Predicción de la Deserción Académica en una Universidad Pública Chilena a través de la Clasificación basada en Árboles de Decisión con Parámetros Optimizados

Patricio E. Ramírez^{(1)*} y Elizabeth E. Grandón⁽²⁾

(1) Universidad Católica del Norte, Escuela de Ingeniería, Larrondo 1281, Coquimbo-Chile.

(e-mail: patricio.ramirez@ucn.cl)

(2) Universidad del Bío-Bío, Departamento de Sistemas de Información, Avenida Collao 1202, Concepción, Chile. (e-mail: egrandon@ubiobio.cl)

* Autor a quien debe ser dirigida la correspondencia

Recibido Oct. 25, 2017; Aceptado Dic. 5, 2017; Versión final Ene. 26, 2018, Publicado Jun. 2018

Resumen

El objetivo de este estudio es presentar una clasificación basada en árboles de decisión (CBAD) con parámetros optimizados para predecir la deserción de los estudiantes universitarios. El estudio analiza 5288 casos de estudiantes pertenecientes a una universidad pública chilena. Para la técnica CBAD se optimizaron los parámetros para mejorar la predicción usando el software RapidMiner. El resultado de la aplicación de esta técnica con parámetros optimizados logra una razón de precisión de un 87.27%. Se concluye que el uso de técnicas de CBAD con optimización de parámetros resulta en una mejor precisión en comparación a otras investigaciones con un número similar de datos.

Palabras clave: deserción estudiantil; clasificación basada en árboles de decisión; optimización

Prediction of Student Dropout in a Chilean Public University through Classification based on Decision Trees with Optimized Parameters

Abstract

The aim of this study is to present a classification based on decision trees (DTBC) with optimized parameters to predict the dropout rate of university students. The study analyses 5288 cases of students belonging to a Chilean public university. For the CBAD technique, the parameters were optimized to improve the prediction using the software RapidMiner. The result of the application of this technique with optimized parameters achieved a precision rate of 87.27%. It is concluded that the use of DTBC technique with parameter optimization results in a better precision compared to other research with similar number of data.

Keywords: student dropout; classification based on decision trees; optimization

INTRODUCCIÓN

Tinto (1989) define la deserción universitaria como el fracaso para alcanzar una meta deseada en pos de la cual un sujeto ingresó a una institución de educación superior. La deserción universitaria genera impactos sociales y económicos en un país. González (2005) detalla las consecuencias sociales que implica la deserción, entre las que se encuentran las expectativas de los estudiantes y sus familias y las consecuencias emocionales sobre los estudiantes. La deserción estudiantil se considera un indicador importante para medir la movilidad social y reflejar la contribución social que realizan las universidades (Grandón y Vargas, 2012). En términos económicos, existe evidencia de que los estudiantes atribuyen su decisión de desertar de sus programas académicos a su situación económica (Acevedo et al., 2015). La deserción provoca importantes brechas salariales entre las personas que terminan sus estudios terciarios con respecto a los que no los terminan (Díaz, 2008; Kovačić, 2010), lo cual provoca una falta de capital humano calificado que retribuya con una mayor productividad al desarrollo económico de un país (Zárate y Mantilla, 2015).

Si bien la deserción es un fenómeno presente en todo sistema educativo (Eckert y Suénaga, 2015; Díaz, 2008), en las universidades latinoamericanas los recientes aumentos en cobertura han generado bajas tasas de retención (Ramírez, 2016). En Chile, por ejemplo, más del 50% de los estudiantes que se matriculan en la educación superior no concluyen el programa en el que se inscribieron inicialmente (Comisión de Financiamiento Estudiantil para la Educación Superior, 2012). Particularmente, se establece que en el país más de un 30% de los estudiantes dejan los estudios en el transcurso del primer año de educación superior. Se constata además que el financiamiento fiscal que reciben las universidades chilenas impacta la calidad del cuerpo docente lo que a su vez influye en la retención de los estudiantes universitarios (Araneda-Guirrman et al., 2013). Este tema constituye una de las grandes preocupaciones del sistema de Educación Superior en Chile, y del Ministerio de Educación en particular (SIES, 2014). Por este motivo las instituciones de educación superior realizan esfuerzos para determinar los factores que influyen en la deserción estudiantil y estudiar formas de predecir la deserción de manera de tomar medidas correctivas y oportunas.

