REPUBLIQUE DU SENEGAL

Un Peuple - Un But - Une Foi

****

MINISTÈRE DE L'ÉCONOMIE DU PLAN ET DE LA COOPÉRATION

******

AGENCE NATIONALE DE LA STATISTIQUE ET DE LA DÉMOGRAPHIE

****

ÉCOLE NATIONALE DE LA STATISTIQUE ET DE L'ANALYSE ÉCONOMIQUE

****

Projet Data Mining, ITS4

PRÉDICTION DE L'OCCURRENCE D'UNE MALADIE CARDIAQUE

Rédigé par :

Khariratou DIALLO, élève Ingénieur des Travaux Statistiques
Danis Rikel JIOGUE TAMATIO, élève Ingénieur des Travaux Statistiques

Supervisé par :

M. Abdoul Aziz NDIAYE
Data Scientist, Enseignant à l'ENSAE - Dakar

Janvier 2019

REMERCIEMENTS

Qu'il nous soit permis d'exprimer notre profonde gratitude à tous ceux qui ont participé à l'élaboration de la présente étude.

Nos remerciements vont à l'endroit de tout le personnel administratif de l'ENSAE, particulièrement au directeur de l'école, M. Abdou DIOUF, à notre responsable de filière Dr. Souleymane FOFANA, au responsable de la filière ISE, M. Idrissa DIAGNE, à Dr. Souleymane DIAKITE, ingénieur statisticien économiste et tout le corps professoral de l'ENSAE, pour leur dévotion à faire de l'excellence, la devise de l'école.

Nous tenons également à remercier l'ensemble des enseignants qui interviennent dans notre formation afin de faire de nous les as de la statistique. Un remerciement tout particulier à l'enseignant du cours de Data Mining, M. Abdoul Aziz NDIAYE qui en plus des cours est un très bon conseillé soucieux de notre réussite scolaire et professionnel.

Nous ne saurions terminer ces remerciements sans y associer les élèves de l'ENSAE, particulièrement nos camarades de classe de la promotion 2016-2017 auprès de qui nous avons beaucoup appris mais aussi tous ceux qui, de près ou de loin, nous ont soutenus avec leurs prières, leurs encouragements.

AVANT PROPOS

L'Ecole Nationale de la Statistique et de l'Analyse Economique (ENSAE) est une école à vocation sous régionale créée en 2008 et située à Dakar. Elle constitue une direction de l'Agence Nationale de la Statistique et de la Démographie (ANSD) et s'intègre dans un réseau coordonné par le Centre d'Appui aux Écoles de Statistique Africaines (CAPESA), avec l'Institut Sous régional de Statistique et d'Économie Appliquée (ISSEA-Yaoundé) et l'École Nationale Supérieure de Statistique et d'Économie Appliquée (ENSEA-Abidjan).

L'École propose des formations qui se déroulent dans trois cycles : les Techniciens Supérieurs de la Statistique (TSS), les Ingénieurs des Travaux Statistiques (ITS) et les Ingénieurs Statisticiens Economistes (ISE).

L'évolution fulgurante des TIC ces dernières années est fortement corrélée avec un accroissement massif de données. Ces données, sont soient structurées (peuvent directement utilisées dans un processus de prise de décision), soient non structurées (nécessitent un traitement préalable avant l'utilisation). Au vu de la place centrale du statisticien en matière de traitement de données, cet accroissement de données surtout non structurées nécessite la maitrise des outils de traitement plus adaptés. Ainsi, c'est dans ce contexte, que le cours de *Data Mining* ou *Fouille de données* a été introduit dans le cycle de formation des ITS. Ce dernier constitue une introduction pour ces élèves aux techniques d'apprentissage supervisées et non supervisées. Dans le cadre de la pratique de ce cours, il nous ait demandé de *Prédire l'occurence de la maladie cardiaque dans un échantillon de 303 patients.*

Table des matières

Liste des tableaux

Table des figures

RÉSUME

L'étude porte sur la prédiction de l'occurrence d'une maladie cardiaque sur un échantillon de 303 patients. Tout d'abord, l'analyse exploratoire décomposée en une analyse univariée et bivariée a permis une meilleure familiarisation avec les données. Il en est ressorti, grâce l'observation de la courbe approximatif de la densité de la loi normale et au résultat du test de *Jarque Berrra* que certaines variables suivent une loi normale . Cependant, l'utilisation du test graphique du boxplot et du test numérique de grubbs a montré qu'excepté la variable âge, toutes les autres variables quantitatives ont des valeurs aberrantes. Une correction, par la suite, de ces valeurs aberrantes a été effectuée par la méthode de l'imputation par la moyenne. L'analyse bivariée a également mis en exergue les liaisons qui existent entre les variables. Par ailleurs, nous avons procédé à une identification des difficultés de l'étude afin de bien poser le problème avant d'aborder la modélisation. Par ailleurs, le choix des variables explicatives pour la prédiction à travers les méthodes backward, forward et stepwise ont précédé l'utilisation des méthodes de modélisation. Différentes techniques de modélisation ont été abordées telles que la régression logistique, l'arbre de décision, le bagging, le Random Forest, le SVM et les réseaux de neurones. Une étude comparative de ces différents modèles à l'aide des critères tels que la courbe de ROC, l'AUC de ROC, le Kappa de cohen, le taux d'erreur de prédiction, la sensibilité et la spécificité a permis de classer la régression logistique comme étant le meilleur modèle. La spécification du modèle retenu de cette technique est la suivante :
$target = oldpeak + cp + ca + thal + exang + sex + chol + trestbps + slope$

INTRODUCTION

contexte de l'étude

Les maladies cardio-vasculaires sont la première cause de mortalité dans le monde : il meurt chaque année plus de personnes en raison de maladies cardio-vasculaires que de toute autre cause. En effet, selon l'Organisation Mondiale de la Santé (OMS), on estime à 17,7 millions le nombre de décès imputables aux maladies cardio-vasculaires, soit 31% de la mortalité mondiale totale. Parmi ces décès, on estime que 7,4 millions sont dus à une cardiopathie coronarienne et 6,7 millions à un AVC (chiffres 2015). En outre, plus des trois quarts des décès liés aux maladies cardiovasculaires interviennent dans des pays à revenu faible ou intermédiaire, toujours selon l'OMS.
Les principaux facteurs de risque de maladie cardio-neurovasculaire sont liés au mode de vie : tabagisme, alimentation déséquilibrée, manque d'activité physique, usage nocif de l'alcool, facteurs psychosociaux tels que le stress. Par ailleurs, d'après l'OMS, les cardiopathies et les accidents vasculaires cérébraux (AVC) pourraient être évités si on adoptait une alimentation saine, si on pratiquait régulièrement une activité physique et si on évitait l'exposition à la fumée de tabac. C'est ce qui nous pousse à nous poser des questions sur les mesures de prédiction des maladies cardiaques. Ainsi, l'on se demande :

1. Quels sont les facteurs explicatifs des maladies cardiaques ?
2. Comment pourrait-on prédire les maladies cardiaques ?

L'objectif fixé à travers cette étude est de prédire l'occurrence d'une maladie cardiaque à partir des caractéristiques des patients. Les objectifs spécifiques sont :

1. Décrire les caractéristiques des patients ;
2. Identifier les facteurs explicatifs de l'occurrence de la maladie cardiaque dans notre échantillon de patient.

De ce fait, afin de répondre à ces questions et d'atteindre nos objectifs, nous allons adopter le plan suivant. Tout d'abord, une analyse préliminaire sur les données sera faite afin de mieux connaitre nos données. Ensuite, . Ensuite nous identifier les principales difficultés de l'étude et la sélection des variables, nous passerons la présentation des différentes techniques de prédiction, Enfin nous procèderons à une étude évaluation des différents modèles suivant les indicateurs de performance retenus.

Chapitre 1

Analyse exploratoire des données

La base soumise dans le cadre de cette étude est composée de **14 variables** et **303 observations**. L'intérêt de cette partie est de nous permettre de mieux nous familiariser à notre base de données. Ainsi, elle a pour objectif de donner une vision globale du jeu de données et découvrir les formes de régularités. Dans la suite, il sera abordé tout d'abord l'analyse univariée, ensuite nous passerons à l'analyse bivariée. Il est à noter que dans ces parties, les tests nécessaires seront également présentés.

1.1 Analyse Univariée

Elle consiste à faire une analyse variable par variable. Cette analyse sera numérique et graphique. Pour ce qui est de l'analyse graphique, nous ferons un histogramme des variables avec la densité d'une loi normale et un boxplot. En ce qui concerne l'analyse numérique, nous utiliserons un test non paramétrique[1] pour tester la normalité de la variable. Le test à utiliser est celui de ***Jarque Bera***[2].

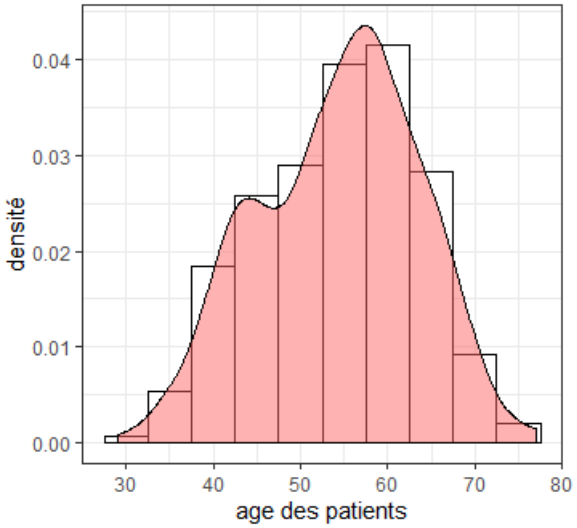
1.1.1 Variables quantitatives

✍ **Age**

L'âge des patients est compris entre 29 ans et 77 ans. La moyenne d'âge dans cet échantillon est de 54,37 ans.

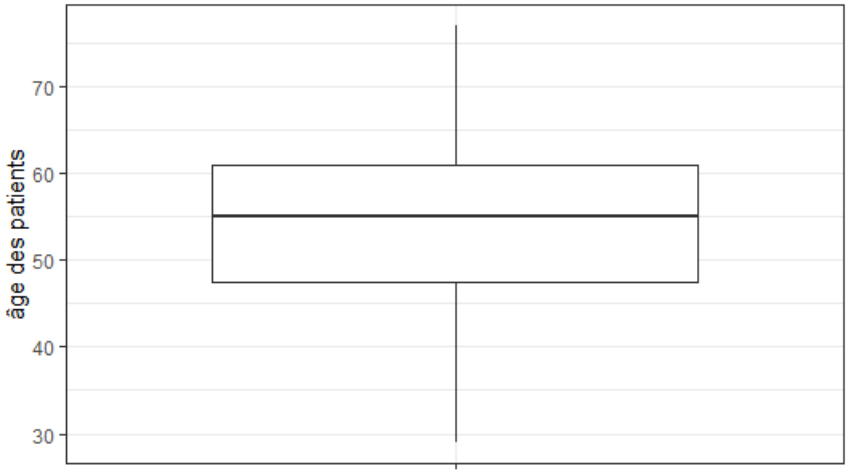
1. L'utilisation du test non paramétrique se justifie par le fait que ce dernier ne repose sur une distribution de loi particulière.
2. Voir description en annexe 1

FIGURE 1 – Histogramme de l'âge des patients



L'observation de l'histogramme permet de constater que dans notre échantillon, les patients âgés entre 56 ans et 60 ans sont les plus nombreux.
Aussi, cette figure nous incite à conclure que la série des âges des patients de notre échantillon n'est pas distribuée suivant une loi normale. Cependant, on s'aidera du test numérique de **Jarque Bera** pour vérifier cette affirmation.
L'utilisation du **test de Jarque Bera** donnne une $p-value = 0.052$. Ceci conduit au rejet de H_0. Autrement dit, le test permet de confirmer la conclusion de l'observation graphique à savoir la non normalité de la série des âges des patients.

FIGURE 2 – Boxplot de l'âge des patients



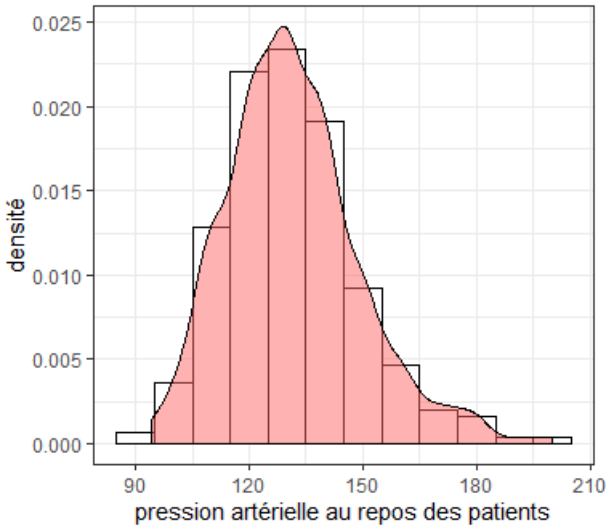
Source : calcul auteurs

L'observation du boxplot permet de constater qu'il n'y a pas de patient ayant un âge aberrant.

✍ **Pression artérielle au repos**

Dans notre échantillon de patients, la pression artérielle moyenne au repos est de 131,6. La plus petite pression observée est de 94 et la plus grande s'élève à 200.

