Université Mohamed V
École Nationale Supérieure d'Informatique et d'Analyse des Systèmes

Prédiction des crises cardiaques: Modèles de Machine/Deep Learning

Rapport de projet

Réalisé par

Bekkai Chamss Doha
Choukhantri Ikram
Outgougua Yahya
N'gar Abdellah

Encadré par

Mme Houda Benbrahim

Année universitaire : 2023-2024

Table des matières

Table des figures

1 Introduction

1.1 Position du problème

L'objet de cette étude est d'explorer la prédiction des crises cardiaques, soulignant la nécessité d'une prévision précoce pour une gestion proactive des risques de santé.

L'application de techniques de machine learning dans ce contexte s'avère essentielle pour analyser de telles tendances , menant à une évaluation plus précise des risques parmi les individus. Les expériences visent à développer des modèles prédictifs capables d'évaluer, catégoriser et anticiper les potentielles crises en fonction des différents paramètres mentionnés, ce qui simplifie le processus d'identification.

L'importance de ces expériences réside dans la possibilité d'aider les individus à mieux comprendre leurs facteurs de risque cardiovasculaire, tout en fournissant aux professionnels de la santé des informations précieuses pour optimiser les processus de prévention et de prise en charge face à de telles situations

1.2 Travaux connexes et état de l'art

Les établissements médicaux peuvent exploiter ces prédictions pour identifier les tendances émergentes et adapter leurs approches de prévention, soulignant ainsi l'importance de ces modèles dans le contexte médical moderne. Ces efforts de prédiction offrent également des applications pratiques dans le domaine de la santé publique, permettant aux professionnels de concevoir des programmes de sensibilisation ciblés et d'élaborer des stratégies de prévention personnalisées. Les implications de cette prédiction vont au-delà du secteur médical, influençant la prise de conscience des risques cardiaques et encourageant des mesures préventives efficaces au sein de la population.

Bien qu'il puisse y avoir quelques différences dans les différentes nuances utilisées dans l'apprentissage automatique, les approches utilisées pour résoudre de tels problèmes sont, en général, similaires aux approches que nous avons utilisées dans ce projet : collecte de données, prétraitement, choix d'un modèle approprié et évaluation pour voir si le résultat final a été bien prédit.

2 Les données

2.1 Jeu de données

Suite à une recherche approfondie sur Internet, nous avons réussi à obtenir un ensemble de données adapté à notre contexte depuis la plateforme Kaggle. Le projet repose sur un ensemble de données détaillé incluant une variété de paramètres médicaux. Ces paramètres englobent des éléments tels que l'âge de la personne (**âge**), le sexe (**sex**), le type de douleur thoracique (**cp**), la tension artérielle au repos (**trbps**), le cholestérol (**chol**), la glycémie à jeun (**fbs**), les résultats électrocardiographiques au repos (**restecg**), la fréquence cardiaque maximale atteinte (**thalachh**), l'angine induite par l'exercice (**exng**), le pic précédent (**oldpeak**), la pente (**slp**), le nombre de vaisseaux majeurs (**caa**), le taux de Thal (**thall**), et enfin, la variable cible (**output**). Chaque paramètre est évalué ou mesuré en fonction de critères spécifiques pour fournir une compréhension détaillée des risques potentiels de maladies cardiaques.
Ces données offrent une perspective exhaustive des paramètres médicaux de chaque individu, simplifiant une analyse approfondie de leur santé cardiovasculaire. Le projet a pour objectif ultime de prédire le risque de crises cardiaques en utilisant la variable cible (**output**), qui prend la valeur 1 en cas de présence d'une crise cardiaque et 0 dans le cas contraire. Cela fournit ainsi un outil inestimable pour une évaluation proactive de la santé cardiovasculaire.

