



INTRODUCCIÓN A LA INTELIGENCIA ARTIFICIAL

Segunda entrega: Predicción de las emisiones de CO2

Presentado a:
Raul Ramos Pollan

Presentado por:
Michael Stiven Zapata Giraldo
1007285842
Estudiante

Daniel Andres Vergara de Leon
1002388181
Estudiante

Universidad De Antioquia

Facultad de Ingeniería

Medellín

2023

Nuestro objetivo con este proyecto es predecir las emisiones de CO2 en lugares sobre todo industriales, con muestras como el dióxido de azufre y monóxido de carbono. Este objetivo se inicia a partir de utilizar factores de tiempo y técnicas como la DOAS (**Differential Optical Absorption Spectroscopy**).

Lo primero que se hizo en esta etapa fue la conexión del dataset de Kaggle con el Google Collab. En Kaggle se descargó el archivo tipo JSON que contiene el usuario y la contraseña para poder empezar a trabajar en Google Collab y poder acceder a los datos. En el Collab se instaló la librería `opendatasets` y luego la importamos. Después de eso se agregó el link de la competencia a otro módulo del Collab para poder iniciar sesión con el usuario y la clave suministrados en archivo JSON que previamente se descargó. En el módulo de "os" agregamos el link de Kaggle perteneciendo al tema de Predict **CO2 Emissions in Rwanda**, y nos muestra los archivos que hay en ese directorio. Todo esto se evidencia en la **figura 1**.

```
[ ] pip install opendatasets --upgrade

[ ] import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import opendatasets as od # downloading datasets from online sources like Kaggle

[ ] dataset_url = 'https://www.kaggle.com/competitions/playground-series-s3e20/data?select=train.csv'
od.download(dataset_url)

[ ] import os
os.chdir('./playground-series-s3e20')
os.listdir()#muestra los archivos que estan en el directorio
```

Figura 1. Módulos de los datos cargados desde Kaggle.

Las características principales para hallar estas predicciones son siete:

- Sulphur Dioxide
- Carbon Monoxide
- Nitrogen Dioxide
- Formaldehyde
- UV Aerosol Index
- Ozone
- Cloud

Las subdivisiones de las anteriores características, se tratan como variables independientes para la predicción, estas subdivisiones suman en total 70 columnas La columna final corresponde a la tabla de predicción de emisión del CO2: `emission`, para 71 columnas y al inicio del dataset están 5 columnas:

1. `ID_LAT_LON_YEAR_WEEK`
2. `latitude`
3. `longitude`
4. `year`
5. `week_no`

Para un total de 76 columnas, del dataset de Kaggle. Para el modelo de predicción se evitó la columna de ID_LAT_LON_YEAR_WEEK, pues es un string y para poder utilizar la función de regresión lineal se debe trabajar con datos numéricos, en este caso se trabajó con valores tipo float64, se evidencia en la figura 3.

```
Dataset information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 79023 entries, 0 to 79022
Data columns (total 76 columns):
#   Column                                                                 Non-Null Count  Dtype
---  -
0   ID_LAT_LON_YEAR_WEEK                                                  79023 non-null object
1   latitude                                                              79023 non-null float64
2   longitude                                                             79023 non-null float64
3   year                                                                  79023 non-null int64
4   week_no                                                               79023 non-null int64
5   SulphurDioxide_SO2_column_number_density                           79023 non-null float64
6   SulphurDioxide_SO2_column_number_density_amf                      79023 non-null float64
7   SulphurDioxide_SO2_slant_column_number_density                    79023 non-null float64
8   SulphurDioxide_cloud_fraction                                       79023 non-null float64
9   SulphurDioxide_sensor_azimuth_angle                               79023 non-null float64
10  SulphurDioxide_sensor_zenith_angle                                 79023 non-null float64
11  SulphurDioxide_solar_azimuth_angle                                79023 non-null float64
12  SulphurDioxide_solar_zenith_angle                                 79023 non-null float64
13  SulphurDioxide_SO2_column_number_density_15km                     79023 non-null float64
14  CarbonMonoxide_CO_column_number_density                           79023 non-null float64
15  CarbonMonoxide_H2O_column_number_density                           79023 non-null float64
16  CarbonMonoxide_cloud_height                                         79023 non-null float64
17  CarbonMonoxide_sensor_altitude                                     79023 non-null float64
18  CarbonMonoxide_sensor_azimuth_angle                               79023 non-null float64
19  CarbonMonoxide_sensor_zenith_angle                                 79023 non-null float64
20  CarbonMonoxide_solar_azimuth_angle                                79023 non-null float64
21  CarbonMonoxide_solar_zenith_angle                                 79023 non-null float64
22  NitrogenDioxide_NO2_column_number_density                           79023 non-null float64
23  NitrogenDioxide_tropospheric_NO2_column_number_density             79023 non-null float64
24  NitrogenDioxide_stratospheric_NO2_column_number_density            79023 non-null float64
25  NitrogenDioxide_NO2_slant_column_number_density                    79023 non-null float64
26  NitrogenDioxide_tropopause_pressure                                 79023 non-null float64
27  NitrogenDioxide_absorbing_aerosol_index                             79023 non-null float64
28  NitrogenDioxide_cloud_fraction                                       79023 non-null float64
29  NitrogenDioxide_sensor_altitude                                     79023 non-null float64
```

