Projet de Machine Learning

Classification des maladies cardiaques

Réalisé par:
DIOP Ousseynou
GADO Seman Giovanni Jocelyn
THIAM Omar
YATABARE Cheikhna Amala

Formatrice :
Mme Mously DIAW

INTRODUCTION

1. CONTEXTE

- Les maladies cardiaques figurent parmi les principales causes de morbidité et de mortalité dans le monde.

- Rien que l'infarctus du myocarde, composante principale cause environ 13 % de la mortalité mondiale totale, ce qui illustre l'ampleur et la gravité de ce problème de santé publique (OMS, 2023).

- Face à ce fléau, la prévention et la détection précoce apparaissent comme des leviers majeurs de santé publique.

- L'usage du Machine Learning pour prédire le risque cardiaque à partir de données simples s'inscrit dans cette logique, en alliant innovation technologique et démarche préventive.

INTRODUCTION

2. OBJECTIFS

L’objectif général de ce projet est de développer un modèle de classification fiable permettant de prédire le risque de maladies cardiaques à partir des données médicales du client.

Les objectifs spécifiques incluent :

- Identifier les variables les plus significatives pour la détection des maladies cardiaques.

- Concevoir un modèle prédictif robuste à l’aide d’algorithmes de Machine Learning.

- Évaluer la précision et la pertinence des prédictions avec des métriques adaptées.

- Déployer le modèle sur une plateforme accessible aux acteurs de la santé et aux patients.

PLAN

Introduction

01 Présentation des données

02 Prétraitement

03 Analyses descriptives

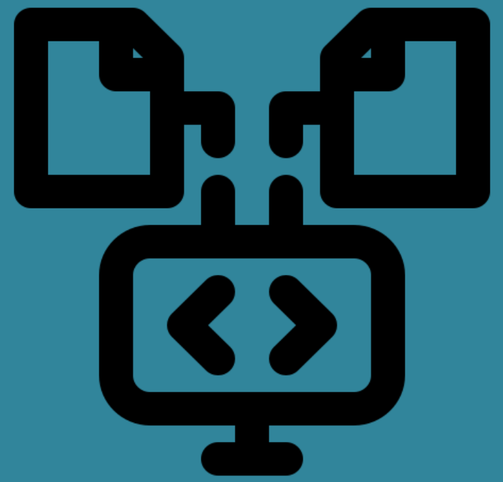
04 Modélisation

05 Choix du meilleur modèle

Déploiement et test de l’application de prédiction

01

Présentation des données



PRÉSENTATION DES DONNÉES

SOURCE

Les données utilisées pour ce projet proviennent de la plateforme *Kaggle*. Ce dataset, vise à évaluer le risque cardiovasculaire de leurs clients. Il est accessible à l'adresse suivante : Classification des maladies cardiaques

- Nombre d’observations : 303
- Nombre total de variables : 14
- 13 variables explicatives dont 7 qualitatives et 6 quantitatives
- Variable cible : Avoir une maladie cardiaque ou non (*NUM)*

02

Prétraitement

Prétraitement

1. CORRECTION DES VALEURS MANQUANTES

Les variables avec des valeurs manquantes sont :

- *ca* : renvoyant au nombre de vaisseaux(1,32%).
- *thal* : renvoyant a la thalassémie (0,66%).

Nous avons imputer la première variable par le KNN (le mieux indiqué dans notre situation).

La seconde variable étant qualitative, nous avons choisi l'amputation par le mode suivant le croisement avec la variable cible.

2. CORRECTION DES VALEURS ABERRANTES

Une valeur aberrante, également connue sous le nom d'outlier en statistiques, est une observation qui se situe à une distance significative des autres valeurs dans un ensemble de données. Dans notre jeu de données, les valeurs aberrantes étant au dessus de troisième quartile ont été ramenées au troisième quartile et celles étant en dessous du premier quartile ont été ramenées au premier quartile.

Prétraitement

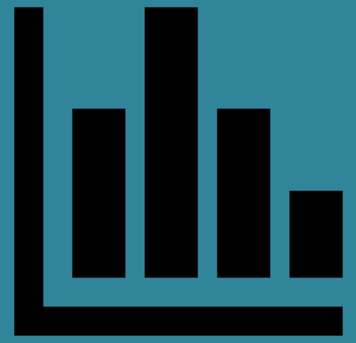
3. ENCODAGE DES VARIABLES QUALITATIVES

La base de données dont nous disposons (fournie par madame) contient des données encodées. Nous avons simplement étiqueté ces données afin de réaliser des statistiques descriptives plus lisibles, en remplaçant les codes par les véritables libellés des variables catégorielles.

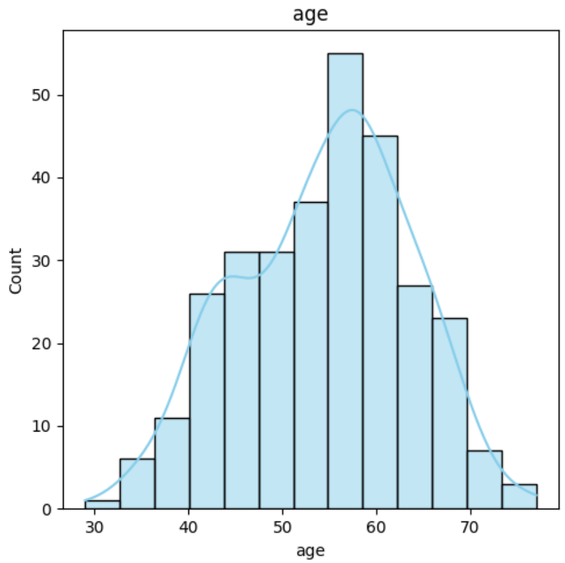
Cette base de données est utilisée dans plusieurs études du genre. Elle est disponible dans le site : Heart Disease - UCI Machine Learning Repository.

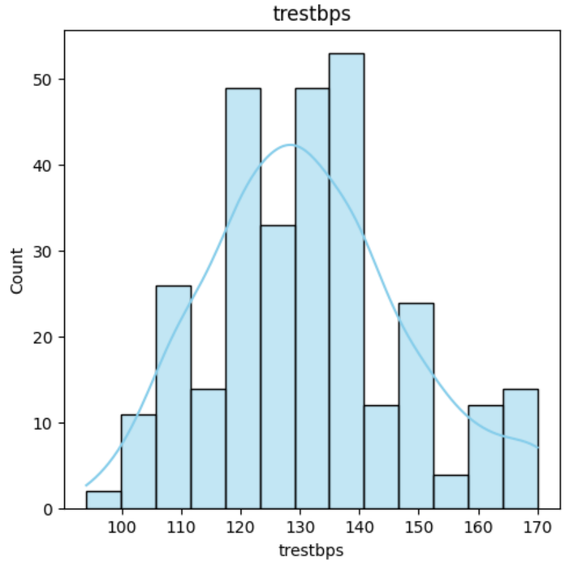
3

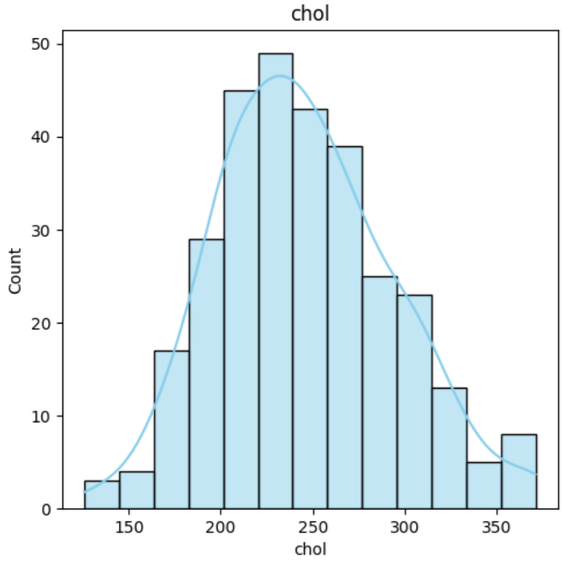
Analyses descriptives

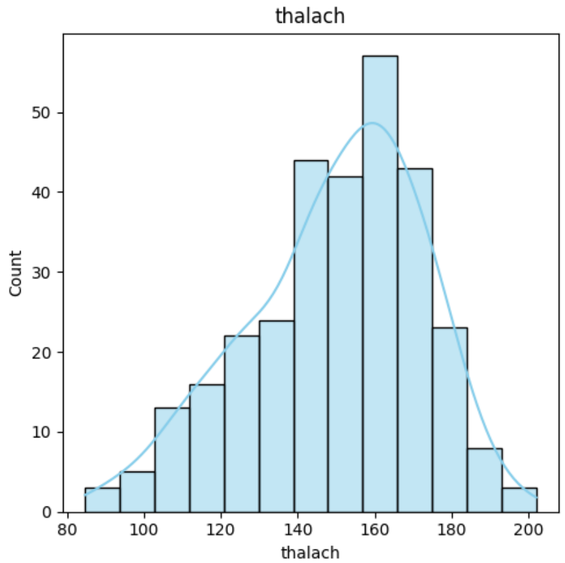


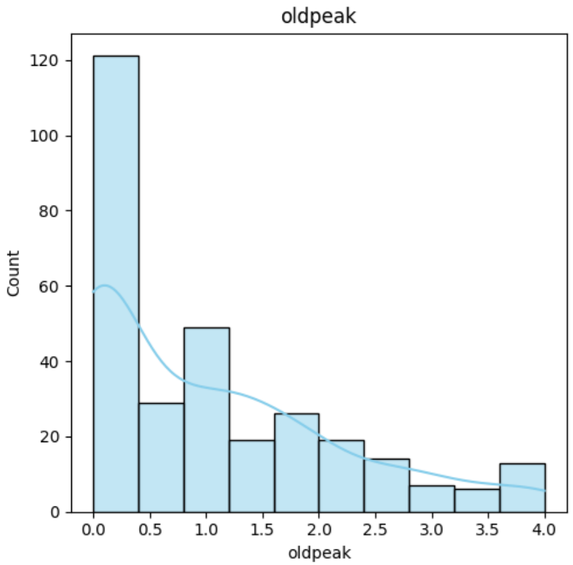
Analyse univariée











Analyse univariée

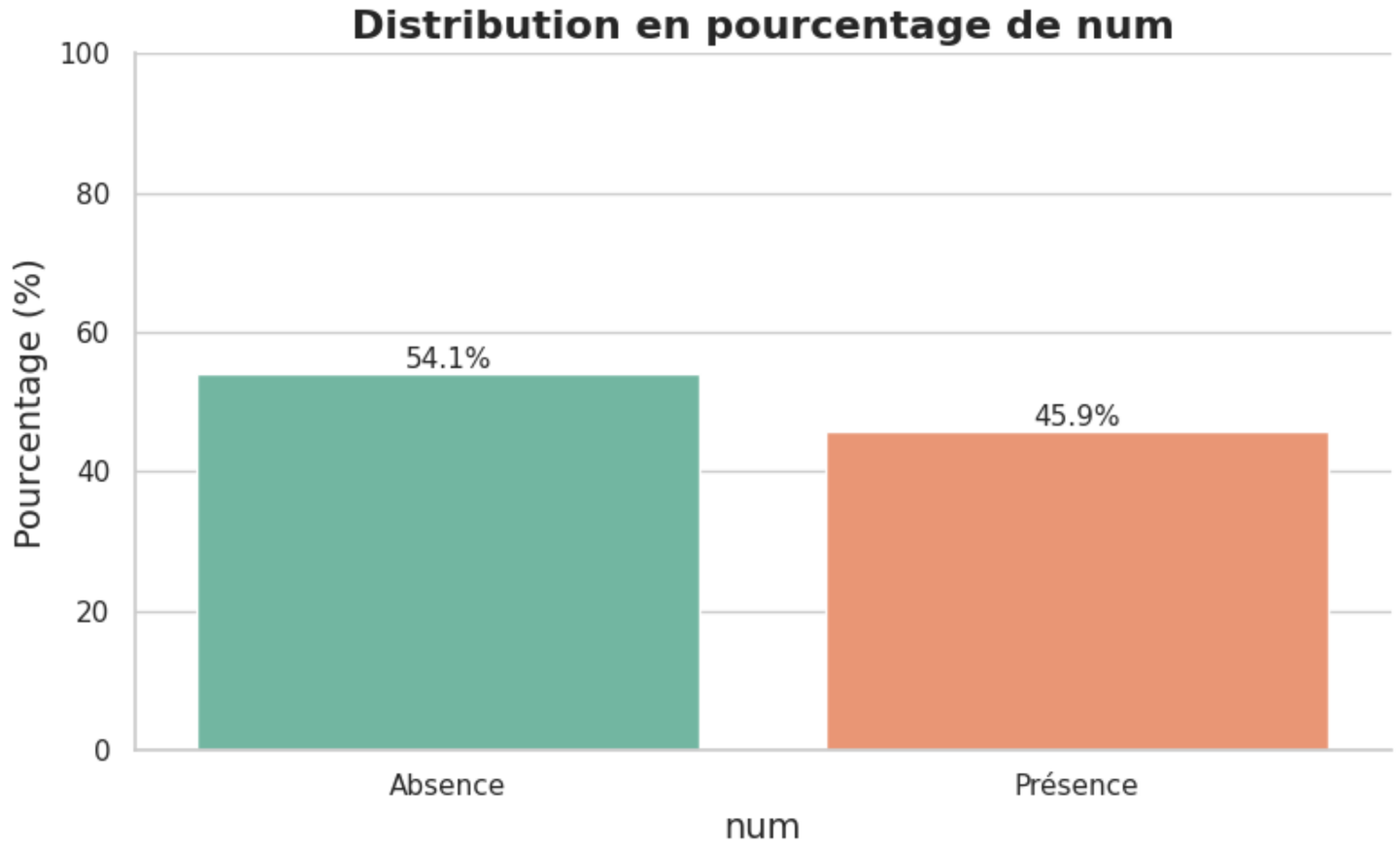
❖ Interprétation

L’analyse des distributions des variables numériques du jeu de données Heart Disease révèle plusieurs caractéristiques intéressantes.

Les variables **age**, **trestbps, chol** *et* **thalach** présentent des distributions globalement symétriques et proches de la loi normale, avec des valeurs centrales respectivement autour de 55 ans, 130 mmHg, 240 mg/dL et 160 bpm.

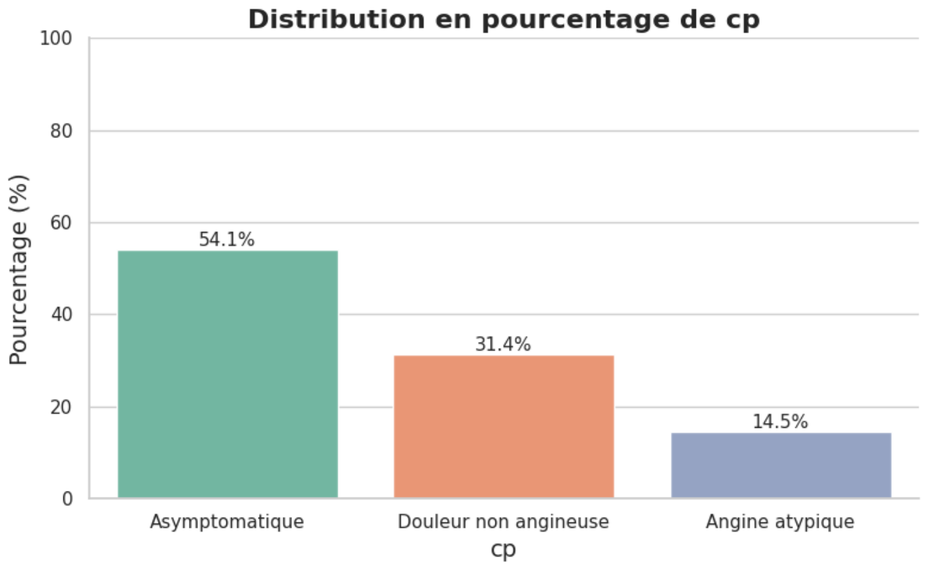
La variable **oldpeak**, qui mesure la dépression du segment ST à l’effort, affiche une distribution fortement asymétrique à droite, la majorité des observations se concentrant autour de 0, indiquant peu ou pas de dépression pour la plupart des patients.

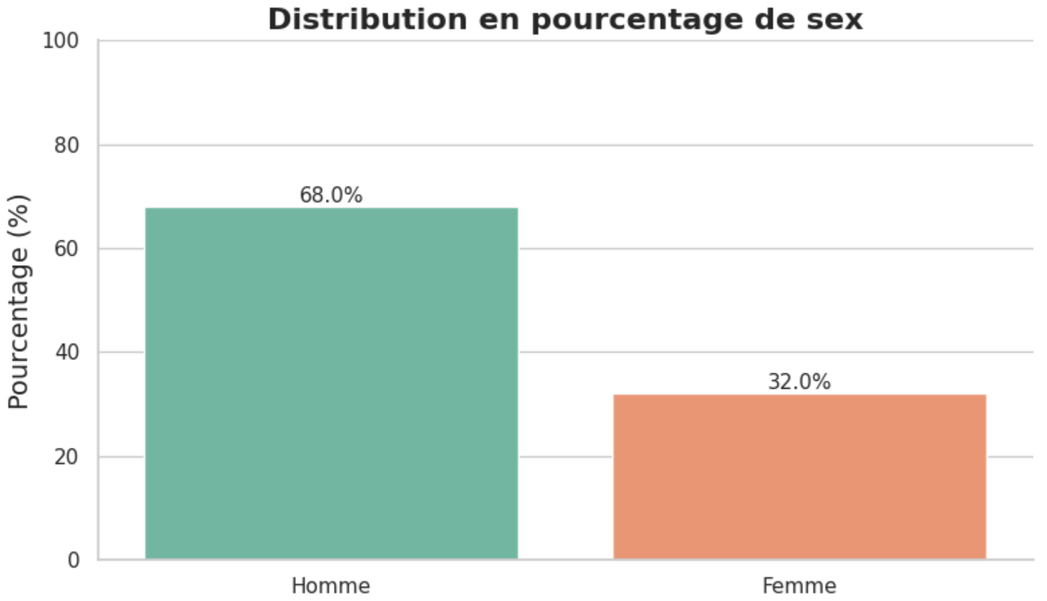
Analyse univariée

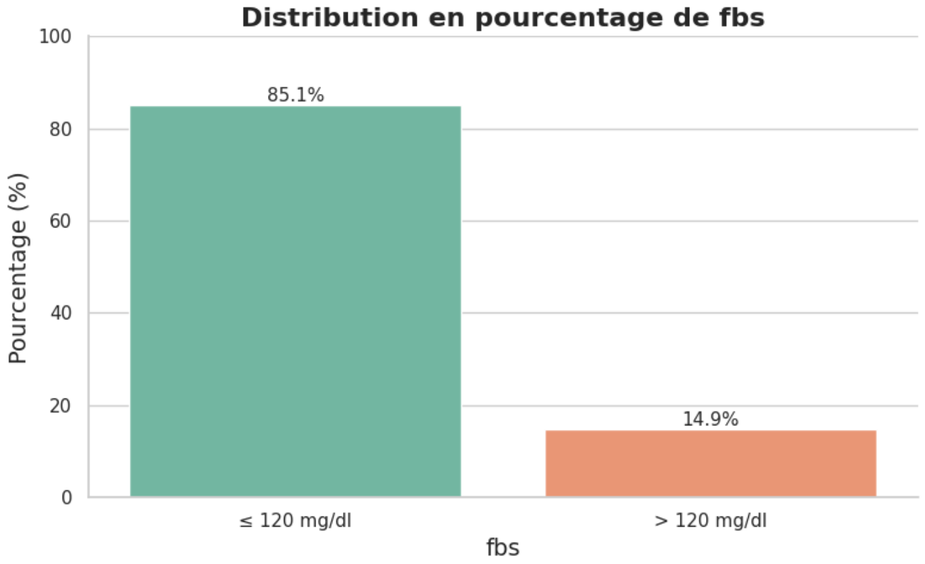


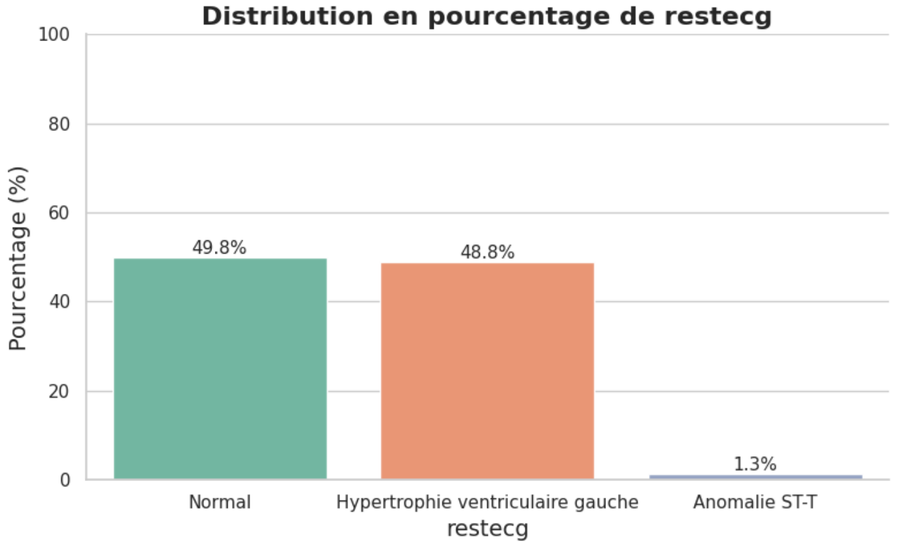
Le graphe ci-contre montre la distribution de notre variable d'intérêt. Les résultats montrent que sur 54,1% des individus il n'y a pas de maladie cardiaque.

Analyse univariée









Analyse univariée

❖ Interprétation

- La variable sex montre une prédominance masculine dans l'échantillon, avec 68 % d'hommes contre 32 % de femmes.

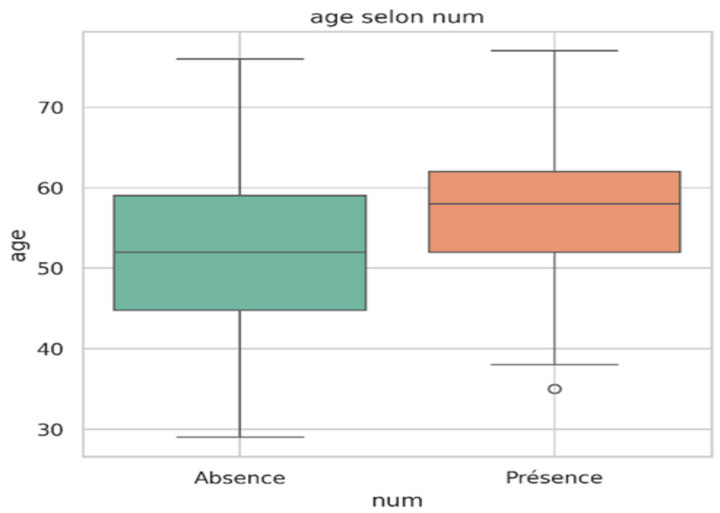
- Concernant la variable cp (type de douleur thoracique), la majorité des patients présentent une douleur asymptomatique (54,1 %), suivie des douleurs non angineuses (31,4 %) et des angines atypiques (14,5 %).

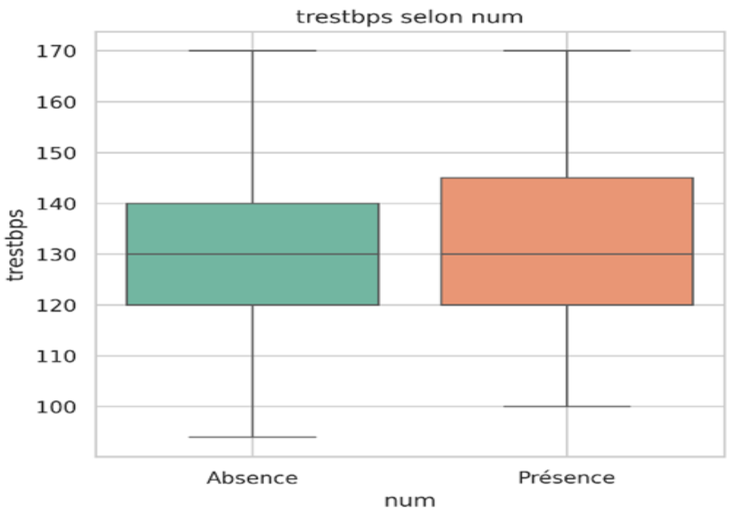
- Pour la glycémie à jeun (fbs), 85,1 % des individus ont un taux ≤ 120 mg/dl, tandis que 14,9 % le dépassent.

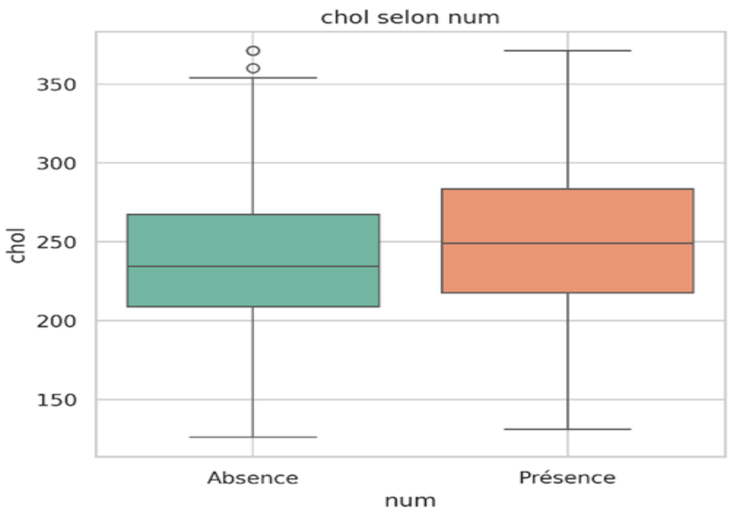
- La répartition de l'électrocardiogramme au repos (restecg) est quasi équilibrée entre les cas normaux (49,8 %) et les hypertrophies ventriculaires gauches (48,8 %), avec une faible proportion d'anomalies ST-T (1,3 %).

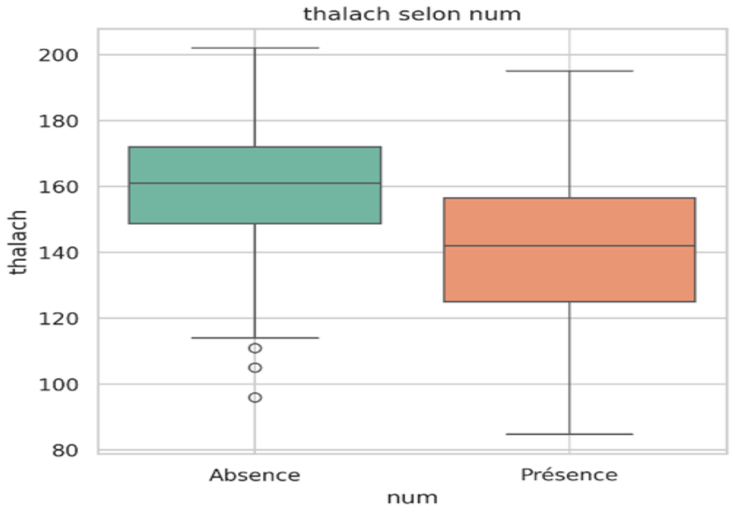
Analyse bivariée

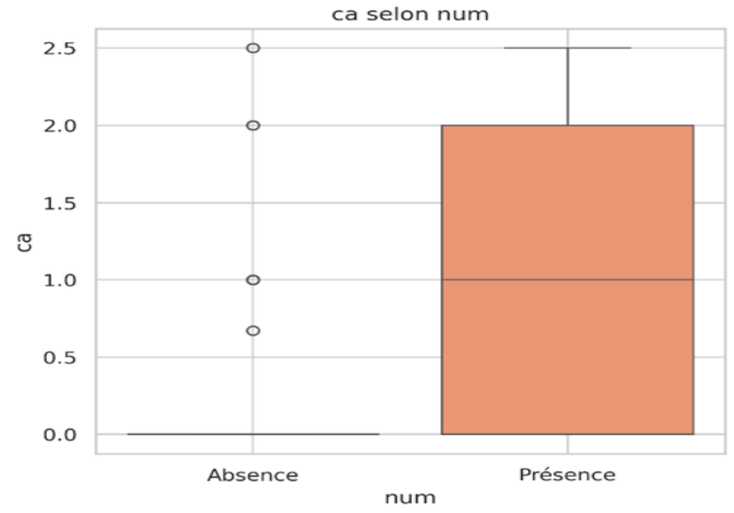
❖ **Boîtes à moustaches des variables quantitatives selon la variable cible**











Analyse bivariée

❖ Interprétation

Nous remarquons que :

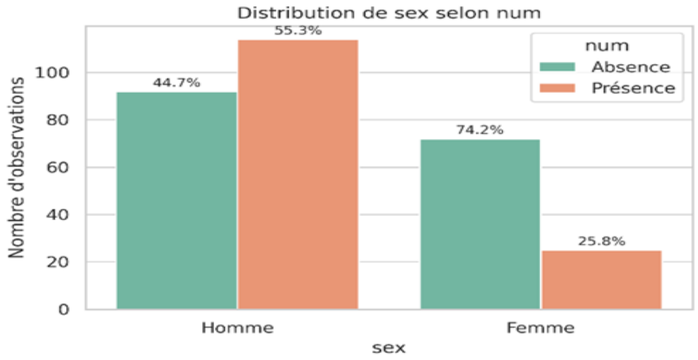
- la médiane d’âge est plus élevée chez les patients atteints de la maladie ; ce qui signifie que les personnes malades ont tendance à être plus âgées que celles non malades.

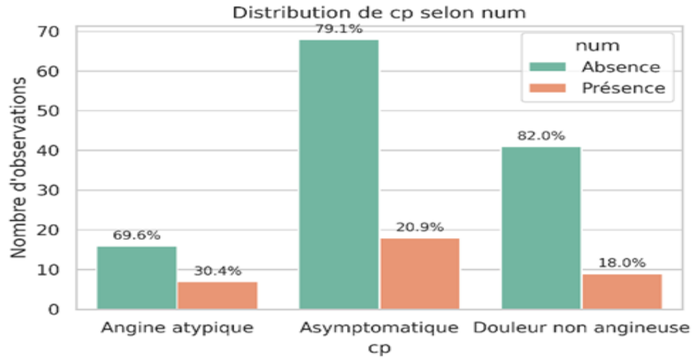
- la variable Oldpeak semble très discriminante : les malades ont des valeurs bien plus importantes ; ce qui peut être un indicateur fort de présence de maladie.

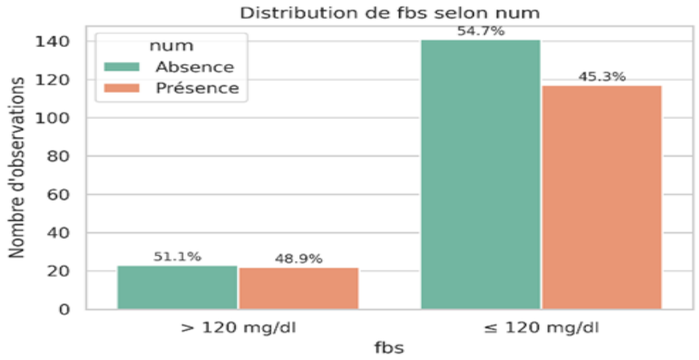
- le nombre de vaisseaux colorés par fluoroscopie prend des valeurs nettement plus élevées chez les malades ; cette variable est fortement discriminante.

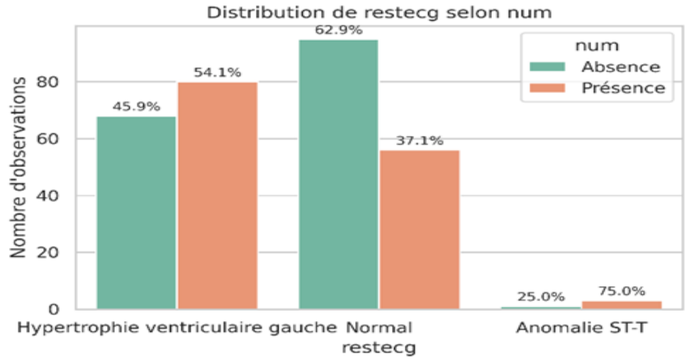
Analyse bivariée

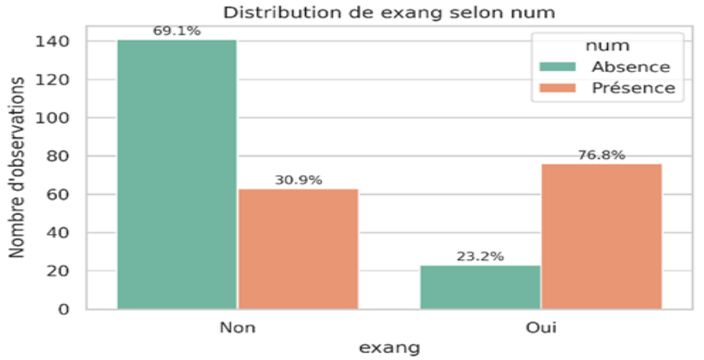
❖ Distribution des variables qualitatives selon la variable cible

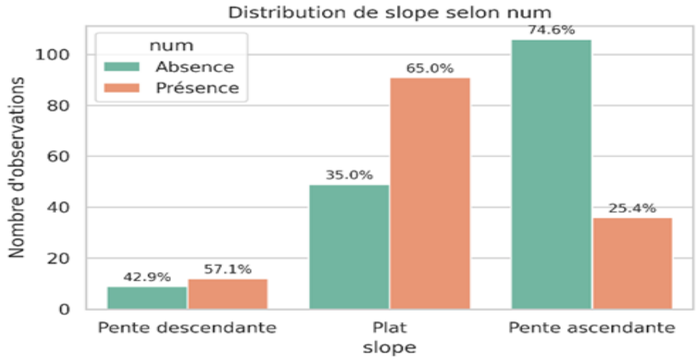


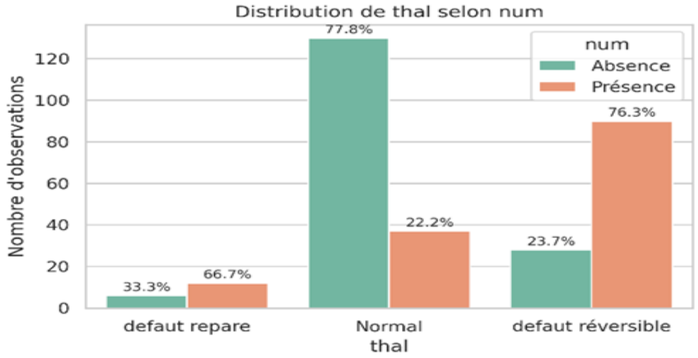












Analyse bivariée

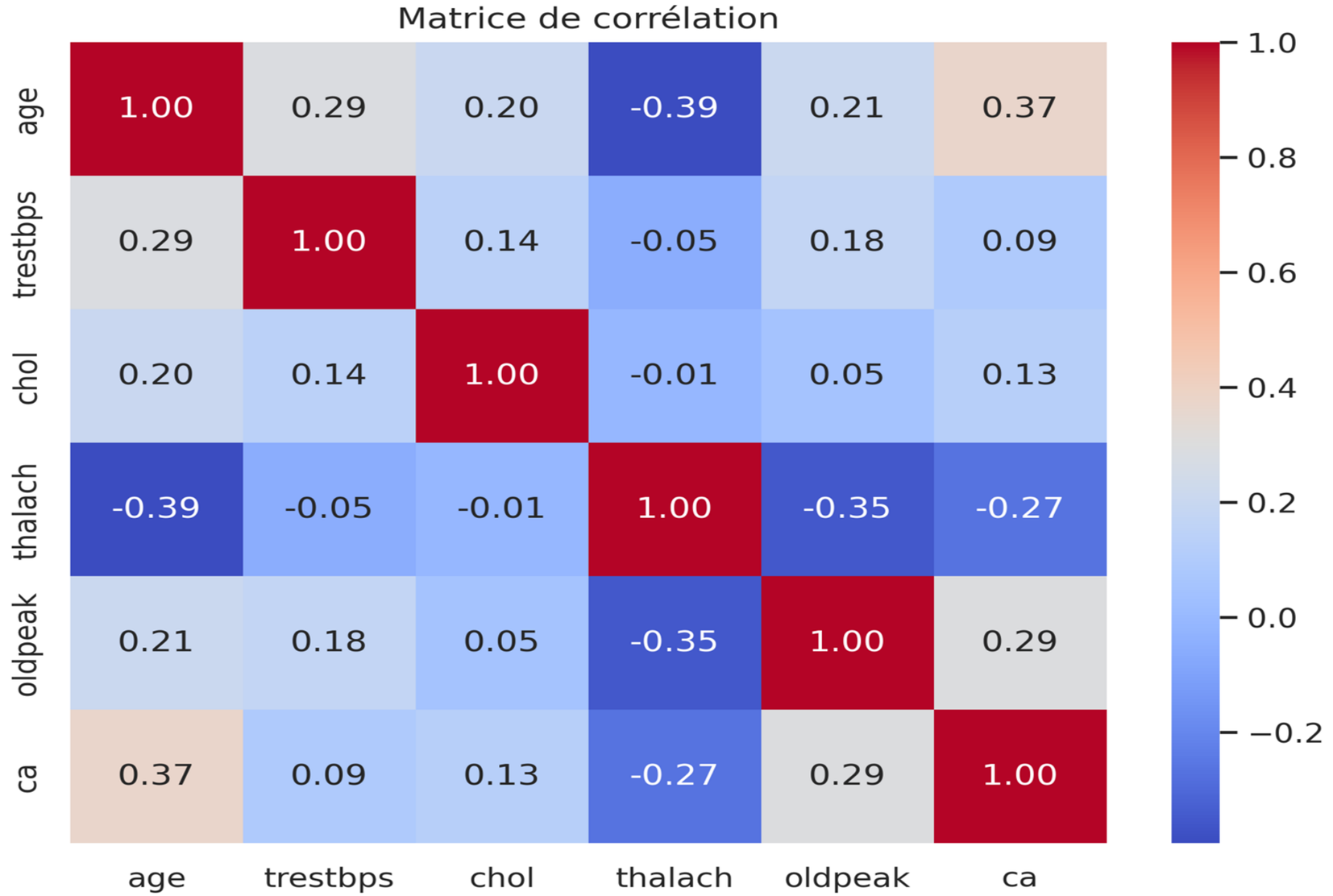
❖ Interprétation

Nous remarquons que :

- les hommes sont plus représentés parmi les malades (55,3 % contre 25,8 %).
- une douleur asymptomatique est fortement associée à la présence de maladie cardiaque.
- les anomalies ST-T sont un fort indicateur de présence de maladie cardiaque.
- l’apparition de l’angine à l’effort est fortement liée à la maladie.
- les individus atteints de défaut réversible ont tendance à être malade.

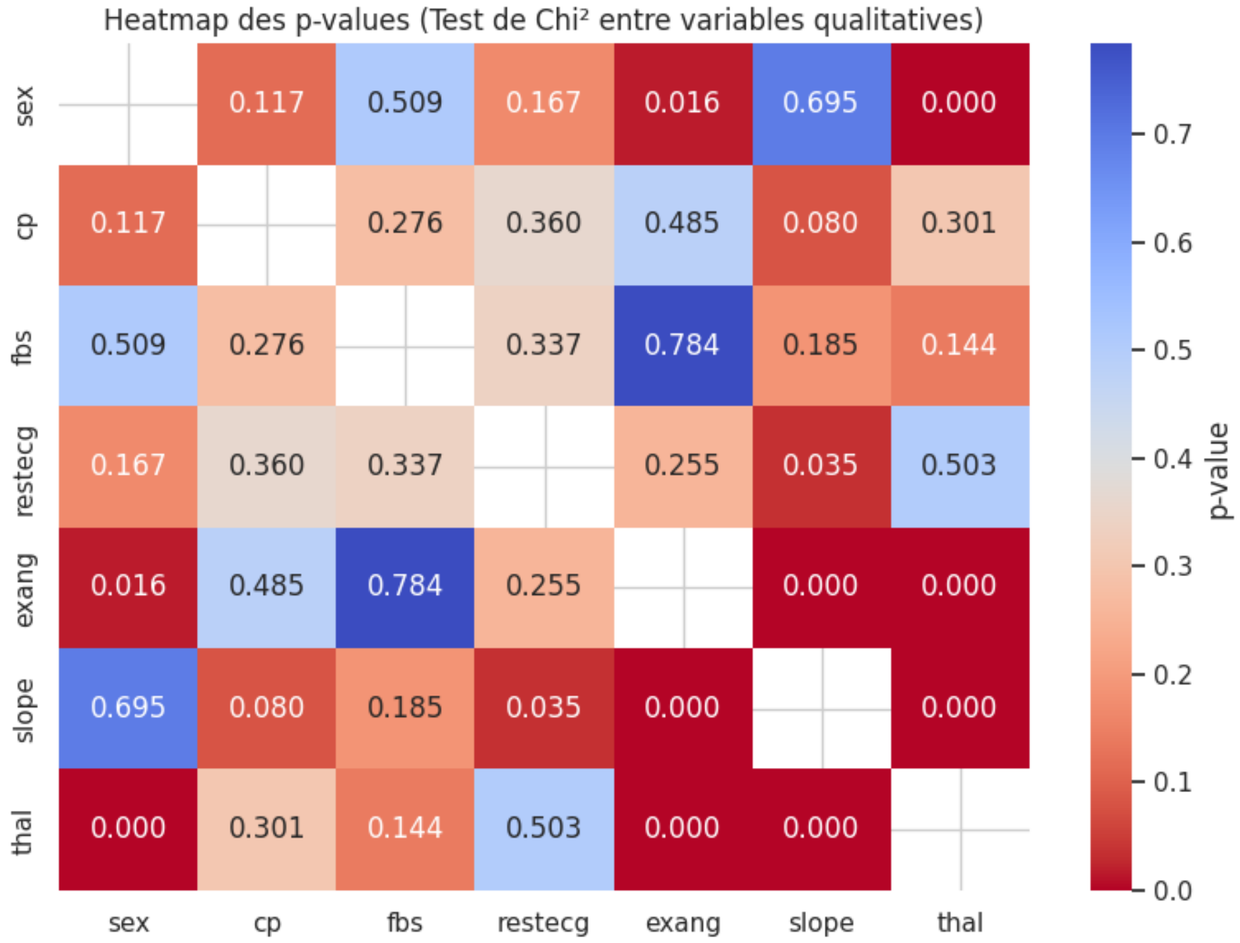
Analyse bivariée

❖ **Matrice de corrélation entre les variables explicatives quantitatives**



Analyse bivariée

❖ Test de Khi-deux entre les variables explicatives qualitatives



Certaines variables explicatives qualitatives présentent des relations significatives entre elles. Pour la partie modélisation, les variables explicatives qualitatives qui ne seront pas prises sont 'exang', 'thal' et 'slope'

4

Modélisation

(x)

Modélisation

1. STANDARDISATION DES DONNÉES

La standardisation consiste à transformer les données de sorte que chaque variable ait une moyenne de 0 et un écart-type de 1. Cela s'effectue en soustrayant la moyenne de chaque variable et en la divisant par son écart-type.

Elle est indispensable car elle nous permet de ramener toutes les données à une même échelle afin de réduire l’impact des valeurs extrêmes et permettre ainsi au modèle de mieux généraliser sur de nouvelles données.

2. MODÈLES TESTÉS

- Logistic Regression
- KNN
- Random Forest
- XGBoost
- Naive Bayes
- Decision Tree
- SVM
- Neural Network

Modélisation

3. MÉTHODOLOGIE

Pour les différents modèles suscités, la méthodologie suivante a été appliquée :

- Initialisation du modèle
- Tests d’hypothèses si nécessaire
- Optimisation des paramètres avec la validation croisée

4. MÉTRIQUES D’ÉVALUATION

- Accuracy
- Precision
- Recall
- F1-Score
- Matrice de confusion

5

Choix du meilleur modèle

Choix du meilleur modèle

1. PRÉSENTATION DES RÉSULTATS

Les performances des différents modèles sont consignées dans le tableau suivant :

Comparaison des Modèles

	Modèle	Accuracy	Précision	Rappel	F1-score	short_name
0	Decision Tree	77.05%	77.67%	77.05%	77.02%	dt
2	Logistic Regression	77.05%	79.31%	77.05%	76.80%	logreg
3	Multi-Layer Perceptron	75.41%	77.08%	75.41%	75.22%	mlp
6	Support Vector Machine	75.41%	78.14%	75.41%	75.05%	svm
1	k-Nearest Neighbors	73.77%	75.84%	73.77%	73.49%	knn
5	Random Forest	73.77%	75.84%	73.77%	73.49%	rf
7	XGBoost	73.77%	75.84%	73.77%	73.49%	xgb
4	Naive Bayes	72.13%	74.61%	72.13%	71.72%	nb

Choix du meilleur modèle

2. MEILLEUR MODÈLE

Le modèle Logistic Regression se distingue par ses :

- Accuracy : 0.85
- Precision : 0.86
- Rappel : 0.85
- F1-Score : 0.85

Par suite, le modèle retenu est **la Logistic Regression**.

Déploiement et test de l’application de prédiction

Classification des maladies cardiaques 📊

Bienvenue dans l'application de classification des maladies cardiaques !

Cette application vous permet d'explorer différentes techniques d'analyse, y compris l'introduction des données, la modélisation, et la visualisation des résultats.

Vous pouvez naviguer entre les différentes pages pour découvrir les insights cachés dans les données.

Prédiction de Maladie Cardiaque

Remplissez les informations du patient pour prédire le risque de maladie cardiaque :

Âge

50

20 100

Sexe

Femme

Sucre à jeun

≤ 120 mg/dl

Résultats ECG au repos

Normal

Prédiction de Maladie Cardiaque

Remplissez les informations du patient pour prédire le risque de maladie cardiaque :

Âge

20 100

Sexe

Femme

Type de douleur thoracique

Angine typique

120

200

Sucre à jeun

< 120 mg/dl

Résultats ECG au repos

Normal

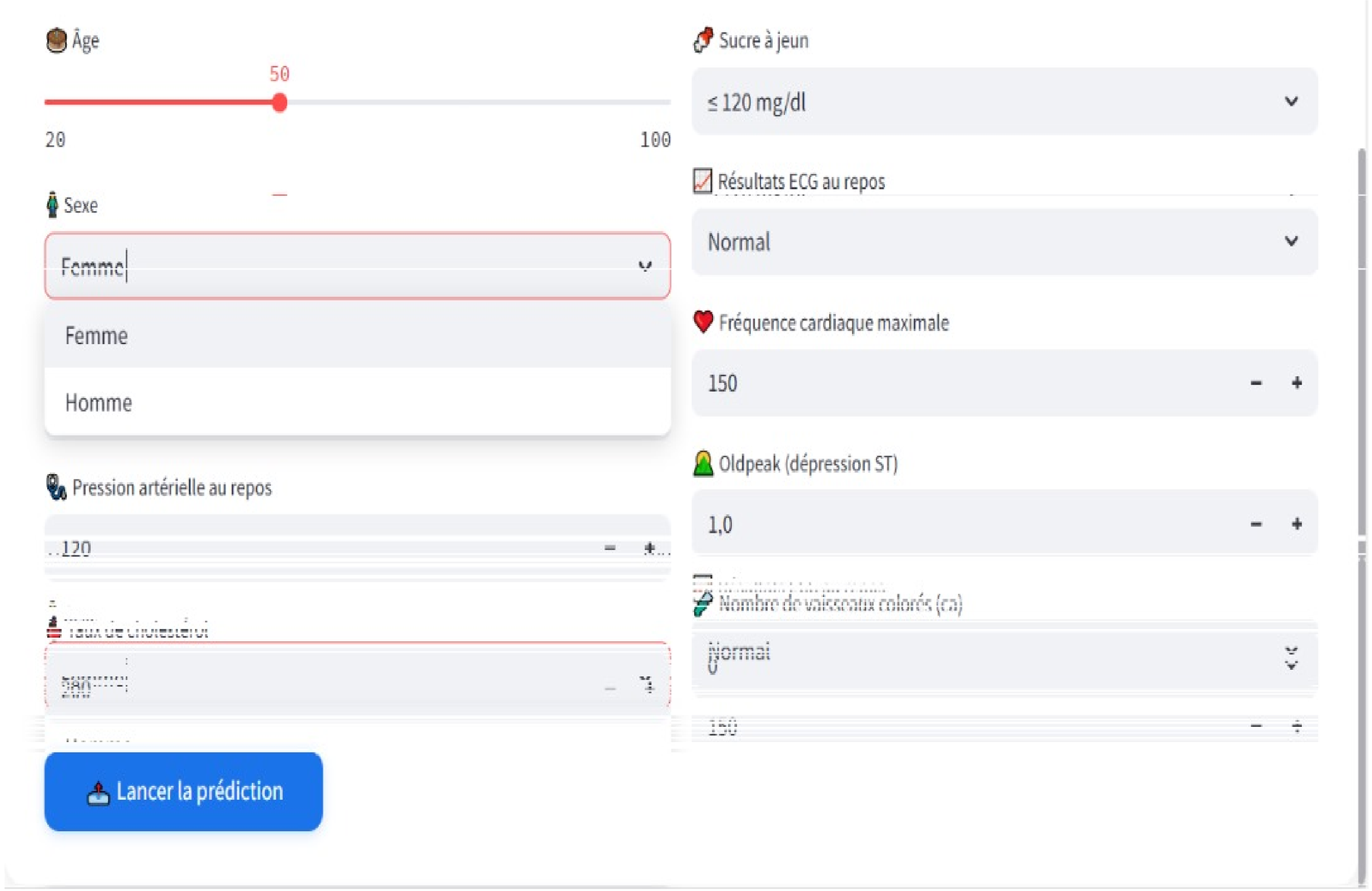
Fréquence cardiaque maximale

150

Oldpeak (dépression ST)

1,0

0


Âge
50
20
100
Sexe
Femme
Femme
Homme
Pression artérielle au repos
120
Sucre à jeun
≤ 120 mg/dl
Résultats ECG au repos
Normal
Fréquence cardiaque maximale
150
Oldpeak (dépression ST)
1,0
Normal
Lancer la prédiction

CONCLUSION

- Nous avons commencé par comprendre les données et le problème à résoudre : prédire le risque de maladies cardiaques à partir de données médicales simples.

- Un prétraitement rigoureux a été effectué : traitement des valeurs manquantes, normalisation, et création de nouvelles variables.
- Plusieurs modèles de classification ont été testés : Régression Logistique, KNN, Arbre de Décision, etc.

- Le modèle **Logistic Regression** s'est révélé le plus performant selon les métriques (précision, rappel, F1-score).

CONCLUSION

- Le modèle final permet d’estimer efficacement le risque cardiaque, à partir de données simples et accessibles.

- Il répond à un enjeu mondial de santé publique, en lien avec les objectifs de l’OMS et les ODD (réduction des maladies non transmissibles).

- Une version déployée sur plateforme rend son usage possible pour la prévention, l’orientation médicale et l’accompagnement personnalisé.

- Ce projet montre comment le Machine Learning peut être un levier utile pour une santé plus prédictive et inclusive

MERCI POUR VOTRE AIMABLE ATTENTION

BIBLIOGRAPHIE

- Cours du professeur
- Kaggle
- Scikit-learn