NoCountry

Proyecto sobre Rendimiento Académico

OBJETO DEL PROYECTO:

ASESORAMIENTO INTEGRAL EN PLAN DE ESTUDIO Y ANÁLISIS SOBRE EL DESEMPEÑO ACADÉMICO DE ESTUDIANTES

Abril del 2024

1. OBJETIVOS

* 1. **Objetivo General**
* Analizar el rendimiento académico de estudiantes en una institución o entorno educativo utilizando datos de calificaciones, asistencia y participación en actividades extracurriculares.

* 1. **Objetivos Específicos**
* Realizar análisis de datos referidos a desempeño académico histórico de alumnos en entornos educativos y/o académicos
* Predecir el desempeño de los alumnos e identificar las principales metricas que lo reflejan
* Dar conclusiones concretas y de valor para la toma de decisiones en base a las politicas educativas del usuario/institucion que las requiera.

1. FUNDAMENTACIÓN

Respecto a la decisión de los datos:

El Instituto Colombiano para la Evaluación de la Educación (ICFES) es una entidad especializada en ofrecer servicios de evaluación de la educación en todos sus niveles, y en particular apoya al Ministerio de Educación Nacional en la realización de los exámenes de Estado y en adelantar investigaciones sobre los factores que inciden en la calidad educativa, para ofrecer información pertinente y oportuna para contribuir al mejoramiento de la calidad de la educación.

Respecto a la elección del stack tecnológico:

El entorno de Google Colaboratory nos brinda los recursos necesarios para el análisis y el acceso a el trabajo colaborativo en conjunto con la facilidad de uso, la conexión nativa con Google Drive, y su compatibilidad con el lenguaje de programación Python (Adecuado para su uso en el análisis de datos)

Respecto al proyecto realizado:

Vertical Tecnológica: Data - BI

* Descripción

La vertical de Data-BI se enfoca en la recopilación, análisis y visualización de datos para tomar decisiones informadas y estratégicas en una organización. Esta vertical abarca desde la recopilación y limpieza de datos hasta el análisis avanzado y la presentación de insights a través de herramientas de Business Intelligence (BI) y Data Visualization.

* Objetivo de Proyecto

El objetivo del proyecto de Data-BI es utilizar datos para generar insights y conocimientos que impulsen la toma de decisiones informadas en una organización. Esto puede incluir la creación de informes ejecutivos, paneles interactivos, modelos predictivos y mucho más.

1. METODOLOGÍA DE TRABAJO

Sprint

El concepto de Sprint es el corazón de las metodologías ágiles que sirven para avanzar de forma ordenada y gradual hacia la consecución de los objetivos, mostrando el progreso de su proyecto en todos sus aspectos: funcional, estético, responsive, etc. La finalidad detrás es que el “cliente” pueda ver que el proyecto avanza y que el equipo no deje todo para el último día de la entrega final.

Cada sprint tendrá una demo con sus TLs y serán 4.

● Semana 1: Jueves.

● Semana 2: Jueves.

● Semana 3: Jueves.

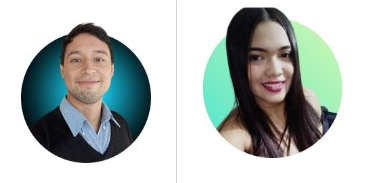
● Semana 4: Último previo al Demo Day

Los principios de Lean Startup promueven la experimentación rápida y el aprendizaje validado. Esto implica estar preparados para enfrentar desafíos y conflictos, resaltando la importancia de la flexibilidad y la adaptabilidad como soft skills fundamentales para el éxito en proyectos tecnológicos.

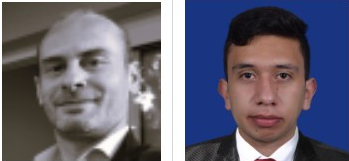
No Country ejercerá el rol de “cliente”, esto tiene como intención simular el proceso de entorno laboral real donde el equipo debe realizar una demostración al cliente. En este contexto cada Team Leader y los diferentes canales de Slack de la comunidad representan los intereses del cliente.

