Informe Entrega Final – Proyecto Ciencia de Datos Aplicada

1. Definición de la problemática y entendimiento del negocio

Objetivo del Proyecto: Desarrollar un modelo predictivo que identifique a los estudiantes con mayor probabilidad de desertar, utilizando técnicas de Machine Learning y análisis de datos históricos. La deserción estudiantil es una problemática crítica, ya que afecta tanto a los estudiantes como a la universidad en términos de retención, reputación y éxito académico.

Contexto de la problemática: La universidad ha identificado la deserción temprana como un desafío que compromete los recursos y el cumplimiento de su misión educativa. A través del modelo predictivo, la universidad podrá implementar medidas proactivas para apoyar a los estudiantes en riesgo, promoviendo un entorno de bienestar integral y ayudando a mejorar sus índices de permanencia y éxito.

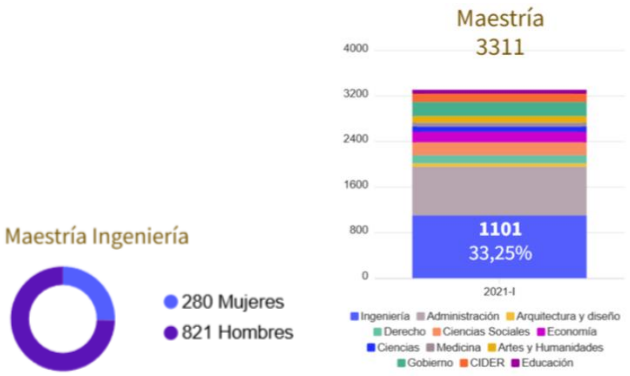


Figura 1. *Datos de 2021-1 para la maestría en Ingeniería y otras*

Relevancia del problema: El alto índice de deserción impacta negativamente tanto en la sostenibilidad del programa de maestría como en el bienestar de los estudiantes. La falta de recursos financieros ha sido identificada como una causa significativa de deserción, alineándose con la hipótesis de que los estudiantes con dificultades económicas son más propensos a abandonar sus estudios.

Indicadores clave (KPIs) del plan de desarrollo 2020-2025 (Vamos a basarnos en los de la facultad de Ingeniería):

- **Reducción de la deserción acumulada:** Meta de reducir el índice de deserción del 29% al 20%.
- **Disminución de la deserción en los primeros 3 semestres:** Reducir la tasa del 10% al 5%.

Estos indicadores son prioritarios y directamente vinculados al proyecto, ya que permiten evaluar el impacto de las estrategias diseñadas para reducir la deserción estudiantil en sus etapas iniciales y acumuladas. Ambos se alinean con el objetivo del modelo predictivo, enfocado en identificar estudiantes en riesgo y facilitar intervenciones tempranas.

En otras palabras, estos KPIs están alineados con el plan estratégico de la universidad, que busca fortalecer el sistema de bienestar estudiantil y reducir los factores de riesgo de deserción a través de intervenciones tempranas y apoyo financiero. [1]

2. Ideación

Producto a desarrollar: El modelo predictivo de deserción estudiantil es un producto de datos diseñado para identificar a los estudiantes en riesgo, permitiendo a la universidad implementar intervenciones efectivas y dirigidas. Este modelo aprovechará la ciencia de datos para ayudar a reducir la deserción y mejorar la experiencia de los estudiantes, creando un entorno de éxito académico.

Usuarios potenciales:

1. Administradores académicos:
   Necesidades: Analizar indicadores globales y tendencias de deserción para desarrollar políticas estratégicas.
   Personalización: Acceso a estadísticas generales, comparativas entre programas y facultades.
2. Consejeros estudiantiles y oficinas de bienestar:
   Necesidades: Identificar estudiantes específicos en riesgo para realizar seguimiento personalizado.

Personalización: Filtros detallados por variables académicas (programa, PGA y % de créditos aprobados) y demográficas (edad).

3. Decanatura y Comité de Alertas Tempranas:
   Necesidades: Identificar estudiantes específicos en riesgo para realizar seguimiento personalizado.
   Personalización: Filtros detallados por variables sociodemográficas y académicas.

Requerimientos del producto:

1. Fuente de datos: Recopilación de información académica, sociodemográfica, y financiera (ej. becas y formas de pago), integrando bases de datos internas de la universidad.
2. Modelo predictivo: Algoritmos de clasificación para identificar el riesgo de deserción basándose en los datos históricos de los estudiantes.
3. Dashboard de visualización: Un panel interactivo adaptado a cada grupo de usuarios, con alertas, filtros y herramientas específicas que faciliten la toma de decisiones y priorización de intervenciones.

Procesos actuales y desafíos:

1. Identificación manual de estudiantes en riesgo: Es reactivo y poco efectivo, ya que depende de alertas tardías.
2. Asignación ineficiente de becas: No siempre se dirige a los estudiantes con mayor riesgo de deserción por dificultades económicas.
3. Falta de seguimiento proactivo: No existen mecanismos predictivos que prioricen los recursos en los estudiantes en riesgo temprano.

Componentes tecnológicos:

1. Integración de datos: Consolidación de bases de datos en formato SQL o BigQuery para obtener un perfil integral de cada estudiante.
2. Motor analítico: Desarrollo del modelo en Python (scikit-learn, TensorFlow) para analizar factores de riesgo y predecir la probabilidad de deserción.
3. Visualización: Implementación de dashboards en herramientas como Power BI o Tableau para seguimiento y visualización de los KPIs y alertas de deserción.

Mockup: Ver Anexo 1.

Este mockup de tablero de control organiza los datos clave en cinco categorías: variables sociodemográficas, financieras, académicas, numéricas y académicas en el tiempo, para ofrecer una visión integral del perfil estudiantil. Cada sección proporciona gráficos específicos, como distribuciones de estado civil, cargos, financiamiento, correlaciones numéricas y evolución del rendimiento académico, lo que facilita la identificación de patrones que influyen en la deserción. Esta estructura permite un análisis claro y directo de los factores críticos, apoyando la toma de decisiones informada para mejorar la retención de estudiantes.

3. Responsible: Consideraciones éticas

En Colombia, existen cuatro normativas que discuten sobre los lineamientos a seguir cuando se está realizando un proyecto que utiliza datos. En primer lugar, el artículo 15 de la constitución colombiana dicta que todas las personas tienen derecho a su intimidad personal y familiar y a su buen nombre, y el Estado debe respetarlos y hacerlos respetar [1]. En este proyecto los datos de los estudiantes de la población de maestría han sido debidamente anonimizados con tal fin. No hay información en el *dataset* entregado por la organización en el que se encuentre información personal de los estudiantes como lo son columnas como: el código del estudiante, cédula, nombres y apellidos.

En segundo lugar, la Ley 1266 del 2008 y la Ley 1581 de 2012 regulan el manejo de datos personales en bases de datos, especialmente en los de tipo financiero, crediticio, comercial y de servicios. Estas dos leyes tienen como principal objetivo proteger los datos personales por lo que dictan la responsabilidad que tienen las entidades de implementar una estrategia robusta para la protección y tratamiento de datos, asegurando la privacidad y seguridad de la información. Por último, la circular externa 002 de 2024, si bien no es de obligatorio seguimiento, sí brinda cuatro principios rectores que todo proyecto que utiliza inteligencia artificial debe seguir [2]. Estas son:

A. Idoneidad: El Tratamiento es capaz de alcanzar el objetivo propuesto
B. Necesidad: No exista otra medida más moderada en cuanto al impacto de las operaciones de Tratamiento en la protección de Datos personales e igual de eficaz para conseguir tal objetivo
C. Razonabilidad: El Tratamiento debe estar orientado a cumplir finalidades constitucionales

D. Proporcionalidad en sentido estricto: Las ventajas obtenidas como consecuencia de la restricción del derecho a la protección de datos no deberán ser superadas por las desventajas de afectar el derecho al Habeas Data

Con el fin de guiarse por la circular, el grupo decidió identificar los posibles riesgos del proyecto, priorizarlos e implementar una metodología con el fin de mitigar los posibles riesgos. Los riesgos identificados asimismo como las posibles estrategias de mitigación se encuentran en la figura 1.

