الجمهورية الجزائرية الديمقراطية الشعبية

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
وزارة التعليم العالي والبحث العلمي

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
تلمسان –جامعة أبي بكر بلقايد –

Université Aboubakr Belkaïd – Tlemcen –
Faculté de TECHNOLOGIE Département de
Génie Biomédical
Laboratoire de Recherche de Génie Biomédical

MEMOIRE de projet de fin d'étude Présenté pour l'obtention du **Diplôme de Master**

En : Génie Biomédical

Spécialité : Électronique et Maintenance Biomédicale

Par :

HASNAOUI Wissem **et GACEM** Nadjat

Sujet :

Modèle intelligent pour la prédiction des anomalies cardiovasculaires

Soutenu publiquement, le 27/ 06 / 2026 , devant le jury composé de :

Dr.DiB Nabil	MCA	Université de Tlemcen	Président
M .Youbi Mohammed Redha	MCB	Université de Tlemcen	Examinateur
Benali Radwane	Prof	Université de Tlemcen	Encadreur

Année universitaire : 2025/2026

Remerciements

Louange à Allah, Le Tout-Puissant et Le Miséricordieux, qui nous a accordé la santé, la patience et la persévérance nécessaires pour mener à bien ce travail. Sans Sa grâce, l'aboutissement de ce projet n'aurait pas été possible.

Nous tenons, en premier lieu, à exprimer notre profonde reconnaissance à nos très chers parents. Leur amour inconditionnel, leur soutien constant et leurs sacrifices tout au long de notre parcours universitaire ont constitué un pilier essentiel de notre réussite.

Nous adressons nos sincères remerciements à Monsieur **BENALI RADWANE**, notre encadrant, pour la qualité de son encadrement, sa disponibilité, ainsi que pour la pertinence de ses conseils et orientations tout au long de ce projet.

Nous exprimons également notre gratitude aux membres du jury, Monsieur **DIB Nabil** et Monsieur **YOUBI Mohamed Redha,** pour l'honneur qu'ils nous ont fait en acceptant d'évaluer ce mémoire, ainsi que pour l'attention et le sérieux qu'ils ont portés à notre travail. Leurs remarques et suggestions seront précieuses pour la poursuite de notre parcours académique et professionnel.

Nous remercions par ailleurs l'ensemble des enseignants du parcours **Électronique Biomédicale** pour la qualité de leur enseignement, leur accompagnement pédagogique et les connaissances qu'ils nous ont transmises tout au long de notre formation.

Enfin, nous adressons nos remerciements à toutes les personnes qui, de près ou de loin, ont contribué à la réalisation de ce travail et nous ont soutenus tout au long de ce projet.

À tous, nous exprimons notre profonde gratitude et notre sincère reconnaissance.

Dédicace

Avant toute chose, je remercie Dieu Tout-Puissant de m'avoir accordé la santé, la patience et la force nécessaires pour mener à bien ce travail.

Je tiens à exprimer ma profonde gratitude à mes chers parents, **Hasnaoui Mohamed** et **Garmala Rachida,** pour leur amour inconditionnel, leurs sacrifices, leur soutien indéfectible et leurs encouragements constants tout au long de mon parcours. Leur confiance, leur patience et leurs précieux conseils ont été une source de motivation qui m'a permis de persévérer et d'atteindre mes objectifs.

Je remercie également ma chère sœur **Mounira** ainsi que mon frère **Wassim**, pour leur affection, leur présence et leur soutien moral qui m'ont accompagnée tout au long de cette aventure académique.

Je tiens à adresser mes sincères remerciements à ma binôme **Nadjet**, pour sa collaboration, son sérieux, son aide précieuse et son esprit d'équipe durant la réalisation de ce mémoire. Ce travail n'aurait pas été le même sans son implication et sa coopération.

Enfin, j'exprime ma reconnaissance à toutes les personnes qui m'ont soutenue et encouragée, de près ou de loin, durant mon parcours universitaire.

À toutes et à tous, j'adresse mes plus sincères remerciements et ma profonde gratitude.

HASNAOUI WISSEM .

Dédicace

Avant toute chose, je rends grâce à Dieu Tout-Puissant de m'avoir accordé la santé, la force, le courage et la patience nécessaires pour mener à bien ce travail.

Je dédie ce travail à mon grand-père **YAHIA GACEM** , qu'Allah lui fasse miséricorde. Je prie pour que Dieu lui accorde Sa clémence et le place parmi les gens du Paradis. Il restera toujours présent dans mon cœur et dans mes prières.

Je tiens à exprimer ma profonde et sincère gratitude à ma très chère mère **Karima GACEM** , pour son amour inconditionnel, ses sacrifices innombrables, ses prières et son soutien indéfectible tout au long de mon parcours. Elle a toujours été ma première source de motivation, de force et d'inspiration.

Je tiens à remercier chaleureusement mes chers oncles **Djilali GACEM** et **Hocine GACEM**, pour leur aide précieuse, leurs conseils avisés et leur soutien permanent. Leur présence et leur bienveillance ont grandement contribué à mon épanouissement personnel et académique.

Je dédie ce travail à mes chers parents, **Aïcha** et **ABDELHAK**, pour leur amour inconditionnel, leurs sacrifices et leur soutien constant tout au long de mon parcours. Leur présence et leurs encouragements ont été ma plus grande force et ma motivation.

Mes sincères remerciements vont également à mes chères sœurs **SALIMA** , **SARAH** et **WIAM** , pour leur affection, leur présence réconfortante et leurs encouragements continus qui m'ont toujours donné la force d'avancer.

Je tiens à adresser une pensée particulière à ma binôme **WISSEM**, pour sa collaboration, son sérieux, son implication et les efforts partagés durant la réalisation de ce mémoire. Travailler à ses côtés a été une expérience enrichissante qui a grandement contribué à l'aboutissement de ce travail.

Je souhaite exprimer toute ma reconnaissance à ma meilleure amie **DJINANE**, pour son amitié sincère, son soutien indéfectible, son écoute attentive et sa présence à mes côtés dans les moments de joie comme dans les moments de difficulté.

J’adresse aussi mes plus sincères remerciements à **Monsieur KOTBI MOHAMED** , pour son accompagnement, sa bienveillance, ses précieux conseils et son soutien tout au long de mon parcours.

Je remercie également mon amie **Asma**, pour son soutien, son amitié sincère et sa présence tout au long de mon parcours universitaire.

Je dédie ce travail à ma petite princesse **Nada,** qui illumine ma vie par sa douceur et son innocence. Je lui souhaite un avenir rempli de bonheur, de réussite et de beaux rêves réalisés.

Enfin, je remercie toute ma famille et mes amies , ainsi que toutes les personnes qui ont contribué, de près ou de loin, à l’aboutissement de ce travail.

À toutes et à tous, j’adresse l’expression de ma profonde reconnaissance et de ma sincère gratitude.

GACEM NADJET

RÉSUMÉ

Ce travail porte sur le développement d'un modèle intelligent de prédiction des anomalies cardiovasculaires basé sur le Machine Learning. Pour cela, nous avons utilisé une base de données cardiologique contenant des informations cliniques de patients, puis nous avons implémenté et comparé quatre algorithmes de classification : SVM, KNN, Random Forest et les réseaux de neurones. Après une optimisation par GridSearchCV et une validation croisée, les résultats ont permis d'identifier le modèle le plus performant. Cette étude montre que le Machine Learning constitue un outil prometteur pour améliorer le dépistage précoce et assister le diagnostic cardiovasculaire.

Mots-clés : Maladies cardiovasculaires, Apprentissage automatique, Diagnostic précoce, SVM, KNN, Random Forest, Réseaux de neurones, Classification médicale.

ABSTRACT

This thesis presents an intelligent model for the prediction of cardiovascular abnormalities using Machine Learning techniques. We used a standard cardiology dataset and compared four classification algorithms: SVM, KNN, Random Forest, and Artificial Neural Networks. After optimizing the models with GridSearchCV and cross-validation, the results identified the best-performing classifier. Overall, this work demonstrates the potential of Machine Learning to support early cardiovascular diagnosis and improve risk prediction.

Keywords : Cardiovascular diseases, Machine learning, Early diagnosis,SVM, KNN, Random Forest, Neural networks, Medical classification.

TABLE DES MATIÈRES

Résumé 6
Abstract..... 6
Liste des figures..... 9
Liste des Tableaux..... 12
Liste des abréviations..... 13
Introduction Générale :..... 14

Chapitre 1 : Généralités sur les maladies cardiovasculaires 17

Introduction..... 18
1.1.Anatomie et physiologie du système cardiovasculaire :..... 18
1.1.1.Structure du cœur :..... 19
1.1.2.Cavités cardiaques..... 19
1.1.3.Valves cardiaques..... 20
1.1.4.Vaisseaux sanguins..... 20
1.1.5.Le myocarde :..... 22
1.1.6.Circulation sanguine..... 23
1.1.7.Cycle cardiaque :..... 24
1.1.8.Activité électrique du cœur :..... 24
1.2.Principales anomalies cardiovasculaires :..... 26
1.2.1.Cardiopathie coronarienne 26
1.2.2.Arythmie :..... 27
1.2.3. Insuffisance cardiaque :..... 28
1.2.4. Accident vasculaire cérébral (AVC):..... 29
1.3.Facteurs de risque cardiovasculaires :..... 31
1.4.Méthodes classiques de diagnostic :..... 33
1.4.1.Électrocardiogramme (ECG) :..... 33
1.4.2.Échocardiographie cardiaque et le doppler :..... 35
1.4.3.Angiographie :..... 36
1.4.4.Test d'effort (épreuve d'effort cardiaque)..... 37
1.5.Limites des méthodes classiques :..... 38
Conclusion :..... 39

Chapitre 2 : Cadre théorique de l'apprentissage automatique appliqué à la classification médicale et présentation des bases de données cardiovasculaires..... 40

Introduction à l'intelligence artificielle en biomédical..... 41
2.1. Machine Learning en médecine prédictive..... 41
2.2. Types d'apprentissage..... 42
2.2.1. L'apprentissage supervisé..... 43
2.2.2. Apprentissage non supervisé..... 43

2.2.3. L'apprentissage par renforcement....44
2.3.Deep Learning (Apprentissage profond).... 45
2.4.Algorithmes de classification utilisés en santé.... 45
2.4.1.Régression logistique....45
2.4.2 SVM....48
2.4.3 Random Forest....50
2.4.4 K plus proches voisins 51
2.4.5 Les réseaux du nuerons 53
2.5.Métriques d'évaluation des modèle.... 54
2.6.Application du Machine Learning dans le domaine de la cardiologie....56
2.8. Analyse exploratoire des données (EDA)....62
2.9. Problèmes liés aux données médicales.... 63
2.9.1.Données manquantes....50
2.9.2Déséquilibre de classe 63
2.9.3Bruit des données....64
Conclusion 64

Chapitre 3: Approche pratique de la prédiction des maladies cardiovasculaires par apprentissage automatique : évaluation des modèles et analyse des résultats..... 66

Introduction....69
3.1.Nettoyage des données :..... 67
3.2.Gestion des valeurs manquantes :....68
3.3.Encodage des variables catégorielles :....69
3.4.Sélection des caractéristiques (Feature Selection) :....69
3.5.Normalisation / Standardisation :.... 70
3.6.Réduction de dimension (PCA – optionnel) :....70
3.7.Division de la base de données (Train / Test / Validation) :....71
3.8.Visualisation et analyse statistique :.... 71
3.8.1.Résultats du script « Nettoyage des données » :.... 72
3.8.2.Résultats du script « Gestion des valeurs manquantes» :....73
3.8.3.Résultats du script « Sélection des caractéristiques » :.... 75
3.8.4.Résultats du script « Normalisation / Standardisation » :.... 77
3.8.5.Résultats du script « Réduction de dimension (PCA) » :....78
3.8.6.Résultats du script « Division de la base de données » :....81
3.9.Description des modèles utilisée :.... 82
3.10.Parametrage et optimisation (Hyperparameter tuning)....82
3.10.1.Paramétrage et optimisation du modèle KNN....85
3.10.2.Paramétrage et optimisation du modèle SVM 86
3.10.3.Paramétrage et optimisation du modèle Random Forst.... 86
3.10.4.Paramétrage et optimisation du modèle MLP....86
3.11.La validation croisée(Cross validation):....89
3.12.Protocole expérimental commun:....91
3.13.Critères de comparaison des modèles :....91
3.14.Analyse résultats de chaque modèle :....94
3.15.Analyse comparative des performances.... 107
3.16. Sélection du modèle optimal et limites des approches concurrentes 107
Conclusion :....111
Conclusion Générale.... 112
Références Bibliographiques.... 115

Liste des figures

Figure 1.1 : Structure anatomique du cœur humain 19
Figure 1.2: Représentation anatomique des quatre valves cardiaques et leurs positions . . 20
Figure 1.3 : Structure et comparaison anatomique des vaisseaux sanguins.....21
Figure 1.4: Structure anatomique de la paroi cardiaque : endocarde, myocarde et péricarde 22
Figure 1.5: Représentation schématique de la double circulation sanguine : la petite circulation (pulmonaire) et la grande circulation (systémique) 23
Figure 1.6: Corrélation temporelle entre les phases mécaniques du cycle cardiaque (systole et diastole) et l'activité électrique (tracé ECG) 25
Figure 1.7: Artère coronaire : de l'état normal à la sténose 27
Figure 1.8: Rythme Cardiaque Normal versus arythmie 28
Figure 1.9: Conséquences anatomiques de la cardiomyopathie dilatée menant à l'insuffisance cardiaque 29
Figure 1.10:Différenciation physiopathologique des accidents vasculaires cérébraux : AVC ischémique (occlusion) versus AVC hémorragique (rupture) 30
Figure 1.11: Classification des facteurs de risque des maladies cardiovasculaires selon leur caractère modifiable 33
Figure 1.12 : Anatomie cardiaque et correspondance avec les ondes de l'électrocardiogramme (ECG)…….36
Figure 1.13 : Echographie cardiaque…….37
Figure 1.14 : Rétrécissement de l'artère coronaire observé par coronarographie……...…..39
Figure 2.1 :Types d'apprentissage 42
Figure 2.2: Illustration de la différence entre le Machine Learning et le Deep Learning ... 45
Figure 2.3: Régression logistique 47
Figure 2.4: Principe de SVM. 48
Figure 2.5:Exemple Random forest 50
Figure 2.6: Exemple de KNN 51
Figure 2.7: Exemple de réseaux du nuerons 53
Figure 2.8: Métriques d'évaluation des modèles 56
Figure 2.9:Taille et aperçu initial de la base de données brute *Cleveland Heart Disease*

Dataset. 59
Figure 2.10:Taille et aperçu initial de la base de données brute Framingham Heart Study Dataset 60
Figure 2.11:Taille et aperçu initial de la base de données brute Dataset Cardiovascular 61
Figure 3.1: Aperçu général et structure des variables du *Cleveland Heart Disease Dataset*. 71
Figure 3.2: Impact du nettoyage des données, avec focus sur la correction des valeurs aberrantes de la variable trestbps 72
Figure 3.3: Impact du traitement des valeurs manquantes sur le *Cleveland Heart Disease Dataset* 74
Figure 3.4: Évolution de la structure de la base de données après application de l'encodage One-Hot 75
Figure 3.5: Analyse et classement de l'importance des variables 76
Figure 3.6: Classement des six variables cliniques les plus discriminantes issues du test statistique de sélection 78
Figure 3.7: Effet de la standardisation par StandardScaler sur la distribution des variables numériques thalach et oldpeak 77
Figure 3.8:Comparaison de la distribution spatiale des patients entre les variables initiales et le plan factoriel PC1 vs PC2 78
Figure 3.9: Évaluation de la variance expliquée cumulée et réduction de la structure mathématique du jeu de données 80
Figure 3.10: Analyse statistique et visuelle du triple partitionnement stratifié du dataset. . 81
Figure 3.11: Résultat de l'optimisation des hyperparamètres du modèle KNN par Grid Search avec validation croisée 83
Figure 3.12: Résultat de l'optimisation des hyperparamètres du modèle SVM par Grid Search avec validation croisée 84
Figure 3.13 : Résultat de l'optimisation des hyperparamètres du modèle Random forest par Grid Search avec validation croisée 86
Figure 3.14: Résultat de l'optimisation des hyperparamètres du modèle Réseaux de Neurones par Grid Search avec validation croisée 86
Figure 3.15 : la technique GridSearch 89
Figure 3.16 : validation croisée 90
Figure 3.17 : Exemple de matrice confusion 92
Figure 3.18: Matrice de confusion du modèle SVM illustrant la répartition des prédictions

correctes et erronées 94
Figure 3.19: Bilan chiffré des métriques d'évaluation du modèle SVM sur l'échantillon de test 95
Figure 3.20: Courbe ROC et mesure de l'aire sous la courbe (AUC) pour l'architecture SVM 96
Figure 3.21: Matrice de confusion du modèle KNN illustrant la répartition des prédictions correctes et erronées 97
Figure 3.22: Bilan chiffré des métriques d'évaluation du modèle KNN sur l'échantillon de test 98
Figure 3.23: Courbe ROC et mesure de l'aire sous la courbe (AUC) pour l'architecture KNN 99
Figure 3.24: Matrice de confusion du modèle Random forest illustrant la répartition des prédictions correctes et erronées 100
Figure 3.25: Bilan chiffré des métriques d'évaluation du modèle Random forestsur l'échantillon de test 101
Figure 3.26: Courbe ROC et mesure de l'aire sous la courbe (AUC) pour l'architecture Random forest 102
Figure 3.27: Matrice de confusion du modèle MLP illustrant la répartition des prédictions correctes et erronées 103
Figure 3.28: Bilan chiffré des métriques d'évaluation du modèle MLP l'échantillon de test 104
Figure 3.29: Courbe ROC et mesure de l'aire sous la courbe (AUC) pour l'architecture MLP 105

Liste des Tableaux

Tableau 2.1 : Synthèse comparative des differentes base de données…….66
Tableau 3.1: Description des modèles.....87
Tableau 3.2: Hyperparamètres du modèle K-Nearest Neighbors (KNN)..... ..89
Tableau 3.3: Hyperparamètres du modèle Support Vector Machine (SVM)……..……90
Tableau 3.4: Hyperparamètres du modèle Random Forest..... 91
Tableau 3.5: Hyperparamètres des Réseaux de Neurones Artificiels..... 93
Tableau 3.6 : Métriques de performance et interprétations cliniques du modèle SVM sur l'échantillon de test..... 101
Tableau 3.7: Métriques de performance et interprétations cliniques du modèle KNN sur l'échantillon de test..... 105
Tableau 3.8: Métriques de performance et interprétations cliniques du modèle Random Forest sur l'échantillon de test..... 108
Tableau 3.9: Métriques de performance et interprétations cliniques du modèle MLP sur l'échantillon de test..... 112
Tableau 3.10: Synthèse comparative des performances des modèles d'apprentissage automatique..... 113

Liste des abréviations

- **IA** : Intelligence Artificielle
- **ML** : Machine Learning (Apprentissage automatique)
- **DL** : Deep Learning (Apprentissage profond)
- **KNN** : K-Nearest Neighbors (K plus proches voisins)
- **SVM** : Support Vector Machine
- **RF** : Random Forest
- **NB** : Naive Bayes
- **LR** : Logistic Regression (Régression logistique)
- **EDA** : Exploratory Data Analysis (Analyse exploratoire des données)
- **CV** : Cross Validation (Validation croisée)
- **ACC** : Accuracy (Précision)
- **ROC** : Receiver Operating Characteristic
- **AUC** : Area Under Curve
- **TP** : True Positive (Vrai positif)
- **TN** : True Negative (Vrai négatif)
- **FP** : False Positive (Faux positif)
- **FN** : False Negative (Faux négatif)
- **CSV** : Comma-Separated Values (Fichier de données)
- **API** : Application Programming Interface

Introduction Générale

Autrefois centrée principalement sur la recherche de remèdes, la médecine moderne, portée par les avancées technologiques et scientifiques, s'oriente désormais vers une approche plus préventive. En effet, prévenir vaut mieux que guérir, et toute prévention efficace repose avant tout sur la capacité à prédire l'apparition des pathologies avant qu'elles ne se manifestent cliniquement.

Au delà se trouve notre sujet , d'où les maladies cardiovasculaires (MCV) sont la première cause de mortalité dans le monde .

Elles regroupent les états pathologiques du cœur et des vaisseaux (infarctus, AVC, insuffisance cardiaque...), liées à l'athérosclérose et aux facteurs de risque modifiables (hypertension, tabac, diabète...) [1] .

En 2022, l'OMS estime 19,8 millions de décès annuels (32% des morts mondiales), dont 85% par crise cardiaque ou AVC ; ce nombre a augmenté de 60% depuis 1990 [1],[2] .

Pourquoi cette augmentation et pourquoi être le "leader mondial" de tous les temps malgré la présence persistante du cancer et du COVID-19 au passé ?

Réellement la cause c'est l'évolution asymptomatique et le silence clinique qui ne cesse qu'à un stade avancé de la pathologie, lorsque des dommages irréversibles ont déjà été causés au myocarde ou aux vaisseaux sanguins.

Vous vous disiez probablement que la médecine à évolué et nous pouvons compter sur les méthodes de diagnostic traditionnelles , mais bien que performantes, elles présentent certaines limitations qui entravent une détection optimale , sans parler de leur interprétation qui repose largement sur l'expertise du clinicien et leur nécessité d'un équipement spécialisé et un accès à des structures médicales équipées, ce qui limite leur disponibilité dans les zones rurales ou défavorisées.

Le réflexe universel de notre ère est de s'orienter vers l'IA pour surmonter ces enjeux et découvrir cette approche préventive , car tous simplement un modèle intelligent formé sur des données représentatives peut surpasser les méthodes traditionnelles de diagnostic , en rendant la médecine plus personnalisée .

Les techniques d'apprentissage automatique permettent d'analyser des quantités

importantes de données médicales complexes , en intégrant de multiples variables (données cliniques, résultats d'examens, facteurs de risque) et en exploitant les signaux des ECG et d'images médicales , du coup , nous nous dirigeons vers une réduction des coûts et une meilleure accessibilité au diagnostic.

À cet effet, ce projet vise à :

- Développer un modèle basé sur des algorithmes d'apprentissage automatique ;
- Comparer plusieurs approches d'apprentissage automatique ;
- Optimiser les paramètres des modèles afin d'améliorer la précision de prédiction ;
- Évaluer les performances à l'aide de métriques appropriées .

Dans ce mémoire , la méthodologie adoptée suit une démarche rigoureuse et structurée en plusieurs phases successives. Dans un premier temps, nous allons voir un aperçu des maladies cardiovasculaires , découvrir les techniques d'intelligence artificielle qui aident la médecine , ainsi que les bases de données disponibles pour la recherche en cardiologie.

La phase suivante consistera à acquérir et à pré-traiter les données cardiovasculaires. À ce stade, il faut nettoyer les données brutes, s'occuper des valeurs manquantes, normaliser les variables et préparer les ensembles d'entraînement et de test. Nous allons nous concentrer sur la qualité des données, car la qualité des données influence directement les performances des modèles que nous développons.

Par la suite, nous implémenterons et entraînerons plusieurs modèles d'apprentissage automatique sur les données. Il y aura une évaluation de chaque algorithme avec des métriques standardisées (la précision, la sensibilité, la spécificité, le score F1 et l'aire sous la courbe ROC). L'analyse comparative des résultats nous permettra de sélectionner le modèle le plus performant et d'identifier les facteurs influençant la qualité des prédictions.

Pour finir, une discussion sur les résultats mettra les observations en lien avec les études déjà publiées.

Cette structure se consiste de trois chapitres principaux, suivis d'une conclusion

générale , comme suit :

Le Chapitre 1 sera consacré à une analyse détaillée des maladies cardiovasculaires, abordant leurs facteurs de risque, leurs manifestations cliniques et les méthodes de diagnostic conventionnelles.

Le Chapitre 2 posera le cadre théorique de l'apprentissage automatique appliqué à la médecine et présente les bases de données cardiovasculaires. Nous y présentons les concepts fondamentaux du Machine Learning, les algorithmes de classification retenus, les métriques d'évaluation, ainsi qu'une analyse exploratoire des données cliniques sélectionnées.

Le Chapitre 3 détaille l'approche pratique de la prédiction des maladies cardiovasculaires. Nous y décrivons toutes les étapes de prétraitement (nettoyage, gestion des valeurs manquantes, encodage et standardisation), le protocole expérimental d'optimisation par validation croisée, ainsi que l'analyse comparative et la discussion des performances de chaque modèle.

Enfin, la **CONCLUSION GÉNÉRALE** synthétisera les principales contributions de ce travail, discutera ses limites et ouvrira des perspectives pour des recherches futures.

Chapitre 1 :

Généralités sur les maladies cardiovasculaires

Chapitre 1

Généralités sur les maladies cardiovasculaires

Introduction :

Les maladies cardiovasculaires, également connues sous le nom de maladies cardiaques et circulatoires, regroupent une catégorie de maladies qui impliquent le cœur et les vaisseaux sanguins , y compris les parties les plus internes du cerveau.

Ces problèmes de santé peuvent entraîner des crises cardiaques, des douleurs thoraciques, comme l'angine de poitrine, ou même des accidents vasculaires cérébraux , certaines formes de MCV affectent même les valves ou le rythme cardiaque .

Pour avoir une bonne compréhension des maladies cardiovasculaires, il est essentiel de connaître le fonctionnement du cœur et de ses vaisseaux, qui comprennent les veines et les artères[3] .

Dans ce chapitre nous essayons de présenter quelques généralités de base sur la l'anatomie et la physiologie du système cardiovasculaire, puis de décrire les principales anomalies cardiovasculaires ainsi que leurs facteurs de risque. Enfin, les méthodes classiques de diagnostic utilisées en pratique clinique seront présentées, ainsi que leurs principales limites.

1.1 Anatomie et physiologie du système cardiovasculaire :

Le système cardiovasculaire, également appelé système circulatoire, est un circuit fermé responsable du transport de l'oxygène, des nutriments et des hormones vers les cellules, ainsi que de l'élimination des déchets métaboliques (CO_2).

Ce système est une unité fonctionnelle composée de trois éléments : le Cœur (la pompe) , les Vaisseaux Sanguins (le réseau) , et le Sang (le fluide) [4].

1.1.1 Structure du cœur

Le cœur est un organe musculaire creux situé dans la cavité thoracique, entre les deux poumons, au niveau du médiastin et repose sur le diaphragme (muscle séparant le thorax de l'abdomen). Il agit comme une pompe assurant la circulation du sang dans tout l'organisme. Sa paroi est constituée de trois couches principales : l'endocarde (couche interne), le myocarde (couche musculaire responsable de la contraction) et le péricarde (enveloppe protectrice). Le myocarde joue un rôle central dans le fonctionnement cardiaque, car il est composé de cellules musculaires spécialisées (cardiomyocytes) capables de se contracter de manière rythmique et autonome. Toute altération de cette structure(Figure 1.1), notamment au niveau du myocarde, peut entraîner des dysfonctionnements cardiaques importants, comme l'insuffisance cardiaque [5].

1.1.3 Cavités cardiaques

Le cœur est divisé en quatre cavités (Figure 1.1) : deux oreillettes (droite et gauche) et deux ventricules (droit et gauche). Les oreillettes reçoivent le sang provenant soit de l'organisme (oreillette droite), soit des poumons (oreillette gauche). Les ventricules, quant à eux, assurent l'éjection du sang vers les poumons (ventricule droit) ou vers l'ensemble du corps (ventricule gauche). Le ventricule gauche possède une paroi plus épaisse en raison de la pression élevée nécessaire pour propulser le sang dans la circulation systémique. Une anomalie dans la structure ou le fonctionnement de ces cavités peut perturber l'efficacité de la pompe cardiaque[6] .

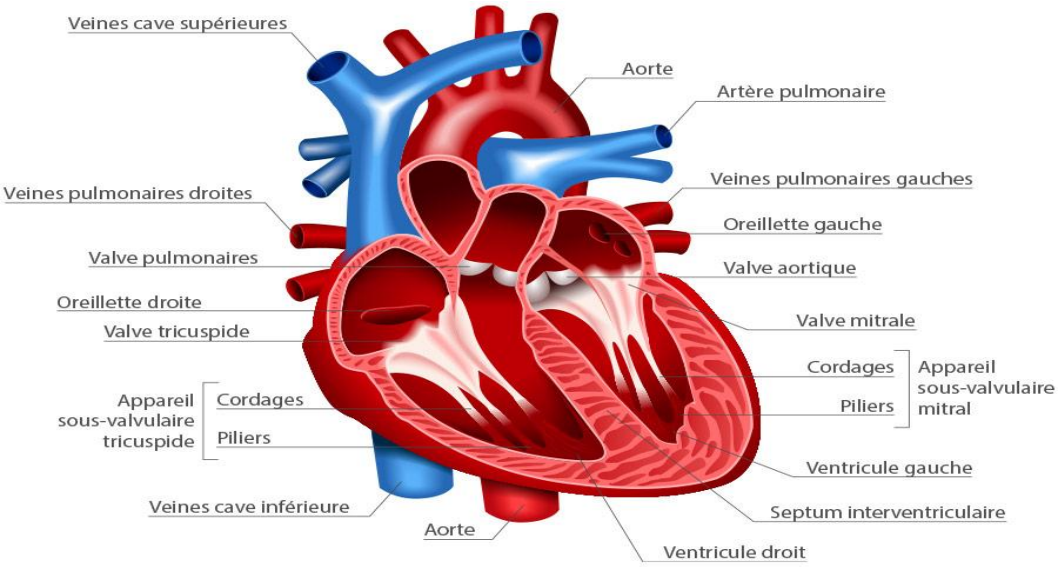


Figure 1.1 : Structure anatomique du cœur humain [7] .

1.1.4 Valves cardiaques

Les valves cardiaques assurent la circulation unidirectionnelle du sang à travers les différentes cavités du cœur. On distingue les valves auriculo-ventriculaires (tricuspide à droite et mitrale à gauche) et les valves sigmoïdes (aortique et pulmonaire)(Figure 1.2) . Elles s'ouvrent et se ferment en fonction des variations de pression entre les cavités. Une défaillance valvulaire (sténose ou insuffisance) peut entraîner des troubles hémodynamiques et contribuer au développement de pathologies cardiovasculaires [8].

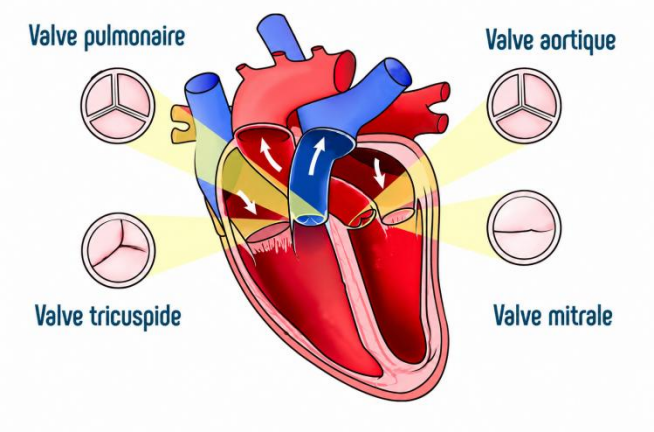


Figure 1.2: Représentation anatomique des quatre valves cardiaques et leurs positions [9] .

1.1.5 Vaisseaux sanguins

Le système vasculaire est constitué de trois types principaux de vaisseaux (Figure 1.3) :

- ✓ Les artères, qui transportent le sang du cœur vers les organes sous haute pression (aorte, artère pulmonaire) .
- ✓ Les veines, qui assurent le retour du sang vers le cœur (veines caves, veines pulmonaires) .
- ✓ Les capillaires, qui permettent les échanges entre le sang et les tissus .
- ✓ Les artères coronaires, en particulier, jouent un rôle essentiel en assurant l'irrigation du myocarde. Leur obstruction est à l'origine de la cardiopathie coronarienne, une des principales maladies cardiovasculaires.

Les vaisseaux sanguins sont constitués de trois parois :

- ✓ L'intima est au contact direct du sang. C'est une membrane mince, lisse, élastique et étanche.
- ✓ La média est constituée de fibres musculaires lisses, de fibre de collagène et de fibres d'élastine. C'est une tunique très élastique.
- ✓ L'adventice est la couche la plus externe , elle reçoit les terminaisons nerveuses.

Les artères possèdent une paroi plus épaisse, avec une abondance d'élasthanne et d'élastine dans les grosses artères (aorte, artères carotides) et davantage de cellules musculaires dans les artères musculaires . Les veines ont une tunique média plus mince, moins d'élastine et des valvules veineuses dans les veines périphériques . Les capillaires, quant à eux, ne comportent qu'une fine couche endothéliale reposant sur une membrane basale, sans muscle lisse ni fibres élastiques [10].

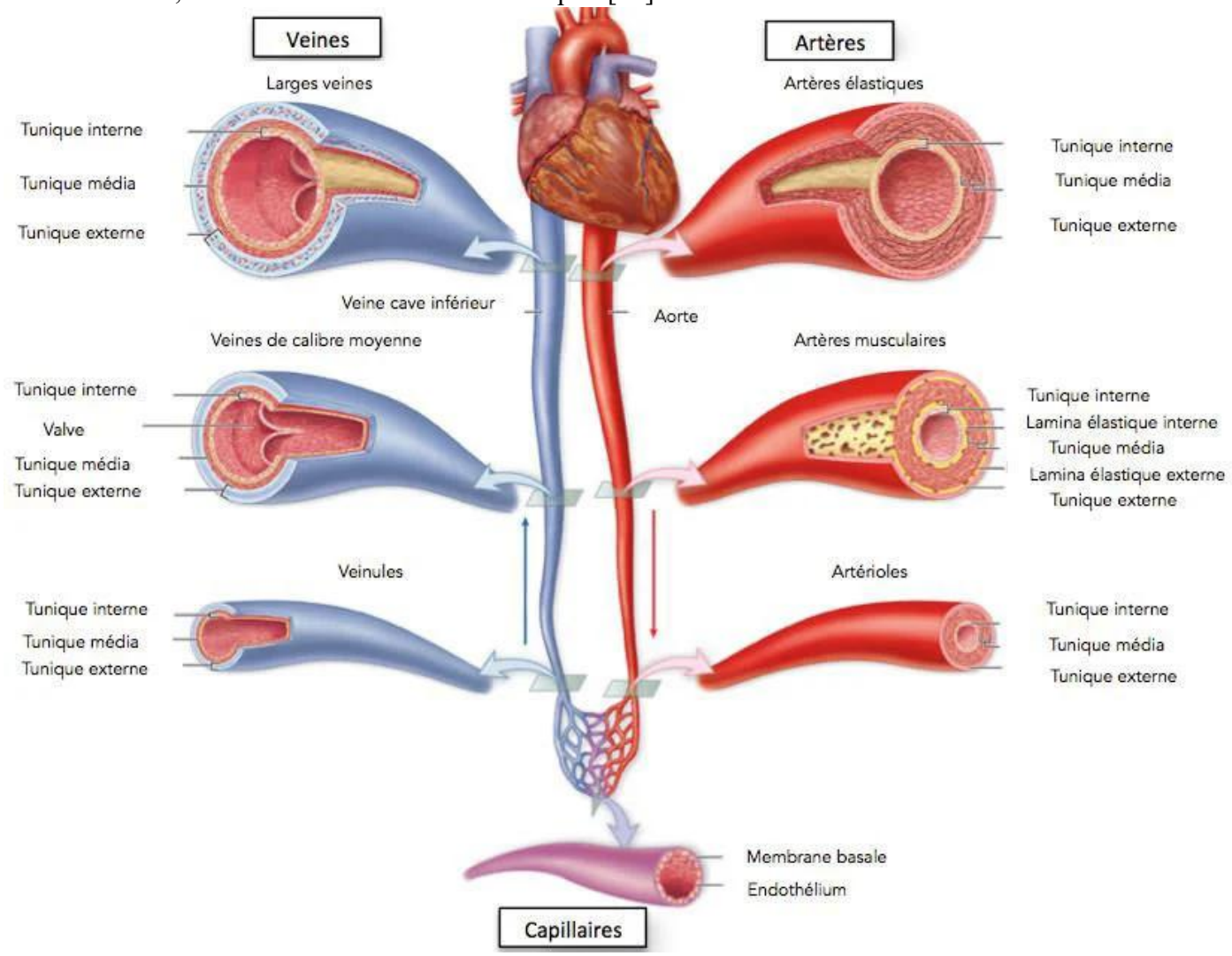


Figure 1.3 : Structure et comparaison anatomique des vaisseaux sanguins (artères, veines et capillaires) [11] .

