Primera Entrega Proyecto Final
Ciencia de Datos Aplicada

Presentado por:

Integrante	Código
Adriana María Ríos	200726240
Andrés Francisco Borda Rincón	201729184
Daniel Felipe Vargas Ulloa	202123899
Diego Alberto Rodríguez Cruz	202110109

Docente(s):
Fabian Camilo Peña
Juan Pablo Reyes
Nancy Marcela Cubillos

Universidad de los Andes
Facultad de Ingeniería
Departamento de Ingeniería de Sistemas y Computación
Maestría en Ingeniería de Información - MINE

2024-II

Definición de la problemática y entendimiento del negocio

Problemática

Una de las principales dificultades es la dificultad de identificar a los estudiantes tempranamente que están en riesgo de enfrentar problemas académicos o abandonar sus estudios. Las intervenciones son a menudo menos efectivas porque las señales de riesgo no aparecen hasta que el estudiante ya ha tomado la decisión de abandonar. Los patrones y las señales de alerta tempranas se pueden detectar utilizando técnicas de analítica y aprendizaje automático avanzados, lo que facilita una posible intervención.

Actualmente, muchas intervenciones para apoyar a los estudiantes son generalizadas y no se ajustan a las necesidades individuales de cada estudiante. Esto puede resultar en un bajo rendimiento de los programas de apoyo. El uso de modelos predictivos y datos permite la personalización de las intervenciones, asegurando que cada estudiante reciba el tipo y nivel de apoyo que necesita para superar sus desafíos específicos.

Muchas veces, las decisiones sobre intervenciones y políticas educativas se toman en base a suposiciones o datos insuficientes. Esto puede conducir a enfoques que no abordan las causas de la deserción. El uso de un enfoque basado en datos aumenta la probabilidad de éxito de las intervenciones implementadas al permitir que las decisiones se basen en análisis detallados y evidencia concreta.

Estrategia institucional

- La Escuela Nacional del Deporte se centra en formar profesionales altamente capacitados en el ámbito del deporte, la salud y la administración.
- Se busca promover la retención y el éxito académico de los estudiantes a través de la implementación de programas que respondan a sus necesidades.

Datos del sector educación

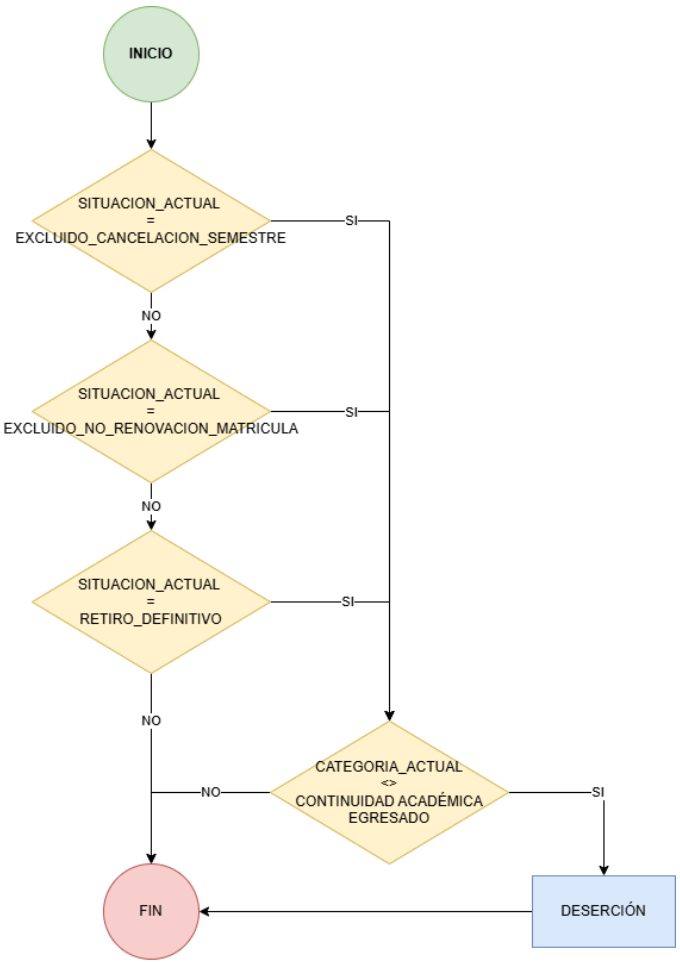
- Existen tendencias crecientes en la deserción estudiantil en instituciones de educación superior, lo que resalta la necesidad de intervenciones basadas en datos para mejorar la retención.
- La analítica de datos en educación se ha vuelto crucial, permitiendo identificar patrones de riesgo y diseñar estrategias efectivas de intervención.

Relevancia

- La deserción estudiantil puede comprometer la reputación de la institución y su capacidad para atraer nuevos estudiantes.
- Identificar y abordar los riesgos académicos de manera temprana es fundamental para mejorar la experiencia educativa y asegurar el éxito de los estudiantes.