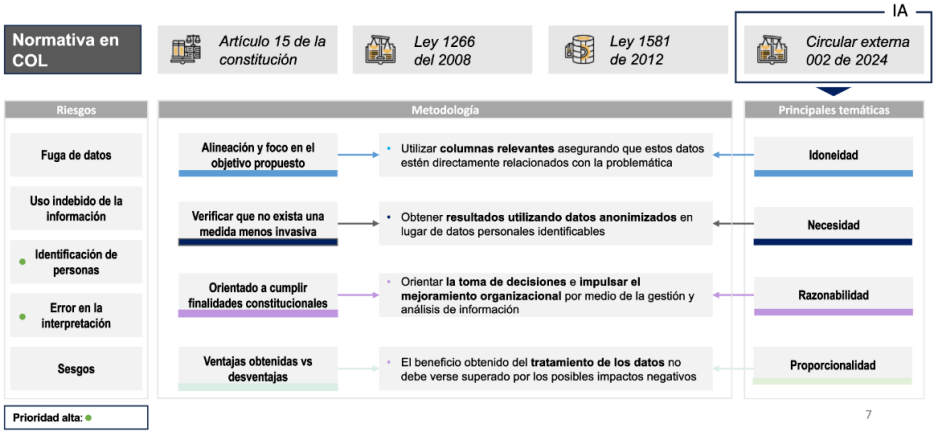


Figura 2. *Metodología de alto nivel para mitigar los riesgos*

Para la idoneidad únicamente se utilizarán las columnas relevantes al caso propuesto dejando por fuera información que no sea necesaria para el objetivo propuesto. En el caso de la necesidad, se plantea verificar que no exista una medida menos invasiva que la propuesta. Para ello, se ha propuesto la opción menos invasiva de anonimizar los datos, lo que garantizará que no sea posible identificar a qué usuario corresponde qué registro. En tercer lugar, para cumplir con la razonabilidad, el proyecto se alinea con la estrategia de la Universidad de los Andes de: “Orientar la toma de decisiones e impulsar el mejoramiento organizacional por medio de la gestión y análisis de información”. Finalmente, para cumplir con proporcionalidad, se identificaron los riesgos más altos los cuales son identificación de las personas y error en la interpretación. Estos se han mitigado a través de a) la anonimización y b) el uso de modelos interpretables como lo es una regresión logística.

4. Enfoque analítico

4.1. Hipótesis

Después de realizar un análisis exhaustivo y un entendimiento profundo del negocio, se han definido las siguientes hipótesis para guiar el desarrollo del modelo predictivo:

- Los estudiantes con promedio académico cercano a 3.25 tienen una mayor probabilidad de desertar de la universidad.
- Los estudiantes que reciben becas y tienen un bajo nivel de aporte económico muestran una mayor probabilidad de desertar

4.2. Preguntas de Negocio

Para validar estas hipótesis y abordar de manera efectiva la problemática de la deserción, se han formulado las siguientes preguntas de negocio:

- ¿Qué factores estructurales (por ejemplo, promedio global por facultad o niveles de beca) son los mayores contribuyentes al riesgo de deserción?
- ¿Cómo impactan las dificultades económicas en la tasa de deserción y qué estrategias se pueden implementar para mejorar la asignación de ayudas financieras?
- ¿Qué estudiantes requieren intervención inmediata basándose en indicadores académicos?

4.3. Técnicas Estadísticas

Para validar las hipótesis y responder a las preguntas planteadas en el contexto de la ideación del modelo predictivo de deserción estudiantil, se propone un enfoque analítico centrado en la identificación y priorización de factores críticos que influyen en la deserción. Estas técnicas permitirán construir una base sólida para el modelo predictivo, asegurando que el análisis inicial se traduzca en estrategias de intervención efectivas::

- **ANOVA (Análisis de Varianza):** Se utilizará para identificar diferencias significativas entre grupos clave de estudiantes, como niveles de ayuda financiera o rangos de promedio académico. Esto permitirá validar si factores como el acceso a becas o el rendimiento académico tienen un impacto medible en la deserción y, por ende, merecen un peso relevante en el modelo.
- **Análisis de Correlación:** Este análisis servirá para identificar y priorizar las variables más relacionadas con la deserción. Por ejemplo, se explorarán asociaciones entre el promedio académico, el porcentaje de créditos aprobados, la edad, y otros indicadores clave. El objetivo es guiar el desarrollo del modelo predictivo hacia factores accionables.
- **PCA (Análisis de Componentes Principales):** Reducirá la dimensionalidad del conjunto de datos, eliminando ruido y destacando las variables con mayor impacto en la deserción. Esto asegurará un modelo más eficiente y comprensible, alineado con la visión de crear intervenciones dirigidas y efectivas.

4.4. Visualización de Datos

El análisis estadístico será complementado con visualizaciones diseñadas para facilitar la interpretación y comunicación de los hallazgos. Estas visualizaciones estarán alineadas con los objetivos del producto:

- **Diagramas de Caja (Boxplots):** Permitirá analizar la distribución de promedios académicos y otras métricas numéricas críticas en relación con la deserción, ayudando a identificar estudiantes en los extremos (altos riesgos).
- **Mapa de Calor (Heatmap):** Este será clave para mostrar gráficamente las correlaciones entre las variables sociodemográficas, académicas y financieras, facilitando la identificación visual de los factores más influyentes en la deserción.
- **Gráficos de Barras Apilados:** Mostrarán la distribución de deserción en diferentes segmentos, como programas académicos, niveles de beca o rangos de edad, ayudando a entender dónde se concentran los riesgos más altos.

4.5. Machine Learning

4.5.1. Planteamiento de Modelos: Enfocado directamente en la problemática de la deserción estudiantil, se plantea la implementación de algoritmos de clasificación robustos, seleccionados por su capacidad de soportar diferentes tipos de datos y escenarios:

- **Random Forest**: Se utilizará como el algoritmo principal debido a su capacidad de manejar datos complejos y proporcionar interpretaciones claras sobre la importancia de las variables, lo que permitirá al equipo universitario priorizar intervenciones específicas.
- **Regresión Logística:** Complementará el análisis, ofreciendo una solución más interpretativa y útil para comunicar los hallazgos a los responsables de políticas educativas.
- **Support Vector Machine (SVM):** Será empleado en escenarios donde se necesite máxima precisión para separar los estudiantes en riesgo de deserción, en caso de que los datos muestren patrones no lineales.

4.5.2. Alineación de Modelo con Usuarios Potenciales: El despliegue del modelo estará diseñado para generar salidas específicas y accionables para cada grupo de usuarios clave (profesores, comités y otros entes administrativos), asegurando que los resultados del análisis se traduzcan directamente en intervenciones efectivas.

5. Recolección de datos

Las bases de datos proporcionadas por la organización contienen información relevante sobre los estudiantes matriculados, su rendimiento académico por periodo, las formas de financiamiento utilizadas, los estudiantes graduados por periodo, y detalles de la oferta académica, incluyendo cursos y franjas horarias, así como la información sociodemográfica de los estudiantes.

Para esta primera etapa, se decidió utilizar solo las bases de datos que incluyen: estudiantes matriculados, estudiantes graduados, desempeño académico por periodo, formas de financiamiento y datos sociodemográficos. Solo se considerarán a los estudiantes que empezaron a estudiar su maestría a partir del 2015 y únicamente el primer programa académico de cada estudiante para simplificar el análisis.