1.1.6 Le myocarde

Le myocarde constitue la couche musculaire intermédiaire du cœur (Figure 1.4) et représente la partie fonctionnelle responsable de la contraction cardiaque. Il est formé de cellules musculaires spécialisées appelées cardiomyocytes, qui possèdent des propriétés uniques : l'automatisme (capacité à générer spontanément des impulsions électriques), l'excitabilité, la conductibilité et la contractilité. Ces propriétés permettent au cœur de fonctionner de manière autonome et rythmique, sans intervention directe du système nerveux central.

Les cardiomyocytes assurent la transmission rapide de l'influx électrique et la synchronisation des contractions. Cette organisation permet au myocarde de se comporter comme un syncytium fonctionnel, garantissant une contraction coordonnée et efficace des différentes cavités cardiaques .L'épaisseur du myocarde varie selon les régions du cœur : elle est plus importante au niveau du ventricule gauche en raison de la pression élevée nécessaire pour propulser le sang dans la circulation systémique, alors qu'elle est plus fine au niveau des oreillettes et du ventricule droit.Le myocarde est également fortement vascularisé par les artères coronaires, qui assurent l'apport en oxygène et en nutriments indispensables à son activité continue. Une diminution de cet apport, due par exemple à une obstruction des artères coronaires, peut entraîner une ischémie myocardique .

Sur le plan fonctionnel, la contraction du myocarde est déclenchée par une activité électrique issue du système de conduction cardiaque. Cette activité permet une contraction coordonnée des oreillettes puis des ventricules, assurant un débit sanguin efficace. Toute altération de cette activité, qu'elle soit d'origine structurelle ou électrique, peut conduire à des dysfonctionnements cardiaques tels que les arythmies ou l'insuffisance cardiaque [12].

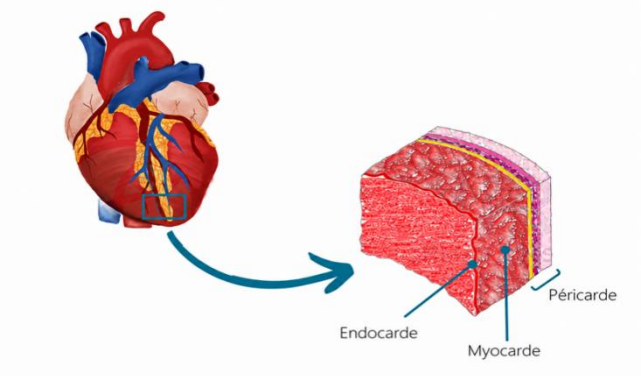


Figure 1.4: Structure anatomique de la paroi cardiaque : endocarde, myocarde et péricarde[13] .

1.1.7 Circulation sanguine

La circulation sanguine assure plusieurs fonctions vitales : le transport de l'oxygène et des nutriments vers les cellules, l'élimination des déchets métaboliques, ainsi que la régulation de la température et de l'équilibre interne de l'organisme (homéostasie).

Elle est organisée en deux circuits complémentaires (Figure 1.5) :

- ✓ La circulation pulmonaire, qui transporte le sang pauvre en oxygène vers les poumons pour son oxygénation .
- ✓ La circulation systémique, qui distribue le sang oxygéné à l'ensemble des organes.

Un dysfonctionnement de cette circulation peut entraîner des pathologies graves telles que les accidents vasculaires cérébraux (AVC) ou les maladies coronariennes [14].

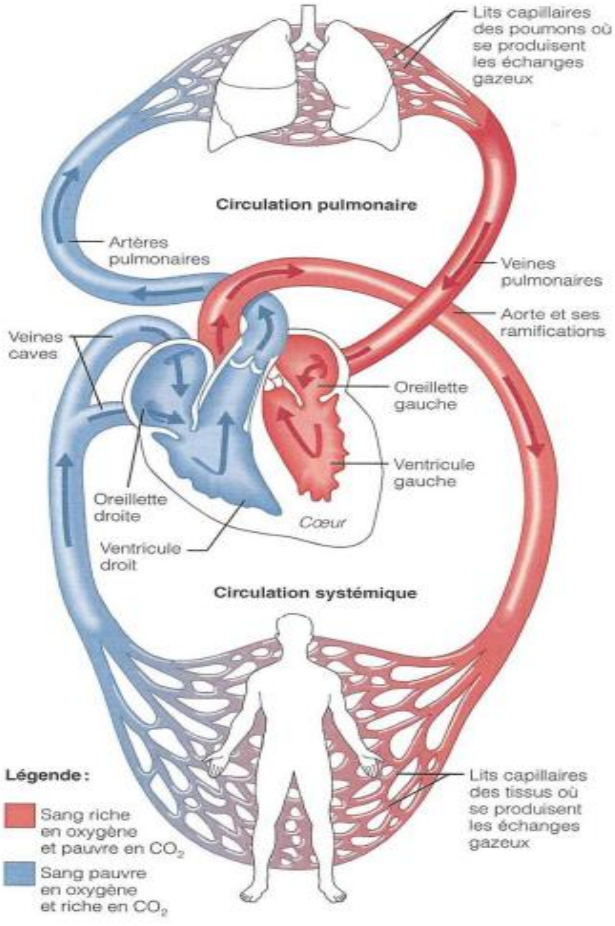


Figure 1.5: Représentation schématique de la double circulation sanguine : la petite circulation (pulmonaire) et la grande circulation (systémique) [15].

1.1.8 Cycle cardiaque

Le cycle cardiaque correspond à l'ensemble des événements mécaniques (Figure 1.6) qui se déroulent au cours d'un battement du cœur. Il comprend deux phases principales :

- ✓ La diastole (phase de remplissage) correspond à la phase de relaxation du muscle cardiaque. Durant cette période, le myocarde est relâché, ce qui permet aux cavités cardiaques de se remplir de sang. Les valves auriculo-ventriculaires (mitrale et tricuspide) sont ouvertes, facilitant le passage du sang des oreillettes vers les ventricules. En revanche, les valves sigmoïdes (aortique et pulmonaire) restent fermées afin d'empêcher tout reflux sanguin depuis les artères. Le remplissage ventriculaire se fait principalement de manière passive, grâce à l'écoulement du sang depuis les oreillettes. En fin de diastole, une contraction des oreillettes intervient, permettant d'optimiser le remplissage des ventricules.
- ✓ La systole (phase d'éjection) correspond à la phase de contraction du cœur, principalement des ventricules. Cette contraction entraîne une augmentation de la pression intraventriculaire, provoquant la fermeture des valves auriculo-ventriculaires, ce qui est à l'origine du premier bruit cardiaque (B1). Lorsque la pression ventriculaire devient supérieure à celle des artères, les valves sigmoïdes s'ouvrent, permettant l'éjection du sang vers l'aorte pour la circulation systémique et vers l'artère pulmonaire pour la circulation pulmonaire. Cette phase assure ainsi la propulsion efficace du sang vers les différents organes [14].

1.1.9 Activité électrique du cœur

Le fonctionnement du cœur est contrôlé par un système de conduction électrique (Figure 1.6) qui assure la coordination des contractions des différentes cavités cardiaques. Le muscle cardiaque se contracte en réponse à des impulsions électriques générées de manière autonome. Le nœud sinusal, considéré comme le stimulateur principal du cœur, est localisé à la jonction entre la veine cave supérieure et l'oreillette droite. Il produit de façon rythmique des impulsions électriques, à une fréquence d'environ 60 à 100 battements par minute, imposant ainsi le rythme sinusal normal. Le signal électrique se propage ensuite vers l'oreillette gauche , puis traverse le muscle de l'oreillette droite jusqu'au nœud auriculo-ventriculaire (nœud AV) .

Le nœud AV joue un rôle essentiel en ralentissant la conduction de l'influx

électrique avant sa transmission au faisceau de His. Ce retard physiologique est indispensable, car il permet aux oreillettes de se contracter et de se vider complètement dans les ventricules avant le début de la contraction ventriculaire. Le faisceau de His se divise ensuite en branches droite et gauche, qui se ramifient en un réseau dense de fibres de Purkinje. Ce système de conduction, appelé réseau His-Purkinje, assure une propagation rapide et homogène de l'influx électrique à travers les ventricules, garantissant ainsi une contraction synchronisée et efficace du muscle cardiaque.

Le tissu nodal, qui comprend le nœud sinusal et le nœud auriculo-ventriculaire, est spécialisé dans la génération spontanée de l'influx électrique (automatisme) ainsi que dans sa propagation (conductibilité). Le nœud sinusal détermine le rythme de base du cœur, tandis que le nœud AV agit comme un relais modulant la transmission du signal, avec une fréquence intrinsèque plus faible, généralement comprise entre 40 et 60 battements par minute[16].

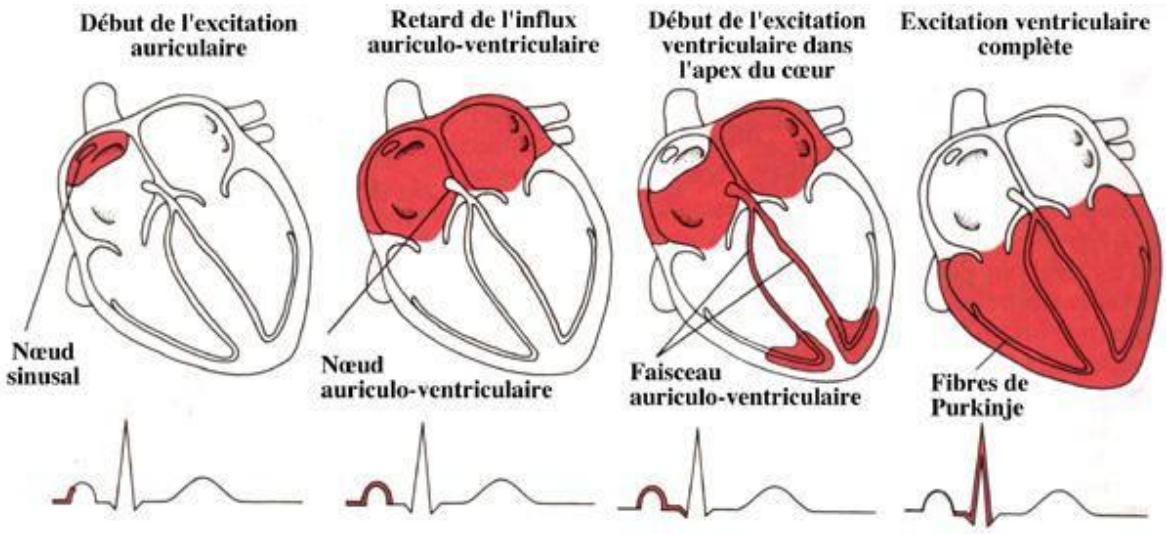


Figure 1.6: Corrélation temporelle entre les phases mécaniques du cycle cardiaque (systole et diastole) et l'activité électrique (tracé ECG) [17].

1.2. Principales anomalies cardiovasculaires :

1.2.1 Cardiopathie coronarienne

La cardiopathie coronarienne est une affection chronique des artères coronaires, caractérisée par l'accumulation progressive de plaques d'athérome au niveau de leur paroi, entraînant un rétrécissement du calibre vasculaire (Figure 1.7) et une diminution du flux sanguin vers le myocarde. Ce processus débute généralement par une altération de l'endothélium vasculaire, favorisée par des facteurs de risque tels que le tabagisme, l'hypertension artérielle, le diabète. Cette lésion permet l'infiltration du LDL cholestérol dans la paroi artérielle ; une fois oxydé, celui-ci déclenche une réaction inflammatoire impliquant des macrophages qui s'accumulent et contribuent à la formation de cellules spumeuses. L'ensemble de ces phénomènes aboutit à la constitution d'une plaque d'athérome qui s'épaissit progressivement, durcit et peut finir par obstruer partiellement ou totalement l'artère. Sur le plan physiopathologique, la réduction du diamètre artériel limite l'apport en oxygène au muscle cardiaque, provoquant une ischémie myocardique. Cliniquement, cela se manifeste le plus souvent par une angine de poitrine, caractérisée par une douleur survenant à l'effort et disparaissant au repos. En cas de rupture de la plaque d'athérome, un thrombus peut se former brutalement et entraîner une occlusion complète de l'artère, responsable d'un infarctus du myocarde. D'autres complications peuvent également survenir, telles que l'insuffisance cardiaque, les troubles du rythme ou la mort subite.Les manifestations cliniques peuvent inclure, outre la douleur thoracique, une dyspnée d'effort, une fatigue inhabituelle, des sueurs ou des nausées. Chez certains patients, notamment les femmes, les personnes âgées ou les diabétiques, les symptômes peuvent être atypiques, rendant le diagnostic plus difficile. Ainsi, la reconnaissance précoce des signes et la prise en charge des facteurs de risque jouent un rôle essentiel dans la prévention des complications graves associées à cette pathologie [18].

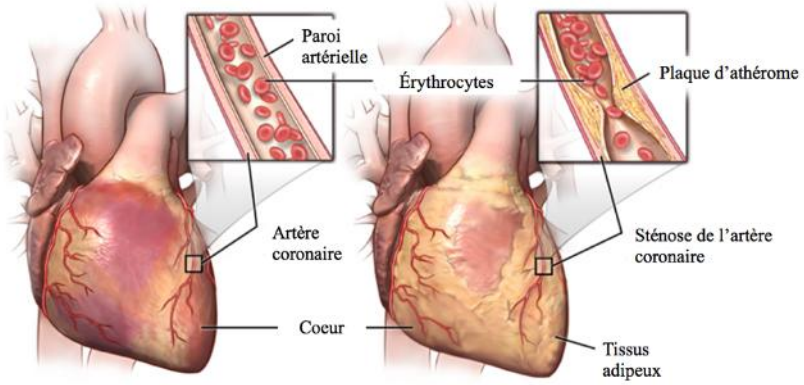


Figure 1.7: Artère coronaire : de l'état normal à la sténose [19].

1.2.2 Arythmie

Une arythmie cardiaque se définit comme toute anomalie du rythme (Figure 1.8) normal du cœur, résultant d'une perturbation de la formation ou de la propagation de l'influx électrique au sein du myocarde. Elle peut se traduire par une modification de la fréquence cardiaque (tachycardie ou bradycardie), de la régularité ou de la conduction de l'activation électrique, normalement initiée au niveau du nœud sinusal et transmise à travers le système de conduction cardiaque.

Sur le plan physiopathologique, elle regroupe les troubles de l'automatisme, de l'activité déclenchée et de la réentrée .

La réentrée (le plus courant, ~70% des cas) est comme un court-circuit électrique : l'impulsion normale tourne en rond dans un circuit défectueux à cause d'une zone de conduction lente ou bloquée comme une cicatrice après infarctus , créant une boucle infinie au lieu d'une onde unique .

L'automaticité anormale fait que des cellules "oubliées" se mettent à battre toutes seules trop vite, comme un faux pacemaker : après une excitation anormale, elles se dépolarisent spontanément via des courants ioniques déréglés généralement a cause des mutations génétiques .

L'activité déclenchée provient d'ondes parasites après le battement normal (post-dépolarisations) et un excès de calcium intracellulaire fait "sursauter" la cellule , déclenchant un nouveau battement prématuré .

Ces perturbations peuvent être induites par divers facteurs, notamment l'ischémie myocardique, les cardiopathies structurelles (fibrose, hypertrophie), les déséquilibres électrolytiques, certaines anomalies génétiques ou encore l'effet de médicaments. Elles entraînent une instabilité électrique du cœur, susceptible de provoquer différentes formes d'arythmies, allant des tachycardies supraventriculaires et de la fibrillation auriculaire aux arythmies ventriculaires graves.

Sur le plan fonctionnel, les arythmies peuvent altérer l'hémodynamique en réduisant le débit cardiaque, en favorisant la stase sanguine et la formation de thrombus, ou en provoquant un collapsus circulatoire dans les formes les plus sévères. Cliniquement, elles se manifestent par des palpitations, une fatigue, une dyspnée, des étourdissements,

des syncopes, bien que certains patients puissent rester asymptomatiques [20].

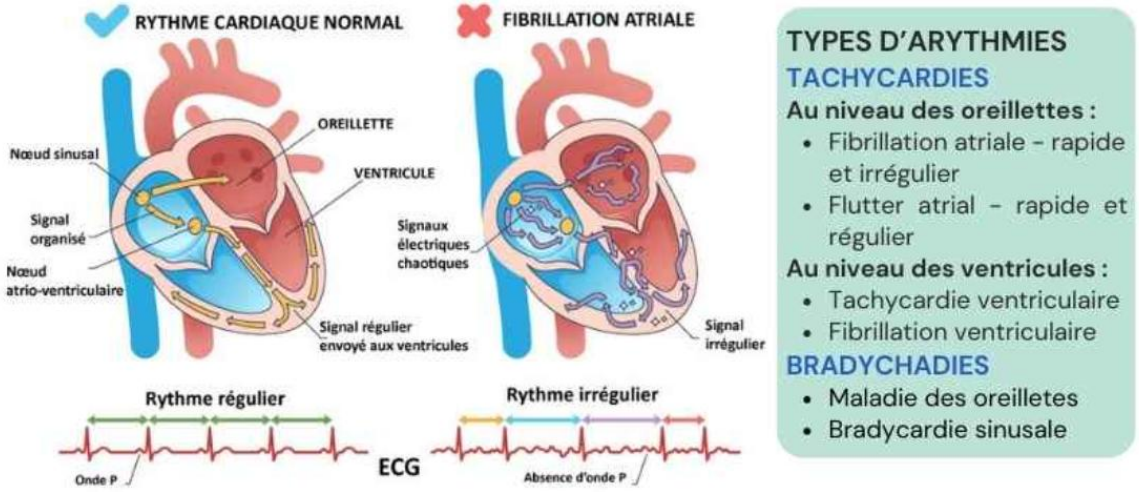


Figure 1.8 : Rythme Cardiaque Normal versus arythmie [21].

1.2.3 Insuffisance cardiaque

L'insuffisance cardiaque (IC) est un syndrome clinique complexe caractérisé par l'incapacité du cœur à assurer un débit sanguin suffisant pour répondre aux besoins métaboliques de l'organisme. Elle ne constitue pas une maladie unique, mais représente le stade final de nombreuses affections cardiovasculaires. Cette défaillance peut résulter d'une altération de la fonction systolique (fraction d'éjection réduite), d'une dysfonction diastolique (fraction d'éjection préservée) ou d'une forme intermédiaire, traduisant des anomalies structurelles ou fonctionnelles du myocarde telles qu'un affaiblissement ou une rigidité des parois cardiaques.

Les principales causes de l'insuffisance cardiaque incluent la cardiopathie ischémique, surtout après un infarctus du myocarde, l'hypertension artérielle, les valvulopathies, les cardiomyopathies ainsi que certains facteurs de risque comme le diabète, l'obésité, le tabagisme ou la consommation excessive d'alcool. D'autres facteurs, tels que les infections, les anomalies congénitales ou certains traitements, peuvent également contribuer à son apparition.Sur le plan physiopathologique, l'IC s'inscrit dans un cercle vicieux associant une diminution du débit cardiaque et l'activation de mécanismes compensatoires, notamment le système rénine–angiotensine–aldostérone. Cette activation entraîne une rétention hydrosodée responsable d'une congestion pulmonaire et périphérique, ainsi qu'un remodelage cardiaque progressif caractérisé par une dilatation des cavités (Figure 1.9) et une aggravation de la dysfonction myocardique. À terme, cette

altération de la circulation sanguine affecte plusieurs organes, notamment les poumons, les reins et le foie.

Cliniquement, l'insuffisance cardiaque se manifeste par une dyspnée, initialement à l'effort puis au repos ou en position allongée (orthopnée), une fatigue importante et une diminution de la capacité fonctionnelle. D'autres signes fréquents incluent des œdèmes des membres inférieurs ou de l'abdomen, une prise de poids rapide liée à la rétention hydrique, une toux nocturne, ainsi que des signes de congestion comme la turgescence jugulaire et les crépitements pulmonaires. Dans certains cas, des douleurs thoraciques ou des troubles du rythme peuvent être associés [22**,** pp. 65–75].

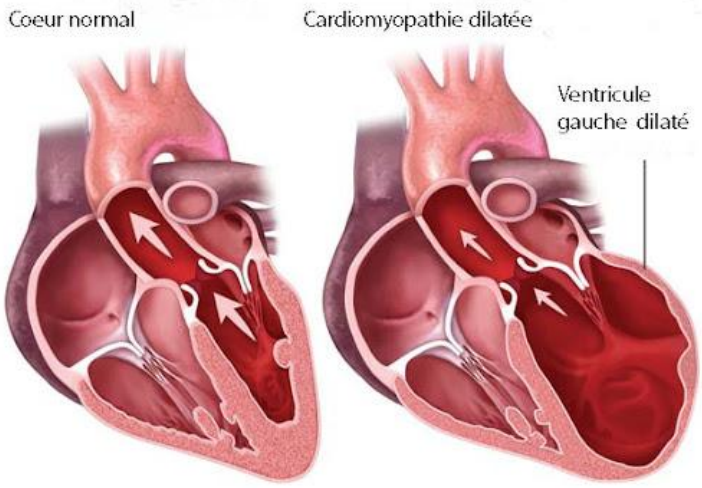


Figure 1.9 : Conséquences anatomiques de la cardiomyopathie dilatée menant à l'insuffisance cardiaque [23].

1.2.4 Accident vasculaire cérébral (AVC)

L'accident vasculaire cérébral (AVC) est une urgence médicale majeure correspondant à une interruption brutale de la circulation sanguine au niveau du cerveau. Il constitue la forme la plus grave des maladies cérébrovasculaires et représente une cause importante de mortalité et de handicap acquis chez l'adulte. On distingue deux types (Figure 1.10) principaux : l'AVC ischémique, le plus fréquent (environ 80 à 85 % des cas), dû à une obstruction d'une artère cérébrale, et l'AVC hémorragique, résultant de la rupture d'un vaisseau sanguin. Dans les deux situations, la privation d'oxygène entraîne une destruction rapide des neurones, pouvant conduire à des séquelles irréversibles ou au décès.

Les causes de l'AVC sont multiples et souvent liées à l'association de plusieurs facteurs de risque. L'hypertension artérielle constitue le principal facteur, suivie par

l'athérosclérose, le diabète, le tabagisme, les dyslipidémies et la fibrillation auriculaire, responsable d'embolies cérébrales. D'autres facteurs tels que l'obésité, la sédentarité, l'apnée du sommeil ou certaines anomalies vasculaires peuvent également intervenir. Les AVC ischémiques peuvent être d'origine athérothrombotique, cardio-embolique ou lacunaire, tandis que les AVC hémorragiques sont souvent associés à une hypertension chronique ou à la rupture d'un anévrisme.

Sur le plan clinique, les manifestations de l'AVC apparaissent de manière brutale et constituent une urgence médicale absolue. Les symptômes initiaux incluent généralement une faiblesse ou une paralysie unilatérale (hémiplégie), une déviation de la bouche, des troubles de la parole ou de la compréhension (aphasie), ainsi que des déficits sensoriels, visuels ou de l'équilibre. Dans les formes hémorragiques, des céphalées intenses, des vomissements ou une altération de la conscience peuvent également être observés.

La gravité des lésions cérébrales dépend directement de leur localisation et de la rapidité de la prise en charge. À long terme, ces atteintes font de l'AVC une cause majeure de mortalité et de dépendance, pouvant laisser des séquelles durables, des troubles de l'humeur comme la dépression, ainsi qu'un risque élevé de récidive [24].

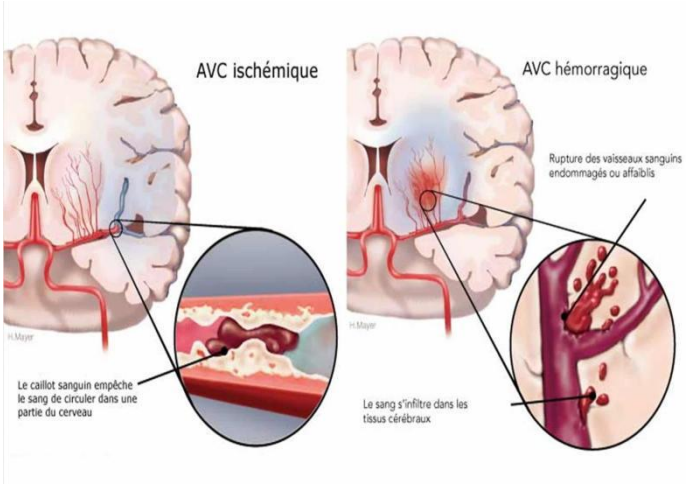


Figure 1.10 : Différenciation physiopathologique des accidents vasculaires cérébraux : AVC ischémique (occlusion) versus AVC hémorragique (rupture) [25].

1.3. Facteurs de risque cardiovasculaires :

Un facteur de risque cardiovasculaire désigne toute caractéristique individuelle, comportementale ou physiologique susceptible d'augmenter la probabilité de survenue d'une maladie cardiovasculaire. Ces facteurs se divisent en deux catégories : les facteurs modifiables, tels que l'hypertension artérielle, le tabagisme, le diabète, la dyslipidémie, l'obésité et la sédentarité, et les facteurs non modifiables, incluant l'âge, le sexe et les antécédents familiaux (Figure 1.11) .

Ces facteurs occupent une place centrale en épidémiologie cardiovasculaire, car ils expliquent la forte prévalence des maladies cardiovasculaires à l'échelle mondiale ainsi que leur position de première cause de mortalité. Leur identification permet non seulement de stratifier le risque individuel, mais également de cibler les populations à haut risque et de mettre en œuvre des stratégies de prévention adaptées, qu'elles soient primaires ou secondaires. Par ailleurs, la prise en charge des facteurs modifiables, notamment par l'arrêt du tabac, le contrôle de la pression artérielle, la gestion du profil lipidique et la promotion de l'activité physique, a démontré une réduction significative de la morbidité et de la mortalité cardiovasculaires. Ainsi, ces facteurs constituent à la fois des indicateurs diagnostiques et des leviers thérapeutiques majeurs [26].

Il est important de souligner que les facteurs de risque cardiovasculaire n'agissent pas de manière isolée, mais interagissent souvent de façon synergique, entraînant une amplification du risque global bien supérieure à la simple addition de leurs effets individuels. Par exemple, le tabagisme potentialise les effets délétères de l'hypertension, du diabète et des dyslipidémies. De même, l'association de plusieurs facteurs, tels que l'hypertension et le diabète ou encore l'obésité et les anomalies métaboliques, augmente de manière significative la sévérité de l'athérosclérose et le risque d'événements cardiovasculaires. Les modèles globaux confirment que la coexistence de plusieurs facteurs modifiables majore considérablement le risque de maladie et réduit l'espérance de vie, tandis que leur contrôle simultané améliore le pronostic et les années de vie sans maladie [27].

L'évaluation du risque cardiovasculaire a évolué au fil du temps, passant d'une approche simplifiée basée sur la sommation des facteurs de risque à des modèles prédictifs plus sophistiqués, tels que les scores issus d'études épidémiologiques (ex. Framingham, SCORE). Ces modèles intègrent plusieurs variables cliniques afin d'estimer la probabilité

de survenue d'un événement cardiovasculaire sur un horizon temporel donné, généralement de 5 à 10 ans. Les approches récentes tendent vers une personnalisation accrue du risque, en intégrant des biomarqueurs et des données issues de l'imagerie médicale [28].

Dans ce contexte, les modèles d'intelligence artificielle s'appuient sur ces mêmes facteurs de risque comme variables d'entrée principales. Grâce à des techniques avancées de sélection de variables, ces approches permettent d'identifier les facteurs les plus discriminants et de mettre en évidence d'éventuels biomarqueurs additionnels, améliorant ainsi la précision des prédictions.

L'intégration de ces variables dans des modèles d'apprentissage automatique, a permis d'améliorer significativement la capacité à prédire les événements cardiovasculaires majeurs. La performance de ces modèles dépend toutefois fortement de la qualité, de la complétude et de la diversité des données utilisées. Plus les données sont riches et multidimensionnelles, plus les modèles sont capables de capturer les interactions complexes entre les facteurs de risque et de fournir des estimations précises.

En pratique, la prise en charge des maladies cardiovasculaires doit reposer sur une approche globale visant la maîtrise simultanée de plusieurs facteurs de risque modifiables. Cette stratégie est essentielle, car leur combinaison engendre un effet multiplicateur sur le risque global, soulignant l'importance d'une intervention précoce, personnalisée et multidimensionnelle.

Le calcul du risque cardiaque a progressé de l'addition brute des caractéristiques à des algorithmes de calcul standardisés (Framingham, SCORE), ouvrant la voie à l'utilisation de l'intelligence artificielle pour une prédiction plus précise et personnalisée [29].

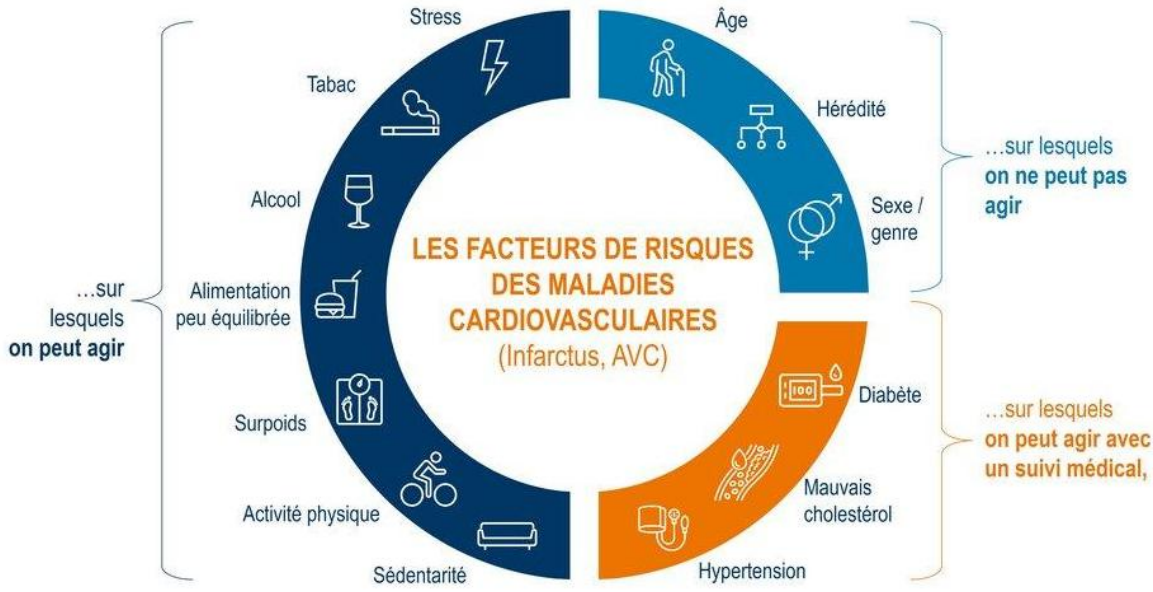


Figure 1.11: Classification des facteurs de risque des maladies cardiovasculaires selon leur caractère modifiable [30] .

1.4. Méthodes classiques de diagnostic :

Le diagnostic des maladies cardiovasculaires repose sur une approche multimodale combinant plusieurs examens complémentaires. Aucun test isolé ne permet une évaluation complète, ce qui nécessite l'association de données cliniques, électriques et d'imagerie. Les explorations électriques, telles que l'ECG, analysent l'activité cardiaque et les troubles du rythme, tandis que les techniques d'imagerie évaluent la structure et la fonction du cœur. Par ailleurs, les examens de perfusion permettent de détecter les anomalies du flux sanguin.

1.4.1 Électrocardiogramme (ECG)

L'électrocardiogramme (ECG) est une méthode diagnostique non invasive permettant d'enregistrer l'activité électrique du cœur (Figure 1.12) à l'aide d'électrodes placées à la surface de la peau, généralement au niveau des membres et du thorax pour un ECG standard à 12 dérivations. Le cœur génère, à chaque battement, des impulsions électriques issues de la dépolarisation et de la repolarisation des cellules myocardiques. Ces signaux, captés par les électrodes, sont amplifiés, filtrés puis retranscrits sous forme de tracés graphiques caractérisés par des ondes spécifiques (ondes P, complexe QRS et onde T). Le tracé enregistré reflétant l'activité cumulée de millions de cellules cardiaques, sa morphologie fournit des informations sur la santé du myocarde. L'enregistrement est rapide dans le cadre d'un ECG de repos (quelques minutes), mais peut également être

prolongé sur 24 à 48 heures à l'aide d'un enregistrement ambulatoire de type Holter.

L'ECG permet d'évaluer le rythme et la fréquence cardiaque, ainsi que la conduction électrique intracardiaque, en identifiant notamment les troubles du rythme (tachycardie, bradycardie, fibrillation auriculaire) et les anomalies de conduction (blocs auriculo-ventriculaires). Il constitue également un outil essentiel dans la détection des signes d'ischémie myocardique, tels que les modifications du segment ST, ainsi que dans l'identification des infarctus, des hypertrophies ventriculaires et de certains déséquilibres électrolytiques.

En pratique, l'ECG est largement utilisé en situation d'urgence, notamment en cas de douleur thoracique, de syncope ou de palpitations, mais aussi dans le cadre du bilan initial de nombreuses pathologies cardiovasculaires, du dépistage chez les sujets à risque et du suivi thérapeutique. Il peut également être associé à une épreuve d'effort afin d'évaluer la réponse cardiaque à une contrainte physique.

Parmi ses principaux avantages, l'ECG se distingue par sa simplicité, sa rapidité d'exécution, son caractère non invasif et son faible coût. Il fournit des informations immédiates essentielles à la prise de décision clinique et constitue un examen de référence en cardiologie [31].

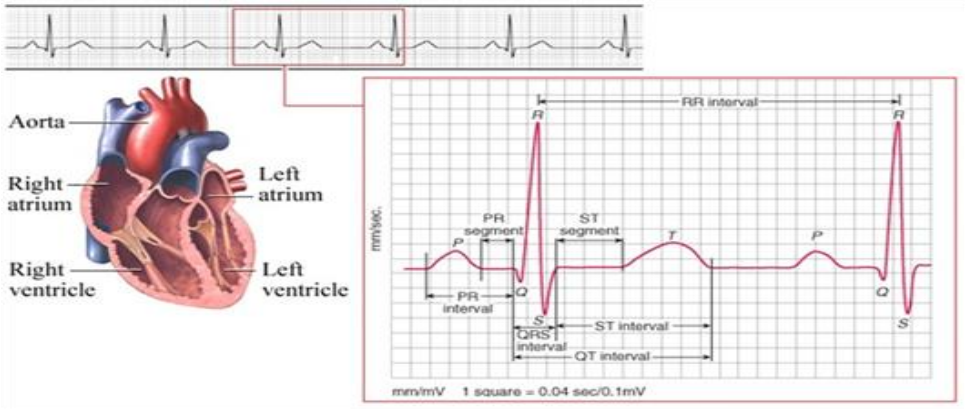


Figure 1.12 : Anatomie cardiaque et correspondance avec les ondes de l'électrocardiogramme (ECG) [32].

1.4.2 Échocardiographie cardiaque et le doppler

L'échocardiographie utilise des ultrasons pour visualiser le cœur (Figure 1.13) en temps réel. Une sonde émet des ondes ultrasonores qui traversent les tissus et sont réfléchies par les structures cardiaques, permettant à un ordinateur de reconstruire des images bidimensionnelles ou tridimensionnelles. Elle permet d'analyser la morphologie du cœur (taille des cavités, épaisseur des parois, valves), d'évaluer la fonction systolique (fraction d'éjection, volumes ventriculaires) et de détecter des anomalies structurelles comme les thrombus. L'examen peut être réalisé par voie transthoracique ou transœsophagienne pour une meilleure résolution, et des modes avancés tels que l'échographie 3D ou l'écho sous stress permettent de compléter l'évaluation. L'échocardiographie est indiquée dans le diagnostic et le suivi de l'insuffisance cardiaque, des valvulopathies, des cardiomyopathies, ainsi que dans la planification et le contrôle des interventions cardiaques. Elle présente l'avantage d'être non invasive, rapide, portable, peu coûteuse, et fournit une évaluation globale de la structure et de la fonction cardiaque.

Le Doppler est une technique complémentaire intégrée à l'échocardiographie qui exploite l'effet Doppler, c'est-à-dire le décalage de fréquence des ultrasons réfléchis par le sang en mouvement. Il permet de mesurer la vitesse et la direction du flux sanguin au sein des cavités et des valves cardiaques. Selon le mode utilisé (continu, pulsé ou couleur), le Doppler fournit des informations sur les gradients de pression, les régurgitations valvulaires, la fonction diastolique et les anomalies de flux intracardiaques. Il est particulièrement utile pour quantifier les insuffisances valvulaires, estimer les pressions pulmonaires et détecter des flux anormaux. Ses principaux avantages résident dans sa capacité à fournir des mesures précises et quantitatives, à améliorer la sensibilité de diagnostique, tout en restant peu invasif et économique [33].