Figura 2. Tipos de datos trabajados en el dataset de Kaggle.

Primer Modelo Predictivo (Regresión lineal)

```
modelo=LinearRegression()
modelo.fit(X=data_1[variables_independientes],y=data_1[variable_objetivo])
```

LinearRegression

LinearRegression()

Figura 3. Modelo de regresión.

La **figura 3** muestra el método de regresión lineal para la predicción de emisiones de CO₂, cuyo resultado se puede ver en la última columna de la **figura 4**. Esto es una primera propuesta de modelo, ya que se buscará otros métodos para acertar un modelo con mejores resultados y el error de medidas y comparaciones sea lo más mínimo posible.

```
data_1[['emission', 'Emission_predict']]
```

	emission	Emission_predict
0	3.750994	108.061534
1	4.025176	71.530307
2	4.231381	59.330956
3	4.305286	96.039428
4	4.347317	79.531064
...
79018	29.404171	69.712717
79019	29.186497	101.025614
79020	29.131205	102.425027
79021	28.125792	63.228106
79022	27.239302	78.245672

79023 rows x 2 columns

Figura 4. Primera propuesta del modelo de predicciones.

Root Mean Squared Error
142.25429866076868

Figura 5. Error cuadrático medio primer modelo predictivo.

En la **figura 5** se hace evidente que el error cuadrático medio se encuentra significativamente por encima del valor objetivo, lo que confirma la necesidad de explorar otro tipo de modelo predictivo con el fin de obtener resultados más precisos y satisfactorios.

Segundo Modelo Predictivo (Regresión de bosques aleatorios)

```
#Dividimos entre train set y de test set
X_train, X_test, y_train, y_test = train_test_split(X, y_true, test_size=0.2, random_state=42)

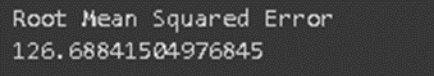
# Entrenando el Random Forest Regressor
rf_model = RandomForestRegressor(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)

RandomForestRegressor
RandomForestRegressor(random_state=42)

# Haciendo predicción en el test set
y_pred = rf_model.predict(X_test)
```

Figura 5. Modelo de Regresión de bosques aleatorios.

En la **figura 5** podemos ver el modelo de regresión de bosques aleatorios implementado en el colab de google, los resultados de este se verán en la **figura 6** haciendo uso del error cuadrático medio entre los datos medidos y los datos reales.



```
Root Mean Squared Error  
126.68841504976845
```

Figura 6. Error cuadrático medio segundo modelo predictivo.

Al observar la **figura 6**, se puede notar que el segundo modelo predictivo ha mostrado una mejora con respecto a los resultados anteriores. A pesar de este avance, aún no se ha logrado alcanzar un valor aceptable de error cuadrático medio según los datos en consideración. Por consiguiente, se continuará la búsqueda de otros modelos predictivos con el fin de mejorar los resultados de manera significativa.