Análisis Comparativo de Técnicas de Predicción para Determinar la Deserción Estudiantil: Regresión Logística vs Árboles de Decisión

Alfredo Pérez
Departamento Sistemas de Información
Universidad del Bío-Bío
Concepción, Chile
luaperez@alumnos.ubiobio.cl

Elizabeth E. Grandón
Departamento Sistemas de Información
Universidad del Bío-Bío
Concepción, Chile
egrandon@ubiobio.cl

Mónica Caniupán
Departamento Sistemas de Información
Universidad del Bío-Bío
Concepción, Chile
mcaniupan@ubiobio.cl

Gilda Vargas
Departamento de Estadística
Universidad del Bío-Bío
Concepción, Chile
gvargas@ubiobio.cl

***Abstract*— Actualmente, la detección de alumnos que puedan desertar un programa de estudios es un tema relevante para las universidades, por lo que existen esfuerzos en examinar las variables que determinan la deserción de los estudiantes. La *deserción* se define de distintas formas, sin embargo, todos los estudios convergen en que para que un alumno deserte de un plan de estudios se deben combinar varias variables. En este trabajo se presenta una comparación de indicadores de rendimiento del modelo de deserción actual de la Universidad del Bío-Bío (UBB), el cual está basado en la técnica de regresión logística y se compara con un nuevo modelo basado en árboles de decisión. El nuevo modelo se obtiene a través de metodologías de minería de datos y fue implementado a través de la herramienta SAP Predictive Analytics. Para entrenar, validar y aplicar el modelo se utilizaron datos reales de las bases de datos de la UBB. La comparación demuestra que el modelo propuesto obtiene una exactitud del 86%, una precisión del 97% con un porcentaje de error del 14% en la predicción de la deserción estudiantil, muy superior al valor actual que entrega el modelo basado en regresión logística. Posteriormente, el modelo de predicción obtenido fue optimizado considerando otras variables logrando mejorar los indicadores de predicción. Las instituciones de educación superior deberían tomar en consideración las variables que más explican el fenómeno de la deserción estudiantil para mejorar la retención de sus estudiantes.**

Keywords—Deserción estudiantil, Minería de datos, SAP, Predictive Analytics, Regresión Logística, Árboles de decisión

I. Introducción

Himmel [1] define la deserción estudiantil como el abandono prematuro de un programa de estudios antes de alcanzar el título o grado académico, y considera un tiempo suficientemente largo como para descartar la posibilidad de que el estudiante se reincorpore al sistema. La deserción estudiantil presenta consecuencias sociales que afectan no solo al estudiante, sino también a sus familias y al país en general [22]. Tal como lo señalan Grandón y Vargas [23], la deserción estudiantil se considera un indicador importante para medir la movilidad social y reflejar la contribución social que realizan las universidades. Como lo indica Zárate [24], desertar de la educación superior genera una falta de capital humano calificado que retribuya al desarrollo económico del país con mayor productividad. Según estudios realizados por la Comisión de Financiamiento Estudiantil para la Educación Superior [24], en Chile más del 50% de los estudiantes que se matriculan en la educación superior no concluyen sus estudios. Más aún, indican que cerca del 30% de los estudiantes abandonan sus estudios durante el primer año de su carrera. Estas razones hacen el estudio de la deserción en universidades un tema importante de ser abordado.

Lo anterior cobra aún mayor importancia considerando el rol social que ejercen ciertas instituciones de educación superior del país. La Universidad del Bío-Bío (UBB), por ejemplo, tiene como uno de sus objetivos aportar con la movilidad social de las personas, desarrollando las competencias necesarias para que los estudiantes puedan asumir, con mejor preparación, los desafíos que la sociedad requiere. La tasa de retención anual total de la UBB, entre los años 2013–2016 es en promedio de un 86%[1]. Por lo tanto, se hace necesario generar un mecanismo que permita determinar el riesgo de deserción de los alumnos, para aplicar medidas correctivas en forma oportuna.

El objetivo de esta investigación es comparar el actual modelo de deserción estudiantil de la UBB, basado en regresión logística con el modelo propuesto en esta investigación, basado en árboles de decisión. Para ello se revisaron modelos propuestos en la literatura y se seleccionaron las variables más importantes. Para el análisis

[1] Anuarios estadísticos de la Universidad del Bío-Bío.

978-1-5386-9233-2/18/$31.00 ©2018 IEEE

de los datos se utilizó la herramienta Predictive Analytics de SAP[2].

La investigación muestra que el nuevo modelo propuesto basado en árboles de decisión mejora la exactitud, precisión, sensibilidad y especificidad de los indicadores obtenidos del modelo de regresión logística. Además, el árbol de decisión generado se optimiza incorporando nuevas variables al modelo, mejorando aún más su rendimiento.

El resto del artículo se organiza de la siguiente forma: En la Sección II se analiza la literatura relacionada a la deserción estudiantil, minería de datos y árboles de decisión. En la Sección III se presenta el actual modelo de predicción de la UBB y el modelo de predicción propuesto en este estudio. En la Sección IV se comparan ambos modelos y finalmente se presenta la discusión y conclusiones en la Sección V.