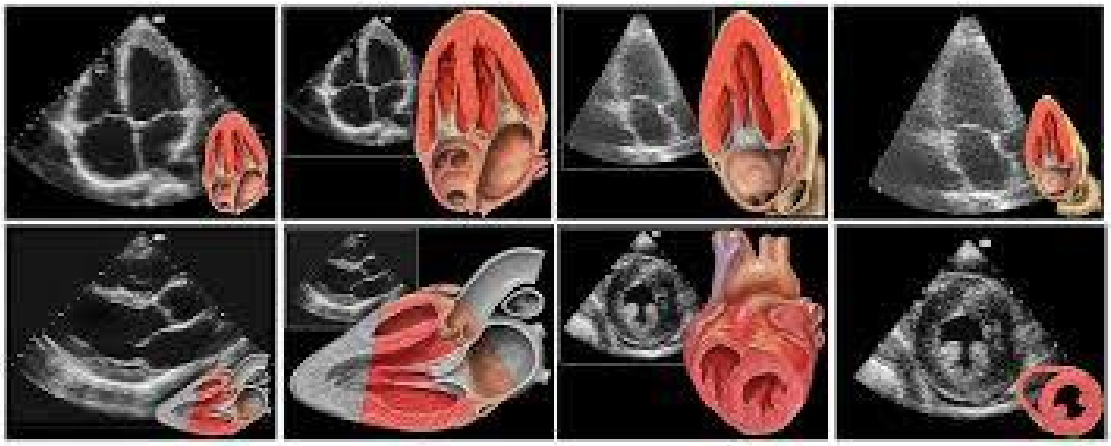


Figure 1.13 : Echographie cardiaque [34].

1.4.3 Angiographie :

L'angiographie est une technique d'imagerie vasculaire basée sur l'injection d'un produit de contraste iodé permettant de visualiser les vaisseaux sanguins à l'aide des rayons X. Elle consiste à introduire un cathéter dans une artère périphérique (fémorale ou radiale), puis à injecter le contraste afin d'obtenir des images en temps réel, sous forme de clichés bidimensionnels ou tridimensionnels. Cet examen permet d'analyser la structure des vaisseaux en détectant les rétrécissements (sténoses), les anévrismes, les dissections ou encore les malformations vasculaires dans différents territoires (cérébral, pulmonaire ou périphérique). L'angiographie est utilisée dans plusieurs contextes cliniques, notamment en situation d'urgence, dans le bilan préopératoire ou pour guider certaines interventions thérapeutiques comme l'embolisation. Elle se distingue par sa **haute précision diagnostique**, sa capacité à visualiser directement la lumière des vaisseaux et son rôle interventionnel. De plus, les techniques modernes permettent une analyse tridimensionnelle et fonctionnelle, améliorant ainsi l'évaluation des lésions vasculaires.

La coronarographie est une forme spécifique d'angiographie dédiée à l'exploration des artères coronaires. Réalisée en salle de cathétérisme, elle consiste à introduire un cathéter jusqu'à l'origine des artères coronaires, puis à injecter un produit de contraste iodé afin de visualiser la circulation coronaire en temps réel. Cette technique permet d'obtenir des images précises de l'anatomie coronarienne et d'évaluer le degré de rétrécissement des artères, notamment en cas d'athérosclérose. La coronarographie permet de détecter les sténoses significatives, les occlusions totales, les thromboses, ainsi que diverses anomalies telles que les fistules ou les spasmes coronariens. Elle offre également une évaluation quantitative du flux sanguin et de la sévérité des lésions. Cet examen est indiqué en cas de suspicion ou de confirmation de maladie coronarienne, notamment lors de douleurs thoraciques, de tests d'effort positifs ou de syndromes coronariens aigus. Il est également utilisé dans le bilan préopératoire et chez les patients à risque lorsque les examens non invasifs sont insuffisants.Elle présente l'avantage majeur de combiner diagnostic et traitement, en permettant la réalisation immédiate d'actes thérapeutiques tels que l'angioplastie ou la pose de stents. Grâce à sa haute résolution et à l'analyse en temps réel, elle offre une évaluation précise de la topographie, de la sévérité et du nombre de lésions coronariennes, ce qui en fait une technique de référence incontournable dans la prise en charge des maladies coronariennes [35].

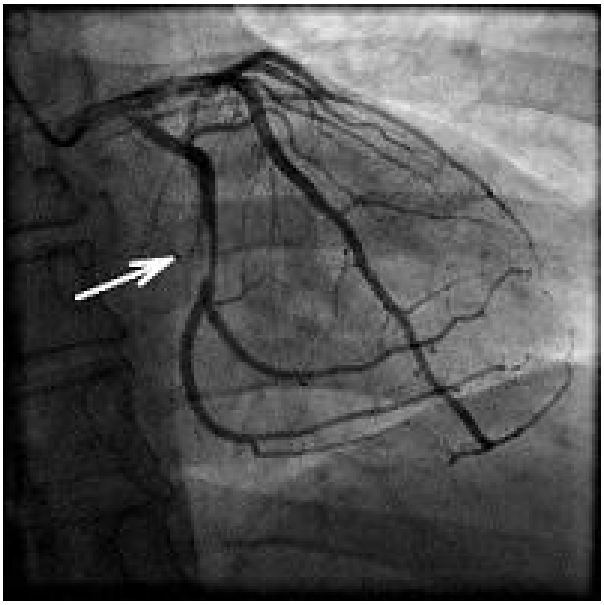


Figure 1.14 : Rétrécissement de l'artère coronaire observé par coronarographie [36] .

1.4.4 Test d'effort (épreuve d'effort cardiaque)

Le test d'effort est un examen fonctionnel permettant d'évaluer la réponse du cœur à un stress physique contrôlé. Il consiste à faire réaliser au patient un effort progressif, généralement sur un tapis roulant ou un vélo ergométrique, tout en assurant une surveillance continue de l'électrocardiogramme (ECG), de la fréquence cardiaque, de la pression artérielle et parfois de la saturation en oxygène. L'augmentation de l'effort entraîne une élévation des besoins en oxygène du myocarde, ce qui permet de mettre en évidence des anomalies non détectables au repos, notamment les déséquilibres de perfusion ou les troubles du rythme. Des variantes pharmacologiques peuvent être utilisées chez les patients incapables de fournir un effort physique.

Cet examen permet de détecter des signes d'ischémie myocardique, tels que les modifications du segment ST ou l'apparition d'arythmies à l'effort. Il évalue également la capacité fonctionnelle du patient , ainsi que la tolérance cardiovasculaire à l'exercice, ce qui aide à différencier une origine cardiaque d'une origine respiratoire ou musculaire des symptômes. Il possède en outre une valeur pronostique importante grâce à l'analyse de paramètres tels que la réponse tensionnelle et le comportement du rythme cardiaque à l'effort.

Le test d'effort est indiqué dans la suspicion ou la confirmation de maladie coronarienne, notamment en cas de douleur thoracique ou de dyspnée d'effort, ainsi que

dans l'évaluation du risque cardiovasculaire avant une intervention chirurgicale ou la pratique sportive. Il est également utilisé pour le suivi thérapeutique, notamment après un infarctus du myocarde, une angioplastie ou dans le cadre de programmes de réadaptation cardiaque.

Parmi ses principaux avantages, le test d'effort se distingue par son caractère **non** invasif, simple, reproductible et peu coûteux. Il permet une évaluation fonctionnelle en conditions proches de la vie réelle et offre une bonne capacité de stratification du risque cardiovasculaire [37].

1.5. Limites des méthodes classiques :

Les méthodes classiques de diagnostic des maladies cardiovasculaires, bien qu'essentielles en pratique clinique, présentent plusieurs limites qui peuvent compromettre la précision et la fiabilité du diagnostic.

L'électrocardiogramme, ou ECG, n'est pas très fiable, surtout chez les femmes et les personnes qui ont des problèmes cardiaques au repos, comme un bloc de branche ou un pacemaker. Dans ces cas, les résultats peuvent être incorrects.De plus,il présente des limites intrinsèques liées à la complexité de l'activité électrique cardiaque et à l'ambiguïté de son interprétation. Sa précision diagnostique peut être affectée par plusieurs facteurs, notamment l'anatomie du patient, le positionnement des électrodes, le bruit du signal ainsi que la présence de tracés similaires associés à des pathologies distinctes [38].

L'échocardiographie, qu'elle soit faite au repos ou après un effort, dépend beaucoup de la personne qui l'interprète et de la qualité de l'image. Si la personne est obèse ou a des problèmes pulmonaires, cela peut rendre les résultats moins précis. Cette méthode ne détecte pas bien les problèmes cardiaques mineurs, comme les sténoses univesiculaires ou les micro-ischémies. De plus, si l'effort n'est pas suffisant ou si la personne a des problèmes avec la fonction de son ventricule, les résultats peuvent être moins fiables [39].

Test d'effort (exercice ± ECG) : la valeur diagnostique varie selon l'âge, le sexe, la forme physique . Chez les femmes, les patients avec incapacité à atteindre 85 % de la fréquence cible ou avec une fonction ventriculaire réduite, le taux de faux-positifs et de faux-négatifs augmente . Le test est surtout fiable pour exclure la maladie mais peu performant pour confirmer une sténose significative [40].

La coronarographie est un procédure invasive avec risques de saignement, infection, néphropathie induite par le produit de contraste, et limitation majeure : elle ne fournit aucune information fonctionnelle [41].

En générale , chaque technique classique possède des limites liées à la sensibilité/spécificité, à la dépendance operatoriale, aux contraintes physiologiques du patient ou à la nature invasive et à l'absence d'évaluation fonctionnelle.

Conclusion :

Nous avons montré dans ce chapitre, que les maladies cardiovasculaires constituent un enjeu majeur de santé publique en raison de leur fréquence élevée et de leur impact sur la morbidité et la mortalité à l'échelle mondiale. La compréhension de l'anatomie et de la physiologie du système cardiovasculaire permet d'appréhender les mécanismes sous-jacents aux principales pathologies .

En plus l'identification des facteurs de risque, qu'ils soient modifiables ou non, joue un rôle essentiel dans la prévention et la stratification des MCV.

Par ailleurs, les méthodes classiques de diagnostic , offrent des outils indispensables pour l'évaluation clinique des patients. Toutefois, malgré leur complémentarité, ces techniques présentent plusieurs limites, notamment en termes de sensibilité, de spécificité, de dépendance à l'opérateur et de capacité à fournir une analyse globale et précoce des anomalies cardiovasculaires.

Dans ce contexte, il devient nécessaire d'explorer de nouvelles approches capables d'exploiter efficacement la complexité des données médicales. L'émergence de l'intelligence artificielle ouvre des perspectives prometteuses pour améliorer la précision diagnostique, anticiper les événements cardiovasculaires et proposer des stratégies de prise en charge plus personnalisées. Ainsi, le chapitre suivant s'inscrit dans cette dynamique en s'intéressant aux modèles intelligents dédiés à la prédiction des anomalies cardiovasculaires.

Chapitre 2 :

Cadre théorique de l'apprentissage automatique appliqué à la classification médicale et présentation des bases de données cardiovasculaires

Chapitre 2

Cadre théorique de l'apprentissage automatique appliqué à la classification médicale et présentation des bases de données cardiovasculaires

Introduction à l'intelligence artificielle en biomédical :

L'intelligence artificielle (IA) prend aujourd'hui une place de plus en plus importante dans le domaine de la santé. Elle permet d'analyser de grandes quantités de données médicales et d'aider les professionnels dans la prise de décision clinique. Parmi ses applications les plus intéressantes, on retrouve la prédiction des maladies cardiovasculaires, qui restent l'une des principales causes de décès dans le monde.

L'apprentissage automatique permet d'exploiter les données des patients afin d'identifier les facteurs de risque et d'anticiper l'apparition de maladies cardiaques avec une précision de plus en plus élevée. Plusieurs algorithmes sont souvent utilisés dans ce domaine, comme la régression logistique, le KNN, le SVM, le Random Forest ainsi que les réseaux de neurones artificiels.

Ce chapitre présente les fondements de l'apprentissage automatique appliqué à la classification médicale, en particulier pour la prédiction des maladies cardiovasculaires. Il introduit les principaux concepts de l'intelligence artificielle en biomédecine, les types d'apprentissage (supervisé, non supervisé et deep learning), ainsi que les principaux algorithmes utilisés tels que la régression logistique, le SVM, le Random Forest, le KNN et les réseaux de neurones.

Il aborde également les principales métriques d'évaluation des modèles, notamment l'accuracy, la précision, le rappel, le F1-score ainsi que la courbe ROC et l'AUC. Enfin, le chapitre présente les bases de données utilisées (Cleveland, Framingham et Kaggle), l'analyse exploratoire des données (EDA) et les principaux défis liés aux données médicales [42] .

2.1 Machine Learning en médecine prédictive :

Le Machine Learning constitue aujourd'hui un pilier essentiel de la médecine prédictive, grâce à sa capacité à analyser de vastes ensembles de données cliniques et à

identifier des relations complexes souvent difficiles à détecter par les méthodes statistiques classiques. Son objectif principal est d'améliorer la prise de décision médicale en permettant une estimation plus précise des risques de développement de maladies.

Dans le domaine de la santé, ces approches reposent sur l'exploitation de données hétérogènes issues des dossiers médicaux des patients, incluant les paramètres biologiques, les antécédents cliniques et les facteurs de mode de vie. Les algorithmes d'apprentissage automatique apprennent à partir de ces données historiques afin de construire des modèles capables de généraliser et de prédire l'apparition de pathologies chez de nouveaux patients.

En médecine cardiovasculaire, le Machine Learning joue un rôle particulièrement important dans la détection précoce des maladies et l'identification des facteurs de risque. Il permet ainsi d'améliorer la prévention, d'optimiser le diagnostic et de contribuer à une prise en charge plus personnalisée des patients[43].

2.2 Types d'apprentissage :

L'apprentissage automatique comprend plusieurs types d'apprentissage, classés selon la structure du problème qu'ils abordent. La figure 2.1 présente les types les plus courants dans les principaux domaines d'application : l'apprentissage supervisé, l'apprentissage non supervisé, et l'apprentissage par renforcement.

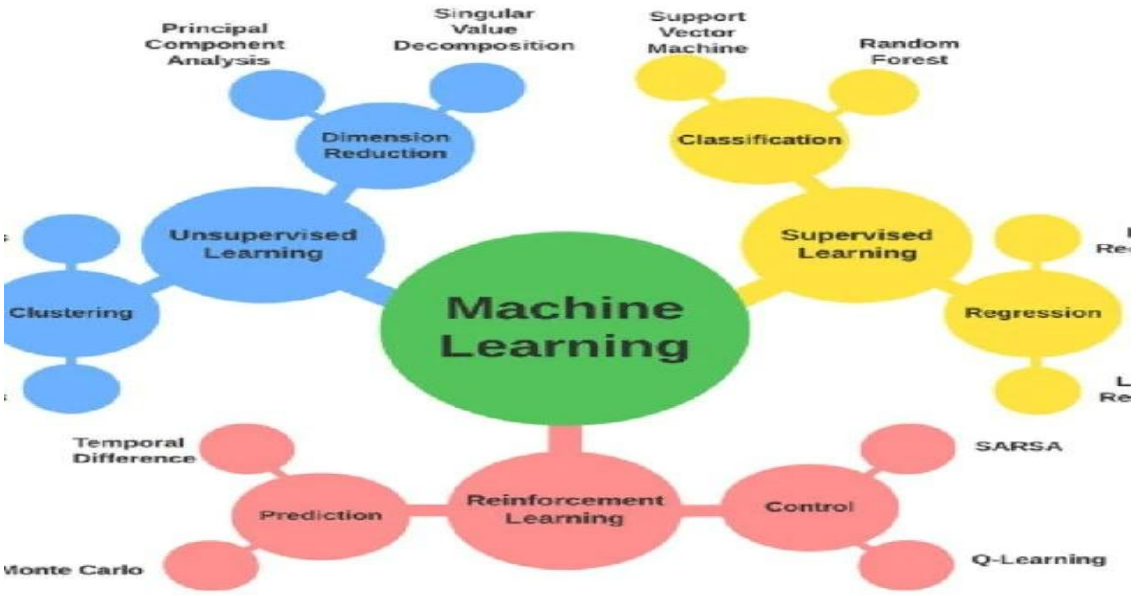


Figure 2.1 : Types d'apprentissage [44]

2.2.1 L'apprentissage supervisé

L'apprentissage automatique supervisé repose sur l'utilisation de données étiquetées afin d'entraîner des algorithmes à généraliser les connaissances acquises à partir d'exemples passés vers de nouvelles observations. Ce paradigme permet ainsi de construire une fonction de prédiction à partir d'un ensemble de données historiques, en ajustant progressivement le modèle en fonction de l'écart entre les valeurs prédites et les valeurs réelles. Cette capacité d'auto-correction contribue à améliorer la performance du modèle au cours de la phase d'apprentissage.

Dans le cadre de l'apprentissage supervisé, on distingue principalement deux catégories de tâches [45] :

a) **La classification** : elle concerne les problèmes où la variable cible est qualitative et discrète. L'objectif consiste à attribuer une étiquette ou une classe à chaque observation. À titre d'exemple, elle peut être utilisée pour prédire la tendance future d'un actif financier (hausse, stabilité ou baisse).

b) **La régression** : elle s'applique lorsque la variable cible est quantitative et continue, visant ainsi à estimer une valeur numérique.

Parmi les principaux algorithmes d'apprentissage supervisé les plus utilisés en pratique [46], on peut citer :

- ✓ La méthode des k plus proches voisins (KNN) .
- ✓ La régression linéaire .
- ✓ La régression logistique .
- ✓ Les machines à vecteurs de support (SVM) .
- ✓ Les arbres de décision ainsi que les forêts aléatoires.

2.2.2 Apprentissage non supervisé

L'apprentissage non supervisé constitue une branche de l'apprentissage automatique dans laquelle l'ensemble de données d'entraînement ne comporte aucune variable cible ou étiquette associée, contrairement à l'apprentissage supervisé. Le modèle est ainsi alimenté uniquement par des données d'entrée, et a pour objectif d'identifier de manière autonome des structures, des régularités ou des relations sous-jacentes présentes

dans les données.

Dans ce contexte, l'algorithme cherche à extraire des informations pertinentes sans guidance explicite, en mettant en évidence des regroupements ou des représentations plus compactes de l'espace des données.

Les principales approches de l'apprentissage non supervisé se répartissent généralement en deux grandes catégories [45], [46] :

a) Le clustering (regroupement)**,** dont l'objectif est de partitionner les données en groupes homogènes selon des critères de similarité. Parmi les méthodes les plus couramment utilisées, on cite :
   •K-means ;
   •le clustering hiérarchique (HCA) ;
   • l'algorithme d'espérance-maximisation (EM).

b) **La réduction de dimension et la visualisation**, qui visent à projeter les données dans un espace de dimension réduite tout en préservant au mieux leur structure intrinsèque. Les méthodes les plus répandues incluent :
   • l'analyse en composantes principales (ACP) ;
   • le Kernel PCA ;
   • le Locally Linear Embedding (LLE) ;
   • le t-distributed Stochastic Neighbor Embedding (t-SNE).

2.2.3 L'apprentissage par renforcement

L'apprentissage par renforcement est une approche de l'apprentissage automatique dans laquelle l'agent apprend par interaction directe avec son environnement, selon un principe d'essai-erreur. À chaque action effectuée, l'agent reçoit un retour sous forme de récompense ou de pénalité, lui permettant d'évaluer la pertinence de ses décisions.

L'objectif principal est de déterminer une politique optimale, c'est-à-dire une stratégie de prise de décision qui maximise la récompense cumulée au cours du temps. L'apprentissage repose ainsi sur un signal de renforcement, qui guide l'agent dans l'amélioration progressive de son comportement.

Cette approche est particulièrement adaptée aux environnements dynamiques et incertains, tels que la prévision financière à haute fréquence, où les conditions du marché évoluent rapidement et rendent difficile la conception manuelle de stratégies optimales [47].

2.3 Deep Learning (Apprentissage profond) :

Le Deep Learning, ou apprentissage profond, est une branche de l'apprentissage automatique basée sur l'utilisation de réseaux de neurones artificiels comportant plusieurs couches cachées. Ces architectures permettent d'apprendre automatiquement des représentations hiérarchiques des données, en transformant progressivement les informations brutes en caractéristiques de plus haut niveau.

Grâce à cette structure multi-couches, le Deep Learning est capable de traiter efficacement des données complexes et de grande dimension, telles que les images, les signaux, le texte ou les données médicales. Il constitue ainsi une approche particulièrement performante pour les tâches de classification, de régression et de reconnaissance de motifs[48].

La figure ci-dessous illustre la différence entre le Machine Learning et le Deep Learning, en mettant en évidence leurs principes de fonctionnement et leurs niveaux de complexité.

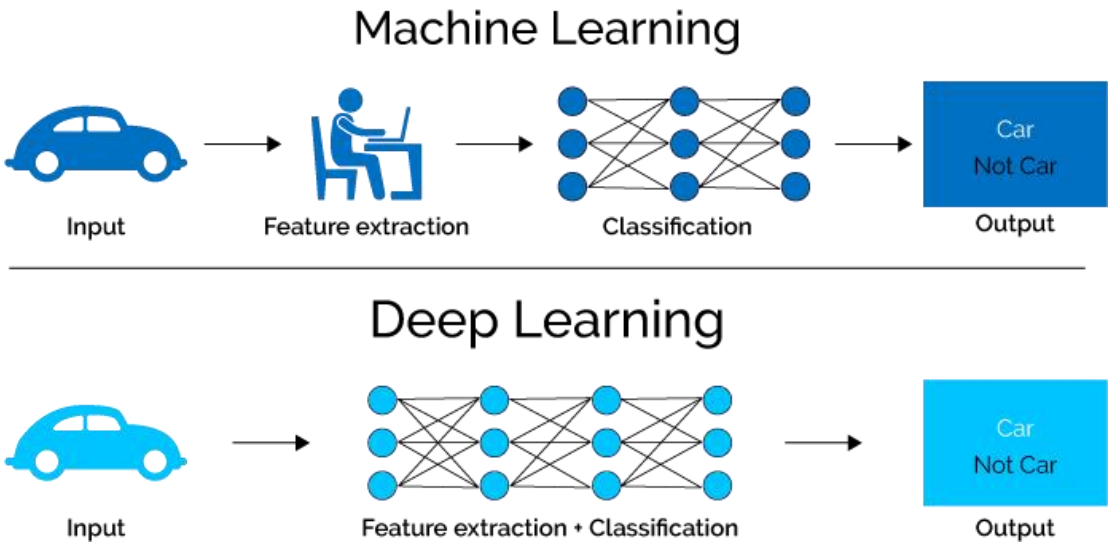


Figure 2.2 : Illustration de la différence entre le Machine Learning et le Deep Learning[49] .

2.4 Algorithmes de classification utilisés en santé :

2.4.1. Régression logistique

Régression logistique constituent une méthode d'apprentissage supervisé largement utilisée pour les problèmes de classification binaire. Ils permettent de modéliser la probabilité

d'appartenance d'une observation à une classe donnée en fonction d'un ensemble de variables explicatives. Grâce à leur simplicité d'interprétation et leur efficacité computationnelle, ils figurent parmi les algorithmes les plus classiques et les plus utilisés en apprentissage automatique et en statistique appliquée.

Le principe fondamental de la régression logistique repose sur l'estimation d'une fonction sigmoïde permettant de transformer une combinaison linéaire des variables explicatives en une probabilité comprise entre 0 et 1. L'objectif est de déterminer les coefficients optimaux du modèle afin de maximiser la vraisemblance des données observées et de fournir une frontière de décision permettant de séparer les différentes classes.

À partir d'un ensemble de données étiquetées, l'algorithme apprend une relation entre les variables d'entrée et la probabilité de sortie, généralement en utilisant la fonction logit comme transformation intermédiaire. Cette approche permet de classifier de nouvelles observations en leur attribuant une probabilité d'appartenance à chaque classe, avec une règle de décision basée sur un seuil (souvent 0,5).

Lorsque les données présentent des relations complexes, la régression logistique peut être étendue à des variantes régularisées (L1, L2) afin d'améliorer la généralisation du modèle et de réduire le surapprentissage. Elle est également adaptée à des problèmes multi-classes via des extensions telles que la régression logistique multinomiale

- **Régression logistique binaire :**
  Dans ce cas, la variable à prédire ne peut prendre que deux valeurs possibles, généralement 0 ou 1. Ce type de modèle est utilisé lorsque l'on souhaite distinguer entre deux situations, comme par exemple oui/non, succès/échec ou présence/absence d'une maladie.

- **Régression logistique multinomiale :**
  Ici, la variable à prédire peut appartenir à plusieurs catégories (trois ou plus), sans qu'il y ait un ordre entre elles. Chaque classe est indépendante des autres. Par exemple, il peut s'agir de différents types de maladies ou de catégories comme « Type A », « Type B » ou « Type C ».

- **Régression logistique ordinale :**

  Dans ce type, la variable à prédire comporte plusieurs catégories qui sont organisées selon un ordre logique. Contrairement au cas multinomial, les classes ont ici une signification hiérarchique. Par exemple, on peut avoir des niveaux comme « mauvais », « moyen », « bon » et « excellent ». [50]

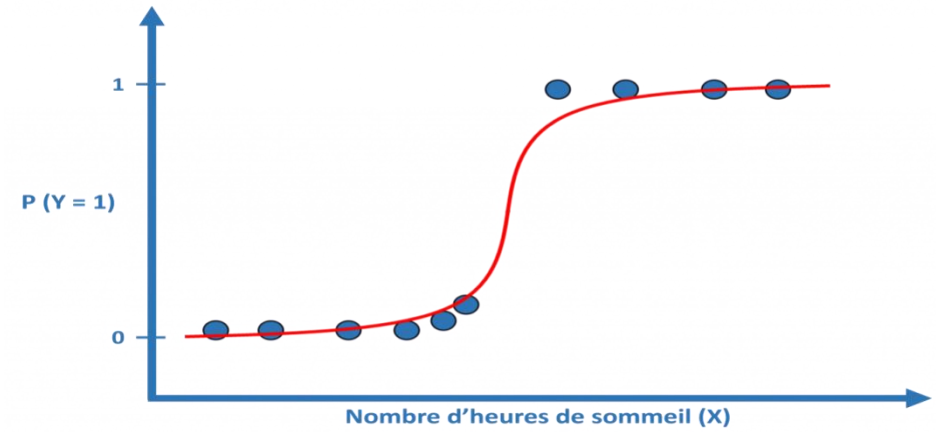


Figure 2.3 : Régression logistique [51] .

2.4.2 SVM

Les machines à vecteurs de support (Support Vector Machines – SVM) constituent une méthode d'apprentissage supervisé largement utilisée pour les tâches de classification et de régression. Grâce à leur capacité à traiter des données complexes et de grande dimension, les SVM figurent parmi les algorithmes les plus performants en apprentissage automatique.

Le principe fondamental des SVM consiste à déterminer un hyperplan optimal permettant de séparer les différentes classes de données avec une marge maximale. L'objectif est de maximiser la distance entre l'hyperplan de séparation et les observations les plus proches, appelées vecteurs de support, afin d'améliorer la capacité de généralisation du modèle.

À partir d'un ensemble de données étiquetées, l'algorithme apprend une frontière de décision capable de classer correctement de nouvelles observations. Dans les cas où les données ne sont pas linéairement séparables, les SVM utilisent des fonctions noyaux (kernels) permettant de projeter les données dans un espace de dimension supérieure où la séparation devient possible.

Les SVM ont été appliqués avec succès dans de nombreux domaines, notamment la reconnaissance de l'écriture manuscrite, la détection de fraudes bancaires, la reconnaissance vocale, la détection faciale ainsi que le domaine biomédical. En particulier, dans la recherche

sur le cancer, ils permettent d'identifier des modèles complexes à partir de données génomiques et cliniques afin d'améliorer le diagnostic, le pronostic et la prédiction de la réponse aux traitements [52].

Le fonctionnement des SVM repose sur plusieurs concepts clés :

- ✓ **Vecteurs de support** : ce sont les points de données les plus proches de la frontière de décision. Ils jouent un rôle essentiel dans la définition du modèle, car ce sont eux qui déterminent la position de l'hyperplan.

- ✓ **Hyperplan** : il s'agit d'une surface de décision qui permet de séparer les données en différentes classes. Dans un espace bidimensionnel, il correspond à une droite, tandis que dans des dimensions supérieures, il devient un hyperplan.

- ✓ **Marge** : la marge représente la distance entre l'hyperplan et les vecteurs de support. L'objectif du SVM est de maximiser cette marge afin d'améliorer la capacité de généralisation du modèle. Une marge plus grande est généralement associée à une meilleure performance de classification. La figure ci-dessous

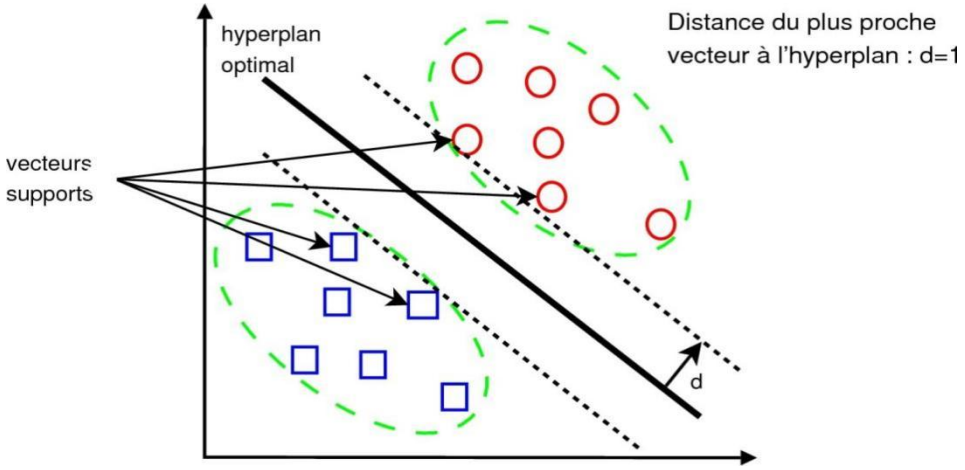


Figure 2.4 : Principe de SVM. [53]

2.4.3 Random Forest (Forêts aléatoires)

La forêt aléatoire (Random Forest) est une méthode d'apprentissage supervisé basée sur l'approche d'ensemble (*ensemble learning*), qui combine plusieurs arbres de décision afin d'améliorer les performances de classification et de réduire le risque de surapprentissage. Les arbres sont construits selon la méthodologie des arbres de classification et de régression (*Classification and Regression Trees – CART*).

Le principe de Random Forest repose sur l'introduction de deux sources d'aléa. Premièrement, chaque arbre est entraîné à partir d'un échantillon **bootstrap** obtenu par tirage aléatoire avec remise des données d'apprentissage. Deuxièmement, lors de la construction de chaque arbre, un sous-ensemble aléatoire de variables explicatives est sélectionné à chaque nœud afin de déterminer la meilleure séparation possible.

Les principaux paramètres de l'algorithme sont le nombre d'arbres composant la forêt ainsi que le nombre de variables considérées lors de chaque division. Toutefois, Random Forest est réputé pour sa robustesse et sa capacité à fournir de bonnes performances même avec les paramètres par défaut.

Dans les problèmes de classification, chaque arbre de la forêt attribue une classe à la nouvelle observation. La décision finale est obtenue par un **vote majoritaire** des prédictions individuelles. Cette stratégie permet d'améliorer la précision du modèle et sa capacité de généralisation.

Par ailleurs, les performances théoriques de Random Forest peuvent être analysées à travers l'erreur de généralisation, laquelle dépend principalement de deux facteurs : la précision des arbres individuels et le degré de corrélation entre ceux-ci. Une forêt composée d'arbres précis et faiblement corrélés tend à offrir de meilleures performances prédictives [54].

La figure illustre le principe de fonctionnement de l'algorithme **Random Forest (Forêt Aléatoire)**. Lorsqu'un nouvel échantillon (« *New Sample* ») est présenté au modèle, celui-ci est transmis simultanément à plusieurs arbres de décision. Chaque arbre analyse les caractéristiques de l'échantillon en suivant un ensemble de règles de décision jusqu'à atteindre une feuille terminale, où une prédiction est produite. Dans l'exemple illustré, trois arbres génèrent respectivement les résultats *Result 1*, *Result 2* et *Result 3*. Comme chaque arbre est construit à partir d'un sous-ensemble aléatoire des données et des variables, les prédictions obtenues peuvent être différentes. Afin de déterminer la prédiction finale,

Random Forest combine les résultats de tous les arbres. Pour les problèmes de classification, la décision est prise par **vote majoritaire** (*Majority Voting*), c'est-à-dire que la classe obtenant le plus grand nombre de votes est sélectionnée. Dans le cas des problèmes de régression, la prédiction finale est calculée en effectuant la moyenne des valeurs prédites par les différents arbres (*Averaging*). Cette approche d'ensemble permet d'améliorer la précision des prédictions, de réduire le risque de surapprentissage (*overfitting*) et d'augmenter la robustesse du modèle par rapport à un arbre de décision unique.

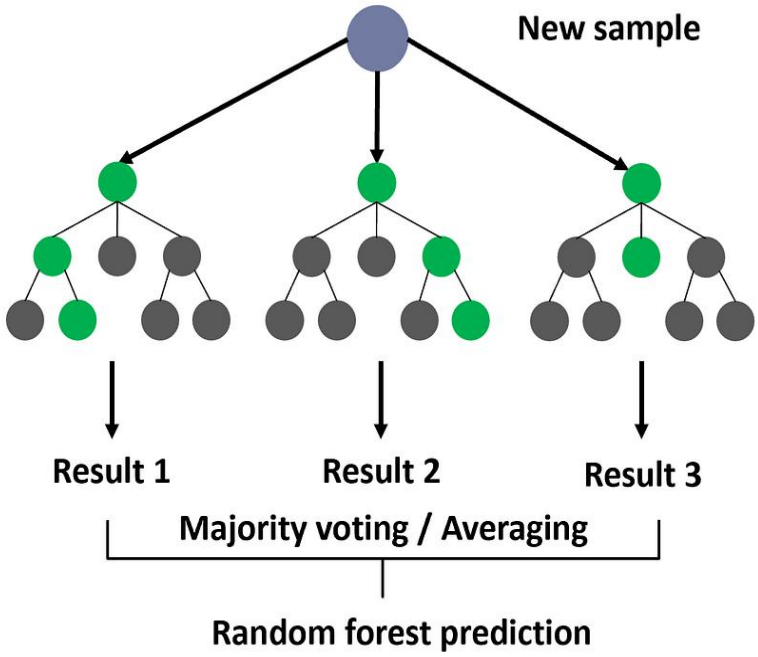


Figure 3.5 : Exemple Random forest [55]

2.4.4 K plus proches voisins (K-Nearest Neighbors, KNN)

L'algorithme des k plus proches voisins (K-Nearest Neighbors, KNN) est une méthode d'apprentissage supervisé non paramétrique, introduite par **Evelyn Fix** et **Joseph Hodges** en 1951. Il est principalement utilisé pour les tâches de classification et de régression.

Le principe du KNN repose sur l'hypothèse selon laquelle les observations présentant des caractéristiques similaires appartiennent généralement à une même catégorie.

Pour prédire la classe ou la valeur associée à une nouvelle observation, l'algorithme identifie les *k* exemples les plus proches dans l'espace des données à l'aide d'une mesure de distance, telle que la distance euclidienne.

Dans les problèmes de classification, la décision est généralement prise par un mécanisme de vote majoritaire parmi les *k* voisins sélectionnés. Dans les problèmes de régression, la prédiction correspond le plus souvent à la moyenne des valeurs observées chez les voisins les plus proches.

Grâce à sa simplicité de mise en œuvre et à son efficacité pour de nombreux problèmes de classification, le KNN demeure l'un des algorithmes les plus utilisés en apprentissage automatique. Toutefois, ses performances peuvent être affectées par le choix du paramètre *k* ainsi que par la présence de données de grande dimension ou de bruit dans le jeu de données [56].

