Informe Entrega 1 – 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.

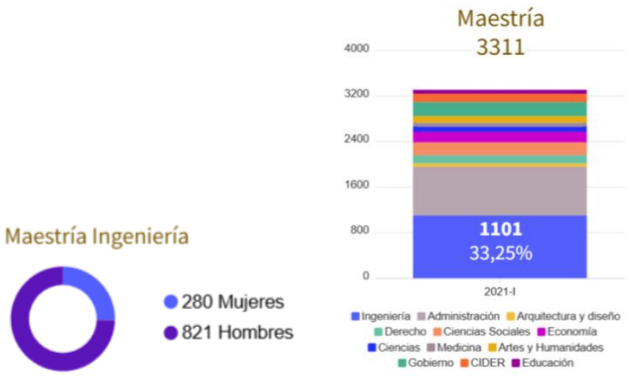


Figura 1. *Datos de 2021-1 para la maestría en Ingeniería y otras*

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%.
- Índice de bienestar PERMA: Mejorar el promedio de bienestar (especialmente en emociones positivas) de 5.85 a 6.5, fortaleciendo el sistema de bienestar estudiantil.
- Índice de experiencia académica: Incrementar el bienestar académico en población estudiantil, especialmente en estudiantes de los primeros semestres (meta de 6.68 a 7.0).

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: Tomarán decisiones estratégicas para políticas de retención y mejora en los servicios de apoyo.
2. Consejeros estudiantiles y oficinas de bienestar: Proporcionarán seguimiento y apoyo personalizado a los estudiantes en riesgo.
3. Oficina de finanzas/becas: Administrará las ayudas económicas y becas para los estudiantes en función de su riesgo de deserción.

4. Decanatura y Comité de Alertas Tempranas: Identificarán y priorizarán los casos de riesgo para intervenciones tempranas.

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 (Regresión Logística, Random Forest) para identificar el riesgo de deserción basándose en los datos históricos de los estudiantes.
3. Dashboard de visualización: Un panel interactivo que permita a los administradores académicos y consejeros ver el estado de los estudiantes en tiempo real, con alertas y filtros que faciliten la 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.

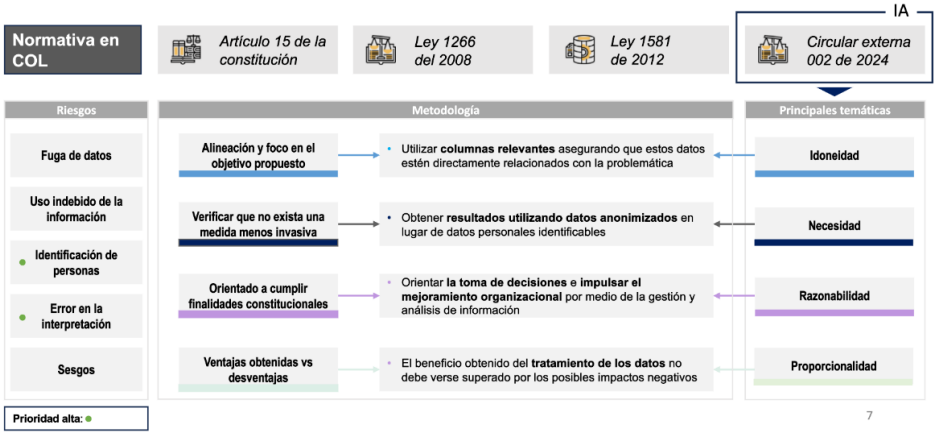


Figura 2. *Metodología de alto nivel para mitigar los riesgos*

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 académicos y sociodemográficos tienen mayor peso en la predicción de la deserción estudiantil y cómo se pueden priorizar para diseñar intervenciones más efectivas?
- ¿Cómo impactan las dificultades económicas en la tasa de deserción y qué estrategias se pueden implementar para mejorar la asignación de ayudas financieras?

4.3. Técnicas Estadísticas

Para validar las hipótesis y responder a las preguntas planteadas, se propone el uso de diversas técnicas estadísticas, cada una adaptada al contexto específico del problema:

- **ANOVA (Análisis de Varianza):** Esta técnica permitirá comparar las tasas de deserción entre distintos grupos de estudiantes, como aquellos con diferentes niveles de beca o con variaciones en su promedio académico. A través del ANOVA, se pueden identificar diferencias significativas entre grupos, lo que ayudará a determinar si ciertas características, como el promedio académico o el nivel de ayuda financiera, afectan de manera considerable la deserción.

