UQAC
Université du Québec
à Chicoutimi

PROJET DE FIN DE SESSION

Étude de classification : Facteurs de risques et maladie coronarienne

Travail présenté à M. Julien Maitre par

Colin Bouchard — BOUC10049302

Étienne Leblanc — LEBE20109205

Frédérike Tremblay — TREF09519705

Comme exigence partielle du cours

Outils de programmation pour la science des données

8PRO408

21 décembre 2020

Département d'informatique et de mathématique

Table des matières

Étude de classification : Facteurs de risques et maladie coronarienne

1. DESCRIPTION DE L'ENSEMBLE DE DONNÉES

Les données obtenues du site de données en libre accès Kaggle sont une portion d'un ensemble de données plus large provenant d'une étude effectuée sur les facteurs de risques coronariens (Rossouw *et al.* 1983; Hamdaoui 2020). Ces données ont été collectées sur des individus caucasiens dans l'objectif d'étudier la prévalence des facteurs réversibles et irréversibles impliqués dans les maladies coronariennes, au niveau de la région du Cap-Occidental en Afrique du Sud (Rossouw *et al.* 1983). Les données ont été collectées dans le cadre de la Coronary Risk Factor Study (CORIS), une étude lancée par le Conseil sud-africain de la recherche médicale, par le Département de la Santé et par le Conseil de la recherche sur le bien-être et en sciences humaines, en 1978 (Rossouw *et al.* 1983; Rossouw *et al.* 1993). C'est en 1979 qu'un sondage fût lancé sur 3 villes se trouvant dans la région du Cap-Occidental (Swellendam, Riversdale et Robertson) permettant de recruter hommes et femmes caucasiens (respectivement, dans l'étude initiale, 3357 hommes et 3831 femmes entre 15 et 64 ans) (Rossouw *et al.* 1983). Un questionnaire sur les facteurs de risques fut donné aux participants à l'étude ainsi qu'une entrevue avec des diététiciens entraînés pour fournir un second questionnaire provenant de l'École de l'Hygiène de Londres (Rossouw *et al.* 1983). Les questions posées et les données amassées ont permis de couvrir plusieurs items socio-économiques, les habitudes en lien avec le tabac, l'historique médical et familial en lien avec les cardiopathies ischémiques, ainsi que des données sur la tension artérielle et des échantillons sanguins (Rossouw *et al.* 1983).

L'ensemble de données a par la suite été réutilisé, en partie, afin de démontrer de manière empirique l'intérêt potentiel de la réduction de dimensionnalité en apprentissage machine, et ce avec différents algorithmes (Hamdaoui 2020). C'est ce dernier sous-ensemble qui sera utilisé dans le présent projet. D'ailleurs, les variables qui sont incluses sont le numéro de l'individu, la

tension artérielle systolique, le nombre de kilogrammes de tabac cumulatif, le cholestérol à lipoprotéines de basse densité, l'adiposité, les comportements de type A, l'obésité, la consommation d'alcool, l'âge et l'historique familial (Hamdaoui 2020).

L'ensemble de données utilisé possède deux classes, c'est-à-dire la présence (1) ou l'absence (0) de maladie coronarienne. Le problème de classification suivant en sera donc un binaire. C'est cette classe qui sera utilisée à titre de variable prédictive dans le cadre du présent travail (Hamdaoui 2020). Le nombre d'instances total de cette dernière classe est de 462 ; plus précisément, le nombre d'instances pour les individus atteint de maladies coronariennes est de 160, tandis que le nombre d'individus non atteint est de 302. Pour ce qui est des valeurs maximales et minimales que peut prendre chacune des variables, il s'agit de se référer au tableau 1. Il faut souligner que l'index, représentant la variable du numéro de l'individu, a été supprimé. Il a aussi été possible de changer les valeurs d'historique familiales présente et absente par leur représentant binaire (respectivement 1 et 0).

Tableau 1. Valeurs minimales et maximales des variables de l'ensemble de données

Variables	Tension artérielle systolique (mmHg)	Tabac (kg)	Cholestérol à lipoprotéines de basse densité (mmol/L)	Adiposité (facteur)	Comportement de type A (indice de Bortner)	Obésité (IMC)	Alcool (ml/jour)	Âge (année)	Historique familial (présent (1) ou absent (0))
Minimum	101.00	0.00	0.98	6.74	13.00	14.70	0.00	15.00	0.00
Maximum	218.00	31.20	15.33	42.49	78.00	46.58	147.19	64.00	1.00

Il est important de mentionner que l'ensemble des variables seront ici utilisées.

2. VÉRIFICATION ET PRÉTRAITEMENT DES DONNÉES

Il est important de mentionner que l'ensemble de données ne possède initialement pas de valeurs manquantes. Il n'a ainsi pas été nécessaire de trouver une façon de les gérer. En ce qui concerne les statistiques descriptives, il est possible de les identifier dans le tableau 2.

Tableau 2. Statistiques concernant la moyenne, la variance et l'écart-type des variables de l'ensemble de données

Variables	Tension artérielle systolique (mmHg)	Tabac (kg)	Cholestérol à lipoprotéines de basse densité (mmol/L)	Adiposité (facteur)	Historique familial (présent (1) ou absent (0))	Comportement de type A (indice de Bortner)	Obésité (IMC)	Alcool (ml/jour)	Âge (année)
Moyenne	138.32684	3.6356	4.740325	25.406732	0.415584	53.103896	26.044113	17.0444	42.816017
Variance	420.099018	21.096	4.288665	60.539271	0.243401	96.383976	17.755101	599.322	213.42161
Écart-type	20.496317	4.593	2.070909	7.780699	0.493357	9.817534	4.21368	24.4811	14.608956

Il s'agissait ensuite de vérifier la distribution des différentes variables de l'ensemble de données (figures 1 à 8). Les figures de la tension artérielle systolique, du tabac, des LDL, du comportement de type A, de l'obésité et de l'alcool possèdent des valeurs aberrantes, ce qui n'est pas observable dans les distributions concernant l'adiposité et l'âge des participants. Les classes concernant les maladies coronariennes n'ont par contre pas été affichées, considérant qu'elles ne représentent que la présence ou l'absence de la maladie. La même chose s'est appliquée ici à la variable d'historique familiale de maladie coronarienne, qui ultimement représente aussi la présence ou l'absence d'un tel historique.

En ce qui concerne la figure 1, il est possible de constater que la médiane des participants possède une tension systolique supérieure à la tension systolique artérielle normale d'un adulte, qui est de 120 mmHg (OMS 2015). Ceci permet de mettre en lumière que certains participants possèdent des tensions systoliques anormales, ce qui pourrait corréler avec la présence dans l'ensemble de données de participants présentant des maladies coronariennes. Pour la distribution des données sur le tabac (figure 2), il est possible de voir que certains participants possèdent des consommations élevées (kg) de la substance. Il faut souligner que de manière générale, une grande consommation de tabac (c'est-à-dire de cigarettes fumées par jour) augmente le risque de coronopathie (Gouvernement du Canada 2011). Les individus présentant ces consommations

élevées sont ainsi plus à risque de développer des maladies coronariennes. De manière générale, le taux de LDL dans le sang des participants est plus élevé qu'une valeur normale de LDL chez l'adulte de moins de 50 ans (figure 3). En effet, cette valeur doit généralement être inférieure à 3,5 mmol/L pour être favorable à la santé (Statistique Canada 2013). L'adiposité présente une distribution assez grande (figure 4) ; il est par contre possible de souligner que les individus qui présenteront une plus grande quantité de tissus adipeux, notamment au niveau abdominal, ont plus de chance de développer des problèmes cardiovasculaires (Statistique Canada 2015). Les comportements de type A sont évalués en fonction de l'échelle de Bortner (Rossouw *et al.* 1983). Cette dernière échelle permet d'évaluer les comportements rendant à risque les individus au développant de maladie coronarienne (Rossouw *et al.* 1983). Les personnes accumulant plus de 55 points sont considérées comme possédant la personnalité de type A (Rossouw *et al.* 1983). La figure 5 permet de mettre en lumière que la distribution de ce comportement est assez étendue et présente même des extrêmes sous la boîte. Pour l'obésité, ou l'indice de masse corporelle (IMC) (figure 6), la distribution permet de voir que plusieurs participants présentent un IMC supérieur à 25, ce qui est associé à un excès de poids (Institut de cardiologie de Montréal 2020). D'ailleurs, pour les IMC au-dessus de 25, les risques de développer des problèmes de santé deviennent de plus en plus élevés (Institut de cardiologie de Montréal 2020). La distribution de la consommation d'alcool met en lumière que de manière générale, la consommation dans la population à l'étude est assez basse ; par contre, plusieurs valeurs extrêmes sont présentes qui soulignent une consommation importante d'alcool (figure 7). Le lien entre une consommation élevée d'alcool et la tension artérielle a déjà été établi : cette dernière augmente, et entraîne donc un risque plus élevé de mourir d'une maladie coronarienne (Foerster *et al.* 2010). Finalement, la distribution de l'âge correspond aux données connues et attendues (figure 8).

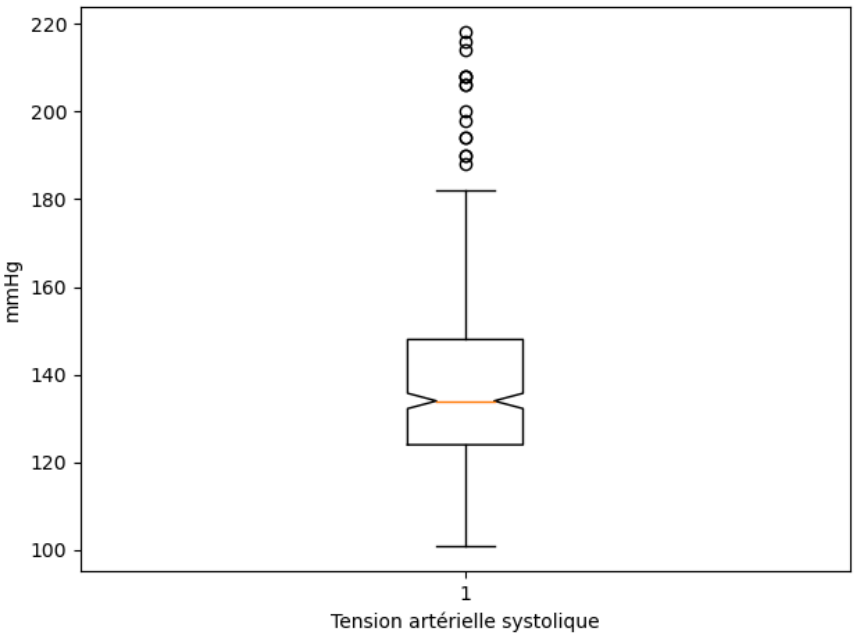

Figure 1. Diagramme en boîte représentant les différentes statistiques (médiane, quantiles, valeurs aberrantes) de la tension artérielle systolique (mmHg) des individus de l'ensemble de données.

