Informe Parcial II Procesamiento de Datos a Gran Escala

Daniela Torres Gómez,Isaac Janica, Daniel Sandoval Higuera

[1]Ciencia de Datos & Pontifica Universidad Javeriana, Colombia

[2]Ingeniería de Sistemas & Pontifica Universidad Javeriana, Colombia

En este documento se desarrolló un proyecto para poner en práctica, con datos reales, los conceptos aprendidos en la clase de procesamiento de datos a gran escala. Exploramos y brindamos información útil acerca los datos de los resultados de las pruebas de estado Saber 11.

Keywords— Saber 11, pyspark, big-data,, procesamiento distribuido, análisis descentralizado, educación, evaluación académica, machine learning, modelos predictivos.

I. INTRODUCTION

Las pruebas Saber 11, en el contexto educativo de Colombia, representan un pilar fundamental en la evaluación del rendimiento académico de los estudiantes de educación media. Estas pruebas, administradas por el Instituto Colombiano para la Evaluación de la Educación (ICFES), abarcan diversas áreas del conocimiento y proporcionan una métrica objetiva para medir el nivel de competencia de los estudiantes en habilidades básicas.

En este informe, nos adentramos en el mundo del procesamiento de datos a gran escala con un enfoque específico en los resultados de las pruebas Saber 11. A través de la aplicación de técnicas de big data y herramientas como pyspark, buscamos analizar exhaustivamente estos datos para extraer información valiosa que pueda contribuir a comprender mejor el panorama educativo colombiano.

La importancia de este proyecto radica en la posibilidad de identificar tendencias, patrones y factores que influyen en el desempeño de los estudiantes en las pruebas Saber 11. Estos insights no solo pueden beneficiar a las instituciones educativas y al sistema de evaluación, sino también a los encargados de formular políticas educativas y a la sociedad en general, al proporcionar una base sólida para la toma de decisiones informadas en el ámbito educativo.

En este informe, presentaremos nuestra metodología, resultados preliminares y reflexiones sobre el impacto potencial de nuestro trabajo en el mejoramiento continuo de la calidad de la educación en Colombia..

II. ENTENDIMIENTO DEL NEGOCIO Y PLANTEAMIENTO PROBLEMÁTICA

El proceso de evaluación del rendimiento académico de los estudiantes mediante las pruebas Saber 11 en Colombia no solo es crucial para las instituciones educativas, sino que también influye en las políticas públicas y en la percepción general sobre la calidad de la educación en el país. Sin embargo, este proceso enfrenta desafíos significativos en términos de análisis y comprensión de los datos generados.

La complejidad de los datos de las pruebas Saber 11, que abarcan múltiples áreas de conocimiento y diversos factores socioeconómicos, requiere un enfoque de procesamiento de datos a gran escala para obtener insights significativos. Por lo tanto, el objetivo principal de este proyecto es utilizar técnicas de big data y herramientas como pyspark para analizar de manera exhaustiva estos datos y abordar las siguientes problemáticas:

Identificación de patrones y tendencias: ¿Qué patrones se pueden identificar en los resultados de las pruebas Saber 11 a lo largo del tiempo (pre-pandemia y post-pandemia)? ¿Existen tendencias claras en el desempeño de los estudiantes en diferentes áreas del conocimiento?

Influencia de variables socioeconómicas: ¿Cómo afectan factores como el nivel socioeconómico y el tipo de institución educativa al rendimiento académico de los estudiantes en las pruebas Saber 11? ¿Existen disparidades significativas que requieran atención específica?

Predicción del rendimiento estudiantil: ¿Es posible desarrollar modelos predictivos utilizando datos históricos de las pruebas Saber 11? ¿Qué variables son más relevantes para predecir el desempeño académico de los estudiantes?

