Proyecto de graduación genera herramienta para pronosticar deserción de estudiantes del TEC

Roberto González Loaiza*
gonzalezloaizar@gmail.com

La deserción es una problemática que afecta a las universidades de Costa Rica y de todo el mundo. Esta situación perjudica el desarrollo personal y profesional de los individuos que no concluyen sus estudios; además, limita el crecimiento económico, social y tecnológico del país.

En Costa Rica, cada estudiante de la educación pública representa una inversión para el Estado. Independientemente de si recibe algún tipo de beca, si cuenta con financiamiento o si paga todos los derechos de estudios, el Estado cubre una parte del costo de formación. Por lo tanto, si el estudiante abandona los estudios será una inversión perdida.

Ante esta problemática, y como requisito para finalizar mis estudios de grado de la carrera de Administración de Tecnologías de Información (ATI), desarrollé el proyecto de graduación "Automatización de dos modelos predictivos de deserción estudiantil del Tecnológico de Costa Rica (TEC), mediante el uso de la herramienta Pentaho". El propósito fue dotar a la institución de una herramienta que permitiera identificar a los estudiantes con alto riesgo de abandonar las aulas y, con ello, facilitar la toma de decisiones y asignación de recursos para ayudar a disminuir los índices de deserción. El estudio se hizo desde la Oficina de Planificación (OPI) del TEC.

Antecedentes

El análisis de la deserción estudiantil es un tema recurrente en instituciones dedicadas a la educación: escuelas, colegios y universidades, entre otros. El TEC como universidad pública realiza esfuerzos para analizar el rendimiento académico, incluyendo la deserción estudiantil. Entre los años 2015 y 2017, la OPI dirigió dos proyectos relacionados con el análisis del rendimiento académico: el primero consistió en el mejoramiento del Sistema de Gestión de Información Académica y Administrativa (SIGI); y el segundo fue un proyecto de análisis exploratorio y predictivo de la deserción estudiantil del TEC.

El proyecto de mejoramiento al SIGI extrae, integra y combina la información de diferentes sistemas institucionales, siendo un insumo para la planificación y toma de decisiones. El SIGI analiza información de distintas áreas: gestión ambiental, desarrollo docente, talento humano, presupuesto, infraestructura física y tecnológica.

El objetivo de la iniciativa llamada "Modelo explicativo y predictivo de la deserción estudiantil en el Tecnológico de Costa Rica de las cohortes 2011, 2012 y 2013", fue crear un modelo que permitiera predecir si los estudiantes van a desertar del TEC tomando como criterio de desertor, un estudiante que abandona la institución por más de cuatro semestres, además de encontrar cuáles son las características o variables que contribuyen en mayor medida a la deserción estudiantil. Este proyecto no llegó a la fase de implementación.

Alcance del estudio

La solución propuesta abarcó cuatro productos para automatizar el pronóstico de la deserción, que se describen a continuación.

Dos modelos predictivos de deserción: la detección de los estudiantes clasificados como posibles desertores se realiza mediante dos modelos predictivos de clasificación binaria. El primero pronostica los estudiantes que van a desertar por uno o más semestres; y el segundo predice los individuos que dejarán de matricular por al menos dos semestres.

El cubo de proyección de deserción estudiantil: contempla una base de datos multidimensional para analizar los resultados de los dos modelos predictivos.

Proceso ETL: un proceso de extracción, transformación y carga de datos, que es el responsable de aplicar periódicamente los dos modelos predictivos y guardar los resultados en una base de datos del SIGI, dejando la información disponible para realizar análisis y visualizar los resultados.

Reporte de proyección de deserción estudiantil: abarca un reporte que muestre únicamente los estudiantes clasificados como posibles desertores, es decir, aquellos que cuenten con 50% o más de probabilidad de desertar en los siguientes períodos. La fuente de datos a consultar son las bases de datos del SIGI.

Unidad de estudio

La investigación abarcó a los estudiantes de grado de los tres centros académicos y dos sedes del TEC que realizaron al menos una

CARTAGO						
ESCUELA DE ADMINISTRACION DE EMPRESAS						
Nombre	Carnet	Sexo	Financiamiento	Beca Mauricio Campos	IDS Distrito	Probabilidad de Desertar
		Mujer	No Financiado	Sin Beca MC	Mayor Desarrollo	90%
		Hombre	No Financiado	Sin Beca MC	Nivel Medio	94%
		Mujer	No Financiado	Sin Beca MC	Nivel Medio	93%

matrícula entre el primer semestre del año 2011 y el segundo semestre del 2016.