1. EQUIPO DE TRABAJO

Nuestro equipo de trabajo se compone de los siguientes integrantes, cada uno abocado a una etapa y cumpliendo un rol esencial en el desarrollo del proyecto:



| Marco Caro  Data Analyst & Engineer | Kathiuska Mangones Ramos  Data Analyst & Engineer |
| --- | --- |



| Pablo Dudiuk  Data Scientist | Sebastian Mora  Analyst BI & Data Engineer |
| --- | --- |

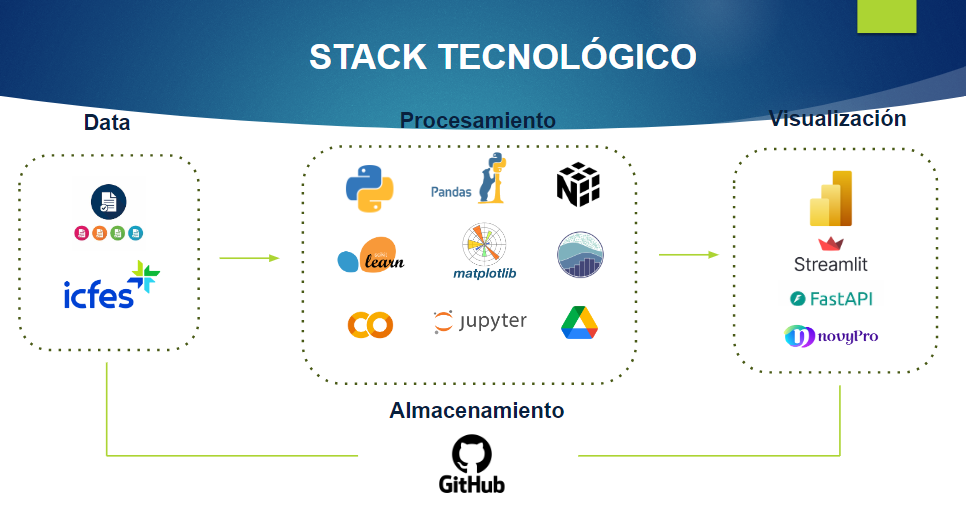
* **Marco Caro. Data Analyst & Engineer**
* **Kathiuska Mangones Ramos. Data Analyst & Engineer:**

* **Pablo Dudiuk. Data Scientist**
* **Sebastian Mora. Analyst BI & Data Engineer**

1. STACK TECNOLÓGICO

El stack tecnológico empleado está, en su mayoría, dentro del universo de paquetes y librerías de Python, el cual ofrece toda la gama de herramientas para desarrollar el ciclo completo de nuestro proyecto. Facilita la ingesta y transformación de los datos, de forma que se podrá trabajar desde el repositorio o almacén de datos que aloja los datos sin procesar, pasando por los entornos para escribir y ejecutar código hasta la generación de la data ya transformada e implementación de machine learning e inteligencia de negocios.

En la imágen se verán las herramientas, programas, lenguajes de programación y librerías contemplados para llevar a cabo cada una de las etapas de la ejecución del proyecto.



Citaremos algunas de las tecnologías, para entrar en el contexto de su funcionamiento.

* **Python:** Es un lenguaje de programación de alto nivel, interpretado y multipropósito. Python será utilizado en todas las etapas del proyecto debido a su simplicidad, versatilidad y a las poderosas bibliotecas que ofrece. Dentro de ella trabajaremos con NumPy, pandas, Matplotlib, Seaborn, Scikit-Learn, Tensor Flow.
* **Numpy:** NumPy es una biblioteca para el lenguaje de programación Python que da soporte para crear vectores y matrices grandes multidimensionales, junto con una gran colección de funciones matemáticas de alto nivel para operar con ellas
* **Pandas:** Pandas es una librería de Python especializada en la manipulación y el análisis de datos. Ofrece estructuras de datos y operaciones para manipular tablas numéricas y series temporales, es como el Excel de Python
* **Matplotlib:** Matplotlib es una [biblioteca](https://es.wikipedia.org/wiki/Biblioteca_(programaci%C3%B3n)) para la generación de gráficos en dos dimensiones, a partir de datos contenidos en listas o arrays en el lenguaje de programación [Python](https://es.wikipedia.org/wiki/Python)
* **Seaborn:** Seaborn es una biblioteca para crear gráficos estadísticos en Python . Está basado en Matplotlib y se integra con las estructuras de Pandas. Es tan poderosa como Matplotlib pero aporta simplicidad y nuevas funciones. Nos permite explorar y comprender datos rápidamente. Se pueden capturar marcos de datos completos y las funciones internas para mapeo semántico y agregación estadística permiten convertir los datos en visualizaciones gráficas.
* **Google Colaboratory:** Colab es un servicio alojado de Jupyter Notebook que no requiere configuración para su uso y proporciona acceso gratuito a recursos informáticos, incluidas GPU y TPU. Colab es especialmente adecuado para el aprendizaje automático, la ciencia de datos y la educación.
* **Jupyter Notebook:** Es un entorno informático [interactivo basado en la web](https://es.wikipedia.org/wiki/Rich_Internet_application) para crear documentos de Jupyter notebook. El término "notebook" puede hacer referencia coloquialmente a muchas entidades diferentes, principalmente la aplicación web Jupyter, el servidor web Jupyter Python o el formato de documento Jupyter según el contexto. Un documento de Jupyter Notebook es un documento [JSON](https://es.wikipedia.org/wiki/JSON), que sigue un esquema versionado y que contiene una lista ordenada de celdas de entrada/salida que pueden contener código, texto (usando [Markdown](https://es.wikipedia.org/wiki/Markdown)), matemáticas, gráficos y texto enriquecidos, generalmente terminado con la extensión ".ipynb".
* **Google Drive:** Es un servicio de alojamiento y sincronización de archivos desarrollado por Google. Lanzado el 24 de abril del 2012, el servicio permite a sus usuarios almacenar archivos en la nube, sincronizar archivos entre dispositivos y compartir archivos.

