PREDICT CO2 EMISSIONS IN RWANDA - IA PROYECT

INTEGRANTES:
Raúl Felipe Berrio
Esteban Machado
Diego Andrés Rodriguez Galeano

MATERIA:
Introducción a la Inteligencia Artificial

PROFESOR:
Raúl Ramos Pollan

UNIVERSIDAD DE ANTIOQUIA
FACULTAD DE INGENIERÍA
2023

CONTENDO

CONTENIDO DE FIGURAS

CONTENIDO DE TABLAS

INTRODUCCIÓN

La inteligencia artificial se ha convertido en una herramienta poderosa y versátil en la actualidad, con la capacidad de desplegarse en una amplia gama de aplicaciones y campos profesionales. Mediante el análisis de datos de diversos procesos y operaciones, la IA tiene la capacidad de extraer un valor significativo y relevante. Al utilizar estos datos para entrenar algoritmos de aprendizaje automático, es posible desarrollar modelos que permiten comprender patrones y tendencias, así como reducir la complejidad de la información sin comprometer su relevancia. Estos modelos, a su vez, tienen la capacidad de ofrecer predicciones precisas y útiles que respaldan la toma de decisiones informada a corto, mediano y largo plazo en diversos contextos empresariales y tecnológicos.

En este estudio se analizará la implementación de algoritmos de Aprendizaje Automático para abordar la monitorización precisa de las emisiones de dióxido de carbono (CO2), un factor crucial en la lucha contra el cambio climático. La capacidad de obtener mediciones precisas del carbono juega un papel fundamental en la comprensión de las fuentes y los patrones de emisión de CO2. Aunque Europa y América del Norte tienen sistemas sólidos para monitorear las emisiones de carbono en tierra, en el continente africano existe una carencia de dichos sistemas.

1. PLANTEAMIENTO DEL PROBLEMA

El problema predictivo se enfoca en la monitorización precisa de las emisiones de dióxido de carbono (CO2), una parte fundamental en la lucha contra el cambio climático. La capacidad de obtener lecturas precisas de carbono permite a los investigadores y gobiernos comprender las fuentes y patrones de emisión de CO2. Aunque Europa y América del Norte cuentan con sistemas extensos para monitorear las emisiones de carbono en tierra, en África hay una falta de tales sistemas disponibles. El objetivo de este desafío es crear modelos de machine learning utilizando datos de emisiones de CO2 de código abierto provenientes de observaciones satelitales Sentinel-5P para predecir las futuras emisiones de carbono. Estas soluciones pueden ayudar a los gobiernos y otras partes interesadas a estimar los niveles de emisiones de carbono en toda África.

1.1. DATASET

Utilizaremos el conjunto de datos GRACED proporcionado por Carbón Monitor(con 76 columnas y más de 79000 datos), basado en observaciones satelitales Sentinel-5P. Este conjunto de datos contiene mediciones de concentración de CO2 y otros gases relacionados en diferentes ubicaciones de África, así como variables temporales y geoespaciales relevantes. El conjunto de datos contiene siete características principales que se extrajeron semanalmente de las observaciones de Sentinel 5P desde enero de 2019 hasta noviembre de 2022. Estas características son:

• Sulfuro de Dióxido (Sulphur Dioxide): Medido por COPERNICUS/S5P/NRTI/L3_SO2.
• Monóxido de Carbono (Carbon Monoxide): Medido por COPERNICUS/S5P/NRTI/L3_CO.
• Dióxido de Nitrógeno (Nitrogen Dioxide): Medido por COPERNICUS/S5P/NRTI/L3_NO2.
• Formaldehído (Formaldehyde): Medido por COPERNICUS/S5P/NRTI/L3_HCHO.
•Índice de Aerosol UV(UV Aerosol Index): Medido por COPERNICUS /S5P /NRTI / L3_AER_AI.
• Ozono (Ozone): Medido por COPERNICUS/S5P/NRTI/L3_O3.
• Nubes (Cloud): Medido por COPERNICUS/S5P/OFFL/L3_CLOUD.

