

Détection des maladies cardiaques à l'aide d'algorithmes d'apprentissage automatique

Rédigé par : MAROUAN MORAKIB
Encadrant : MORAD NACHAOUI

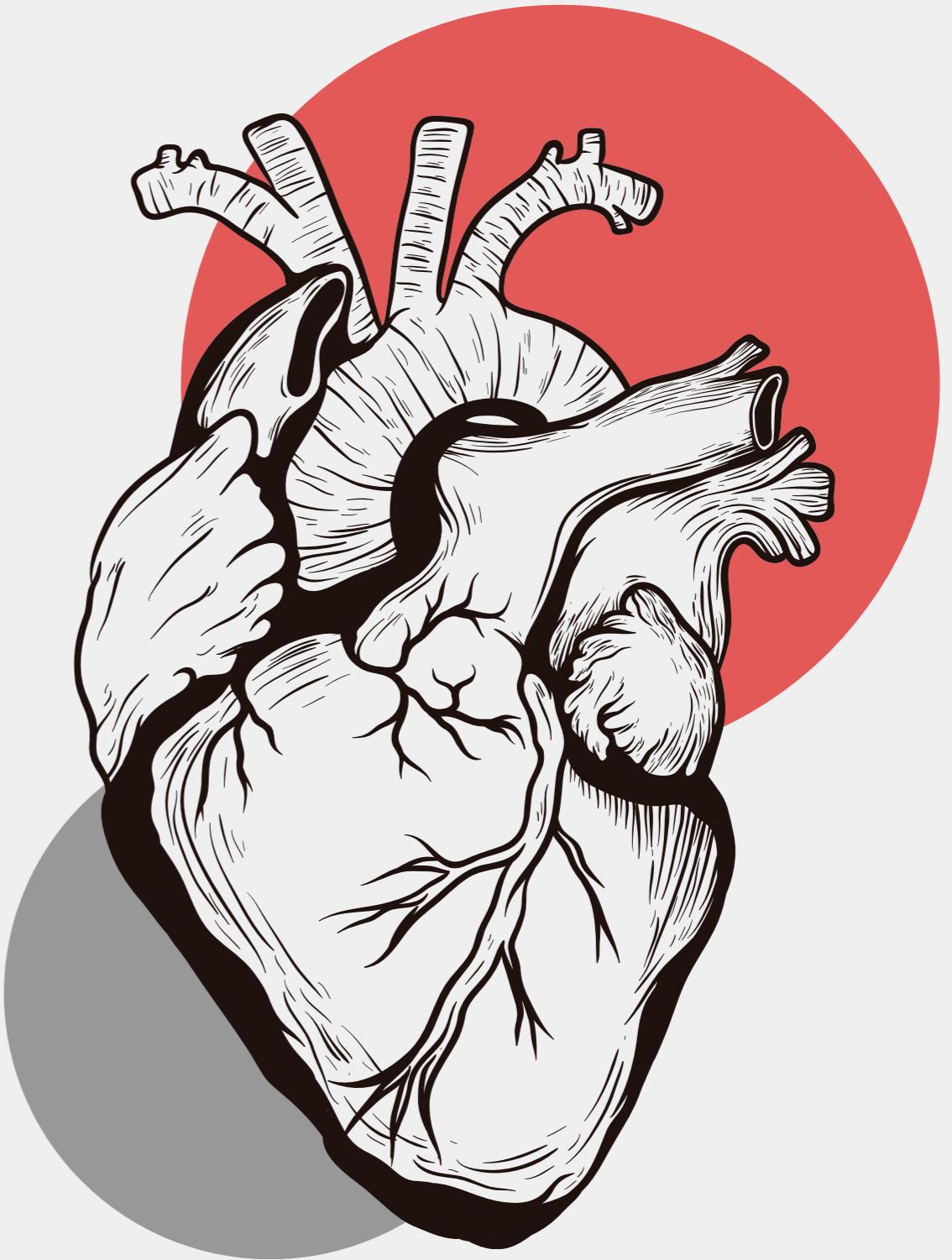
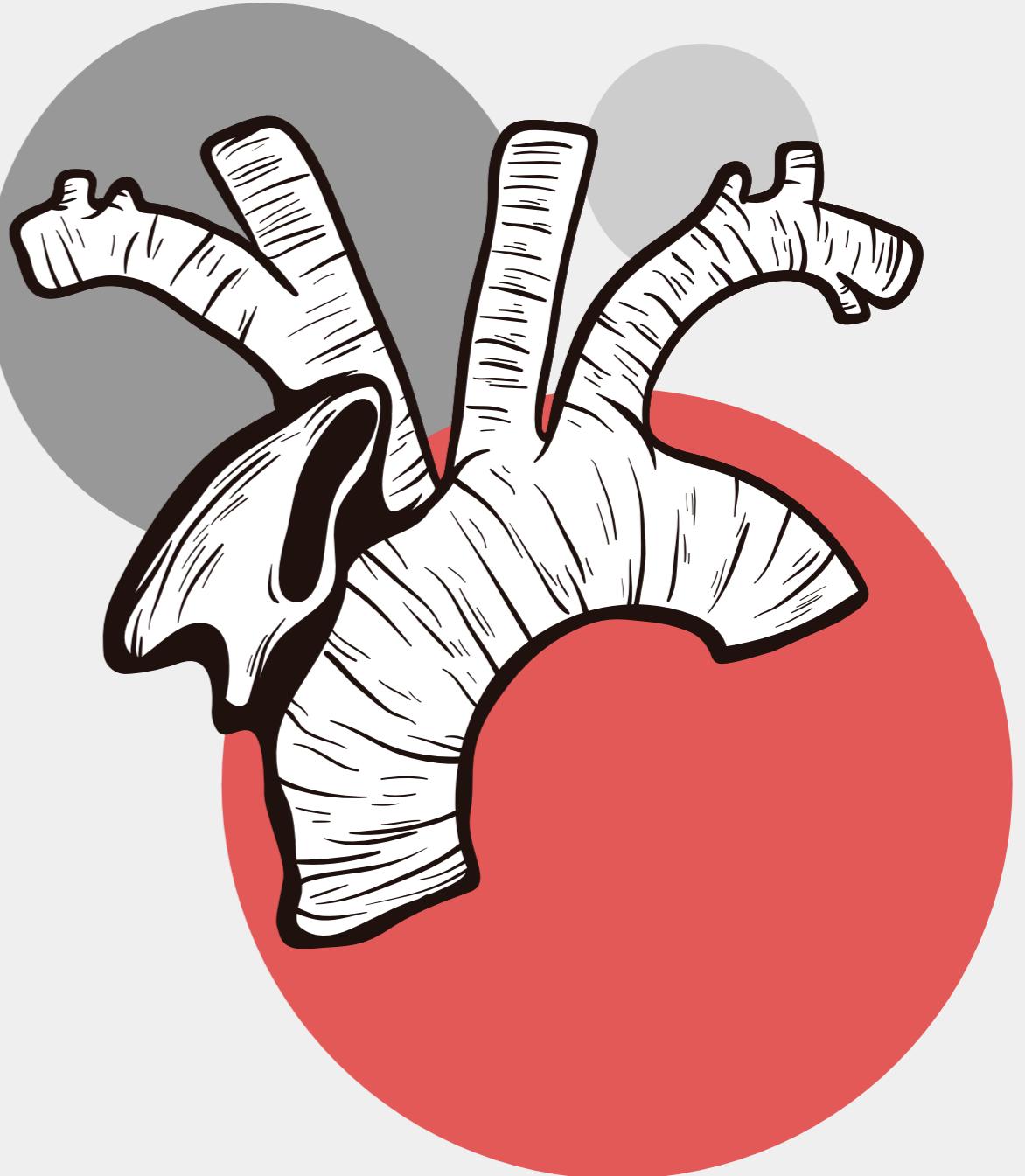


