AJEA - Actas de Jornadas y Eventos Académicos de UTN
DOI: https://doi.org/10.33414/ajea.1146.2022

Deserción de Estudiantes en Carreras de Ingeniería: Análisis Multivariable Utilizando Minería de Datos Educativa

Marcela Andrea Vera, Mariel Alejandra Ale, Luciana Ballejos
mavera@frsf.utn.edu.ar, male@frsf.utn.edu.ar, lballejos@frsf.utn.edu.ar
CIDISI – Centro de I+D de Ingeniería en Sistemas de Información
UTN – Facultad Regional Santa Fe

Resumen

La alta tasa de deserción estudiantil en las carreras universitarias es una problemática actual en las universidades argentinas y de toda Latinoamérica. El objetivo de este trabajo es analizar mediante algoritmos de minería de datos este fenómeno para obtener conocimiento que permita a las autoridades generar estrategias que disminuyan la cantidad de alumnos que no finalizan sus estudios de grado. En particular, en este trabajo se utilizó la metodología CRISP-DM para guiar las diferentes etapas y se implementaron los modelos K-Means y Perceptron Multicapa, ambos ampliamente utilizados en el contexto de la minería de datos para datos académicos. Finalmente, se encontró una interrelación entre la probabilidad de abandono y la cantidad de materias que el alumno debe recursar, además de generar un modelo predictivo de deserción con una precisión cercana al 90%.

1. Introducción

En Sudamérica en general y en Argentina en particular, la tasa de graduación de las carreras universitarias es baja. En promedio, la tasa bruta de matrícula en educación superior de América Latina y el Caribe creció del 17 por ciento en 1991 al 21 por ciento en el año 2000 y al 40 por ciento en el año 2010 [1]. Sin embargo, finalizan sus estudios superiores sólo un 50% de los matriculados. Según los últimos datos en Argentina, alrededor del 30% de los alumnos finalizan sus estudios en tiempo y forma. Y estos números empeoran en las carreras de Ingeniería, donde el porcentaje de egreso apenas llega al 20% según datos del Consejo Federal de Decanos de Ingeniería (CONFEDI) del año 2019 [2].

Cuando comparamos el porcentaje de egresados en otros países de América encontramos que, según datos estadísticos, en Brasil se gradúa un 55,8% y en México y Perú se alcanzan tasas de graduación cercanas al 70% [3]. Teniendo en cuenta esta problemática, se desarrolló este trabajo con el objetivo de detectar características comunes en aquellos alumnos que tienen una alta probabilidad de abandono de las carreras de ingeniería en la Universidad Tecnológica Nacional-Facultad Regional Santa Fe (UTN FRSF). Se utilizaron como soporte los recursos de los diferentes modelos descriptivos de Minería de Datos Educativos [4], definiendo "abandono de carrera" como la situación en la que se encuentran aquellos alumnos que luego de dos años consecutivos no realizan ningún tipo de actividad académica (inscripción a materias, regularización y/o aprobación directa de materias, inscripción a exámenes, aprobación o no aprobación de exámenes).

Luego, el estudio se amplió con el uso de modelos predictivos que permitan predecir en las carreras de ingeniería de la Regional Santa Fe, a partir de los datos de alumnos en los primeros años de las carreras, cuáles serán los que tendrán una alta probabilidad de abandonar sus estudios superiores. Esto es importante, ya que la determinación de estos perfiles de alumnos permitirá a la facultad definir políticas específicas y estrategias que reviertan el alto porcentaje de alumnos que abandonan las carreras.

Para realizar este trabajo, se utilizó la metodología Cross Industry Standard Process for Data Mining (CRISP-DM) que desglosa el proyecto en diferentes fases [5]. Además, se utilizaron los modelos de agrupamiento K-Means y de clasificación perceptrón de múltiples niveles, que permite aprender en modelos complejos [6]. Se seleccionaron ambos modelos considerando los objetivos planteados en el estudio realizado, ya que nos permiten descubrir asociaciones entre los datos, además ambos modelos han sido ampliamente usados en el contexto de la minería de datos educativos, con resultados ampliamente difundidos.

El algoritmo K-Means permite clasificar individuos, partiendo de grupos de individuos parecidos entre sí, según un conjunto de variables de entrada [7]. En este trabajo, las variables de entrada fueron definidas en conjunto con el Sector de Dirección Académica, teniendo en cuenta aquellas que se muestran asociadas a la deserción en las carreras de ingeniería.

AJEA - Actas de Jornadas y Eventos Académicos de UTN
DOI: https://doi.org/10.33414/ajea.1146.2022

Existen muchos trabajos que han utilizado las técnicas de minería de datos educativa para realizar diferentes análisis de desempeño académico, deserción escolar, y factores asociados a estas problemáticas. Entre los trabajos que han realizado estudios de la deserción escolar utilizando técnicas de minería de datos educativa, la propuesta de Urbina-Nájera [8] utiliza algoritmos de selección de atributos para detectar los factores más importantes que afectan la deserción escolar en institutos de Educación Superior y árboles de decisión que permitan definir patrones para alertar una inminente deserción.

La técnica de Árboles de Decisión fue también utilizada para detectar las causas de la deserción escolar en la carrera de Ingeniería en Computación en el estado de México en el trabajo de Aguirre Mendiola y otros [9].

