Sujet Data Mining: Heart Disease

Filière : Data Science et Cloud Computing
École : ENSA Oujda

Encadrant : Naji Abdelwahab

Nassrou-eddine Belhaid

27 octobre 2024

Table des matières

Chapitre 1

Introduction

1.1 Contexte du Projet

Il s'agit d'un jeu de données de type multivarié, ce qui signifie qu'il comprend ou implique une variété de variables mathématiques ou statistiques distinctes, pour une analyse numérique multivariée. Il est composé de 14 attributs, qui incluent l'âge, le sexe, le type de douleur thoracique, la pression artérielle au repos, le cholestérol sérique, la glycémie à jeun, les résultats électrocardiographiques au repos, la fréquence cardiaque maximale atteinte, l'angine induite par l'exercice, le "oldpeak" — dépression du segment ST induite par l'exercice par rapport au repos, la pente du segment ST au pic d'exercice, le nombre de vaisseaux majeurs et la thalassémie. Cette base de données comprend en réalité 76 attributs, mais toutes les études publiées utilisent un sous-ensemble de 14 attributs. La base de données de Cleveland est la seule utilisée à ce jour par les chercheurs en apprentissage automatique. Une des principales tâches sur ce jeu de données est de prédire, sur la base des attributs fournis pour un patient, si une personne particulière est atteinte d'une maladie cardiaque ou non, et l'autre tâche consiste à expérimenter pour diagnostiquer et découvrir divers insights qui pourraient aider à mieux comprendre le problème.

1.2 Objectif et Problématique

L'objectif principal de ce projet est de créer un modèle de prédiction pour identifier la présence et la gravité des maladies cardiaques chez les patients en utilisant les 14 attributs fournis dans le jeu de données. La particularité de ce dataset est que l'attribut *num* sert à indiquer la présence et le niveau de gravité de la maladie cardiaque, selon une classification en quatre classes distinctes.

Pour aborder ce problème, nous allons mettre en œuvre des **arbres de décision**. Ce modèle est particulièrement adapté pour comprendre les relations entre les différentes caractéristiques médicales et la présence ou l'absence de maladie. Les arbres de décision permettent une interprétation simple et visuelle, ce qui peut être particulièrement utile pour expliquer les résultats aux professionnels de santé.

En résumé, ce projet vise à :

- **Prédire la présence ou l'absence de maladie cardiaque** à l'aide d'attributs médicaux, avec la possibilité de détailler les prédictions selon quatre niveaux de gravité.
- **Explorer et interpréter les facteurs de risque** en utilisant des arbres de décision pour mieux comprendre l'importance relative de chaque attribut.

1.2.1 Description de la Problématique

L'attribut cible *num* se présente comme suit :
— **0** : Pas de maladie cardiaque (absence de la maladie).
— **1-4** : Présence de la maladie cardiaque, avec des niveaux de gravité variés.
Les valeurs de 1 à 4 sont interprétées de la manière suivante :
— **1** : Maladie cardiaque légère.
— **2** : Maladie cardiaque modérée.
— **3** : Maladie cardiaque sévère.
— **4** : Maladie cardiaque très sévère.

1.3 Présentation des Données

Ce jeu de données comprend plusieurs attributs qui fournissent des informations importantes pour l'analyse et la prédiction des maladies cardiaques. Les colonnes sont décrites comme suit :

— **id** : Identifiant unique pour chaque patient dans le jeu de données, utilisé principalement pour le suivi et l'indexation.
— **age** : Âge du patient en années. L'âge est un facteur significatif dans l'évaluation du risque de maladies cardiaques, car la probabilité de certaines conditions cardiaques augmente avec l'âge.
— **origin** : Indique le lieu d'étude ou l'origine du patient, se référant à l'emplacement ou à l'institution où les données ont été collectées.
— **sex** : Genre du patient, catégorisé comme Masculin ou Féminin. Le genre peut influencer la présentation et les facteurs de risque des maladies cardiaques.
— **cp** : Type de douleur thoracique ressentie par le patient, qui peut être classifiée comme suit :
  — Angine typique
  — Angine atypique
  — Douleur non-anginale
  — Asymptomatique
— **trestbps** : Pression artérielle au repos du patient en mm Hg mesurée lors de l'admission à l'hôpital. Une pression artérielle élevée est un facteur de risque courant pour les maladies cardiaques.
— **chol** : Niveau de cholestérol sérique du patient en mg/dl. Des niveaux de cholestérol élevés sont associés à un risque accru de maladies cardiaques.
— **fbs** : Indique si le patient a un taux de sucre dans le sang à jeun supérieur à 120 mg/dl (Vrai/Faux). Des niveaux de sucre à jeun élevés peuvent indiquer un diabète, qui est un facteur de risque pour les maladies cardiaques.
— **restecg** : Résultats électrocardiographiques au repos, classifiés comme suit :
  — Normal
  — Anomalie des ondes ST-T
  — Hypertrophie ventriculaire gauche
— **thalach** : Fréquence cardiaque maximale atteinte par le patient lors des tests. La fréquence cardiaque maximale peut fournir des informations sur la condition physique et la fonction cardiovasculaire.
— **exang** : Indique si le patient a ressenti une angine induite par l'exercice (Vrai/-Faux).