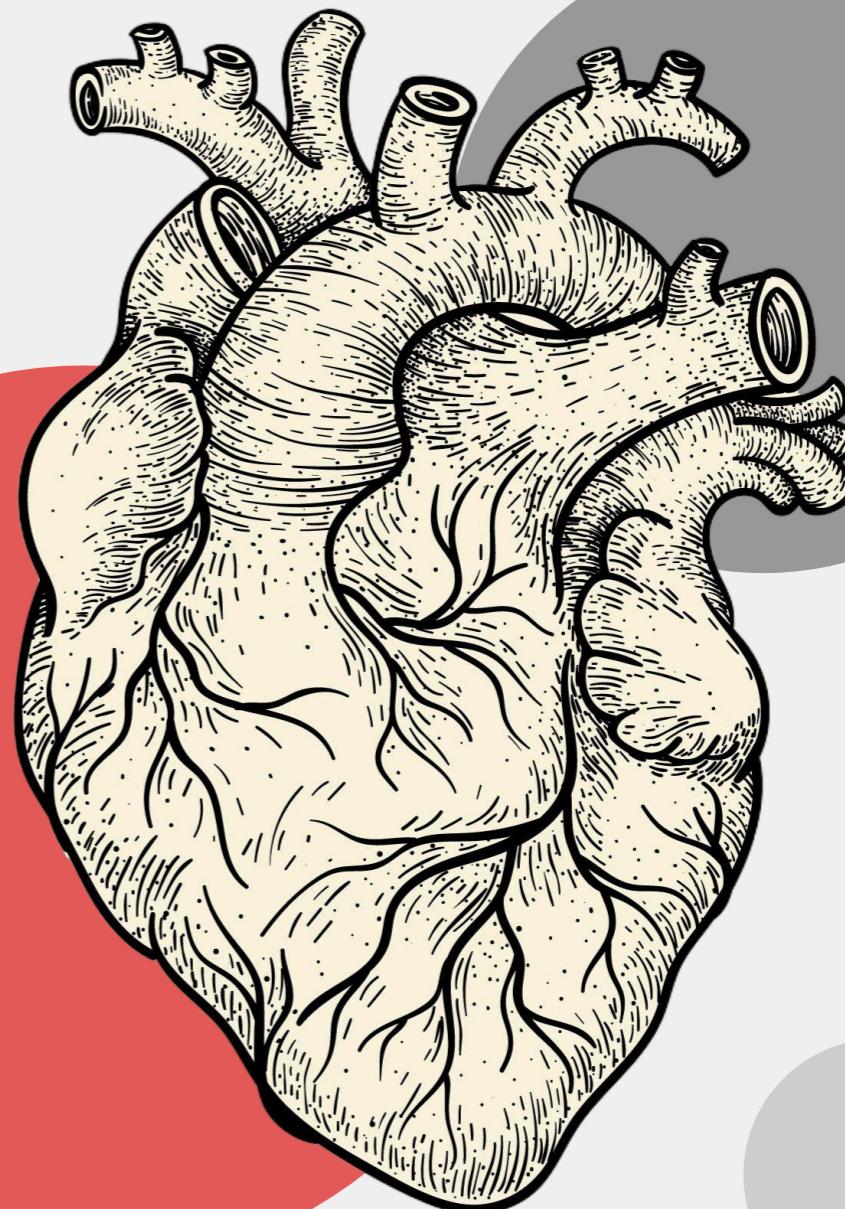
Table des matières

- 01 Introduction**
- 02 Objectif**
- 03 Méthodologie**
- 04 Résultats**
- 05 Conclusion**



01

Introduction



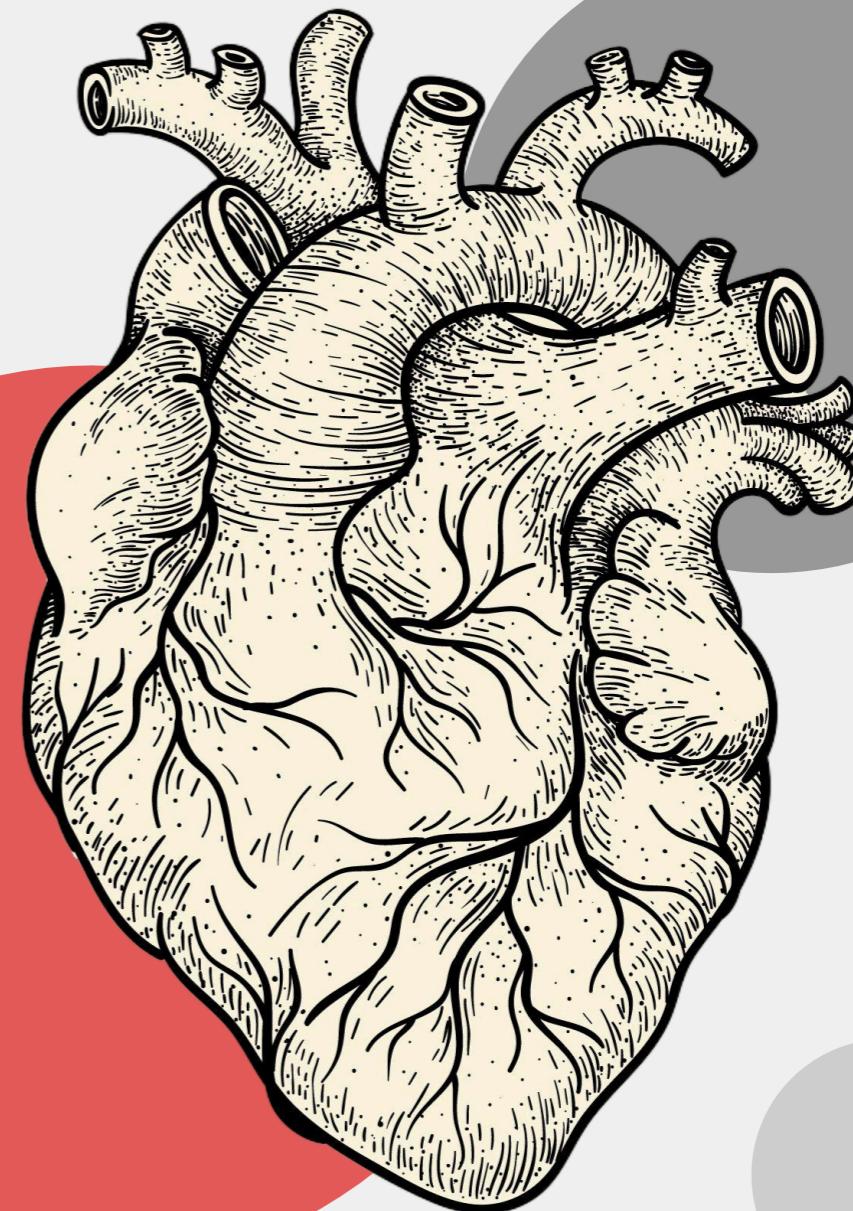
Introduction

Les maladies cardiaques sont la principale cause de décès dans le monde, représentant environ 31% de tous les décès mondiaux. La détection précoce et la gestion efficace de ces maladies sont essentielles pour améliorer les résultats pour les patients.



02

Objective



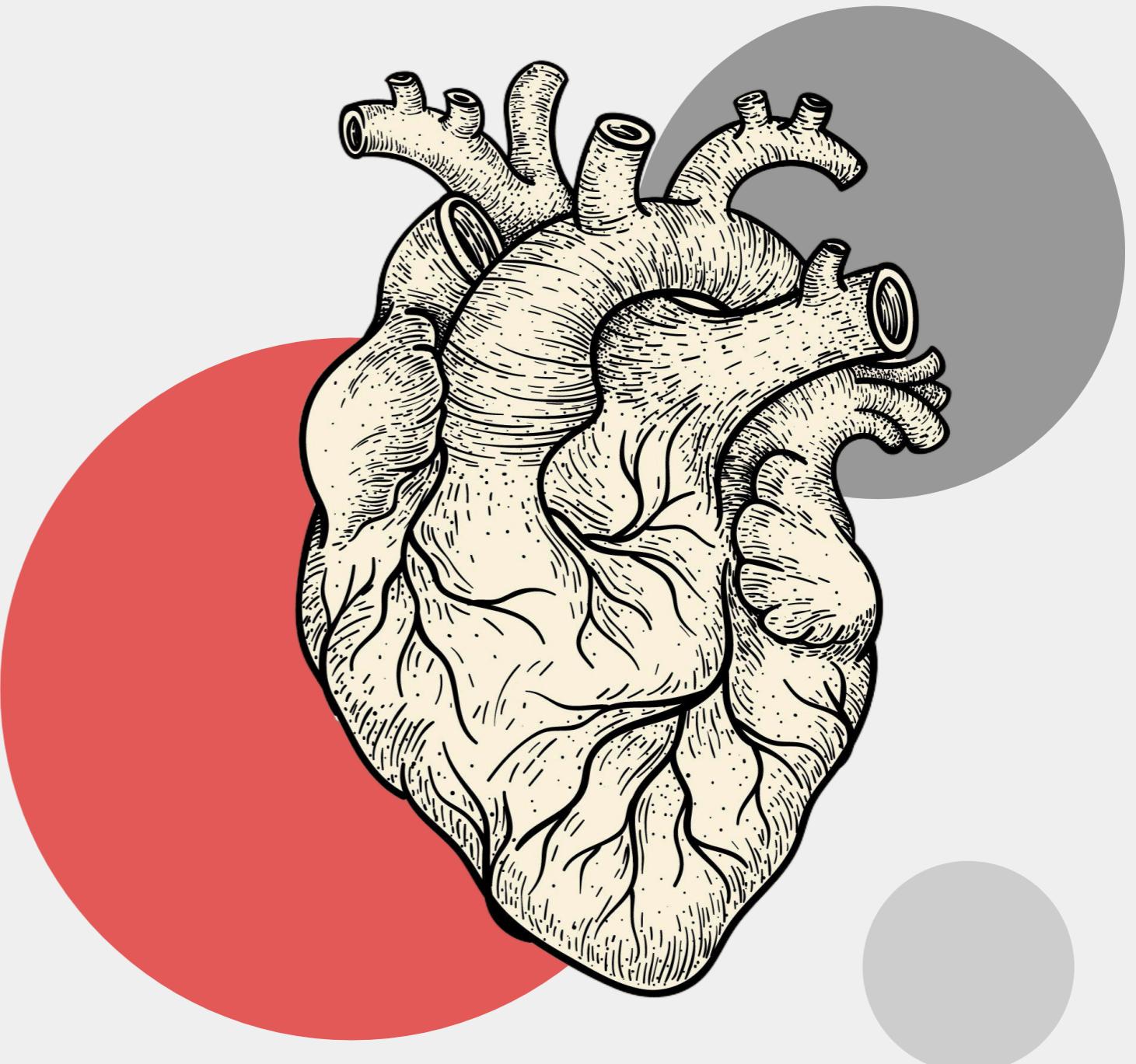
Objective

Développer un modèle d'intelligence artificielle pour détecter les maladies cardiaques en utilisant des algorithmes de machine learning. L'étude se concentre sur les défis et les limites de l'apprentissage automatique dans le domaine médical, ainsi que sur la qualité des données.



