Inteligencia Artificial

Entrega#1

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

1) Planteamiento del problema

El problema predictivo que se busca resolver se centra en la capacidad de monitorear con precisión las emisiones de dióxido de carbono (CO2), lo cual es un paso fundamental en la lucha contra el cambio climático. La obtención de mediciones precisas de las emisiones de carbono permite a los investigadores y gobiernos comprender las fuentes y los patrones de emisión de carbono. Si bien Europa y América del Norte cuentan con sistemas extensos para monitorear las emisiones de carbono en tierra, en África existen pocas infraestructuras disponibles para llevar a cabo dicho monitoreo.

El objetivo de este desafío es crear modelos de machine learning utilizando datos de emisiones de CO2 de código abierto obtenidos a través de observaciones satelitales del satélite Sentinel-5P. Estos modelos tendrán la tarea de predecir las futuras emisiones de carbono en regiones de África. Estas soluciones pueden ayudar a los gobiernos y otros actores a estimar los niveles de emisión de carbono en todo el continente africano, incluso en lugares donde no es posible realizar un monitoreo directo en el terreno.

Este desafío tiene como objetivo contribuir a la comprensión y mitigación del cambio climático al proporcionar métodos efectivos para prever las emisiones de CO2 en regiones donde la información es limitada o inaccesible.

2) Dataset

El desarrollo se hará en base a la competencia llamada “Predict CO2 Emissions in Rwanda”(enlace: https://www.kaggle.com/competitions/playground-series-s3e20/overview) Se seleccionaron aproximadamente 497 ubicaciones únicas en diversas áreas de Ruanda, distribuidas en zonas de tierras de cultivo, ciudades y centrales eléctricas. Los datos para esta competencia están divididos en función del tiempo; se incluyen los años 2019 a 2021 en los datos de entrenamiento, y la tarea es predecir los datos de emisiones de CO2 para el año 2022 hasta noviembre.

Se extrajeron siete características principales semanalmente de las observaciones del Sentinel-5P desde enero de 2019 hasta noviembre de 2022. Cada característica (Dióxido de Azufre, Monóxido de Carbono, etc.) contiene subcaracterísticas como la densidad vertical de la columna, que es la densidad de la columna vertical a nivel del suelo, calculada mediante la técnica DOAS. Puede obtener más información sobre cada característica en los enlaces proporcionados, incluyendo cómo se miden y las definiciones de las variables. Se le proporcionan los valores de estas características en el conjunto de prueba, y su objetivo es predecir las emisiones de CO2 utilizando información temporal y estas características.

El dataset que se va a utilizar contiene 3 archivos .csv:

Train.csv: Con los datos de entrenamiento y la columna target.

Test.csv: Con el conjunto de pruebas, a las cuales se les tendrá que hallar el target para cada fila.
Sample submission.csv: Un ejemplo de cómo se debe enviar el archivo.

3) Métricas

Para el desarrollo del proyecto se utilizará el error cuadrático medio de la raíz (Root Mean Squared Error, RMSE). Comúnmente utilizada en estadísticas y en la evaluación de modelos en machine learning para medir la calidad de las predicciones o estimaciones en comparación con los valores reales. El RMSE es especialmente útil cuando estamos trabajando con problemas de regresión, donde tratamos de predecir valores numéricos continuos en lugar de clases discretas. La fórmula para calcular el RMSE es la siguiente:

$$RMSE = \sqrt{(\frac{1}{N}) * \sum_{i=1}^{N} (Y_i - \widehat{Y}_i)^2}$$

Podríamos utilizar otras métricas para trabajar la regresión, en caso tal de que el RMSE propuesto por la competencia sea complicado. La elección de la métrica depende de la naturaleza del problema y de los objetivos específicos. Es importante comprender las características de las métricas y considerar el contexto antes de seleccionar la métrica adecuada para la evaluación del modelo de regresión.

Un ejemplo de RMSE. Si obtenemos un RMSE de 5 ppm (partículas por millón), significa que, en promedio, nuestras predicciones del modelo tienen un error de aproximadamente 5 ppm en comparación con las mediciones reales de CO2. Cuanto menor sea el RMSE, mejor será el rendimiento del modelo en la predicción de las concentraciones de CO2. Este proceso nos permite evaluar cuán bien se desempeña nuestro modelo en la predicción de CO2 y nos proporciona una medida cuantitativa de la calidad de esas predicciones.

4) Primer criterio

El objetivo de este desafío es crear un modelo de aprendizaje automático que utilice los datos de emisiones de fuente abierta (de observaciones del satélite Sentinel-5P) para predecir las emisiones de carbono.
Debemos establecer un umbral de precisión, que nuestro modelo debe cumplir para considerarse como exitoso en producción. por ejemplo establecemos como objetivo un RMSE por debajo de 5 ppm (partículas por millón) de CO2, para considerar que el modelo es aceptable en producción.