FIGURE 3 – Boxplot de la pression artérielle des patients



Source : calcul auteurs

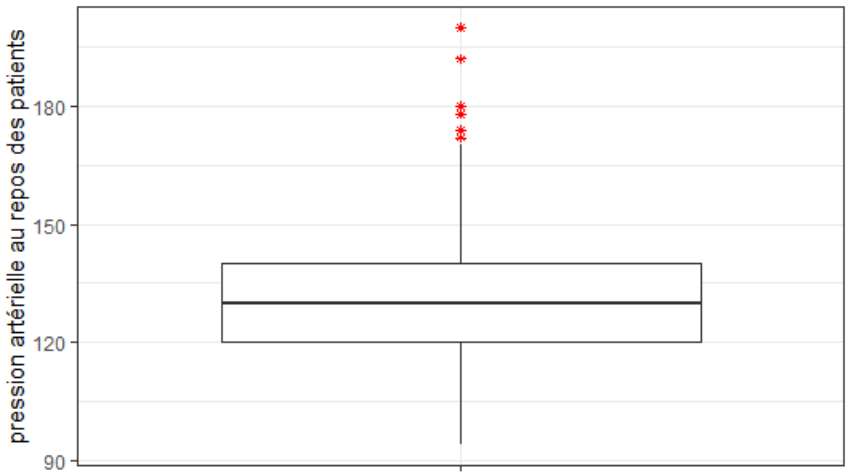
L'histogramme permet de constater que la plupart de notre échantillon ont une pression artérielle au repos comprise entre 120 et 140.
La courbe de densité de la loi normale permet de conclure à la non normalité de la série car on n'a pas une bonne symétrie de la variable. Toutefois, cette conclusion n'est que hâtive, car l'observation graphique n'est pas robuste.
L'utilisation du **test de Jarque Bera** permet d'avoir une $p - value = 1.893e - 08$. Ainsi, le test conduit à accepter H_0, c'est-à-dire accepter la normalité de la distribution de la pression. Ceci nous permet de constater qu'avec une erreur de 5%, la distribution de la pression artérielle est normalement distribuée. L'observation du BoxPlot permet de rendre compte de l'existence de outliers dans la série d'observations de la pression artérielle. Ils sont au nombre de 6 et sont représentés dans le tableau suivant :

TABLE 1 – Valeurs aberrantes de la pression artérielle normale (BoxPlot)

Observations	172	174	178	180	192	200
fréquence	1	1	2	3	1	1

Compte tenu du fait que les outliers tels que présentés par le graphique du BoxPlot sont les observations n'appartenant pas à l'intervalle : $I = [Q_1 - 1.5 * mediane; Q_3 - 1.5 * mediane[$.
Aussi, le BoxPlot est réputé être très sensible aux valeurs aberrantes. Fort de ce constat et prenant en compte la dsitribution normale de la série de pression artérielle, nous allons procéder à un test statistique pour définir clairement les valeurs aberrantes avec

FIGURE 4 – Boxplot de la pression artérielle des patients



Source : calcul auteurs

une erreur bien définie. Le test numérique de détection d'ouliers utilisé dans le document est celui de ***Grubbs***[3]. Ce dernier repose sur la normalité de la série, et nous sommes bien dans cette situation.

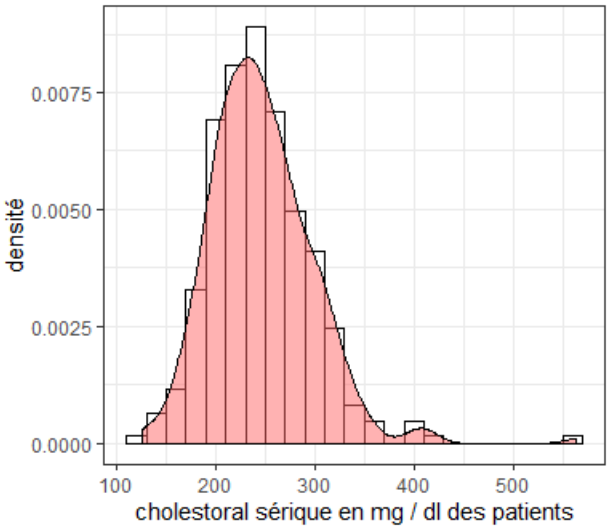
Le rendu du test de grubbs permet de constater que la série présente une seule valeur aberrante qui est sa valeur maximale. Ainsi, on peut affirmer avec un risque de 5% que la valeur aberrante de la série selon le test de grubbs est de **200**.

✍ **cholestérol sérique en mg / dl**

Dans la série des observations du cholestérol sérique de notre échantillon de patient, la valeur minimale est de 126 mg/dl, la valeur maximale est de 564 mg/dl. La moyenne de la série est de 246.3 mg/dl.

3. Voir annexe 2

FIGURE 5 – histogramme du cholestérol sérique en mg/dl des patients



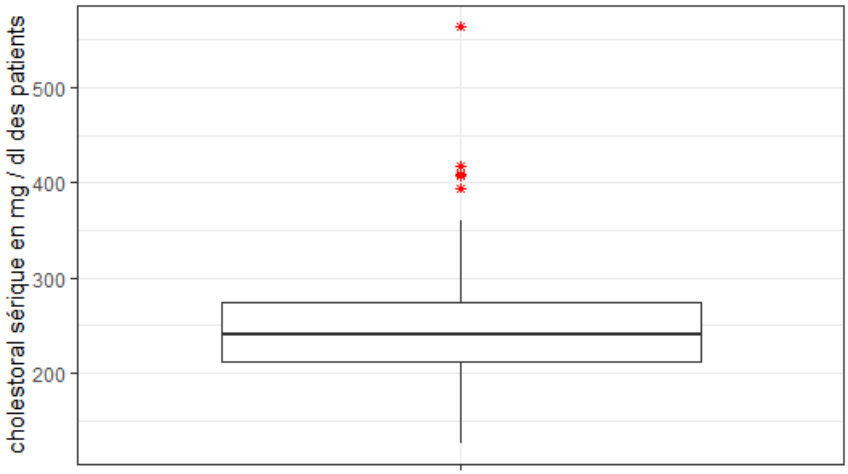
Source : calcul auteurs

L'observation de l'histogramme permet de soupçonner l'existence de certaines valeurs aberrantes qui semblent particulièrement se démarquer des autres.
Aussi, la forme de représentation de la densité de la loi normale par rapport à nos données permet de soupçonner que notre série suit une loi normale. L'utilisation du **test de Jarque Bera** permet d'obtenir une $p - value = 2.2e - 16$. Donc, le test conduit à l'acceptation de H_0. Ainsi, comme pressenti avec l'observation graphique, nous concluons que notre série du cholestérol sérique est distribuée suivant une loi normale avec un risque de se tromper de 5%.
Dans la suite, l'utilisation d'un test graphique (BoxPlot) va permettre de confirmer ou d'infirmer ce soupçon d'existence de valeurs aberrantes tel que supposé dans l'analyse de l'histogramme.
Le BoxPlot permet de constater la présence des valeurs aberrantes suivantes :

TABLE 2 – Valeurs aberrantes du cholestérol sérique en mg/dl des patients

Observations	394	407	409	417	564
fréquence	1	1	1	1	1

FIGURE 6 – Boxplot du cholestérol sérique en mg/dl des patients



Source : calcul auteurs

Compte tenu du fait que le **test de Jarque Bera** permet de conclure à la normalité de la série, nous utilisons **le test de Grubbs** pour détecter de façon numérique les outliers de la série. ***Le test de Grubbs*** considère que cette série ne présente qu'une seule valeur aberrante qui est sa valeur maximale : **564**.

✍ **fréquence cardiaque maximale atteinte**

La série de fréquence cardiaque maximale atteinte de notre échantillon de patients a une moyenne de 149.6.

TABLE 3 – Résumé de la fréquence maximale cardiaque atteinte

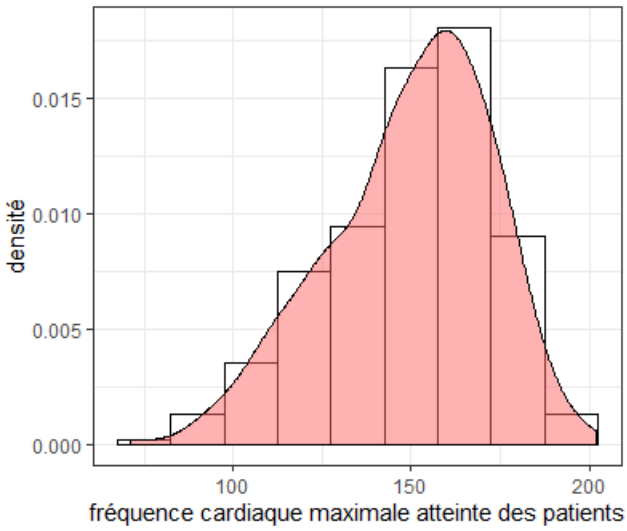
Min	Q1	Me	Mean	Q3	Max
71	133.5	153	149.6	166	202

L'observation de l'histogramme permet de constater une concentration de la fréquence cardiaque maximale de patients entre 150 et 160.

La représentation de la densité de la loi normale, permet de soupçonner que cette série est distribuée suivant une loi normale.

L'utilisation du **test de Jarque Bera** renvoie une $p-value = 0.0007021 < 5\%$, donc on accepte H_0. Ainsi, selon le **test de Jarque Bera**, on consent à la normalité des données de fréquence cardiaque avec une erreur de 5%.

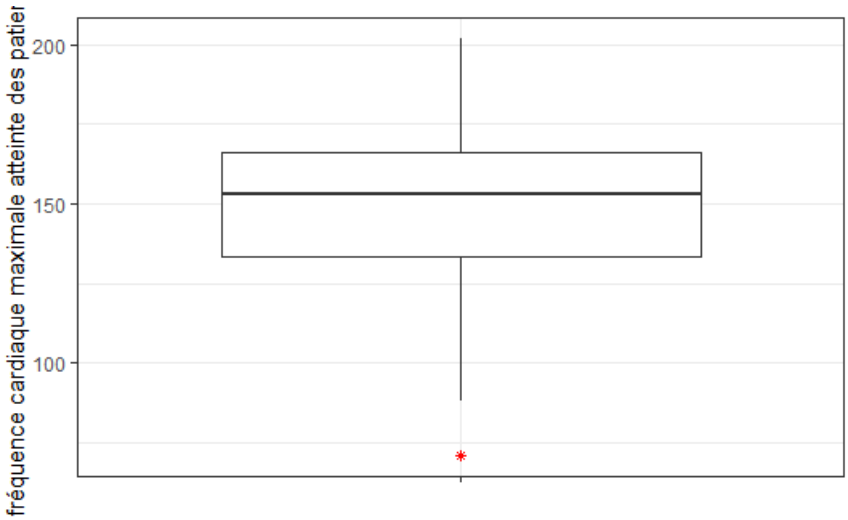
FIGURE 7 – histogramme de la fréquence cardiaque maximale des patients



Source : calcul auteurs

Le BoxPlot de la série est donné par le graphique suivant :

FIGURE 8 – BoxPlot de la fréquence cardiaque maximale des patients



Source : calcul auteurs

Le Boxplot permet de constater que cette série ne présente qu'une seule valeur aberrante qui est sa valeur minimale.
Le test de **Grubbs** vient confirmer que cette valeur minimale est bien un outlier car $p - value = 0.08031 > 5\%$ ce qui conduit au rejet de H_0.

✍ **Dépression ST induite par l'exercice par rapport au repos**

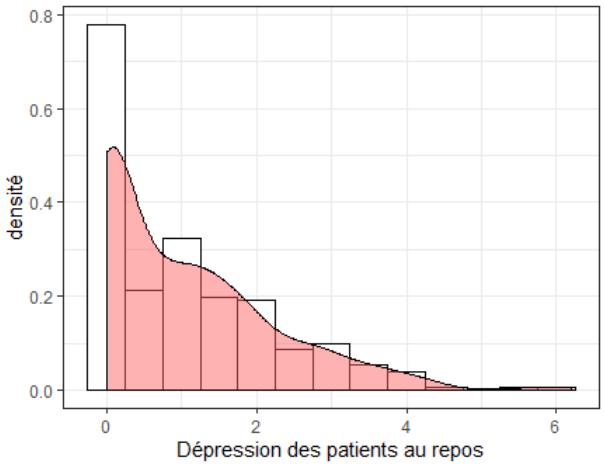
La série d'observations des patients sur la dépression ST induite par l'exercice par rapport au repos présente une moyenne égale 1.4. Le résumé exhaustif de ladite série est représenté dans le tableau ci-joint :

TABLE 4 – Résumé de la dépression ST induite par l'exercice par rapport au repos

Min	Q1	Me	Mean	Q3	Max
0	0	0.8	1.4	1.6	6.2

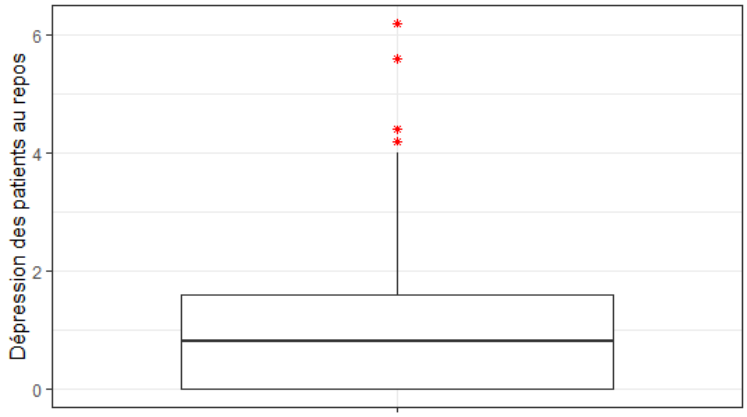
L'histogramme de la série est donnée par le graphique suivant :

FIGURE 9 – histogramme de la ST patients au repos



L'observation de l'histogramme permet de constater une concentration des patients ayant une dépression ST au repos autour de 0. Cependant, on note également la présence d'un individu qui se démarque présentant une valeur de 6.2.
La représentation de la densité de la normale sur notre graphe permet de soupçonner que notre série n'est pas distribuée suivant une loi normale.
L'utilisation du **test de Jarque Bera** permet d'avoir : $p - value = 1.893e - 08 < 5\%$. Donc on accepte H_0. Ainsi, selon le test numérique de **Jarque Bera**, notre série est normalement distribuée avec un risque de 5% de se tromper.