- **Análisis de Correlación:** Este análisis servirá para explorar la relación entre variables como el promedio académico, el estrato socioeconómico, y la probabilidad de deserción. Al identificar qué factores tienen una mayor asociación con la deserción estudiantil, se podrá priorizar las variables más relevantes para el modelo predictivo, guiando de esta manera las estrategias de intervención.
- **PCA (Análisis de Componentes Principales):** Con el objetivo de simplificar el modelo y reducir la dimensionalidad del conjunto de datos, el PCA permitirá identificar las variables más influyentes en la deserción. Esta técnica facilitará la reducción de ruido en el análisis, centrándose en las características que tienen mayor impacto en el comportamiento de los estudiantes.

4.4. Visualización de Datos

También, se proponen las siguientes técnicas de visualización de datos, y cómo se pueden aplicar para complementar el análisis

- **Diagramas de Caja (Boxplots):** Son ideales para el análisis univariado de las variables numéricas, ya que permiten visualizar de manera gráfica la distribución de cada variable, así como la identificación de valores atípicos. Los boxplots ayudarán a entender la dispersión de los promedios académicos y otros indicadores en relación con la deserción.
- **Mapa de Calor (Heatmap):** Podría servir para representar gráficamente el análisis gráfico del análisis de correlación realizado en las técnicas estadísticas, lo cual ayuda nuevamente a visualizar las correlaciones entre variables críticas del negocio.
- **Gráficos de Barras Apilados:** Podría ayudar a mostrar las proporciones de los estudiantes que desertan en diferentes grupos.

4.5. Machine Learning

Para abordar directamente la problemática de la deserción, se propone el desarrollo de un modelo de clasificación que permita predecir la probabilidad de deserción de cada estudiante. Al ser una tarea de clasificación binaria (deserta o no deserta), se consideran los siguientes algoritmos de Machine Learning:

- **Random Forest**: Este algoritmo de clasificación es conocido por su capacidad de manejar grandes cantidades de datos y por su robustez frente al sobreajuste. Utiliza múltiples árboles de decisión para mejorar la precisión de la predicción y permite identificar la importancia de cada variable en la predicción de deserción.
- **Regresión Logística:** Es una técnica simple pero efectiva para problemas de clasificación binaria. La regresión logística permitirá modelar la relación entre las variables independientes (promedio académico, nivel de ayuda económica, etc.) y la probabilidad de deserción, proporcionando interpretabilidad en los coeficientes de cada variable.
- **Support Vector Machine (SVM):** Este algoritmo es especialmente útil cuando se trata de encontrar un límite claro entre las clases. SVM buscará maximizar el margen que separa a los estudiantes en riesgo de deserción de aquellos que no lo están, siendo una alternativa robusta cuando las clases son complejas de separar.

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 registros a partir del 2015 en adelante 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.

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

6. Entendimiento de los datos

6.1. Análisis exploratorio

El análisis exploratorio comenzó con 21,190 estudiantes, consolidando una base de datos unificada de 90,352 registros, distribuidos en 31 columnas (8 numéricas y 23 categóricas). Este análisis inicial se enfoca en maestrías cuyos períodos terminan en 10 y 20, excluyendo períodos intersemestrales.

6.1.1. Análisis Univariado

<u>Variables sociodemográficas</u>

El 61.02% de los estudiantes son del sexo masculino y el 38.94% son del sexo femenino, mientras que el 0.03% no informó. El 50% de los estudiantes entran a la maestría teniendo entre 30-39 años. La mayoría trabajan en el sector privado (66.3 %) y ocupa cargos de ingeniero (21.26%). Además, el 79.1% son solteros y el 99.52% reside en Colombia (82.56% en Bogotá).

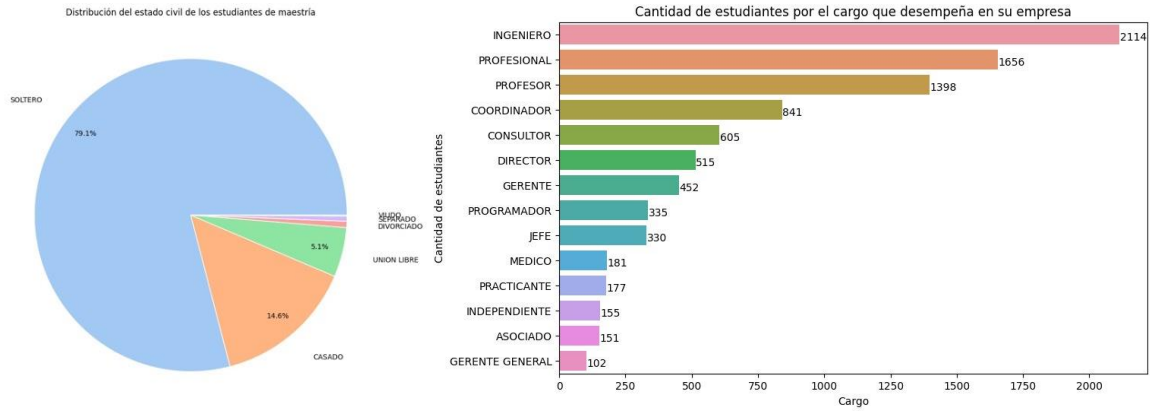


***Figura** 3. (Izquierda) Diagrama de pie que muestra la distribución del estado civil de los estudiantes de maestría. (Derecha)Gráfico de barras que muestra la cantidad de estudiantes por el cargo que desempeña en su empresa.*

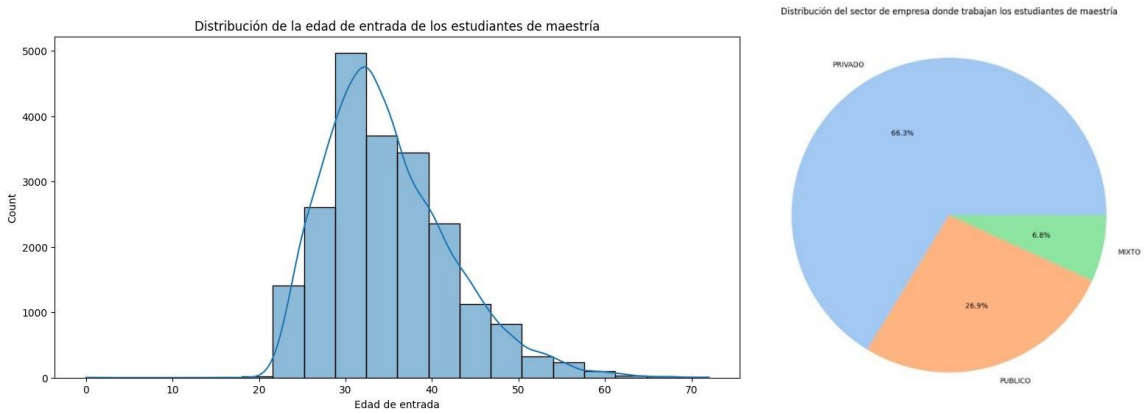


***Figura** 4. (Izqueirda) Histograma que muestra la distribución de la edad de entrada de los estudiantes de maestría. (Derecha) Diagrama de pie que muestra la distribución del sector de empresa donde trabajan los estudiantes de maestría.*

Variables académicas

- La mayoría de los estudiantes hacen una maestría en la Facultad de Ingeniería (37.28%) y Administración (24.87%).

