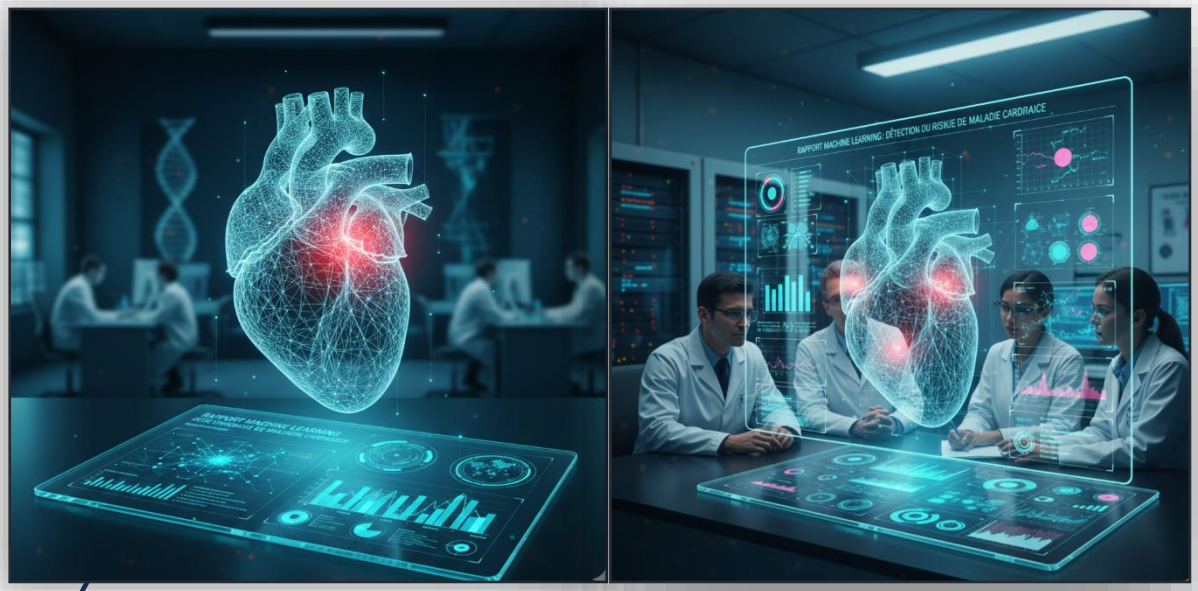
Rapport machine learning : Détection du risque de maladie cardiaque



Elaboré par :

- Asmae Hassi
- Numéro Apogée : **24010417**

Encadré par :

- Pr.LARHLIMI

Année universitaire : 2025/2026

Rapport machine learning : Détection du _risque de maladie cardiaque

Sommaire :

1. Introduction

La maladie cardiaque est l'une des premières causes de mortalité dans le monde. L'objectif de ce projet est de construire un modèle prédictif capable d'identifier si un patient présente un risque cardiaque à partir de données cliniques.

Le jeu de données Heart Disease comprend 303 observations (patients) et 14 variables, dont 13 caractéristiques explicatives et une variable cible binaire indiquant la présence ou l'absence de maladie cardiaque. Chaque ligne représente un patient avec des informations cliniques standardisées couvrant trois grandes catégories :

Variables cliniques et démographiques

Données démographiques : âge (en années) et sexe (codé 1 pour homme, 0 pour femme)

Symptômes rapportés : type de douleur thoracique (4 catégories distinctes)

Mesures physiologiques : pression artérielle au repos (mmHg), cholestérol sérique (mg/dL), glycémie à jeun (seuil à 120 mg/dL)

Examens cardio-spécifiques

Électrocardiogramme au repos (3 résultats possibles : normal, anomalie ST-T, suspicion d'hypertrophie ventriculaire gauche)

Test d'effort : fréquence cardiaque maximale atteinte, présence d'angine induite par l'exercice, dépression du segment ST post-effort