En este contexto, este estudio tiene por objetivo presentar una clasificación basada en árboles de decisión con parámetros optimizados para predecir la deserción de los estudiantes universitarios. El restante texto del artículo se organiza como sigue. Primero se presenta una revisión de la literatura en relación a la deserción estudiantil y a la técnica de clasificación basada en árboles de decisión. A continuación, en las secciones materiales y métodos, se detalla la fuente de información y la herramienta utilizada para el análisis. En la sección resultados se presenta el proceso de optimización y la aplicación de la clasificación basada en árboles de decisión. Seguidamente se entrega la discusión de los resultados y, finalmente, se presentan las conclusiones del estudio.

REVISIÓN DE LA LITERATURA

A continuación se presenta una revisión del fenómeno de deserción estudiantil en la educación terciaria, seguido de una exposición de la clasificación basada en árboles de decisión y su utilización para predecir dicha deserción.

Deserción estudiantil

El fenómeno de la deserción estudiantil en la educación terciaria se ha estudiado principalmente utilizando dos líneas de investigación. Una de ellas corresponde a la aplicación de modelos teóricos que explican, entre otros, los factores que influyen en la intención de desertar de los programas académicos. Por ejemplo, Fichten et al. (2016) utilizó la teoría del comportamiento planeado (TPB – Azjen, 1991) para predecir la graduación y deserción de estudiantes universitarios en Canadá. Otros modelos, dentro de esta misma línea, incorporan factores externos a los percibidos por el individuo, como es el caso del modelo de Spady (1970) que, con un enfoque sociológico, considera factores relacionados al medio familiar como antecedente de la permanencia del estudiante en la universidad. Incluyendo además elementos organizacionales, el modelo de Tinto (1975) ha sentado las bases para estudios posteriores que incluyen las experiencias institucionales, tales como la incorporación de actividades extracurriculares y las interacciones entre los estudiantes, como decisoras al momento de explicar el fenómeno de la deserción estudiantil. Este es el caso de Bean (1985) quien, en base a la teoría de Tinto (1975), propuso un modelo integrado que explica la deserción. Entre los factores considerados por Bean se encuentran los académicos, psicosociales, ambientales y de socialización.

La otra línea de investigación con la cual se ha estudiado el fenómeno de la deserción estudiantil tiene relación al uso de minería de datos (MD) a través de modelos analíticos, ya sea descriptivos o predictivos. Los modelos descriptivos intentan encontrar patrones escondidos en los datos mientras que los modelos predictivos buscan predecir futuros comportamientos (Mishra et al., 2014). En general, la MD se utiliza tanto para extraer información útil desde grandes volúmenes de datos como para visualizarlos de una manera fácil de interpretar

(Song y Ying, 2015). De este modo, a diferencia de la línea de investigación anterior, este enfoque no requiere formular hipótesis previas para explicar el fenómeno de la deserción estudiantil, sino más bien de aplicar algoritmos automatizados para detectar patrones que sirvan para predecir el fenómeno en cuestión (Azoumana, 2013). Aunque la MD es un campo conocido en ciencias de la vida y en ámbitos comerciales, el número de estudios que la han aplicado en el contexto educacional es limitado (Dutt et al., 2017). Algunos de estos corresponden a las investigaciones de Mishra et al. (2014), Pradeep y Thomas (2015), Kotsiantis et al., (2003), y Yukselturk, Ozekes y Turel (2014) que han utilizado, entre otras técnicas, árboles de decisión, K-vecino más cercano, regresión logística, redes bayesianas y redes neuronales para describir o predecir la deserción de estudiantes en la educación terciaria.