03

Méthodologie



Algorithmes utilisés



1. K-Nearst Neighbors (KNN)

KNN, ou k-nearest neighbors(k plus proches voisins), est un algorithme d'apprentissage supervisé utilisé pour la classification et la régression. C'est l'un des algorithmes les plus simples et les plus intuitifs utilisés en apprentissage automatique.



Plus précisément :



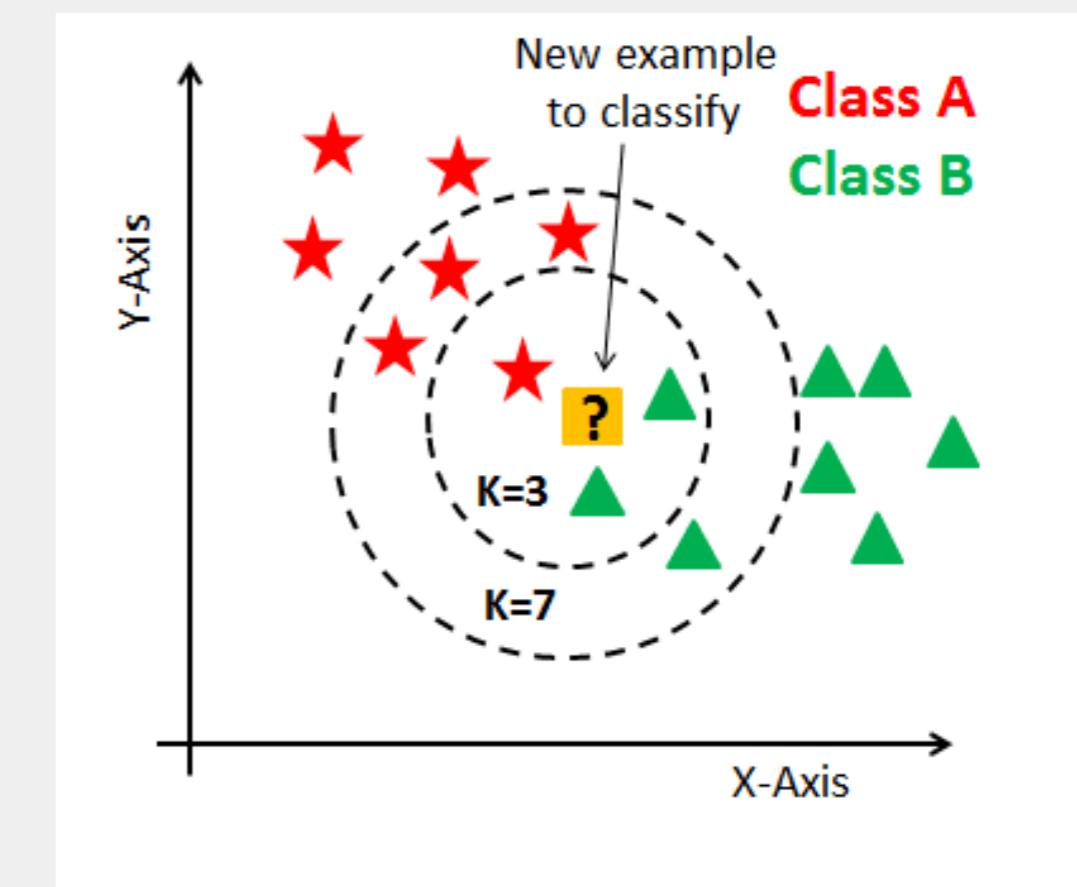
Apprentissage

KNN stocke simplement les exemples et leurs étiquettes de classe dans la mémoire.



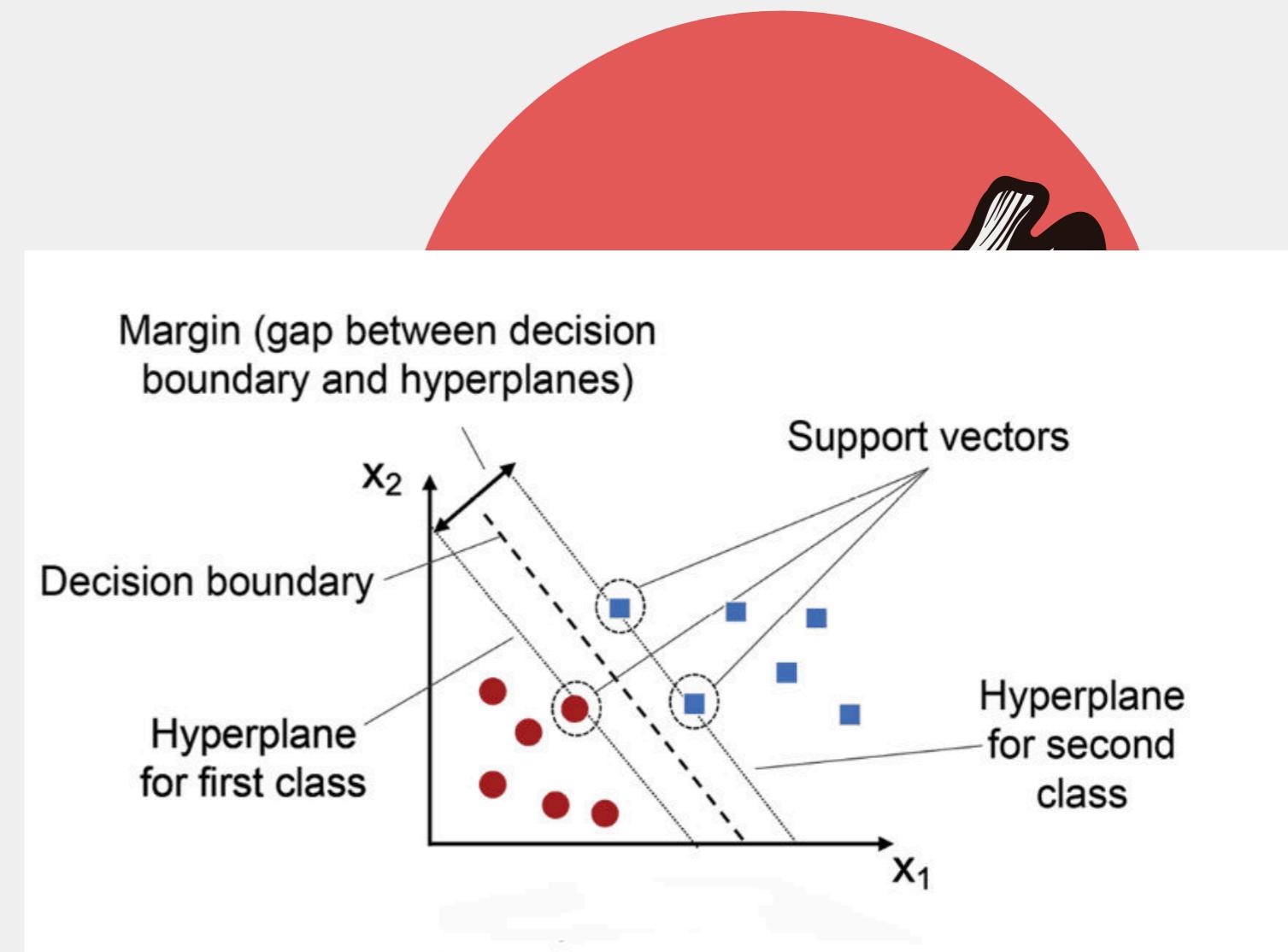
Prédiction pour un nouvel exemple

Calcul de la distance entre le nouvel exemple et tous les exemples dans le jeu de données, suivi de la sélection des k exemples les plus proches en fonction de cette distance. Ensuite, un vote majoritaire est effectué parmi les étiquettes de classe des k voisins pour attribuer cette classe au nouvel exemple.