Entre otros, es importante el trabajo de Ujkani y otros [10], en el cual se utilizaron cuatro técnicas de aprendizaje automático con la evaluación de un total de dieciséis algoritmos utilizando el software Weka para predecir la matriculación de alumnos en la universidad, teniendo en cuenta no sólo los exámenes de ingreso, sino también la actuación en la escuela secundaria

En las siguientes secciones se describe la metodología utilizada para realizar el estudio y cada una de las etapas, actividades y tareas realizadas, junto con las conclusiones a las que se arribó al finalizar este estudio.

2. Metodología

2.1 Marco teórico

Según el sitio Web Kdnuggets, CRISP-DM continúa siendo la metodología más usada en minería de datos [18], a pesar de las criticas de diversos autores, respecto a la informalidad de sus primeras fases.

La metodología CRISP-DM [5] dispone de una guía de usuario y un modelo de referencia. Se describe en términos de un modelo jerárquico de procesos, consistente en un conjunto de tareas organizadas en cuatro niveles de abstracción: fase, tarea genérica, tarea especializada e instancia de procesos. En el nivel superior, el proceso de minería de datos se organiza en un número de fases; cada fase consta de varias tareas genéricas de segundo nivel.

Las fases o niveles que se identifican en esta metodología son: Comprensión del negocio, Comprensión de los datos, Preparación de los datos, Modelado, Evaluación e implementación, tal como se muestra en la Figura 1. Cada una de estas fases (nivel 1) se compone de tareas genéricas (nivel 2), éstas a su vez se dividen en tareas específicas (nivel 3). Además, se encuentra la instancia del proceso que describe las actividades específicas a efectuar (nivel 4).

Esta secuencia de fases no es necesariamente rígida. Cada fase es estructurada con tareas de un segundo nivel, donde se describen las acciones a desarrollar ante situaciones específicas, pero no se indica "cómo" realizarlas. Estas fases son:

Comprensión del negocio: Se enfoca en la comprensión de los objetivos de proyecto y exigencias desde una perspectiva de negocio, convirtiendo luego este conocimiento de los datos en la definición de un problema de minería de datos y en un plan preliminar diseñado para alcanzar los objetivos.

Comprensión de los datos: Se realiza para reconocer los datos, identificar sus problemas de calidad, descubrir los primeros conocimientos ocultos en ellos y/o descubrir subconjuntos interesantes para formar hipótesis.

Preparación de datos: Cubre todas las actividades necesarias para construir el conjunto de datos final de los datos en bruto iniciales. Las tareas incluyen la selección de

tablas, registros y atributos, así como la transformación y la limpieza de datos para las herramientas que modelan.

Modelado: En esta fase se seleccionan y aplican varias técnicas de modelado y se calibran sus parámetros a valores óptimos.

Evaluación: Antes de proceder al despliegue final del modelo, es importante evaluarlo y revisar los pasos ejecutados para crearlo, para comparar el modelo correctamente obtenido con los objetivos de negocio.

Desarrollo: Esta fase puede ser tan simple como la generación de un informe o tan compleja como la realización repetida de un proceso cruzado de minería de datos a lo largo de la empresa [5].

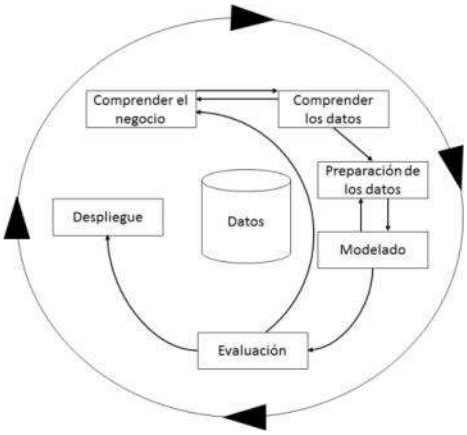

Figura 1. Niveles de la metodología CRISP-DM

2.2 Implementación Metodología

En la fase de comprensión del negocio o del problema, se deben definir los objetivos del trabajo a realizar y convertirlos en un plan de proyecto. Si lo que se desea es obtener el mayor provecho de Data Mining (DM), se debe entender el problema que se quiere resolver.

En general, en las carreras de ingeniería en nuestro país y en América Latina, existe una tasa de abandono muy alta si la comparamos con las Europa y América del Norte. Ésta debe ser analizada para poder ser minimizada. Con el

AJEA - Actas de Jornadas y Eventos Académicos de UTN
DOI: https://doi.org/10.33414/ajea.1146.2022

objetivo de avanzar sobre esta línea, en este trabajo se propuso descubrir qué relación existe entre el abandono de la carrera y la reprobación de exámenes finales, reprobación de exámenes finales de materias homogéneas, nivel de estudios de los padres, cantidad de materias recursadas y otras variables que se enumeran en este trabajo y se toman como variables de entrada.

Se consideran como materias homogéneas, aquellas que son comunes a todas las ingenierías, según definición de la Resolución 68/1994 del Consejo Superior de la Universidad Tecnológica Superior (UTN) [11].

Podemos identificar como objetivo del negocio la relación existente entre las diferentes variables a ser capturadas desde el sistema de gestión académica y la alta probabilidad de abandono del alumno de la carrera. Estas variables son: cantidad de materias aprobadas, cantidad de materias homogéneas aprobadas, cantidad de exámenes reprobados, cantidad de exámenes de materias homogéneas reprobados, vive en Santa Fe, nivel de estudios del padre, nivel de estudios de la madre, cantidad de materias recursadas, año de ingreso, carrera y año de última actividad académica.