1.2. MÉTRICA

Métrica de Machine Learning: La métrica de desempeño para evaluar los modelos será el Error Cuadrático Medio (RMSE, por sus siglas en inglés).

$$\epsilon = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(log(p_i + 1) - log(a_i + 1)\right)^2}$$

El RMSE mide la raíz cuadrada de la diferencia entre los valores predichos y los valores reales, lo que proporciona una medida de cuán cerca están las predicciones de los valores reales de emisiones de CO2.

1.3 VARIABLE OBJETIVO

En el contexto del análisis presentado, la variable objetivo es "emission". Esta variable representa la cantidad de un determinado contaminante o sustancia que es liberada en un entorno específico. El objetivo del análisis es entender los factores y características asociadas con las emisiones en una ubicación particular y durante una determinada semana.

Se explorarán distintas variables que podrían estar relacionadas con las emisiones, como la ubicación geográfica, las condiciones climáticas, la densidad poblacional y otros indicadores relevantes. La predicción precisa de las emisiones puede permitir la identificación de patrones y tendencias, así como el desarrollo de estrategias efectivas para controlar y reducir las emisiones contaminantes.

2. EXPLORACIÓN DE VARIABLES

La exploración de variables comienza con la integración de la información distribuida en varios conjuntos de datos, incluyendo "building_metadata", "weather_train" y "train". Estos conjuntos de datos se combinan para formar un único conjunto de datos consolidado, el cual se nombra simplemente como "train" y se utiliza como punto de partida para el análisis.

Una vez que se ha creado este conjunto de datos consolidado, se procede a analizar una serie de variables que se consideran de interés para el estudio. Estas variables pueden incluir características específicas de los edificios, datos climáticos relevantes y mediciones de consumo de energía. El objetivo de este análisis inicial es identificar patrones, tendencias y posibles relaciones entre estas variables, lo que podría proporcionar información valiosa para futuros análisis y la construcción de modelos predictivos.

2.1 ANÁLISIS DE LA VARIABLE OBJETIVO

Como parte del análisis de las variables, se realiza un examen detallado del comportamiento de la variable objetivo. En la Figura 1 se representa la distribución de esta variable, revelando una asimetría significativa hacia la izquierda. Para abordar esta asimetría y lograr una distribución más adecuada, se propone la aplicación de una transformación logarítmica.

