Nom : BELKACEM

Prénom : Liza

N étudiant : 19008250

Pour la réalisation de ce projet j'ai travaillé sur deux sujets différents afin de mieux maitriser la fouille de données.

Sujet 01:

Objectif : Détecter le cancer du sein à partir de données.
Cette analyse vise à observer quelles caractéristiques sont les plus utiles pour prédire le cancer malin ou bénin et à voir les tendances générales qui peuvent nous aider dans la sélection des modèles et la sélection des hyperparamètres.
L'objectif est de classer si le cancer du sein est bénin ou malin. Pour y parvenir, j'ai utilisé des méthodes de classification d'apprentissage automatique pour adapter une fonction qui peut prédire la classe discrète de nouvelles entrées.

Problématique :
Le cancer du sein est un cancer courant et meurtrier chez les femmes du monde entier car il est souvent diagnostiqué en retard, et la détection précoce de ce dernier peut améliorer considérablement le pronostic et les chances de survie en favorisant le traitement clinique auprès des patientes tôt
il est alors important pour moi de travailler sur ce sujet.

Input: les données d'entré sont un fichier csv récolté sur kaggle; contenant 569 lignes et 33 colonnes.
569 lignes de données,signifiant qu'il y a 569 patients
33 colonnes; signifiant qu'il y a 33 caractéristiques ou points de données pour chaque patient.

output: le modèle qui a obtenu les meilleurs résultats sur les données de test étaient le classificateur de forêt aléatoire avec un score de précision d'environ 96,5%
le but étant de faire la prédiction / classification sur les données de test et de montrer à la fois la classification / prédiction du modèle Random Forest Classifier et les valeurs réelles du patient qui montrent plutôt ou non qu'il a un cancer.

Stratégie : Importation des librairies et packages
Chargement des données
Exploration des données
Traitement des données
Création d'un comptage de patient ayant des cellules begnignes et malignes
visualisation des comptages

Visualisation du type de colonnes
Modification des valeur de la colonne diagnostic
Création d'un nuage à point (une variable de la même ligne de données correspond à la valeur d'une autre variable.)
Impression du nouvel ensemble de données
Visualisation de la corréaltion
Configuration des données (données traitement et test)
Standardisation des valeur avec standardscaler
Régression logistique
Classificateur d'arbre de décision
Classificateur de forêt aléatoire
Création d'un modèle conteant tous ces modèles
Création de la matrice de confusion et la précision

Sujet 2

Objectif : prédire les causes d'accidents de la route

Problématique :
Les accidents de la route au RU sont assez réccurents et très meurtrier c'st pourquoi il est important de diagnostiquer les causes principales de c phénomère afin de le réduire au maximum.

Input : Les données d'entré sont trois fichiers csv récolté sur kaggle sur les accidents de la route au Royaume-Uni allant de 2014 à 2016
https://data.gov.uk/dataset/efe5505-941f-45bf-b576-4c1e09b579a1/road-traffic-accidents

Output
Le but étant de faire la prédiction sur les données de test et de montrer les causes principales des accidents de la route au Royaume unis
Les différents modèles sont arrivé à conclure que Class_Pedestrian
Road Surface_Dry
Road Surface_Wet ou Damp.
sont les causes de ce phénomène

Stratégie:
Importer les librairies et packages
Prétraitement de données
Lecture des fichiers
Fusion de fichier
Suppression de colonnes inutiles ou manquantes
Lister les objets valeur des colonnes
Convertir les fonction de type
Division du jeu de données en eux x et y
Standardisation des données avec standardscaler
Choisir le nombre de composants pour PCA (ceux qui contiennt 90% de la variance on a trouvé 12)
Arbre de decison
Recherche de la profondeur de l'arbre qui renvoie la meilleure Précision du modèle
Définir le modèle avec max_depth = 6
Cross validation
Trouver les fonctionnalité les plus
Affichage de l'arbre
Random forest
Neural network
Logistic regression matrice de correlation
Redefinition du modèle
Cross validation

Affichage des colonnes avec leurs coefficients respectif

Conclusion

Nous pouvons conclure que les trois caractéristiques les plus importantes qui affectent la gravité d'un accident automobile sont:
1/ Class_Pedestrian
2/ Road Surface_Dry
3/ Road Surface_Wet ou Damp.