USO DE DATA MINING EN LA PREDICCIÓN DE LA DESERCIÓN UNIVERSITARIA

Predicción de los factores que influyen en la deserción universitaria

Claudio Henríquez, Ingeniero Civil en Informática, Magister en Ingenieria informática

RESUMEN

Actualmente en la Facultad de Ingeniería se puede observar una alta deserción, lo cual se generaliza a nivel país respecto de la deserción de alumnos que ingresan al sistema universitario, convirtiéndose en una problemática, tanto para las instituciones académicas, el gobierno, los mismos estudiantes y sus familias. Bajo este contexto, se resumirá en este trabajo todos los proyectos que han generado algún modelo de Deserción para alguna carrera o para la misma Facultad de Ingeniería de la Universidad Central de Chile.

El primer estudio realizado, abordó la creación de modelos predictivos de deserción para la carrera de Ingeniería Civil en Computación e Informática (Bustamante-Morales), considerando modelos de primer año y segundo año, posteriormente surgió un análisis a nivel Facultad, considerando modelos predictivos (Hurtado) y modelos descriptivos (Ausset-González), este último enfocado principalmente al área de Ciencias Básicas. Por último, se realizó un estudio orientado a la carrera de Ingeniería Civil Industrial, utilizando un modelo econométrico y un modelo predictivo.

Por lo cual la siguiente investigación otorga una herramienta de apoyo a la toma de decisiones. En el presente estudio se aplicó minería de datos sobre la base de datos de los estudiantes de la Facultad de Ingeniería de la Universidad Central de Chile, con el fin de clasificar a los estudiantes entre no desertores y posibles desertores a través de un modelo predictivo.

PALABRAS CLAVES
deserción universitaria, modelo predictivo, minería de datos.

INTRODUCCIÓN

Para las casas de Estudio de Educación Superior es fundamental lograr que sus estudiantes logren superar las barreras que se le van presentando durante el primer año y continúen estudiando hasta finalizar el plan de estudios. Aquí es donde se presenta el concepto de deserción estudiantil, se deben identificar los factores que influyen y así lograr disminuirla dentro del sistema educativo.

Para la Universidad Central de Chile, todos los estudiantes que suspenden sus estudios son posibles desertores. Para la casa de estudios, la definición de deserción es "el abandono voluntario y definitivo de los estudios, este puede ser explicado por diferentes causas, financiera, académica, vocacional, familiar, salud y servicio".

La falta de conocimiento previo sobre las variables que pueden influir en que un alumno deserte de la carrera que cursa es una de las mayores problemáticas por su complejidad de predecir, por lo tanto se hace necesario intervenir con algún tipo de mecanismo que pueda otorgar respuestas satisfactorias a las causas que provocan la deserción.

Son pocas las herramientas predictivas que pueden permitir a las universidades detectar cuáles son los estudiantes que tienen tendencia a abandonar una carrera o cambiarse a otra institución de educación superior.

Dadas las características de la problemática y la capacidad de obtención de datos asociados a los estudiantes, se pueden utilizar técnicas de minería de datos con el fin de detectar posibles desertores a partir de un modelo de deserción para la Facultad de Ingeniería de la Universidad Central de Chile.

Debido a que la Universidad Central de Chile lleva un registro histórico de los datos de cada estudiante que ingresa, es posible conseguir esta información con el fin de obtener conocimiento útil y no trivial dentro del conjunto de datos y de esta forma crear una herramienta de ayuda para la toma de decisiones dentro de la Facultad.

CONCEPTOS Y MODELOS

La deserción universitaria es uno de los temas que más interesa a los planificadores e investigadores en educación. El establecimiento de la definición de lo considerado como deserción universitaria ha generado bastante discusión. Luis González define deserción como: "el proceso de abandono, voluntario o forzoso de la carrera en la que se matricula un estudiante, por la influencia positiva o negativa de circunstancias internas o externas a él o ella" (González, 2005).

