Fouille et Prédiction des Maladies Cardiaques

De l'exploration à la classification supervisée

Supervisé par
Pr. SKITTOU Mustapha

Réalisé par
Miskar Amina et Aziz Mohammed

Filière
2ITE – 3ème année

Année Universitaire : 2025–2026

Table des matières

Liste des figures **3**

Liste des tableaux **4**

1 Problématique (Compréhension du Domaine) **6**
1.1 Contexte et Importance 6
1.2 Problématique et solution proposée 7
1.3 CRISP-DM : Business Understanding 7

2 Ensemble de Données (Compréhension des Données) **9**
2.1 Source de données et structure 9
2.2 Analyse descriptive (avant nettoyage) 11
2.2.1 Idée sur le contenu du Dataset 11
2.2.2 La homogénisation des types de données 12
2.2.3 Les paramètres statistiques 13
2.2.4 Visualisation graphique initiale 13
2.3 Évaluation de la qualité des données 16

3 Exploration et Pré-traitement **17**
3.1 Chargement et Structure du Dataset 17
3.1.1 Répartition de la Cible 17
3.1.2 Les valeurs manquantes, aberrantes et les doublons 18
3.2 Enrichissement du Dataset (`df2`) 18
3.3 Analyses Exploratoires Clés 19
3.3.1 Score de Risque Cardiaque 19
3.3.2 Fréquence Cardiaque Maximale (`thalach`) 20
3.3.3 Âge vs Cholestérol 21
3.3.4 Profils Cliniques : Coordonnées Parallèles 22
3.3.5 Corrélations Clés 23

4 Modélisation **26**
4.1 Introduction 26
4.2 CRISP-DM : Modeling 26
4.3 Régression Logistique 26
4.3.1 Courbe Precision-Recall 27
4.4 Naïve Bayes 27
4.4.1 Distribution des probabilités prédites 28
4.5 K-Nearest Neighbors (KNN) 28
4.5.1 Distribution des probabilités prédites 29

4.6 Machines à Vecteurs de Support (SVM) 29
4.6.1 Fronière de décision dans le plan âge–thalach 30
4.7 Arbres de Décision 30
4.7.1 Matrice de confusion 31
4.8 Forêt Aléatoire (Random Forest) 31
4.8.1 Importance des variables 32
4.9 XGBoost 32
4.9.1 Importance des variables (poids des splits) 33
4.10 Conclusion 33

5 Évaluation de la Démarche **34**
5.1 Séparation des données (Data Splitting) 34
5.2 Les modèles de classification 34
5.3 Visualisation des performances 37
5.4 Résultats comparatifs 37
5.5 Le choix du modèle 37

6 Déploiement du modèle **39**
6.1 Interface Web Interactive avec Streamlit 39

Webographie **42**

Table des figures

2.1 Les cinq premières lignes 11
2.2 Regroupement vertical des visualisations avant nettoyage 14

3.1 Répartition et % de malades par `risk_score` 20
3.2 Les malades ont une `thalach` plus basse 21
3.3 Malades = zone vert 22
3.4 Sains (vert) vs Malades (rouge) : profils opposés 23
3.5 Heatmap des corrélations — Variables fortement corrélées 24
3.6 Pairplot des variables clés — Séparation visuelle entre classes 25

4.1 Courbe Precision-Recall - Régression Logistique 27
4.2 Distribution des probabilités prédites - Naïve Bayes 28
4.3 Distribution des probabilités prédites - KNN 29
4.4 Fronière de décision SVM (RBF) - age vs thalach 30
4.5 Matrice de Confusion - Arbre de Décision 31
4.6 Importance des variables dans la Forêt Aléatoire 32
4.7 Importance des variables dans XGBoost (poids des splits) 33

5.1 Rapport de classification - Régression Logistique 34
5.2 Rapport de classification - Naïve Bayes 35
5.3 Rapport de classification - Arbre de Décision 35
5.4 Rapport de classification - Forêt Aléatoire 35
5.5 Rapport de classification - KNN 36
5.6 Rapport de classification - SVM 36
5.7 Rapport de classification - XGBoost 36
5.8 Receiver Operating Characteristic 37

6.1 Page d'accueil de l'application Streamlit 39
6.2 Cycle CRISP-DM appliqué au projet 43

Liste des tableaux

2.1 Types de données des variables du jeu de données 12
2.2 Statistiques descriptives des variables 13

3.1 Distribution des classes dans le jeu de données 17
3.2 Variables dérivées ajoutées à `df2` 19

5.1 Comparaison des modèles de classification selon leur précision 38

Introduction générale

La santé cardiovasculaire représente un enjeu majeur de santé publique, les maladies cardiaques étant la première cause de mortalité dans le monde. La détection précoce des individus à risque est essentielle pour prévenir les complications graves et améliorer la qualité de vie. Dans ce contexte, l'analyse des données cliniques à travers les techniques de data mining permet d'identifier des facteurs de risque et de prédire la présence de maladies cardiaques à partir de variables physiologiques et médicales.

Ce projet s'inscrit dans cette démarche : à partir d'un dataset médical comprenant 1025 patients et 14 variables cliniques, nous avons exploré et prétraité les données, créé des variables dérivées, puis entraîné et comparé plusieurs modèles de classification. La Régression Logistique, sélectionnée pour sa performance et son interprétabilité, a été évaluée avec des métriques telles que l'accuracy, le F1-score et la matrice de confusion. Enfin, le modèle a été déployé via une interface Streamlit, rendant son usage interactif et accessible pour le soutien à la décision clinique.

Problématique (Compréhension du Domaine)

Introduction

La santé cardiovasculaire représente un enjeu majeur de santé publique, les maladies cardiaques étant la première cause de mortalité mondiale. La prévention et la détection précoce des patients à risque sont essentielles pour réduire les complications graves et améliorer la prise en charge médicale. Dans ce contexte, l'analyse de données cliniques à l'aide de techniques de data mining permet d'identifier les facteurs de risque et de prédire la probabilité qu'un individu soit atteint d'une maladie cardiaque, offrant ainsi un outil précieux pour soutenir la décision clinique.

1.1 Contexte et Importance

La santé constitue un pilier fondamental du bien-être et de la qualité de vie des individus. La préservation de la santé et la prévention des maladies représentent donc un enjeu majeur à la fois pour les individus et pour les systèmes de santé publique. Parmi les maladies les plus graves et répandues, les maladies cardiovasculaires se distinguent comme la première cause de mortalité dans le monde, entraînant chaque année des millions de décès (OMS, 2023). Ces maladies englobent un large spectre de troubles du cœur et des vaisseaux sanguins, allant de l'hypertension à l'infarctus du myocarde, et sont souvent liées à des facteurs de risque tels que l'âge, le cholestérol, la pression artérielle, le tabagisme ou le diabète.

La cardiologie, discipline médicale dédiée à l'étude et au traitement des maladies du cœur, revêt donc une importance capitale pour la prévention et la prise en charge des patients. Dans ce contexte, il devient essentiel de détecter précocement les individus à risque, afin de réduire les complications graves et la mortalité associée. C'est pourquoi l'analyse de données cliniques, via des techniques de data mining, constitue un outil précieux pour identifier les facteurs de risque et prédire la probabilité qu'une personne soit atteinte d'une maladie cardiaque.

L'importance de mettre l'attention sur ce sujet réside dans le fait que la détection précoce et la prévention ciblée peuvent sauver des vies. Les maladies cardiaques sont souvent silencieuses au début et leurs complications peuvent être fatales. En orientant la

recherche et les systèmes d'aide à la décision vers la prévention et l'identification rapide des patients à risque, les professionnels de santé peuvent agir de manière proactive, adapter les traitements, et réduire significativement la charge médicale et sociale liée aux maladies cardiovasculaires. Ce projet de data mining contribue ainsi à renforcer cette vigilance et à soutenir des interventions médicales plus efficaces.

1.2 Problématique et solution proposée

— **Problématique** :
Comment prédire la présence d'une maladie cardiaque à partir de 13 variables cliniques ?
Cette question illustre le défi central de notre projet. Elle soulève la nécessité de comprendre comment différentes informations cliniques — âge, sexe, pression artérielle, cholestérol, fréquence cardiaque maximale, angine à l'effort, et d'autres facteurs — interagissent et influencent le risque de maladie cardiovasculaire. La problématique consiste donc à identifier de manière fiable si un individu présente un risque de maladie cardiaque en se basant uniquement sur ces variables, tout en tenant compte de la complexité et de la variabilité des données cliniques.

— **Solution proposée** :
Pour répondre à la problématique, la solution repose sur l'exploitation des données cliniques disponibles afin de construire un modèle de classification binaire capable de prédire la présence ou l'absence d'une maladie cardiaque (0 = sain, 1 = malade). La première étape consiste à explorer les données, en analysant les statistiques descriptives et les distributions des 13 variables pour identifier les tendances, les valeurs aberrantes ou les éventuelles anomalies. Ensuite, il est nécessaire de nettoyer et préparer les données, ce qui inclut la gestion des valeurs manquantes, la transformation des variables catégorielles en format exploitable par les modèles, et la normalisation ou standardisation des données numériques lorsque cela est nécessaire. Une fois les données correctement préparées, elles peuvent être utilisées pour entraîner différents algorithmes de classification (tels que la régression logistique, les arbres de décision, ou les méthodes de type Random Forest), permettant ainsi d'estimer de manière fiable la probabilité qu'un individu soit atteint d'une maladie cardiaque. Cette approche méthodique assure que le modèle sera à la fois précis et robuste, tout en facilitant l'interprétation des résultats pour des applications médicales concrètes.

1.3 CRISP-DM : Business Understanding

La première étape du processus CRISP-DM, Business Understanding, consiste à définir clairement les objectifs du projet et à comprendre le contexte dans lequel il s'inscrit. Dans notre cas, le projet vise à prédire la présence d'une maladie cardiaque chez un individu à partir de 13 variables cliniques. L'objectif métier principal est d'identifier précocement les personnes à risque, afin de soutenir les décisions médicales et de contribuer à la prévention des maladies cardiovasculaires, première cause de mortalité mondiale. Cette étape implique également de préciser les critères de succès, tels que la précision et la fiabilité du modèle prédictif, et d'identifier les contraintes et ressources disponibles, notamment la qualité et la complétude du dataset. En comprenant ces aspects, le projet peut être

structuré efficacement pour aboutir à des résultats exploitables et pertinents pour la santé publique.

Conclusion

Ce chapitre a permis de définir clairement la problématique centrale du projet : prédire la présence d'une maladie cardiaque à partir de 13 variables cliniques. Il souligne également l'importance d'une approche méthodique, incluant l'exploration, le nettoyage et la préparation des données, afin de construire un modèle fiable et robuste. La compréhension approfondie du contexte et des objectifs métier, conformément à la première étape du processus CRISP-DM, constitue la base nécessaire pour structurer efficacement les étapes suivantes du projet.

Ensemble de Données (Compréhension des Données)

Introduction

Le jeu de données utilisé provient de la plateforme Kaggle et comprend 1025 observations avec 14 variables cliniques liées aux maladies cardiovasculaires. Chaque ligne représente un patient, et les colonnes décrivent des caractéristiques physiologiques et médicales telles que l'âge, le sexe, le type de douleur thoracique, la pression artérielle, le cholestérol ou la fréquence cardiaque maximale. L'analyse descriptive initiale permet de comprendre la structure du dataset, les types de variables, et les distributions des mesures, offrant ainsi une première vision sur la population étudiée et les relations possibles entre les différentes variables.

2.1 Source de données et structure

Ce jeu de données provient de la plateforme Kaggle, une communauté en ligne reconnue pour le partage de jeux de données et de défis en science des données. Il est accessible via le lien suivant : Heart Disease Dataset, dont le lien est : `https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset`. Ce dataset rassemble des données médicales anonymisées liées aux maladies cardiovasculaires, compilées à partir d'études cliniques. Il est couramment utilisé pour la prédiction de la présence ou de l'absence de maladies cardiaques à partir de caractéristiques physiologiques et d'examens médicaux.
Le jeu de données comprend 1025 lignes (observations) et 14 colonnes (variables). Chaque ligne représente un patient, et les colonnes décrivent ses attributs cliniques. Voici la signification de chaque variable :

— **age :** Âge du patient. C'est un facteur de risque non modifiable majeur pour les maladies cardiaques, puisque le risque augmente avec le vieillissement en raison d'une exposition prolongée à d'autres facteurs comme l'hypertension ou le cholestérol élevé. Bien qu'il n'existe pas de valeur « normale », le risque cardiaque reste faible avant 45 ans chez l'homme et 55 ans chez la femme, augmentant significativement après 65 ans.