Para la Escuela Nacional del Deporte - ENDEPORTE, se considera que un estudiante ha desertado su ciclo académico de acuerdo con las siguientes reglas:

- Si la variable denominada “situación actual” contiene cumple las siguientes condiciones “excluido cancelación semestre” o “excluido no renovación matricula” o “retiro definitivo” y,
- La variable denominada “categoría actual” el valor es diferente de “continuidad académica - egresado” se considera “Deserción”.



Objetivo general

Desarrollar un sistema integral de análisis de datos que permita identificar tempranamente a los estudiantes en riesgo de enfrentar problemas académicos o de abandonar sus estudios, con el fin de diseñar y aplicar intervenciones personalizadas y efectivas que mejoren la retención estudiantil, optimicen el uso de recursos institucionales y potencien el éxito académico de todos los estudiantes.

Objetivos del Proyecto

- Recopilar y analizar datos históricos sobre el rendimiento académico, asistencia, participación en actividades extracurriculares y factores socioeconómicos de los estudiantes para identificar variables clave que influyan en la deserción
- Desarrollar un modelo predictivo para identificar estudiantes en riesgo de enfrentar problemas académicos o de deserción.
- Implementar estrategias de intervención personalizadas basadas en los datos recolectados.
- Evaluar la eficacia del modelo mediante métricas de rendimiento, como la precisión, la tasa de falsos positivos y la tasa de verdaderos positivos, asegurando que las predicciones sean fiables y útiles para la toma de decisiones.

Métricas de Negocio (KPIs) para Evaluación:

- Eficiencia en la asignación de recursos basada en los análisis de datos (por ejemplo, mejora en la retención en cursos críticos).

Ideación

El producto en datos va a ir enfocado en los modelos de análisis de la deserción estudiantil.

Usuarios potenciales

ROL	Descripción
Estudiantes	Buscan apoyo académico
Docentes	Necesitan herramientas para la identificación de estudiantes en riesgo en riesgo de deserción
Coordinadores académicos	Requieren datos para la toma de decisiones
Consejeros estudiantiles/Apoyo Académico	Pueden utilizar datos socioeconómicos y académicos para brindar apoyo emocional y financiero a los estudiantes.

Procesos actuales

Área	
Registro y control	Se registran calificaciones, sin embargo, con los datos obtenidos no se tiene un análisis de estos.
Apoyo académico	Se está sin un modelo o datos que indiquen las acciones a tomar para evitar la deserción estudiantil.

Dolores actuales

- Identificación tardía de los estudiantes que desertan
- Falta una estrategia personalizada a las necesidades de cada estudiante
- Falta de información de eventos o actividades en las que participan durante su jornada académica.

Para el desarrollo del producto de datos destinado a monitorear la deserción de estudiantes en la Escuela del Deporte, se deben definir los siguientes componentes tecnológicos:

1. **Motor de Análisis Predictivo:** Un modelo de aprendizaje automático identificará patrones de riesgo de deserción basado en el historial académico y otros factores relevantes.
2. **Interfaz de Visualización:** Un dashboard interactivo mostrará métricas clave, tales como el número total de estudiantes, la tasa de deserción esperada y el impacto en el promedio general, además de permitir la visualización detallada por semestre, estrato y asignaturas, como se ilustra en el mockup proporcionado.

El mockup del dashboard simula cómo se presentarán los datos, destacando las métricas de riesgo, distribución por notas y cursos, y segmentación por género y estrato. Este enfoque permitirá identificar y abordar los factores que contribuyen a la deserción de manera proactiva (dashboard disponible en el repositorio para mejor visualización).

Responsabilidad

Uso de datos

En Colombia, la **ley 1581 de 2012** regula la protección y uso de datos personales. Esta ley tiene como objeto desarrollar el derecho constitucional que tienen todas las personas a conocer, actualizar y rectificar las informaciones que se hayan recogido sobre ellas en bases de datos o archivos, así como garantizar también el derecho a la información. A continuación, los principios rectores presentes en el artículo 4 de la ley y las implicaciones que tienen en el presente estudio:

Principios	Implicaciones en el estudio
Principio de legalidad	Dado el uso de datos personales, el estudio debe ser una actividad reglada según la ley 1581 de 2012.
Principio de finalidad	Garantizar que la finalidad del estudio corresponda a una necesidad dentro del marco legal colombiano.
Principio de libertad	Autorización por parte de los estudiantes para el tratamiento de los datos personales, dicha autorización es dada por el estudiante en su proceso de matrícula.
Principio de veracidad o calidad	Los datos del estudio son veraces y de calidad ya que provienen del área de la entidad encargada de administrar esta información.
Principio de transparencia	Los datos suministrados por la entidad para el estudio son los mismos que existen en la oficina de control y registro académico de la entidad, los cuales están disponibles para consulta de los estudiantes.
Principio de acceso	Restringir el acceso a los datos solo para participantes autorizados, adicionalmente la información o los resultados del estudio no deben divulgarse públicamente.
Principio de seguridad	La información suministrada por la entidad debe estar almacenada en repositorios privados y con autenticación de la universidad y el proyecto Caoba.
Principio de confidencialidad	Se debe garantizar la custodia de la información antes, durante y después del estudio según lo estipulado en el acuerdo de confidencialidad con la entidad.