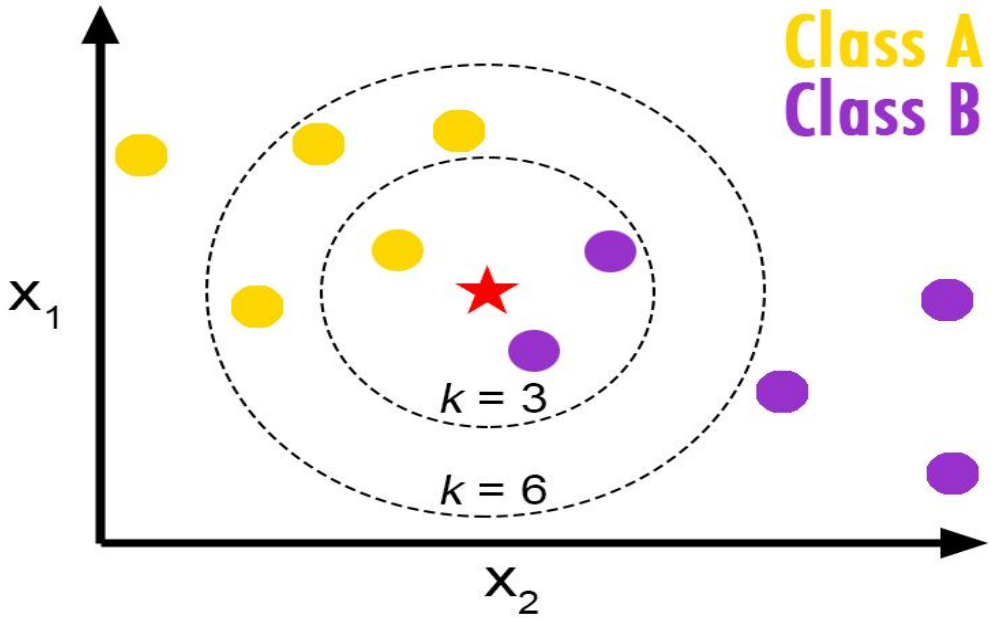


Figure 2.6 : Exemple de KNN[57]

La Figure 2.6 illustre le fonctionnement de l'algorithme des **k plus proches voisins (KNN)** dans un problème de classification binaire comportant deux classes : la **classe A** (points jaunes) et la **classe B** (points violets). L'étoile rouge représente une nouvelle observation dont la classe est inconnue et qui doit être prédite à partir des données d'apprentissage.

Dans un premier temps, lorsque **k = 3**, l'algorithme considère les trois voisins les

plus proches de l'observation à classifier, délimités par le premier cercle. Parmi ces trois voisins, deux appartiennent à la **classe B** et un seul à la **classe A**. Selon le principe du vote majoritaire, la nouvelle observation est donc affectée à la **classe B**.

Dans un second temps, lorsque **k = 6**, l'algorithme prend en compte les six voisins les plus proches, représentés à l'intérieur du second cercle. Cette fois-ci, la majorité des voisins appartient à la **classe A**. Par conséquent, l'observation est classée dans la **classe A**.

Cette illustration met en évidence l'influence du paramètre **k** sur le résultat de la classification. En effet, une valeur faible de *k* rend l'algorithme plus sensible aux observations locales et au bruit, tandis qu'une valeur plus élevée permet d'obtenir une décision plus stable en tenant compte d'un plus grand nombre de voisins. Le choix optimal de ce paramètre constitue donc une étape essentielle pour garantir les performances du modèle KNN.

2.4.5 Les réseaux du nuerons

Les réseaux de neurones artificiels (Artificial Neural Networks – ANN), notamment les perceptrons multicouches (Multilayer Perceptron – MLP), constituent une famille de méthodes d'apprentissage supervisé largement utilisée dans le domaine de l'intelligence artificielle et du Deep Learning. Grâce à leur capacité à modéliser des relations non linéaires complexes, ils sont particulièrement efficaces pour le traitement de données de grande dimension et fortement corrélées.

Le principe fondamental des réseaux de neurones repose sur une architecture composée de plusieurs couches de neurones artificiels : une couche d'entrée, une ou plusieurs couches cachées, et une couche de sortie. Chaque neurone effectue une transformation mathématique des données en appliquant des poids synaptiques, une fonction d'activation et un biais, permettant ainsi l'apprentissage progressif de représentations de plus en plus abstraites.

L'apprentissage du réseau s'effectue généralement par rétropropagation de l'erreur (backpropagation) associée à un algorithme d'optimisation tel que la descente de gradient. L'objectif est de minimiser la fonction de coût en ajustant itérativement les poids du réseau afin d'améliorer la précision des prédictions.Grâce à leur profondeur et à leur capacité d'apprentissage hiérarchique, les réseaux de neurones sont capables d'extraire

automatiquement des caractéristiques pertinentes à partir des données brutes, sans nécessiter une ingénierie manuelle des variables.Les réseaux de neurones ont été appliqués avec succès dans de nombreux domaines tels que la reconnaissance d'images, le traitement du langage naturel, la reconnaissance vocale ainsi que le domaine biomédical. En particulier, dans la recherche sur les maladies cardiovasculaires et le cancer, ils permettent d'analyser des données complexes afin d'améliorer la classification, le diagnostic précoce et la prédiction des issues cliniques[58]

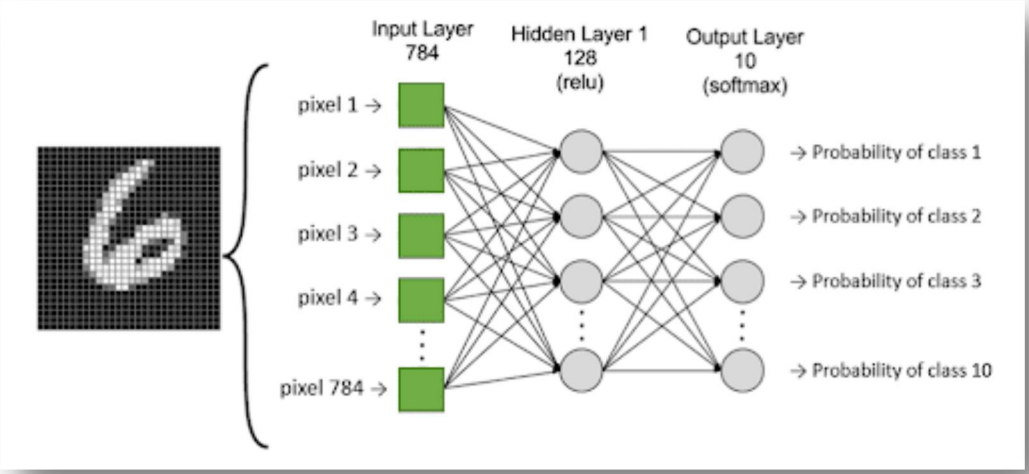


Figure 2.7 : Exemple de réseaux du nuerons[59]

Un perceptron multicouche (Multi Layer Perceptron : MLP) : Un perceptron multicouche (MLP) est une classe de réseau neuronal artificiel à action directe.

Un MLP se compose d'au moins trois couches de nœuds : couche d'entrée, couche cachée et couche de sortie.

a) La couche d'entrée (input layer) : Un ensemble de neurones qui portent le signal d'entrée. Tous les neurones de cette couche sont ensuite reliés à ceux de la couche suivante.

b) La couche cachée (hidden layer) : Ou plus souvent les couches cachés (couche cachée 1, couche cachée 2, ...). Il s'agit du cœur de perceptron.

c) La couche de sortie (output layer) : Cette couche représente le résultat nal du réseau, c'est-à-dire sa prédiction .

2.5 Métriques d'évaluation des modèle :

Les mesures d'évaluation de performance constituent des outils fondamentaux pour apprécier la qualité d'un modèle d'apprentissage automatique. Elles permettent d'analyser de manière quantitative les résultats du modèle en comparant ses prédictions aux valeurs réelles observées.

Parmi les principales métriques utilisées en classification,

- ✓ **l'accuracy** (ou exactitude) représente la proportion d'échantillons correctement classés par rapport à l'ensemble des données. Elle fournit ainsi une vision globale de la performance du modèle.

- ✓ Accuracy=$\frac{TP+TN}{TP+TN+FP+FN}$ (2. 1)

- ✓ **Précision** (Precision) La précision mesure la proportion des prédictions positives qui sont réellement correctes. Elle indique la fiabilité du modèle lorsqu'il prédit une classe positive, en évaluant le nombre de vrais positifs parmi l'ensemble des résultats prédits positifs.

- ✓ Précision =$\frac{TP}{TP+FP}$ (2. 2)

- ✓ **Le rappel** (reccall), également appelé sensibilité, mesure la capacité du modèle à identifier correctement les instances positives. Il correspond à la proportion des cas positifs réels qui sont effectivement détectés par le modèle, ce qui est particulièrement important dans des domaines sensibles comme la santé

- ✓ Rappel=$\frac{TP}{TP+FN}$ (2. 3)

- ✓ **La F-mesure** (ou score F1) combine la précision et le rappel en une seule métrique. Elle permet d'obtenir un compromis entre ces deux indicateurs, offrant ainsi une évaluation plus équilibrée des performances, notamment lorsque les données sont déséquilibrées.

- ✓ F-mesure=$2 * \frac{Reccal \times \text{Précision}}{Reccal \times \text{Précision}}$ (2. 4)

- ✓ **Spécificité** mesure la capacité du modèle à identifier correctement les instances

négatives. Elle correspond à la proportion des cas négatifs réels qui sont effectivement bien classés comme négatifs par le modèle.

- ✓ Spécificité $= \frac{TN}{TP+FP}$ (2. 5)

- ✓ **ROC–AUC (Receiver Operating Characteristic – Area Under the Curve)**

La courbe ROC (*Receiver Operating Characteristic*) est un outil graphique largement utilisé pour évaluer les performances d'un modèle de classification binaire. Elle représente l'évolution du taux de vrais positifs (*True Positive Rate* ou sensibilité) en fonction du taux de faux positifs (*False Positive Rate*) pour différents seuils de décision. Cette courbe permet ainsi d'analyser la capacité du modèle à distinguer les individus appartenant à la classe positive de ceux appartenant à la classe négative.

L'aire sous la courbe ROC, appelée **AUC (Area Under the Curve)**, constitue un indicateur synthétique de la performance globale du modèle. Sa valeur est comprise entre 0 et 1 : une AUC proche de 1 traduit une excellente capacité de discrimination, tandis qu'une valeur proche de 0,5 indique des performances comparables à celles d'une classification aléatoire. Plus la courbe ROC se rapproche du coin supérieur gauche du graphique et plus l'AUC est élevée, plus le modèle est considéré comme efficace pour distinguer correctement les différentes classes.

En complément, la courbe ROC et l'AUC sont généralement interprétées conjointement avec d'autres métriques d'évaluation, telles que l'exactitude (*Accuracy*), la précision (*Precision*), le rappel (*Recall*), le score F1 et la matrice de confusion. Cette approche permet d'obtenir une évaluation plus complète et plus fiable des performances du modèle, particulièrement dans les applications médicales où les conséquences des erreurs de classification peuvent être significatives.

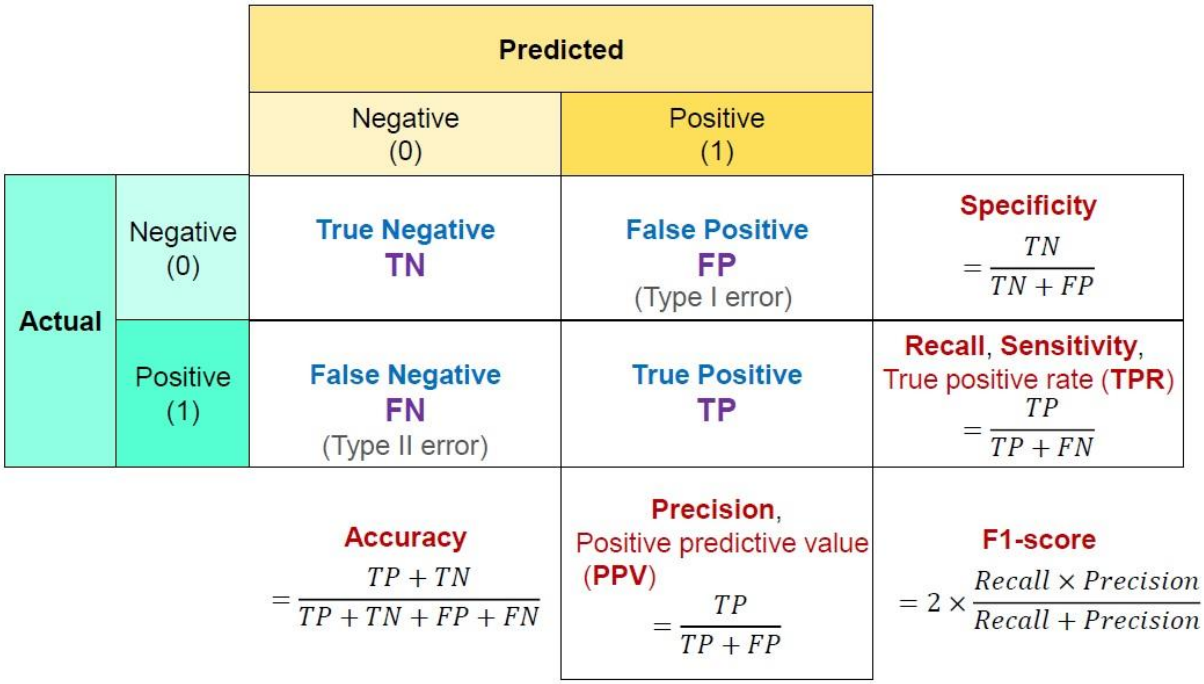


Figure 2.8 : Métriques d'évaluation des modèles[60]

2.6 Application du Machine Learning dans le domaine de la cardiologie :

L'application du **Machine Learning (ML)** en cardiologie constitue aujourd'hui un axe de recherche majeur dans le domaine de la santé numérique. Elle vise à exploiter les données cliniques et biomédicales afin d'améliorer la précision du diagnostic, la prédiction des maladies cardiovasculaires ainsi que l'aide à la décision médicale. En effet, les maladies cardiovasculaires représentent l'une des principales causes de mortalité dans le monde, ce qui rend l'utilisation des méthodes intelligentes particulièrement pertinente pour leur détection précoce et leur suivi.

Dans ce contexte, les algorithmes d'apprentissage supervisé sont largement utilisés pour la classification et la prédiction du risque cardiovasculaire. Des modèles tels que la **régression logistique**, les **machines à vecteurs de support (SVM)**, les **forêts aléatoires (Random Forest)** ainsi que les **réseaux de neurones multicouches (MLP)** permettent d'analyser un ensemble de variables cliniques (âge, pression artérielle, cholestérol, glycémie, fréquence cardiaque, etc.) afin de déterminer la probabilité qu'un patient développe une pathologie cardiaque. Ces modèles sont entraînés sur des bases de données médicales historiques contenant des cas étiquetés, ce qui leur permet de généraliser leurs prédictions sur de nouveaux patients.

Par ailleurs, le ML est également appliqué à l'analyse des signaux électrocardiographiques (ECG). Grâce à des techniques d'extraction de caractéristiques et

de classification automatique, il devient possible d'identifier des anomalies du rythme cardiaque telles que la fibrillation auriculaire, les tachycardies ou encore les bradycardies. Cette automatisation contribue à réduire le temps d'analyse et à améliorer la fiabilité du diagnostic.

En outre, dans le domaine de l'imagerie médicale cardiaque, les approches basées sur le **deep learning**, notamment les réseaux de neurones convolutionnels (CNN), offrent des performances remarquables pour la segmentation des structures cardiaques et la détection des anomalies morphologiques. Ces méthodes permettent une analyse plus fine et plus rapide des images issues de l'échocardiographie ou de l'IRM cardiaque.

Enfin, l'intégration des techniques de Machine Learning en cardiologie s'inscrit dans une démarche de médecine prédictive et personnalisée. Elle permet non seulement d'améliorer la précision diagnostique, mais également d'optimiser la prise en charge thérapeutique en adaptant les décisions médicales au profil spécifique de chaque patient [61].

2.7 Les bases de donnée cardiovasculaire :

Les maladies cardiovasculaires constituent aujourd'hui l'une des principales causes de mortalité à l'échelle mondiale. Elles regroupent un ensemble de pathologies affectant le cœur et les vaisseaux sanguins, notamment l'infarctus du myocarde, l'hypertension artérielle ainsi que l'insuffisance cardiaque. Leur apparition est généralement associée à plusieurs facteurs de risque tels que l'âge, le tabagisme, le diabète, l'hypercholestérolémie, ainsi que divers comportements et habitudes de vie.

Avec les avancées récentes dans le domaine de l'intelligence artificielle et de l'apprentissage automatique, l'exploitation des données médicales à grande échelle est devenue une approche prometteuse pour la détection précoce de ces pathologies. En effet, le développement de modèles prédictifs permet d'identifier les patients à risque à partir de leurs données cliniques, facilitant ainsi une prise en charge anticipée et plus efficace.

Dans ce contexte, plusieurs bases de données publiques sont largement utilisées dans la recherche scientifique, notamment celles issues de l'**UCI Machine Learning Repository**, de la **Framingham Heart Study**, ainsi que des plateformes telles que **Kaggle**. Ces ensembles de données constituent une ressource essentielle pour l'analyse des facteurs de risque et le développement de modèles fiables dédiés à la prédiction des maladies cardiovasculaires.

Ainsi, ce travail s'inscrit dans une démarche de **data science appliquée au domaine de la santé**, visant à améliorer la prévention, la détection précoce et l'aide à la décision clinique grâce aux techniques d'apprentissage automatique.

- **Cleveland Heart Disease Dataset**

Le dataset **Cleveland Heart Disease**, intégré à la base de données de l'**UCI Machine Learning Repository**, est largement utilisé dans les travaux de recherche en apprentissage automatique appliqué à la cardiologie. Il s'agit d'un ensemble de données médicales regroupant des informations cliniques issues d'examens réalisés sur des patients.

Ce dataset contient plusieurs variables médicales pertinentes, notamment l'âge, le sexe, la pression artérielle au repos, le taux de cholestérol, ainsi que différents résultats d'examens cardiaques. Il est principalement utilisé dans des problèmes de classification binaire visant à déterminer la présence ou l'absence d'une maladie cardiaque, ce qui en fait une référence dans les études de prédiction cardiovasculaire [62].

Les principales variables du dataset sont :

- âge
- sexe
- type de douleur thoracique (chest pain type)
- pression artérielle au repos
- cholestérol sérique
- glycémie à jeun
- résultats électrocardiographiques au repos
- fréquence cardiaque maximale atteinte
- angine induite par l'effort
- dépression du segment ST (oldpeak)

Taille de la base de données : (303, 14)

Aperçu du dataset complet (Cleveland) :

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
0	63	1	1	145	233	1	2	150	0	2.3	3	0.0	6.0	0
1	67	1	4	160	286	0	2	108	1	1.5	2	3.0	3.0	2
2	67	1	4	120	229	0	2	129	1	2.6	2	2.0	7.0	1
3	37	1	3	130	250	0	0	187	0	3.5	3	0.0	3.0	0
4	41	0	2	130	204	0	2	172	0	1.4	1	0.0	3.0	0

Figure 2.9 : Taille et aperçu initial de la base de données brute Cleveland Heart Disease Dataset.

- **Framingham Heart Study Dataset**

Le dataset **Framingham Heart Study** provient d'une étude épidémiologique longitudinale de grande envergure menée aux États-Unis, débutée en 1948 dans la ville de Framingham (Massachusetts). Cette étude a pour objectif principal d'identifier les principaux facteurs de risque associés au développement des maladies cardiovasculaires.

Ce jeu de données contient des informations collectées sur plusieurs années auprès de patients suivis dans le temps, permettant ainsi l'analyse de l'évolution de leur état de santé et la compréhension des mécanismes de progression des maladies cardiaques [63] .

Les principales variables incluent :

- âge
- sexe
- tabagisme (smoking)
- diabète
- hypertension artérielle
- taux de cholestérol
- indice de masse corporelle (BMI)
- traitements médicaux
- variable cible : risque de maladie cardiovasculaire à 10 ans

Ce dataset est particulièrement important en data science médicale, car il permet la modélisation de la prédiction à long terme et l'identification des facteurs les plus influents dans le développement des pathologies cardiovasculaires.

✅ Dataset chargé: (4240, 16)

Les 5 premières lignes :

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
0	1	39	4.0	0	0.0	0.0	0	0	0	195.0	106.0	70.0	26.97	80.0	77.0	0
1	0	46	2.0	0	0.0	0.0	0	0	0	250.0	121.0	81.0	28.73	95.0	76.0	0
2	1	48	1.0	1	20.0	0.0	0	0	0	245.0	127.5	80.0	25.34	75.0	70.0	0
3	0	61	3.0	1	30.0	0.0	0	1	0	225.0	150.0	95.0	28.58	65.0	103.0	1
4	0	46	3.0	1	23.0	0.0	0	0	0	285.0	130.0	84.0	23.10	85.0	85.0	0

Figure 2.10 : Taille et aperçu initial de la base de données brute Framingham Heart Study Dataset

- **Dataset Cardiovascular Disease (Kaggle)**

Le dataset **Cardiovascular Disease** disponible sur la plateforme **Kaggle** est largement utilisé dans les applications d'intelligence artificielle pour la prédiction des maladies cardiovasculaires. Il contient des données cliniques et physiologiques relatives à des patients, utilisées principalement dans des tâches de classification binaire.

Chaque observation représente un patient décrit par plusieurs caractéristiques telles que l'âge, le sexe, la pression artérielle, le taux de cholestérol et la glycémie. La variable cible indique la présence ou l'absence d'une maladie cardiovasculaire [64] .

Les principales variables sont :

- âge et sexe
- pression artérielle
- cholestérol
- glycémie
- indice de masse corporelle (selon les versions du dataset)
- habitudes de vie et paramètres de santé

La variable cible est binaire :

- 1 : présence d'une maladie cardiovasculaire
- 0 : absence de maladie cardiovasculaire

Ce dataset est particulièrement utilisé en raison de sa simplicité, de sa structure

✅ Dataset chargé: (70000, 13)

Les 5 premières lignes :

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	2	18857	1	165	64.0	130	70	3	1	0	0	0	1
3	3	17623	2	169	82.0	150	100	1	1	0	0	1	1
4	4	17474	1	156	56.0	100	60	1	1	0	0	0	0

équilibrée et de son intérêt dans les applications de classification en santé.**Figure 2.11:** Taille et aperçu initial de la base de données brute Dataset Cardiovascular

- **Étude comparative des différentes bases de données :**

Critère	Framingham	UCI Heart	Kaggle CVD
Taille	4 621 patients	303 patients	70 000 patients
Variables médicale	30+ variables (ECG, lipides détaillés, facteurs cliniques avancés	14 variables (signes cliniques basiques)	~12 variables (features simplifiées)
Qualité des données	Données de référence NIH (très fiable)	Bonne qualité mais dataset limité	Données anonymisées et hétérogènes
Suivi temporel	Suivi longitudinal sur ~30 ans	Aucun suivi temporel	Aucun suivi temporel
Validation clinique	Utilisé dans plus de 3000 études médicales	Utilisé dans ~100+ études ML	Faible validation clinique directe
Complexité pour IA	Très riche, réaliste, difficile à modéliser	Moyenne, adaptée aux tests ML	Simple, optimisée pour machine learning

Tableau 2.1: Synthèse comparative des différentes base de données

- **Notre choix de base de donne pour le modèle :**

Dans le cadre de ce travail, nous avons retenu la base de données Cleveland Heart Study, largement reconnue dans la littérature scientifique en apprentissage automatique appliqué à la cardiologie. Ce jeu de données regroupe des informations cliniques collectées auprès de patients, incluant plusieurs variables pertinentes telles que l'âge, la pression artérielle, le taux de cholestérol, ainsi que divers facteurs de risque cardiovasculaire (tabagisme, type de douleur thoracique, fréquence cardiaque, entre autres).

Le choix de cette base repose sur sa fiabilité, sa richesse descriptive et son utilisation fréquente comme benchmark dans les travaux de recherche en prédiction des maladies cardiovasculaires.

Elle permet ainsi l'entraînement et l'évaluation de modèles d'apprentissage automatique supervisé dans le but de classifier efficacement la présence ou l'absence de pathologies cardiovasculaires, tout en garantissant des résultats comparables aux études existantes.

2.8. Analyse exploratoire des données (EDA) :

L'analyse exploratoire des données (*Exploratory Data Analysis – EDA*) constitue une étape fondamentale du processus de fouille de données et d'apprentissage automatique. Elle vise à examiner et à comprendre les caractéristiques du jeu de données avant la phase de modélisation. Cette étape permet d'identifier la structure des données, de détecter les valeurs manquantes, les valeurs aberrantes ainsi que les éventuelles incohérences pouvant affecter la qualité des analyses.

Dans le contexte de la prédiction des maladies cardiovasculaires, l'EDA consiste à étudier les principales variables cliniques telles que l'âge, la pression artérielle, le taux de cholestérol, la glycémie ou encore la fréquence cardiaque. Elle repose sur l'utilisation de statistiques descriptives et de techniques de visualisation permettant de mettre en évidence les distributions des variables ainsi que les relations existantes entre celles-ci.

Cette phase préliminaire joue un rôle essentiel dans la préparation des données, puisqu'elle contribue à améliorer la qualité des données utilisées et à optimiser les performances des modèles prédictifs développés par la suite [65].

2.9. Problèmes liés aux données médicales

Les données médicales présentent plusieurs défis susceptibles d'influencer la qualité des analyses et la performance des modèles d'apprentissage automatique. Parmi les principales difficultés rencontrées figurent les données manquantes, le déséquilibre de classe et le bruit des données [66] .

- **Données manquantes**

Les données manquantes représentent l'un des problèmes les plus fréquents dans les bases de données médicales. Elles surviennent lorsque certaines informations relatives aux patients ne sont pas disponibles ou n'ont pas été correctement enregistrées. Cette situation peut être due à l'absence de certains examens médicaux, à des erreurs de saisie ou à des pertes d'informations lors de la collecte des données.

Dans les bases de données cardiovasculaires, les valeurs manquantes peuvent concerner des variables importantes telles que la pression artérielle, le taux de cholestérol ou la glycémie. Leur présence peut entraîner une diminution de la qualité des analyses et affecter la capacité des modèles à produire des prédictions fiables.

Afin de remédier à ce problème, plusieurs approches peuvent être utilisées, notamment la suppression des observations incomplètes ou l'imputation des valeurs manquantes à l'aide de méthodes statistiques telles que la moyenne, la médiane ou des techniques plus avancées basées sur l'apprentissage automatique.

- **Déséquilibre de classe**

Le déséquilibre de classe se produit lorsque les différentes catégories d'un jeu de données ne sont pas représentées de manière équivalente. Dans le domaine cardiovasculaire, il est fréquent que le nombre de patients sains soit supérieur à celui des patients atteints d'une maladie cardiaque.

Cette situation peut entraîner un biais d'apprentissage, les modèles ayant tendance à favoriser la classe majoritaire au détriment de la classe minoritaire. Par conséquent, les cas pathologiques, pourtant les plus importants à identifier dans un contexte médical, risquent d'être mal détectés.

Pour atténuer ce problème, différentes stratégies peuvent être mises en œuvre, telles que le sur-échantillonnage (*oversampling*) de la classe minoritaire, le sous-échantillonnage (*undersampling*) de la classe majoritaire, l'attribution de poids spécifiques aux classes ou encore l'utilisation d'algorithmes conçus pour gérer les jeux de données déséquilibrés.

- **Bruit des données**

Le bruit correspond aux erreurs, imprécisions ou incohérences présentes dans les données collectées. Dans les bases de données médicales, il peut résulter d'erreurs de saisie, de mesures incorrectes, de dysfonctionnements des dispositifs médicaux ou encore de variations liées aux procédures de collecte.

La présence de bruit peut masquer les relations réelles entre les variables cliniques et la pathologie étudiée, ce qui réduit la précision des modèles d'apprentissage automatique et leur capacité de généralisation.

Afin de limiter l'impact du bruit, il est nécessaire d'effectuer un prétraitement rigoureux des données comprenant le nettoyage des données, la détection et le traitement des valeurs aberrantes, ainsi que la normalisation ou la standardisation des variables. Ces opérations permettent d'améliorer la qualité des données et de renforcer la robustesse des modèles prédictifs.

Conclusion

Ce chapitre a permis de présenter les concepts fondamentaux de l'intelligence artificielle et de l'apprentissage automatique appliqués au domaine biomédical. L'étude des différents types d'apprentissage, ainsi que des principaux algorithmes de classification utilisés dans le secteur de la santé, a mis en évidence leur importance dans l'analyse et la prédiction des pathologies cardiovasculaires.

Les critères d'évaluation des modèles prédictifs ont également été exposés afin de souligner l'importance d'une mesure rigoureuse des performances lors du développement de solutions basées sur l'intelligence artificielle. De plus, les diverses applications du Machine Learning en cardiologie ont démontré le potentiel de ces technologies pour assister les professionnels de santé dans le diagnostic précoce et la prise de décision clinique.

Par ailleurs, une attention particulière a été accordée aux bases de données

cardiovasculaires les plus utilisées dans la littérature scientifique, ainsi qu'aux étapes essentielles de préparation des données. L'analyse exploratoire des données a notamment été présentée comme une phase indispensable pour comprendre les caractéristiques des données et identifier les éventuelles anomalies pouvant influencer les résultats de la modélisation.

Enfin, les principaux défis liés aux données médicales, notamment les valeurs manquantes, le déséquilibre des classes et le bruit, ont été examinés en raison de leur impact sur la fiabilité et la précision des modèles d'apprentissage automatique. Les notions développées dans ce chapitre constituent ainsi le socle théorique sur lequel repose la démarche méthodologique adoptée dans ce mémoire. Elles serviront de fondement à la phase expérimentale consacrée à la conception, à l'entraînement et à l'évaluation des modèles de prédiction des maladies cardiovasculaires présentés dans les chapitres suivants.

Chapitre 3 :

Approche pratique de la prédiction des maladies cardiovasculaires par apprentissage automatique : évaluation des modèles et analyse des résultats

Chapitre 3

Approche pratique de la prédiction des maladies cardiovasculaires par apprentissage automatique : évaluation des modèles et analyse des résultats

Introduction :

Ce troisième chapitre est dédié à la mise en œuvre pratique et expérimentale de notre système d'aide au diagnostic. Après avoir posé les bases théoriques, nous détaillons ici l'ensemble du pipeline technique, depuis la préparation rigoureuse des données cliniques du Cleveland Heart Disease Dataset jusqu'à l'évaluation fine des algorithmes de Machine Learning. À travers un protocole standardisé incluant la sélection, la standardisation et l'optimisation des hyperparamètres, nous comparons les performances de quatre architectures : les SVM, les forêts aléatoires (Random Forest), les k-plus proches voisins (KNN) et les réseaux de neurones. L'objectif final est d'identifier le modèle optimal capable de prédire avec la plus grande précision la présence d'anomalies cardiovasculaires, tout en analysant ses apports et ses limites en contexte médical.
Pour amorcer cette démarche expérimentale, la première étape indispensable consiste à traiter la base de données brute afin d'en garantir la fiabilité. C'est l'objet du Nettoyage des données, une phase critique pour éliminer les bruits et les incohérences physiologiques susceptibles d'altérer l'apprentissage de nos modèles.

3.1 Nettoyage des données :

Le nettoyage des données est une étape fondamentale du pré-traitement, visant à identifier et corriger les erreurs, incohérences et anomalies d'un jeu de données. Cela garantit la fiabilité et la qualité des données avant l'analyse, crucial dans la prédiction des anomalies cardiovasculaires. Des données incorrectes, comme un âge négatif ou une pression artérielle irréaliste, peuvent fausser les résultats. Les imperfections comme les valeurs hors plage ou le bruit nuisent à la performance des modèles d'apprentissage automatique[67],[68] .

Concrètement, le processus de nettoyage repose sur plusieurs opérations principales. Il commence par la détection et la suppression des doublons afin d'éviter la redondance des informations (dont le model apprend trop à partir des mêmes données). Ensuite, une

vérification de la validité des valeurs est réalisée, notamment en contrôlant les bornes physiologiques (comme la fréquence cardiaque ou la pression artérielle) ainsi que les formats des variables (par exemple les codes médicaux ou les types de données)[69].

Le traitement des valeurs aberrantes (outliers) constitue également une étape que l'on a effectué , celles-ci sont généralement détectées à l'aide de méthodes statistiques comme l'IQR (intervalle interquartile) ou le Z-score, puis soit supprimées, soit remplacées par des valeurs plus plausibles[70],[71]. Par ailleurs, les erreurs de format sont corrigées en convertissant les types de données (par exemple de texte vers numérique) et en éliminant les caractères indésirables [72].

Dans la pratique, ces opérations sont souvent mises en œuvre à l'aide d'outils comme les bibliothèques Python **Pandas** et **NumPy**, qui permettent de manipuler efficacement les données, de détecter les anomalies et d'automatiser les corrections[73],[74].

3.2 Gestion des valeurs manquantes :

La gestion des valeurs manquantes consiste à identifier et traiter les données absentes dans un dataset, souvent représentées par des valeurs nulles ou **NaN**. Ces absences peuvent être dues à des erreurs de saisie, des pannes de capteurs ou des informations non renseignées lors de la collecte des données. Leur traitement est une étape essentielle du prétraitement, car la majorité des algorithmes d'apprentissage automatique ne supportent pas directement les valeurs manquantes. Leur présence peut provoquer des erreurs de calcul ou introduire des biais dans le modèle, ce qui est particulièrement critique dans le domaine cardiovasculaire où une donnée manquante (par exemple la tension artérielle) peut fausser la prédiction et conduire à des décisions erronées[75] .

La gestion des données manquantes repose sur plusieurs stratégies adaptées selon la nature des variables. La méthode la plus simple consiste à supprimer les lignes ou colonnes contenant des valeurs manquantes via la méthode «dropna», mais cette approche n'est recommandée que lorsque le taux de données manquantes est très faible(souvent inférieur à 5 %) afin d'éviter une perte importante d'information. Pour les variables numériques continues, lorsque la distribution est stable et peu influencée par les valeurs extrêmes, on privilégie l'imputation par la moyenne ou la médiane, ce qui permet de conserver l'ensemble des observations tout en limitant les biais. Pour les variables catégorielles, l'imputation par la modalité la plus fréquente est utilisée afin de préserver la cohérence des

données. Cette combinaison de méthodes permet d'obtenir un dataset complet et exploitable pour les étapes suivantes du pré-traitement [76] ,[77] .

3.3 Encodage des variables catégorielles :

L'encodage des variables catégorielles correspond au processus de transformation des données qualitatives en représentations numériques exploitables par les algorithmes d'apprentissage automatique , car la majorité des modèles de machine learning ne peuvent traiter que des données numériques. En l'absence d'un encodage approprié, les variables textuelles peuvent soit provoquer des erreurs d'exécution , soit introduire des biais, notamment en suggérant un ordre artificiel entre des catégories qui n'en possèdent pas. Dans ce travail, l'encodage de type One-Hot Encoding est privilégié pour les variables nominales, car il consiste à créer une variable binaire pour chaque catégorie sans introduire de relation d'ordre, généralement à l'aide de la fonction get_dummies() de la bibliothèque Pandas[78],[79].

3.4 Sélection des caractéristiques (Feature Selection) :

La sélection de caractéristiques (ou Feature Selection) vise à identifier et conserver uniquement les variables les plus pertinentes et informatives d'un dataset, tout en éliminant celles qui sont redondantes, non discriminantes ou bruitées. L'idée est de diminuer la dimensionnalité sans perdre en performance du modèle , car elle permet de simplifier ce dernier , d'accélérer son entraînement, d'améliorer sa précision en se concentrant sur les variables clés, et en meme temps d'éviter le surapprentissage (overfitting) et de rendre les résultats plus interprétables, ce qui est particulièrement essentiel dans les domaines médicaux où l'identification des facteurs clés est primordiale.

Il y en a trois méthodes de sélection : Les méthodes Filter reposent sur des tests statistiques (choisieses dans notre code) pour évaluer chaque variable indépendamment de tout modèle ; elles sont rapides et simples. Les méthodes Wrapper, utilisent un modèle de machine learning pour tester différentes combinaisons de variables et sélectionner celles qui optimisent les performances . Enfin, les méthodes Embedded intègrent directement la sélection des variables dans le processus d'apprentissage du modèle [80],[81].

3.5 Normalisation / Standardisation :

Dans le contexte cardiovasculaire, la pression artérielle (en mmHg), le taux de cholestérol (en mg/dL) et l'âge (en années) sont représentés sur des échelles très différentes. Ce décalage peut présenter une problématique de sorte que "les variables ayant les valeurs les plus élevées risquent de dominer les autres lors de l'apprentissage " , ce qui peut entraîner des résultats erronés et perturbe plusieurs algorithmes.

Pour pallier cet déséquilibre, on applique La normalisation et la standardisation dont ce sont des techniques de mise à l'échelle des données numériques pour les ramener à une plage commune (normalisation) ou à une distribution centrée (standardisation). Elles éliminent les biais dus aux différences d'unités ou d'échelles entre variables.

Dans ce mémoire, la standardisation (StandardScaler) est privilégiée pour les données médicales, car elle conserve la distribution des variables et s'adapte à leur variabilité(des plages habituelles mais aussi des cas pathologiques), tandis que la normalisation Min-Max est utilisée lorsque les variables présentent des bornes connues ou nécessitent une mise à l'échelle dans un intervalle spécifique [82],[83].