Para poder avanzar en este punto, se definió -en acuerdo con la Dirección Académica de la Facultad Regional Santa Fe (FRSF)-, que el objetivo específico es detectar características comunes entre los alumnos que tienen alta probabilidad de abandonar la carrera. Se realizaron entrevistas y analizaron los datos académicos de los alumnos, además de considerar la ordenanza Nº 1549 de la UTN del año 2016 (UTN, Ordenanza 1549, 2016), que describe las siguientes definiciones:

- Estudiante Activo: Estudiante que en un ciclo lectivo aprueba, cursa, se presenta a rendir examen final o asiste regularmente a clases al menos en una asignatura.
- Estudiante Pasivo: Quien no cumple la condición de estudiante activo.

Considerando estos conceptos, se acordó definir como Alumno con Alta Probabilidad de Abandono aquel alumno que durante dos años consecutivos es Estudiante Pasivo.

En el ámbito de la UTN no existe una definición acordada por reglamento en relación con cuándo considerar que un alumno “abandona” la carrera. Sin embargo, el alumno no está obligado a informar esta situación a su casa de estudios en las universidades nacionales, pudiendo retomar los estudios en cualquier momento, lo que dificulta el análisis de esta variable.

Teniendo en cuenta la definición generada en conjunto con la Secretaría Académica de Alumno con Alta Probabilidad de Abandono, el objetivo principal del estudio será encontrar características comunes en estos alumnos con la finalidad de dar soporte a la Dirección Académica de la facultad en la implementación de procesos que disminuyan la tasa de deserción en las carreras de ingeniería.

A partir de la definición de un objetivo principal, podemos indicar como un objetivo secundario del trabajo de DM es analizar si existe relación entre la cantidad total de exámenes reprobados, la cantidad total de exámenes reprobados de materias homogéneas, la cantidad de recursados de materias, la cantidad de recursados de materias homogéneas, además de otros factores que colaboren en el aumento de la probabilidad de que el alumno abandone la carrera, en función de ser un alumno pasivo durante dos años consecutivos. En la elección de las variables de entrada, se trabajó sobre conocimientos adquiridos en la Secretaría Académica y aquellos factores que se consideran influyentes sobre la decisión de abandonar las carreras universitarias.

En la fase de comprensión de datos, de acuerdo con CRISP-DM, comprende la recolección inicial de datos y su análisis inicial, que permite comprenderlos. En esta fase se identifica la calidad de los datos y se verifican las primeras hipótesis.

Se parte en este estudio de un supuesto ampliamente difundido en la Universidad, que indica que la dificultad en aprobar las materias homogéneas de la carrera influye directamente en la deserción del alumno. Teniendo en cuenta esta primera hipótesis de trabajo, se generaron los datos de entrada para el estudio.

En nuestro caso, en esta fase se hizo un estudio de los datos de la Base de Datos que utiliza el Sistema de Gestión Académica de nuestra Universidad (Sysacad), identificando los datos necesarios para el estudio a realizar.

Se decidió considerar los datos de todos los alumnos de las cinco carreras de ingeniería de nuestra facultad: Mecánica, Eléctrica, Industrial, Civil y Sistemas de Información. Además, se tomaron los datos de los alumnos ingresantes a partir del año 2008, ya que ese fue el año en el cual se reformularon todos los planes de estudio de las ingenierías en la UTN.

A partir de este consenso generado respecto de los datos a estudiar, se generaron las consultas para obtener los datos requeridos. Posteriormente, se realizó una exploración de los mismos, realizando verificaciones que permitan validar las consultas generadas. En estas verificaciones se utilizó el Sistema de Gestión Académica para realizar verificaciones cruzadas de los datos obtenidos mediante las consultas SQL y las historias académicas de los alumnos.

Se trabajó con un conjunto de datos de 3028 registros generados al finalizar el año académico 2019 (marzo de 2020), con un registro por alumno ingresante en las carreras de ingeniería entre los años 2008 al 2016. Se descartaron de este trabajo los años académicos 2020 y 2021, ya que, debido al aislamiento definido en nuestro país, se desarrollaron las actividades académicas de forma virtual, y se flexibilizaron las condiciones académicas necesarias para el cursado y para los exámenes finales.

Este trabajo es una primera aproximación al análisis de esta problemática, ya que se pretende extender este estudio

AJEA - Actas de Jornadas y Eventos Académicos de UTN
DOI: https://doi.org/10.33414/ajea.1146.2022

a todas las facultades regionales de nuestro país, con lo cual, la cantidad de registros con la que se trabajará es significativamente mayor.

Como parte del proceso de verificación de la calidad de datos, además, se obtuvieron datos sobre la cantidad de ingresantes por carrera de ingeniería y por año académico, y la cantidad de egresados por carrera y año académico. Los datos obtenidos se compararon con los reportes generados para Rectorado cada año. Esto permitió verificar la consistencia de la información que se ha generado.

El sistema de gestión académica desde el cual se tomaron los datos se utiliza desde el año 2017, por lo cual toda la información fue verificada, controlando la correctitud e integridad con los informes generados a rectorado cada año, desde el 2008 al 2020.