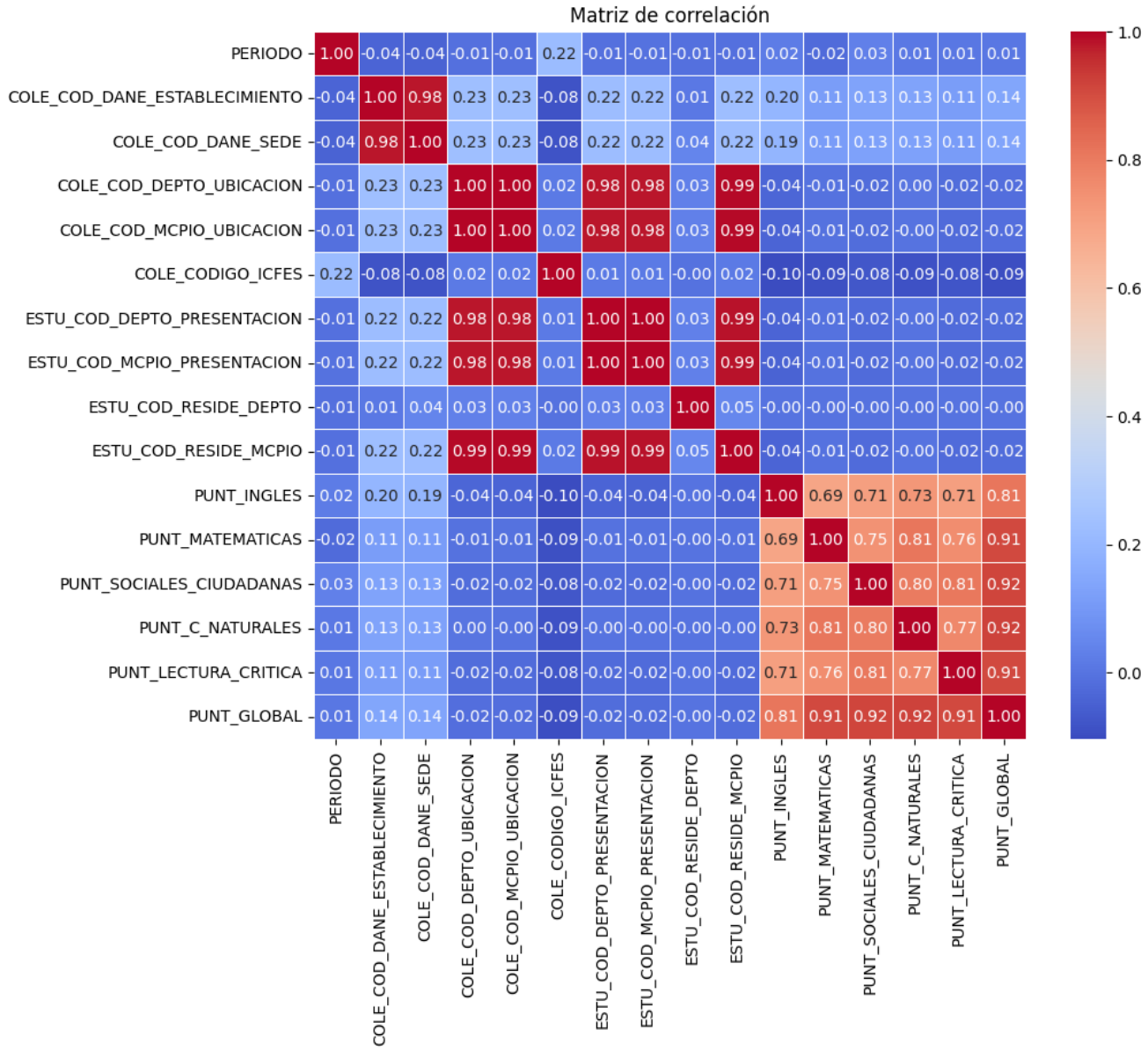
Al abordar estas problemáticas, buscamos no solo proporcionar una visión integral del rendimiento académico en Colombia, sino también generar insights accionables que puedan informar políticas educativas y estrategias de intervención destinadas a mejorar la calidad y equidad en la educación del país. En este informe, detallaremos nuestra metodología para abordar estas cuestiones y presentaremos los resultados preliminares obtenidos hasta el momento..

