Entrega Final
Modelo Predictivo Deserción Estudiantil
Escuela Nacional del Deporte
Proyecto Alianza Caoba
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

PRIMERA ENTREGA
Definición de la problemática y entendimiento del negocio

Definición de deserción de la entidad:

La Escuela Nacional del Deporte – ENDEPORTE al ser una entidad educativa de carácter público, se acoge la definición general de Deserción Estudiantil dada por el Ministerio de Educación Nacional de Colombia el cual indica que la Deserción Estudiantil puede entenderse como: el abandono del sistema escolar por parte de los estudiantes, provocado por la combinación de factores que se generan tanto al interior del sistema como en contextos de tipo social, familiar, individual y del entorno. La entidad, alineada con la definición general de deserción del Ministerio de Educación Nacional, considera que un estudiante ha desertado en los siguientes casos específicos:

1. El estudiante presenta una solicitud de **retiro definitivo** de sus estudios ante la entidad.
2. El estudiante presenta una solicitud de **cancelación de semestre** ante la entidad y no registra matricula en los siguientes 2 semestres.
3. Una vez finalizado el semestre, el estudiante **no renueva matricula** durante los siguientes 2 semestres.

ENDEPORTE ha proporcionado una regla de clasificación de la deserción, la cual se puede consultar en el Anexo 1: Regla.

Métricas de negocio (KPIs) para evaluación:

Tasa de deserción anual (TDA): Mide el porcentaje de estudiantes que abandonaron sus estudios durante el año con respecto al total de estudiantes.

$$TDA = \frac{\#\ Estudiantes\ desertores}{\#\ Total\ de\ Estudiantes} x100$$

Tasa de efectividad de las intervenciones (TEI): Mide la diferencia entre los estudiantes intervenidos y que desertaron con respecto al total de los estudiantes intervenidos.

$$TEI = \left(1 - \frac{\#\ Estudiantes\ desertores}{\#\ Estudiantes\ Intervenidos}\right) x100$$

Ideación

Consultamos con la entidad y definitivamente los estudiantes, aunque son el objeto de estudio, no hacen parte de los usuarios potenciales del producto de datos, ya que ésta será una herramienta de uso administrativo para ENDEPORTE. Por lo anterior se actualiza la tabla con los usuarios potenciales del producto de datos.

ROL	Descripción
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.

Responsabilidad

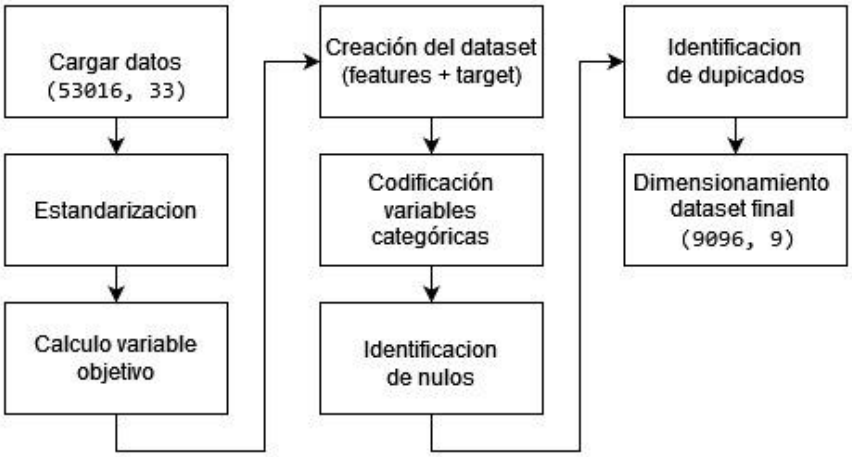
Al presente proyecto aplica la ley colombiana de protección datos 1581 de 2012. Con respecto al uso de la Inteligencia Artificial se aplicarán los lineamientos indicados en el documento *CONPES 3975 POLÍTICA NACIONAL PARA LA TRANSFORMACIÓN DIGITAL E INTELIGENCIA ARTIFICIAL.*
El detalle de estos documentos se puede consultar en Anexo 2: Responsabilidad.

Enfoque analítico

Pregunta de Negocio
¿Cuál es la probabilidad de deserción de los estudiantes, que factores influyen y en qué proporción?

Métricas para evaluar la calidad del modelo
Consideramos que la métrica de **Recall** se relaciona con el **KPI** de **Tasa de Deserción Anual (TDA)** ya que a mayor Recall, el modelo identifica un mayor número de estudiantes en riesgo de deserción que puede intervenir para lograr su retención, acción que debe reflejarse en la disminución de la TDA. La métrica de **Precisión** se relaciona con el **KPI** de **Tasa de efectividad de las intervenciones (TEI)** ya que una mayor tasa de precisión corresponde a una menor cantidad de falsos positivos en el modelo, lo que asegura que las intervenciones se realizaran sobre los estudiantes que verdaderamente están en riesgo de deserción lo que apalanca también la optimización de los recursos de la entidad. Dado que, en el presente contexto, una mayor efectividad de las intervenciones disminuirá la tasa de deserción (TDA), consideramos que debe existir un balance entre la Precisión y el Recall del modelo por lo que las métricas más apropiadas para evaluación del modelo serian: f1 score y/o Coeficiente de Gini.

ENTREGA FINAL
Preparación de datos



- **Cargar datos:** Sube la información demográfica de estudiantes de dos cohortes (2017-1 y 2021-1) que han finalizado sus estudios de manera exitosa y de quienes han desertado, para un dataset inicial con 53.016 registros y 33 columnas.
- **Estandarización:** Se ajusta el dataset inicial de acuerdo con las definiciones dadas por la entidad: excluir especializaciones y tomar la información demográfica correspondiente al último periodo académico de cada estudiante ya que el código de estudiante se repite tantas veces como periodos académicos haya cursado.
- **Calculo variable objetivo:** Se crea la variable objetivo: **target** de acuerdo con la siguiente definición: si la variable denominada “situación actual” contiene las siguientes condiciones “excluido cancelación semestre” o “excluido no renovación matricula” o “retiro definitivo” y el valor de la variable “categoría actual” es diferente de “continuidad académica - egresado” se considera “**Deserción**”.
- **Creación del dataset (features + target):** Las variables de entrada (features) seleccionadas son: ESTRATO, SEMESTRE, PROMEDIOSEMESTRE, EDAD2, PROGRAMA, GENERO, CIUDADRESIDENCIA, CIUDADNACIMIENTO, cuya alta relación con la Deserción (target) se demostró en el análisis exploratorio realizado para la primera entrega del proyecto.
- **Codificación de variables categóricas:** Se utiliza la función Labelencoder para convertir a valores numéricos las siguientes variables categóricas: PROGRAMA, GENERO, CIUDADRESIDENCIA, CIUDADNACIMIENTO.
- **Identificación de nulos:** Se aplica el siguiente código para identificar el porcentaje de nulos de las columnas del dataset: *df[[target_var]+features].isnull().sum() / df.shape[0]*, sin embargo, se observa que solo la columna estrato presenta valores nulos y en un porcentaje muy bajo (menor a 1%) por lo que se conserva esta variable de entrada sin ser necesario imputar los valores nulos.
- **Identificación de duplicados:** El porcentaje total de duplicados en el dataset es de 5,36% dado que es un porcentaje bajo, se procede a eliminarlos.