2.2 Preprocessing des donnéess

Cleaning des données :
La présence de duplications dans notre ensemble de données souligne l'importance cruciale de les éliminer.

```
[ ]  data.duplicated().sum()

     1
```

Figure 1 – Présence d'enregistrements dupliqués

```
[ ]  data[data.duplicated()]
```

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
164	38	1	2	138	175	0	1	173	0	0.0	2	4	2	1

Figure 2 – Supression d'enregistrements dupliqués

Distrubition des données :

Nous pouvons clairement voir que nos valeurs cibles sontion à peu près équitablement réparties

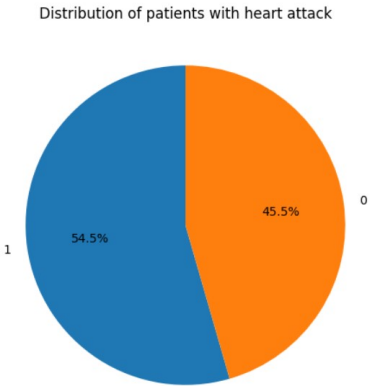

Figure 3 – Distrubition des patients

Matrice de corrélation :

À partir de cette carte de corrélation, nous pouvons constater que chol et fbs n'ont pas une corrélation significativement plus élevée avec la variable de sortie (output). Nous procédons donc à leur suppression.

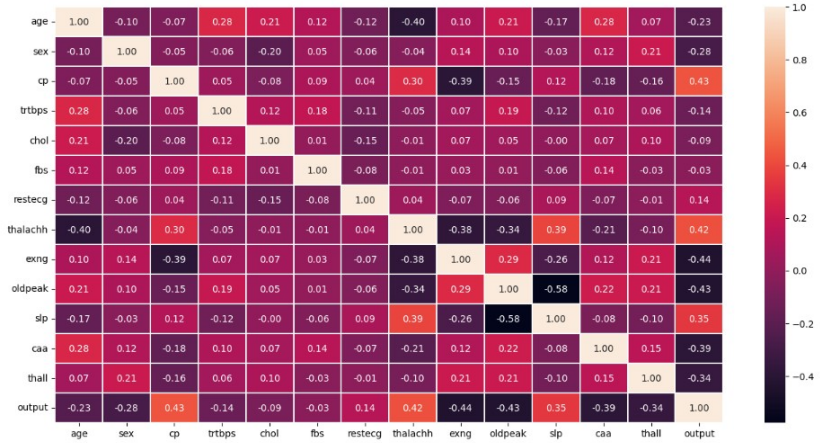

Figure 4 – Nombre de chaque classe

3 Application des algorithmes

3.1 Apprentissage supervisé

Dans cette section, nous avons appliqué les algorithmes d'apprentissage supervisé dédiés pour la classification vu la nature de notre problème. Les algorithmes utilisés sont : KNN, NB, DT, SVM et MLP (Ce dernier est un algorithme de Deep Learning).

3.1.1 K plus proches voisins (KNN)

	precision	recall	f1-score	support
0	0.70	0.76	0.73	37
1	0.82	0.78	0.80	54
accuracy			0.77	91
macro avg	0.76	0.77	0.76	91
weighted avg	0.77	0.77	0.77	91

Figure 5 – Mesure de performance de KNN

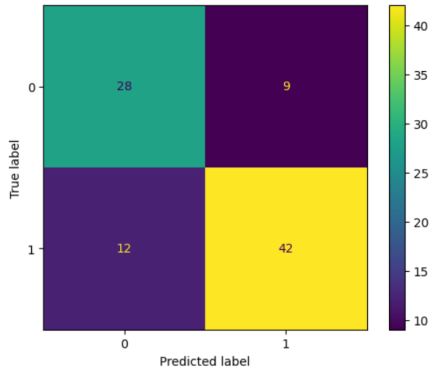

Figure 6 – Matrice de confusion de KNN

3.1.2 Naive Bayes (NB)

	precision	recall	f1-score	support
0	0.83	0.81	0.82	37
1	0.87	0.89	0.88	54
accuracy			0.86	91
macro avg	0.85	0.85	0.85	91
weighted avg	0.86	0.86	0.86	91