5.1. Estudiantes matriculados

Esta fuente nos permite identificar los estudiantes matriculados en cada periodo y su programa académico. Las variables seleccionadas incluyen el periodo, el código anonimizado del estudiante, la edad, el estado civil y detalles sobre el programa académico.

Nombre de variable	Tipo de dato	Descripción
PERIODO	str	Semestre en curso
CODIGO_ESTUDIANTE_ANON	str	Código de estudiante único anonimizado
EDAD	int	Edad del estudiante
TIPO_ESTADO_CIVIL	str	Estado civil del estudiante
FACULTAD_PROGRAMA_1	str	Facultad del programa del estudiante
DEPARTAMENTO_PROGRAMA_1	str	Departamento del programa del estudiante
PROGRAMA_1	str	Programa académico del estudiante

5.2. Desempeño académico

Esta fuente contiene el rendimiento académico general de los estudiantes por cada periodo. Se calculó una variable derivada, el "porcentaje de créditos aprobados", para mitigar la inflación de promedios generada durante la pandémica, cuando se permitió retirar materias hasta el final del semestre. Las variables consideradas incluyen el periodo, el código del estudiante anonimizado, los créditos intentados y aprobados por semestre, el promedio semestral, los créditos acumulados intentados y aprobados y el promedio global.

Nombre de variable	Tipo de dato	Descripción
PERIODO	str	Semestre en curso
CODIGO_ESTUDIANTE_ANON	str	Código de estudiante único anonimizado
CREDITO_SEM_INTENTADO_INSTITUCIONAL	float	Créditos intentados en ese semestre
CREDITO_SEM_APROBADO_INSTITUCIONAL	float	Créditos aprobados en ese semestre
PROMEDIO_SEM_GLOBAL	float	Promedio del semestre
CREDITO_INTENTADO_GLOBAL	float	Créditos intentados en la carrera hasta ese momento de tiempo (acumulado).
CREDITO_APROBADO_GLOBAL	float	Créditos aprobados en la carrera hasta ese momento de tiempo (acumulado).
PROMEDIO_GLOBAL	float	Promedio global del estudiante.

$$\%\ creditos\ aprobados = \frac{CREDITO\_APROBADO\_GLOBAL}{CREDITO\_INTENTADO\_GLOBAL}$$

5.3. Forma de financiación

Esta base de datos informa cómo cada estudiante financia su matrícula por semestre. Se seleccionó solo el tipo de financiamiento que más contribuyó al pago total de la matrícula, priorizando simplicidad en el análisis. Las variables incluyen el periodo, el código de estudiante anonimizado y los tipos y clasificaciones de financiamiento.

Nombre de la variable	Tipo de dato	Descripción
PERIODO	str	Semestre en curso
CODIGO_ESTUDIANTE_ANON	str	Código del estudiante único anonimizado
CLASIFICACION_BECA	str	Clasificación general del tipo de financiación del estudiante
TIPO_BECA	str	Clasificación más específica del tipo de financiación.
PORCENTAJE_APORTE_ITEM	float	Porcentaje del semestre que financió con ese TIPO_BECA

5.4. Información sociodemográfica

Esta base de datos proporciona una visión general de la situación personal del estudiante más allá del ámbito académico. Las columnas seleccionadas incluyen sexo biológico, el estrato, tipo de vivienda, pertenencia a minoría étnica, afiliación al Sisbén y detalles laborales como sector y cargo.

Nombre de la variable	Tipo de dato	Descripción
CODIGO_ESTUDIANTE_ANON	str	Código del estudiante único anonimizado
DESCRIPCION_SEXO	str	Sexo biológico del estudiante
ESTRATO	str	Estrato del estudiante

TIPO_VIVIENDA	str	Tipo de vivienda en la que vive.
PERTENECE_MINORIA_ETNICA	str	Establece si el estudiante pertenece o no a una minoría étnica.
SISBEN	str	Pertenece o no al Sisbén
PUN_SISBEN	str	Puntaje en el Sisbén
NOMBRE_SECTOR_EMPRESA_ESTUDIANTE	str	Sector de la empresa en la que trabaja el estudiante (PUBLICO, PRIVADO, MIXTO)
CARGO_ESTUDIANTE	str	Cargo que desempeña el estudiante en su trabajo
FECHA_GRADO_PS	datetime	Fecha de grado del pregrado
PORCENTAJE_TIEMPO_PARA_ESTUDIO	float	Porcentaje de tiempo que el estudiante puede dedicarle al estudio a la semana
CIUDAD_RESIDENCIA_UG	str	Ciudad de residencia del estudiante
DEPARTAMENTO_RESIDENCIA_UG	str	Departamento de residencia del estudiante
PAIS_RESIDENCIA_UG	str	País de residencia del estudiante

5.5. Estudiantes graduados

Esta fuente proporciona información sobre los estudiantes que se han graduado en cada periodo, el programa académico y la facultad de la cual se han graduado. Aunque estos datos no se incluirán directamente en el modelo, se usarán para distinguir si un estudiante dejó de matricularse porque se graduó, evitando así interpretarlo erróneamente como deserción.

Nombre de la variable	**Tipo de dato**	**Descripción**
PERIODO	str	Semestre en curso
CODIGO_ESTUDIANTE_ANON	str	Código del estudiante único anonimizado
PROGRAMA	str	Programa académico del cual se graduó el estudiante
FACULTAD	str	Facultad de la que se graduó el estudiante
DEPARTAMENTO	str	Departamento del que se graduó el estudiante

5. Entendimiento de los datos

6.1. Análisis exploratorio

El análisis exploratorio comenzó con 15,533 estudiantes, consolidando una base de datos unificada de 60,038 registros, distribuidos en 32 columnas (9 numéricas y 23 categóricas).

6.1.1. Análisis Univariado

Variables sociodemográficas

El 59.22% de los estudiantes son del sexo masculino y el 40.74% son del sexo femenino, mientras que el 0.03% no informó. El 50% de los estudiantes entran a la maestría teniendo entre 29-38 años. La mayoría trabajan en el sector privado (64.91%) y ocupa cargos de ingeniero (18.27%). Además, el 80.70% son solteros y el 99.59% reside en Colombia (83.62% en Bogotá). Si desea ver el detalle de estas estadísticas ir al notebook eda.ipynb entre las celdas 31 y 51.

Variables académicas

La mayoría de los estudiantes hacen una maestría en la Facultad de Ingeniería (33.55%) y Administración (24.55%). Si desea ver los porcentajes y cantidades del resto de facultades diríjase al notebook eda.ipynb a la celda 29 y 30).

6.1.2. Análisis Bivariado

Variables sociodemográficas

Se puede observar que los desertores se concentran entre las edades 31 y 40, mientras que los no desertores entre los 30 y 39. Con respecto al sexo y el sector no se ve una tendencia. Por otro lado, se puede ver que las personas con estado civil SEPARADO y VIUDO son los que más desertan con un porcentaje de 12.7% y 20% respectivamente; sin embargo, son muy pocas personas por lo que no se podría declarar una tendencia. (Para ver más información ir al .pbix maestria_caracterizacion página “edad”, “sexo y sector” y “estado civil”).

Variables financieras y académicas

La mayoría de los estudiantes financian sus estudios mediante recursos propios o préstamos (OTRAS FORMAS DE PAGO). Por otra parte, en general, el desempeño académico por parte de los estudiantes a lo largo de los periodos es muy bueno. En promedio, los estudiantes de maestría tienen un porcentaje de créditos aprobados de aproximadamente el 97.5%, aunque hay casos atípicos que requieren atención debido a su posible riesgo de deserción. Para ver más el detalle diríjase al notebook eda.ipynb a las celdas 52-64.