En la etapa de preparación de los datos se procesan los datos para prepararlos para las técnicas de minería de datos. Entre otras tareas, se realiza la selección de los datos, limpieza, generación de variables auxiliares, integración de datos de diferentes orígenes y, de ser necesario, cambios de formato.

Teniendo en cuenta las restricciones y supuestos acordados con la Dirección Académica, se obtiene un lote de 3028 registros de alumnos de ingeniería ingresantes a partir del año 2008 hasta el año 2016, con el siguiente formato:

1. *Carrera*: en la cual se encuentra inscripto el alumno ingresante.
2. *Año de Ingreso*: a la carrera de grado.
3. *Nivel de estudio del padre*: primario incompleto (1), primario completo (2), secundario incompleto (3), secundario completo (4), terciario/universitario incompleto (5), terciario/Universitario completo, posgrado incompleto, posgrado completo.
4. *Nivel de estudio de la madre*: categorizado de igual forma que el anterior.
5. *Vive en Santa Fe*: indica si vive en la ciudad de Santa Fe.
6. *Sexo*: Femenino o Masculino.
7. *Egresado*: indica si ya rindió todas las materias correspondientes para la obtención del título de grado.
8. *Cantidad de recursadas:* cantidad de veces que recursó materias del plan de estudio.
9. *Cantidad de recursadas de materias homogéneas*: cantidad de veces que recursó materias homogéneas.
10. *Cantidad total de exámenes aprobados*: Cantidad de exámenes finales aprobados.
11. *Cantidad total de exámenes de materias homogéneas aprobadas*: Cantidad total de exámenes finales de materias homogéneas aprobadas.
12. *Cantidad total de exámenes reprobados*.
13. *Cantidad total de exámenes de materias homogéneas reprobadas*.
14. *Año de última actividad académica*: Último año en que el alumno realizó algún tipo de actividad académica.
15. *Alumno con alta probabilidad de abandono*: Indica si el alumno tiene o no alta probabilidad de abandono.

Se han formateado los datos de forma que todos sean numéricos. Aquellos datos con valor booleano, tales como *Vive en Santa Fe*, *Egresado*, *Alumno con alta probabilidad de abandono*, se han modificado a formato numérico, siendo Si=1 y No=0.

Esta información, que en muchos casos es información derivada, se ha generado mediante consultas y procedimientos que exploraron y calcularon en caso de ser necesario, los datos de la Base de Datos del Sistema de Gestión Académica.

En el caso de la variable *Egresado*, indica si el alumno aprobó todas las materias necesarias para obtener el título de grado. En el caso de nuestro estudio, hemos trabajado con los títulos finales de las carreras de grado, sin considerar los títulos intermedios.

Se trabajó sobre este conjunto de variables, indicadas desde Dirección Académica, ya que se consideran como factores que pueden afectar el abandono de las carreras. Se consideraron otras variables (cursado simultáneo de otras carreras, situación económica del grupo familiar, etc.), pero al no contar con estos datos actualmente en el Sistema de Gestión Académico, se resolvió trabajar sobre las variables indicadas y considerar los demás en los próximos estudios a realizar

En la fase de Modelado, se seleccionan las técnicas de modelado teniendo en cuenta los siguientes ítems:

- Técnica apropiada para el problema.
- Disponer de datos adecuados.
- Que cumpla con los requisitos del problema.
- Tiempo adecuado para obtener un modelo.
- Conocimiento de la técnica.

Dentro de este proceso de Modelado de datos, se generaron varias notebooks de Python, donde se realizó la exploración y análisis y se construyeron modelos que permitieran descubrir agrupamientos y patrones de comportamiento.

Se utilizaron las bibliotecas de Python: PySpark y Seaborn. Con estas herramientas, se trabajo en Google Colab, para realizar en una primera parte una exploración de los datos, que nos permita descubrir características relevantes en cada agrupamiento de datos. En una segunda

AJEA - Actas de Jornadas y Eventos Académicos de UTN
DOI: https://doi.org/10.33414/ajea.1146.2022

etapa, se genero una red MLP que permitió tener una precisión superior al 85% en el etiquetado de los alumnos con alta probabilidad de abandono.

Una primera notebook verificó la consistencia de los datos obtenidos, comparando cantidad de materias aprobadas, con la cantidad de materias homogéneas aprobadas, cantidad de exámenes reprobados con cantidad de exámenes reprobados de materias homogéneas. Además, se verificó la cantidad de materias aprobadas en aquellos alumnos egresados que se correspondiera con el número de materias del plan de estudio, tanto materias obligatorias como electivas.

Luego, se generó otra notebook para realizar un análisis de los datos, indicando cantidad de registros con alumnos con alta probabilidad de abandono, promedio de cantidad de materias aprobadas, promedio de cantidad de materias reprobadas, cantidad de alumnos que viven en Santa Fe, cantidad de registros con más de 5 exámenes aprobados, lo que nos permitió realizar inspección de los datos. A continuación, en la Tabla 1, se muestra parte de la información obtenida a partir de esta inspección de datos.