Figure 7 – Mesure de performance de NB

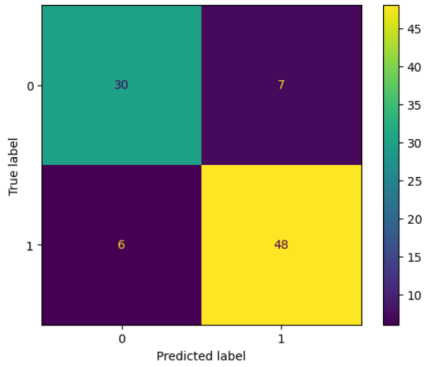

Figure 8 – Matrice de confusion de NB

3.1.3 Decision Trees (DTs)

	precision	recall	f1-score	support
0	0.80	0.76	0.78	37
1	0.84	0.87	0.85	54
accuracy			0.82	91
macro avg	0.82	0.81	0.82	91
weighted avg	0.82	0.82	0.82	91

Figure 9 – Mesure de performance de DTs

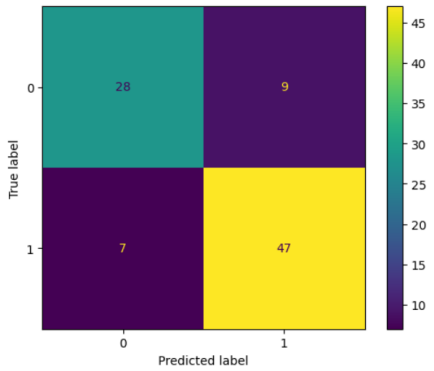

Figure 10 – Matrice de confusion de DTs

3.1.4 Support Vector Machine (SVM)

	precision	recall	f1-score	support
0	0.35	0.16	0.22	37
1	0.58	0.80	0.67	54
accuracy			0.54	91
macro avg	0.47	0.48	0.45	91
weighted avg	0.49	0.54	0.49	91

Figure 11 – Mesure de performance de SVM

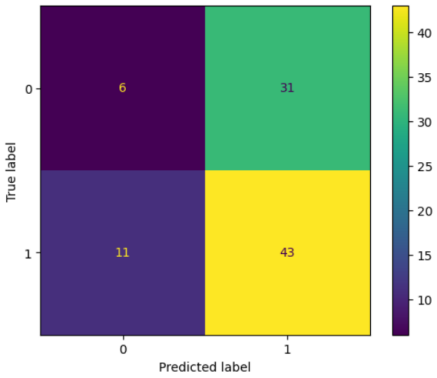

Figure 12 – Matrice de confusion de SVM

Analyse des resultats :

L'évaluation des performances des quatre algorithmes de classification - K-NN, Naive Bayes, Decision Tree, et SVM - révèle des caractéristiques distinctes dans la tâche de prédiction de la présence ou de l'absence de maladies cardiaques. Le modèle Naive Bayes affiche une remarquable précision, recall, et F1-Score pour les deux classes, suggérant une capacité élevée à discriminer entre les individus en bonne

santé et ceux atteints de maladies cardiaques. Le Decision Tree présente également des performances solides, avec un équilibre entre précision et recall. En revanche, bien que le modèle K-NN ait une précision et un recall acceptables, la matrice de confusion met en évidence des erreurs significatives, en particulier des faux positifs. Le Naive Bayes se distingue par une précision et un recall exceptionnels pour les deux classes, indiquant une fiabilité élevée dans la classification des individus. Enfin, le SVM montre des performances inférieures, marquées par un faible recall pour la classe 0, indiquant une propension à sous-estimer les individus en bonne santé.