II. Estado del Arte

A. *Deserción Estudiantil*

En los últimos 30 años, el sistema de educación superior en Chile experimento un crecimiento significativo, pasando de una matrícula de pregrado de 165 mil estudiantes, a principios de los años ochenta, a una matrícula de más de 1 millón de estudiantes el año 2012. Desde el año 1981 hasta el año 2005 se presentó un incremento exponencial de creación de instituciones universitarias, pasando de 8 a 62 universidades. En los inicios de este proceso de apertura del mercado universitario había cerca de 118.000 estudiantes en las universidades, al año 2005 la cifra se eleva sobre los 450.000. De acuerdo a cifras del consejo nacional de educación, el año 2012 cerca del 50% de quienes se matricularon en las universidades o centros de formación técnica no completaron el programa en que se matricularon[3]. Por lo tanto, la deserción es un problema de gran envergadura, el cual ha sido estudiado por diversos investigadores en Chile [23], [26-29].

Durante bastante tiempo, la deserción universitaria fue considerada como un fenómeno normal e, incluso, como una muestra de la exigencia de permanecer en las carreras y universidades. Sin embargo, hoy en día este problema se ve como un signo de ineficiencia y como un gran costo para el país, para los estudiantes y las instituciones de educación superior, lo que conlleva a convertirse en un problema que hay que comprender para poder combatirlo. La Tabla I presenta la tasa de deserción (promedio de deserción anual) de la UBB, de los campus de Chillán y Concepción organizada por facultad, para el período 2013-2016. Como se puede apreciar en la Tabla I, existen variaciones en los índices de deserción en los diferentes años, por lo que se hace imprescindible conocer los determinantes que hace que estos indicadores varíen. Del mismo modo, se necesita detectar en forma temprana los posibles alumnos desertores para poder trabajar o integrarlos a estos alumnos a algún programa de tutoría focalizado, con el objetivo de disminuir los índices de deserción.

Los diversos estudios de la literatura establecen que las variables asociadas a la deserción están relacionadas con el ámbito académico, socioeconómico, institucionales y personales. Las causas más determinantes en la deserción de estudiantes en primer año universitario son: problemas vocacionales, situación económica familiar, rendimiento académico, entre otros. Por otro lado, existen diversos modelos que explican la deserción estudiantil.

TABLE I. Tasa de deserción de alumnos de primer año en la UBB entre los años 2013 al 2016.

Facultad	Porcentajes de deserción por años			
	2013	**2014**	**2015**	**2016**
Arquitectura	14,72%	26,56%	22,22%	14,29%
Ciencias	22,58%	26,67%	14,85%	22,22%
Educación y Humanidades	7,86%	12,90%	9,78%	7,61%
Ingeniería	11,56%	12,65%	21,10%	13,97%
Ciencias de la Salud y de los Alimentos	8,96%	7,32%	10,56%	12,02%
Ciencias Empresariales	9,46%	9,31%	8,71%	12,86%
Promedio UBB	**12,52%**	**15,90%**	**14,54%**	**13,83%**

De acuerdo a Braxton et al. [2], existen cinco grandes enfoques para analizar la deserción y retención, dependiendo del énfasis que se les otorga a las variables explicativas, ya sea individuales, institucionales o del medio familiar: (i) Enfoque Psicológico: orientado a los rasgos de personalidad que diferencian a los individuos, caracteriza el comportamiento, actitudes, desarrollo intelectual, intenciones conductuales, entre otras variables. (ii) Económico: distingue la sustentabilidad económica del individuo, como también el costo y beneficio que se obtienen de la institución académica. (iii) Sociológico: enfatiza la influencia de factores externos, que se suman a los psicológicos, como por ejemplo la falta de integración y el ambiente familiar, los cuales influencian positiva o negativamente al individuo. (iv) Organizacionales: analiza la deserción desde el punto de vista de la institución universitaria, analizando los servicios que ésta ofrece a los estudiantes, como la calidad de educación que se entrega, beneficios, organizaciones y recursos disponibles. (v) Interacciones: analiza la interacción del individuo y la institución como organización. Esta interacción es el significado que cada estudiante le atribuye a su relación con las dimensiones formales e informales de la organización académica, como por ejemplo, las experiencias durante la permanencia del individuo en la institución.

[2] https://www.sap.com/latinamerica/products/predictive-analytics.html

[3] https://centroestudios.mineduc.cl/wp-content/uploads/sites/100/2017/06/A2N15_Desercion_escolar.pdf

Fishbein y Ajzen [3] manifiestan que las intenciones de una persona son el resultado de sus creencias, las que influyen en sus actitudes y llevan a manifestar su comportamiento. De acuerdo a esto, se puede considerar que la deserción es el resultado de un debilitamiento de las intenciones iniciales de continuar estudios secundarios mientras que la persistencia es el fortalecimiento de las mismas. Esta concepción puede explicar tanto la deserción voluntaria, como la retención en un programa de estudio.