En el año 1975, el sociólogo Vincent Tinto publicó su modelo de deserción universitaria, en el cual se incluyen distintos factores determinantes con los cuales el estudiante enfrenta sus estudios universitarios, entre ellos las metas que persigue para su educación y sus compromisos institucionales (ver Figura 1), otros investigadores proponen modelos basados en el de Tinto pero con características distintas dependiendo de cada una de las universidades en donde se ha enfocado el estudio de la deserción (Narvaez, 2013).

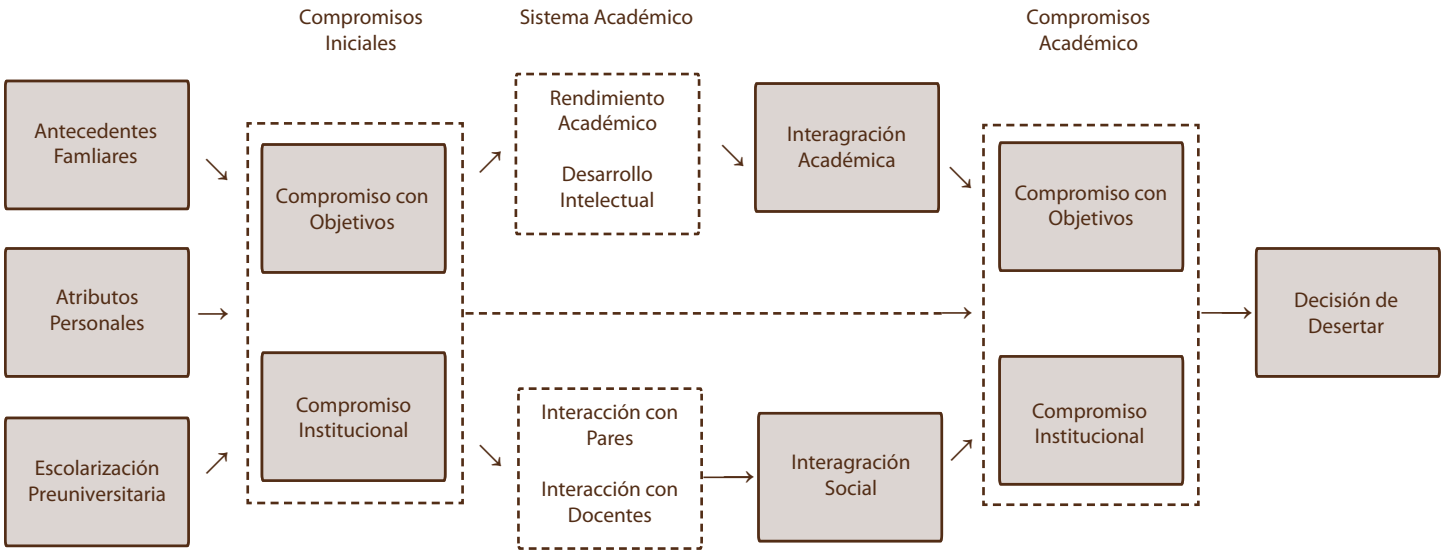

FIGURA 1

Modelo de deserción de Tinto (Narvaez, 2013)

Así como Tinto, existen otros autores que proponen modelos de deserción, tales como Ethington, Spady y Bean, centrándose este último en 4 grandes factores de relevancia, como lo son los factores académicos, factores psicosociales, factores ambientales y factores de socialización como se puede apreciar en la figura 2 (Narvaez, 2013).

