Wassila SEGHIER
***M2 - SID* Année 2021-2022**

Découverte et mise en oeuvre de Spark

Dans ce projet d'initialisation au Spark, j'ai commencé par créer un compte Databricks , puis j'ai fait l'ingestion de données.

Le dataset utilisé est sous forme d'un fichier CSV qui regroupe des données sur les déplacements à vélos, vous trouverez le fichier ici :

https://s3.amazonaws.com/baywheels-data/index.html

j'ai choisi par hasard le fichier 2017_fordgobike_tripdata.csv.

Après la phase d'importation des données, j'ai pu analyser celles-ci avec quelques requêtes SQL avec SparkSQL.

Par exemple la longitude des station et la latitude selon le nom de la station, j'ai compter le nombre de stations sur la base de données qui est de 272 et le type d'utilisateur client et abonné.

Puis j'ai attaqué la partie Machine learning, où j'ai construit un modèle pyspark pour détecter si l'utilisateur est un client ou un abonné. Il a fallu d'abord équilibrer les données pour obtenir un bon modèle et cela en réduisant le nombre de Suscriber.

J'ai utilisé les Random Forest comme algorithme d'apprentissage (supervisé), j'ai divisé le corpus en un ensemble d'apprentissage et un ensemble de test, et cela après une phase de transformation des données pour cela j'ai utilisé un vecteur assembleur pour combiner les caractéristique que j'ai jugé utiles pour l'apprentissage, je n'ai gardé que les caractéristiques numériques, puis j'ai utilisé un string Indexer pour coder les étiquettes.

Ensuite j'ai entrainé mon modèle, le Random Forest à obtenu environ 73% d'accuracy, j'ai à la fin visualisé la matrice de confusion du modèle.