Universidad Católica
de Santiago del Estero
República Argentina

Nuevas Propuestas

ISBN 2683-8044
77 Páginas - Año XLIII - Volumen Nro. 63
Ediciones UCSE 2024
Revista incluida en Catálogo Latindex v1.0

NUEVAS PROPUESTAS, REVISTA DE LA UCSE
Nro.63 - DICIEMBRE 2024

2. Inteligencia Artificial aplicada a Educación Superior: detección de causas de deserción en carreras universitarias

Artificial Intelligence for detection of University dropout

Marcela Andrea Vera
Docente UCSE DAR - UTN
marcela.vera@ucse.edu.ar

Alejandro Aguirre
Docente UCSE DAR
alejandro.aguirre@ucse.edu.ar

Resumen

La problemática de la deserción universitaria representa un flagelo no resuelto en América Latina [1], necesitándose de manera urgente contar con un mejor entendimiento de sus causas, que permitan desarrollar políticas de retención tanto dentro de las universidades, como a nivel público en los diferentes niveles del estado. En ese contexto, la Inteligencia Artificial puede jugar un papel importante a la hora de encontrar patrones y relaciones ocultas en los datos, que permitan reconocer las verdaderas causas y poder establecer predicciones que desemboquen en decisiones más efectivas.

El objetivo general de este trabajo es generar modelos que permitan descubrir aquellos estudiantes que posean altas probabilidades de desertar una carrera universitaria, permitiendo establecer y aplicar acciones de retención. Se utilizaron técnicas de minería de datos y machine learning, usando como fuente de datos la información brindada por el Sistema de Gestión Académica de la Universidad Católica de Santiago del Estero.

Palabras clave: Inteligencia artificial, educación, minería de datos, machine learning, aprendizaje automático, deserción universitaria.

Abstract

The university dropout is an unresolved problem in Latin America [1], and it is urgent to understand its causes, which will allow the development of retention policies both in the university and government. In this context, Artificial Intelligence may play an important role in finding hidden patterns and relationships in data, which will allow us to recognize the true causes and make predictions that lead to more effective decisions.

The general objective of this work is to generate models that allow us to discover those students who have a higher probability of dropping out of a university studies, allowing retention actions. Data mining and machine learning techniques were used, using data provided by the Academic Management System of the Catholic University of Santiago del Estero.

Keywords: Artificial intelligence, education, data mining, machine learning, university dropout

1. Introducción

El fenómeno de la deserción o abandono universitario, empezó a tratarse en Argentina como parte del proceso de evaluación institucional, iniciado en la década del 90. A partir de allí se comenzó a tomar dimensión de la problemática, y reconocer que se trata de una situación que debía ser comprendida, para proponer políticas universitarias que colaboren en disminuir este fenómeno. En la actualidad, se establece un indicador que se trata de la tasa de deserción estudiantil en educación superior, representando uno de los indicadores más utilizados para evaluar la eficiencia de los procesos de enseñanza y aprendizaje de las instituciones terciarias y universitarias. Actualmente alrededor del 30 % de los alumnos de Argentina finalizan sus estudios en tiempo y forma. [1]. La Universidad Católica de Santiago del Estero, no escapa a este fenómeno del crecimiento de los números de la deserción en las diferentes carreras que se dictan.

Éste fenómeno se extiende a nivel regional, ya que Sudamérica completa se caracteriza por tener bajas tasas de graduación en carreras universitarias. Si bien la tasa bruta promedio de matrícula en educación superior de América Latina y el Caribe creció del 17 % en 1991, al 21 % en el año 2000, y al 40 por ciento en el año 2010, finalizan sus estudios superiores solo un 50 % de los matriculados. [2]

Por otro lado, la aplicación de Inteligencia Artificial proporciona el potencial necesario para abordar algunos de los desafíos mayores de la educación actual, ya que va más allá de la automatización. Con su capacidad para personalizar, puede adaptar la educación para satisfacer las necesidades individuales de los estudiantes o los maestros, un avance que trasciende la mera eficiencia técnica. [13]

Finalmente, el objetivo general de nuestro trabajo es mediante el uso de diferentes técnicas y modelos de aprendizaje automático, generar modelos predictivos para reconocer a los alumnos que tengan una alta probabilidad de abandonar sus estudios, de forma que se pueda atender a esta problemática antes de la deserción del mismo.

