

Predicción de Concentraciones de CO2 en Ruanda: Un Enfoque de Machine Learning para la Gestión del Cambio Climático

1. Introducción

Este informe presenta el proyecto Kaggle: concursos para desarrollar un modelo de aprendizaje automático basado en el conjunto de datos. El objetivo principal de este proyecto es crear un modelo que pueda predecir futuras emisiones de dióxido de carbono (CO2) utilizando datos recopilados por las observaciones del satélite SENTINEL-5P. Se ha compilado un conjunto de datos que cubre alrededor de 497 ubicaciones diferentes en diferentes regiones de Ruanda, incluidas áreas urbanas, áreas agrícolas y centrales eléctricas. Este conjunto de datos contiene variables importantes como "latitud", "longitud", "año" y "número de semanas" registradas cada semana desde enero de 2019 hasta noviembre de 2022. La variable objetivo de nuestro estudio son las "emisiones", que representan las emisiones de dióxido de carbono. . de estos lugares. A continuación se muestra un enlace al conjunto de datos específico con el que trabajamos.

2. Objetivo

El objetivo de este proyecto es desarrollar un modelo de pronóstico de emisiones de CO2 para Ruanda con una precisión objetivo del 75% o mejor. Se basa en tener muchas variables importantes como "latitud", "longitud", "año", "SulfurDioxide_SO2_slant_column_number_density", "SulfurDioxide_SO2_column_number_density". Estas variables nos permiten identificar patrones de emisiones de CO2 en diferentes regiones y en diferentes momentos.

Además de ser importante para la previsión, este modelo, una vez implementado con éxito, puede proporcionar información valiosa a las agencias ambientales. Las personas, las agencias gubernamentales y las empresas pueden utilizar esta información para tomar decisiones informadas que mejoren la calidad de vida en Ruanda. Estas decisiones pueden incluir la promoción de medios de transporte respetuosos con el medio ambiente en las ciudades, la promoción de alternativas como la bicicleta y la reducción del consumo de energía tanto a nivel doméstico como industrial.

Cabe señalar que este proyecto también ofrece oportunidades de ganar dinero. Algunas de las oportunidades incluyen la capacidad de ofrecer suscripciones, asociarse con agencias gubernamentales o empresas interesadas en reducir las emisiones de carbono y brindar servicios de consultoría en sostenibilidad y neutralidad de carbono utilizando el modelo desarrollado. Estas habilidades adicionales pueden contribuir a la sostenibilidad y viabilidad a largo plazo de este proyecto.

3. Exploración y análisis del DataSet

Descripción general del conjunto de datos

Inicialmente, realizamos una revisión rápida del conjunto de datos para comprender su estructura y los valores contenidos en el mismo. Simultáneamente, identificamos las columnas y verificamos la integridad de la carga de datos. En una primera observación, notamos la presencia de valores faltantes en la fila con la identificación "ID_-0.510_29.290_2019_03".

```
[25] df.head()
```