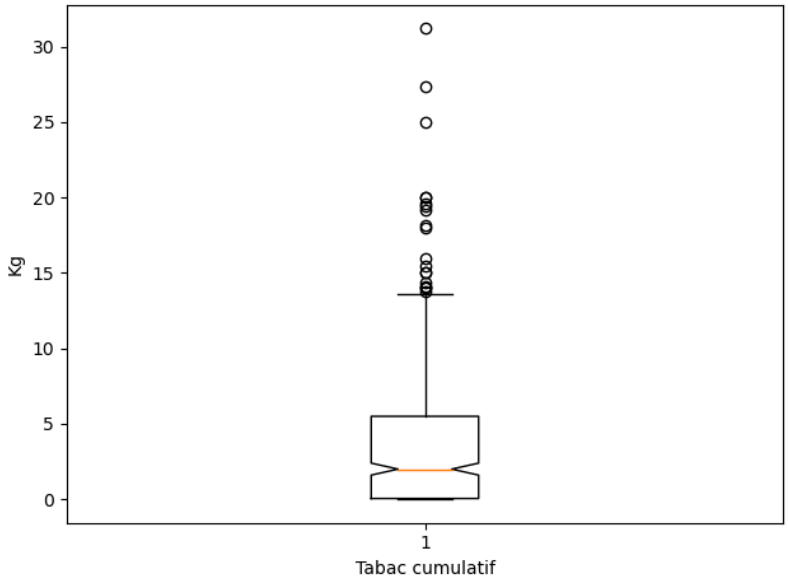

Figure 2. Diagramme en boîte représentant les différentes statistiques (médiane, quantiles, valeurs aberrantes) de la quantité de tabac utilisé (kg) par les individus de l'ensemble de données.

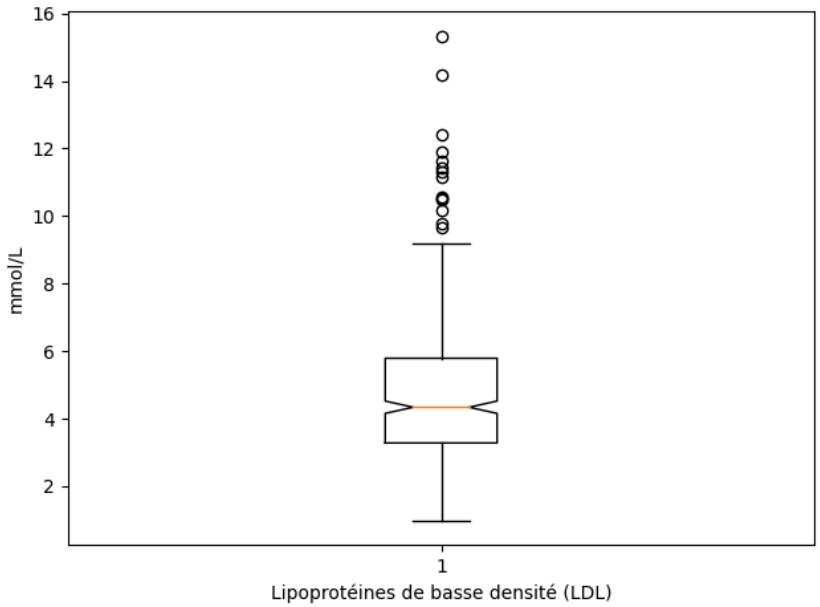

Figure 3. Diagramme en boîte représentant les différentes statistiques (médiane, quantiles, valeurs aberrantes) pour la mesure des lipoprotéines de basse densité (mmol/L) des individus de l'ensemble de données.

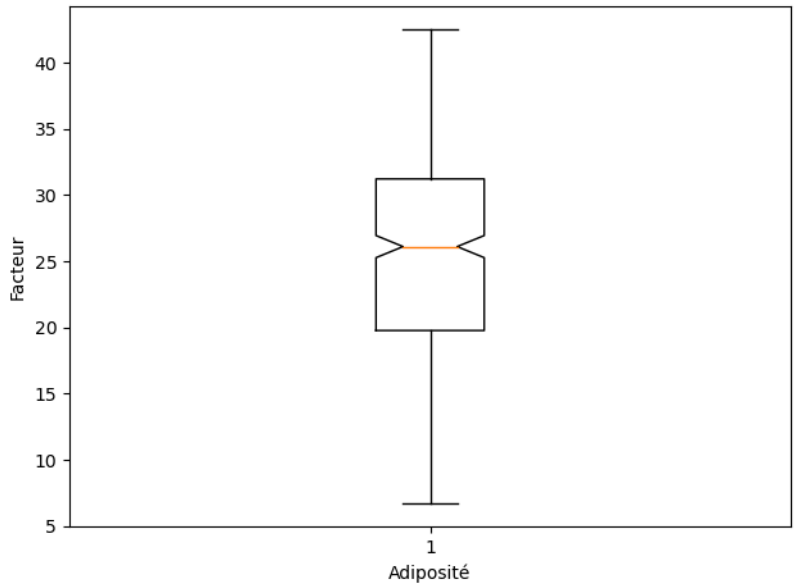

Figure 4. Diagramme en boîte représentant les différentes statistiques (médiane, quantiles, valeurs aberrantes) pour concernant l'adiposité (facteur) des individus de l'ensemble de données.

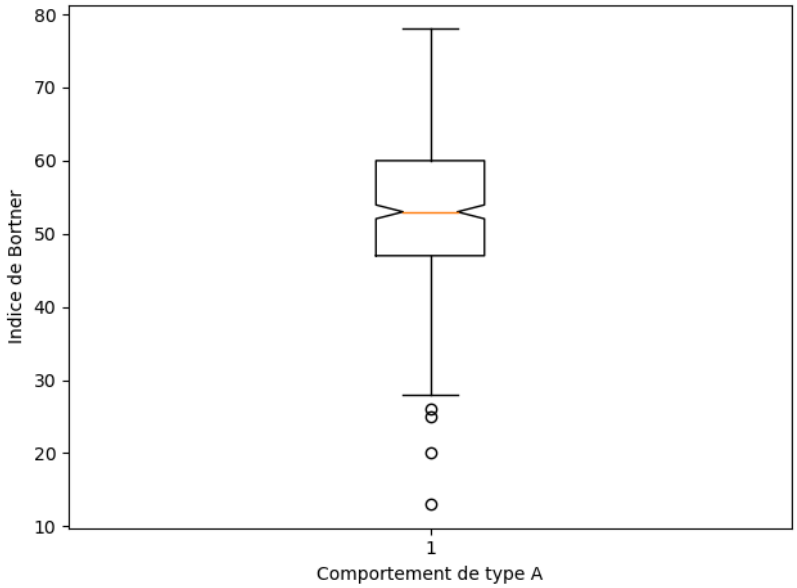

Figure 5. Diagramme en boîte représentant les différentes statistiques (médiane, quantiles, valeurs aberrantes) pour concernant les comportements de type A (selon l'indice de Bortner) des individus de l'ensemble de données.

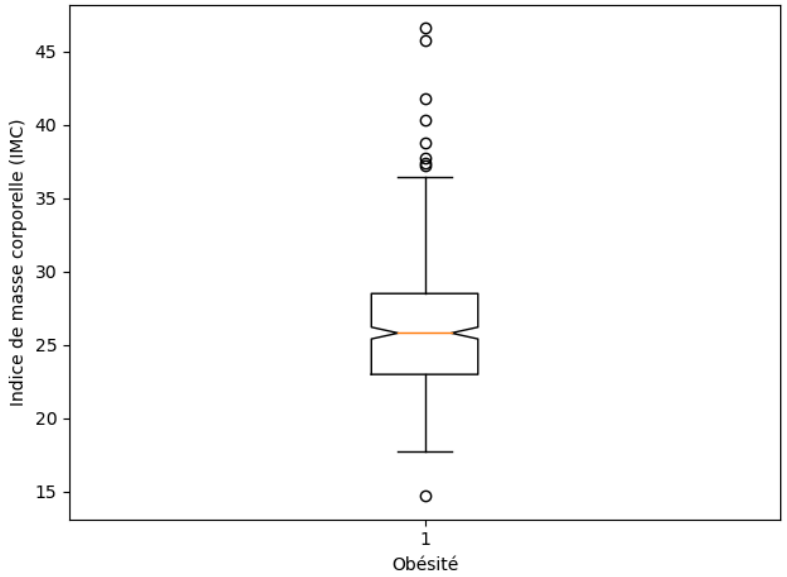

Figure 6. Diagramme en boîte représentant les différentes statistiques (médiane, quantiles, valeurs aberrantes) pour concernant l'obésité (selon l'indice de masse corporelle [IMC]) des individus de l'ensemble de données.

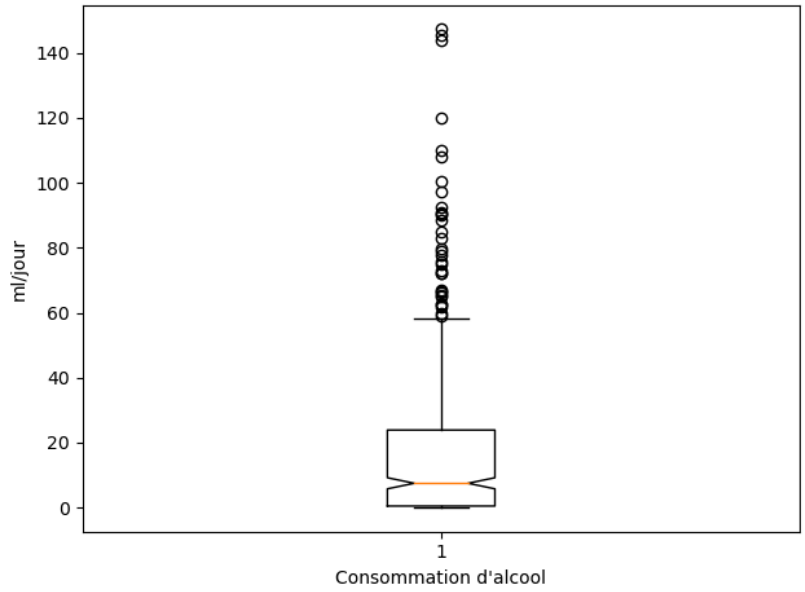

Figure 7. Diagramme en boîte représentant les différentes statistiques (médiane, quantiles, valeurs aberrantes) pour concernant la consommation d'alcool (ml/jour) des individus de l'ensemble de données.

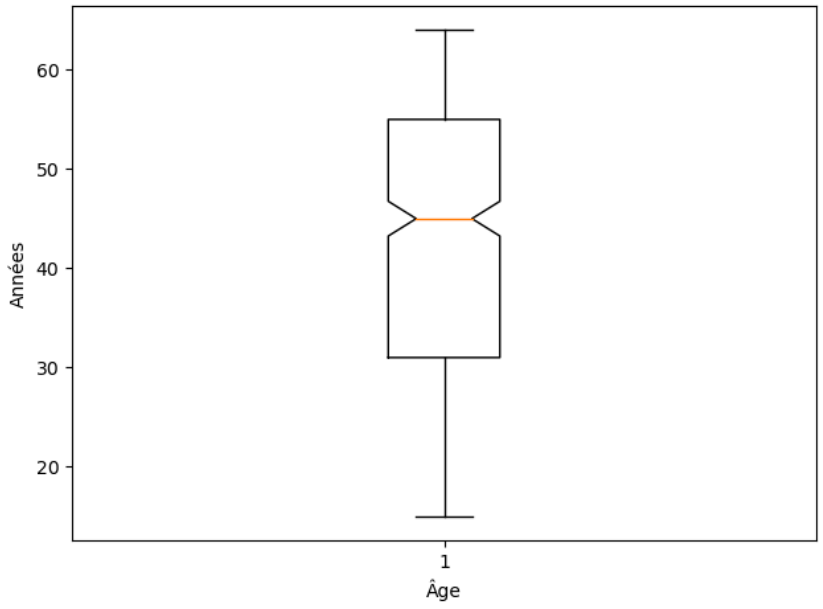

Figure 8. Diagramme en boîte représentant les différentes statistiques (médiane, quantiles, valeurs aberrantes) pour concernant l'âge (années) des individus de l'ensemble de données.

