Université de Mostaganem
M2, IA4IOT, 2024-2025
Module Gestion des données massives

Projet (MongoDB + Apache Spark)

Notes Importantes :

1. Le Projet doit être remis au plus tard le **13/12/2024** à 23 :59.
2. Le Projet doit être déposé sur la plateforme sous la forme d'un ***dossier archive*** (.zip, .rar,…)
3. Le nom du dossier doit être de la forme ***Nométudinant1_ Nométudinant2*** (dans le cas des binômes) ou ***Nométudinant*** (dans le cas d'un monôme).
4. Le dossier doit contenir : ***le projet de la base de données + le projet Apache Spark.***
5. Un projet non consulté est considéré comme non fait et la note sera par conséquent **zéro**.

Enoncé du projet :

Soit le fichier **`healthcare-dataset-stroke-data.csv`** contenant l'enregistrement des états de santé d'un ensemble de patients d'un établissement de santé. Le fichier contient 5110 lignes pour 11 colonnes. On ne dénombre pas de données manquantes.

Partie A : Conception de la BDD et requêtes

Dans MongoDB, réaliser les opérations suivantes :

1. Créer la base de données Dossier_Patients.
2. Créer et remplir l'enregistrement de chaque patient dans la BDD sachant que la taille d'une collection ne doit pas dépasser 4 000 000 bytes et le nombre de documents maximum est 5111 sachant que la BDD ne doit pas tolérer la suppression des documents.
3. Créer un indexe dans l'ordre croissant sur l'âge du patient et un indexe dans l'ordre décroissant sur la valeur du bmi (Body mass index).
4. Exécuter les requêtes ci-dessous :
    a. Affichage de tous les enregistrements.
    b. Affichage de la Pression Sanguine, et Sucre dans le sang de tous les hommes adultes qui fument.
    c. Affichage des 10 premiers états de santé de patients qui souffrent de Problème cardiaque et qui vivent dans un lieu urban.
    d. Affichage des 20 prochains états de santé après avoir ignoré les 5 premiers patients diabétiques.
    e. Trouver les états de santé des patients mariés et qui ne sont pas des travailleurs indépendants et qui ont un bmi (Body mass index) supérieur à 20 mais inférieur à 40.
    f. Trouver les identificateurs des états de santé des patients qui ne vivent pas dans des lieu Rurale et qui ne fument pas. Le résultat doit afficher les états de santé des patients du plus jeune au plus vieux.
    g. Ajouter l'état de santé suivants : 44688, Female, 44, 0, 0, Yes, Govt_job, Urban, 85.28, 26.2, Unknown, 0.
    h. Trouver les identificateurs des états de santé des patients qui ont passé par un accident vasculaire cérébral (stroke) et qui souffrent d'un seul problème : d'une hypertension ou d'un diabète.

Partie B : distribution Spark de CNN

L'objectif de cette partie est la prédiction d'AVC (accident cardio-vasculaire « stroke ») à l'aide d'un réseau de neurones (CNN) en utilisant la distribution d'Apache Spark. Pour ce faire, il faut accomplir les opérations ci-dessous en utilisant les opérations RDD citées dans la fiche TP4.

1. Importer les données.
2. Suppression des données manquantes.
3. Suppression de la colonne id.
4. Encoder les variables non numériques (un encodage par label).
5. Suppressions des données manquantes.
6. Entrainer le réseau (extraire les features et les labels).
7. Séparer les données d'apprentissage train_set et les données de test test_set.
8. Afficher les répartitions des labels entre le data_train et data_test.
9. Création du réseau (crée le modèle avec les paramètres comme la fonctions de perte ou la métrique utilisées... .)
10. Lancer l'analyse de la matrice de confusion.