PROYECTO IA - INFORME #2

Raúl Felipe Berrio – Esteban Machado – Diego Andrés Rodríguez

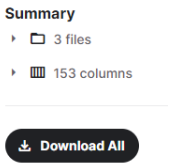
UDEA

2023

Este informe presenta los avances del proyecto que busca desarrollar un modelo de pronóstico de emisiones utilizando datos geoespaciales y temporales. Se describirán las etapas realizadas hasta ahora, como la recopilación y preprocesamiento de datos, la exploración de estos y los avances en la construcción del modelo predictivo.

DATOS

1. Descargar la base de datos de kaggle con enlace: *https://www.kaggle.com/competitions/playground-series-s3e20/data*



2. Unir los datos a drive:
    a. Copiar los archivos descargados a su Google Drive en la unidad principal. No es necesario crear carpetas adicionales ni cambiar los nombres de los archivos, ya que esto podría afectar la lectura de los datos en Google Colab.

```
# Especifica la ruta al archivo CSV en tu Google Drive
train_csv_path = ("/content/drive/MyDrive/Proyecto/train.csv")
test_csv_path = ("/content/drive/MyDrive/Proyecto/test.csv")
```

3. Correr el codigo y permitir a colab ingresar al drive

```
# Montar Google Drive en Google Colab
from google.colab import drive
drive.mount('/content/drive')
```

4. Estos archivos también se podrán descargar desde el GitHub de cada participante en caso de ser necesario.

PROGRESO ALCANZADO.

1. Recopilación y Preprocesamiento de Datos

El proyecto inició con la recopilación y preprocesamiento de datos, utilizando conjuntos de datos de entrenamiento y prueba descargados previamente de Kaggle y almacenados en Google Drive. Los datos se cargaron en DataFrames de Pandas para su análisis posterior.

2. Exploración de Datos

Se realizó una exploración inicial de los datos para comprender mejor su estructura y características. Esto incluyó la visualización de estadísticas descriptivas, la identificación de valores faltantes y la observación de tendencias temporales en las emisiones. Se observó que las emisiones tienden a aumentar en ciertas semanas de los años 2019, 2020 y 2021, lo que podría estar relacionado con eventos específicos. Además en este análisis vemos que un factor externo tal como el COVID pudo a ver afectado las emisiones del año 2020.

Se realizó un análisis de correlación que reveló las variables numéricas con mayor correlación con la variable objetivo "emisión". Sin embargo, se observó que la mayoría de las variables tienen correlaciones relativamente bajas, lo que sugiere que podría ser necesario agregar características adicionales o utilizar enfoques más sofisticados para lograr una alta precisión en la predicción de emisiones.

3. Exploración Geoespacial

Se crearon visualizaciones geoespaciales utilizando la biblioteca Folium para mostrar la distribución de las emisiones en el mapa. Esto proporcionó una perspectiva geográfica de las emisiones y resaltó áreas específicas de interés.

4. Modelado Inicial e ingeniería de las características

En el proceso de ingeniería de características, se creó una nueva columna denominada "location" en los conjuntos de datos de entrenamiento y prueba. Esta columna combina la latitud y la longitud de cada registro mediante una concatenación, la información de ubicación se crea concatenando la latitud (latitude) y la longitud (longitude) de cada fila con un guion bajo "_" como separador. Esto crea una cadena única que representa la ubicación específica de cada registro. Por ejemplo, si la latitud es 40.7129 y la longitud es 74.0060, la ubicación se representaría como "40.7129_74.0060".

Posteriormente, se convirtió la columna "location" en una variable categórica utilizando el método ".astype('category')". Finalmente, se creó un nuevo DataFrame llamado "train_agg" que contiene la ubicación, el número de semana y la emisión promedio correspondiente.

Esto se logró mediante la agrupación de los datos de entrenamiento por ubicación y número de semana, calculando la media de las emisiones para cada grupo.

El resultado es una tabla que muestra la emisión promedio para diferentes ubicaciones y semanas, lo que podría ayudar a identificar patrones o tendencias en las emisiones en función de la ubicación y el tiempo.

Si bien no se ha desarrollado un modelo predictivo completo en este informe preliminar, se tendrá en cuenta planear la utilización de técnicas de aprendizaje automático, como la regresión de bosques aleatorios, para construir un modelo de pronóstico de emisiones. y se planea continuar con los siguientes pasos:

- Abordar el problema de los valores faltantes en los datos.
- Desarrollar y evaluar modelos de pronóstico utilizando técnicas de aprendizaje automático.
- Refinar la ingeniería de características para mejorar la precisión del modelo.
- Investigar en profundidad las relaciones espaciales y temporales en los datos.

En conclusión. Se han realizado avances en la recopilación y preprocesamiento de datos, la exploración de datos y la preparación para el modelado. Aunque aún queda mucho trabajo por hacer, se espera que este proyecto proporcione un modelo sólido para predecir las emisiones en función de datos geoespaciales y temporales.