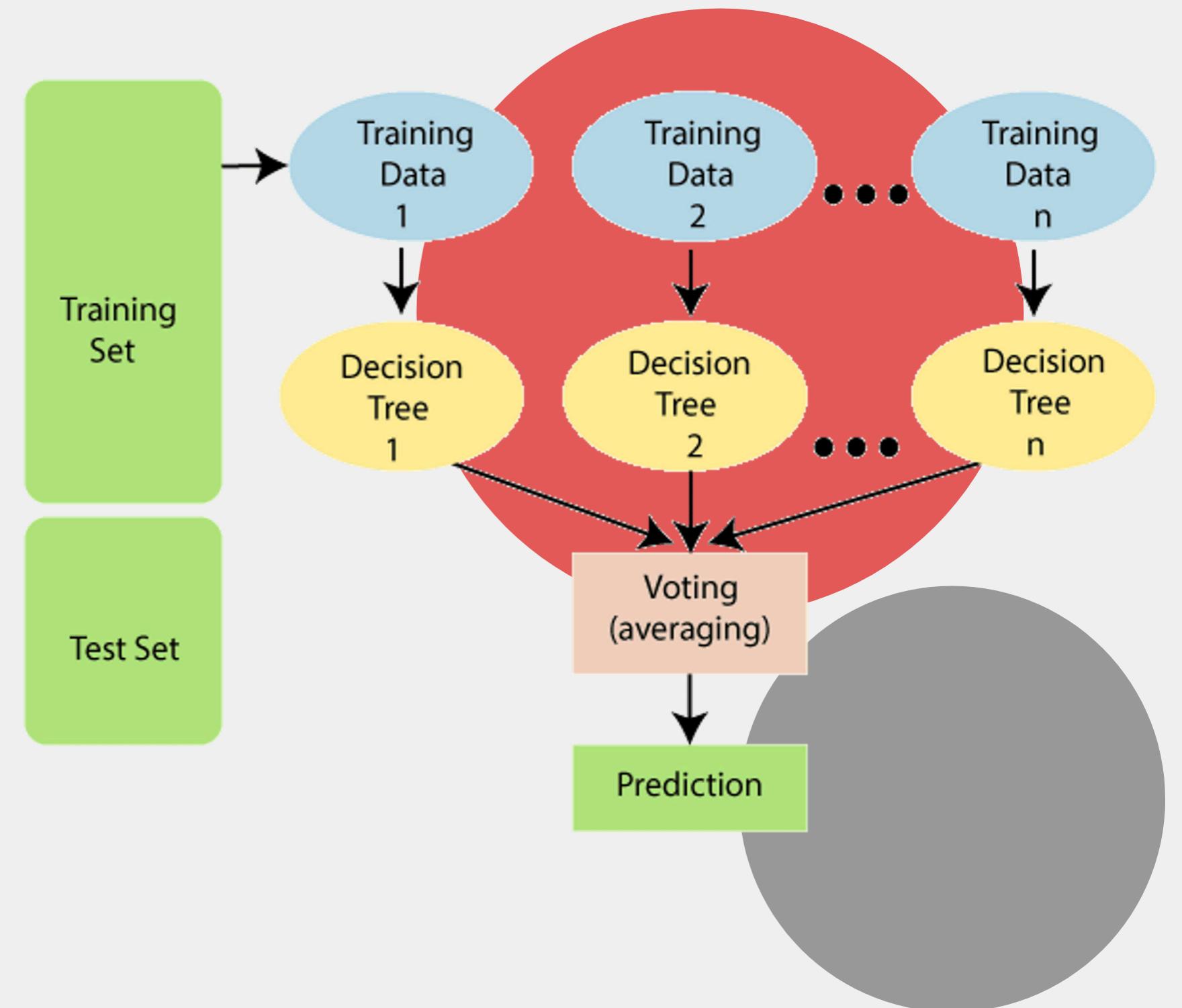
2. Support Vector Machine (SVM)

La Machine à Vecteurs de Support (SVM) est un algorithme d'apprentissage automatique supervisé utilisé pour les tâches de classification et de régression. Elle vise à trouver un hyperplan ou une ligne optimal(e) dans un espace à N dimensions qui maximise la distance entre les différentes classes.



3. Random Forest

Random Forest est un ensemble d'arbres de décision qui combine les prédictions de plusieurs arbres pour améliorer la précision et réduire le surajustement.



Principe de base des forêts aléatoires :



Ensemble d'arbres de décision :

Une forêt aléatoire construit plusieurs arbres de décision sur différents sous-échantillons du jeu de données et utilise la moyenne ou la majorité des votes pour améliorer la précision prédictive et contrôler l'overfitting.



Bagging (Bootstrap Aggregating) :

Chaque arbre est entraîné sur un échantillon bootstrap différent du jeu de données. Un échantillon bootstrap est un sous-ensemble aléatoire du jeu de données d'entraînement sélectionné avec remplacement.



Sélection aléatoire des caractéristiques :

Lors de la construction de chaque arbre, chaque division ne considère qu'un sous-ensemble aléatoire des caractéristiques. Cela ajoute de la diversité entre les arbres, ce qui améliore la robustesse de l'ensemble.

Validation des résultats



Validation des résultats :

La validation des résultats est une étape cruciale dans le développement de modèles d'apprentissage automatique. Pour évaluer la performance du modèle, plusieurs mesures statistiques sont utilisées, notamment :



Mesures Statistiques



Matrice de confusion

La matrice de confusion est un outil essentiel pour évaluer les performances d'un modèle de classification.

Rappel

Le rappel (recall en anglais) est une métrique importante pour évaluer les performances d'un modèle de classification.



Précision

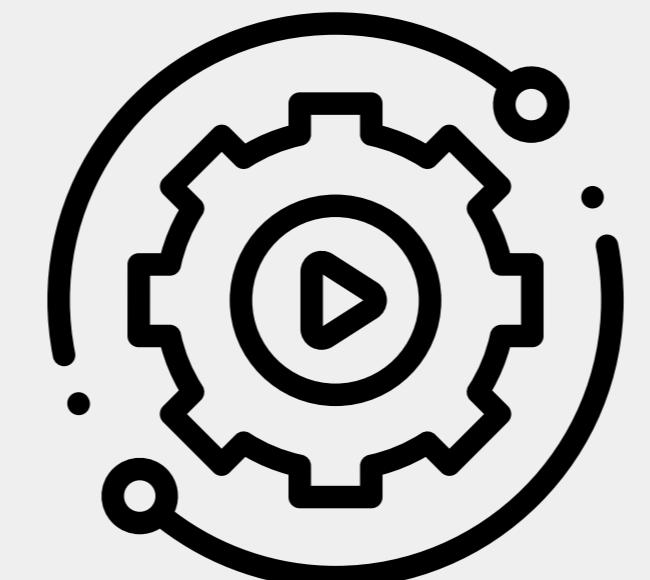
La précision est une mesure de la pertinence des prédictions positives du modèle.

F1 Score

Le score F1 est une mesure de performance qui combine la précision et le rappel d'un modèle de classification.



Simulation setup





Description de la base de données

La base de données utilisée dans cette étude est un ensemble de données publiques appelé "Heart Disease UCI" disponible sur le site web de l'UCI Machine Learning Repository. Elle contient des informations sur 303 patients, avec 14 variables pour chaque patient.

Les variables de la base de données :

- age : Âge du patient (en années)
- sex : Sexe du patient (1 = homme, 0 = femme)
- cp : Douleur thoracique (4 valeurs possibles)
- trestbps : Pression artérielle au repos (en mmHg)
- chol : Taux de cholestérol sérique (en mg/dl)
- fbs : Glycémie à jeun (1 = vrai, 0 = faux)
- restecg : Résultats de l'électrocardiogramme de repos (3 valeurs possibles)
- thalach : Fréquence cardiaque maximale atteinte lors de l'exercice (en battements par minute)
- exang : Douleur thoracique induite par l'exercice (1 = oui, 0 = non)
- oldpeak : Dépression ST induite par l'exercice par rapport au repos (en mm)
- slope : Pente de la dépression ST (3 valeurs possibles)
- ca : Nombre de vaisseaux sanguins principaux colorés par fluoroscopie (0-3)
- Thal : Résultats du test thallium (3 valeurs possibles)
- target : Présence ou absence de maladie cardiaque (1 = présence, 0 = absence)