FIGURE 10 – BoxPlot de la ST patients au repos



Source : calcul auteurs

L'observation du BoxPlot permet de constater que les valeurs aberrantes sont au nombre de 5. Elles sont représentées dans le tableau suivant :

L'utilisation du **test de Grubbs** permet de conclure que cette série présente 2 valeurs

TABLE 5 – Valeurs aberrante issue du boxplot du de la dépression ST des patients des patients

Observations	6.2	5.6	4.4	4.2
fréquence	1	1	1	2

aberrantes à savoir **5.6 et 6.2**.

1.1.2 Variables qualitatives

✍ **Sexe**

Notre échantillon est majoritairement représenté par les hommes. En effet, ces derniers représentent 68,32% de l'échantillon.

TABLE 6 – Répartition de l'échantillon suivant le sexe

Femme	96
Homme	207

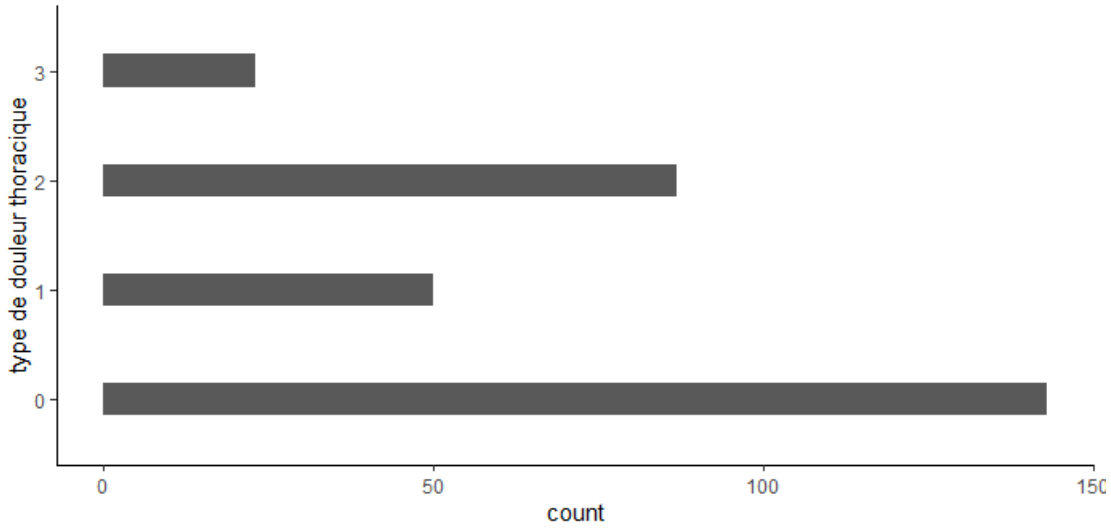
✍ **type de douleur thoracique**

La série de type de douleur thoracique est repartie suivant 4 modalités. Celle la plus représentée est : la douleur thoracique de type **angine de poitrine**, soit 47,2% de

TABLE 7 – Répartition des types de douleurs thoraciques

0	143
1	50
2	87
3	23

FIGURE 11 – BarPlot du type de douleur thoracique des patients

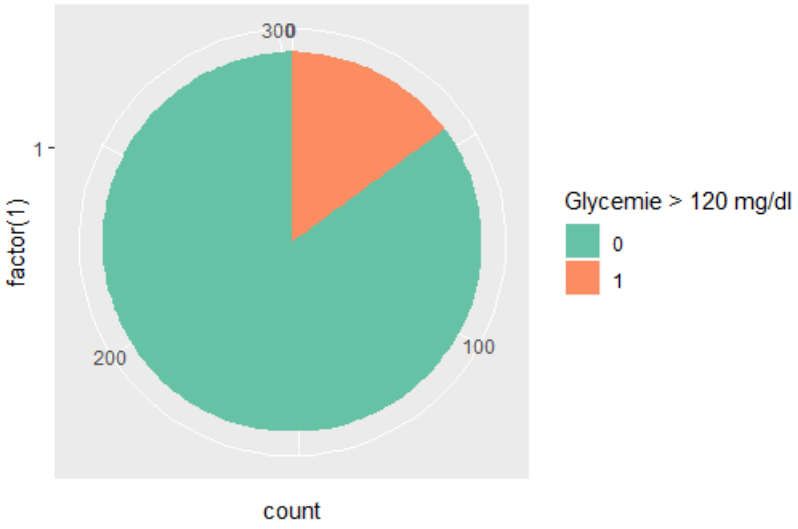


Source : calcul auteurs

notre échantillon de patients. Elle est suivie de douleurs non angulaires qui représentent 28.7%.

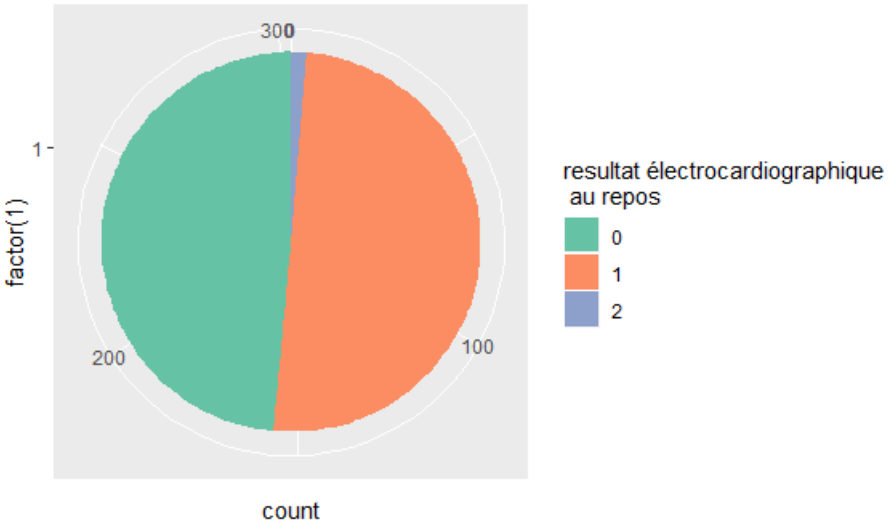
✍ **glycémie à jeun> 120 mg / dl** Dans notre échantillon de patients, la quasi-totalité possède une glycémie à jeun > 120 mg/dl. Cette proportion est de : 85.15%.

FIGURE 12 – Diagramme en Camembert des patients ayant un taux de glycémie



✍ **résultats électrocardiographiques au repos**
Les résultats électrocardiographiques des patients de l'échantillon sont regroupés en trois modalités. Les patients présentant les résultats de la modalité 0 et 1 sont les plus représentés. Ils occupent respectivement les proportions 48.51% et 50.17% de l'échantillon total.

FIGURE 13 – Diagramme en Camembert des patients ayant un taux de glycémie

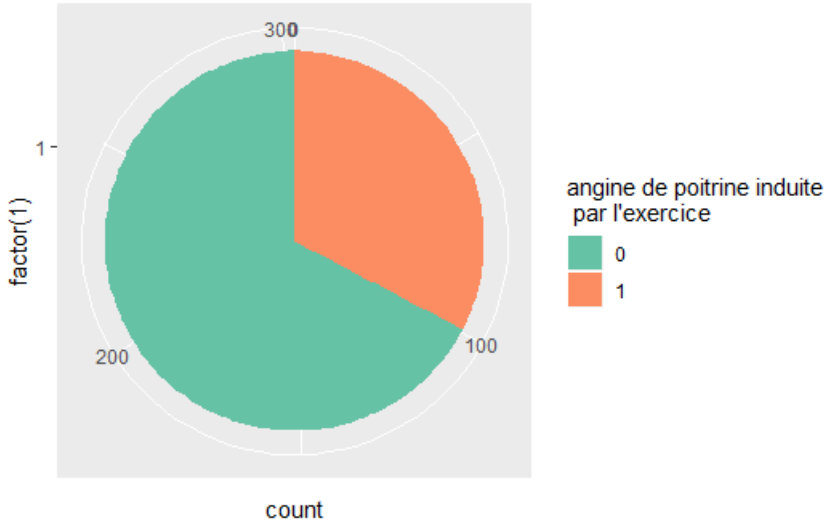


Source : calcul auteurs

✍ **angine de poitrine induite par l'exercice** 67.33% de notre échantillon présentent l'angine de poitrine induite par l'exercice. Il est important de remarquer que l'analyse de **la variable type de douleur** a permis de constater que l'angine de poitrine est

le type de douleur le plus observé dans l'échantillon.

FIGURE 14 – Diagramme en Camembert des patients ayant un taux de glycémie

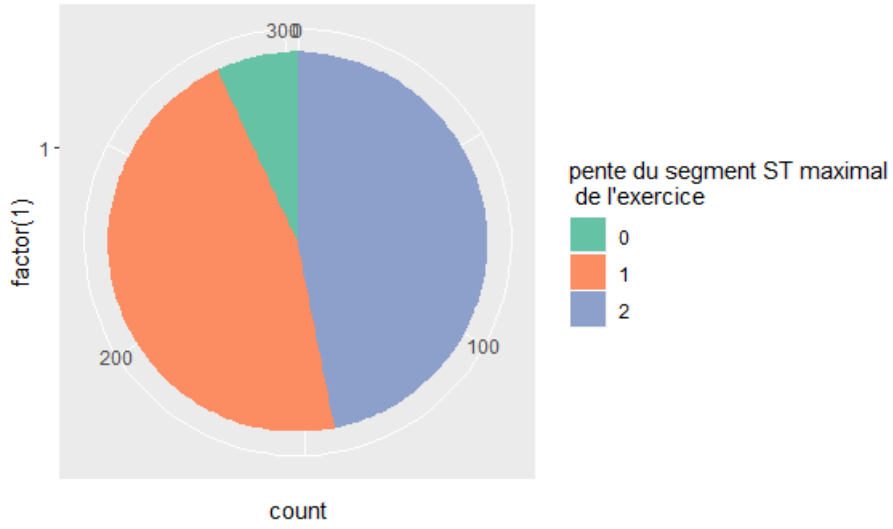


Source : calcul auteurs

✍ **pente du segment ST maximal de l'exercice**

Cette Variable présente 3 modalités dont la moins représentée est la modalité 0. Pour cette dernière, seulement 6.93% de la population la possède.

FIGURE 15 – Diagramme en Camembert des patients ayant un taux de glycémie



Source : calcul auteurs

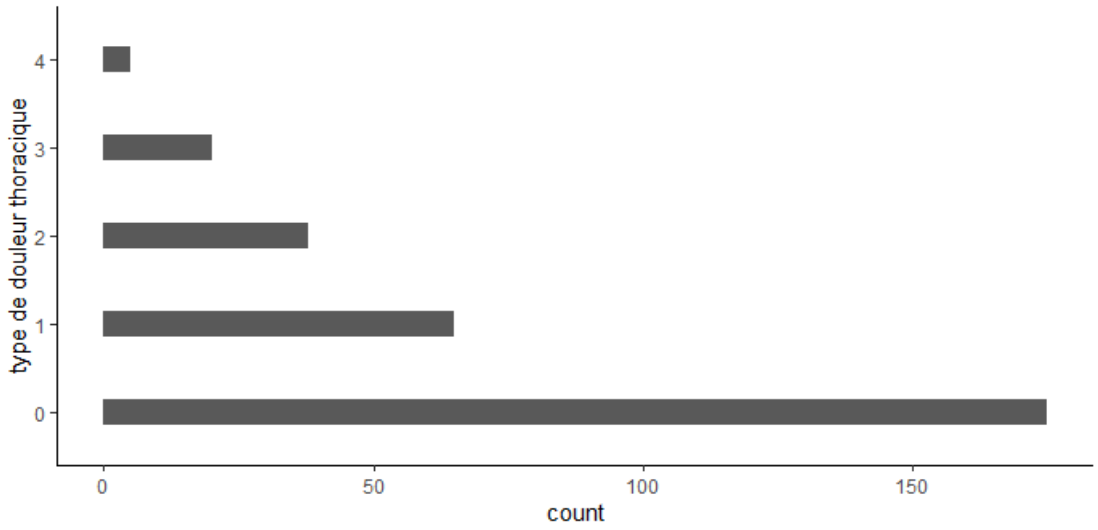
✍ **nombre de vaisseaux principaux (0-3) colorés par une fluoroscopie**

Cette variable présente 04 modalités dont la plus représentée est la modalité 0. Cette

dernière a une proportion de 57.76% de l'échantillon des patients.

Le barPlot de la distribution de notre variable :

FIGURE 16 – Diagramme en bar du nombre de vaisseaux principaux (0-3) colorés par une fluoroscopie des patients



Source : calcul auteurs

✍ **thal**

Compte tenu des modalités de cette variable, elle peut être considérée comme étant une variable caractérisant l'état du patient.