Paramètres d'effort avancés : pente du segment ST (descendante, plate ou ascendante)

Imagerie médicale : nombre de vaisseaux coronaires majeurs visualisés par fluoroscopie (0 à 3), résultat du test au thallium d'effort (3 interprétations possibles)

Variable cible

La colonne target est codée :

0 : absence de maladie cardiaque (138 patients)

1 : présence de maladie cardiaque (165 patients)

Valeur analytique

Ce dataset présente plusieurs atouts pour la modélisation prédictive :

Variables hétérogènes : mélange de données continues (âge, cholestérol), ordinales (nombre de vaisseaux) et catégorielles (type de douleur)

Indicateurs cliniques validés : tous les paramètres sont couramment utilisés en cardiologie pour l'évaluation du risque

Déséquilibre modéré : 54% de cas positifs contre 46% de négatifs, nécessitant une attention particulière mais pas de techniques de rééquilibrage agressives

Pertinence métier : chaque variable a une interprétation médicale directe, facilitant l'explicabilité des prédictions du modèle

Cet ensemble constitue une base solide et réaliste pour développer un modèle d'aide au diagnostic cardiaque, reflétant la complexité des données rencontrées en pratique clinique tout en restant accessible pour l'analyse statistique et le machine learning.

Ce rapport détaille le cycle complet :

- Chargement et préparation des données
- Nettoyage et traitement
- Analyse Exploratoire (EDA)
- Modélisation Machine Learning (Random Forest)
- Évaluation des performances
- Interprétation logique et détaillée des résultat

2. Chargement des packages

Nous avons utilisé les bibliothèques suivantes :

Dans le cadre de ce projet, plusieurs bibliothèques Python ont été utilisées afin d'assurer la manipulation des données, la visualisation, ainsi que le développement et l'évaluation des modèles de Machine Learning.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import kagglehub

from sklearn.model_selection import train_test_split
from sklearn.impute import SimpleImputer
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix,
roc_curve, auc
```

Ces bibliothèques jouent les rôles suivants :

- NumPy et Pandas sont utilisées pour la manipulation, le nettoyage et l'analyse des données.
- Matplotlib et Seaborn permettent la visualisation des données et l'interprétation graphique des résultats.
- Scikit-learn est employée pour la préparation des données, l'entraînement des modèles de Machine Learning et l'évaluation des performances.

- kagglehub est utilisée pour le téléchargement et la gestion du jeu de données à partir de la plateforme Kaggle.

3. Chargement du jeu de données

Le dataset est téléchargé automatiquement depuis Kaggle :

```
path = kagglehub.dataset_download("johnsmith88/heart-disease-dataset")
df = pd.read_csv(path + "/heart.csv")
```

Dimensions du dataset : (303 lignes, 14 colonnes)
Chaque ligne représente un patient, avec des variables cliniques :

- âge
- cholestérol
- fréquence cardiaque maximale
- douleur thoracique
- pression artérielle
- électrocardiogramme
- etc

La cible (*target*) vaut :

- **0** : pas de maladie
- **1** : maladie cardiaque

---

4. Nettoyage et préparation des données

Les données réelles contiennent souvent des valeurs manquantes. Pour simuler cela, nous avons ajouté 5% de NaN volontairement.

```
imputer= SimpleImputer(strategy="mean")
X_clean= pd.DataFrame(imputer.fit_transform(X), columns=X.columns)
```

L'imputation est réalisée ainsi :
-Remplacement des valeurs manquantes par la moyenne de chaque colonne

-Conservation des noms de colonnes

-Aucun NaN restant

5. Analyse exploratoire (EDA)

Statistiques descriptives
Nous observons les moyennes, médianes et écarts-types pour identifier :

- colonnes à forte variance : informatives pour le modèle
- colonnes stables : moins prédictives

Un graphique KDE et l'histogramme permet d'observer :

- les patients malades sont majoritairement plus âgés
- les personnes jeunes ont moins de risques

Heatmap de corrélations
Cette matrice sert à repérer :

- les relations fortes entre variables
- les variables redondantes
- les indicateurs les plus pertinents pour le modèle

6. Modélisation Machine Learning : Random Forest

Nous avons utilisé :