Inteligencia Artificial

Con respecto al uso de la Inteligencia Artificial, en Colombia actualmente no existe una ley que la regule, sin embargo, existen avances al respecto y se registran en el

documento *CONPES 3975 POLÍTICA NACIONAL PARA LA TRANSFORMACIÓN DIGITAL E INTELIGENCIA ARTIFICIAL*. En el documento CONPES 3975 se encuentran las directrices para implementar leyes que permitan regular el desarrollo de la IA en Colombia, siguiendo los principios de la OCDE (Organización para la Cooperación y el Desarrollo Económico). De acuerdo con el estudio de la OCDE sobre gobernanza publica para el uso estratégico y responsable de la inteligencia artificia, en el capítulo 4 considera las acciones necesarias para desarrollar un abordaje responsable, fiable y centrado en el ser humano.

Acciones	Implicaciones en el estudio
Ética de los datos	Cumpliendo los principios de la ley 1581 de 2012 estaríamos alineados con esta recomendación, sin embargo, consideramos que es importante tener en cuenta el principio sobre la *adopción de un enfoque de riesgo con respecto a la automatización de las decisiones*, es decir, sugerir a la entidad cotejar los resultados del modelo con el criterio de las personas expertas en la problemática de deserción de la entidad.
IA Fiable	Consideramos que el objetivo del estudio está alineado con esta acción ya que el propósito es garantizar el bienestar de una población de estudiantes que lo necesitan respetando su diversidad y equidad.
Imparcialidad y atenuación del sesgo	Consideramos que la imparcialidad y la atenuación del sesgo son garantizadas desde los datos provistos por la entidad, ya que corresponde a información real, imparcial y sin sesgo de discriminación.
Transparencia y Explicabilidad	Esta acción debe ser aplicada por la entidad en lo referente a hacer público tanto las decisiones que se tomen con base en el algoritmo y el proceso que tiene que llevar a cabo la entidad para hacerlo.
Seguridad y Protección	Esta acción debe ser evaluada por la entidad ya que implica desarrollar un proceso de gestión de riesgo que garantice un uso apropiado y seguro del algoritmo.
Asegurar un abordaje inclusivo y centrado en el usuario.	La inclusión se garantiza con la multidisciplinariedad que existe en los datos de entrenamiento del algoritmo ya que corresponde a datos de estudiantes con diferentes antecedentes educativos, también son diversos ya que corresponde a estudiantes de diferentes géneros, razas, edades, niveles socioeconómicos y orígenes geográficos.

Enfoque analítico

Pregunta de Negocio

¿Cuál es la probabilidad de deserción de los estudiantes de carreras universitarias y tecnológicas del Instituto Nacional del Deporte en Colombia?

Técnicas estadísticas de visualización

- **Gráficos de barras:** Se utilizará para representar la frecuencia de las variables categóricas como programa, género y estrato.
- **Histogramas:** Se utilizará para representar la frecuencia de variables cuantitativas como el promedio del semestre, promedio general y edad.
- **Gráficos de torta:** Se utilizará para visualizar la distribución de estudiantes que desertan vs los que no desertan.
- **Diagramas de cajas:** Se utilizará para visualizar las medidas de tendencia central y outliers de las variables cuantitativas.
- **Gráficos de correlación:** Se utilizará para identificar correlación entre las variables cuantitativas de semestre, promedio general y edad.

Técnicas de Machine Learning

Se va a utilizar una técnica de aprendizaje supervisado, dado que contamos con datos históricos etiquetados con la variable de interés Deserción, clasificada en estudiantes que presentaron Deserción y los que no (No Deserción). Dado el contexto de la problemática, el cual corresponde a identificar estudiantes que probablemente desertan de los que no, consideramos que un clasificador es el algoritmo de aprendizaje supervisado que mejor se ajusta a la necesidad. Teniendo en cuenta que los datos de entrenamiento están desbalanceados con respecto a la variable de interés, es decir el 88,8% de los datos corresponden a estudiantes que presentaron NO DESERCIÓN y el restante 11,2% a estudiantes que presentaron DESERCIÓN, consideramos que para mitigar el efecto de datos desbalanceados podemos trabajar con los algoritmos de clasificación: Random Forest Balanceado y Regresión Logística con Ponderación de Clases.