3.6 Réduction de dimension (PCA – optionnel) :

La PCA (Analyse en Composantes Principales) est une technique mathématique qui transforme un grand nombre de variables corrélées en un petit nombre de variables non corrélées appelées "composantes principales", en les projetant sur un espace de dimensions inférieures , tout en conservant le maximum d'information du dataset original.

Elle est utilisée pour réduire la dimensionnalité des données lorsque plusieurs variables sont corrélées. Cette technique permet de compresser l'information essentielle, de diminuer le bruit et les redondances, d'accélérer l'entraînement des modèles, d'améliorer la généralisation et de faciliter la visualisation des données.

La méthode consiste à centrer les données, calculer les relations entre variables, puis projeter les données sur les composantes principales les plus importantes. Le nombre de composantes est choisi de manière à conserver environ 90 % à 95 % de la variance expliquée par la base de donnée [84].

3.7 Division de la base de données (Train / Test / Validation) :

La division du dataset consiste à séparer les données en plusieurs ensembles distincts , généralement entraînement, validation et test , afin d'entraîner, ajuster et évaluer le modèle de manière objective. Cette étape permet de mesurer la capacité de généralisation du modèle sur des données jamais vues, d'éviter le surapprentissage et d'obtenir une évaluation fiable des performances. Dans ce mémoire, la séparation est réalisée avec «train_test_split» de scikit-learn en utilisant la stratification «(stratify=y)» pour conserver la répartition des classes, avec des proportions d'environ 70 % pour l'entraînement, 15 % pour la validation et 15 % pour le test [85].

3.8 Visualisation et analyse statistique :

Avant de lancer l'entraînement de nos modèles de Machine Learning, un passage par l'analyse de nos données s'impose. Pour ce travail, nous nous appuyons sur le Cleveland Heart Disease Dataset, une base de données médicale incontournable issue du répertoire UCI. Ce jeu de données rassemble les dossiers cliniques de 303 patients, décrits à travers 14 variables clés comme l'âge, le taux de cholestérol, la pression artérielle ou encore la nature des douleurs thoraciques. Ici, tout l'enjeu est de réussir à prédire la présence ou l'absence d'une anomalie cardiaque (notre variable cible num). Afin de mesurer concrètement l'apport de notre protocole de préparation (étapes 1 à 7), nous présentons une comparaison visuelle et statistique des données, avant et après leur prétraitement.

```
Taille de la base de données : (303, 14)
--------------------------------------------------
Aperçu du dataset complet (Cleveland) :
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
0	63	1	1	145	233	1	2	150	0	2.3	3	0.0	6.0	0
1	67	1	4	160	286	0	2	108	1	1.5	2	3.0	3.0	2
2	67	1	4	120	229	0	2	129	1	2.6	2	2.0	7.0	1
3	37	1	3	130	250	0	0	187	0	3.5	3	0.0	3.0	0
4	41	0	2	130	204	0	2	172	0	1.4	1	0.0	3.0	0

Figure 3.1: Aperçu général et structure des variables du *Cleveland Heart Disease Dataset*.

3.8.1 Résultats du script « Nettoyage des données » :

Les résultats de notre script de nettoyage et l'analyse visuelle de la variable trestbps (pression artérielle au repos) valident l'efficacité de notre protocole de prétraitement :

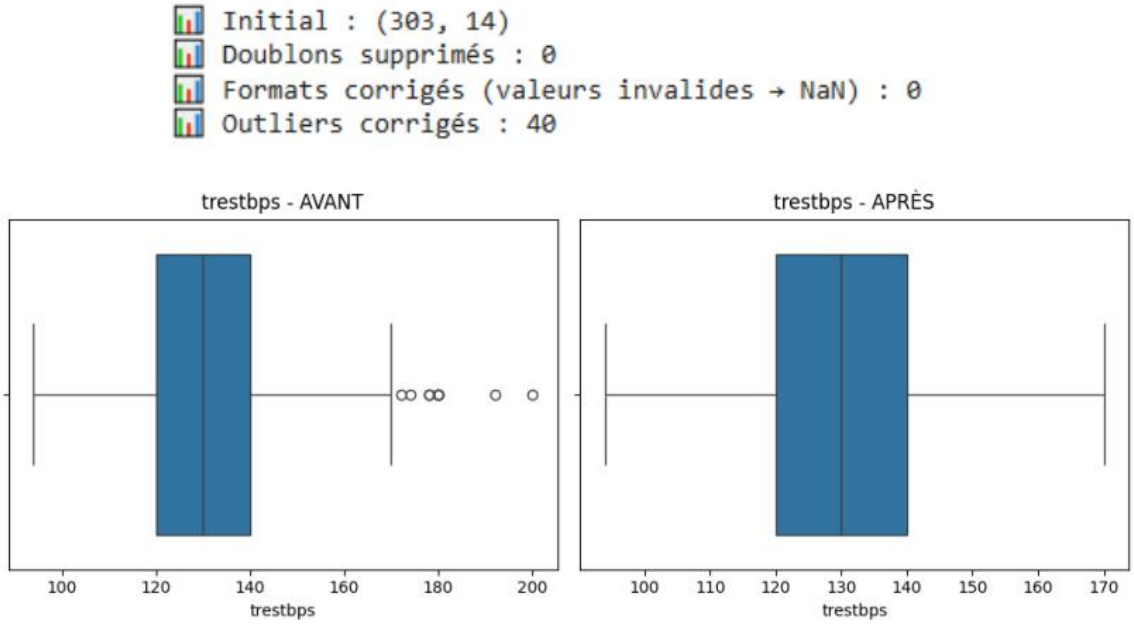


Figure 3.2: Impact du nettoyage des données, avec focus sur la correction des valeurs aberrantes de la variable trestbps .

Légende : Le bilan textuel (en haut) confirme la correction de 40 outliers. Les boxplots (en dessous) montrent la suppression des valeurs extrêmes supérieures à 170 mmHg, stabilisant la distribution avant l'entraînement des modèles.

Le volume initial de la base de données est préservé à (303, 14), confirmant qu'aucun doublon n'était présent (0) ou supprimé et qu'aucune valeur n'affichait de format invalide exigeant une conversion en valeur manquante (NaN = 0). Cela témoigne de la bonne qualité initiale du Cleveland Heart Disease Dataset.

Notre algorithme a identifié et corrigé avec succès 40 valeurs aberrantes sur l'ensemble des variables. L'impact de cette correction est parfaitement visible sur les diagrammes en boîte (boxplots) de la variable trestbps.

- **Avant nettoyage :** Le graphique de gauche met en évidence plusieurs points isolés au-delà de la valeur de 170 mmHg (allant jusqu'à 200mmHg). Ces valeurs anormalement élevées faussent la réalité des chiffres et risquent de tromper nos modèles de Machine Learning pendant leur apprentissage.

- **Après nettoyage :** Le graphique de droite montre une distribution parfaitement nettoyée et resserrée sur des bornes physiologiques plus stables (maximum stabilisé à 170 mmHg). Les points isolés ont disparu, ce qui élimine le bruit statistique tout en conservant la structure centrale de la distribution (la médiane et l'écart interquartile restent inchangés).

Il est important de préciser que ce même processus (incluant la suppression des doublons, la correction des formats et le traitement des valeurs aberrantes) a été appliqué avec succès à toutes les autres variables de notre jeu de données ; par souci de concision, leurs graphiques "avant/après" sont mis à disposition en **Annexe B**.

3.8.2 Résultats du script « Gestion des valeurs manquantes» :

La Figure 3.3 confirme l'efficacité du traitement appliqué pour éliminer les données absentes du jeu de données :

- **État initial (Avant traitement) :** Le premier bilan textuel ainsi que la carte de chaleur (heatmap) de gauche mettent en évidence la présence de 6 valeurs manquantes (les lignes jaunes) localisées de manière isolée sur deux variables spécifiques : 4 valeurs manquantes pour la variable ca (nombre de vaisseaux colorés par fluoroscopie) et 2 valeurs manquantes pour la variable thal (type de thalassémie). Les autres variables cliniques sont totalement complètes.
- **Résultat final (Après traitement) :** Le second bilan textuel et la carte de chaleur de droite (devenue entièrement violette et uniforme) démontrent que toutes les données manquantes ont été corrigées avec succès. Chaque variable affiche désormais un compteur de valeurs manquantes égal à 0 .

La visualisation des valeurs manquantes montre que le jeu de données contenait initialement un faible nombre de données absentes, localisées uniquement dans les variables *ca* et *thal*. Après l'étape de prétraitement, aucune valeur manquante n'est observée, indiquant que les données manquantes ont été correctement traitées. Le jeu de données final est donc complet et prêt pour les étapes suivantes.

VALEURS MANQUANTES (AVANT)

age	0
sex	0
cp	0
trestbps	0
chol	0
fbs	0
restecg	0
thalach	0
exang	0
oldpeak	0
slope	0
ca	4
thal	2
num	0

VALEURS MANQUANTES (APRÈS)

age	0
sex	0
cp	0
trestbps	0
chol	0
fbs	0
restecg	0
thalach	0
exang	0
oldpeak	0
slope	0
ca	0
thal	0
num	0

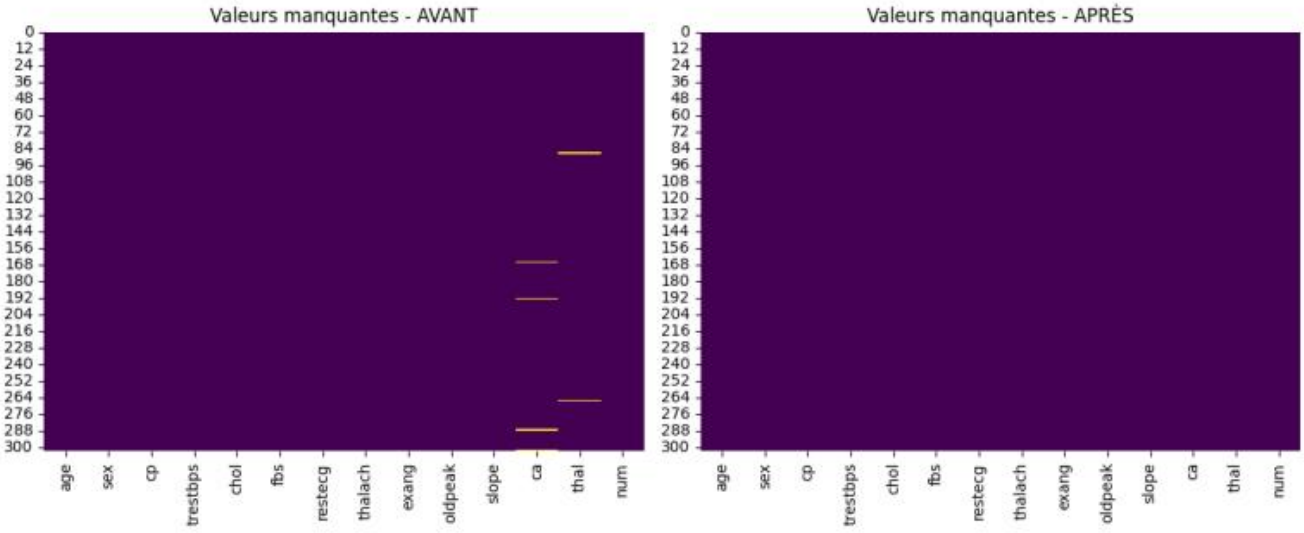


Figure 3.3 : Impact du traitement des valeurs manquantes sur le *Cleveland Heart Disease Dataset* .

3.8.3 Résultats du script « Encodage des variables catégorielles» :

La Figure 3.4 démontre visuellement et statistiquement l'impact de la transformation de nos variables qualitatives en valeurs numériques :

- **Variables ciblées :** L'encodage a été appliqué de manière sélective sur 4 variables spécifiques du jeu de données : cp (type de douleur thoracique), restecg (résultats de l'électrocardiogramme au repos), slope (pente du segment ST à l'effort) et thal (type de thalassémie).

- **Modification de la structure (Shape) :** Le bilan montre que la taille du jeu de données est passée de (303, 14) à (303, 19). Les 303 dossiers de patients sont rigoureusement conservés, tandis que le nombre total de caractéristiques passe à 19 colonnes.

- Lorsqu'on applique le *One-Hot Encoding*, chaque catégorie unique d'une variable textuelle ou nominale est convertie en une colonne indépendante. Dans ces nouvelles colonnes, la présence d'une caractéristique est codée par la valeur 1 (Vrai) et son absence par la valeur 0 (Faux).

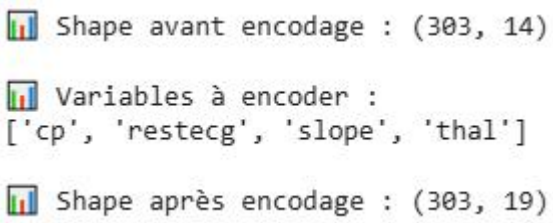


Figure 3.4 : Évolution de la structure de la base de données après application de l'encodage One-Hot

3.8.4 Résultats du script « Sélection des caractéristiques » :

La méthode de sélection des caractéristiques appliquée a permis de classer les variables selon leur capacité à discriminer les patients atteints d'une maladie cardiovasculaire des patients sains, comme illustre la figure 3.5. Plus le score attribué à une variable est élevé, plus son influence sur la variable cible est
importante .

📌 Classement complet des variables importantes :

Feature	Score
cp_4.0	109.490091
thal_7.0	90.396519
ca	86.042821
exang	69.020891
oldpeak	68.400247
thalach	63.587697
slope_2.0	43.602005
cp_3.0	33.189667
sex	24.978872
cp_2.0	19.841940
age	15.769637
restecg_2.0	7.948641
trestbps	6.462598
thal_6.0	3.346715
chol	3.284557
restecg_1.0	1.382024
slope_3.0	1.150628
fbs	0.192237

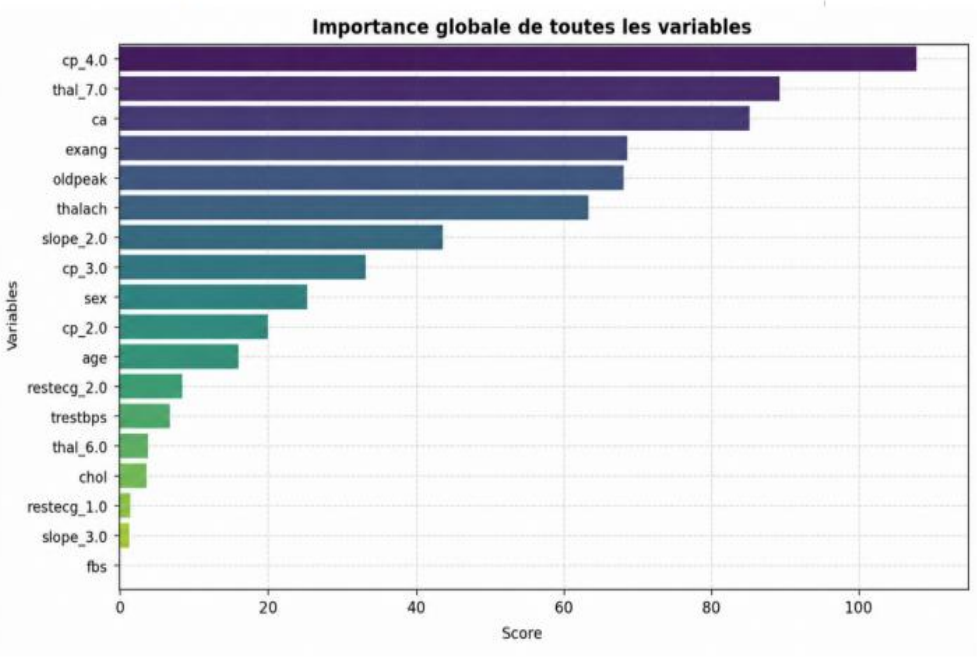


Figure 3.5: Analyse et classement de l'importance des variables

On observe une forte décroissance des scores après les premières variables, indiquant que l'essentiel de l'information prédictive est concentré dans un nombre limité de caractéristiques.

Ces variables(figure 3.6) apparaissent comme les facteurs les plus pertinents pour prédire la présence d'une maladie cardiovasculaire dans le base de données étudiée.

À l'inverse, certaines variables présentent une faible contribution à la discrimination des classes, leur faible score suggère qu'elles apportent peu d'information discriminante .

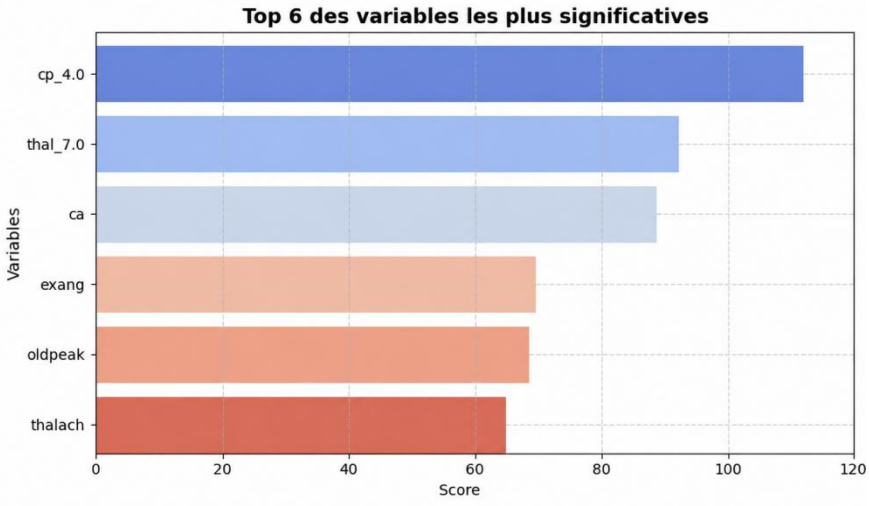


Figure 3.6 : Classement des six variables cliniques les plus discriminantes issues du test statistique de sélection

Bien que l'analyse par la méthode Filter a permis d'établir avec succès la hiérarchie de ces scores, nous avons choisi de ne pas appliquer de réduction de caractéristiques pour l'entraînement final de nos modèles, car notre dimensionnalité est initialement faible : Le jeu de données de Cleveland ne comporte que 14 variables initiales (portées à 19 après encodage One-Hot). Ce volume reste extrêmement restreint. La sélection de caractéristiques trouve sa véritable utilité dans des bases de données de haute dimensionnalité (comme des données génomiques ou imagerie) .

3.8.5 Résultats du script « Normalisation / Standardisation » :

La figure 3.7 traduit visuellement l'efficacité de notre protocole de mise à l'échelle sur thalach et oldpeak, qui constituent les deux seules variables numériques parmi les six caractéristiques sélectionnées précédemment.

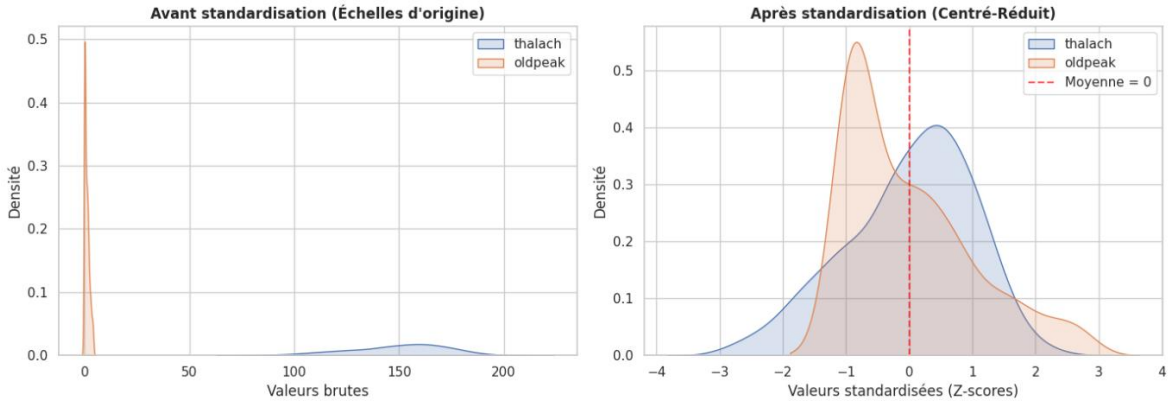


Figure 3.7: Effet de la standardisation par StandardScaler sur la distribution des variables numériques thalach et oldpeak

Avant standardisation :

- thalach présente des valeurs brutes beaucoup plus grandes, autour de 100 à 180.
- oldpeak reste concentrée sur une plage très faible, proche de 0 à 3.
- Cette différence d'échelle signifie qu'un algorithme sensible aux distances ou aux coefficients pourrait donner plus d'importance à thalach qu'à oldpeak.

Après standardisation :

- Les deux variables sont transformées en scores centrés-réduits.
- La ligne verticale rouge indique une moyenne égale à 0, ce qui confirme que la standardisation recentre les données.
- Les distributions gardent leur forme générale, mais elles deviennent comparables en amplitude, ce qui évite qu'une variable domine simplement parce qu'elle a des valeurs plus grandes.

La **standardisation** est utile en contexte médical parce qu'elle permet de comparer des variables comme thalach et oldpeak (mesures physiologiques) sur **une même base**, sans supprimer leur **variabilité naturelle**. Elle est particulièrement adaptée lorsque les données n'ont pas les mêmes unités ou les mêmes ordres de grandeur, ce qui est fréquent dans les variables cardiovasculaires.

3.8.6 Résultats du script « Réduction de dimension (PCA) » :

Les résultats de l'Analyse en Composantes Principales (ACP) et leur représentation graphique(figure 3.8) mettent en évidence l'efficacité de cette méthode pour réorganiser et synthétiser les données, tout en facilitant la distinction entre les patients sains et les patients malades. Il convient toutefois de préciser que l'objectif premier de l'ACP n'est pas uniquement de simplifier cette distinction visuelle, mais avant tout de réduire la dimensionnalité du jeu de données en éliminant les redondances et en compressant l'information essentielle:

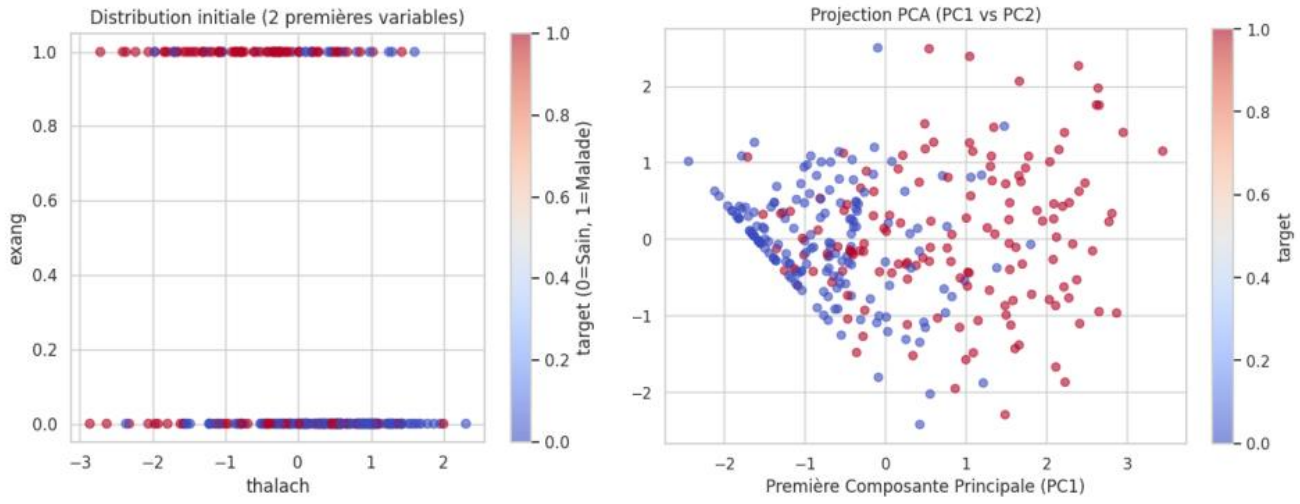


Figure 3.8: Comparaison de la distribution spatiale des patients entre les variables initiales et le plan factoriel PC1 vs PC2

- La Distribution initiale (Graphique à gauche) confrontant thalach et exang montre des données écrasées, bloquées et alignées sur deux lignes horizontales fixes (car la variable exang vaut soit 0, soit 1). Les points bleus (sains) et rouges (malades) y sont complètement mélangés. Cette forte superposition reflète la complexité multidimensionnelle des données et montre que les informations discriminantes sont

réparties sur plusieurs variables, rendant difficile la séparation visuelle des deux classes dans un espace bidimensionnel.

Pour un algorithme de classification, cette structure est un véritable cauchemar car il est très difficile d'y tracer une frontière nette.

- À l'inverse, la Projection PCA (Graphique à droite) fait disparaître cette rigidité. L'ACP a combiné l'ensemble des variables d'origine (âge, cholestérol, tension, etc.) pour générer quatre nouveaux axes mathématiques (les composantes principales). Afin de pouvoir projeter ces données en deux dimensions sur notre écran, seuls les deux premiers axes imaginaires (PC1 et PC2) qui capturent la majorité de l'information (67%) ont été utilisés. Les données explosent alors en un nuage bien étalé où une tendance naturelle se dessine : les patients sains (en bleu) se regroupent à gauche, tandis que les patients malades (en rouge) se concentrent à droite. Si l'on devait décrire une image mentale, c'est comme si l'ACP agit ici comme des "lunettes magiques" : elle révèle une séparation naturelle basée sur les caractéristiques physiques communes des patients.

Toutefois, un certain chevauchement subsiste entre les deux groupes, ce qui est attendu puisque l'ACP est une méthode non supervisée dont l'objectif principal est de maximiser la variance des données et non d'optimiser la séparation des classes. Ces résultats montrent néanmoins que l'ACP parvient à condenser l'information la plus pertinente dans un espace de dimension réduite, facilitant ainsi l'exploration visuelle et l'analyse des données.

Le tracé de la figure 3.9 montre la quantité d'information capturée par chaque nouvel axe (composante).

- La première composante (PC1) contient à elle seule près de 48 % de l'information, et la deuxième (PC2) monte le total à environ 67 %. La ligne rouge en pointillés indique que le seuil académique de 90 % de l'information globale (variance expliquée) est franchi exactement à la 4ème composante.

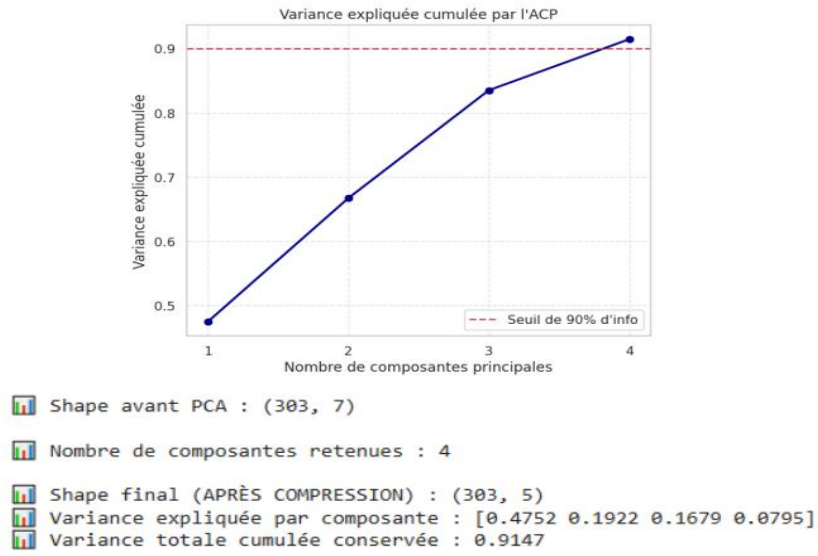


Figure 3.9: Évaluation de la variance expliquée cumulée et réduction de la structure mathématique du jeu de données

- Le bilan textuel confirme que le volume final des données après compression est de (303, 5). Cela signifie que l'algorithme a remplacé nos variables d'origine par seulement 4 composantes principales (suffisantes pour garder 90 % de l'information médical utile), plus la colonne de notre variable cible (target).

Bien que cette analyse exploratoire par ACP démontre avec succès sa capacité à simplifier et à séparer visuellement nos données, nous avons choisi de ne pas utiliser cette approche pour l'entraînement final de nos modèles de Machine Learning.

Cette décision se justifie par deux raisons :

- **Une dimensionnalité initialement faible :** Le jeu de données de Cleveland ne comporte que 14 caractéristiques d'origine (portées à 19 après l'encodage One-Hot). Ce volume reste très restreint. La réduction de dimension par ACP trouve sa véritable utilité dans des bases de données massives (comme la génétique ou l'imagerie) contenant des centaines ou des milliers de variables. Ici, notre ordinateur n'a pas besoin de cette compression pour accélérer ses calculs.
- **La préservation de l'interprétabilité médicale :** L'ACP crée des axes mathématiques abstraits (PC1, PC2) qui mélangent les données. En faisant cela, elle détruit la

signification clinique directe des variables d'origine.

3.8.7 Résultats du script « Division de la base de données » :

Les résultats de notre script de partitionnement et l'analyse visuelle de la distribution des classes(figure 3.10)valident l'efficacité de notre protocole de division du dataset :

```
📊 Dimensions finales :
└─ Train      (70%) : (212, 5)
└─ Validation (15%) : (45, 5)
└─ Test       (15%) : (46, 5)
```

```
📊 Distribution des classes dans chaque échantillon :

[Train Set]
target
0    54.25
1    45.75
Name: proportion, dtype: float64

[Validation Set]
target
0    53.33
1    46.67
Name: proportion, dtype: float64

[Test Set]
target
0    54.35
1    45.65
Name: proportion, dtype: float64
```

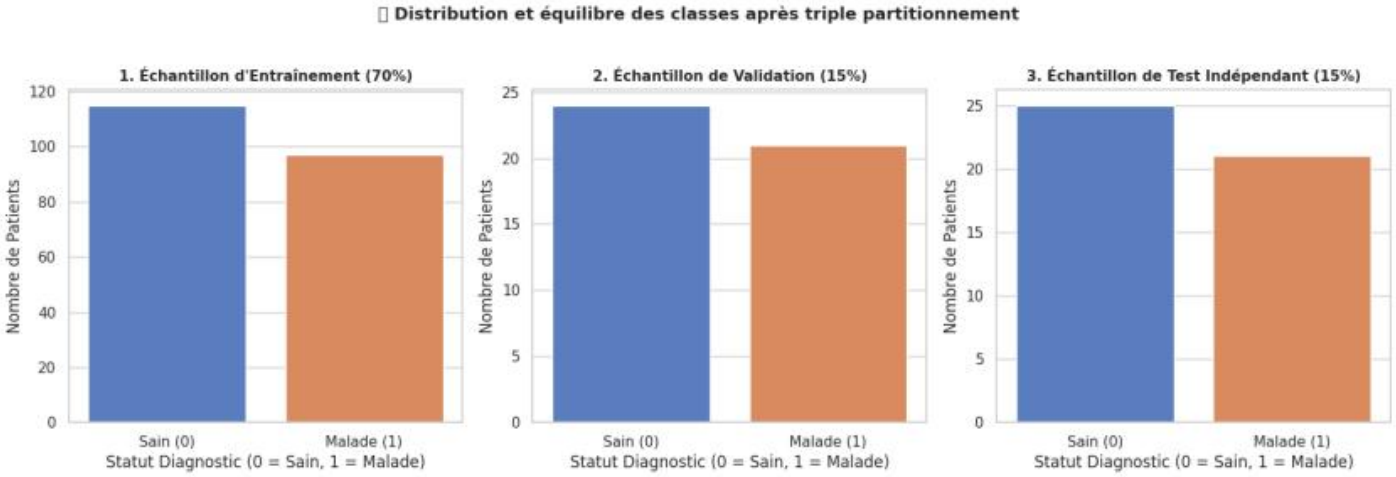


Figure 3.10: Analyse statistique et visuelle du triple partitionnement stratifié du dataset

- Les 303 dossiers de patients ont été segmentés selon les proportions cibles fixées. L'ensemble d'entraînement (Train) reçoit la majorité des données avec 212 patients (70 %) pour construire les règles d'apprentissage. Les ensembles de validation et de test se partagent équitablement le reste avec respectivement 45 patients (15 %) pour l'ajustement des hyperparamètres et 46 patients (15 %) pour l'évaluation finale indépendante.

- Les pourcentages textuels et les diagrammes en barres montrent que la proportion de patients sains (Classe 0, environ 54 %) et malades (Classe 1, environ 46 %) reste

globalement conservée dans chacun des sous ensembles, ce qui démontrent l'efficacité de l'argument de stratification lors du processus de partitionnement.

- Cette conservation des proportions entre les différentes partitions garantit que chaque sous-ensemble demeure représentatif de la population initiale.

3.9 Description des modèles utilisée :

Le tableau ci-dessous présente une synthèse des modèles de classification utilisés dans cette étude ainsi que leur rôle dans le système de prédiction des maladies cardiovasculaires.

modèle utilisée	Description	Rôle dans l'étude
Random Forest	Ensemble d'arbres de décision combinés par vote majoritaire	Améliore la précision et réduit le sur apprentissage
SVM	Algorithme basé sur la séparation optimale des classes par un hyperplan	Permet une classification efficace des patients
KNN	Algorithme basé sur la distance entre les observations	Classification par similarité des patients
RESEAUX DU NURONS	Modèle inspiré du cerveau humain composé de plusieurs couches	Capture les relations complexes dans les données

Tableau 3.1: Description des modèles

3.10 Paramétrage et optimisation (Hyperparameter tuning) :

L'hyperparamétrage des modèles, ou plus simplement le réglage des hyperparamètres, c'est le fait d'ajuster les petits leviers d'un algorithme de Machine Learning pour qu'il donne de meilleurs résultats. Ces fameux hyper paramètres, on les fixe avant même que le modèle commence à apprendre, et ils influencent pas mal son comportement. Par exemple, dans KNN, le nombre de voisins qu'on regarde, dans Random Forest, la profondeur des arbres, ou encore le paramètre C dans un SVM.

3.10.1 Paramétrage et optimisation du modèle K-Nearest Neighbors (KNN) :

L'optimisation du KNN consiste à tester plusieurs valeurs de *k*, différentes métriques de distance et méthodes de pondération. Des techniques telles que la validation croisée et le *Grid Search* sont souvent utilisées afin d'identifier automatiquement la meilleure combinaison de paramètres .

Hyperparametre	Description	Impact sur le modèle
Nombre de voisins (k)	Nombre d'observations les plus proches utilisées pour classer une nouvelle donnée.	Une faible valeur de k peut entraîner un surapprentissage, tandis qu'une valeur élevée peut provoquer un sous-apprentissage et réduire la précision.
Métrique de distance	Méthode utilisée pour mesurer la proximité entre les observations (euclidienne, Manhattan, Minkowski, etc.).	Influence directement la qualité de la classification et la capacité du modèle à distinguer les différentes classes.
Pondération des voisins (weights)	Définit l'importance accordée à chaque voisin lors de la prise de décision.	Une pondération basée sur la distance donne davantage d'influence aux voisins les plus proches et peut améliorer les performances.
Paramètre p (Minkowski)	Paramètre contrôlant le type de distance utilisé dans la métrique de Minkowski.	Modifie la façon dont les distances sont calculées et peut influencer la précision de la classification.
Algorithme de recherche des voisins	Méthode utilisée pour identifier les voisins les plus proches (Brute Force, KD-Tree, Ball Tree).	Affecte principalement le temps d'exécution et l'efficacité du modèle sur les grands ensembles de données.

Tableau 3.2: Hyperparamètres du modèle K-Nearest Neighbors (KNN)

```
Fitting 5 folds for each of 28 candidates, totalling 140 fits

============================================================
✅ Meilleure combinaison trouvée : {'knn__metric': 'manhattan', 'knn__n_neighbors': 21, 'knn__weights': 'distance'}
============================================================
```

Figure 3.11 : Résultat de l'optimisation des hyperparamètres du modèle KNN par Grid Search avec validation croisée

3.10.2 Paramétrage et optimisation du modèle Support Vector Machine (SVM) :

L'optimisation du SVM repose principalement sur la recherche des meilleures valeurs de C, gamma et du type de noyau. Cette recherche est souvent réalisée à l'aide des méthodes *Grid Search* et *Cross Validation* afin d'obtenir le modèle le plus performant

Hyperparamètre	Description	Impact sur le modèle
Paramètre C	Contrôle le compromis entre la minimisation des erreurs et la généralisation.	Une valeur élevée peut entraîner un surapprentissage, tandis qu'une faible valeur améliore la généralisation.
Kernel	Fonction permettant de projeter les données dans un espace de dimension supérieure.	Détermine la forme de la frontière de décision.
Gamma (γ)	Définit l'influence des exemples d'apprentissage sur la frontière de décision.	Une valeur élevée produit des frontières complexes, tandis qu'une faible valeur génère des frontières plus lisses.
Degree	Paramètre utilisé avec le noyau polynomial.	Augmente la complexité du modèle lorsque sa valeur est élevée.
Coef 0	Paramètre utilisé avec les noyaux polynomial et sigmoid.	Influence la forme de la frontière de décision

Tableau 3.3 : Hyperparamètres du modèle Support Vector Machine (SVM)

```
Fitting 5 folds for each of 56 candidates, totalling 280 fits

================================================
Meilleure combinaison : {'svm__C': 1, 'svm__gamma': 0.05}
================================================
```

Figure 3.12 : Résultat de l'optimisation des hyperparamètres du modèle SVM par Grid Search avec validation croisée

3.10.3 Paramétrage et optimisation du modèle Random Forest :

L'optimisation du Random Forest consiste à ajuster simultanément plusieurs paramètres afin d'obtenir le meilleur équilibre entre précision et temps de calcul. Les méthodes de validation croisée sont largement utilisées pour évaluer les différentes configurations.

Hyperparamètre	Description	Impact sur le modèle
Nombre d'arbres (n_estimators)	Nombre total d'arbres de décision dans la forêt.	Un nombre élevé améliore généralement la stabilité et la précision du modèle.
Profondeur maximale (max_depth)	Limite la profondeur des arbres de décision.	Une profondeur excessive peut provoquer un surapprentissage.
min_samples_split	Nombre minimal d'échantillons requis pour diviser un nœud.	Réduit la complexité des arbres lorsqu'il est élevé.
min_samples_leaf	Nombre minimal d'échantillons présents dans une feuille terminale.	Améliore la généralisation du modèle et réduit le surapprentissage.
max_features	Nombre de variables sélectionnées aléatoirement à chaque division.	Augmente la diversité des arbres et améliore la robustesse du modèle.
Bootstrap	Utilisation ou non de l'échantillonnage avec remise.	Favorise la diversité des arbres et améliore les performances globales.

Tableau 3.4 : Hyperparamètres du modèle Random Forest

```
Fitting 5 folds for each of 288 candidates, totalling 1440 fits

======================================================
✅ Meilleure combinaison trouvée pour Random Forest :
 {'rf__criterion': 'gini', 'rf__max_depth': None, 'rf__min_samples_leaf': 4, 'rf__min_samples_split': 10, 'rf__n_estimators': 300}
======================================================
```

Figure 3.13 : Résultat de l'optimisation des hyperparamètres du modèle Random forest par Grid Search avec validation croisée

3.10.4 Paramétrage et optimisation des Réseaux de Neurones :

L'optimisation des réseaux de neurones consiste à ajuster simultanément plusieurs hyperparamètres, tels que le nombre de couches cachées, le nombre de neurones par couche, le taux d'apprentissage (*Learning Rate*), la taille des lots (*Batch Size*) et le nombre d'époques (*Epochs*). L'objectif est d'obtenir un modèle capable d'apprendre efficacement les relations présentes dans les données tout en évitant le surapprentissage. Afin d'identifier la configuration optimale, différentes combinaisons d'hyperparamètres sont évaluées à l'aide de techniques telles que la validation croisée (*Cross-Validation*) et la recherche sur grille (*Grid Search*). Ces méthodes permettent d'améliorer les performances du modèle et sa capacité de généralisation sur de nouvelles données.

