

09/02/2025

Projet RCP209 :

Prédire la présence ou l'absence
de maladie du cœur



Sourou Alain NOUNAWON
CNAME PARIS

Table des matières

1	Présentation du projet.....	4
1.1	Contexte.....	4
1.2	Les données.....	4
1.3	Méthodologie.....	5
2	Exploration préparatoire des données	5
2.1	Analyse univariée	5
2.1.1	Variables quantitatives.....	5
2.1.2	Variables qualitatives	8
2.2	Analyse bivariée.....	9
2.2.1	Lien entre variables quantitatives.....	9
2.2.2	Lien entre variables quantitatives et qualitatives	10
2.2.3	Lien entre variables qualitatives	12
3	Construction de différents modèles	13
3.1	Division « Train » / « Test ».....	14
3.2	Modèle de régression logistique	14
3.2.1	Entraînement sur la base « Train »	14
3.2.2	Prédiction sur la base « Test »	14
3.2.3	Evaluation du modèle.....	14
3.3	Modèle Forêt aléatoire	16
3.3.1	Entraînement du modèle Random Forest sur la base « Train »	16
3.3.2	Prédiction du modèle Random Forest (RF) sur la base « Test ».....	17
3.3.3	Evaluation du modèle RF	17
3.4	Modèle XgBoost	18
3.4.1	Entraînement sur la base « Train »	18
3.4.2	Prédiction du modèle XgBoost sur la base « Test »	18
3.4.3	Evaluation du modèle XgBoost avec les meilleurs paramètres.....	18
3.5	Modèle K Plus Proches Voisins (KNN)	19
3.5.1	Entraînement du modèle KNN sur la base « Train »	19
3.5.2	Prédiction du modèle KNN sur la base « Test »	19
3.5.3	Evaluation du modèle KNN	19
4	Comparaison des modèles et choix du modèle final.....	20
4.1	Comparaison.....	20
4.2	Choix du modèle final.....	20
5	Conclusion.....	20
	Annexes	22

Les tableaux :

Tableau 1 : Variables renommées et description	4
Tableau 2 : Résumé numérique des variables quantitatives	6
Tableau 3 : Tests de normalité de Kolmogorov-Smirnov et Jarque-Bera.....	7
Tableau 4 : Valeurs extrêmes.....	7
Tableau 5 : Analyse bivariée entre variables quantitatives et variable cible - résultats des tests de Student et de Mann-Whitney	11
Tableau 6 : Liaison entre variables qualitatives explicatives et la Target - résultats du test de Chi2 et V de Cramer	13
Tableau 7 : Indicateurs de performance du modèle de régression logistique avec toutes les variables.	15
Tableau 8 : Indicateurs de performance du modèle de régression logistique sans les variables source de multi-colinéarité	15
Tableau 9 : Indicateurs de performance du modèle Random Forest non optimisé	17
Tableau 10 : Indicateurs de performance du modèle Random Forest optimisé.....	17
Tableau 11 : Indicateurs de performance du modèle XgBoost optimisé	18
Tableau 12 : Indicateurs de performance du modèle KNN optimisé.....	19

Les figures :

Figure 1 : Histogrammes de « age », « rest_blood_pres », « serum_chol » , « max_heart_rate », et « oldpeak ».....	6
Figure 2 : Répartition en effectif (à gauche) en proportion (à droite) de la variable grp_imc	8
Figure 3 : Nuage de points entre « age » et « max_heart_rate » (à gauche) et entre « age » et « rest_blood_pres » (à droite)	9
Figure 4 : Matrice de corrélation et p-value de Pearson.....	10
Figure 5 : Classement de la relation entre les variables quantitatives et la variable cible selon le test de Mann-Whitney	11
Figure 6 : Heatmap des pourcentages de liaison entre les variables qualitatives et la « Target ».....	12
Figure 7 : Classement de la relation entre les variables qualitatives et la Target selon le V de Cramer	13
Figure 8 : Matrice de confusion et courbe ROC de regression logistique avec toutes les variables.....	15
Figure 9 : Matrice de confusion et courbe ROC de régression logistique sans les variables source de multi-colinéarité	15
Figure 10 : Test Modèle RF optimisé - Matrice de confusion, courbe ROC et importance des variables	17
Figure 11 : Test Modèle XgBoost optimisé - Matrice de confusion, courbe ROC et importance des variables	19
Figure 12 : Test Modèle KNN optimisé - Matrice de confusion, courbe ROC sur le seuil 0.5	20

Annexes :

Annexe 1 : Tableau et information sur variables à l'origine	22
Annexe 2 : Histogramme et courbe de densité (à gauche), Boxplot (au milieu) et QQ Plot (à droite) de « age », « rest_blood_pres », « serum_chol » , « max_heart_rate », et « oldpeak »	22
Annexe 3 : Résumé numérique des variables qualitatives	23

Annexe 4 : Répartition en effectif (à gauche) en proportion (à droite) de quelques modalités.....	24
Annexe 5 : Boxplots entre les différentes variables quantitatives et la target, tests statistique.....	25
Annexe 6 : Tableau de contingence, Heatmap des pourcentages de liaison, test Chi-2 et V de Cramer	27
Annexe 7 : Coefficients de la régression logistique (à droite : sans les variables source de multicolinéarité : 'rest_blood_pres', 'age', 'max_heart_rate', 'serum_chol')	29
Annexe 8 : Indicateurs de performance du modèle de d'arbre de décision.....	29
Annexe 9 : Arbre de décision du modèle optimisé et importance de variables	29
Annexe 10 : Indicateurs de performance des différents modèles étudiés	30

1 Présentation du projet

1.1 Contexte

Pour valider l'UE [RCP209](#) du CNAM (Conservatoire national des arts et métiers), en plus du passage de l'examen écrit, une étude de cas donnant lieu à la rédaction d'un rapport, doit être réalisée en mettant œuvre les techniques vues au cours permettant d'effectuer l'évaluation et la sélection d'un ou plusieurs modèles décisionnels adaptés au problème, précédée d'une exploration préparatoire des données.

L'étude menée ici, consiste à **prédirer la présence ou l'absence de maladie du cœur à partir de treize (13) attributs représentant l'état physiologique du patient sur un jeu de données contenant 270 observations**. Les données étudiées sont issues de [Kaggle](#) : "heart.dat". La variable à expliquer étant qualitative et binaire, le présent rapport fait l'objet d'une analyse prédictive et concerne la classification, relevant de l'apprentissage supervisé.

1.2 Les données

Le jeu de données "heart.dat" provient de [Kaggle](#) et se présente à l'origine à l'*Annexe 1*. On retient :

- 14 variables : 5 quantitatives et 9 qualitatives dont la variable cible ;
- 270 individus : âgés de 29 à 77 ans.

Il faut préciser que les variables ont été renommées :

N°	Variable (renommée)	Description
1	<i>age</i>	Âge du patient.
2	<i>sex</i>	Sexe du patient (1 = homme, 0 = femme).
3	<i>chest_pain_type</i>	Type de douleur thoracique (1 = typique, 2 = atypique, 3 = non-angineuse, 4 = asymptomatique).
4	<i>rest_blood_pres</i>	Pression artérielle au repos (en mm Hg).
5	<i>serum_chol</i>	Taux de cholestérol sérique (mg/dl).
6	<i>fast_blood_sugar</i>	Glycémie à jeun > 120 mg/dl (1 = vrai; 0 = faux).
7	<i>rest_ecg_res</i>	Résultats de l'électrocardiogramme au repos (0 = normal, 1 = ST-T anormal, 2 = hypertrophie ventriculaire).
8	<i>max_heart_rate</i>	Fréquence cardiaque maximale atteinte.
9	<i>ex_angina</i>	Angine induite par l'exercice (1 = oui; 0 = non).
10	<i>oldpeak</i>	Dépression ST induite par l'exercice par rapport au repos.
11	<i>slope_peak</i>	Pente du segment ST lors de l'exercice (1 = pente descendante, 2 = plate, 3 = ascendante).
12	<i>nb_maj_ves</i>	Nombre de vaisseaux principaux colorés par fluoroscopie (0-3).
13	<i>thal</i>	Résultat de l'examen sanguin (3 = normal, 6 = fixe, 7 = réversible).
14	<i>target</i>	Variable cible indiquant la présence (2) ou l'absence (1) de maladie cardiaque.

Tableau 1 : Variables renommées et description

La variable « *nb_maj_ves* » est initialement de type numérique ; elle a été recodée en catégorielle en raison de sa description et cela fait plutôt sens.

1.3 Méthodologie

Le projet a été réalisé sous le langage Python à l'aide de *Jupyter NoteBook* en suivant la ligne directive :

- définition de la problématique (rappel) : prédire la présence ou l'absence de maladie du cœur à partir de 13 attributs représentant l'état physiologique du patient ;
- collecte de données : [Kaggle](#) ;
- exploration préparatoire des données : analyse univariée et bivariée des données, *traitement des données manquantes et aberrantes, sélection de variables* ;
- division « Train » / « Test » ;
- construction de différents modèles ;
 - entraînement sur la base « Train » avec différents modèles
 - prédition sur la base « Test »
- évaluation des modèles et choix du modèle final ;
- déploiement du modèle : pas d'application développée dans ce projet.

2 Exploration préparatoire des données

Il est important de souligner qu'après la collecte des données, on remarque que le jeu de données ne contient aucune donnée manquante et compte peu de données aberrantes (extrêmes) qui seront éventuellement traitées avec une des solutions parmi : la suppression, la transformation avec le logarithme ou Box-Cox en fonction de leur représentation dans le jeu de données.

Quant à la sélection de variables notamment utile pour les modèles simples tels que la régression logistique, il n'est pas nécessaire de l'effectuer ici, au vu de la quantité faible de variables ; en revanche, elle est plus ou moins employée à travers l'analyse bivariée entre les variables explicatives et la variable à expliquer et les variables explicatives quantitatives entre elles pour vérifier l'intensité des relations et pouvoir éviter par exemple la multi colinéarité dans les modèles.

Ce chapitre se concentre donc sur l'analyse univariée et bivariée en mettant l'accent sur la relation des variables avec la cible et celle de variables explicatives quantitatives entre elles.

2.1 Analyse univariée

2.1.1 Variables quantitatives

Pour les variables numériques, on examine principalement :

- les statistiques descriptives : moyenne, médiane, écart-type, etc. ;
- la distribution via les boîtes à moustaches (boxplots), diagrammes Quantile-Quantile

2.1.1.1 Résumé numérique des variables quantitatives

Sur 270 individus observés, le *Tableau 2* présente les principaux indicateurs statistiques des variables quantitatives : « *age* », « *rest_blood_pres* », « *serum_chol* » , « *max_heart_rate* », et « *oldpeak* ».