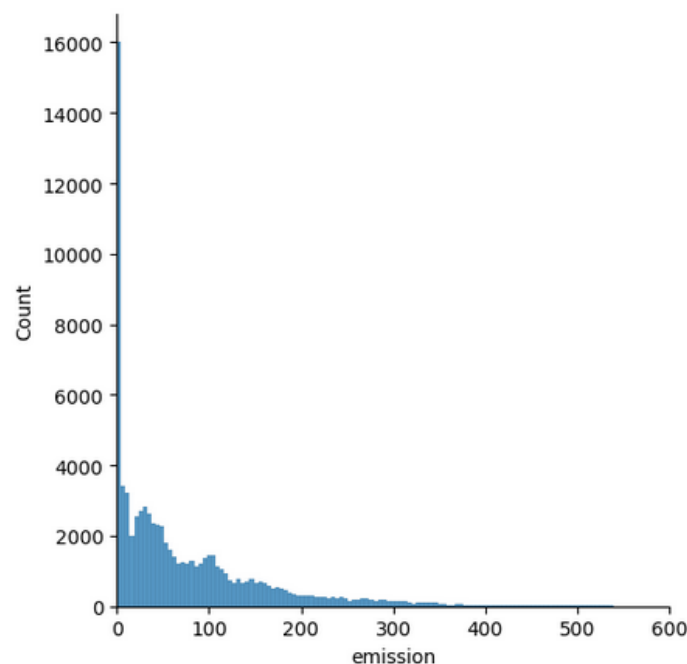
| ID_LAT Lon_YEAR_WEEK | Latitude | longitude | year | week_no | SulphurDioxide_SO2_column_number_density | SulphurDioxide_SO2_column_number_density_anf | SulphurDioxide_SO2_slant_column_number_density | SulphurDioxide_cloud_fraction |
|--------------------------|----------|-----------|------|---------|--|--|--|-------------------------------|
| ID_-0.510_29.290_2019_00 | -0.51 | 29.29 | 2019 | 0 | -0.000108 | 0.603019 | -0.000065 | 0.255668 |
| ID_-0.510_29.290_2019_01 | -0.51 | 29.29 | 2019 | 1 | 0.000021 | 0.728214 | 0.000014 | 0.130988 |
| ID_-0.510_29.290_2019_02 | -0.51 | 29.29 | 2019 | 2 | 0.000514 | 0.748199 | 0.000385 | 0.110018 |
| ID_-0.510_29.290_2019_03 | -0.51 | 29.29 | 2019 | 3 | NaN | NaN | NaN | NaN |
| ID_-0.510_29.290_2019_04 | -0.51 | 29.29 | 2019 | 4 | -0.000079 | 0.676296 | -0.000048 | 0.121164 |

5 rows x 9 columns

Por otro lado, se aplican un par de funciones para entender un poco más el DataSet para describir estadísticas y poder entender un poco más el comportamiento del DataSet.

Análisis de la variable objetivo “emission”

En primera instancia se analiza la distribución de la variable objetivo “emission” para identificar posibles sesgos en ella y así poder corregir con alguna transformación si es el caso



en la gráfica se logra observar que la distribución de la variable objetivo es asimetría ya que tiene la cola derecha muy larga por lo que se procede a calcular la asimetría “slewness” de la variable objetivo

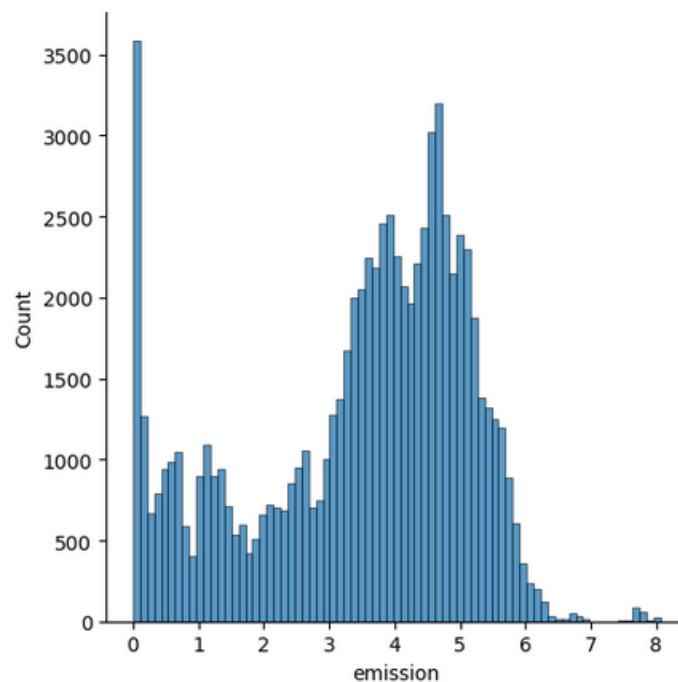
```
[27] print('La asimetira (Slewness) de la variable objetivo es:', df['emission'].skew())
```

La asimetira (Slewness) de la variable objetivo es: 10.173825825101622

- Asimetría: 10.173825825101622

la asimetría de la variable objetivo en nuestro caso de estudio es 10.173825825101622, es un valor bastante alto y positivo lo que significa que la distribución de la variable objetivo está sesgada hacia la derecha “sesgo positivo”, esto indica que, en promedio los valores de la variable objetivo son mayores que la mediana, lo que indica que en la mayoría de las observaciones tienen valores más bajos pero un pequeño porcentaje de las observaciones tienen valores extremadamente altos que influyen en la distribución de la variable objetivo.