Tabla 1. Datos obtenidos en la lectura e inspección

Consulta	Resultado
Cantidad Total de Registros	3028
Cantidad de Registros con más de 5 exámenes reprobados	1133 (37,41%)
Cantidad de Registros con alta probabilidad de abandono	1123 (37,08%)
Cantidad de Registros que han reprobado al menos 1 materia homoguenea	2081 (68,72)
Cantidad promedio de exámenes de materias homogéneas reprobados	3,10
Cantidad de Alumnos egresados	513 (16,94%)

3. Resultados

Realizando un primer análisis de los datos obtenidos (Tabla 2), se puede observar que los alumnos que tienen alta probabilidad de abandono tienen, en promedio, un número bajo de materias aprobadas, de exámenes reprobados, de materias homogéneas aprobadas y exámenes de materias homogéneas reprobadas, en relación con aquellos alumnos que no egresaron aún y no tienen alta probabilidad de abandono, ya que no han sido pasivos durante dos años académicos consecutivos

Se realiza un análisis de correlación para diferentes atributos y se obtienen los datos que se muestran en la Tabla 3, donde se detallan los valores obtenidos en correlación con el atributo *Abandona* y el atributo *Materias Homogéneas Reprobadas*.

Cabe aclarar que, respecto a este análisis de correlación realizado, se trabajó con el archivo completo y luego extrayendo las variables que pudieran actuar por su alta correlación como predictoras: *Año de última actividad académica, exámenes aprobados y Egresado*. El análisis de correlación incluyendo todas las variables y extrayendo las posibles variables predictoras como: anio_ult_actividad y examenes_aprobados, no arrojó diferencias significativas que indiquen un sesgo en el modelo generado. Cuando la correlación es negativa, esto indica que: en tanto el valor en estudio crece, el correlacionado decrece.

Tabla 2. Comparativa de promedio de exámenes en alumnos con alta probabilidad de abandono y activos

Alumnos con Alta Probabilidad de Abandono			
Carrera	**Cantidad Promedio Materias Aprobadas**	**Cantidad Promedio Homogéneas Aprobadas**	**Cantidad Promedio Exámenes Homogéneas Reprobados**
Sist. de Información	5,77	2,16	1,03
Eléctrica	5,89	4,44	1,83
Mecánica	6,17	4,09	2,97
Industrial	10,16	4,86	2,16
Civil	5,48	3,64	2,76
Alumnos No Egresados Activos (sin alta probabilidad de Abandono)			
Carrera	**Cantidad Promedio Materias Aprobadas**	**Cantidad Promedio Homogeneas Aprobadas**	**Cantidad Promedio Exámenes Homogeneas Reprobadas**
Sist. de Información	23,27	7,88	2,79
Eléctrica	28,56	11,24	3,53
Mecánica	23,81	10,75	5,22
Industrial	30,83	10,43	4,15
Civil	28,01	11,26	5,11

Por último, se generó una nueva notebook, para trabajar con agrupamiento de los datos. Se aplicó un modelo de clustering, ya que se pretende encontrar características comunes entre los alumnos con alta probabilidad de abandono.

El algoritmo de Clustering K-Means es uno de los más usados para encontrar grupos ocultos, o sospechados en teoría, sobre un conjunto de datos no etiquetado [12]. Esto puede servir para confirmar o desterrar alguna teoría asumida en relación con los datos. Y también puede ayudar a descubrir relaciones entre conjuntos de datos que de manera manual no se reconocen, que es la situación de nuestro estudio, en donde los datos están etiquetados, pero nos interesa descubrir relaciones entre los agrupamientos. Una vez que el algoritmo se ha ejecutado y se han obtenido las etiquetas, será fácil clasificar nuevos valores o muestras entre los grupos obtenidos. Para poder trabajar con este algoritmo, se trabajo el lote de datos eliminando la etiqueta.

K-Means utiliza un proceso iterativo en el que se van ajustando los grupos para producir el resultado final. Para ejecutar el algoritmo se debe pasar como entrada el

AJEA - Actas de Jornadas y Eventos Académicos de UTN
DOI: https://doi.org/10.33414/ajea.1146.2022

conjunto de datos y un valor de K. El conjunto de datos serán las características o features para cada punto. Las posiciones iniciales de los K centroides serán asignadas de manera aleatoria en cualquier punto del conjunto de datos de entrada y luego, de forma iterativa, se va modificando la posición de este centroide en función de los puntos más cercanos [13].

Tabla 3. Correlación entre los diferentes valores de entrada

Atributo	Correlación con *Abandona*	Correlación con *Materias Homogéneas Reprobadas*
Anio_ingreso	-0.199	-0.098
Nivel_est_padre	-0.203	-0.020
Nivel_est_madre	-0.243	-0.025
Vive_Santa_Fe	0.004	-0.054
Sexo	-0.025	-0.012
Egresado	-0.346	-0.101
Cant_recursadas	-0.217	0.298
Cant_homo_recursadas	-0.037	0.268
Examenes_aprobados	-0.694	0.091
Mat_homog_aprobadas	-0.707	0.263
Examenes_reprobados	-0.314	0.848
Mat_homog_reprobadas	-0.190	1.000
Anio_ult_actividad	0.999	-0.190
Abandono	1.000	-0.190

Al trabajar con K-Means se realizaron diferentes pruebas con diferentes valores de K. Los valores obtenidos con K mayor a 2 no fueron significativos para el estudio. Se muestran en la tabla 4 los valores obtenidos al trabajar con K=2.