Para alcanzar dicho objetivo, se utilizaron algoritmos de aprendizaje automático o Machine Learning, que es una rama de la inteligencia artificial que se centra en desarrollar sistemas que aprenden, o que mejoran el rendimiento, en función de los

datos que consumen, y que luego serán utilizados para predecir o ayudar en la toma de decisiones observando nuevos datos [3].

Éste proyecto de investigación provee información clave sobre la problemática planteada, brindando la información necesaria para generar estrategias que permitan evaluar la calidad de los contenidos brindados, y brindar apoyo académico a los estudiantes que lo requieran para lograr finalizar la carrera que esté estudiando. Conocer cuáles son las variables que influyen en la deserción y detectar de forma anticipada cuáles son los alumnos con mayores probabilidades de abandonar la carrera, permitirá definir políticas universitarias específicas que atiendan estas problemáticas.

Los resultados que se obtengan podrán ser de utilidad para otras instituciones similares, permitiendo contribuir en la retención de alumnos y de esta manera poder mejorar la cantidad de egresados, lo que además afecta directamente a las partidas presupuestarias.

2. Metodología

Al formar parte de un proyecto de investigación sobre Minería de Datos, el trabajo se desarrolló utilizando la metodología CRISP-DM, una de las metodologías más utilizadas en proyectos de este tipo.

Esta metodología posee una secuencia de fases no necesariamente rígida, cada una de ellas estructurada con tareas genéricas de segundo nivel, y que a su vez también se pueden dividir en un tercer nivel para describir acciones o tareas específicas a realizar [4]. Estas fases son:

1- **Comprensión del Negocio**: se deben definir los objetivos del trabajo y convertirlos en un plan de proyecto, para así poder comprender el problema que se quiere resolver.

2- **Comprensión de datos**: comprende la recolección inicial de datos y su análisis inicial, permitiendo comprenderlos. Se busca identificar la calidad de los datos, además de verificar las primeras hipótesis.

3- **Preparación de los datos**: se procesan los datos, preparándolos para las técnicas de mine-ría de datos. Entre otras tareas, se realiza la selección y limpieza de los datos, generación de variables auxiliares, integración de datos de diferentes orígenes y, de ser necesario, cambios de formato.

4- **Modelado**: se seleccionan las técnicas de modelado, teniendo en cuenta que sean apropiadas para el problema, que se disponga de datos adecuados, que se cumpla con los requisitos del problema, que se invierta el tiempo adecuado para obtener un modelo y que se tenga un buen conocimiento de la técnica.

5- **Evaluación**: se evalúan el o los modelos teniendo en cuenta los criterios de éxito plantea-dos. Se debe revisar el proceso, observando los resultados obtenidos, para poder repetir los pasos que sean necesarios y revisar si se ha cometido algún error. Esta es la última fase incluida en el presente trabajo.

6- **Despliegue**: una vez que el modelo fue construido y validado, se transforma en conocimiento dentro de la organización, y habilita la toma de decisiones para mejorar los procesos de negocios. En el caso del proyecto del que forma parte este trabajo, se pretende que los resultados sean comunicados a las áreas tácticas y estratégicas de la Universidad Católica de Santiago del Estero, para poder implementar estrategias que disminuyan el abandono por parte de los alumnos de las carreras de grado.

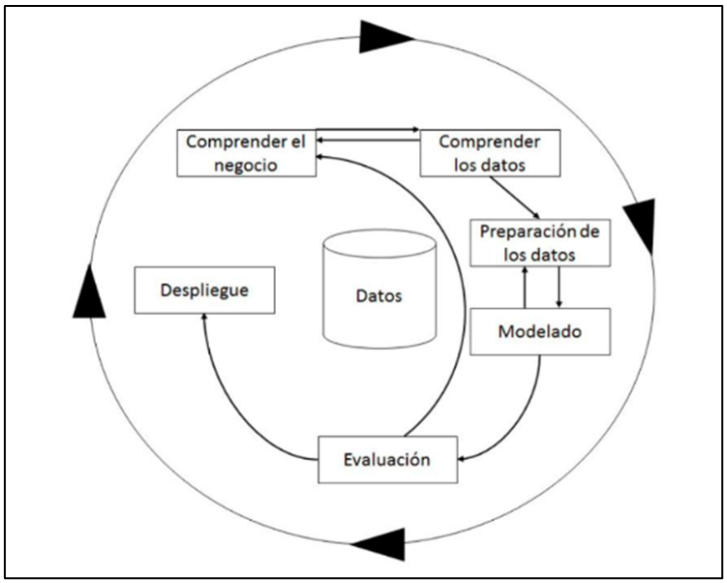

Fig. 1. Fases de la metodología CRISP-DM.