```
Fitting 5 folds for each of 48 candidates, totalling 240 fits

======================================================
✅ Meilleure combinaison trouvée pour MLP : {'mlp__activation': 'relu', 'mlp__alpha': 0.01, 'mlp__hidden_layer_sizes': (100,), 'mlp__solver': 'adam'}
======================================================
```

Figure 3.14: Résultat de l'optimisation des hyperparamètres du modèle Réseaux de Neurones par Grid Search avec validation croisée

Hyperparamètre	Description	Impact sur le modèle
hidden_layer_sizes	Définit le nombre de couches cachées ainsi que le nombre de neurones dans chaque couche.	Influence directement la capacité du réseau à apprendre des relations complexes entre les données.
Activation	Fonction d'activation utilisée dans les couches cachées (ReLU, Tanh, Logistic)	Détermine la capacité du réseau à modéliser des relations non linéaires.
Solver	Algorithme d'optimisation utilisé pour ajuster les poids du réseau (Adam, SGD, LBFGS).	Affecte la vitesse de convergence et les performances globales du modèle
learning_rate_init	Valeur initiale du taux d'apprentissage utilisée lors de l'entraînement.	Contrôle l'amplitude des mises à jour des poids et la stabilité de l'apprentissage
learning_rate	Stratégie d'adaptation du taux d'apprentissage (constant, adaptive, invscaling).	Influence la vitesse de convergence et la qualité de l'optimisation.
Alpha	Paramètre de régularisation L2 appliqué aux poids du réseau.	Réduit le risque de surapprentissage en limitant la complexité du modèle.
batch_size	Nombre d'exemples utilisés avant chaque mise à jour des poids.	Affecte la vitesse d'entraînement et la stabilité de l'optimisation.
max_iter	Nombre maximal d'itérations ou d'époques d'apprentissage.	Un nombre insuffisant peut empêcher la convergence du modèle.

Tableau 3.5 : Hyperparamètres des Réseaux de Neurones Artificiels

3.10.5 Techniques et méthodes d'optimisation des modèles d'apprentissage automatique

L'optimisation des modèles d'apprentissage automatique représente une étape essentielle visant à améliorer leurs performances et leur capacité de généralisation. Elle consiste principalement à ajuster les hyperparamètres des modèles afin de déterminer la configuration la plus adaptée aux données étudiées. Plusieurs techniques peuvent être utilisées dans ce processus, notamment la recherche sur grille (*Grid Search*), la validation croisée (*Cross-Validation*) et l'arrêt anticipé (*Early Stopping*). Ces méthodes permettent d'évaluer différentes configurations, de sélectionner les meilleurs paramètres et de limiter le phénomène de surapprentissage (*Overfitting*), garantissant ainsi des modèles plus fiables et plus performants.

La **Grid Search** (ou recherche par grille) est une technique d'optimisation des hyperparamètres utilisée en apprentissage automatique (*Machine Learning*) permettant d'identifier la meilleure combinaison de paramètres d'un modèle afin d'améliorer ses performances. Cette méthode consiste à définir un ensemble de valeurs possibles pour chaque hyperparamètre, puis à évaluer systématiquement toutes les combinaisons générées. Pour chaque configuration, le modèle est entraîné et évalué à l'aide d'une métrique de performance telle que l'accuracy, la précision, le rappel ou le score F1. La combinaison produisant les meilleurs résultats est ensuite sélectionnée comme configuration optimale du modèle.

Cependant, lorsque le nombre d'hyperparamètres ou de valeurs testées devient important, la recherche par grille peut nécessiter un temps de calcul élevé. Afin d'obtenir une évaluation plus fiable et de limiter le risque de surapprentissage (*overfitting*), elle est généralement associée à une méthode de validation croisée (*Cross-Validation*).[86]

Cette image(la figure 3.15) représente le principe de **Grid Search (Grid Search CV)** utilisé pour l'**optimisation des hyperparamètres** des modèles de Machine Learning.

Dans cette représentation :

- **Axe horizontal (Hyperparameter 2)** : représente les différentes valeurs possibles d'un deuxième hyperparamètre du modèle (par exemple : nombre d'arbres dans Random Forest, valeur de C dans SVM, nombre de voisins K dans KNN…).

- **Axe vertical (Hyperparameter 1)** : représente les différentes valeurs possibles d'un premier hyperparamètre.
- **Chaque cercle bleu** correspond à une **combinaison unique de paramètres testée** par l'algorithme. Par exemple :
    - Hyperparamètre 1 = valeur A
    - Hyperparamètre 2 = valeur B

→ le modèle est entraîné et évalué avec cette combinaison.

Le fonctionnement du **Grid Search** consiste à tester **toutes les combinaisons possibles** des hyperparamètres dans une grille définie à l'avance. Les flèches dans l'image illustrent le parcours systématique de toutes les cases de la grille.

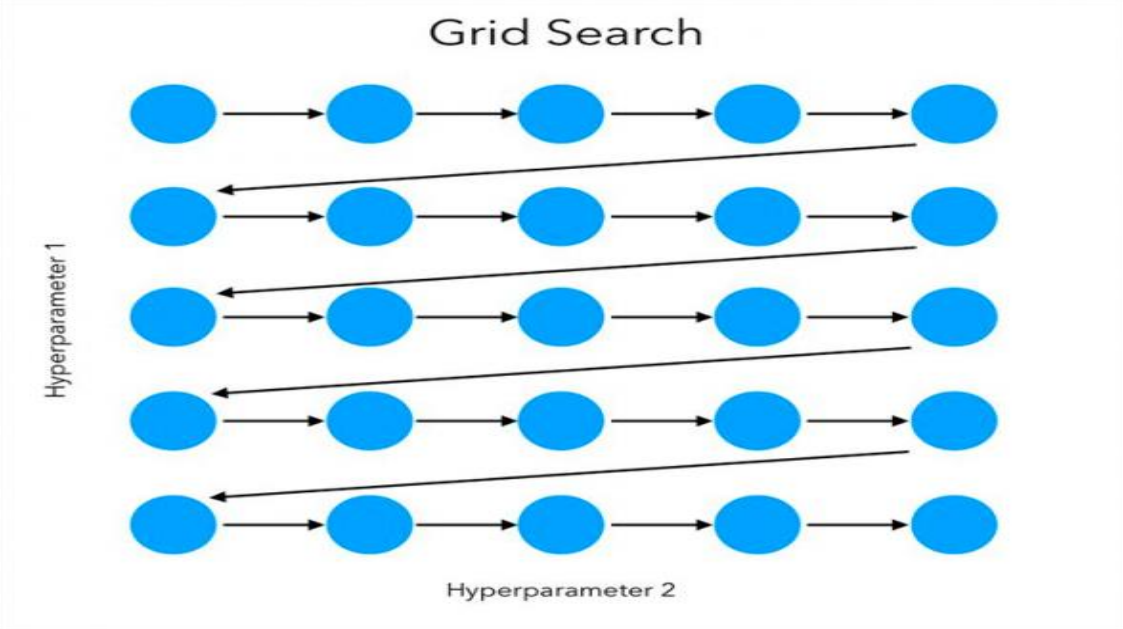


Figure 3.15 : la technique GridSearch[87]

3.11 La validation croisée(Cross validation) :

La **validation croisée** (*Cross-Validation*) est une méthode d'évaluation utilisée en apprentissage automatique (*Machine Learning*) permettant d'estimer de manière plus fiable la capacité de généralisation d'un modèle sur de nouvelles données. Elle consiste à diviser l'ensemble des données disponibles en plusieurs sous-ensembles appelés *folds*. Le modèle est ensuite entraîné sur une partie des sous-ensembles et évalué sur le sous-ensemble

restant. Cette opération est répétée plusieurs fois en changeant à chaque itération la partie utilisée pour le test.

La méthode la plus couramment utilisée est la **validation croisée K-Fold**, où les données sont réparties en *K* groupes de taille approximativement égale. Lors de chaque itération, un groupe est utilisé pour la validation tandis que les *K-1* groupes restants servent à l'apprentissage. Les performances obtenues durant chaque itération sont ensuite moyennées afin d'obtenir une estimation globale et plus robuste des performances du modèle.

Cette technique permet de réduire la dépendance du modèle à une seule séparation des données et contribue à limiter le risque de **surapprentissage (*overfitting*)**, en fournissant une meilleure estimation de la capacité du modèle à généraliser sur des données non observées[88].

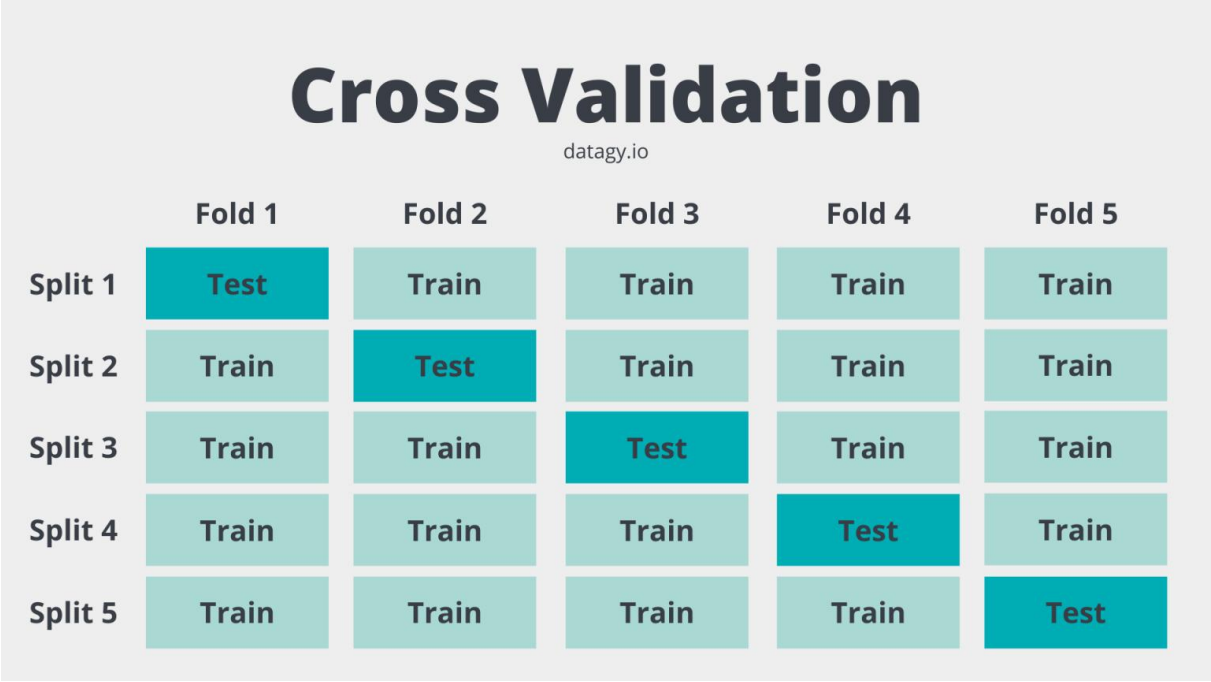


Figure 3.16 : validation croisée

3.12 Protocole expérimental commun:

Dans notre travail, nous avons adopté un protocole expérimental structuré pour l'analyse et la prédiction des maladies cardiovasculaires à partir d'une base de données dédiée Clevland. La démarche commence par le chargement et le nettoyage des données, suivi par la gestion des valeurs manquantes, l'encodage des variables catégorielles, la sélection des caractéristiques pertinentes et la normalisation des données. Une réduction de dimension par PCA peut également être appliquée afin d'optimiser la représentation des données. Ensuite, la base de données est divisée en ensembles d'apprentissage, de validation et de test afin d'entraîner et d'évaluer les modèles. Plusieurs algorithmes de Machine Learning sont étudiés, notamment SVM, Random Forest, KNN et les réseaux de neurones. Enfin, une optimisation des hyperparamètres à l'aide de Grid Search ainsi qu'une validation croisée (*Cross-Validation*) sont utilisées pour améliorer les performances des modèles et garantir une évaluation fiable.

3.13 Critères de comparaison des modèles :

- **Matrice de confusion**

La matrice de confusion, également appelée matrice d'erreur, est un outil d'évaluation des performances d'un modèle de classification. Elle est représentée sous la forme d'une matrice de taille $N \times N$, où N correspond au nombre de classes du problème.

Elle permet de comparer les valeurs réelles (ground truth) avec les valeurs prédites par le modèle, offrant ainsi une vision globale de ses performances ainsi que des différents types d'erreurs commises.

Dans un problème de classification binaire, la matrice de confusion est composée de quatre éléments principaux :

a) **Vrais positifs (True Positives, TP)** :

Ce sont les cas où le modèle prédit correctement la classe positive. Autrement dit, la valeur réelle est positive et la prédiction est également positive.

b) **Vrais négatifs (True Negatives, TN)** :

Ce sont les cas où le modèle prédit correctement la classe négative. La valeur réelle est négative et la prédiction est également négative.

c) **Faux positifs (False Positives, FP)** :

Ils correspondent à une erreur de type I. Le modèle prédit une classe positive alors que la valeur réelle est négative.

d) **Faux négatifs (False Negatives, FN)** :

Ils correspondent à une erreur de type II. Le modèle prédit une classe négative alors que la valeur réelle est positive.

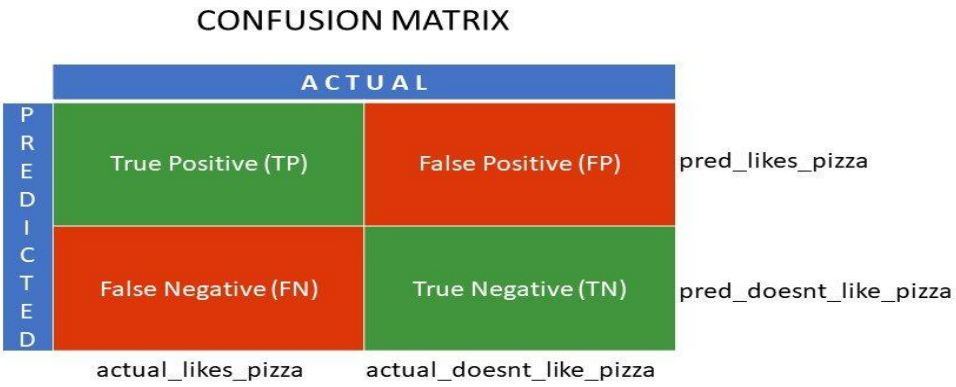


Figure 3.17 : Exemple de matrice confusion [89]

- La **courbe ROC (Receiver Operating Characteristic)** est un graphique utilisé pour **évaluer les performances d'un modèle de classification binaire** . Elle montre la capacité du modèle à **séparer correctement les deux classes** en fonction de différents seuils de décision.

Elle représente :

Axe X : False Positive Rate (FPR) = Taux de faux positifs

→ proportion des personnes classées positives par erreur parmi les personnes réellement négatives.

Axe Y : True Positive Rate (TPR) = Sensibilité / Rappel

→ proportion des vrais positifs correctement détectés par le modèle.

Pour comprendre la signification de ce graphique, il convient d'analyser sa trajectoire selon des règles d'interprétation précises :

a) **Regarder la position de la courbe**

Une bonne courbe ROC se rapproche du **coin supérieur gauche** du graphique, ça veux dire que le modèle détecte beaucoup de vrais positifs avec peu de faux positifs.

Une mauvaise courbe reste proche de la **diagonale, ça veux dire que l**e modèle est presque équivalent à un choix aléatoire.

b) **Analyser l'AUC (Area Under the Curve)**

L'AUC représente la surface sous la courbe ROC. Elle donne une mesure globale de performance :

- **AUC = 1** → modèle parfait (séparation totale entre les classes)
- **AUC ≈ 0,9 – 1** → excellent modèle
- **AUC ≈ 0,8 – 0,9** → bon modèle
- **AUC ≈ 0,7 – 0,8** → modèle acceptable
- **AUC = 0,5** → modèle aléatoire (aucune capacité de discrimination)
- **AUC < 0,5** → modèle moins performant qu'un hasard

C) **Comparer plusieurs modèles**

Si plusieurs courbes ROC sont présentes :

La courbe qui est **plus proche du coin supérieur gauche** est généralement la meilleure.

Le modèle avec la **plus grande AUC** possède la meilleure capacité de classification [90].

Ces mesures permettent d'évaluer l'efficacité des modèles, leur capacité à détecter correctement les anomalies cardiaques ainsi que leur niveau de fiabilité.

Les performances obtenues par chaque algorithme seront présentées et analysées dans les sections suivantes afin de déterminer le modèle le plus adapté à la prédiction des anomalies cardiaques.

3.14 Analyse résultats de chaque modèle :

3.14.1 SVM :

Sur les 61 patients du groupe de test (représentant 20 % de la population totale de Cleveland), notre modèle SVM a correctement classifié 52 patients et a commis 9 erreurs comme le montre la matrice de confusion (figure 3.18) .

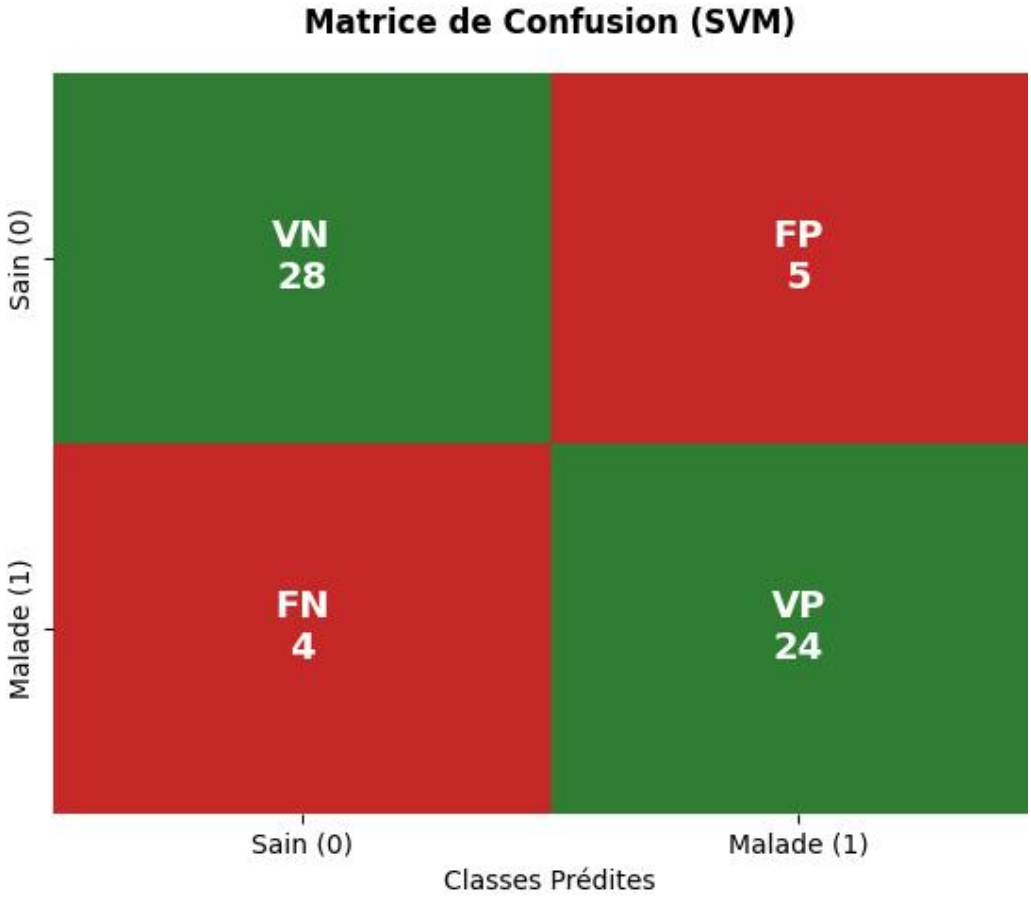


Figure 3.18: Matrice de confusion du modèle SVM illustrant la répartition des prédictions correctes et erronées

- Vrais Négatifs (VN) = 28 : Patients effectivement sains et correctement identifiés comme sains par le modèle.
- Faux Positifs (FP) = 5 : Patients sains incorrectement classés comme malades.
- Faux Négatifs (FN) = 4 : Patients malades incorrectement classifiés comme sains .
- Vrais Positifs (VP) = 24 : Patients effectivement malades et correctement identifiés comme tels.

Les informations fournies par la matrice de confusion sont complétées par des métriques quantitatives telles que l'exactitude, la précision, la sensibilité, la spécificité et le score F1, représentées dans la figure 3.19, afin d'obtenir une évaluation plus complète .

```
--- MÉTRIQUES D'ÉVALUATION ---
Taux de classification (Accuracy)     : 0.8525
Précision                             : 0.8276
Sensibilité (Recall)                  : 0.8571
F1-score                              : 0.8421
Spécificité                           : 0.8485
```

Figure 3.19 : Bilan chiffré des métriques d'évaluation du modèle SVM

Métrique	Valeur	Interprétation médicale
Accuracy (Exactitude)	0.8525 (85,25%)	Notre modèle prédit correctement 85% des cas (malades + sains).Il identifie correctement 52 patients malades sur 61 et fournit un diagnostic correct, ce qui est considérée comme acceptable pour un outil d'aide au dépistage cardiovasculaire.
Précision (Precision)	0.8276 (82,76%)	Parmi les patients prédits malades, 83% le sont réellement.Ainsi qu'un faible taux de faux positifs (5) sert à éviter des examens invasifs ou complémentaires inutiles d'où les faux alertes sont acceptables en contexte médical.
Sensibilité (Recall)	0.8571 (85,71%)	Notre modèle détecte 86% des vrais malades (24 sur 28) . Cette capacité de détection est essentielle en cardiologie, bien que quelques cas (4 faux négatifs) puissent encore échapper au modèle.
F1-score	0.8421 (84,21%)	Équilibre acceptable entre précision et sensibilité, car on veut **détecter les vrais malades** (Recall) et en même temps **éviter les fausses alertes** (Precision) .
Spécificité	0.8485 (84,85%)	Notre modèle présente une spécificité de 85 %, ce qui signifie qu'il identifie correctement la majorité des patients sains (28 sur 33). Cette capacité d'exclusion permet de limiter les fausses alertes .

Tableau 3.6: Métriques de performance et interprétations cliniques du modèle SVM sur l'échantillon de test.

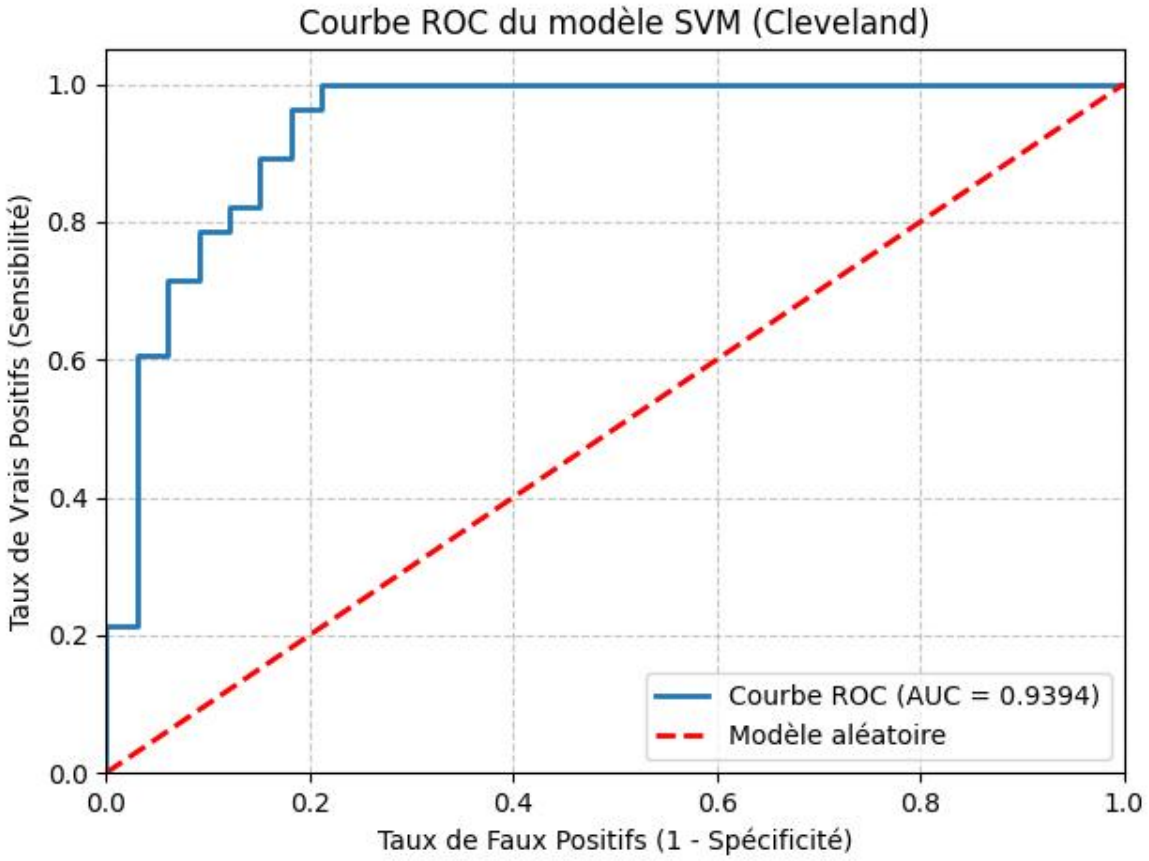


Figure 3.20 : Courbe ROC et mesure de l'aire sous la courbe (AUC) pour l'architecture SVM

Dans le cadre de l'évaluation de ce modèle et pour garantir une validation rigoureuse des résultats obtenus, l'analyse de la courbe ROC (figure 3.20) révèle une aire sous la courbe (AUC) de 0.9394, ce qui démontre une excellente capacité du classifieur SVM à discriminer les patients sains des patients atteints d'une pathologie cardiaque. Visuellement, la trajectoire de la courbe bleue s'élève rapidement vers le coin supérieur gauche, ce qui témoigne d'un compromis optimal : le modèle maximise le taux de vrais positifs (sensibilité) tout en maintenant un très faible taux de faux positifs. De plus, l'éloignement marqué de cette courbe par rapport à la diagonale rouge (représentant un modèle aléatoire à l'AUC de 0.5) confirme la robustesse statistique et la haute valeur prédictive de notre approche sur le jeu de données de Cleveland.

3.14.2 KNN :

Sur les 61 patients du groupe de test (représentant 20 % de la population totale de Cleveland), notre modèle KNN a correctement classifié 54 patients et a commis 7 erreurs comme le montre la matrice de confusion (figure 3.21) .

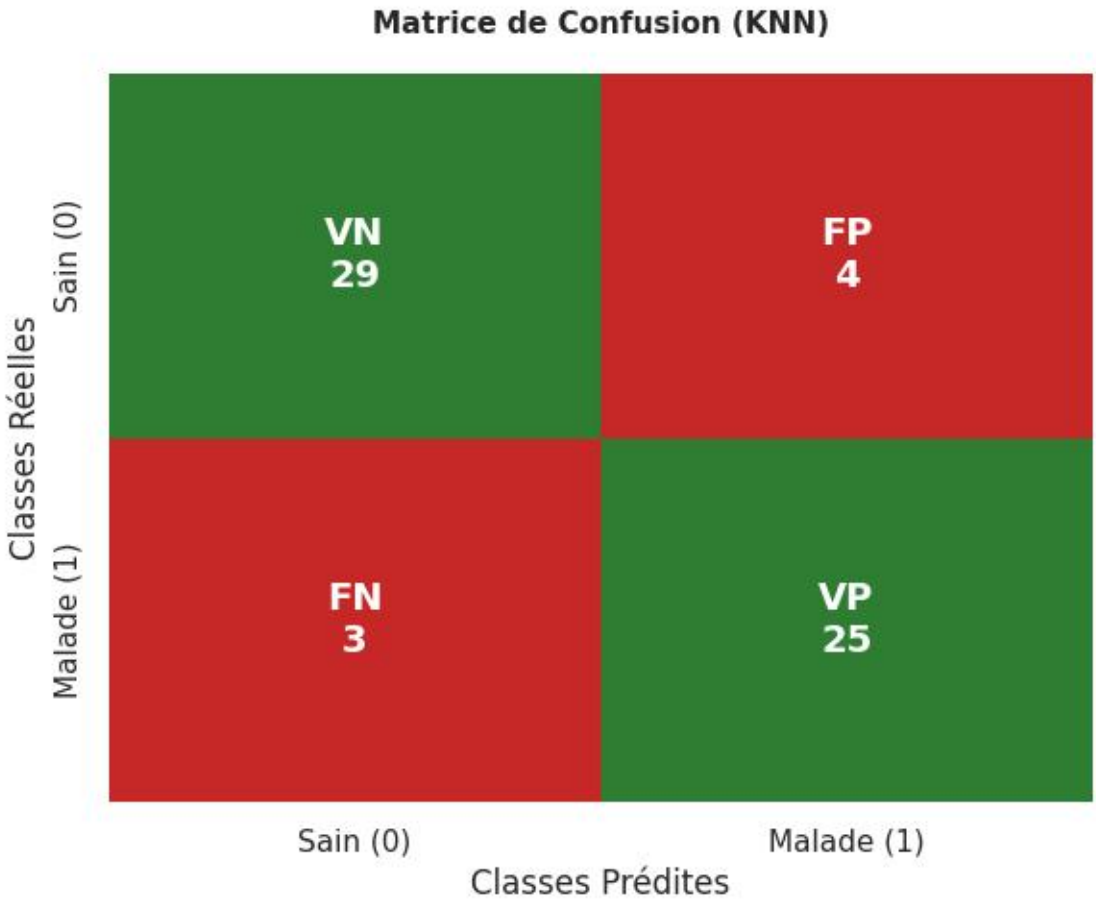


Figure 3.21: Matrice de confusion du modèle KNN illustrant la répartition des prédictions correctes et erronées

- Vrais Négatifs (VN) = 29 : Patients effectivement sains et correctement identifiés comme sains par le modèle.
- Faux Positifs (FP) = 4 : Patients sains incorrectement classés comme malades.
- Faux Négatifs (FN) = 3 : Patients malades incorrectement classifiés comme sains.
- Vrais Positifs (VP) = 25 : Patients effectivement malades et correctement identifiés comme tels.

Les informations fournies par la matrice de confusion sont complétées par des métriques quantitatives telles que l'exactitude, la précision, la sensibilité, la spécificité et le score F1, représentées dans la figure 3.22, afin d'obtenir une évaluation plus complète .