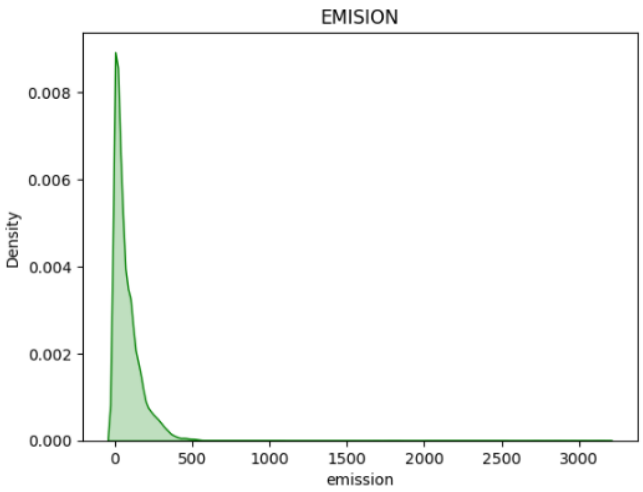

Figura 2.1 Distribución de la variable objeto

2.2 DATOS FALTANTES

DATOS FALTANTES DEL CONJUNTO TRAIN

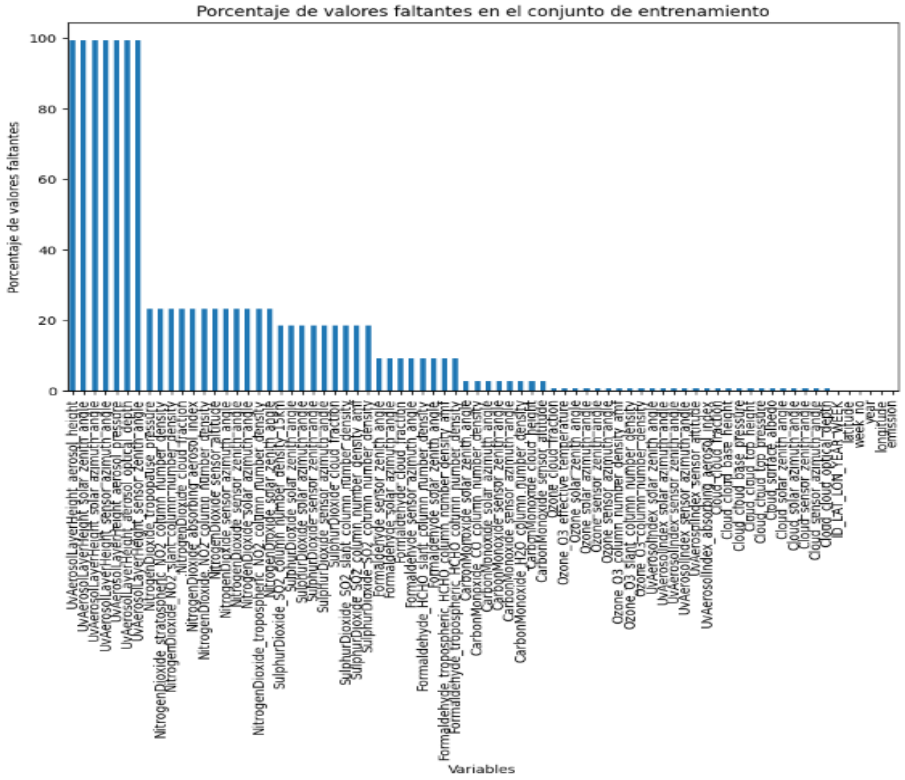

Figura 2.2 Datos faltantes del conjunto train

DATOS FALTANTES PARA EL CONJUNTO TEST

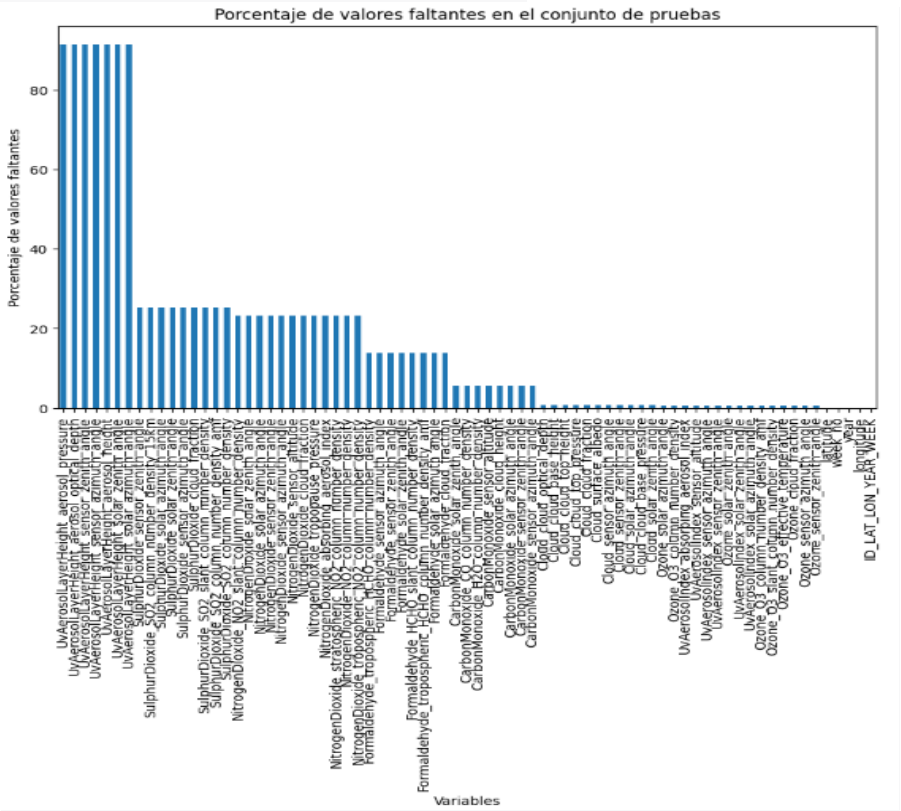

Figura 2.3 Datos faltantes del conjunto test

2.3 CORRELACIÓN

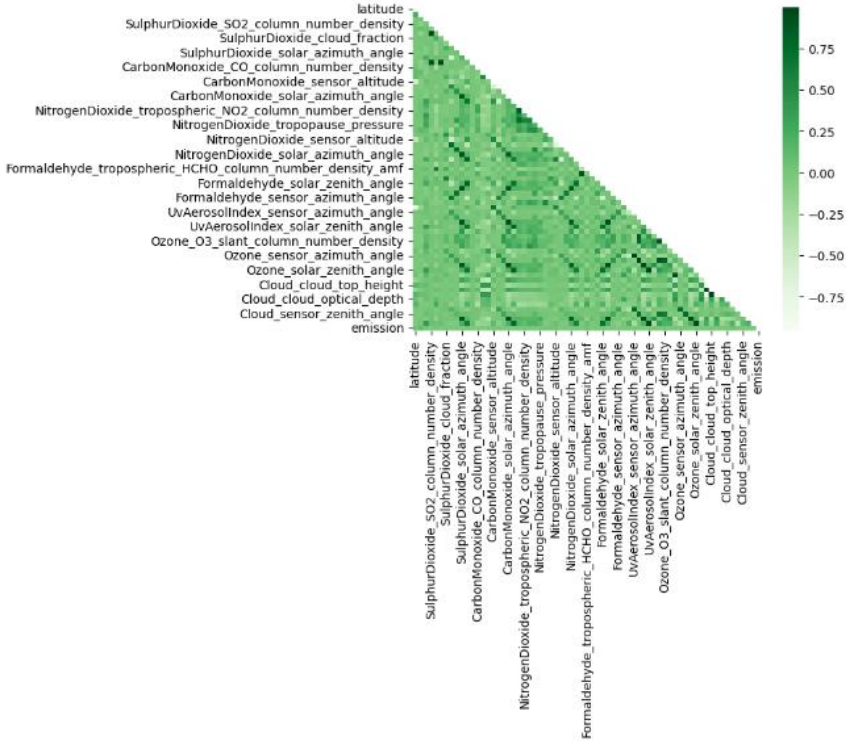

Figura 2.4 Correlación

Se evidencia la correlación entre todas las variables numéricas en el conjunto de datos. Luego se utiliza para representar esta matriz de correlación como un mapa de calor, donde los valores más altos de correlación se muestran en tonos más oscuros del color especificado, y los valores más bajos se muestran en tonos más claros.
La máscara triangular superior se utiliza para ocultar la mitad superior de la matriz de correlación, ya que la matriz de correlación es simétrica y la mitad superior es un reflejo de la mitad inferior

2.4. EXPLORACIÓN GEOESPACIAL

Se crearon visualizaciones geoespaciales utilizando la biblioteca Folium para mostrar la distribución de las emisiones en el mapa. Esto proporcionó una perspectiva geográfica de las emisiones y resaltó áreas específicas de interés.

