SCHNEIDER ELECTRIC HACKATHON

DATA SCIENCE

BELÉN SANTAMARÍA BALFAGÓN

RETO

La UE aporta el 18 % de las emisiones totales de gases de efecto invernadero. Sin embargo, en los últimos años se ha decidido a tomar la iniciativa en la lucha contra el cambio climático. Por este motivo, se ha fijado el objetivo de alcanzar cero emisiones de carbono en 2050. Con este fin, se han puesto en marcha una gran cantidad de recursos para ayudar a lograr este objetivo en los próximos años, y, para conseguirlo, se necesitará la ayuda de todos.

Esta es la razón de la realización de este reto, predecir el tipo de emisiones que producirán ciertas instalaciones de la unión europea a partir de indicadores como su localización, sector, tipo de actividad que desarrollan o condiciones medioambientales como el viento o la temperatura.

EXTRACCIÓN DE DATOS

En este apartado se detallará como se ha automatizado la extracción de los datos a partir de unas funciones que utilizan datos almacenados en constantes y diccionarios.

Los datos provenían de varias fuentes, en primer lugar, dos csv con distintos separadores, a continuación, tres json que había que extraer mediante peticiones a la API y por último un conjunto de pdfs almacenados en un zip.

La parte con mayor dificultad en este paso consistía en extraer la información de los pdfs de forma estructurada. Para ello, se ha extraído el texto contenido en cada documento. Estos textos extraídos no tenían la forma correcta ya que incluían espacios adicionales e incluían los títulos de las secciones del informe. Además, los nombres de los campos de la información eran diferentes a los de los csv y json. Por ello, se ha separado la cadena de texto obtenida por saltos de línea, se han eliminado todos los espacios, se ha convertido el texto en minúscula y se han eliminado los campos no deseados como los títulos de las secciones del informe. A continuación, se han sustituido los nombres y valores de las variables por los equivalentes extraídos de los datos contenidos en los csv y json.

Una vez extraídos y unificados los datos de las diferentes fuentes se han concatenado en un dataframe común para poder manipular todos los datos disponibles.

PREPROCESADO

En el preprocesamiento se ha tratado de determinar qué columnas eran innecesarias por no aportar información adicional.

En primer lugar, se han comparado las variables que se muestran a continuación:

Variable 1	**Variable 2**
facilityName	FacilityInspireID
EPRTRAnnexIMainActivityLabel	EPRTRAnnexIMainActivityCode
eprtrSectorName	EPRTRSectorCode
City	CITY ID

Tras comparar los valores que toman se ha determinado que contienen la misma información, por lo que se ha eliminado una de cada una de ellas.

Además, se han rellenado los valores nulos que contenía la columna EPRTRAnnexIMainActivityLabel.

A continuación, se han observado los valores únicos que toma cada columna. Como las columnas targetRelease y CONTINENT tan solo contenían un valor se han eliminado.

Por último, se han combinado las columnas reportingYear, MONTH y DAY en una única variable que almacena la fecha completa.

MODELO

Para entrenar el modelo, en primer lugar, se han normalizado los datos convirtiendo las variables categóricas en numéricas y aplicando MinMaxScaler a todas las variables.

El modelo elegido se compone de un bloque que se repite tres veces. Este bloque se compone por una capa convolucional con 64 filtros y un tamaño de kernel 3, seguida de una normalización y una capa relu. Estos bloques permiten extraer las características relevantes de los datos de entrada. Finalmente se aplica una capa densa para realizar la tarea de clasificación de la emisión de gases.

PREDICCIONES

Para realizar las predicciones se ha realizado el mismo preprocesamiento de los datos descrito en los apartados anteriores, eliminando las columnas necesarias y normalizando los datos.