Por otro lado, se puede observar que el 70.26% de los estudiantes que desertan financian su semestre con OTRAS FORMAS DE PAGO, mientras que el 61.76% de los estudiantes que no desertan financian con OTRAS FORMAS DE PAGO. Es decir, que los que desertan financian con OTRAS FORMAS DE PAGO casi 10 puntos porcentuales más que los que no. Adicionalmente, se observa que las facultades de las cuales más se deserta son CIDER y MEDICINA con un porcentaje de 18.61% y 18.24% respectivamente. Para ver más al detalle estas tendencias ir al .pbix llamado maestria_caracterizacion en la página “financiación” y “facultad”.

6.1.3. Análisis Multivariado

Variables académicas en el tiempo

En la figura, se puede observar que el valor del promedio global es más estable que el promedio semestral, debido a que el primero acumula el desempeño a lo largo de la carrera. Por otra parte, el comportamiento anormal del periodo 202010 se puede atribuir a la llegada de la pandemia, ya que ese semestre las calificaciones no se dieron en formato numérico, sino Aprobado/Reprobado, por lo que muchos estudiantes no tuvieron promedio ese semestre.

Adicionalmente, como se comentó en el análisis bivariado, se puede ver que el desempeño académico de un estudiante de maestría promedio es bueno. La mayoría de los estudiantes de maestría tienen un promedio por encima de 4.0.

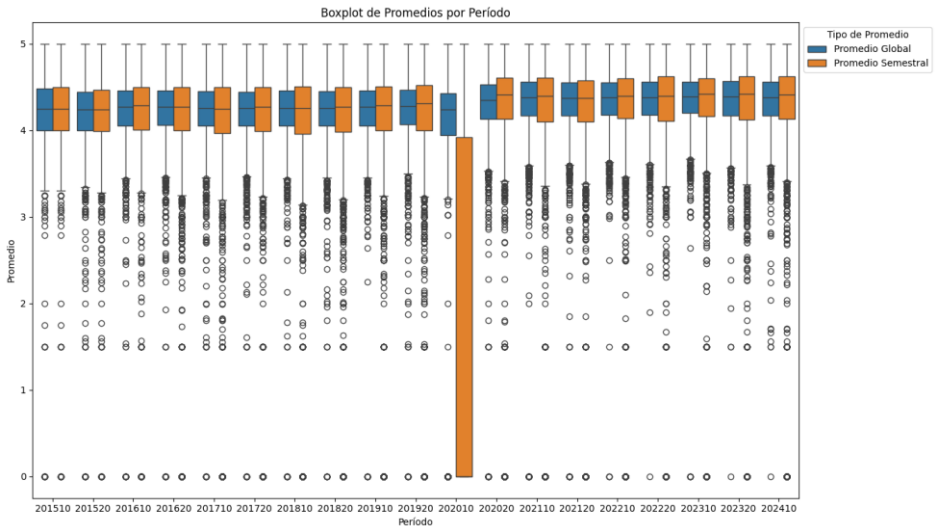


Figura 3. Diagrama de cajas y bigotes que muestra la distribución del promedio global y semestral de los estudiantes de maestría por periodo.

Por otro lado, se observa que tanto el promedio global como el semestral de los estudiantes graduados (NO DESERTORES) es más elevado que el de los desertores; no obstante, los promedios de los desertores son mayores de 3.7 en promedio, lo cual no es malo. Un comportamiento similar sucede con el porcentaje de créditos aprobados, con la diferencia de que los porcentajes de créditos aprobados de los desertores son mayores a 80% en promedio. (Para más detalles diríjase al .pbix llamado maestria_caracterizacion página PGA, Promedio Semestral y % créditos aprobados).

Variables numéricas

Existe una correlación entre el porcentaje de créditos aprobados global y el semestral. También se encontró una leve correlación entre el promedio global y el semestral (0.36) y entre el promedio global y el porcentaje de créditos aprobados (0.35). Esto sugiere la conveniencia de descartar algunas variables para evitar redundancia. Es importante señalar que en este análisis no se incluyeron las variables sobre el número de créditos aprobados e intentados, ya que estas fueron representadas por la medida derivada del porcentaje de créditos aprobados.

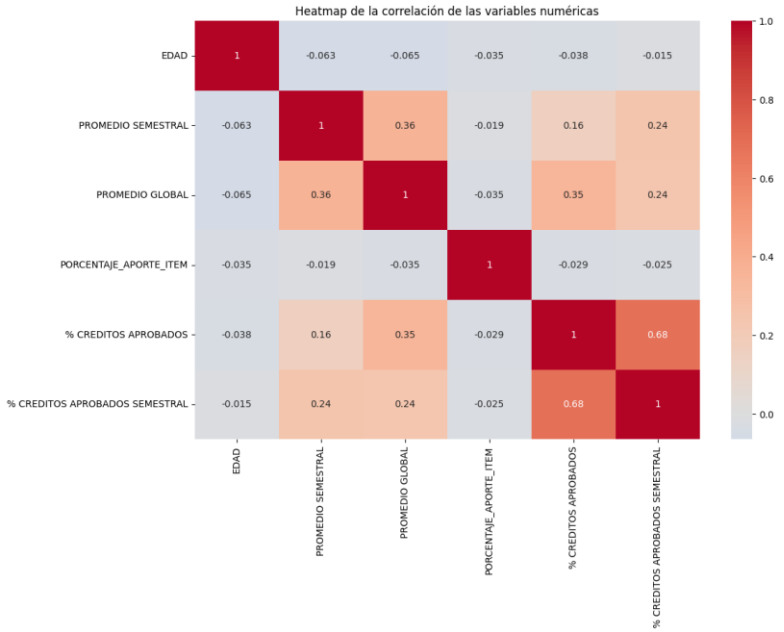


Figura 4. Mapa de calor de la correlación de las variables numéricas.

6.2. Calidad de los datos

Para evaluar la calidad de los datos, se verificaron tres dimensiones principales: completitud, consistencia y validez.

6.2.1. Completitud

Con respecto a la completitud, se revisó el porcentaje de valores nulos por parte de cada variable. Se encontró que, la mayoría de las columnas presentan un bajo porcentaje de valores nulos, aunque algunas columnas (estrato, tipo de vivienda, Sisbén) se eliminarán debido a su alto nivel de datos faltantes.

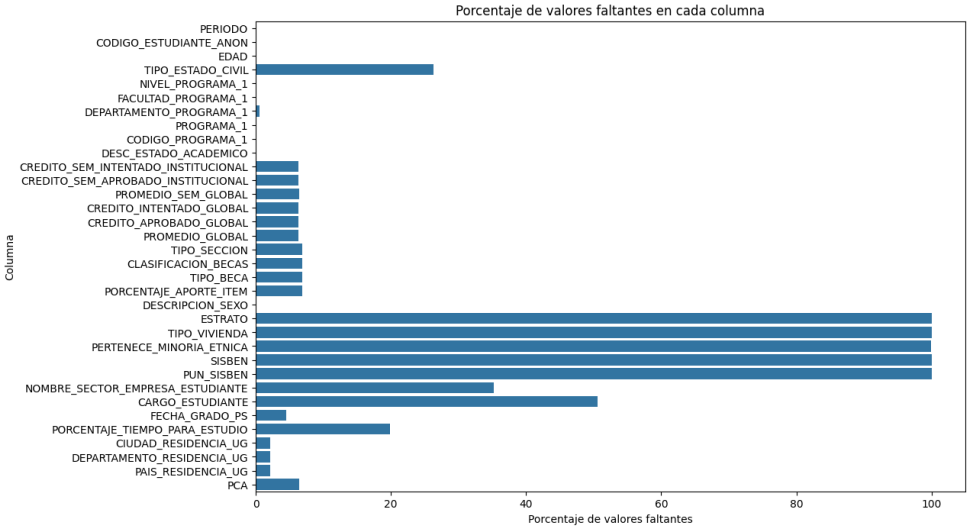


Figura 10. Gráfica de barras que muestra el porcentaje de valores faltantes o nulos por variable.

