

**Inteligencia Artificial**

**Informe**

**Entrega#2**



**Therry Jones Bent O'Neill**

**Julian Mateo Mena Urrego**

**Miguel Angel Rivera Florez**

**Universidad de Antioquia**

**Facultad de Ingeniería**

**Departamento de Ingeniería Eléctrica**

**Medellín**

**2023-1**

## Descripción del progreso alcanzado

### Preprocesamiento del dataset.

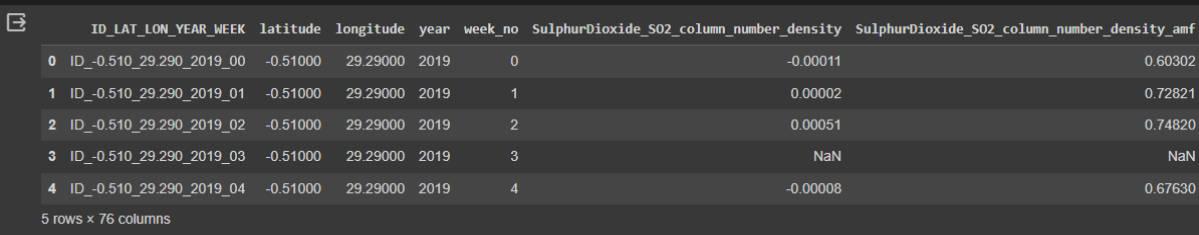
El dataset “Predict CO2 Emissions in Rwanda” está compuesto de los siguientes archivos:

- sample\_submission.csv
- test.csv
- train.csv

Antes de realizar modelos predictivos es necesario el preprocesamiento de los datos.

Preprocesar los datos se trata de cualquier tipo de procesamiento que se realiza con los datos brutos para transformarlos en datos que tengan formatos que sean más fáciles de utilizar.

Para el primer paso en el preprocesamiento se tomaron los datos presentes en el archivo “train.csv” que está compuesto de 76 columnas y 5 filas. La primera columna es el ID con un prefijo de latitud, longitud, año y semana\_no, Cada fila del train contiene cuatro columnas de índice (latitud, longitud, año y semana\_no), y las otras son 70 características y la emisión que es el objetivo.



	ID_LAT_LON_YEAR_WEEK	latitude	longitude	year	week_no	SulphurDioxide_S02_column_number_density	SulphurDioxide_S02_column_number_density_amf
0	ID_-0.510_29.290_2019_00	-0.51000	29.29000	2019	0	-0.00011	0.60302
1	ID_-0.510_29.290_2019_01	-0.51000	29.29000	2019	1	0.00002	0.72821
2	ID_-0.510_29.290_2019_02	-0.51000	29.29000	2019	2	0.00051	0.74820
3	ID_-0.510_29.290_2019_03	-0.51000	29.29000	2019	3	NaN	NaN
4	ID_-0.510_29.290_2019_04	-0.51000	29.29000	2019	4	-0.00008	0.67630

5 rows x 76 columns

**Figura 1. Composición del archivo “train.csv”**

Las 70 características vienen en 8 grupos distintos tamaños las cuales son: Dióxido de azufre, Monóxido de carbono, Dióxido de nitrógeno Formaldehído, Índice de aerosol UV, Ozono, Nube.

El grupo de características contienen diferentes sub características, pero en cada grupo aparecen cuatro sub características:

- 'sensor\_azimuth\_angle',
- 'sensor\_zenith\_angle',

- ángulo azimut solar,
- ángulo cenital solar",

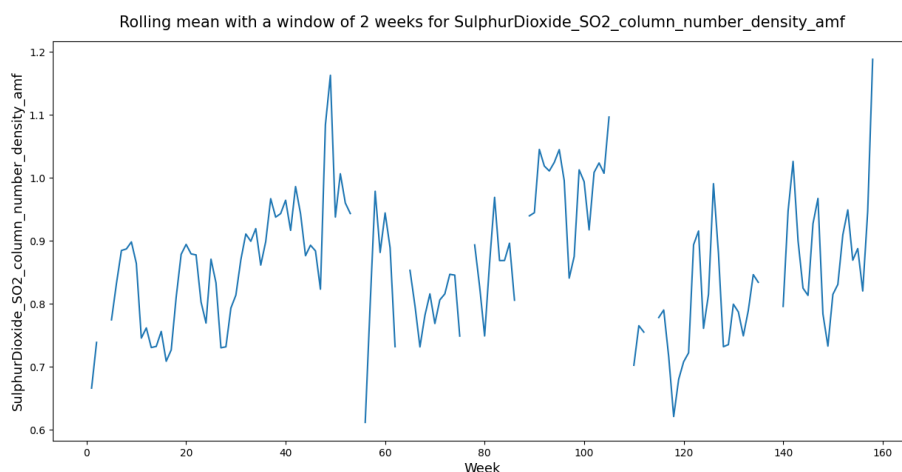
Las subcaracterísticas son necesarias para interpretar las mediciones. Como los datos fueron medidos por un satélite que no está directamente por encima del lugar medido, el ángulo del satélite afecta a la medición. Además, el ángulo del sol y las nubes influyen en las mediciones.