Este trabajo incluyó desde la fase de comprensión de datos hasta la evaluación de los modelos, la fase de despliegue queda por fuera del alcance de este proyecto de investigación.

Fases del Proceso:

2.1. Comprensión del Negocio

La deserción o evasión universitaria es un término que conceptualiza el abandono de los estudios por parte de un estudiante, generando una importante repercusión en la institución que lo albergaba, porque puede presentar mal uso de recursos, ociosidad de profesores y personal y mal uso de la infraestructura disponible. Resulta esencial

trabajar en la retención de los estudiantes, que consiste en acciones para fidelizarlos a lo largo de su trayectoria de estudios, sobre todo en un momento en que el mercado valora cada vez más la formación continua [14].

El punto de partida es, a partir de los datos registrados en el sistema de alumnado de UCSE, relevar las razones por las cuales los alumnos han abandonado la carrera. La problemática principal es la falta de información acerca de esas razones, por lo que se requieren complementarla con instrumentos como pueden ser el envío de encuestas a los alumnos que han abandonado una carrera.

2.2. Comprensión de datos

Inicialmente se comenzó a trabajar con un dataset que fue generado a partir de tres archivos en formato ".xlsx" que se obtuvieron en base al sistema de gestión académica de la Universidad Católica de Santiago del Estero (UCSE). Estos archivos contienen datos sobre los alumnos de la universidad, como carrera, sede, fecha de ingreso, etc., datos de instituciones educativas de nivel secundario a las que asistieron estos alumnos, como nombre de la escuela, ciudad de la misma, si es de gestión privada o estatal, de ámbito rural o urbano, etc., y datos de las diversas carreras que estos están haciendo o hicieron en la universidad, detallando aspectos como el plan, si este está vigente, la unidad académica, la duración en años, etc.

La generación de estos datasets, se realizó en conjunto con el área de informática de UCSE, teniendo en cuenta aspectos que los directores de las diferentes carreras de grado, nos indicaban como variables que podían tener relación la deserción de los estudiantes. Adicionalmente, con el objetivo de trabajar con un conjunto de datos un más ordenado, consistente y con menos valores nulos o erróneos, se tomó la decisión de realizar una encuesta a alumnos y ex alumnos de UCSE de cualquier carrera y sede.

Estos datos se combinaron para generar un solo dataset con toda la información relacionada. Además, se decidió en conjunto con el equipo de trabajo, que el dominio a analizar sería únicamente las carreras de grado, y los alumnos que hayan iniciado sus estudios a partir del año 2000, quedando disponibles, tras la selección de registros en base al nuevo dominio, un total de 25925 registros y 31 columnas para comenzar con el análisis y preprocesamiento de los datos, donde los datos personales fueron ofuscados, de forma de no permitir identificar personas [5].

2.2.1. Variable de salida

Luego, se identificó y seleccionó a la variable 'estado_academico' que representa el estado de un alumno con respecto a la carrera que cursó o está cursando, como la variable objetivo (target).

Inicialmente esta variable contaba con 17 estados posibles, tomados desde el Sistema de Gestión Académica, que se agruparon 3 valores diferentes, representativos del

estado académico del alumno, resultando el 49,47 % del dataset con estado 'Activo', el 39,44 % con estado 'Dado de baja' y el 11,1 % con estado 'Egresado' [5].

2.2.2. Variables de entrada y correlaciones Las demás variables de entrada que se usaron en las fases siguientes fueron: sexo, carrera, anio_nacimiento, sede, unidad_academica, localidad_residencia, hizo_curso_ingreso, nacionalidad, localidad_nacimiento, cant_materias_aprobadas, cant_examenes_reprobados, cant_materias_recursadas, fue_becado, nivel_ultima_cursada, anio_ultima_cursada, tipo_escuela, ambito_escuela, cant_asignaturas_carrera, duracion_carrera, plan_estudio, plan_vigente, anio_ingreso, anio_egreso.

Se generó un mapa de correlación para poder observar qué variables tenían relación con la variable de salida. Se debe destacar aquellas variables que presentaron una correlación más alta, ya sea positiva o negativa, con el target mencionado, tal como se puede observar en la Tabla 1: 'fue_becado', 'nivel_ultima_cursada' y 'cant_materias_aprobadas', entre otras [5].

Tabla 1
Variables de entrada con mayores correlaciones con la variable objetivo.