Clasificación basada en árboles de decisión

La clasificación basada en árboles de decisión (CBAD) es probablemente una de las técnicas de MD más utilizadas (Hofmann y Klinkenberg, 2013). Si bien en investigación de operaciones los árboles de decisión describen modelos jerárquicos de decisiones y sus consecuencias, en MD se refieren a un modelo predictivo (Rokach y Maimon, 2014). De hecho, como parte de un grupo de modelos analíticos orientados a realizar predicciones, esta técnica no paramétrica clasifica una población en un modelo de segmentos de tipo ramas que construyen un árbol invertido, y luego este modelo se utiliza para predecir una variable objetivo (Song y Ying, 2015). Las principales ventajas de la CBAD se asocian a que puede manejar eficientemente grandes y complejos conjuntos de datos (Song y Ying, 2015), y a su vez, es fácil de usar y sus resultados sencillos de entender (Hofmann y Klinkenberg, 2013).

La noción básica detrás de la CBAD es el enfoque de dividir para conquistar. Siguiendo esta idea, en cada paso el algoritmo de CBAD divide los datos en diferentes segmentos, considerando que cada segmento debe representar lo mejor posible una de las potenciales clases (asociadas a la clasificación analizada). El resultado final será una estructura de árbol invertido, donde cada nodo interno representa una prueba para el valor de un atributo particular de los datos y cada hoja representa la decisión para una clase particular (Kotu y Deshpande, 2014). Posteriormente, y para predecir futuras clasificaciones, cada caso nuevo es evaluado utilizando este árbol como base de decisiones. En específico, la estrategia de la CBAD para la construcción del árbol de decisión es *top-down* en una división recursiva (Hofmann y Klinkenberg, 2013). El primer paso es seleccionar uno de los atributos para el nodo raíz. Luego se crea una rama para cada posible valor del atributo y se dividen los registros en subconjuntos de acuerdo a estos posibles valores. Se repiten estos pasos recursivamente para cada rama, pero usando solo los registros incluidos en esa rama. Se puede detener el proceso si todos los registros tienen la misma clase.

Existen básicamente dos preguntas a responder en cada paso del proceso de construcción del árbol en la CBAD: dónde dividir los datos y cuándo detener la división (Kotu y Deshpande, 2014). Dado que el objetivo de la técnica es obtener el árbol más pequeño posible - pues un árbol pequeño consigue explicar mejor los datos, y además conduce a menos errores para predecir el resultado de futuros registros en comparación con un árbol más complejo - se utiliza la siguiente heurística para seleccionar, en cada paso, el nuevo atributo para discriminar: elegir el atributo que produce los subconjuntos de datos más homogéneos (Hofmann y Klinkenberg, 2013). Hay varias opciones para medir la homogeneidad o pureza de los subconjuntos, las más utilizadas son el Índice Gini y la Ganancia de Información (Breiman, 1996). La primera opción es la probabilidad de no sacar dos registros de la misma clase de un nodo, y la segunda se basa en seleccionar el atributo que produce la división con la menor entropía ponderada (mientras más uniforme es la probabilidad de los registros, mayor es su entropía). Por otra parte, para detener el proceso de división se pueden dar diversas circunstancias, tales como, no existe un atributo que satisfaga el mínimo umbral de Ganancia de Información, se alcanzó un máximo determinado de profundidad del árbol, o hay menos de un número determinado de registros en una rama. En cualquiera de estos casos, el procedimiento se detiene (Kotu y Deshpande, 2014).