```
--- MÉTRIQUES D'ÉVALUATION ---
Taux de classification (Accuracy) : 0.8852
Précision                         : 0.8621
Sensibilité (Recall)              : 0.8929
F1-Score                          : 0.8772
Spécificité                       : 0.8788
```

Figure 3.22: Bilan chiffré des métriques d'évaluation du modèle KNN

Métrique	Valeur	Interprétation médicale
Accuracy (Exactitude)	0,8852 (88,52%)	Ce modèle atteint une exactitude globale de 88,5 % (54 patients correctement diagnostiqués sur 61).Cette performance, supérieure à celle du SVM, est considérée comme bonne pour un outil de dépistage cardiovasculaire.
Précision (Precision)	0,8621 (86,21%)	Le modèle KNN présente une précision de 86 %, ce qui signifie que la majorité des patients prédits malades le sont réellement (25 sur 29). Le faible nombre de faux positifs (4) limite les fausses alertes et réduit les examens complémentaires inutiles, rendant les prédictions du modèle presque fiables.
Sensibilité (Recall)	0,8929 (89,29%)	89 % traduit que notre modèle détecte correctement la majorité des patients réellement malades (25 sur 28).Bien que ces performances soient robustes, cela implique néanmoins que trois patients pathologiques (faux négatifs) échappent au diagnostic du classifieur.
F1-score	0,8772 (87,72%)	Cela indique que le modèle détecte efficacement les patients malades tout en limitant les fausses alertes, et c'est un très bon compromis pour le diagnostic cardiovasculaire.
Spécificité	0,8788 (87,88%)	La majorité des patients sains sont identifie correctement (29 sur 33) , ce score limite le risque d'engorger les hôpitaux ou de stresser les patients avec de fausses alertes .

Tableau 3.7: Métriques de performance et interprétations cliniques du modèle KNN sur l'échantillon de test.

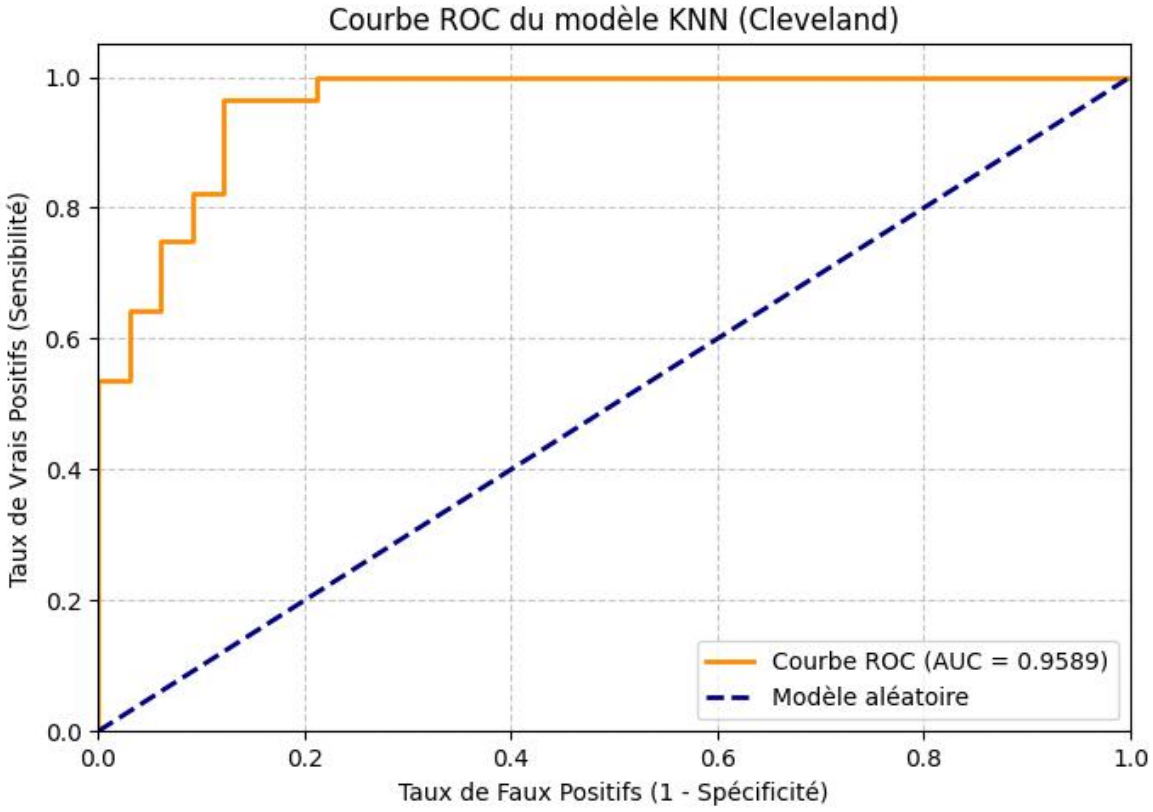


Figure 3.23: Courbe ROC et mesure de l'aire sous la courbe (AUC) pour l'architecture KNN

Au-delà des premiers indicateurs de performance, l'analyse de la courbe ROC (figure 3.23) permet d'évaluer la fiabilité globale du modèle à différents seuils. Avec une aire sous la courbe (AUC) s'élevant à 0.9589, ce graphique met en évidence l'excellente capacité globale du classifieur KNN à discriminer les patients sains des patients atteints d'une pathologie cardiaque.

Visuellement, la trajectoire de la courbe orange s'élève de manière abrupte vers le coin supérieur gauche, ce qui témoigne d'un compromis hautement performant : le modèle maximise rapidement le taux de vrais positifs (sensibilité) tout en maintenant un taux de faux positifs minimal. De plus, l'éloignement prononcé de cette courbe par rapport à la diagonale pointillée bleue (représentant un classifieur aléatoire à l'AUC de 0.5) souligne la robustesse statistique et le fort potentiel d'aide au diagnostic médical de cette approche KNN sur les données encodées de Cleveland.

3.14.3 Random forest :

Sur les 61 patients du groupe de test (représentant 20 % de la population totale de Cleveland), notre modèle **Random Forest** a correctement classifié **56 patients** et a commis **5 erreurs**, comme le montre la matrice de confusion présentée à la(figure 3.24).

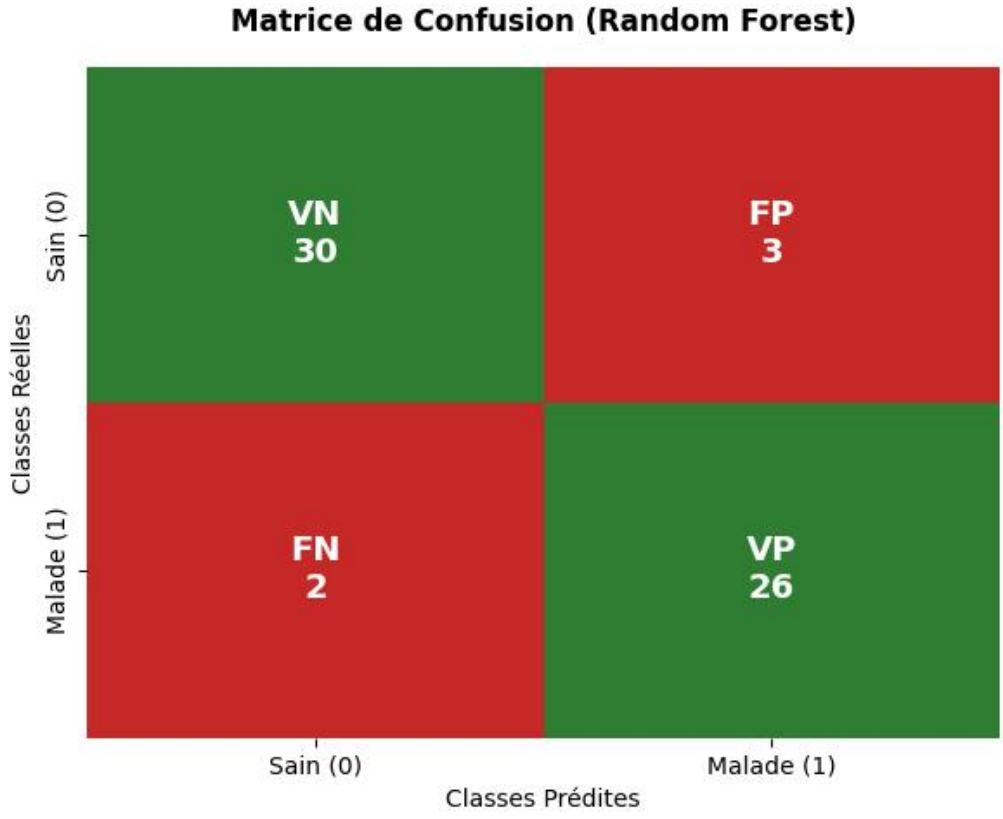


Figure 3.24 : Matrice de confusion du modèle Random forest illustrant la répartition des prédictions correctes et erronées

- Vrais Négatifs (VN) = 30 : Patients effectivement sains et correctement identifiés comme sains par le modèle.
- Faux Positifs (FP) = 3 : Patients sains incorrectement classés comme malades.
- Faux Négatifs (FN) = 2 : Patients malades incorrectement classifiés comme sains.
- Vrais Positifs (VP) = 26 : Patients effectivement malades et correctement identifiés comme tels.

Les informations fournies par la matrice de confusion sont complétées par des du modèle métriques quantitatives telles que l'exactitude, la précision, la sensibilité, la spécificité et le score F1, représentées dans la figure 3.25, afin d'obtenir une évaluation plus complète des performances.

```
--- MÉTRIQUES D'ÉVALUATION ---
Taux de classification (Accuracy) : 0.9180
Précision                         : 0.8966
Sensibilité (Recall)              : 0.9286
F1-Score                          : 0.9123
Spécificité                       : 0.9091
```

Figure 3.25: Bilan chiffré des métriques d'évaluation du modèle Random forestsur l'échantillon de test

Métrique	Valeur	Interprétation médicale
Accuracy (Exactitude)	0,9180 (91,80 %)	Notre modèle prédit correctement **91,80 %** des cas, qu'il s'agisse de patients malades ou sains. Cette performance élevée montre sa capacité à fournir un diagnostic fiable et à distinguer efficacement les individus atteints d'une maladie cardiovasculaire des sujets sains.
Précision (Precision)	0,8966 (89,66 %)	Parmi les patients prédits comme malades, **89,66 %** le sont réellement. Cette valeur élevée traduit un faible nombre de faux positifs, limitant ainsi les examens complémentaires inutiles et les inquiétudes injustifiées chez les patients sains.
Sensibilité (Recall)	0,9286 (92,86 %)	Le modèle identifie correctement **92,86 %** des patients réellement atteints d'une maladie cardiovasculaire. Cette capacité de détection est particulièrement importante dans le domaine médical, car elle réduit le risque de passer à côté d'un patient malade.
F1-score	0,9123 (91,23 %)	Le F1-score élevé reflète un excellent équilibre entre la précision et la sensibilité. Il confirme que le modèle est capable de détecter efficacement les patients malades tout en maintenant un faible taux de fausses alertes.
Spécificité	0,9091 (90,91 %)	Le modèle identifie correctement **90,91 %** des patients sains. Cette valeur indique une bonne capacité à éviter les faux diagnostics positifs et à distinguer les individus non atteints de la maladie

Tableau 3.8: Métriques de performance et interprétations cliniques du modèle Random Forest sur l'échantillon de test

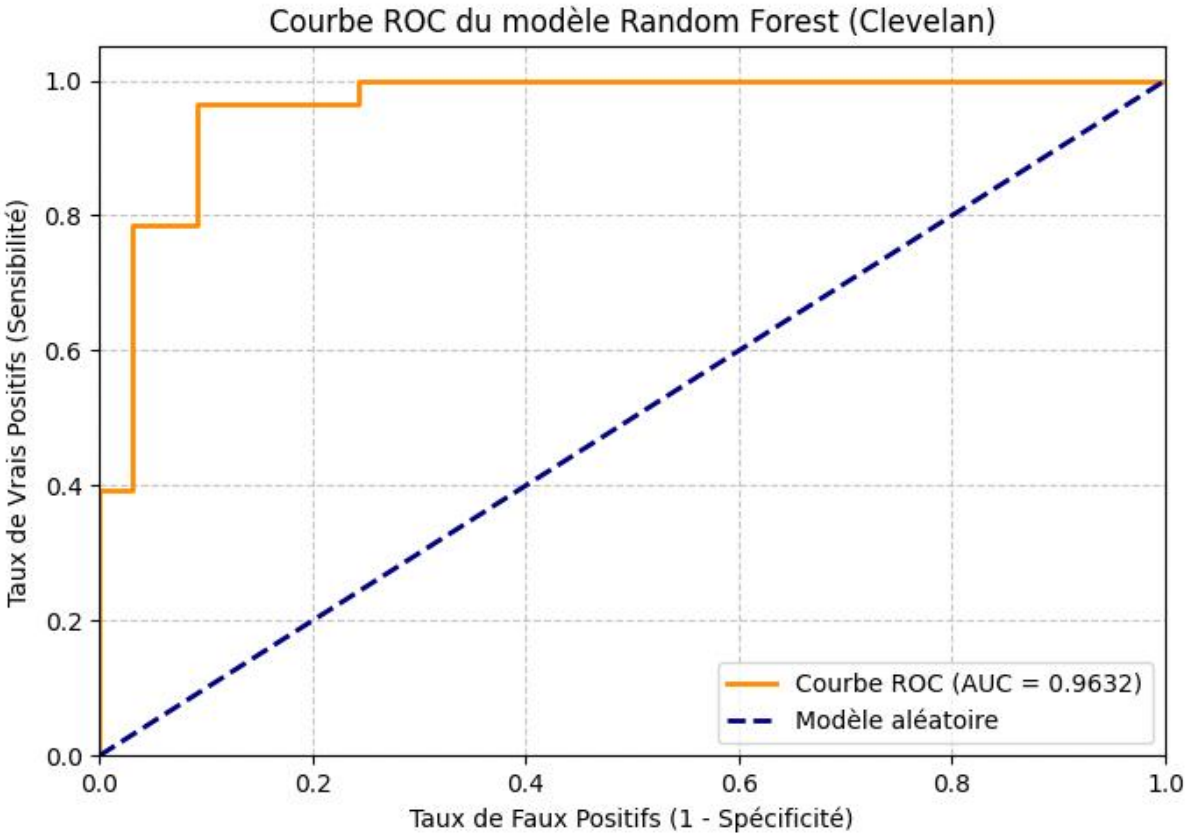


Figure 3.26: Courbe ROC et mesure de l'aire sous la courbe (AUC) pour l'architecture Random forest

La Figure 3.26 présente la courbe ROC obtenue à partir du modèle Random Forest. L'évaluation de cette courbe révèle une aire sous la courbe (AUC) égale à **0,9632**, ce qui témoigne d'une très forte capacité du modèle à distinguer les patients présentant une anomalie cardiaque de ceux ne présentant aucun signe de maladie cardiovasculaire.

L'allure de la courbe montre une nette proximité avec le coin supérieur gauche du graphique, traduisant une performance élevée en matière de classification. Cette configuration indique que le modèle atteint simultanément un taux élevé de détection des cas positifs et un faible taux de faux positifs. En comparaison avec la diagonale de référence correspondant à une classification aléatoire, la courbe ROC obtenue se situe significativement au-dessus, démontrant ainsi la pertinence du modèle développé.

Ces résultats mettent en évidence l'efficacité de l'algorithme Random Forest dans l'exploitation des données médicales de la base Cleveland. La valeur élevée de l'AUC confirme non seulement la qualité de la discrimination entre les différentes classes, mais également la stabilité et la capacité de généralisation du modèle.

3.14.4 Les réseaux du nuerons :

Sur les 61 patients du groupe de test (représentant 20 % de la population totale de Cleveland), notre modèle MLP a correctement classifié 50 patients et a commis 11 erreurs, comme le montre la matrice de confusion (figure 3.27)

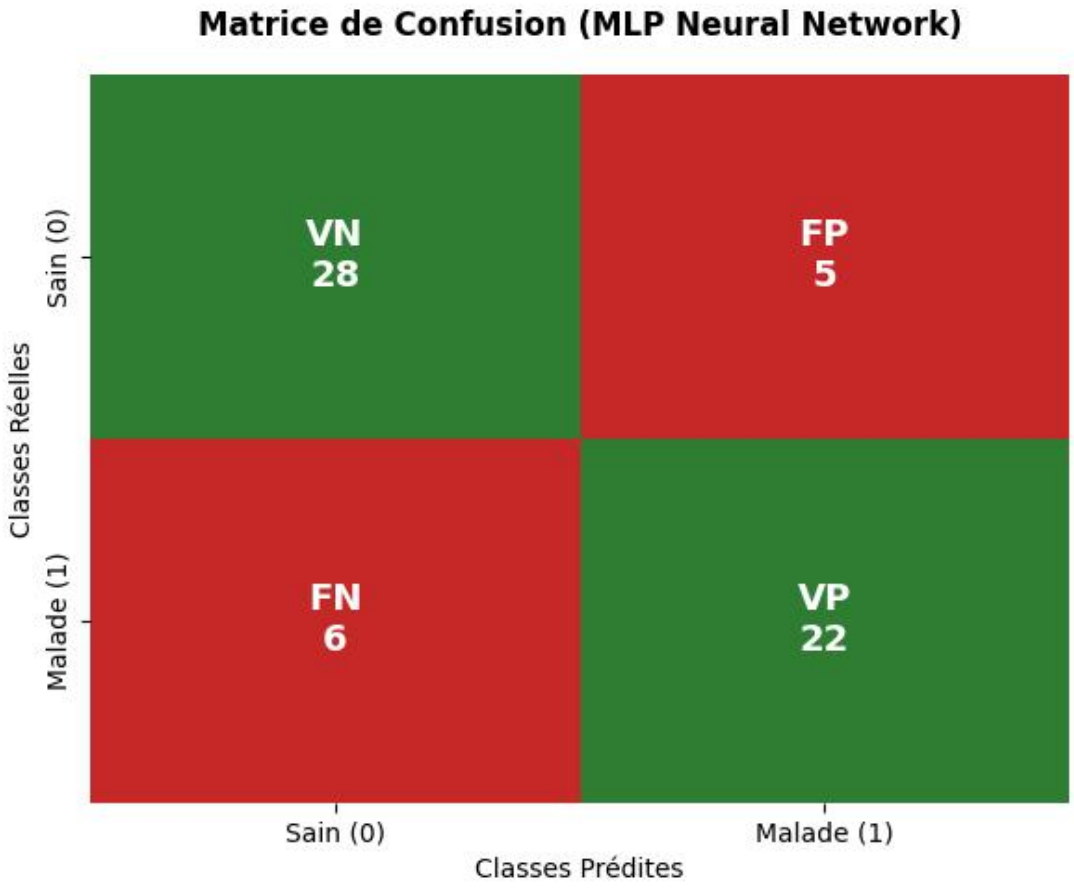


Figure 3.27: Matrice de confusion du modèle MLP illustrant la répartition des prédictions correctes et erronées

- Vrais Négatifs (VN) = 28 : Patients effectivement sains et correctement identifiés comme sains par le modèle.
- Faux Positifs (FP) = 5 : Patients sains incorrectement classés comme malades.
- Faux Négatifs (FN) = 6 : Patients malades incorrectement classifiés comme sains.
- Vrais Positifs (VP) = 22 : Patients effectivement malades et correctement identifiés comme tels.

Les informations fournies par la matrice de confusion sont complétées par des du modèle métriques quantitatives telles que l'exactitude, la précision, la sensibilité, la spécificité et le score F1, représentées dans la figure 3.28, afin d'obtenir une évaluation plus complète des performances.