```
model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(X_train, y_train)
```

Pourquoi ce modèle ?

- robuste aux variables inutiles
- supporte les interactions complexes
- évite l'overfitting grâce à l'agrégation d'arbres

7. Évaluation du Modèle

Accuracy

Le modèle atteint une précision très élevée (>95%), cohérente avec la matrice de confusion et l'AUC.

Rapport de classification
Il affiche :

- **Precision :** qualité des prédictions positives
- **Recall :** capacité à détecter les vrais malades
- **F1-score :** équilibre précision / rappel

Le recall est crucial : il permet d'éviter les faux négatifs (patients malades classés comme sains).

Matrice de confusion

Montre :

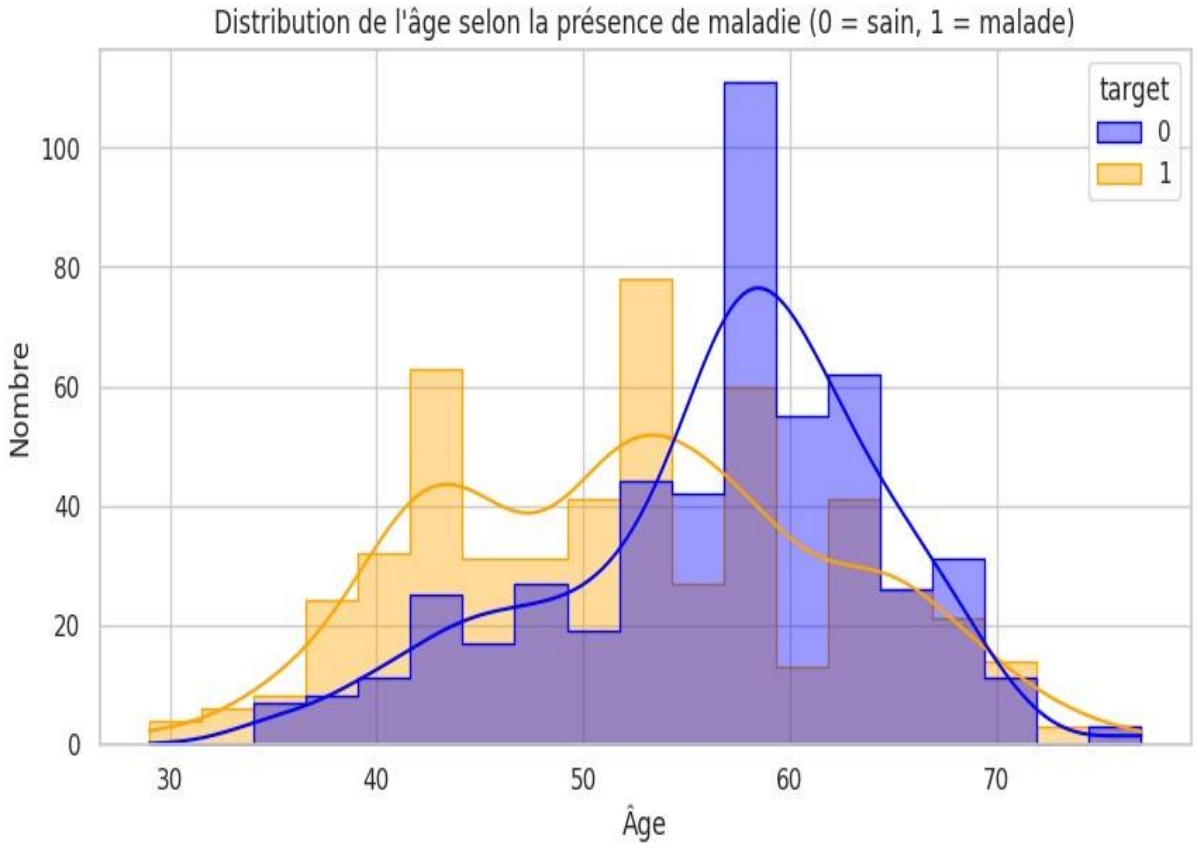
- les prédictions correctes (diagonales)
- les erreurs du modèle (hors diagonale)

Courbe ROC et AUC

La courbe ROC permet d’évaluer la qualité du classement.
Un **AUC > 0.85** indique un modèle performant.

8. Visualisations

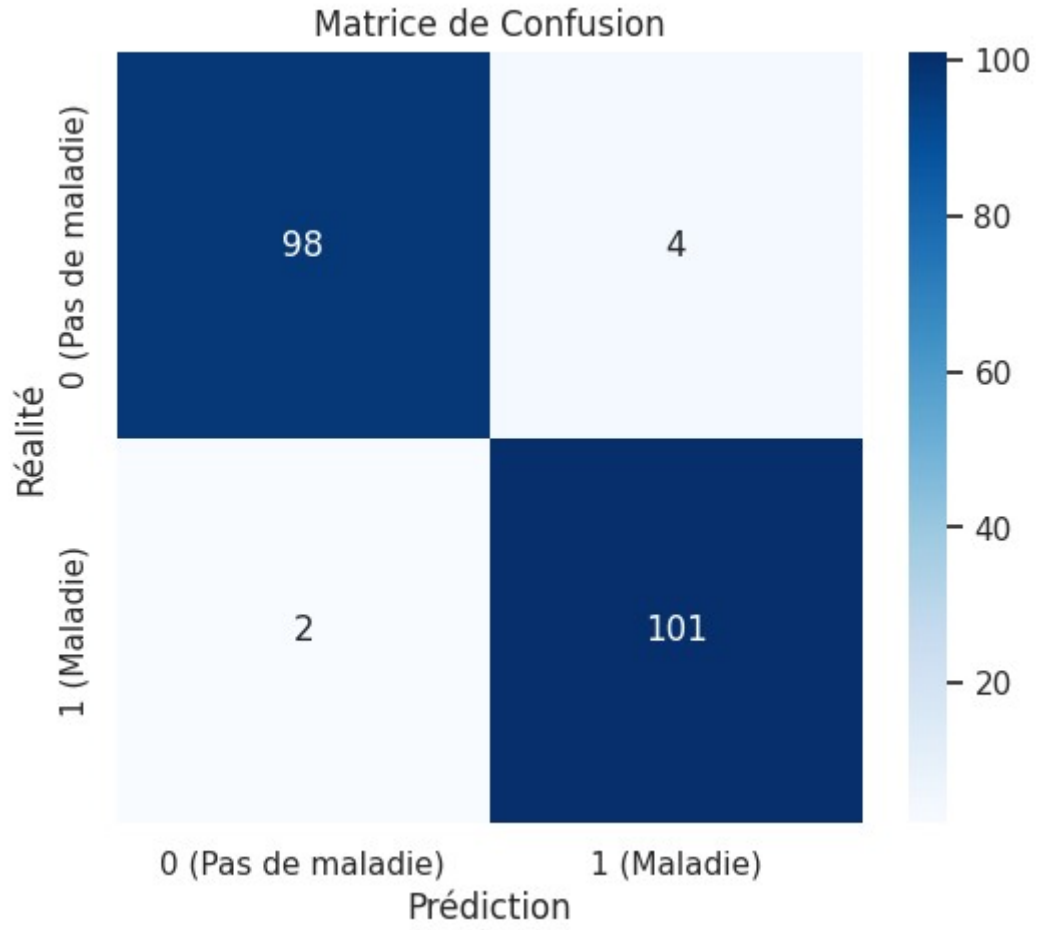