Un problema normal en el uso de CBAD es el sobreajuste del árbol de decisiones, este fenómeno ocurre cuando el modelo trata de memorizar los casos de entrenamiento en lugar de generalizar la relación entre los atributos de entrada y la clasificación resultante (Kotu y Deshpande, 2014). Este sobreajuste normalmente clasifica muy bien el conjunto de casos de entrenamiento, pero se comporta mal en la clasificación de nuevos casos. Para prevenir el sobreajuste, es posible restringir el crecimiento del árbol o reducirlo, a este proceso se le llama poda. La pre-poda del árbol de decisión es cuando se utiliza un criterio de restricción durante el crecimiento del árbol (como por ejemplo el máximo de profundidad). Por otra parte, la post-poda es cuando no se restringe el número de ramas y se permite que el árbol crezca tan profundamente como los datos lo permitan, y luego se podan las ramas que no cambian las tasas del error de clasificación. Esta última forma de poda no pierda ninguna relación, por pequeña pero significativa, entre los valores de atributos y las clases, sin embargo, requiere cálculos adicionales que pueden ser desperdiciados cuando el árbol necesita ser podado.

Para medir el nivel de la predicción realizada por una CBAD se usa una tabla de contingencia que describe los casos acertados y los casos erróneos de la predicción. A esta tabla de contingencia se le llama matriz de confusión, y en específico contiene información acerca de las clasificaciones reales y aquellas predichas por la CBAD, de modo que 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 técnica de CBAD ha sido utilizada en numerosos estudios que se han enfocado en la predicción del éxito académico de estudiantes en instituciones de educación superior. Enfocados en estudiantes de Ingeniería Eléctrica de una universidad Holandesa, Dekker et al., (2009) predijeron la tasa de deserción después del primer semestre de su carrera e identificaron factores críticos de éxito asociados a ese programa de estudios. A través de árboles de decisión lograron predecir la deserción estudiantil con una precisión de un 81%. Kovačić (2010) estudió los factores que predicen el éxito de estudiantes en una institución de educación superior de Nueva Zelanda. Utilizando una muestra de 450 estudiantes que cursaron una clase de Sistemas de Información, encontraron que la etnia, programa académico y nivel de avance fueron los factores que predecían la deserción con una precisión de 60,5%. Por su parte, Pal (2012) aplicó distintos algoritmos de clasificación basados en árboles de decisión para predecir la tasa de deserción de estudiantes universitarios en India. Considerando una muestra de 1.650 estudiantes recolectó datos durante un periodo de 5 años y encontró que el algoritmo C4.5 predijo la deserción con un porcentaje de precisión del 80,8% siendo el atributo notas de enseñanza media el más efectivo en la evaluación.

En una línea similar de investigación, Natek y Zwilling (2014) predijeron las tasas de aprobación de asignaturas cursadas por estudiantes universitarios. Los autores sometieron a consideración los datos de 106 individuos inscritos en asignaturas del área informática y comprobaron que los algoritmos de árboles de decisión son igualmente eficientes para predecir futuros comportamientos en muestras pequeñas. Con el uso de distintos paquetes de software para ejecutar los algoritmos de análisis, encontraron un 90% de precisión en la predicción de la aprobación de las asignaturas. En un estudio más reciente, Sivakumar et al. (2016), al igual que Pal (2012), propusieron un algoritmo de decisión para predecir la tasa de deserción de estudiantes universitarios en India. Para ello utilizaron una muestra de 240 estudiantes y consideraron 32 potenciales atributos que podrían influir en la decisión de desertar. El resultado del estudio arrojó 11 atributos relevantes entre los que se encuentran la participación del estudiante en actividades extracurriculares, su satisfacción con el curso, ambiente en el campus universitario, e infraestructura. El algoritmo mostró un porcentaje de precisión de un 92,5%.