Variable	Moyenne	Médiane	Max	Min	Ecart-type	Q1	Q3	IQR	Asymétrie	Aplatissement
<i>age</i>	54.433333	55.0	77.0	29.0	9.109067	48.0	61.0	13.0	-0.163615	-0.544815
<i>rest_blood_pres</i>	131.344444	130.0	200.0	94.0	17.861608	120.0	140.0	20.0	0.722618	0.923097
<i>serum_chol</i>	249.659259	245.0	564.0	126.0	51.686237	213.0	280.0	67.0	1.183721	4.895599
<i>max_heart_rate</i>	149.677778	153.5	202.0	71.0	23.165717	133.0	166.0	33.0	-0.527737	-0.103072
<i>oldpeak</i>	1.050000	0.8	6.2	0.0	1.145210	0.0	1.6	1.6	1.262893	1.759317

Tableau 2 : Résumé numérique des variables quantitatives

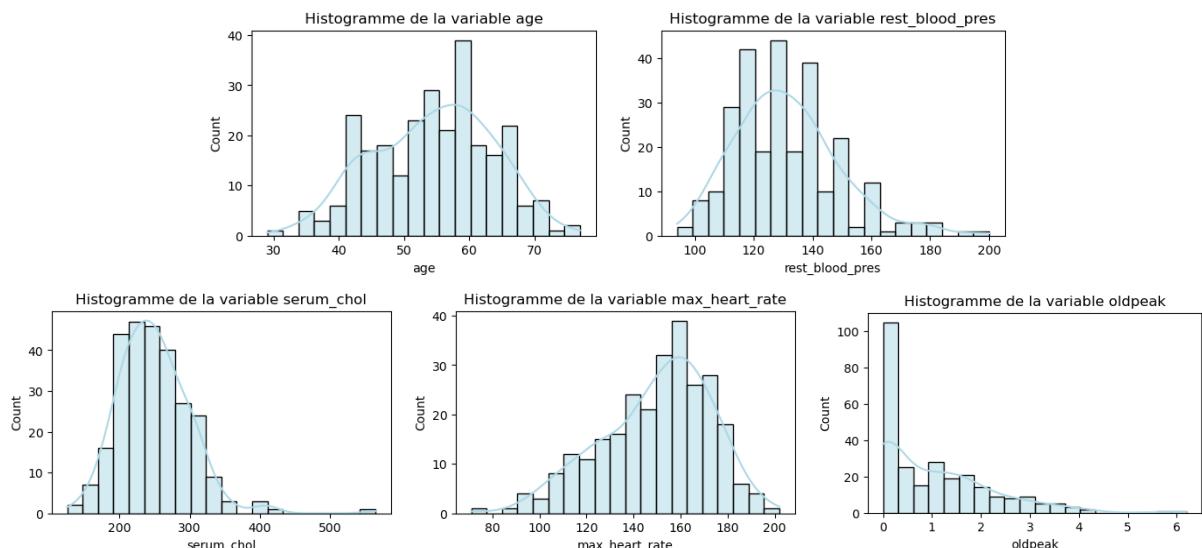
L'individu moyen issu des 270 observations est âgé de 54 ans, a une pression artérielle au repos « *rest_blood_pres* » de 131 ; son taux de cholestérol « *serum_chol* » envoisine 245 mg/dl avec une fréquence cardiaque maximale atteignant 150. Sa dépression ST induite par l'exercice par rapport au repos est évaluée à 1.05. Parmi les individus, on constate que :

- 25% ont moins de 48 ans, moins de 213 mg/dl de taux de cholestérol « *serum_chol* » ;
- 50% ont entre 48 et 61 ans, avec une fréquence cardiaque maximale « *max_heart_rate* » située entre 133 et 166 ;
- 25% ont plus de 61 ans, et plus de 140 mm Hg comme pression artérielle au repos « *rest_blood_pres* ».

Selon les indicateurs de forme, asymétrie et aplatissement, il semble que les distributions des variables quantitatives ne suivent pas la loi normale.

2.1.1.2 Graphes des variables quantitatives

La *Figure 1* illustre les histogrammes des variables « *age* », « *rest_blood_pres* », « *serum_chol* » , « *max_heart_rate* », et « *oldpeak* ».



*Figure 1 : Histogrammes de « *age* », « *rest_blood_pres* », « *serum_chol* » , « *max_heart_rate* », et « *oldpeak* »*

La même interprétation que le résumé numérique se précise à travers ces histogrammes. A l'aide de l'Annexe 2 qui illustre les courbes de densité, les boîtes à moustaches (Boxplot) et les diagrammes Quantile-Quantile (QQ Plot) des variables « age », « rest_blood_pres », « serum_chol », « max_heart_rate », et « oldpeak », on peut ajouter que les variables autres que l'âge « age » comportent quelques valeurs extrêmes : « rest_blood_pres : 9 », « serum_chol : 5 », « max_heart_rate : 1 », « oldpeak : 4 ».

Les diagrammes Quantile-Quantile permettant de comparer la distribution d'une variable avec une distribution théorique (généralement la distribution normale), montrent que les variables quantitatives ne suivent pas une loi normale car tous les points n'appartiennent pas à la droite en rouge, autrement les déviations par rapport à la ligne droite indiquent des écarts par rapport à la distribution normale. L'âge « age » tendrait mieux vers une loi normale. Cela se confirme d'ailleurs à l'aide du test de normalité de Kolmogorov-Smirnov et de Jarque Bera :

Variable	Kolmogorov-Smirnov Stat	KS p-value	Jarque-Bera Stat	JB p-value
age	0.066496	1.755233e-01	4.680560	9.630065e-02
rest_blood_pres	0.100370	8.052478e-03	32.027587	1.109936e-07
serum_chol	0.050955	4.695539e-01	319.750186	3.690875e-70
max_heart_rate	0.077910	7.147221e-02	12.564759	1.868948e-03
oldpeak	0.179607	4.308994e-08	103.669017	3.080059e-23

Tableau 3 : Tests de normalité de Kolmogorov-Smirnov et Jarque-Bera

- le test de normalité de Shapiro-Wilk n'est pas adapté à un jeu de données de plus de 50 observations comme notre cas ;
- les tests de normalité de Kolmogorov-Smirnov et de Jarque-Bera traduisent que seule la variable « age » suit une loi normale ($p\text{-value} > 0.05$), les autres variables ayant des $p\text{-value} < 0.05$ dans l'un des deux tests rejettent l'hypothèse nulle (H_0 : les données suivent une distribution normale). En effet, les variables « serum_chol » et « max_heart_rate » suivent une loi normale pour Kolmogorov-Smirnov mais pas pour Jarque-Bera au sens de l'asymétrie et de l'aplatissement ;
- en revanche, selon le théorème « Central Limit », on peut supposer que toutes les variables quantitatives suivent la loi normale au vu du nombre d'observations.

2.1.1.3 Détection de valeurs extrêmes

Les boîtes à moustaches présentées à l'Annexe 2 montrent la présence de valeurs extrêmes sur les variables quantitatives sauf « age ». Voici le nombre détecté dans le jeu de données et leur représentation en pourcentage :

Variable	Nombre de valeurs extrêmes	Pourcentage (%)
age	0	0
rest_blood_pres	9	3
serum_chol	5	2
max_heart_rate	1	0
oldpeak	4	1

Tableau 4 : Valeurs extrêmes

Ces valeurs extrêmes ne seront pas traitées dans le cadre de cette étude en raison de leur faible représentation et du manque d'information par rapport au métier.

2.1.2 Variables qualitatives

Le jeu de données présente plus de variables catégorielles pour lesquelles, on analyse :

- la répartition des catégories qui permet de comprendre la diversité au sein d'une variable ;
- la fréquence des catégories à travers les diagrammes en barres (Barplots).

2.1.2.1 Résumé numérique des variables qualitatives

Le résumé numérique exhaustif des variables qualitatives peut être consulté à l'

Annexe 3. On note par exemple 32% de femmes (87) contre 68% d'hommes (183).

2.1.2.2 Graphes des variables qualitatives

La *Figure 2* met en évidence la répartition des modalités des variables « *sex* », « *thal* » (*examen sangin*), « *rest_ecg_res* » (résultats de l'électrocardiogramme au repos), « *target* » (variable cible), et « *oldpeak* ».

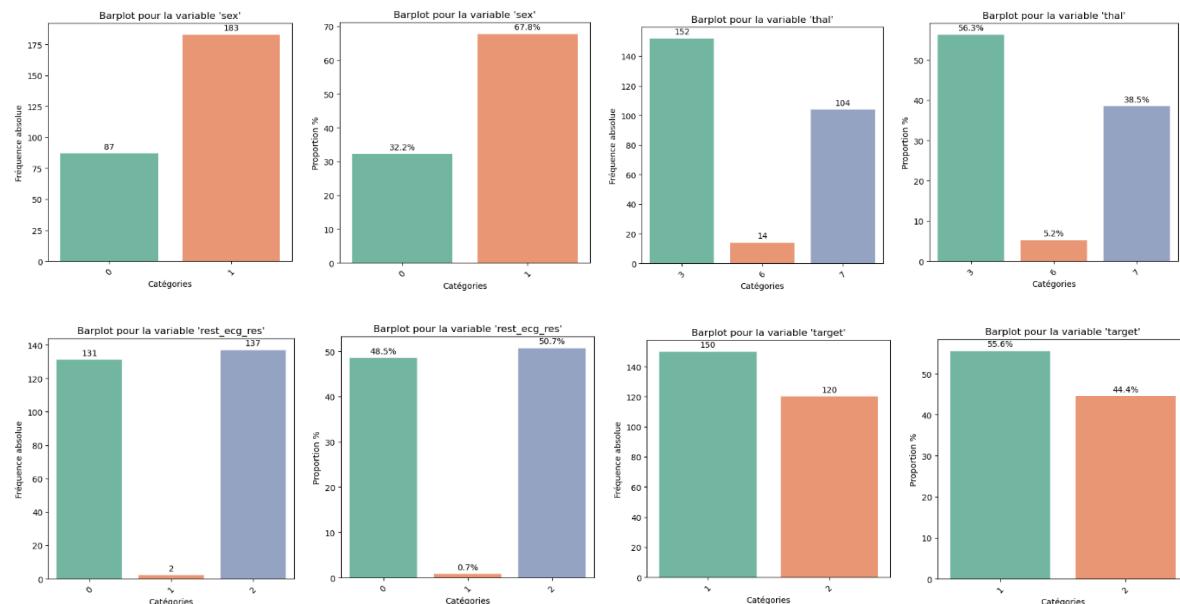


Figure 2 : Répartition en effectif (à gauche) en proportion (à droite) de la variable grp_imc

Elle permet de visualiser la répartition des modalités et d'identifier la catégorie dominante ou les catégories les plus fréquentes. On note à travers la variable cible « *target* », que 44% des individus du jeu de données ont une maladie cardiaque et 56% ne souffrent pas de maladie du cœur. La répartition des deux modalités de cette variable permet une modélisation décisionnelle pour prédire la présence ou l'absence de maladie cardiaque.

Existe-t-il des modalités rares dans le jeu de données ?

2.1.2.3 Détection de modalité rare

Lorsqu'on observe la variable « *rest_ecg_res* » ayant trois modalités (0 = normal, 1 = ST-T anormal, 2 = hypertrophie ventriculaire), la modalité « 1 » représente 0,7%, donc largement moins de 5% et il serait pertinent de l'associer dans la modalité « 2 » au vu de sa rareté,

autrement effectuer un recodage pour avoir au final deux modalités (normal, anormal) pour cette variable ; ce qui a été fait par la suite.

L'analyse bivariée permettra d'avoir une première intuition de la liaison ou relation entre les variables en mettant l'accent sur celle avec la variable cible, afin d'éviter d'intégrer dans le modèle les variables apportant quasiment la même information.

2.2 Analyse bivariée

2.2.1 Lien entre variables quantitatives

2.2.1.1 Nuage de points

La *Figure 3* illustre le nuage de points entre « age » et « max_heart_rate » (à gauche) et entre « age » et « rest_blood_pres » (à droite)

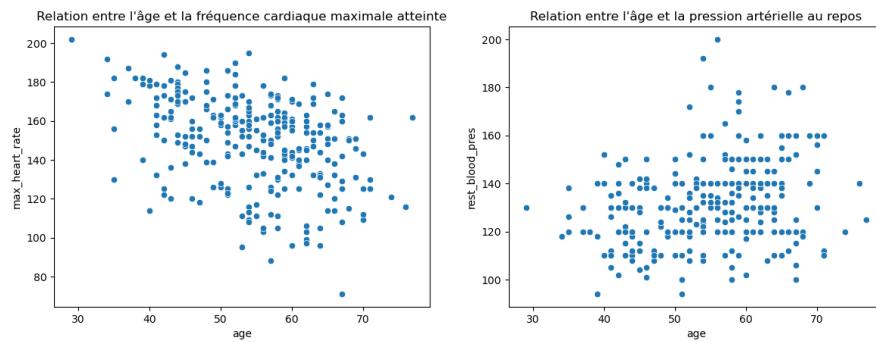


Figure 3 : Nuage de points entre « age » et « max_heart_rate » (à gauche) et entre « age » et « rest_blood_pres » (à droite)

Le nuage de points à gauche illustre la relation entre l'âge et la fréquence cardiaque maximale atteinte, celui à gauche montre la relation entre l'âge et la pression artérielle au repos. Qu'observe-t-on ?

