SEGUNDA 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. Introducción

La generación de modelos predictivos es una parte fundamental de la inteligencia artificial (IA). Implica la creación de algoritmos y sistemas capaces de analizar datos históricos y predecir eventos futuros. Estos modelos utilizan una variedad de técnicas, como aprendizaje automático y redes neuronales, para identificar patrones y relaciones en los datos.

Los modelos predictivos de IA tienen aplicaciones en una amplia gama de campos, desde pronósticos meteorológicos y análisis financiero hasta diagnóstico médico y recomendaciones personalizadas. Son esenciales para tomar decisiones basadas en datos y anticipar tendencias, lo que permite a las organizaciones optimizar operaciones y ofrecer mejores servicios. En resumen, la generación de modelos predictivos de IA desempeña un papel crucial en la toma de decisiones informadas y en la creación de un futuro más eficiente y preciso.

En este proyecto trabajaremos con el entrenamiento modelos predictivos usando la herramienta colab. Nuestro dataset fue sacado de la competencia generada por kaggle, escogimos una que cumpliera los requisitos necesarios para el proyecto y esa fue Predecir las emisiones de C02 en Ruanda, donde hay aproximadamente 497 ubicaciones únicas de múltiples 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 están incluidos en los datos de entrenamiento y nuestra tarea es predecir los datos de emisiones de CO2 desde 2022 hasta noviembre. Se usarán diferentes librerías de python para poder obtener la mejor predicción para nuestro dataset.

2. Informe de avance del proyecto

El avance del proyecto se divide en las siguientes partes:

a. Lectura de datos

Se realiza la lectura de datos por medio de pandas a través de un repositorio en Gitlab y se usa wget para descargarlos.

b. Procesamiento de datos

Para el procesamiento de los datos se importan todas las librerías requeridas. Del dataset original se remueven siete columnas que no son necesarias para el análisis pues no tienen datos. Para cumplir el requerimiento del número mínimo de columnas categóricas se añaden dos columnas: “month_no” y “Quartile”. La columna de “month_no” calcula el mes del año a partir de la columna “week_no”. La otra columna, “Quartile”, calcula el cuartil a partir del año correspondiente de la columna “year”.

c. Limpieza de datos

Para llenar los espacios nulos del dataset se utiliza un promedio entre el valor de la celda anterior y la siguiente a la celda que tiene el valor nulo. Para las columnas de NitrogenDioxide se obtuvieron valores nulos en la primera fila

debido a que el primer valor era nulo y no tenía un valor anterior con el cual realizar el promedio..

d. Generación del modelo

El modelo planteado utiliza todas las variables de los gases provenientes del dataset para calcular las emisiones totales producidas. Se usa el algoritmo de Random Forest Regressor el cual es un modelo de aprendizaje supervisado que pertenece a la familia de modelos basados en árboles y se utiliza para problemas de regresión. La idea central detrás de Random Forest es construir múltiples árboles de decisión y combinar sus predicciones para obtener un modelo más robusto y generalizable. En este caso se utiliza este algoritmo pues se tienen un árbol de variables independientes las cuales deben ser tenidas en cuenta para el cálculo de las emisiones.

e. Entrenamiento del modelo

El modelo se entrena con los datos del archivo 'train.csv', constituyendo las variables de entrenamiento y de test del mismo, donde se usa el 30% de las columnas para el test y el 70% restante para el entrenamiento. Luego, se realiza la predicción según los datos obtenidos anteriormente. Se calcula el error absoluto de la predicción y de la emisión real para saber qué tan alejada está la predicción de la realidad. Finalmente, se añade una gráfica para visualizar la diferencia entre la predicción y la emisión real, ver Figura 1.

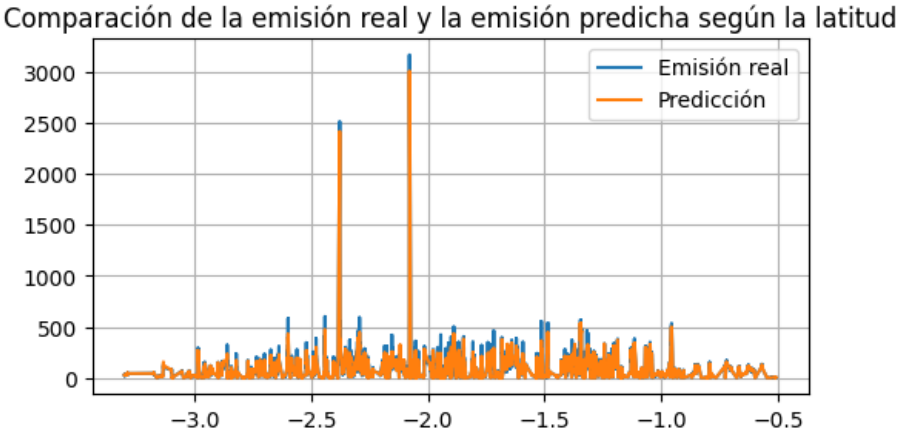

Figura 1. Comparativa de la emisión real y la emisión predicha.

f. Análisis de la métrica

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.

Se obtuvo un resultado de 21.99 en la métrica lo que nos quiere decir que, en promedio, las predicciones del modelo están a 21.99 unidades de los valores reales.

3. Conclusiones

Para la realización del modelo predictivo se empezó con la búsqueda de la mejor forma para subir los dataset que se descargaron de kaggle para la cual se analizaron varias alternativas pero al final se optó por enlazar colab por medio de un link publico a un repositorio en gitlab. Luego se prosiguió con el procesamiento de datos, limpieza de datos, generación del modelo y por último el entrenamiento de este por medio de la librería randomforest. Después de un análisis de los resultados de predicción con los datos reales se obtuvo una métrica de evaluación de 21.76 unidades. Por lo tanto podemos ver que se obtuvo un error medio cuadrático un poco alto pero se espera que para la próxima entrega contemos con un valor menor.

Bibliografía

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