ÉCOLE NATIONALE SUPÉRIEURE D'ARTS ET MÉTIERS
UNIVERSITÉ HASSAN II DE CASABLANCA

Mini-projet Big Data : Assurance

Dataset : assurance.csv

Travail à faire :

- Analyse générale (structure de dataset ; nombe de lignes…)

En utilisant Spark SQL créer les requêtes suivantes :

- Calculer le nombre de femmes assurées
- Calculer le nombre de femmes ayant un véhicule âgé de plus de 2 ans **(`> 2 Years`)**
- Calculer la moyenne des **Annual_Premium** pour les hommes ayant des véhicules endommagés
- Créer un dataFrame contenant les colonnes suivantes : **(id , Gender, Age , Vehicle_Age , Annual_Premium , Response)**
- Enregistrer le nouveau dataframe au format json

Prétraitement :

- Créer une fonction udf permettant de convertir les valeurs de la colonne **Vehicle_Damage** en 1 et 0
- Créer une fonction udf pour encoder la colonne **Vehicle_Age** en valeurs numérique
- Créer une fonction udf pour encoder la colonne **Gender** en valeurs numé rique

Classification

- Créer un modèle de classification en utilisant les algorithmes suivants :
  - Régression Logistique
  - Naïve Bayes
  - SVM
- Faites une comparaison des résultats obtenus par ces trois algorithmes