* **Power BI:** Es una plataforma de análisis que permite visualizar y compartir datos de manera efectiva para tomar decisiones informadas. Ofrece una amplia gama de herramientas para la preparación de datos, visualización de datos, análisis y colaboración en un solo lugar.
* **Streamlit**: Es una biblioteca de python que permite crear aplicaciones web interactivas para el análisis de datos y la visualización de manera rápida y sencilla. Ofrece una amplia gama de widgets para la entrada de datos, gráficos interactivos y capacidades de visualización, lo que la convierte en una herramienta poderosa para la creación rápida de prototipos y la implementación de aplicaciones de análisis y ciencia de datos.
* **Github:** GitHub es una forja (o repositorio) para alojar proyectos utilizando el sistema de control de versiones Git. Se utiliza principalmente para la creación de código fuente de programas de ordenador.
* **NovyPro:** NoviPro es una plataforma donde los desarrolladores de Business Intelligence pueden mostrar sus paneles de Live BI reales y compartirlos con su red.

1. Extraction, Transformation, Loading (ETL)

En la etapa de transformación de los datos se abordarán los siguientes aspectos básicos para realizar el tratamiento previo y limpieza de datos para poder llevar a cabo la carga de datos al data warehouse

- Verificación De tipo de dato de columna

- Dimensionalidad de los datos

- Valores nulos

- Verificación gráfica de outliers

- Indagación de consistencia de los datos, máximos, mínimos.

* 1. Informe ETL

El informe ETL es un informe de las condiciones preliminares en que encontramos la data original (una vez extraída de su fuente)

Presenta múltiples funciones creadas en base a realizar un análisis amplio y profundo de los distintos aspectos a analizar de las condiciones en que están nuestros datos.

Estas funciones son:

* **verificar\_tipo\_datos:**

* **informe\_dataframe:** Esta función obtiene un dataframe, y realiza un informe analizando y explorando algunas caracteristicas del dataframe centrandose principalmente en características a nivel general de nuestro dataframe y realizando un procesamiento de algunos datos obteniendo metricas e informacion devuelve un informe que contiene:

-Dimensiones del DataFrame

-Número de datos

-Filas y Columnas

-Tipo de columnas

-Cantidad de registros duplicados

-Métricas Generales

* **informe\_columna:** Esta funcion obtiene un dataframe y el nombre de una de sus columnas, y realiza un informe analizando y explorando algunas caracteristicas de la feature, centrandose principalmente en caracteristicas a nivel general y realizando un procesamiento de algunos datos obteniendo metricas e informacion.

Dependiendo el tipo de dato contenido en la feature/columna, devolvera informacion ligeramente diferente:

Para tipo object:

-Número de datos nulos

-Cantidad de valores unicos en la columna

-Valores unicos en la columna (Primeros 5 valores, en caso de exceder los 5, en caso contrario, devuelve todos los valores unicos)

-Moda de la columna

-Distribucion de frecuencias

Para tipo datetime64[ns]:

-Número de datos nulos

-Cantidad de valores unicos en la columna

-Valores unicos en la columna (una muestra de 4 valores como ejemplo, y el rango que abarcan (desde que valor hasta que valor))

-Moda de la columna

-Distribucion de frecuencias

-Valor maximo y minimo

Para tipo numerico (int, float):

-Número de datos nulos

-Valores unicos en la columna (una muestra de 5 valores como ejemplo, y el rango que abarcan (desde que valor hasta que valor))

-Moda de la columna

-Estadisticos Principales de la columna

-Valores extremos

-Distribucion de frecuencias

-Valor maximo y minimo

* **visualizar\_nulos:** Toma como parámetro un dataframe pandas y realiza la construcción de 4 gráficos principales en base al análisis de sus datos con valor np.nan o nulos:

Matrix Plot: Una cuadrícula donde cada fila representa una observación y cada columna representa una variable en el DataFrame.