Prétraitement des données



Importation des données



Transformation de variables catégorielles en variables numériques



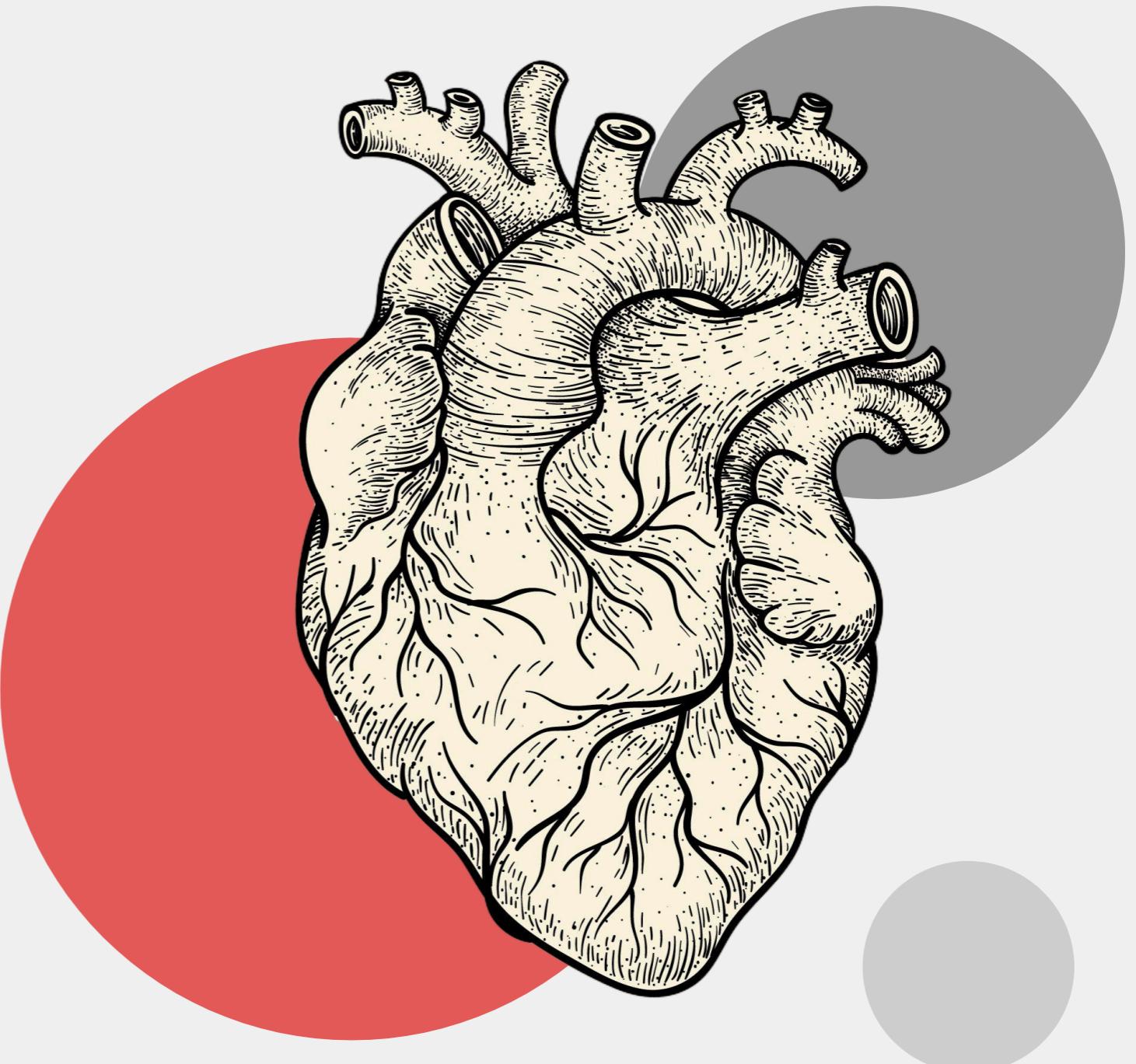
Normalisation des variables numériques



Séparation des données en ensembles d'entraînement et de test

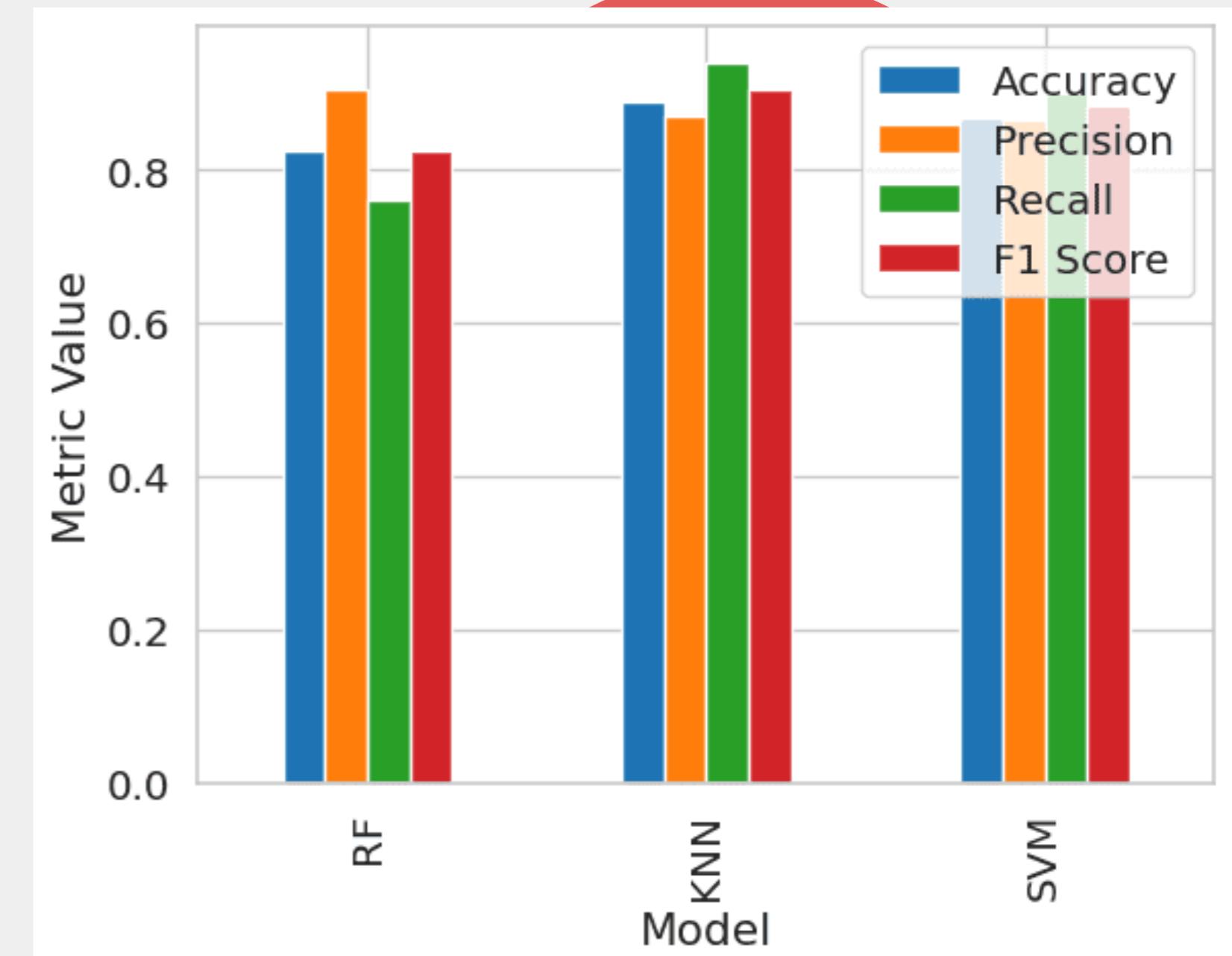
04

Resultats



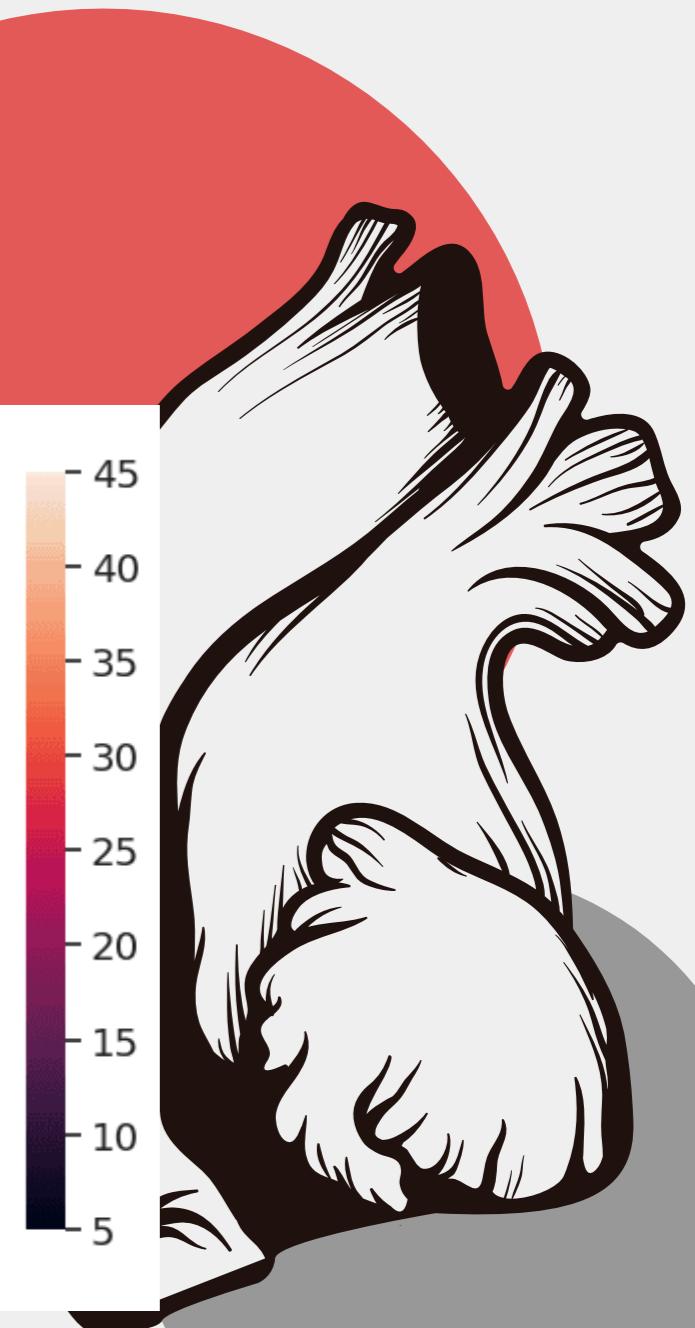
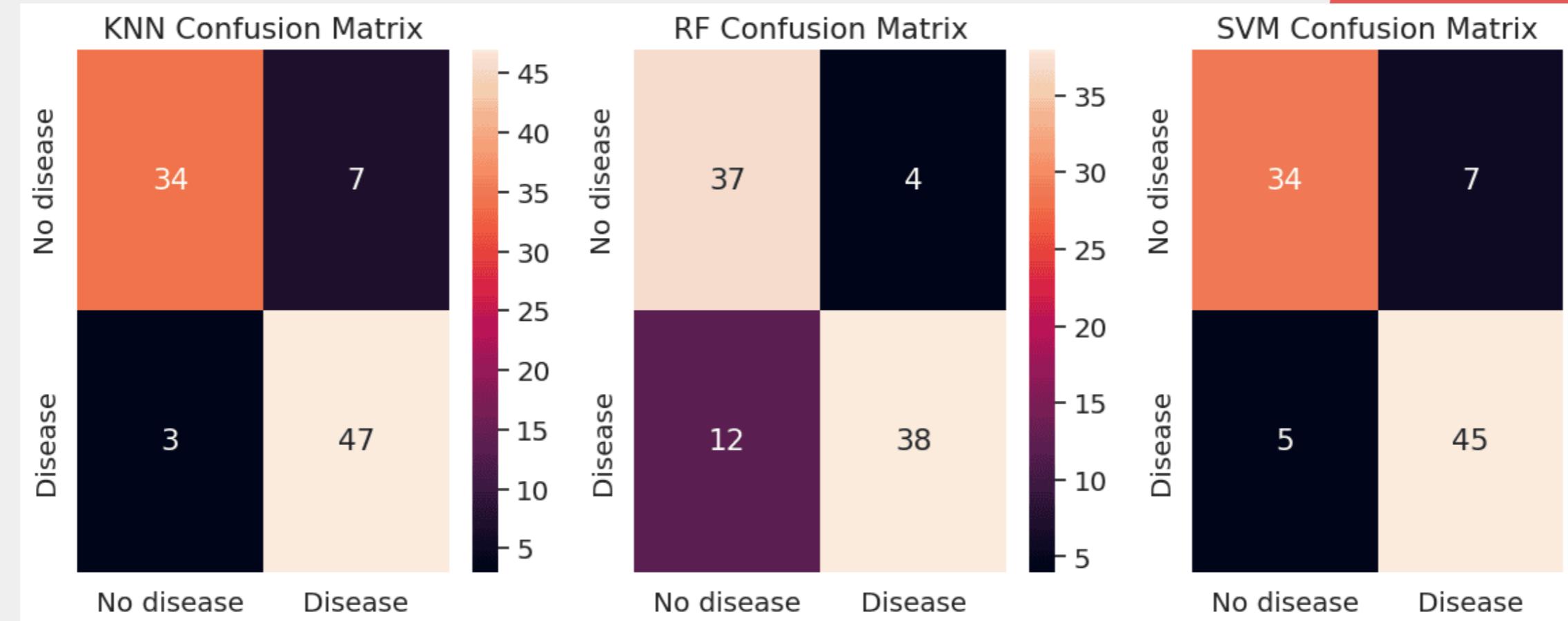
Résultats obtenus avec notre modèle

