ATICA 2020

Aplicación de Tecnologías de la Información y Comunicaciones Avanzadas y Accesibilidad

Luis Bengochea
Gerardo Contreras Vega
(Editores)

OBRAS COLECTIVAS
TECNOLOGÍA 32

UAH

Desarrollo de un modelo predictivo para descubrir factores que inciden en la deserción de alumnos en la Facultad de Humanidades, Universidad Nacional del Nordeste

Viviana Moschner[1], Paola Britos[2],
[1]Universidad Nacional del Nordeste (Argentina)
Maestría en Tecnologías de la Información
[2]Universidad Nacional de Rio Negro. Laboratorio de Informática Aplicada. Río Negro (Argentina)

Resumen. La baja proporción de egresos anuales es una situación que preocupa a las autoridades de las instituciones de educación superior de la Argentina. Desde 1983, año en que se eliminaron los cupos de ingreso, hubo un considerable aumento en la matrícula de las universidades públicas. Sin embargo, el ingreso directo a las facultades, sólo posterga el fracaso, con costos elevados para las instituciones y para la sociedad en general. El crecimiento de la matrícula no se vio reflejado en el número de egresados e incluso, por el contrario, la brecha existente entre ingresos y egresos anuales, aumentó. También se comprobó que la aplicación de técnicas y herramientas utilizadas en Ciencia de datos en el ámbito educativo tienen resultados positivos y que el uso de las mismas permite predecir situaciones de deserción estudiantil. En éste contexto, el presente proyecto propone desarrollar un método predictivo que permita descubrir factores comunes, en la población estudiantil de la Facultad de Humanidades de la UNNE, que hayan abandonado la carrera Profesorado en Ciencias de la Educación o bien presenten un marcado rezago, con el objetivo de brindar conocimiento que permita a las autoridades, a desarrollar y aplicar nuevas estrategias para disminuir las situaciones de abandono/deserción y rezago en la educación superior.

Palabras clave: Deserción en la educación superior, modelo predictivo, Ciencia de datos

1. Introducción

En la República Argentina las universidades cuentan desde hace décadas, con el sistema de información académica, SIU guaraní, desarrollado oportunamente por el consorcio SIU, dependiente del CIN (Consejo Interuniversitario Nacional). El mismo registra una abundante y amplia gama de información de la población estudiantil, desde el ingreso y hasta el egreso de los alumnos.

Por otro parte, la disciplina en auge, conocida como Ciencia de datos, nos ofrece la posibilidad de descubrir patrones interesantes presentes en grandes cantidades de información, sin necesidad de tener planteada una hipótesis previa.

El principal objetivo de este trabajo es aplicar diferentes herramientas de Ciencia de datos, con el propósito de obtener un modelo predictivo que permita reconocer con antelación, a los estudiantes que están en riesgo de abandonar la carrera. Con la creación

de un modelo predictivo de deserción estudiantil, se busca generar conocimiento de causales de abandono o rezago, detectar al alumno en riesgo, para poder aplicar diferentes estrategias que aumenten la retención estudiantil de forma genuina.

1.1. Ciencia de datos

La Ciencia de datos es considerada actualmente como la principal herramienta para la explotación de datos y la generación de conocimiento. Tiene como objetivo la búsqueda de modelos que describan patrones y comportamientos a partir de los datos con el fin de tomar decisiones o hacer predicciones [1]. Es un área que experimentó un importante crecimiento al extenderse el acceso a grandes volúmenes de datos e incluso su tratamiento en tiempo real, requiriendo de técnicas nuevas y superadoras, que puedan tratar con los problemas prácticos como escalabilidad, robustez ante errores, adaptabilidad con modelos dinámicos. Abarca a varios grupos de investigación de diferentes áreas, como computación, estadística, matemáticas e ingeniería, todas trabajan en la elaboración de nuevos algoritmos, técnicas de computación e infraestructuras para la captura, almacenamiento y procesado de grandes masas de datos [2].

1.2. Explotación de la información

La Explotación de información es una sub-disciplina de los sistemas de información que brinda a la Inteligencia de negocios, las herramientas para transformar la información en conocimiento [3], la misma se define como la búsqueda de patrones interesantes y de regularidades importantes en grandes masas de información.
Cada proceso de Explotación de información aplica un conjunto de técnicas de Minería de datos, la mayoría provenientes del campo del aprendizaje. Por ello se concluye que los términos Minería de datos y Explotación de información no deben utilizarse para referirse al mismo cuerpo de conocimientos [3], ya que la Minería de datos se relaciona con los algoritmos necesarios para transformar los datos en conocimiento mientras que la Explotación de la información lo hace con los procesos y las metodologías propias de la ingeniería que son necesarias para lograr este objetivo. Es por esto, que la Minería de datos se aproxima a la operatoria propia de la Programación, y la Explotación de información se acerca más a los procesos de la Ingeniería de software.