— **oldpeak** : Dépression ST induite par l'exercice par rapport au repos. La dépression ST peut indiquer une ischémie, c'est-à-dire une réduction du flux sanguin vers le muscle cardiaque pendant l'exercice.
— **slope** : Pente du segment ST peak lors de l'exercice. Cela peut servir d'indicateur de la gravité de la maladie des artères coronaires.
— **ca** : Nombre de vaisseaux majeurs (0-3) colorés par fluoroscopie. Un nombre élevé de vaisseaux colorés peut indiquer une maladie coronarienne plus sévère.
— **thal** : Résultat du test de stress au thallium, classé comme suit :
  — Normal
  — Défaut fixe
  — Défaut réversible
— **num** : Attribut prédit, indiquant probablement la présence ou la gravité de la maladie cardiaque sur la base des autres caractéristiques du jeu de données.

Chapitre 2

Préparation des Données

2.1 Exploration Préliminaire des Données

	id	age	sex	dataset	cp	trestbps	chol	fbs	restecg	thalch	exang	oldpeak	slope	ca	thal	num
0	1	63	Male	Cleveland	typical angina	145.0	233.0	True	lv hypertrophy	150.0	False	2.3	downsloping	0.0	fixed defect	0
1	2	67	Male	Cleveland	asymptomatic	160.0	286.0	False	lv hypertrophy	108.0	True	1.5	flat	3.0	normal	2
2	3	67	Male	Cleveland	asymptomatic	120.0	229.0	False	lv hypertrophy	129.0	True	2.6	flat	2.0	reversable defect	1
3	4	37	Male	Cleveland	non-anginal	130.0	250.0	False	normal	187.0	False	3.5	downsloping	0.0	normal	0
4	5	41	Female	Cleveland	atypical angina	130.0	204.0	False	lv hypertrophy	172.0	False	1.4	upsloping	0.0	normal	0

FIGURE 2.1 – romws :920 *columns :16

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 920 entries, 0 to 919
Data columns (total 16 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   id        920 non-null    int64
 1   age       920 non-null    int64
 2   sex       920 non-null    object
 3   dataset   920 non-null    object
 4   cp        920 non-null    object
 5   trestbps  861 non-null    float64
 6   chol      890 non-null    float64
 7   fbs       830 non-null    object
 8   restecg   918 non-null    object
 9   thalch    865 non-null    float64
 10  exang     865 non-null    object
 11  oldpeak   858 non-null    float64
 12  slope     611 non-null    object
 13  ca        309 non-null    float64
 14  thal      434 non-null    object
 15  num       920 non-null    int64
dtypes: float64(5), int64(3), object(8)
memory usage: 115.1+ KB
```

FIGURE 2.2 – Data-type

1. Le dataset contient 920 entrées et 16 caractéristiques pertinentes aux études sur les maladies cardiaques.
2. Des données complètes sont disponibles pour 'id', 'age', 'sex', 'dataset', 'cp', et la variable de résultat 'num'.
3. Des données manquantes significatives se trouvent dans 'ca' avec 66,41
4. Des données manquantes significatives se trouvent dans 'thal' avec 52,83
5. 'Slope' a également une quantité considérable de données manquantes, avec seulement

611 entrées non nulles.
6. Les variables 'trestbps', 'chol', 'fbs', 'thalch', 'exang' et 'oldpeak' ont quelques valeurs manquantes, mais dans une moindre mesure.

2.2 Statistique descriptive

	id	age	trestbps	chol	thalch	oldpeak	ca	num
count	920.000000	920.000000	861.000000	890.000000	865.000000	858.000000	309.000000	920.000000
mean	460.500000	53.510870	132.132404	199.130337	137.545665	0.878788	0.676375	0.995652
std	265.725422	9.424685	19.066070	110.780810	25.926276	1.091226	0.935653	1.142693
min	1.000000	28.000000	0.000000	0.000000	60.000000	-2.600000	0.000000	0.000000
25%	230.750000	47.000000	120.000000	175.000000	120.000000	0.000000	0.000000	0.000000
50%	460.500000	54.000000	130.000000	223.000000	140.000000	0.500000	0.000000	1.000000
75%	690.250000	60.000000	140.000000	268.000000	157.000000	1.500000	1.000000	2.000000
max	920.000000	77.000000	200.000000	603.000000	202.000000	6.200000	3.000000	4.000000

FIGURE 2.3 – statistique

Nous avons les données de 920 individus avec 16 colonnes, qui incluent les attributs suivants : ['age', 'sex', 'cp', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach', 'exang', 'oldpeak', 'slope', 'ca', 'thal', 'num', 'dataset'].

— **'num'** est l'attribut prédit qui indique le niveau de maladie cardiaque.
— **'dataset'** est la source des données.
— **'age'** est l'âge de la personne.
— **'trestbps'** est la pression artérielle au repos.

Nos données montrent une moyenne d'âge des individus de 53 ans, avec un âge maximum de 77 ans, tandis que nous avons également un âge minimum de 28 ans pour un patient dans notre dataset, avec une maladie cardiaque suspectée. La pression artérielle au repos moyenne est de 132, avec un maximum de 200.

1. Il n'y a pas de maladies cardiaques trouvées chez 25 % des patients avec une moyenne d'âge de 47,0 ans.
2. Une présence légère de maladies cardiaques est trouvée chez 50 % des patients avec une moyenne d'âge de 54,0 ans.
3. Une présence modérée de maladies cardiaques est trouvée chez 75 % des patients avec une moyenne d'âge de 60,0 ans ou plus.

2.3 Data cleaning

2.3.1 Missing value detection

FIGURE 2.4 – statistique

ca	66.41
thal	52.83
slope	33.59
fbs	9.78
oldpeak	6.74
trestbps	6.41
exang	5.98
thalch	5.98
chol	3.26
restecg	0.22
cp	0.00
dataset	0.00
id	0.00
age	0.00
sex	0.00
num	0.00

(a) Pourcentage