- il y a une tendance générale où, à mesure que l'âge augmente, la fréquence cardiaque maximale atteinte tend à diminuer ;
- et lorsque l'âge augmente, la pression artérielle au repos tend également à augmenter. C'est d'ailleurs en général l'intuition que l'on a ;
- on note une dispersion de quelques points. Cela reflète une diversité de fréquence cardiaque maximale atteinte ou pression artérielle au repos pour un même âge, ce qui est cohérent avec la variabilité des caractéristiques humaines.

En conclusion, la corrélation entre l'âge et la fréquence cardiaque maximale atteinte semble être négative, mais elle n'est pas parfaite ; et la corrélation entre l'âge et la pression artérielle au repos semble être positive, mais elle n'est pas parfaite. Le calcul de la corrélation aidera à quantifier la force de cette relation.

2.2.1.2 Corrélation entre les variables et test de Pearson

Selon le théorème « Central Limit », on peut supposer que toutes les variables quantitatives suivent la loi normale au vu du nombre d'observations. Ainsi, il est utilisé ici le coefficient de corrélation de Pearson pour quantifier les relations entre variables quantitatives.

La *Figure 4*, appelée « heatmap de corrélation » à gauche, montre la corrélation entre les variables quantitatives, accompagnée du test de corrélation « p-value » à droite.

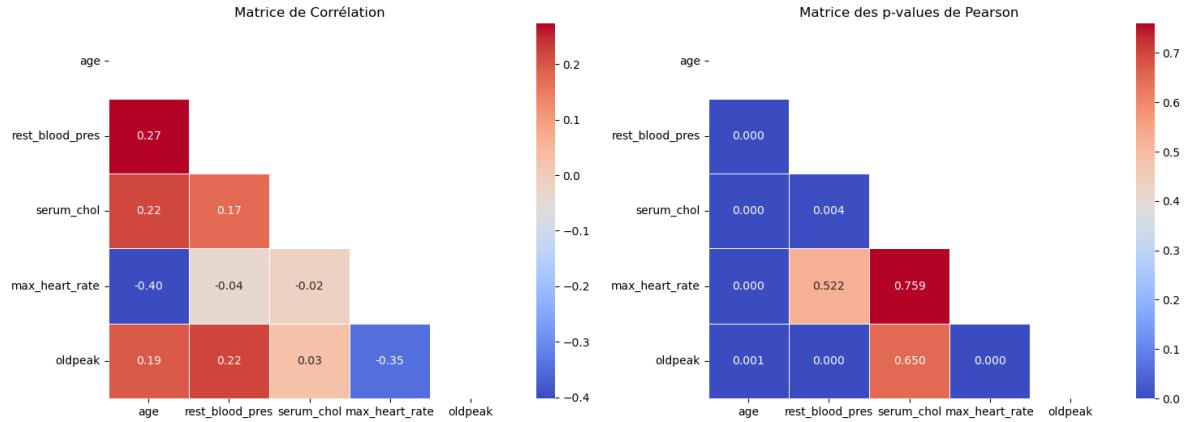


Figure 4 : Matrice de corrélation et p-value de Pearson

A gauche, la matrice de corrélation révèle au moyen des coefficients de corrélation de Pearson qu'il existe une relation linéaire notamment entre « *age* » et « *max_heart_rate* » (fréquence cardiaque maximale atteinte) d'une part et, « *oldpeak* » (dépression ST induite par l'exercice par rapport au repos) et « *max_heart_rate* » d'autre part.

A droite, avec le test d'hypothèse de Pearson, on observe que :

- toutes les cases bleues indiquant une $p\text{-value} < 0.05$ expliquent qu'il y a une relation linéaire entre les variables même si les coefficients paraissent faibles (intensité faible de la relation) ;
 - exemple : il y a une corrélation significative (relation linéaire négative et faible) entre « *age* » et « *max_heart_rate* » ; ces deux variables incluses dans le modèle de regression logistique par exemple provoqueront des problèmes causés par la multicolinéarité.
- Les $p\text{-value} > 0.05$ traduisent qu'il n'y a pas de relation linéaire entre les variables. Ainsi, les variables « *max_heart_rate* », « *rest_blood_pres* », « *serum_chol* » prises ensemble ne seraient pas source de problème de multicolinéarité.

2.2.2 Lien entre variables quantitatives et qualitatives

Sous la contrainte du nombre de pages du présent document, l'accent est mis ici sur le lien entre les variables quantitatives (toutes explicatives) et la variable qualitative cible afin d'avoir une visibilité sélective de variables. Se reporter au *Jupyter NoteBook* pour l'analyse bivariée exhaustive.

2.2.2.1 Résumé numérique par modalité des variables qualitatives

Le résumé numérique par modalité de la variable cible « *Target* » en fonction des variables quantitatives peut être consulté à l'*Annexe 5*, et plus encore dans le *Jupyter NoteBook* sur l'ensemble des variables du jeu de données.

2.2.2.2 Boîtes à moustaches bivariées

Les boîtes à moustaches bivariées entre toutes les variables quantitatives explicatives et la *Target* peuvent être observées à l'*Annexe 5*, et de manière exhaustive avec les autres variables qualitatives dans le *Jupyter NoteBook*.

En se basant sur ces boîtes représentant les différentes modalités, on a comme intuition que la dépression ST induite par l'exercice par rapport au repos « *oldpeak* », l'âge « *age* », le taux de cholestérol sérique « *serum_chol* », la pression artérielle au repos « *rest_blood_pres* » auraient une influence remarquable sur la présence ou l'absence d'une maladie cardiaque ; la fréquence cardiaque maximale atteinte quant à elle, aurait une liaison beaucoup plus faible avec le « *Target* ».

Que peut-on conclure avec les tests d'hypothèses de relations bivariées entre les variables quantitatives et qualitatives ?

2.2.2.3 Tests d'hypothèses sur les relations bivariées entre les variables quantitatives et qualitatives

D'après le *Tableau 5*, les tests statistiques de Student et de Wilcoxon-Mann-Witney montrent et confirment que les distributions des groupes en fonction de la « *target* » sont significativement différentes en raison des p-value < 0,05.

	Variables	Student Stat	t p-value	Mann-Whitney Stat	MWW p-value
Target (2 vs 1)	oldpeak	7.531875	7.677946e-13	13170.5	2.979412e-11
	age	3.556964	4.434804e-04	11366.0	2.051984e-04
	rest_blood_pres	2.574999	1.056095e-02	10699.5	7.700853e-03
	serum_chol	1.945677	5.273889e-02	10367.5	3.154067e-02
	max_heart_rate	-7.543813	7.119583e-13	4611.5	5.841045e-12

Tableau 5 : Analyse bivariée entre variables quantitatives et variable cible - résultats des tests de Student et de Mann-Whitney

On observe avec la *Figure 5* ci-dessous, selon le test de Mann-Whitney, le classement de la relation entre les variables quantitatives et la variable cible.

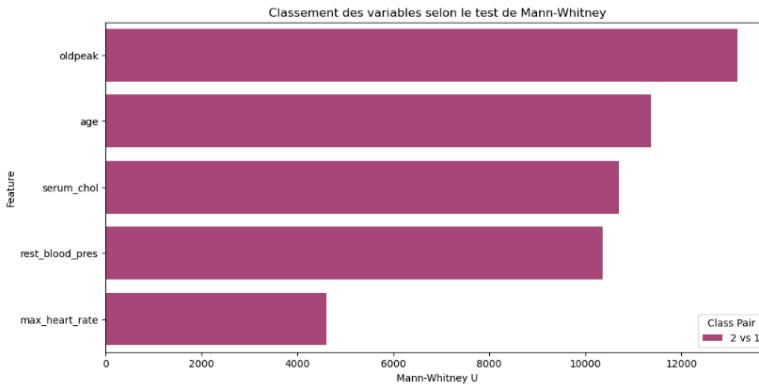


Figure 5 : Classement de la relation entre les variables quantitatives et la variable cible selon le test de Mann-Whitney

2.2.3 Lien entre variables qualitatives

On met surtout en évidence ici, le lien entre les variables qualitatives explicatives et la variable qualitative cible. Cela aidera en effet à la sélection de variables si besoin dans les modèles.

2.2.3.1 Tableaux de contingence et Heatmap des pourcentages de liaison

On s'intéresse ici aux heatmap des pourcentages de liaison avec la variable cible « Target » appuyés sur leurs tableaux de contingence :

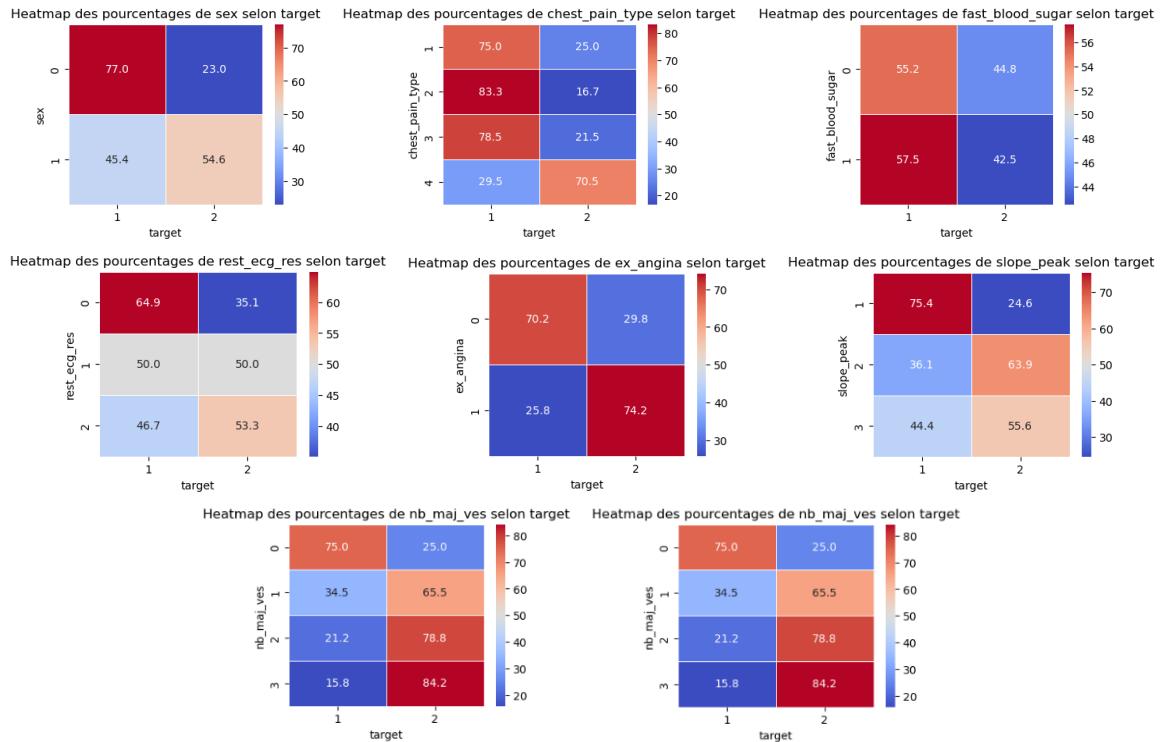


Figure 6 : Heatmap des pourcentages de liaison entre les variables qualitatives et la « Target »

On note une forte liaison dans les cases orange et rouges entre la « *Target* » et les modalités concernées.

Les tableaux de contingence entre la variable cible « *Target* » et les variables qualitatives peut être consulté à l'*Annexe 5*, et plus encore dans le *Jupyter NoteBook* sur l'ensemble des variables qualitatives du jeu de données.

