INFORME ESTADISTICA INFERENCIAL

Alejandro Jaime Martinez U00174986

Andrés Fabian Leal Archila

Introducción

Este informe se basa en un análisis exploratorio de datos de un conjunto de datos titulado "produccioncrudo.csv".

El objetivo de este análisis es comprender la estructura de los datos, identificar valores faltantes y obtener un resumen estadístico del conjunto de datos. El proceso se lleva a cabo utilizando las bibliotecas de Python pandas, numpy, matplotlib.pyplot y seaborn.

Desarrollo del Documento

El análisis comienza cargando el archivo producioncrudo.csv en un DataFrame de pandas. A continuación, se realizan los siguientes pasos:

Información del dataset: Se utiliza el método df.info() para mostrar la información básica del dataset, incluyendo el número de entradas, las columnas, la cantidad de valores no nulos y los tipos de datos de cada columna. El dataset contiene 48,416 entradas y 15 columnas.

Primeras filas: Se muestran las primeras filas del dataset con df.head() para tener una visión inicial de los datos.

Resumen estadístico: Se genera un resumen estadístico con df.describe(), que proporciona medidas como el conteo, la media, la desviación estándar, los valores mínimo y máximo, y los cuartiles para las columnas numéricas.

Valores faltantes: Se identifican y cuentan los valores nulos en cada columna utilizando df.isnull().sum(). Las columnas con valores faltantes son TipoContrato (437 nulos), Longitud (7,927 nulos), Latitud (7,927 nulos) y Geolocalizacion (7,927 nulos).

Visualización de nulos: Se crea un mapa de calor (heatmap) con seaborn para visualizar gráficamente la distribución de los valores faltantes en el dataset.

Conclusiones

El análisis exploratorio inicial del archivo producioncrudo.csv revela que el dataset está bien estructurado con una mezcla de tipos de datos enteros, flotantes y de objetos. Los datos están relacionados con la producción de crudo, incluyendo campos como vigencia, mes, departamento, municipio, operadora, contrato, campo, geolocalización y la producción en barriles. Sin embargo, se identificó un número significativo de valores faltantes en las columnas relacionadas con la geolocalización y, en menor medida, en la columna TipoContrato. Estos valores faltantes deberán ser tratados en futuras etapas del análisis para garantizar la integridad y la calidad de cualquier modelo o visualización posterior.