3.1.5 Multilayer perceptron (MLP)

Cette méthode fait partie du Deep Learning (étant un type de réseau neuronal artificiel à couches multiples).

	precision	recall	f1-score	support
0	0.53	0.65	0.58	40
1	0.67	0.55	0.60	51
accuracy			0.59	91
macro avg	0.60	0.60	0.59	91
weighted avg	0.61	0.59	0.59	91

Figure 13 – Performance de MLP

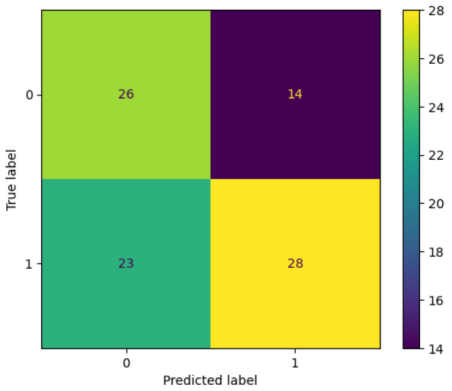

Figure 14 – Matrice de confusion pour MLP

Les meilleurs hyperparametères sont donc (à l'aide de GridSearch) : 'activation' : **'tanh'**, 'alpha' : **0.01**, 'hidden layer sizes' : **(20,)**, 'learning rate' : **'constant'**, 'max iter' : **500**, 'solver' : **'adam'**

3.2 Apprentissage non supervisé

3.2.1 K-Means

Visualiser des clusters

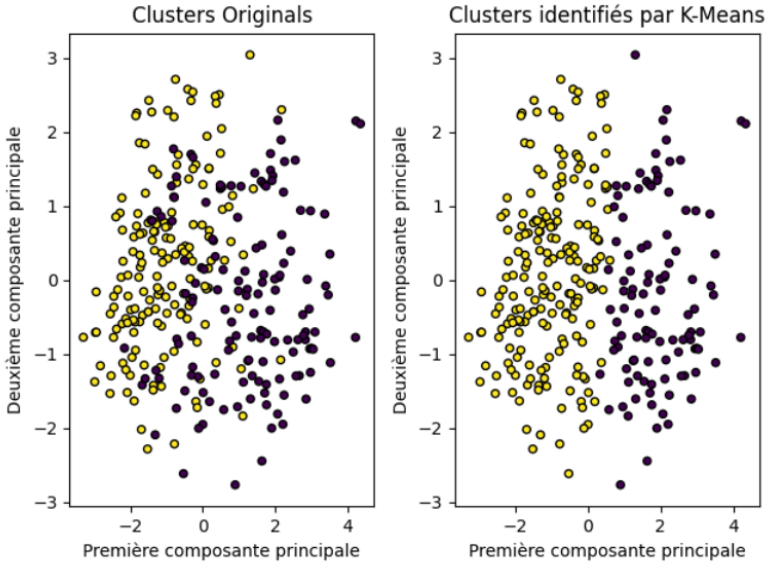

Figure 15 – Visualisation des Clusters

Determination du nombre de clusters

- Méthode du Coude (Elbow Method)

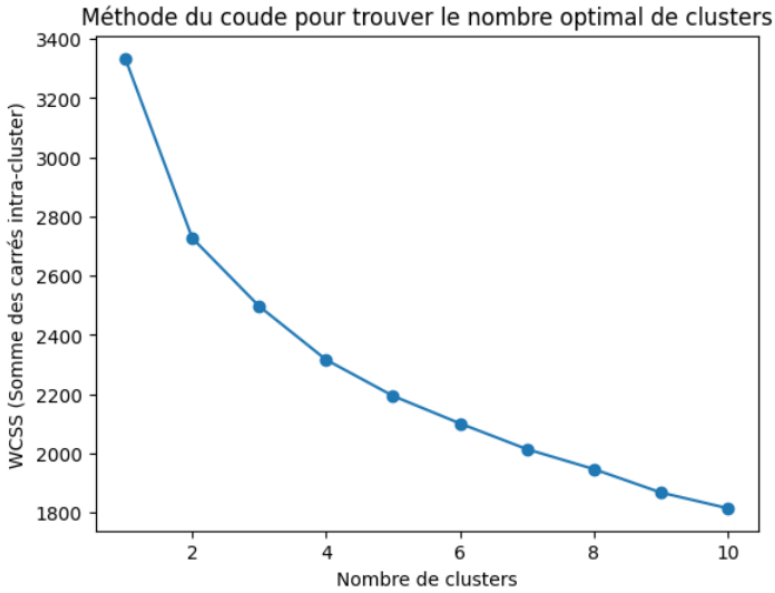

Figure 16 – Courbe representative de la méthode du Coude

Dans ce cas, on peut dire que le point du coude se situe entre 2 et 3 clusters. Cela signifie que le nombre optimal de clusters serait probablement 2 ou 3.