Métricas para evaluar la calidad del modelo

Dada la naturaleza de la problemática, es importante identificar, si no todos, la mayor parte de los estudiantes que puedan desertar y que efectivamente estén en riesgo y necesiten un apoyo, esto último, permite optimizar los recursos de la entidad. Lo anterior sugiere buen balance entre las métricas de precisión y recall del modelo, por lo tanto, consideramos que las métricas más apropiadas para evaluación del modelo serian: f1 score, R2 y/o Coeficiente de Gini.

Recolección de datos

Para el proyecto se dispone de dos fuentes de datos principales, cada una con características particulares que aportarán información relevante para el análisis.

La primera fuente de datos es el archivo *Demograficos.xlsx*, el cual contiene información académica y demográfica de cada estudiante. Entre los datos recopilados se incluyen el promedio académico, el programa al que pertenece, la edad, la ciudad de origen, y si presenta alguna discapacidad, entre otros atributos. Este conjunto de datos será de gran utilidad para caracterizar el perfil de los estudiantes y analizar posibles correlaciones entre sus características demográficas y su rendimiento académico. Además, permitirá identificar patrones específicos en subgrupos de estudiantes, lo que contribuirá a un análisis más detallado respecto a que tipo de estudiante es más propenso a abandonar sus estudios.

La segunda fuente de datos es el archivo *HistoricoNotas.xlsx*, que contiene un registro histórico de las calificaciones de cada estudiante en los cursos que ha tomado. Este dataset será fundamental para evaluar el desempeño académico a lo largo del tiempo y determinar tendencias en las calificaciones. También permitirá realizar un análisis longitudinal del progreso académico de los estudiantes, identificando cursos o periodos específicos en los que se presenten variaciones significativas en el rendimiento. Al combinar esta información con los datos demográficos, será posible desarrollar modelos que consideren tanto el contexto personal como el histórico del desempeño.

Para ambas fuentes, se presenta un archivo adjunto con el diccionario de datos que construimos para cada una de las fuentes de datos (no se incluye en este documento dado el gran espacio que ocuparían).

Entendimiento de los datos

Introducción al Dataset

La población objetivo de este análisis son estudiantes de pregrado de los cuales analizaremos su periodo más frecuente. El enfoque en el último periodo de cada estudiante permite capturar la situación más actual de su trayectoria académica. El objetivo de este enfoque es, capturar las características que causan el retiro en los estudiantes en el semestre en que se retiran.

Variables principales

- Variable de Interés: DESERCIÓN - Categórica, con valores "Deserción" y "No deserción", indicando si un estudiante ha desertado.

La variable de interés se generó según la definición institucional, presentada anteriormente. Esta variable objetivo presenta una distribución desbalanceada, con más casos negativos, siendo solo 11.2% los estudiantes que desertan.

Evaluación de la calidad de los datos

Para asegurar la confiabilidad de los análisis y garantizar que los resultados reflejen la realidad de la población estudiada, se realizó una evaluación exhaustiva de la calidad de los datos en varias dimensiones clave.

A continuación, se detallan los hallazgos y las acciones realizadas en cada dimensión:

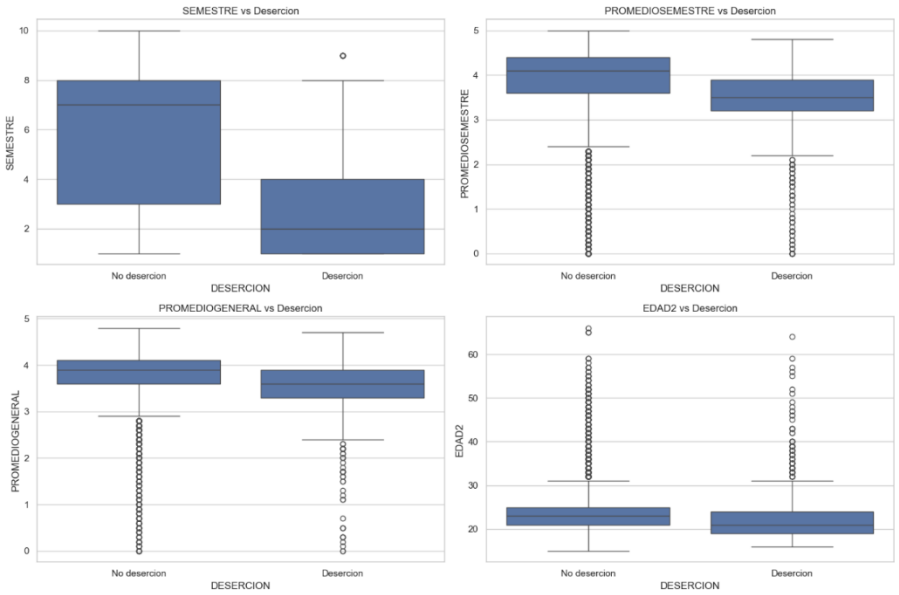
- Unicidad
  - Se verificó la existencia de registros duplicados en el dataset y no se encontraron entradas repetidas. Esto asegura que cada registro es único y representa a un estudiante distinto, evitando así que cualquier análisis se vea sesgado por la presencia de datos duplicados.
- Completitud
  - Columnas Incompletas: Se identificaron columnas con datos incompletos que no cumplían con los estándares mínimos de calidad para el análisis.
  - Columna del SISBÉN: La columna relacionada con el puntaje del SISBÉN presenta una tasa de incompletitud del 82%. Dado el alto porcentaje de valores faltantes, este factor no podrá ser tomado en cuenta para el análisis.
  - Otras Columnas con Valores Faltantes: Se revisaron las demás columnas y se constató que había unas con un porcentaje de valores faltantes muy bajo y estos registros fueron eliminados.
- Consistencia
  - Valores Anómalos: Se detectaron valores inconsistentes en algunas de las variables numéricas y categóricas, los cuales podrían afectar la precisión del análisis. La variable de edad contenía valores con errores y digitación como 120 años, el cual debió ser corregido.