III. ANÁLISIS EXPLORATORIO DE DATOS

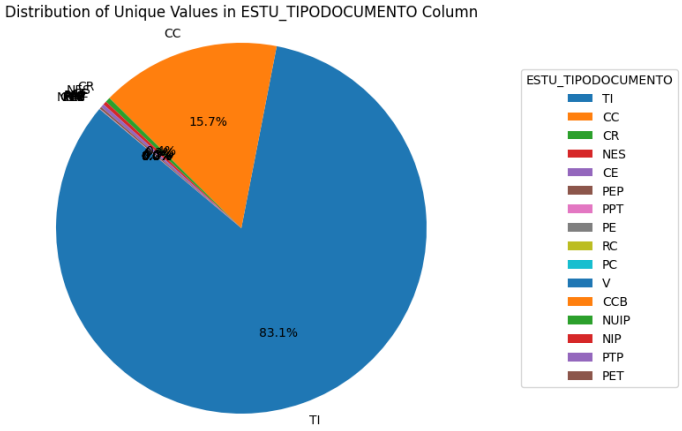
Se cuentan con unos datos iniciales con 51 columnas. Nos enfocamos en los datos numéricos ya que estos nos permiten observar correlaciones más fácilmente.

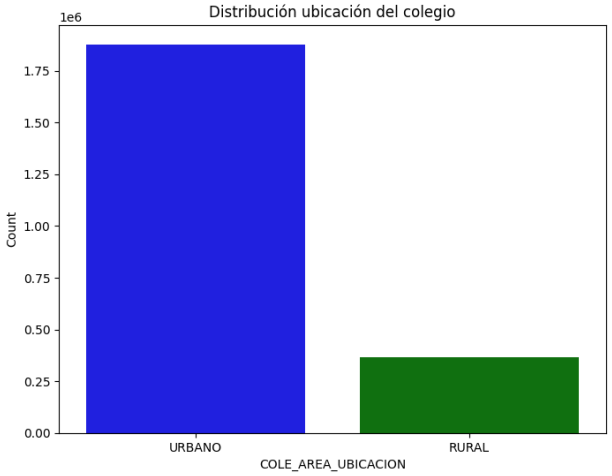
Realizamos gráficas, pre-pandemia, post-pandemia y ambas juntas.

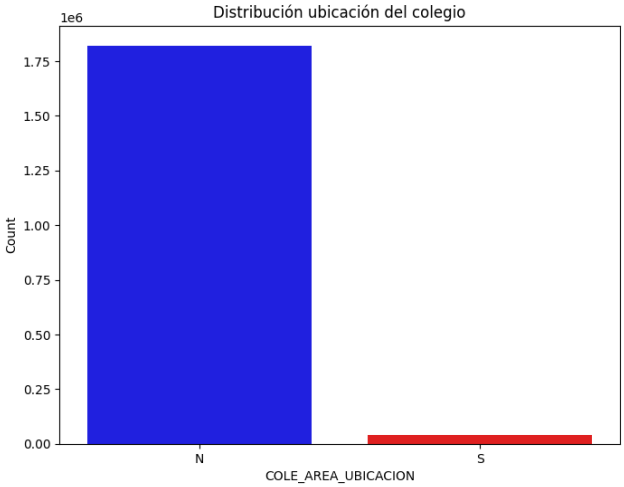
A continuación mostramos una matriz de correlación, únicamente de las variables que contienen solo números:

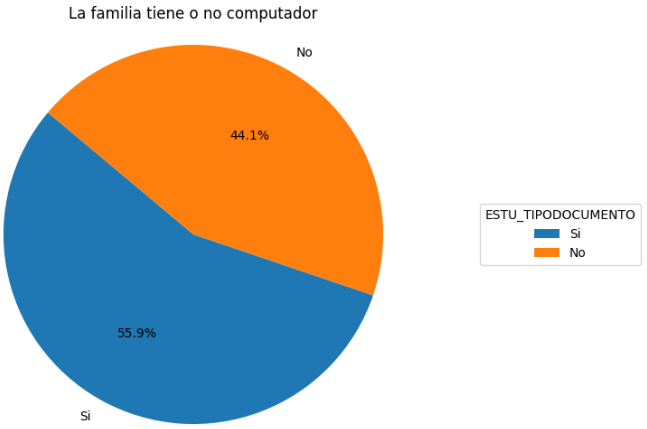


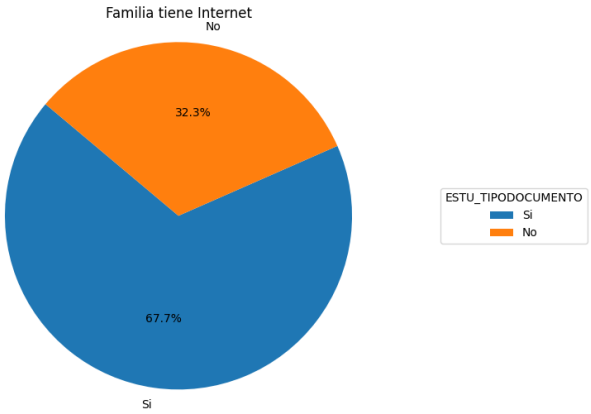
Ahora revisaremos algunas de las variables que nos parecieron interesantes:

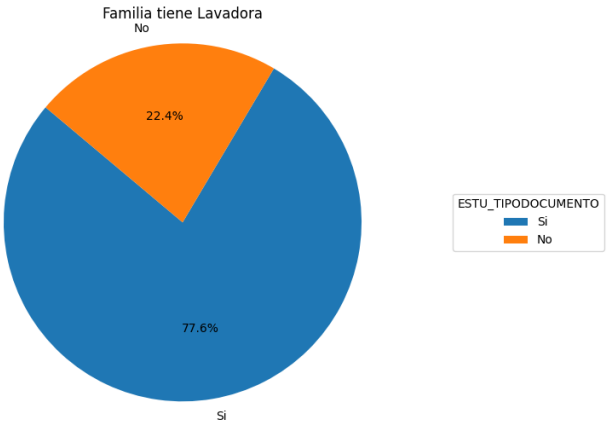

Distribution of Unique Values in ESTU_TIPODOCUMENTO Column
CC
15.7%
83.1%
TI
ESTU_TIPODOCUMENTO
TI
CC
CR
NES
CE
PEP
PPT
PE
RC
PC
V
CCB
NUIP
NIP
PTP
PET