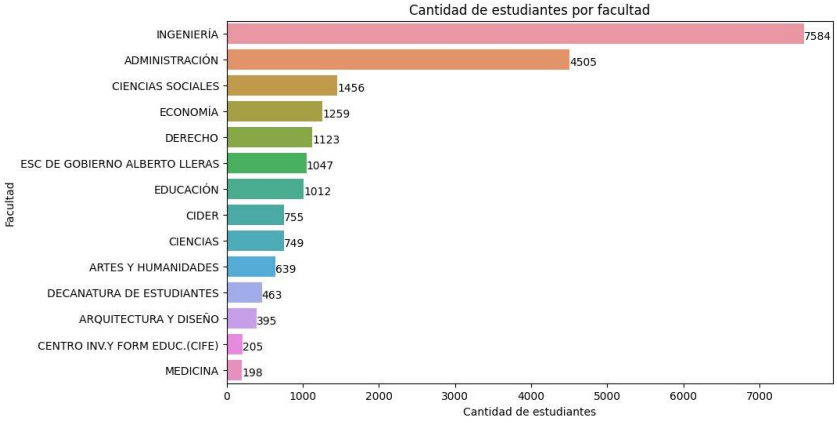


***Figura** 5. Gráfica de barras que muestra la cantidad de estudiantes de maestría por facultad.*

6.1.2. Análisis Bivariado

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.

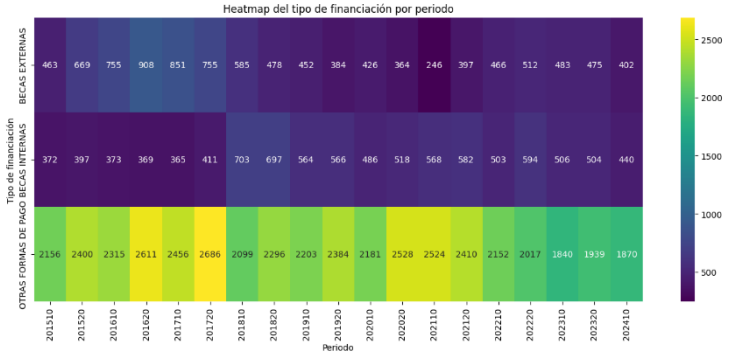


***Figura** 6. Mapa de calor del tipo de financiación de los estudiantes de maestría por periodo.*

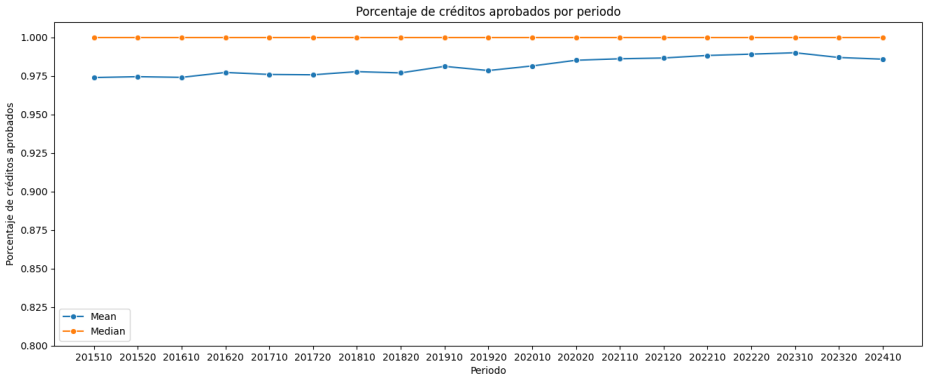


Figura 7. Gráfico de líneas que muestra el promedio y la mediana del porcentaje de créditos aprobados por periodo.

6.1.3. Análisis Multivariado

Variables académicas en el tiempo

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.

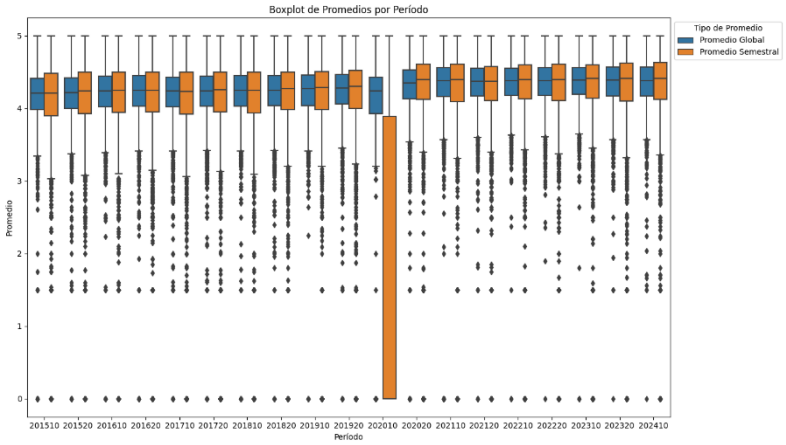


Figura 8. Diagrama de cajas y bigotes que muestra la distribución del promedio global y semestral de los estudiantes de maestría por periodo.