Ethington [4] propone un modelo que incorpora la teoría sobre las "conductas de logro" de Eccles [5], que incluyen como atributos la perseverancia, la elección y el desempeño. Una de las premisas de Eccles es que el rendimiento académico previo de un estudiante influencia el desempeño futuro [24]. Además, establece que el apoyo y el estímulo que un estudiante recibe de su familia a su vez inciden sobre su nivel de aspiraciones. Ethington empíricamente muestra que el nivel de aspiraciones tiene un efecto directo sobre los valores, observando que las expectativas de éxito estaban explicadas por el autoconcepto académico y la percepción de las dificultades de los estudios. Además, establece que las expectativas de éxito de los estudiantes influyen sobre la persistencia en la universidad.

Por otra parte, el modelo propuesto por Spady [6] enfatiza la influencia de factores externos al individuo que influyen en la retención estudiantil. Spady basa su teoría en el suicidio de Durkheim [7], que corresponde al resultado de la ruptura del individuo con el sistema social por su imposibilidad de integrarse a la sociedad. Spady sostiene que el resultado de la falta de integración de los estudiantes en el entorno de la educación superior, así como también, el medio familiar son una de las fuentes que expone a los estudiantes a influencias, expectativas y demandas, las que a su vez afectan su nivel de integración social en la universidad. Las influencias negativas implican un rendimiento académico insatisfactorio, un bajo nivel de integración social, satisfacción y compromiso institucional, y por lo tanto, una alta posibilidad de que el estudiante decida abandonar sus estudios. Por el otro lado, si los efectos o influencias son positivas y congruentes con la situación inicial, el estudiante logra un desarrollo académico y social acorde con sus expectativas y las de la institución, entonces permanece en el programa hasta lograr su meta. Spady además postula que la percepción del estudiante de su capacidad o incapacidad de solventar los costos asociados a los estudios universitarios es también un factor importante en la decisión de desertar.

Tinto [8] expande el modelo de Spady [6], incorporando la teoría del intercambio que radica en el principio de que los seres humanos evitan las conductas que implican algún tipo de costo para ellos y buscan recompensas en las relaciones, interacciones y estados emocionales. Tinto afirma que los estudiantes que actúan de acuerdo a dicha teoría en la construcción de su integración social y académica desertarán si los beneficios de permanecer en la institución académica son percibidos como mayores que los costos personales. Su modelo considera que, a medida que el alumno evoluciona a través del programa de estudios, diversas variables contribuyen a reforzar su adaptación dentro de la institución. Estas variables abarcan los antecedentes socioeconómicos y culturales a nivel familiar, los valores que la familia sustenta, los atributos personales del estudiante y su experiencia académica preuniversitaria. A su vez incluye en su modelo la integración académica tanto en el rendimiento, como en el desarrollo intelectual que tenga el estudiante.

Finalmente, Bean [9-11] incorpora al modelo de Tinto [8] características del modelo de productividad en el contexto las organizaciones laborales. Señala que la deserción universitaria es similar a la productividad y destaca la importancia de las intenciones conductuales de permanecer o abandonar los estudios superiores. Además, postula que las creencias y aptitudes son afectadas por diferentes componentes de la institución universitaria, tales como, la calidad de cursos, programas, docentes y pares que otorga dicha institución. Al mismo tiempo destaca que la satisfacción con los estudios es similar a la satisfacción con el trabajo, lo cual influye directamente en las intenciones de continuar o abandonar los estudios universitarios.

B. Minería de Datos y Árboles de Decisión

La minería de datos (o Data Mining, en inglés) permite obtener patrones en grandes bases de datos de manera automática y generar un modelo predictivo o descriptivo de los datos [12][13]. La minería de datos utiliza diversas técnicas para obtener patrones, entre las cuales se encuentran los árboles de decisión, que son una de las técnicas más utilizadas [18].

Los árboles de decisión permiten generar modelos predictivos y descriptivos [17]. De acuerdo a Song y Ying [19] esta técnica no paramétrica clasifica una población en un modelo de segmentos de tipo ramas que construyen un árbol invertido y de esta forma predice una variable objetivo. Este árbol se recorre desde la raíz hasta las hojas siguiendo las condiciones especificadas en cada uno de los nodos intermedios. Se utiliza principalmente para labores de clasificación y agrupamiento de datos. Una de las ventajas de esta técnica es que puede manejar eficientemente grandes y complejos conjuntos de datos [19]. Si bien esta investigación utiliza árboles de decisión, el estudio puede ser replicado utilizando otras técnicas de predicción tales como redes neuronales, máquinas de vectores de soporte (SVM por sus siglas en inglés), u otras [12].

Existen diversas herramientas de minería de datos, tanto de software libre, por ejemplo WEKA [4], como de software propietario, por ejemplo SAP Predictive Analytics, el cual fue elegido en este estudio. La empresa SAP distribuye uno de los sistemas de planificación de recursos empresariales (ERP por sus siglas en inglés) más utilizados por las instituciones [21], lo

[4] https://www.cs.waikato.ac.nz/ml/weka/

que trae consigo la generación de muchos datos transaccionales que pueden servir de base para análisis predictivos de distinta naturaleza.