Selección de variables

Para seleccionar las variables que se utilizaran en el modelo se realizara un análisis univariado y multivariado de las relaciones entre las variables.

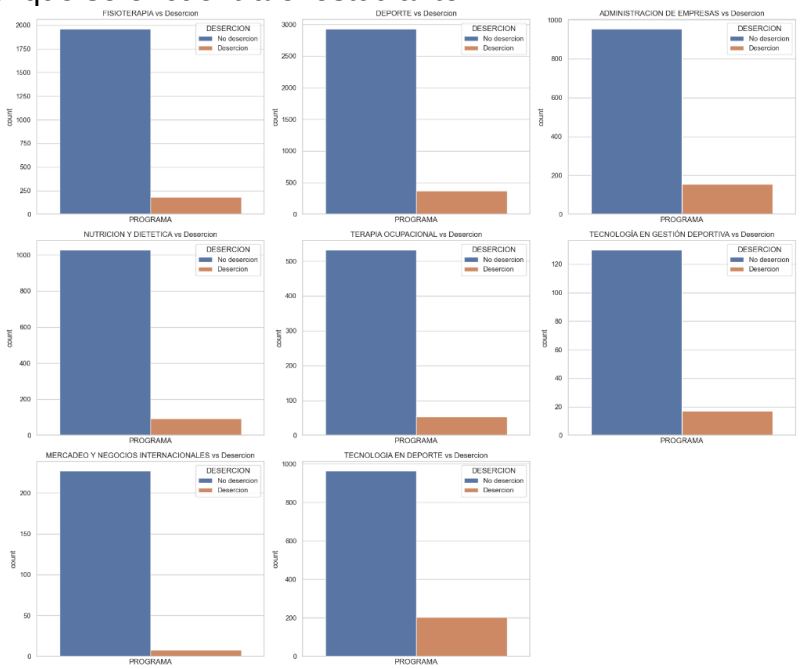
- Análisis univariado

Empezando con las variables numéricas. Se analizo como se distribuyen estas frente a la variable de interés.

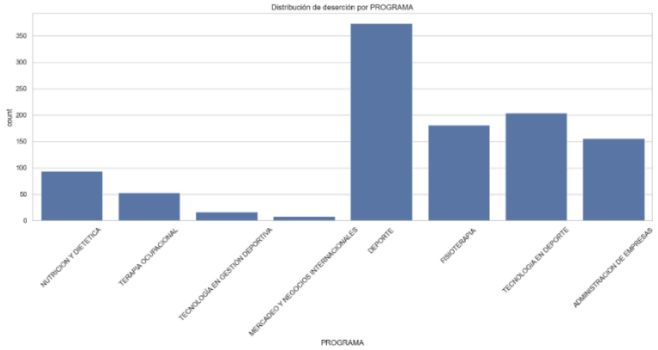


Analizando gráficamente la distribución de las dos poblaciones frente a estas tres variables podemos observar como la media en los cuatro casos es diferente mostrando una posible relación entre estas variables y la variable de interés. Podemos observar que los estudiantes que desertaron están principalmente en semestres tempranos, con promedios más bajos.