Ya que la asimetría de la variable objetivo es un valor bastante alto es necesario aplicar una transformación para corregir dicho fenómeno en este caso aplicamos la transformación logarítmica (Logaritmo Natural), se volvió a graficar la distribución de la variable objetivo y se evidenciaron los siguientes resultados.



se observa una distribución más clara aunque aun así siendo transformada logarítmicamente se denota la cola a la derecha de la distribución original, también se observa con en la distribución original que el dato más recurrente es el 0, a demás para esta distribución transformada también le calculamos su asimetría “slewnees” y nos da el siguiente resultado

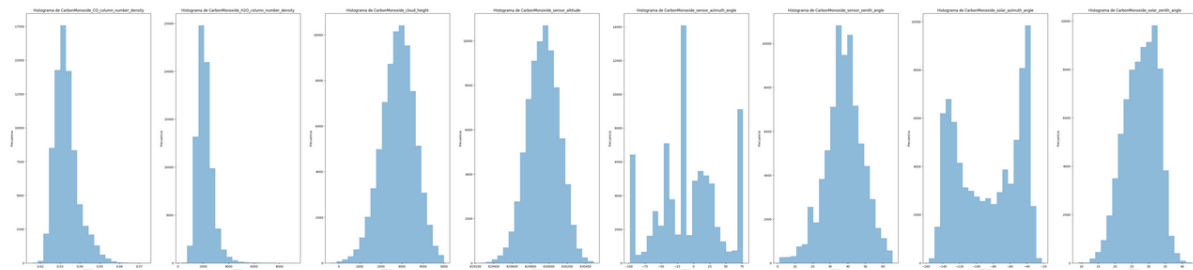
```
[45] print('La asimetira (Slewness) de la variable objetivo despues de una tranformacion logaritmica es:', df['emission'].skew())  
La asimetira (Slewness) de la variable objetivo despues de una tranformacion logaritmica es: -0.6091471436943189
```

- *Asimetría:* -0.6091471436943189

Después de aplicar la transformada logarítmica a la variable objetivo su asimetría se redujo significativamente, lo que indica que la distribución se volvió más simétrica o incluso en este caso sesgada hacia la izquierda, esta asimetría negativa indica que los valores tienden a agruparse hacia la izquierda de la distribución y que los valores extremadamente altos se atenuaron.

Exploración de variables

Se realiza la identificación del tipo de datos de cada columna para tomar decisiones acerca de cómo visualizar cada una de las columnas, al realizar la identificación nos damos cuenta que las variables que componen el DataSet son de tipo “float64” y “int64” es decir todas las variables son numéricas por lo que es conveniente visualizarlas con histogramas.



Cabe aclarar que por falta de conocimiento en el tema no fue posible identificar las variables más relevantes del DataSet por lo que se optó por graficar cada variable para conocer su distribución y dar claridad del comportamiento de cada una de ellas.

Análisis de datos faltantes

Al examinar nuestros datos, observamos que a algunas variables les faltan valores, lo cual es un aspecto importante de nuestro análisis. Profundizando en este problema, encontramos que algunas de estas variables muestran un alarmante porcentaje de datos faltantes, un 99,44% para ser exactos. Este hallazgo plantea serias preocupaciones sobre la integridad y confiabilidad de nuestros resultados.

Una cantidad significativa de datos faltantes puede afectar negativamente a nuestro análisis estadístico. Esto se debe a que los datos faltantes pueden afectar nuestras conclusiones, afectar la representatividad de la muestra y, en última instancia, reducir la precisión de los modelos o las conclusiones extraídas de estos datos incompletos.

| | Total | Percent |
|---|-------|-----------|
| UvAerosolLayerHeight_aerosol_height | 78584 | 99.444466 |
| UvAerosolLayerHeight_sensor_azimuth_angle | 78584 | 99.444466 |
| UvAerosolLayerHeight_aerosol_pressure | 78584 | 99.444466 |
| UvAerosolLayerHeight_aerosol_optical_depth | 78584 | 99.444466 |
| UvAerosolLayerHeight_sensor_zenith_angle | 78584 | 99.444466 |
| ... | ... | ... |
| latitude | 0 | 0.000000 |
| longitude | 0 | 0.000000 |
| week_no | 0 | 0.000000 |
| year | 0 | 0.000000 |
| emission | 0 | 0.000000 |