Investigaciones realizadas en Latino América también se han enfocado en predecir la deserción estudiantil utilizando clasificación basada en árboles de decisión. Por ejemplo, Eckert y Suénaga (2015) analizaron los factores que influyen en la deserción estudiantil utilizando diferentes criterios de representación y aplicación de algoritmos de clasificación - árboles de decisión, redes bayesianas y reglas. Utilizaron como muestra los datos de 855 estudiantes de la carrera de Ingeniería Informática que ingresaron entre los años 2000 y 2009 a una universidad Argentina. Encontraron como variables influyentes en la deserción las asignaturas aprobadas, cantidad y resultado de asignaturas cursadas, procedencia y edad de ingreso del estudiante. Los autores concluyen que durante el primer año de la carrera es donde adquieren mayor importancia las acciones de contención, apoyo y tutoría de los estudiantes. La combinación de estos criterios obtuvo porcentajes de aciertos, de entre un 76% y un 80% de los casos clasificados correctamente. Finalmente, y en un estudio aún más reciente, Miranda y Guzmán (2017) encontraron que las razones socioeconómicas y el puntaje de ingreso a la universidad son las variables que mejor predicen la deserción de estudiantes. Su investigación incluyó 9.195 sujetos provenientes de carreras de Ingeniería entre los años 2000 y 2013 pertenecientes a una universidad chilena. Según el árbol de decisión construido se clasifica correctamente un 81,3% de los casos en la muestra de entrenamiento y un 82,2% en la muestra de contraste.

MATERIALES Y MÉTODOS

Como fuente de información para realizar al análisis se utilizó una base de datos de 5.288 estudiantes de una universidad estatal chilena. Estos datos corresponden a cuatro cohortes consecutivas de estudiantes pertenecientes a 44 programas de pregrado de las áreas de humanidades, artes, educación, ingeniería, y salud. La Tabla 1 describe la deserción académica de la muestra en relación a los años de avance.

Tabla 1: Deserción académica de la muestra

Deserción	Año 1	Año 2	Año 3	Año 4	Total (%)
No	1.343	1.050	975	821	4.189 (79)
Si	52	221	397	429	1.099 (21)
Total	1.395	1.271	1.372	1.250	5.288 (100)

Los atributos seleccionados para el análisis se relacionan a variables demográficas del estudiante (edad y género), antecedentes de su ingreso a la universidad (puntaje de la prueba de selección universitaria y puntaje asociado a las notas de enseñanza media), aproximaciones a su situación económica (nivel de ingreso familiar y tipo de colegio de enseñanza media), y datos de su rendimiento académico (años de avance, promedio de notas y desviación estándar de notas). La Tabla 2 describe los atributos utilizados en el análisis de la CBAD.

Tabla 2: Atributos para el análisis de la CBAD

<i>Atributo</i>	<i>Tipo</i>	<i>Media</i>	<i>Desv. Est.</i>
Años de Avance	Numérico	2,5	1,1
Edad	Numérico	19,9	2,2
Nivel de Ingreso Familiar (1 a 6)	Numérico	1,4	0,7
Puntaje Prueba de Selección	Numérico	568,9	40,7
Puntaje de Notas Enseñanza Media	Numérico	566,4	85,3
Promedio de Notas	Numérico	4,5	0,9
Desviación Estándar de Notas	Numérico	1,0	0,4
Género	Nominal	<i>N</i>	%
• Femenino		2.941	55,6
• Masculino		2.346	44,4
Colegio de Enseñanza Media	Nominal	<i>N</i>	%
• Privado		2.013	38,1
• Público		322	6,1
• Subvencionado		2.894	54,7
Deserción	Nominal	<i>N</i>	%
• No		4.189	79,2
• Sí		1.099	20,8
Total		5.288	100,0

Para realizar este análisis de CBAD se utilizó la herramienta RapidMiner Studio 7.5 (RapidMiner, 2017), que implementa el algoritmo C4.5 para aprendizaje automático. De acuerdo a lo señalado por Chen et al. (2012), en base a una evaluación de expertos, al número de citas en la literatura y a una encuesta a la comunidad académica, el algoritmo C4.5 fue seleccionado como el más influyente en DM.