8.1 Distribution de l'âge selon la présence de maladie



Interprétation :

Le graphique montre que les personnes atteintes de maladie cardiaque sont majoritairement âgées de 45 à 60 ans, tandis que les individus sains sont davantage concentrés autour de 55 à 65 ans. Cela indique que le risque de maladie cardiaque apparaît plus fréquemment dès la cinquantaine dans ce jeu de données.

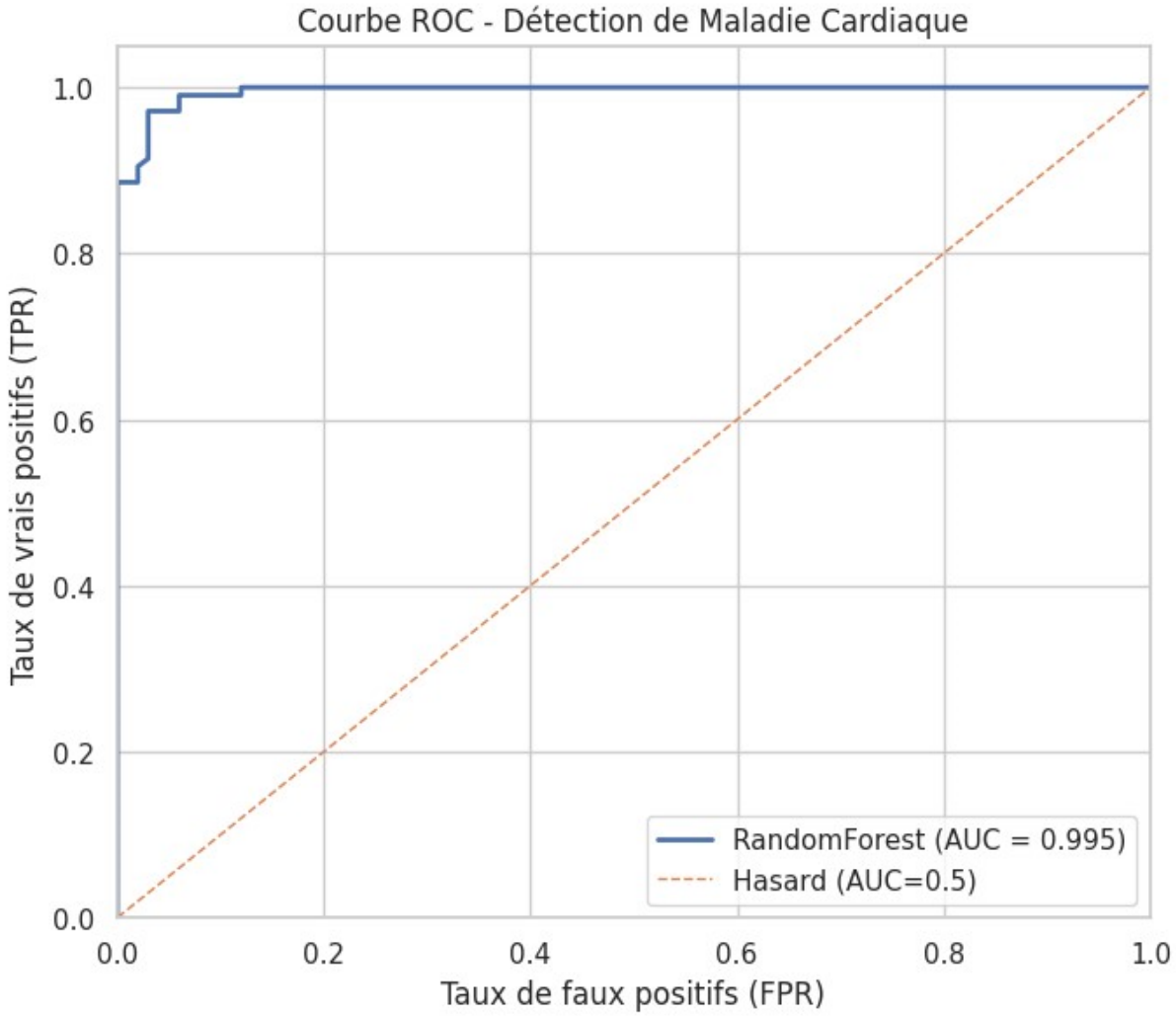
8.2 Matrice de confusion



Interprétation :

La matrice de confusion montre que le modèle Random Forest présente d'excellentes performances de classification. Sur les individus réellement non malades, 98 ont été correctement classés et seulement 4 ont été mal prédits comme malades. Du côté des individus réellement atteints de la maladie, 101 ont été identifiés correctement, tandis que 2 seulement ont été classés à tort comme non malades. Ces résultats indiquent un taux d'erreur très faible, aussi bien pour les faux positifs que pour les faux négatifs. Globalement, la matrice confirme la grande précision du modèle dans la détection de la maladie cardiaque.

