Projet AOSI – FDD

Introduction

Le but de ce projet est de vous initier d'une part à l'architecture RMI (Remote Method Invocation) qui permet d'appeler, exécuter et renvoyer le résultat d'une méthode exécutée dans une machine différente de celle de l'objet l'appelant. Cette machine peut être ou virtuelle ou une machine différente et doit être accessible via un réseau.

D'autre part, vous allez vous intéresser à la fouille de données en réalisant un prétraitement d'un dataset en vue de le préparer pour mettre en place un modèle capable de prédire/classer efficacement de nouveaux enregistrements (nouvelles données). Pour se faire, il est nécessaire de bien comprendre le dataset en réalisant une analyse statistique entre-autres.

Fouille de données

Durant cette partie, vous devez examiner un recensement de revenus réalisé aux États-Unis en 1994. Le but étant de prédire si le revenu annuel d'un individu sera plus que 50.000$ en se basant sur plusieurs attributs. Ce dataset est composé de 14 variables prédictives (colonnes) et d'une variable à prédire (classe) ainsi que de plus de 32000 enregistrements. Le dataset peut être téléchargé à partir de l'adresse https://archive.ics.uci.edu/ml/datasets/Adult. Le travail demandé devra être réalisé en utilisant le langage Python. Vous devez aussi rédiger un mini rapport pour décrire et expliquer les différentes étapes réalisées ainsi qu'une présentation où vous présenter brièvement les points importants de votre travail.

Exploration du dataset

1. Décrire/Expliquer les attributs (colonnes) du dataset, le type de données et valeurs possibles (ensemble de valeurs ou l'intervalle).
2. Donnez les maximum, minimum, moyenne et écart type des attributs numériques.
3. Donner le nombre d'individus qui gagnent plus de 50.000$ et ceux qui gagnent moins ainsi que leurs pourcentages.
4. Donner le nombre d'hommes et de femmes.
5. Donner le nombre d'hommes qui gagnent plus de 50.000$ et ceux qui gagnent moins ainsi que leurs pourcentages.
6. Donner le nombre de femmes qui gagnent plus de 50.000$ et ceux qui gagnent moins ainsi que leurs pourcentages.
7. Créer un nuage de points montrant la distribution de l'âge des enregistrements (âge vs nombre d'entrées). Quelles conclusions pouvez-vous en tirer ?
8. Créer un graphique de barres empilées présentant le nombre d'individus qui plus et moins de 50.000$ par tranches d'âge (par exemple nombre d'individus âgés de 20 à 30 ans qui gagnent plus ou moins de 50.000$, ceux qui sont âgés de 31 à 40 ans, …).
9. Refaire les questions 6 et 7, en vous intéressant au niveau d'études au lieu de l'âge.
10. Refaire les questions 6 et 7, en vous intéressant au statut de l'employé (self-employed, private, federal-gov, …) au lieu de l'âge.
11. Refaire les questions 6 et 7, en vous intéressant au poste occupé (adm-clerical, sales, craft-repair, …) au lieu de l'âge.
12. Créer un graphique de barres empilées présentant la relation entre le nombre d'heures travaillées et le revenu.
13. Créer un graphique de barres empilées présentant la relation entre la race et le revenu.

14. Créer un graphique de barres empilées présentant la relation entre le sexe et le revenu.

Prétraitement du dataset

Dans cette étape, vous devez préparer le dataset afin de pouvoir l'utiliser lors de l'étape suivante afin de mettre en place un modèle capable de prédire/classer de nouveaux enregistrements. Le prétraitement comporte plusieurs étapes parmi lesquelles :

1. Sélection d'attributs : cette étape consiste à identifier les attributs utiles et d'éliminer ceux qui ne comportent pas d'informations importantes
2. Réduction de dimension : dans certains cas le nombre de colonnes du dataset est important et l'objectif et de tenter de trouver une corrélation entre les attributs et de les fusionner ou les éliminer, …,
3. Données manquantes/erronées : dans cette sous-étape on tente de retrouver les données manquantes, erronées/aberrantes et de soit les supprimer ou les corriger,
4. Transformation des données : en fonction de la méthode de classification à utiliser, il est nécessaire de convertir les valeurs (attributs) nominales en valeurs numériques ou vice-versa,
5. Normalisation/Standardisation des données : lorsque les intervalles des attributs sont différents, une normalisation/standardisation est requise afin d'éviter que l'algorithme d'apprentissage utilisé ne favorise un attribut aux dépends d'autres,
6. Séparation des données : afin d'évaluer le modèle créé, une étape d'évaluation est effectuée à base de données destinées pour cette tâche. Ainsi, le dataset est séparé en deux groupes, un qui est utilisé pour créer le modèle et l'autre pour le tester.

A base de définitions données précédemment, vous devez déterminer quelles sont les étapes requises pour préparer le dataset, tout en justifiant ce besoin et en décrivant les tâches réalisées. Note qu'une suppression pure et simple des données manquantes et/ou aberrantes ne sera pas prise en considération lors de l'évaluation.

Prédiction et classification

Maintenant que le dataset est prêt, vous devez utiliser des méthodes de classification afin de mettre en place un modèle qui permette de classer de nouveaux individus.

1. Considérez les algorithmes suivants :
   - Régression logistique
   - Naïve Bayes
   - Arbres de décision
   - K plus proches voisin
2. Utilisez le sous-ensemble réservé pour les tests et donnez le taux de précision pour chacun des classifieurs.
3. Calculez et affichez la heatmap de la matrice de confusion