```
['trestbps',
 'chol',
 'fbs',
 'restecg',
 'thalch',
 'exang',
 'oldpeak',
 'slope',
 'ca',
 'thal']
```

(b) les colonnes dont on a Missing value

FIGURE 2.5 – Informations

2.3.2 Traitement des valeurs manquantes

Observation : Bien que certaines colonnes contiennent des valeurs manquantes, dépassant même 50% des données totales, certaines d'entre elles, comme *ca* et *thal*, sont cruciales pour notre prédiction. Par conséquent, il est essentiel de ne pas les supprimer. Nous adopterons une stratégie d'imputation rigoureuse pour combler ces valeurs manquantes, garantissant ainsi l'intégrité de l'analyse.

- Le jeu de données présente des lacunes significatives dans des caractéristiques clés, *ca* et *thal* étant les plus touchées, avec respectivement 66,41% et 52,83% de données manquantes.
- La colonne *slope* montre également un taux élevé de valeurs manquantes, atteignant 33,59%.
- Ces niveaux élevés d'incomplétude posent un défi majeur pour toute analyse prédictive, car ils pourraient biaiser les résultats si les données sont mal imputées.

Pour répondre à ce problème, nous mettons en œuvre une méthode d'imputation avancée en deux étapes :

1. **Imputation des colonnes numériques :** Nous utilisons l'algorithme *Iterative Imputer* pour estimer les valeurs manquantes dans les colonnes numériques. Cette technique repose sur des modèles itératifs, ajustés pour chaque colonne, ce qui permet de minimiser le biais induit par des méthodes traditionnelles telles que la moyenne, la médiane ou le mode.
2. **Imputation des colonnes catégorielles :** Nous utilisons un *Random Forest Classifier* pour prédire et remplir les valeurs manquantes des colonnes catégorielles. Ce modèle de classification permet d'exploiter la complexité des interactions entre les caractéristiques pour estimer avec précision les valeurs manquantes.

Cette approche différenciée nous permet de gérer les données manquantes de manière optimale, en préservant autant que possible la structure et l'information contenue dans le jeu de données, tout en limitant les biais potentiels qui pourraient affecter les résultats prédictifs.

2.3.3 Résultats de l'Imputation Avancée

Dans notre étude, nous avons appliqué des méthodes d'imputation pour traiter les valeurs manquantes dans notre jeu de données. Nous avons utilisé deux approches distinctes en fonction du type de variable : une pour les variables catégorielles et une autre pour les variables continues.

Résultats pour les Variables Continues

Pour les variables continues, nous avons calculé des indicateurs de performance tels que l'erreur absolue moyenne (MAE), la racine de l'erreur quadratique moyenne (RMSE) et le coefficient de détermination (R^2) pour évaluer l'efficacité de notre modèle d'imputation. Voici quelques résultats clés :

- **Tension artérielle au repos (trestbps)** :