Figura 2.5 Mapa geoespacial de África

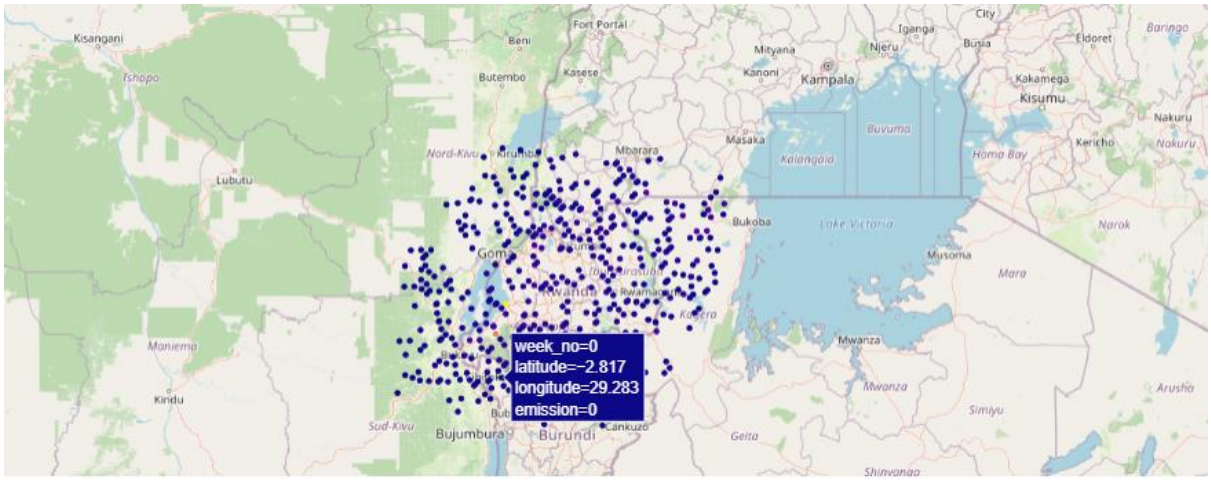

Figura 2.6 Mapa geoespacial de Rwanda

3. MODELADO

Se crea una nueva columna llamada 'location' que combina las coordenadas de latitud y longitud para cada registro en el conjunto de datos. Esto podría ser útil para identificar patrones de emisiones en ubicaciones específicas y analizar cómo varían según la geolocalización.

La conversión de la columna 'location' en un tipo de dato categórico es útil para reducir el uso de memoria y acelerar ciertas operaciones, especialmente cuando se trabaja con conjuntos de datos grandes. Esto optimiza el manejo de datos categóricos durante el análisis y modelado subsiguientes.

El Data Frame agregado 'train_agg' se crea para calcular y almacenar la media de las emisiones agrupadas por ubicación y número de semana. Este paso puede ayudar a identificar tendencias o patrones generales de emisiones en diferentes ubicaciones a lo largo del tiempo.

	location	week_no	emission
0	-0.51_29.29	0	3.608051
1	-0.51_29.29	1	4.016319
2	-0.51_29.29	2	4.138755
3	-0.51_29.29	3	4.184737
4	-0.51_29.29	4	4.247569

Tabla 3.1 DataFrame Train_agg

Se divide un conjunto de datos de entrenamiento en dos subconjuntos distintos, uno que abarca el período anterior al brote de la pandemia de COVID-19 y otro que abarca el período durante la pandemia. La finalidad de esta división es facilitar un análisis diferenciado de los datos y permitir la evaluación de posibles cambios, tendencias o impactos ocasionados por la pandemia en las variables de interés.

Al dividir el conjunto de datos en dos segmentos, uno que representa el periodo pre-COVID y otro que representa el periodo de COVID, se pueden realizar comparaciones entre las dos etapas y analizar cómo ciertas variables pueden haber sido afectadas por el brote de la pandemia. Esto puede ayudar a identificar patrones, tendencias o cambios significativos en

los datos, lo que a su vez puede ser útil para comprender mejor el impacto de la pandemia en diferentes aspectos o fenómenos relacionados con los datos en cuestión.

```
train_nocovid = train[(train.year == 2019) |
                      ((train.year == 2020) & (train.week_no <= 8))  |
                      ((train.year == 2020) & (train.week_no >= 32)) |
                      (train.year == 2021)]
train_covid = train[((train.year == 2020) & (train.week_no > 8)) &
                    ((train.year == 2020) & (train.week_no < 32))]

assert train_nocovid.shape[0] + train_covid.shape[0] == train.shape[0]
```

Figura 3.1 Segmentación de datos de entrenamiento en periodos de COVID y no COVID

Se crea una instancia del objeto AutoML, que se utiliza para ejecutar el proceso de aprendizaje automático automático. Se definen varias configuraciones en el diccionario automl_settings, que incluyen el presupuesto de tiempo máximo de 200 segundos, la métrica de evaluación establecida como RMSE (Root Mean Square Error), la tarea identificada como una tarea de regresión, y el uso de todos los núcleos de la CPU disponibles para el proceso. Además, se especifica el método de evaluación como validación cruzada.
Posteriormente, el método fit se utiliza para ajustar el modelo a los datos de entrenamiento, X_train y y. Estas configuraciones y el proceso automatizado proporcionado por FLAML pueden simplificar significativamente el proceso de desarrollo y optimización de modelos de regresión, lo que permite una mayor eficiencia y rapidez en la identificación del mejor modelo para un conjunto de datos determinado.