Variable	**Estado Académico**	**Correlación**	**Relación**
Fue becado	Activo	0.49	Directa o positiva
Fue becado	Dado de baja	-0.49	Indirecta o negativa
Nivel última cursada	Dado de baja	-0.31	Indirecta o negativa
Nivel última cursada	Egresado	0.55	Directa o positiva
Cant. materias aprobadas	Dado de baja	-0.34	Indirecta o negativa
Cant. materias aprobadas	Egresado	0.68	Directa o positiva

2.3. Preparación de los datos

Una vez finalizado el análisis inicial, se continuó con la fase de preprocesamiento de los datos. Esta fase fue dividida en varias partes, incluyendo eliminación de inconsistencias, tratamiento de los valores nulos, creación de nuevas variables que se decidieron a partir del análisis inicial, y se les realizó el mismo análisis que a las variables en la etapa anterior. Por último, se analizaron los valores de correlación con la variable de salida y las nuevas variables creadas y/o algunas de las variables que hayan sido actualizadas a través de los procedimientos anteriores [6].

2.3.1. Nuevas variables y correlaciones

A continuación se procedió a la generación de las nuevas variables, en base a las variables originales: 'duracion_en_carrera', 'edad_ingreso', 'vive_ciudad_de_sede', 'se_mudo', 'carrera_plan', 'carrera_plan_vigente', 'estudia_plan_actual', 'porcentaje_ultima_cursada','porcentaje_materias_aprobadas', 'porcentaje_materias_recursadas', y 'porcentaje_exito_examenes'. A partir de éstas nuevas variables y tras el tratamiento de las originales, se generó un gráfico de correlación en el que se pudo detectar las variables que tuvieron mayores valores de correlación con respecto a la variable de salida 'estado_academico', tal como se puede observar en la Tabla 2: 'duracion_en_carrera', 'anio_ingreso', 'anio_ultima_cursada', 'porcentaje_ultima_cursada', y 'porcentaje_materias_aprobadas', entre otras [6].

Tabla 2

Variables de entrada nuevas con mayores correlaciones con la variable objetivo

Variable	**Estado académico**	**Correlación**	**Relación**
Duración en carrera	Dado de baja	-0.52	Indirecta o negativa
Año ingreso	Activo	0.63	Directa o positiva
Año última cursada	Dado de baja	-0.67	Indirecta o negativa
Año última cursada	Activo	0.74	Directa o positiva
Porcentaje última cursada	Egresado	0.60	Directa o positiva
Porcentaje materias aprobadas	Egresado	0.73	Directa o positiva

2.4. Modelado

Es importante aclarar que el dataset utilizado contiene únicamente datos académicos de un alumno en un momento estático del tiempo, sin considerar la evolución o cambios a lo largo de su trayectoria en la universidad, lo cual pudo haber influido y/o sesgado a los modelos predictivos. Esto también podría afectar la performance en un futuro, en caso de querer replicarlos con otros datos de entrada, ya que se encontrarían limitaciones debido a este posible sesgo introducido durante el entrenamiento con el presente dataset, llegando a generar resultados de menor exactitud.

Por otra parte, también se debe mencionar que se modificó la variable de salida, agrupando los estados 'Activo' y 'Egresado' para llegar a una variable booleana que indique si el alumno se dio de baja (abandono la carrera) o no [7].

2.4.1. Modelos y métricas

En la etapa 4 del modelo CRISP, se seleccionan y desarrollan las técnicas de modelado, teniendo en cuenta que sean las más apropiadas para el problema, que

dispongan de los datos adecuados y que cumplan con los requisitos del problema. Los modelos seleccionados para aplicar aprendizaje supervisado sobre el dataset fueron KNeighbors, Random Forest, Gradient Boosting y MultiLayer Perceptron [7].

KNeighbors permite clasificar valores a partir de los datos más similares o cercanos aprendidos durante el entrenamiento. El único parámetro que utiliza es la cantidad de datos "vecinos" a tener en cuenta para clasificar los grupos [8].

Gradient Boosting combina secuencialmente modelos más débiles, para crear otro más fuerte, ajustando los estimadores en cada iteración, usando los errores del modelo anterior como variable a predecir. Los valores obtenidos se suman y se llega a un resultado más realista [9].

Random Forest consiste en la creación de múltiples árboles de decisión para luego combinar sus resultados tomando valores mayoritarios o promedios, logrando reducir el sesgo del modelo y mejorando la generalización y robustez de las predicciones [10].