En ce qui concerne le prétraitement des données, il s'agissait ici d'ajuster le nombre d'instances pour les classes de maladie coronarienne à une valeur maximale. La méthode ici utilisée se nomme SMOTE, pour « synthetic minority over-sampling technique ». Le papier de Chawla *et al*. (2002) met en lumière cette technique de construction de classificateurs pour les ensembles de données possédant des classes non balancées. Le non-balancement des classes peut entraîner le classificateur à surévaluer le nombre d'instances faisant partie de la classe ayant initialement le plus d'instances ; il est donc essentiel d'ajuster cet élément de l'ensemble de données initiales afin d'augmenter la sensibilité de la classification (Chawla *et al.* 2002). La classe minoritaire de l'ensemble de données utilisé est ici les individus atteints de maladie coronarienne. L'algorithme a donc créé de nouvelles instances similaires à celles déjà dans cette classe afin de les ajouter à cette dernière, permettant d'obtenir un ensemble de données balancées (Chawla *et al.* 2002). Cette étape n'a été appliquée que sur les données d'entraînement, ce qui fait en sorte que dans chacune des séparations de données, ces dernières sont toujours balancées. Les données de tests n'ont ainsi pas été affectées par cette opération.

Il a par la suite été possible de normaliser les valeurs. Pour ce faire, la méthode de MinMaxScaler de Scikit Learn a permis de transformer les différentes valeurs des instances dans un intervalle de valeur, ici entre 0 et 1 (Scikit-learn 2020a). Cette méthode permet de modifier les valeurs afin qu'elles s'inscrivent dans un intervalle fixe et ainsi, d'éliminer l'effet de la moyenne (Bramer 2016). Cette étape-ci, comme pour la PCA, est appliquée initialement sur l'ensemble de données d'entraînement qui après vérification par l'algorithme, déterminera quelles sont les fonctions les plus appropriées afin d'effectuer la PCA et la normalisation. C'est ensuite, lorsque les données d'entraînement seront fournies à l'arbre de classification, que ces mêmes fonctions sélectionnées pour les données d'entraînement seront appliquées sur l'ensemble de tests : cette étape sera effectuée notamment pour que les données contenues dans l'ensemble de tests soient comparables avec les données d'entraînement fournies à l'arbre de classification.

Finalement, il a été possible d'appliquer une analyse des principales composantes (PCA) afin d'effectuer une réduction de dimensionnalité, grâce à la méthode PCA() de Scikit Learn provenant du module décomposition (Scikit-Learn 2020b). Cette étape est généralement appliquée, car elle permet de réduire la complexité du modèle, d'éviter le surapprentissage et

d’augmenter les performances (Li 2019). L’analyse des principales composantes est une méthode d’extraction de caractéristiques, qui permet de diminuer l’ensemble de données tout en conservant l’information la plus pertinente (Li 2019). L’application de l’analyse des principales composantes cherchera à identifier les directions de variance maximale au niveau des données possédant de grandes dimensions ; ceci permettra de générer un sous-ensemble avec des dimensions égales ou plus petites que celles de l’ensemble original (Li 2019). Au final, l’ensemble de données possédera un nombre réduit de colonnes grâce à une combinaison des variables d’origine (PCA).

3. ÉTUDE STATISTIQUE ET ANALYSES/INTERPRÉTATION DE CETTE ÉTUDE STATISTIQUE

Il s'agit, tout d'abord, d'effectuer une étude statistique pour chaque variable par rapport aux différentes classes, qui sont respectivement les individus atteints de maladies coronariennes et les individus non atteints. Il faut souligner que la variable dépendante correspond ici à l'état des classes, c'est-à-dire à la présence de maladies coronariennes ou à l'absence. Cette variable est d'ordre qualitatif. Les autres variables, autres que celle de l'historique familial, sont toutes des variables quantitatives, en plus d'être d'ordre indépendant. L'historique familial est aussi indépendant, mais elle demeure une variable qualitative. En considérant ces informations, il s'agira de se diriger vers une régression logistique. Cette dernière analyse statistique a pour but de vérifier les relations possibles de retrouver entre la variable dépendante (ou à expliquer), ici la présence ou l'absence de maladies coronariennes, et les variables indépendantes (explicatives) (Sanharawi et Naudet 2013), qui se trouvent à être la tension artérielle systolique, la consommation de tabac, le taux de LDL, l'adiposité, le comportement de type A, l'IMC, la consommation d'alcool, l'âge et l'historique familial. La variable dépendante de ce test se doit en effet d'être qualitative, alors que les variables explicatives peuvent être qualitatives ou quantitatives (Sanharawi et Naudet 2013). La régression logistique se présente comme un choix logique lorsqu'il s'agit de déterminer les facteurs de risque et les facteurs de protection physiopathologiques sous-jacents à une maladie (Sanharawi et Naudet 2013).

Ce test nécessite certaines conditions d'application au préalable. Il s'agira de vérifier la colinéarité entre les variables indépendantes et quantitatives en premier lieu (Sanharawi et Naudet 2013). Par la suite, la corrélation pourra être aussi vérifiée (Sanharawi et Naudet 2013). De plus, il faut généralement au moins 10 fois plus d'évènements positifs pour la variable dépendante (ici, au nombre de 160) que de variables explicatives (au nombre de 9) : ceci permet de souligner qu'il y a, dans l'ensemble de données, un nombre suffisant d'évènements pour la variable dépendante, permettant d'appliquer le modèle (Sanharawi et Naudet 2013).

En ce qui a trait à la colinéarité, ce test statistique vérifie la corrélation entre des variables indépendantes qui leur permet d'exprimer une relation linéaire dans un modèle de régression (Enders 2013). Si ces variables dépendantes sont corrélées, elles ne pourraient pas de façon

indépendante prédire la valeur de la variable à expliquer. Leur état de corrélation, dans ce cas-ci, diminuerait leur signification statistique (Enders 2013) : il deviendrait en effet difficile de distinguer l'effet de chacune des variables indépendantes sur la variable dépendante. Une des façons de vérifier la colinéarité est un test de multicolinéarité en utilisant le facteur VIF (« Variance Inflation Factor »), qui se calcul selon la règle suivante (GeeksforGeeks 2020) :

$$VIF = \frac{1}{1-R^2}$$

R^2 est ici le coefficient de détermination de régressions linéaires (GeeksforGeeks 2020). Le tableau 3 permet de confirmer que la plupart des VIF sont très grands. Comme ce facteur dépend de la corrélation entre les variables, il est possible de conclure qu'un plus grand VIF souligne une plus grande corrélation entre les variables indépendantes (GeeksforGeeks 2020). Considérant ses résultats, il a donc été décidé de poursuivre les analyses avec une régression logistique univariée pour chacune des variables indépendantes avec la variable dépendante.

Tableau 3. Résultats du calcul du facteur VIF pour les variables indépendantes

Variables	Tension artérielle systolique (mmHg)	Tabac (kg)	Cholestérol à lipoprotéines de basse densité (mmol/L)	Adiposité (facteur)	Historique familial (présent (1) ou absent (0))	Comportement de type A (indice de Bortner)	Obésité (IMC)	Alcool (ml/jour)	Âge (année)
VIF	39.559488	2.1001	7.955978	38.211907	1.847693	23.987961	68.510957	1.59273	20.375501

En ce qui a trait à la corrélation entre les variables indépendantes et la variable dépendante, il s'agira de sélectionner un test de corrélation qui permet d'utiliser une variable indépendante quantitative continue ainsi qu'une variable dépendante qualitative binaire. Il peut s'agir, en fait, de la corrélation bisérienne ponctuelle, qui répond à ces caractéristiques (SciPy 2014). Cette corrélation prend des valeurs entre -1 et +1 ; la valeur 0 indique qu'il n'existe pas de corrélation. Une corrélation de +1 indique une corrélation positive, tandis qu'une corrélation de -1 en indique une négative (SciPy 2014). L'hypothèse nulle dans le cadre de la corrélation utilisée ici présente est qu'il n'y a pas de corrélation existante entre les deux variables utilisées, tandis que l'hypothèse alternative stipule qu'il existe une corrélation.

En fonction des résultats obtenus dans le tableau 4, il est possible de remarquer que la plupart des coefficients se rapprochent de zéro ; ils indiquent tout de même une corrélation légèrement

positive. Conséquemment, et comme presque l'ensemble des $p < 0,05$, ce qui souligne que les résultats sont significatifs, il est possible de rejeter l'hypothèse nulle et d'accepter l'hypothèse alternative qui spécifie qu'il existe bel et bien une corrélation entre les variables indépendantes et la variable dépendante. Il faut par contre accepter l'hypothèse nulle dans le cas de l'alcool ($p > 0,05$), spécifiant qu'il n'existe pas de corrélation entre la consommation d'alcool et les maladies coronariennes au sein de notre ensemble de données.

Tableau 4. Corrélation bisérienne ponctuelle entre les variables indépendantes continues et la variable dépendante binaire

Corrélations		Coefficient (R)	P value
Tension artérielle systolique	**Maladies coronariennes**	0.19235411	3.15E-05
Tabac		0.299717538	4.82E-11
Cholestérol à lipoprotéines de basse densité		0.263052685	9.46E-09
Adiposité		0.254121391	3.05E-08
Comportement de type A		0.103155829	0.026612042
Obésité		0.100095075	0.031473168
Alcool		0.062530679	0.179687358
Âge		0.372973337	1.07E-16
Historique familial		0.272372726	6.58E-10

Considérant les résultats de colinéarité ainsi que les résultats de la corrélation bisérienne ponctuelle, il est possible de poursuivre les analyses avec la régression logistique univariée. Il s'agira ainsi de produire une régression pour chacune des variables indépendantes avec la variable dépendante (Sanharawi et Naudet 2013). L'hypothèse nulle est que les variables indépendantes n'ont pas n'effet sur les maladies coronariennes, ce qui se traduirait par une pente de zéro. L'hypothèse alternative est que l'augmentation des différentes variables indépendantes ont un effet sur les maladies coronariennes, ce qui se traduirait par une pente n'égalant pas zéro.

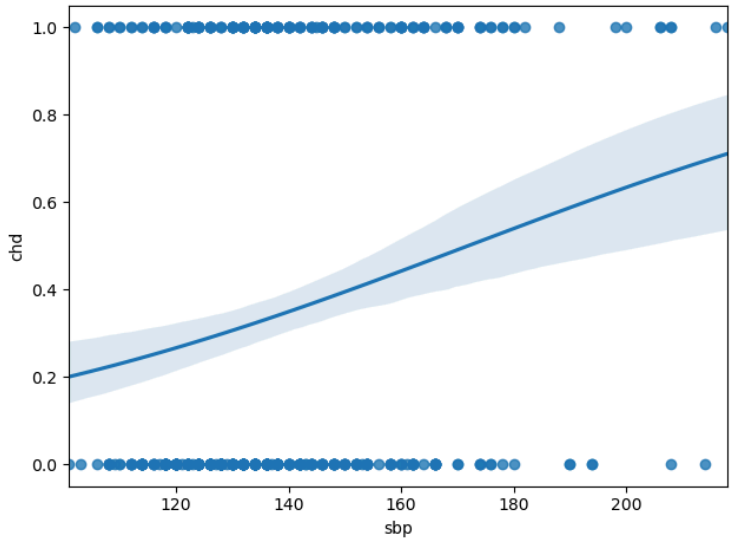

Figure 9. Régression logistique univariée entre la variable indépendante de la tension artérielle systolique (sbp) et la variable dépendante binaire de la présence ou de l'absence de maladies coronariennes (chd).

