Análisis de la deserción en las carreras universitarias de UCSE

Construcción de modelos predictivos utilizando técnicas de aprendizaje automático

Paula Cassina[1], Florencia Giay[1], Gonzalo Knoll[1], Marcela Vera[2]

1 Ingeniería en Informática, Universidad Católica de Santiago del Estero, Departamento Académico Rafaela, Santa Fe, Argentina

2 Profesor Asociado, Universidad Católica de Santiago del Estero, Departamento Académico Rafaela. Profesor Adjunto, Universidad Tecnológica Nacional, Facultad Regional Santa Fe

Abstract. El presente trabajo, forma parte de un proyecto de investigación y desarrollo de la Universidad Católica de Santiago del Estero (UCSE), aborda la problemática de la deserción de alumnos en las carreras de dicha universidad, generando modelos que permitan descubrir aquellos estudiantes que se encuentren en una situación en la que posean altas probabilidades de desertar; con el objetivo de poder prever estos casos, e intentar accionar ante estas circunstancias, permitiéndoles a las autoridades definir estrategias para evitarlo y así poder disminuir la deserción.
Para lograr alcanzar estos objetivos se aplicaron técnicas de minería de datos y machine learning centradas en el aprendizaje automático, las cuales permitieron generar modelos predictivos para discernir y predecir de la manera más certera posible, si un estudiante abandonará una determinada carrera o no. Los modelos utilizados fueron KNeighbors, Random Forest, Gradient Boosting y Multilayer Perceptron, usando como fuente de datos, un dataset generado a partir de información brindada por el Sistema de Gestión Académica de la Universidad Católica de Santiago del Estero.

Keywords: Minería de datos educacionales. Aprendizaje automático. Modelos predictivos.

1 Introducción

1.1 Problemática

Según el anteproyecto [1], Sudamérica se caracteriza por tener bajas tasas de graduación en carreras universitarias, ya que, 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 [2], finalizan sus estudios superiores

solo un 50% de los matriculados. En Argentina alrededor del 30% de los alumnos finalizan sus estudios en tiempo y forma.

"A partir de la década del '80 se inició en las universidades de todo el mundo la preocupación por la calidad del servicio educativo que prestan. Esto dio lugar a procesos de evaluación a fin de detectar las debilidades y fortalezas institucionales y generar acciones correctivas de las deficiencias encontradas. En Argentina, en la década del '90, el Estado Nacional incluye en su agenda de política educativa la evaluación de la calidad del accionar universitario, y la mayoría de las universidades nacionales inician procesos de evaluación institucional. Dentro de este proceso de evaluación, se comenzó a revisar el fenómeno de la deserción o abandono universitario, como un problema que debía ser comprendido, para proponer políticas universitarias que colaboren en disminuir este fenómeno".

"En la actualidad, la tasa de deserción estudiantil en educación superior es uno de los indicadores más utilizados a nivel internacional para evaluar la eficiencia de los procesos de enseñanza y aprendizaje de las instituciones terciarias y universitarias" [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.

1.2 Objetivos

Como resultado principal de nuestro trabajo mediante el uso de diferentes técnicas y modelos de aprendizaje automático, se pretende 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.

1.3 Marco teórico

Para alcanzar los objetivos planteados, se utilizaron algoritmos de aprendizaje automático, definido como 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].

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 que son: Comprensión del Negocio, Comprensión de datos, Preparación de los datos, Modelado, Evaluación y Despliegue [4]. Este trabajo incluyó desde la fase de comprensión de datos hasta la evaluación de los modelos.

2 Análisis exploratorio

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. 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.

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 serían 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 [5].

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 target.

Inicialmente esta variable contaba con 17 estados posibles, que se decidieron clasificar y agrupar en 3, quedando 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 Variables de entrada

Las demás variables de entrada que se usaron en las fases siguientes fueron:

- 'sexo': representa el sexo del alumno con 'F' (femenino) y 'M' (masculino).
- 'carrera': representa la carrera que cursa o cursó un alumno.
- 'anio_nacimiento': representa el año de nacimiento de un alumno.
- 'sede': representa la ubicación de la institución en la cual un alumno cursó o está cursando una carrera.
- 'unidad_academica': representa la facultad a la que pertenece una determinada carrera, es decir una subdivisión en la que se divide una universidad donde se dictan estudios especializados en alguna materia o rama del saber. Sin embargo, en el dataset, solo la sede de Santiago del Estero está dividida en facultades.
- 'localidad_residencia': representa el lugar de residencia del alumno al cursar.
- 'hizo_curso_ingreso': representa si un alumno realizó el curso de ingreso o no para una determinada carrera; mostrando 'CURSO DE INGRESO' en el caso de que lo haya hecho y 'SIN CURSO DE INGRESO' en el caso de que no.
- 'nacionalidad': representa el país de nacimiento de un alumno.
- 'localidad_nacimiento': representa la localidad de nacimiento de un alumno.
- 'cant_materias_aprobadas': representa la cantidad de materias o asignaturas aprobadas o acreditadas de un alumno para una determinada carrera y plan.
- 'cant_examenes_reprobados': representa la cantidad de exámenes reprobados de un alumno para una determinada carrera y plan.

- ‘cant_materias_recursadas’: representa la cantidad de materias o asignaturas recursadas por un alumno en una determinada carrera y plan.
- ‘fue_becado’: representa si un alumno fue becado con ‘1’, o si no lo fue con ‘0’.
- ‘nivel_ultima_cursada’: representa el año más avanzado de una carrera en el que un alumno haya cursado o esté cursando por lo menos una materia. Indica el nivel más alto a pesar de estar haciendo materias.
- ‘anio_ultima_cursada’: representa el último año calendario en que un alumno cursó una determinada carrera.
- ‘tipo_escuela’: representa el tipo escuela secundaria a la cual asistió un alumno, siendo "Estatal" en caso de que se trate de una escuela pública, "Social/Cooperativa" en caso de que sea de gestión social o cooperativa, o bien "Privada" en caso de que sea gestión privada.
- ‘ambito_escuela’: representa si la escuela secundaria a la cual asistió un alumno se encuentra en zona urbana o rural, mediante "Urbano" y "Rural".
- ‘cant_asignaturas’: representa la cantidad de materias o asignaturas correspondientes a una determinada carrera y plan.
- ‘duracion_carrera’: representa la cantidad de años que, en teoría, dura una determinada carrera con cierto plan de estudio.
- ‘plan_estudio’: representa el plan de estudio (año) de la carrera que cursó o está cursando un alumno.
- ‘plan_vigente’: representa el último plan actualizado de la carrera que cursó o está cursando un alumno, o si ya no se dicta lo describe como "No vigente".
- ‘anio_ingreso’: representa el año en el que un alumno comienza una carrera.
- ‘anio_egreso’: representa el año de egreso de los alumnos que hayan finalizado una determinada carrera.

Para cada variable, se realizó un análisis que comprendía diversas tareas como:

1. Definir la variable, qué representa y cuáles son sus posibles valores.
2. Verificar el tipo de dato.
3. Contar los valores nulos, calculando la proporción del conjunto total de datos, y establecer la forma en la que podían ser rellenados.
4. Obtener la distribución de los posibles valores de la variable, realizando una breve conclusión acerca de la misma.
5. Generar un conjunto de gráficos que permitan visualizar y en consecuencia realizar un análisis de la relación que tiene esta variable con la variable de salida, ‘estado_academico’.
6. A partir de todo esto, decidir si la variable será considerada o no en la generación de los modelos, y si fuera necesario, realizar un análisis sobre posibles transformaciones para preparar la variable para las próximas etapas.
7. También definir la creación de nuevas variables con el objetivo de mejorar la información que recibirán los modelos.
8. Por último, buscar posibles inconsistencias, y seleccionar la manera en la que estas serán resueltas o tratadas [5].

2.3 Correlaciones

Luego de realizar el análisis sobre cada una de las variables, se generó un mapa de correlación para poder observar qué variables tenían relación con la variable de salida. Se debe destacar a 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].