2.2.3.2 Tests d'hypothèses sur les relations bivariées entre les variables qualitatives

Le *Tableau 6* affiche les tests statistiques de Chi2 et le V de Cramer entre la variable « *target* » et les variables qualitatives explicatives.

	Variable	Chi2 Statistic	P-Value	V de Cramer
Target	thal	74.569346	6.419071e-17	0.525531
	chest_pain_type	68.588207	8.560988e-15	0.504014
	nb_maj_ves	62.863092	1.436620e-13	0.482521
	ex_angina	45.691873	1.383958e-11	0.411375
	slope_peak	40.370391	1.712699e-09	0.386678
	sex	22.667256	1.926226e-06	0.289746
	rest_ecg_res	8.979452	1.122372e-02	0.182366
	fast_blood_sugar	0.009171	9.237061e-01	0.005828

Tableau 6 : Liaison entre variables qualitatives explicatives et la Target - résultats du test de Chi2 et V de Cramer

Les résultats de ces tests statistiques montrent et confirment que la variable « *target* » et les variables qualitatives explicatives sauf « *fast_blood_sugar* » sont significativement dépendantes en raison des p-value < 0,05. Ce qui aidera à privilégier les liaisons les plus fortes lors de la modélisation.

On observe avec la *Figure 7* ci-dessous selon le test de Mann-Whitney, le classement de la relation entre les variables quantitatives et la « *Target* ».

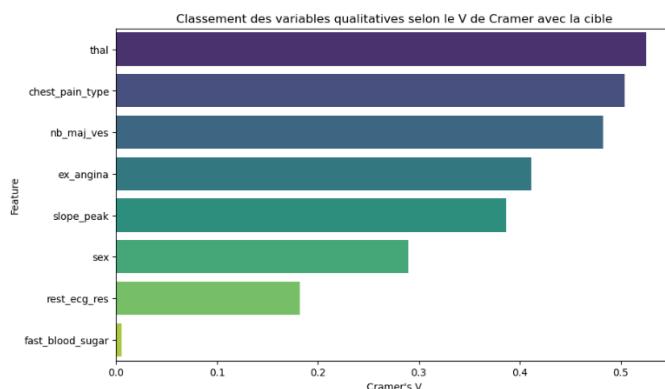


Figure 7 : Classement de la relation entre les variables qualitatives et la Target selon le V de Cramer

3 Construction de différents modèles

Dans le *Jupyter NoteBook*, plusieurs modèles ont été développés à savoir :

- **régression logistique** ;
- régression logistique avec pénalisation Ridge, Lasso et ElasticNet ;
- arbre de décision (son sur-apprentissage est élevé sans optimisation) ;
- bagging : **forêt aléatoire** (Random Forest) ;
- boosting : Adaboost, LightGBM, CatBoost, **XgBoost** ;
- **K-Nearest Neighbors (KNN)**.

En raison du nombre de pages limité de ce rapport, quatre parmi eux, ceux indiqués en gras ci-dessus sont mis en exergue dans le présent document. Se référer au *Jupyter NoteBook* pour voir la construction et l'évaluation des autres.

3.1 Division « Train » / « Test »

Il est important de diviser le jeu de données en deux parties : la base d'apprentissage et la base de test afin d'évaluer les performances des modèles. Dans le cadre de l'étude, la première représente 80% et la seconde, la base de test, 20%.

Le prétraitement de données a été effectué suite à cette division en appliquant la technique du « *One-hot Encoding* » sur les variables qualitatives explicatives puisque les modèles reconnaissent et comprennent le format binaire ‘0’ et ‘1’.

Il faut noter que le package « Scikit Learn » a été utilisé pour construire les modèles.

3.2 Modèle de régression logistique

La régression logistique est une méthode statistique utilisée pour prédire la probabilité qu'un événement survienne, en se basant sur une ou plusieurs variables explicatives. Contrairement à la régression linéaire, où la variable dépendante est continue, la régression logistique est utilisée lorsque la variable cible est binaire ou catégorielle.

3.2.1 Entraînement sur la base « Train »

Le modèle de régression logistique a été entraîné en incluant la constante « intercept » et avec un nombre d'itérations fixé à 1050 :

- d'abord en considérant toutes les variables : arrêt de l'algorithme à l'itération 1049 ;
- ensuite en retirant une à une les variables source de multi colinéarité en utilisant la fonction VIF (Facteur d'inflation de la variance) : arrêt de l'algorithme à l'itération 22.

Les coefficients obtenus après l'entraînement de la régression logistique sont montrés à l'*Annexe 7*, ainsi que les variables source de multi colinéarité.

3.2.2 Prédiction sur la base « Test »

La prédiction avec le modèle de régression logistique a été effectuée sous trois différents seuils de probabilité 0.50, 0.45, 0.40 en utilisant d'une part le modèle avec toutes les variables et d'autre part celui sans les variables source de multi colinéarité ('rest_blood_pres', 'age', 'max_heart_rate', 'serum_chol') pour optimiser.

3.2.3 Evaluation du modèle

Il est important d'évaluer tout modèle pour vérifier le sur-apprentissage ou le sous-apprentissage. Les métriques utilisées pour la régressions logistique (et tout modèle de classification) sont : Accuracy, Precision, Recall, F1 et AUC .

Ci-dessous les résultats :

- avec toutes les variables :

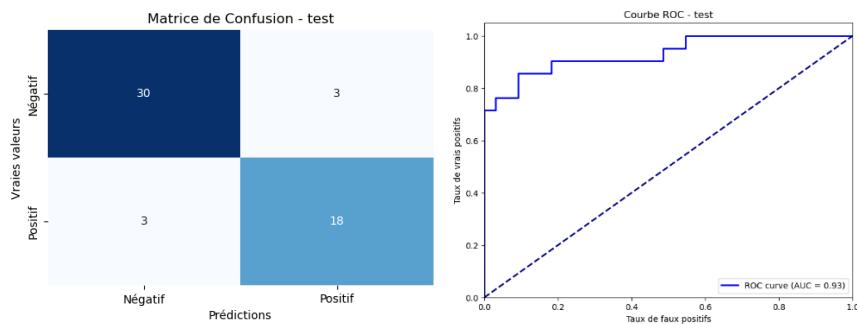


Figure 8 : Matrice de confusion et courbe ROC de régression logistique avec toutes les variables

Seuil	Echantillon	Accuracy	Precision	Recall	F1	AUC
0.5	Train	0.88	0.88	0.85	0.87	0.93
	Test	0.89	1.00	0.71	0.83	0.93
0.45	Train	0.87	0.84	0.87	0.86	0.93
	Test	0.89	0.94	0.76	0.84	0.93
0.40	Train	0.86	0.81	0.89	0.85	0.93
	Test	0.89	0.86	0.86	0.86	0.93

Tableau 7 : Indicateurs de performance du modèle de régression logistique avec toutes les variables

On remarque que ce modèle n'est pas optimal surtout avec les indicateurs « Precision » et « Recall » sous les seuils 0.50 et 0.45 et ne change pas.

- sans les variables source de multi colinéarité ('rest_blood_pres', 'age', 'max_heart_rate', 'serum_chol') : on obtient un bon compromis Precision vs Recall avec un seuil de probabilité de 0.40

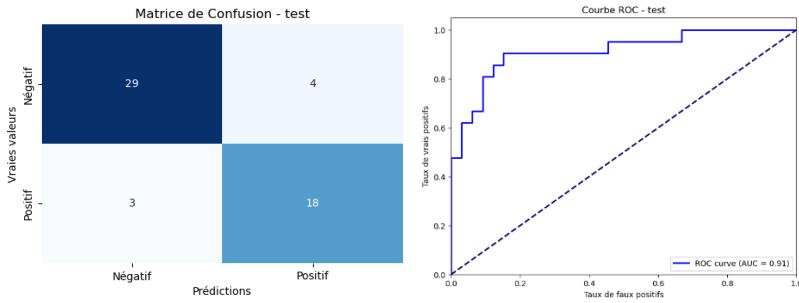


Figure 9 : Matrice de confusion et courbe ROC de régression logistique sans les variables source de multi-colinéarité

Seuil	Echantillon	Accuracy	Precision	Recall	F1	AUC
0.5	Train	0.86	0.87	0.82	0.84	0.93
	Test	0.81	0.87	0.61	0.72	0.91
0.45	Train	0.86	0.84	0.86	0.85	0.93
	Test	0.81	0.82	0.67	0.74	0.91
0.40	Train	0.87	0.83	0.90	0.86	0.93
	Test	0.87	0.82	0.86	0.84	0.91

Tableau 8 : Indicateurs de performance du modèle de régression logistique sans les variables source de multi-colinéarité

Ce modèle sans les variables source de multi colinéarité est plus réaliste du point de vue de l'AUC. Sous le seuil de 0.40, on a un bon compromis entre les indicateurs « Precision » et « Recall ».

3.3 Modèle Forêt aléatoire

En raison des inconvénients de l'arbre de décision tels que le surajustement observé lors de l'évaluation (cf. l'Annexe 8 et le Jupyter Notebook), l'instabilité, le biais de sélection de variable), il est nécessaire d'utiliser des techniques de régularisation ou d'autres méthodes d'apprentissage automatique comme les forêts aléatoires qui prennent en compte plusieurs arbres de décision et appliquent un vote à la majorité dans le cas d'une classification.

3.3.1 Entraînement du modèle Random Forest sur la base « Train »

Lorsque l'on utilise une forêt aléatoire, il est essentiel de faire des choix judicieux pour certains paramètres afin d'optimiser les performances du modèle. Deux paramètres importants à considérer sont :

- *max_features* : permet de contrôler la complexité du modèle et d'éviter le surajustement (overfitting) en réglant la profondeur maximale de chaque arbre de décision dans la forêt. **L'approche courante consistant à fixer *max_features* à la racine carrée du nombre de variables utilisées** a été employée dans la modélisation. Cela permet de limiter la profondeur des arbres et de prévenir une croissance excessive, tout en maintenant un bon équilibre entre sous-apprentissage et surapprentissage.

- *n_estimators* : Il s'agit du nombre d'arbres de décision dans la forêt aléatoire. Un nombre suffisamment grand d'arbres permet d'améliorer les performances du modèle. Cependant, il est important de noter que l'augmentation de *n_estimators* entraîne également une augmentation du temps de calcul. On cherche plutôt à trouver un compromis entre les performances du modèle et le temps de calcul disponible.

L'estimateur optimal du modèle Random Forest non optimisé a comme paramètres :
`max_features=4, n_estimators=300`

```
RandomForestClassifier
RandomForestClassifier(max_features=4, n_estimators=300, oob_score=True,
random_state=42)
```

L'estimateur optimal du modèle Random Forest optimisé a comme paramètres :
`max_depth=2, n_estimators=300`

```
RandomForestClassifier
RandomForestClassifier(max_depth=2, n_estimators=300, oob_score=True,
random_state=42)
```

3.3.2 Prédiction du modèle Random Forest (RF) sur la base « Test »

La prédiction a été effectuée sous trois différents seuils de probabilité 0.50, 0.45, 0.40 avec :

- le modèle RF non optimisé suivant les paramètres `max_features=4, n_estimators=300` ;
- le modèle RF optimisé suivant les paramètres `max_depth=2, n_estimators=300`.

Qu'a-t-on observé à l'évaluation ?

