République Algérienne Démocratique & Populaire
Ministère de l'Enseignement Supérieur & de la Recherche Scientifique
Ecole Supérieure en Informatique 08 mai 1945 de Sidi Bel Abbes

Année universitaire : 2023-2024 **Semestre 2**

2 SC : IASD **Module : Big Data Technologies**

TP 4 : Spark MLlib

Sous Google Colab et en utilisant la bibliothèque MLlib :

1) Stocker le contenu de tous les fichiers du dossier **tp4_data** dans un DataFrame.
2) Afficher le schéma du Dataframe obtenu.
3) Remplir les valeurs manquantes (NaN) avec la valeur 0.
4) Ajouter une nouvelle colonne nommée "day_of_week". La valeur de la colonne est le jour de la semaine correspondant à la date de chaque ligne dans la colonne "InvoiceDate".
5) Diviser les données en un ensemble d'apprentissage et un ensemble de test. Effectuer la division en se basant sur l'attribut **InvoiceDate** : l'ensemble d'apprentissage contient les achats effectués avant 2010-12-13 et l'ensemble de test contient les achats effectués durant ou après 2010-12-13.
6) Créer un **StringIndexer** qui permet de transformer les jours de semaine présents dans la colonne **day_of_week** en valeurs numériques correspondantes.
7) En utilisant le **StringIndexer**, Spark peut par exemple représenter samedi par 6 et lundi par 1. Cependant, avec ce schéma de numérotation, nous indiquons implicitement que samedi est supérieur à lundi (par des valeurs numériques pures). Ceci est évidemment incorrect. Comment résoudre ce problème ?
8) Créer un **VectorAssembler** contenant trois attributs : UnitPrice, Quantity, et day_of_week_encoded.

   Remarque : day_of_week_encoded est le résultat de la question 7.

9) Créer un pipeline configuré avec les résultats des étapes 6, 7 et 8.
10) Notre StringIndexer doit savoir combien de valeurs uniques il y a à indexer, comment résoudre ce problème ?
11) Transformer les données de l'ensemble d'apprentissage en se basant sur les étapes (stages) du pipeline.
12) Créer une instance de KMeans, on supposse que le nombre de clusters est 20.
13) Lancer l'apprentissage de KMeans en se basant sur l'ensemble obtenu dans l'étape 11.
14) Effectuer des prédictions sur l'ensemble de test.
15) Calculer le coefficient de silhouette.