Les résultats que nous avons obtenus avec notre modèle sont très encourageants. Le modèle K-voisins le plus proche (KNN) a atteint une précision (Accuracy) de 89 %, ce qui est supérieur aux autres modèles. Le modèle Random Forest a obtenu une précision de 84 %, tandis que le modèle Support Vector Machine (SVM) a obtenu une précision de 86 %.



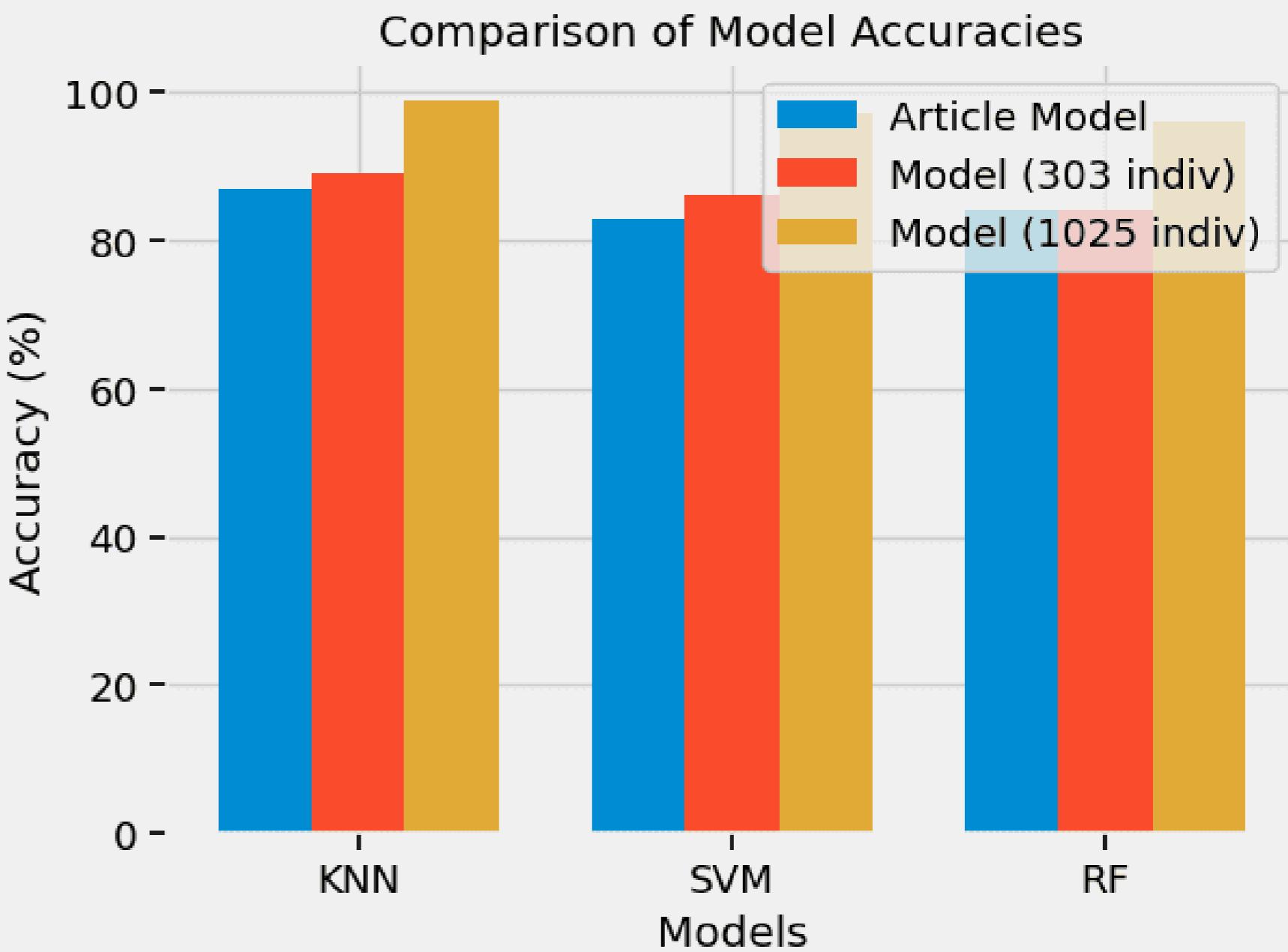
Résultats obtenus avec notre modèle

Matrice de confusion

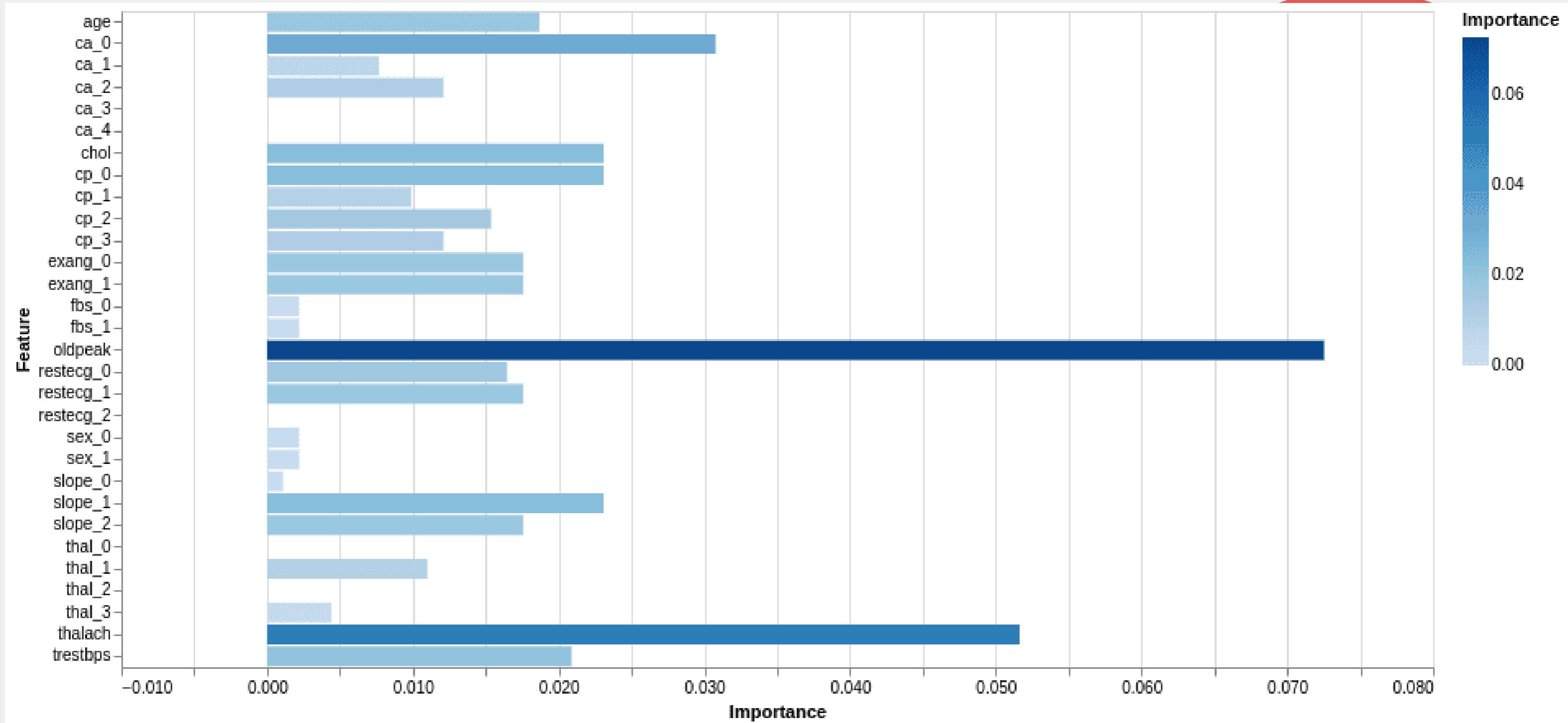


Comparaison avec les résultats de l'article

Les performances de notre modèle sur les deux jeux de données surpassent celles présentées dans l'article.