75 rows x 2 columns

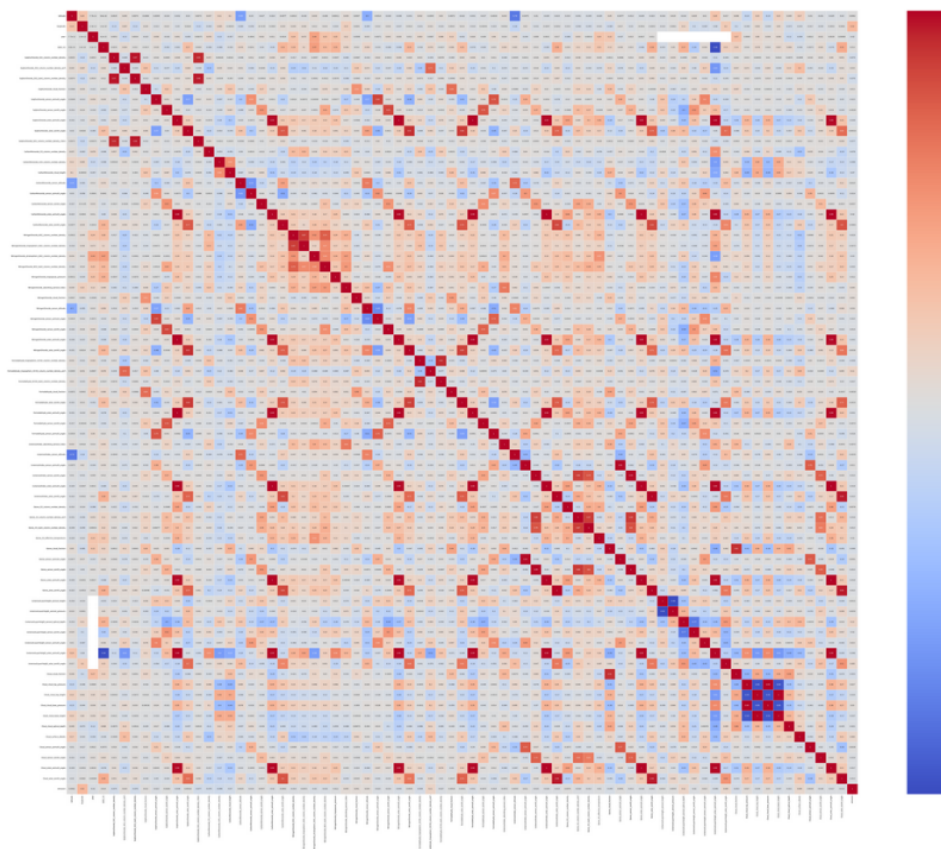
Analisis de correlacion

Al realizar el análisis de correlación nos permite identificar el tipo de correlación que existe entre las variables del DataSet ya sea una correlación negativa perfecta (Variables inversamente proporcionales) o una correlación positiva perfecta (Variables directamente proporcional) o si indica falta de correlación.

- **Matriz de correlación:**

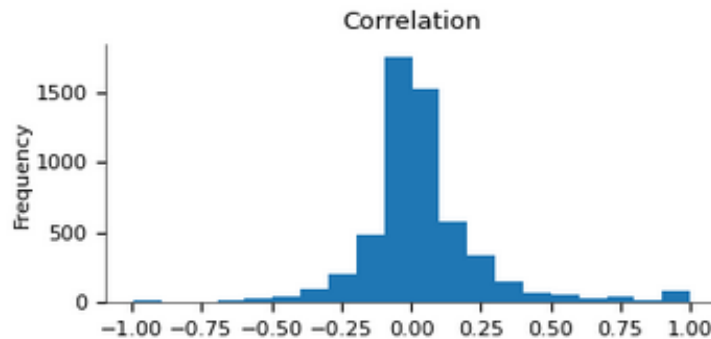
Al realizar la matriz de correlación logramos identificar varias cosas a simple vista, dejando de lado la diagonal principal ya que representa la correlación de la variable con ella misma, por otro lado en primera instancia podemos observar algunas relaciones positivas perfectas o fuertes lo que nos quiere decir que esas variables tienen un comportamiento muy similar, por otro lado también se observan algunas correlaciones negativas perfectas o débiles, además se logran observar algunos patrones curiosos en la matriz que serían un espacio de cuadrícula en la parte inferior derecha.