3.3.3 Evaluation du modèle RF

- Modèle RF non optimisé : on note un surapprentissage

Seuil	Echantillon	Accuracy	Precision	Recall	F1	AUC
0.5	Train	1.00	1.00	1.00	1.00	1.00
	Test	0.78	0.80	0.57	0.67	0.87
0.45	Train	1.00	1.00	1.00	1.00	1.00
	Test	0.78	0.80	0.57	0.67	0.87
0.40	Train	1.00	1.00	1.00	1.00	1.00
	Test	0.74	0.67	0.67	0.67	0.87

Tableau 9 : Indicateurs de performance du modèle Random Forest non optimisé

- Modèle RF optimisé : le bon compromis Precision vs Recall s'obtient avec le seuil de probabilité de 0.45

Seuil	Echantillon	Accuracy	Precision	Recall	F1	AUC
0.5	Train	0.85	0.88	0.79	0.83	0.93
	Test	0.78	0.80	0.57	0.67	0.85

0.45	Train	0.84	0.82	0.85	0.83	0.93
	Test	0.76	0.70	0.67	0.68	0.85
0.40	Train	0.82	0.76	0.88	0.82	0.93
	Test	0.72	0.62	0.71	0.67	0.85

Tableau 10 : Indicateurs de performance du modèle Random Forest optimisé

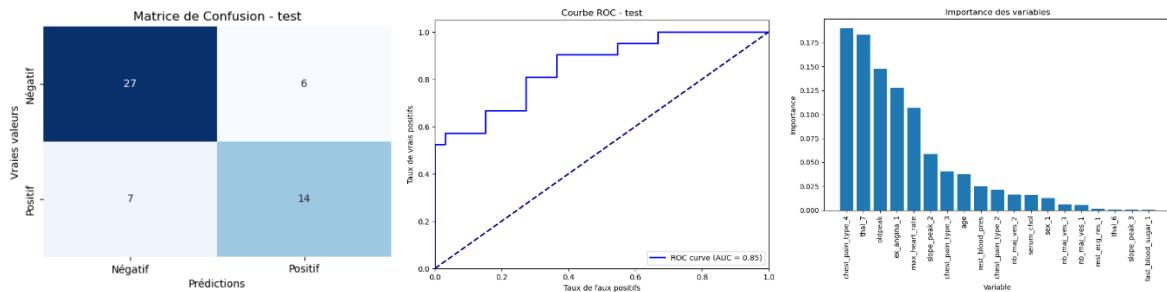


Figure 10 : Test Modèle RF optimisé - Matrice de confusion, courbe ROC et importance des variables

3.4 Modèle XgBoost

Contrairement au Random Forest qui apprend les arbres indépendamment, XGBoost corrige les erreurs des arbres précédents. Il utilise en effet le gradient boosting optimisé pour une meilleure convergence, ce qui le rend plus précis, souvent plus performant et rapide que Random Forest, surtout sur de grands jeux de données, car il optimise la façon dont les arbres sont construits et corrigés.

3.4.1 Entraînement sur la base « Train »

Le modèle XgBoost a été entraîné avec le GridSearch composé des paramètres :

- profondeur maximale de chaque arbre « *max_depth* » : [1,2,3,4]
- taux d'apprentissage « *learning_rate* » : [0.01, 0.1, 0.2]
- nombre d'arbres à construire « *n_estimators* » : [100, 200, 300,500,1000]

Il en ressort, après un temps d'exécution de 10 secondes, le meilleur modèle avec « *learning_rate* » = 0.1, « *max_depth* » = 2, « *n_estimators* » = 100

3.4.2 Prédiction du modèle XgBoost sur la base « Test »

La prédiction a été effectuée sous trois différents seuils de probabilité 0.50, 0.45, 0.40 avec le meilleur modèle obtenu suivant les paramètres sus-cités.

Que donne son évaluation ?

3.4.3 Evaluation du modèle XgBoost avec les meilleurs paramètres

Le meilleur compromis Precision vs Recall s'obtient en ajustant le seuil de probabilité. Tout dépend de l'objectif de l'application. Si par exemple les faux négatifs sont coûteux dans le cadre d'un diagnostic médical par exemple (le cas de la présente étude), il est plutôt pertinent d'augmenter le rappel en réduisant le seuil.

Seuil	Echantillon	Accuracy	Precision	Recall	F1	AUC
0.5	Train	0.94	0.94	0.94	0.94	0.98
	Test	0.76	0.75	0.57	0.65	0.85
0.45	Train	0.92	0.89	0.94	0.92	0.98
	Test	0.76	0.72	0.62	0.67	0.85
0.40	Train	0.92	0.89	0.95	0.92	0.98
	Test	0.78	0.71	0.71	0.71	0.85

Tableau 11 : Indicateurs de performance du modèle XgBoost optimisé

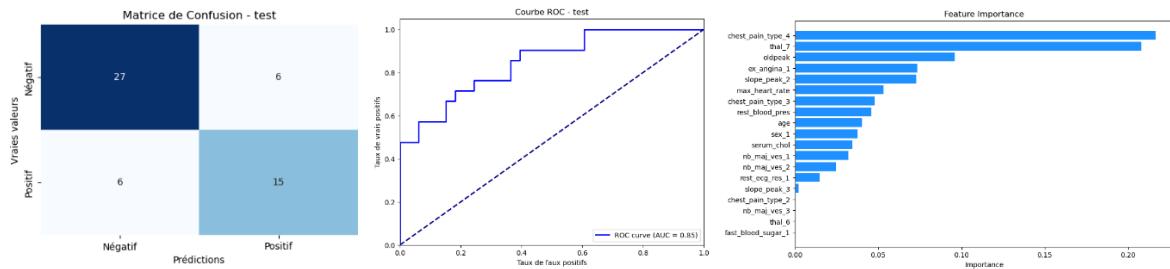


Figure 11 : Test Modèle XgBoost optimisé - Matrice de confusion, courbe ROC et importance des variables

3.5 Modèle K Plus Proches Voisins (KNN)

Le modèle K Plus Proches Voisins est un modèle non paramétrique qui classe une observation en fonction de la classe majoritaire de ses k voisins les plus proches dans l'espace des caractéristiques.

3.5.1 Entraînement du modèle KNN sur la base « Train »

Le modèle KNN a été implémenté et entraîné avec le paramètre « `n_neighbors` », nombre de voisins à considérer pour faire la prédiction. On recherche le meilleur modèle avec la grille : « `n_neighbors` » : [1, 2, 3, 4, 5, 7, 9]

3.5.2 Prédiction du modèle KNN sur la base « Test »

La prédiction sous le modèle KNN a été réalisée avec le meilleur paramètre « *n_neighbors* » obtenu, à savoir 5 avec les résultats d'évaluation ci-après.

3.5.3 Evaluation du modèle KNN

Lorsqu'on diminue le seuil de probabilité, on augmente le « Recall » mais on perd trop en « Precision ».

Seuil	Echantillon	Accuracy	Precision	Recall	F1	AUC
0.5	Train	0.77	0.78	0.70	0.74	0.84
	Test	0.63	0.52	0.62	0.57	0.74
0.45	Train	0.77	0.78	0.70	0.74	0.84
	Test	0.63	0.52	0.62	0.57	0.74
0.40	Train	0.74	0.66	0.86	0.75	0.84
	Test	0.65	0.53	0.95	0.68	0.74

Tableau 12 : Indicateurs de performance du modèle KNN optimisé

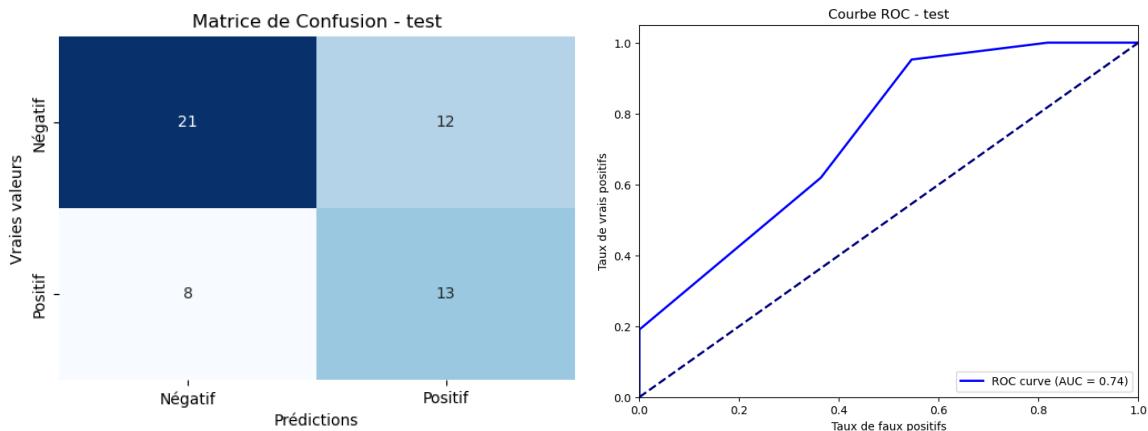


Figure 12 : Test Modèle KNN optimisé - Matrice de confusion, courbe ROC pour le seuil 0.5

4 Comparaison des modèles et choix du modèle final

4.1 Comparaison

Sur la base des indicateurs illustrés à l'*Annexe 10*, les modèles étudiés offrent des performances différentes.

Le modèle de régression logistique pénalisée Ridge L2 se retrouve avec les meilleurs indicateurs de performance pour généraliser. Suit la régression logistique simple mais en excluant les variables quantitatives explicatives ('rest_blood_pres', 'age', 'max_heart_rate', 'serum_chol') source de multi-colinéarité.

Les modèles non paramétriques XgBoost, Random Forest, sont dans l'ordre meilleurs que l'arbre de décision qui surapprend trop et le modèle KNN.

4.2 Choix du modèle final

Au vu du jeu de données comportant relativement un faible nombre d'observation, et suivant les écarts entre l'apprentissage et la prédiction, le modèle de régression logistique me paraît

suffisant pour généraliser. On gagnera ainsi en temps de calcul et sur l'avantage d'interprétation qu'il offre.

5 Conclusion

Le projet a permis de pratiquer les modèles vus durant la formation RCP209 et d'observer leurs avantages et inconvénients. Il est important de noter que le seuil de probabilité a une forte influence sur l'ajustement des différents modèles ; ainsi il urge de bien connaître les données pour juger du bon compromis entre la « Precision » et le « Recall » ou la courbe ROC.

Le projet m'a personnellement permis de pratiquer et de monter en compétence sur l'analyse descriptive au cours l'exploration des données.

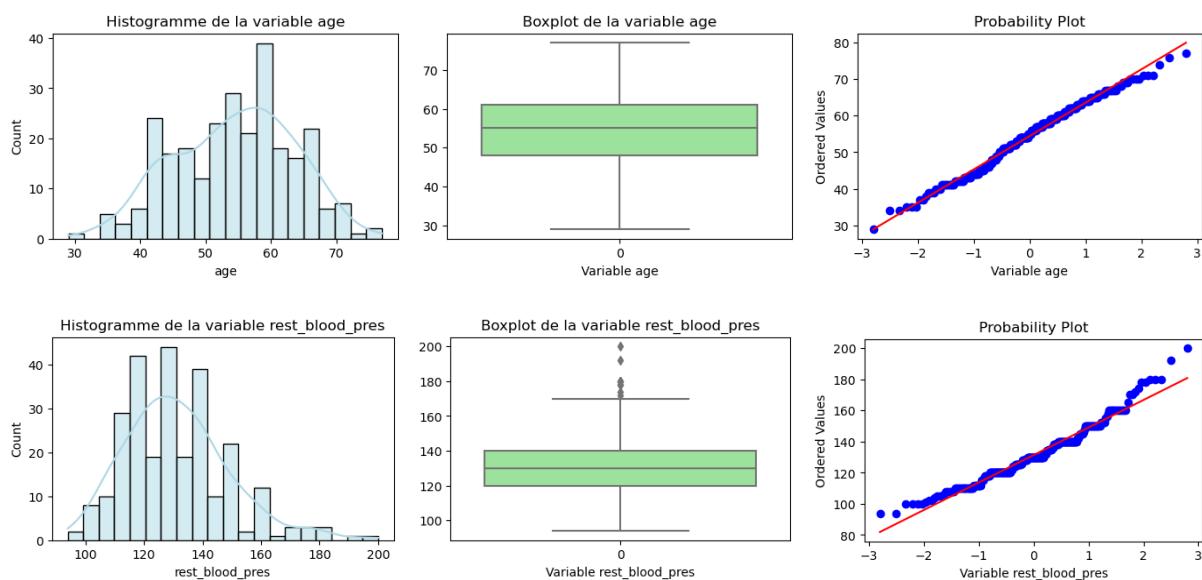
Les réseaux de neurones n'ont pas été entraînés car il est inutile de les appliquer à des jeux de données de faible nombre d'observations. Ils seront mis en pratique à travers d'autres projets.