La distribution de ses modalités est ainsi représentée :

TABLE 8 – Répartition de la variable Thal

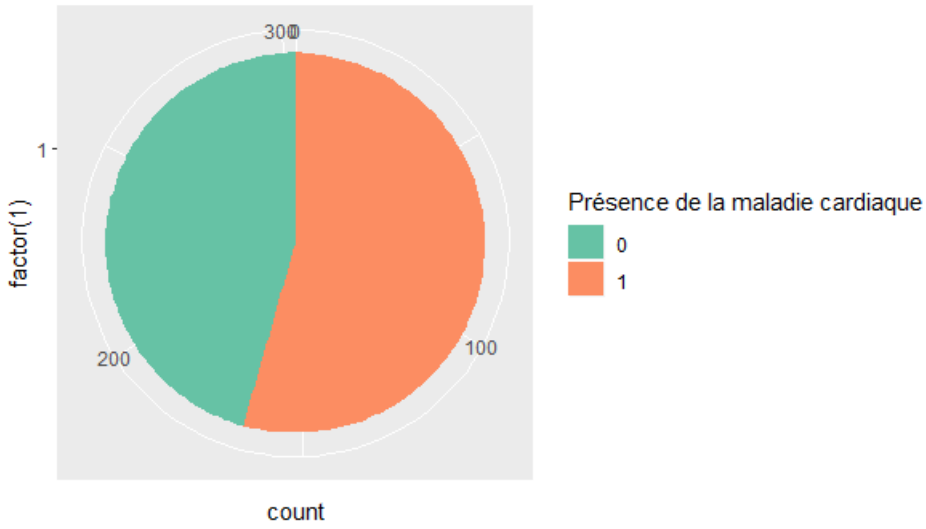
0	2
1	18
2	166
3	117

✍ **présence d'une maladie cardiaque**

Parmi les patients, 54.46% présentent une maladie cardiaque. Ce pourcentage nous permet de conclure que notre échantillon n'est pas déséquilibré, c'est-à-dire que l'on tend vers une égalité entre le nombre de patients qui présente une maladie et ceux qui sont sains. Cette information est très importante car elle nous servira dans la suite lors de la modélisation.

Une fois l'analyse univariée terminée, il est important de retenir que nous avons certaines variables qui présentent des outliers. Comme méthode de détection de ces outliers,

FIGURE 17 – Diagramme circulaire de la présence d'une maladie cardiaque dans la population



nous avons priorisé la méthode de détection par l'utilisation du test de grubbs car elle permet d'avoir un intervalle de confiance et aussi une erreur clairement définie de se tromper, ceci pour des variables normalement distribuées. En amont avec l'analyse bi-variée à venir, nous allons procéder à la correction de ces outliers en prenant en compte à la fois l'analyse uni-variée et bi-variée.

1.2 Analyse bi-variée

Cette section va nous permettre d'identifier les différentes interrelations existant entre nos variables. A ce niveau, nous allons principalement mettre le focus sur la liaison entre la variable " Présence d'une maladie cardiaque" et les autres variables. Dans la suite, l'analyse des interrelations se fera en trois parties suivant la nature des variables en présence.

1.2.1 Variable quantitative * Variable quantitative

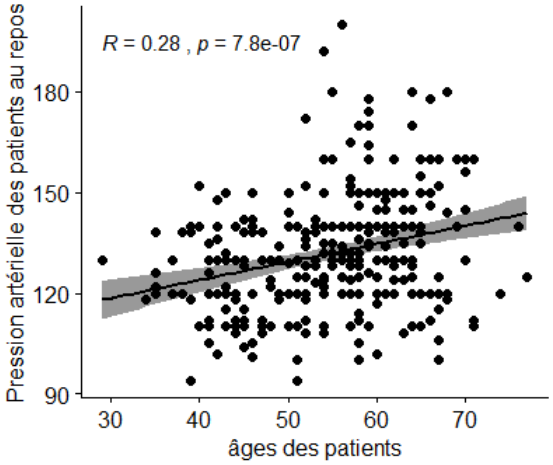
Pour ces variables, nous allons faire une représentation du nuage de points du couple de variables ensuite calculer le coefficient de corrélation de pearson [4] et enfin tester la significativité de ce dernier.

✍ **âge * pression artérielle au repos**

La représentation graphique de ces variables est donnée dans le graphe suivant :

4. voir en annexe

FIGURE 18 – scatter plot âge * pression artérielle au repos

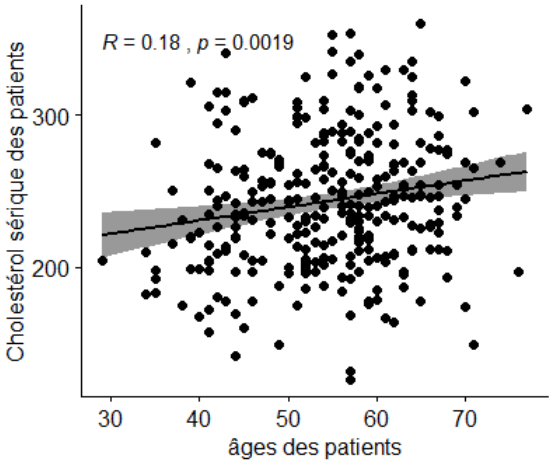


L'observation du graphe permet de noter une faible association entre l'âge des patients et leur pression artérielle au repos. Egalement, le coefficient de corrélation de pearson permet de conclure que ces deux variables sont très faiblement corrélées.

✍ **âge * cholestoral sérique en mg / dl**

Le graphe ci-contre permet de conclure que l'âge des patients et leur cholestoral sérique en mg / dl sont faiblement liés. le coefficient de corrélation de pearson de ces variables s'élève à 0.21 et est significatif. Donc ces deux variables sont faiblement corrélées.

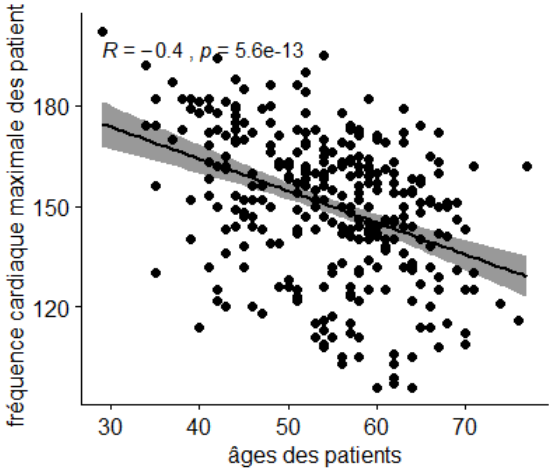
FIGURE 19 – scatter plot âge * cholestoral sérique en mg / dl



✍ **âge * fréquence cardiaque maximale des patients atteints**

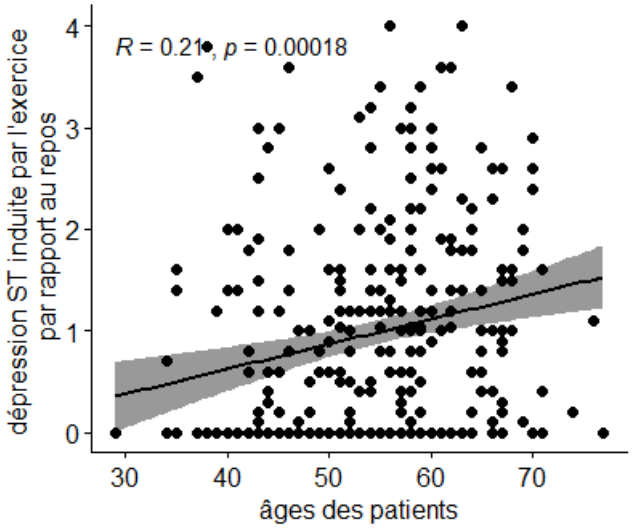
L'observation du graphe permet de constater une concentration des patients ayant un âge compris entre 40 ans et 60 ans et le ceux ayant une fréquence cardiaque maximale comprise entre 140 et 170. Le signe négatif du coefficient de corrélation de pearson permet de constater ces deux variables varient en sens inverses. Sa faible valeur permet de retenir que très peu de patients vérifient ces critères.

FIGURE 20 – scatter plot âge * fréquence cardiaque maximale des patients atteint



✍ **âge * dépression ST induite par l'exercice par rapport au repos**
L'observation du graphe ne permet pas de se douter de l'existence d'une relation linéaire entre les deux modèles. Ceci est également justifié par 'utilisation du coefficient de corrélation.

FIGURE 21 – scatter plot âge * dépression ST induite par l'exercice par rapport au repos

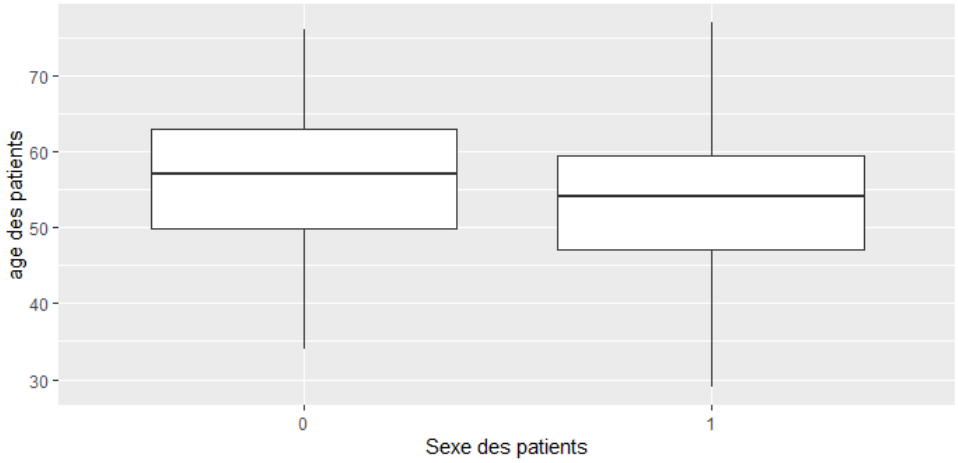


1.2.2 Variable quantitative * Variable qualitative

L'intérêt de cette partie est de voir si les valeurs de la variable quantitative se répartissent différemment selon la catégorie d'appartenance de la variable qualitative. Aussi, nous allons nous intéresser à l'homogénéité interne dans chaque groupe de variable formé par les modalités de la variable qualitative. Comme outil statistique, nous utiliserons le BoxPlot.

✍ **âge * sexe** L'observation du boxPlot permet de constater que dans le groupe des hommes et celui des femmes, il n'y a pas de valeurs aberrantes. Cependant, l'âge minimal des hommes est inférieur à celui des femmes et l'âge médian des femmes est aussi plus élevé.

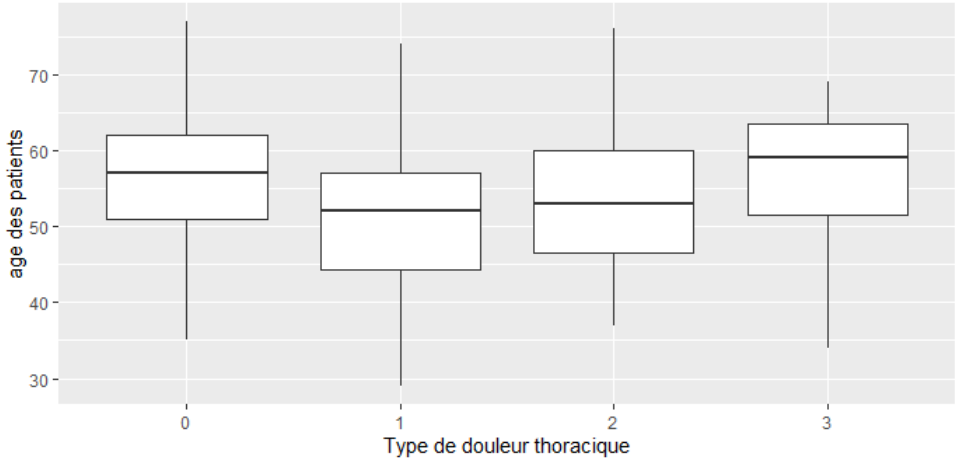
FIGURE 22 – BoxPlot Âge * sexe



✍ **Âge * type de douleur thoracique**

L'observation du BoxPlot permet de constater que dans chaque type de douleur thoracique, il n'y a pas de patients présentant un âge aberrant, cela témoigne de l'homogénéité interne de chacun de ces groupes. Aussi, l'individu le moins âgé se trouve dans le groupe de patients ayant une douleur thoracique non atypique (modalité 1).