Los valores nulos están marcados en blanco en la cuadrícula.

Bar Plot: Un gráfico de barras vertical que muestra la cantidad de valores nulos en cada columna del DataFrame.

Cada barra vertical representa una columna, y la altura de la barra indica la cantidad de valores nulos en esa columna.

Heatmap: Similar a la matriz, pero con colores para resaltar las áreas con más valores nulos.

Los valores nulos pueden estar representados con colores más oscuros para destacar las áreas con más ausencias de datos.

Dendrogram: Un diagrama de árbol que agrupa las variables y observaciones basadas en la similitud de los patrones de valores nulos.

Las ramas del árbol se fusionan según la similitud de los patrones de valores nulos entre las variables y observaciones.

**análisis\_frecuencia\_palabras:** Toma como parametro un dataframe pandas, selecciona las columnas tipo 'object', para realizar una serie de normalizacion y procesamiento de los datos y generar un grafico de barras por cada columna seleccionada que representa las palabras mas frecuentes en la columna

**boxplots numéricas:** Toma como parámetro un pandas.DataFrame y devuelve un boxplot de cada columna de tipo numérico (variables cuantitativas) que refleja la distribución de los valores de las

**hist\_duplicados**: Esta función toma como parámetro un pandas.DataFrame. Da salida a un gráfico de barras/histograma que refleja la cantidad de registros duplicados y no duplicados presentes en el dataframe

En el Caso que no se registren duplicados, dará salida a un mensaje confirmándolo.

* 1. Desarrollo ETL

Se cargan los datos y se verifica el estado general del dataset:

* 83 columnas
* 100k nulos como máximo
* 563853 registros para el año 2023
* 0 duplicados

Se procede a verificar las columnas en el dataset y ver su relevancia para el proyecto, esto se premedito en conjunto de todo el equipo, concluyendo con la eliminación de las siguientes columnas:

* ESTU\_FECHANACIMIENTO
* ESTU\_CONSECUTIVO
* ESTU\_COD\_RESIDE\_DEPTO
* ESTU\_COD\_RESIDE\_MCPIO
* FAMI\_SITUACIONECONOMICA
* ESTU\_TIPOREMUNERACION
* COLE\_CODIGO\_ICFES
* COLE\_COD\_DANE\_ESTABLECIMIENTO
* COLE\_CALENDARIO
* COLE\_COD\_DANE\_SEDE
* COLE\_COD\_MCPIO\_UBICACION
* COLE\_COD\_DEPTO\_UBICACION
* COLE\_DEPTO\_UBICACION
* ESTU\_PRIVADO\_LIBERTAD
* ESTU\_COD\_MCPIO\_PRESENTACION
* ESTU\_DEPTO\_PRESENTACION
* ESTU\_COD\_DEPTO\_PRESENTACION
* ESTU\_ESTADOINVESTIGACION
* PERCENTIL\_ESPECIAL\_GLOBAL

Se procede a verificar el peso de los valores de las columnas categóricas y ver su relevancia dentro de las columnas, se concluye lo siguiente:

* Se eliminan valores de la columna ESTU\_TIPODOCUMENTO: CE, CR, NES, PEP, PC, PE, PPT y DNI
* Se eliminan las columnas ESTU\_NACIONALIDAD, ESTU\_PAIS\_RESIDE, COLE\_GENERO, ESTU\_ESTADOINVESTIGACION y ESTU\_ESTUDIANTE debido a que estas columnas poseen un valor con un porcentaje mayor al 95% de la población o son constantes.
* ESTU\_NSE\_INDIVIDUAL y ESTU\_INSE\_INDIVIDUAL se eliminan nulos ya que eso una variable importante para poder rellenar valores faltantes en otras columnas
* Se eliminan las columnas FAMI\_TIENEINTERNET,FAMI\_TIENESERVICIOTV, FAMI\_TIENECOMPUTADOR, FAMI\_TIENELAVADORA, FAMI\_TIENEHORNOMICROOGAS, FAMI\_TIENEAUTOMOVIL, FAMI\_TIENEMOTOCICLETA y FAMI\_TIENECONSOLAVIDEOJUEGOS debido a su alta dependencia de otras columnas “ESTU\_NSE\_INDIVIDUAL”

Se codifican las siguientes columnas:

* FAMI\_ESTRATOVIVIENDA:
  + Personas con NSE en (4,3) y INSE sobre 70 serán 'ESTRATO 4'
  + Personas con NSE en (4,3) y INSE entre 55 y 70 serán 'ESTRATO 3'
  + Personas con NSE en (2,3) y INSE menor 55 serán 'ESTRATO 2'
  + Personas con NSE 1 serán 'ESTRATO 1'
* FAMI\_EDUCACIONMADRE (se aplicara de manera similar al padre):
  + NSE 1 = 'Primaria incompleta'
  + NSE 2 = 'Primaria completa'
  + NSE 3 = 'Secundaria (Bachillerato) completa'
  + NSE 4 = 'Educación profesional completa'
* FAMI\_COMELECHEDERIVADOS, FAMI\_COMECEREALFRUTOSLEGUMBRE, FAMI\_COMECARNEPESCADOHUEVO:
  + 1 = NSE 4
  + 2 = NSE 3
  + 3 = NSE 2
  + 4 = NSE 1
* COLE\_BILINGUE:
  + Seran S si COLE\_NOMBRE\_ESTABLECIMIENTO o COLE\_NOMBRE\_SEDE tienen ['BILINGÜE', 'BILINGUE', 'BILINGUAL', 'BRITÁNICO', 'BRITANICO', BRITANIC', 'AMERICAN', 'AMERICANO'] en su nombre.

Se procede a codificar numéricamente, reemplazar y rellenar nulos de las columnas restantes con el fin de reducir su falta de información lo máximo posible y generar agrupaciones para reducir la granularidad de algunas variables. Las codificaciones finales son:

| FAMI\_ESTRATOVIVIENDA | |
| --- | --- |
| Estrato 1 | 1 |
| Estrato 2 | 2 |
| Estrato 3 | 3 |
| Estrato 4 | 4 |
| Estrato 5 | 5 |
| Estrato 6 | 6 |
| FAMI\_CUARTOSHOGAR | |
| Uno | 1 |
| Dos | 2 |
| Tres | 3 |
| Cuatro | 4 |
| Cinco | 5 |
| Seis o más | 6 |
| FAMI\_EDUCACIONMADRE/PADRE | |
| No sabe | Ninguno |
| No aplica | Ninguno |
| FAMI\_COMELECHEDERIVADOS, COMECARNEPESCADOHUEVOS, COMECEREALFRUTOSLEGUMBRE | |
| Todos o casi todos los días | 1 |
| 1 o 2 veces por semana | 2 |
| 3 a 5 veces por semana | 3 |
| Nunca o rara vez comemos eso | 4 |
| FAMI\_NUMLIBROS | |
| MÁS DE 100 LIBROS | 1 |
| 11 A 25 LIBROS | 2 |
| 0 A 10 LIBROS | 3 |
| 26 A 100 LIBROS | 4 |
| ESTU\_DEDICACIONLECTURADIARIA | |
| No leo por entretenimiento | 3 |
| 30 minutos o menos | 4 |
| Entre 30 y 60 minutos | 6 |
| Entre 1 y 2 horas | 1 |
| Más de 2 horas | 2 |
| ESTU\_DEDICACIONINTERNET | |
| 0 | 1 |
| Más de 30 horas | 2 |
| Menos de 10 horas | 3 |
| Entre 11 y 20 horas | 4 |
| Entre 21 y 30 horas | 5 |
| COLE\_CARACTER | |
| NO APLICA | ACÁDEMICO |
| FAMI\_TRABAJOLABORPADRE/MADRE | |
| Es dueño de un negocio grande, tiene un cargo de nivel directivo o gerencial | 1 |
| Es agricultor, pesquero o jornalero | 2 |
| Trabaja como profesional (por ejemplo médico, abogado, ingeniero) | 3 |
| Trabaja en el hogar, no trabaja o estudia | 4 |
| Es dueño de un negocio pequeño (tiene pocos empleados o no tiene, por ejemplo tienda, papelería, etc. | 5 |
| No sabe | 6 |
| Es operario de máquinas o conduce vehículos (taxista, chofer) | 7 |
| Pensionado | 8 |
| Tiene un trabajo de tipo auxiliar administrativo (por ejemplo, secretario o asistente) | 9 |
| Es vendedor o trabaja en atención al público | 10 |
| Trabaja por cuenta propia (por ejemplo plomero, electricista) | 11 |
| No aplica | 6 |
| Trabaja como personal de limpieza, mantenimiento, seguridad o construcción | 12 |

1. Exploratory Data Analysis (EDA)

* Preparación de Datos: Los datos recopilados se sometieron a un proceso de limpieza y preprocesamiento para garantizar su calidad y coherencia.
* Análisis Exploratorio de Datos: Se llevó a cabo un análisis exploratorio de los datos para comprender su estructura y distribución. Se utilizaron técnicas estadísticas y herramientas de visualización para identificar patrones, tendencias y relaciones entre las variables.
* Análisis Descriptivo: Se realizaron análisis descriptivos para examinar el rendimiento académico en función de diversas variables, como el tipo de institución educativa, la ubicación geográfica, el nivel educativo de los padres, la jornada escolar, etc.