C. SAP Predictive Analytics

SAP PA es un software para realizar análisis de datos que además de estar basado en el lenguaje de programación R[5], integra algoritmos predictivos, minería de datos y aprendizaje automático. SAP PA permite crear modelos predictivos con el fin de descubrir información oculta en los datos y de esta manera poder hacer predicciones sobre eventos futuros. R es un entorno y lenguaje de programación de código libre y uno de los más utilizados por la comunidad estadística en el campo de la minería de datos [14]. La ventaja de SAP PA de incorporar R es que se puede crear componentes a medida en el caso que los algoritmos estándar no se ajusten al problema en cuestión. En esos casos la herramienta admite nuevos desarrollos en SQLScript y R.

Dentro de SAP PA, se cuenta con dos modos:

- Automático: para un usuario que realiza minería de datos exploratoria, masiva y automática, dejando al software orientar las variables, y modificaciones en el algoritmo.
- Experto: para un usuario que realiza un tipo de minería de datos más compleja, manual, extensible, que manipula la caracterización de los datos y la manipulación de los modelos predictivos.

SAP entrega su propia metodología de trabajo para tratar y crear los modelos de predicción. En el modo automático propone las siguientes fases:

1. **Carga de datos**: se determinan las fuentes de datos y el tipo de información a utilizar. Es la fase donde los datos relevantes para el análisis son extraídos desde las fuentes de datos, luego son unificados y cargados a SAP PA.
2. **Preparación de datos**: se estandarizan los datos dependiendo de las necesidades y criterios. Se modifican datos que necesitan ser caracterizados o sustituidos y se clasifica las variables.
3. **Procesamiento de datos**: es una de las fases más delicadas debido a que el usuario se interioriza en el algoritmo que aplicará, sus modalidades y atributos a modificar, para que de esta manera se pueda obtener un resultado con mayor veracidad y este modelo sea aplicable. Se puede usar una combinación de algoritmos para analizar los datos.
4. **Visualización**: en esta etapa, con las variables y resultados ya obtenidos de los modelos, se procede a automatizar de forma que se puedan mostrar los resultados para su precisa interpretación.

[5] https://www.r-project.org/

III. Modelos de Deserción Estudiantial de la UBB

A. Modelo Actual

Retamal y Rubilar [15] propusieron un modelo de deserción, el cual se basa en el método ScoreCard, aplicando la estimación de parámetros a través de un modelo de regresión logística [16]. El ScoreCard es un método estadístico ampliamente usado en las empresas como bancos y/o de retail para otorgar créditos o préstamos a sus clientes, mediante la clasificación de los individuos, por algún criterio. Por ejemplo, clasificar a los clientes considerando el riesgo de otorgar un crédito en clientes “aptos” y “no-aptos” [16]. La población total en que se basó el estudio consistió en 4519 estudiantes de las cohortes 2014 y 2015. Los autores consideraron inicialmente 31 variables las cuales, luego de un análisis de los indicadores, valor de la información (IV) y peso de la evidencia (WoE) para obtener el poder de predicción de cada atributo, se redujeron a 19 variables para ser ingresadas al modelo de predicción. Finalmente se seleccionaron 10 variables para el modelo.

Se elaboró el ScoreCard mediante puntajes asignados a los atributos de cada variable y se calcularon en función de las estimaciones de los parámetros de la regresión logística. De esta forma se pudo obtener la puntuación asociada a los estudiantes dada sus características de ingreso y así determinar qué alumnos presentan un alto riesgo de desertar. Para medir la eficiencia del ScoreCard [15], se utilizó el índice de Gini y la prueba de Kolmogorov-Smirnov. Al mismo tiempo se determinaron los puntos de corte para poder clasificar a los estudiantes en desertores, posibles desertores, y no desertores.

Se desarrollaron cinco modelos experimentales de los cuales se seleccionó el que presentó mejores índices de bondad de ajuste y capacidad predictiva. El modelo pudo estimar con exactitud el 60% de los casos, con una especificidad de un 69%, una sensibilidad de 58% y una precisión del 90%. Los autores sugieren investigar nuevas variables que puedan influir en la deserción y buscar nuevas técnicas para el desarrollo de modelo de deserción universitaria. Considerando lo anterior, y debido al gran volumen de datos que hoy en día se maneja en la UBB se hace necesario investigar nuevas técnicas basadas en herramientas de inteligencia de negocios (BI por sus siglas en inglés), con el fin de aumentar la capacidad predictiva de los modelos.

B. Modelo de Deserción Predictivo basado en Árboles de Decisión

El modelo predictivo propuesto se basa en árboles de decisión, implementado en SAP PA 3.3. Para la construcción del modelo, y con el fin de comparar los resultados del modelo actual basado en la técnica de regresión logística, se consideró la misma población, compuesta por 4.519 registros, correspondientes a alumnos de las cohortes 2014 y 2015. Estos datos fueron obtenidos de la Dirección de Admisión y Registro Académico, Dirección General de Análisis Institucional y Dirección de Desarrollo Estudiantil de la UBB.

Se utilizó el motor de base de datos MySql[6] versión 5.6.24 para almacenar y manipular los diferentes conjuntos de datos. El 82,5% de los estudiantes no desertó durante la fecha de desempeño definida, lo cual corresponde a 3.726 estudiantes, mientras que el 17,5% si lo hizo (793 estudiantes).