```
--- MÉTRIQUES D'ÉVALUATION ---
Taux de classification (Accuracy) : 0.8197
Précision                         : 0.8148
Sensibilité (Recall)              : 0.7857
F1-Score                          : 0.8000
Spécificité                       : 0.8485
```

Figure 3.28: Bilan chiffré des métriques d'évaluation du modèle MLP l'échantillon de test

Métrique	Valeur	Interprétation médicale
Accuracy (Exactitude)	0,8197 (81,97 %)	Le modèle classe correctement **81,97 %** des patients, qu'ils soient malades ou sains. Cette performance indique une bonne capacité globale de classification pour la prédiction des maladies cardiovasculaires.
Précision (Precision)	0,8148 (81,48 %)	Parmi les patients prédits comme malades, **81,48 %** sont effectivement atteints de la maladie. Cette valeur montre que le modèle présente un taux modéré de faux positifs et fournit des prédictions relativement fiables.
Sensibilité (Recall)	0,7857 (78,57 %)	Le modèle détecte correctement **78,57 %** des patients réellement malades. Bien que satisfaisante, cette valeur indique que certains cas de maladie peuvent encore ne pas être identifiés par le modèle.
F1-score	0,8000 (80,00 %)	Le F1-score reflète un bon équilibre entre la précision et la sensibilité. Il confirme que le modèle maintient des performances relativement homogènes dans l'identification des patients malades.
Spécificité	0,8485 (84,85 %)	Le modèle reconnaît correctement **84,85 %** des patients sains. Cette performance traduit une bonne capacité à limiter les faux positifs et à éviter de classer à tort des sujets sains comme malades.

Tableau 3.9 : Métriques de performance et interprétations cliniques du modèle MLP sur l'échantillon de test.

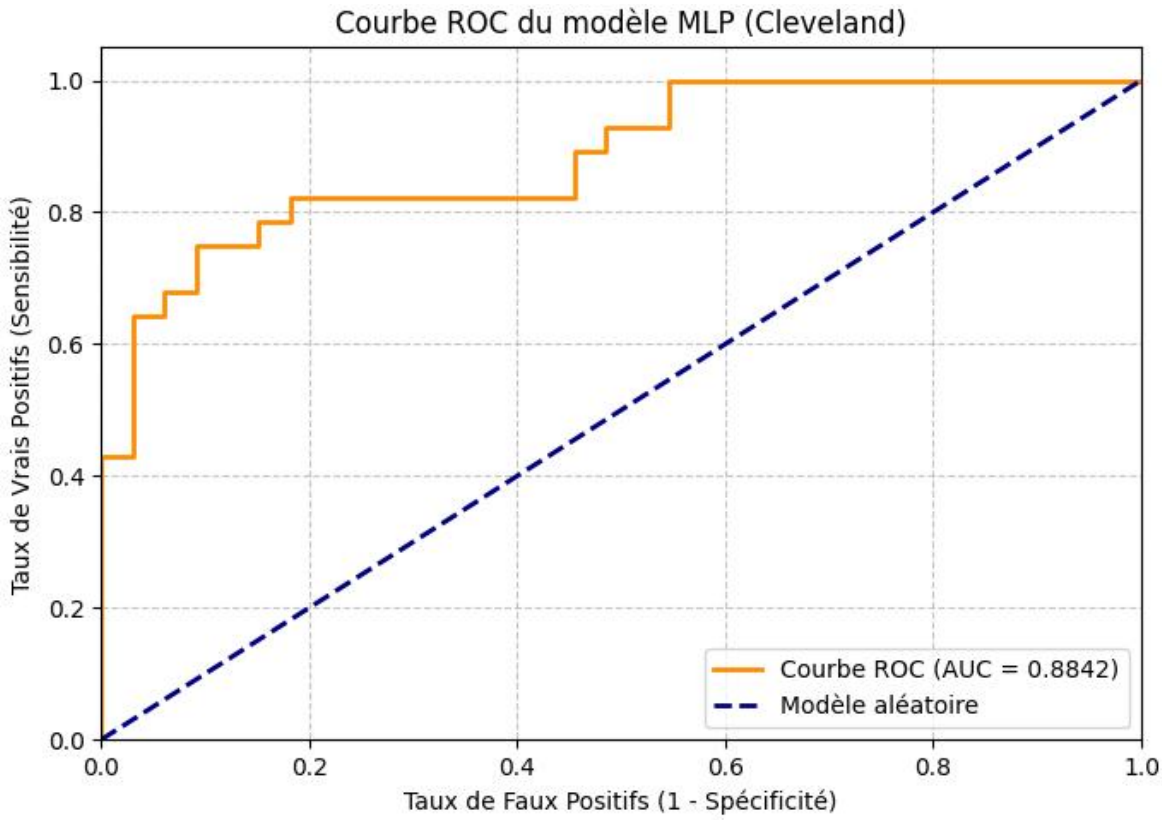


Figure 3.29: Courbe ROC et mesure de l'aire sous la courbe (AUC) pour l'architecture MLP

L'examen de la courbe ROC associée au modèle de réseau de neurones multicouche (MLP), illustrée à la Figure 3.29, met en évidence une aire sous la courbe (AUC) de **0,8842**. Cette valeur reflète une aptitude significative du modèle à différencier les individus atteints d'une maladie cardiovasculaire de ceux ne présentant aucune anomalie cardiaque. La courbe obtenue demeure largement au-dessus de la diagonale de référence caractéristique d'un classificateur aléatoire, attestant ainsi de la qualité prédictive du modèle. Les performances enregistrées traduisent une bonne capacité de généralisation et une classification fiable des observations. En outre, l'équilibre observé entre le taux de détection des cas positifs et la limitation des faux positifs renforce la pertinence du MLP dans un contexte d'aide au diagnostic cardiovasculaire. Ces résultats confirment l'intérêt de cette approche d'apprentissage profond pour la prédiction des maladies cardiovasculaires à partir des données cliniques de Cleveland.

3.15 Analyse comparative des performances :

Après avoir analysé individuellement le comportement de chaque classifieur, nous présentons ici une confrontation globale de leurs performances. À travers notre approche comparative, nous visons à mettre en lumière le modèle offrant le compromis diagnostique le plus robuste et le plus fiable pour la prédiction des anomalies cardiovasculaires.

Métrique	MLP	SVM	KNN	Random Forest
Accuracy (Exactitude)	0,8197 (81,97%)	0,8525 (85,25%)	0,8852 (88,52%)	0,9180 (91,80%)
Précision	0,8148 (81,48%)	0,8276 (82,76%)	0,8621 (86,21%)	0,8966 (89,66%)
Sensibilité (Recall)	0,7857 (78,57%)	0,8571 (85,71%)	0,8929 (89,29%)	0,9286 (92,86%)
F1-Score	0,8000 (80%)	0,8421 (84,21%)	0,8772 (87,72%)	0,9123 (91,23%)
Spécificité	0,8485 (84,85%)	0,8485 (84,85%)	0,8788 (87,88%)	0,9091 (90,91%)
AUC-ROC	0,8842 (88,42 %)	**0,9394 (93,94%)**	**0,9589 (95,89%)**	**0.9632** (96,32%)

Tableau 3.10: Synthèse comparative des performances des modèles d'apprentissage automatique

- **La capacité de détection clinique (Axe Sensibilité) :**

D'un point de vue médical, l'enjeu prioritaire de notre étude réside dans la capacité du système à identifier correctement les patients atteints d'anomalies cardiovasculaires afin d'éviter tout retard de prise en charge. Sous cet angle, l'analyse comparative révèle que le modèle Random Forest s'impose comme le plus performant avec une sensibilité remarquable de (92,86%), suivi par le KNN (89,29%) et le SVM (85,71%). En maximisant le taux de vrais positifs, la forêt aléatoire réduit l'occurrence des faux négatifs. Cette caractéristique est cruciale en environnement hospitalier, car elle sécurise le parcours de prise en charge des patients à haut risque en réduisant au strict minimum la probabilité qu'un malade échappe au diagnostic initial.
À l'inverse, le modèle MLP affiche la sensibilité la plus faible (82,14%), confirmant sa fragilité à repérer les cas pathologiques sur cet échantillon.

- **Contrôle des fausses alertes (Axe Spécificité) :**

Parallèlement à la détection des malades, la réduction des faux positifs constitue un indicateur clé pour prévenir l'anxiété inutile des patients et limiter les coûts liés à la prescription d'examens cardiologiques complémentaires et invasifs. L'évaluation de la spécificité confirme la même progression entre les classifieurs. Le modèle **Random Forest** démontre à nouveau la meilleure aptitude à exclure les sujets sains du protocole d'alerte avec un score de 90,91%. Les modèles KNN (87,88%) et SVM (84,85%) maintiennent un filtrage honorable, tandis que le réseau MLP s'établit à 81,82%. La supériorité de la forêt aléatoire garantit une allocation optimale des ressources médicales et une plus grande confiance des praticiens dans les alertes générées par notre outil intelligent.

- **Équilibre diagnostique et pouvoir discriminant (F1-Score et AUC-ROC) :**

Pour évaluer la fiabilité globale et la robustesse de **nos** approches, **nous avons analysé** conjointement le score F1 et l'aire sous la courbe ROC (AUC-ROC). le score F1 valide le compromis statistique direct entre la précision sensibilité. À cet égard, **notre** classifieur Random Forest se distingue par l'équilibre le plus solide **(91,23%)**, surpassant le **KNN (87,72%), et c**e dernier se montre à son tour plus performant que le **SVM (84,21%)** et le réseau **MLP (80,70%)**.De surcroît, l'examen du pouvoir discriminant à travers l'AUC-ROC confirme cette supériorité, dont la courbe ROC domine toutes les autres avec un score de **96,32%**. Mathématiquement, cela prouve qu'il existe une probabilité de près de

96,32% pour que l'algorithme attribue un score de risque plus élevé à un patient malade qu'à un patient sain. Enfin, le KNN (0,9589) et le SVM (0,9394) se révèlent également très discriminants, tandis que le MLP les superforme de manière notable avec une AUC de seulement 0,8119.

3.16 Sélection du modèle optimal et limites des approches concurrentes :

L'analyse comparative des performances désigne le **Random Forest** comme le modèle le plus adapté à la prédiction des maladies cardiovasculaires. Sur l'ensemble des métriques évaluées (Accuracy, précision, sensibilité, F1-score, spécificité et AUC-ROC), le Random Forest enregistre les scores les plus élevés, avec une avance particulièrement marquée sur la **sensibilité** (taux de vrais positifs) et l'**AUC-ROC** (pouvoir discriminant global).

Plusieurs caractéristiques **propres au modèle** expliquent cette supériorité :

- Supériorité absolue sur toutes les métriques **:** Le Random Forest surpasse tous les autres modèles avec une **Accuracy globale de 91,80 %** et un **F1-Score de 91,23 %**, ce qui prouve sa robustesse et son excellent équilibre général.

- Maximisation de la Sensibilité (Recall = 92,86 %) **:** C'est l'argument le plus crucial dans le domaine médical. Un Recall de 92,86 % signifie que le modèle minimise drastiquement les faux négatifs. Il est capable de détecter correctement près de 93 % des patients réellement malades, évitant ainsi de renvoyer chez lui un patient à risque sans surveillance.

- Excellente capacité de discrimination (AUC-ROC = 96,32 %) **:** Avec le score AUC-ROC le plus élevé, le Random Forest démontre une capacité quasi-parfaite à distinguer un patient sain d'un patient atteint d'une pathologie cardiaque .

- Fiabilité des prédictions positives (Précision = 89,66 %) **:** Lorsqu'il diagnostique une maladie cardiaque, le modèle a raison dans près de 90 % des cas, ce qui limite les fausses alertes et le stress inutile pour les patients sains.

- **Robustesse au bruit et aux valeurs aberrantes .**

- **Gestion des relations non linéaires** : le modèle capture naturellement les interactions complexes entre les facteurs de risque cardiovasculaire (âge, pression artérielle, cholestérol, etc.).

- **Importance des caractéristiques** : le classement **automatique** des variables les plus discriminantes renforce l'interprétabilité clinique.

- **Passage à l'échelle** : l'algorithme conserve de bonnes performances même lorsque le volume de données augmente, sans prétraitement coûteux.

La supériorité statistique du Random Forest ne doit pas occulter ses contraintes techniques. Afin de valider la faisabilité de son intégration dans un environnement médical réel, il est essentiel d'analyser ses limites et ses exigences opérationnelles.

- **Généralisation limitée** : Les résultats obtenus le sont sur un jeu de données spécifique, prétraité et équilibré dans le cadre de ce travail. Le modèle fonctionne bien sur cet ensemble, mais cela ne garantit pas son efficacité sur une population réelle et hétérogène. Une **validation clinique** rigoureuse sur des patients réels, dans un cadre hospitalier, avec des données collectées en conditions réelles serait indispensable pour prouver sa viabilité et sa robustesse avant tout déploiement en pratique médicale.

- La taille du modèle (nombre d'arbres, profondeur) peut devenir importante, rendant son stockage et son déploiement plus coûteux. Par ailleurs, bien que plus interprétable qu'un réseau de neurones, le Random Forest reste moins transparent qu'un arbre de décision unique .

Bien que la méthode des k-Nearest Neighbors est simple, elle présente plusieurs limites importantes.

- Elle peut devenir coûteuse en temps de calcul, car chaque nouvelle observation doit être comparée à l'ensemble des données d'apprentissage, ce qui ralentit les prédictions lorsque le volume de données est élevé.

- De plus, ses performances dépendent fortement du choix du paramètre k : une valeur trop faible rend le modèle sensible au bruit, tandis qu'une valeur trop élevée peut entraîner une perte de précision en lissant excessivement les différences entre les classes.

- Le kNN est également sensible aux valeurs aberrantes, car une observation

incorrecte peut influencer directement la décision finale. Enfin, il est affecté par la “malédiction de la dimension”, où l’augmentation du nombre de variables rend les distances moins pertinentes et dégrade ainsi la qualité des résultats. Il exige également une normalisation stricte des données pour éviter qu'une variable n'écrase les autres.

L'évaluation du modèle SVM met en évidence certaines limites qu'il convient de souligner :

- La performance du SVM est extrêmement dépendante du choix du noyau et des paramètres. Bien que cette sélection ait été **automatisée par notre code**, elle exige une phase de recherche coûteuse en temps et en ressources pour atteindre une configuration optimale.

- Sensibilité au bruit et au déséquilibre : L'algorithme est vulnérable aux valeurs aberrantes qui perturbent la trajectoire de sa frontière de séparation (l'hyperplan). Sans ajustement spécifique, il tend également à favoriser la classe majoritaire, ce qui dégrade sa sensibilité et augmente le risque de faux négatifs.

- Coût computationnel élevé : En raison d'une complexité mathématique , le temps d'apprentissage du SVM explose lorsque le volume de données augmente, limitant sa capacité à passer à l'échelle sur de très grandes bases de données hospitalières.

- Difficulté d'interprétation clinique : Le fonctionnement du SVM repose sur des projections géométriques abstraites. Contrairement aux arbres de décision, il n'offre pas de règles logiques claires, ce qui rend ses prédictions difficiles à justifier et à interpréter pour un personnel médical.

L'analyse du modèle MLP met en lumière des faiblesses structurelles importantes, particulièrement visibles lors de son application à notre recherche :

- Les réseaux de neurones exigent de grandes quantités de données pour être entraînés efficacement. Sur un échantillon restreint comme le dataset de Cleveland (303 patients), le MLP a tendance à mémoriser le bruit et les spécificités des données d'entraînement, ce qui dégrade ses performances sur le jeu de test.

- Le MLP fonctionne comme une "boîte noire" complexe où l'information traverse plusieurs couches de neurones cachées. Cette absence de transparence rend l'explication clinique des diagnostics impossible, un défaut majeur dans un contexte médical où la traçabilité et la justification des décisions sont indispensables.

- Complexité de configuration et coût d'entraînement .
- Stricte dépendance à la mise à l'échelle des données (standardisation) .

Conclusion :

Ce troisième chapitre révélant notre partie pratique, a permis de concrétiser notre démarche prédictive à travers le déploiement d'un protocole expérimental rigoureux, allant du prétraitement des variables cliniques jusqu'à l'optimisation fine des modèles . La confrontation systématique de nos quatre classifieurs a rendu un verdict clair : le modèle Random Forest s'impose comme la solution la plus performante et la plus sécurisante sur le plan médical, affichant une sensibilité de 92,86% et une aire sous la courbe de 0,9632. Toutefois, notre étude met également en lumière l'excellent comportement du modèle KNN, qui se présente comme une alternative intéressante en termes d'interprétabilité et d'équilibre entre performance et complexité.

En montrant que les forêts aléatoires et le KNN sont meilleurs que le SVM et le réseau de neurones (MLP), ce chapitre livre les clés techniques du modèle intelligent optimal.Une fois validé sur les données de test, ce classifieur se positionne comme un support de diagnostique robuste.

Conclusion Générale

Conclusion Générale

Ce projet vise à mettre l'apprentissage automatique au service de la santé cardiovasculaire, un enjeu critique face à des pathologies qui demeurent la menace sanitaire la plus meurtrière au monde. Pour ouvrir la voie à un dépistage plus précoce, notre étude s'est d'abord penchée sur la nature subtile de ces anomalies afin de combler les lacunes prédictives des méthodes de diagnostic classiques.

Pour y répondre, nous avons bâti un protocole rigoureux, transformant la donnée brute en un levier d'apprentissage automatique. Chaque étape, du nettoyage des bases de données à l'optimisation de la distribution des classes par stratification, a été pensée pour garantir la fiabilité des modèles. L'encodage par One-Hot Encoding a ainsi permis de donner une voix numérique aux variables cliniques catégorielles. Au cœur de cette expérimentation, quatre piliers algorithmiques majeurs ont été mis à l'épreuve : les réseaux de neurones (MLP), les machines à vecteurs de support (SVM), les K plus proches voisins (KNN) et les forêts aléatoires (Random Forest).Au-delà des équations et des codes, cette étude met en lumière le formidable potentiel de l'intelligence artificielle en tant que boussole diagnostique. Loin de vouloir remplacer le geste clinique, les modèles que nous avons développés se dessinent comme une première ligne de défense. Ils offrent aux praticiens un outil de dépistage précoce et intelligent, capable de repérer les signaux faibles, de prioriser les patients à haut risque et d'orienter les examens approfondis vers ceux qui en ont le plus urgemment besoin.

Les simulations numériques valident concrètement cette approche computationnelle en confirmant la supériorité du Random Forest. Ce modèle champion se distingue par des scores remarquables : une exactitude (Accuracy) de 91,80 %, une sensibilité (Recall) de 92,86 % et une aire sous la courbe (AUC-ROC) de 96,32 %. De telles métriques prouvent la robustesse de l'algorithme pour discriminer les profils sains des sujets pathologiques.

Les contributions de ce mémoire se déclinent sur trois niveaux distincts :

- **Scientifique :** Par une comparaison objective de quatre algorithmes majeurs et la mise en place d'un protocole expérimental standardisé et reproductible.
- **Médical :** En proposant une solution d'aide au diagnostic à haute sensibilité, directement intégrable au sein des systèmes d'aide à la décision clinique pour optimiser le temps médical et réduire les coûts par la prévention.
- **Pédagogique :** En documentant chaque étape de traitement pour offrir un

guide méthodologique clair, facilitant l'appropriation de l'IA par le corps médical.

L'enjeu à présent est d'appréhender la portée de ces contributions et de définir comment un tel outil peut s'intégrer harmonieusement dans le quotidien des cardiologues comme une aide à la décision, sans jamais se substituer à leur expertise humaine. L'objectif ultime n'est pas d'automatiser l'acte médical, mais d'offrir un second regard technologique fiable, capable de faire gagner un temps précieux aux praticiens lors du tri clinique.

Malgré ces performances, notre étude se heurte à des contraintes techniques et méthodologiques. La base de données reste restreinte à 303 patients et limitée à 19 caractéristiques cliniques, ce qui exclut des biomarqueurs cruciaux (troponine, BNP) ou un suivi longitudinal. De plus, le léger déséquilibre des classes (54 % sains, 46 % malades), l'absence de validation externe (Framingham, Kaggle) et le manque de données locales de patients algériens freinent la généralisation immédiate du modèle.

Pour franchir le cap d'une application médicale réelle, nos perspectives de recherche s'articulent autour de plusieurs axes :

- **Parcours de soins et Déploiement :** Intégrer l'outil sous forme de questionnaire intelligent pour orienter les profils critiques vers des examens approfondis (ECG, échographie), tout en ciblant une certification réglementaire (normes CE) et une interface logicielle ergonomique.
- **Évolutions techniques :** Explorer l'apprentissage multi-sources (Multi-Source Learning) pour fusionner des bases hétérogènes, le Deep Learning multimodal (génétique, imagerie) et le Federated Learning pour garantir la confidentialité des données entre établissements.
- **Confiance :** Développer l'IA explicable (Explainable AI) pour rendre les prédictions transparentes pour les médecins.

Bien que les résultats soient prometteurs, nous soulignons que ces modèles doivent être utilisés comme outils d'appoint, jamais comme substitut à l'examen clinique du cardiologue. C’est précisément dans cette synergie entre l’intuition clinique humaine et la puissance computationnelle de la machine que réside l’avenir d’une cardiologie moderne, préventive et personnalisée.

Références Bibliographiques

[1]World Health Organization (WHO), “Maladies cardiovasculaires,” Fact sheet, Jun. 2021. Available: https://www.who.int/fr/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) .

[2] F. Pinto, “Deaths from cardiovascular disease surged 60% globally over the last 30 years: Report,” May 2023. Available: https://faustopinto.com/2023/05/deaths-from-cardiovascular-disease-surged-60-globally-over-the-last-30-years-report/?lang=fr

[3] Fédération mondiale du cœur (WHF), “Qu'est-ce qu'une maladie cardiovasculaire ?,” Available: https://world-heart-federation.org/fr/what-is-cvd/

[4] Santé.fr, “Le système cardio-vasculaire,” Available: https://www.sante.fr/le-systeme-cardio-vasculaire .

[5] A. Y. Akgun and J. M. Bateman, “Anatomy, Thorax, Heart,” in *StatPearls* [Internet], Treasure Island, FL, USA: StatPearls Publishing, 2025. Available: https://www.ncbi.nlm.nih.gov/books/NBK535355/.

[6] M. A. Alghamdi and R. S. Alotaibi, “Cardiovascular System,” in *StatPearls* [Internet], Treasure Island, FL, USA: StatPearls Publishing, 2025. Available: https://www.ncbi.nlm.nih.gov/books/NBK541098/ .

[7] Santé sur le Net, “Structure anatomique du cœur (image),” Oct. 2015. Available: https://www.sante-sur-le-net.com/wp-content/uploads/2015/10/structure-anatomique-du-c%C5%93ur.jpg.

[8] Hôpitaux Universitaires de Genève (HUG), “Fonctionnement des valvules,” Atlas de santé, disponible : https://www.hug.ch/atlas-sante/anatomie/fonctionnement-des-valvules.

[9] Centre Hospitalier Universitaire de Rouen, “Schéma de la valve aortique,” image en ligne. Available: https://www.chu-rouen.fr/wp-content/uploads/sites/66/2023/04/schema-valve-aortique.png

[10] F. Bessaguet, D. Henrion, and A. Desmoulière, “Organisation générale des vaisseaux sanguins,” *Actualités Pharmaceutiques*, vol. 61, no. 619, pp. 53–56, 2022, doi: 10.1016/j.actpha.2022.07.046.

[11] H. M. Reisner, “Présentation simplifiée du système cardiovasculaire avec les différents types des vaisseaux sanguins élastiques et musculaires,” figure, 2015. Available:

https://www.researchgate.net/profile/Zeinab-El-Dirani-2/publication/332961044/figure/fig8/AS:794597303263233@1566458168633/Presentation-simplifiee-du-systeme-cardiovasculaire-avec-les-differents-types-des.jpg .

[12] A. M. T. Bui and J. M. Hake, "Cardiac Physiology," in *StatPearls*, Treasure Island, FL, USA: StatPearls Publishing, 2025. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK572070/.

[13] Masterclass Formations Naturopathe, "Endocarde, myocarde et péricarde," image en ligne, 2021. Available: https://masterclass.formations-naturopathe.eu/wp-content/uploads/2021/03/endo-myo-pericarde.png

[14] R. Chaudhry, J. H. Miao, and A. Rehman, "Physiology, Cardiovascular," *StatPearls [Internet]*, Treasure Island (FL), USA: StatPearls Publishing, 2025. Available: https://www.ncbi.nlm.nih.gov/books/NBK493197/.

[15] SMART-Servier Medical Art, "Cardiovascular system illustration," *SVT Tanguy Jean (via Smart-Servier resources)*, [Online]. Available: https://www.svt-tanguy-jean.com/uploads/1/2/0/4/120408978/tb-10-circulation.pdf

[16] H. Dobrzyński, R. H. Anderson, and A. Atkinson, "Structure, function and clinical relevance of the cardiac conduction system, including the atrioventricular ring and outflow tract tissues," *Pharmacology & Therapeutics*, vol. 139, no. 2, pp. 260–288, 2013, doi: 10.1016/j.pharmthera.2013.04.010.

[17] "Image (sans titre)," Available: https://static.wixstatic.com/media/a8e1f9_3d42d6e130604918a75dc30457b4c472.jpg/v1/fill/w_600,h_272,al_c,q_80,enc_avif,quality_auto/a8e1f9_3d42d6e130604918a75dc30457b4c472.jpg.

[18] J. M. Smith and A. Patel, "Physiology, Cardiac Function," in *StatPearls [Internet]*, Treasure Island (FL), USA: StatPearls Publishing, 2025. Available: https://www.ncbi.nlm.nih.gov/books/NBK564304/.

[19] "CCIB Education (image)," CCI Belle Donne, 2020. Available: https://ccibelledonne.com/wp-content/uploads/2020/11/ccib_education_1.png

[20] S. M. Brown and A. A. Smith, "Anatomy, Thorax, Heart," in *StatPearls*

[Internet], Treasure Island (FL), USA: StatPearls Publishing, 2025. Available: https://www.ncbi.nlm.nih.gov/books/NBK558923/.

[21] Cap Retraite, "Infographie : Arythmie cardiaque," 2025. Available: https://www.capretraite.fr/wp content/uploads/2025/08/infographie_arythmie_cardiaque.jpg.

[22] S. Ahayan and S. Ghulam, *Guide IPA PCS PPCSP : Infirmière en pratique avancée. Pathologies chroniques stabilisées. Prévention et polypathologies courantes en soins primaires*, Issy-les-Moulineaux, France: Elsevier Masson, 2025. Available: https://books.google.dz/books?hl=fr&lr=&id=jPMmEQAAQBAJ&oi=fnd&pg=PA65&dq=Comprendre+l%27insuffisance+cardiaque&ots=0p2iYJvm4T&sig=7rwbTJAs6Sbvizjeggyq HZcv3-o&redir_esc=y#v=onepage&q=Comprendre%20l'insuffisance%20cardiaque&f=false

[23] H. Alkhareef et al., "Comparison of healthy left ventricle during systole with dilated left ventricle (figure)," in *A Review of Heart Failure Disease Management*, 2017. Available: https://www.researchgate.net/profile/Hamad-Alkhareef-2/publication/321587487/figure/fig2/AS:568580728791047@1512571617850/Comparison-of-healthy-left-ventricle-during-systole-with-dilated-left-ventricle-Medscape.png.

[24] A. S. Khaku and P. Tadi, "Cerebrovascular Disease," in *StatPearls* [Internet]. Treasure Island (FL), USA: StatPearls Publishing, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK430927/.

[25] Cœur et Santé Déodatien, "AVC ischémique et hémorragique (image)," 2020. Available:https://coeuretsante.deodatie.fr/wp-content/uploads/2020/06/avc_ischemique-et-hemorragique-scaled.jpg.

[26] K. K. Teo and T. Rafiq, "Cardiovascular Risk Factors and Prevention: A Perspective From Developing Countries," *Canadian Journal of Cardiology*, vol. 37, no. 5, pp. 733–743, May 2021, doi: 10.1016/j.cjca.2021.02.009. https://pubmed.ncbi.nlm.nih.gov/33610690/.

[27] S. Yusuf, S. Reddy, S. Ôunpuu et S. Anand, "Global burden of cardiovascular diseases: Part I: General considerations, the epidemiologic transition, risk factors, and impact of urbanization," *Circulation*, vol. 104, no. 22, pp. 2746–2753, 2001, doi: 10.1161/hc4601.099487. https://pmc.ncbi.nlm.nih.gov/articles/PMC8109623/

[28] A. Trimaille, “Les outils d’estimation du risque cardiovasculaire, le SCORE2 de l’ESC,” *Revue Francophone des Laboratoires*, vol. 2022, no. 543, pp. 35–42, 2022, doi: 10.1016/S1773-035X(22)00213-1. [Online]. Available: https://doi.org/10.1016/S1773-035X(22)00213-1

[29] C. Li, X. Liu, P. Shen, Y. Sun, T. Zhou, W. Chen, Q. Chen, H. Lin, X. Tang et P. Gao, “Improving cardiovascular risk prediction through machine learning modelling of irregularly repeated electronic health records,” *European Heart Journal – Digital Health*, vol. 5, no. 1, pp. 30–40, 2023, doi: 10.1093/ehjdh/ztad058. https://pmc.ncbi.nlm.nih.gov/articles/PMC10802828/

[30] Centre de Gestion de l’Aisne (CDG 02), *Les maladies cardio-vasculaires*, [En ligne].Available:https://www.cdg02.fr/CDG02_WEB/doc/5/MALADIES_CARDIO_VASCUALAI.pdf.

[31] Y. Sattar and L. Chhabra, “Electrocardiogram,” *StatPearls*, StatPearls Publishing, Treasure Island (FL), USA, Jun. 5, 2023. Available: https://www.ncbi.nlm.nih.gov/books/NBK549803/

[32]Data Sciences International, “Heart (image),” Available: https://www.datasci.com/images/default-source/cardiovascular-ecg/heart.png?sfvrsn=55ceaf51_9.

[33] M. K. Herbst, J. Velasquez, G. Adnan, and M. C. O'Rourke, “Cardiac Ultrasound,” *StatPearls*, StatPearls Publishing, Treasure Island (FL), USA, Nov. 21, 2022. Available: https://www.ncbi.nlm.nih.gov/books/NBK470584/

[34]“Image (source originale non identifiée),” Available: https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcSG4K6X2mYca0uj_Fdzo8qyGdXk13cuATIO8A&s,

[35] D. J. Omeh and E. Shlofmitz, “Angiography,” *StatPearls*, StatPearls Publishing, Treasure Island (FL), USA, Aug. 7, 2023. Available: https://www.ncbi.nlm.nih.gov/books/NBK557477/

[36] Santé sur le Net, “Rétrécissement d’une artère coronaire (image),” Oct. 2017. Available:https://www.sante-sur-le-net.com/wp content/uploads/2017/10/retrecissement-artere-coronaire.jpg.

[37] C. Carluccio, F. Bressan, M. Pizzolato, A. De Antoni, S. Ungaro, D. Balla, A. Cipriani, M. De Lazzari, M. Perazzolo Marra, H. Vago, D. Corrado, A. Zorzi and F. Graziano, "Exercise Stress Testing in Clinical Cardiology: A Practical Guide to Performance and Interpretation," *Journal of Clinical Medicine*, vol. 15, no. 4, p. 1656, 2026, doi: 10.3390/jcm15041656. https://www.mdpi.com/2077-0383/15/4/1656

[38] J. N. de Alencar and G. S. Barbosa, "The inverse problem of electrocardiography: From a clinical point of view," *JRSM Cardiovascular Disease*, vol. 15, 2026, doi: 10.1177/20480040261457700. https://www.researchgate.net/publication/406083687_The_inverse_problem_of_electrocardiography_From_a_clinical_point_of_view.

[39] S. Janower, "Limites de l'échographie cardiaque de stress," dans *Imagerie en coupes du cœur et des vaisseaux*, L. Boyer et P. Guéret, Eds., Paris, France : Springer, 2013, doi: 10.1007/978-2-8178-0435-4_9. Disponible : https://link.springer.com/chapter/10.1007/978-2-8178-0435-4_9.

[40] M. Romanens, "Apports et limites de l'ECG d'effort dans la pratique médicale quotidienne : un update," *Revue Médicale Suisse*, vol. 2, pp. 638–645, 2006. https://www.revmed.ch/view/642235/5025692/RMS_56_638.pdf.

[41] R. J. Myerson, "Clinical evaluation of anticoagulant therapy in atrial fibrillation," *Journal of Cardiovascular Electrophysiology*, vol. 13, no. 12, pp. 1238–1244, 2002, doi: 10.1046/j.1540-8167.2002.01238.x. https://pubmed.ncbi.nlm.nih.gov/12555359/.

[42] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, no. 1, pp. 44–56, 2019.

[43] Mihaela van der Schaar et al., "How artificial intelligence and machine learning can help healthcare systems respond to COVID-19," *Machine Learning*, vol. 110, no. 1, pp. 1–14, 2021.

[44] M. Çakır and O. I. Çetin, "A Bibliometric Analysis on Machine Learning Applications in the Finance Sector," Conference Paper, Dec. 2022. [Online]. Available: ResearchGate

[45] R. Mifdal, *Application des techniques d'apprentissage automatique pour la prédiction de la tendance des titres financiers*, Thèse de doctorat, École de technologie supérieure, 2019.

[46] H. Bellahmer, *Implémentation et évaluation d'un modèle d'apprentissage automatique pour l'estimation de la valeur marchande de propriétés immobilières*, Thèse de doctorat, Université Mouloud Mammeri de Tizi-Ouzou, 2020.

[47] Richard S. Sutton and Andrew G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.

[48] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
https://www.deeplearningbook.org/

[49] Kaizen, Machine Learning illustration, Kaizen, 2024. [En ligne]. Disponible : https://kaizen.com/wp-content/uploads/2024/07/image-64.png

[50] David W. Hosmer, Stanley Lemeshow, and Rodney X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Hoboken, NJ, USA: Wiley, 2013.

[51] Unknown (Bing Images), *Illustration / image de machine learning*, Bing Images, 2026. [En ligne]. Disponible : https://thfvnext.bing.com/th/id/OIP.x-1u5_YGsSk0W7WsK82HqwHaEo

[52] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.

[53] Medium, *Machine Learning illustration*, Medium, 2024. [En ligne]. Disponible : https://miro.medium.com/v2/resize:fit:1200/1*1woLesOTZE0t3EswiT0mnA.png

[54] Corinna Cortes and Vladimir N. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[55] Unknown (Bing Images), *Illustration (machine learning concept)*, Bing Images, 2026. [En ligne]. Disponible : https://thfvnext.bing.com/th/id/OIP.pZbu5wvKp2QU3-aNfVQiGwHaEg

[56] T. M. Cover and P. E. Hart, "Nearest Neighbor Pattern Classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967.

[57] Sysblog, "K-Nearest Neighbors (KNN) illustration," Université Paris Diderot

– Informatique, 2020. [Online]. Available: https://sysblog.informatique.univ-paris-diderot.fr/wp-content/uploads/2020/03/knn2.png

[58] Simon Haykin, *Neural Networks and Learning Machines*, 3rd ed. Upper Saddle River, NJ, USA: Pearson, 2009.

[59] DataScientest, *Machine Learning workflow diagram*, DataScientest, 2020. [En ligne]. Disponible : https://datascientest.com/wp-content/uploads/2020/06/DP_2.png

[60] ICHI.PRO, "Evaluation Metrics in Machine Learning," [Online]. Available: https://ichi.pro

[61] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," *BMC Medical Informatics and Decision Making*, 2020.

[62] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, *Heart Disease Dataset*. Irvine, CA, USA: UCI Machine Learning Repository, 1989. DOI: 10.24432/C52P4X.

[63] T. R. Dawber, G. F. Meadors, and F. E. Moore, "Epidemiological approaches to heart disease: The Framingham Study," *American Journal of Public Health*, vol. 41, no. 3, pp. 279–286, 1951.

[64] Kaggle, *Heart Disease Dataset*. Available: https://www.kaggle.com.

[65] J. W. Tukey, *Exploratory Data Analysis*. Reading, MA, USA: Addison-Wesley, 1977.

[66] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[67]O. Sami, Y. Elsheikh, and F. Almasalha, "The Role of Data Pre-processing Techniques in Improving Machine Learning Accuracy for Predicting Coronary Heart Disease," International Journal of Advanced Computer Science and Applications, vol. 12, no. 6, pp. 816–824, 2021, doi: 10.14569/IJACSA.2021.0120695.

[68] H. Benhar, A. Idri, and J. L. Fernández-Alemán, "Data preprocessing for heart disease classification: A systematic literature review," Computer Methods and Programs in Biomedicine, vol. 195, p. 105635, 2020, doi: 10.1016/j.cmpb.2020.105635.

[69] M. Pica, "Cycle Data de l'ARHN," Maison des Sciences de l'Homme Lyon,

France, 2023, Available: https://hal.science/hal-04800035v1/document

[70] H. P. Vinutha, B. Poornima, and B. M. Sagar, "Detection of outliers using interquartile range technique from intrusion dataset," in Advances in Intelligent Systems and Computing, vol. 701, pp. 511–518, 2018, doi: 10.1007/978-981-10-7563-6_53.

[71] C. M. Salgado, C. Azevedo, H. Proença, and S. M. Vieira, "Noise Versus Outliers," in Secondary Analysis of Electronic Health Records, Springer, Cham, 2016, doi: 10.1007/978-3-319-43742-2_14.

[72] L. Berti-Équille, "Qualité des données," Techniques de l'Ingénieur, ref. H3700, 2018. doi: 10.51257/a-v1-h3700. Available: https://www.techniques-ingenieur.fr/base-documentaire/technologies-de-l-information-th9/big-data-42659210/qualite-des-donnees-h3700/

[73] P. Gupta and A. Bagchi, "Introduction to Pandas," in Python Data Analytics: With Pandas, NumPy, and Matplotlib, Apress, 2024, doi: 10.1007/978-3-031-43725-0_5.

[74] F. Nelli, "The NumPy Library," in Python Data Analytics: With Pandas, NumPy, and Matplotlib, 3rd ed., Apress, 2023, doi: 10.1007/978-1-4842-9532-8_3.

[75] K. Benhamza, R. Benselim, H. Naidja, and H. Seridi, "A comprehensive survey of imputation methods in medical missing data analysis," *Applied Intelligence*, vol. 55, p. 778, 2025, doi: 10.1007/s10489-025-06602-2.

[76] M. Cottrell, S. Ibbou, and P. Letrémy, "Traitement des données manquantes au moyen de l'algorithme de Kohonen," INRIA, 2010. Available: https://inria.hal.science/inria-00494698v1

[77] G. Cottrell, M. Cot, and J.-Y. Mary, "L'imputation multiple des données manquantes aléatoirement : concepts généraux et présentation d'une méthode Monte-Carlo," *Rev. Épidémiol. Santé Publique*, vol. 57, no. 5, pp. 361–372, 2009, doi: 10.1016/j.respe.2009.04.011.

[78] F. Clerici and N. Nobani, "Categorical variable encoding methods for tabular data: a benchmarking study," *Int. J. Data Sci. Anal.*, vol. 22, art. no. 55, 2026, doi: 10.1007/s41060-025-00886-w.

[79] Scikit-learn Developers, "OneHotEncoder — sklearn.preprocessing," Scikit-learn Documentation, Available: https://scikit-

learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html

[80]H. Wang, E. Doumard, C. Soulé-Dupuy, P. Kémoun, J. Aligon, and P. Monsarrat, "Explanations as a New Metric for Feature Selection: A Systematic Approach," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 8, pp. 4131–4142, 2023, doi: 10.1109/JBHI.2023.3279340

[81]P.-E. Jouve and N. Nicoloyannis, "Une approche filtre pour la sélection de variables en apprentissage non supervisé," *EGC 2005, RNTI-E-3*, pp. 31–42, 2005, Available: https://editions-rnti.fr/?inprocid=1000209

[82]N. S. Muhammad Noor Mathivanan, E. Foo Zhi Xian, D. Foo Yong Xi, and C. Hiang Kiat, "Impact of Feature Standardization on Heart Disease Prediction: A Comparative Analysis of Logistic Regression and Support Vector Machine Models," *Malaysian Journal of Computing*, vol. 10, no. 1, pp. 2159–2175, 2025, doi: 10.24191/mjoc.v10i1.6835.

[83]Scikit-learn Developers, "StandardScaler — sklearn.preprocessing," *Scikit-learn Documentation*, Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

[84] Greenacre, M., Groenen, P. J. F., Hastie, T., et al., "Principal component analysis," *Nature Reviews Methods Primers*, vol. 2, no.100, 2022, doi: 10.1038/s43586-022-00184-w.

[85]Li Yang and Abdallah Shami, "On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice," *arXiv preprint arXiv:2007.15745*, 2020, doi: 10.48550/arXiv.2007.15745.

[86] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[87]Kumar, "Hyperparameter Tuning with Sklearn GridSearchCV and RandomizedSearchCV," *Machine Learning Knowledge*, Oct. 5, 2021. [Online]. Available: https://machinelearningknowledge.ai/hyperparameter-tuning-with-sklearn-gridsearchcv-and-randomizedsearchcv/.

[88] Unknown (Bing Images), *Illustration of machine learning concept*, Bing Images, 2026. [En ligne].

[89] Geekflare, *Confusion Matrix (basic_cm-edited)*, Geekflare, 2022. [En ligne]. Disponible : https://geekflare.com/fr/wp-content/uploads/2022/07/basic_cm-edited.jpg

[90]. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006, d] T oi: 10.1016/j.patrec.2005.10.010.

[91] Python Software Foundation, *Python Documentation*. [En ligne]. Disponible : Python Documentation.

[92] E. Bisong, "Google Colaboratory," in *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, Berkeley, CA, USA: Apress, 2019, pp. 59–64, doi: 10.1007/978-1-4842-4470-8_7.

[9**3]** C. R. Harris *et al*., "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020. Disponible : NumPy Documentation

[94] NumPy Developers, *NumPy Documentation*. [En ligne]. Disponible : NumPy Documentation

[95] J. D. Hunter, "Matplotlib: A 2D Graphics Environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[96] M. Waskom, *Seaborn: Statistical Data Visualization*. [En ligne]. Disponible : Seaborn Documentation.

[97] F. Pedregosa *et al*., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[98] D. Dua and C. Graff, *UCI Machine Learning Repository: Heart Disease Data Set*. University of California, Irvine, 1989. [En ligne]. Disponible : UCI Heart Disease Dataset

Annexes

Annexe A:

➢ Python :

Python est un langage de programmation interprété, reconnu pour sa simplicité et sa grande efficacité dans le développement rapide d'applications. Il est largement utilisé dans de nombreux domaines grâce à ses structures de données de haut niveau et à son approche claire de la programmation orientée objet.

Sa syntaxe lisible, sa flexibilité ainsi que son caractère multiplateforme en font un outil particulièrement adapté au développement de scripts et d'applications scientifiques. De plus, Python dispose d'une bibliothèque standard riche comprenant de nombreux modules prêts à l'emploi. **[91]**



➢ Google Colab :

Google Colaboratory (Colab) est un environnement de développement en ligne basé sur Jupyter Notebook. Il permet d'écrire et d'exécuter du code Python directement dans le navigateur, sans installation locale.

Colab offre des ressources de calcul gratuites, notamment l'accès aux GPU et TPU, ce qui le rend particulièrement adapté aux projets en apprentissage automatique et en science des données. Il facilite également le partage et la collaboration en temps réel sur les notebooks. **[92]**



- **Pandas :**

Pandas est une bibliothèque open source conçue pour la manipulation et l'analyse de données en Python. Elle fournit des structures de données performantes, notamment les DataFrames, qui facilitent le traitement des données structurées.

Elle est largement utilisée dans différents domaines tels que la finance, les statistiques ou encore l'analyse de données. Dans ce projet, Pandas a été utilisé pour manipuler et exploiter les données issues de fichiers CSV. **[93]**

- **NumPy :**

NumPy (Numerical Python) est une bibliothèque fondamentale dédiée au calcul scientifique. Elle permet de manipuler efficacement des tableaux multidimensionnels et propose un ensemble de fonctions mathématiques optimisées pour les opérations numériques.

Elle est souvent utilisée en complément de Pandas afin d'améliorer les performances des traitements sur les données. **[94]**

- **Matplotlib :**

Matplotlib est une bibliothèque de visualisation de données en Python permettant de créer des graphiques 2D de haute qualité. Elle est particulièrement utile pour l'exploration et la représentation visuelle des données. **[95]**

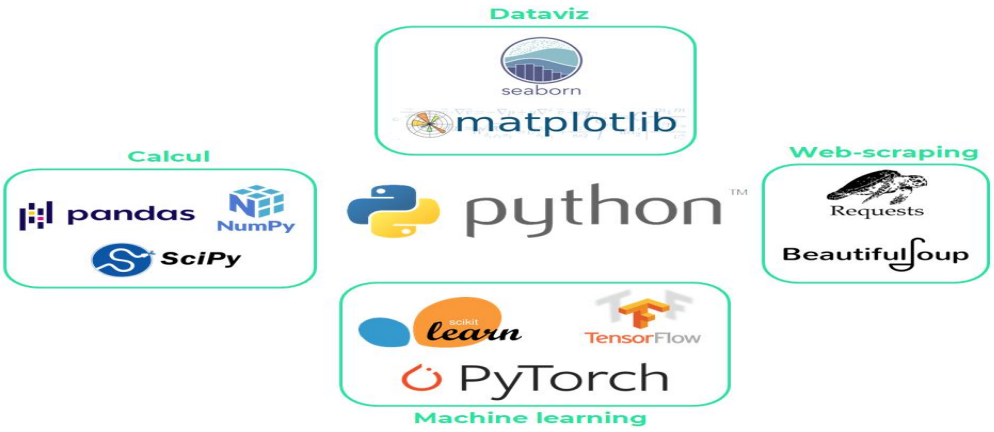
➢ Seaborn :

Seaborn est une bibliothèque de visualisation statistique basée sur Matplotlib. Elle permet de générer des graphiques plus esthétiques et informatifs tout en facilitant l'analyse exploratoire des données. Elle s'intègre naturellement avec les structures de Pandas. **[96]**

➢ Scikit-learn :

Scikit-learn est une bibliothèque dédiée à l'apprentissage automatique en Python. Elle regroupe plusieurs algorithmes pour la classification, la régression et le clustering.

Elle est appréciée pour sa simplicité d'utilisation, sa documentation claire et sa cohérence, ce qui la rend adaptée aussi bien à la recherche académique qu'aux applications industrielles. **[97]**



Variables cliniques du dataset de Cleveland :

Le dataset de Cleveland regroupe un ensemble de variables cliniques utilisées pour l'étude et la prédiction des maladies cardiovasculaires. Ces données décrivent différentes caractéristiques des patients, allant des informations générales jusqu'aux résultats d'examens médicaux.

On y retrouve tout d'abord des informations de base comme l'âge et le sexe du patient. Le jeu de données inclut également des éléments liés à l'état cardiaque, notamment le type de douleur thoracique, la pression artérielle au repos, le taux de cholestérol ainsi que la glycémie à jeun.

D'autres variables proviennent d'examens médicaux plus spécifiques, comme les résultats de l'électrocardiogramme au repos, la fréquence cardiaque maximale atteinte lors

d'un effort, ou encore la présence d'angine induite par l'exercice. Le dataset contient aussi des indicateurs importants tels que la dépression du segment ST, la pente de ce segment, le nombre de vaisseaux sanguins visibles et les résultats du test de thalassémie.

Enfin, la variable cible permet d'indiquer si le patient est atteint ou non d'une maladie cardiovasculaire.

Les principales variables cliniques de cette base sont :

a) **Âge (age)** : âge du patient en années
b) **Sexe (sex)** : sexe du patient (0 = femme, 1 = homme)
c) **Type de douleur thoracique (cp)** : catégorie de la douleur ressentie
d) **Pression artérielle au repos (trestbps)** : tension artérielle en mm Hg
e) **Cholestérol sérique (chol)** : taux de cholestérol en mg/dl
f) **Glycémie à jeun (fbs)** : si glycémie > 120 mg/dl (0 ou 1)
g) **Électrocardiogramme au repos (restecg)** : résultats ECG
h) **Fréquence cardiaque maximale (thalach)** : fréquence maximale atteinte
i) **Angine induite par l'exercice (exang)** : présence ou non (0 ou 1)
j) **Dépression ST (oldpeak)** : dépression induite par l'exercice
k) **Pente du segment ST (slope)** : forme du segment ST
l) **Nombre de vaisseaux (ca)** : vaisseaux principaux colorés (0 à 3)
m) **Thalassémie (thal)** : anomalie sanguine détectée
n) **Classe cible (target)** : présence ou absence de maladie cardiaque**[98]**

Annexe B :

Afin d'attester de la rigueur de notre protocole de prétraitement, cette section rassemble les analyses visuelles "avant/après" obtenues pour l'ensemble des variables numériques du dataset. Les Figures B.1 à B.4 illustrent comment l'application des filtres statistiques a permis de stabiliser les distributions en corrigeant les valeurs aberrantes (outliers) sans altérer la cohérence globale des données cliniques.

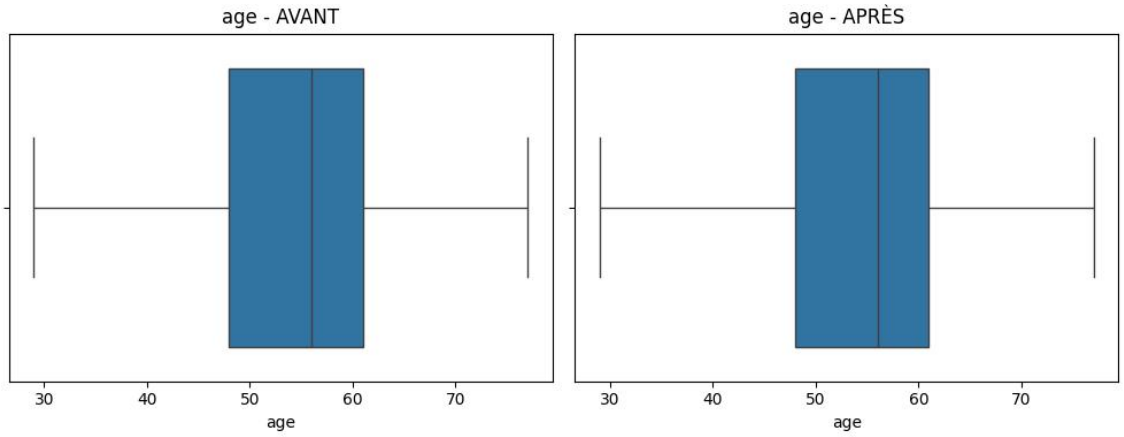


Figure B.1 : Impact du nettoyage des données sur la variable age.

Légende : Les boxplots comparatifs confirment l'absence initiale de valeurs aberrantes pour l'âge des patients. La distribution reste rigoureusement identique avant et après traitement, validant la cohérence biologique de cette caractéristique.

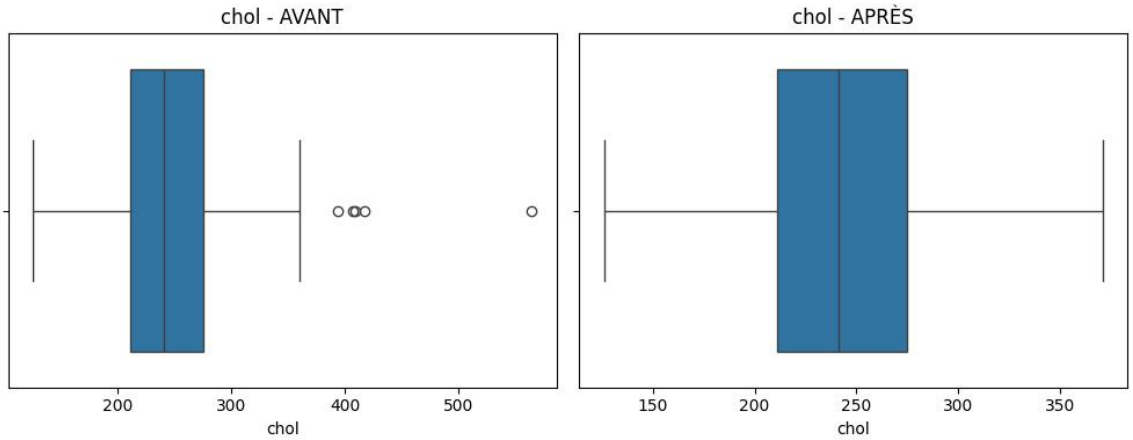


Figure B.2 : Impact du nettoyage des données, avec focus sur la correction des valeurs aberrantes de la variable chol.

Légende : Les boxplots montrent la suppression des valeurs extrêmes et pathologiques supérieures à 370 mg/dL, stabilisant la distribution avant l'entraînement des modèles .

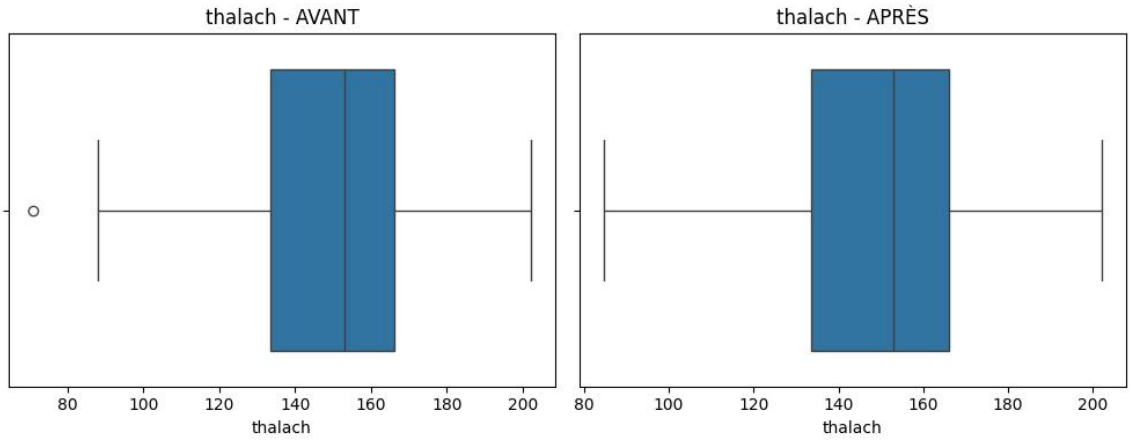


Figure B.3 : Impact du nettoyage des données, avec focus sur la correction des valeurs aberrantes de la variable thalach.

Légende : Les boxplots montrent la suppression de la valeur aberrante inférieure isolée située en dessous de 80 bpm, permettant de stabiliser et de resserrer la distribution sur des bornes physiologiques optimales.

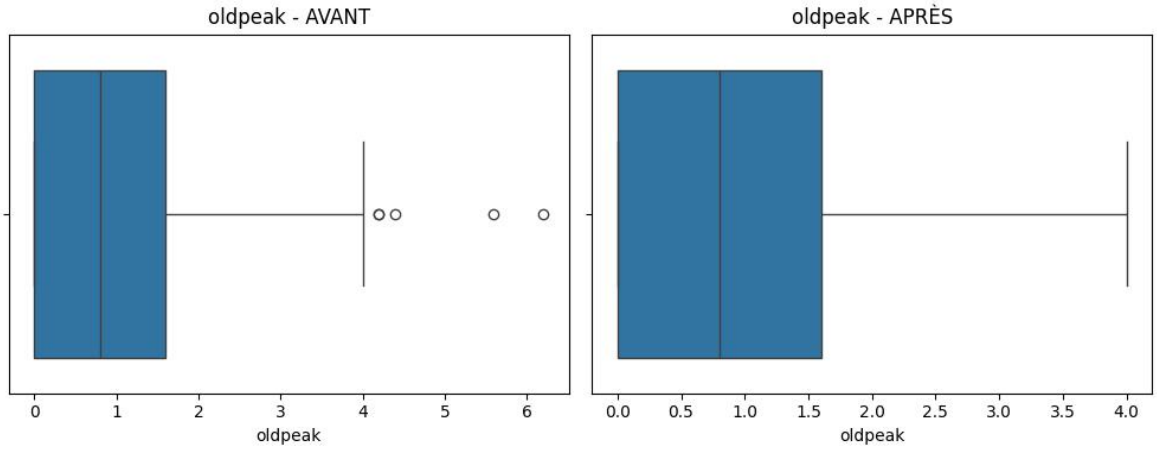


Figure B.4 : Impact du nettoyage des données, avec focus sur la correction des valeurs aberrantes de la variable oldpeak.

Légende : Les boxplots montrent la suppression des valeurs extrêmes supérieures à 4.0 (qui s'étalaient initialement jusqu'à plus de 6.0), stabilisant la distribution de cette mesure électrique avant l'entraînement des modèles.