6.2.2. Consistencia

Para evaluar la consistencia de los datos, se verificó que los valores de cada variable estuvieran expresados en una misma unidad y que cada columna presentara el tipo de dato correspondiente. Se encontró que el 3.13% de las columnas (1 de 32) no cumple con el tipo de dato adecuado debido a que los valores no están en una unidad consistente; esto ocurre en la columna **PORCENTAJE_TIEMPO_PARA_ESTUDIO**.

Además, se identificaron inconsistencias en un registro dentro de los 68,038: en un caso, los créditos aprobados globalmente superan los créditos intentados globalmente (**CREDITO_APROBADO_GLOBAL >**

CREDITO_INTENTADO_GLOBAL), y los créditos aprobados en el semestre exceden los créditos intentados en el semestre (**CREDITO_SEM_APROBADO_INSTITUCIONAL > CREDITO_SEM_INTENTADO_INSTITUCIONAL**). Este comportamiento es incoherente, ya que no es posible aprobar más créditos de los que se están cursando.

6.2.3. Validez

Para evaluar la validez de los datos, se verificó que los valores de las variables numéricas estuvieran dentro del rango esperado y que los valores de las variables categóricas fueran coherentes. En cuanto a las variables numéricas, se encontró que el 11.11% (1 de 9) presenta valores fuera del rango esperado; esto ocurrió en la columna **EDAD**, donde aparecieron edades de 124 en 3 de los 60,038 registros. Por otro lado, el 8.69% de las variables categóricas (2 de 23) contienen valores no válidos o inconsistentes, específicamente en las columnas **PORCENTAJE_TIEMPO_PARA_ESTUDIO** y **TIPO_BECA**.

6. Preparación de los datos

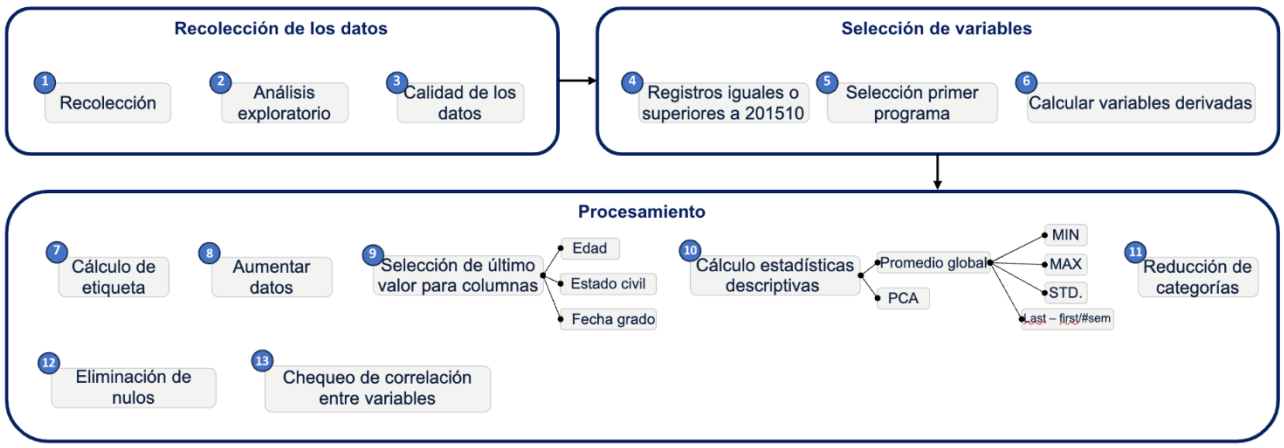


Figura 11. Diagrama de bloques del proceso de preparación de datos

En la figura 11 se observa el proceso seguido para la preparación de los datos. Primero, se realizó la recolección de los datos, detallada en la sección de este documento titulada *"Recolección de datos"*.

Tras la recolección, se llevó a cabo un análisis exploratorio de los datos y una evaluación de su calidad, seguida de una selección inicial de variables o *feature selection*. En esta etapa, se incluyeron únicamente los registros de estudiantes que comenzaron su maestría a partir de 2015, ya que la información financiera solo está disponible desde ese año. Además, se consideró únicamente el primer programa de cada estudiante para simplificar el modelo. También se calcularon variables derivadas, como:

- **PCA** (porcentaje de créditos aprobados, explicado anteriormente).
- **PORCENTAJE_BECA_INTERNA**, que indica el porcentaje de financiación semestral proporcionado por una beca interna (otorgada por la universidad).
- **Cumulative_Missed_Semesters**, que representa el número total de semestres en los que un estudiante no se matriculó.
- **Times_returned**, que refleja las veces que un estudiante dejó de matricularse y luego regresó.

Seguido a esto, se descartaron algunas variables. Las variables eliminadas y sus respectivas razones se detallan en la siguiente tabla:

Variable(s)	Razón
ESTRATO, TIPO_VIVIENDA, PERTENECE_MINORIA_ETNICA, SISBEN, PIN_SISBEN, CARGO_ESTUDIANTE, NOMBRE_SECTOR_EMPRESA_ESTUDIANTE	Alto porcentaje de valores nulos.
TIPO_SECCION, TIPO_BECA	Se decidió usar la categoría más general CLASIFICACION_BECAS para reducir complejidad del modelo. Y estas variables están relacionadas entre sí, por lo que solo vale la pena usar una de las tres.
CREDITO_INTENTADO_GLOBAL, CREDITO_APROBADO_GLOBAL	Al calcular el porcentaje de créditos aprobados (PCA), estas variables se volvieron redundantes.

PROMEDIO_SEM_GLOBAL, CREDITO_SEM_INTENTADO_INSTITUCIONAL, CREDITO_SEM_APROBADO_INSTITUCIONAL	Se decidió usar información global de los semestres para reflejar mejor la trayectoria académica y evitar multicolinealidad.
DESC_ESTADO_ACADEMICO	Falta de definiciones claras para todas las categorías.
CODIGO_PROGRAMA_1, PROGRAMA_1, DEPARTAMENTO_PROGRAMA_1	Se decidió usar la categoría más general FACULTAD_PROGRAMA_1 para reducir la complejidad del modelo. Y estas variables están relacionadas entre sí, por lo que solo vale la pena usar una de las cuatro.
NIVEL_PROGRAMA_1	Es una constante: todos los estudiantes son de maestría (**MA**).

Esto dejó como resultado la siguiente lista de variables seleccionadas: PERIODO, CODIGO_ESTUDIANTE_ANON, EDAD, TIPO_ESTADO_CIVIL, FACULTADO_PROGRAMA_1, PROMEDIO_GLOBAL, CLASIFICACION_BECAS, PORCENTAJE_BECA_INTERNA, FECHA_GRADO_PS, DEPARTAMENTO_RESIDENCIA_UG, PORCENATAJE_APORTE_ITEM, PCA, Cumulative, Missed_Semesters y Times_returned.

Luego, se procedió al procesamiento de los datos. Primero, se definió la etiqueta o variable objetivo (*target*). Para ello, se consideró como desertor a cualquier estudiante de maestría que haya tenido cuatro o más semestres consecutivos sin matricularse y que no haya vuelto a hacerlo hasta la fecha actual. Con base en esta definición:

- Los estudiantes etiquetados como **NO DESERTOR** son aquellos que ya se graduaron de su maestría.
- Los **DESERTORES** cumplen con la definición anterior.
- Los estudiantes **ACTIVOS** son aquellos que aún están matriculados.

Para entrenar y evaluar el modelo, solo se usaron los datos de estudiantes graduados y desertores.

