Rapport projet machine learning

realisé par : Imane Kasmi,Hiba Mimouni

1. Définition de l'objectif de la premiere patie du Dataset Diabète

Objectif : Développer un modèle de machine learning capable de prédire si une personne est atteinte de diabète en utilisant un dataset contenant diverses mesures médicales comme age , nombre de grosses,etc ...

1.1. Source des données

Dataset : Le dataset "Diabetes" est disponible sur Kaggle. Ce dataset contient des mesures médicales de différentes personnes et une colonne indiquant si elles sont atteintes de diabète (Outcome).
Description des attributs du dataset :

- Pregnancies : Nombre de grossesses
- Glucose : Niveau de glucose dans le sang
- BloodPressure : Mesure de la pression artérielle
- SkinThickness : Épaisseur de la peau
- Insulin : Niveau d'insuline dans le sang
- BMI : Indice de masse corporelle
- DiabetesPedigreeFunction : Pourcentage de diabète
- Age : Âge
- Outcome : Résultat final (1 = Oui, 0 = Non)

1.2.Analyse Exploratoire des Données pour le Dataset de Diabète

L'analyse exploratoire des données (EDA) permet de mieux comprendre la structure et les caractéristiques d'un dataset. Voici un résumé des résultats obtenus et des étapes suivies pour l'EDA du dataset de diabète.

- Nous avons chargé le dataset et affiché les premières lignes pour obtenir une vue d'ensemble des données.
- Nous avons vérifié le nombre de lignes et de colonnes pour comprendre la taille du dataset. Le dataset contient 768 observations et 9 colonnes (8 caractéristiques et 1 variable cible).

- *Nous avons calculé les statistiques descriptives pour chaque caractéristique pour obtenir des informations sur la distribution des données. Ces statistiques montrent :Les valeurs moyennes (mean), médianes (50%), et les dispersions (std) des différentes caractéristiques.Les valeurs minimales (min) et maximales (max).*
  La présence de valeurs nulles pour certaines caractéristiques, comme Insulin et SkinThickness.
- *Nous avons examiné la répartition des classes de la variable cible Outcome.Il y a 500 cas non diabétiques et 268 cas diabétiques, ce qui indique un déséquilibre dans les classes.*
- *Nous séparons ensuite les caractéristiques (X) de la variable cible (Y).Les caractéristiques (X) incluent toutes les colonnes sauf Outcome, et les étiquettes (Y) incluent uniquement la colonne Outcome.*

1.3.Description de la phase de Pre-Processing des données

Dans cette section, nous avons effectué une analyse exploratoire des données pour comprendre la structure et la distribution des variables. Ensuite, nous avons utilisé ColumnTransformer pour prétraiter les données en imputant les valeurs manquantes avec la moyenne et en mettant à l'échelle les caractéristiques numériques.

1.4.Justification du Choix des Algorithmes

- **SVM** est adapté aux problèmes de classification binaire comme celui-ci, où les classes peuvent être linéairement séparables après transformation. Les attributs comme Glucose, BMI, et Age pourraient potentiellement être séparés linéairement pour distinguer les patients diabétiques des non-diabétiques.
- **Random Forest** est robuste face aux données bruitées et peut gérer efficacement les ensembles de données avec des attributs divers.Avec des attributs variés comme Insulin, SkinThickness, et DiabetesPedigreeFunction, Random Forest peut capturer des relations non linéaires entre les caractéristiques et la cible.
- **Le KNN** a été inclus pour sa simplicité et sa capacité à identifier des groupes similaires basés sur la proximité des voisins, adapté à des attributs comme Glucose et BMI.
- **la Logistic Regression** a été utilisée pour son interprétabilité et sa capacité à modéliser la probabilité de diabète en fonction des caractéristiques disponibles.