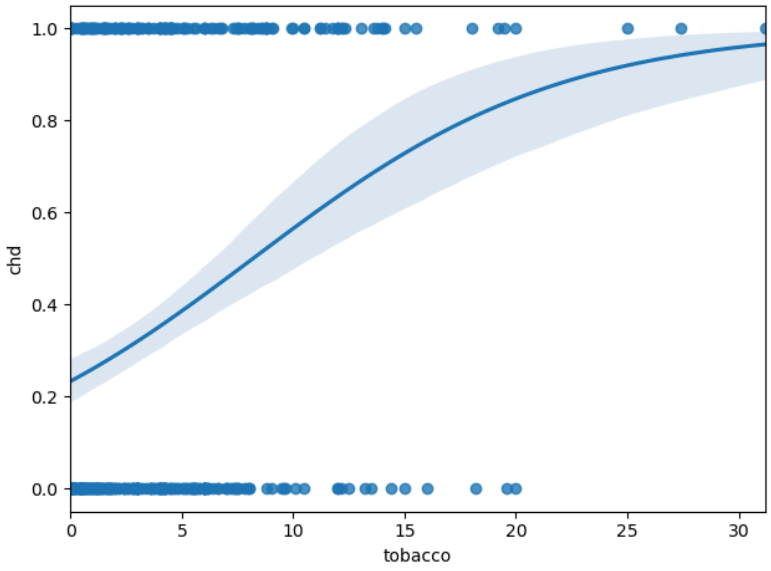

Figure 10. Régression logistique univariée entre la variable indépendante de consommation de tabac (tobacco) et la variable dépendante binaire de la présence ou de l'absence de maladies coronariennes (chd).

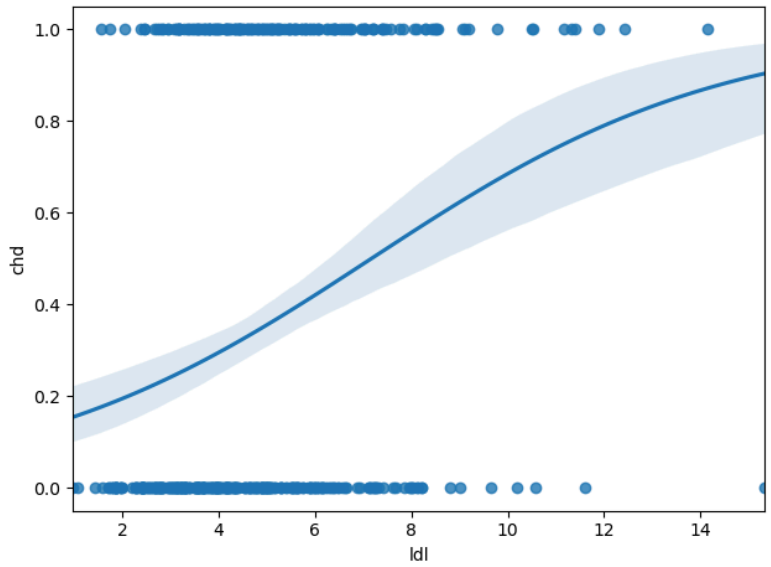

Figure 11. Régression logistique univariée entre la variable indépendante de la quantité de lipoprotéines de basse densité (LDL) et la variable dépendante binaire de la présence ou de l'absence de maladies coronariennes (chd).

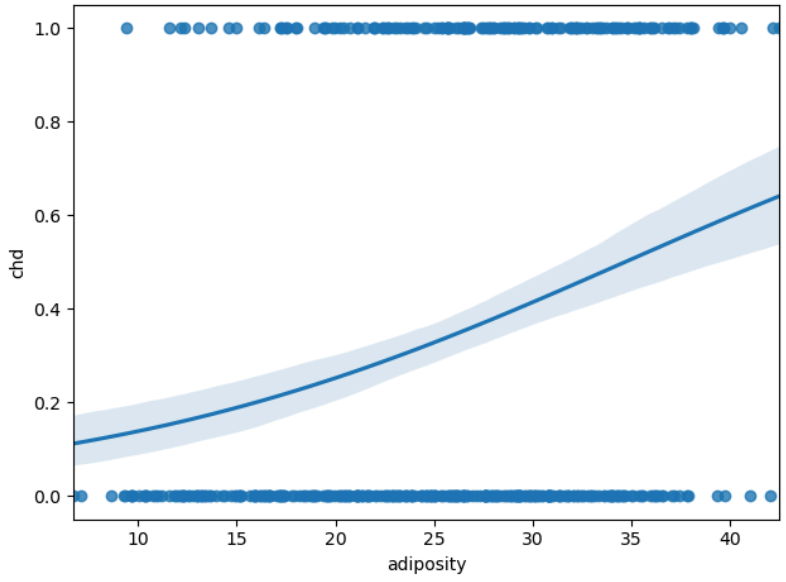

Figure 12. Régression logistique univariée entre la variable indépendante de l'adiposité (adiposity) et la variable dépendante binaire de la présence ou de l'absence de maladies coronariennes (chd).

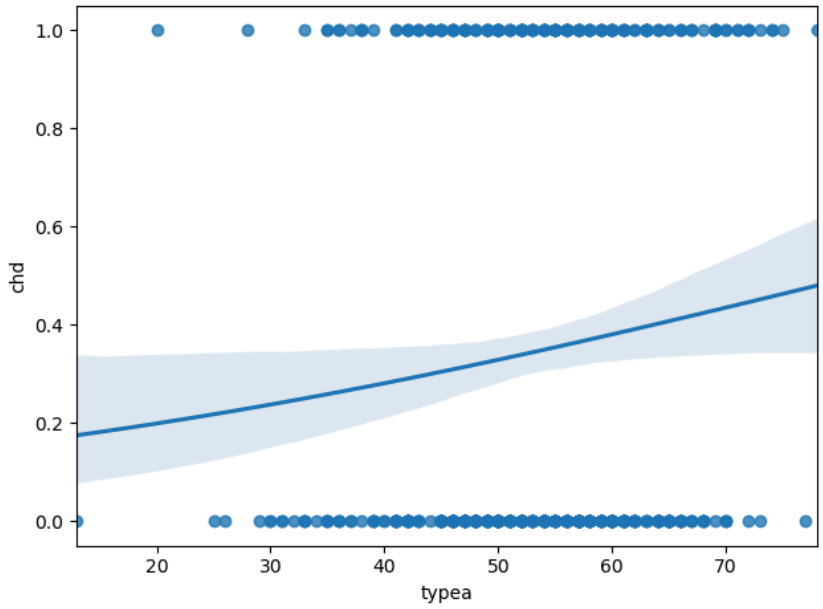

Figure 13. Régression logistique univariée entre la variable indépendante du comportement de type A (typea) et la variable dépendante binaire de la présence ou de l'absence de maladies coronariennes (chd).

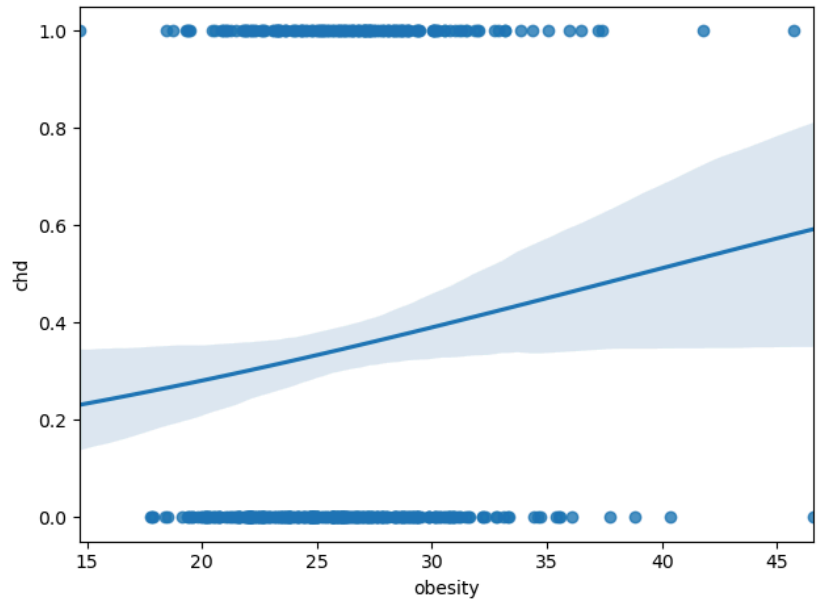

Figure 14. Régression logistique univariée entre la variable indépendante de l'obésité (obesity) et la variable dépendante binaire de la présence ou de l'absence de maladies coronariennes (chd).

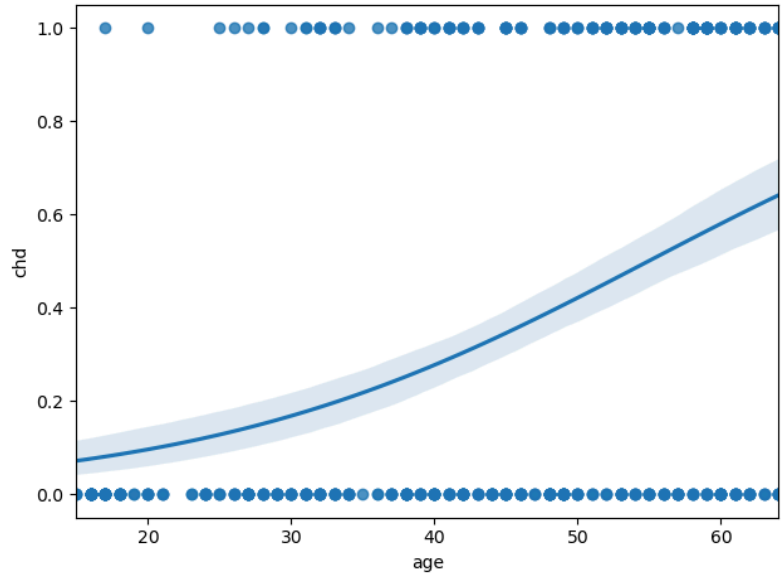

Figure 15. Régression logistique univariée entre la variable indépendante de l'âge (age) et la variable dépendante binaire de la présence ou de l'absence de maladies coronariennes (chd).

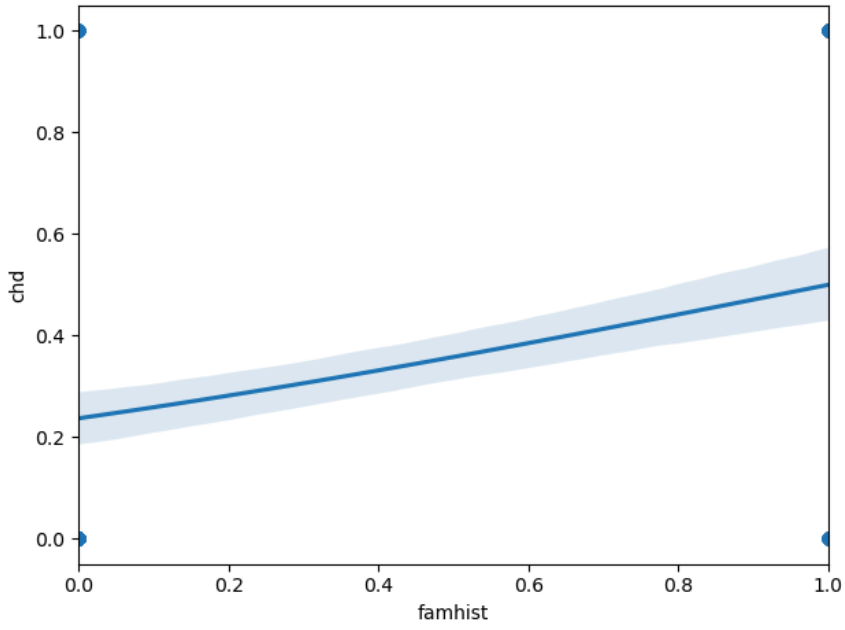

Figure 16. Régression logistique univariée entre la variable indépendante de l'historique familial (famhist) et la variable dépendante binaire de la présence ou de l'absence de maladies coronariennes (chd).