Los resultados obtenidos de la aplicación de este algoritmo se agruparon en dos clusters y son los que se muestran en la Tabla 4, donde se puede ver que en el Cluster 1 quedó la mayoría de los alumnos que tienen alta probabilidad de abandono y ningún alumno egresado. En este Cluster sólo hay 167 alumnos que no están caratulados como alumnos con alta probabilidad de abandono. Los valores de los features de la Tabla 4 son los promedios por cluster. En el Cluster 0 han quedado todos los alumnos egresados y sólo 24 alumnos con alta probabilidad de abandono.

Tabla 4. Comparativa de datos promedios en cada Cluster generado

	Cluster 1	Cluster 0
Cantidad registros	1266	1762
Nivel Estudio Padre	4,29	5,78
Nivel Estudio Madre	4,45	6,01
Vive en Santa Fe	0,43	0,44
Cantidad Recursadas	4,41	5,63
Egresado	0,00	0,29
Materias Aprobadas	5,94	34,27
Exámenes Reprobados	2,70	7,46
Materias Homogéneas Aprobabas	3,40	11,60
Exámenes Reprobados Materias Homog	1,89	3,97
Abandona	0,86	0,01

Otro dato interesante es que en el Cluster 1, que se puede caratular como *Cluster de alumnos con alta probabilidad de abandono*, el promedio de nivel de estudios tanto del padre como de la madre es más bajo que en el Cluster 0, que es el *Cluster de Alumnos egresados y con baja probabilidad de abandono de la carrera*. Además, se puede encontrar en el Cluster 1 que la cantidad de veces que el alumno recursa una materia, si se tiene en cuenta la cantidad de materias aprobadas, es significativamente más alto que en el Cluster 0.

La Figura 2 muestra el gráfico de los Clusters generados a partir de la aplicación del algoritmo K-Means. En este caso, se utilizó como una herramienta inicial, que nos permita descubrir posibles relaciones entre las variables.

Finalmente, se corrió un modelo predictivo que permita, a partir del lote de datos generados, predecir cuáles son alumnos con alta probabilidad de abandono en el futuro. Para esto se utilizó un modelo de redes neuronales seleccionando específicamente el perceptrón multicapa que tiene la capacidad de resolver problemas que no son linealmente separables.

Los sistemas neuronales artificiales tienen limitaciones y sólo poseen un parecido superficial con las redes neuronales biológicas. Las redes neuronales, en relación con el procesamiento de información, heredan tres características básicas de las redes neuronales biológicas:

- Paralelismo masivo.
- Respuesta no lineal de las neuronas frente a las entradas recibidas.
- Procesamiento de información a partir de múltiples capas de neuronas.

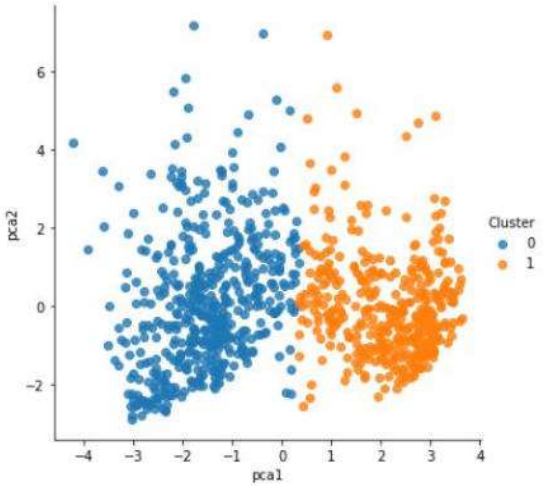

Figura 2. Gráfico de distribución del Cluster 0 y 1

AJEA - Actas de Jornadas y Eventos Académicos de UTN
DOI: https://doi.org/10.33414/ajea.1146.2022

Una de las principales propiedades de estas redes es su capacidad de aprender y generalizar a partir de ejemplos reales: la red aprende a reconocer la relación que existe entre un conjunto de entradas y sus correspondientes salidas. Finalizado el aprendizaje, cuando a la red se le presenta una nueva entrada (siendo ésta incompleta), ella es capaz de generalizarla ofreciendo una salida en base a la relación funcional establecida en el aprendizaje.

En este trabajo se aplicó el perceptrón multicapa, que es un método formalizado por Rumelhart y otros [14]. Ellos generaron un método para que una red del tipo perceptrón multicapa aprendiera la asociación que existe entre un conjunto de patrones de entrada y sus salidas correspondientes. Este método, conocido como backpropagation error (propagación del error hacia atrás) -también denominado método de gradiente decreciente-, ya había sido descrito anteriormente por otros autores, aunque fue el Parallel Distributed Processing Group (grupo PDP), quien realmente lo popularizó [14 - 17].

Para la realización del trabajo se consideró un 70% de los datos para entrenamiento de la red, y el 30% restante para testeo. Luego, se creó el perceptrón multicapa con una primera capa con 7 neuronas, que son los 7 features del conjunto de datos que se tomó como entrada para este algoritmo. Se colocaron 3 capas ocultas con 7, 6 y 7 neuronas cada una respectivamente. El número de neuronas de la última capa se corresponde con el número de etiquetas. En este caso son dos (Alumno con alta probabilidad de abandono, Alumnos sin alta probabilidad de abandono).

