Rapport de Projet : Détection Précoce du Cancer du Poumon

**1. Définition de l'objectif du projet**

Objectif

L'objectif de ce projet est de développer un programme complet utilisant des techniques d'apprentissage automatique et de traitement de données pour la détection précoce du cancer du poumon. En appliquant plusieurs pipelines de classification sur un dataset spécifique, nous visons à comparer les performances de différents algorithmes afin d'identifier le plus efficace pour cette tâche critique.

**Problématique**

La détection précoce du cancer du poumon est cruciale pour augmenter les chances de survie des patients. Cependant, en raison de la nature souvent asymptomatique des stades initiaux, il est difficile de diagnostiquer la maladie à temps. En utilisant des données cliniques et des techniques d'apprentissage automatique, nous pouvons potentiellement identifier des modèles qui signalent la présence de cancer à un stade précoce.

**2. Source des données**

Dataset :

Le dataset utilisé pour ce projet est surveylungcancer.csv, qui contient les informations suivantes :

Total d'attributs : 16

Nombre d'instances : 284

Attributs :

Gender: M(male), F(female)

Age: Age of the patient

Smoking: YES=2, NO=1

Yellow fingers: YES=2, NO=1

Anxiety: YES=2, NO=1

Peer\_pressure: YES=2, NO=1

Chronic Disease: YES=2, NO=1

Fatigue: YES=2, NO=1

Allergy: YES=2, NO=1

Wheezing: YES=2, NO=1

Alcohol: YES=2, NO=1

Coughing: YES=2, NO=1

Shortness of Breath: YES=2, NO=1

Swallowing Difficulty: YES=2, NO=1

Chest pain: YES=2, NO=1

Lung Cancer: YES, NO

**3. Analyse exploratoire des données (EDA)**

Statistiques Descriptives :

Genre : Répartition presque égale entre les hommes et les femmes.

Âge : La distribution de l'âge montre une majorité de patients dans la tranche d'âge de 40-70 ans.

Habitudes : Une majorité des patients ont des habitudes telles que fumer, boire de l'alcool et des symptômes comme la toux, l'essoufflement, etc.

Visualisation :

Des graphiques tels que des histogrammes, des diagrammes de dispersion et des matrices de corrélation ont été utilisés pour comprendre la distribution des données et les relations entre les variables.

Détection des Outliers

Les valeurs aberrantes ont été identifiées et traitées pour éviter leur impact négatif sur les performances des modèles.

**4. Prétraitement des données**

Étapes :

Nettoyage des Données : Traiter les valeurs manquantes, les doublons et les incohérences.

Encodage des Variables Catégorielles : Utiliser des techniques d'encodage telles que One-Hot Encoding pour la variable Gender.

Normalisation et Standardisation : Appliquer des transformations pour assurer que toutes les variables ont une échelle similaire.

Division des Données : Séparer les données en ensembles d'entraînement et de test (80%/20%).

Justification

Valeurs Manquantes : Utilisation de l'imputation pour gérer les valeurs manquantes afin d'éviter les biais.

Encodage : Nécessaire car les algorithmes d'apprentissage automatique ne peuvent pas traiter les variables catégorielles directement.

Normalisation/Standardisation : Pour une convergence plus rapide et plus fiable des algorithmes.

**5. Algorithmes Utilisés**

Algorithmes de Classification :

Nous avons implémenté plusieurs algorithmes de classification :

Régression Logistique : Simple et efficace pour les problèmes de classification binaire.

Forêt Aléatoire (Random Forest) : Puissant pour la gestion des données complexes et des relations non linéaires.

Support Vector Machines (SVM) : Efficace pour les espaces de grande dimension.

k-Nearest Neighbors (k-NN) : Simple à comprendre et à implémenter.

Réseaux de Neurones (Deep Learning) : Capables de capturer des relations complexes dans les données.

Justification des Choix :

Chaque algorithme a ses propres avantages et inconvénients. Par exemple, la régression logistique est rapide et interprétable, tandis que les forêts aléatoires et les réseaux de neurones peuvent capturer des relations plus complexes mais nécessitent plus de temps de calcul et de ressources.

**6. Grille d'Évaluation des Algorithmes**

Métriques Utilisées

Accuracy : Pourcentage de prédictions correctes.

Précision : Nombre de vrais positifs sur le nombre total de positifs prédits.

Rappel (Recall) : Nombre de vrais positifs sur le nombre total de positifs réels.

F1-Score : Moyenne harmonique de la précision et du rappel.

AUC-ROC : Surface sous la courbe ROC pour évaluer la performance globale.

Résultats

Les résultats obtenus pour chaque algorithme sont comparés en utilisant les métriques ci-dessus, ce qui permet de déterminer l'algorithme le plus performant pour notre problématique.

**7. Comparaison avec une Autre Approche**

Pour une comparaison rigoureuse, nous avons appliqué une approche basée sur des règles (Rule-Based Approach) et des algorithmes de clustering pour voir si les clusters naturels correspondent aux diagnostics. Cette comparaison nous a permis de valider l'efficacité de nos modèles d'apprentissage automatique.

**8. Conclusion et Perspectives**

Conclusion :

Ce projet a démontré l'efficacité de l'utilisation des techniques d'apprentissage automatique pour la détection précoce du cancer du poumon. Le modèle de forêt aléatoire s'est avéré être le plus performant, avec une précision et un rappel élevés.

Perspectives :

Pour améliorer encore les performances, nous pourrions :

Collecter plus de données pour améliorer la robustesse des modèles.

Essayer d'autres algorithmes et techniques de prétraitement.

Explorer des méthodes d'ensemble pour combiner plusieurs modèles et améliorer les résultats.

**9. Références et Bibliothèques Utilisées**

Pandas : Manipulation des données.

Scikit-learn : Implémentation des algorithmes d'apprentissage automatique.

TensorFlow/Keras : Pour les réseaux de neurones.

Matplotlib/Seaborn : Visualisation des données.

Numpy : Calcul scientifique de base.