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

Variable	Estado_academico	Correlación	Relación
Fue_becado	"Activo"	0,49	Directa o positiva
Fue_becado	"Dado de baja"	-0.49	Indirecta o negativa
Nivel_utlima_cursada	"Dado de baja"	-0.31	Indirecta o negativa
Nivel_ultima_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 postiva

3 Preprocesamiento

Una vez finalizado el análisis inicial, se continuó con la fase de preprocesamiento de los datos. Esta fase fue dividida en varias partes:

1. Primero se realizó la eliminación y/o corrección de registros por inconsistencias detectadas en la etapa anterior.
2. Después, se siguió con el tratamiento de los valores nulos en cada variable, rellenando o eliminando según cada caso, teniendo en cuenta la importancia de la variable, la posibilidad de llenarlos sin introducir sesgo, entre otros criterios.
3. Luego, se crearon 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.
4. Por último, se hicieron el resto de los cambios y procesamiento de las variables, como cambiar tipos de datos, modificar valores por otros que sirvan mejor para los modelos, etc.
5. También 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].

3.1 Nuevas variables

Las nuevas variables creadas en esta fase, en base a las variables originales, fueron:

- 'duracion_en_carrera': representa el tiempo de permanencia de los alumnos en una carrera, calculado en base a 'estado_academico', restando 'anio_ingreso' a 'anio_egreso' cuando el estado era 'Egresado', restando 'anio_ingreso' a la fecha en la que se obtuvieron los datos cuando el estado era 'Activo', y restando 'anio_ingreso' a 'anio_ultima_cursada' cuando el estado era 'Dado de baja'.

- ‘edad_ingreso’: representa la edad de un alumno al momento de ingresar a la universidad, que se calculó restando ‘anio_nacimiento’ a ‘anio_ingreso’.
- ‘vive_ciudad_de_sede’: representa si un alumno vive en la misma localidad que la sede mientras está estudiando, completando esta variable con el valor 1, en el caso de que ‘localidad_residencia’ coincida con ‘sede’, y con 0 de no ser así.
- ‘se_mudo’: representa si el alumno se tuvo que mudar para estudiar, completando esta nueva variable con el valor 0 si 'localidad_nacimiento' y 'localidad_residencia' coinciden, y con 1 de no ser así.
- ‘carrera_plan’: representa la combinación de ‘carrera’ y ‘plan_estudio’.
- ‘carrera_plan_vigente’: representa la combinación de ‘carrera’ y ‘plan_vigente’.
- ‘estudia_plan_actual’: representa si el plan de la carrera que hizo o está haciendo el alumno es el actual, completando esta nueva variable con el valor 1 si 'plan_estudio' y 'plan_vigente' coinciden, y con 0 de no ser así.
- ‘porcentaje_ultima_cursada’: representa la relación porcentual que tienen los valores de 'nivel_ultima_cursada' con respecto a la variable 'duracion_carrera'.
- ‘porcentaje_materias_aprobadas’: representa la relación porcentual de los valores de 'cant_materias_aprobadas' con respecto a la variable 'cant_asignaturas'.
- ‘porcentaje_materias_recursadas’: representa la relación porcentual de los valores de 'cant_materias_recursadas' con respecto a la variable 'cant_asignaturas'.
- ‘porcentaje_exito_examenes’: representa la relación porcentual de los valores de 'cant_examenes_reprobados' con respecto a 'cant_materias_aprobadas' [6].

3.2 Correlaciones

Como se mencionó anteriormente, a partir de las nuevas variables creadas y tras el tratamiento de las originales, también se generó un gráfico de correlación, a partir del cual se pudieron 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].

Table 2. Variables de entrada nuevas con mayores correlaciones con la variable objetivo.

Variable	Estado_academico	Correlación	Relación
Duracion_en_carrera	“Dado de baja”	-0.52	Indirecta o negativa
Anio_ingreso	“Activo”	0.63	Directa o positiva
Anio_ultima_cursada	“Dado de baja”	-0.67	Indirecta o negativa
Anio_ultima_Cursada	“Activo”	0.74	Directa o positiva
Porcentaje_ultima_cursada	“Egresado”	0.6	Directa o positiva
Porcentaje_materias_aprobadas	“Egresado”	0.73	Directa o postiva

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 menos acertados y realistas.

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 o no [7].

4.1 Modelos

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].

4.2 Métricas

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].

4.3 Sets

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

4.4 Grupos de variables

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].

4.5 Transformadores

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].

4.6 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].

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].

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].

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].

5.3 Valores de las métricas en validation

Luego, se analizaron los valores promedio de las métricas en el set de validation, detallados en la Tabla 3, 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].