1.3. Trabajos relacionados

El abandono de las carreras preocupa a las universidades Argentinas y latinoamericanas entre otras. Así es, que se han realizado estudios empíricos con el objetivo de encontrar patrones de comportamiento, en forma automática, utilizando los datos de los sistemas de gestión académica de las instituciones. La Universidad Nacional de Misiones, UNAM-Argentina, utilizando la información recabada por el sistema de gestión académica, y valiéndose del uso de algoritmos TDIDT, ha realizado investigaciones con el fin de identificar variables que inciden en la deserción,[4]. La Universidad Nacional del Nordeste de Buenos Aires, UNNOBA, elaboró un trabajo muy interesante, en el mismo se aplicaron técnicas de Explotación de la información, utilizando datos del SIU, para detectar alumnos en riesgo de abandono y además se desarrolló un tablero de control que permite a los docentes , que no dominan las técnicas

de Minería de datos, visualizar la situación de las distintas cohortes y de cada alumno en detalle, [5].
En otros países latinoamericanos también se han desarrollado investigaciones para entender el problema de la deserción en la educación superior, en Chile se elaboró un estudio con el objetivo de presentar una clasificación basada en árboles de decisión con parámetros optimizados para predecir la deserción de los estudiantes universitarios, [6].
En la Universidad Simón Bolívar, Barranquilla, Colombia, se realizó un trabajo, en el que también se optó por la inducción de árboles de decisión, porque además de ser la técnica más común dentro las técnicas de clasificación de datos, representa una gran ventaja con respecto a las demás técnicas de clasificación debido a que se puede poder representar el conocimiento extraído en un conjunto de reglas de decisión de fácil entendimiento,[7].

1.4. Deserción

Son varias las definiciones de indicadores de deserción, Tinto [8] plantea : "El estudio de la deserción de la educación superior es extremadamente complejo, pues implica no solo una variedad de perspectivas sino también una amplia gama de diferentes tipos de abandono. Probablemente ninguna definición pueda captar en su totalidad la complejidad de este fenómeno universitario", también indica que existe una gran variedad de comportamientos denominados con el rotulo común de deserción; mas no debe definirse con este término a todos los abandonos de estudios, ni todos ellos merecen intervención institucional, agrega que "solo algunos de los abandonos de la educación superior son producidas por bajo desempeño académico pues la mayor parte de las deserciones son voluntarias. Los estudiantes que abandonan la universidad a menudo tienen niveles de rendimiento superiores a los que persisten", Tinto [8], considera desertor al estudiante que no presenta actividad académica durante tres semestres académicos consecutivos.
En esta investigación se considera deserción a un periodo suficientemente largo como para que un estudiante decida retomar sus estudios, equivalente a 2 (dos) años.

2. Materiales y métodos

2.1. Metodología

Las metodologías de Explotación de la información son herramientas que permiten llevar a cabo el proceso de Minería de datos en forma sistemática y no trivial, éstas definen las fases del proceso y además describen las tareas a realizarse y el modo de llevarlas a cabo. Entre las metodologías disponibles que se pueden aplicar en la ejecución de los proyectos de Ciencia de datos se destacan: KDD, CRISP-DM, P^3TQ SEMMA y MoProPEI.
En esta investigación se optó por utilizar la metodología creada recientemente por argentinos. MoProPEI (Modelo de Procesos para Proyectos de Explotación de información), debido a que la misma ofrece una guía precisa del desarrollo del proyecto, considerando los aspectos de gestión y los técnicos, con el fin de generar piezas de conocimiento que sirvan para la toma de decisiones. El modelo cuenta con una estructura jerárquica dividida en cuatro niveles: Subprocesos, Fases, Actividades y Tareas; cada uno de los cuales presenta un mayor nivel de especificidad. [9].

2.2. Población objetivo

La información que se utiliza proviene del SIU Guaraní, sistema que se implementó, en la Facultad de Humanidades, dependiente de la Universidad Nacional del Nordeste, en el año 2003. El mismo recaba datos de tipo personal, laboral, económico y académico, de alumnos que registran su ingreso desde el año 1983. Específicamente se utiliza la información de los estudiantes del Profesorado en Ciencias de la Educación, cohortes **2010 a 2018**, debido a que es una de las propuestas que registra el mayor número de ingresos y egresos anuales.