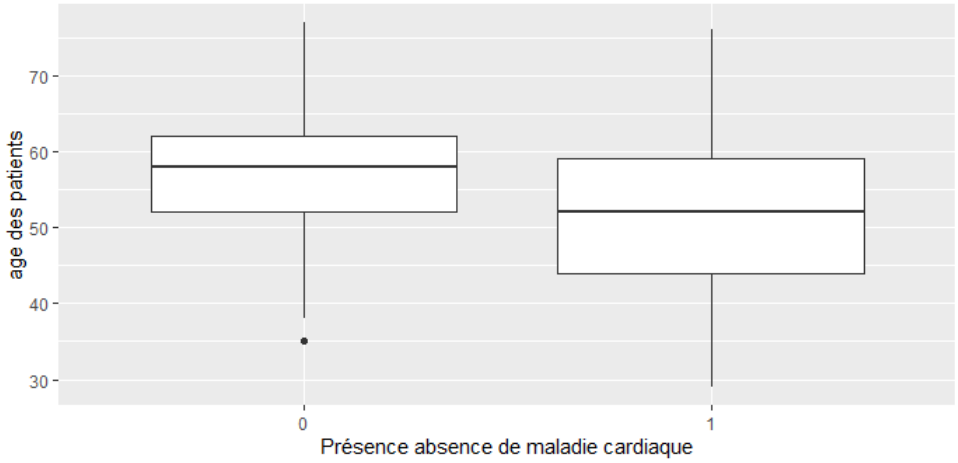
FIGURE 23 – BoxPlot Âge * type de douleur thoracique



✍ **présence ou absence de maladie cardiaque * âge**

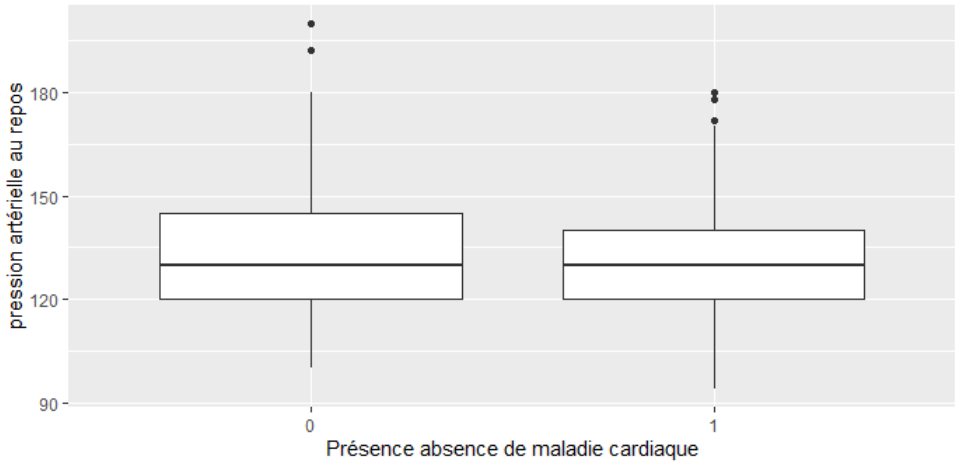
L'observation du BoxPlot permet de constater une forte présence des patients ayant une faible âge et atteint de la maladie cardiaque. Par ailleurs, dans le groupe des patients n'ayant pas la maladie cardiaque, on constate la présence d'une valeur aberrante qui est l'individu le moins âgé de ce groupe.

FIGURE 24 – BoxPlot Âge * présence ou absence de maladie cardiaque



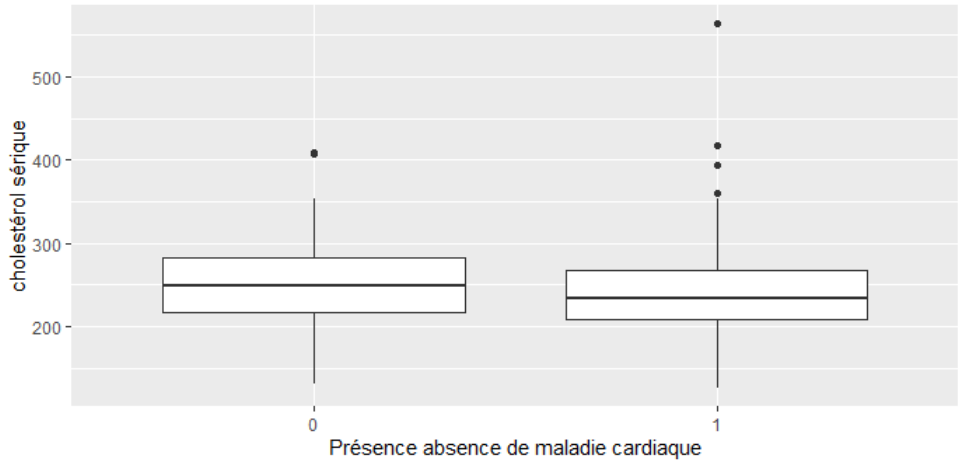
✍ **présence ou absence de maladie cardiaque * pression artérielle au repos**
L'analyse du BoxPlot de ces 2 variables permet de constater la présence de valeurs aberrantes dans chaque de groupe de patient. Les patients ayant une plus forte pression artérielle au repos sont ceux n'ayant pas la maladie cardiaque.

FIGURE 25 – BoxPlot pression artérielle au repos * présence ou absence de maladie cardiaque



✍ **présence ou absence de maladie cardiaque * cholestoral sérique en mg / dl**
Le BoxPlot permet de constater la présence d'outliers dans chaque groupe de patients. Dans le groupe de patients malades, on note une forte présence de faibles taux de cholestérol sérique. Ceci est aussi observé dans le groupe des non malades. Les valeurs considérées comme outliers dans chaque groupe sont aussi ceux observées lors de l'étude univariée de la variable.

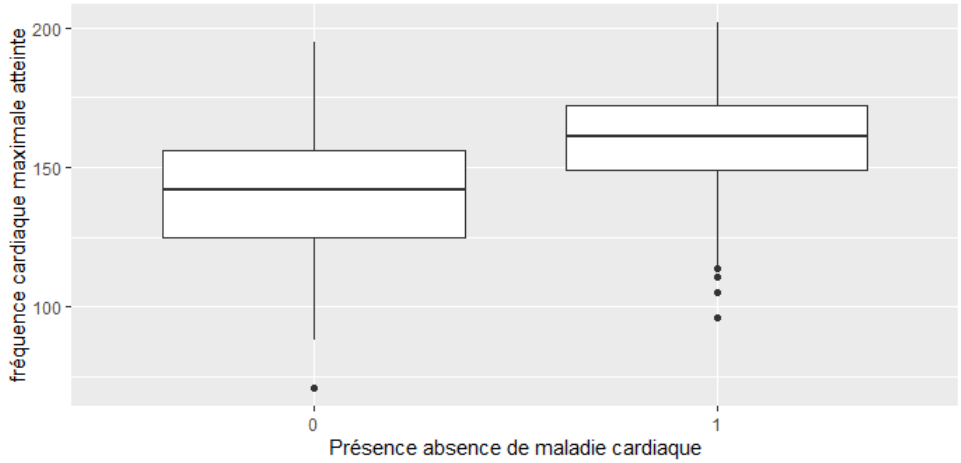
FIGURE 26 – BoxPlot cholestérol sérique * présence ou absence de maladie cardiaque



✍ **présence ou absence de maladie cardiaque * fréquence cardiaque maximale atteinte**

A l'observation du BoxPlot, on note une présence de valeurs aberrantes dans chaque groupe de patients. Cependant, la fréquence cardiaque maximale atteinte est plus élevée dans le groupe de patients atteints par la maladie cardiaque.

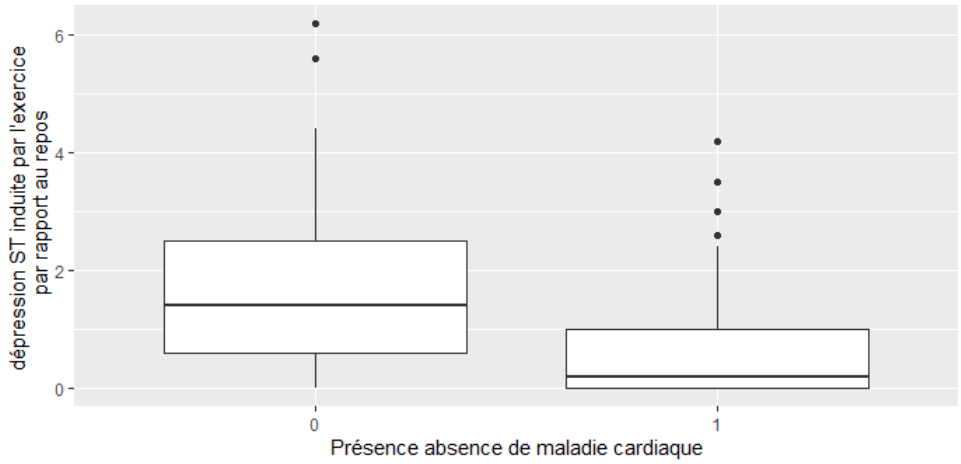
FIGURE 27 – BoxPlot présence ou absence de maladie cardiaque * fréquence cardiaque maximale atteinte



✍ **présence ou absence de maladie cardiaque * dépression ST induite par l'exercice par rapport au repos**

Le boxplot permet de noter la présence de valeurs aberrantes dans chaque groupe. Aussi, chez les malades, la dépression ST induite par l'exercice par rapport au repos est relativement plus faible.

FIGURE 28 – BoxPlot présence ou absence de maladie cardiaque * dépression ST induite par l'exercice par rapport au repos



1.2.3 Variable qualitative * variable qualitative

Pour ces variables, nous allons nous intéresser à la relation existante entre ces dernières au moyen d'un **test du Chi2** qui permet de tester l'existence de liaison entre 2 variables qualitatives.

Les hypothèses du test :

H_0 : *indpendance entre les 2 variables*

H_1 : *Liaison entre les 2 variables*

✍ **sexe * type de douleurs thoracique**

L'analyse du tableau de contingence des deux variables permet de constater que les hommes sont les plus représentés pour tous les types de douleurs.

TABLE 9 – Add caption

	Femme	Homme
angine de poitrine typique	39	104
angine de poitrine atypique	18	32
douleur non angulaire	35	52
asymptomatique	4	19

En ce qui concerne la relation entre les deux variables, l'utilisation du test du chi2 tel que décrit précédemment donne une **pv-value = 0.078**. Ceci conduit à accepter H_0 donc les deux variables sont indépendantes.

✍ **sexe * présence de la maladie cardiaque**

L'utilisation du test du chi2 permet de conclure à une forte liaison entre les 02 variables car la **p-value = 1.877e-06** $< 5\%$.

TABLE 10 – tableau croisé du sexe suivant la présence ou non de la maladie cardiaque
Le tableau croisé entre les deux variables permet de constater que chez les patients de sexe masculin, la maladie cardiaque est plus présente.

	pas present	present
Homme	24	72
Femme	114	93

✍ **présence de maladie cardiaque * type de douleur thoracique**

L'observation du tableau croisé permet de se rendre compte que chez les patients n'ayant pas la maladie cardiaque, le type de douleur le plus présent est la douleur non angulaire. Chez ceux ayant la maladie cardiaque, la douleur de type angine de poitrine typique est la plus prononcée.

L'utilisation du test du chi2 permet de conclure à une liaison entre les 2 variables. En effet **p-value = 2.2e-16 < 5%**.

TABLE 11 – Tableau croisé type de douleur thoracique * présence ou absence de maladie cardiaque

	Present	Pas present
Angine de poitrine	39	104
angine de poitrine atypique	41	9
douleur non angulaire	69	18
asymptomatiques	16	7

✍ **présence de la maladie cardiaque * glycémie à jeun> 120 mg / dl**

L'observation du tableau croisé permet de constater que pour les patients ayant une glycémie < 120 mg/dl, la présence et l'absence de maladie cardiaque y sont fortement représentées. Ainsi, cette variable ne permet pas de caractériser la présence ou l'absence de maladie cardiaque. Le test du chi2 permet de confirmer le soupçon identifié ci-haut à

TABLE 12 – Add caption

	glycemie > 120	glycemie < 120
present	22	116
absence	23	142

savoir une indépendance entre la variable présence ou absence de maladie et la glycémie dans le corps > 120mg/dl. En effet **p-value = 0.7444 > 5%**. D'où on accepte H_0.

✍ **présence ou absence de maladie cardiaque * résultats électrocardiographiques au repos**

Le tableau croisé permet de constater que chez les patients présentant la maladie cardiaque, la modalité 1 de la variable résultats électrocardiographiques au repos est la plus représentée.

Le test de chi2 permet de conclure à une indépendance entre les 2 variables. En effet, **p-value = 0.006661 < 5%**.

TABLE 13 – Add caption

	0	1	2
absence	79	56	3
présence	68	96	1

✍ **présence ou absence de maladie cardiaque * angine de poitrine induite par l'exercice**

Les patients présentant une angine de poitrine non induite par l'exercice physique ont le plus de maladies cardiaques. Tandis que chez ceux dont l'angine de poitrine est causée par la maladie cardiaque, la maladie est moins présente.

L'utilisation du test du chi2 permet de conclure à une liaison entre les 02 variables. En effet **p-value = 7.454e-14 < 5%**.

TABLE 14 – tableau croisé de présence ou absence de maladie cardiaque * angine de poitrine induite par l'exercice

	non	oui
absence	62	76
presence	142	23

✍ **présence ou non de la maladie cardiaque * pente du segment ST maximal de l'exercice**

Le tableau croisé de ces deux variables permet de constater que plus la pente du segment ST maximale de l'exercice est élevée chez un patient, plus il est susceptible d'avoir la maladie cardiaque.

L'utilisation du test du chi2 permet de conclure à une liaison entre les 2 variables. En effet, **p-value = 4.831e-11 < 5%**.

TABLE 15 – tableau croisé présence ou non de la maladie cardiaque * pente du segment ST maximal de l'exercice

	0	1	2
absence	12	91	35
presence	9	49	107

✍ **présence ou non de la maladie cardiaque * nombre de vaisseaux principaux (0-3) colorés par une fluoroscopie**

Le tableau croisé des 2 variables permet de se rendre compte que chez les patients ayant la maladie cardiaque, le nombre de vaisseaux principaux (0-3) colorés par une fluoroscopie est nul.

L'utilisation du test du Chi2 permet de conclure à une liaison entre les 2 variables. En effet, **p-value = 2.712e-15 < 5%**.

TABLE 16 – Tableau croisé présence ou non de la maladie cardiaque * nombre de vaisseaux principaux (0-3) colorés par une fluoroscopie

	0	1	2	3	4
absence	45	44	31	17	1
presence	130	21	7	3	4

✍ **présence ou non de la maladie cardiaque * thal**

L'observation du tableau croisé entre les 2 variables permet de constater que chez les patients atteints de maladie cardiaque, la modalité 2 de la variable Thal est la plus représentée.

L'utilisation du test du chi2 permet de conclure à une forte liaison entre les 02 variables. En effet, **p-value = 2.2e-16** $<$ **5%**.