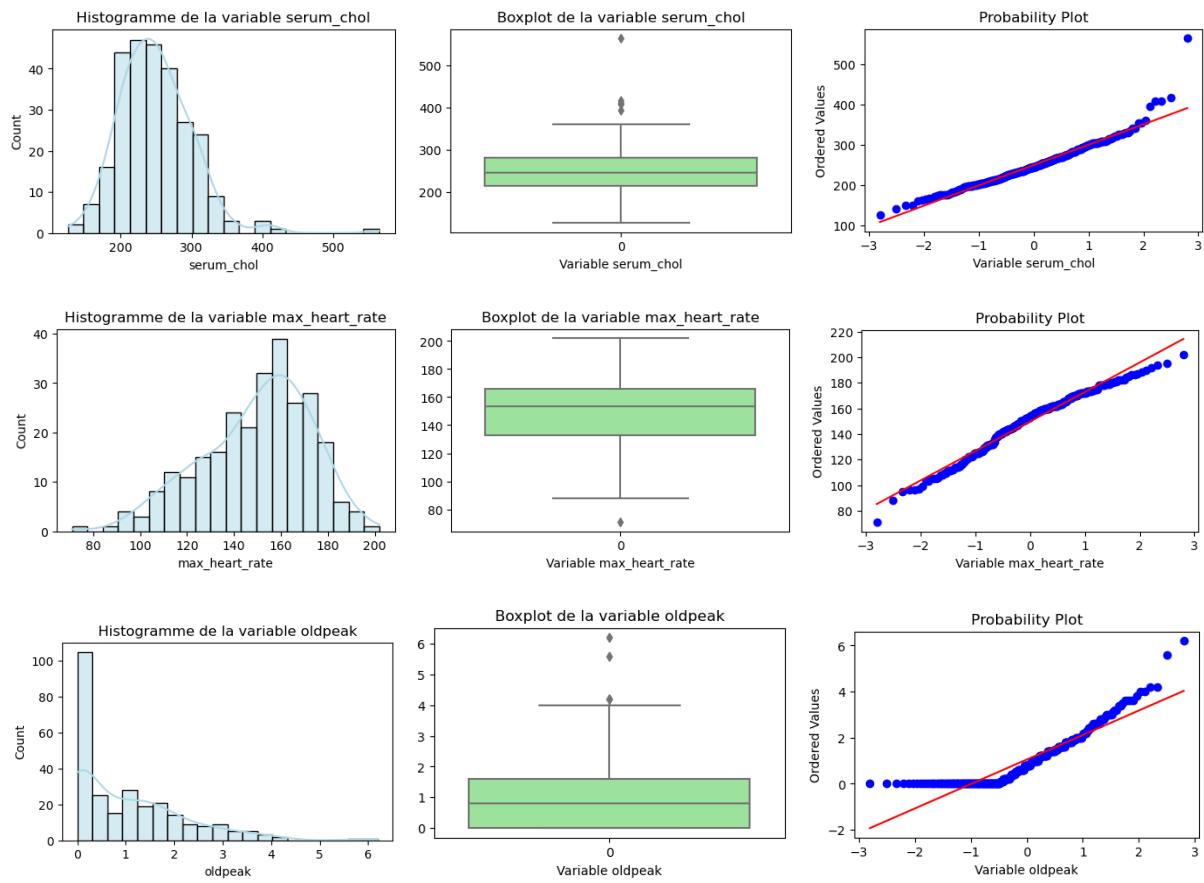
Annexes

Annexe 1 : Tableau et information sur variables à l'origine

N°	Variable	Type	Information
1	age	Real	age
2	sex	Binary	sexe
3	cp	Nominal	chest pain type (4 values)
4	trestbps	Real	resting blood pressure
5	chol	Real	serum cholesterol in mg/dl
6	fbs	Binary	fasting blood sugar > 120 mg/dl
7	restecg	Nominal	resting electrocardiographic results (values 0, 1, 2)
8	thalach	Real	maximum heart rate achieved
9	exang	Binary	exercise induced angina
10	oldpeak	Real	oldpeak = ST depression induced by exercise relative to rest
11	slope	Ordered	the slope of the peak exercise ST segment
12	ca	Real	number of major vessels (0-3) colored by flourosopy
13	thal	Nominal	thal: 3 = normal; 6 = fixed defect; 7 = reversible defect
14	target	Binary	Target (Variable to be predicted) : Absence (1) or presence (2) of heart disease

Annexe 2 : Histogramme et courbe de densité (à gauche), Boxplot (au milieu) et QQ Plot (à droite) de « age », « rest_blood_pres », « serum_chol », « max_heart_rate », et « oldpeak »



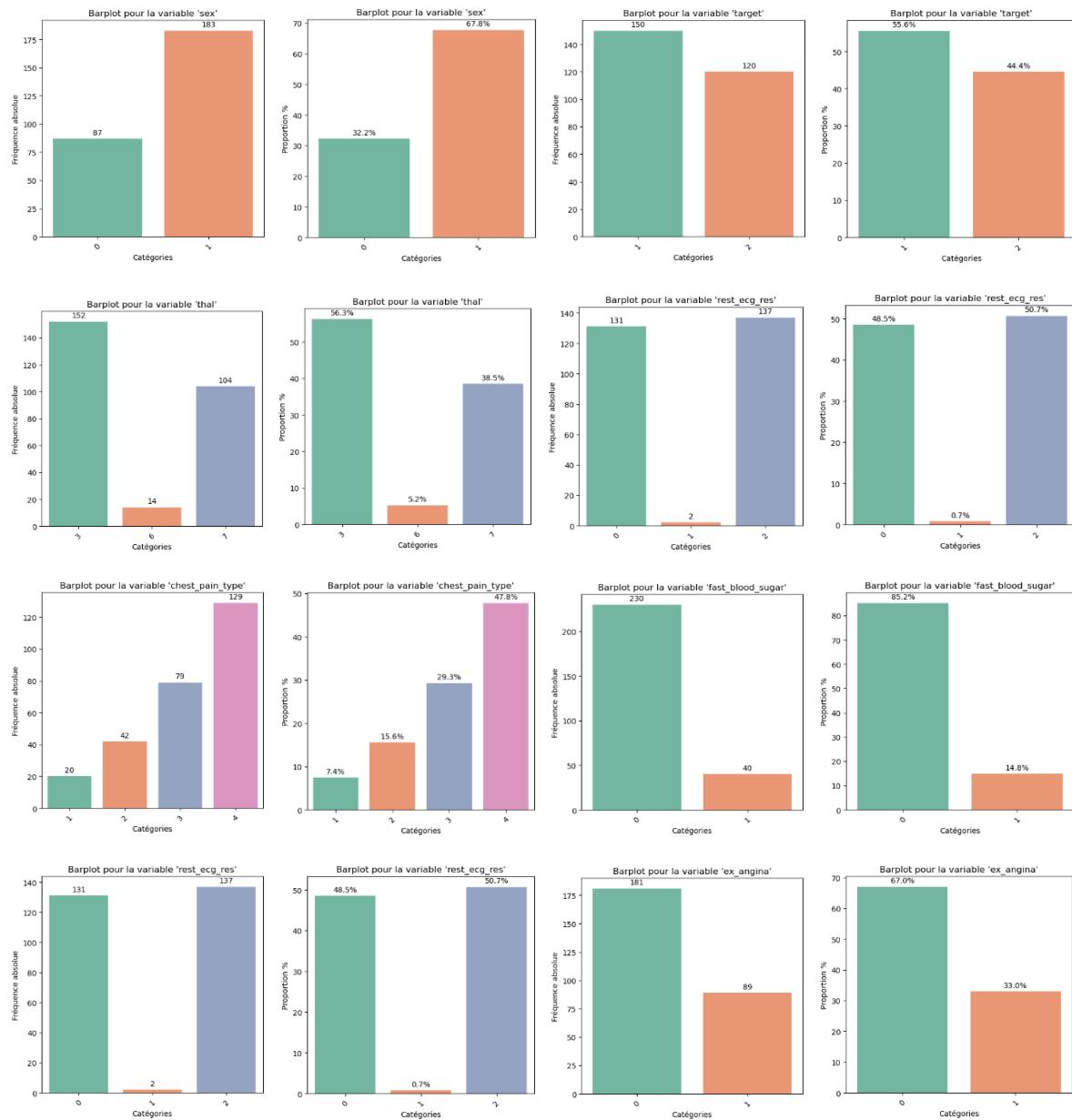


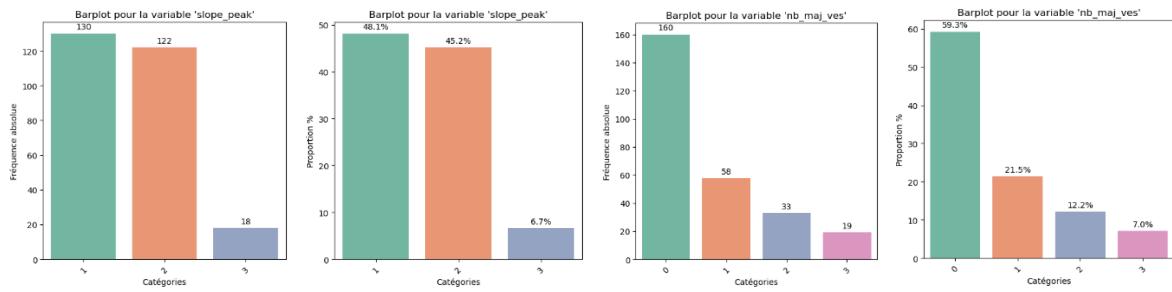
Annexe 3 : Résumé numérique des variables qualitatives

Variable	Catégorie	Fréquence absolue	Proportion	Mode
sex	1	183	67.777778	1.0
	0	87	32.222222	NaN
chest_pain_type	4	129	47.777778	4.0
	3	79	29.259259	NaN
	2	42	15.555556	NaN
	1	20	7.407407	NaN
fast_blood_sugar	0	230	85.185185	0.0
	1	40	14.814815	NaN
rest_ecg_res	2	137	50.740741	2.0
	0	131	48.518519	NaN
	1	2	0.740741	NaN
ex_angina	0	181	67.037037	0.0
	1	89	32.962963	NaN
slope_peak	1	130	48.148148	1.0
	2	122	45.185185	NaN
	3	18	6.666667	NaN
nb_maj_ves	0	160	59.259259	0.0
	1	58	21.481481	NaN
	2	33	12.222222	NaN

	3	19	7.037037	NaN
thal	3	152	56.296296	3.0
	7	104	38.518519	NaN
	6	14	5.185185	NaN
target	1	150	55.555556	1.0
	2	120	44.444444	NaN

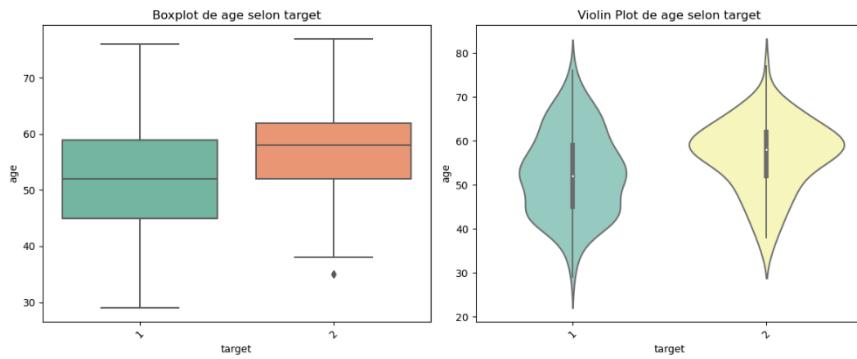
Annexe 4 : Répartition en effectif (à gauche) en proportion (à droite) de quelques modalités