La conceptualización de los alumnos desertores se basó en los siguientes criterios: aquellos que ingresaron a una carrera y al año inmediatamente posterior se inscribieron en una carrera distinta, aquellos que efectuaron un retiro definitivo, tuvieron pérdida de carrera, no inscribieron asignaturas en los 3 semestres subsiguientes por causas distintas al retiro temporal y aquellos que no inscribieron asignaturas pero porque efectuaron un retiro temporal (se desconoce si volverán a estudiar en semestres subsiguientes).

Como resultado de la revisión bibliográfica y de acuerdo a la disponibilidad de los datos, se definieron las variables de la Tabla II, que formaron parte del modelo predictivo, donde deserción se consideró como la variable predictiva objetivo.

TABLE II. VARIABLES DEL MODELO PREDICTIVO

Atributo	**Características**	
	Tipo de dato	***Valores***
Deserción	Booleano	0: Retenido 1: Desertado
Genero	Entero	1: Masculino. 2: Femenino
Edad	Entero	17-34
tipo_colegio	Entero	1-5
tipo_enseñanza	Entero	1-3
psu_mate	Entero	320-850
psu_leng	Entero	320-850
Nem	Entero	320-850
ranking	Entero	320-850
Bea	Booleano	0-1
orden_postulacion	Entero	1-8
años_acreditacion	Entero	0-5
becas_internas	Booleano	0-1
becas_arancel_midenduc	Booleano	0-1
Sede	Texto	Chillán-Concepción
gratuidad	Booleano	0-1
Hijos	Entero	0: ninguno 1-5
carrera	Numero	Código

El algoritmo de árboles de decisión de SAP PA genera un modelo de minería de datos mediante la creación de una serie de divisiones en el árbol. Estas divisiones se representan como nodos. El algoritmo agrega un nodo al modelo cada vez que una columna de entrada tiene una correlación significativa con la columna o variable de predicción. La forma en que el algoritmo determina una división varia si predice una variable del tipo continua o una variable del tipo discreta. El algoritmo utilizado fue el R-CNR TREE el cual realiza automáticamente un análisis de minería de datos para seleccionar características que aporten o contribuyan de mayor manera a mejorar el rendimiento y la calidad del análisis. Es importante evitar atributos irrelevantes debido a que puede aumentar significativamente los tiempos de ejecución y procesamiento.

La Tabla III indica los parámetros utilizados para la configuración del algoritmo R-CNR TREE. Para realizar el análisis, la población se dividió en tres submuestras. La primera con el 80% de los registros la cual sirvió para determinar la clasificación basada en el árbol de decisión, la segunda con el 10% de los registros se utilizó para probar la capacidad de predicción del modelo y la tercera con el 10% restante para la realización del test.

TABLE III. PARÁMETROS DEL ALGORITMO R-CNR TREE

	Atributo	**Valor**
Propiedades	Output Mode	Tendencia
	Algorithm Type	Clasificación
	Target Variable	Deserción
	Predicted Column Name	Pred_Deserción
Avanzado	Missing Values	Omitir
	Minimum Split	4
	Complexity Parameter	0.001
	Maximum Depth	Null
	Split Criteria	Gini
	Cross Validation	Null
	Prior Probability	Null
	Use Surrogate	No aplica
	Surrogate Style	No aplica
	Maximum Surrogate	No aplica

La Fig. 1 muestra el árbol de decisión obtenido:

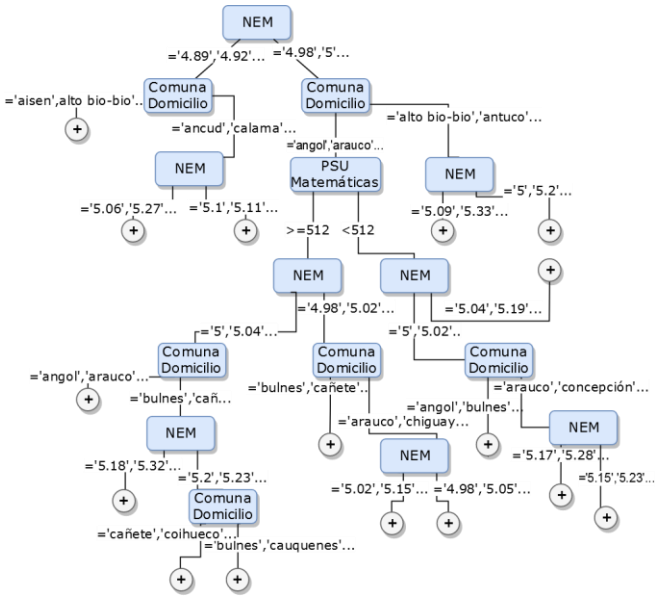

Fig. 1: Árbol de decisión obtenido con SAP PA

[6] https://www.mysql.com/

Para medir el nivel de la predicción realizada se utilizó una tabla de contingencia que describe los casos acertados y los casos erróneos de la predicción. A esta tabla de contingencia se la llama matriz de confusión y contiene información acerca de las clasificaciones reales y aquellas estimadas por el árbol [20]. De este modo, en cada columna de la matriz se muestra el número de predicciones de cada clase, mientras que en cada fila se muestran las instancias de la clase real. El ratio de precisión de la predicción se calcula dividiendo la suma de los aciertos de clasificación por el total de casos.