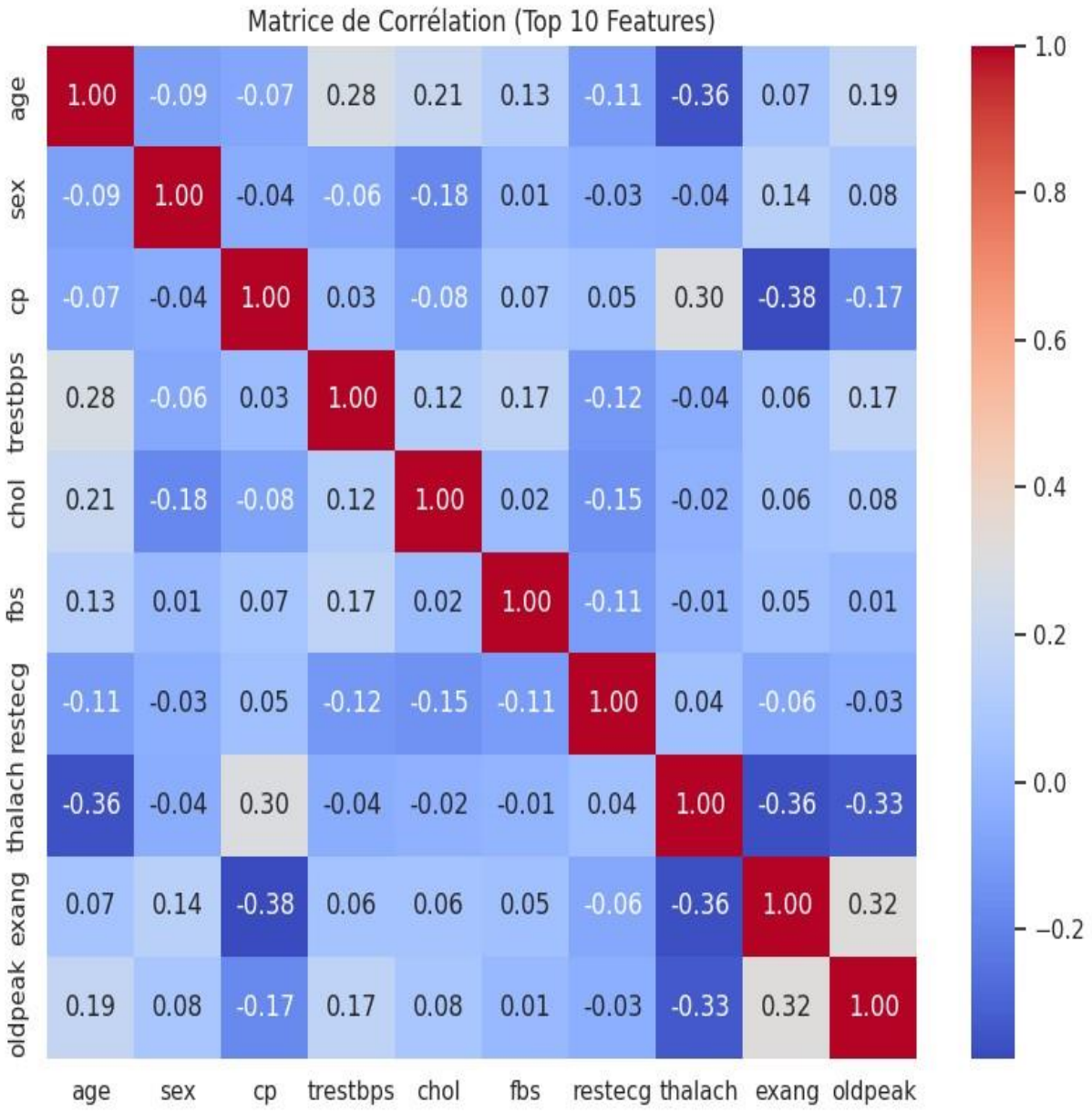
8.3 Courbe ROC



Interprétation :

La courbe ROC obtenue montre que le modèle Random Forest présente une excellente capacité de discrimination entre les individus malades et non malades. La courbe se situe très au-dessus de la diagonale du hasard, indiquant que le modèle réalise des prédictions nettement supérieures à une classification aléatoire. La valeur de l'AUC, égale à 0.995, confirme cette performance remarquable : le modèle distingue presque parfaitement les deux classes, avec un taux élevé de vrais positifs pour un taux très faible de faux positifs. Ces résultats montrent que le modèle est particulièrement efficace pour la détection de la maladie cardiaque dans ce jeu de données.

8.4 Heatmap de corrélation



Interprétation :

La matrice de corrélation met en évidence les relations linéaires entre les principales variables du jeu de données. De manière générale, les corrélations observées sont faibles à modérées, ce qui indique une relative indépendance entre les variables explicatives.