  - MAE : 13.18
  - RMSE : 17.26
  - R^2 : 0.0697

Ce résultat indique une précision modérée, suggérant que le modèle peut être amélioré.

— **Cholestérol (chol)** :
  — MAE : 45.26
  — RMSE : 64.50
  — R^2 : 0.6704

  Le modèle a montré une meilleure performance avec un R^2 relativement élevé, ce qui indique une meilleure capacité à prédire cette variable.

— **Fréquence cardiaque maximale atteinte (thalch)** :
  — MAE : 16.80
  — RMSE : 21.66
  — R^2 : 0.3176

  Les résultats montrent un certain degré d'imprécision, ce qui souligne la nécessité d'une analyse plus approfondie des facteurs influençant cette variable.

schema explicatif

FIGURE 2.6 – diagramme de l'inputation

Résultats pour les Variables Catégorielles

Pour les variables catégorielles, nous avons évalué la performance de notre imputation en termes de précision :

— **Fumeur (fbs)** :
  — Précision : 79.52%

  Cela démontre une bonne capacité du modèle à prédire cette variable, bien que des améliorations soient toujours possibles.

— **Exercice anginal (exang)** :
  — Précision : 79.19%

  Une performance similaire à celle de la variable fbs, indiquant une fiabilité raisonnable dans nos prédictions.

— **Slope et ca** :
  — Précisions respectives : 69.11% et 62.90%

Ces résultats soulignent des opportunités d'amélioration pour mieux capturer les tendances sous-jacentes.

— **Thal** :
  — Précision : 74.71%
  Ce taux de précision suggère que le modèle a une compréhension décente de la variable, mais qu'il existe encore un potentiel d'optimisation.

Conclusion

Dans l'ensemble, nos résultats d'imputation montrent des performances variées selon les types de variables, avec des succès notables dans les variables catégorielles et des performances plus mitigées dans le cas des variables continues. Ces analyses soulignent l'importance d'une approche méthodologique rigoureuse lors de la gestion des valeurs manquantes, et ouvrent la voie à des améliorations potentielles par le biais d'une sélection plus fine des caractéristiques et de l'optimisation des modèles.

2.4 Etude de la correlation et la distribution

Dans cette section, nous examinons la corrélation entre les différentes colonnes numériques de notre jeu de données. Nous avons utilisé une matrice de corrélation pour visualiser les relations entre les variables.

	age	trestbps	chol	thalch	oldpeak
age	1.000000	0.254993	-0.084600	-0.375914	0.292071
trestbps	0.254993	1.000000	0.089094	-0.115725	0.184404
chol	-0.084600	0.089094	1.000000	0.221788	0.042483
thalch	-0.375914	-0.115725	0.221788	1.000000	-0.177503
oldpeak	0.292071	0.184404	0.042483	-0.177503	1.000000

FIGURE 2.7 – Matrice de Corrélation

Observation : Il n'existe pas de corrélation hautement positive entre les colonnes numériques. Cependant, nous notons une certaine corrélation négative entre les colonnes `trestbps`, `thalch`, `age` et `chol`.

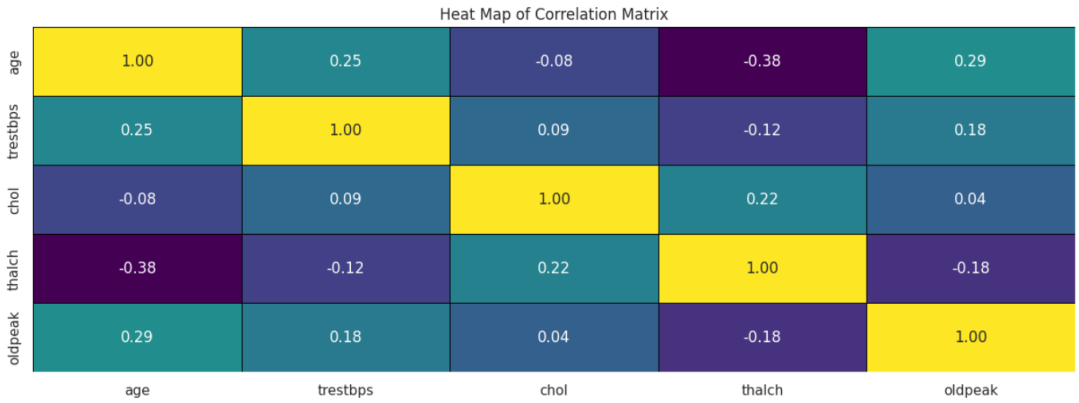

FIGURE 2.8 – Visualisation de la Corrélation

2.4.1 Interprétation des Graphiques

En regardant les distrubtuion on peut dire que :

— **Âge :** La distribution de l'âge semble être approximativement normale, avec un pic autour de 55 ans. Il y a quelques valeurs aberrantes plus âgées.

— **Restres :** La distribution des restres est également approximativement normale, avec une légère asymétrie positive (la queue est plus longue à droite).

— **Cholestérol (chol) :** La distribution du cholestérol est légèrement asymétrique positive, avec un pic autour de 125-150. Il y a quelques valeurs aberrantes élevées.

— **Thalach :** La distribution de thalach est approximativement normale, avec un pic autour de 150.

— **Oldpeak :** La distribution de oldpeak est fortement asymétrique positive, avec la plupart des valeurs concentrées autour de 0 et quelques valeurs élevées.

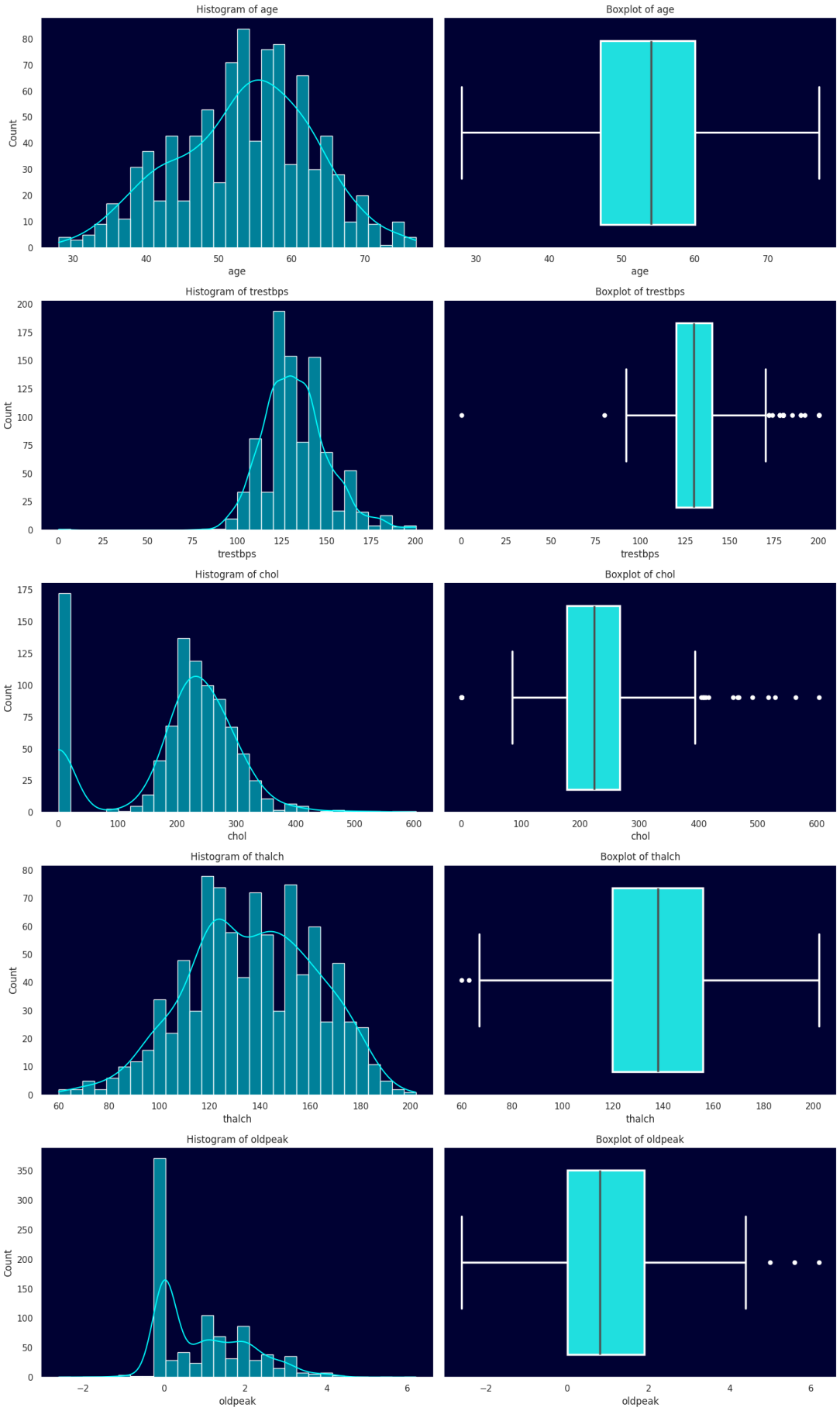

FIGURE 2.9 – Visualisation de la distribution

2.4.2 traitement des valeurs abberantes

La prochaine étape consiste à éliminer les valeurs aberrantes en utilisant la méthode de l'intervalle interquartile (IQR). Les valeurs aberrantes peuvent fausser les résultats d'un ensemble de données, surtout dans les analyses statistiques qui supposent une distribution normale. Une fonction a été définie pour détecter et retirer les valeurs aberrantes dans chaque colonne numérique du DataFrame. Cette fonction calcule les premier (Q1) et troisième (Q3) quartiles, ainsi que l'IQR, et détermine les limites inférieure et supérieure des points de données acceptables. Les valeurs en dehors de 1,5 fois l'IQR en dessous de Q1 ou au-dessus de Q3 sont considérées comme des valeurs aberrantes et sont supprimées. Ce processus de nettoyage a été appliqué à chaque colonne numérique, à l'exception de la colonne 'chol', qui a été omise. En raffinant l'ensemble de données de cette manière, on peut potentiellement améliorer la précision du modèle, car les données restantes seront plus représentatives de la tendance sous-jacente sans la distorsion causée par les valeurs extrêmes.

Résultat

Nombre de valeurs aberrantes détectées dans oldpeak : 0,33 %
Nombre de valeurs aberrantes détectées dans thalch : 0,22 %
Nombre de valeurs aberrantes détectées dans chol : 20,11 %
Nombre de valeurs aberrantes détectées dans trestbps : 2,72 %
Nombre de valeurs aberrantes détectées dans age : 0,0 %

2.5 Encodage des Variables Catégorielles

Choix du Label Encoding dans les Modèles de Classification

Le Label Encoding est une méthode efficace pour transformer des variables catégorielles en valeurs numériques. Cette approche est particulièrement bénéfique dans le cadre des modèles de machine learning, tels que les arbres de décision.

2.5.1 Avantages du Label Encoding

L'utilisation du Label Encoding présente plusieurs avantages :

— **Simplicité** : Le Label Encoding convertit chaque catégorie en un entier unique, ce qui est simple à comprendre et à mettre en œuvre.

— **Compatibilité avec les Arbres de Décision** : Les arbres de décision, tels que le Random Forest, peuvent gérer les variables catégorielles de manière intrinsèque. Cependant, en utilisant le Label Encoding, les arbres de décision peuvent interpréter les catégories sans introduire de complexité supplémentaire.

— **Meilleure Performance** : En encodant les catégories en valeurs numériques, on permet au modèle d'apprendre des relations non linéaires, ce qui peut améliorer les performances du modèle sur des ensembles de données complexes.

2.5.2 Comparaison avec d'autres Méthodes

Il est souvent recommandé de comparer les performances des modèles en utilisant différentes techniques d'encodage. Par exemple :

— **One-Hot Encoding** : Bien qu'il soit utile pour éviter d'introduire un ordre implicite entre les catégories, le One-Hot Encoding peut entraîner une augmentation significative de la dimensionnalité, ce qui peut affecter la performance des modèles.
— **Comparaison des Modèles** : En utilisant le Label Encoding, nous pouvons facilement implémenter d'autres modèles tels que la régression logistique ou les réseaux de neurones. En comparant les métriques de performance (précision, rappel, F1-score), nous pouvons déterminer le modèle le plus adapté à notre problème spécifique.

2.5.3 Resultat et Conclusion

	id	age	sex	dataset	cp	trestbps	chol	fbs	restecg	thalch	exang	oldpeak	slope	ca	thal	num
0	1	63	1	0	3	145.0	233.0	1	0	150.0	0	2.3	0	0.0	0	0
1	2	67	1	0	0	160.0	286.0	0	0	108.0	1	1.5	1	3.0	1	2
2	3	67	1	0	0	120.0	229.0	0	0	129.0	1	2.6	1	2.0	2	1
3	4	37	1	0	2	130.0	250.0	0	1	187.0	0	3.5	0	0.0	1	0
4	5	41	0	0	1	130.0	204.0	0	0	172.0	0	1.4	2	0.0	1	0

FIGURE 2.10 – dataset apres encodage

Le choix du Label Encoding est une étape cruciale dans le prétraitement des données pour les modèles de machine learning. Sa simplicité et son efficacité en font une option privilégiée, surtout lorsqu'il est associé à des algorithmes tels que les arbres de décision. De plus, il offre la flexibilité nécessaire pour tester et comparer divers modèles, permettant ainsi d'optimiser les performances.

2.6 Scaling des donnees

2.7 Data Scaling

Le choix des variables à mettre à l'échelle est déterminé par la nature des données et les exigences des modèles de machine learning. Les variables sélectionnées pour la mise à l'échelle sont `age`, `trestbps`, `chol`, `thalch`, et `oldpeak`, tandis que les colonnes catégorielles ne le sont pas.

1. **Type de Données :** Les variables choisies sont toutes des variables numériques continues :
   — `age` : Âge du patient.