La Tabla IV presenta la matriz de confusión obtenida después de correr el algoritmo.

TABLE IV. MATRIZ DE CONFUSIÓN

		Predicción de Deserción		
		Sí	**No**	**Total**
Deserción Real	**Sí**	**3621** Verdaderos positivos (VP)	**541** Falsos Negativos (FN)	**4162** Total Positivos
	No	**105** Falsos Positivos (FP)	**252** Verdaderos Negativos (VN)	**357** Total Negativos
Total		**3726** Total clasificados positivos	**793** Total clasificados negativos	**4519**

IV. COMPARACIÓN DE MODELOS Y DISCUSIÓN

La Tabla V muestra un resumen de las descripciones de las técnicas utilizadas para la comparación: regresión logística y árboles de decisión.

TABLE V. REGRESIÓN LOGÍSTICA VS ARBOL DE DECISIÓN

Descripción	**Regresión Logística**	**Árbol de Decisión**
Cantidad de Variables	10	10
Algoritmo utilizado	Regresión Logística	Árbol de Decisión
Variable Categórica Objetivo	Retenido o Desertado	Retenido o Desertado
Agrupación de contenido a través de criterios	Criterio de agrupación WoE	No se utilizó criterio de agrupamiento en variables
Software Utilizado	R Project Estadístico	SAP Predictive Analytics

La comparación de los modelos se realizó mediante un análisis de rendimiento obtenido de la matriz de confusión, dado que es una métrica común obtenida tanto del análisis de regresión como de los árboles de decisión. La Tabla VI muestra las variables utilizadas en los distintos algoritmos organizados por sus contribuciones dentro del modelo de regresión, en una escala al 0-1 (100%).

TABLE VI. CONTRIBUCIÓN DE CADA VARIABLE EN EL MODELO

Variables	**Contribución Modelo Regresión**	**Contribución Árbol de Decisión**
Código Carrera	0.38	0.04
PSU Matemáticas	0.19	0.07
NEM	0.10	0.45
Orden de Postulación	0.05	0.01
Becas Arancel MINEDUC	0.05	0.01
Vive año actual	0.04	0.04
Motivación trabajar	0.04	0.02
Genero	0.02	0.01
Comuna domicilio	0.02	0.35
Gratuidad	0.00	0.00

A continuación se presentan algunos indicadores [25] obtenidos desde la matriz de confusión para cada modelo a evaluar.

TABLE VII. INDICADORES DEL MODELO

Evaluación	**Modelo de Regresión**	**Árbol de Decisión**
Exactitud (VP+VN)/Total	60%	86%
Tasa de error (FP+FN)/Total	40%	14%
Sensibilidad (VP/Total Positivos)	58%	87%
Especificidad (VN/Total Negativos)	69%	70%
Precisión (VP/Total clasificados positivos)	90%	97%
Predicción negativa (VN/Total clasificados negativos)	25%	32%

Como se puede apreciar, en todos los casos, los valores obtenidos por la técnica de árbol de decisión son superiores a los resultados obtenidos por la regresión logística. Cabe destacar la exactitud del árbol de decisión que alcanza un 86% comparada con un 60% obtenida con el uso de regresión logística. De manera similar, la precisión obtenida es 97% a través del árbol de decisión y de un 90% a través de la regresión logística, mientras que la especificidad aumenta de un 69% a 70%.

La Tabla VIII muestra una comparación de los indicadores especificidad, sensibilidad y AUC el cual genera un trazado de la sensibilidad (tasa de verdaderos positivos) en varios umbrales versus la especificidad (tasa de falsos negativos) [25].

TABLE VIII. COMPARACIÓN DE INDICADORES

Modelo	**AUC**	**Especificidad**	**Sensibilidad**
Regresión Logística	0.735	69%	58%
Árbol de decisión	0.661	70%	87%

Como se puede apreciar, ambos valores de AUC son superiores a 0.5, valor mínimo aceptado de acuerdo a [25]. A pesar de obtener un valor AUC menor que el modelo de regresión logística, el modelo de árbol de decisión propuesto presenta una mayor especificidad y sensibilidad, lo que indica que el modelo clasifica mejor a los estudiantes desertores y no desertores.

A. *Modelo de Árbol de Decisión Optimizado*

Finalmente, el modelo de predicción obtenido fue optimizado considerando otras variables sugeridas en la literatura (PSU Lenguaje, ranking, edad, duración de la carrera, entre otras). De esta forma se ejecutó SAP PA considerando el algoritmo R-CNR Tree de árbol de decisión para el modelo de réplica y optimizado, como también un comparador de modelos. La partición de la muestra correspondió a un 80% para el entrenamiento, un 10% para la validación del modelo y un 10% para el test. La Fig. 2 muestra el diseñador de SAP PA.

