République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur
et de la Recherche Scientifique
ECOLE SUPÉRIEURE EN INFORMATIQUE
8 Mai 1945 - Sidi-Bel-Abbès

الجمهورية الجزائرية الديمقراطية الشعبية
وزارة التعليم العالي والبحث العلمي
المدرسة العليا للإعلام الآلي
8 ماي 1945 - سيدي بلعباس

Examen TP : Big Data

1. Partie 1 : Installation

a. Installer : ***jdk-8, spark-cluster, findspark, pyspark***

b. Importez [***findspark, pyspark***] et initialisez ***SparkContext***

2. Partie 2 : SPARK SQL

a. Charger les DataSets ***dataset1.csv*** et ***dataset2.csv*** dans deux dataFrames ***d1_df*** et ***d2_df***, respectivement.

b. Afficher les deux dataFrames.

c. Nettoyer les deux dataFrames en gardant uniquement ceux ils n'ont pas d'attributs manquants.

d. Afficher des statistiques de l'attribut ***BMD*** **(***bone mineral density***)** du deuxième dataFrame ***d2_df.***

e. Afficher les différentes valeurs distinctes pour les attributs « ***sex, medication»*** du dataFrame ***d1_df.***

f. Transformer les valeurs des attributs « ***sex, medication»*** en valeurs numériques.

3. Partie 3 : SPARK MLLIB

On souhaite créer un modèle de régression qui prédit le «***BMD*** » des patients en se basant sur les features :[**age, sex, medication, BMI**], où le **BMI**= [**weight**/(**height**)**2] , tel que, le ***weight*** en kg et ***height*** en mètres

a. A partir des deux dataFrames ***d1_df et d2_df*** créer le dataset correspondant.

b. Spliter le dataset en train_df et test_df [80% , 20%]

c. Trainer le modèle

d. Afficher les coefficients et les évaluations (***erreurs***)

e. Prédire le BMD de certains nouveaux patients