2.3. Selección de los datos

Se utilizan 12 (doce) de las tablas relacionadas de la base de datos y se extraen diferentes atributos, censales, académicos y administrativos.
Se extrae un set de 1280 registros, el 76% son mujeres y el 24% varones, **Gráfico 1,** al momento de la obtención de los datos, el 38% de ellos tiene entre 19 y 25 años, el 40% entre 26 y 30, el 16% entre 31 y 40 y el resto más de 40 años, **Gráfico 2**.
Respecto a la edad de ingreso a la carrera el 56% tenía entre 18 y 20 años, el 25% entre 21 y 25, el 14% entre 26 y 35 y el 5% 36 años o más, **Gráfico 3.**
En cuanto a la procedencia el 80% declara proceder de la provincia del Chaco, el 15% de la provincia de Corrientes y el 5% restante de otras provincias, **Gráfico 4.** De los procedentes de la provincia del Chaco, el 50% procede de la ciudad capital, el 20% de localidades muy próximas y el 30% de localidades del interior de la provincia, **Gráfico 5**. El 51% de los alumnos posee título secundario Bachiller, el 42% Educación Polimodal y el 4% Técnico, **Gráfico 6.**

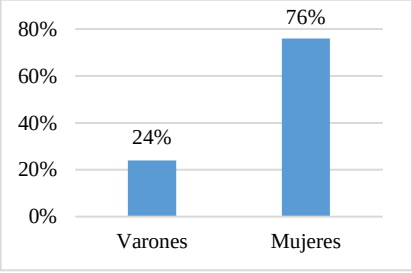

Gráfico 1. Alumnos por género

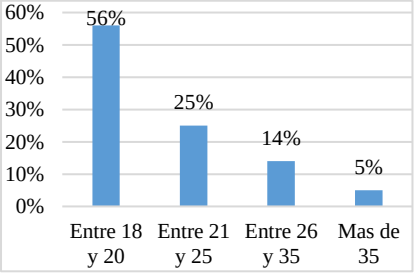

Gráfico 3. Por edad al ingreso

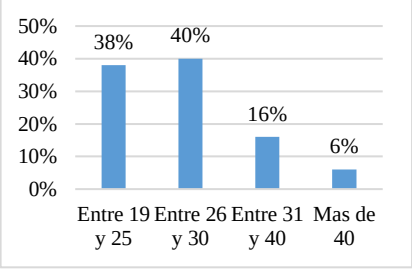

Gráfico 2. Por edad al obtener datos

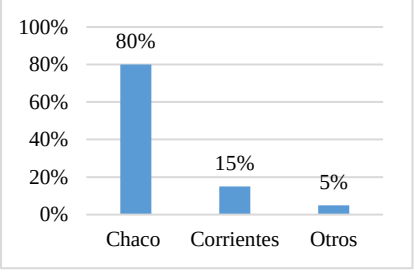

Gráfico 4. Por procedencia

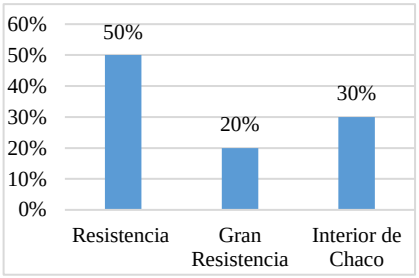

Gráfico 5. Por localidad (Chaco)

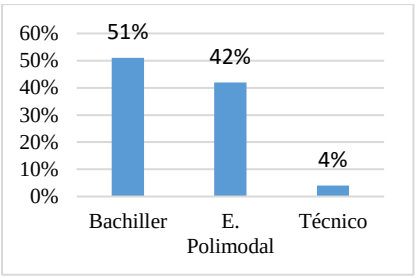

Gráfico 6. Por título secundario

2.4. Limpieza y transformación de los datos

Se realiza un preprocesamiento con el objetivo de mejorar la calidad de los datos, se usa el nodo *selección de características* de la herramienta SPSS-Modeler, que identifica los campos más importantes para predecir determinados resultados. Luego se procede a la ejecución del algoritmo *auditoría de datos*, para evaluar la calidad de los atributos. Los valores nulos y/o faltantes, son imputados usando el algoritmo CRT, a fin de optimizar el proceso de Explotación de información.
Algunas variables fueron descartadas ya que no brindaban información sustantiva al objetivo planteado, se crearon nuevos atributos partir de los ya existentes y se transformaron varios de ellos, agrupándolos.