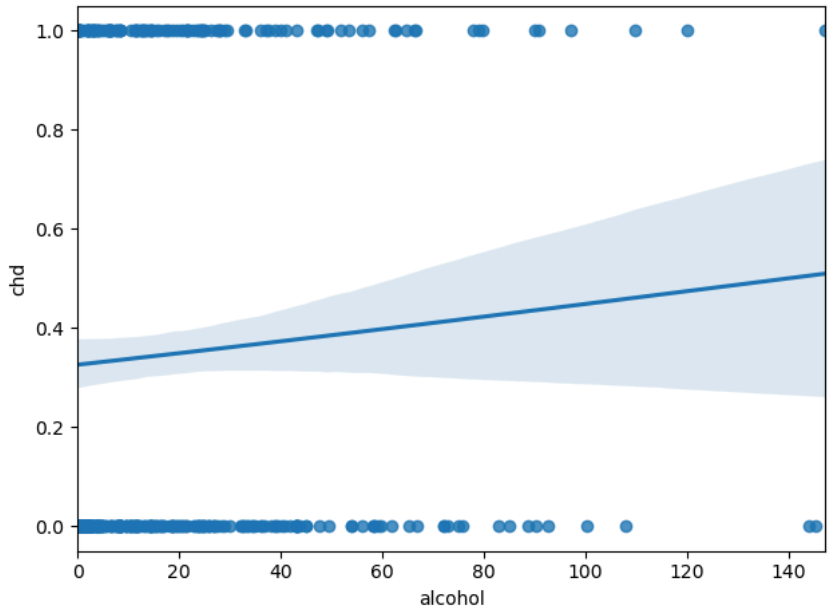

Figure 17. Régression logistique univariée entre la variable indépendante de la consommation d'alcool (alcohol) et la variable dépendante binaire de la présence ou de l'absence de maladies coronariennes (chd).

Le tableau 5 permet de vérifier le pourcentage d'explication des modèles de régression logistique grâce au coefficient de détermination. Dans l'ensemble, les modèles montrés aux figures 9 à 17 expliquent entre 63 et 70 % de la variation observée chez la présence ou l'absence de maladies coronariennes. La relation qui semble être la plus fortement expliquée par le modèle de régression logistique est l'utilisation de tabac (tableau 5) : il est donc possible d'émettre l'hypothèse qu'une utilisation de tabac influence l'apparition d'un phénotype de maladies coronariennes. Comme il existe une certaine variation au sain des pourcentages d'explication, il est possible de conclure que ce n'est pas tous les modèles qui expliquent au même degré la variation de la variable dépendante.

Il faut mentionner que malgré l'application de régression logistique (figures 9 à 17), cette dernière est une version améliorée d'une régression linéaire (Sucky 2020). Cette dernière formule, à des fins de simplification, sera utilisée pour l'interprétation : Sucky (2020) démontre d'ailleurs cette méthode. Les deux types de régression font partie des modèles linéaires généralisés (« Generalized Linear Models » [GLM]) (Pennsylvania State University 2018).

Il est possible de souligner que l'ensemble des pentes des régressions sont positives ; considérant que les pentes n'égalent pas zéro, il serait possible de rejeter l'hypothèse nulle et d'accepter l'hypothèse alternative qui stipulait que l'augmentation des variables indépendantes avait un effet sur l'apparition de maladies coronariennes. Cette affirmation est d'ailleurs confirmée par pratiquement l'entièreté des p values (ceux qui sont < 0,05), sauf pour l'alcool. Dans ce dernier cas, il est possible d'accepter l'hypothèse nulle et d'assumer que l'alcool ne semble pas influencer l'apparition de maladies coronariennes. Ceci corrobore les résultats obtenus, d'ailleurs, avec les corrélations bisérienne ponctuelles (tableau 4).

Tableau 5. Formule de la régression linéaire selon la formule de modèles linéaires généralisés, p-value et coefficient de détermination de la régression logistique des variables indépendantes en relation avec la variable dépendante

Variables	Formule de la régression linéaire (GLM)	P-value	Coefficient de détermination avec les maladies coronariennes
Tension artérielle systolique	y = 0,0195x - 3,3527	0,000	0.666666667
Tabac	y= 0,1453x - 1,1894	0,000	0.701298701
Cholestérol à lipoprotéines de basse densité	y = 0,2747x - 1,9687	0,000	0.666666667
Adiposité	y = 0,0741x - 2,5692	0,000	0.677489177
Comportement de type A	y = 0,0226x - 1,8447	0,027	0.653679654
Obésité	y = 0,0494x - 1,9283	0,033	0.653679654
Alcool	y = 0,0052x - 0,7260	0,182	0.651515152
Âge	y = 0,0641x - 3,5217	0,000	0.67965368
Historique familial	y = 1,1690 - 1,1690	0,000	0.653679654

Il est possible d'émettre l'hypothèse, face aux résultats obtenus suite aux analyses statistiques, que certains facteurs influenceront plus la classification que d'autres. Notamment, il pourrait être attendu que l'utilisation de tabac est une variable importante de catégorisation.

4. UNE ÉTUDE COMPARATIVE DE CLASSIFICATION

L'étude comparative de classification se divisera en deux sections. Dans un premier temps, une description de la méthodologie sera présentée. Bien que l'algorithme de classification nous ait été fourni pour ce projet, nous avons effectué plusieurs modifications et manipulations qui nécessitent une explication. Dans un deuxième temps, une analyse et une interprétation des résultats de l'étude comparative de classification seront effectuées.

4.1 Méthodologie de l'étude comparative

L'objectif de cette section du projet était de manipuler quelques paramètres du modèle de classification et ainsi d'analyser les changements dans les métriques de performance. Parmi les paramètres que nous avons modifiés, nous en retrouvons certains qui sont inhérents au modèle et d'autres qui y sont externes. Les deux tableaux suivants en font la liste exhaustive et répertorient les valeurs de ces paramètres que nous avons utilisées dans le cadre de ce projet :

Tableau 6. Paramètres inhérents au modèle d'arbre de décision et leurs valeurs possibles

Paramètres	Valeurs possibles
Critère de sélection de coupure	Gini ou entropie
Stratégie de coupure	La meilleure ou aléatoire
Nombre minimum d'instances par nœuds	2 à 200
Nombre minimum d'instances par feuille	1 à 100
Profondeur maximale	Aucune

Tableau 7. Paramètres externes aux modèles et leurs valeurs possibles

Paramètres	Valeurs possibles
Type de modèle	Machine à vecteurs de support (SVM) ou arbre de décision (DT)
Type d'estimation	Une séparation (splits) ou validation croisée (K-Fold)
Ratio test/total (dans le cas de « splits »)	10%, 20%, 30%, 40%, 50%, 60%, 70%, 80% ou 90%
Nombre de plis (dans le cas de « K-Fold »)	4, 6, 8 ou 10
Normalisation avec MinMaxScaler()	Oui ou Non
Nombre de composantes de l'analyse en composantes principales	3, 4 ou 5

Certains de ces paramètres ont été testés quelques fois manuellement, sans toutefois intégrer notre réelle étude comparative. C'est le cas du critère de sélection de coupure, de la stratégie de coupure et de la profondeur maximale de l'arbre de décision. En ce qui a trait au critère de sélection, nos résultats ne semblaient pas varier de manière assez significative pour l'intégrer dans notre étude. Nous avons donc conservé la valeur « gini » pour l'entièreté du projet. Quant à la stratégie de coupure, seulement quelques itérations ont été nécessaires pour nous convaincre de garder la valeur de ce paramètre à « meilleur » au lieu de « aléatoire ». Enfin, nous n'avons établi aucune limite quant à la profondeur de l'arbre, car nous la contrôlons indirectement avec le nombre d'instances minimum par nœud et par feuille.

En ce qui concerne le reste des paramètres, ils ont tous pris part à une étude comparative en deux temps. D'abord, nous avons effectué une analyse comparative des métriques de performance qu'avec les deux paramètres inhérents au modèle d'arbre de décision, c'est-à-dire le nombre d'instances minimum par nœud et par feuille. Le contrôle de ces paramètres porte le nom anglophone de pruning. Cette méthode limite la création de règles intéressantes qui peuvent subvenir lorsque l'arbre croît trop profondément ou de manière trop complexe (Patel et Upadhyay 2012). Pour accomplir cette méthode, une boucle itérative a répété à 100 reprises la classification avec l'étendue des valeurs possibles du nombre d'instances par nœud (de 2 à 200 et de 1 à 100, respectivement pour les nœuds et les feuilles). Les résultats de la variation des performances ont été enregistrées dans un DataFrame en format pickle. Ce dernier nous a permis de générer des graphiques afin de mieux interpréter les résultats qui seront présentés dans la section suivante.

Une fois le nombre d'instances minimum par nœud et par feuille étudiée et sélectionnée, nous avons effectué une recherche des meilleures combinaisons possibles de paramètres externes aux modèles. Dans le même ordre d'idée que le paragraphe précédent, nous avons construit une boucle itérative qui calcule les métriques de performance pour toutes les combinaisons possibles des paramètres du tableau 7. Encore une fois, les résultats ont été enregistrés dans un DataFrame en format pickle qui nous a servi à générer une multitude de graphiques représentant, sous différents angles, notre étude comparative.

4.2 Analyses et interprétation de l'étude comparative de classification

En premier lieu, avant d'entrer dans les détails de comparaison paramétrique, il serait pertinent d'observer et d'interpréter les résultats graphiques d'un seul arbre de décision. L'arbre qui a permis de générer les figures 18 à 21 est construit selon les paramètres présentés dans le tableau 8.

Tableau 8. Paramètres utilisés pour la démonstration de résultats d'un seul arbre de décision

Paramètres	Valeurs
Nombre minimum d'instances par nœuds	50
Nombre minimum d'instances par feuille	25
Type de modèle	Arbre de décision (DT)
Type d'estimation	Une séparation (splits)
Ratio test/total (dans le cas de « splits »)	20%
Normalisation avec MinMaxScaler()	Oui
Nombre de composantes de l'analyse en composantes principales	4

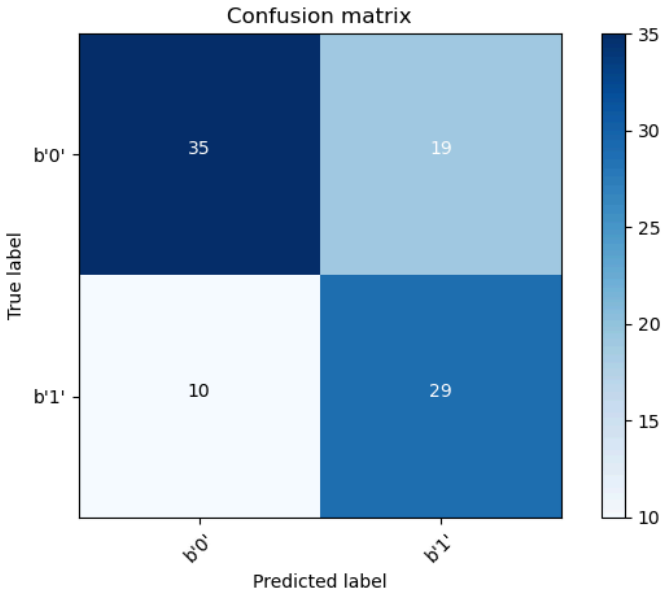

Figure 18. Matrice de confusion formant la relation entre les valeurs réelles et les valeurs prédites.