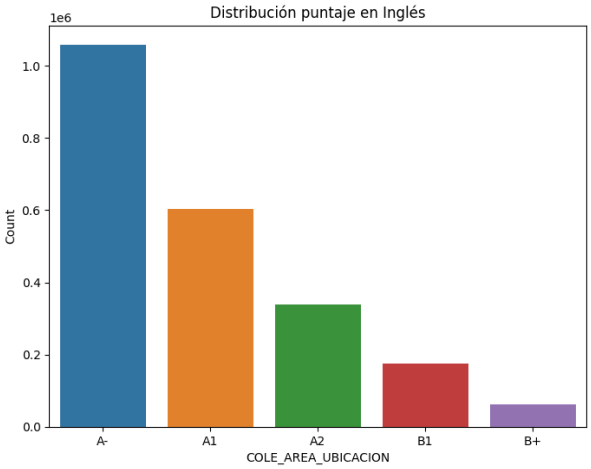

Distribución ubicación del colegio
1e6
Count
URBANO
RURAL
COLE_AREA_UBICACION

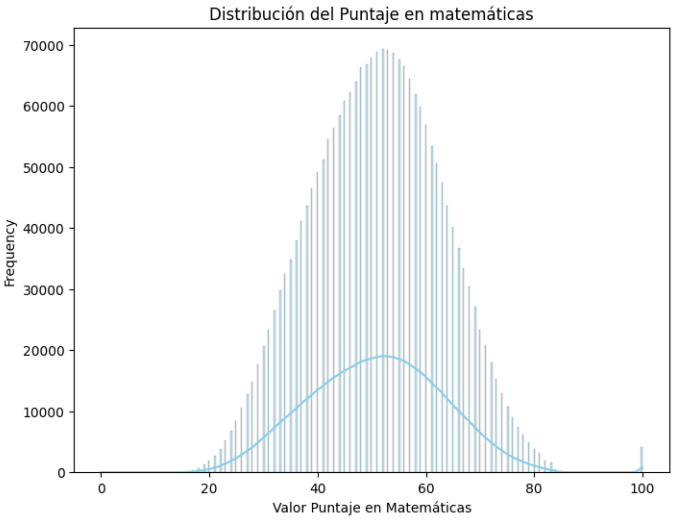

Distribución ubicación del colegio
1e6
Count
N
S
COLE_AREA_UBICACION

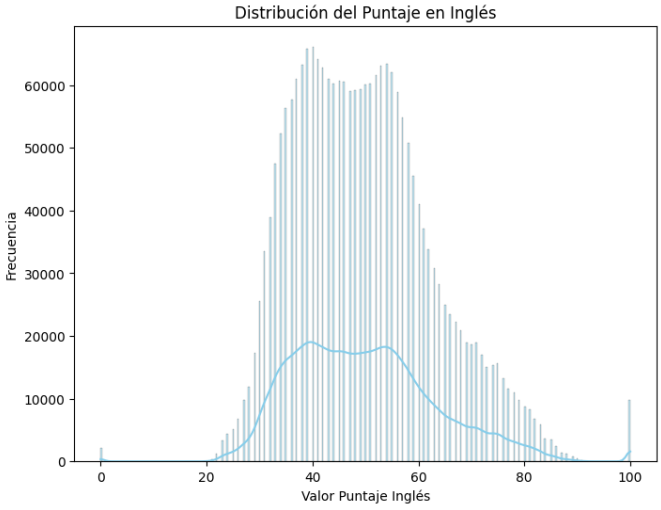

La familia tiene o no computador
No
44.1%
55.9%
Si
ESTU_TIPODOCUMENTO
Si
No

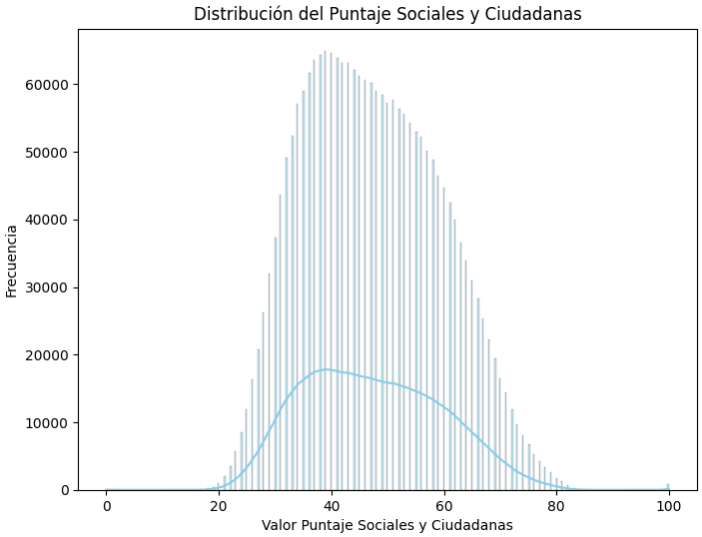

Familia tiene Internet
No
32.3%
67.7%
Si
ESTU_TIPODOCUMENTO
Si
No

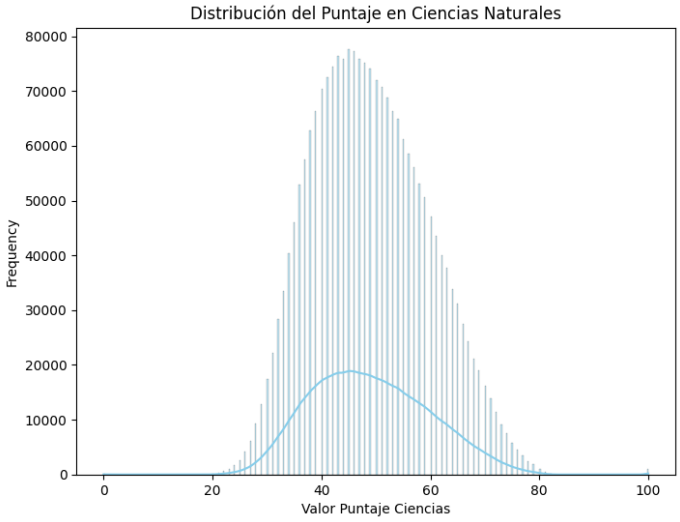