RESULTADOS

En este apartado se muestran los resultados de los dos procedimientos requeridos para el análisis. Primero se entregan los hallazgos del proceso de búsqueda de valores óptimos de parámetros para la CBAD, y luego se detallan los resultados de aplicar la CBAD con esos parámetros en los datos de la muestra.

Optimización

Se calcularon los valores óptimos para los parámetros asociados a criterios para seleccionar los atributos para la división, la profundidad máxima del árbol, y el nivel de confianza utilizado para el cálculo del error pesimista de la poda. La Tabla 3 muestra en la columna 1 los parámetros analizados por el proceso de optimización, en la columna 2 el rango (inicio y término) de los valores analizados para los parámetros numéricos, y la cantidad de pasos usados dentro de ese rango, en la columna 3 la lista de posibilidades para los parámetros no numéricos, y finalmente, en la columna 4 el resultado asociado al valor óptimo entre todas las 800 posibilidades evaluadas.

Tabla 3: Parámetros optimizados

<i>Parámetro</i>	<i>Rango/Pasos</i>	<i>Lista</i>	<i>Resultado</i>
Criterio de selección de atributos para división		Precisión Índice Gini Ratio de Ganancia Ganancia de Información	Índice Gini
Profundidad máxima	De 1 a 20 / 20		16
Nivel de confianza utilizado para el cálculo del error pesimista de la poda	De 0,05 a 0,5 / 9		0,15

Resultados de la CBAD

A partir de los parámetros determinados en el proceso anterior se realizó la CBAD. La Figura 1 muestra el modelo determinado por la CBAD.

Fig. 1: Modelo de predicción de la deserción académica

Para efectos de realizar el análisis predictivo se dividieron aleatoriamente los datos en dos submuestras. La primera con el 70% de los registros la cual sirvió para determinar la CBAD, y la segunda con el 30% restante de registros para probar la capacidad de predicción del modelo. En base a la matriz de confusión presentada en la Tabla 4, el ratio de precisión de la predicción alcanzó un 87,27%.

En concreto, la aplicación de la CBAD indica tres factores académicos que explican en la deserción estudiantil: el promedio de notas, los años de avance en la carrera y el puntaje en la prueba de selección. Por tanto, factores demográficos como el género y la edad, o económicos, como el nivel de ingreso familiar y el tipo de colegio de enseñanza media, no explican esta deserción. Igualmente, los factores académicos puntaje de notas enseñanza media y desviación estándar de notas no afectan la predicción de deserción estudiantil en la muestra.

Tabla 4: Matriz de confusión para la predicción de deserción

		Predicción de Deserción		Total
		Sí	No	
Deserción Real	Sí	172	44	216
	No	158	1.213	1.371
Total		330	1.257	1.587

DISCUSIÓN

En relación a los resultados deseamos destacar dos elementos: la precisión de predicción lograda en este estudio y los atributos encontrados relevantes en la predicción. Con respecto a la precisión de predicción, se puede apreciar que los resultados de este estudio superan a aquellos obtenidos en otras investigaciones y

que han utilizado el algoritmo C4.5 a través de la técnica CBAD. La precisión obtenida (87.27%) supera estudios que han considerado muestras de tamaño mediana o grande y que fueron brevemente descritos en la revisión de la literatura (Miranda y Guzmán, 2017; Pal, 2012; Eckert y Suénaga, 2015; Dekker et al., 2009; y Kovačić, 2010). Asociado a esta precisión, es necesario indicar que otras técnicas de clasificación podrían ser una posibilidad para mejorar esta precisión. En particular, los modelos de redes neuronales artificiales son una alternativa a la CBAD para predecir la deserción universitaria. En ese contexto, el estudio de Miranda y Guzmán (2017) que compara la aplicación de un modelo de redes neuronales artificiales con la aplicación de una CBAD para estimar los determinantes de la deserción estudiantil universitaria, señala que la CBAD posee una clasificación correcta superior al modelo de redes neuronales artificiales (82% versus 80%), pero una precisión 1% inferior (72% versus 73%).