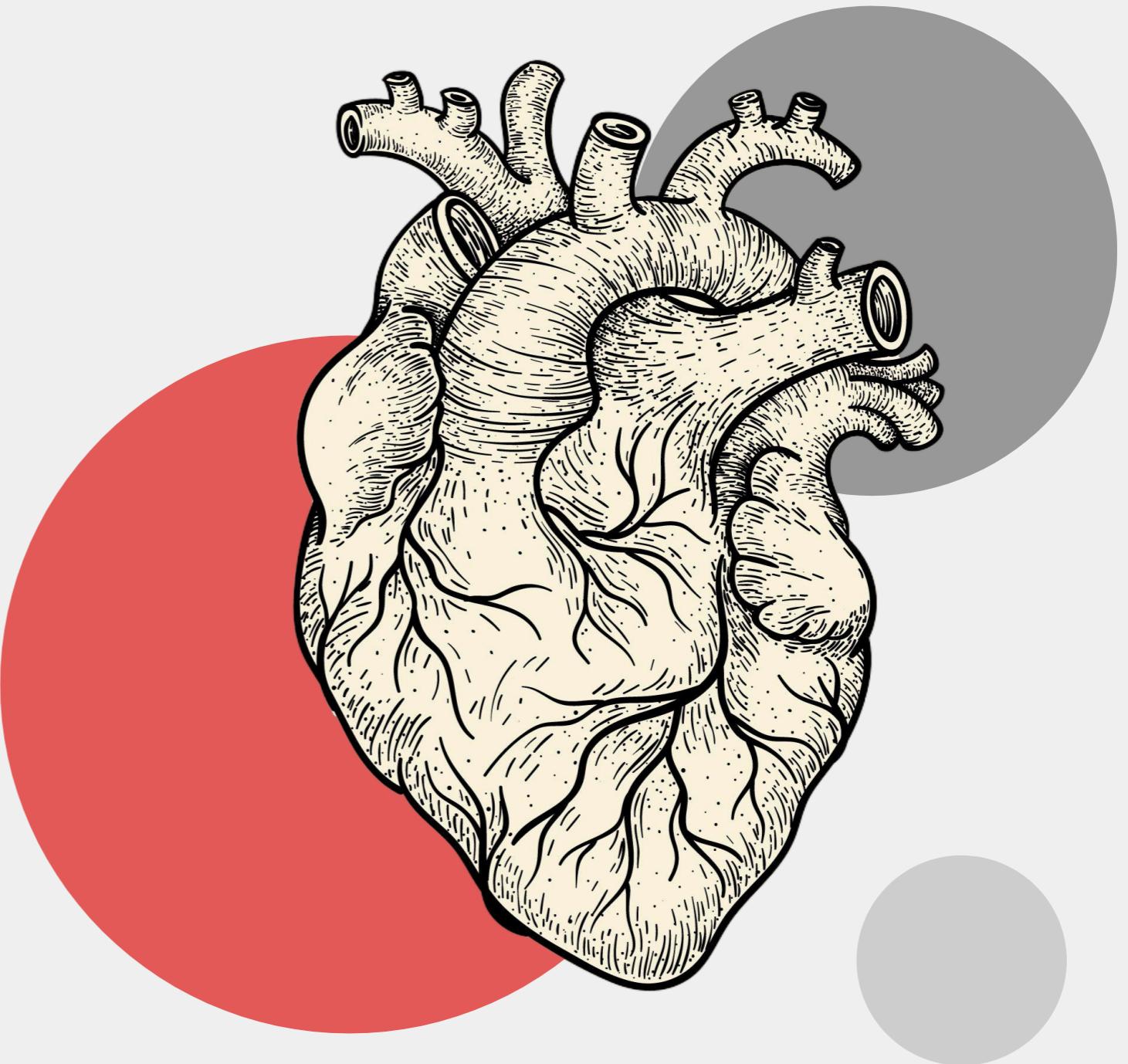


Principales caractéristiques de la prédition des maladies cardiaques



05

Conclusion



Implications et Contributions de l'Étude

- Identification des patients à risque élevé
- Identifiant les facteurs de risque et les biomarqueurs associés aux maladies cardiaques.