Table 3. Valores obtenidos en las métricas para el set de validation con los modelos entrenados.

Modelo	Grupo Variables	Promedio Accuracy	Promedio Precission	Promedio Recall	Promedio F1
Gradient Boosting	Grupo 1	0.918812	0.922871	0.868705	0.894921
Gradient Boosting	Grupo 2	0.920864	0.924018	0.873021	0.897759
Gradient Boosting	Grupo 3	0.917835	0.922546	0.866428	0.893578
Gradient Boosting	Grupo 4	0.882321	0.870071	0.827927	0.848476
Gradient Boosting	Grupo 5	0.916760	0.932075	0.853051	0.890798
KNeighbors	Grupo 1	0.887549	0.877686	0.833794	0.855140
KNeighbors	Grupo 2	0.893948	0.874755	0.856257	0.865385
KNeighbors	Grupo 3	0.901226	0.892278	0.855059	0.873271
KNeighbors	Grupo 4	0.859655	0.844978	0.792870	0.818089
KNeighbors	Grupo 5	0.907528	0.910916	0.850939	0.879895
MultiLayer Perceptron	Grupo 1	0.914220	0.930668	0.848086	0.887225
MultiLayer Perceptron	Grupo 2	0.913585	0.933587	0.843449	0.885951
MultiLayer Perceptron	Grupo 3	0.915148	0.919467	0.862355	0.889973
MultiLayer Perceptron	Grupo 4	0.877632	0.866986	0.818793	0.841992
MultiLayer Perceptron	Grupo 5	0.912413	0.941041	0.832289	0.883093
Random Forest	Grupo 1	0.918177	0.929518	0.859590	0.893183
Random Forest	Grupo 2	0.917786	0.925821	0.862564	0.893071
Random Forest	Grupo 3	0.919105	0.926952	0.864879	0.894833
Random Forest	Grupo 4	0.881198	0.871834	0.822476	0.846420
Random Forest	Grupo 5	0.917688	0.934167	0.853349	0.891916

5.4 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.

Si bien Accuracy fue designada como la métrica principal a analizar, también se hizo un análisis con las demás y es por esto que se pudieron tener en cuenta en este análisis. Por eso se agregó que justamente estos 3 modelos también fueron de los que obtuvieron los mayores valores en validation para F1. Estos modelos además fueron los que tuvieron mejores valores para la métrica Recall, con excepción de RandomForest-Grupo3.

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].

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

Set	Promedio Accuracy	Promedio Precission	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

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

Set	Promedio Accuracy	Promedio Precission	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

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

Set	Promedio Accuracy	Promedio Precission	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

5.5 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].

6 Conclusiones

6.1 Conclusiones generales

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.

6.2 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. M. Vera, “Análisis de la deserción en las carreras universitarias de UCSE: construcción de un modelo predictivo utilizando técnicas de aprendizaje automático”, Universidad Católica de Santiago del Estero, 2021.
2. M. M. Ferreyra, C. Avitabile, J. Botero Álvarez, et al., “Momento decisivo: La educación superior en América Latina y el Caribe”, Washington DC: Grupo Banco Mundial, 2017.
3. “¿Qué es el aprendizaje automático?” Oracle | Cloud Applications and Cloud Platform. [En línea]. Disponible: https://www.oracle.com/ar/artificial-intelligence/machine-learning/what-is-machine-learning/. [Último acceso: 10 Febrero 2024]
4. P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, R. Wirth, "CRISP-DM 1.0 Step-by-step data mining guide", Editorial SPSS, 2000.
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 línea]. Disponible: https://www.geeksforgeeks.org/ml-gradient-boosting/. [Último acceso: 10 Febrero 2024]
10. “Random Forest”. Interactive Chaos. [En línea]. Disponible: https://interactivechaos.com/es/wiki/random-forest. [Último acceso: 10 Febrero 2024]
11. “Multi-Layer Perceptron Learning in Tensorflow”. GeeksforGeeks. [En línea]. Disponible: https://www.geeksforgeeks.org/multi-layer-perceptron-learning-in-tensorflow/. [Último acceso: 10 Febrero 2024]
12. “Métricas de Clasificación”, Roberto Díaz. [En línea]. Disponible: https://www.themachinelearners.com/metricas-de-clasificacion. [Último acceso: 15 Febrero 2024].