A continuación, se realizó un aumento de datos. La razón detrás de esta decisión es que los datos originales reflejan la trayectoria completa del estudiante. Sin embargo, para que el modelo sea capaz de predecir antes de que ocurra la deserción o graduación, se duplicaron los datos de estudiantes graduados y desertores, pero utilizando únicamente la primera mitad de su carrera. Por ejemplo, si un estudiante se matriculó durante cuatro semestres, solo se duplicaron los primeros dos. Este enfoque busca simular la historia de un estudiante activo. Los datos duplicados se registraron como estudiantes distintos, modificando el código de estudiante para diferenciarlos.

Finalmente, se adecuaron los datos al formato necesario para ingresarlos al modelo, de manera que cada registro representara un único estudiante. Para capturar toda la información relevante en un solo registro, se realizaron los siguientes pasos:

- Para las variables **TIPO_ESTADO_CIVIL, FECHA_GRADO_PS y DEPARTAMENTO_RESIDENCIA_UG**, se tomó la información original del formulario de admisión.
- Para las variables **EDAD, FACULTAD_PROGRAMA_1, PROMEDIO_GLOBAL, PCA, Cumulative_Missed_Semesters y Times_returned**, se seleccionó el valor correspondiente al último semestre matriculado.
- Para **CLASIFICACION_BECAS**, se calculó el tipo de financiación que más aportó durante la maestría, basado en **PORCENTAJE_APORTE_ITEM**, y esta última variable fue descartada posteriormente.
- Se calcularon estadísticas descriptivas para **PROMEDIO_GLOBAL** y **PCA**, como mínimo, máximo, desviación estándar de los cambios semestre a semestre y tendencia ((último - primero) / número de semestres matriculados).
- Para **PORCENTAJE_BECA_INTERNA**, se sumaron los porcentajes a lo largo de los semestres y se dividieron entre el total de semestres, generando una nueva columna llamada **PORCENTAJE_BECA_INTERNA_CARRERA**.

Además, se realizaron las siguientes transformaciones adicionales:

- De **FECHA_GRADO_PS**, se extrajo únicamente el año de graduación.
- Las categorías de **DEPARTAMENTO_RESIDENCIA_UG** se simplificaron en **BOGOTÁ D.C.** y **FUERA DE BOGOTÁ D.C.**, para reducir la complejidad del modelo.

Se eliminaron los registros donde **PROMEDIO_GLOBAL** y/o **PCA** estaban vacíos, ya que carecían de información relevante. Para los casos donde la desviación estándar de los cambios en estas variables era nula, se imputó un valor de cero, asumiendo que no había variaciones registradas. Los nulos restantes fueron eliminados.

Por último, se revisó la correlación entre variables numéricas para evitar multicolinealidad o redundancia al entrenar el modelo. Se descartaron las siguientes variables: **PROMEDIO_GLOBAL_min, PROMEDIO_GLOBAL_max, PCA_min, PCA_max, YEAR_GRADUATION** y **PCA**. El código del estudiante también fue eliminado al ser un identificador único sin relevancia predictiva.

7. Estrategia de validación y selección de modelo

7.1. Estrategia de Experimentación

7.1.1 Construcción de Modelos

El desarrollo del modelo predictivo se basa en datos históricos de estudiantes, tanto graduados como desertores, con el objetivo de identificar a los estudiantes activos con mayor riesgo de deserción. Este enfoque busca aprovechar diferentes técnicas de Machine Learning para construir modelos robustos y confiables que permitan predecir el riesgo de manera precisa.

Los modelos tenidos en cuenta fueron

- **Random Forest:** Este modelo es conocido por su capacidad de manejar grandes volúmenes de datos y una posible resistencia al overfitting, se explorarán y se optimizarán los siguientes hiperparámetros clave
    - N_estimators: número de árboles en el bosque
    - Criterion: Métrica para medir la calidad de las divisiones de los árboles
    - Max_depth: Profundidad máxima permitida por cada ñarbol
- **Support Vector Machine:** Se utilizó al ser un modelo particularmente efectivo en problemas de clasificación binaria en datos complejos, durante la experimentación, se ajustarán los siguientes parámetros
    - C: Controla el equilibrio entre margen amplio y error de clasificación
    - Degree: Grado del polinomio que utiliza un kernel polinómico
    - Kernel: Tipo de función kernel a utilizar
- **Adaboost:** Este método se utiilizó ya que es útil para mejorar el desempeño de clasificadores débiles, para este caso se configurarán
    - N_estimators: Número de clasificadores base que se combinarán
    - Learning_rate: Tasa de aprendizaje para ajustar el peso de cada clasificador en el ensamble
- **Regresión Logística:** Se utilizó ya que es altamente interpretable y efectivo para problemas de clasificación binaria, los hiperparámetros a ajustar incluyen:
    - C: Parámetro de regularización que controla el ajuste del modelo
    - Penalty: Tipo de penalización (L2 o ninguna)
    - Degree; Posible inclusión de términos no polinómicos para capturar relaciones no lineales.

7.1.2. Optimización de Rendimiento

Para maximizar la eficacia de cada modelo, se implementará un proceso de búsqueda de hiperparámetros, en concreto la técnica de Grid Search será utilizada para identificar las combinaciones óptimas de hiperparámetros. Este proceso se realizará sobre el conjunto de entrenamiento, asegurando tener el mejor rendimiento posible en esta separación de datos

7.1.3 Selección de Modelo

La selección de modelo será un proceso riguroso, basado en el desempeño de cada técnica de clasificación en el contexto del problema. Para ello se utilizará el F1-Score como la métrica principal de evaluación. Esto permitirá equilibrar precisión y sensibilidad, asegurando que el modelo pueda identificar estudiantes en riesgo de deserción con un mínimo de falsos positivos y negativos.

Una vez identificados los mejores modelos, se llevará a cabo una comparación mediante pruebas conocidas como A/B Testing. Esto consistirá en aplicar los modelos seleccionados a datos independientes para evaluar su desempeño en condiciones reales y determinar cuál es más adecuado para cumplir con los objetivos del proyecto.

7.2. Separación y Creación de Pipeline para el Modelo

7.2.1 Separación de Datos

División en Entrenamiento y Prueba

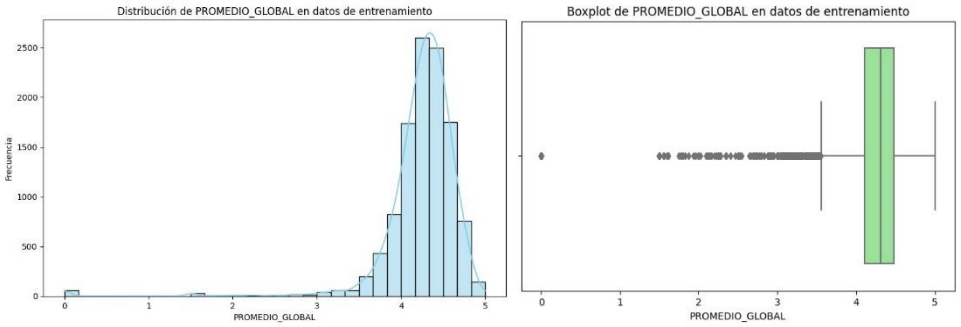
En esta etapa, se realizó la división de datos y se definieron los pasos necesarios para preparar los datos en el pipeline y así garantizar la calidad del entrenamiento y evaluar los modelos predictivos de manera adecuada.

En primer lugar, se realizó la división de datos, los datos se separaron en dos conjuntos. 80% para entrenamiento y 20% para prueba. Este enfoque asegura que el modelo pueda aprender de una parte de los datos mientras se evallúa de manera objetiva en datos no vistos.