Después de realizar un análisis exhaustivo de los datos, se encontraron varios hallazgos significativos que proporcionan una visión clara del rendimiento académico en las pruebas ICFES 2023: Tamaño de la Población: La muestra analizada consistió en un total de 405,869 estudiantes. Este tamaño de muestra considerable proporciona una base sólida para sacar conclusiones estadísticamente significativas sobre el rendimiento académico en las pruebas ICFES. Promedio de Puntaje: El puntaje promedio obtenido en las pruebas fue de 258.8 puntos sobre un máximo de 500. Este valor proporciona una medida general del desempeño académico de la población estudiantil evaluada. En promedio, se observa una tendencia interesante en relación con el nivel económico y el rendimiento académico. Los estudiantes con un nivel económico más alto tienden a obtener puntajes promedio más altos en las pruebas. Este hallazgo sugiere una posible correlación entre el nivel económico y el acceso a recursos educativos adicionales o a una educación de mayor calidad, lo que puede influir positivamente en el desempeño académico. Por otro lado, los estudiantes con un nivel económico más bajo tienden a tener puntajes promedio más bajos, lo que podría indicar desafíos adicionales en términos de acceso a recursos educativos y apoyo académico. Estos resultados subrayan la importancia de abordar las disparidades socioeconómicas en la educación para garantizar oportunidades equitativas para todos los estudiantes. Al examinar los datos poblacionales, se observa que la población rural representa el 13.88%, mientras que la población urbana constituye el 86.12%. A pesar de esta disparidad en la distribución, los puntajes promedio obtenidos fueron del 48.03% y 51.97%, respectivamente. En cuanto al tipo de institución, aquellas no oficiales o privadas registraron un puntaje superior del 53.32%, a pesar de representar solo el 22.46% de la población. En cuanto al género, el 45.76% de la población corresponde al género femenino, mientras que el 54.24% al masculino. Los puntajes promedio fueron del 49.18% y 50.82%, respectivamente, lo que sugiere una diferencia mínima entre géneros en términos de rendimiento académico. Un hallazgo significativo revelado en el análisis es la correlación entre la calidad educativa de los padres y el desempeño académico de sus hijos. Se observó que a medida que aumentaba el nivel educativo de los padres, mejoraron los resultados de las pruebas de sus hijos. Este patrón sugiere una influencia positiva de la educación de los padres en el rendimiento académico de los estudiantes.

1. Machine Learning
   1. Feature Engineer:

Debido a complejidad del dataset realizado se procede a realizar agrupaciones y reducciones de términos de baja relevancia para reducir con mayor peso las granularidades:

* Se eliminan las columnas de puntajes por materia y se convierten sus respectivas columnas de percentiles en cuartiles.
* Se reducen los valores de bajo peso de las columnas de la residencia del municipio, departamento y presentación
* Se eliminan las columnas: 'COLE\_NOMBRE\_ESTABLECIMIENTO', 'COLE\_NOMBRE\_SEDE', 'ESTU\_MCPIO\_PRESENTACION' debido a su alta granularidad y su baja relevancia respecto con el objetivo del proyecto
* Se agrupan las columnas de comida en una sola
* Se agrupan las columnas de dedicación, número de libros y horas de trabajo en una sola
* Se eliminan valores de bajo peso en las columnas ‘ALIMENTACION’ (Columna de agrupación de comida se eliminan los que están por debajo de 0.9%), 'ESTU\_DEPTO\_RESIDE' (del 1.5%), 'ESTU\_MCPIO\_RESIDE' (del 0.04%) y 'COLE\_MCPIO\_UBICACION' (0.07%)

Se obtiene un dataset de:

* 37 columnas
* 180413 registros
* 0 nulos
* 0 duplicados

**Análisis de Clustering con K-Means**

Este análisis utilizó el algoritmo de clustering K-Means para agrupar datos relacionados con el desempeño académico y el nivel socioeconómico de los estudiantes. Se siguieron los siguientes pasos:

1. **Selección de Características**: Se seleccionaron características relevantes, como el período, el estrato de la vivienda, el número de cuartos en el hogar, el puntaje global y el nivel socioeconómico individual del estudiante.

2. **Normalización de Datos**: Los datos fueron normalizados para asegurar que todas las características tuvieran la misma escala.

