Rapport de Mini-Projet : Business Intelligence

Analyse et Prédiction des Maladies Cardiaques

Sellam Tahani Asma

20 décembre 2025

Table des matières

1 Introduction

Ce rapport présente les étapes de réalisation du TP portant sur l'analyse de données de santé. L'objectif est de prédire la présence de maladies cardiaques en comparant plusieurs jeux de données et en entraînant des modèles d'apprentissage automatique.

2 Exploration et Sélection des Données

Dans cette phase, deux jeux de données ont été analysés pour déterminer le plus pertinent pour l'étude.

2.1 Tableau Comparatif des Datasets

L'analyse structurelle a donné les résultats suivants :

Dataset	Lignes	Colonnes	Valeurs Manquantes	Doublons
Dataset 1	303	14	0	1
Dataset 2	920	16	1759	0

TABLE 1 – Comparaison des structures de données.

Justification : Le Dataset 1 a été choisi comme dataset principal pour sa taille exploitable et sa structure plus propre (absence de valeurs manquantes massives par rapport au Dataset 2).

3 Prétraitement des Données

Pour préparer les données à la modélisation, les étapes suivantes ont été appliquées :

- **Gestion des doublons :** Suppression des lignes identiques.
- **Imputation :** Remplissage des valeurs manquantes par la *mode* pour les variables catégorielles et la *médiane* pour les variables numériques.
- **Encodage :** Utilisation du `LabelEncoder` pour transformer les variables textuelles en valeurs numériques.
- **Normalisation :** Mise à l'échelle des données avec `StandardScaler` pour uniformiser les ordres de grandeur.

4 Modélisation et Résultats

Trois algorithmes ont été testés. Voici les performances obtenues sur l'ensemble de test :

5 Visualisations et Analyses

5.1 Matrice de Corrélation

La matrice de corrélation permet d'identifier les relations entre les variables d'entrée et la cible.

Modèle	Accuracy	Précision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.8197	0.8203	0.8197	0.8198	0.9267
Random Forest	0.8689	0.8709	0.8689	0.8689	0.9397
XGBoost	0.8361	0.8380	0.8361	0.8362	0.8998

TABLE 2 – Comparaison des performances des modèles.

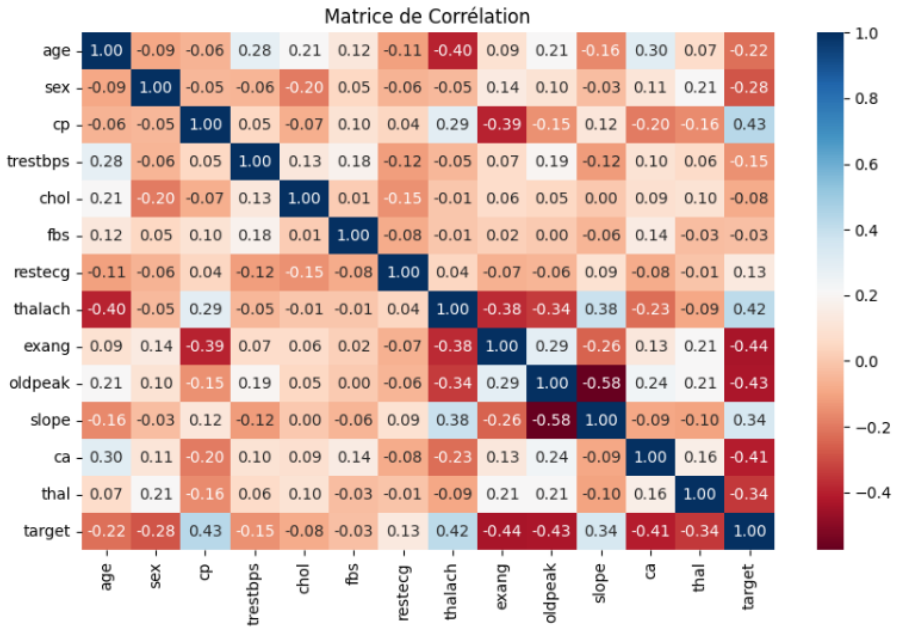

FIGURE 1 – Matrice de corrélation (Heatmap) générée par Seaborn.

5.2 Importance des Variables (Feature Importance)

Le modèle Random Forest a permis d'identifier les variables les plus prédictives.

5.3 Évaluation du Meilleur Modèle

La courbe ROC et la matrice de confusion confirment la robustesse du modèle sélectionné.