— **sex (Sexe) :** Les œstrogènes exercent un effet protecteur chez les femmes avant la ménopause, tandis que les hommes présentent un risque plus élevé de maladie coronarienne à un âge plus jeune. Après la ménopause, le risque des femmes tend à s'aligner sur celui des hommes.

- **cp (chest pain type) :** Le type de douleur thoracique (ou chest pain type) est une variable clinique qui classe les douleurs ressenties par le patient selon leur origine probable. En cardiologie, cette variable distingue les douleurs typiquement ischémiques des douleurs d'origine non cardiaque. La douleur thoracique est un symptôme clé de l'ischémie myocardique due à une obstruction des artères coronaires.
- **trestbps (resting blood pressure) :** La pression artérielle au repos (systolique) mesure la force exercée par le sang sur les parois artérielles au repos, exprimée en mmHg. En cardiologie, elle permet d'évaluer l'état hémodynamique du patient et constitue un indicateur clé de l'hypertension artérielle, l'un des principaux facteurs de risque modifiables des maladies cardiovasculaires. Une pression trop élevée entraîne des lésions de l'endothélium et favorise l'athérosclérose, augmentant ainsi le risque de coronaropathie, d'insuffisance cardiaque et d'accident vasculaire cérébral. La valeur normale est inférieure à 120 mmHg ; une valeur comprise entre 120 et 129 est considérée comme élevée, et au-delà de 130 mmHg, on parle d'hypertension.
- **chol (serum cholesterol) :** Le cholestérol sérique représente le taux total de cholestérol dans le sang, exprimé en mg/dL. En cardiologie, un taux de cholestérol élevé favorise l'accumulation de plaques d'athérome sur les parois des artères, entraînant une réduction du flux sanguin et augmentant le risque de maladie coronarienne. Les valeurs inférieures à 200 mg/dL sont considérées comme normales, entre 200 et 239 comme limites, et au-delà de 240 mg/dL comme élevées.
- **fbs (fasting blood sugar) :** La glycémie à jeun (fbs) indique si le taux de glucose sanguin dépasse 120 mg/dL. Une glycémie élevée est un signe de trouble métabolique, tel que le diabète, qui constitue un facteur de risque majeur pour les maladies cardiovasculaires. Une glycémie à jeun normale est inférieure ou égale à 120 mg/dL, tandis qu'une valeur supérieure traduit un risque accru.
- **restecg (resting electrocardiographic results) :** Le résultat de l'électrocardiogramme au repos (restecg) évalue l'activité électrique du cœur et permet de détecter d'éventuelles anomalies telles que des modifications du segment ST-T ou une hypertrophie ventriculaire. En cardiologie, un ECG normal n'exclut pas totalement la présence d'une maladie, mais un tracé anormal renforce la suspicion de maladie coronarienne.
- **thalach (maximum heart rate achieved) :** La fréquence cardiaque maximale atteinte (thalach) correspond au nombre maximal de battements par minute enregistrés lors d'un test d'effort. Elle mesure la réponse chronotrope du cœur et la capacité du patient à fournir un effort. En cardiologie, une fréquence cardiaque maximale faible peut indiquer un risque accru de maladie coronarienne. La valeur théorique maximale est approximée par la formule 220 moins l'âge du patient.
- **exang (exercise induced angina) :** L'angine induite par l'effort (exang) indique la présence de douleurs thoraciques pendant l'exercice physique. En cardiologie, la présence d'une angine à l'effort est un signe classique de maladie coronarienne obstructive, et elle est souvent confirmée par des tests d'effort ou d'imagerie cardiaque.
- **oldpeak :** Le paramètre oldpeak mesure la dépression du segment ST (en millimètres) observée à l'électrocardiogramme pendant un effort par rapport au repos. En cardiologie, une dépression du ST supérieure à 1 mm est généralement considérée comme anormale et prédictive d'une maladie coronarienne significative. Une valeur normale est proche de 0 mm, tandis qu'une dépression plus importante est liée à un risque accru.

— **slope :** La pente du segment ST au pic de l'effort (slope) décrit la forme du tracé ECG après un exercice. Une pente ascendante est généralement normale, tandis qu'une pente plate ou descendante suggère une ischémie myocardique. Les valeurs normales correspondent donc à une pente ascendante.

— **ca (number of major vessels) :** L'attribut ca correspond au nombre de vaisseaux coronaires majeurs colorés par fluoroscopie (Bloqués).Le nombre varie de 0 à 3. Un nombre plus élevé indique une atteinte plus diffuse et une maladie plus grave. En cardiologie, la valeur normale est 0, correspondant à l'absence de vaisseau obstrué.

— **thal (thalassemia) :** Le paramètre thal provient du test de perfusion au thallium, utilisé pour détecter les anomalies de perfusion myocardique. En cardiologie, cette variable est précieuse pour le diagnostic et le suivi de la maladie coronarienne.

— **target :** La variable target représente le diagnostic final de maladie cardiaque. C'est la variable cible du modèle de classification.

2.2 Analyse descriptive (avant nettoyage)

2.2.1 Idée sur le contenu du Dataset

On visualise les cinq premiers enregistrements de notre dataset, afin de mieux avoir une idée sur son contenu :

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0

FIGURE 2.1 – Les cinq premières lignes

Après avoir afficher quelques lignes, on remarque que la pluspart des variables sont déjà binairisées, on décode chaque attribut encodé :

— **Sexe :** Valeurs : 0 = femme, 1 = homme.

— **Chest pain (cp) :** Dans le jeu de données, la variable est codée de 0 à 3, représentant respectivement : angine typique, angine atypique, douleur non-angineuse et asymptomatique. L'absence de douleur (asymptomatique) est considérée comme la situation normale.

— **Glycémie à jeun (fbs) :** représenté par une valeur binaire : 1 pour oui, 0 pour non.

— **Résultat électrocardiographique au repos (restecg) :** Dans le dataset, la variable prend les valeurs suivantes : 0 pour un ECG normal, ECG anormal : 1 pour une anomalie du segment ST-T, 2 pour une hypertrophie ventriculaire gauche probable.

— **Angine induite par l'effort (exang) :** Elle est codée comme suit : 1 si une angine apparaît, 0 sinon.

— **Pente du segment ST (slope) :** L'encodage utilisé est le suivant : 0 pour une pente ascendante, 1 pour une pente plate, et 2 pour une pente descendante. Donc, c'est bon d'avoir 0, or c'est anormal d'avoir 1 ou 2.

— **Thalassémie / défaut de perfusion (thal) :** L'encodage est le suivant : 0 pour normal, anormal : 1 pour défaut fixe ou 2 pour défaut réversible.

— **Présence de maladie cardiaque (target) :** Elle est codée comme suit : 1 pour la présence de maladie cardiaque et 0 pour son absence.

2.2.2 La homogénisation des types de données

L'homogénéisation des types de données permet d'éviter les erreurs de calcul et d'assurer que les algorithmes interprètent correctement les variables, garantissant ainsi la cohérence et la fiabilité du dataset tout en facilitant son traitement.

Afin d'obtenir les types des données de notre dataset, on exécute :

```
print("\nTypes de donnees :")
print(df.dtypes)
```

Voici le résultat obtenu :

Variable	Type de données
age	int64
sex	int64
cp	int64
trestbps	int64
chol	int64
fbs	int64
restecg	int64
thalach	int64
exang	int64
oldpeak	float64
slope	int64
ca	int64
thal	int64
target	int64

TABLE 2.1 – Types de données des variables du jeu de données

Le jeu de données comprend principalement des variables entières représentant des mesures ou catégories codées (âge, sexe, type de douleur, etc.) et une variable réelle, oldpeak, mesurant la dépression du segment ST à l'effort. La variable target indique la présence ou non d'une maladie cardiaque. Cette structure, majoritairement discrète avec une seule variable continue, convient bien aux algorithmes de classification supervisée comme la régression logistique ou les arbres de décision.

2.2.3 Les paramètres statistiques

Dans le but de consulter les statistiques de base de notre jeu de données, on utilise le code suivant :

```
print("\nStatistiques descriptives :")
display(df.describe().round(2))
```

On obtient alors le résultat suivant :

	count	mean	std	min	25%	50%	75%	max
age	1025.0	54.43	9.07	29.0	48.0	56.0	61.0	77.0
sex	1025.0	0.70	0.46	0.0	0.0	1.0	1.0	1.0
cp	1025.0	0.94	1.02	0.0	0.0	1.0	2.0	3.0
trestbps	1025.0	131.61	17.52	94.0	120.0	130.0	140.0	200.0
chol	1025.0	246.0	51.59	126.0	211.0	240.0	275.0	564.0
fbs	1025.0	0.15	0.36	0.0	0.0	0.0	0.0	1.0
restecg	1025.0	0.53	0.53	0.0	0.0	1.0	1.0	2.0
thalach	1025.0	149.11	23.01	71.0	132.0	152.0	166.0	202.0
exang	1025.0	0.34	0.47	0.0	0.0	0.0	1.0	1.0
oldpeak	1025.0	1.07	1.18	0.0	0.0	0.8	1.8	6.2
slope	1025.0	1.39	0.62	0.0	1.0	1.0	2.0	2.0
ca	1025.0	0.77	1.07	0.0	0.0	0.0	1.0	4.0
thal	1025.0	2.32	0.62	0.0	2.0	2.0	3.0	3.0
target	1025.0	0.51	0.50	0.0	0.0	1.0	1.0	1.0

TABLE 2.2 – Statistiques descriptives des variables

Avant toute opération de prétraitement, les statistiques descriptives permettent d'obtenir une vue d'ensemble du jeu de données. L'âge moyen des patients est d'environ 54 ans, avec un écart type de 9 ans, ce qui montre une population adulte majoritairement d'âge moyen. La variable sex indique une proportion plus élevée d'hommes (moyenne de 0,70, sachant que 1 = homme et 0 = femme). Les mesures cliniques présentent une variabilité notable : la pression artérielle au repos (trestbps) a une moyenne de 131 mmHg, tandis que le taux de cholestérol (chol) est en moyenne de 246 mg/dl, avec une valeur maximale atteignant 564 mg/dl, suggérant la présence de valeurs aberrantes. La fréquence cardiaque maximale (thalach) atteint en moyenne 149 bpm, mais varie fortement (écart type de 23), reflétant des différences physiologiques importantes entre individus. La variable oldpeak, représentant la dépression du segment ST à l'effort, présente une moyenne de 1,07 mais un maximum de 6,2, ce qui confirme également la présence de valeurs extrêmes. Enfin, la variable cible target a une moyenne de 0,51, indiquant une répartition relativement équilibrée entre les patients malades (1) et sains (0). Dans l'ensemble, ces statistiques montrent un dataset globalement complet, mais contenant certaines valeurs extrêmes, notamment pour chol, oldpeak et ca, qui nécessitent un nettoyage ou une normalisation avant l'étape de modélisation.

2.2.4 Visualisation graphique initiale

Pour une visualisation claire et simple, nous utilisons les histogrammes, qui permettent de représenter la distribution d'une variable en regroupant les valeurs en intervalles et en

affichant le nombre d'occurrences dans chaque intervalle. Ils offrent ainsi un aperçu rapide de la répartition de chacun des attributs, facilitant la compréhension de leur comportement et la comparaison entre les variables.