La figure 18 représente une matrice de confusion. Il s'agit d'un outil très important dans l'interprétation de modèles de classification. L'axe des ordonnées définit la valeur réelle de l'instance alors que l'axe des abscisses définit la valeur de prédiction de cette instance. Ainsi, la figure 18 nous véhicule l'information que l'arbre a classée 35 vrais positifs, 19 faux négatifs, 29 vrais négatifs et 10 faux positifs (dans ce cas-ci, un positif représente notre classe chd 0). C'est aussi depuis ces valeurs que sont calculées la majorité des métriques de performance, notamment le pourcentage d'instances correctement classées (accuracy), la sensibilité, la spécificité, la précision, le F-score et le Kappa de Cohen. De plus, l'importance qu'on accorde à certaines métriques dépendra du contexte. En effet, chaque métrique présente un aspect différent du modèle et celui-ci peut parfois s'avérer performant sur un et nettement moins sur un autre (Caruana et Niculescu-Mizil 2006).

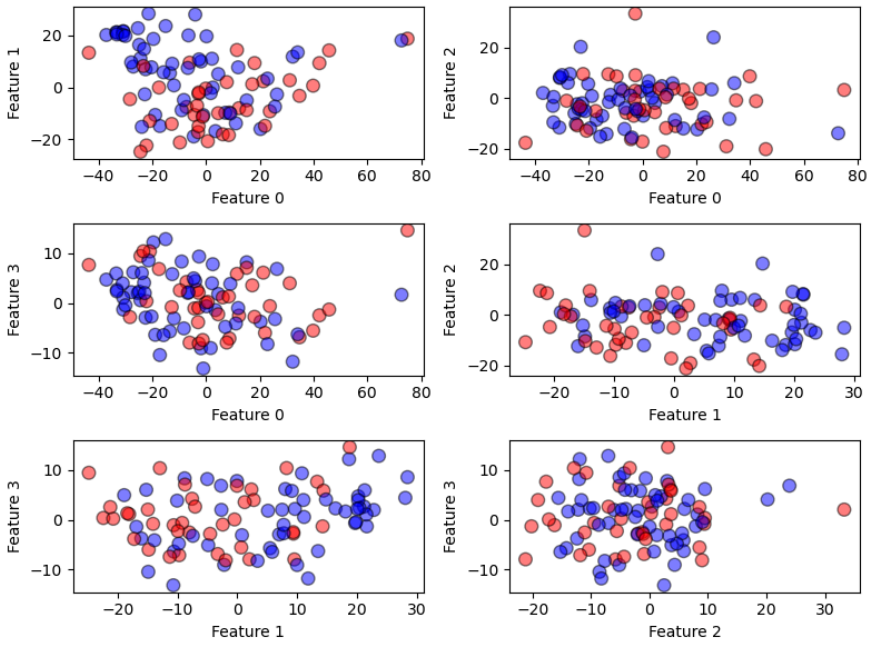

Figure 19. Rassemblement de graphiques en nuage de points représentant la relation entre les composantes principales.

La figure 19 représente simplement la distribution entre chacune des composantes principales qui sont issues de la fonction *PCA*. La couleur distingue les deux classes de la variable dépendante, c'est-à-dire la présence ou non de maladie cardiovasculaire. Une distinction spatiale claire entre des points d'une couleur et de l'autre se traduira probablement par un gain d'information important en termes de classification. Par exemple, dans la figure 19, il est notable que les instances rouges de la composante 1 ne dépassent pas un certain seuil (environ 10) alors que beaucoup d'instances bleues le font. Cette caractéristique sera sans doute importante dans la classification et apportera un gain d'information plus élevé que les autres.

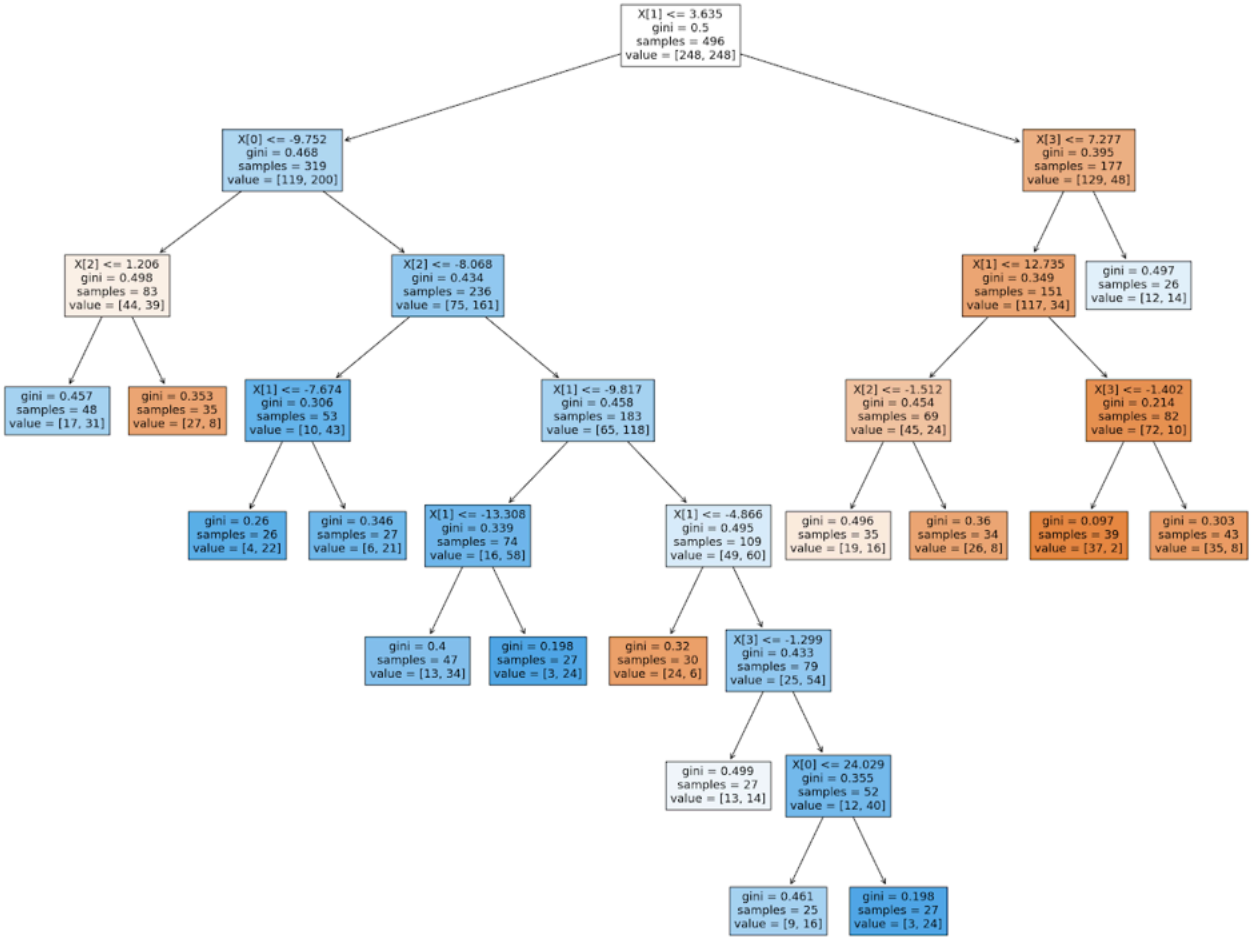

Figure 20. Visualisation de l'arbre de décision actuel.

La figure 20 nous permet de visualiser l'arbre de classification tel qu'il s'est formé à la suite de son entraînement. D'abord, les couleurs bleu et orange distinguent encore une fois nos deux classes de la variable dépendante. D'ailleurs, plus cette couleur est saturée, plus ce nœud ou cette feuille est constitué que d'une seule classe. En d'autres termes, cela veut dire que le nœud ayant créé une feuille (ou un autre nœud) saturée a établi une coupure bien performante. En outre, il est aussi possible de voir les composantes utilisées pour chaque coupure, le nombre d'instances par classe présent et l'indice *gini*, qui est une mesure de qualité de la coupure (à quel point le nœud a réussi à bien séparer les deux classes). Enfin, il est intéressant de noter, comme nous l'avons ciblé dans la figure précédente, que nous retrouvons la composante 1 comme étant la plus discriminatoire au niveau de la racine de l'arbre.

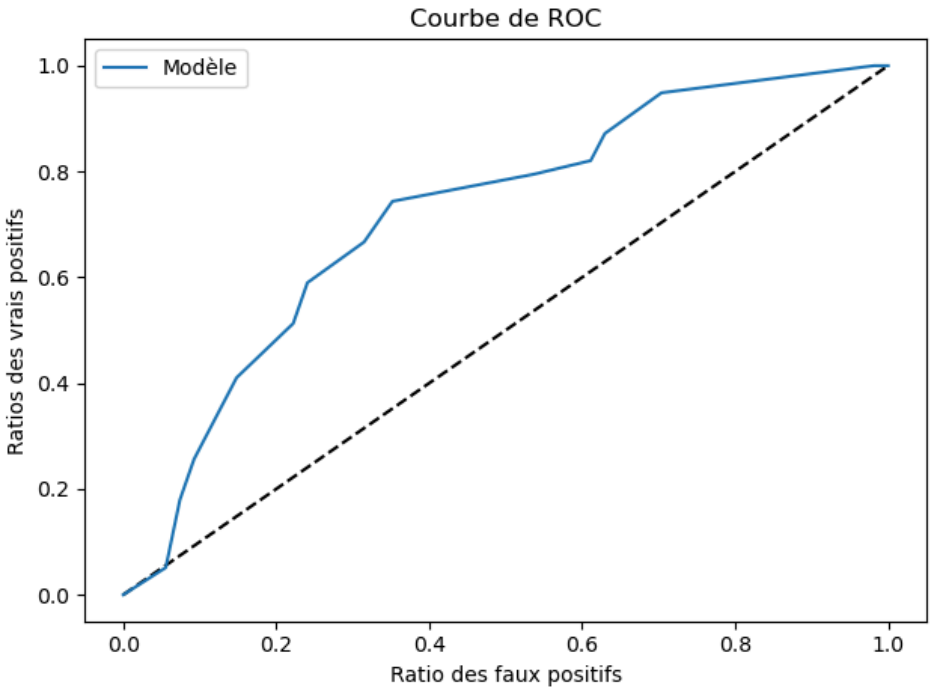

Figure 21. Courbe de caractéristique de fonctionnement du récepteur (ROC en anglais).

Enfin, la figure 21 trace la courbe de caractéristique de fonctionnement du récepteur (« Receiver operating characteristic ») de cet arbre de classification. Cette courbe est créée en observant le taux de vrais positifs et de faux négatifs en fonction de différents seuils de classification dans les feuilles de l'arbre (Fawcett 2006). Contrairement à un modèle de régression logistique, l'objectif de la courbe ROC dans un contexte d'arbre de décision n'est pas de trouver le seuil de classification optimal, car celui-ci reste par défaut 50 %. En revanche, ce graphique peut être une excellente métrique de performance, et ce en calculant l'aire sous cette courbe (*AUC*). Ainsi, la figure 21 démontre que notre modèle est relativement performant, car sa courbe est éloignée de la diagonale pointillée qui représente un modèle complètement aléatoire.