Variables numéricas

Se puede ver que 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 (0.37) y entre el promedio global y el porcentaje de créditos aprobados (0.38). 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.

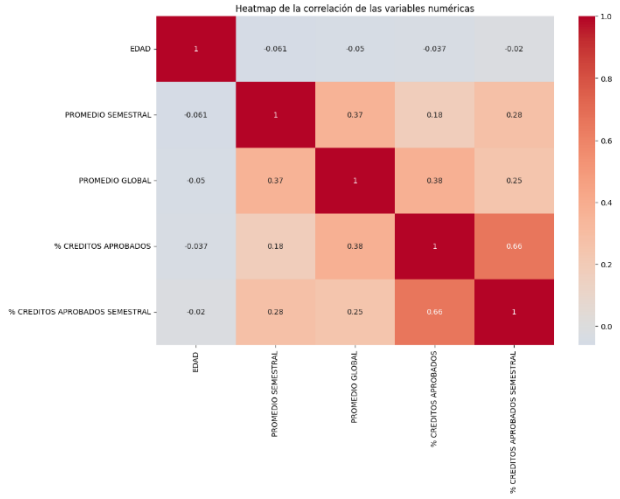


Figura 9. Mapa de calor de la correlación de las variables numéricas.

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.

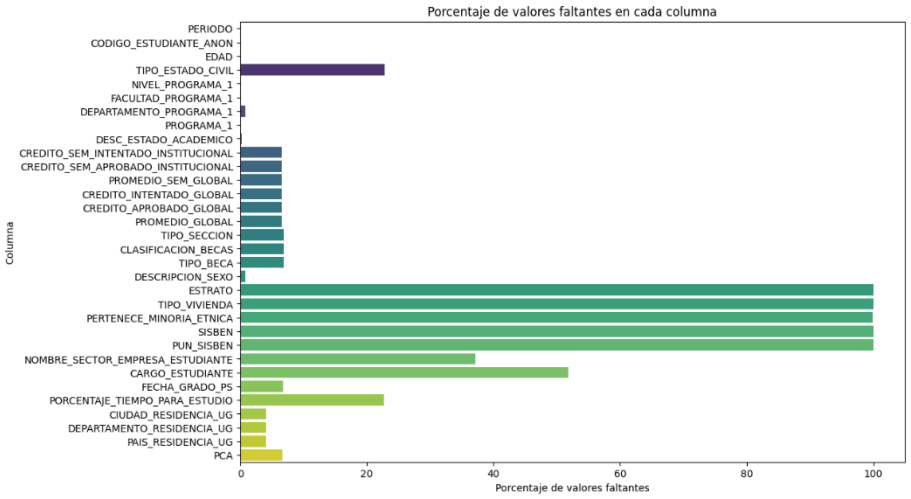


Figura 10. Gráfica de barras que muestra el porcentaje de valores faltantes o nulos por variable.

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 3.23% de las columnas (1 de 31) 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 90,352: 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 14.28% (1 de 7) presenta valores fuera del rango esperado; esto ocurrió en la columna **EDAD**, donde aparecieron edades de 1934 y 124 en 4 de los 90,352 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**.

7. Conclusiones iniciales

Dado que aún no se cuenta con una definición clara de la deserción ni con una etiqueta específica que permita identificar directamente a los estudiantes que han desertado, el análisis preliminar de los datos permite identificar ciertos factores potencialmente relevantes. Observamos que la mayoría de los estudiantes utiliza financiamiento propio o préstamos para cubrir sus estudios, mientras que aquellos que reciben apoyo financiero en forma de becas podrían enfrentar desafíos económicos que impacten su continuidad académica. Además, el rendimiento académico, especialmente en términos de promedio y porcentaje de créditos aprobados, muestra variabilidad que podría asociarse a situaciones de riesgo académico, aunque aún es necesario validar estos supuestos en función de la definición de deserción que se acuerde con el cliente. La calidad de los datos sugiere algunas áreas de mejora, como la eliminación de valores atípicos en variables como la edad y la corrección de inconsistencias en las métricas de créditos, para asegurar que el modelo predictivo sea preciso y confiable.