- Coefficient Silhouette :

```
Silhouette Score pour nombre de clusters 2 : 0.19201851636008116
Silhouette Score pour nombre de clusters 3 : 0.1326455156425572
Silhouette Score pour nombre de clusters 4 : 0.14211198435395972
Silhouette Score pour nombre de clusters 5 : 0.13519499317365175
Silhouette Score pour nombre de clusters 6 : 0.12034011534652757
Silhouette Score pour nombre de clusters 7 : 0.13400540192794824
Silhouette Score pour nombre de clusters 8 : 0.12777775750444376
Silhouette Score pour nombre de clusters 9 : 0.13150729657564664
Silhouette Score pour nombre de clusters 10 : 0.12953691190793473
```

Figure 17 – Resultat de l'application du Score Silhouette

Il semble que le score de silhouette soit le plus élevé pour 2 clusters, indiquant que la meilleure séparation des données est obtenue lorsque les individus sont regroupés en deux clusters distincts.

- Coefficient de Davies-Bouldin :

```
Coefficient de Davies-Bouldin pour nombre de clusters 2 : 1.9904430415961256
Coefficient de Davies-Bouldin pour nombre de clusters 3 : 2.269439809329281
Coefficient de Davies-Bouldin pour nombre de clusters 4 : 2.0440161877291105
Coefficient de Davies-Bouldin pour nombre de clusters 5 : 2.164001979656839
Coefficient de Davies-Bouldin pour nombre de clusters 6 : 2.0800471629047075
Coefficient de Davies-Bouldin pour nombre de clusters 7 : 2.0265553533784018
Coefficient de Davies-Bouldin pour nombre de clusters 8 : 2.0242305924243897
Coefficient de Davies-Bouldin pour nombre de clusters 9 : 1.9036329516363668
Coefficient de Davies-Bouldin pour nombre de clusters 10 : 1.8465234532729977
```

Figure 18 – Resultat de l'application du coefficient de Davies-Bouldin

Les valeurs les plus basses du coefficient de Davies-Bouldin sont généralement associées à une meilleure qualité de clustering. Ainsi, dans ce cas, le nombre optimal de clusters semble être 10, car il donne le coefficient de Davies-Bouldin le plus bas.

3.2.2 Clustering hiérarchique

```
Silhouette Score pour nombre de clusters 2 : 0.3070303987508227
Silhouette Score pour nombre de clusters 3 : 0.2116815305351958
Silhouette Score pour nombre de clusters 4 : 0.12458822807742909
Silhouette Score pour nombre de clusters 5 : 0.05613765784171341
Silhouette Score pour nombre de clusters 6 : 0.039729590759482944
Silhouette Score pour nombre de clusters 7 : 0.09397366263274165
Silhouette Score pour nombre de clusters 8 : 0.1012979158652774
Silhouette Score pour nombre de clusters 9 : 0.11485494579971113
Silhouette Score pour nombre de clusters 10 : 0.10197251566823658
```

Figure 19 – Resultat de l'application du score silhouette

Dans ce cas, le score de silhouette est le plus élevé pour 2 clusters, indiquant que la séparation en deux groupes distincts est relativement bonne.

4 Conclusion

Cette étude se concentre sur la prédiction des crises cardiaques à partir d'un ensemble de données, mettant en avant plusieurs facteurs pouvant causer de telles crises. Cette problématique revêt une importance cruciale à notre ère, où la fréquence des attaques devient beaucoup plus importante.

L'application du Machine Learning dans ce contexte se révèle essentielle en permettant une meilleure anticipation des arrêts cardiaques : les caractères qui en mènent le plus sont détectés à l'aide de lappliquation des modèles prédictifs sur les données, contibuant ainsi à prévoir, et donc de prévenir, d'éventuelles attaques futures.