TABLE 17 – Tableau croisé présence ou non de la maladie cardiaque * thal

	0	1	2	3
absence	1	12	36	89
présence	1	6	130	28

1.3 Traitement des valeurs aberrantes

L'analyse uni-variée et bi-variée nous a permis d'identifier les valeurs aberrantes présentes dans notre base d'étude.En effet, dans l'analyse uni-variée, nous avons constaté grâce aux BoxPlot plusieurs valeurs aberrantes pour certaines variables à l'instar de la variable cholestérol sérique, dépression ST etc. Compte tenu de la normalité de ces dernières en se référant au test de **Jarque Bera**, nous avons utilisé le test de **Grubbs** qui a permis de réduire significativement le nombre de ces outliers par variable. Cependant, dans l'analyse bi-variée, lors du croisement variables qualitatives et quantitatives, nous avons constaté à l'aide du boxplot que les outliers détectés lors de l'analyse uni-variée étaient considérés comme aberrants dans les groupes de variables qualitatives formés par les modalités de ces dernières. Fort de ce constat, nous allons procéder à la correction des valeurs aberrantes détectées par le boxPlot, ceci pour la principale raison que nous ne voulons pas d'outliers à l'intérieur des groupes formés par les modalités des variables qualitatives. La méthode de correction adoptée dans le document est la correction par **la méthode de correction par la moyenne**. Celle ci se justifie pour les raisons suivantes :

- réduit l'influence de valeurs aberrantes ;
- facile à mettre en œuvre.

Comme inconvénient de ladite méthode :

- peut modifier les corrélations ;
- possibilité de changer la distribution des variables.

Chapitre 2

Identification de difficultés de l'étude et sélection de variables

L'analyse exploratoire nous a permis de mieux connaitre notre base de donnée, ceci à travers la découverte des relations bivariées pouvant exister entre les différentes variables du modèle.
Dans ce chapitre il est essentiellement question pour nous dans un premier temps d'identifier les principales difficultés pour l'atteinte des objectifs fixés. Dans un second temps, nous allons procéder à une sélection des variables susceptibles d'expliquer notre variable d'intérêt. Il est important de noter que cette procédure est une préalable pour les modèles de régression notamment le modèle de régression logistique.

2.1 Les différentes difficultés pour la réalisation de l'étude

Dans l'optique d'atteindre les objectifs de cette étude, il est important de cerner l'ensemble des difficultés auxquelles nous seront confrontés dans la suite. Ainsi, comme difficultés, nous avons :

- Identification des techniques d'apprentissage à appliquer au modèle ;
- Comprendre le fondement théorique et le contexte d'application des modèles retenus ;
- Trouver les codes des différentes techniques pour les mettre en œuvre sur R et surtout la spécification du modèle comme paramètre en entrée ;
- Identification des indicateurs de performance pour évaluer les différents modèles ;
- Trouver les codes pour la compilation de différents indicateurs sur R ;
- Compte tenu de la spécificité de chacun des indicateurs retenus, sur quel indicateur se baser pour choisir le modèle final ;
- Tracer une courbe synthétique de ROC pour les différents modèles retenus.

2.2 Choix des variables explicatives pour la prédiction

A présent, nous devons sélectionner l'ensemble de variables indépendantes qui vont nous permettre d'expliquer la présence de la maladie cardiaque dans l'échantillon de patients soumis à notre analyse. Dans l'idéal, les variables explicatives doivent répondre à deux caractéristiques :

- être indépendantes entre elles ;
- être le plus possible liées avec la variable à expliquer.

La sélection de variable s'impose à nous car celle permet d'affiner le modèle à mettre sur-pied. Ceci à travers pour 03 principales raisons :

— Elle permet de rendre le modèle plus opérationnel (on pourra poser peu de question pour prédire si un patient est atteint de maladie cardiaque ou pas) ;
— Elle conduit à la réduction du nombre de variables indépendantes ;
— Elle aide le modèle à être plus robuste en cas de généralisation (ceci dû à une nombre réduit de variables explicatives).

Compte tenu de l'absence d'avis d'expert pour nous guider dans le choix des variables explicatives, nous nous limiterons aux résultats fournis par cette méthode. Toutefois, il importe de noter que cette sélection de variable ne sera valable que pour certains modèles, tels que les modèles de régression que l'on abordera dans la suite. D'autres modèles à l'instar des arbres de décision sont dotés de leur propre mécanisme de sélection de variables suivant leurs critères.

Dans ce document, la méthode de sélection de variable utilisée est **la méthode de sélection par optimisation** proposée par le logiciel R. Elle repose sur le principe de trouver le sous-ensemble de variables prédictives qui minimise un critère. Les critères considérés sont le critère d'information AIC d'AIKAIKE et le BIC de Schwartz.

$AIC = -2LL + 2 * (J + 1)$

$BIC = -2LL + ln(n) * (J + 1)$ avec $-2LL$ *qui reprsente la dviance*;

$(J + 1)$ *est le nombre de paramtres estimer*,

J *le nombre de variables explicatives*

Cette méthode présente trois alternatives à savoir une sélection **forward**, **backward** et un sélection **stepwise (both)** qui combine les deux dernières.

2.2.1 Méthode forward

Principe : « part du modèle trivial, puis rajoute une à une les variables explicatives jusqu'à ce que l'on déclenche la règle d'arrêt.»

Après exécution de la commande **stepAIC** du package MASS de R avec l'option forward, on obtient comme modèle optimale suivant la méthode forward le modèle suivant :

$target = oldpeak + cp + ca + thal + exang + sex + trestbps + slope + chol$

2.2.2 Méthode Backward

Principe : « part du modèle complet, incluant la totalité des descripteurs, puis enlève une à une les variables non significatives.»

Pour cette méthode on utilise la même fonction que précédemment avec comme option "backward". Le modèle obtenu est le suivant :

$target = sex + cp + trestbps + chol + exang + oldpeak + slope + ca + thal$

2.2.3 Méthode stepwise

Principe : « consiste à vérifier si chaque ajout de variable ne provoque pas le retrait d'une explicative qui aurait été intégrée précédemment.»
L'option "both" de la fonction stepAIC permet d'avoir le rendu de cette méthode. Selon elle, le modèle optimal est le suivant :
$target = oldpeak + cp + ca + thal + exang + sex + trestbps + slope + chol$

En somme, nous constatons que les trois méthodes donnent le même résultat :

1. 9 variables explicatives qui sont identiques ;

Donc pour l'application de la régression logistique, le modèle que l'on retiendra est le suivant :
$target = oldpeak + cp + ca + thal + exang + sex + chol + trestbps + slope$

Chapitre 3

Présentation des différentes techniques

Le Data Mining reconnu sous l'appellation, fouille des données intervient après la conception de l'entrepôt de données. La technique utilisée en Data Mining est **ECD : Extraction des connaissances des données**. Ainsi, pour aboutir à l'extraction de ces connaissances, nous avons deux principaux types d'apprentissage : l'apprentissage supervisé et le non supervisé. L'apprentissage non supervisé qui utilise les techniques descriptives est généralement utilisé pour définir les classes. L'apprentissage supervisé quant à lui utilise les techniques prédictives. Dans ces techniques prédictives, suivant la nature de la variable dépendante (qualitative ou quantitative), on retrouve les techniques d'association et de régression. Compte tenu du fait que nous nous situons dans un cas de prédiction d'une variable binaire, les techniques à mettre en œuvre sont ceux appartenant à la famille de techniques d'apprentissage supervisé. Dans cette famille, nous allons appliquer les méthodes suivantes :

- Régression logistique ;
- Arbre de décision (CARD) ;
- Bagging ;
- Random Forest ;
- Support Vecteur Machine ;
- Réseau de neurones.

3.1 Régression logistique

3.1.1 Brève description

L'apprentissage par **Régression Logistique** est une technique d'apprentissage supervisée. Elle a pour objectif de prédire une variable catégorielle Y à partir d'une collection de descripteurs. Suivant le nombre de modalités de la variable Y, on est soit dans le cas binaire ou polynomiale. Ici, nous sommes dans le cas binaire ceci au vu de notre variable dépendante *Le patient présente t'il une maladie cardiaque* $(0, 1)$. Avec technique de prévision, il est possible au préalable de sélectionner les variables indépendantes du modèle. Ainsi, la sélection fait dans le précédent chapitre nous sera d'une très grande utilité, car c'est le modèle issu de l'une des méthodes (*Forward, Backward, stepwise* qui sera utilisé

pour spécifier notre modèle. Comme l'ensemble des trois méthodes conduit au même modèle, alors notre modèle de régression se fera sur ce dernier qui est ainsi spécifié :
$target = oldpeak + cp + ca + thal + exang + sex + chol + trestbps + slope$

3.1.2 Application

L'utilisation de la fonction *glm* du package **stats** implémenté par défaut sur r permet d'avoir le rendu de cette méthode. Les variables retenues pour cette technique sont celles obtenues lors de la sélection préliminaire en utilisant les critères de **forward**, **backword**, **stepwise**. Ainsi, la spécification du modèle retenu pour cette technique est suivante :
$target = oldpeak + cp + ca + thal + exang + sex + chol + trestbps + slope$
Comme argument de la fonction, il est important de préciser que $family = binomial$ pour nous permettre d'être dans le cas logistique. La fonction *summary* permet d'avoir le détail du résultat du modèle. Comme résultat, on constate que :

- Pour la variable sexe, on remarque que relativement aux femmes, les hommes sont plus favorables à ne pas avoir la maladie cardiaque. Aussi, cette dernière est significative au seuil de 5%.
- En considérant la modalité 0 comme référence, les patients ayant un nombre de vaisseaux principaux (0-3) colorés par une fluoroscopie correspondant aux modalités $1, 2, 3$ contribuent négativement au modèle.Autrement dit, plus leurs valeurs pour les modalités $1, 2, 3$ sont élevées relativement à la modalité 0 plus ces derniers sont enclins à ne pas avoir de maladie cardiaque (target = 0).

La matrice de confusion obtenue par cette méthode est donnée par le tableau suivant :

TABLE 18 – Matrice de confusion issue de la régression logistique

Maladie cardiaque	absent	présent
absent	42	7
présent	6	36

3.2 Arbre de décision

3.2.1 Brève description

Les arbres de décision sont utilisés pour la prédiction ou l'explication d'une variable cible catégorielle (Y) à partir d'un ensemble de variables explicatives (Xi). Le résultat est un ensemble de règles simples qui permettent de réaliser des prévisions ou de segmenter la population. Cependant, suivant le niveau de segmentation et la profondeur de segmentation, les performances de l'arbre varient. De ce fait, il est important de bien définir les spécificités des paramètres de l'arbre pour obtenir un arbre optimal tout en veillant au caractère aléatoire de l'échantillon. Ainsi, grâce au pruning (élagage de l'arbre) il est possible d'identifier et de supprimer les branches qui représentent du "bruit".

3.2.2 Application

Pour construire notre arbre, nous utiliserons le package *rpart*. Il est important de noter que l'arbre comporte ses propres algorithmes d'optimisation, donc nous lui donnons la base brute après correction des valeurs aberrantes. Pour plus affiner notre modèle, nous allons spécifier quelques paramètres :
MINSPLIT = 10 indique qu'on ne doit pas segmenter un nœud avec moins de 10 observations ;
MINBUCKET = 1 nous permet de refuser toute segmentation où un des nœuds enfants aurait moins d'une observation ;
coefficient de pénalité (cp) = 0.01 pour gérer la profondeur de l'arbre, car fonction du niveau de profondeur le biais et la variance de notre estimateur sont impactés. Cette valeur $cp = 0.01$ est celle fournie par défaut, pour ne pas avoir à afficher une structure inutilement grande.
La définition de ces deux paramètres ne suffit pas pour obtenir un arbre optimal, car la technique d'arbre de décision nous donne la possibilité d'optimisé le modèle, ceci grâce au **pruning ou élagage**. Nous allons par suite procéder au post(élagage de notre modèle. La commande *printcp* permet d'obtenir les cp optimaux. En effet, cette commande décrit des séquences d'arbres en mettant en relation les cp avec l'erreur calculée sur l'échantillon d'apprentissage ; le nombre de segmentation (minsplit), l'erreur en validation croisé. Le choix du modèle optimal après post élagage est basé sur l'arbre qui minimise l'erreur en validation croisée. Ainsi, Le rendu de cette fonction nous conduit à retenir l'arbre numéro 4 qui propose un $cp \in [0.034, 0.066]$, $minslip = 3$. Par suite la commande *prune()* permet d'obtenir cet arbre.

Avec ces critères optimaux, on se rend compte après prédiction que le taux d'erreur de ce modèle est très élevé ($tauxerreur = 0.33$). Ceci est dû au fait que nous avons trop réduit l'arbre, en l'occurrence le nombre minimal d'individu dans chaque nœud. Nous retiendrons finalement $cp \in [0.011, 0.033]$ ceci conduit à retenir $minslip = 3$ avec une erreur en validation croisé $xerror = 0.416$.
La matrice de confusion obtenue avec l'arbre optimale est la suivante :

TABLE 19 – Matrice de confusion issue de l'arbre de décision

Maladie cardiaque	absent	présent
absent	34	15
présent	7	35

3.3 Bagging

3.3.1 Brève description

Il est utilisé pour résoudre un compromis entre biais et variance. En général, les méthodes d'estimation s'intéressent de façon uni-latéral à l'un ou à l'autre des paramètres suivant les objectifs de l'étude. La particularité de la technique du bagging est de rechercher un double objectif de minimisation de la variance tout en veillant à garder un biais faible de l'estimateur. Ainsi, Il permettent d'obtenir de bons résultats pour les techniques d'estimations connues pour donner des estimateurs de faibles biais et de grandes variances, comme les arbres de décision. En effet, en créant des arbres individuels plus profonds (moins biaisés), on améliore sa qualité de prédiction.
principe : : « On tire indépendamment plusieurs échantillons boostrap et on applique une règle de base (arbre de décision) sur chacun de ces échantillons.»