Après avoir bien saisi l'interprétation d'un arbre de décision, nous avons effectué une étude comparative uniquement sur le nombre d'instances par nœud et par feuille. Chaque itération de l'arbre de classification a été faite avec un nombre d'instances par nœud deux fois plus élevé que le nombre d'instances par feuilles, permettant à un nœud divisant deux groupes d'instances avec un ratio de 50 %/50 % de conserver ses deux feuilles. Les paramètres utilisés dans cette étude

comparative sont décrits à l'aide du tableau 9. Finalement, pour chaque valeur possible du nombre minimum d'instances par nœud/feuille, la moyenne de performance de 15 itérations sera calculée comme résultat de cette combinaison de paramètres.

Tableau 9. Paramètres et valeurs de l'étude comparative du nombre d'instances par nœud/feuille

Paramètres	Valeurs
Nombre minimum d'instances par nœuds	2 à 200
Nombre minimum d'instances par feuille	1 à 100
Type de modèle	Arbre de décision (DT)
Type d'estimation	Une séparation (splits)
Ratio test/total (dans le cas de « splits »)	20%
Normalisation avec MinMaxScaler()	Oui
Nombre de composantes de l'analyse en composantes principales	4

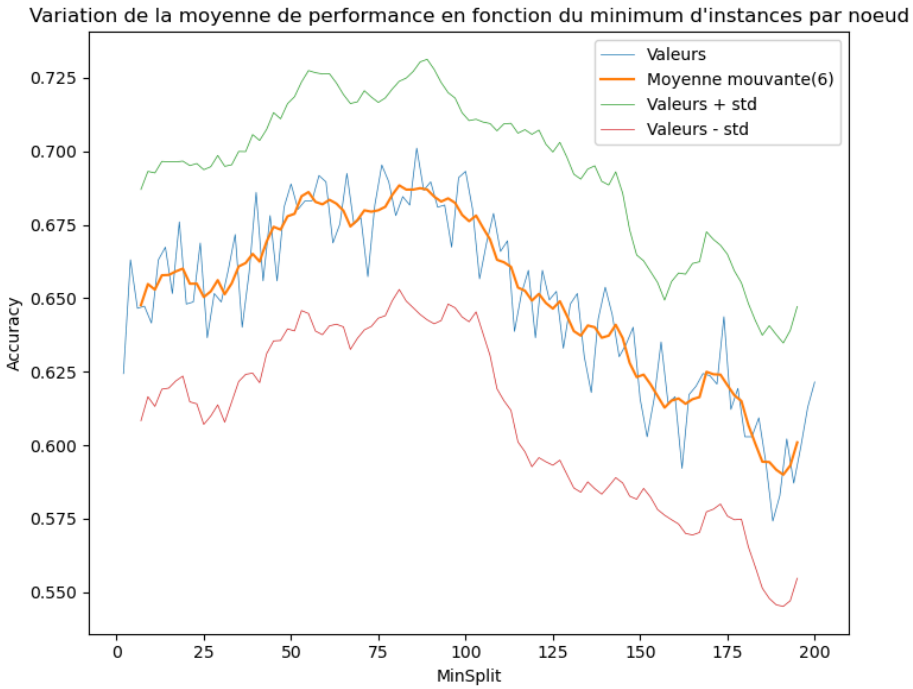

Figure 22. Variation de la moyenne de performance en fonction du minimum d'instances par nœud.

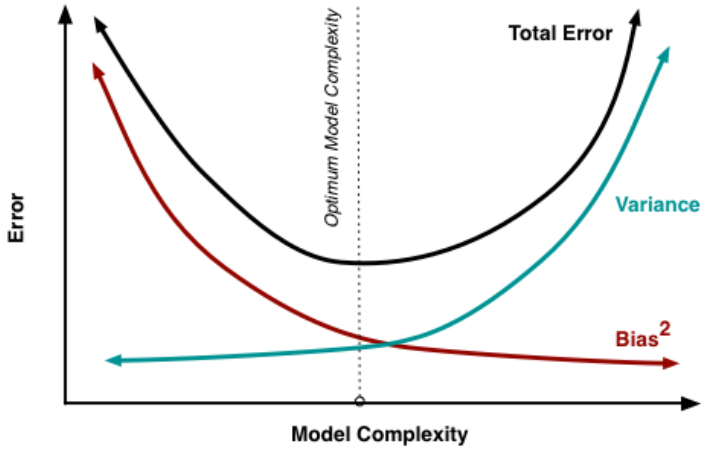

Figure 23. Réduction optimisée de la quantité d'erreurs par le contrôle du biais et de la variance (Fortmann-Roe 2012).

L'axe des ordonnées de la figure 22 mesure la performance de chaque itération alors que l'axe des abscisses fait l'état du nombre d'instances minimales par nœud. Les résultats sont tracés

exactement à l'aide de la ligne bleue et plus grossièrement avec la ligne rouge (moyenne mouvante). Les lignes orange et vertes représentent l'addition et la soustraction de l'écart-type des 15 résultats par valeur possible de *MinSplit* (axe des X). Hypothétiquement, ce graphique est une belle représentation du phénomène de la variance et du biais, autrement nommé sous-apprentissage et surapprentissage (fig. 23). En effet, lorsque le nombre d'instances minimum par nœud est très bas, l'arbre se développera d'une manière très complexe. En d'autres mots, il épousera les détails précis de l'ensemble de données d'entraînement, dont les règles apprises pourraient fort probablement ne pas s'appliquer à de nouvelles données. Donc il est possible de voir que, dans un premier temps, plus la variable *MinSplit* augmente, plus la performance augmente. Ensuite, un seuil est atteint et lorsque la variable *MinSplit* dépasse 90 instances minimum par nœud, la performance recommence à chuter, cette fois à cause du phénomène de variance qui signifie que l'arbre devient trop général dans sa classification. Enfin, nous avons décidé de poursuivre l'étude avec un nombre minimum d'instances par nœud et par feuille de, respectivement, 60 et 30.

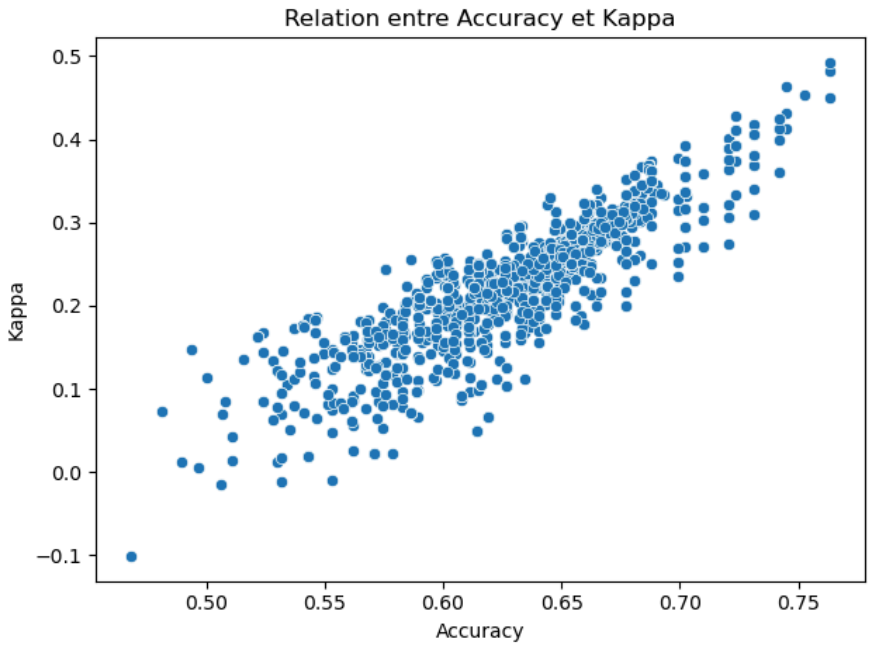

Figure 24. Relation entre les métriques Accuracy et Kappa.

Comme expliqué dans la méthodologie, nous avons ensuite effectué une comparaison de performance entre toutes les combinaisons possibles de paramètres externes au modèle (tableau 7). Cela a généré 156 différentes itérations de classification. Comme métrique de performance, nous avons préservé le pourcentage d'instances correctement classées (*accuracy*) et le Kappa de Cohen. La différence entre ces mesures de performance est que le taux de classification prend en compte le succès de classification de toutes les classes simultanément, alors que le Kappa de Cohen se calcule indépendamment pour chacune des classes et agrège les résultats (García *et al.* 2009). Ainsi, le Kappa est moins sensible au changement aléatoire causé par un débalancement d'instances par classe (García *et al.* 2009). Cependant, nous avons vérifié la corrélation de ces deux métriques avec un test de corrélation de Pearson et le coefficient semble assez élevé ($\rho = 0.85$). En conséquence, nous allons performer les analyses principalement avec le taux de classement général. Leur relation linéaire peut être visualisée à l'aide de la figure 24.

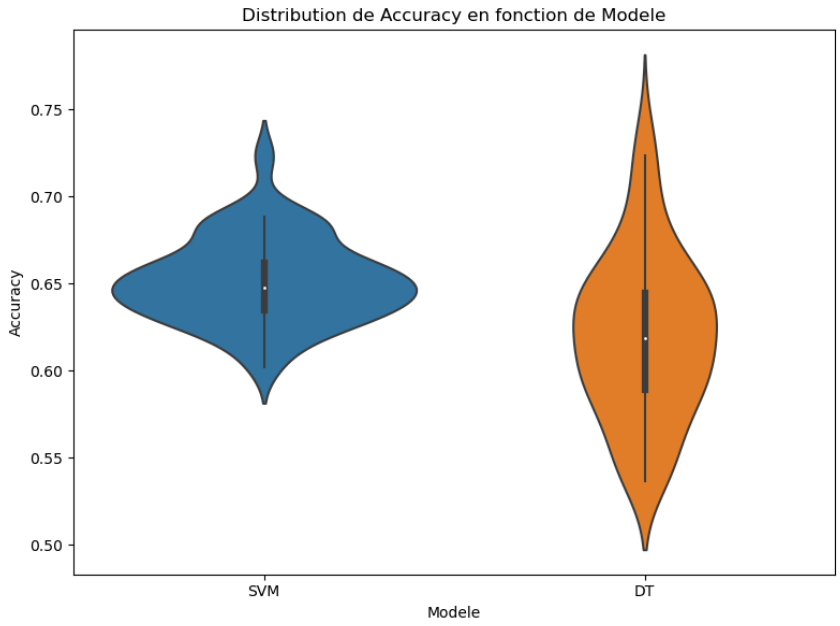

Figure 25. Distribution de performance en fonction du modèle (SVM ou DT).

La première comparaison démontrée à la figure 25 met en lumière les différences de distributions de performance en fonction du modèle choisi, soit machine à vecteurs de support (SVM) ou arbre de décision (DT). Par ailleurs, il est important de noter que toutes les variations des autres

paramètres sont présentes dans ce graphique. Ce dernier graphique dévoile donc quelques distinctions entre ces deux modèles. En premier lieu, la médiane de SVM est plus élevée que celle de DT. En revanche, le maximum de DT est plus élevé. En effet, la performance de DT varie davantage que celle de SVM. Nous pourrions potentiellement en conclure que les arbres de décisions sont plus sensibles à certains paramètres que les machines à vecteurs de support ou bien que les arbres de décisions tendent à varier davantage d'une itération à l'autre.