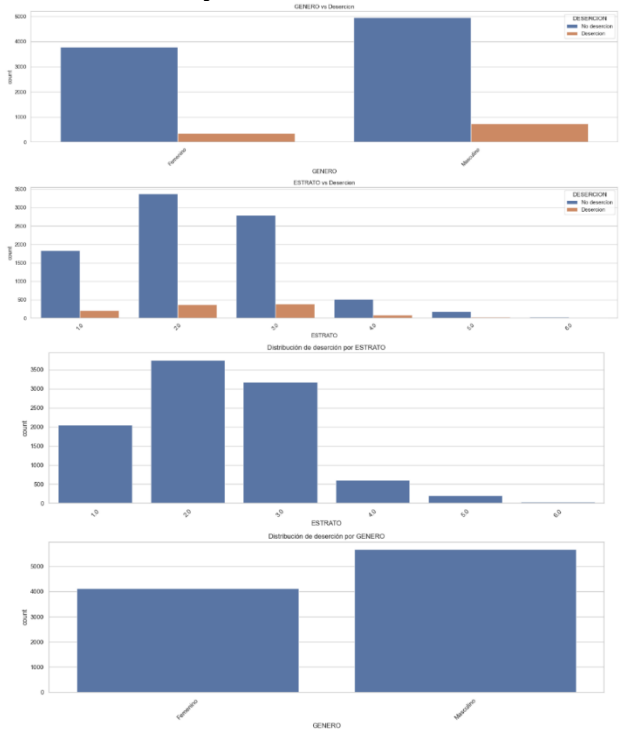
Observando ahora la distribución de la variable que contiene la información sobre el programa en el que se encuentra el estudiante.



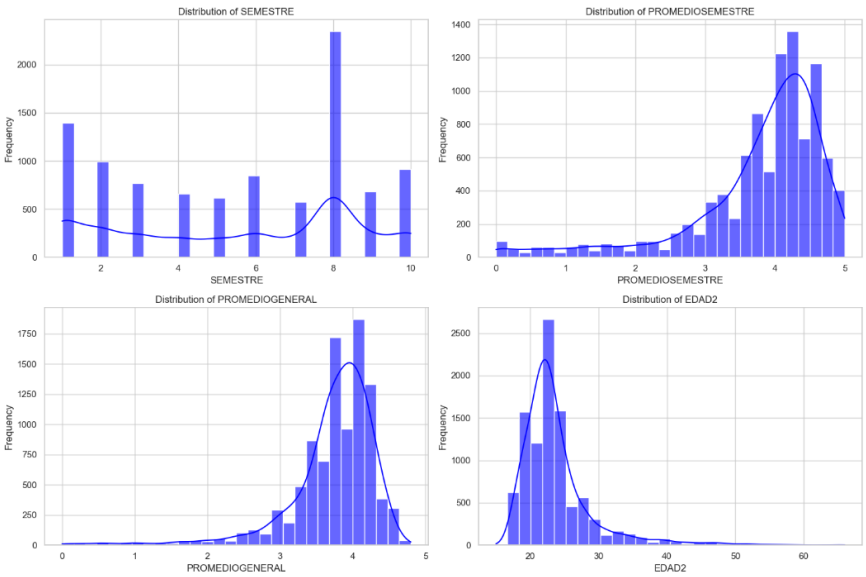
Mirando ahora esta distribución, pero solo registros de estudiantes en deserción.



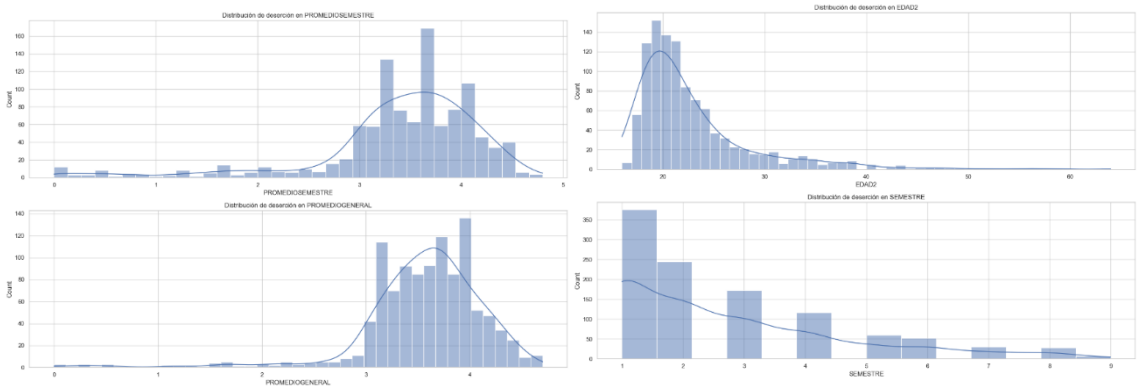
En las gráficas podemos determinar que el programa de Deporte tiene una mayor deserción que otros programas mientras que mercadeo y tecnología en gestión deportiva la deserción es muy baja. Esto nos indica que esta variable también tiene una relación con la variable de interés y nos permite identificar algunos programas que tienen una mayor problemática.
Ahora, observando la distribución en las variables demográficas de los estudiantes podemos analizar como estas influyen en la deserción de los estudiantes.



