Université Mohammed Premier
École Nationale des Sciences Appliquées

-Oujda-

RAPPORT DE PROJET DE FIN DE SEMESTRE

Filière : Génie Informatique

Encadré par :
Mr.Zakaria Haja

Réalisé par:
Hajar Abdellaoui
Abir Allaoui
Hafsa Daanouni
Nadia Azam

SOMMAIRE

INTRODUCTION

Dans le cadre de ce mini-projet, nous avons entrepris une étude visant à prédire le risque de décès (DEATH_EVENT) à partir de plusieurs caractéristiques médicales chez des patients. Pour ce faire, nous avons exploré l'utilisation de quatre modèles de machine learning différents : la régression logistique, la méthode des k plus proches voisins (KNN), la forêt aléatoire (Random Forest) et l'arbre de décision. Notre objectif principal était d'évaluer la performance de ces modèles dans la prédiction du DEATH_EVENT et d'identifier celui qui offre les meilleures performances.
Dans ce rapport, nous présenterons les étapes suivies pour mener à bien cette étude, notamment la collecte et la préparation des données, l'exploration et la visualisation des données, la sélection des modèles, l'entraînement des modèles, l'évaluation de leur performance et enfin, une discussion sur les résultats obtenus. Cette étude vise à fournir des insights précieux pour la prise de décision médicale en identifiant les facteurs prédictifs associés au risque de décès chez les patients.

DESCRIPTION DU DATASET

Titre du jeu de données :
Ensemble de données médicales sur les patients atteints de maladies cardiovasculaires : Analyse des facteurs de risque de décès.

Origine et source des données :
Les données ont été extraites du jeu de données "Heart Failure Prediction - Clinical Records" disponible sur Kaggle. Ce jeu de données a été compilé à partir de diverses sources, y compris des enquêtes de santé publique, des bases de données médicales et des études de recherche. Les informations ont été anonymisées pour protéger la confidentialité des participants.

Objectif du jeu de données :
L objectif de cette base de données pourrait être d'analyser les facteurs de risque associés aux événements de décès chez les patients atteints de maladies cardiaques. En utilisant les différentes variables fournies, on peut tenter de comprendre les facteurs qui influencent la mortalité chez ces patients, ce qui pourrait aider à améliorer les stratégies de prévention et de prise en charge médicale

Description des attributs :

- **Age:** L'âge des patients.
- **Anaemia:** Indique si le patient souffre d'anémie (0 pour non, 1 pour oui).
- Creatinine Phosphokinase: Niveau de l'enzyme créatine phosphokinase dans le sang (CPK).
- **Diabetes:** Indique si le patient est diabétique (0 pour non, 1 pour oui).
- **Ejection Fraction:** Pourcentage de volume de sang éjecté par le ventricule gauche lors de chaque contraction cardiaque.
- **High Blood Pressure:** Indique si le patient souffre d'hypertension (0 pour non, 1 pour oui).
- **Platelets:** Nombre de plaquettes dans le sang.
- **Serum** Creatinine: Niveau de créatinine sérique dans le sang, qui est un indicateur de la fonction rénale.
- **Serum Sodium:** Niveau de sodium sérique dans le sang.
- **Sex:** Genre du patient (0 pour femme, 1 pour homme).
- **Smoking:** Indique si le patient fume (0 pour non, 1 pour oui).
- **Time:** Temps de suivi en jours.
- **Death Event:** Indique si le patient est décédé pendant la période de suivi.(0 pour non, 1 pour oui).

.

DESCRIPTION DES BIBLIOTHÈQUES UTILISÉES

Voici une brève description de chaque bibliothèque utilisée :

1.**pandas** (import pandas as pd) : Pandas est une bibliothèque Python qui offre des structures de données et des outils d'analyse de données faciles à utiliser. Elle est largement utilisée pour la manipulation et l'analyse de données tabulaires.

2.**numpy** (import numpy as np) : NumPy est une bibliothèque Python destinée à la manipulation d'arrays (tableaux) et à la réalisation de calculs numériques. Elle fournit des fonctions pour travailler efficacement avec des tableaux multidimensionnels.

3.**seaborn** (import seaborn as sns) : Seaborn est une bibliothèque de visualisation de données basée sur Matplotlib. Elle fournit une interface de haut niveau pour créer des graphiques attrayants et informatifs.

4.**Scikit-learn** (sklearn) : est une bibliothèque Python d'apprentissage automatique qui offre une large gamme d'algorithmes pour la classification, la régression, le clustering, et des outils pour l'évaluation et la sélection de modèles. Elle est largement utilisée pour sa simplicité, sa cohérence et son intégration aisée avec d'autres bibliothèques Python.

5. **train_test_split** (from sklearn.model_selection import train_test_split) : Cette fonction de scikit-learn est utilisée pour diviser un ensemble de données en ensembles de données d'entraînement et de test, ce qui est essentiel pour évaluer les performances des modèles d'apprentissage automatique.

6.**matplotlib.pyplot** (import matplotlib.pyplot as plt) : Matplotlib est une bibliothèque de visualisation de données en Python. Pyplot est un module de Matplotlib qui fournit une interface similaire à celle de MATLAB pour la création de graphiques.

EXPLORATION ET VISUALISATION DES DONNÉES

Chargement du jeu de données

Après avoir téléchargé le jeu de données à partir du fichier CSV "heart_failure_clinical_records.csv", nous avons effectué une première inspection pour comprendre la structure et le contenu des données. Ce jeu de données comprend un total de 5000 entrées et 13 colonnes, chacune représentant une caractéristique médicale spécifique des patients. Ces caractéristiques incluent des variables telles que l'âge, la présence d'anémie, le niveau d'enzymes dans le sang, le statut du diabète, la fraction d'éjection cardiaque, la pression artérielle élevée, le nombre de plaquettes, les niveaux de créatinine et de sodium sériques, le sexe, le statut du tabagisme, la durée de suivi et l'événement de décès.

Qualité des données et exploration des variables

Après avoir examiné les informations sur les colonnes à l'aide de la méthode info() de Pandas, nous avons constaté que notre jeu de données ne contient aucune valeur manquante, ce qui indique une bonne qualité des données. Cela nous a permis de passer à l'étape suivante de notre analyse en explorant plus en détail les relations entre les différentes variables et leur impact potentiel sur le résultat de l'étude, à savoir l'événement de décès.

Réserve d'un sous-ensemble de données pour une évaluation indépendante

Nous avons entrepris plusieurs étapes de prétraitement pour préparer nos données à l'analyse. Parmi ces étapes, nous avons décidé de réserver un sous-ensemble de 1000 entrées pour une utilisation ultérieure dans notre étude. Cette décision stratégique a été prise dans le but spécifique de constituer un ensemble de données de test distinct, sur lequel nous pourrons évaluer les performances de nos modèles de machine learning après avoir déterminé le meilleur modèle à utiliser.

EXPLORATION ET VISUALISATION DES DONNÉES

Pour cela, nous avons initialement supprimé les entrées correspondant aux lignes de 4000 à 5000 de notre jeu de données. Ensuite, nous avons extrait ces 1000 entrées supprimées et les avons conservées dans un sous-ensemble distinct pour une utilisation ultérieure. Cette approche nous permettra de réaliser une évaluation indépendante de nos modèles sur un ensemble de données inédit, garantissant ainsi une évaluation impartiale de leur performance une fois que nous aurons sélectionné le meilleur modèle.

Cette réserve de données nous offre également une opportunité précieuse d'évaluer la capacité de généralisation de nos modèles, en les testant sur des données réelles qu'ils n'ont pas encore vues. Cela nous permettra de déterminer si nos modèles sont capables de maintenir leur performance lorsqu'ils sont confrontés à de nouvelles observations, ce qui est essentiel pour garantir leur utilité dans des situations réelles.
Cette approche méthodologique réfléchie nous permettra de mener une analyse rigoureuse et complète de nos modèles de machine learning, en fournissant des résultats fiables et exploitables pour notre étude sur les risques de décès chez les patients atteints de maladies cardiovasculaires.

Séparation des variables explicatives et de la variable cible

Après avoir nettoyé notre jeu de données, nous avons procédé à la séparation des variables en variables explicatives (X) et la variable cible (y) à prédire. Les variables explicatives comprennent toutes les caractéristiques médicales des patients, tandis que la variable cible représente l'événement de décès, que nous cherchons à prédire à l'aide de nos modèles de machine learning.

PRÉTRAITEMENT DES DONNÉES DU DATASET

Avant d'entraîner nos modèles de machine learning, nous avons effectué plusieurs étapes de prétraitement des données pour garantir la qualité et la pertinence de nos données.

Traitement des valeurs manquantes

Une première étape importante consiste à traiter les valeurs manquantes dans notre jeu de données. Après une inspection initiale, nous avons constaté que notre ensemble de données comprenait 5000 entrées et aucune valeur manquante dans aucune des colonnes. Par conséquent, aucune action de traitement des valeurs manquantes n'était nécessaire.

Encodage des valeurs catégorielles

Nous avons ensuite vérifié si nos caractéristiques contenaient des variables catégorielles nécessitant un encodage. Après examen, nous avons constaté que toutes nos caractéristiques étaient numériques, sans aucune variable catégorielle requérant un encodage supplémentaire.

Mise à l'échelle des données numériques

Enfin, nous avons mis à l'échelle nos données numériques pour les ramener à une plage commune. Pour ce faire, nous avons utilisé le MinMaxScaler de la bibliothèque scikit-learn pour mettre à l'échelle nos caractéristiques numériques entre 0 et 1. Cette étape est essentielle pour garantir que les caractéristiques avec des plages de valeurs différentes ne dominent pas inutilement l'entraînement de nos modèles de machine learning.

Les données ainsi prétraitées sont désormais prêtes à être utilisées pour entraîner nos modèles de machine learning et à subir une analyse plus approfondie pour prédire avec précision l'événement de décès chez les patients atteints de maladies cardiovasculaires. Ces étapes de prétraitement sont essentielles pour garantir la qualité et la pertinence de nos données dans le cadre de notre étude. Elles nous permettent de passer à l'étape suivante de notre analyse en entraînant et en évaluant nos modèles de machine learning pour prédire avec précision l'événement de décès chez les patients atteints de maladies cardiovasculaires.

SÉLÉCTION DES MODÈLES UTILSÉS:

Dans cette étape, nous avons exploré plusieurs algorithmes d'apprentissage automatique pour leur capacité à prédire la probabilité de décès pendant une opération cardiaque en utilisant les données médicales fournies. Nous avons sélectionné les modèles suivants pour l'évaluation :

1- **Régression Logistique** :
Un modèle de régression linéaire utilisé pour la classification binaire, qui modélise la relation entre les variables explicatives et la probabilité de la classe cible.

2- **k-Nearest Neighbors (k-NN)** :
Un algorithme non paramétrique utilisé pour la classification qui attribue une classe à une observation en se basant sur les classes majoritaires de ses voisins les plus proches dans l'espace des caractéristiques.

3 **-Arbre de decision**:
Les arbres de décision sont des modèles d'apprentissage supervisé polyvalents et faciles à interpréter, souvent utilisés pour la classification et la régression.

4- **Random Forest** :
Un modèle d'apprentissage ensembliste basé sur des arbres de décision, connu pour sa capacité à capturer les relations non linéaires dans les données et à gérer efficacement les caractéristiques complexes.

Chaque modèle a été entraîné sur l'ensemble d'entraînement et évalué par la suite sur l'ensemble de test pour comparer leurs performances en termes de précision, de rappel, de F1-score et d'autres mesures de performance.

ENTRAÎNEMENT DES MODÈLES

Division des données en ensembles d'entraînement et de test :

Avant d'entraîner nos modèles, nous avons divisé notre ensemble de données en deux parties distinctes : l'ensemble d'entraînement et l'ensemble de test. Nous avons réservé une portion des données (80%) pour l'ensemble d'entraînement, sur lequel les modèles seront entraînés, et le reste des données pour l'ensemble de test, qui sera utilisé pour évaluer la performance des modèles.

Ajustement des hyperparamètres des modèles :

Nous avons ajusté les hyperparamètres de chaque modèle afin d'optimiser leurs performances. Pour ce faire, nous avons utilisé des techniques telles que la recherche sur grille (grid search) en utilisant des outils tels que GridSearchCV ou RandomizedSearchCV de la bibliothèque Scikit-learn. Cette étape nous a permis d'explorer un ensemble de combinaisons d'hyperparamètres et de sélectionner celle qui produit les meilleurs résultats.

Entraînement des modèles sur les données d'entraînement :

Une fois les hyperparamètres optimisés, nous avons entraîné chaque modèle sur l'ensemble de données d'entraînement en utilisant la fonction fit() de la bibliothèque Scikit-learn. Pendant l'entraînement, les modèles ont appris à partir des caractéristiques fournies et de la variable cible (DEATH_EVENT), ajustant ainsi leurs paramètres internes pour minimiser l'erreur de prédiction.

ÉVALUATION DE LA PERFORMANCE

Après avoir entraîné et évalué quatre modèles différents sur notre ensemble de données médicales, nous avons procédé à une comparaison détaillée de leurs performances pour prédire la probabilité de décès pendant une opération cardiaque. Chaque modèle a été entraîné avec les meilleures configurations d'hyperparamètres trouvées lors de la recherche sur grille, et ses performances ont été évaluées sur l'ensemble de test. Les résultats obtenus sont résumés ci-dessous :

Expressions :

Accuracy : mesure la proportion de prédictions correctes parmi l'ensemble des prédictions.

Precision : mesure la proportion de vrais positifs parmi les prédictions positives.

Recall : mesure la proportion de vrais positifs identifiés correctement parmi tous les vrais positifs.

F1 Score : une mesure harmonique de la précision et du rappel, qui donne une vue globale de la performance du modèle.

Résultats :

Modèle de regression logistique :

	precision	recall	f1-score	support
0	0.85	0.90	0.87	540
1	0.76	0.67	0.72	260
accuracy			0.83	800
macro avg	0.81	0.79	0.80	800
weighted avg	0.82	0.83	0.82	800

la précision pour la classe 0 est de 0.89, ce qui signifie que 89% des prédictions positives pour cette classe sont correctes. De manière similaire, pour la classe 1, la précision est de 0.77, ce qui indique que 77% des prédictions positives pour cette classe sont correctes.

ÉVALUATION DE LA PERFORMANCE

Modèle de l’arbre de décision:

	precision	recall	f1-score	support
0	0.99	0.99	0.99	540
1	0.98	0.97	0.97	260
accuracy			0.98	800
macro avg	0.98	0.98	0.98	800
weighted avg	0.98	0.98	0.98	800

Après l'entraînement du modèle et son évaluation sur un ensemble de données de test, nous avons observé une performance impressionnante avec une précision globale de 98,25%. Cette précision élevée indique que le modèle est capable de classifier avec précision les patients en fonction de leur risque de décès.

En examinant le rapport de classification, nous constatons que le modèle présente une précision de 99% pour la classe "pas de décès" (0) et de 97% pour la classe "décès" (1). De plus, les valeurs de rappel pour ces deux classes sont également élevées, ce qui signifie que le modèle est capable de capturer la grande majorité des occurrences de chaque classe. Ces résultats sont confirmés par les scores F1, qui combinent à la fois la précision et le rappel pour fournir une mesure globale de la performance du modèle.

ÉVALUATION DE LA PERFORMANCE

Modèle du K plus proche voisin:

	precision	recall	f1-score	support
0	0.98	0.98	0.98	540
1	0.95	0.95	0.95	260
accuracy			0.97	800
macro avg	0.97	0.97	0.97	800
weighted avg	0.97	0.97	0.97	800

Nous avons utilisé le modèle KNN pour prédire l'événement de décès chez les patients atteints de maladies cardiovasculaires. Après avoir entraîné le modèle et l'avoir évalué sur un ensemble de données de test, nous avons constaté une précision globale de 97%. Cette précision indique que le modèle est capable de classifier avec précision les patients en fonction de leur risque de décès.

En analysant le rapport de classification, nous observons que le modèle présente une précision de 98% pour la classe "pas de décès" (0) et de 95% pour la classe "décès" (1). De plus, les valeurs de rappel pour ces deux classes sont également élevées, ce qui signifie que le modèle est capable de capturer la grande majorité des occurrences de chaque classe. Les scores F1, qui combinent à la fois la précision et le rappel, confirment également la performance globale du modèle.

ÉVALUATION DE LA PERFORMANCE

Modèle du Random Forest:

	precision	recall	f1-score	support
0	0.99	0.99	0.99	540
1	0.99	0.98	0.98	260
accuracy			0.99	800
macro avg	0.99	0.99	0.99	800
weighted avg	0.99	0.99	0.99	800

Nous avons utilisé le modèle Random Forest pour prédire l'événement de décès chez les patients atteints de maladies cardiovasculaires. Après avoir entraîné le modèle et l'avoir évalué sur un ensemble de données de test, nous avons observé une précision globale de 99%. Cette précision indique que le modèle est capable de classifier avec précision les patients en fonction de leur risque de décès.

En examinant le rapport de classification, nous constatons que le modèle présente une précision de 99% pour la classe "pas de décès" (0) et de 99% pour la classe "décès" (1). De plus, les valeurs de rappel pour ces deux classes sont également élevées, ce qui signifie que le modèle est capable de capturer la grande majorité des occurrences de chaque classe. Les scores F1, qui combinent à la fois la précision et le rappel, confirment également la performance globale du modèle.

SÉLECTION DU MEILLEUR MODÈLE

Pour déterminer le meilleur modèle pour notre tâche de prédiction de la probabilité de décès pendant une opération cardiaque, nous avons évalué quatre modèles différents : Régression Logistique, Arbre de Décision, k-NN et La forêt aléatoire. Nous avons utilisé la validation croisée stratifiée pour évaluer les performances de chaque modèle sur plusieurs plis de données.

- **Régression Logistique** : Les scores de validation croisée stratifiée pour ce modèle sont les suivants : [0.871875 0.8640625 0.8734375 0.8484375 0.8375]. La moyenne des scores est de 0.8590625000000001.

- **Arbre de Décision** : Nous avons obtenu les scores suivants pour la validation croisée stratifiée avec l'Arbre de Décision :[0.978125 0.984375 0.9890625 0.984375 0.9765625]. La moyenne des scores est de 0.9824999999999999.

- **k-NN** : Les scores de validation croisée stratifiée pour le modèle k-NN sont les suivants :[0.9671875 0.965625 0.9578125 0.96875   0.96875]. La moyenne des scores est de 0.965625.

- **La forêt aléatoire** : Enfin, pour le modèle Random Forest, nous avons obtenu les scores suivants : [0.990625 0.99375 0.9875   0.990625 0.9875]. La moyenne des scores est de 0.99.

Conclusion :
Sur la base de ces résultats, le modèle **Random Forest** a présenté les performances les plus élevées avec une moyenne de 99% pour la précision sur l'ensemble des plis de validation croisée. Par conséquent, nous sélectionnons le modèle Random Forest comme le meilleur modèle pour notre tâche de prédiction.

ÉVALUATION DU MEILLEUR MODÈLE SUR L'ENSEMBLE DE TEST

Une fois le meilleur modèle sélectionné, nous avons évalué sa performance sur un ensemble de données indépendant réservé à cet effet. Pour ce faire, nous avons extrait les caractéristiques de cet ensemble de test dans une variable que nous avons nommée x_features_test, et les étiquettes cibles dans une variable appelée y_target_test.

Le modèle Random Forest, qui a été sélectionné comme le meilleur modèle, a été utilisé pour prédire les étiquettes sur l'ensemble de test. Les prédictions résultantes sont stockées dans une variable y_predict_result. Pour évaluer la performance du modèle, nous avons calculé plusieurs mesures :

- **Accuracy** (Exactitude) : L'exactitude du modèle mesure la proportion de prédictions correctes parmi l'ensemble des prédictions. Pour notre modèle, l'accuracy est de 0.7.

- **Mean Absolute Error** (MAE) : Cette mesure quantifie l'erreur moyenne entre les valeurs prédites et les valeurs réelles. Pour notre modèle, le MAE est de 0.3.

Ces métriques nous fournissent une évaluation globale de la performance du modèle Random Forest sur l'ensemble de test, nous permettant de mieux comprendre sa capacité à généraliser sur de nouvelles données.

CONCLUSION

Dans ce projet, nous avons développé et évalué des modèles de classification pour prédire l'événement de décès (DEATH_EVENT) à partir de caractéristiques médicales. Nous avons testé la régression logistique, le KNN, l'arbre de décision, et la forêt aléatoire. Après la prétraitement des données et l'entraînement des modèles, nous avons évalué leurs performances avec des mesures telles que l'accuracy, la précision, le rappel et le score F1. La forêt aléatoire a montré les meilleures performances globales avec une précision de 99%, suivie de l'arbre de décision et du KNN. La régression logistique, bien qu'efficace, a été surpassée par les autres modèles. En conclusion, la forêt aléatoire est le modèle le plus performant pour prédire les événements de décès dans notre dataset.