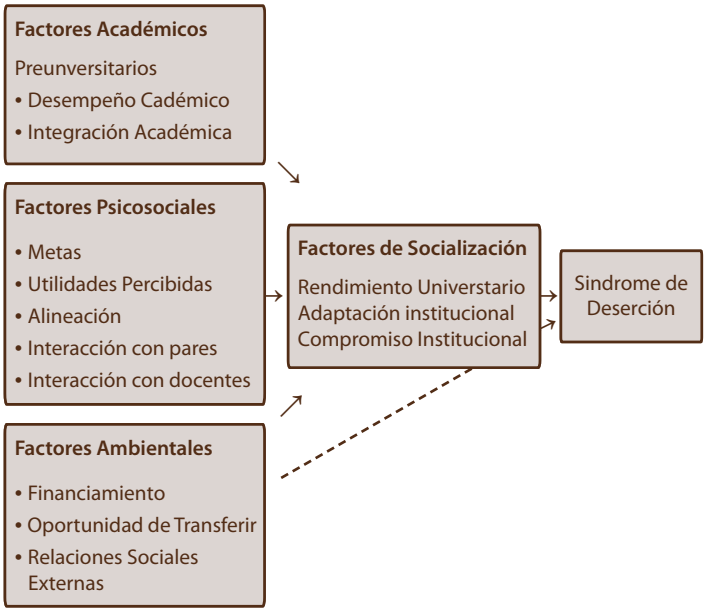

FIGURA 2

Modelo de deserción de Bean (Narvaez, 2013)

De acuerdo con el Modelo de Bean, en la Facultad de Ingeniería de la Universidad Central se ha utilizado un esquema bastante similar, principalmente se ha utilizado Minería de Datos para el estudio de la deserción de los estudiantes de la Facultad de Ingeniería, en donde el foco de la investigación se ha determinado bajo una secuencia de proyectos de título, en donde se han desarrollado modelos de deserción tanto predictivos como descriptivos, además de considerar análisis a nivel de carreras como a nivel de Facultad.

Los modelos desarrollados en la facultad se plantean en términos de factores académicos, factores ambientales y factores de socialización de acuerdo al modelo de Bean, faltando considerar factores psicosociales debido a la falta de datos al respecto, lo que a partir del modelo de tutorías de la Facultad de Ingeniería de la Universidad Central se intenta generar y que a través de un proyecto de título desarrollado por Caucott y Orellana pretende sistematizar la captura de estos datos.

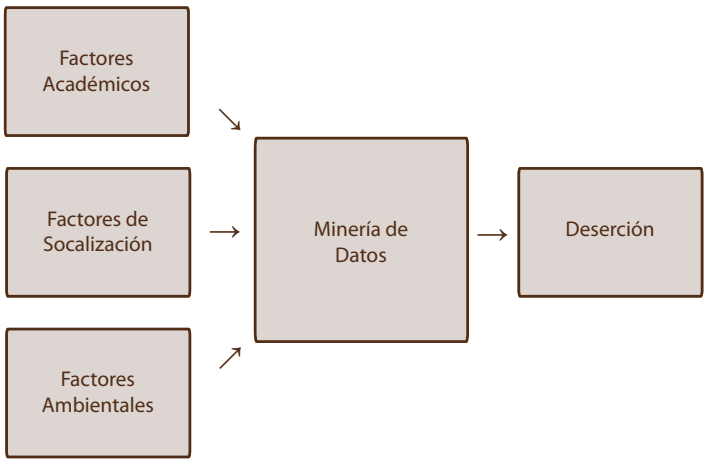

FIGURA 3

Modelo de deserción FING con Minería de Datos

Los datos obtenidos fueron otorgados por el área de Análisis Institucional de la Universidad Central de Chile y corresponden a estudiantes de la Facultad de Ingeniería desde el año 1985. Dentro de los atributos considerados en los distintos proyectos de título y acorde a los factores del modelo de deserción se tiene:

Factores Ambientales

- Domicilio de estudiante (comuna, región)
- Domicilio de colegio (comuna, región)
- Tipo de colegio
- Año Egreso Enseñanza Media

Factores Académicos

- Año ingreso a Universidad Central de Chile
- Notas PSU o PAA separadas por prueba
- Notas NEM

Factores de socialización

- Plan de Estudio
- Histórico de todas las notas finales de ramos cursados por cada estudiante
- Titulados

La minería de datos es la ciencia de extracción de conocimiento no trivial en grandes cantidades de datos, utilizando máquinas de autoaprendizaje y estadística. Para poder aplicar esta herramienta se aplica el proceso de extracción de conocimiento en bases de datos, más conocido como KDD (del inglés Knowledge Discovery in Data Base), este proceso se indica en la figura 4. Para este caso particular, los datos asociados a factores académicos, ambientales y de socialización son la entrada de este proceso, variando en algunos detalles de acuerdo al conjunto objetivo de estudiantes a analizar, que para este caso se ha realizado a la carrrera de Ingeniería Civil en Computación e informática, a la Facultad de Ingeniería y a Ciencias Básicas de la Facultad de Ingeniería, también se tiene un estudio en proceso de la carrera de Ingeniería Civil Industrial.