Los datos de entrada para el entrenamiento fueron normalizados, dentro del rango [1, -1]. Se trabajo para las capas internas una función de activación sigmoide, y para la última capa se utilizó la función de activación softmax. Además, se ajusto el modelo según el data set de entrenamiento.

El modelo se entrenó con esta configuración y se obtuvo que en el set de datos de testeo la exactitud obtenida es de un 85,02%. Si se modifica la arquitectura de la red neuronal a una sola capa oculta con 5 neuronas, se obtiene que la exactitud en el conjunto de testeo aumenta a un 86,88%. La exactitud con esta última arquitectura mejora tanto con el set de datos de entrenamiento como con el set de datos de testeo. Por todo esto, se puede afirmar que se logró desarrollar una arquitectura cuya exactitud es cercana al 90%.

En la Tabla 5 se muestran los valores de exactitud obtenidos, tanto para el set de entrenamiento, como en el de testeo para las diferentes arquitecturas que se utilizaron.

Tabla 5. Precisión en los diferentes modelos generados

Arquitectura Perceptrón Multicapa	Exactitud Modelo Entrenamiento	Exactitud Modelo Testeo
3 capas ocultas con 7, 6 y 7 neuronas cada una	87,64%	85,02%
1 capa oculta con 5 neuronas	88,35%	86,99%
1 capa oculta con 3 neuronas	84,14%	80,76%

En la etapa de Evaluación, se debe examinar el/los modelo/s teniendo en cuenta los criterios de éxito planteados. Se debe revisar el proceso teniendo en cuenta los resultados obtenidos, de forma de poder repetir los pasos que sean necesarios para detectar si se ha cometido algún error.

En base a estos criterios, se adopta el modelo del perceptrón multicapa con sólo una capa oculta con 5 neuronas, ya que de las diferentes arquitecturas que se probaron fue la que dio como resultado mayor exactitud, tanto para el set de datos de aprendizaje, como para el set de datos de testeo. Este modelo es un modelo predictivo, que permite detectar cuáles serían los alumnos con alta probabilidad de abandono, ya desde momentos iniciales, cuando el alumno comienza el cursado de las carreras de ingeniería.

Luego, en la fase de Despliegue, una vez que el modelo fue construido y validado, se transforma en conocimiento dentro de la organización, además de habilitar la toma de decisiones para mejorar los procesos de negocios.

En el caso de este trabajo, los resultados se comunican a las áreas tácticas y estratégicas de la UTN FRSF, de forma tal que, en el futuro, se puedan implementar estrategias que disminuyan el abandono de la carrera por parte de los alumnos de ingeniería.

Este trabajo constituye una primera aproximación al estudio de las causas por las cuales los alumnos de ingeniería abandonan la carrera. Se debe seguir trabajando con diferentes modelos de minería de datos educacionales y aprendizaje automatizado, que permitan realizar un análisis más profundo y detallado de los datos considerados.

4. Discusión

En este trabajo, hemos presentado un análisis inicial las características de los alumnos de carreras de grado, que presentan a partir de su actuación académica, alta probabilidad de abandonar sus estudios de grado universitario.

Los resultados obtenidos, en los que se ha trabajado con un data set inicial limitado a los alumnos de cinco carreras de ingeniería de la Facultad Regional Santa Fe, ingresantes entre el año 2007 al 2016, nos permiten obtener un modelo predictivo con una exactitud cercana al 87%. Consideramos que realizar este estudio, con un conjunto de datos mayor (todas las facultades regionales), permitirá obtener conocimiento valioso, y con una exactitud en los modelos generados considerablemente mayor.

AJEA - Actas de Jornadas y Eventos Académicos de UTN
DOI: https://doi.org/10.33414/ajea.1146.2022

Hemos explorado, dentro de los diferentes modelos de aprendizaje automático, diferentes arquitecturas de la la red neuronal perceptrón multicapa (MLP), quedando para trabajos posteriores, realizar un análisis completo con otros modelos como: arboles de decisión, redes bayesianas, entre otros.

Consideramos necesario, trabajar con diferentes modelos que permitan realizar un análisis comparativo respecto a la exactitud tanto en el modelo de testeo como de entrenamiento en cada caso, de forma que se pueda definir el modelo adecuado.

Otro aspecto por considerar es sumar mayor cantidad de variables al estudio, que las personas conocedoras del dominio califican importantes, ya que pueden influir directa o indirectamente en la decisión de abandonar la carrera de grado. Para esto, es imprescindible trabajar en conjunto con las autoridades de los sectores académicos de cada facultad, para realizar un relevamiento completo y correcto de estas variables.

Una vez identificadas, se deben definir los procesos de captura de esa información, para luego agregarla, siguiendo los procesos adecuados, que permitan obtener un set de datos iniciales completo, correcto y consistente.

5. Conclusiones

Se puede afirmar, como una primera conclusión del trabajo realizado, que las técnicas de minería de datos educativos brindan valioso conocimiento para entender la problemática de la deserción en las carreras de ingeniería. La posibilidad que brindan de encontrar patrones de comportamiento entre los alumnos sirve para definir políticas de apoyo al estudiante a fin de minimizar la alta probabilidad de abandono de la carrera universitaria.

Además, utilizando modelos predictivos que tengan un alto porcentaje de exactitud como el generado en este trabajo con valor cercano al 90%, la Universidad puede planificar estrategias desde el ingreso mismo de los alumnos, que permitan disminuir de forma efectiva el alto porcentaje de abandono. Específicamente, en este trabajo, se ha detectado que los alumnos con alta probabilidad de abandono recursan un número mayor de materias que el resto del alumnado, y esto puede ser modificado con sistemas de tutorías de alumnos de años superiores, cursos especiales con apoyo docente, etc.

Otra característica detectada es que el nivel de estudio de los padres en aquellos alumnos con alta probabilidad de abandono es menor al nivel de estudio del resto.

Sin embargo, sería importante considerar en futuros trabajos otras variables de interés tales como, por ejemplo, datos socioeconómicos de los alumnos, edad, orientación vocacional recibida y otras, a fin de realizar un análisis más completo del abandono de las carreras.

Por otro lado, el uso de la metodología CRISP-DM, permitió gestionar de forma clara el proceso de minería de datos, ya que el enfoque iterativo que plantea brinda un proceso de mejora continua, que asegura la calidad de los resultados obtenidos.

La aplicación de la Minería de Datos Educativos (EDM) puede ser muy flexible y práctica para descubrir patrones descriptivos de los estudiantes al facilitar el análisis desde diversas perspectivas. Aún más, con los resultados se pueden planificar estrategias de intervención acordes a las necesidades detectadas.

El objetivo principal que se persigue con este trabajo como miembros y actores principales de la comunidad académica universitaria es continuar trabajando para mejorar e igualar las posibilidades de todos los ingresantes, para que puedan finalizar sus estudios de grado.

Referencias

[1] Ferreyra, M., Avitabile, C., Botero, J., Haimovich, F., y Urzúa, S. (2017). At a Cross-roads: Higher Education in Latin America and the Caribbean. Directions in Development. World Bank. https://doi.org/10.1596/978-1-4648-0971-2.

[2] CONFEDI. Día de la Ingenieria Argentina: Se necesitan más graduados, pero de calidad certificada. (https://confedi.org.ar/en-el-dia-de-la-ingenieria-argentina-se-necesitan-mas-graduados-por-ano-pero-de-calidad-certificada/) (2019).

[3] INFOBAE. (https://www.infobae.com/educacion/2018/09/12/brasil-ya-duplica-a-la-argentina-en-cantidad-de-graduados-universitarios-por-habitante/).

[4] Kumar, M., Singh, A. J., & Handa, D. Literature Survey on Student's Performance Pre-diction in Education using Data Mining Techniques. International Journal of Education and Management Engineering, 7(6), 40–49. https://doi.org/10.5815/ijeme.2017.06.05. (2017).

[5] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. CRISP-DM 1.0 Step-by-step data mining guide. Editorial SPSS. (2000).

[6] Menacho, C.H. Predicción del rendimiento académico aplicando técnicas de minería de datos. Anales Científicos, 78(1), 26. https://doi.org/10.21704/ac.v78i1.811 (2017).

[7] García Gutiérrez, J.A. Líneas de investigación en minería de datos en aplicaciones en ciencia e ingeniería: Estado del arte y perspectivas. Universidad Nacional de Educación a Distancia. (2016).

[8] Urbina-Nájera, A.B., Camino-Hampshire, J.C., & Cruz-Barbosa, R. (2020). Deserción escolar universitaria: Patrones para prevenirla aplicando minería de datos educativa. RELIEVE, 26(1), art. 4.

[9] Aguirre Mendiola, J., Valdovino Rosas, R, Velazquez, J., Eleuterio, R. y Romero, J. Análisis de deserción escolar con minería de datos. (2015). DOI: 10.13053/rcs-93-1-6.

[10] Ujkani, B., Minkovska, D. y Stoyanova, L. A Machine Learning Approach for Predict-ing Student Enrollment in the University. Proc. XXX International Scientific Conference

AJEA - Actas de Jornadas y Eventos Académicos de UTN
DOI: https://doi.org/10.33414/ajea.1146.2022

Electronics - ET2021, September 15 - 17, 2021, Sozopol, Bulgaria.

[11] UTN. Consejo Superior. Reglamento de Estudio para todas las Carreras de Grado en la Universidad Tecnológica Nacional. (http://csu.rec.utn.edu.ar/docs/php/salida.php3?tipo=ORD&numero=1549&anio=0&facultad=CSU) (2016).

[12] Aggarwal Charu C., y Reddy Chandan K. Data clustering: algorithms and applications. Chapman & Hall, CRC Press. (2014).

[13] Haykin, S. Neural Networks: A Comprehensive Foundation (2 edición). Prentice Hall. (1998). ISBN 0-13-273350-1.

[14] Rumelhart, D.E., Hinton, G.E. y Williams, R.J. Learning internal representations by error propagation. En D.E. Rumelhart y J.L. McClelland (Eds.), Parallel distributed processing (pp. 318-362). Cambridge, MA: MIT Press. (1986).

[15] Werbos, P.J. Beyond regression: new tools for prediction an analysis in behavioral sciences. Tesis doctoral no publicada. Harvard University. (1974).

[16] Parker, D. Learning logic (Informe técnico Nº TR-87). Cambridge: Center for Computational Research in Economics and Management Science. (1985).

[17] Le Cun, Y. A learning procedure for asymetric threshold network. Proceedings of Cognitiva, 85, 599-604. (1985).

[18] KDnuggets. Poll: Data Mining Methodology. Octubre 2014. Available: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html.