4. ITERACIONES DE DESARROLLO

Descripcion estadistica
Conteo de datos almacenados:

- Proporciona la cantidad de registros válidos para cada variable en el conjunto de entrenamiento.
- Puede ayudar a identificar la presencia de valores faltantes o inconsistencias en los datos.

Media aritmética:

- Ofrece la media aritmética para cada variable.
- Indica el valor promedio de las variables, lo que es crucial para comprender el comportamiento general de las emisiones.

Desviación estándar:

- Muestra la dispersión de los datos alrededor de la media para cada variable.
- Puede revelar la variabilidad en las emisiones y ayudar a evaluar la consistencia de los datos.

Estos resultados son esenciales para el proceso de exploración de datos y el entendimiento de la distribución de las variables clave en el conjunto de entrenamiento. Por ejemplo, si la variable objetivo "emisiones" tiene una desviación estándar significativa, esto podría indicar una variabilidad considerable en las emisiones a lo largo del tiempo o en diferentes ubicaciones.
Valor mínimo:

- En el caso de las emisiones de gases, el valor mínimo representaría la menor cantidad de emisiones observadas. Esto podría ser útil para identificar situaciones con emisiones excepcionalmente bajas.

Percentil 25% (Q1):

- Indicaría el punto en el cual el 25% de las observaciones tienen emisiones inferiores. Esto ayuda a comprender la distribución de las emisiones en el cuartil inferior.

Percentil 50% (Q2 o mediana):

- Representaría la mediana de las emisiones. Indica el valor medio y puede ser útil para entender la cantidad de emisiones típica.

Percentil 75% (Q3):

- Indicaría el punto en el cual el 75% de las observaciones tienen emisiones inferiores. Ayuda a analizar la distribución en el cuartil superior.

Valor máximo:

- Representaría la mayor cantidad de emisiones observada. Puede ayudar a identificar situaciones con emisiones excepcionalmente altas.

La interpretación de estos estadísticos específicos para las emisiones de gases permite comprender la variabilidad en los niveles de emisión y puede ser crucial para identificar posibles patrones, establecer límites para valores atípicos y tomar decisiones informadas sobre la gestión ambiental.

```
Conteo de datos:
 latitude                                      79023.0
longitude                                     79023.0
week_no                                       79023.0
SulphurDioxide_SO2_column_number_density      79023.0
SulphurDioxide_SO2_column_number_density_amf  79023.0
                                               ...
Cloud_sensor_azimuth_angle                    79023.0
Cloud_sensor_zenith_angle                     79023.0
Cloud_solar_azimuth_angle                     79023.0
Cloud_solar_zenith_angle                      79023.0
emission                                      79023.0
Name: count, Length: 67, dtype: float64

Media aritmética:
 latitude                                      -1.891072
longitude                                     29.880155
week_no                                       26.000000
SulphurDioxide_SO2_column_number_density       0.000043
SulphurDioxide_SO2_column_number_density_amf   0.830091
                                                ...
Cloud_sensor_azimuth_angle                    -10.796402
Cloud_sensor_zenith_angle                      40.441157
Cloud_solar_azimuth_angle                     -86.787376
Cloud_solar_zenith_angle                       27.928478
emission                                       84.146988
Name: mean, Length: 67, dtype: float64

Desviación estándar:
 latitude                                       0.694522
longitude                                      0.810375
week_no                                       15.297155
SulphurDioxide_SO2_column_number_density       0.000246
SulphurDioxide_SO2_column_number_density_amf   0.167669
                                                ...
Cloud_sensor_azimuth_angle                     30.281658
Cloud_sensor_zenith_angle                       6.408721
Cloud_solar_azimuth_angle                      37.721592
Cloud_solar_zenith_angle                        4.390443
emission                                      147.601014
Name: std, Length: 67, dtype: float64
```