Además de los distintos grupos de estudiantes objetivo, se ha considerado distintos tipos de modelo de minería de datos, ya sea un modelo predictivo (clasificación) o un modelo descriptivo.

Las fases de selección, procesamiento y transformación de datos fueron llevados a cabo a través de un proceso ETL (Extracción, Transformación y Carga) a través del software Kettle Data Integration y la fase de Minería de Datos a través del software Weka o el software Data Miner.

DESARROLLO DEL MODELO EN FING

El primer proyecto de título fue desarrollado en el año 2012 por Bustamante y Morales quienes aplicaron Minería de Datos sobre los datos de los estudiantes de la carrera de Ingeniería Civil en Computación e Informática de la Universidad Central de Chile, desarrollando un modelo predictivo que caracterizó conductas de deserción en los estudiantes de la carrera (Bustamante y Morales, 2012).
El proceso ETL llevado a cabo con los datos se muestra en la figura

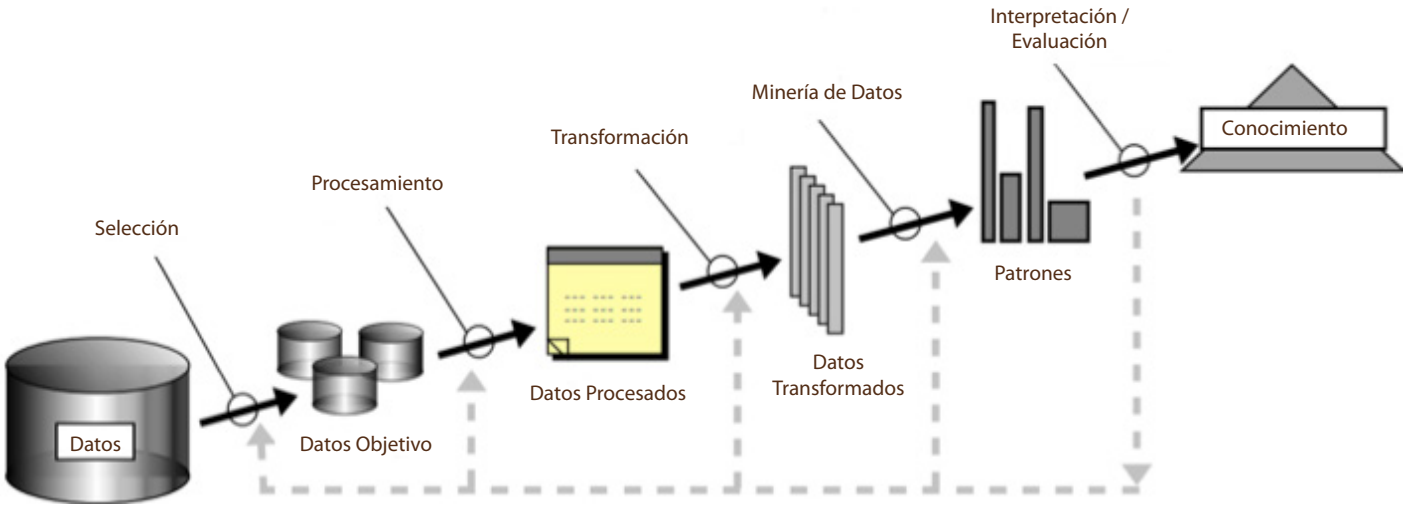

FIGURA 4
Fases del Proceso KDD

5, el cual automatiza el tratamiento de datos. Considerando la cantidad de datos, se desarrolló un modelo de deserción de primer año sobre un total de 182 estudiantes y un modelo de deserción de segundo año sobre un total de 37 estudiantes. Sobre las variables que crearon tranformaciones con el objetivo de mejorar la calidad de los datos, principalmente considerando promedios de notas por áreas o promedios acumulados, además de considerar variables de cantidad de reprobación. Dentro de los cálculos realizados están:

- **Promedio** acumulado notas de asignaturas de ciencias básicas
- **Promedio** acumulado notas de asignaturas de ingles
- **Promedio** acumulado notas de asignaturas de Ciencias de la computación e informática
- **Promedios** acumulados por asignaturas semestrales
- **Promedio** acumulado primer año, primer semestre
- **Promedio** acumulado primer año, segundo semestre
- **Promedio** acumulado segundo año, primer semestre
- **Ramos** reprobados
- **Ramos** reprobados primer año, primer semestre
- **Ramos** reprobados primer año, segundo semestre
- **Ramos** reprobados segundo año, primer semestre
- **Porcentaje** de ramos reprobados
- **Cantidad** de veces que dio las asignaturas por semestre
- **Cantidad** de veces primer semestre
- **Cantidad** de veces segundo semestre
- **Cantidad** de veces tercer semestre

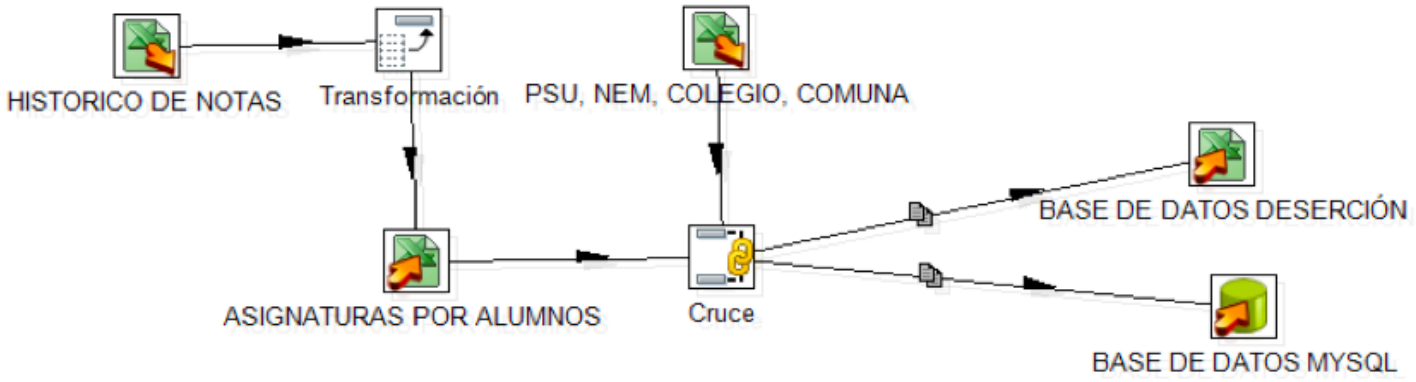

FIGURA 5

ETL para el modelo de deserción de la carrera de Ingeniería Civil en Computación e Informática (Bustamante y Morales, 2012)

Finalmente las variables a considerar en este modelo para estudiantes de primer año se muestran en la tabla 1, luego de análizar la relevancia de los atributos respecto de la variable de deserción y considerando los resultados de los estadísticos Chi Squared, Gain Ratio e Info Gain, el resto de las variables no fueron consideradas.

Atributos	ChiSquared	GainRatio	InfoGain
Promedio ACUM	587.492	0.16797	0.25952
Promedio ACUM Primer Año 1s	548.654	0.16539	0.27748
Porcentaje Repr 1s	536.034	0.25019	0.2419
Promedio CB	511.991	0.21884	0.21812
Ra Reprobados 1 años 1s	436.308	0.06433	0.201
Promedio ACUM Ciencias Ingeniería	325.061	0.19229	0.15514
Matemáticas	306.378	0.11101	0.15153
Especifica	217.781	0.09735	0.09261
Pon Total	137.126	0.18285	0.07974
Ponderación Lenguaje Mateamática NEM	118.522	0.174009	0.06919
Ponderación Lenguaje Matemáticas	20.674	0.129952	0.0955

TABLA 1

Selección de datos de entrada al modelo de deserción de la carrera de Ingeniería Civil en Computación e Informática (Bustamante y Morales, 2012)

Las herramientas de Minería de Datos seleccionas fueron Árboles d e Decisión, Clasificador Bayesiano y Máquinas de Soporte Vectorial (SVM), los cuales se muestran en la tabla 2 con sus respectivas matrices de confusión.