Familia tiene Lavadora
No
22.4%
77.6%
Si
ESTU_TIPODOCUMENTO
Si
No

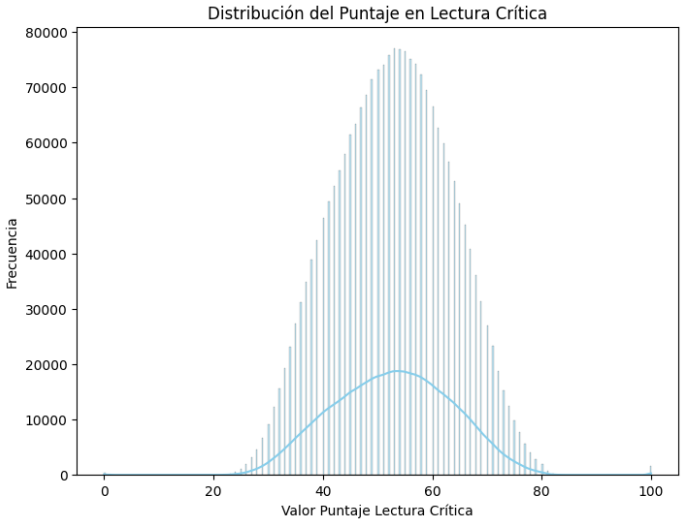

Distribución puntaje en Inglés
1e6
Count
A-
A1
A2
B1
B+
COLE_AREA_UBICACION

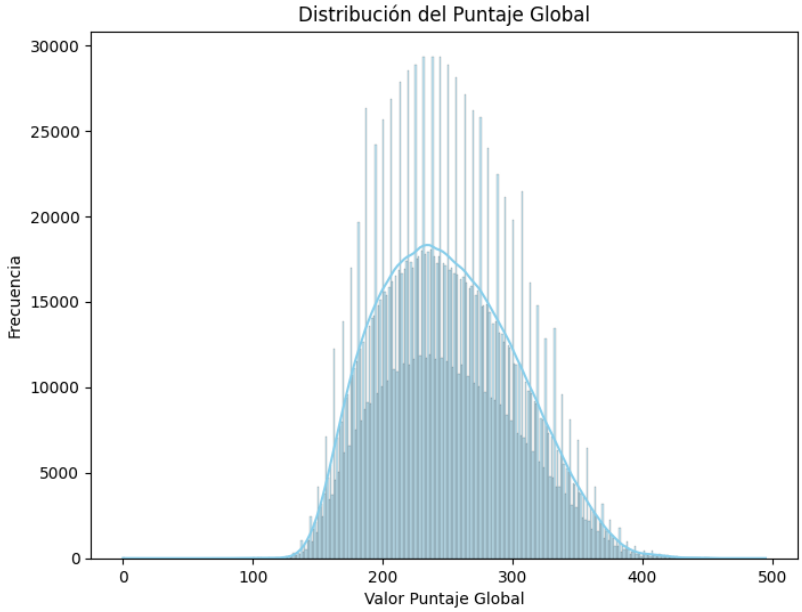

Distribución del Puntaje en matemáticas
Frequency
Valor Puntaje en Matemáticas











Se puede observar que los puntajes de las pruebas siguen una distribución normal.

IV. LIMPIEZA DE DATOS

V. MODELADO

VI. CONCLUSIONES

REFERENCIAS: