Introducion

....

La détection précoce et précise est cruciale pour améliorer les taux de survie

Objective

Notre étude vise à développer une application basée sur Python pour la détection des maladies cardiaques
en utilisant des algorithmes d'apprentissage automatique. Nous prendrons en compte les limites et les défis
potentiels de l'utilisation de l'apprentissage automatique pour la détection des maladies cardiaques.

Un diagnostic et un traitement précoces peuvent améliorer considérablement les résultats pour les patients atteints de maladies cardiaques. Les modèles d'apprentissage automatique peuvent être utilisés pour identifier les patients à risque élevé de développer une maladie cardiaque, ce qui permet aux médecins de prendre des mesures préventives.

Methodologie

Dans le contexte de la détection des maladies cardiaques à l'aide de l'apprentissage automatique, plusieurs algorithmes peuvent être considérés pour résoudre ce problème.

kNN :

L'idée de base derrière KNN est de prédire la classe d'un nouvel exemple en trouvant les k exemples les plus proches dans l'ensemble de données d'apprentissage et en effectuant un vote majoritaire parmi ces voisins.

SVM :

Dans le contexte de la détection des maladies cardiaques, les SVM peuvent être utilisées pour générer un modèle qui classe de nouveaux exemples dans différentes catégories, ce qui peut être utile pour prédire le risque de maladie cardiaque chez un individu.

Elle vise à trouver un hyperplan ou une ligne optimale dans un espace à N dimensions qui maximise la distance entre les différentes classes.

Leur objectif principal est de maximiser la marge, qui représente la distance entre l'hyperplan et les échantillons les plus proches de chaque classe.Cela permet de bien séparer les données et d'assurer une classification précise.

RF :

C'est un choix populaire en raison de sa capacité à gérer des ensembles de données complexes.

Bootstrapping : ensures that we are not using the same data for every tree , to help our model to be less sensitive to the original train data

feature selection : help to reduce correlation btw the trees , bcz if we use the same feature then ost of your trees
will have the same decision nodes and will act similarly (increase variance)

- many features cosider : researchers found that values close to the log or sqrt of the total nmbr of features work well.

Validation :

Elle vise à garantir que les modèles fonctionnent correctement et qu'ils peuvent être utilisés de manière fiable pour prédire le risque de maladie cardiaque chez les patients.

MC :

Elle compare les prédictions du modèle à la réalité, permettant d'identifier les prédictions justes et fausses pour chaque classe.

Precision :

Elle est définie comme le rapport entre le nombre de vrais positifs (échantillons positifs correctement prédits) et le nombre total de prédictions positives (échantillons positifs prédits).

Rappel :

Il mesure la capacité du modèle à identifier correctement tous les éléments positifs

f1 :

Il combine la précision et le rappel en une seule valeur comprise entre 0 et 1

Simultation Setup

Dataset :

Les variables comprennent des informations démographiques, des antécédents médicaux, des résultats d'examens médicaux et le diagnostic final de la présence ou non d'une maladie cardiaque.

L'ensemble de données contient des
informations sur les patients, y compris leur âge, leur sexe, leur tension artérielle, leur taux de cholestérol,
etc.

Vars :

Resultats
Resultats Obtenus :

```
models = ['KNN', 'SVM', 'RF']
accuracy_model1 = [87, 83, 84]
accuracy_model2 = [89, 86, 84]
```

```
# RF :
# ACC 0.8241758241758241
# PREC 0.9047619047619048
# RECAL 0.76
# F1 0.8260869565217391
# KNN :
# Acc 0.8901098901098901
```

```
# Prec 0.8703703703703703
# recal 0.94
# F1 0.9038461538461539
# SVM:
# - Precision: 0.8653846153846154
# - Recall: 0.9
# - F1 Score: 0.8823529411764707
# - Accuracy: 0.8681318681318682
```

parametre de model montioner dans la phase de resultat obtenu :

1. K-Nearest Neighbors (KNN)

- **Nombre de voisins (k)** : Une grille de recherche a été utilisée pour tester des valeurs de k allant de 1 à 25.
- **Distance de similarité** : Distance euclidienne utilisée pour calculer la similarité entre les échantillons.
- **Sélection du meilleur k** : Le k avec le score de précision le plus élevé sur l'ensemble de test a été sélectionné.

Résultats : Le modèle KNN a atteint une précision de 89%.

3. Support Vector Machine (SVM)

- **Noyau** : Radial Basis Function (RBF)
- **C (régularisation)** : Valeurs testées entre 0,1 et 100. La meilleure valeur trouvée était 100.
- **Gamma** : Valeurs testées entre 0,001 et 1. La meilleure valeur trouvée était 0,01.
- **Méthode de validation** : Validation croisée

Résultats : Le modèle SVM a obtenu une précision de 86%.

2. Random Forest (RF)

- **max_depth** : 70
- **max_features** : 'auto'
- **min_samples_leaf** : 4
- **min_samples_split** : 10
- **n_estimators** : 400
- **random_state** : 42

Résultats : Le modèle RF a obtenu une précision de 84%.

ona utiliser diff tech d'optimisation par example recherche en grille pour trouver les config optimales

Le noyau radial est souvent utilisé dans les SVM pour les problèmes de classification non linéaire, car il permet de transformer les données d'entrée en un espace de dimension supérieure où elles peuvent être séparées linéairement. Il est également utilisé dans d'autres algorithmes d'apprentissage automatique, tels que les réseaux de neurones et les méthodes de clustering.

MC :
Matrice de Confusion KNN :

Vrais Positifs (TP) : 47 (Prédiction correcte de la maladie)

Faux Positifs (FP) : 7 (Prédiction de la maladie alors qu'il n'y en a pas)

Faux Négatifs (FN) : 3 (Non prédiction de la maladie alors qu'elle est présente)

Vrais Négatifs (TN) : 34 (Prédiction correcte de l'absence de maladie)

Matrice de Confusion RF (Random Forest) :

Vrais Positifs (TP) : 38

Faux Positifs (FP) : 4

Faux Négatifs (FN) : 12

Vrais Négatifs (TN) : 37

Matrice de Confusion SVM (Support Vector Machine) :

Vrais Positifs (TP) : 45

Faux Positifs (FP) : 7

Faux Négatifs (FN) : 5

Vrais Négatifs (TN) : 34

Comparaison :

L’article a rapporté une précision d’environ 84 % pour le modèle Random Forest. notre modèle KNN a obtenu une précision légèrement supérieure, ce qui suggère qu’il pourrait être plus performant pour prédire le risque de maladie cardiaque dans notre ensemble de données.

Features selection :

L'objectif de cette analyse est d'identifier les caractéristiques les plus importantes pour prédire les maladies cardiaques à l'aide d'un classificateur K-Nearest Neighbors (KNN).

Present the top 3 features:
oldpeak
thalach
ca_0

Détails approfondis

Oldpeak (oldpeak)

Description : Dépression ST induite par l'exercice par rapport au repos (mesure ECG).
Pertinence clinique :
Indique une ischémie, qui est une caractéristique commune de la maladie coronarienne.
Une importance élevée suggère une valeur diagnostique significative.

Fréquence Cardiaque Maximale Atteinte (thalach)

Description : Fréquence cardiaque maximale atteinte pendant l'effort physique.
Pertinence clinique :
Une fréquence cardiaque maximale plus faible peut indiquer une mauvaise forme cardiovasculaire et des problèmes cardiaques potentiels.
Critique pour diagnostiquer la maladie cardiaque.

Nombre de Vaisseaux Majeurs Colorés par Fluoroscopie (ca_0)

Description : Nombre de vaisseaux sanguins majeurs colorés par fluoroscopie (0-3), spécifiquement ca_0 pour aucun vaisseau.
Pertinence clinique :
L'absence d'obstruction visible des vaisseaux pourrait être une condition significative liée à la santé cardiaque.

Implications et contributions :

Cela permet aux médecins de prendre des mesures préventives,
telles que des changements de style de vie ou des médicaments,

Limites :

Cette étude présente des limitations. Premièrement, la taille des données était petite. Deuxièmement, les modèles d'apprentissage automatique ne sont pas parfaits et peuvent faire des erreurs, donc Ils doivent être des outils d'aide à la décision, pas des substituts au jugement clinique.

Conclusion :

Pour conclure cette présentation, nous avons démontré que l'utilisation des modèles d'apprentissage automatique présente un potentiel considérable pour la détection et la prévention des maladies cardiaques. Le modèle KNN, en particulier, a atteint une précision impressionnante de 89%, surpassant d'autres approches comme Random Forest et SVM.Cependant, des recherches supplémentaires sont nécessaires pour développer et valider ces modèles, bien que les résultats préliminaires soient prometteurs.

il serait également intéressant de valider ces résultats sur des données de test supplémentaires ou sur des jeux de données différents pour confirmer la robustesse du modèle.

Comparaison

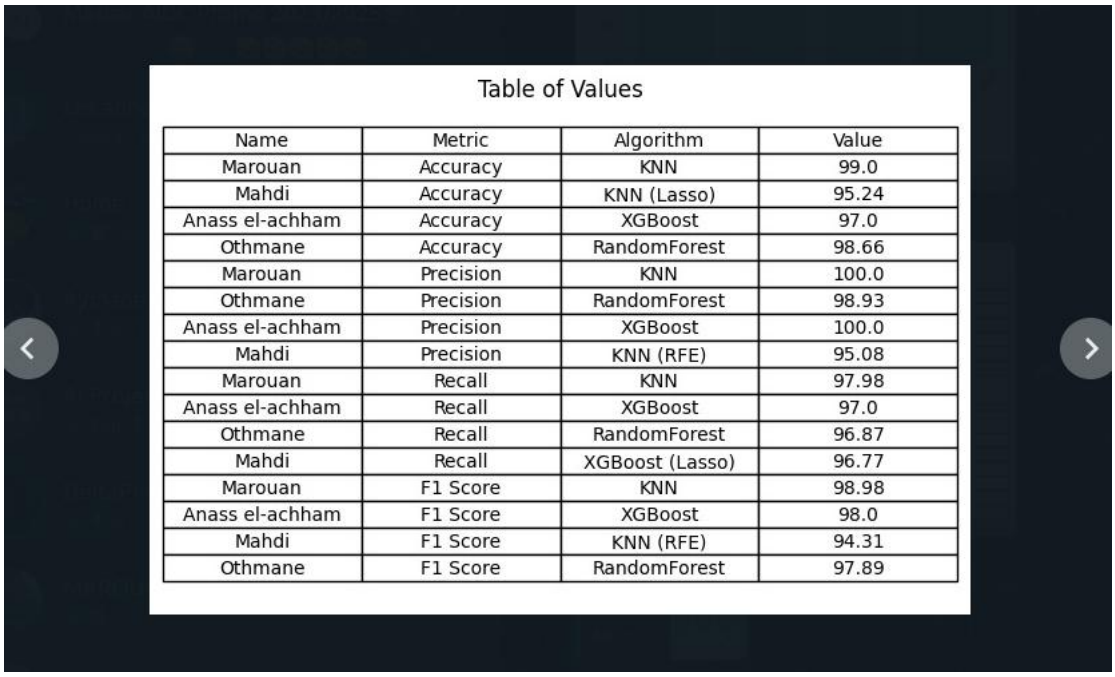

Table of Values

Name	Metric	Algorithm	Value
Marouan	Accuracy	KNN	99.0
Mahdi	Accuracy	KNN (Lasso)	95.24
Anass el-achham	Accuracy	XGBoost	97.0
Othmane	Accuracy	RandomForest	98.66
Marouan	Precision	KNN	100.0
Othmane	Precision	RandomForest	98.93
Anass el-achham	Precision	XGBoost	100.0
Mahdi	Precision	KNN (RFE)	95.08
Marouan	Recall	KNN	97.98
Anass el-achham	Recall	XGBoost	97.0
Othmane	Recall	RandomForest	96.87
Mahdi	Recall	XGBoost (Lasso)	96.77
Marouan	F1 Score	KNN	98.98
Anass el-achham	F1 Score	XGBoost	98.0
Mahdi	F1 Score	KNN (RFE)	94.31
Othmane	F1 Score	RandomForest	97.89

Accuracy (Précision) :
La précision mesure le pourcentage de prédictions correctes parmi l'ensemble des prédictions. C'est une métrique globale qui donne une idée générale de la performance du modèle. Dans le contexte des maladies cardiovasculaires, une haute précision signifie que le modèle peut correctement identifier les patients atteints et non atteints de MCV.

Precision (Précision) :
La précision est le ratio des vrais positifs par rapport au total des positifs prédits (vrais positifs + faux positifs). Elle est cruciale lorsque le coût d'une fausse alarme (faux positif) est élevé. Pour les MCV, une haute précision garantit que les patients identifiés comme malades le sont réellement, évitant ainsi des traitements ou des interventions inutiles.

Recall (Rappel) :
Le rappel est le ratio des vrais positifs par rapport au total des positifs réels (vrais positifs + faux négatifs). C'est essentiel dans les cas où manquer un cas positif est critique. Dans le domaine des MCV, un haut rappel signifie que le modèle peut identifier la majorité des patients atteints, ce qui est vital pour une intervention précoce et un traitement adéquat.

F1 Score :

Le F1 score est la moyenne harmonique de la précision et du rappel. Il offre un équilibre entre les deux métriques, particulièrement utile lorsque les classes sont déséquilibrées. Pour les MCV, un bon F1 score indique que le modèle maintient un bon équilibre entre éviter les faux positifs et ne pas manquer les vrais positifs.

Comparaison avec conclusion :

En raison des différentes techniques de prétraitement et des variations dans les datasets utilisés, nous avons obtenu plusieurs résultats différents. Par conséquent, cette comparaison n'est pas vraiment efficace. Cependant, nous souhaitons souligner l'importance du dataset, en particulier le traitement des données (valeurs manquantes, type de données, outliers, smoothing), la représentativité de chaque caractéristique, ainsi que les méthodes d'apprentissage (régularisation, grid search, random search, algorithmes utilisés, etc.).

La solution à ce problème réside dans la nécessité de disposer d'un nouveau dataset, réel et de qualité, qui pourra être utilisé pour les recherches dans le domaine des maladies cardiovasculaires. Ce dataset servirait de référence, comme le dataset CIFAR pour la vision par ordinateur. De plus, nous envisageons de collaborer avec des experts du domaine pour remédier à ces problèmes et garantir des résultats plus fiables et pertinents.