Por el lado de la variable género, gráficamente no encontramos que exista una relación marcada entre esta y la variable de interés. SI bien hay más retiros de estudiantes masculinos esto se debe a que en general hay más estudiantes de este género en la institución. Para el caso del estrato si fue posible observar una diferencia en como esta variable afecta la variable de interés y se encontró que en los estratos 2 y 3 se encuentra la mayoría de los retiros.
Ahora para comparar estas distribuciones exploraremos los dos conjuntos de estudiantes que tenemos en las diferentes variables que estamos analizando.
No deserción:

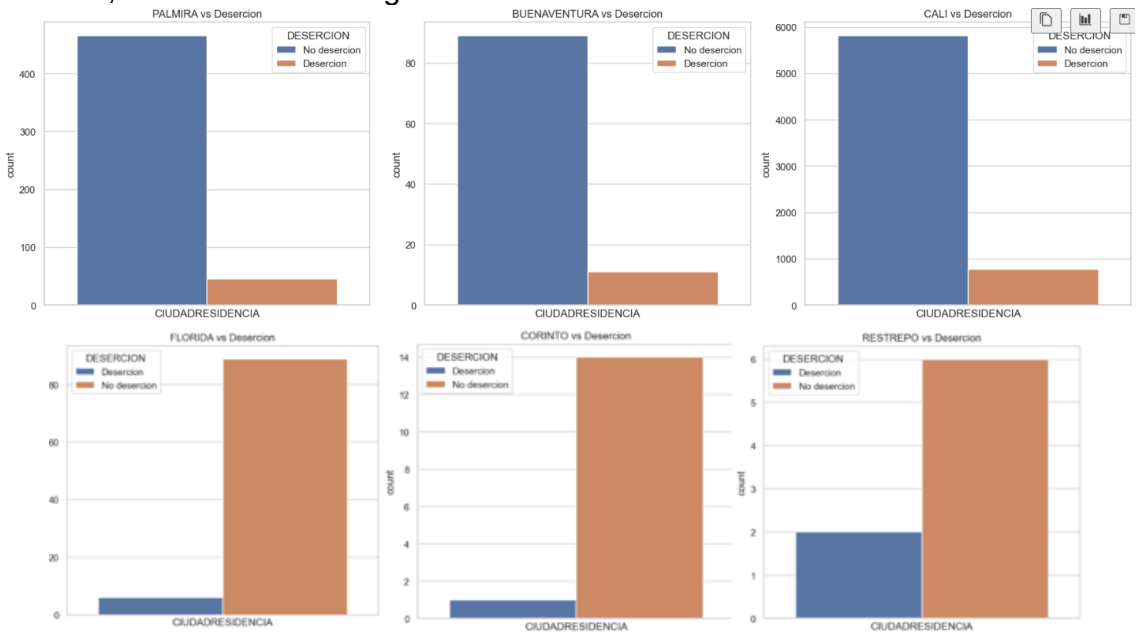


Deserción:



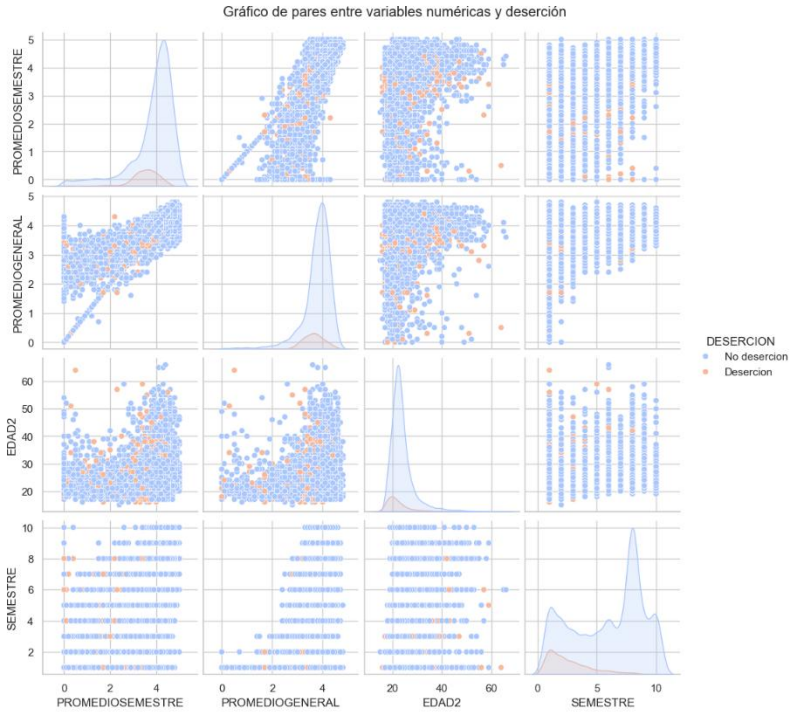
En estas graficas comparando las distribuciones de nuestras dos poblaciones se evidencia como estas en efecto afectan la variable de interés. En la distribución de semestre podemos ver como la distribución de los desertores está fuertemente ladeada hacia los semestres tempranos a comparación de la distribución más homogénea de nuestra otra población. En el caso de los promedios también podemos observar cómo estas distribuciones en la población de interés tienen una media menor y además están más ladeados hacia las notas más bajas. Por último, analizando las distribuciones de edad, aunque la media este alrededor del mismo punto parece que en la población que deserta hay más varianza y esta esta más ladeada hacia edades mayores.