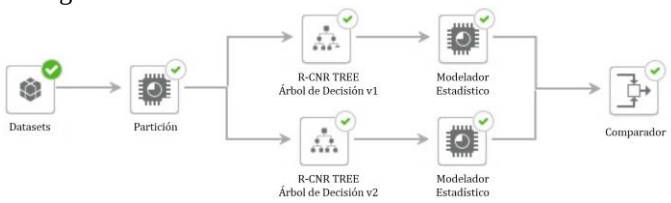

Fig. 2: Diseñador de SAP PA

La Tabla IX muestra las variables consideradas junto a su contribución al modelo y la Tabla X presenta la matriz de comparación de resultados de los indicadores. Como se puede apreciar, a pesar de que la precisión baja en un 1%, la exactitud del modelo aumenta a un 93% y la especificidad a un 82%, bajando la tasa de error a un 7%.

TABLE IX. VARIABLES Y SU CONTRIBUCIÓN AL MODELO

Árbol de Decisión (Réplica)	**Contribución**	**Árbol de Decisión Optimizado**	**Contribución**
Código Carrera	0.04	NEM	0.25
PSU Matemáticas	0.07	PSU Matemáticas	0.10
NEM	0.45	PSU Lenguaje	0.12
Orden de Postulación	0.01	Ranking	0.10
Becas Arancel MINEDUC	0.01	Código Carrera	0.08
Vive año actual	0.04	Edad	0.05
Motivación trabajar	0.02	Duración de carrera	0.05
Genero	0.01	Tipo de enseñanza	0.05
Comuna domicilio	0.35	Orden de postulación	0.03
Gratuidad	0.00	Años de acreditación	0.04
		Tipo de colegio	0.03
		Genero	0.02
		Gratuidad	0.01
		Sede	0.02
		Becas Arancel MINEDUC	0.02
		BEA	0.00
		Discapacidad	0.00

TABLE X. COMPARACIÓN DE INDICADORES

Evaluación	**Árbol de Decisión (Réplica)**	**Árbol de Decisión Optimizado**
Exactitud	86%	93%
Tasa de error	14%	7%
Sensibilidad	87%	95%
Especificidad	70%	82%
Precisión	97%	96%
Predicción negativa	32%	75%

V. DISCUSIÓN Y CONCLUSIONES

La presente investigación tuvo por objetivo comparar dos técnicas de predicción para la deserción estudiantil, regresión logística y árboles de decisión. Las variables que más explicaron el fenómeno de la deserción utilizando los algoritmos de regresión logística fueron la carrera a la cual está adscrito el alumno, el puntaje PSU de matemáticas y las notas de enseñanza media (NEM), mientras que las que más explicaron el fenómeno utilizando los algoritmos de árbol de decisión fueron NEM, la comuna de domicilio del alumno y el puntaje PSU de matemáticas, en ese orden respectivamente. Algunas de las variables que tienen una mínima contribución en ambos modelos son el orden de postulación, las becas de arancel y la motivación por trabajar. Coincidentemente, la gratuidad resultó no tener ninguna incidencia en la predicción de la deserción estudiantil cuando se comparan ambas técnicas. En general, los indicadores obtenidos reflejan que la técnica de árboles de decisión predice la deserción estudiantil con una mayor exactitud y precisión y con menor porcentaje de error que la técnica de regresión logística.

La optimización del modelo de árbol de decisión a través de la incorporación de nuevas variables, permitió validar los resultados. Es destacable mencionar el comportamiento de la variable NEM, una de las con mayor incidencia en ambos modelos, estando valorizada sobre un 25%, siendo el nodo padre tanto en el árbol de réplica como en el árbol optimizado. Al analizar los resultados de la matriz de confusión es posible ver la especificidad aumentada de un 70% en el modelo de réplica a un 82% en el modelo optimizado, aumentando de esta manera a un 82% de alumnos clasificados correctamente como desertores del total de la muestra. La sensibilidad aumentó de un 87% a un 95% para los casos de los alumnos no desertores. La optimización del modelo se consigue al agregar 7 variables más, lo que involucra un mayor costo tanto en la recolección de datos como en el procesamiento, por lo cual las instituciones de educación superior deberían evaluar si están dispuestas a invertir más recursos en la adquisición de más variables en pos de una mejor exactitud, sensibilidad y especificidad del modelo ya que su precisión se mantiene más o menos constante. El modelo obtenido puede ser utilizado por las instituciones de educación superior para centrar sus esfuerzos en las variables que más explican el fenómeno de deserción, incorporando programas de seguimiento y

acompañamiento a estudiantes con bajo NEM y que provengan de comunas aledañas y rurales de la región.

Una limitación del estudio corresponde a la no incorporación en los modelos de ciertas variables encontradas importantes en la literatura ya que no se encontraban disponibles. Por ejemplo, variables psicológicas, rendimiento académico del semestre y si el alumno posee problemas personales que se encuentren afectando su rendimiento. Futuros estudios podrían considerar este tipo de variables para ver cómo se comportan en este tipo de modelos.

Otras posibles investigaciones podrían considerar crear nuevos métodos o funciones obtenidas de R que ayuden a mejorar la calidad y precisión del modelo. Además, se podría utilizar otros modelos de clasificación como redes neuronales, clústeres o redes bayesianas. Finalmente se podría crear una evaluación de algoritmos utilizando un conjunto de datos conformado por distintos clasificadores y analizar con cuales se obtiene mejores resultados.