1.5.Résultats des Modèles

SVM

- *Accuracy (Training): 78.66%*
- *Accuracy (Test): 77.27%*
- *Cross-validation Mean Accuracy: 77.35%*
- *Cross-validation Standard Deviation: 0.02*

Random Forest

- *Accuracy (Training): 100%*
- *Accuracy (Test): 72.73%*
- *Cross-validation Mean Accuracy: 75.53%*
- *Cross-validation Standard Deviation: 0.04*

KNN

- *Accuracy (Training): 82.74%*
- *Accuracy (Test): 71.43%*
- *Cross-validation Mean Accuracy: 73.57%*
- *Cross-validation Standard Deviation: 0.02*

Logistic Regression

- *Accuracy (Training): 78.50%*
- *Accuracy (Test): 75.97%*
- *Cross-validation Mean Accuracy: 77.09%*
- *Cross-validation Standard Deviation: 0.02*

1.6.Comparaison des Performances

- ***Accuracy (Test):*** *Le SVM a la meilleure précision sur les données de test avec 77.27%, suivi de la régression logistique avec 75.97%, KNN avec 71.43%, et enfin Random Forest avec 72.73%. Bien que Random Forest ait une précision de 100% sur les données d'entraînement, son score sur les données de test est moins élevé, indiquant un potentiel surapprentissage (overfitting).*
- ***Cross-validation:*** *En termes de validation croisée, SVM et la régression logistique montrent des performances similaires avec des moyennes d'exactitude autour de 77%. KNN suit avec une moyenne d'exactitude de 73.57%, tandis que Random Forest est légèrement inférieur avec 75.53%. La régression logistique et SVM montrent également une faible variance avec des écarts-types de précision autour de 0.02.*

2. Détection de la maladie de Parkinson

L'objectif de ce projet est de développer un modèle de machine learning capable de détecter la maladie de Parkinson en se basant sur diverses caractéristiques extraites de l'enregistrement vocal des patients. Ce modèle aidera à diagnostiquer la maladie à un stade précoce, ce qui peut améliorer les options de traitement et la qualité de vie des patients.

2.1. Source des données

Dataset : Le dataset "Parkinson Disease Detection "est disponible sur Kaggle.Ce dataset contient diverses mesures vocales de sujets atteints de la maladie de Parkinson et de sujets sains, utilisées pour créer des modèles de classification afin de détecter la présence de la maladie.

Description des Attributs du Dataset

- *name : Nom du sujet et numéro d'enregistrement*
- *MDVP:Fo(Hz) : Fréquence fondamentale vocale moyenne*
- *MDVP:Fhi(Hz) : Fréquence fondamentale vocale maximale*
- *MDVP:Flo(Hz) : Fréquence fondamentale vocale minimale*
- *MDVP:Jitter(%), MDVP:Jitter(Abs), MDVP:RAP, MDVP:PPQ, Jitter:DDP : Mesures de variation de la fréquence fondamentale*
- *MDVP:Shimmer, MDVP:Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, MDVP:APQ, Shimmer:DDA : Mesures de variation de l'amplitude*
- *NHR, HNR : Mesures du rapport bruit/tonalité*
- *status : État de santé du sujet (1 = Parkinson, 0 = sain)*
- *RPDE, D2 : Mesures de complexité dynamique non linéaire*
- *DFA : Exposant de mise à l'échelle fractale du signal*
- *spread1, spread2, PPE : Mesures non linéaires de variation de la fréquence fondamentale*

2.2.Justification du Choix des Algorithmes

1. *Support Vector Machine (SVM)*
   - *Justification : Les SVM sont efficaces pour les espaces de grande dimension et restent efficaces même si le nombre de dimensions est supérieur au nombre d'échantillons. Utiliser un noyau linéaire simplifie l'interprétation et est bien adapté pour des problèmes de classification binaire comme celui-ci.*
2. *Random Forest*
   - *Justification : Les forêts aléatoires sont robustes contre le surapprentissage et peuvent gérer des jeux de données avec de nombreuses caractéristiques. Elles offrent également une bonne performance avec des données déséquilibrées.*
3. *K-Nearest Neighbors (KNN)*
   - *Justification : KNN est simple et intuitif. Il ne fait aucune supposition forte sur les données et peut capturer des relations complexes dans les données. Cependant, il peut être sensible aux échelles des caractéristiques, d'où la normalisation préalable des données.*

4.Logistic Regression

Justification : La régression logistique est un modèle de base efficace pour la classification binaire. Elle est interprétable et offre une bonne performance lorsque les relations entre les caractéristiques et la cible sont linéaires.