3.3.2 Application

L'utilisation de la fonction *bagging* du package *adabag*, permet d'obtenir les résultats de la méthode en passant en argument la base brute avec prise en compte de la correction des valeurs aberrantes.
Comme spécification des paramètres de ladite fonction, nous retiendrons les paramètres d'optimalités obtenus avec le modèle issu de la technique d'arbre de décision. En effet, l'argument *control* de la fonction *bagging* est programmé identique à celui de l'arbre de décision grâce à *rpart.control*. Compte tenu du fait que la technique du **bagging** correspond à plusieurs échantillons de boostrap, nous retiendrons le nombre de tirage

mfinal = 100. Ce choix est motivé par le fait que plus le nombre de tirage est grand, plus le caractère aléatoire du tirage est robuste. Cependant comme limite à ce choix, il est important de noter que ce dernier peut conduire à faire planter le programme. On obtient la matrice de confusion suivante :

TABLE 20 – Matrice de confusion issue du bagging

Maladie cardiaque	absent	présent
absent	36	13
présent	7	35

3.4 Forêt Aléatoire (Random Forest)

3.4.1 brève description

La méthode de Random Forest est une modification du bagging qui permet de construire un grand nombre d'arbres de décision non corrélés entre eux dans le but de les agréger ensuite.

principe : Random Forest : « Dans la construction de chaque arbre, pour découper chaque noeud, on tire aléatoirement un nombre m de variables. Ces arbres obtenus qui sont non corrélés sont ensuite agrégées.»
Selon *Breiman, (2001)*, en pratique le Random Forest améliore les performances du bagging. Ceci est imputable à son caractère doublement aléatoire (alétoire pour le tirage des m variables et lors du bagging simple).

3.4.2 Application

L'utilisation de la fonction *randomForest* du package **randomForest** de r nous permet d'obtenir le rendu de cette technique. La spécification du modèle considéré contient toutes les variables du modèle sans considération de la sélection préalable des variables. Aussi, ces variables sont corrigées de leurs oultiers.
Pour la mise en oeuvre de cette fonction, nous passerons en argument le modèle spécifié de la base apprentissage. L'argument *ntree* est laissé à sa valeur par défaut à savoir $ntree = 500$. Ceci pour augmenter la robustesse du caractère aléatoire du nombre d'arbres à construire. Donc, cette méthode construit 500 arbres de décision aléatoires avec une taille d'échantillon aléatoire (cette dimension d'aléatoire est imputable au bagging).
La matrice de confusion obtenue est la suivante :

TABLE 21 – Matrice de confusion issue du Random Forest

Maladie cardiaque	absent	présent
absent	38	11
présent	7	35

3.5 SVM

3.5.1 brève description

Les SVM sont une technique d'apprentissage supervisé d'une variable explicative binaire. L'idée est de rechercher une règle de décision basée sur une séparation par hyperplan de marge optimale. Le principe de l'algorithme est d'intégrer lors de la phase d'apprentissage une estimation de sa complexité pour limiter le phénomène de sur-apprentissage (over-fitting).
Principe de l'algorithme : il repose sur la recherche de bonnes performances du modèle, tant en qualité de prédiction qu'en terme de complexité du modèle obtenu. il peut être résumé en trois étapes :

- On cherche l'hyperplan comme solution d'un problème d'optimisation sous-contraintes. La fonction a optimiser intègre un terme de qualité de prédiction et un terme de complexité du modèle ;
- ensuite, on utilise le noyau de kernel qui a pour effet de coder une transformation non linéaire des données ;
- Numériquement, toutes les équations s'obtiennent en fonction de certains produits scalaires utilisant le noyau et certains points de la base de données (ce sont les Support Vecteurs).

3.5.2 Application

L'utilisation de la fonction *svm* du package **e1071** de r permet d'avoir le rendu de cette méthode. Pour cette fonction, nous considérons comme spécification du modèle, l'ensemble des variables de la base de donnée. Ainsi, l'étape de sélection de variables précédemment abordée n'est pas d'une utilité pour ce modèle. La base à considérer est la base d'apprentissage. Concernant les paramètres du modèle, la principale spécification intervient au niveau de l'argument *kernel* qui permet de définir la fonction de noyau à utiliser. Une première application de ce modèle sans utilisation de la fonction noyaux nous permet de constater un fort taux d'erreur. Fort de ce constat, l'utilisation de la fonction

de noyau linéaire ($kernel = linear$) permet d'avoir une erreur beaucoup plus faible. Aussi, il est à noter que lorsque l'on utilise cette fonction de noyau, le paramètre gamma de la fonction de svm est $gamma = \frac{1}{dimensiondelabase}$. Donc nous allons nous en tenir à cette fonction. La matrice de fusion du modèle est donnée par le tableau suivant :

TABLE 22 – Matrice de confusion issue des SVM

Maladie cardiaque	absent	présent
absent	37	12
présent	4	38

3.6 Les réseaux de neurones

3.6.1 brève description

Un réseau de neurone ou réseau neuronal a une architecture calquée sur celle du cerveau, organisée en neurone et en synapses, et se présente comme un ensemble de noeuds connectés entre eux. Nous principalement 4 types de réseaux de neurones à savoir :

- *Le perceptron multicouche (PMC)* : adapté pour prédire une variable cible continue ou discrète ;
- *La fonction à fonction radiale de base (RBF)* : utilisée pour prédire une variable cible continue ou discrète ;
- *Le réseau de Kohonen* : effectue des analyses typologiques (clustering, recherche de segments) ;
- Réseau par estimation de densité de probabilité : utilisé pour le classement et la régression ;
- *Analyse discriminante généralisée* : les SVM sont utilisés ici pour prédire une variable cible discrète.

Fort de cette typologie, le constat clair est que le *PMC et RBF* sont les plus indiqués pour notre étude. Par suite, nous allons utiliser le Perceptron MultiCouche, ce choix s'explique par le fait que :

➥ le PMC a une meilleure capacité de généralisation [5]. En effet, cette avantage est précieuse car l'intérêt même de l'utilisation de cette technique réside dans sa capacité à pouvoir être efficace et présenter de bonnes performances avec de nouveaux jeux

5. Stéphane Tufféry, Mai 2017

de données. Aussi, compte tenu du manque d'avis d'expert du domaine pour mieux préciser la sélection des variables, nous avons tout intérêt à avoir un modèle général ;

➥ Le PMC est facilement applicable sur R.

Principe de fonctionnement : "Un nœud reçoit des valeurs en entrée et renvoie 0 à n (n : taille de l'échantillon d'apprentissage = 212, dans le cas de l'étude) valeurs en sortie. Toutes ces valeurs sont normalisées pour être comprises entre 0 et 1 (parfois entre -1 et 1, selon les bornes de la fonction de transfert)."
Il est important de noter qu'avant d'avoir la valeur de sortie, il y'a des étapes intermédiaires, notamment l'utilisation des fonctions pour transformer les valeurs d'entrées en celles de sorties. Comme fonction nous avons la **fonction de combinaison**, qui calcule une première valeur à partir des nœuds connectés en entrée et du poids des connexions. Dans notre cas de PMC, la fonction utilisée est la somme pondérée $\sum_{i=1}^{i=n} n_i p_i$. Afin de déterminer une valeur en sortie, une seconde fonction, appelée **fonction de transfert ou d'activation** est appliquée à cette valeur. Les nœuds de la couche d'entrée sont triviaux, dans la mesure où ils ne combinent rien, et ne font que transmettre la valeur de la variable qui leur correspond.
Afin que cette technique de réseau de neurones puisse être utilisée de façon efficiente, nous allons nous attelés à lui donner une bonne structure. En effet, **la structure du réseau de neurones** (architecture ou topologie) du réseau de neurones désigne le nombre de couches et de nœuds, la façon dont sont inter-connectés les différents nœuds (choix des fonctions de combinaisons et de transfert) et le mécanisme d'ajustement des poids.

3.6.2 Application

Le package *nnet* a été utilisé dans cette étude pour obtenir le résultat de cette méthode. Dans ce package, la fonction *nnet* a été utilisée. Il est important de noter que ce package se limite à l'utilsation du perceptron à un couche. Cependant, avec les options d'optimisations telles que proposées par la package $e1071$ [6] des SVM permet de nous garantir la robustesse de nos résultats.
La matrice de confusion du modèle est ainsi fournie.

TABLE 23 – Matrice de confusion issue des réseaux de neurones

Maladie cardiaque	absent	présent
absent	37	12
présent	6	36

6. Car la librairie *nnet* ne comporte pas d'option d'optimisation

Chapitre 4

Étude comparative des différents modèles

Dans la précédente partie, nous avons passé en revue quelques techniques de prédiction qui cadrent avec l'objectif de notre étude. Dans cette partie, il est essentiellement question pour évaluer la performance de chacun des modèles précédemment construits et sur la base des indicateurs de performance retenus, de choisir le modèle qui cadre le plus avec notre étude c'est-à-dire celui qui présente les indicateurs optima pour avec l'échantillon considéré pour cette étude.
Comme indicateurs de performance, ceux que nous avons retenu sont les suivants :

- La courbe de ROC ;
- L'AUC de la courbe de ROC ;
- le Kappa de Cohen ;
- le taux d'erreur de prédiction ;
- La sensibilité ;
- La spécificité ;

Dans la suite, après une brève présentation de ces indicateurs, nous allons passer au choix du meilleur modèle en conformité avec notre échantillon.

4.1 Critères d'évaluation

Il est très important d'évaluer les modèles de prédiction. En effet, cette évaluation va permettre de :

- Se donner une idée des performances du modèle en déploiement (fiabilité et robustesse du modèle) ;
- Comparer les modèles candidats ;
- Savoir si le modèle est globalement significatif.

Par suite, nous passerons en revue les différents indicateurs de performance précédemment cités.

4.1.1 La courbe de ROC

La courbe de ROC (Receiver Operator Curve) permet de représenter la $sensibilite = f(1 - specificite)$. Cette courbe est un outil d'évaluation et de comparaison des modèles. Son importance réside dans sa puissance. En effet, sa portée va largement au-delà des

indicateurs issus de l'analyse de la matrice de confusion. Aussi, elle représente un outil graphique qui permet de visualiser les performances. En effet, un seul coup d'oeil devrait permettre de voir le(s) modèle(s) susceptibles(s) de nous intéresser. Par ailleurs, la courbe fournit des résultats valables même si l'échantillon test n'est pas représentatif. Ce qui n'est pas le cas pour les indicateurs obtenus grâce à la matrice de confusion.
Pour sa construction, on s'aidera des fonctions *auc* et *roc* du package *pROC* de r.

4.1.2 L'AUC de ROC

L'AUC (Area Under Curve) de ROC appelé généralement l'Aire en dessous de la courbe de ROC, est un indicateur d'évaluation des modèles qui varient entre 0 *et* 1. En effet, l'AUC est une probabilité qui indique l'évènement que la fonction score place un patient sain devant un patient ayant la maladie cardiaque. Un score de 1 indique une prédiction parfaite, un score de 0.5 caractérise un pouvoir de prédiction aussi bon que la chance (aléatoire, le modèle de prédiction ne sert à rien) et une AUC inférieure à 0.5 caractérise une prédiction moins bonne que la chance. Cette statistique qui ne dépend pas d'un seuil est proche de la *Statistique de Mann-Whithney* (le U de *Statistique de Mann-Whithney*), donc elle est à ce titre une statistique de rang.
Il est à noter que cet indicateur est peu sensible au mauvais classement de l'échantillon. Cependant, comme évoquer lors de l'analyse univariée de la variable *Target*, notre échantillon n'est pas sujette à ce cas. Donc, par suite, l'efficacité relative ou pas de cet indicateur ne pourra être imputable à la prédisposition de l'échantillon.
L'utilisation des fonctions de compilation de la courbe de roc permet d'avoir l'AUC.