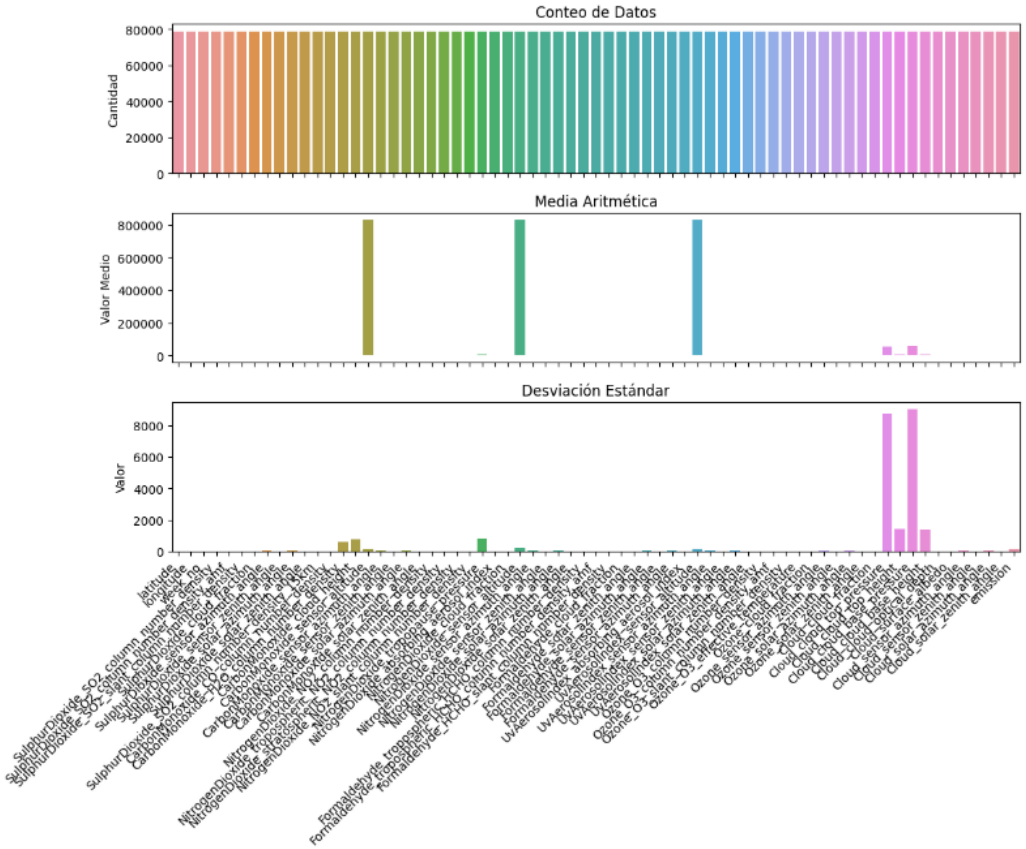

Figura 4.1 Grafica descripción estadística 1

```
Estadísticas de las emisiones:
count    79023.000000
mean        84.146988
std        147.601014
min          0.000000
25%         10.193523
50%         46.897280
75%        112.321268
max       3167.768000
Name: emission, dtype: float64
```