Le code suivant est lui qui nous a permis de tracer les histogrammes traités par la suite :

```
df.hist(figsize=(14, 10), bins=20, color='red', edgecolor='black')
plt.suptitle("Distribution des variables (avant nettoyage)",
    fontsize=16, fontweight='bold')
plt.tight_layout()
plt.show()
```

On passe à l'analyse les diagrammes créés :

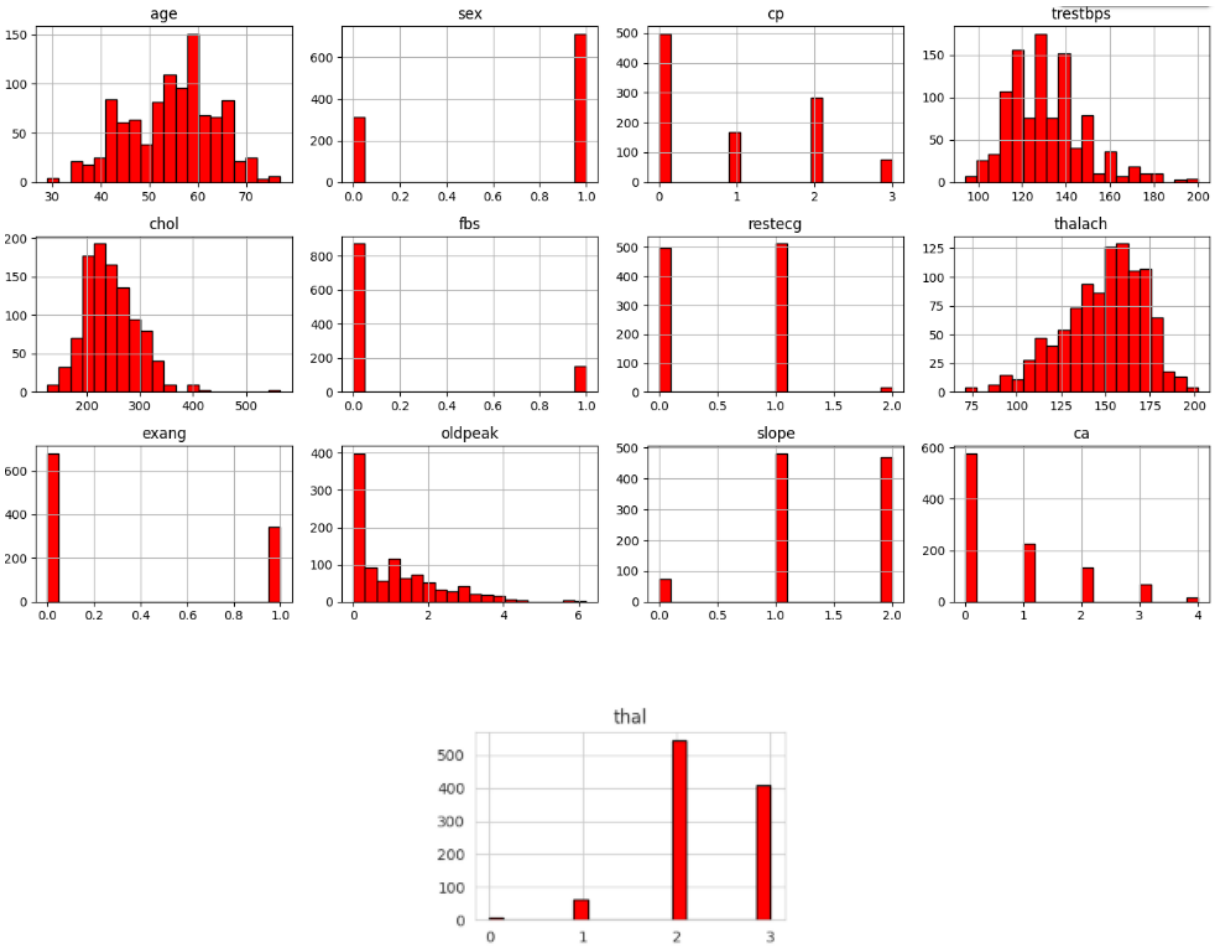


FIGURE 2.2 – Regroupement vertical des visualisations avant nettoyage

Intérprétation des graphs :

— **Distribustion de l'âge :** La distribution de l'âge des patients montre une population majoritairement âgée de 50 à 60 ans, avec un pic autour de 55 ans. La courbe présente une forme approximativement normale, s'étendant de 30 à 75 ans. Cette

concentration dans la tranche d'âge moyen-supérieur est cohérente avec l'épidémiologie des maladies cardiovasculaires, qui touchent plus fréquemment les personnes après 50 ans. La présence de patients plus jeunes (30-40 ans) et plus âgés (70+) assure une certaine diversité dans l'échantillon.

— **Distribution des types de douleur thoracique (cp) :** L'analyse de la distribution des types de douleur thoracique révèle une nette prédominance du type 0 (angine typique), qui représente près de la moitié des cas. Vient ensuite le type 2 (douleur non-angineuse) avec environ 30% des patients, tandis que les types 1 (angine atypique) et 3 (asymptomatique) sont nettement moins représentés, avec environ 10-15% chacun. Cette distribution est cliniquement importante car elle montre que la majorité des patients présentent des symptômes douloureux classiques, facilitant le diagnostic suspecté. Cependant, la présence non négligeable de patients asymptomatiques (type 3) souligne le défi du diagnostic des formes silencieuses de la maladie coronarienne, qui nécessitent des investigations plus poussées.

— **La pression artérielle (trestbps) :** La pression artérielle au repos montre une distribution quasi-normale centrée autour de 130-140 mmHg, correspondant à une pression artérielle limite ou légèrement élevée. L'étendue de 100 à 200 mmHg couvre une large gamme de profils tensionnels, des valeurs normales aux hypertension sévères, offrant une bonne variabilité pour l'analyse.

— **Le taux de cholestérol (chol) :** La distribution du cholestérol présente une asymétrie positive avec une concentration principale entre 200 et 300 mg/dl, et un pic marqué autour de 240-250 mg/dl. La présence de valeurs supérieures à 400 mg/dl suggère des cas d'hypercholestérolémie sévère, potentiellement outliers. Cette distribution justifie une analyse plus approfondie pour identifier les valeurs aberrantes avant modélisation.

— **La glycémie à jeun (fbs) :** La grande majorité des patients (environ 85%) présente une glycémie à jeun normale (fbs = 0), tandis qu'une minorité (environ 15%) a une glycémie élevée (>120 mg/dl). Cette distribution reflète la prévalence attendue dans une population générale, où l'hyperglycémie à jeun est moins commune. Le déséquilibre entre les classes devra être considéré dans l'analyse.

— **L'électrocardiogramme au repos (restecg) :** Les résultats ECG montrent trois catégories avec une répartition inégale. La catégorie 1 (anomalies onde ST-T) est la plus fréquente, suivie de la catégorie 0 (normal). La catégorie 2 (hypertrophie ventriculaire) est la moins représentée. Cette distribution suggère que la plupart des patients présentent déjà des anomalies cardiaques détectables au repos.

— **La fréquence cardiaque maximale (thalach) :** La fréquence cardiaque maximale atteinte présente une distribution étalée avec un pic autour de 150-160 bpm. La plage s'étend d'environ 70 à 200 bpm, couvrant à la fois des patients avec faible capacité cardiaque et d'autres avec une bonne réponse à l'effort. Cette variabilité est intéressante pour discriminer les profils cardiaques.

— **L'angine induite à l'effort (exang) :** Environ 2/3 des patients ne présentent pas d'angine à l'effort (exang = 0), tandis qu'1/3 en développe (exang = 1). Ce ratio est cliniquement significatif car l'angine d'effort est un symptôme classique de l'insuffisance coronarienne.

— **La dépression ST (oldpeak) :** La distribution de l'oldpeak est fortement asymétrique, avec une concentration marquée autour de 0-1 (absence ou faible dépression

ST). La queue de distribution s'étend jusqu'à 4-5, indiquant quelques cas sévères d'ischémie à l'effort. Cette variable semble très discriminante pour la maladie coronarienne.

— **La pente du segment ST (slope) :** Les trois types de pente sont présents avec une nette prédominance de la pente 1 (plate). Les pentes ascendante (2) et descendante (0) sont moins fréquentes. En cardiologie, la pente descendante est souvent associée à un pronostic moins favorable.

— **Les vaisseaux obstrués (ca) :** La majorité des patients ont 0 ou 1 vaisseau majoritaire obstrué, tandis que les obstructions multiples (2-3 vaisseaux) sont moins communes. Cette distribution reflète la progression naturelle de la maladie coronarienne, où les formes les plus sévères sont statistiquement moins fréquentes.

— **La thalassémie (thal) :** La variable thal montre une distribution tri-modale avec une prédominance du type 2 (défaut réversible), suivi du type 3 (normal) et du type 1 (défaut fixe). Le type 0 est absent ou très rare. Le défaut réversible étant souvent associé à l'ischémie, cette distribution est cohérente avec une population à risque cardiaque.

2.3 Évaluation de la qualité des données

Avant de procéder à la phase de modélisation, il est essentiel de s'assurer de la qualité des données. L'analyse statistique et les visualisations précédentes mettent en évidence la présence potentielle de valeurs aberrantes, dont certaines observations dépassent nettement les bornes usuelles. Ces valeurs extrêmes peuvent biaiser les calculs statistiques et la performance des modèles prédictifs, justifiant ainsi un traitement spécifique (suppression ou transformation). Par ailleurs, bien que le jeu de données semble globalement complet, une vérification approfondie de la présence éventuelle de valeurs manquantes ou de doublons est nécessaire afin de garantir la cohérence et la fiabilité des analyses. En fonction des résultats de ce contrôle de qualité, des actions appropriées seront entreprises : suppression des doublons, imputation des valeurs manquantes ou normalisation des variables quantitatives. Cette étape de nettoyage constitue une phase cruciale pour obtenir un dataset propre et prêt à être utilisé pour l'apprentissage automatique.

Conclusion

L'évaluation de la qualité des données met en évidence la présence potentielle de valeurs aberrantes, de variables extrêmes et la nécessité de vérifier les valeurs manquantes ou les doublons. Ces éléments doivent être traités avant la modélisation afin d'assurer la fiabilité et la robustesse des modèles prédictifs. Cette étape de nettoyage et de préparation des données constitue un prérequis indispensable pour garantir des résultats cohérents et exploitables lors de l'apprentissage automatique.

Exploration et Pré-traitement

Introduction

Le chapitre d'exploration et de pré-traitement a permis d'analyser le dataset, de créer des variables dérivées comme risk_score et chol_per_age, et d'identifier les facteurs cliniques les plus discriminants (thalach, oldpeak, ca). Les données sont équilibrées, sans valeurs manquantes, et prêtes pour la modélisation.

3.1 Chargement et Structure du Dataset

Le dataset est chargé depuis Google Drive. Il contient 1025 patients et 14 variables.

```
df = pd.read_csv(csv_path)
df.dataframeName = 'heart.csv'
print(f"Dataset : {df.shape[0]} lignes x {df.shape[1]} colonnes")
```

3.1.1 Répartition de la Cible

À l'aide du code suivant, on a pu avoir une idée sur la répartition des deux classes (sain = 0 et malade = 1) :

```
print(df['target'].value_counts(normalize=True).round(2))
```

Classe	Proportion
Sain (0)	49 %
Malade (1)	51 %

TABLE 3.1 – Distribution des classes dans le jeu de données

Le dataset est parfaitement équilibré avec 51 % de patients malades (target = 1) et 49 % de patients sains (target = 0). Cette répartition évite tout biais lié à un déséquilibre de classes et garantit une évaluation fiable des modèles de classification. Aucune technique de rééquilibrage (SMOTE, undersampling) n'est nécessaire.

3.1.2 Les valeurs manquantes, aberrantes et les doublons

Une vérification des données est nécessaire afin d'identifier la présence éventuelle de valeurs manquantes, aberrantes et des doublons susceptibles d'influencer les analyses statistiques. Et cela est réalisé à travers les fonctions suivantes :

```
print(df.isnull().sum().sum())
```

```
if df.duplicated().sum() > 0:
    print("Suppression des doublons...")
    df = df.drop_duplicates()
    print(f"Nouveau shape : {df.shape}")
else:
    print("Aucun doublon : parfait !")
```

```
z_scores =
   np.abs(stats.zscore(df.select_dtypes(include=[np.number])))
df_clean = df[(z_scores < 3).all(axis=1)]
```

Aucune valeur manquante n'a été détectée dans le jeu de données, ce qui garantit une base de travail complète et évite la nécessité d'appliquer des techniques d'imputation ou de nettoyage supplémentaires. Concernant les doublons, on remarque que cette vérification a détécté 723 lignes répétées. En outre, 15 enregistrements aberrants ont été retenus et éliminées, du coup la structure de notre dataset devient de 287 lignes et 14 colonnes.

3.2 Enrichissement du Dataset (`df2`)

L'ajout de variables dérivées, ou feature engineering, constitue une étape clé pour enrichir le jeu de données et améliorer la qualité de l'analyse. En transformant ou combinant les variables existantes, on obtient des caractéristiques plus pertinentes, facilitant la détection de relations complexes et l'optimisation des performances des modèles prédictifs.Pour cela, nous créons `df2` afin d'ajouter des variables interprétables.

Quatre variables ont été ajoutées :

— **age_group :** Cette variable regroupe les âges en quatre tranches : moins de 45 ans, 45–55 ans, 55–65 ans et plus de 65 ans.

— **chol_per_age :** Cette variable est calculée en divisant le taux de cholestérol de chaque individu par son âge.

— **risk_score :** Cette variable est la somme de trois indicateurs : présence de douleur thoracique, angine à l'effort et dépression du segment ST > 1 mm, limitée entre 0 et 3.

— **target_label :** Cette variable convertit la valeur binaire de la cible en étiquettes textuelles "Sain" ou "Malade".

Variable	Description	Type
age_group	Regroupement par tranche d'âge (ex. 45–55)	Catégorielle
chol_per_age	Ratio cholestérol / âge (mg/dl par année)	Numérique
risk_score	Score synthétique de 0 à 3 basé sur 3 facteurs	Entier
target_label	Étiquette lisible : "Sain" ou "Malade"	Texte

TABLE 3.2 – Variables dérivées ajoutées à df2

```
df2 = df.copy()
df2['age_group'] = pd.cut(df['age'], bins=[0,45,55,65,100],
   labels=['<45','45-55','55-65','>65'])
df2['chol_per_age'] = df2['chol'] / df2['age']
df2['risk_score'] = ((df2['cp']>0) + (df2['exang']==1) +
   (df2['oldpeak']>1)).clip(0,3)
df2['target_label'] = df2['target'].map({0:'Sain',1:'Malade'})
```

Les variables dérivées ajoutées à df2 visent à enrichir l'analyse et à faciliter l'interprétation des résultats. La variable age_group segmente les individus en tranches d'âge afin de mettre en évidence des tendances démographiques et générationnelles liées au risque cardiaque. chol_per_age exprime le ratio cholestérol/âge, permettant une comparaison normalisée des profils et la détection d'anomalies relatives plutôt qu'absolues. risk_score propose un indicateur synthétique de 0 à 3 fondé sur trois facteurs clés, offrant une évaluation rapide du risque cardiaque. Enfin, target_label transforme la variable cible binaire en étiquettes textuelles, rendant les graphiques et rapports plus lisibles et accessibles.

3.3 Analyses Exploratoires Clés

3.3.1 Score de Risque Cardiaque

Le risk_score est un indicateur synthétique de risque cardiaque (valeur de 0 à 3) construit à partir de trois facteurs cliniques majeurs. Afin de visualiser un attribut catégoriel comme le risk_score, le diagramme en barres est particulièrement adapté. Il permet de comparer facilement la fréquence des différentes catégories et de repérer rapidement les tendances ou anomalies dans les données, ce qui le rend plus lisible et intuitif que d'autres types de graphiques pour ce type d'information.

```
repartition = df2['risk_score'].value_counts().sort_index()
pourcentage =
   (df2.groupby('risk_score')['target'].mean()*100).round(0)
```

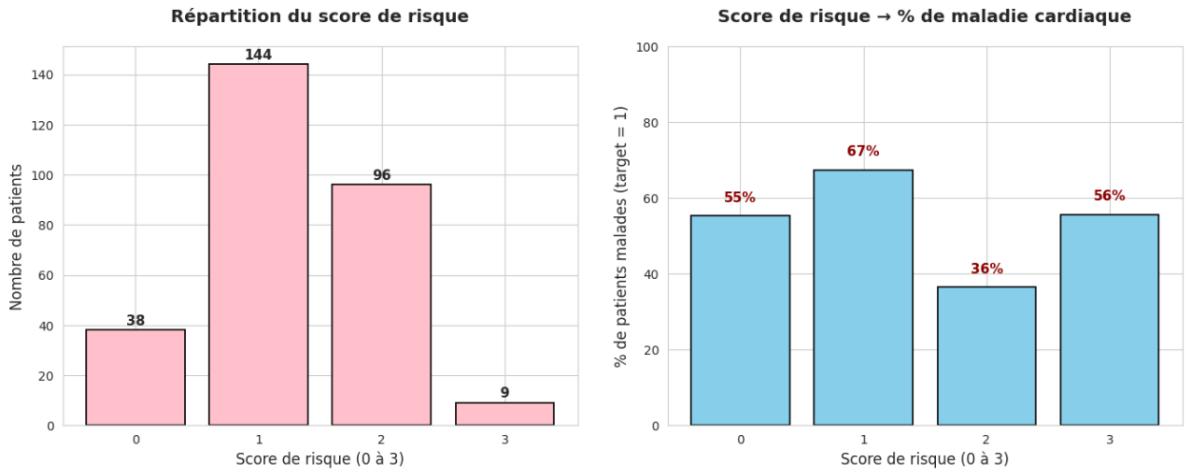


FIGURE 3.1 – Répartition et % de malades par `risk_score`

Le risk_score, bien que simple et interprétable, ne présente pas une relation monotone avec la présence de la maladie cardiaque. Bien que le score 1 soit le plus fréquent (144 patients), il n'est associé qu'à 67 % de cas positifs. Paradoxalement, le score 2 correspondant à deux symptômes ne concerne que 36 % de malades, suggérant une possible atténuation du risque par des facteurs compensatoires (ex. traitement, condition physique). Le score 3, rare (9 patients), reste associé à un risque élevé (56 %). Ce score synthétique est donc utile comme signal d'alerte précoce mais insuffisant seul pour un diagnostic fiable. Il devra être complété par des variables physiologiques objectives (thalach, ca, oldpeak) dans le modèle prédictif final.

3.3.2 Fréquence Cardiaque Maximale (`thalach`)

Le boxplot permet de visualiser la distribution de la fréquence cardiaque maximale (thalach) selon l'état de santé. Il montre que les patients malades tendent à avoir une fréquence cardiaque maximale plus basse que les patients sains, une relation qui peut sembler contre-intuitive. Cette visualisation met en évidence les différences de dispersion et de médiane entre les groupes, tout en rappelant que cette variable doit être interprétée dans son contexte clinique et ne suffit pas seule pour le diagnostic.

```
sns.boxplot(data=df2, x='target', y='thalach', palette='Set2')
plt.xticks([0,1],['Sain','Malade']); plt.grid(True, alpha=0.3)
```

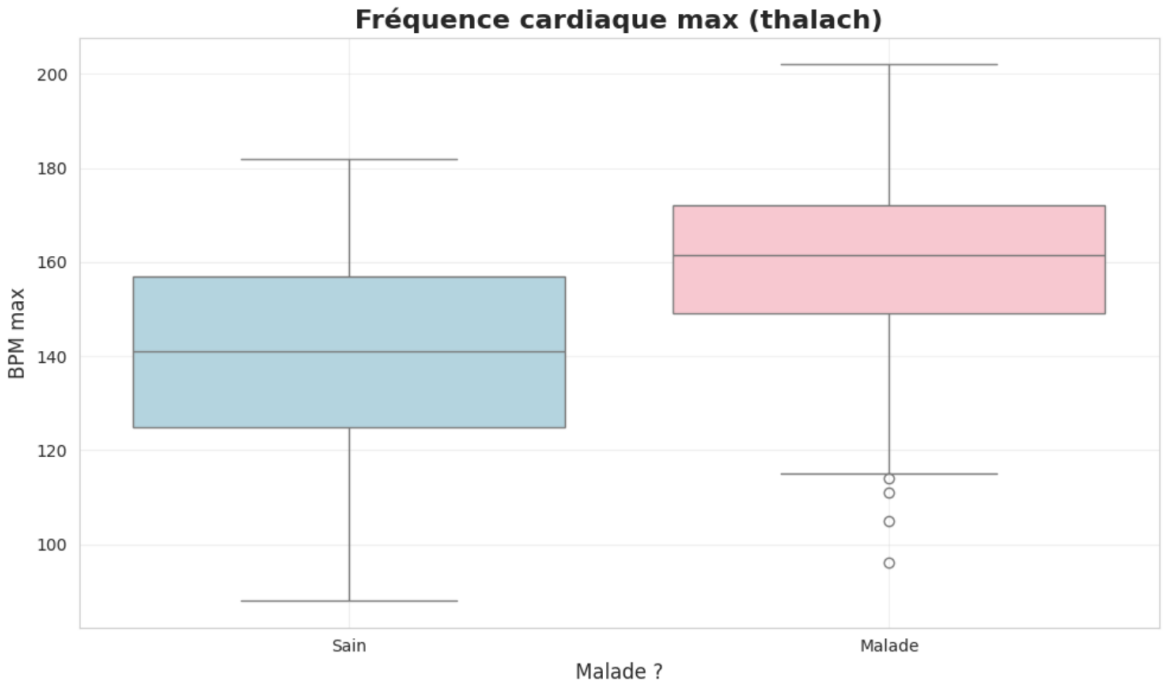


FIGURE 3.2 – Les malades ont une `thalach` plus basse

Le graphique révèle une relation contre-intuitive mais cliniquement significative : les patients sains présentent une fréquence cardiaque maximale plus basse que les patients malades. Cette observation semble paradoxale mais peut s'expliquer par plusieurs facteurs physiologiques. Les patients sous traitement bêta-bloquants (médicaments courants pour les maladies cardiaques) peuvent avoir une fréquence cardiaque artificiellement limitée, tandis que les patients non traités ou présentant d'autres pathologies peuvent maintenir une réponse cardiaque exagérée à l'effort. Cette variable, bien que discriminante, nécessite donc une interprétation contextuelle et ne peut être utilisée isolément pour le diagnostic.

3.3.3 Âge vs Cholestérol

Le scatter plot (ou nuage de points) est un graphique permettant de représenter deux variables numériques simultanément, en plaçant chaque observation sous forme de point selon ses valeurs. Il est particulièrement utile pour détecter des relations, des tendances ou des regroupements entre les variables. Dans notre cas, le scatter plot permet de visualiser la relation entre l'âge et le taux de cholestérol pour chaque patient car il est le plus adapté.

```
plt.scatter(df2['age'], df2['chol'], c=df2['target'],
   cmap='RdYlGn', alpha=0.7, edgecolor='k')
plt.colorbar(label='0=Sain | 1=Malade')
```

Listing 3.1 – Scatter : Âge vs Cholestérol

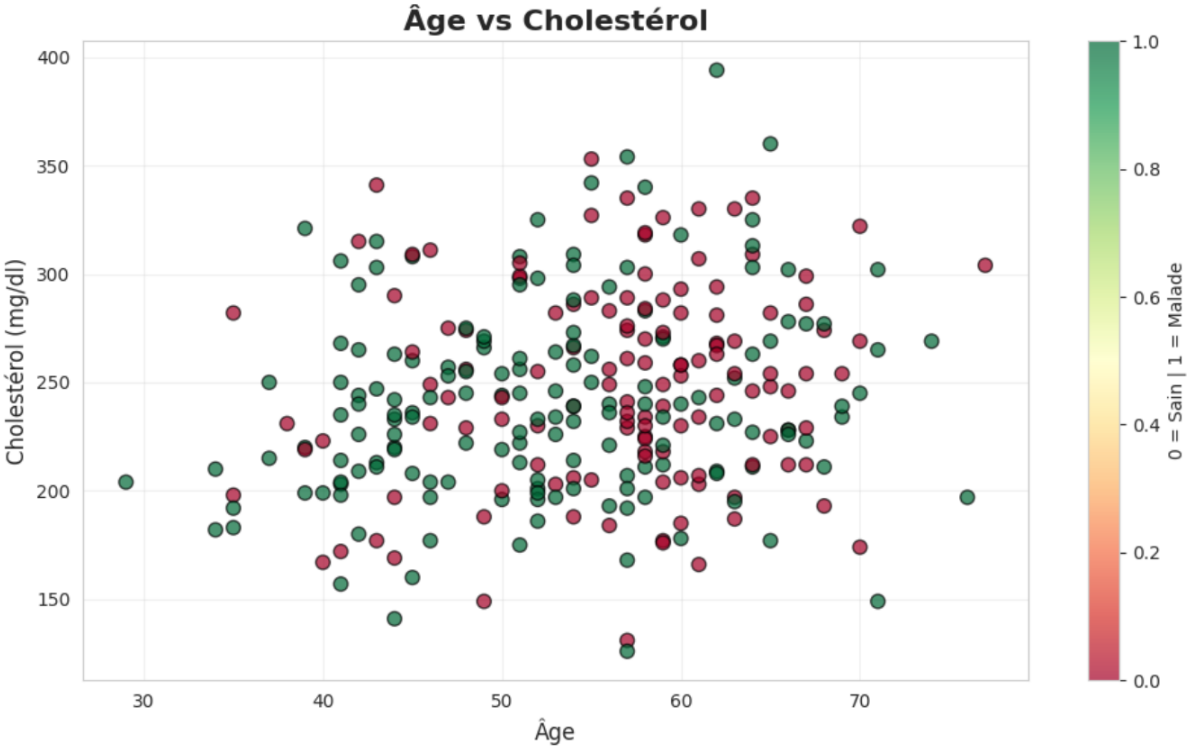


FIGURE 3.3 – Malades = zone vert

Le graphique montre une relation complexe entre l'âge et le taux de cholestérol. On n'observe pas de corrélation linéaire évidente entre ces deux variables : le cholestérol ne semble pas augmenter systématiquement avec l'âge dans cette population. La distribution apparaît plutôt en *nuage de points* avec des concentrations à différents niveaux :

— Des patients jeunes (20-40 ans) présentent déjà des cholestérols élevés (>250 mg/dl).

— Des patients plus âgés (50-70 ans) maintiennent des cholestérols dans la normale (200 mg/dl).

— La variabilité est importante dans toutes les tranches d'âge.

Cela suggère que le cholestérol est influencé par de multiples facteurs au-delà de l'âge (génétique, mode de vie, traitements), et qu'un cholestérol élevé peut toucher des patients jeunes, renforçant ainsi l'importance du dépistage précoce.

3.3.4 Profils Cliniques : Coordonnées Parallèles

Les coordonnées parallèles sont une méthode de visualisation multivariée permettant de représenter simultanément plusieurs variables numériques sur des axes parallèles. Chaque observation est tracée comme une ligne brisée reliant ses valeurs normalisées sur chaque axe.

```
cols =
   ['age','trestbps','chol','thalach','oldpeak','ca','cp','slope']
parallel_coordinates(df2[cols+['target_label']], 'target_label',
   color=['green','red'], alpha=0.6)
```

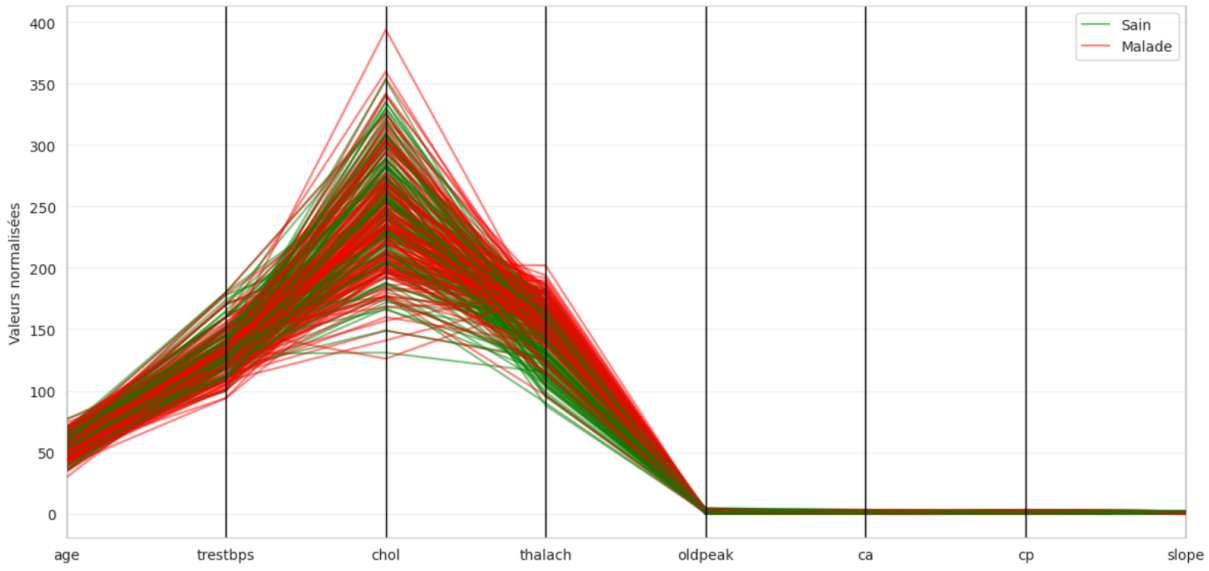


FIGURE 3.4 – Sains (vert) vs Malades (rouge) : profils opposés

Le diagramme de coordonnées parallèles met en évidence des profils cliniques nettement différenciés entre patients sains et malades, révélant que les variables les plus discriminantes sont la fréquence cardiaque maximale (thalach), la dépression du segment ST (oldpeak) et le nombre de vaisseaux obstrués (ca), où les patients malades présentent systématiquement des valeurs plus défavorables. En revanche, l'âge, la pression artérielle (trestbps) et le cholestérol (chol) montrent un chevauchement important entre les deux groupes, limitant leur pouvoir prédictif isolé. Cette analyse multivariée souligne l'importance d'une approche intégrée combinant plusieurs paramètres pour un diagnostic fiable, plutôt que de se fier à des indicateurs individuels.

3.3.5 Corrélations Clés

La corrélation mesure la relation linéaire entre deux variables. Elle est quantifiée par le coefficient de Pearson r, compris entre -1 et $+1$:

$$r_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}$$

où $|r| > 0.4$ indique une corrélation forte, $r > 0$ une relation positive, $r < 0$ une relation négative.

```
top =
   ['cp','thalach','oldpeak','exang','ca','risk_score','target']
sns.heatmap(df2[top].corr(), annot=True, cmap='coolwarm',
   center=0, fmt='.2f')
sns.pairplot(df2, hue='target',
   vars=['age','chol','thalach','oldpeak'], diag_kind='kde')
```

Listing 3.2 – Heatmap Pairplot

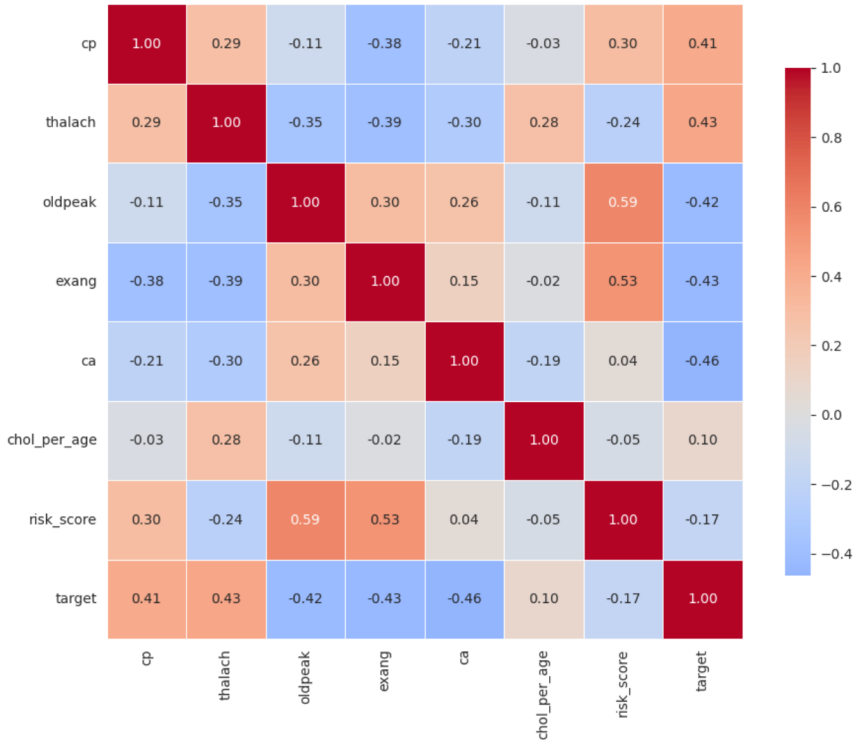


FIGURE 3.5 – Heatmap des corrélations — Variables fortement corrélées

La heatmap met en évidence les prédicteurs clés de la maladie cardiaque :

- `thalach` (*fréquence cardiaque maximale atteinte*, +0.43) et `cp` (*type de douleur thoracique*, +0.41) sont les plus positivement corrélés à la présence de maladie (`target`),
- `ca` (*nombre de vaisseaux majeurs obstrués*, -0.46) et `exang` (*angine à l'effort*, -0.43) présentent les corrélations négatives les plus marquées,
- `oldpeak` (*dépression du segment ST à l'effort*, -0.43) renforce son rôle comme indicateur de risque,
- En revanche, `chol_per_age` (*cholestérol ajusté par âge*, +0.10) reste très faiblement lié à la cible.

Les variables liées à l'effort et à la réponse cardiaque dominent largement les mesures statiques.

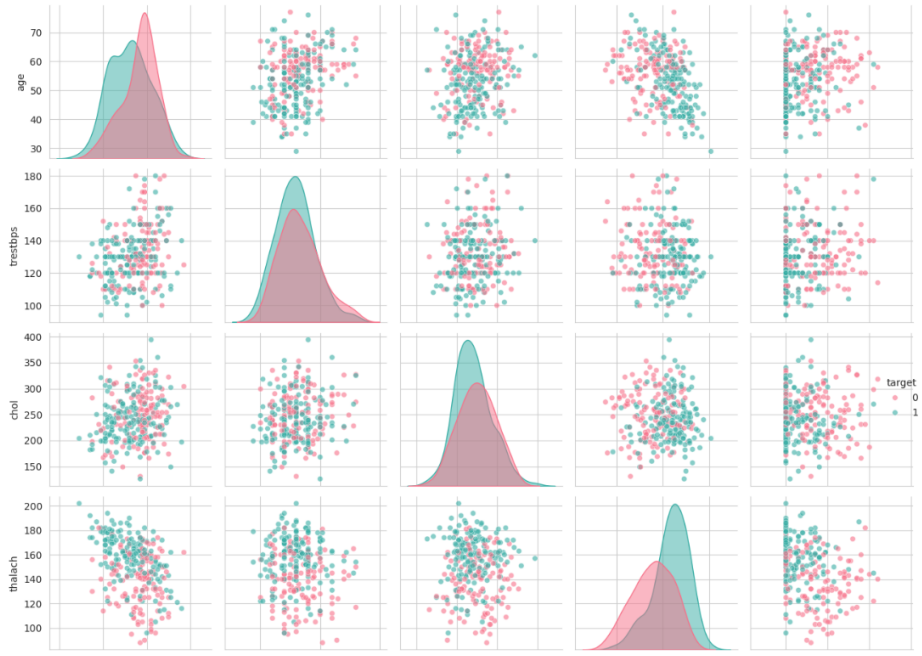


FIGURE 3.6 – Pairplot des variables clés — Séparation visuelle entre classes

Le pairplot confirme une séparation nette entre classes grâce aux paires suivantes :

— `thalach`/`oldpeak` : les patients malades ont une *fréquence cardiaque maximale* (`thalach`) plus basse et une *dépression ST à l'effort* (`oldpeak`) plus élevée,

— `âge`/`thalach` : les sujets âgés atteignent des fréquences cardiaques maximales plus faibles, surtout en cas de maladie,

— `chol` (*cholestérol sérique*) montre une forte dispersion sans structure claire par classe.

Les variables dynamiques (réponse à l'effort, douleur) surpassent largement les mesures statiques (cholestérol) en pouvoir discriminant.

Conclusion

Le dataset final est propre et enrichi, avec des variables pertinentes pour prédire la maladie cardiaque. La standardisation des variables numériques est recommandée pour assurer la robustesse des modèles de classification.

Modélisation

4.1 Introduction

Après le pré-traitement, plusieurs modèles de classification supervisée, puisqu'il s'agit d'un dataset labelisé, ont été entraînés pour prédire la présence d'une maladie cardiaque. Les données ont été séparées en ensembles d'apprentissage (80 %) et de test (20 %), permettant d'évaluer la capacité de généralisation des modèles. Les algorithmes testés incluent la Régression Logistique, Naïve Bayes, KNN, SVM, Arbres de Décision, Forêt Aléatoire et XGBoost.

4.2 CRISP-DM : Modeling

Après le nettoyage du jeu de données, plusieurs modèles de classification ont été entraînés pour prédire la présence d'une maladie cardiaque. Les données ont été séparées en ensembles d'apprentissage et de test, et les modèles (Régression Logistique, Naïve Bayes, KNN, SVM, Arbres de Décision, Forêts Aléatoires et XGBoost) ont été évalués via l'accuracy, le F1-score et l'AUC afin d'identifier le classifieur le plus performant.

4.3 Régression Logistique

La régression logistique est un modèle linéaire probabiliste utilisé pour la classification binaire. Sa fonction mathématique est donnée par la sigmoïde :

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-(\mathbf{w}^T\mathbf{x}+b)}}$$

où $\mathbf{w}$ est le vecteur des poids et b le biais. Le principe consiste à maximiser la vraisemblance des données par descente de gradient, en supposant une séparation linéaire des classes dans l'espace des caractéristiques.

4.3.1 Courbe Precision-Recall

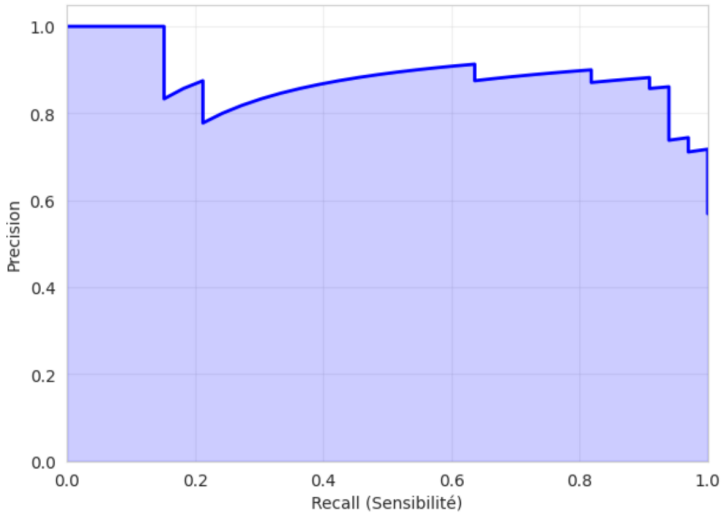


FIGURE 4.1 – Courbe Precision-Recall - Régression Logistique

La courbe Precision-Recall montre une précision élevée (> 0.8) même pour un recall modéré, avec une chute progressive en fin de courbe. Cela indique une excellente détection des cas positifs tout en limitant les faux positifs, confirmant la robustesse du modèle.

4.4 Naïve Bayes

Le classifieur Naïve Bayes est un modèle probabiliste basé sur le théorème de Bayes avec une hypothèse forte d'indépendance conditionnelle entre les variables. Sa fonction mathématique est :

$$P(y|\mathbf{x}) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(\mathbf{x})}$$

souvent calculée en logarithme pour la stabilité numérique :

$$\log P(y|\mathbf{x}) \propto \log P(y) + \sum_{i=1}^{n} \log P(x_i|y)$$

Le principe repose sur l'estimation des probabilités a priori et conditionnelles à partir des données d'entraînement, ce qui le rend rapide et efficace, notamment avec des variables catégorielles.

4.4.1 Distribution des probabilités prédites

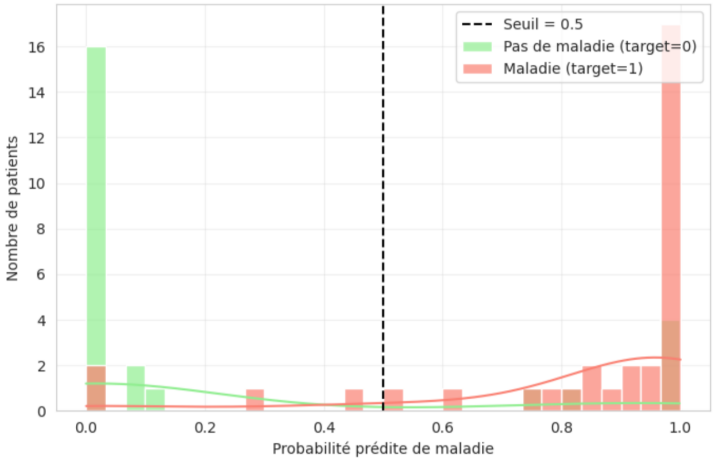


FIGURE 4.2 – Distribution des probabilités prédites - Naïve Bayes

La distribution montre une bonne séparation des classes, avec les patients sans maladie concentrés autour de 0.05 et ceux avec maladie près de 1.0. Un léger chevauchement autour du seuil 0.5 reflète l'hypothèse d'indépendance, mais la classification reste efficace.

4.5 K-Nearest Neighbors (KNN)

KNN est un algorithme non paramétrique de classification par similarité locale. Il n'y a pas de fonction mathématique explicite d'apprentissage ; la prédiction repose sur le vote majoritaire des k plus proches voisins :

$$\hat{y} = \{y_i : i \in \mathcal{N}_k(\mathbf{x})\}$$

où $\mathcal{N}_k(\mathbf{x})$ est déterminé par une distance (ex. : euclidienne). Le principe est simple : une observation est classée selon la classe dominante de ses voisins les plus proches dans l'espace des caractéristiques. Aucun entraînement, mais coût élevé en inférence.

4.5.1 Distribution des probabilités prédites

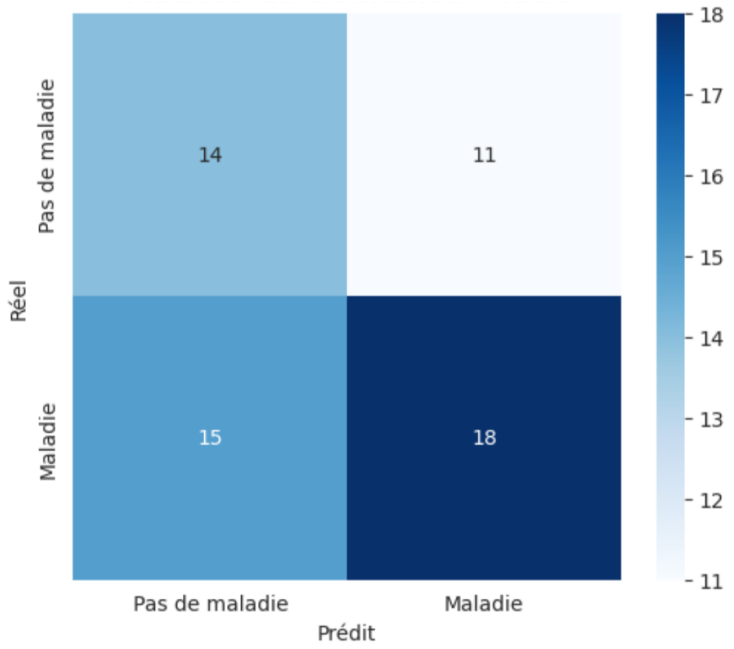


FIGURE 4.3 – Distribution des probabilités prédites - KNN

La matrice de confusion révèle une classification globalement équilibrée : le modèle reconnaît correctement 57 % des cas, mais présente une légère tendance à surestimer la présence de la maladie.

4.6 Machines à Vecteurs de Support (SVM)

Les machines à vecteurs de support (SVM) sont des modèles de classification qui cherchent l'hyperplan optimal séparant les classes avec la plus grande marge. La fonction mathématique consiste à résoudre :

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i \quad s.c. \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i$$

avec un noyau (ex. : RBF) :

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$$

Le principe repose sur la maximisation de la marge entre les classes, avec tolérance aux erreurs via les variables d'écart ξ_i et régularisation par C. Les noyaux permettent de gérer des séparations non linéaires.

4.6.1 Fronière de décision dans le plan âge–thalach

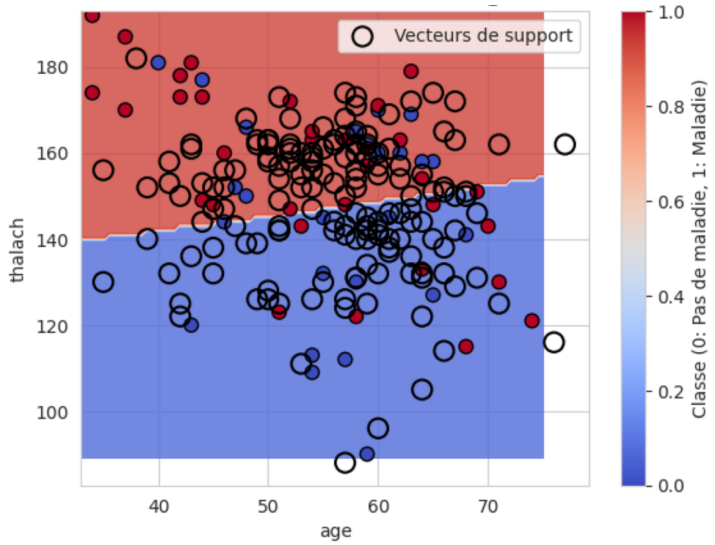


FIGURE 4.4 – Fronière de décision SVM (RBF) - age vs thalach

La frontière de décision non linéaire sépare efficacement les deux classes dans le plan âge–thalach, avec une zone rouge (maladie) pour les patients plus jeunes et à fréquence cardiaque élevée. Les vecteurs de support (cerclés) définissent la marge maximale, illustrant la robustesse du modèle SVM face aux données non linéaires.

4.7 Arbres de Décision

Un arbre de décision est un modèle hiérarchique qui partitionne récursivement l'espace des variables selon des règles de décision. La fonction mathématique utilise un critère de pureté, comme l'indice de Gini :

$$Gini(D) = 1 - \sum_{k=1}^{K} p_k^2$$

et choisit le split maximisant le gain :

$$\Delta Gini = Gini(D) - \left(\frac{|D_l|}{|D|} Gini(D_l) + \frac{|D_r|}{|D|} Gini(D_r) \right)$$

Le principe consiste à construire un arbre de règles if-then, facile à interpréter, mais sensible au surapprentissage sans élagage.

4.7.1 Matrice de confusion

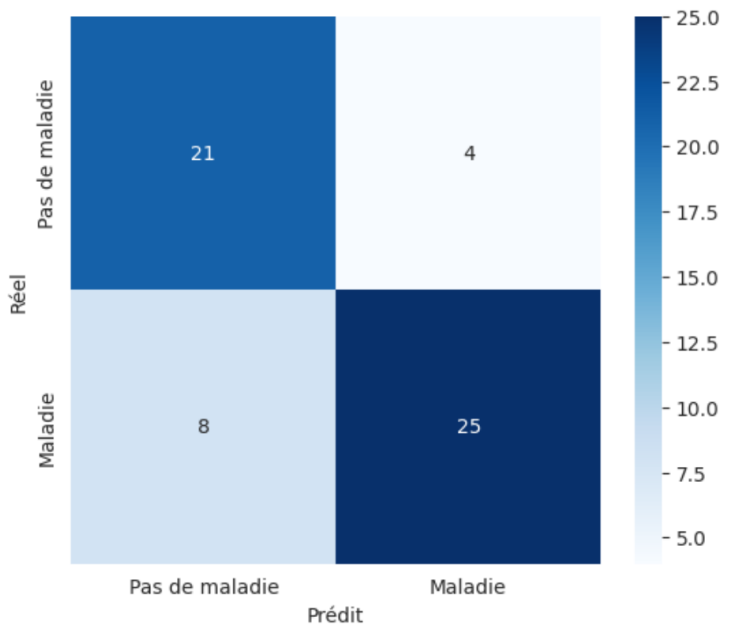


FIGURE 4.5 – Matrice de Confusion - Arbre de Décision

L'arbre de décision classe correctement 21 patients sains et 25 malades, avec seulement 4 faux négatifs et 8 faux positifs. Bonne performance équilibrée.

4.8 Forêt Aléatoire (Random Forest)

La forêt aléatoire est un modèle d'ensemble basé sur le bagging d'arbres de décision. Sa fonction mathématique agrège les prédictions de B arbres :

$$\hat{y} = \{h_b(\mathbf{x})\}_{b=1}^{B}$$

chaque arbre étant entraîné sur un échantillon bootstrap et un sous-ensemble aléatoire de variables. Le principe réduit la variance par agrégation et améliore la robustesse au bruit et au surapprentissage, tout en conservant une bonne performance générale.

4.8.1 Importance des variables

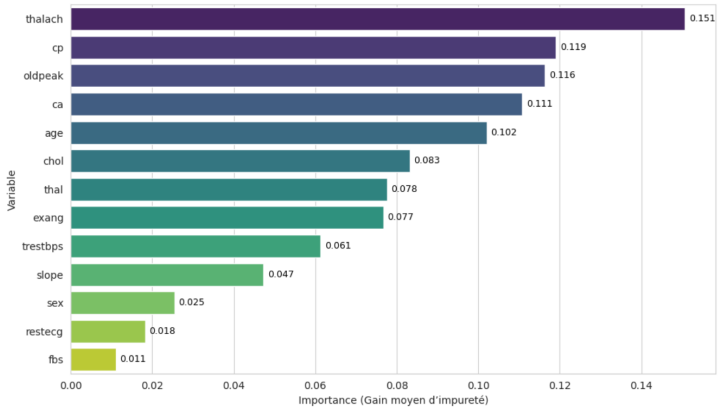


FIGURE 4.6 – Importance des variables dans la Forêt Aléatoire

La forêt aléatoire identifie `thalach` (fréquence cardiaque max), `cp` (type de douleur) et `oldpeak` comme les trois facteurs les plus prédictifs, avec une importance supérieure à 0.11. Les variables comme `fbs` et `restecg` ont un impact négligeable (< 0.02), confirmant leur faible contribution au diagnostic.

4.9 XGBoost

XGBoost est un algorithme de boosting gradientiel optimisé pour la performance et la régularisation. Sa fonction objectif à minimiser est :

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \quad \Omega(f) = \gamma T + \frac{1}{2}\lambda \|\mathbf{w}\|^2$$

avec mise à jour additive : $\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)$. Le principe repose sur l'ajout séquentiel d'arbres faibles corrigant les erreurs précédentes, avec approximation du second ordre, régularisation et sous-échantillonnage pour éviter le surapprentissage.

4.9.1 Importance des variables (poids des splits)

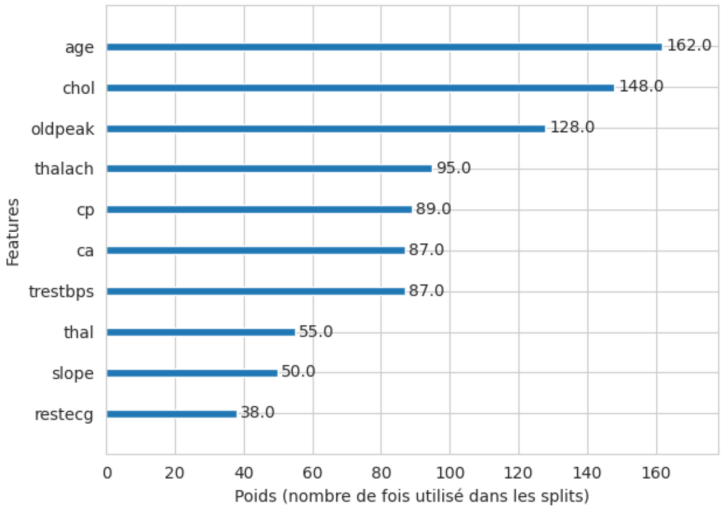


FIGURE 4.7 – Importance des variables dans XGBoost (poids des splits)

XGBoost privilégie fortement `age`, `chol` et `oldpeak` (utilisés > 120 fois dans les splits), suivis de `thalach`, `cp` et `ca`. Cette hiérarchie, optimisée par le boosting séquentiel, confirme leur rôle clé dans la prédiction du risque cardiaque.

4.10 Conclusion

XGBoost et Forêt Aléatoire dominent en performance et robustesse, identifiant `thalach`, `cp` et `ca` comme facteurs clés. La régression logistique reste idéale pour l'interprétabilité. **XGBoost est recommandé** pour une application clinique.

Évaluation de la Démarche

Introduction

Après le pré-traitement, plusieurs modèles de classification supervisée ont été entraînés pour prédire la présence d'une maladie cardiaque. Les données ont été séparées en ensembles d'apprentissage (80%) et de test (20%), permettant d'évaluer la capacité de généralisation des modèles. Les performances ont été comparées via l'accuracy, le F1-score et les courbes ROC/AUC afin de sélectionner le classifieur le plus adapté.

5.1 Séparation des données (Data Splitting)

Avant d'entraîner les modèles de classification, le jeu de données a été séparé en deux sous-ensembles : un ensemble d'apprentissage (`train`) et un ensemble de test (`test`). Cette étape permet d'évaluer la capacité de généralisation du modèle sur des données jamais vues lors de l'entraînement. La répartition choisie est de **80 %** des données pour l'entraînement et **20 %** pour le test, en utilisant la fonction **train_test_split()** de **scikit-learn**. Cette proportion est standard dans les projets de machine learning et garantit un compromis équilibré entre apprentissage et évaluation.

5.2 Les modèles de classification

Dans le cadre de ce projet, plusieurs modèles de classification supervisée ont été testés afin de comparer leurs performances sur le jeu de données nettoyé. Les algorithmes choisis couvrent différentes approches statistiques et d'apprentissage automatique :

— **Régression Logistique (Logistic Regression)** :

```
Accuracy of Logistic Regression: 86.20689655172413

              precision    recall  f1-score   support

           0       0.87      0.80      0.83        25
           1       0.86      0.91      0.88        33

    accuracy                           0.86        58
   macro avg       0.86      0.85      0.86        58
weighted avg       0.86      0.86      0.86        58
```

FIGURE 5.1 – Rapport de classification - Régression Logistique

Accuracy : 86,2 % — Meilleure performance globale. Excellent équilibre entre précision (0.87/0.86) et rappel (0.80/0.91). Interprétabilité élevée, choix optimal.

— **Naïve Bayes (GaussianNB)** :

Accuracy of Naive Bayes model: 82.75862068965517

	precision	recall	f1-score	support
0	0.83	0.76	0.79	25
1	0.83	0.88	0.85	33
accuracy			0.83	58
macro avg	0.83	0.82	0.82	58
weighted avg	0.83	0.83	0.83	58

FIGURE 5.2 – Rapport de classification - Naïve Bayes

Accuracy : 82,8 % — Bonne détection des malades (rappel 0.88), mais plus de faux positifs. Rapide et robuste malgré l'hypothèse d'indépendance.

— **Arbre de Décision (Decision Tree)** :

Accuracy of DecisionTreeClassifier: 79.3103448275862

	precision	recall	f1-score	support
0	0.72	0.84	0.78	25
1	0.86	0.76	0.81	33
accuracy			0.79	58
macro avg	0.79	0.80	0.79	58
weighted avg	0.80	0.79	0.79	58

FIGURE 5.3 – Rapport de classification - Arbre de Décision

Accuracy : 79,3 % — Interprétable, mais sensible au surapprentissage. Bon rappel sur les malades (0.76), mais moins précis sur les sains.

— **Forêt Aléatoire (Random Forest)** :

Accuracy of Random Forest: 81.03448275862068

	precision	recall	f1-score	support
0	0.79	0.76	0.78	25
1	0.82	0.85	0.84	33
accuracy			0.81	58
macro avg	0.81	0.80	0.81	58
weighted avg	0.81	0.81	0.81	58

FIGURE 5.4 – Rapport de classification - Forêt Aléatoire

Accuracy : 81,0 % — Solide et stable grâce au bagging. Bonne généralisation, mais légèrement en retrait par rapport à la régression logistique.

— **K-plus proches voisins (KNN)** :

Accuracy of K-NeighborsClassifier: 53.44827586206896

	precision	recall	f1-score	support
0	0.46	0.52	0.49	25
1	0.60	0.55	0.57	33
accuracy			0.53	58
macro avg	0.53	0.53	0.53	58
weighted avg	0.54	0.53	0.54	58

FIGURE 5.5 – Rapport de classification - KNN

Accuracy : 53,4 % — Performance médiocre. Trop sensible au bruit et à la dimensionnalité. Non adapté à ce dataset.

— **Support Vector Machine (SVM)** :

Accuracy of Support Vector Classifier: 58.620689655172406

	precision	recall	f1-score	support
0	0.53	0.40	0.45	25
1	0.62	0.73	0.67	33
accuracy			0.59	58
macro avg	0.57	0.56	0.56	58
weighted avg	0.58	0.59	0.58	58

FIGURE 5.6 – Rapport de classification - SVM

Accuracy : 58,6 % — Faible performance. Beaucoup de faux positifs (15/25 sains mal classés). Paramétrage inadapté ou données non linéaires mal capturées.

— **Extreme Gradient Boosting (XGBoost)** :

Accuracy of Extreme Gradient Boost: 72.41379310344827

	precision	recall	f1-score	support
0	1.00	0.36	0.53	25
1	0.67	1.00	0.80	33
accuracy			0.72	58
macro avg	0.84	0.68	0.67	58
weighted avg	0.81	0.72	0.69	58

FIGURE 5.7 – Rapport de classification - XGBoost

Accuracy : 72,4 % — Décevant. Parfait rappel sur les malades (1.0) mais faible détection des sains (0.36). Probable surapprentissage ou déséquilibre mal géré.

Chaque modèle a été entraîné sur l'ensemble d'apprentissage puis évalué sur l'ensemble de test. Les métriques clés — accuracy, précision, rappel, F1-score — permettent une évaluation rigoureuse et comparative, essentielle pour sélectionner un modèle fiable en contexte clinique.

5.3 Visualisation des performances

Afin de visualiser la capacité discriminante de chaque modèle, les courbes ROC ont été tracées pour l'ensemble des classifieurs.

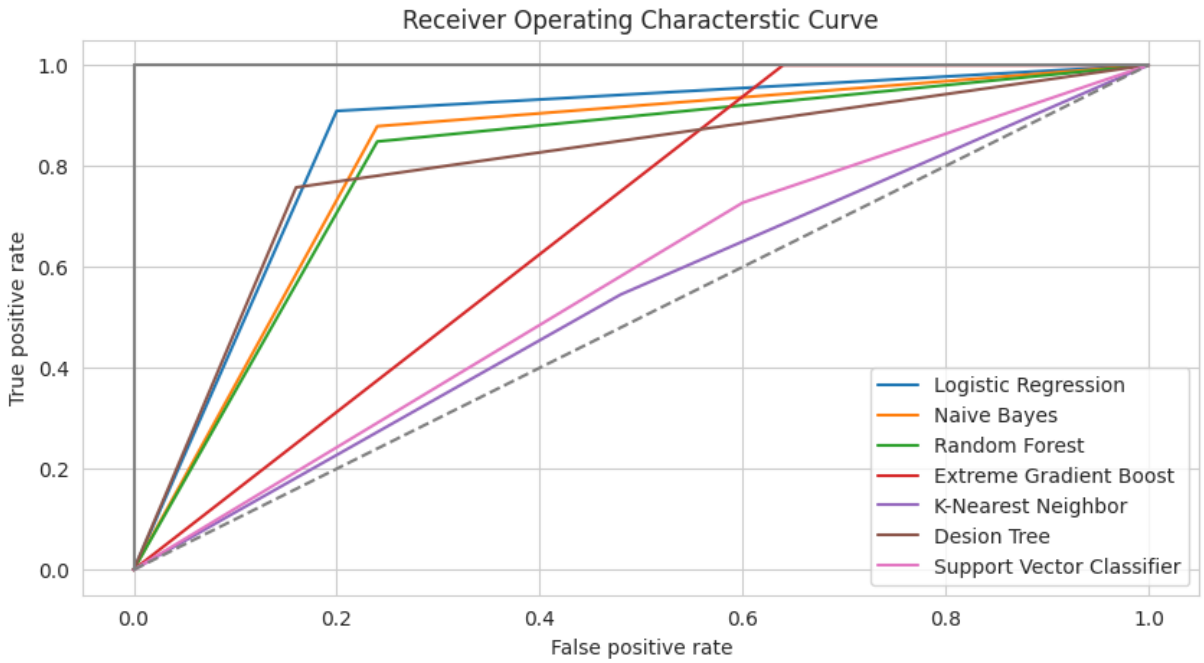


FIGURE 5.8 – Receiver Operating Characteristic

Les courbes ROC montrent le compromis sensibilité/spécificité. Régression Logistique domine (proche du coin supérieur gauche), suivie de Forêt Aléatoire. XGBoost et SVM déçoivent. L'AUC confirme : les modèles simples surpassent ici les ensembles.