2.5. Clasificación

En esta investigación se considera Posible Desertor al alumno que no registra actividad académica por un periodo de 2 (dos) años.

Trabajando con los expertos, se decidió hacer la clasificación en base al atributo calidad (indica si es egresado o alumno activo), total de asignaturas aprobadas, regularizadas, total de inscripciones al cursado y años de inactividad, se construye la variable a predecir, Calidad real, categorizando a la población estudiantil en:

- **Egresado (E)**: Persona que después de haber revestido la calidad de estudiante en una carrera, completa todos los cursos y requisitos reglamentarios de la carrera a la que pertenece y que ha tramitado su título de graduado, (calidad=E).
- **Deserción Precoz (SA)**: Alumno que se inscribió a la carrera, no registra actividad académica, no se inscribió a cursar ninguna asignatura, (calidad=A, total de inscripciones al cursado = 0).
- **Deserción Temprana (DT)**: Alumno que se inscribió a la carrera, registra inscripciones a cursar, no regularizó/aprobó/reprobó ninguna materia, (calidad=A, total inscripciones al cursado>0, Total aprobadas=0, Total regularizadas=0).
- **Posible Desertor (PD)**: Alumno con materias aprobadas y con la última actividad registrada antes del 2018, calidad=A, Total aprobadas>0, años inactivo >2).

- **Activo Rezagado (R):** Alumno que registra asignaturas aprobadas, pero no son, en cantidad, las esperadas según la cohorte a la que pertenece, calidad=A, total aprobadas>0).

2.6. Modelado

Para cumplir con el objetivo de ésta investigación se utiliza la herramienta SPSS-Modeler, [10]. Se seleccionó el algoritmo Redes Neuronales, debido a que en el marco de la investigación arrojó mejores resultados, que otras técnicas de clasificación supervisada, como KNN y CHAID.
Redes neuronales: Las redes neuronales nacen a partir de los intentos de los investigadores por establecer un sistema que logrará representar las características de funcionamiento del sistema nervioso de las personas.
C5.0: Este forma parte de la familia de los Árboles Inducidos de Arriba hacia Abajo (TDIDT). Pertenece a los métodos inductivos, los cuales aprenden a partir de ejemplos preclasificados.

3. Los resultados

Después de varias pruebas con el set de datos seleccionado y con el fin de obtener resultados más lógicos y esperados se decidió eliminar registros de los alumnos que no registran actividad académica, no regularizaron, aprobaron o reprobaron asignaturas, con el propósito de evitar la generación de ruido.
Al set con los datos imputados se aplica el algoritmo Redes Neuronales, seleccionando el modelo perceptrón multicapa, obteniéndose la siguiente Matriz de Confusión, con una muy buena precisión, 97,4%, **Figura 1.**
El modelo clasificó incorrectamente, sólo a 20 alumnos sobre un total de 755 registros, con una exactitud (*accuracy*) de 0,97.

Clasificación para CALIDAD_R

Porcentaje correcto global = 97,4%

Observado	Predichos		
	E	PD	R
E	98,2%	0,0%	1,8%
PD	0,6%	97,8%	1,5%
R	1,6%	4,0%	94,4%

Porcentaje de filas
100,00
80,00
60,00
40,00
20,00
0,00

Figura 1. Matriz de confusión

Al resultado de este proceso se aplica el algoritmo C5.0, el mismo genera reglas que nos permitirá conocer factores comunes entre los alumnos clasificados como posibles desertores, rezagados y egresados. Del conjunto de reglas obtenidas, las más interesantes por cantidad de ocurrencias y precisión son las siguientes:
El alumno que ingresó con 19 años o menos, uno de los padres tiene estudios universitarios, no tiene familiares a cargo, la madre tiene estudios Secundarios

completos o mayor, y financia sus estudios con ayuda familiar, entonces es Egresado**,** **Figura 2**.

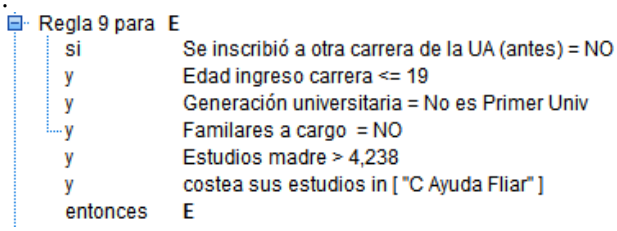

Figura 2. Regla para Egresado

El alumno que ingresó con 19 años de edad o menos, es la primera generación universitaria, y estudia con la ayuda económica de familiares y planes sociales, entonces es Posible desertor, **Figura 3**.