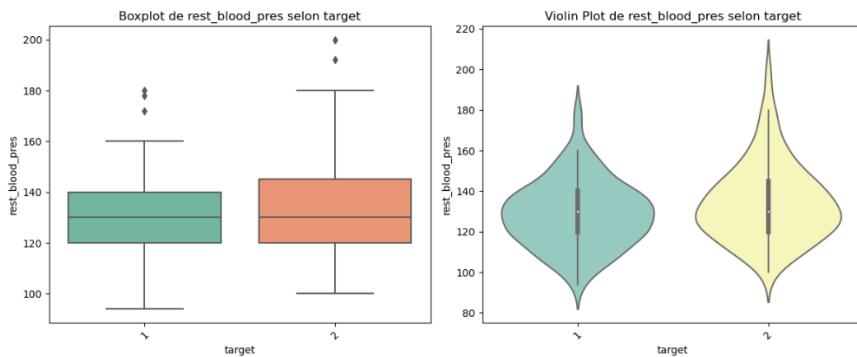
Annexe 5 : Boxplots entre les différentes variables quantitatives et la target, tests statistique

◆ Analyse de age en fonction de target (Cat 1 : 150, Cat 2 : 120)



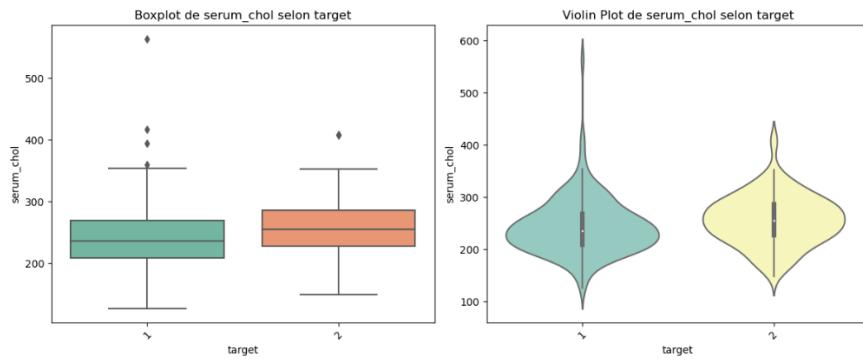
- Test t de Student : Statistique=3.6199, p-value=0.0004
- Test de Wilcoxon : Statistique=11366.0000, p-value=0.0002

◆ Analyse de rest_blood_pres en fonction de target (Cat 1 : 150, Cat 2 : 120)



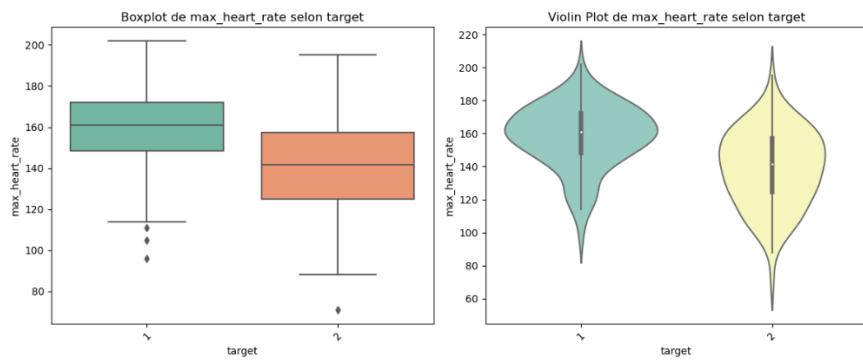
- Test t de Student : Statistique=2.5330, p-value=0.0120
- Test de Wilcoxon : Statistique=10367.5000, p-value=0.0315

◆ Analyse de serum_chol en fonction de target (Cat 1 : 150, Cat 2 : 120)



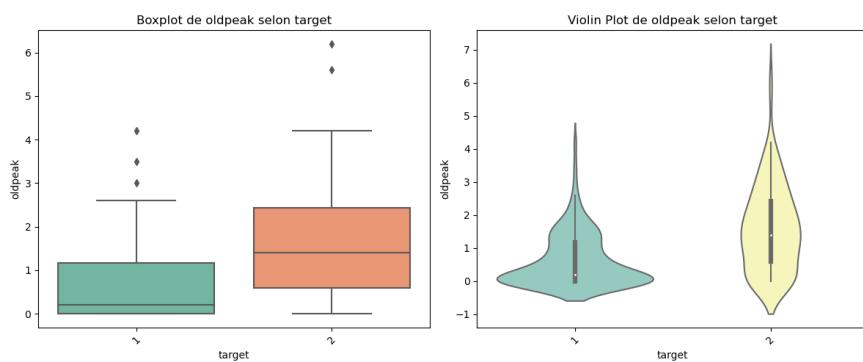
- Test t de Student : Statistique=1.9715, p-value=0.0497
 - Test de Wilcoxon : Statistique=10699.5000, p-value=0.0077
-

◆ Analyse de max_heart_rate en fonction de target (Cat 1 : 150, Cat 2 : 120)



- Test t de Student : Statistique=-7.3939, p-value=0.0000
 - Test de Wilcoxon : Statistique=4611.5000, p-value=0.0000
-

◆ Analyse de oldpeak en fonction de target (Cat 1 : 150, Cat 2 : 120)



- Test t de Student : Statistique=7.1719, p-value=0.0000
 - Test de Wilcoxon : Statistique=13170.5000, p-value=0.0000
-

Annexe 6 : Tableau de contingence, Heatmap des pourcentages de liaison, test Chi-2 et V de Cramer

◆ Analyse entre sex et target

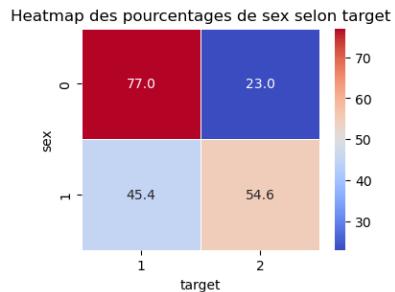


Table de contingence :

target 1 2

sex

	target 1	target 2
0	67	20
1	83	100

Test du Khi-2 : $\chi^2 = 22.6673$, p-value = 0.0000, avec V de Cramer : 0.28974609262433276

◆ Analyse entre chest_pain_type et target

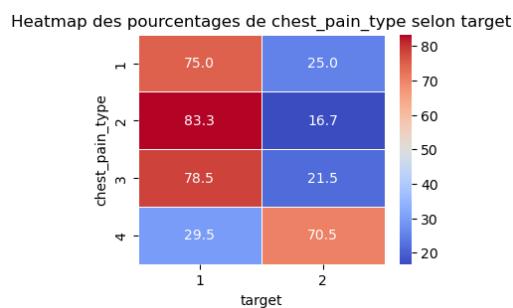


Table de contingence :

target 1 2

chest_pain_type

	target 1	target 2
1	15	5
2	35	7
3	62	17
4	38	91

Test du Khi-2 : $\chi^2 = 68.5882$, p-value = 0.0000, avec V de Cramer : 0.5040142800216132

◆ Analyse entre fast_blood_sugar et target

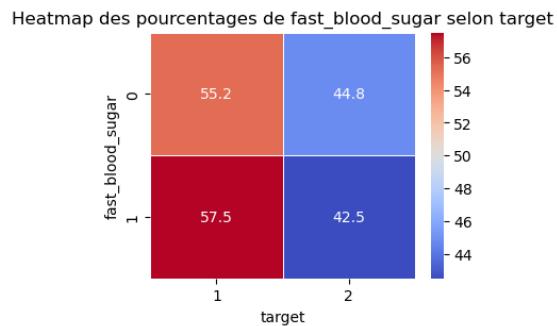


Table de contingence :

target 1 2

fast_blood_sugar

	target 1	target 2
0	127	103
1	23	17

Test du Khi-2 : $\chi^2 = 0.0092$, p-value = 0.9237, avec V de Cramer : 0.005828155051501955

◆ Analyse entre rest_ecg_res et target

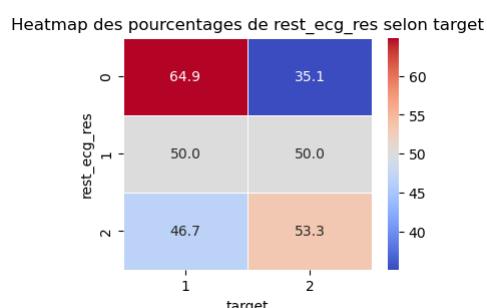


Table de contingence :

target 1 2

rest_ecg_res

	target 1	target 2
0	85	46
1	1	1
2	64	73

Test du Khi-2 : $\chi^2 = 8.9795$, p-value = 0.0112, avec V de Cramer : 0.18236564813733264

◆ Analyse entre ex_angina et target

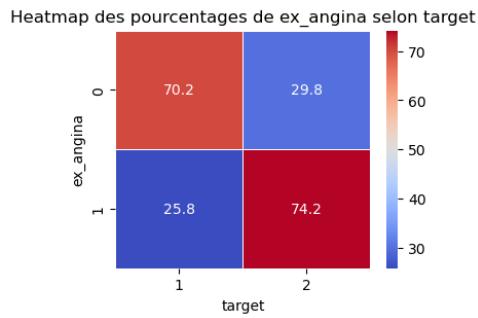


Table de contingence :

	1	2
ex_angina	127	54
0	23	66
1		

Test du Khi-2 : $\chi^2 = 45.6919$, p-value = 0.0000, avec V de Cramer : 0.41137471678963966

◆ Analyse entre slope_peak et target

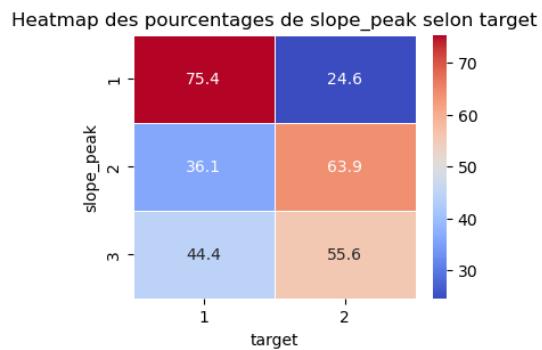


Table de contingence :

	1	2
slope_peak	98	32
1	44	78
2		
3	8	10

Test du Khi-2 : $\chi^2 = 40.3704$, p-value = 0.0000, avec V de Cramer : 0.38667811726606066

◆ Analyse entre nb_maj_ves et target

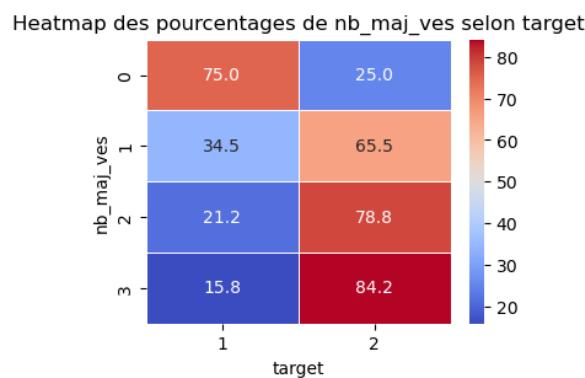


Table de contingence :

	1	2
nb_maj_ves	120	40
0	20	38
1	7	26
2	3	16
3		

Test du Khi-2 : $\chi^2 = 62.8631$, p-value = 0.0000, avec V de Cramer : 0.4825207418263913

◆ Analyse entre thal et target

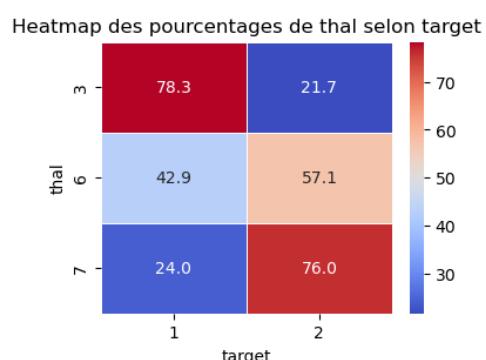


Table de contingence :

