Predicción del tipo de polución

Luis Rodriguez
Imunky

CARGA Y LIMPIEZA DE DATOS

- El lenguaje escogido para la realización del reto fue python.
- Procedimos a cargar la data de los archivos en formatos json, pdf y csv utilizando las librerías pandas y PyPDF2 de python.
- Se hizo un one-hot-encoding para tratar las variables categóricas.
- Descartamos las columnas ‘facilityName’, ‘CITY_ID’, ‘targetRelease’, ‘CONTINENT’ y ‘’FacilityInspireID’ debido a que eran categóricas y a la gran cantidad de valores diferentes que tenían.
- Se realizó una normalización de la data para tener la misma desviación típica y media en todas las columnas.

MODELO PREDICTIVO

- El modelo escogido para el entrenamiento de la data fue randomforest
- Utilizamos el 80% de la data para el entrenamiento del modelo y el 20% para el testeo.
- Escogimos los parámetros más óptimos del randomforest haciendo un GridSearch a la data.
- Medimos la precisión de la predicción del modelo utilizando el f1-score y obtuvimos 0.39 de precisión en el data de entrenamiento.