El notebook utilizado en esta sección se encuentra en el repositorio Ciencia-de-Datos-Aplicada-Proyecto-Final/data/analytics bajo el nombre de preparacion_de_datos.ipynb

Estrategia y validación del modelo

La estrategia de preparación de los datos contempla dividir el conjunto en entrenamiento, validación y prueba, implementar preprocesamiento adecuado, manejar el desbalance de clases y, finalmente, verificar que los conjuntos resultantes mantengan la consistencia y distribución original del problema. Para los modelos candidatos, Random Forest y Regresión Logística, se generaron conjuntos de datos con transformaciones específicas para cada uno. En ambos casos, se aplicó la técnica de sobre muestreo SMOTE para equilibrar las clases y mitigar el sesgo causado por el desbalanceo en los datos. La regresión logística es un modelo probabilístico basado en la función sigmoide y es sensible a la escala de los datos, ya que las variables con magnitudes mayores pueden influir desproporcionadamente en los coeficientes. Para evitar esto, se aplicó la técnica de estandarización a los datos antes de entrenar el modelo. Siguiendo la técnica de data splitting se construyeron los siguientes conjuntos:

1. **Conjunto de entrenamiento**: 60% de los datos y se utiliza para entrenar el modelo
2. **Conjunto de validación**: Representa el 20% de los datos y se emplea para ajustar los hiper parámetros del modelo y evaluar su desempeño intermedio.
3. **Conjunto de prueba**: También representa el 20% de los datos y se utiliza exclusivamente para la evaluación final del modelo.

La división busca garantizar que el modelo generalice bien a nuevos datos, minimizando el riesgo de sobreajuste (overfitting). Una vez transformados los datos se realizaron verificaciones para garantizar que las distribuciones de las características en los conjuntos de entrenamiento, validación y prueba se mantuvieran representativas del conjunto original. Esto incluyó análisis gráficos y estadísticos, como histogramas y estadísticas descriptivas, para asegurar que no hubiera sesgos significativos en la segmentación de los datos. Se verificó que las transformaciones no afectaron la distribución ni las relaciones entre variables; las gráficas se presentaran en el Anexo 3: Estrategia y validación del modelo.

El notebook utilizado en esta sección se encuentra en el repositorio Ciencia-de-Datos-Aplicada-Proyecto-Final/data/analytics bajo el nombre de seleccion_de_modelo.ipynb

Construcción y Evaluación del modelo

Con base a los datos proporcionados, se han evaluado 3 modelos obteniendo:

Modelo	Target	Precisión	Recall	F1-Score	Accuracy
Regresión Logística	0	0.28	0.77	0.42	0.70
	1	0.95	0.69	0.80	
Árbol de decisión	0	0.77	0.77	0.77	0.94
	1	0.96	0.96	0.96	
Random Forest	0	0.43	0.50	0.46	0.84
	1	0.92	0.90	0.91	

El target es la variable objetivo para la deserción, donde:

- 0: Corresponde a deserción
- 1: Corresponde a no deserción

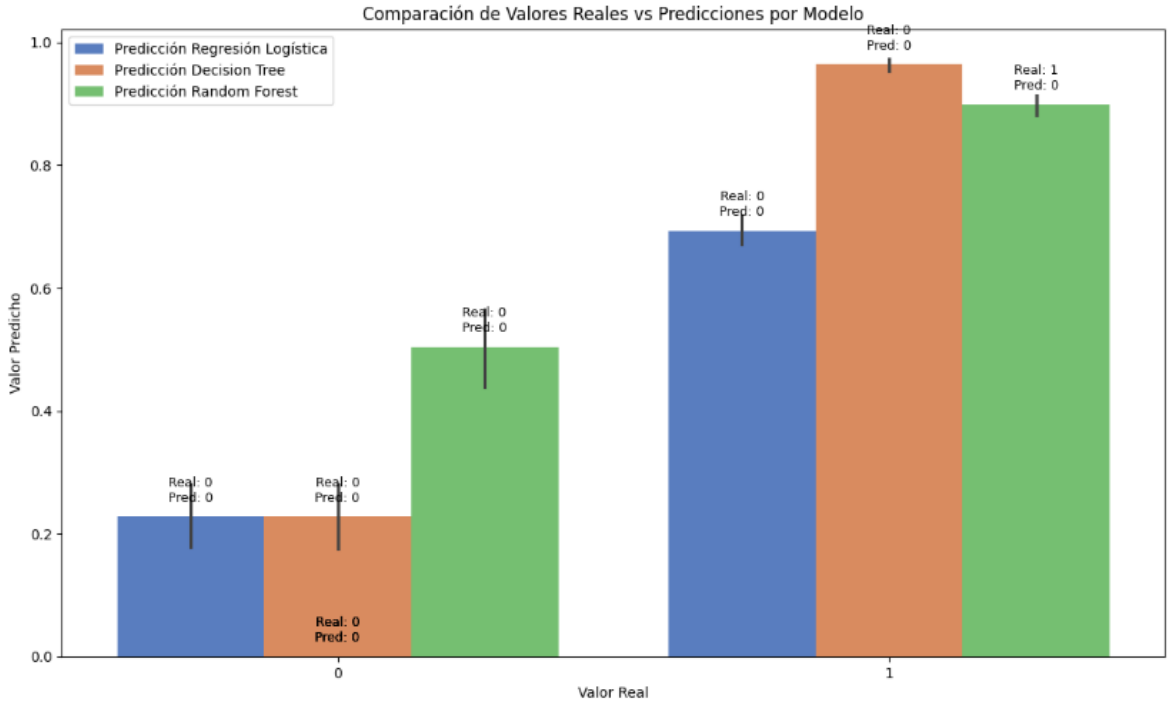
Con base en los valores obtenidos:

- Se evidencia que el árbol de decisión es el modelo más equilibrado para el target 0, con una alta precisión, recall y f1-score del 0.77; así mismo, se evidencia una exactitud del 0.94.
- Con el valor recall de los modelos, se evidencia que el árbol de decisión tiene un valor alto, lo que corresponde a que identifica mejor a los estudiantes que desertan.

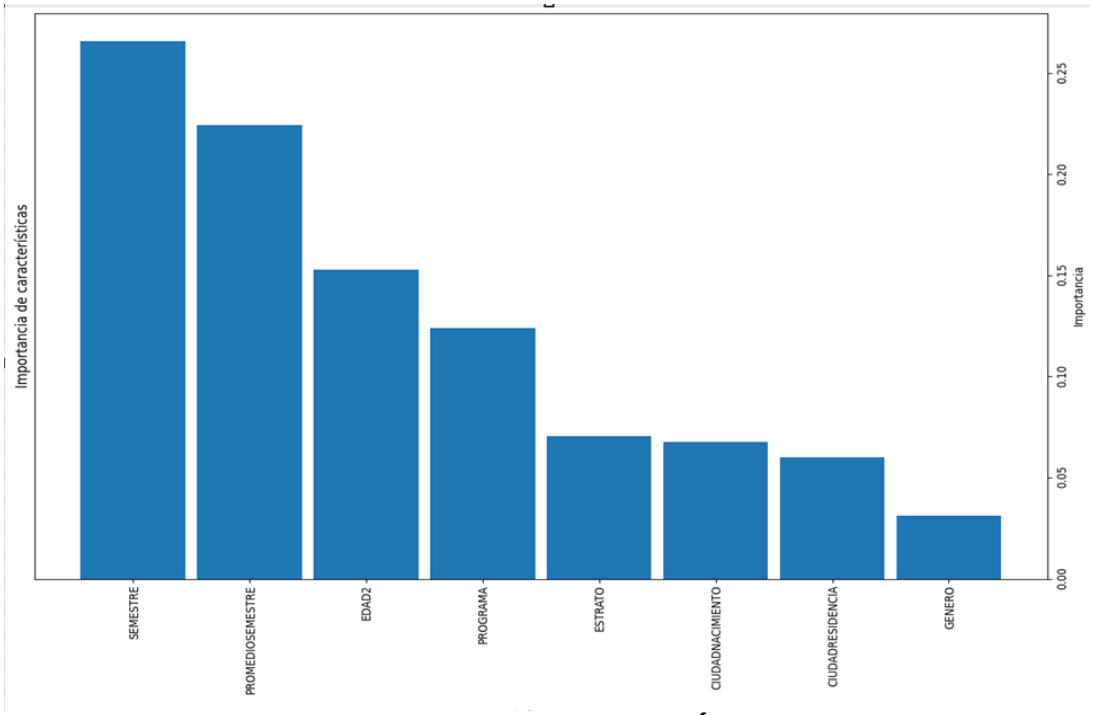
Al realizar una tabla con los valores predichos de los modelos, comparados con el valor real, se evidencia:

Valor Real	Predicción Regresión Logística	Predicción Decision Tree	Predicción Random Forest
1	1	1	1
1	1	1	1
1	1	1	1
0	1	0	1
0	0	0	0
1	1	1	1
1	0	1	0
1	1	1	1
1	1	1	1
1	1	1	1

Se sigue soportando que los valores de árbol de decisión siguen siendo acertados en su predicción.



Para ver el orden de los notebook y esquema de fuentes usadas para generación del modelo, se puede consultar el Anexo 4: Esquema del modelo



Gráfica: Feature Importance Modelo Árbol de decisión.

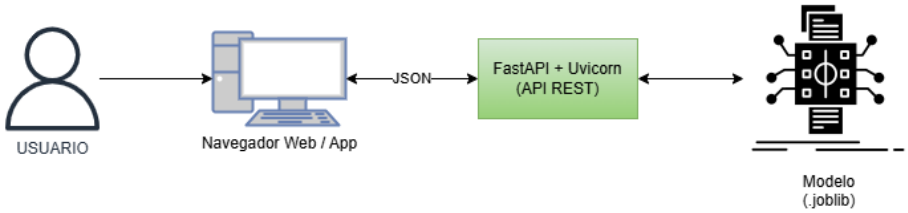
Según el grafico de feature importance, el árbol de decisión otorga mayor importancia a las variables de semestre, promedio de semestre, edad y programa, de acuerdo con la retroalimentación recibida por la entidad, esto puede estar relacionado con la dificultad que presentan los estudiantes al cursar materias de ciencias básicas durante los primeros 4 semestres en programas de formación como: Deporte, Tecnología en Deporte, Fisioterapia y Administración de Empresas, lo cual es coherente con el análisis univariado de estas mismas variables realizado en el análisis exploratorio de la primera entrega del proyecto.

Se observa que la ciudad de residencia tiene el séptimo lugar de importancia dentro de las 8 variables, lo cual contrasta, con la conclusión inicial referente a que la ciudad de residencia y el tiempo de desplazamiento hasta la entidad tienen alta influencia sobre la deserción. Lo anterior sugiere que la ciudad de residencia es importante en la deserción, sin embargo, las dificultades académicas que pueden presentar los estudiantes durante los primeros semestres influyen aún más en la deserción que la ciudad de residencia.

El notebook utilizado en esta sección se encuentra en el repositorio [Ciencia-de-Datos-Aplicada-Proyecto-Final/data/analytics] bajo el nombre de modelo.ipynb

Construcción del producto de datos.

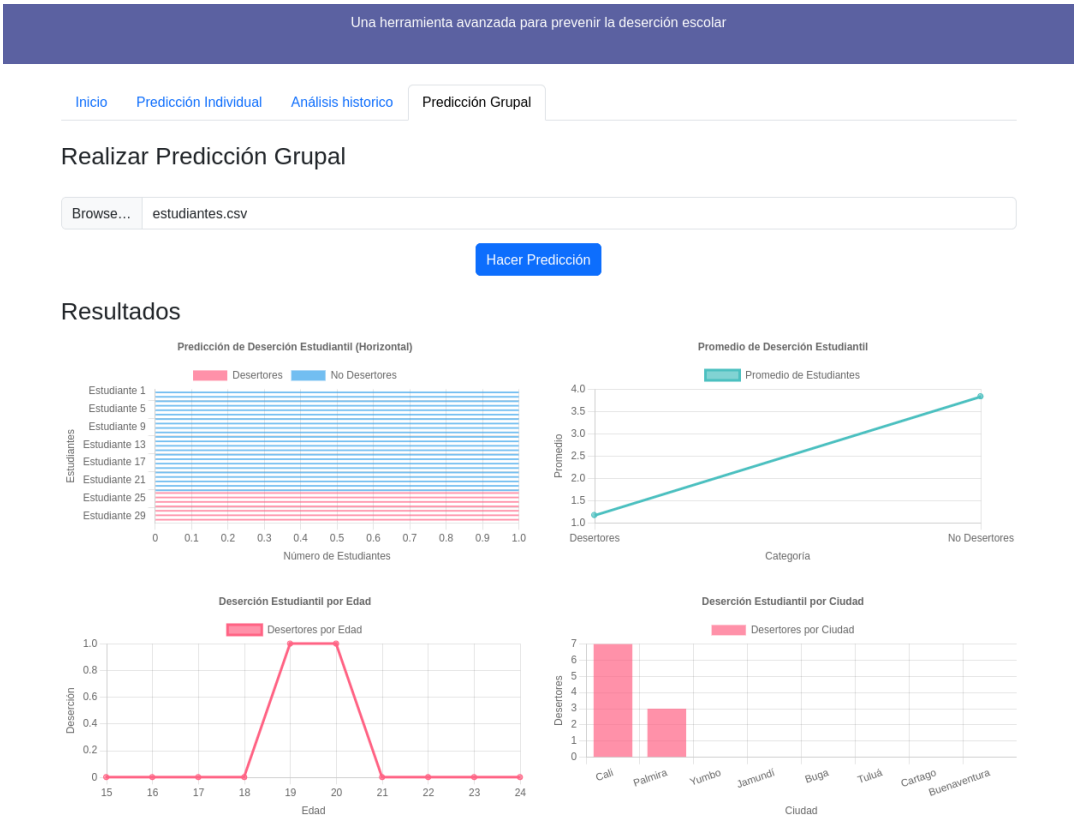
El siguiente diagrama ilustra la arquitectura del producto de datos desarrollado para el proyecto de analítica. La solución utiliza una API REST implementada con FastAPI y desplegada mediante Uvicorn.



Para ver los pasos de despliegue del modelo, consultar el Anexo 5: Despliegue.

Una vez la aplicación ha sido desplegada, se proporciona a la universidad una herramienta web intuitiva para cargar los datos de los estudiantes actualmente inscritos en cursos de la institución. Por medio del modelo construido, la herramienta permite predecir cuáles de estos estudiantes podrían estar en riesgo de desertar.

En la vista presentada, se observa un ejemplo de predicción grupal. Los resultados muestran que los estudiantes en riesgo de deserción tienen, en su mayoría, un promedio de calificaciones de 1.4, se encuentran en el rango de edad entre 19 y 20 años, y residen principalmente en Cali. Estas características permiten a la institución identificar patrones y tomar acciones preventivas para reducir la deserción estudiantil.



Adicionalmente, la institución puede explorar los factores individuales que contribuyen al riesgo de deserción de cada estudiante. Para ello, se diseñó una vista específica, como la mostrada en esta imagen, que detalla las características del estudiante en cuestión y resalta las posibles causas de su riesgo de deserción.
Por ejemplo, este estudiante (Estudiante 14) presenta un estrato socioeconómico bajo (1), un promedio semestral de 2.1, cursa el tercer semestre, tiene 20 años, pertenece al programa de Deporte, y reside en Cali. Con esta información, la institución puede tomar acciones personalizadas, como ofrecer apoyo académico o financiero.
Esta vista es fácilmente accesible haciendo clic en cualquiera de los estudiantes mostrados en la gráfica superior izquierda de la vista de predicción grupal anterior, lo que permite un análisis más detallado y estratégico para cada caso.

De acuerdo con las recomendaciones y necesidades de la Escuela Nacional del Deporte, esta herramienta está diseñada para que, antes de finalizar cada semestre, la institución pueda subir los datos de los estudiantes inscritos y, mediante el modelo de clasificación desarrollado, identificar aquellos que están en riesgo de desertar. Al utilizar la herramienta, los miembros de la institución podrán caracterizar a estos estudiantes, analizar los factores comunes que los colocan en esta situación y también acceder a detalles individuales sobre las posibles causas de deserción. Esto permite a la entidad tomar decisiones estratégicas y personalizadas para intervenir oportunamente, asegurando que estos estudiantes cuenten con el apoyo necesario y aumentando la probabilidad de que se matriculen el siguiente semestre, reduciendo así las tasas de deserción.

Retroalimentación por parte de la organización

A lo largo del desarrollo del proyecto de analítica, se llevaron a cabo múltiples interacciones con los stakeholders de la Escuela Nacional del Deporte, que fueron fundamentales para definir la problemática a abordar, acordar los detalles del producto de datos y validar los resultados obtenidos.

A continuación, se resumen las principales reuniones y los acuerdos alcanzados en cada una:

Reunión 1: Clarificación de objetivos y entrega del conjunto de datos

En esta primera sesión, el objetivo principal fue definir de manera clara la problemática que el proyecto buscaría abordar: la deserción estudiantil. Se discutió la importancia de contar con un sistema que permita identificar de manera anticipada a los estudiantes en riesgo y entender los factores que contribuyen a la deserción. La organización proporcionó el conjunto inicial de datos, que incluía información demográfica y académica de los estudiantes, sentando las bases para las etapas iniciales de entendimiento y preparación de los datos. Además, se acordaron las métricas clave que se utilizarían para evaluar el éxito del proyecto.

Reunión 2: Presentación de avances y acuerdo sobre el producto de datos

En esta sesión intermedia, se presentaron los resultados preliminares del análisis exploratorio de datos y el entendimiento inicial de los factores asociados a la deserción. También se mostraron las primeras versiones de los modelos predictivos desarrollados, lo que permitió una discusión sobre su desempeño inicial y las oportunidades de mejora. Durante esta reunión, se acordó con los stakeholders que el producto de datos incluiría una herramienta web con vistas para predicción individual y grupal, diseñada específicamente para satisfacer las necesidades de la institución.

Reunión 3: Presentación de resultados finales y retroalimentación

En la reunión final, se presentó el producto terminado, incluyendo la herramienta web y los resultados del modelo de clasificación. La entidad confirmó que el producto construido cumplía con sus expectativas, destacando su utilidad práctica para identificar y caracterizar a los estudiantes en riesgo de desertar. Además, se discutieron posibles mejoras futuras, enfocadas en integrar nuevas fuentes de datos, como encuestas de satisfacción estudiantil y datos sobre el desempeño individual en cada curso. Esta retroalimentación permitió comprobar el éxito del proyecto y reflejó un alto grado de satisfacción con los resultados obtenidos.

Conclusiones

- Se construyó un prototipo funcional que, mediante el uso de técnicas de machine learning, permitió evaluar el rendimiento de varios modelos en la predicción de deserción estudiantil. Los resultados obtenidos han proporcionado información valiosa sobre la capacidad de cada modelo para predecir correctamente los casos de deserción (0) y no deserción (1), y se identificó el modelo más efectivo para este propósito.

- Para obtener resultados aún más precisos y efectivos, es necesario la incorporación de más datos, especialmente aquellos relacionados con el contexto socioeconómico de los estudiantes, actividades extracurriculares, historial académico detallado y otros factores, podría enriquecer los modelos y mejorar su capacidad predictiva.

- Se debe mejorar la calidad en los datos con el fin de asegurar que los datos estén completos, sin valores nulos o inconsistencias, es crucial para obtener predicciones más precisas.

- El Árbol de Decisión (Decision Tree) resultó ser el modelo más efectivo con una accuracy de 0.94, lo que indica que tiene un buen rendimiento general en cuanto a las métricas de precisión, recall y F1 score todas puntuaron 0.77, sin embargo, se requieren algunos ajustes adicionales como, por ejemplo, la inclusión de más información para optimizar su capacidad en la predicción de deserción, lo que podría mejorar aún más su desempeño. No obstante, de acuerdo con la retroalimentación recibida por la Escuela Nacional del Deporte, la solución alcanzada cumple con todos los requisitos esperados para solucionar su problemática y darles la información suficiente para plantear acciones apropiadas para cada alumno en riesgo de desertar.

- De acuerdo con la métrica de Recall se estima que el modelo identifique el 77% de todos los estudiantes que pueden desertar, quienes serán intervenidos para que finalicen sus estudios normalmente, con esta acción se espera que el número de estudiantes desertores tienda a disminuir y por ende también disminuya el KPI de Tasa de Deserción de la entidad (TDA).

- La métrica de precisión del 77% indica que el porcentaje de verdaderos positivos detectados por el modelo, es aproximadamente 3 veces mayor que el porcentaje de falsos positivos, lo que asegura que las intervenciones se realicen sobre los estudiantes que verdaderamente están en riesgo de deserción lo que apalanca también la optimización de los recursos de la entidad. Lo anterior se reflejaría en un mejor desempeño del KPI de Tasa de efectividad de las intervenciones (TEI).

- De acuerdo con la gráfica de feature importance del modelo, se observa que los aspectos que tienen mayor relación con la deserción en su orden de importancia son:

    1. Aspectos académicos relacionados con la posible dificultad que tienen los estudiantes durante los primeros semestres al cursar materias de ciencias básicas. Lo anterior se evidencia en la importancia que otorga el modelo a las variables de semestre, promedio de semestre y programa.

2. Aspectos socioeconómicos de los estudiantes relacionados posiblemente con el nivel de ingresos familiares, apoyo económico con que cuenta, lo cual es evidenciado en la importancia que otorga el modelo a la variable estrato.
3. Aspectos geográficos como la ciudad de origen lo cual sugiere que los estudiantes foráneos tienen un mayor riesgo de deserción posiblemente relacionado con la falta del apoyo familiar en la ciudad y/o los costos adicionales que debe cubrir para su mantenimiento.
4. La ciudad de residencia es otro aspecto geográfico de importancia para el modelo y posiblemente relacionado con el tiempo de desplazamiento de los estudiantes que residen en ciudades aledañas a Cali.

Bibliografía

Biblioguías. (Octubre 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

Congreso de la República de Colombia. (17 de Octubre de 2012). *Ley 1581 de 2012 - Gestor Normativo*. Obtenido de Función Pública: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=49981

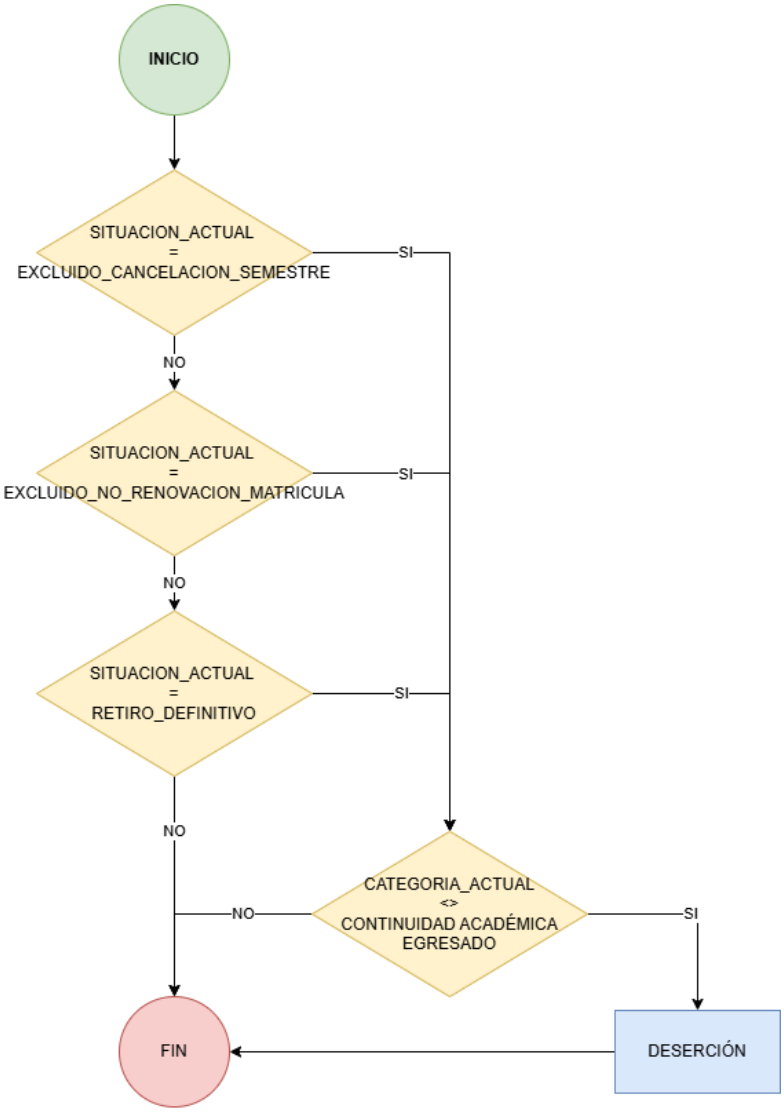
Departamento Nacional de Planeación. (08 de Noviembre de 2019). *Documento Conpes 3975*. Obtenido de https://colaboracion.dnp.gov.co/CDT/Conpes/Econ%C3%B3micos/3975.pdf

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

ANEXOS

Anexo 1: Regla

ENDEPORTE ha mencionado la siguiente regla de acuerdo con la información proporcionada para clasificar una deserción:



Anexo 2: Responsabilidad

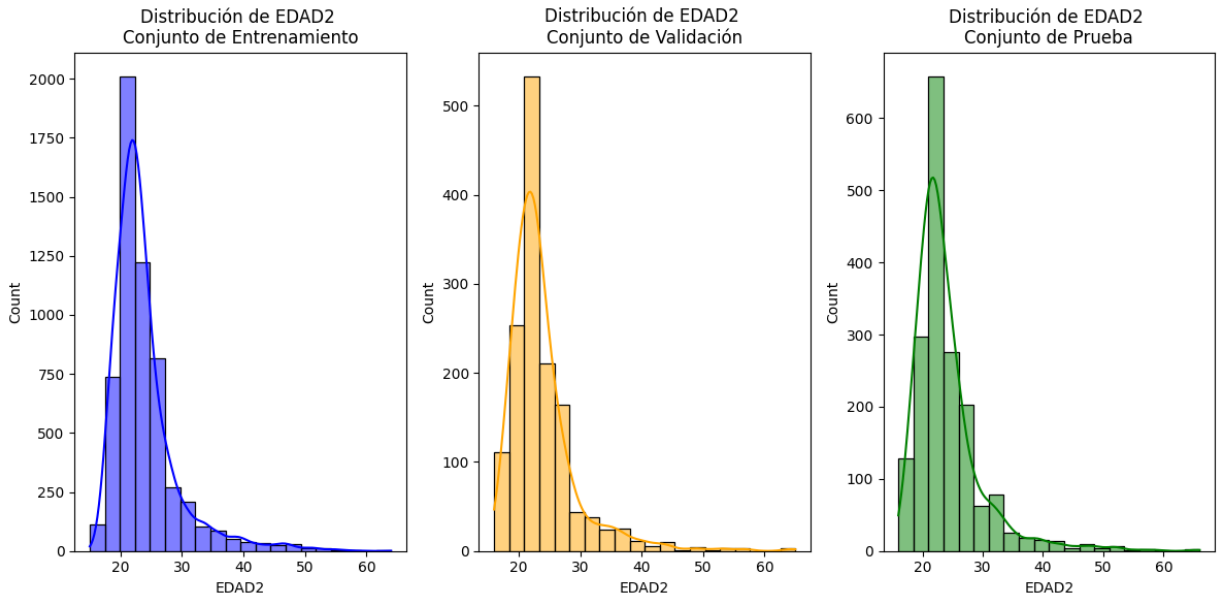
Principios	Implicaciones en el estudio – Ley 1582 de 2012
Principio de legalidad	Dado el uso de datos personales, el estudio debe ser una actividad regulada según la ley 1581 de 2012.
Principio de finalidad	Garantizar que la finalidad del estudio corresponda a una necesidad dentro del marco legal colombiano.
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.

Acciones	Implicaciones en el estudio – Lineamientos IA – CONPES 3975
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.

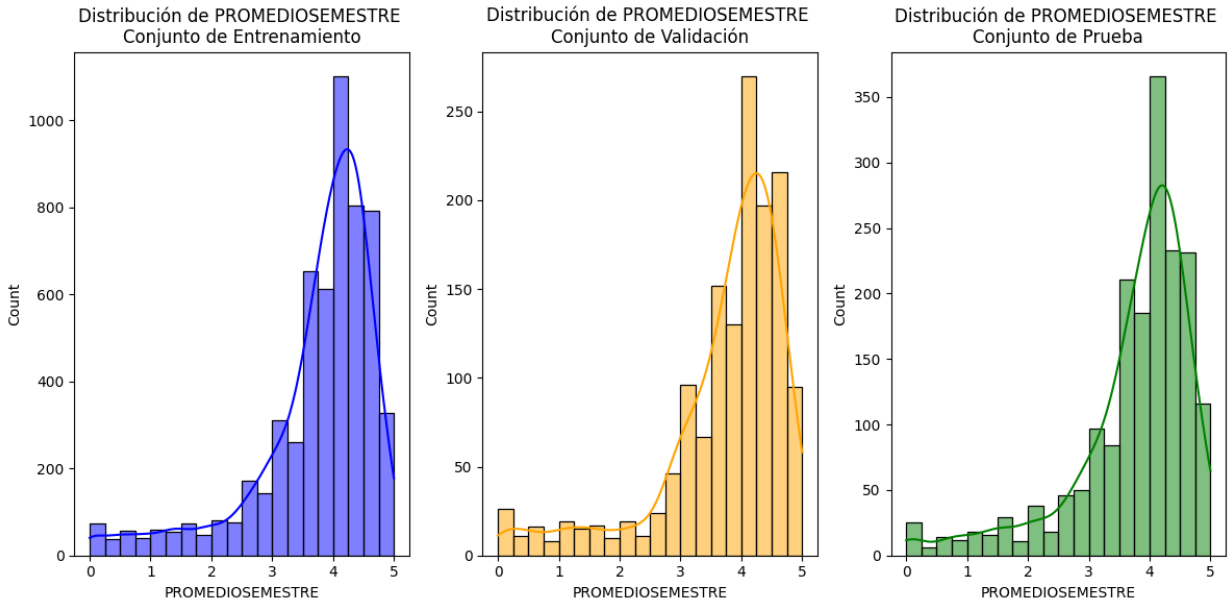
Seguridad y Protección	Esta acción debe ser evaluada por la entidad ya que implica desarrollar un proceso de gestión de riesgo que garantice un uso apropiado y seguro del algoritmo.
Asegurar un abordaje inclusivo y centrado en el usuario.	La inclusión se garantiza con la multidisciplinariedad que existe 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.

Anexo 3: Estrategia y validación del modelo

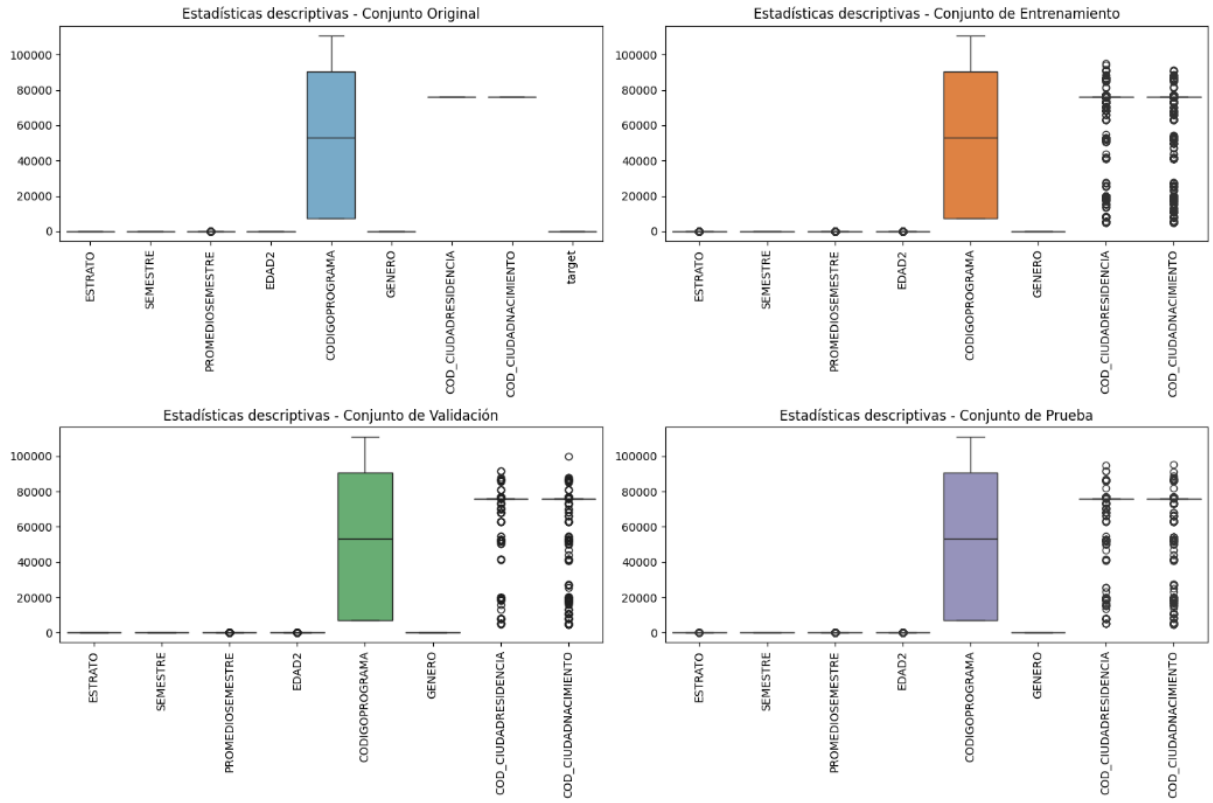
- Validación Distribuciones de las variables



Grafica 1: Distribución train/test/validation | Edad.

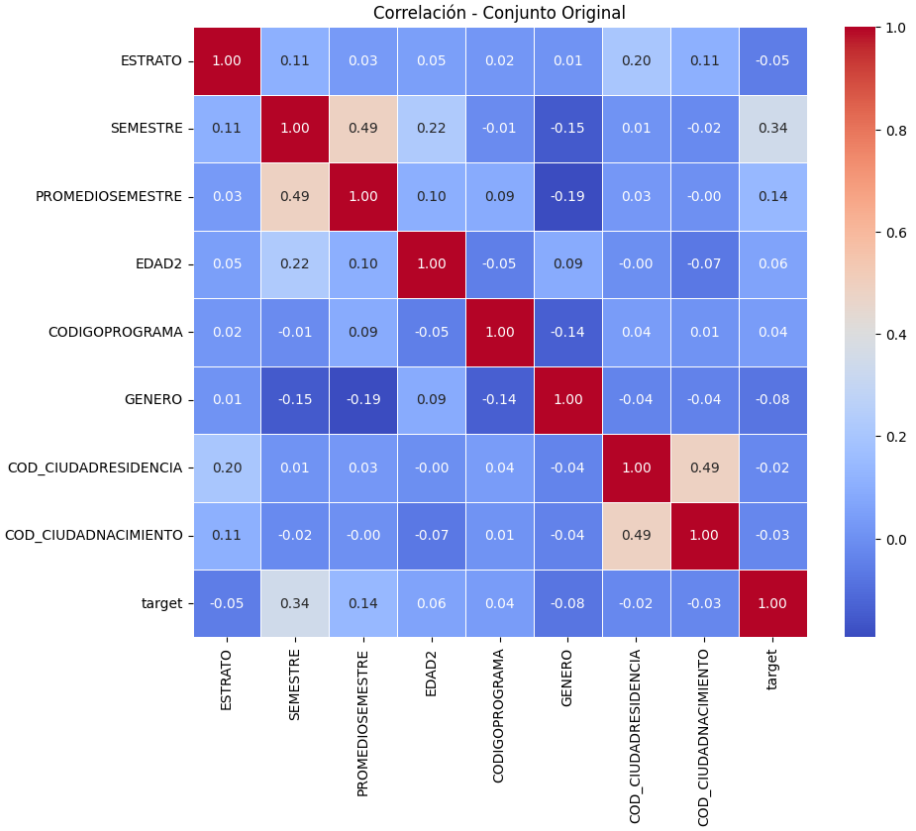


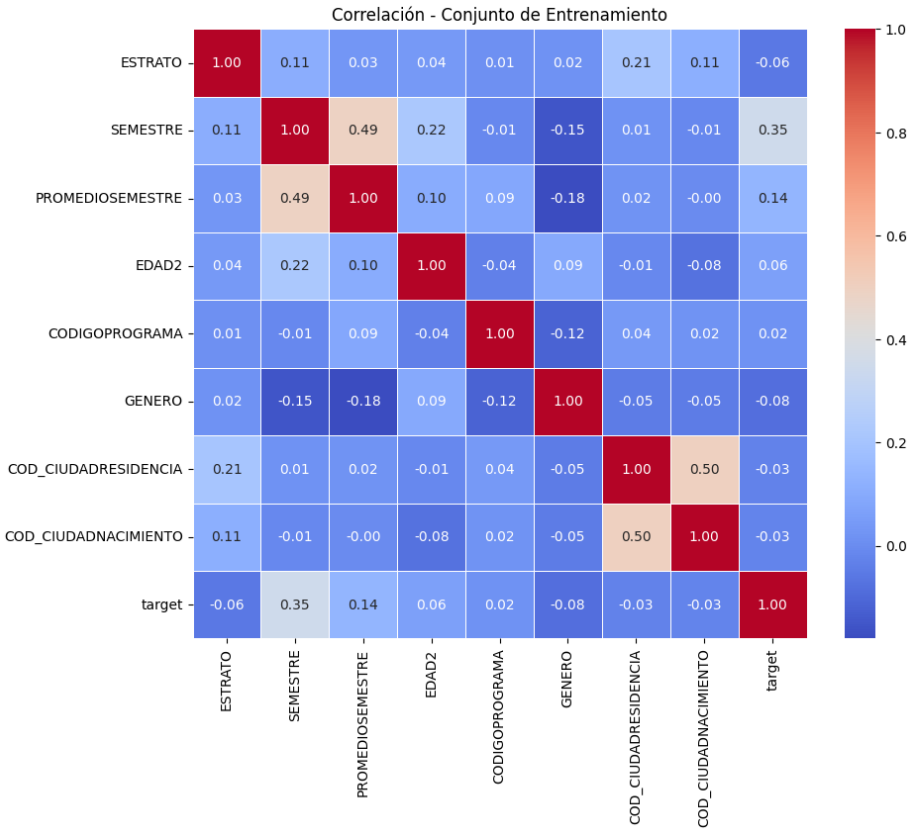
Grafica 2: Distribución train/test/validation variable PromedioSemestre.



Grafica 3: Distribución train/test/validation variables categóricas.

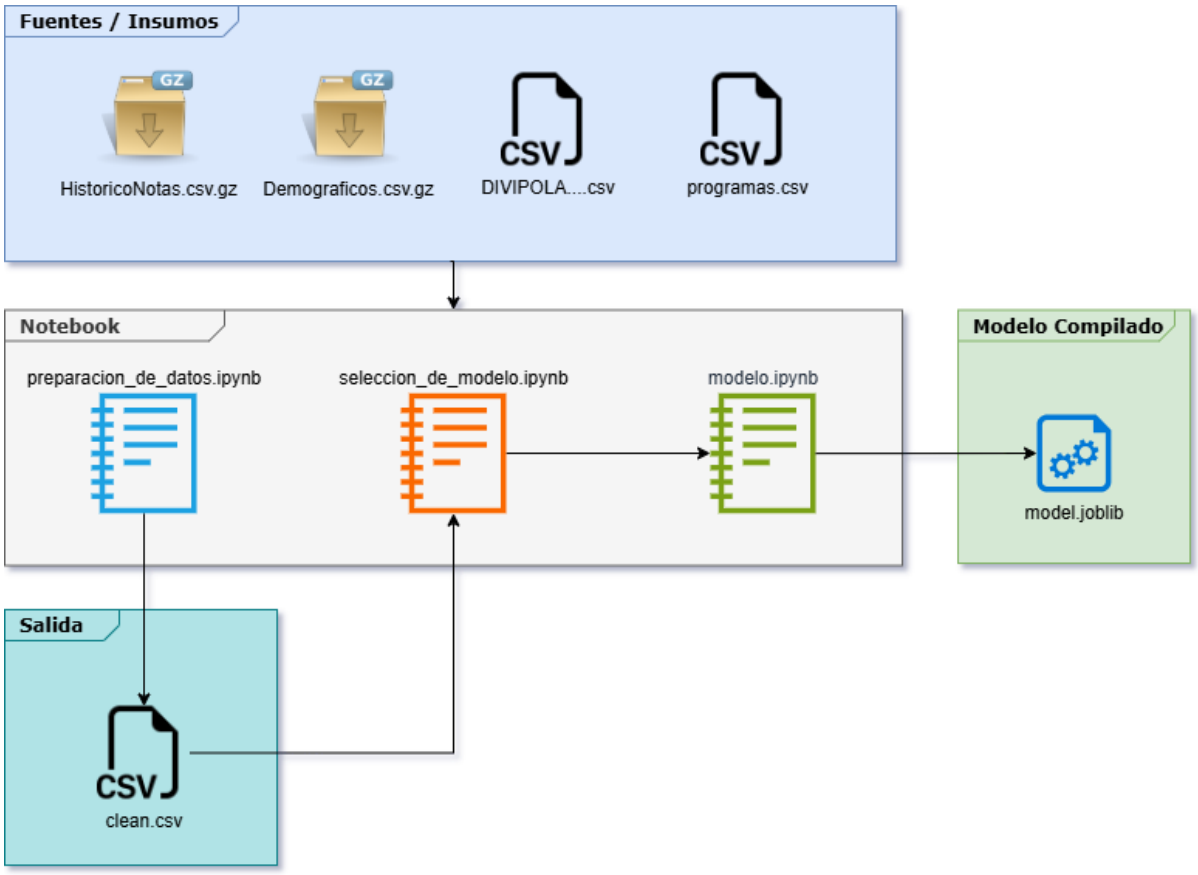
- Validación relaciones





Grafica 4: Relaciones train/test/validation y conjunto original.

Anexo 4: Esquema del modelo



1. Recepción de fuentes para el modelo, datos proporcionados por ENDEPORTE (Histórico Notas, Demográficos) e información complementaria de fuentes externas (DIVIPOLA).
2. Lectura y cargue de fuentes suministradas, procesamiento de datos, limpieza y generación de archivo de salida.
3. Visualización, ajuste y balanceo de datos.
4. Entrenamiento de modelos, generación de archivo compilado para su despliegue posterior.

Anexo 5: Despliegue

De acuerdo con las fuentes presentes en el repositorio Ciencia-de-Datos-Aplicada-Proyecto-Final , para el despliegue se debe seguir los siguientes pasos:

1. Instalar un ambiente virtual
   `pip install virtualenv`
2. Crear un entorno virtual:
   `python -m venv env`
3. Activar el entorno:
   a. En Windows:
      `.\env\Scripts\Activate.ps1`
   b. En macOS/Linux:
      `source env/bin/activate`
4. Instalar las dependencias
   `pip install -r requirements.txt`
5. Posteriormente, ingresar a la carpeta deploy y activar el servidor:
   `uvicorn main:app --reload`
6. Una vez el servidor presente el mensaje de inicio correcto, similar al siguiente:

```
INFO:     Uvicorn running on http://127.0.0.1:8000 (Press CTRL+C to quit)
INFO:     Started reloader process [15952] using StatReload
INFO:     Started server process [10220]
INFO:     Waiting for application startup.
INFO:     Application startup complete.
```

7. Luego de evidenciar el servidor en ejecución, se debe ingresar a un navegador e ingresar a la siguiente URL: http://127.0.0.1:8000/inicio , donde se carga una interfaz para diligenciar la información.
   Si se requiere consumir el modelo por un servicio REST, se puede probar el servicio con una herramienta como Postman, enviando:

```
POST /predict HTTP/1.1
Host: 127.0.0.1:8000
Content-Type: application/json
[
   {
      "ESTRATO": 1,
      "SEMESTRE": 1,
      "PROMEDIOSEMESTRE": 3.8,
      "EDAD2": 18,
      "PROGRAMA": 53212,
      "GENERO": 1,
      "CIUDADRESIDENCIA": 76001,
      "CIUDADNACIMIENTO": 76001
   }
]
```

Los valores de:

- PROGRAMA corresponden a los códigos presentes en el archivo ubicado en data/raw/programas.csv
- CIUDADRESIDENCIA el código corresponde al Codigo_Municipio presente en el archivo de DIVIPOLA ubicado en data/raw/DIVIPOLA-_C_digos_municipios_20241127.csv
- CIUDADNACIMIENTO el código corresponde al Codigo_Municipio presente en el archivo de DIVIPOLA ubicado en data/raw/DIVIPOLA-_C_digos_municipios_20241127.csv