Cabe resaltar que el satélite mide la concentración de varios gases en la atmósfera, pero no mide el CO2. Nuestro objetivo consiste en predecir la emisión de CO2 para cada lugar y punto en el tiempo basándonos en la concentración de los demás gases. No sabemos cómo se miden las emisiones de CO2. El satélite Sentinel-5P no las mide.

Para analizar los datos del train, filtramos una ubicación y graficamos la media correspondiente a las emisiones.

```
1 # 1. Crear una ubicación única desde latitud y longitud
2 train['location'] = train['latitude'].round(2).astype(str) + '_' + train['longitude'].round(2).astype(str)
3
4 # 2. Filtrar el conjunto de datos para la ubicación deseada
5 example_loc = train[train['location'] == '-0.51_29.29']
6
7 # 3. Calcular la media móvil con una ventana de 2 semanas
8 rolling_mean = example_loc['SulphurDioxide_SO2_column_number_density_amf'].rolling(window=2).mean()
9
10 # 4. Visualizar los resultados
11 plt.figure(figsize=(15, 7))
12 rolling_mean.plot()
13 plt.title('Rolling mean with a window of 2 weeks for SulphurDioxide_SO2_column_number_density_amf', y=1.02, fontsize=15)
14 plt.xlabel('Week', y=1.05, fontsize=13)
15 plt.ylabel('SulphurDioxide_SO2_column_number_density_amf', x=1.05, fontsize=13)
16 plt.show()
```

**Figura 2. filtro de ubicación.**



**Figura 1. Gráfica de la media.**

observamos que hay datos faltantes por qué la gráfica no presenta continuidad y hay espacios entre partes de la gráfica.

## Completación de datos

Excepto las columnas de índice (latitude, longitude, year and week\_no) y emisión, en todas las columnas faltan valores, se puede asumir que faltan valores si durante toda una semana el satélite nunca pudo realizar una medición fiable de un lugar.

Como faltan datos en el train y en el test, no podemos simplemente eliminar las filas con datos que faltan, sino que necesitamos atribuir los valores que faltan por lo cual usamos las mediciones de lugares cercanos para atribuir los valores que faltan.

```
4 with pd.option_context("display.min_rows", 14):
5     # Mostrar la suma de valores faltantes ordenados
6     display(train.isna().sum().sort_values())
7
8 # Dejar un espacio en blanco entre las visualizaciones
9 print()
10
11 # Para el conjunto de prueba (test)
12 with pd.option_context("display.min_rows", 14):
13     # Mostrar la suma de valores faltantes ordenados
14     display(test.isna().sum().sort_values())
```

**Figura 2. Completación de datos faltantes.**

A el resultado de la completación de los datos que será una serie de pandas con las combinaciones únicas de latitud y longitud como índice y el valor promedio de las emisiones como datos.

```
1 #El resultado de esta operación será una serie de pandas con las combinaciones únicas de latitud
2 # y longitud como índice y el valor promedio de las emisiones como datos
3 train.groupby(['latitude', 'longitude']).emission.mean().sort_values()
```

**Figura 3. Media de datos completados.**

## Próximamente

Intentaremos construir un modelo de referencia simple que se base en la ubicación y el patrón anual, que no necesariamente se utilice mediciones satelitales ni realizar extrapolaciones. La predicción sería el promedio de las emisiones de los últimos años para la misma ubicación y semana, no se va a requerir imputación de valores. Además trataremos de mirar cómo fueron las emisiones en el segundo trimestre de 2020 debido el COVID-19 y mirar su evolución en el segundo trimestre de 2021.