MultiLayer Perceptron consiste en una potente red neuronal de múltiples capas conectadas entre sí, de manera que las salidas de algunas neuronas se conviertan en la entrada de otras. Está compuesta por una capa de entrada (con una neurona por cada variable de entrada) y una capa de salida (que entrega el resultado), conectadas entre sí por una o más capas escondidas (en las que se realizan todos los cálculos) [11].

En cuanto a las métricas seleccionadas para la evaluación de estos modelos, principalmente se utilizó Accuracy, que mide el porcentaje de casos clasificados correctamente. Por otra parte, Precision, Recall y F1, también fueron estudiadas para complementar el análisis del rendimiento de dichos modelos [7]. Precision mide el porcentaje de valores que se han clasificado como positivos que son realmente positivos, Recall mide cuántos valores positivos son correctamente clasificados y F1-Score combina Precision y Recall con el objeto de obtener valores más objetivos [12].

2.4.2. Sets, grupos de variables y transformadores

En primer lugar, el dataset se dividió en 2 sets, set de test con un 20 % de los datos para hacer las predicciones finales, y set de train y validation con un 80 % con el que, mediante validación cruzada, se entrenaron los modelos y se hicieron las primeras predicciones [7].

Previo al entrenamiento de estos modelos, se conformaron 5 grupos de variables para comparar el rendimiento de estos según distintos conjuntos de información. Los grupos fueron:

- Grupo 1: todas las variables.
- Grupo 2: sin las nuevas variables creadas.

- Grupo 3: sólo las variables relacionadas a la carrera y al rendimiento académico. Por ejemplo: 'carrera', 'cant_materias_aprobadas', 'nivel_ultima_cursada'.
- Grupo 4: todas las variables, utilizando aquellas creadas en la etapa de preprocesamiento en lugar de las variables que representan, o en las que se basan. Por ejemplo: 'porcentaje_ultima_cursada' en lugar de 'nivel_ultima_cursada'.
- Grupo 5: selección de variables más relevantes mediante SelectFromModel [7].

Luego, se asignó el debido preprocesamiento sobre las variables acorde a su tipo, para poder escalar los distintos grupos de variables, usando herramientas como OneHotEncoder, TargetEncoding y MinMaxScaller [7].

2.4.3. Entrenamiento

El entrenamiento y predicción con los diversos modelos fue llevado a cabo mediante GridSearchCV utilizando Accuracy como métrica. Y para cada tipo de modelo se definieron una serie de hiperparámetros y valores con los cuales evaluó GridSearchCV.

Tras definir los grupos de variables y los modelos, se definieron pipelines que combinaron cada tipo de modelo con cada uno de los grupos de variables.

Finalmente, se entrenaron cada uno de estos pipelines con el set de train y validation, y a partir de los resultados obtenidos, se llevaron a cabo una serie de análisis y evaluaciones, con el objetivo de seleccionar 3 modelos, con los que luego se hicieron predicciones usando el set de test, generando las conclusiones [7].

2.5. Evaluación

Una vez finalizado el modelado, se realizó la etapa 5 del modelo CRISP-DM, es decir, la evaluación. En esta etapa se evalúan los modelos teniendo en cuenta los valores de las métricas seleccionadas previamente [7].

2.5.1. Sesgo

En un primer análisis se detectaron valores de Accuracy muy cercanos a 1 en el set de validation para todos los modelos entrenados. Ante la sospecha de que estos estaban siendo sesgados, se analizó la importancia que diez (10) de los modelos les dieron a las variables según cada grupo. La mayoría de ellos les dieron una importancia considerable a las variables 'duracion_en_carrera' y 'anio_ultima_cursada'. Además, algunos de los hallazgos detectados en la fase de análisis exploratorio demostraron valores de correlación relativamente altos entre estas dos variables y la variable de salida. Por esto, se decidió entrenar nuevamente los modelos, sin incluir dichas variables y así evitar el posible sesgo que se podría haber estado generando en los mismos. Tras ello se pudo ver que los nuevos valores de Accuracy en validation bajaron de 1 a 0,9 aproximadamente, y luego de comprobar nuevamente la importancia de las variables para los mismos diez (10) modelos, pero

entrenados sin 'duracion_en_carrera' y 'anio_ultima_cursada', se pudo verificar que ya no estaban siendo sesgados por alguna variable en particular, ya que ninguna de ellas recibió la suficiente importancia como para sospechar de otro posible caso de bias [7].

