregando una guía fundamental para llevar a cabo proyectos de este tipo, su enfoque iterativo plantea el proceso de minería de datos como un proceso de mejora continua que permite asegurar la calidad del trabajo realizado.

Si bien los modelos generados para esta investigación aplicada tienen una buena capacidad para predecir la deserción de estudios (77% de exactitud), determinando que las variables que mayor injerencia tienen en la deserción correspondían al desempeño académico en el primer año (aprobación y promedio final), se hace necesario continuar profundizando en el problema. De los resultados obtenidos emergen preguntas como: ¿si las variables académicas de ingreso no tienen un efecto importante en la deserción, como el desempeño académico afecta a los estudiantes en su primer año? ¿Cuáles son las variables que afectan el rendimiento de los estudiantes y por lo tanto gatillan la deserción de estudios?, mediante trabajos futuros donde también se utilicen técnicas de minería de datos es posible estudiar el fenómeno de la aprobación y reprobación de cursos, y así determinar los factores que la afectan.

Como otro trabajo futuro se propone sistematizar las predicciones realizadas por la herramienta WEKA y llevarlas a una base de datos, donde la información sea distribuida y gestionada mediante un sistema de tipo CRM *(Costumer Relationship Management)* que se encuentre integrado al sistema académico de la institución y que permita llevar un registro de las gestiones exitosas (recuperar un estudiante marcado como desertor), de esta manera se podrá monitorear el éxito de las gestiones de control de la deserción de forma automatizada.

REFERENCIAS BIBLIOGRÁFICAS

Bayesian Networks. (2007) *Encyclopedia of Statistics in Quality y Reliability*, Wiley y Sons.

Berger, J. and J. Milem (2000) *Organizational Behavior in Higher Education and Student Outcomes*. En: J. Smart (Ed.), Higher Education*:* Handbook of theory and research 15. Springer Science & Business Media.

Braxton, J. M., Milem, J. F., y Sullivan, A. S. (2000) The Influence of Active Learning on the College Student Departure Process: Toward a Revision of Tinto's Theory. *Journal of Higher Education*, 71 (5), 569–590.

Carletta, J. (1996) Assesing agreement on clasification tasks: the kappa statistic. *Computational linguistic*, 22 (2), 249-254.

Chand, T. y Manoj, J. (2013) WEKA Approach for Comparative Study of Classification Algorithm. *International Journal of Advanced Research in Computer and Communication Engineering*, 2 (4), 1925-1931.

Díaz, C. (2008) Modelo Conceptual para la Deserción Estudiantil Universitaria Chilena. *Estudios Pedagógico*s, 34 (2), 65-86.

Er, E. (2012) Identifying At-Risk Students Using Machine Learning Techniques: A Case Study. *International Journal of Machine Learning and Computing,* 2 (4), 476-480.

Fawcett, T. (2003) *ROC Graphs: Notes and Practical Considerations for Data Mining Researchers* [en línea] Disponible en: http://www.hpl.hp.com/techreports/2003/HPL-2003-4.pdf [Consulta: 08 enero 2016]

Fernández, O., Martínez-Conde, M., Melipillán, R. (2009) Estrategias de Aprendizaje y Autoestima: su Relación con la Permanencia y Deserción Universitaria. *Estudios Pedagógicos*, 34 (1), 27-45.

Fishbein, M. *y* Ajzen, I. (1975) *Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research.* Reading, MA: Addison-Wesley.

Fisher Angulo, E. (2012) *Modelo para la automatización del proceso de determinación de riesgo de deserción en estudiantes universitarios* [en línea] Disponible en: http://repositorio.uchile.cl/handle/2250/111188 [Consultado 04 diciembre 2015]

Herzog, S. (2006) Estimating Student Retention and Degree - Completion Time: Decision Trees and Neural Networks Vis-a-Vis Regression. *New Directions For Institutional Research*, 131, 17-33.

Himmel K. Erika. (2002) Modelo de Análisis de la Deserción estudiantil en la Educación Superior", *Calidad en la Educación*, 17, 91-108.

Lopez, G. y Fernández, P. (1999) *Medidas de concordancia: el índice de Kappa* [en línea] Disponible en: http://fisterra.com/mbe/investiga/kappa/kappa2.pdf [consulta: 06 octubre 2015]

Medrano, L., Galleano, C., Galera, M. y Del Valle, R. (2010). Creencias Irracionales, Rendimiento y Deserción Académica en Ingresantes Universitarios*. LIBERAVIT*, 16 (2), 183-191.

Nguyen, D. y Bernard, W. (1990) Neural Networks for Self-Learning Control Systems [en línea]. *IEEE Control Systems Magazine*. Disponible en: https://web.stanford.edu/class/ee373b/NNselflearningcontrolsystems.pdf [Consulta: 03 noviebre 2015]

Piatetsky, G. (2014) *Kdnuggets* [en línea]. Disponible en: http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html [Consulta: 08 ocbre 2015]

Prieto, A. (2002) Indecisión Vocacional: perdidas y "perdidos" en la educación superior. *Calidad en la Educación*, 17, 145-163.

SIES. (2014). -. MINEDUC, *Principales resultados evolución de la retención de 1er año* 2009-2013.

Spady, W. G. (1970) Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange*, 1 (1) 64-85.

Spositto, O., Etcheverry, M. , Ryckeboer, H. y Bossero, J. (2010) Aplicación de Técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil. *Memorias de la Novena Conferencia Iberoamericana de Sistemas. Cibernética e Informática (CISCI 2010)*, Orlando, Florida. EEUU.

St. John, E., Hu, S., Simmons, A., Faye, D., Weber, J. (2004) What Difference Does a Major Make? The Influence of College Major Field on Persistence by African American and White Student. *Research in Higher Education*, 45 (3), 209-232.

Timaran, R., Calderon, A., Romero y Jimenez Toledo, J. (2013) Aplicación de la Minería de datos en la extracción de Perfiles de Deserción Estudiantil. *Ventana informática*, 28, 31-47.

Tinto, V. (1975) Dropout from Higher Education: A theoretical synthesis of recent research. *Review of Educational Research*, 45 (1), 89-125.

Tinto, V. (1993) *Leaving College; Rethinking the Causes and Cures of Student Attrition*. Second Edition. Chicago. The University of Chicago Press.

Vuk, M. y Tomaz, C. (2006) ROC Curve, Lift Chart and Calibration Plot. *Metodolosky zvezki,* 3 (1), 89-108.

Weka (2013) *Waikato Environment for Knowledge Analysis* [en línea] Disponible en: http://www.cs.waikato.ac.nz/ml/weka/index.html [Consulta: 05 diciembre 2015]

Weka (2015) *Waikato Environment for Knowledge Analysis* [en línea] Disponible en: http://weka.sourceforge.net/doc.stable/ [Consulta: 20 octubre 2015]

Zhang, Y., Oussena, S., Clark, T. y Kim, H. (2010) Use Data Minng To Improve Student Retention in Higher Education- A Case Study [en línea]. *ICEIS 2010 - Proceedings of the 12th International Conference on Enterprise Information Systems, Volume 1*. Disponible en: http://repository.uwl.ac.uk/723/ [Consulta: 06 noviembre 2015]