De acuerdo con los resultados expuestos en la Tabla 2, el Árbol de Decisión fue el mejor clasificador con una efectividad 75,4%, ya que es el que mejor aprende de la conducta desertora.

Un segundo proyecto en el 2013 aplicó Minería de Datos sobre los datos de los estudiantes de la Facultad de Ingeniería de la Universidad Central de Chile, desarrollando un modelo predictivo que caracterizó conductas de deserción en los estudiantes de la Facultad (Hurtado, 2013). En este estudio se propone la creación de un modelo de deserción al termino del primer año y un modelo de deserción al termino del segundo año, diferenciados principalmente por las asignaturas involucradas.

Al hacer un análisis comparativo de todas las carreras, sus mallas y asignaturas, se definieron los ramos comunes, los cuales se denominarán "Plan Común Ingenierías", ver la Tabla 3.

N°	Plan Común
1	Cálculo I
2	Cálculo II
3	Cálculo III
4	Álgebra I
5	Álgebra II
6	Ecuaciones Diferenciales
7	Introduccón a la Fisica
8	Mecánica
9	Ondas, Óptica y Calor
10	Electricidad y Magnetismo
11	Química General
12	Probabilidades y Estadística
13	Métodos Numéricos
14	Inglés Para Ingeniería I
15	Inglés Para Ingeniería II
16	Inglés Comunicacional
17	Transversal Institucional I
18	Transversal Institucional II

TABLA 3

Asignaturas comunes de las carreras de la Facultad de Ingeniería de la UCEN

Técnica	Arbol de Desición		Clasificador Bayesiano		SVM		
	No Desertor	Desertor	No Desertor	Desertor	No Desertor	Desertor	Total real
No Desertor	20	11	27	4	26	5	31
Desertor	4	26	8	22	7	23	30
Total Predicho	24	37	35	26	33	28	61

TABLA 2

Matrices de confusión para las 3 técnicas de minería de datos utilizadas en el modelo de deserción de primer año

Los modelos de deserción para la Facultad de Ingeniería de la Universidad Central de Chile se muestran en las figuras 6 y 7 respectivamente, luego de validar la relevancia de cada atributo basado en los test estadísticos Chi Squared, Gain Ratio e Info Gain.