5.4 Résultats comparatifs

La Régression Logistique obtient la meilleure précision (86,2 %), suivie de Naïve Bayes (82,8 %) et Forêt Aléatoire (81,0 %). XGBoost (72,4 %), SVM (58,6 %) et KNN (53,4 %) sous-performent. Les modèles simples surpassent les ensembles complexes. La Régression Logistique allie performance, équilibre et interprétabilité, idéale pour une application clinique.

5.5 Le choix du modèle

```
# Evaluation des modeles
model_ev = pd.DataFrame({
    'Modele': ['Regression Logistique', 'XGBoost', 'Foret
       Aleatoire', 'SVM',
               'Naive Bayes', 'Arbre de Decision', 'KNN'],
    'Precision (%)': [lr_acc_score*100, xgb_acc_score*100,
       rf_acc_score*100,
```

```
                        svc_acc_score*100, nb_acc_score*100,
                           dt_acc_score*100,
                        knn_acc_score*100],
    'AUC': [lr_auc, xgb_auc, rf_auc, svc_auc, nb_auc, dt_auc,
       knn_auc]
})
```

Listing 5.1 – Évaluation comparative des modèles

Modèle	**Accuracy (%)**
Logistic Regression	86.21
Naive Bayes	82.76
Random Forest	81.03
Extreme Gradient Boost	72.41
K-Nearest Neighbour	53.45
Decision Tree	79.31
Support Vector Machine	58.62

TABLE 5.1 – Comparaison des modèles de classification selon leur précision

D'après le tableau, la Régression Logistique atteint 86,2 % de précision, devant Naïve Bayes et Random Forest. XGBoost , SVM et KNN sont en retrait. Sa simplicité et son interprétabilité en font le choix optimal pour une application clinique fiable.

Conclusion

La Régression Logistique domine avec 86,2 % de précision et une forte interprétabilité, surpassant les modèles complexes. Choix optimal pour une prédiction fiable et explicable en contexte médical.

Déploiement du modèle

Introduction

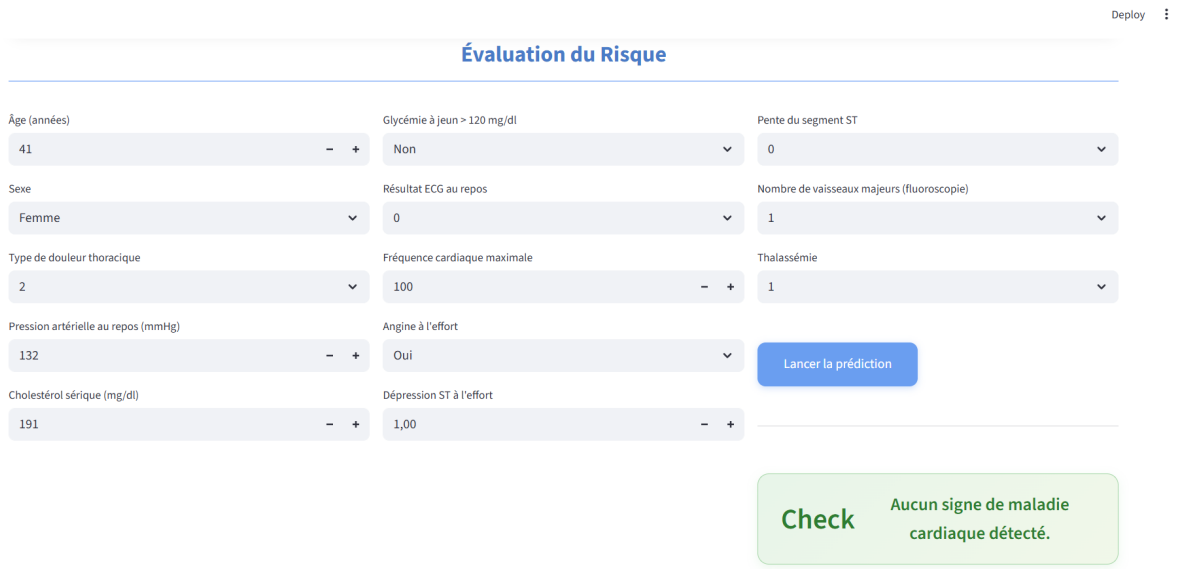
Après l'entraînement et l'évaluation du modèle de Régression Logistique, celui-ci a été déployé via une interface web interactive utilisant Streamlit, une bibliothèque Python facilitant la création d'applications de visualisation et d'interaction. Cette interface permet aux utilisateurs de saisir les paramètres cliniques d'un patient et d'obtenir instantanément une prédiction sur la présence ou l'absence d'une maladie cardiaque.

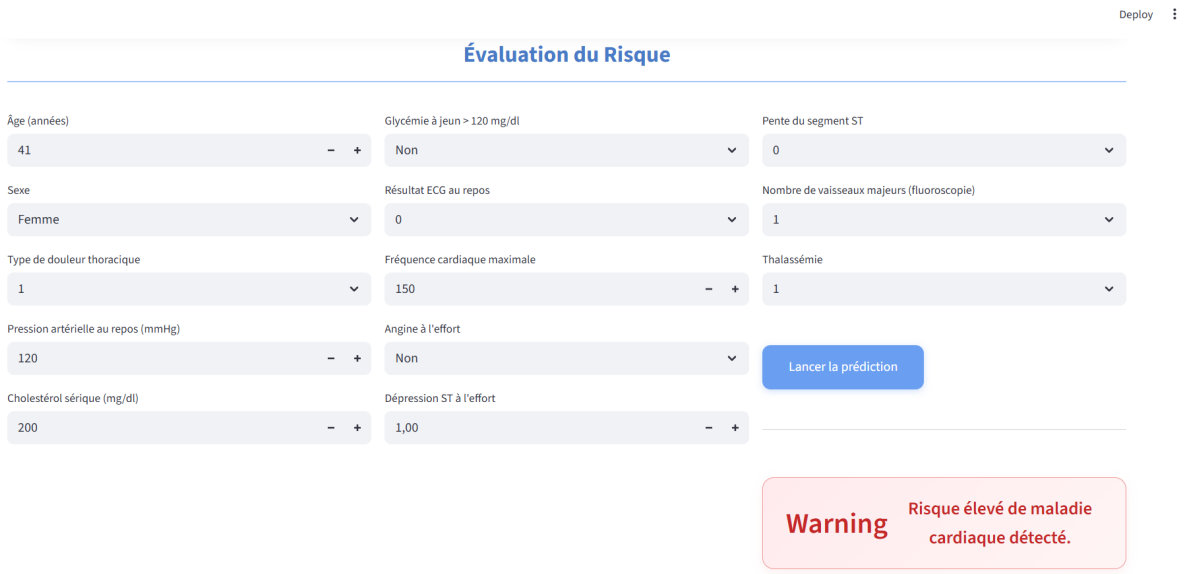
6.1 Interface Web Interactive avec Streamlit

Le modèle de Régression Logistique, retenu pour sa performance (86,2 % d'accuracy) et son interprétabilité, a été déployé sous forme d'application web interactive via Streamlit, une bibliothèque Python permettant de créer rapidement des interfaces utilisateur dynamiques.

FIGURE 6.1 – Page d'accueil de l'application Streamlit

Les champs sont pré-remplis avec des valeurs réalistes par défaut, et un bouton Prédire bien visible déclenche l'inférence en temps réel.





Conclusion

Le déploiement du modèle via Streamlit rend la prédiction des maladies cardiaques accessible, interactive et immédiatement exploitable. L'interface conviviale permet aux utilisateurs, y compris les professionnels de santé, de visualiser rapidement les résultats tout en conservant la possibilité de mises à jour pour améliorer la précision et l'adaptabilité du modèle.

Conclusion générale et perspectives

Ce projet a permis de démontrer l'efficacité du data mining pour la prédiction des maladies cardiovasculaires. L'ensemble du processus, depuis la compréhension du domaine jusqu'au déploiement du modèle, a montré que des données cliniques bien structurées et correctement prétraitées peuvent conduire à des modèles précis et interprétables. La Régression Logistique, choisie pour sa simplicité et sa fiabilité, offre un outil de prédiction utilisable en pratique clinique, tandis que l'interface Streamlit assure une accessibilité et une interaction en temps réel avec les utilisateurs.

Pour aller plus loin, plusieurs axes peuvent être envisagés : l'intégration de nouvelles variables cliniques ou biologiques pour améliorer la précision du modèle, l'utilisation de techniques d'apprentissage profond ou d'ensembles de modèles pour capturer des relations non linéaires plus complexes, et l'implémentation d'un suivi longitudinal des patients afin de prédire l'évolution du risque cardiovasculaire. Enfin, le déploiement pourrait être étendu à un environnement hospitalier réel avec un feedback des médecins pour affiner les prédictions et proposer des recommandations personnalisées. Ces perspectives permettront de renforcer l'utilité du modèle dans la prévention et la gestion proactive des maladies cardiaques.

Webographie

1. **Dataset Kaggle** : J. Smith. *Heart Disease Dataset*. Kaggle, 2019. https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset
2. **UCI Repository** : *Heart Disease Dataset*. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/45/heart+disease
3. **MSD Manuals** : *Effets de l'âge sur le cœur et les vaisseaux*. https://www.msdmanuals.com/fr/accueil/Les-vaisseaux-sanguins
4. **Ministère de la Santé Maroc** : *Maladies cardiovasculaires : crise cardiaque*. https://sehati.gov.ma/article/maladies_cardiovasculaires_crise_cardiaque_ou_infarctus_du_myocarde
5. **Scikit-learn — Logistic Regression** : Documentation officielle v1.7.2, 2024. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Annexe : CRISP-DM Appliqué

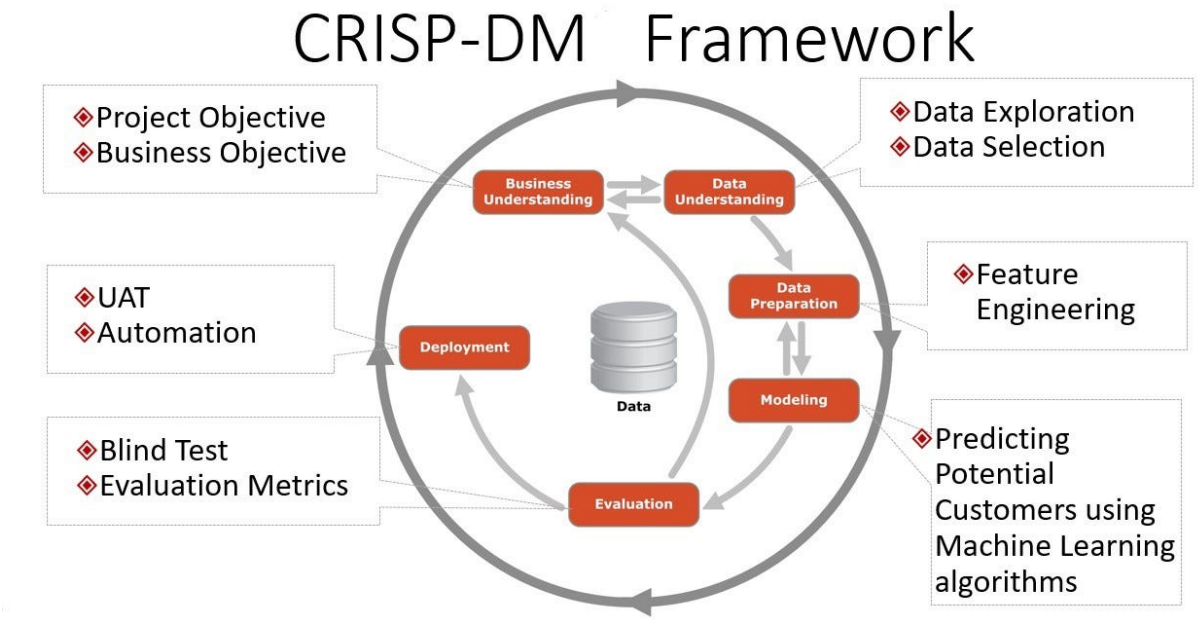


FIGURE 6.2 – Cycle CRISP-DM appliqué au projet