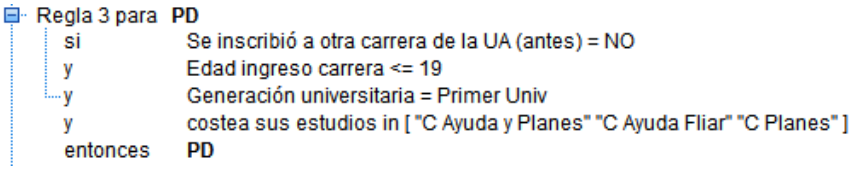

Figura 3. Regla Posible desertor

El alumno que ingresó con 19 años de edad o menos, es la primera generación universitaria, el padre tiene estudios de nivel secundario completo o mayor y costea los estudios con su trabajo, entonces es alumno Rezagado, **Figura 4.**

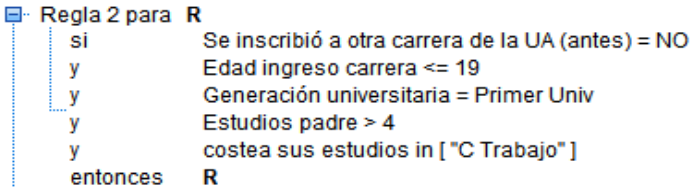

Figura 4. Regla alumno Rezagado

4. Conclusiones

Desde hace varios años en la Universidad Nacional del Nordeste, de Argentina, se implementan diferentes estrategias para disminuir el abandono estudiantil. Para ello, entre las principales, se otorgan becas económicas y se promueven sistemas de tutorías, pero no se ha logrado mitigar de manera sensible el alto porcentaje de deserción.

En base a los patrones obtenidos en la presente investigación, con la aplicación de los algoritmos de clasificación e inducción, se puede afirmar que es posible obtener modelos predictivos, que generen conocimiento, en los que las autoridades de las Unidades Académicas puedan respaldarse para individualizar al alumno en riesgo de abandono y desarrollar nuevas políticas en base a la información específica generada

con éstas herramientas, con el fin de prevenir el abandono y aumentar la retención estudiantil.

Como líneas a ser abordadas en el futuro, se propone, adaptar el modelo obtenido, para su aplicación a todas las carreras de la universidad, UNNE, Argentina y la elaboración e implementación de un tablero de control que permita, a los docentes y expertos, detectar con antelación al alumno rezagado y/o en riesgo de abandono.

5. Referencias

[1] R. E. López Briega, "Ciencia de datos," *Ciencia de datos - Libro online de IAAR*, 2017. [Online]. Available: https://iaarbook.github.io/datascience/.

[2] P. Altaria, J. M. Molina, A. Berlanga, and M. A. Patri-, *Ciencia de Datos*. .

[3] R. García-Martínez and P. Britos, "Towards an Information Mining Engineering," *Towards an Information Mining Engineering. En Software Engineering, Methods, Modeling and Teaching*, pp. 83–99, 2011.

[4] J. G. A. Pautsch, H. D. Kuna, and A. E. Godoy, "Resultados Preliminares del Proceso de Minería de Datos Aplicado al Análisis de la Deserción en Carreras de Informática Utilizando Herramientas Open Source Objetivo principal Revisión conceptual," no. Md, pp. 1027–1036.

[5] C. C. Russo, "Minería de datos aplicada a estrategias para minimizar la deserción universitaria en carreras de Informática de la UNNOBA," *Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología*, 2019.

[6] P. E. Ramírez and E. E. Grandón, "Predicción de la Deserción Académica en una Universidad Pública Chilena a través de la Clasificación basada en Árboles de Decisión con Parámetros Optimizados," *Formación universitaria*, 2018.

[7] K. Amaya, E. Barrientos, and J. Heredia, "Modelo predictivo de deserción estudiantil utilizando técnicas de minería de datos," *Mining Techniques*, 2014.

[8] V. Tinto, "Summary for Policymakers," in *Climate Change 2013 - The Physical Science Basis*, Intergovernmental Panel on Climate Change, Ed. Cambridge: Cambridge University Press, 1989, pp. 1–30.

[9] S. Martins, P. Pesado, and R. García-Martínez, "Propuesta de Modelo de Procesos para una Ingeniería de Explotación de Información: MoProPEI," *Revista Latinoamericana de Ingenieria de Software*, vol. 2, no. 5, p. 313, 2015.

[10] IBM, "IBM SPSS Modeler CRISP-DM," *IBM Corporation*, 2016.