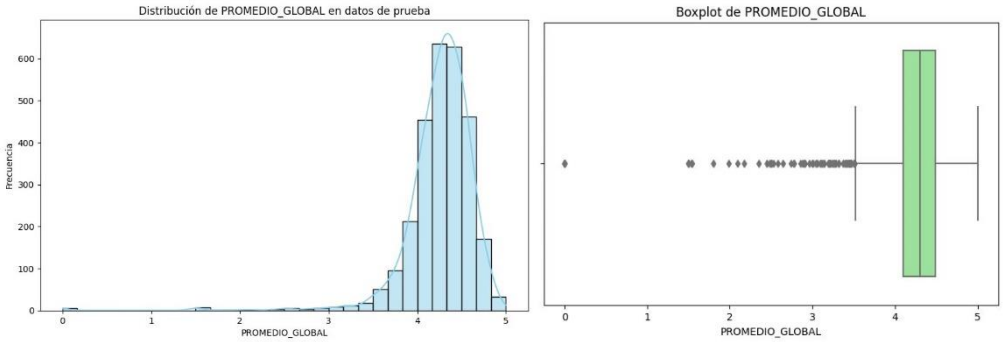
Conservación de Distribución de los Datos

Se verificó que la separación de los datos en el conjunto de entrenamiento y prueba respetara la distribución de las variables a analizar, para ello, se creó un reporte de perfilamiento tanto para datos de entrenamiento como datos de prueba (que se puede encontrar en el repositorio de GitHub como train_report.html y test_report.html), como ejemplo para demostrar que si se conserva la distribución, se mostrará la distribución tanto en histograma como en diagrama de cajas de la variable PROMEDIO_GLOBAL

Distribución en Datos de Entrenamiento



Distribución en Datos de Prueba



En estas dos imágenes, se comprueba que la distribución de esta variable tanto en entrenamiento como en prueba son muy similares.

Haciendo esta comparación en otras variables como el número de semestres (num_semesters) o la edad (EDAD), presentan resultados similares, por lo que se puede asegurar de forma general que en efecto se hace la conservación de distribuciones al realizar la separación de datos.

7.2.2 Creación del Pipeline

Por otro lado, ya dentro de la creación del pipeline, se realizó la codificación de varialbes, para la variable objetivo se utilizó label encoding, y para las demás variables categóricas, se utilizó One-Hot Encoding, asegurando que el modelo pueda interpretar dichas variables correctamente.

Para la normalización, se utilizó la técnica de StandardScaler para normalizar las variables numéricas, lo que garantiza que las características tengan una escala uniforme y evita que las variables con los valores más grandes dominen el modelo.

Por último, con respecto a otras técnicas avanzadas de entrenamiento, se utilizaron técnicas para prevenir el desbalanceo entre clases, para ello se implementó en primer lugar el StratifiedKFold, el cual es usado para dividir los datos de forma

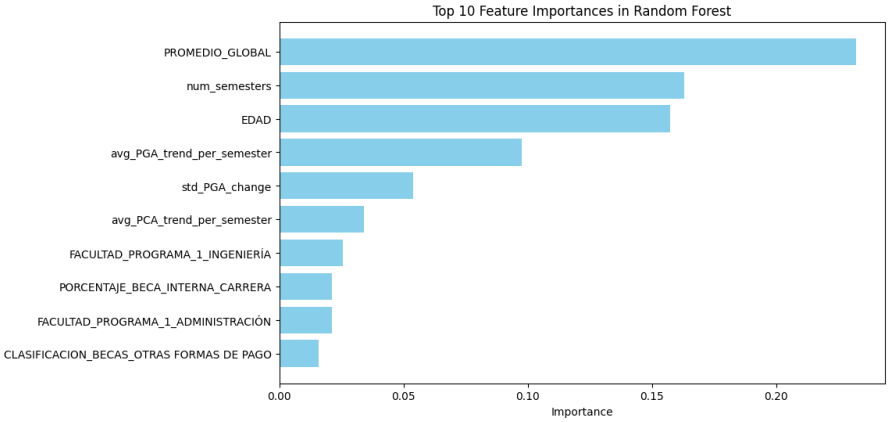
estratificada, asegurando que cada partición conserve la proporción original de las clases en el conjunto de datos, la segunda técnica implementada fue SMOTE (Synthetic Minority Oversampling Technique), el cual también se aplica para equilibrar las clases en el conjunto de datos, creando ejemplos sintéticos y mejorando el desempelño de modelo al predecir ambas clases de manera justa

8. Construcción y evaluación del modelo

De acuerdo con lo mencionado en el anterior literal, se evaluaron los resultados para los cuatro modelos variando los hiperparámetros seleccionados. En el caso de Random Forest el mejor modelo se obtuvo con el criterio *entropy* y con un *max_depth* de 40 y 500 estimadores. En el caso de SVM el mejor modelo se obtuvo con un C de 0.1, de grado 3 y un kernel poly. En el caso de Adaboosting se obtuvo un classifier_estimator de DecisionTreeClassifier con *max_depth* de 20 y *min_samples_split* de 10. Además, en el adaboost el mejor learning rate fue de 0.1 y el número de estimadores de 500. Finalmente, para la regresión logística los mejores hiperparámetros se obtuvieron un C de 0.1, sin penalización y con un grado polinomial de 1. A continuación, en la siguiente tabla se observan los resultados obtenidos por los modelos tanto para train como para test. Los resultados mostrados son aquellos obtenidos en weighted average. Para ver los resultados macro dirigirse a notebook *eda_training*.

	Train				**Test**			
	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
Random forest	1.00	1.00	1.00	1.00	0.89	0.89	**0.89**	0.89
SVM	0.89	0.87	0.88	0.87	0.88	0.86	0.87	0.86
Adaboost	1.00	1.00	1.00	1.00	0.89	0.90	**0.89**	0.90
Logistic Regression	0.89	0.77	0.81	0.77	0.89	0.77	0.81	0.77

Se seleccionó como métrica el f1-score dado que es la combinación armónica entre la precisión como el recall. Esta medida le da importancia a ambos tipos de errores tanto a los falsos positivos (Identificar incorrectamente a un estudiante como en riesgo de deserción) como a los falsos negativos (No identificar a un estudiante que realmente está en riesgo de deserción). Entre estos se puede observar que los dos modelos con mejores resultados en esta métrica fueron random forest y adaboost. De estos dos, se decidió realizar una prueba A/B para ver si hay diferencia significativa en el f-1 score obtenido entre RF y Adaboosting. Para ello se aplicó una prueba tipo t-student para comparar los F1-scores de dos modelos (Random Forest y Adaboost). Se obtuvo un p-value de 0.10 el cuál es superior de 0.05 por lo que no hay una diferencia estadísticamente significativa en el f1 entre los modelos. Por tanto, si bien no hay una diferencia significativa con el otro modelo, RF le puede dar interpretabilidad al negocio sobre cuáles son los features que determinan a los estudiantes que desiertan de los que no. Además, adaboosting obtuvo un f-1 score de 0.45 sobre los desertores y RF obtuvo un f1 score de 0.47 por lo que al final se seleccionó RF.

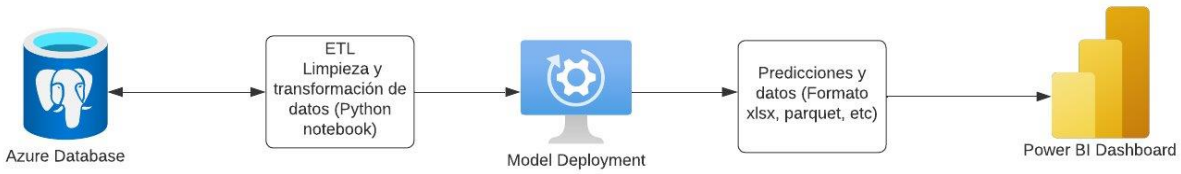


A continuación, en la figura superior, se observan el top 10 features más importantes obtenidos por random forest. El PROMEDIO_GLOBAL emerge como el predictor más potente, indicando que el rendimiento académico general es crucial para identificar estudiantes en riesgo, lo que sugiere la necesidad de un sistema de seguimiento temprano de calificaciones.