2.3.Analyse Exploratoire des Données

L'analyse exploratoire des données (EDA) permet de mieux comprendre la structure et les caractéristiques d'un dataset. Voici un résumé des résultats obtenus et des étapes suivies pour l'EDA du dataset de maladie de Parkinson

- ***Chargement des Données***

Les données du projet de détection de la maladie de Parkinson ont été chargées à partir d'un fichier CSV contenant 195 enregistrements et 24 caractéristiques. Le jeu de données ne comporte aucune valeur manquante, ce qui simplifie le processus de prétraitement.

- ***Aperçu Initial***

Un aperçu des premières lignes du jeu de données et des informations sur les colonnes a été effectué pour comprendre la structure et le contenu des données.

- ***Vérification des Valeurs Manquantes***

Aucune valeur manquante n'a été trouvée dans le jeu de données, ce qui est un bon indicateur de la qualité des données.

- ***Statistiques Descriptives***

Les statistiques descriptives fournissent un résumé des principales caractéristiques des données, y compris les moyennes, les écarts-types, les valeurs minimales et maximales, et les quartiles.

- *Distribution de la Variable Cible*

La variable cible status indique l'état de santé du sujet, où 1 représente un sujet atteint de la maladie de Parkinson et 0 un sujet sain. La distribution montre un déséquilibre avec plus de cas de Parkinson (147) que de sujets sains (48).

1.3.Description de la phase de Pre-Processing des données

- ***Séparation des Caractéristiques et de la Cible***

Les caractéristiques X et la cible y sont séparées. La colonne name est supprimée car elle n'apporte pas d'information utile pour la classification.

- ***Division des Données***

Les données sont divisées en ensembles d'entraînement (80%) et de test (20%) pour évaluer les performances des modèles de manière indépendante.

- ***prétraitement des données***

La phase de prétraitement des données commence par la séparation des caractéristiques et de la cible, suivie de la division du jeu de données en ensembles d'entraînement et de test. Ensuite, un ColumnTransformer est utilisé pour standardiser les caractéristiques numériques, en imputant les valeurs manquantes avec la moyenne et en normalisant les données. Ce pipeline assure une transformation cohérente et robuste des données, préparant efficacement le jeu de données pour l'entraînement des modèles de classification.

1.4.Résultats des modèles

Modèle	Exactitude Entraînement	Exactitude Test	Cross-Validation Mean Accuracy	Cross-Validation Std Dev
SVM	0.8846	0.8718	0.7949	0.0959
Random Forest	1.0000	0.7949	0.7897	0.0497
KNN	0.9679	0.7692	0.7795	0.0754
Logistic Regression	0.8718	0.8205	0.8103	0.0771

Analyse des performances :

- Exactitude de Test : Pour évaluer la performance sur les donnees, l'algorithme SVM obtient un score de 0.8718, ce qui est légèrement supérieur à celui de la Régression Logistique (0.8205), du KNN (0.7692) et du Random Forest (0.7949).
- Cross-Validation Mean Accuracy : Bien que SVM ait une exactitude de test élevée, sa moyenne de validation croisée (0.7949) est inférieure à celle de Random Forest (0.7897). Cependant, il est important de noter que la différence est relativement faible, avec une plus grande variance pour SVM par rapport à Random Forest.

Conclusion :

En se basant sur l'exactitude de test et la performance générale sur les différentes métriques, SVM semble être l'algorithme le plus performant dans cette grille d'évaluation, avec une exactitude de test élevée et une performance globale compétitive en validation croisée malgré une légère variance.