6 Déploiement et Accès Distant

La dernière étape du projet consiste à rendre le modèle prédictif accessible via une interface utilisateur interactive et à l'exposer sur le web pour des tests en conditions réelles.

6.1 Interface Utilisateur avec Streamlit

Pour le déploiement, nous avons utilisé **Streamlit**. Cette bibliothèque permet de transformer des scripts Python en applications web interactives. L'application développée permet à un utilisateur de :

— Saisir les paramètres cliniques (âge, cholestérol, type de douleur thoracique, etc.).

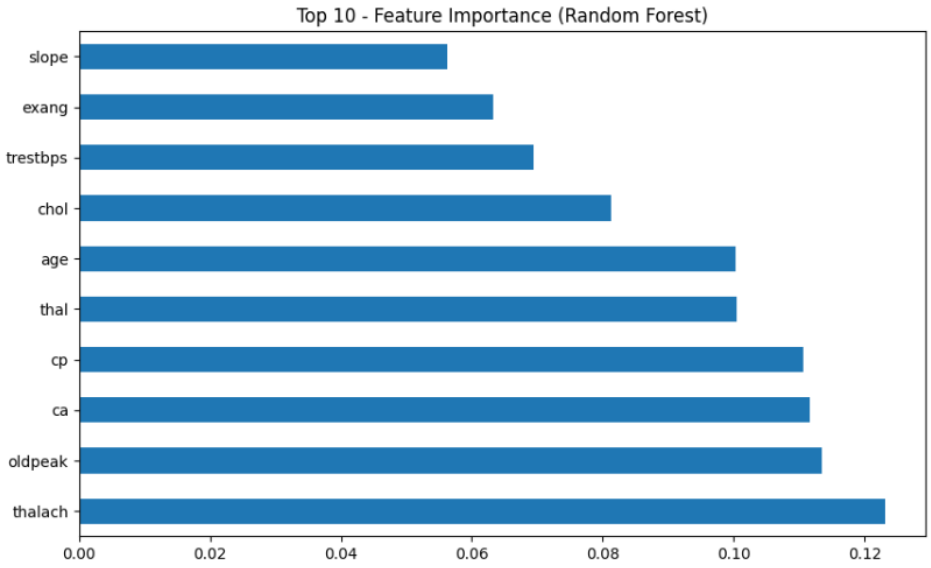

FIGURE 2 – Top 10 des variables les plus importantes.

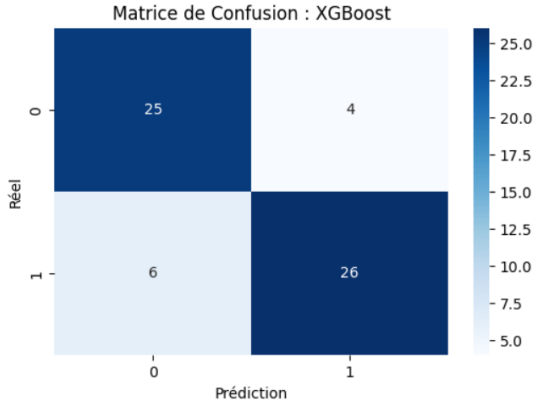

FIGURE 3 – Matrice de Confusion.

— Lancer la prédiction via le modèle **Random Forest** préalablement sauvegardé.
— Visualiser instantanément le résultat du diagnostic (Présence ou Absence de maladie).

6.2 Exposition via Ngrok

Étant donné que le développement a été effectué dans un environnement cloud (Google Colab), l'application s'exécute par défaut sur un serveur local inaccessible depuis l'extérieur. Pour résoudre ce problème, nous avons intégré **Ngrok**.

Fonctionnement de Ngrok :

1. **Tunneling :** Ngrok crée un tunnel sécurisé entre le port local de l'application (généralement le port 8501 pour Streamlit) et une URL publique sécurisée (HTTPS).
2. **Accessibilité :** Grâce à ce tunnel, n'importe quel utilisateur disposant du lien généré par Ngrok peut accéder à l'application sans que celle-ci ne soit hébergée sur un serveur web traditionnel.
3. **Test en temps réel :** Cela permet de tester l'outil de diagnostic sur différents appareils (smartphone, tablette) comme s'il était déjà en production.

7 Conclusion

L'analyse montre que les modèles d'ensemble comme **Random Forest** offrent les meilleures performances pour ce type de diagnostic médical avec une précision de 87%. Les variables clés identifiées permettent de mieux comprendre les facteurs de risque cardiovasculaire.