La cantidad de semestres cursados aparece como el segundo factor más influyente, señalando que existen períodos críticos durante la carrera donde el riesgo de deserción podría incrementarse. La edad se posiciona como el tercer factor más relevante, lo que podría indicar patrones de deserción diferentes según los grupos etarios. Asimismo, hay otros factores como los indicadores de tendencia PGA, incluyendo el promedio por semestre y su variabilidad, que también muestran una importancia significativa. Esto sugiere que no solo el rendimiento actual es importante, sino también cómo este evoluciona a lo largo del tiempo.

Por último, con respecto a mejoras de los modelos, en particular, los modelos que se podrían mejorar son Random Forest y Adaboost, ya que muestran signos de overfitting (rendimiento perfecto en training) y más bajo en test. Estos podrían mejorarse enfocándose principalmente en técnicas de regularización y variar un poco más los hiperparámetros. Para Random Forest, se podría experimentar con valores más restrictivos de max_depth (actualmente en 40), aumentar el min_samples_leaf para asegurar que cada hoja tenga más observaciones, o reducir el número de estimadores (actualmente en 500) si esto no impacta significativamente el rendimiento. En el caso de Adaboost, se podría probar con un learning rate aún más bajo que 0.1 para que el aprendizaje sea más gradual, reducir la profundidad máxima del DecisionTreeClassifier base (actualmente en 20), o implementar early stopping para prevenir el sobreajuste.

9. Construcción del producto de datos



Azure Database:

Es el repositorio centralizado que almacena los datos académicos y sociodemográficos de los estudiantes. Estos datos se consolidan desde diferentes fuentes internas de la universidad, incluyendo historiales académicos, detalles de becas y formas de financiación, y características sociodemográficas como edad, estrato y estado civil. La base de datos está diseñada para facilitar consultas eficientes y la integración con el sistema ETL.

ETL (Python Notebook):

Este componente realiza la extracción, transformación y carga de datos (ETL). Utilizando notebooks en Python, el sistema ejecuta procesos como la limpieza de datos, detección y manejo de valores nulos, estandarización y anonimización, en línea con las normativas éticas y legales aplicables. Además, transforma los datos para calcular métricas derivadas clave, como el porcentaje de créditos aprobados y tendencias de rendimiento académico. Estos procesos aseguran que los datos sean consistentes y preparados para alimentar el modelo de Machine Learning.

Model Deployment:

El modelo predictivo, desarrollado con frameworks como scikit-learn y TensorFlow, se despliega en un entorno local. Este modelo utiliza algoritmos de clasificación para identificar estudiantes en riesgo de deserción, basándose en variables como promedio académico, nivel de beca y características sociodemográficas. El despliegue asegura que el modelo sea escalable y accesible para realizar predicciones en tiempo real o en lotes, según las necesidades de los usuarios.

Predicciones y Datos:

Los resultados del modelo predictivo, como el nivel de riesgo de deserción de cada estudiante, se exportan en formatos compatibles como .xlsx, .csv y Parquet. Estos formatos permiten la integración con sistemas externos y la visualización en herramientas de análisis. Además, las predicciones pueden ser retroalimentadas al sistema para mejorar continuamente el modelo y su precisión.

Power BI Dashboard:

El dashboard interactivo de Power BI se utiliza para presentar visualizaciones de los indicadores clave de desempeño (KPIs) relacionados con la deserción estudiantil. Este panel está diseñado para diferentes grupos de usuarios, como

administradores académicos y consejeros estudiantiles, y ofrece funcionalidades como filtros por facultad, edad, PGA, etc. Las visualizaciones incluyen gráficos de barras apilados y diagramas de caja para facilitar la toma de decisiones informadas.

Alineación con el proyecto:

Esta arquitectura respalda directamente el objetivo del proyecto al consolidar datos relevantes, garantizar su calidad, generar predicciones accionables y visualizar los resultados de manera clara. Cada componente está diseñado para responder a las necesidades específicas de los usuarios finales, como la identificación temprana de estudiantes en riesgo y la asignación eficiente de recursos para reducir la deserción. Además, cumple con las normativas de privacidad y los principios éticos mencionados en el documento.

A continuación, se presenta el link del tablero de control desplegado con los datos y las predicciones ya realizadas por el modelo entrenado: (Link al PowerBI) https://app.powerbi.com/links/QIiyYX-GXy?ctid=fabd047c-ff48-492a-8bbb-8f98b9fb9cca&pbi_source=linkShare

10. Retroalimentación por parte de la organización

A lo largo del semestre, se llevaron a cabo reuniones y discusiones con los stakeholders del proyecto, quienes proporcionaron retroalimentación clave para refinar y orientar los objetivos del modelo predictivo, entre los principales comentarios e ideas recibidas, se destacan

- **La definición de desertor:** Se estableció que un estudiante será considerado como desertor si ha pasado 4 semestres consecutivos sin matricularse y no ha regresado a la universidad hasta la fecha, este criterio permitió estandarizar la definición de deserción, lo cual es crucial para garantizar la consistencia en el anállisis y las predicciones del modelo
- Se decidió que el análisis en cuanto a maestrías se limitará únicamente a los programas de maestrías presenciales, por lo que quedan excluidos programas como MISW, MAIA, MAGI, MIID, MBAV o M-RDSV. Esta delimitación asegura que los resultados sean más específicos y relevantes para los estudiantes en un formato presencial, quienes enfrentan condiciones diferentes a los programas virtuales o híbridos.
- En vista a que el modelo en primeras iteraciones produjo un overfitting, se recomendó incluir información adicional sobre cada estudiante. En primer lugar, se sugirió incluir el número total de matrículas o el total de periodos que el estudiante estuvo matriculado en la universidad y en segundo lugar, el número de reingresos o de veces que un estudiante ha regresado a la universidad después de interrumpir sus estudios.
- Por último, se sugirió crear un dashboard interactivo que permita identificar a los estudiantes con mayor probabilidad de deserción, visualizar los atributos de estos estudiantes y mostrar tendencias y patrones a través de gráficos dinámicos que faciliten la toma de decisiones por parte de profesores y otros administrativos.

11. Conclusiones

El modelo Random Forest fue seleccionado debido a su precisión (F1-Score de 0.89) y su capacidad de interpretación, ya que permite identificar factores clave como el promedio global, los semestres cursados, la edad y las tendencias en el rendimiento académico. Además, el proceso de limpieza y preparación de datos eliminó variables con altos porcentajes de nulos y valores atípicos; asimismo, se generaron variables derivadas, como el porcentaje de créditos aprobados (PCA), que mejoraron significativamente la capacidad predictiva del modelo. Por otro lado, el dashboard en Power BI integra estas predicciones para monitorear indicadores clave, identificar estudiantes en riesgo y priorizar intervenciones, optimizando recursos como el tiempo. En cuanto a las consideraciones éticas, el proyecto cumplió con normativas de protección de datos mediante la anonimización y la mitigación de riesgos, garantizando un enfoque ético. Por último, se recomienda automatizar la actualización de datos, incluir series temporales y enfocar las intervenciones en estudiantes con desafíos financieros, ya que todo esto consolida una herramienta robusta que no solo mejora la retención estudiantil, sino que también posiciona a la universidad como líder en el uso de tecnología avanzada para el bienestar de sus estudiantes.

12. Autoevaluación

	Jairo	Santiago	Catalina	Lina
Jairo	5.0	5.0	5.0	5.0
Santiago	5.0	5.0	5.0	5.0
Catalina	5.0	5.0	5.0	5.0
Lina	5.0	5.0	5.0	5.0

13. Bibliografía

[1] Universidad de los Andes. (2025). Plan de desarrollo institucional 2020-2025. https://pdi2125.uniandes.edu.co/planes-desarrollo-facultad

14. Anexo

Anexo 1. Mockup de dashboard