El segundo elemento que debemos destacar se refiere a los atributos que impactan la deserción estudiantil. Al igual que lo obtenido en los estudios de Mishra et al. (2014) y Pradeep y Thomas (2015), el promedio de notas es uno de los factores que más explica el fenómeno de la deserción en estudiantes de la educación terciaria. En una línea similar, Eckert y Suénaga (2015) encontraron que la cantidad y el resultado obtenido en las asignaturas aprobadas impactan en gran medida la deserción estudiantil. Los otros atributos importantes encontrados en esta investigación corresponden a los años de avance en la carrera y el puntaje de ingreso a la universidad. Este último además valida el estudio de Miranda y Guzmán (2017) desarrollado en Chile quienes encontraron que el puntaje de ingreso es uno de los atributos que más explica la deserción. Resulta interesante señalar que estos tres atributos se agrupan dentro de los factores académicos propuestos por Bean (1985). Finalmente, y al igual que los hallazgos de Mishra et al. (2014), las condiciones socioeconómicas tienen solo un efecto marginal en la explicación del fenómeno de la deserción estudiantil.

CONCLUSIONES

Una CBAD se utilizó en este estudio para determinar variables que predicen la deserción estudiantil en una universidad estatal chilena. El procedimiento utilizado determinó el conjunto de parámetros de la CBAD que aumentaban su ratio de precisión; estos parámetros se usaron como base para realizar el procedimiento de predicción. El resultado refleja una precisión de predicción de un 87,27%.

Las variables determinadas por el análisis son el promedio de notas, los años de avance en la carrera y el puntaje en la prueba de selección. Estos hallazgos son consistentes con investigaciones previas.

Una limitación de este estudio es no considerar el área de conocimiento de los programas de pregrado para predecir la deserción. En futuros trabajos esto puede ser una interesante idea a explorar.

REFERENCIAS

- Acevedo, D., J. D. Torres, y D. Tirado, Análisis de la deserción estudiantil en el programa de ingeniería de alimentos de la Universidad de Cartagena durante el periodo académico 2009-2013, *Formación Universitaria*, 8(1), 35-42 (2015)
- Ajzen, I., The theory of planned behavior; *Organizational Behavior and Human Decision Processes*, 50, 179-211 (1991)
- Araneda-Guirrman, C., E. Rodríguez-Ponce y L. M. Pedraja-Rejas, Relación entre el financiamiento fiscal, la calidad del cuerpo académico y la retención de estudiantes universitarios en Chile, *Formación Universitaria*, 6(6), 55-64 (2013)
- Azoumana, K., Análisis de la deserción estudiantil en la Universidad Simón Bolívar, Facultad Ingeniería de Sistemas, con técnicas de minería de datos, *Pensamiento Americano*, 41-51 (2013)
- Breiman, L., Some properties of splitting criteria, *Machine Learning*, 24(1), 41-47 (1996)
- Chen, H., R. Chiang, y V. Storey, Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, 36(4), 1165-1188 (2012)
- Comisión de Financiamiento Estudiantil para la Educación Superior, Análisis y recomendaciones para el sistema de financiamiento estudiantil, Ministerio de Educación (2012)
- Dekker, G., M. Pechenizkiy y J. Vleeshouwers, Predicting students drop out: A case study, *Proceedings of the 2nd International Conference on Educational Data Mining*, 41-50 (2009)
- Díaz, C., Modelo conceptual para la deserción estudiantil universitaria chilena, *Estudios Pedagógicos XXXIV*, 2, 65-86 (2008)
- Dutt, A., M. A. Ismail y T. Herawan, A Systematic Review on Educational Data Mining. *IEEE Access*, 99, 1-1 (2017)
- Eckert, K. B. y R. Suénaga, Análisis de deserción-permanencia de estudiantes universitarios utilizando técnica de clasificación en minería de datos. *Formación Universitaria*, 8(5), 3-12 (2015)
- Fichten, C., R. Amsel, M. Jorgensen, M. Nguyen, J. Budd, A. Havel, L. King, S. Jorgensen, y J. Asuncion, Theory of Planned Behavior: Sensitivity and Specificity in Predicting Graduation and Drop-Out among College and University Students, *International Journal of Learning, Teaching and Educational Research*, 15(7), pp. 38 (2016)

