PRIMERA ENTREGA PROYECTO

POR:
Samuel Gacía Bonilla
Jhon Fredy Hoyos Cardenas
Juan José Bustamante Betancur

MATERIA:
Introducción a la Inteligencia Artificial para las Ciencias e Ingenierías

PROFESOR:
Raul Ramos Pollan

UNIVERSIDAD DE ANTIOQUIA
FACULTAD DE INGENIERÍA
Medellín 2023

1. Planteamiento del problema

Los niveles de emisión de carbono son muy importantes a la hora de el estudio del calentamiento global, por lo tanto, la predicción de estas emisiones es muy importante. La precisión en estas lecturas permite a las entidades encargadas conocer diferentes tipos de información que puede ser muy relevante a la hora de realizar el análisis respectivo. Mientras que Europa y Norteamérica cuentan con amplios sistemas de seguimiento de las emisiones de carbono sobre el terreno, en África se dispone de muy pocos.

El objetivo es crear un modelo automatizado sobre los datos de emisiones de CO_2 procedentes de observaciones hechas por el satélite Sentinel-5P para predecir futuras emisiones de carbono. Estas soluciones pueden ayudar a los gobiernos y a otros actores a estimar los niveles de emisiones de carbono en África, incluso en lugares donde no es posible el seguimiento sobre el terreno. Para este problema, se utilizarán datos específicamente de Ruanda.

2. Dataset

Se hizo una búsqueda de un dataset, en la plataforma Kaggle, que reuniera los requerimientos necesarios para poder llevar a cabo el proyecto (aquí se puede encontrar la página de la competencia).

Para el dataset se seleccionaron aproximadamente 497 ubicaciones únicas de múltiples zonas de Ruanda, con una distribución en torno a tierras de cultivo, ciudades y centrales eléctricas. Los datos para esta competición se dividen por tiempo; los años 2019 - 2021 se incluyen en los datos de entrenamiento (train.csv), y el objetivo es predecir los datos de emisiones de CO_2 de 2022 a noviembre.

Se extrajeron siete características principales semanalmente del Sentinel-5P desde enero de 2019 hasta noviembre de 2022. Cada característica (dióxido de azufre, monóxido de carbono, etc.) contiene subcaracterísticas como column_number_density, qué es la densidad de columna vertical a nivel del suelo, calculada mediante la técnica DOAS. Se dan los valores de estas características en el conjunto de pruebas y se busca predecir las emisiones de CO_2utilizando información temporal, así como estas características.

- Dióxido de azufre
- Monóxido de carbono
- Dióxido de nitrógeno
- Formaldehído
- Índice de aerosol
- Ozono
- Nube

Resumiendo, se cuenta con tres archivos, uno de entrenamiento (train.csv), otro de prueba (test.csv) y un último archivo de comparación (sample_submission.csv) con el

cual se podrá conocer la exactitud de las predicciones. Ambos archivos de entrenamiento y de testeo (train.csv y test.csv) contienen la siguiente información:

- **ID_LAT_LON_YEAR_WEEK -** El id, la latitud y la longitud del lugar específico, el año y la semana en la que se presenta la emisión.
- **latitude -** Latitud del lugar de estudio.
- **longitude -** Longitud del lugar de estudio.
- **year -** Año en el que se realizó el estudio.
- **week_no -** Número de la semana
- **SulphurDioxide -** Variables con características de interés para el dióxido de sulfuro.
- **CarbonMonoxide -** Variables con características de interés para el monóxido de Carbono.
- **NitrogenDioxide -** Variables con características de interés para el Dióxido de Nitrógeno.
- **Formaldehyde -** Variables con características de interés para el Formaldehído.
- **Ozone -** Variables con características de interés para el Ozone.
- **Cloud -** Variables con características de interés para el Cloud.
- **emission -** Es el resultado de la emisión predicha por el modelo. (solo en el archivo train.csv)
- **Quartile -** Define el cuartil al cual pertenece la época del año.
- **Month -** Mes estimado de acuerdo al número de semanas.

El archivo final es el archivo sample_submission.csv el cual tendrá la siguiente información:

ID_LAT_LON_YEAR_WEEK - El id, la latitud y la longitud del lugar específico, el año y la semana en la que se presenta la emisión que identifica las características en el train.csv.
emission - Es el resultado de la emisión predicha por el modelo.

3. Métricas

La métrica de evaluación es el error medio cuadrático o root mean squared error (RMSE) y está definido como:

$$RSME = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

donde N es el número total de observaciones, $\hat{y}_i$ es el valor predecido y y_i es el valor original para cada instancia i.

El error medio cuadrático (RMSE) es una métrica que se utiliza para medir la diferencia promedio entre las predicciones de un modelo y los valores reales. Se calcula tomando la raíz cuadrada de la media de los errores al cuadrado. Cuanto menor sea el RMSE, mejor se ajusta el modelo a los datos, ya que indica que las predicciones son cercanas a los valores reales.

4. Desempeño

El objetivo de este proyecto es crear modelos de aprendizaje automático que utilicen datos de emisiones de fuente abierta (de observaciones del satélite Sentinel-5P) para predecir las emisiones de carbono.

Se tienen aproximadamente 497 ubicaciones únicas de varias áreas de Ruanda, distribuidas entre tierras agrícolas, ciudades y plantas de energía. Los datos de esta competición están divididos por tiempo; Los años 2019 - 2021 que están incluidos en los datos de entrenamiento y se busca predecir los datos de emisiones de CO2.

Se espera que las predicciones de CO_2 puedan ayudar a los gobiernos y a otros actores a estimar los niveles de emisiones de carbono en África, incluso en lugares donde no es posible el seguimiento sobre el terreno.

5. Bibliografía

- Predict CO2 emissions in Rwanda | Kaggle. (s. f.). https://www.kaggle.com/competitions/playground-series-s3e20