Algunas correlaciones para destacar podrían la variable “cloud_cloud_top_height” y “cloud_cloud_top_pressure” la cual tienen una correlación de “- 0.95%” es decir que son inversamente proporcionales y se podría interpretar como la dispersión de gases, ya que la presión y el tamaño de las nubes influyen en la dispersión y mezcla de gases en la atmósfera.



Por otro lado podemos ver que la distribución de la matriz de correlación es una distribución simétrica esto podría significar que hay presencia de independencia entre variables ya que la

falta de correlación entre las variable es lineal lo que denota que las variables son independientes entre sí, esto tiene aspectos buenos y malos como punto positivo es que facilita los análisis estadísticos y negativos es que en la vida real rara vez se tienen datos con estas características es fundamental indagar de la naturaleza de los datos.



- **Correlación con respecto a la variable objetivo “emission”**

Al analizar la correlación de las variables del DataSet con respecto a la variable objetivo “emission”, a simple vista podemos observar que las correlaciones se trata de “correlación débil” y en la mayoría de los casos se trata de una “correlación muy débil”, esto indica que la variables no están fuertemente relacionadas con la variable objetivo por lo tanto, se pueden considerar variables independientes.

Cabe destacar que la correlación no debe ser el único factor que determine la importancia de una variable, es posible que una variable con una correlación muy baja aporte información valiosa en combinación con otras variables para predecir la variable objetivo es decir que en conjunto se vuelven altamente significativas.

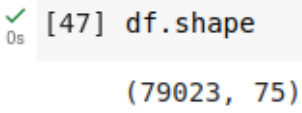
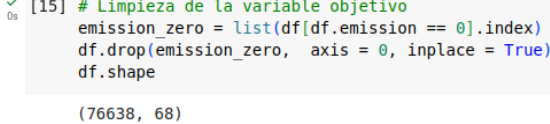
| | emission |
|--|-----------|
| emission | 1.000000 |
| longitude | 0.318397 |
| UvAerosolLayerHeight_aerosol_optical_depth | 0.118135 |
| Formaldehyde_tropospheric_HCHO_column_number_density_amf | 0.113583 |
| UvAerosolLayerHeight_solar_zenith_angle | 0.084772 |
| ... | ... |
| SulphurDioxide_SO2_column_number_density | -0.074261 |
| SulphurDioxide_SO2_slant_column_number_density | -0.075850 |
| SulphurDioxide_SO2_column_number_density_15km | -0.076736 |
| CarbonMonoxide_CO_column_number_density | -0.077760 |
| CarbonMonoxide_H2O_column_number_density | -0.097528 |

75 rows x 1 columns

4. Limpieza y reparación del DataSet

Limpieza de la variable objetivo “emission”

En la exploración de los datos, al analizar la variable objetivo se encontró que presentaba algunos datos faltantes, por lo tanto se remueven las filas que contienen dichos valores 0, ya que estos valores no indican ninguna medición y no tiene sentido utilizarlos en el entrenamiento del modelo.

| Antes | Después |
|---|---|
|  <pre>[47] df.shape (79023, 75)</pre> |  <pre>[15] # Limpieza de la variable objetivo emission_zero = list(df[df.emission == 0].index) df.drop(emission_zero, axis = 0, inplace = True) df.shape (76638, 68)</pre> |

Cómo se logra observar en las imágenes después de realizar la limpieza de la variable objetivo se eliminaron se eliminaron 2.385 filas que contienen valores 0 en la variable objetivo.