Las próximas acciones incluyen, primero, definir de manera clara y consensuada la métrica de deserción junto con el cliente, lo cual es fundamental para poder entrenar y evaluar el modelo predictivo adecuadamente. Asimismo, se explorará la posibilidad de incorporar series de tiempo para capturar la evolución en el desempeño de los estudiantes, añadiendo un componente temporal que podría mejorar la predicción. Finalmente, se evaluarán técnicas de aprendizaje supervisado, como Regresión Logística y Random Forest, para identificar la opción que mejor equilibre precisión y capacidad de interpretación, facilitando así la identificación de estudiantes en potencial riesgo una vez se disponga de la etiqueta de deserción.

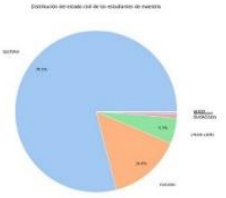
8. Bibliografía

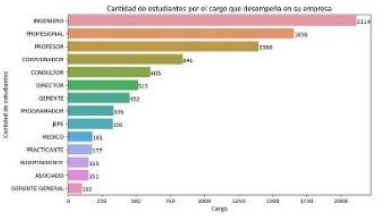
[1] Universidad de los Andes. (2025). Plan de desarrollo institucional 2020-2025. https://pdi2125.uniandes.edu.co/planes-desarrollo-facultad

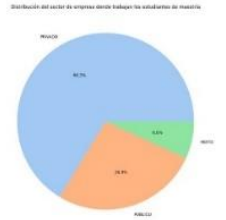
9. Anexo

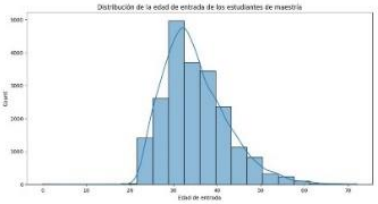
Anexo 1. Mockup de dashboard

Variables sociodemográficas

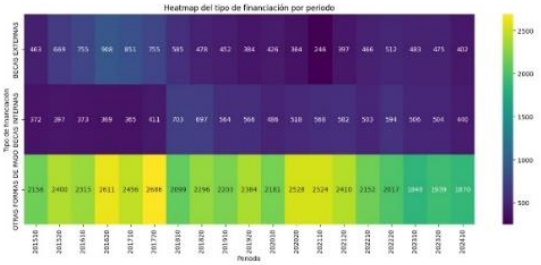


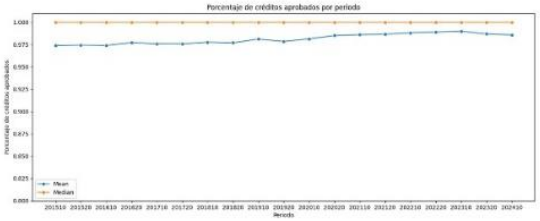




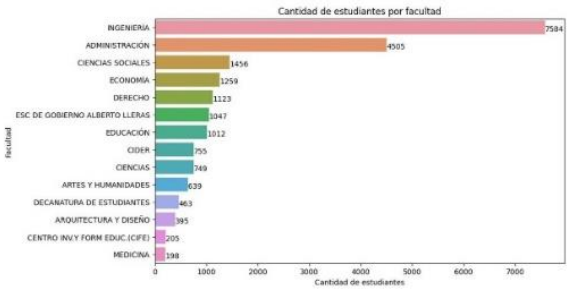


Variables financieras y académicas

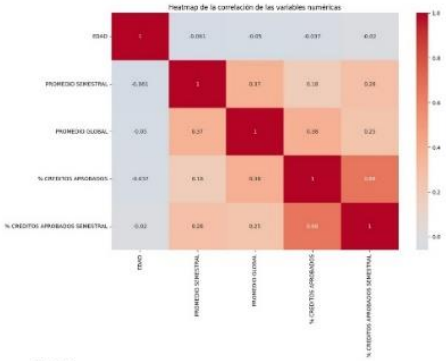




Variables académicas



Variables numéricas



Variables académicas en el tiempo