Schneider Electric Hackhaton

Data Science

Members: Ivan Luque Garcia y Joan Ignasi Martí Franco

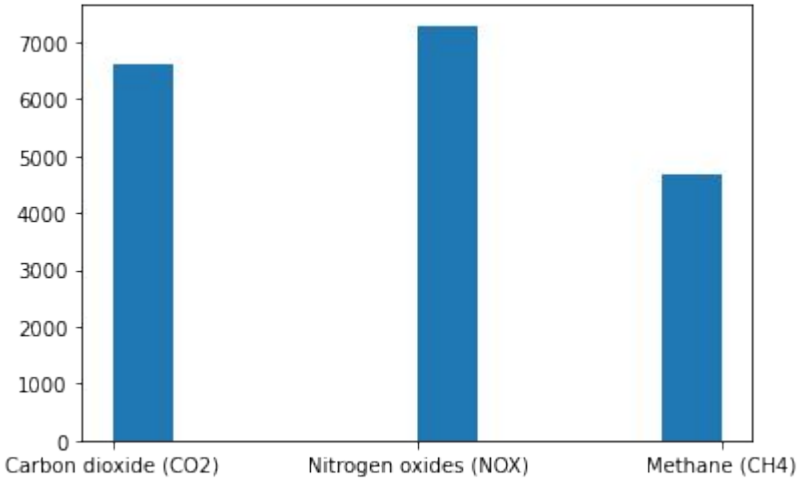
Extracción de datos

- Primero se empezó pasando a pandas los los los csv, teniendo que hacer una sustitución de los punto y coma en el segundo.
- Después se extrajeron los json.
- El siguiente paso fue extraer los datos del zip y luego de los pdfs, que se logró pero como los datos había que procesarlos mucho acabamos no teniendo tiempo y no se usaron.
- Finalmente se concatenaron los pandas de los csv y los json, se les aplicó un encoding y se guardaron en un fichero conjuntamente.

Visualización de los datos

- Primero se visualizó una muestra de los datos para ver en qué estado estaban, y se pudo ver que había algunas filas con algún campo nulo.

- Y al hacer una gráfica de los outputs se pudo ver que los datos estaban un poco desbalanceados.



Procesado y modelo

- Para intentar reducir el número de variables utilizadas se utilizó la pca, con eso se consiguió pasar de 20 features a 13, lo que permitió trabajar mejor con los modelos y facilitó los cálculos de los modelos al tener menos variables con las que lidiar.

- Uno de los modelos que se intentó utilizar fue una cnn, pero los resultados no acaban de dar bien y al final se optó por pulir el otro modelo que estábamos utilizando .

Procesado y modelo

- El modelo que se decidió probar fue el random forest, que también fue el primero que se utilizó para el reto, primero de todo se le quitaron las filas con valores null, después en la separación de train y test se hizo 93,5% train 6,5% test, entonces se le aplicó oversampling y finalmente se ejecutó el modelo, obteniendo 71% en precision, recall y f1-score.