	1	2
thal	119	33
3	6	8
6	25	79
7		

Test du Khi-2 : $\chi^2 = 74.5693$, p-value = 0.0000, avec V de Cramer : 0.5255309359151155

Annexe 7 : Coefficients de la régression logistique (à droite : sans les variables source de multicolinéarité : 'rest_blood_pres', 'age', 'max_heart_rate', 'serum_chol')

Variable	Coefficient	Odds Ratio	Variable	Coefficient	Odds Ratio	Facteur d'inflation de la variance (VIF) :	Feature	VIF	Facteur d'inflation de la variance (VIF) :	Feature	VIF
0 sex_1	1.096643	2.994099	0 sex_1	0.848095	2.3305193	0 sex_1	4.091454	3.589662	0 sex_1	3.589662	
1 chest_pain_type_2	0.177905	1.194712	1 chest_pain_type_2	0.123560	1.131518	1 chest_pain_type_2	3.057885	1.320068	1 chest_pain_type_2	1.320068	
2 chest_pain_type_3	-0.341159	0.710946	2 chest_pain_type_3	-0.476434	0.620994	2 chest_pain_type_3	4.186571	1.681568	2 chest_pain_type_3	1.681568	
3 chest_pain_type_4	1.437701	4.211002	3 chest_pain_type_4	1.409077	4.092178	3 chest_pain_type_4	7.673652	3.656816	3 chest_pain_type_4	3.656816	
4 fast_blood_sugar_1	-0.310674	0.732953	4 fast_blood_sugar_1	-0.165463	0.847501	4 fast_blood_sugar_1	1.337603	1.249620	4 fast_blood_sugar_1	1.249620	
5 rest_ecg_res_1	0.136761	1.146554	5 rest_ecg_res_1	0.235997	1.266170	5 rest_ecg_res_1	2.371011	2.118524	5 rest_ecg_res_1	2.118524	
6 ex_angina_1	0.414442	1.513525	6 ex_angina_1	0.488166	1.629326	6 ex_angina_1	2.477852	2.397938	6 ex_angina_1	2.397938	
7 slope_peak_2	0.752226	2.121717	7 slope_peak_2	0.911869	2.488969	7 slope_peak_2	3.083124	2.938441	7 slope_peak_2	2.938441	
8 slope_peak_3	0.039964	1.040773	8 slope_peak_3	0.061995	1.063957	8 slope_peak_3	1.775617	1.747374	8 slope_peak_3	1.747374	
9 nb_maj_ves_1	1.124248	3.077902	9 nb_maj_ves_1	1.067243	2.907533	9 nb_maj_ves_1	1.684798	1.439373	9 nb_maj_ves_1	1.439373	
10 nb_maj_ves_2	0.331185	3.785527	10 nb_maj_ves_2	1.189237	3.284573	10 nb_maj_ves_2	1.638788	1.475937	10 nb_maj_ves_2	1.475937	
11 nb_maj_ves_3	0.568636	1.765957	11 nb_maj_ves_3	0.762210	2.143007	11 nb_maj_ves_3	1.361973	1.280867	11 nb_maj_ves_3	1.280867	
12 thal_6	0.060629	1.068258	12 thal_6	0.166036	1.180618	12 thal_6	1.279700	1.257842	12 thal_6	1.257842	
13 thal_7	1.226606	3.409639	13 thal_7	1.268486	3.555464	13 thal_7	2.682751	2.655867	13 thal_7	2.655867	
14 age	-0.006278	0.993741	14 age	45.918516	61.979401	14 age	45.918516	14 age	3.476550	14 age	3.476550
15 rest_blood_pres	0.021724	1.021962	15 rest_blood_pres	0.004476	1.004486	15 rest_blood_pres	61.979401	15 rest_blood_pres	37.694751	15 rest_blood_pres	37.694751
16 serum_chol	0.004476	1.004486	16 serum_chol	29.100700	29.100700	16 serum_chol	29.100700	16 serum_chol	1.00	16 serum_chol	1.00
17 max_heart_rate	-0.011800	0.988270	17 max_heart_rate	37.694751	37.694751	17 max_heart_rate	37.694751	17 max_heart_rate	1.00	17 max_heart_rate	1.00
18 oldpeak	0.551886	1.736525	18 oldpeak	0.636699	1.890231	18 oldpeak	3.671844	18 oldpeak	3.476550	18 oldpeak	3.476550

Annexe 8 : Indicateurs de performance du modèle de d'arbre de décision

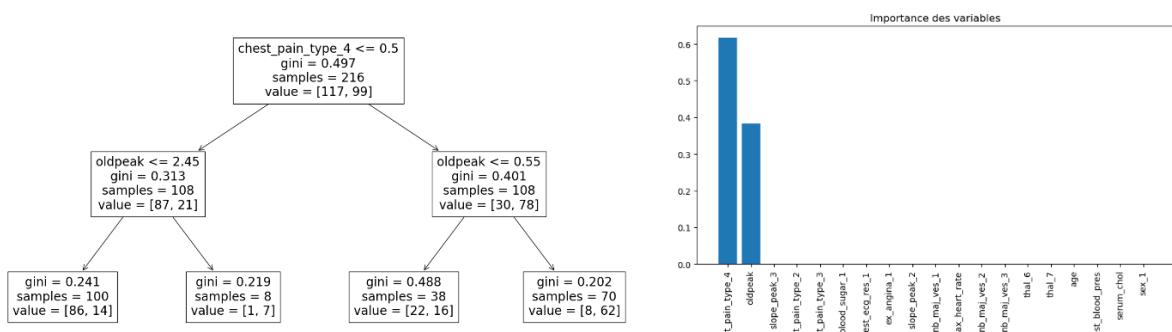
Sans optimisation

Seuil	Echantillon	Accuracy	Precision	Recall	F1	AUC
0.5	Train	1.00	1.00	1.00	1.00	1.00
	Test	0.72	0.65	0.62	0.63	1.00
0.45	Train	1.00	1.00	1.00	1.00	1.00
	Test	0.72	0.65	0.62	0.63	1.00
0.40	Train	1.00	1.00	1.00	1.00	1.00
	Test	0.72	0.65	0.62	0.63	1.00

Avec optimisation

Seuil	Echantillon	Accuracy	Precision	Recall	F1	AUC
0.5	Train	0.82	0.88	0.70	0.78	0.86
	Test	0.65	0.56	0.43	0.49	0.67
0.45	Train	0.82	0.88	0.70	0.78	0.86
	Test	0.65	0.56	0.43	0.49	0.67
0.40	Train	0.79	0.73	0.86	0.79	0.86
	Test	0.67	0.57	0.62	0.59	0.67

Annexe 9 : Arbre de décision du modèle optimisé et importance de variables



Annexe 10 : Indicateurs de performance des différents modèles étudiés

Régression logistique

Avec toutes les variables

Seuil	Echantill.	Accurac.	Preciso.	Recall	F1	AUC
0.5	Train	0.88	0.88	0.85	0.87	0.93
	Test	0.89	1.00	0.71	0.83	0.93
0.45	Train	0.87	0.84	0.87	0.86	0.93
	Test	0.89	0.94	0.76	0.84	0.93
0.40	Train	0.86	0.81	0.89	0.85	0.93
	Test	0.89	0.86	0.86	0.86	0.93

Sans les variables source de multicollinearité : 'rest_blood_pres', 'age', 'max_heart_rate', 'serum_chol'

Seuil	Echantill.	Accurac.	Preciso.	Recall	F1	AUC
0.5	Train	0.86	0.87	0.82	0.84	0.93
	Test	0.81	0.87	0.61	0.72	0.91
0.45	Train	0.86	0.84	0.86	0.85	0.93
	Test	0.81	0.82	0.67	0.74	0.91
0.40	Train	0.87	0.83	0.90	0.86	0.93
	Test	0.87	0.82	0.86	0.84	0.91

Ridge L2

Seuil	Echantill.	Accurac.	Preciso.	Recall	F1	AUC
0.5	Train	0.88	0.88	0.85	0.87	0.93
	Test	0.89	1.00	0.71	0.83	0.93
0.45	Train	0.87	0.84	0.87	0.86	0.93
	Test	0.89	0.94	0.76	0.84	0.93
0.40	Train	0.86	0.81	0.89	0.85	0.93
	Test	0.89	0.86	0.86	0.86	0.93

Lasso L1

Seuil	Echantill.	Accurac.	Preciso.	Recall	F1	AUC
0.5	Train	0.81	0.85	0.73	0.78	0.90
	Test	0.83	0.93	0.62	0.74	0.86
0.45	Train	0.82	0.83	0.78	0.80	0.90
	Test	0.76	0.72	0.62	0.67	0.86
0.40	Train	0.82	0.78	0.84	0.81	0.90
	Test	0.74	0.67	0.67	0.67	0.86

XgBoost

Seuil	Echantill.	Accurac.	Preciso.	Recall	F1	AUC
0.5	Train	0.83	0.87	0.74	0.80	0.90
	Test	0.83	0.93	0.62	0.74	0.86
0.45	Train	0.84	0.84	0.81	0.82	0.90
	Test	0.74	0.68	0.62	0.65	0.86
0.40	Train	0.83	0.79	0.85	0.82	0.90
	Test	0.74	0.67	0.67	0.67	0.86

KNN

KNN avec optimisation

Seuil	Echantill.	Accurac.	Preciso.	Recall	F1	AUC
0.5	Train	0.77	0.78	0.70	0.74	0.84
	Test	0.63	0.52	0.62	0.57	0.74
0.45	Train	0.77	0.78	0.70	0.74	0.84
	Test	0.63	0.52	0.62	0.57	0.74
0.40	Train	0.74	0.66	0.86	0.75	0.84
	Test	0.65	0.53	0.95	0.68	0.74

Arbre de décision

Arbre de décision sans optimisation

Seuil	Echantill.	Accurac.	Preciso.	Recall	F1	AUC
0.5	Train	1.00	1.00	1.00	1.00	1.00
	Test	0.72	0.65	0.62	0.63	1.00
0.45	Train	1.00	1.00	1.00	1.00	1.00
	Test	0.72	0.65	0.62	0.63	1.00
0.40	Train	1.00	1.00	1.00	1.00	1.00
	Test	0.72	0.65	0.62	0.63	1.00

Arbre de décision avec optimisation

Seuil	Echantill.	Accurac.	Preciso.	Recall	F1	AUC
0.5	Train	0.82	0.88	0.70	0.78	0.86
	Test	0.65	0.56	0.43	0.49	0.67
0.45	Train	0.82	0.88	0.70	0.78	0.86
	Test	0.65	0.56	0.43	0.49	0.67
0.40	Train	0.79	0.73	0.86	0.79	0.86
	Test	0.67	0.57	0.62	0.59	0.67

Random Forest

Random Forest sans optimisation

Seuil	Echantill.	Accurac.	Preciso.	Recall	F1	AUC
0.5	Train	1.00	1.00	1.00	1.00	1.00
	Test	0.78	0.80	0.57	0.67	0.87
0.45	Train	1.00	1.00	1.00	1.00	1.00
	Test	0.78	0.80	0.57	0.67	0.87
0.40	Train	1.00	1.00	1.00	1.00	1.00
	Test	0.74	0.67	0.67	0.67	0.87

Random Forest avec optimisation

Seuil	Echantill.	Accurac.	Preciso.	Recall	F1	AUC
0.5	Train	0.85	0.88	0.79	0.83	0.93
	Test	0.78	0.80	0.57	0.67	0.85
0.45	Train	0.84	0.82	0.85	0.83	0.93
	Test	0.76	0.70	0.67	0.68	0.85
0.40	Train	0.82	0.76	0.88	0.82	0.93
	Test	0.72	0.62	0.71	0.71	0.85