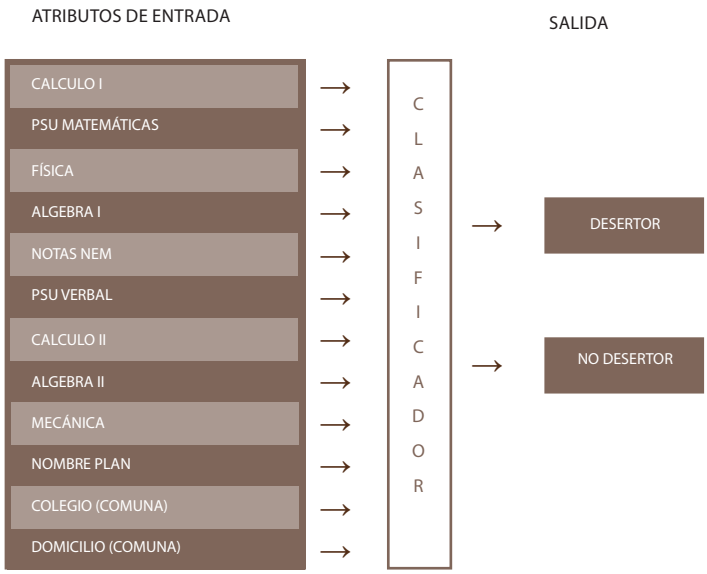

FIGURA 6

Modelo predictivo de deserción para primer año

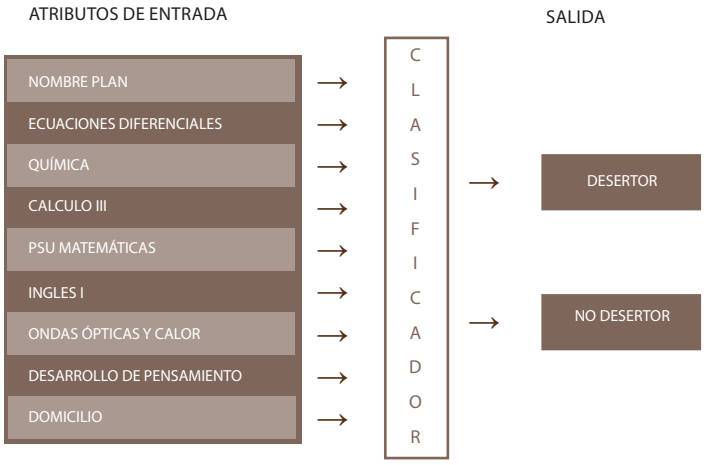

FIGURA 7

Módelo predictivo de deserción para segundo año

Los modelos definidos para estudiantes que han cursado primer año entregan los resultados presentados en la Tabla 4, la cual muestra la matriz de confusión generada posterior al entrenamiento y validación para cada una de las tres técnicas de minería de datos utilizadas en esta investigación, las cuales son: árboles de decisión, Máquinas de Soporte Vectorial (SVM) y Redes Neuronales Artificiales (RNA). En este caso el mejor modelo está asociado a la técnica de Redes Neuronales Artificiales, con un 61,5% de efectividad en los desertores. En este caso particular, además se comprobó con los 55 desertores propuestos por el modelo al siguiente año, lo cual generó un 100% de efectividad ya que los 55 estudiantes desertaron efectivamente.

La técnica de Redes Neuronales Artificiales se verificó comprobando si los 55 desertores propuestos por el modelo lograban completar su segundo año, lo cual reflejó que el 100% de ellos desertaron posteriormente.

CONCLUSIONES

La deserción estudiantil es un factor de importancia para la Facultad de Ingeniería de la Universidad Central de Chile y pese a sus esfuerzos aún se mantiene en porcentajes que superan el 25% en primer año, por lo cual adquiere relevancia explorar nuevas formar de extraer conocimiento que ayuden a tomar mejores decisiones para lograr una mejor retención de estudiantes. En este sentido, la aplicación de minería de datos ha logrado entregar parte de esa información desconocida, reconociendo un porcentaje de estudiantes de manera anticipada.

Si bien las técnicas predictivas permiten anticiparse, cabe destacar que los mejores resultados se obtuvieron con técnicas denominadas de caja negra, lo que no nos permite saber el por qué sucede la deserción. Sin embargo, estas técnicas permiten predecir un mayor porcentaje de estudiantes.

Las técnicas de minería de datos son una gran herramienta de ayuda a la detección de deserción, sin embargo, no sirven de nada si no se acompañan de acciones de acompañamiento a los estudiantes, las cuales deben ser apoyadas a nivel de escuelas, departamentos o la unidad respectiva de apoyo académico. Dentro de los trabajos futuros, está la definición de acciones y la medición de efectividad éstas, con el objetivo de disminuir la deserción.

REFERENCIAS

Balmori, E., De la Garza, M., Reyes, E. (2009): El Modelo De Deserción De Tinto Como Base Para La Planeación Institucional: El Caso De Dos Instituciones De Educación Superior Tecnológica.

Bustamante, M., Morales, O. (2012): Modelo predictivo de deserción universitaria.

González, L. (2005): Estudio sobre la repitencia y deserción en la educación superior chilena. Santiago.

Hurtado, P. (2015): Modelo predictivo de deserción de estudiantes de la Facultad de Ingeniería de la Universidad Central de Chile.

Técnica	Arbol de Desición		SVM		RNA		
	No Desertor	Desertor	No Desertor	Desertor	No Desertor	Desertor	Total real
No Desertor	568	30	556	42	543	55	598
Desertor	214	100	156	158	121	193	314

TABLA 4

Matrices de confusión para las 3 técnicas de minería de datos utilizadas en el modelo de deserción de primer año