   — `trestbps` : Pression artérielle au repos (mm Hg).
   — `chol` : Taux de cholestérol sanguin (mg/dl).
   — `thalch` : Fréquence cardiaque maximale atteinte (bpm).
   — `oldpeak` : Déclin du segment ST (indicateur de la santé cardiaque).
   Les modèles de machine learning fonctionnent mieux lorsque les caractéristiques numériques sont dans une plage comparable.
2. **Sensibilité aux Échelles :** Les modèles comme les réseaux de neurones, la régression logistique et les k plus proches voisins (KNN) sont sensibles à l'échelle des données. Des échelles différentes peuvent conduire à des résultats biaisés.

3. **Variabilité des Données :** Les colonnes choisies présentent une variabilité importante et impactent significativement les résultats du modèle. La mise à l'échelle assure une contribution équitable des variables lors du calcul des distances entre points de données.
4. **Prise en Compte des Relations :** Les variables continues peuvent interagir entre elles et avec d'autres variables. La mise à l'échelle maintient ces relations proportionnelles, évitant ainsi que certaines caractéristiques dominent à cause de leurs valeurs.

Objectif : L'objectif de cette approche est de gagner en aisance dans l'implémentation des arbres de décision, tout en comparant leurs performances avec d'autres modèles de machine learning. Cela permettra d'évaluer l'efficacité des arbres de décision par rapport à d'autres méthodes et de déterminer la meilleure approche pour le problème en question.

2.7.1 visualisation de la distribution apres le scaling

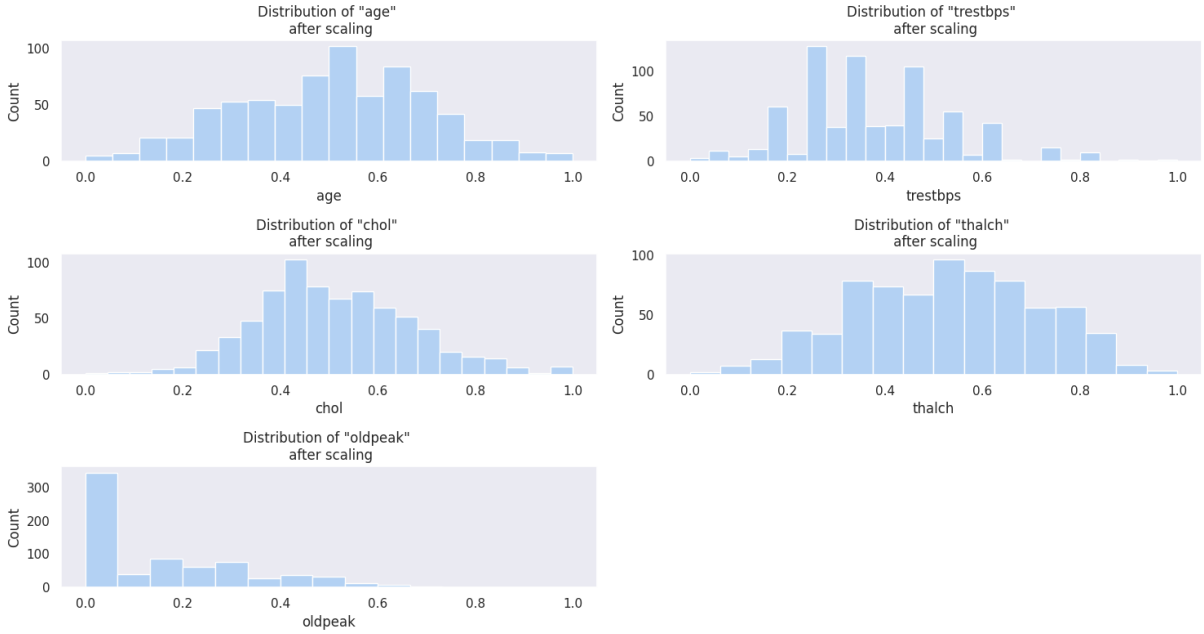

FIGURE 2.11 – distribution

— **Distribution plus uniforme :** Après le scaling, les distributions des variables sont devenues plus homogènes, se répartissant de manière plus uniforme sur l'intervalle $[0, 1]$.
— **Pas d'unité :** Les valeurs après le scaling n'ont plus d'unité spécifique, elles sont désormais des valeurs numériques comprises entre 0 et 1, facilitant ainsi les calculs.

Chapitre 3

Construction du Modèle de Classification

3.1 Arbre de Décision

3.1.1 Séparation des Données

Dans cette étape, nous avons séparé le jeu de données en ensembles d'entraînement et de test. Cela permet de valider les performances du modèle en utilisant des données non vues pendant l'entraînement. Une portion de 80% des données a été attribuée à l'entraînement et 20% aux tests.

3.1.2 Configuration de la Grille des Hyperparamètres

Nous avons défini une grille d'hyperparamètres pour la recherche aléatoire. Les hyperparamètres choisis sont :

- **max_depth** : La profondeur maximale de l'arbre, choisie entre 3 et 4.
- **criterion** : La fonction de mesure de la qualité de la séparation, parmi `gini` et `entropy`.
- **max_features** : Le nombre maximal de caractéristiques utilisées, sélectionné parmi `auto`, `sqrt`, et `log2`.