Por último, analizaremos el lugar donde residen los estudiantes.



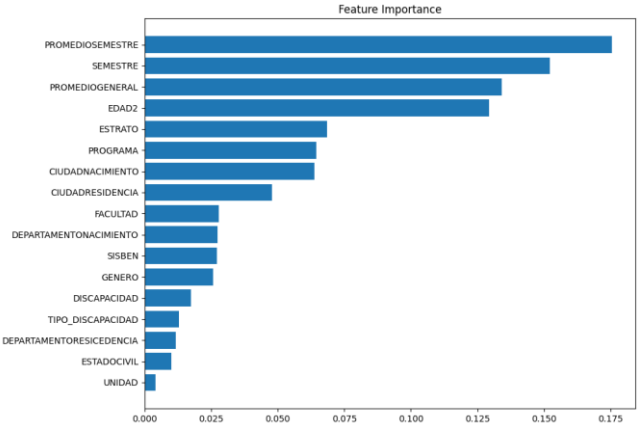
Aquí podemos identificar como las poblaciones cercanas a Cali tienen mucha más deserción. Esto nos indica que los estudiantes que viven en la ciudad en la que está la institución y no deben realizar grandes desplazamientos para atender a clase tienen menores probabilidades de deserción.

- Análisis multivariado



Viendo el resultado de esta grafica bivariado podemos ver que no podemos identificar ningún clúster de datos. Sin embargo, si se hace evidente la alta correlación entre los datos de promedio general y semestral (como es de esperarse) y de esto podemos concluir que solo se debe utilizar una dado que esta información es redundante.

Para complementar la selección de variables primero se realizó un algoritmo de Random Forest para así tener un primer indicio de cuales características son las más influyentes en la variable de interés.



Basado en todos los análisis anteriores estas son las variables elegidas:

- PROGRAMA: Programa académico del estudiante.
- GENERO: Género del estudiante.
- ESTRATO: Nivel socioeconómico del estudiante.
- SEMESTRE: Semestre académico actual del estudiante.
- CIUDADRESIDENCIA: Departamento y ciudad de residencia del estudiante.
- CIUDADNACIMIENTO: Departamento y ciudad de nacimiento del estudiante.
- PROMEDIOSEMESTRE: Promedio alcanzado en el último semestre.
- EDAD: Edad del estudiante.

Conclusiones

- Se encontró una alta correlación entre el promedio del último semestre, semestre cursado y estrato sobre la tasa de deserción de los estudiantes de ENDEPORTE.
- Se evidenció que los estudiantes residentes en municipios aledaños a Cali tienen una mayor tasa de deserción, posiblemente relacionada con el tiempo de desplazamiento hasta la entidad.
- Los estudiantes que desertaron mostraron un promedio en su último semestre que se sitúa 0.47 desviaciones estándar por debajo del promedio general de todos los estudiantes. Asimismo, se observó que la mayor tasa de deserción ocurre durante los primeros tres semestres, lo que podría atribuirse a la dificultad de las materias de ciencias básicas que se cursan en ese periodo.
- Con el análisis realizado, se evidencia que los estratos 2 y 3 presentan una tasa de deserción significativamente mayor, lo que sugiere la necesidad de implementar estrategias de apoyo específicas para estos grupos.

Validación Estadística:

- Se comprobó la significancia estadística de todas las conclusiones para las cuales se pudo realizar una prueba, con un p-value menor al 0.05.
- En efecto, todas las conclusiones para las cuales fue posible llevar a cabo una prueba resultaron estadísticamente significativas.

Acciones próximas:

- Compartir los hallazgos con la entidad, con el fin de ser orientados en el refinamiento del modelo.
- Complementar el conocimiento que tiene la entidad sobre la problemática de deserción y así apoyar la toma de decisiones al respecto.
- Experimentar diferentes técnicas de modelado estadístico que permitan identificar el modelo que mejor se ajuste al contexto de los datos y la naturaleza del problema a solucionar.

Bibliografía

Biblioguías. de 2024). *Gestión de datos de investigación: Protección, derechos y acceso a los datos*. Obtenido de https://biblioguias.cepal.org/c.php?g=495473&p=4396793

Institución Universitaria Escuela Nacional del Deporte. de 2024). *Institución Universitaria Escuela Nacional del Deporte: Misión*. Obtenido de https://endeporte.edu.co/publicaciones/1/mision/