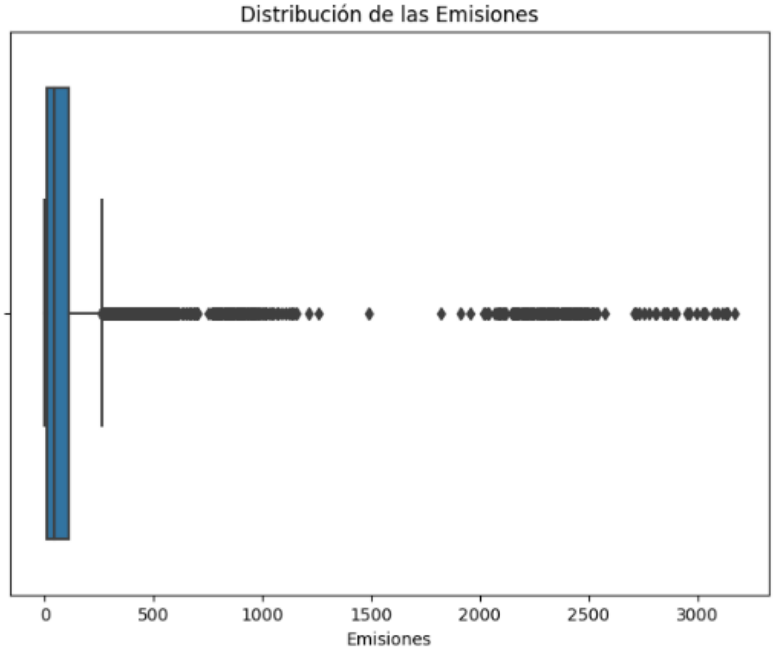

Figura 4.2 Grafica descripción estadística 2

5. RETOS Y CONSIDERACIONES DE DESPLIEGUE

- **Recopilación de datos en tiempo real:** Asegurarse de que los datos en tiempo real estén disponibles y se integren de manera efectiva en el modelo para garantizar pronósticos precisos y actualizados.
- **Mantenimiento del modelo:** Implementar un proceso de mantenimiento regular para actualizar el modelo con nuevos datos y mejorar su precisión con el tiempo.
- **Escalabilidad:** Asegurarse de que el modelo sea escalable y pueda manejar grandes volúmenes de datos de manera eficiente, especialmente si la aplicación se implementa en un entorno de producción a gran escala.
- **Monitoreo y actualización:** Establecer un sistema de monitoreo continuo para supervisar el rendimiento del modelo en el tiempo y garantizar que las predicciones se mantengan precisas y actualizadas.
- **Interpretabilidad del modelo:** Comprender y comunicar de manera efectiva cómo funciona el modelo y qué características contribuyen más a las predicciones puede ser crucial para ganar la confianza de los usuarios y las partes interesadas.
- **Implementación en entornos de producción:** Utilizar herramientas y prácticas adecuadas, como Docker u otras soluciones de contenedores, para garantizar un despliegue eficiente y efectivo en un entorno de producción.

6. CONCLUSIONES

- El preprocesamiento de datos incluye la creación de una categoría de ubicación basada en la combinación de latitud y longitud, lo que sugiere que la ubicación geográfica puede ser un factor importante en la predicción de emisiones.
- Se realiza un análisis de las emisiones promedio en función de la ubicación y el número de semana, lo que implica una comprensión más profunda de las tendencias de emisión en diferentes ubicaciones y a lo largo del tiempo.
- La inclusión de variables como '**latitude**', '**longitude**', '**week_no**' y '**emission**' sugiere que estas son características clave que se consideran en el análisis de emisiones.
- La transformación de variables categóricas a tipo 'category' puede ser útil para ciertos modelos y análisis posteriores.
- El uso de algoritmos de aprendizaje automático, como **RandomForestRegressor** y otros modelos de ensamblaje, puede ser beneficioso para predecir las emisiones en función de las características procesadas.

[flaml.automl.logger: 11-09 18:14:37] {2630} INFO - retrained model:
LGBMRegressor(colsample_bytree=0.8282214365115305,
learning_rate=0.27303282394961403, max_bin=1023,
min_child_samples=10, n_estimators=1, n_jobs=-1, num_leaves=16,
reg_alpha=0.007704104902643932, reg_lambda=0.02708521372554917,
verbose=-1)
[flaml.automl.logger: 11-09 18:14:37] {1930} INFO - fit succeeded
[flaml.automl.logger: 11-09 18:14:37] {1931} INFO - Time taken to find the best model:
114.9193994998931

En el proceso de optimización automática utilizando FLAML para abordar un problema de regresión, se identificó y entrenó un modelo LightGBMRegressor. Los hiperparámetros específicos del modelo, como la tasa de aprendizaje, el número de hojas y la cantidad de árboles, fueron ajustados automáticamente por FLAML para optimizar el rendimiento en función de la métrica de error cuadrático medio (RMSE). El proceso de ajuste fue exitoso, y el mejor modelo fue encontrado en un tiempo total de aproximadamente 114.91 segundos. Este enfoque automatizado de optimización de modelos demuestra ser eficiente para encontrar configuraciones óptimas en un tiempo limitado, facilitando la implementación de modelos de regresión de alta calidad.