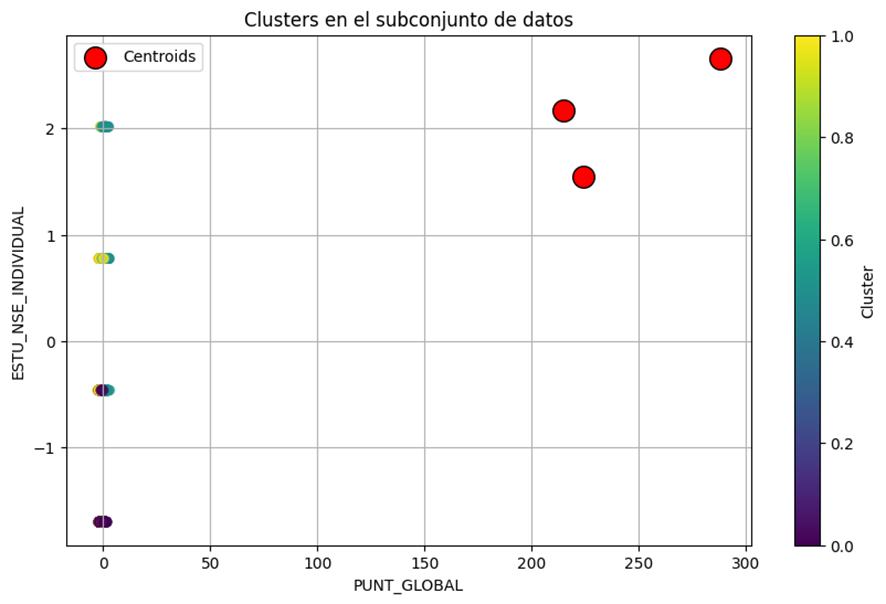
3. **División en Subconjuntos**: Debido al tamaño de los datos, se dividieron en subconjuntos más pequeños para facilitar el proceso de clustering.

4. **Aplicación de K-Means**: Se aplicó el algoritmo K-Means con un número de clusters predefinido de 3 a cada subconjunto de datos. Se evaluó la cohesión de los clusters mediante el puntaje de silueta.

5. **Visualización de los Clusters**: Los clusters resultantes se visualizaron en un gráfico de dispersión, mostrando los centroides de cada cluster.

6. **Estadísticas Resumidas**: Se calcularon estadísticas resumidas para cada cluster, incluyendo la media de los puntajes globales y del nivel socioeconómico individual, así como el número de datos en cada cluster.

Los resultados revelaron la existencia de tres clusters distintos, cada uno con características promedio diferentes de puntaje global y nivel socioeconómico individual. Por ejemplo, el Cluster 1 mostró puntajes globales y niveles socioeconómicos más altos en comparación con los otros clusters.



Cluster 0:

PUNT\_GLOBAL mean: -0.66

ESTU\_NSE\_INDIVIDUAL mean: -1.03

Count: 2587

Cluster 1:

PUNT\_GLOBAL mean: 0.56

ESTU\_NSE\_INDIVIDUAL mean: 0.35

Count: 1730

Cluster 2:

PUNT\_GLOBAL mean: -0.84

ESTU\_NSE\_INDIVIDUAL mean: -0.25

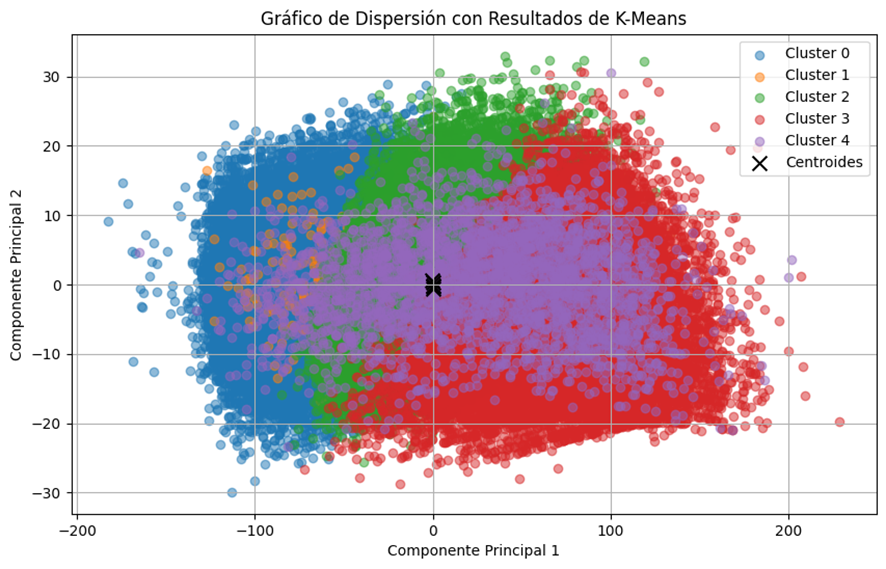
Count: 1552

Cluster 0: Este cluster tiene un valor medio de PUNT\_GLOBAL negativo y un valor medio de ESTU\_NSE\_INDIVIDUAL también negativo. Esto podría indicar un grupo de estudiantes con puntajes más bajos en la prueba global y un nivel socioeconómico individual más bajo.