4.1.3 Le Kappa de cohen

Cet indicateur permet de chiffrer l'accord entre les observations. Il estime un taux de concordance entre les observations et leur prédiction, en tenant en compte des erreurs d'omission et de commission. Plus il est élevé, plus le modèle est de bonne qualité.
En considérant la matrice de confusion (MF) suivante :

Maladie cardiaque	absent	présent
absent	a	b
présent	c	d

Le *k de cohen* est donné par la formule suivante : $K = \frac{P_{boe_pred} - P_{hazard}}{1 - P_{hazard}}$
avec $P_{boe_pred} = \frac{a+d}{a+b+c+d}$ $P_{hazard} = P_{malade} + P_{absent}$
$P_{malade} = \frac{c+d}{a+b+c+d} * \frac{b+d}{a+b+c+d}$ $P_{absent} = \frac{a+b}{a+b+c+d} * \frac{a+c}{a+b+c+d}$

4.1.4 Le taux d'erreur de prédiction

Il est un indicateur synthétique pertinent. Il estime la probabilité de mal classer un individu de la population.
En utilisant la matrice de confusion MF précédemment définie, la formule est la suivante :

$taux_erreur = \frac{b+c}{a+b+c+d}$

4.1.5 La sensibilité

Elle représente la probabilité d'avoir un patient malade sachant qu'il est malade.
En utilisant la matrice de confusion MF elle est donnée par :

$sens = \frac{d}{d+c}$

4.1.6 La spécificité

la **Spécificité** est un indicateur qui renseigne sur la probabilité pour un patient de ne pas avoir une maladie cardiaque sachant qu'il n'a pas la maladie.
En utilisant la matrice de confusion MF, elle est donnée par :

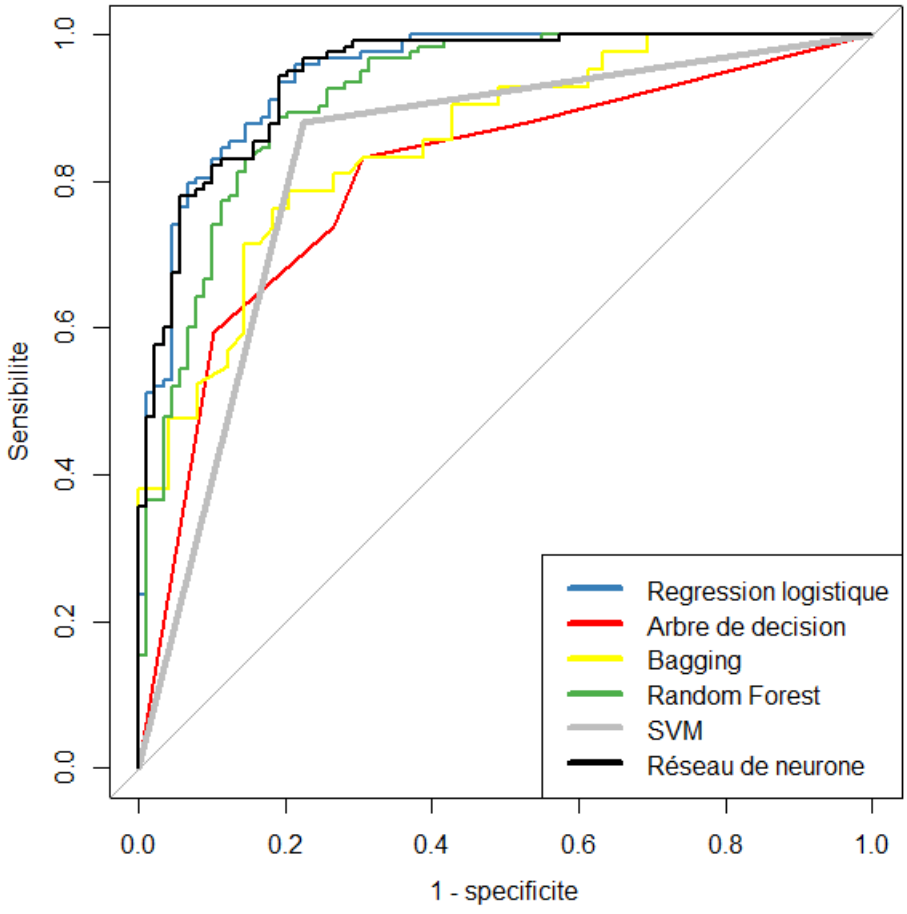
$spe = \frac{b}{a+b}$

4.2 Application et résultat

Pour synthétiser cette étape de choix du modèle, nous allons utiliser un tableau synthétique et la courbe de roc comportant le graphe superposé des 6 modèles.

4.2.1 La courbe de ROC

FIGURE 29 – Courbe de ROC des différents modèles



Interprétation

Pour cette analyse nous allons principalement nous intéresser à l'enveloppe convexe. Cette dernière est formée par les courbes qui, à un moment ou à un autre, n'ont aucune courbe « au-dessus » d'elles.
Ainsi, à l'observation de la courbe de ROC de nos différents modèles permet de constater que l'enveloppe convexe est formée par :

— La courbe du modèle de régression de logistique ;
— La courbe du réseau de neurone.

Pour ces deux modèles, on constate que leur performance s'alterne. En effet, sur l'intervalle 0.1 – 0.4 le modèle de régression logistique à une performance.

4.2.2 Tableau des indicateurs des modèles

Pour construire ce tableau, on s'aidera de la fonction *confusion Matrix* du package *caret* de r. Cette fonction renvoie l'ensemble des indicateurs précédemment. Il est à noter que cette fonction prend les mêmes arguments que ceux utilisés pour la matrice de confusion à savoir la *variable de prédiction* et *L'observation de la variable cible de l'échantillon d'apprentissage*. Cependant, suivant le modèle, certaines modifications sont à apporter à la *variable de prédiction* .

TABLE 24 – Tableau synthétiques des indicateurs de performance des différentes techniques utilisées

	AUC de ROC	K de cohen	Taux d'erreur	Sensibilité	Spécificité
Régression logistique	0.945	0.6915	0.1538	0.8723	0.8182
Arbre de décision	0.805	0.5201	0.2418	0.8293	0.70
Bagging	0.853	0.5185	0.2418	0.8140	0.7083
Random Forest	0.912	0.6047	0.1978	0.8444	0.7609
SVM	0.830	0.6498	0.1758	0.8837	0.7708
Réseaux de neurone	0.944	0.606	0.2198	0.861	0.75

Interprétation

En ce qui concerne le tableau des indicateurs de performance, il est à noter que :

— **Pour l'AUC :** Le modèle de régression logistique et celui du réseau de neurone présentent les AUC les plus élevés ; Ceci est tout à fait normal compte tenu du fait ces deux constituent l'enveloppe convexe de la courbe de ROC ;
— **Pour le K de cohen :** Le meilleur modèle est celui de régression logistique suivi du modèle des SVM ;

— **Le taux d'erreur :** le modèle de régression logistique et celui des SVM présentent les plus faibles taux d'erreurs. C'est que relativement aux autres modèles de l'étude, ces dernières donnent une meilleure prédiction de l'état de santé du patient en conformité avec la matrice de confusion ;
— **Sensibilité :** La meilleure sensibilité est donnée par le modèle de regression logistique suivie de celle du réseau de neurones. Ainsi, ces dernières prédisent bien le classement des individus malades ;
— **Spécificité :** Suivant cet indicateur les deux meilleurs modèles sont : la régression logistique et le SVM. Donc, elles ont tendance à bien classer les patients qui ne présentent pas de maladie cardiaque.

4.3 Modèle retenu

Les précédents résultats (analyse de la courbe de ROC, l'analyse du tableau synthétique) permettent de constater que **le modèle de régression logistique, le réseau de neurone et les SVM** présentent de meilleures performances.
Compte tenu de l'avantage de l'utilisation de la courbe de roc à savoir : de permettre d'aller au delà des interprétations des indicateurs issus de la matrice de confusion, qui sont fortement dépendantes de l'échantillon de test, et de prendre en compte les coûts de mauvais classement, nous allons nous en tenir aux meilleurs modèles suivant cette courbe plus précisément à l'AUC. Un indicateur synthétique de la courbe (qui permet d'éviter les entrelacement des 2 courbes) est l'AUC.
Suivant l'AUC, le meilleur pour modèle pour cette étude est celui obtenu grâce à la méthode de **régression logistique**.

Conclusion

En définitive, à l'issue de cette étude qui consistait à prédire l'occurrence d'une maladie cardiaque à l'aide des caractéristiques d'un échantillon de 303 patients, plusieurs résultats ont été obtenus. D'abord, l'analyse exploratoire sur les données de l'échantillon a montré que certaines variables suivent la loi normale par observation de l'histogramme et de la courbe de la loi normale en plus du test de Jarque Berra. Il ressort également de l'observation du boxplot que toutes les variables présentaient des valeurs aberrantes que le test numérique de grubbs a d'ailleurs confirmées. Il s'en est suivi le traitement de ces valeurs aberrantes par la méthode de l'imputation par la moyenne. L'analyse univariée a été complétée par l'analyse bivariée où les liaisons entre les variables quantitatives et qualitatives ont été passées en revue. En outre, les différentes difficultés de l'étude ont été les suivantes : la compréhension du fondement théorique et le contexte d'application des modèles retenus, l'identification des indicateurs de performances pour évaluer les différents modèles et la réalisation de la courbe synthétique de ROC. Par suite, les différentes techniques à savoir *la régression logistique*, *l'arbre de décision*, *le bagging*, *le Random Forest*, *le SVM et les réseaux de neurones* ont été présentées puis appliquées aux données. Enfin, à l'issu de cela, dans notre cas d'application la régression logistique est le meilleur modèle pour la prédiction de l'occurrence de la maladie cardiaque.

Comme principale **limite de l'étude**, on note le manque d'avis d'expert du domaine pour apporter ses compétences métiers à cette étude. De plus la faible taille de l'échantillon constitue une autre limite de l'étude. En effet, disposer d'une taille d'échantillon plus grande pourrait augmenter les performances, la robustesse et la généralisation de nos modèles.

Par ailleurs, cette étude nous a permis non seulement de mettre en pratique les connaissances théoriques acquises en classe mais également d'apprendre de nouvelles techniques d'apprentissages supervisée telles que : *le Random Forest*, *le réseau de neurones*. L'un des constats les plus marquants de cette étude qui nous a subjugué est le fait que nos modèles respectent le soubassement théorique de ces derniers. En effet, le Random Forest est considéré comme étant une amélioration du bagging (aléa au niveau du nombre d'arbres), lui même considéré comme une amélioration de l'arbre de décision (bootstrap au niveau de la taille de l'échantillon d'apprentissage). Cette relation ressort clairement dans notre modèle grâce à l'observation des indicateurs de performance tels que l' AUC et le taux d'erreur. En plus, cette étude à contribuer à améliorer significativement nos qualités en terme de travail de groupe.

Références bibliographiques

1. M. Abdoul Aziz NDIAYE, Enseignant à l'ENSAE-Dakar ;
2. Fabrice ROSSI, enseignant à TELECOM Paris
3. Ricco RAKOTOMALALA, enseignant à l'université de Lyon 2 ;
4. Stéphane Tufféry
5. Sébastien Gadat, Laboratoire de Statistique et probabilité

Annexes

Annexe 1 : Test de normalité de Jarque-bera

Test de normalité Jarque-Bera

Il est fondé sur les coefficients d'asymétrie et d'aplatissement. Il évalue les écarts simultanés de ces coefficients avec les valeurs de référence de la loi normale. Sa formulation est très simple et devient intéressant lorsque les effectifs sont élevés.

La statistique de test de Jarque-Bera est la suivante :

$$T = n\left(\frac{{b_1}^2}{6} + \frac{(b_2 - 3)^2}{24}\right)$$

Elle est distribuée asymptotiquement suivant une loi de Khi 2 à 2 degré de libertés.

Les hypothèses du test :

$$\begin{cases} H_0 : La\ série\ suit\ une\ loi\ normale \\ H_1 : la\ série\ ne\ suit\ pas\ une\ loi\ normale \end{cases}$$

On l'obtient sur R grâce à la fonction :

```
jarqueberaTest(x, title = NULL, description = NULL)
```

Contenu dans le package « fBasics ».

Annexe 2 : Test de détection des valeurs aberrantes de Grubbs

Test de détection de valeurs aberrantes de Grubbs

Le test de Grubbs, également connu sous le nom de test du maximum des résidus normalisés, recherche à déterminer si la valeur maximale ou minimale de X est un outlier. Ainsi, avec la p-value renvoyer par le test on peut réitérer jusqu'à la détection de toutes les valeurs aberrantes présentes dans la série.

La statistique de test est la suivante :

$$G = max_{i=1,2,3,\dots} \frac{|x_i - \overline{x}|}{}$$

Les hypothèses du test :

$$\begin{cases} H_0 : La\ valeur\ testée\ n'est\ pas\ un\ outlier \\ H_1 : la\ valeur\ testée\ est\ un\ oultier \end{cases}$$

On l'obtient sur R grâce à la fonction :

```
grubbs.test(x, type = 10, opposite = FALSE, two.sided = FALSE)
```

Contenu dans le package « outliers».

NB : Il est important de noter que le test de Grubbs permet de tester une seule valeur soir le min ou le max de la série entrée en paramètre (en précisant l'option `opposite` de la fonction. Ainsi, pour tester plusieurs valeurs suivant que l'on teste les valeurs maxima ou minima de la série comme oultier, nous avons utilisé la boucle While dans dans r.

Annexe 3 : Coefficient de corrélation de pearson, test de student

Coefficient de corrélation de pearson & test de significativité de student

Il constitue une mesure de l'intensité de liaison linéaire entre 2 variables. Il peut être égal à zéro alors qu'il existe une liaison fonctionnelle entre les variables. Toutefois, il repose sur la normalité des variables. Hors dans notre cas, nos variables ne suivent pas toutes la loi normale, donc on serait tenté de penser aux coefficients de **spearman** ou **kendall** qui eux, sont non paramétriques, mais sont compilés sur les rangs des observations de chaque variable.

Compte tenu du fait que nous nous intéressons à l'existence d'une liaison linéaire entre nos variables à des fins de construction de notre modèle de prédiction, nous allons nous limiter à l'utilisation du **coeficient de corrélation de pearson**. Il est égal au rapport de la covariance et de la racine carré des écarts types de chaque variable.

$$r(x,y) = \frac{cov(x,y)}{\sqrt{x}\sqrt{y}}$$

Afin de se prononcer sur la valeur de ce dernier, il est important de tester s'il est significativement différent de 0 ou non. Ainsi, pour le faire le test généralement utilisé est celui de student :

Les hypothèses du test sont les suivantes :

$$\begin{cases} H_0 : le\ coefficient\ de\ corrélation\ est\ égale\ à\ 0 \\ H_1 : le\ coefficient\ de\ corrélation\ est\ différent\ de\ égale\ à\ 0 \end{cases}$$

On l'obtient sur R grâce à la fonction :

```
cor.test(x, method = "pearson")
```

Contenu dans le package « stats», package standart de r.