En total se analizó el comportamiento de 15 190 individuos, de los cuales dos terceras partes son hombres. Las carreras con mayor número de estudiantes fueron: Administración de Empresas, 15,46%; Ingeniería en Computación, 14,21%; y Producción Industrial, 10,08%.

En la sede de Cartago se realizaron el 79% de las matrículas, seguido por la sede San Carlos con 11% y el restante 10% se dividió en los centros académicos de Alajuela, San José y Limón.

Se encontró que el 28% de los estudiantes dejaron de matricular por al menos un semestre y el 25% por al menos dos semestres.

Variables del estudio

Las variables utilizadas para construir los dos modelos predictivos se dividen en dos tipos: independientes y dependientes. La variable que se busca predecir es conocida como variable dependiente; esta es pronosticada en función de un conjunto de variables conocidas como variables independientes o predictoras[1].

Para cada modelo predictivo de deserción se definió una variable dependiente, la cual toma dos valores: "Desertor" y "No desertor". Para el modelo predictivo de deserción de al menos un semestre, la variable dependiente toma el valor de "Desertor" cuando el estudiante dejó de matricular por al menos un período entre los años 2011 y 2017 y no aparece como graduado. El valor de "No desertor" lo adquiere en cualquier condición que no cumpla lo especificado anteriormente. En el caso del modelo predictivo de deserción de dos semestres, la variable dependiente toma el valor de "Desertor" cuando el estudiante dejó de matricular por al menos dos períodos entre los años 2011 y 2017 y además no aparece como graduado. El valor de "No desertor" lo adquiere en cualquier condición que no cumpla lo especificado anteriormente.

La escogencia de las variables independientes o predictoras se basó en los estudios previos de la deserción universitaria en el TEC y en otras universidades del mundo. En total se utilizan 20 variables: Año; Semestre; Grado del plan de estudio; Escuela de formación; Promedio del semestre; Cantidad de cursos matriculados; Cantidad de cursos aprobados; Cantidad de semestres de no matricular en períodos anteriores; Detalle de si cuenta con beca Mauricio Campos; Detalle de si cuenta con financiamiento; Edad; Sexo; Permanencia; Sede; Año de ingreso a la institución; Jornada de la carrera; Detalle de si ingresó a la primera opción; Detalle de si solicitó cambio de carrera y cuál fue la resolución; Cantidad de cursos pendientes para graduarse; y Cantidad de semestres en la carrera.

Para los dos modelos predictivos de deserción se utilizaron las mismas variables predictoras.

Técnicas de minería de datos utilizadas

La minería de datos se define como un proceso que reúne un conjunto de técnicas y herramientas de diversas ciencias, especialmente estadística e informática, para extraer conocimiento oculto y patrones no observables en grandes volúmenes de datos[2].

En la investigación se utilizó la técnica de minería de datos conocida como clasificación binaria, la cual pretende, a partir de conjunto de datos de ejemplo, clasificar nuevos individuos o eventos. La particularidad de esta técnica radica en que la variable a predecir es una variable cualitativa.

Los algoritmos de clasificación utilizados fueron los bosques aleatorios y los árboles de decisión. Los árboles de decisión han sido utilizados en diferentes universidades alrededor del mundo para trabajar el problema de la deserción[3]. El algoritmo de bosques aleatorios se eligió debido a que obtuvo la mayor precisión en estudios previos del pronóstico de la deserción en el TEC.

Para evaluar el desempeño de los dos modelos predictivos se separaron los datos en dos muestras: la primera, los datos de entrenamiento, los cuales estuvieron comprendidos por el 75% de los registros; y la segunda, los datos de prueba, que sirven para evaluar la confiabilidad de las predicciones y representaban el 25% de los registros. Además, se eligió la métrica F-score para determinar cuál algoritmo y modelo era más preciso.

Precisión de los dos modelos predictivos

En ambos modelos predictivos el algoritmo de bosques aleatorios fue más preciso que el de árboles de decisión. Para el modelo predictivo de deserción de al menos un semestre, los bosques aleatorios obtuvieron un 77% de índice F-Score, mientras los árboles de decisión un 66%. En cuanto al modelo predictivo de deserción de al menos dos semestres, los bosques aleatorios obtuvieron un 78% y los árboles de decisión un 65%.

El modelo predictivo de deserción de al menos un semestre captó el 73% de la desertores y la precisión fue de un 80%. Es decir, de cada 100 desertores, este modelo identificó 73 y de cada 100 individuos que el modelo clasificó como desertores, 80 fueron acertados y erró en 20 oportunidades.

El modelo predictivo de deserción de al menos dos semestres captó el 76% de los desertores y la precisión fue de un 81%. Es decir, de cada 100 desertores este modelo identificó 76 y de cada 100 individuos que el modelo clasificó como desertores, 81fueron aciertos y erró en 19 ocasiones.

Visualización de los resultados

Los resultados de los dos modelos predictivos se visualizan mediante dos herramientas: el cubo de proyección de deserción estudiantil y el reporte de proyección de deserción estudiantil. Ambos productos fueron incorpo-

rados a la plataforma del SIGI; por lo tanto, están disponibles para ser consultadas a través de un explorador web.

El reporte de proyección de deserción estudiantil muestra aquellos estudiantes clasificados como posibles desertores, de acuerdo con el modelo predictivo de deserción de al menos dos semestres. El reporte se personaliza de acuerdo con tres parámetros: sede, carrera y período de pronóstico. Se incluye el detalle del nombre; carné; sexo; detalle de si recibió financiamiento; detalle de si recibió beca Mauricio Campos; Índice de Desarrollo Social (IDS) del distrito de procedencia; y la probabilidad de desertar. En la imagen de la página anterior se observa un ejemplo del cuerpo del reporte.

El cubo de proyección de deserción estudiantil permite realizar análisis a partir de la agrupación de resultados; es decir, el propósito de esta herramienta es comprender y analizar las proyecciones realizadas por los dos modelos según un conjunto de características.

Las dimensiones incluidas en el cubo de proyección de deserción estudiantil son: período de ingreso; estudiante; carrera; tipo de bachillerato; grado académico; plan de estudios; área de conocimiento; proceso de admisión; dirección en tiempo lectivo; dirección permanente; sede; forma de ingreso; beca Mauricio Campos; financiamiento; tiempo; solicitud de cambio de carrera; probabilidad de desertar al menos un semestre; y probabilidad de desertar al menos dos semestres. Las dimensiones mencionadas se pueden combinar para analizar el comportamiento de las proyecciones de la deserción, colaborando con el descubrimiento de información valiosa por medio de la interpretación de las agregaciones.

Futuros proyectos

Como parte del estudio se sugiere la implementación de una serie de iniciativas que podrían contribuir a la mejora de la precisión de los dos modelos predictivos. A continuación se describen:

- Construir un modelo predictivo de graduación, el cual clasifique a los estudiantes de acuerdo con la probabilidad de finalizar con éxito el plan de estudios donde se encuentra matriculado al finalizar el semestre.

- Implementar la técnica de minería de datos conocida como reglas de asociación, para identificar de manera individual las principales causas de la deserción.

- Crear un procedimiento de almacenado que sea capaz de identificar, de forma masiva, si un estudiante cumplió con los requisitos para graduarse.

- Desarrollar un proyecto de calidad de datos con el objetivo de analizar la completitud, conformidad, consistencia, precisión, duplicidad e integridad de las bases de datos del SIGI.

Es importante rescatar que el TEC se convirtió en la primera universidad pública de Costa Rica en implementar una iniciativa que permite pronosticar la deserción estudiantil mediante el uso de técnicas de minería de datos. Esto se debió al esfuerzo conjunto del personal de la OPI y el estudiante Roberto González, quienes contaron con el apoyo del Ph.D. Martín Solís, de la Escuela de Administración de Empresas.

Como parte de la divulgación del estudio se realizaron presentaciones ante los Consejos de Rectoría, Vicerrectoría de Docencia y Vicerrectoría de Vida Estudiantil y Servicios Académicos.

El proyecto fue dirigido por la OPI; por consiguiente, para obtener acceso a las proyecciones y a cualquier funcionalidad del SIGI, el proceso debe tramitarse mediante dicha oficina.

Referencias bibliográficas

[1]Tan, P. Steinbach, M. y Kumar, V (2006). Introduction to DATA MINING. [Introducción a la minería de datos]. Boston: Pearson Education, Inc.

[2]Han, J. Kamber, M. y Pei, J. (2011). Data mining concepts and techniques [Conceptos y técnicas de minería de datos]. EE.UU.: Morgan Kaufmann.

[3]Amaya, Y., Barrientos, E. y Heredia, D. (2014) Modelo predictivo de deserción estudiantil utilizando técnicas de minería de datos. Universidad Francisco de Paula Santander, Ocaña, Colombia y Universidad Simón Bolívar, Barranquilla, Colombia.

*Roberto González Loaiza es profesional del área de tecnologías de información. Labora en la empresa BD Consultores, donde se especializa en proyectos de calidad e integración de datos. Además, cursa una especialidad para obtener el título de especialista en big data y minería de datos.