Cluster 1: En este cluster, tanto el PUNT\_GLOBAL como el ESTU\_NSE\_INDIVIDUAL tienen valores medios positivos. Esto sugiere que estos estudiantes tienen puntajes más altos en la prueba global y un nivel socioeconómico individual más alto en comparación con otros clusters.

Cluster 2: Este cluster muestra un valor medio negativo para PUNT\_GLOBAL y un valor medio negativo para ESTU\_NSE\_INDIVIDUAL, aunque menos negativo que el Cluster 0. Esto podría representar un grupo intermedio de estudiantes con puntajes más bajos en la prueba global pero con un nivel socioeconómico individual algo más alto en comparación con el Cluster 0.

Análisis con base de datos reducida:



En el gráfico los centroides están juntos, porque los datos estén altamente superpuestos y además la dimensionalidad de los datos sea alta, lo que dificulta la visualización de los clusters de manera clara y separada. Esto hace que los datos sean difíciles de separar en un espacio bidimensional, como en este caso, donde los datos se han reducido a solo dos componentes principales.

El análisis realizado consistió en aplicar el algoritmo de clustering K-Means a un conjunto de datos que incluían características relacionadas con el desempeño académico y el contexto socioeconómico de los estudiantes. A continuación, se presenta un resumen de los pasos y resultados principales:

1. **Preprocesamiento de Datos**:
   * Se seleccionaron características relevantes, como el período, el estrato de la vivienda, el número de cuartos en el hogar, el puntaje global y el nivel socioeconómico individual del estudiante.
   * Los datos fueron normalizados para asegurar que todas las características tuvieran la misma escala.
2. **Clustering con K-Means**:
   * Se aplicó el algoritmo K-Means con un número de clusters predefinido de 3.
   * Se dividieron los datos en subconjuntos más pequeños para facilitar el proceso de clustering.
   * Se calcularon estadísticas resumidas para cada cluster, incluyendo la media de los puntajes globales y del nivel socioeconómico individual, así como el número de datos en cada cluster.
3. **Visualización de Resultados**:
   * Se utilizó PCA (Análisis de Componentes Principales) para reducir la dimensionalidad de los datos a dos componentes principales.
   * Se creó un gráfico de dispersión que mostraba la distribución de los datos en función de los dos componentes principales, coloreando los puntos según el cluster al que pertenecían.
   * Se observó que en el gráfico los centroides de los clusters estaban juntos, lo que indicaba una posible superposición de los clusters en el espacio de las características reducidas.

En resumen, el análisis reveló la existencia de tres clusters distintos con características promedio diferentes de puntaje global y nivel socioeconómico individual. Sin embargo, la visualización de los resultados sugiere que la separación entre los clusters podría no ser completamente clara en el espacio de dos dimensiones después de la reducción de la dimensionalidad.

1. Deploy del Modelo y Dashboard

El despliegue del dashboard se realizó en la web de NoviPro, que realiza el servicio de sincronización con una cuenta de Power BI service, y proporciona los recursos y el entorno para el funcionamiento dinámico del dashboard con todas sus funcionalidades:

URL dashboard: https://www.novypro.com/project/proyecto-del-an%C3%A1lisis-de-desempe%C3%B1o-acad%C3%A9mico

1. Conclusiones

El análisis del rendimiento académico en las pruebas ICFES 2023 ha proporcionado insights valiosos sobre diversos factores que pueden influir en los resultados de los estudiantes. Se ha encontrado que variables como el nivel socioeconómico, el tipo de institución educativa, el género y el nivel educativo de los padres están relacionados con el desempeño académico. Estos hallazgos resaltan la importancia de abordar las disparidades educativas y socioeconómicas para promover la equidad en la educación y mejorar los resultados estudiantiles. Además, subrayan la necesidad de políticas y programas dirigidos a mejorar el acceso a recursos educativos de calidad y a apoyar el desarrollo académico de todos los estudiantes, independientemente de su origen socioeconómico o género.