- **min_samples_split** : Le nombre minimal d'échantillons requis pour diviser un nœud, choisi parmi 2, 4, et 6.

3.1.3 Recherche Aléatoire des Hyperparamètres

Une recherche aléatoire (*Randomized Search*) a été effectuée en utilisant la méthode `RandomizedSearchCV` de `scikit-learn`. Cette recherche a permis d'optimiser les hyperparamètres du modèle d'arbre de décision en utilisant une validation croisée avec 5 sous-échantillons. Les meilleurs hyperparamètres trouvés seront présentés dans la figure 3.1.

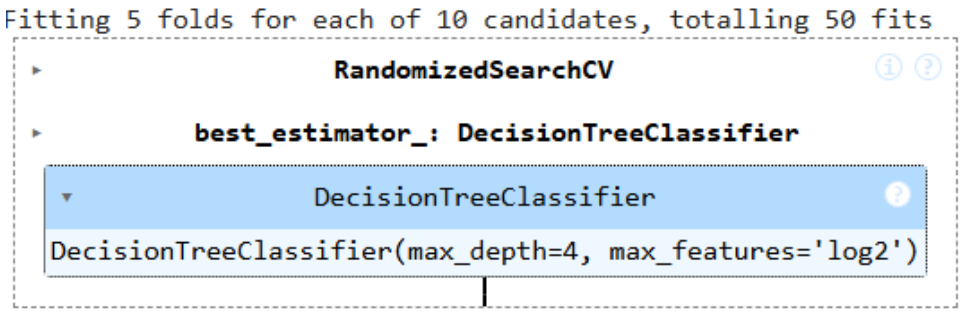

FIGURE 3.1 – Résultats de la recherche aléatoire des hyperparamètres

3.1.4 Création et Entraînement de l'Arbre de Décision

L'arbre de décision a été créé en utilisant les meilleurs hyperparamètres trouvés lors de la recherche. Le modèle a ensuite été entraîné sur l'ensemble de données d'entraînement. La figure **??** montre l'entraînement de l'arbre de décision.

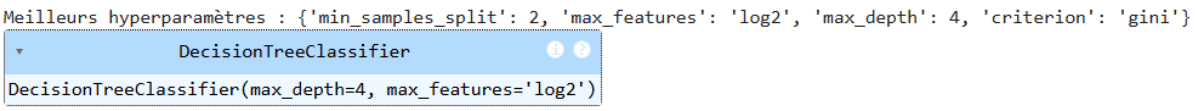

FIGURE 3.2 – Identification du nœud le plus pertinent

3.1.5 Visualisation de l'Arbre de Décision

Après l'entraînement, l'arbre de décision a été visualisé afin de comprendre les règles de séparation des données. La visualisation est montrée dans la figure 3.3.

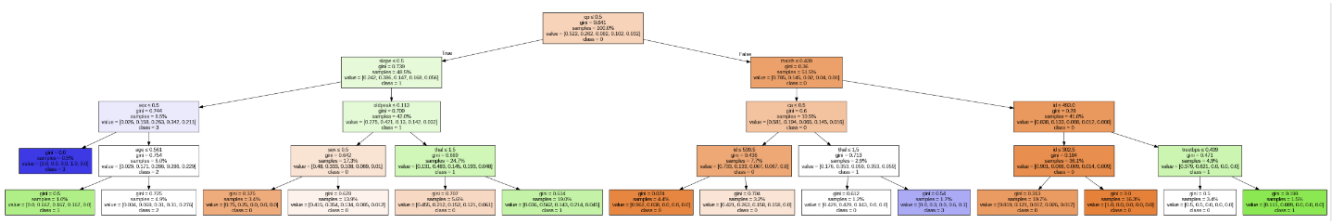

FIGURE 3.3 – Visualisation de l'arbre de décision

3.1.6 Identification du Nœud le Plus Pertinent

Une analyse a été effectuée pour identifier le nœud le plus pertinent de l'arbre de décision. Le nœud le plus pertinent correspond à celui qui offre la meilleure séparation des données à un niveau élevé de l'arbre. La figure 3.4 présente l'identification de ce nœud.

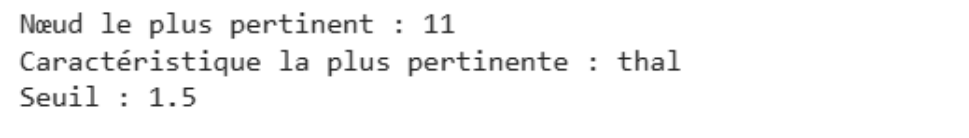

FIGURE 3.4 – Identification du nœud le plus pertinent

Résultats de la Classification avec un Arbre de Décision

Configuration Optimale

Les meilleurs hyperparamètres obtenus à partir de la recherche aléatoire (RandomizedSearchCV) sont les suivants :

- **Profondeur maximale (max_depth) :** 4
- **Caractéristique maximale (max_features) :** sqrt
- **Nombre minimal d'échantillons pour une séparation (min_samples_split) :** 4

Performance du Modèle

La précision globale (*accuracy*) du modèle sur l'ensemble de test est de 0.694. La matrice de confusion et le rapport de classification sont présentés ci-dessous :

Matrice de Confusion

	Classe 0	**Classe 1**	**Classe 2**	**Classe 3**	**Classe 4**
Classe 0	73	5	1	0	0
Classe 1	16	27	0	0	0
Classe 2	5	6	2	1	0
Classe 3	4	4	1	0	0
Classe 4	1	1	0	0	0

Rapport de Classification

	Précision	**Rappel**	**F1-score**	**Support**
Classe 0	0.74	0.92	0.82	79
Classe 1	0.63	0.63	0.63	43
Classe 2	0.50	0.14	0.22	14
Classe 3	0.00	0.00	0.00	9
Classe 4	0.00	0.00	0.00	2
Précision Globale	0.69			
Moyenne Macro	0.37	0.34	0.33	147
Moyenne Pondérée	0.63	0.69	0.65	147

3.2 AUTRES MODEL POUR choisir le meilleur modele