Eliminación de columnas con datos faltantes

En la exploración de los datos, al analizar la cantidad y porcentaje de datos faltantes por variable en el DataSet se identificó algunas variables con un porcentaje de datos faltantes del 99.44%, se toma la decisión de eliminar las variables que contengan una cantidad considerable de datos faltantes. Como criterio se decide que las variables que tengan un porcentaje igual o mayor al 50% se eliminará del DataSet

| Antes | | | Después | | |
|--|-------|-----------|---|-------|-----------|
| | Total | Percent | | Total | Percent |
| UvAerosolLayerHeight_aerosol_height | 78584 | 99.444466 | NitrogenDioxide_solar_azimuth_angle | 17333 | 22.616718 |
| UvAerosolLayerHeight_sensor_azimuth_angle | 78584 | 99.444466 | NitrogenDioxide_sensor_altitude | 17333 | 22.616718 |
| UvAerosolLayerHeight_aerosol_pressure | 78584 | 99.444466 | NitrogenDioxide_absorbing_aerosol_index | 17333 | 22.616718 |
| UvAerosolLayerHeight_aerosol_optical_depth | 78584 | 99.444466 | NitrogenDioxide_tropopause_pressure | 17333 | 22.616718 |
| UvAerosolLayerHeight_sensor_zenith_angle | 78584 | 99.444466 | NitrogenDioxide_NO2_slant_column_number_density | 17333 | 22.616718 |
| ... | ... | ... | ... | ... | ... |
| latitude | 0 | 0.000000 | latitude | 0 | 0.000000 |
| longitude | 0 | 0.000000 | longitude | 0 | 0.000000 |
| week_no | 0 | 0.000000 | week_no | 0 | 0.000000 |
| year | 0 | 0.000000 | year | 0 | 0.000000 |
| emission | 0 | 0.000000 | emission | 0 | 0.000000 |
| 75 rows x 2 columns | | | 68 rows x 2 columns | | |

Como se observa en las imágenes Antes de realizar la eliminación de las variables se contaba con 75 variables, después de hacer la eliminación de las variables con el 50% o más de datos

faltantes el DataSet queda con 68 variables, por lo que concluimos que se eliminaron 7 variables que presentaban altos porcentajes de datos faltantes.

Rellenado de datos faltantes

Anteriormente se eliminaron las variables con un porcentaje igual o mayor al 50% de datos faltantes pero esto no es suficiente para tener un DataSet óptimo para hacer el entrenamiento de un modelo confiable ya que aún hay muchos datos faltantes en las variables que aún hacen parte del DataSet por lo que se opta por rellenar estos datos faltantes con la “*mediana*” de cada variable.

La decisión de realizar el relleno de los datos faltantes con la “*mediana*” fue por su robustez frente a datos atípicos ya que la “*mediana*” es una medida de tendencia central que es menos sensible a los valores atípicos o extremos que la media. Esto significa que, si tienes valores extremos del DataSet, la mediana puede proporcionar una mejor estimación del valor central sin verse afectada por esos valores extremos.

| Antes | Después |
|---|---|
| <pre> 0s df.isnull().sum() latitude 0 longitude 0 year 0 week_no 0 SulphurDioxide_S02_column_number_density 14609 ... Cloud_sensor_azimuth_angle 484 Cloud_sensor_zenith_angle 484 Cloud_solar_azimuth_angle 484 Cloud_solar_zenith_angle 484 emission 0 Length: 75, dtype: int64 </pre> | <pre> 0s [18] df.isnull().sum() latitude 0 longitude 0 year 0 week_no 0 SulphurDioxide_S02_column_number_density 0 .. Cloud_sensor_azimuth_angle 0 Cloud_sensor_zenith_angle 0 Cloud_solar_azimuth_angle 0 Cloud_solar_zenith_angle 0 emission 0 Length: 68, dtype: int64 </pre> |

Como se observa en las imágenes antes de realizar el relleno de datos faltantes el DataSet contaba con Variables de 14.689 datos faltantes, después de realizar el relleno podemos observar que los datos de cada una de las variables están completos.

En conclusión realizando estos procedimientos de limpieza del DataSet contamos con mayor calidad de datos, simplificación del análisis, menos sesgo en los modelos entre otros aspectos positivos que se ganan realizando una limpieza a el DataSet

[Video Entrega 2](#)