2.5.2. Diferencia de las métricas entre train y validation

Luego de esto, se analizaron los valores de las métricas en train y validation. En primer lugar, se observaron las diferencias de las métricas entre ambos sets. Esto es de importancia ya que una gran diferencia entre los sets (alto valor para train y bajo valor para validation) podría implicar que un modelo está sobreentrenando. Se notó que ningún modelo obtuvo grandes diferencias, a excepción de KNeighbors, que, en comparación de los demás, obtuvo diferencias de entre 0.08 y 0.15 para los grupos 1, 2 y 3 en la mayoría de las métricas. A estos les siguieron Random Forest para los grupos 2 y 4, que obtuvieron diferencias de entre 0.05 y 0.08 en la mayoría de las métricas. Esto sugirió que dichos modelos tuvieron más probabilidades de estar sobreentrenado con algunos de los datos de train y no generalizando tan bien como se esperaría con los datos de validation. Por otra parte, los 5 modelos que menores diferencias obtuvieron en todas las métricas fueron Gradient Boosting para el grupo 4, y MultiLayer Perceptron para los grupos 2, 5, 4 y 1, en ese orden. Sin embargo, los valores de las diferencias para los modelos, a excepción de los que mayores valores obtuvieron, variaron tan poco de unos a otros que las variaciones podrían considerarse despreciables [7].

Luego, se analizaron los valores promedio de las métricas en el set de validation, ya que un alto valor podría llegar a indicar un mejor rendimiento para un determinado modelo, comparado con los demás. Se notó que todos los modelos y grupos obtuvieron valores relativamente altos, llegando a alrededor de 0,90 para la mayoría de las métricas. Sin embargo, hubo algunos modelos que obtuvieron para algunas de las métricas valores en torno a 0,85. Los modelos y grupos que menores valores obtuvieron en todas las métricas fueron KNeighbors, MultiLayer Perceptron y Gradient Boosting solo con el grupo 4 de variables. Por otra parte, los 3 modelos y grupos que mayores valores obtuvieron fueron diversos en cada métrica. En Accuracy los mayores valores fueron obtenidos por Gradient Boosting para los grupos 1 y 2, y Random Forest para el grupo 1. En Precision fueron MultiLayer Perceptron para los grupos 2 y 5, y Random Forest para el grupo 5. En Recall fueron Gradient Boosting para los grupos 1, 2 y 3. Y en F1, también fueron Gradient Boosting pero para los grupos 1 y 2, y Random Forest para el grupo 3. No obstante, los valores de validation para los modelos y grupos, a excepción de los que menores valores obtuvieron, variaron tan poco de unos a otros que estas variaciones podrían considerarse despreciables [7].

2.5.3 Valores de las métricas en test

Como se mencionó anteriormente, las diferencias entre los mejores valores obtenidos

por algunos modelos podrían considerarse despreciables. Sin embargo, para hacer predicciones con el set de test, se debieron seleccionar algunos. Es por esto que, a pesar de las pequeñas diferencias, se eligieron los 3 modelos que mayores valores de Accuracy obtuvieron en el set de validation: GradientBoosting-Grupo2, RandomForest-Grupo3 y GradientBoosting-Grupo1.

Tras la predicción de estos modelos con el set de test, se pudo observar cuál o cuáles fueron los que obtuvieron los mayores valores para las métricas, los que se encuentran detallados en las Tablas 4, 5 y 6. El modelo que mayores valores obtuvo en todas las métricas fue Gradient Boosting entrenado con el Grupo 2 de variables. En segundo lugar, quedó Gradient Boosting entrenado con el Grupo 1 de variables en las métricas Accuracy, Precision y F1, quedando tercero en Recall. Y por último, Random Forest entrenado con el Grupo 3 de variables, quedó tercero en las métricas Accuracy, Precision y F1, y segundo en Recall. Sin embargo, la diferencia entre los valores de test fue mínima, al punto en que podría considerarse despreciable, y, en general, todos los modelos obtuvieron valores de 0,92 aproximadamente en Accuracy, lo que quiere decir que los estimadores acertaron en un 92 % de las predicciones. Además, estos valores obtenidos en el set de test, resultaron ser mayores a los valores obtenidos en el set de validation para casi todas las métricas y modelos [7].

Tabla 4
Valores obtenidos en las métricas para cada set en el modelo GradientBoosting-Grupo2

Set	**Promedio Accuracy**	**Promedio Precision**	**Promedio Recall**	**Promedio F1**
Train	0.941918	0.951330	0.900198	0.925055
Validation	0.920864	0.924018	0.873021	0.897759
Test	0.922431	0.926234	0.874877	0.899823

Tabla 5
Valores obtenidos en las métricas para cada set en el modelo RandomForest-Grupo3

Set	**Promedio Accuracy**	**Promedio Precision**	**Promedio Recall**	**Promedio F1**
Train	0.956341	0.968863	0.919924	0.943758
Validation	0.919105	0.926952	0.864879	0.894833
Test	0.919891	0.921762	0.872915	0.896673

Tabla 6
Valores obtenidos en las métricas para cada set en el modelo GradientBoosting-Grupo1

Set	Promedio Accuracy	Promedio Precision	Promedio Recall	Promedio F1
Train	0.943762	0.953225	0.903083	0.927476
Validation	0.918812	0.922871	0.868705	0.894921
Test	0.921258	0.925560	0.872424	0.898207

2.5.4 Matrices de confusión

Un análisis complementario que se hizo sobre las predicciones de los modelos seleccionados fue a través de la generación de matrices de confusión.

Para el modelo GradientBoosting-Grupo2 hubo 1783 instancias positivas (alumnos que se dieron de baja de la carrera) y el modelo las clasificó correctamente como positivas, y 142 casos de falsos positivos, es decir instancias que eran negativas (alumnos que no se dieron de baja de la carrera), pero que el modelo las clasificó incorrectamente como positivas. Por otra parte, hubo 2938 instancias negativas y el modelo las clasificó correctamente como negativas, y 255 casos de falsos negativos, es decir, instancias que eran positivas pero que el modelo las clasificó incorrectamente como negativas.

Para el modelo RandomForest-Grupo3 hubo 1779 instancias que eran positivas y el modelo las clasificó correctamente, y 151 casos de falsos positivos. Por otra parte, hubo 2929 instancias que eran negativas y el modelo las clasificó correctamente, y 259 casos de falsos negativos.

Para el modelo GradientBoosting-Grupo1 hubo 1778 instancias que eran positivas y el modelo las clasificó correctamente, y 143 casos de falsos positivos. Por otra parte, hubo 2937 instancias que eran negativas y el modelo las clasificó correctamente, y 260 casos de falsos negativos.

En resumen, los 3 modelos tuvieron mayor proporción de falsos negativos que falsos positivos, ya que obtuvieron en promedio un 12,7 % de falsos negativos (alumnos que SÍ se dieron de baja predichos como que NO se dieron de baja) y un 4,7 % de falsos positivos (alumnos que NO se dieron de baja predichos como que SI se dieron de baja) [7].

Conclusiones

El proceso de análisis exploratorio y de preprocesamiento de los datos realizado en esta investigación, reveló información valiosa sobre los conjuntos de datos analizados, destacando las variables con mayor correlación, como 'anio_ultima_cursada',

'nivel_ultima_cursada', 'cant_materias_aprobadas', 'porcentaje_ultima_cursada', 'porcentaje_materias_aprobadas', 'duracion_en_carrera' y 'fue_becado'. Estas fases permitieron comprender en profundidad la estructura de los datos con los que se trabajó, enfrentando desafíos y complicaciones varias debido a la presencia de inconsistencias, errores y gran cantidad de valores nulos, que requirieron ser abordados estratégicamente para evitar posibles sesgos en las fases posteriores. Para esto, se utilizaron diversas técnicas como eliminación directa, cálculos estadísticos (moda, media, etc.), consultas a los responsables de los datos, entre otras.

En cuanto a los modelos desarrollados, los algoritmos Gradient Boosting y Random Forest fueron los más efectivos, alcanzando los mayores valores en las métricas de evaluación seleccionadas, ya que particularmente Gradient Boosting entrenado con los grupos 1 y 2 de variables, y Random Forest entrenado con el grupo 3 de variables, alcanzaron alrededor de un 92 % de Accuracy.

En resumen, este trabajo ha permitido no solo comprender la estructura presente en el conjunto de datos y detectar variables importantes, sino también desarrollar modelos predictivos en base al mencionado dataset generado a partir del sistema académico, con un nivel de confianza superior al 90 %, para la identificación temprana de estudiantes en riesgo de abandono.

La implementación futura de estos modelos y el aprovechamiento del conocimiento generado contribuirán al entendimiento de los factores que inciden en el abandono universitario, y permitir intervenciones más informadas y efectivas en el contexto educativo.

Líneas futuras Si bien el desarrollo de esta investigación ha generado un amplio conocimiento acerca de la deserción universitaria en UCSE, hay muchos caminos a seguir para mejorar y potenciar estos resultados. Algunas líneas futuras que se plantean para aumentar el impacto de este trabajo en el ámbito educativo son:

- Profundizar en la experimentación mediante la exploración de una gama más amplia de modelos e hiperparámetros, con el objetivo de mejorar los valores de las métricas obtenidos en los modelos predictivos.
- Recolectar registros temporales de los estudiantes, es decir, información del progreso del alumno a lo largo del tiempo, para reentrenar y ajustar los modelos predictivos con datos más completos y representativos. Esto podría ser trabajado, por ejemplo, a partir de la incorporación de series temporales, que permitirían obtener y analizar datos de los alumnos en momentos determinados y así poder tener más información sobre su paso por la carrera.
- Seleccionar el modelo predictivo más efectivo para avanzar a la fase de despliegue según CRISP-DM, lo que implicaría traducir los resultados del modelo en conocimiento accionable dentro de la institución, permitiendo así

la toma de decisiones estratégicas dirigidas a reducir la problemática de la deserción estudiantil.

- Colaborar con otras instituciones educativas para compartir buenas prácticas, experiencias, conocimiento y cualquier otro tipo de dato y/o hallazgos sobre la retención estudiantil. Esta colaboración podría generar y enriquecer futuros datasets con una mayor diversidad de contextos educativos, contribuyendo a la creación de modelos predictivos más generalizables y efectivos.

Referencias

1. Cassina, P., Giay, F., Knoll, G., & Vera, M. (2024). Análisis de la deserción en las carreras universitarias de UCSE: Construcción de modelos predictivos utilizando técnicas de aprendizaje automático. JAIIO, Jornadas Argentinas de Informática, 10(5), 23-36. https://revistas.unlp.edu.ar/JAIIO/article/view/17986

2. Ferreyra, M. M., Avitabile, C., Botero Álvarez, J., Haimovich Paz, F., & Urzúa, S. (2017). Momento decisivo: la educación superior en América Latina y el Caribe. http://hdl.handle.net/10919/83253

3. ¿Qué es el aprendizaje automático? Oracle Cloud Applications and Cloud Platform. [En línea]. Disponible: urlhttps://www.oracle.com/ar/artificial-intelligence/machine-lear-ning/what-is-machine-learning/. [Último acceso: 10 Febrero 2024]

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

5. Cassina, P., Giay, F., Knoll, G. "Trabajo Final de Carrera: Análisis Exploratorio". Google Colaboratory. [Notebook]. No Disponible. [Último acceso: 1 Marzo 2024]

6. Cassina, P., Giay, F., Knoll, G. "Trabajo Final de Carrera: Preprocesamiento de Datos". Google Colaboratory. [Notebook]. No Disponible. [Último acceso: 1 Marzo 2024]

7. Cassina, P., Giay, F., Knoll, G. "Trabajo Final de Carrera: Modelado y Evaluación". Google Colaboratory. [Notebook]. No Disponible. [Último acceso: 1 Marzo 2024]

8. “Clasificar con K-Nearest-Neighbor ejemplo en Python”. Aprende Machine Learning, 10 Julio 2018. [En línea]. Disponible: https://www.aprendemachinelearning.com/clasificar-con-k-nearest-neighbor-ejemplo-en-python/. [Último acceso: 10 Febrero 2024.]

9. “Gradient Boosting in ML”. GeeksforGeeks. [En liínea]. Disponible: https://www.geeksforgeeks.org/ml-gradient-boosting/. [Último acceso: 10 Febrero 2024]

10. “Random Forest. Interactive Chaos”. [En liínea]. Disponible: https://interactivechaos.com/es/wiki/random-forest. [Último acceso: 10 Febrero 2024]

11. “Multi-Layer Perceptron Learning in Tensorflow. GeeksforGeeks”. [En liínea]. Disponible: https://www.geeksforgeeks.org/multi-layer-perceptron-learning-in-tensorflow/. [Último acceso: 10 Febrero 2024]

12. “Métricas de Clasificación”, Roberto Díaz. [En liínea]. Disponible: `https://www.themachinelearners.com/metricas-de-clasificacion`. [Último acceso: 15 Febrero 2024].

13. “¿Cómo integrar a la inteligencia artificial en la educación de manera responsable?” Disponible: `https://blogs.iadb.org/educacion/es/inteligencia-artificial-educacion/` [Último acceso: 24 Julio 2024]

14. “Deserción universitaria: ¿cuáles son las razones y cómo prevenirla?” `https://www.sydle.com/es/blog/desercion-universitaria-639a22f22ff02745fa4eface` [Último acceso: 25 Julio 2024]

Regresar al Sumario