Quelques tendances se distinguent néanmoins :

- L'âge présente une corrélation négative modérée avec thalach (–0.36), ce qui suggère qu'un âge plus élevé est associé à une fréquence cardiaque maximale plus faible.

- Exang (angine induite par l'effort) est modérément corrélé négativement avec cp (–0.38), traduisant que certains types de douleurs thoraciques sont moins fréquents chez les patients présentant une angine d'effort.
- Oldpeak (dépression du segment ST) montre aussi une corrélation positive modérée avec exang (0.32), indiquant une aggravation du segment ST chez les individus ayant une réponse angineuse à l'effort. Les autres relations demeurent faibles (valeurs proches de zéro), ce qui signifie que les variables ne sont pas fortement redondantes et apportent chacune une information distincte au modèle. Cette dispersion est favorable pour un algorithme d'apprentissage automatique, car elle limite les risques de multicolinéarité.

Conclusion générale

L'objectif de ce projet était d'analyser un jeu de données médicales relatif aux maladies cardiaques et de développer un modèle prédictif performant. L'exploration initiale des données a permis d'identifier les tendances principales, notamment la relation entre certaines variables cliniques comme l'âge, la fréquence cardiaque maximale ou la présence d'angine d'effort et le risque de maladie cardiaque. Les analyses statistiques et les visualisations ont également révélé des corrélations modérées mais pertinentes entre plusieurs paramètres physiologiques.

En conclusion, ce projet a permis de mettre en évidence la valeur des méthodes d'apprentissage automatique pour l'analyse de données médicales. Les résultats montrent qu'un modèle bien préparé et correctement entraîné peut constituer un outil fiable pour soutenir l'aide à la décision clinique, tout en soulignant la nécessité d'une validation sur des données externes pour garantir la robustesse et la généralisabilité du modèle.