- González, L. E., 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)
- Grandón, E. y 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, 275-294 (2012)
- Hofmann, M. y R. Klinkenberg, RapidMiner: Data mining use cases and business analytics applications, CRC Press (2013)
- Kotsiantis, S. B., C. J. Pierrakeas, y P. E. Pintelas, Preventing student dropout in distance learning using machine learning techniques, Educational Software Development Laboratory Palade, Howlett and Jain (Eds.): KES 2003, LNAI 2774, 267-274 (2003)
- Kotu, V. y B. Deshpande, Predictive analytics and data mining: concepts and practice with rapidminer, Morgan Kaufmann (2014)
- Kovačić, Z. J., Early prediction of student success: Mining student enrollment data, Proceedings of Informing Science & IT Education Conference, 647-665 (2010)
- Miranda, M. A. y J. Guzmán, Análisis de la Deserción de Estudiantes Universitarios usando Técnicas de Minería de Datos, Formación Universitaria, 10(3), 61-68 (2017)
- Mishra, T., D. Kumar, y S. Gupta, Mining students' data for performance prediction. Fourth International Conference on Advanced Computing & Communication Technologies, doi 10.1109/ACCT.2014.105 (2014)
- Natek, S. y M. Zwilling, Student data mining solution—knowledge management system related to higher education institutions, Expert Systems with Applications, 41(14), 6400-6407 (2014)
- Pal, S., Mining educational data to reduce dropout rates of engineering students, International Journal of Information Engineering and Electronic Business, 4(2), 1 (2012)
- Pradeep, A. y J. Thomas, Predicting college students dropout using EDM techniques. International Journal of Computer Applications, 123(5), 0975-8887 (2015)
- Ramírez, P., Antecedentes del estudio en estudiantes universitarios: Validación de un modelo de ecuaciones estructurales basado en la teoría del comportamiento interpersonal en Chile, HOLOS, 7, 366-377 (2016)
- Rapidminer, RapidMiner Studio 7.5., Recuperado desde <http://www.rapidminer.com> (2017)
- Rokach, L. y O. Maimon, Data mining with decision trees: theory and applications, World Scientific (2014)
- SIES, Deserción en la educación superior en Chile, Ministerio de Educación (2014)
- Spady, W.G. Dropouts from higher education: An interdisciplinary review and synthesis. Interchange, 1(1), 64-85 (1979)
- Sivakumar, S., S. Venkataraman, y R. Selvaraj, Predictive modeling of student dropout indicators in educational data mining using improved decision tree, Indian Journal of Science and Technology, 9(4), 1-5 (2016)
- Song, Y.Y. y L.U. Ying, Decision tree methods: applications for classification and prediction, Shanghai Archives of Psychiatry, 27(2), 130-135 (2015)
- Tinto, V., Definir la deserción: una cuestión de perspectivas, Revista de Educación Superior, XVIII 3 (71) 33-51 (1989)
- Tinto, V. Dropout from higher education: A theoretical synthesis of recent research. Review of Educational Research, 45, 89-125 (1975)
- Zárate, R. y E. Mantilla, La deserción estudiantil UIS, una mirada desde la responsabilidad social universitaria, Zona Próxima, 21, 120-133 (2014)
- Yukselturk, E., S. Ozekes, y Y. K. Turel, Predicting dropout students: an application of data mining methods in an online education program. European Journal of Open, Distance and e-learning. 17(1), DOI: 10.2478/eurodl-2014-0008. Publicado en línea el (2014)