Limitation

- Taille de l'échantillon limitée
- Nécessité de validations supplémentaires



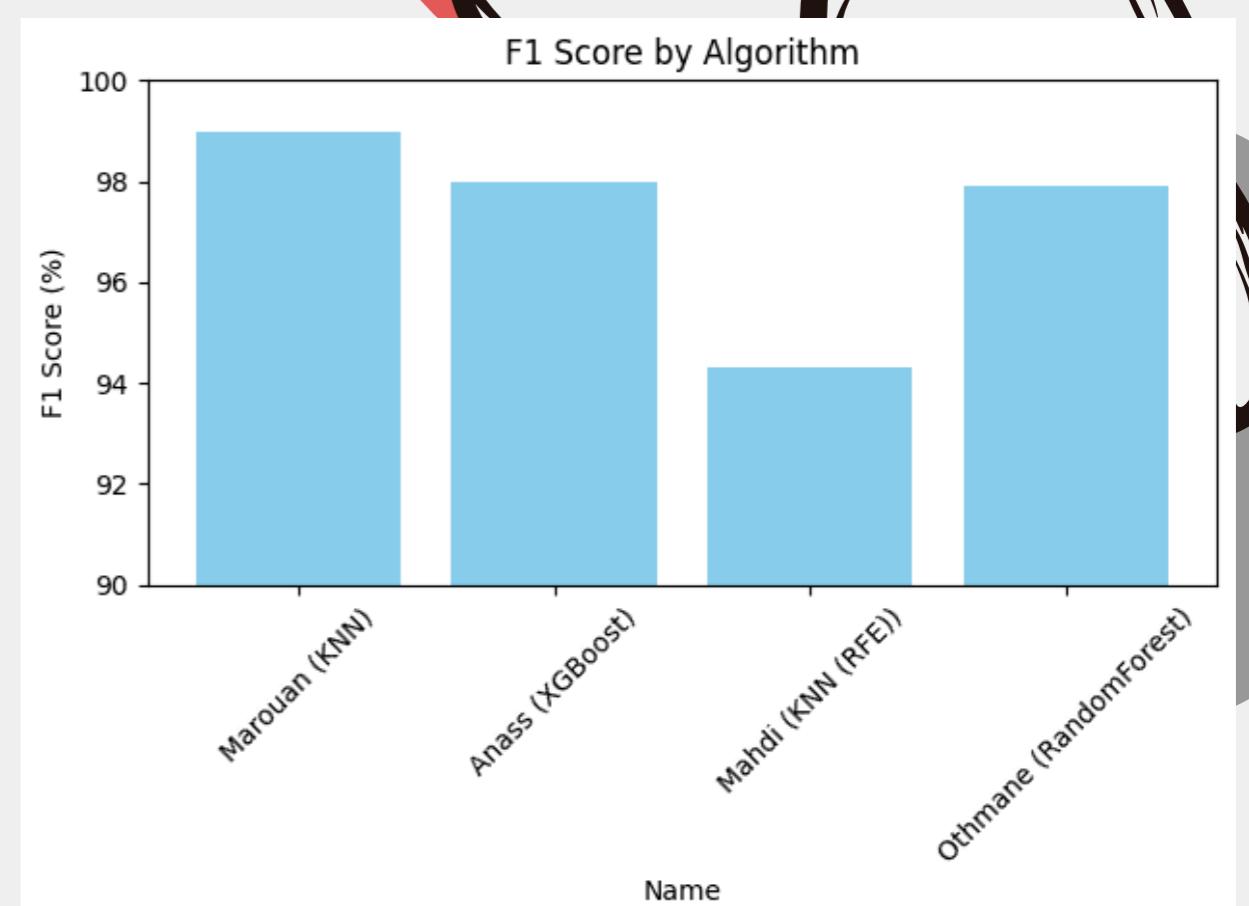
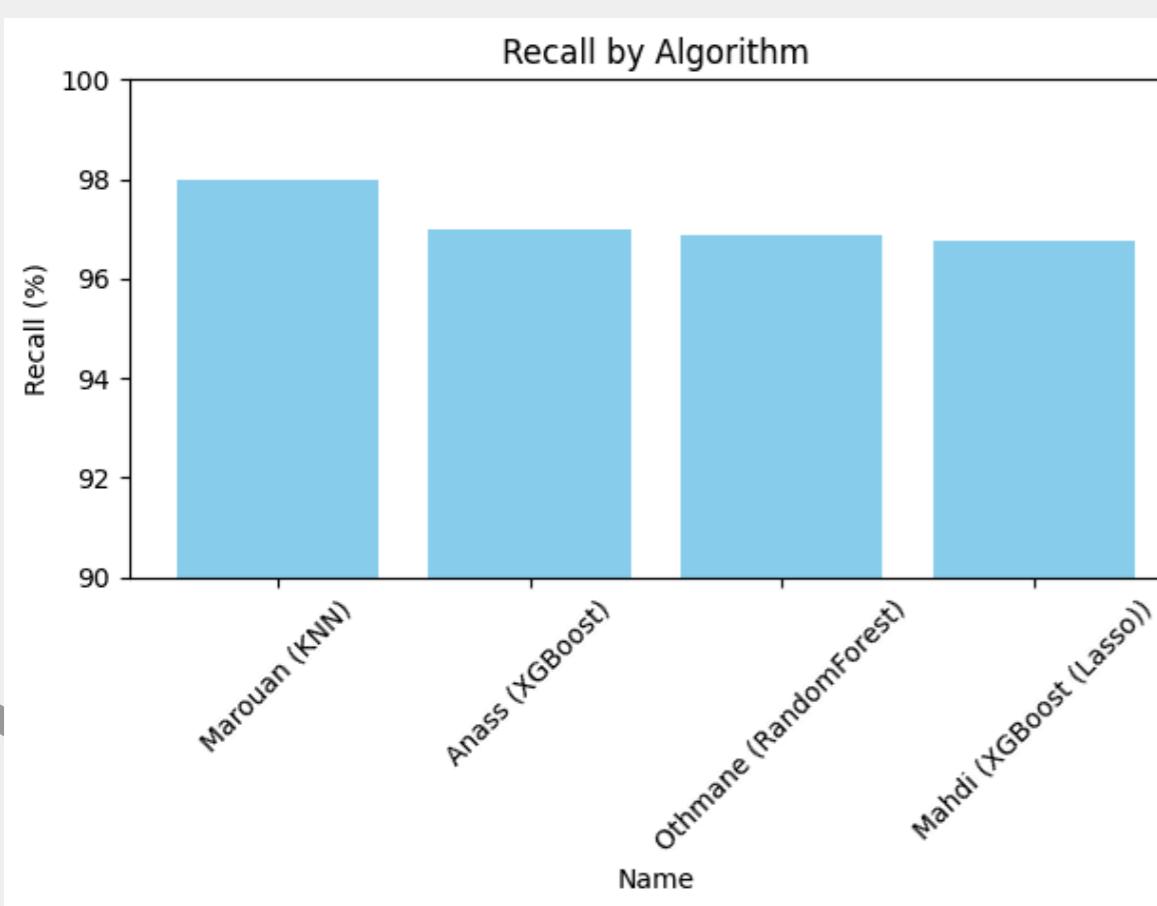
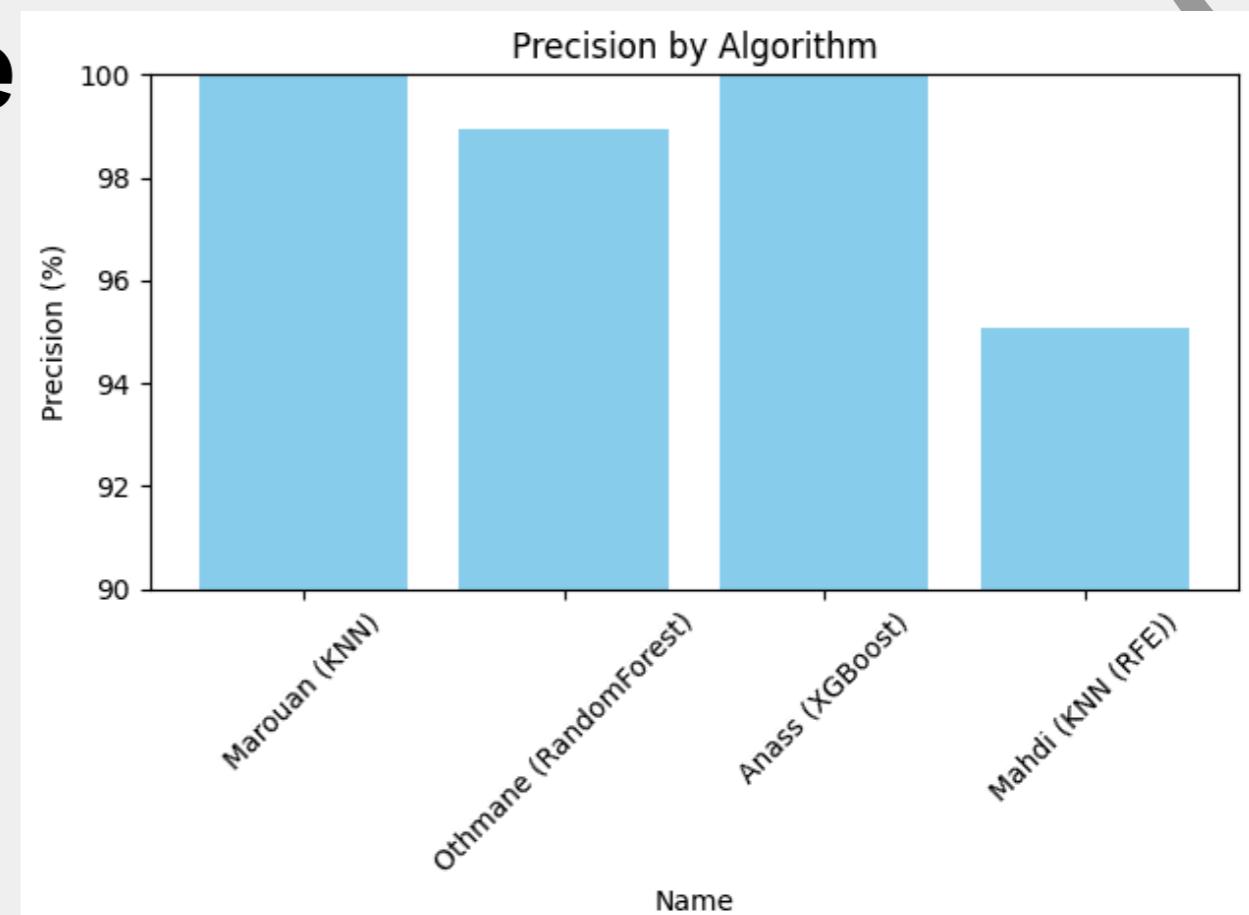
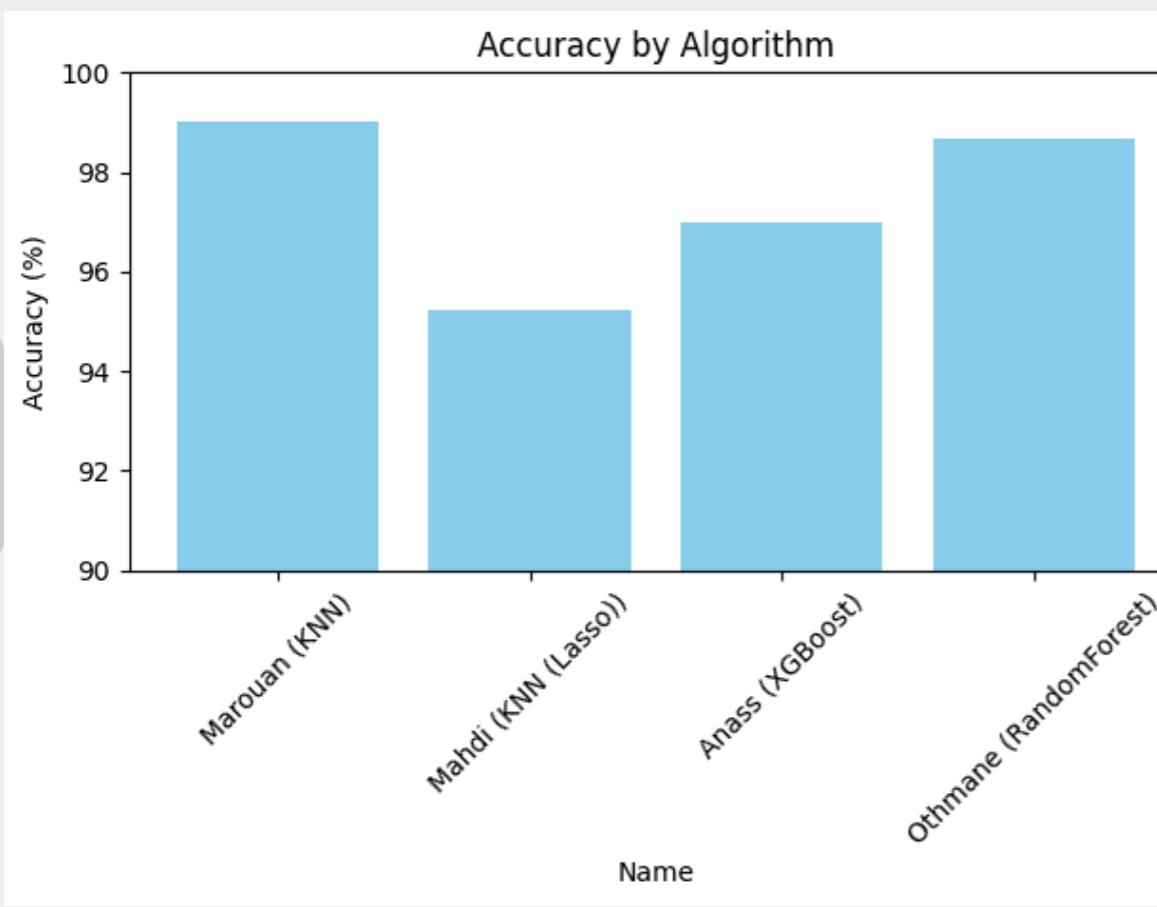
Conclusion Générale



Comparaison Globale des Approches sur Même Sujet



Comparaison des Performances des Algorithmes par Métrique



Comparaison des Performances des Algorithmes par Métrique

Table of Values

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



**MERCI
Pour Votre Attention**