REFERENCIAS

[1] E. Himmel, "Models of analysis of student desertion in higher", Universidad Católica de Chile, 2002.

[2] J. Braxton, A. Sullivan, and R. Johnson, "Appraising Tinto's theory of college student departure," in J.C. Smart (Ed.), Higher education: Handbook of theory and research, (vol. XII), Agathon Press, 1997.

[3] M. Fishbein, and I. Ajzen, "Belief, attitude, intention and behavior: An Introduction to theory and Research", Addison-Wesley, 1975.

[4] C. Ethington, "A psychological model of student persistence", Research in Higher Education, 31(3), 1990, pp. 266-269.

[5] J. Eccles, "Expectancies, values and academic behaviors". J. T. Spence (Ed.), Achievement and achievement motives: Psychological and sociological approaches, 1983, pp. 75-146.

[6] W. Spady, "Dropouts from higher education: An interdisciplinary review and synthesis", Interchange, 1(1), 1970, pp. 64-85.

[7] E. Durkheim, "Suicide: A Study in Sociology", Routledge Classics, Volume 117, 2002.

[8] V. Tinto, "Dropout in higher education: A theoretical synthesis of recent research", Review of Educational Research, 45, (1), 1975, pp. 89-125.

[9] J. Bean, "The application of model of turnover in work organizations of the student attrition process", Review of Higher Education, 6(2), 1983, pp. 129-148.

[10] J. Bean, "Integration effects based on class level in an explanatory model of college student dropout syndrome", American Education research Journal, 22(1), 1985, pp. 35-64.

[11] J. Bean, "Dropout and turnover: The synthesis and test of a casual model of student attrition", Research in Higher Education, 12(4), 1986, pp. 155-187.

[12] J. Han, and M. Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann, 2001.

[13] D. Larose, "Discovering Knowledge in Data: An introduction to data mining", John.Wiley & Sons, 2004.

[14] E. Paradis, "R for beginners", University of Hawaii, 2003, pp. 4-16.

[15] J. Retamal, and A. Rubilar, "Desarrollo de un modelo de deserción para los estudiantes de primer año de la Universidad del Bío-Bío", tesis de pregrado, Universidad del Bío-Bío, 2017.

[16] S. Rayo, J. Lara, and D. Camino, "A Credit Scoring Model for Institutions of Microfinance under the Basel Ii Normative", Journal of Economics, Finance and Administrative Science, 15(28), 2010, pp. 89-124.

[17] L. Rokach and O. Maimon, "Data mining with decision trees: theory and applications", World Scientific, 2014.

[18] M. Hofmann and R. Klinkenberg, "RapidMiner: Data mining use cases and business analytics applications", CRC Press, 2013.

[19] Y. Song and L. Ying, "Decision tree methods: applications for classification and prediction" Shanghai Archives of Psychiatry, vol. 27, issue 2, 2015, pp. 130-135.

[20] P. Ramirez-Correa and E. Grandón, "Predicción de la deserción académica a través de la clasificación basada en árboles de decisión con parámetros optimizados", Formación Universitaria, 11(3), 2018, pp. 3-10.

[21] M. Pinto, P. Ramírez-Correa and E. Grandón, "Antecedentes del éxito de los ERP en las grandes empresas chilenas: Un modelo factorial exploratorio", Información Tecnológica, 3(28), 2017, pp. 139-146.

[22] L. González, "Estudio sobre la repitencia y deserción en la educación superior chilena", Digital observatory for higher education in Latin America and The Caribbean, IESALC – UNESCO, 2005.

[23] E. Grandón and G. Vargas, "Movilidad social intergeneracional: Una mirada de la contribución social de las universidades. El caso de los titulados de la Universidad del Bío-Bío, Chile." Centro Interuniversitario de desarrollo CINDA, Colección Gestión Universitaria, 2012, pp. 275-294.

[24] R. Zárate and E. Pinilla, "La deserción estudiantil UIS, una mirada desde la responsabilidad social universitaria, Zona Próxima, 21, 2014, pp. 120-133.

[25] T. Fawcett, "An introduction to ROC analysis", Pattern Recognition Letters 27, 2006, 861–874.

[26] M. Caniupán, and G. Toro, "Aplicación del proceso data mining en el área docente.", Proceedings of the 1st Ibero-American Symposium on Software Engineering and Knowledge Engineering (JIISIC´01), June 13-15, Buenos Aires, Argentina, 2001, pp. 121-126.

[27] C. Gutiérrez-Soto, P. Oliva, A. Paredes, and P. Campos, "Una Aplicación de la Minería de Datos en la Educación Superior". Actas de las Jornadas Chilenas de ciencias de Computación, 10-15 Noviembre, Punta Arenas, Chile, 2008.

[28] C. Díaz, "Modelo conceptual para la deserción estudiantil universitaria chilena", Estudios Pedagógicos XXXIV, 2, 2008, pp. 65-86.

[29] M. A. Miranda, and J. Guzmán, "Análisis de la Deserción de Estudiantes Universitarios usando Técnicas de Minería de Datos", Formación Universitaria, 10(3), 2017, pp. 61-68