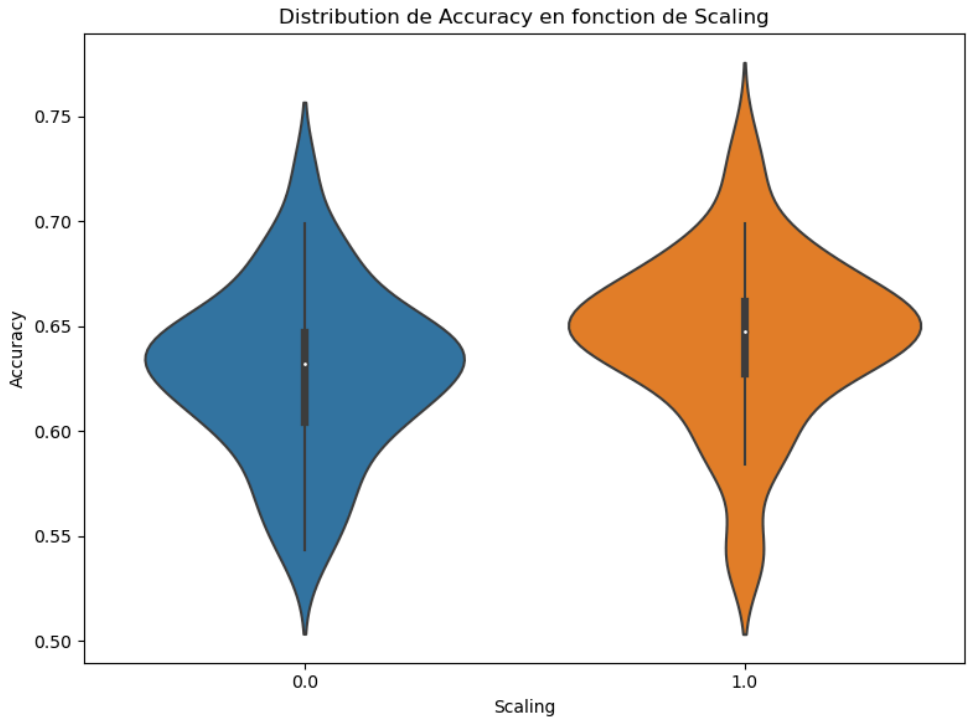

Figure 26. Distribution de la performance en fonction de la normalisation.

Brièvement, la figure 26 permet de constater que la normalisation effectuée avec la fonction MinMaxScaler() augmente en général les performances de notre modèle. En revanche, cette différence n'est pas significativement élevée. Nos données étaient peut-être, dès le départ, sur une échelle relativement similaire.

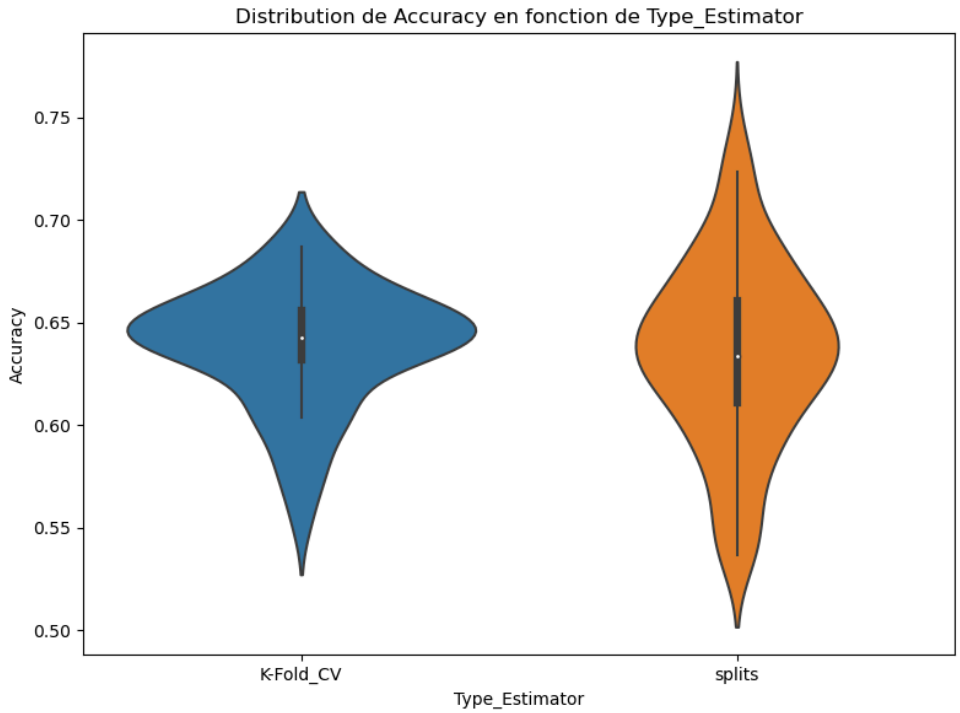

Figure 27. Distribution de la performance en fonction du type d'estimation.

Cette fois, la figure 27 représente la distribution de la performance en fonction du type d'estimation du modèle. Sans surprise, la validation croisée (K-Fold_CV) est nettement plus stable en performance que les séparations uniques variant entre 10 % et 90 %. Cela pourrait être dû à deux phénomènes. D'une part, la validation croisée utilisait des ratios de test différents que les séparations uniques. En effet, les différentes itérations faites avec la validation croisée avaient un nombre de plis de 10, 8, 6, et 4. Ainsi, cela se traduit à un ratio de test d'environ 10 %, 12 %, 16 % et 25 %. D'autre part, la nature de la validation croisée est de tester plusieurs fois l'ensemble d'entraînement avec des tranches différentes de celui-ci et de reconstituer une performance globale avec une moyenne des résultats par tranche. Nécessairement, cette opération aura tendance à stabiliser la performance et de fournir un résultat plus représentatif de l'ensemble de données.

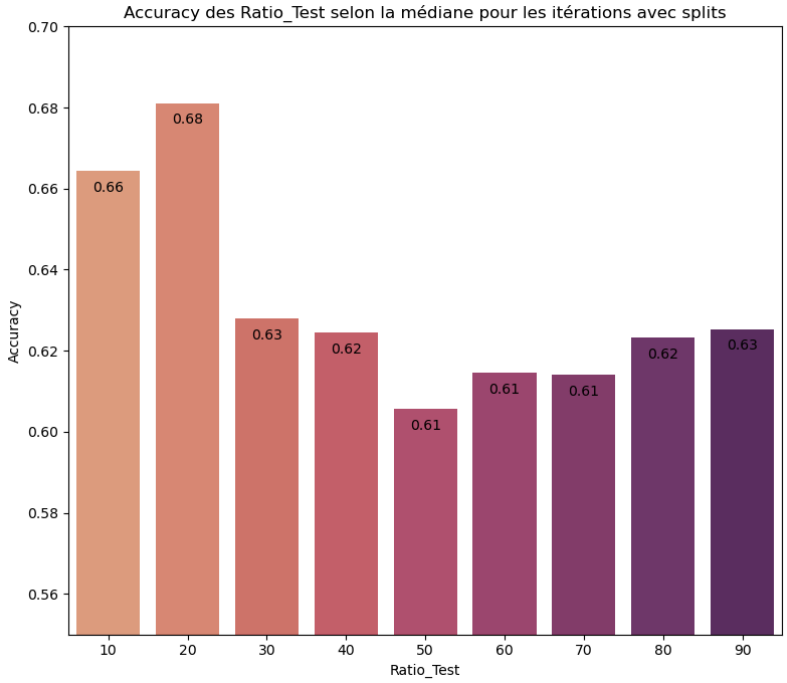

Figure 28. Classement des ratios de test en fonction d'accuracy.

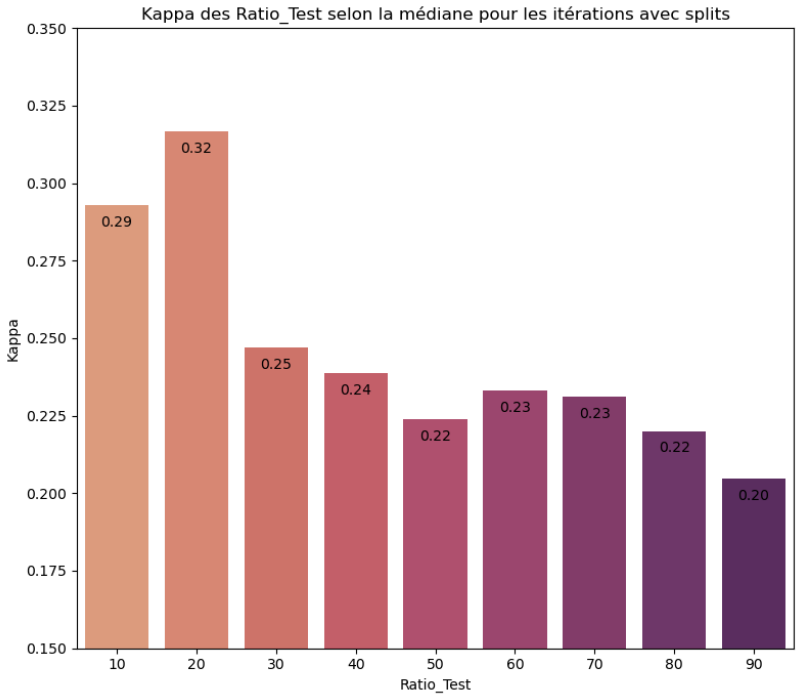

Figure 29. Classement des ratios de test en fonction de Kappa.

Par ailleurs, les figures 28 et 29 mettent en lumière le classement des ratios de données de test en fonction de l'*accuracy* et du Kappa de Cohen. Dans les deux cas, les ratios les plus performants sont 10 % et 20 %, et ce par une marge assez grande. En effet, la différence entre le ratio de 20 % et de 30 % est pour le moins curieuse. Cela a peut-être un lien avec quelques instances aux valeurs bien distinctes qui se sont retrouvées dans l'ensemble de tests pour le ratio de 30 % alors qu'elles étaient dans celui d'entraînement pour le ratio de 20 %. Dans un autre ordre d'idées, la comparaison entre le taux de classement général (*accuracy*) et le Kappa de Cohen est dans ce cas-ci porteuse d'informations. C'est-à-dire qu'une divergence est présente entre ces deux métriques pour les ratios de test dépassant 70 %. Une interprétation possible serait que malgré un très grand ratio de test (donc un petit ratio d'entraînement), l'arbre de décision est tout de même apte à classer adéquatement une des deux classes, mais beaucoup moins pour l'autre. En effet, cette situation se traduit par un taux de classement relativement standard, mais avec un Kappa de Cohen bien plus faible. C'est probablement pour cela que, par exemple, le ratio de test de 90 % s'en sort avec une *accuracy* de 63 % et avec un Kappa, nettement plus faible que les autres ratios, de 0,2.

Finalement, il serait tout de même pertinent d'observer les paramètres de la meilleure itération de notre étude comparative. Le tableau 10 en fait la présentation.

Tableau 10. Valeurs des paramètres de l'itération avec la meilleure Accuracy

Paramètres	Valeurs
Accuracy	0.763441
Kappa	0.49256
MinSplit	60
Modele	Arbre de décision (DT)
NbCompPCA	4
Ratio_Test	20
Scaling	Oui
Type_Estimator	splits

5. CONCLUSION

Au niveau de la description de l'ensemble de données, il a été possible d'établir que celui utilisé est en fait un sous-ensemble d'une étude plus grande (CORIS) sur les facteurs présumés comme ayant une influence sur le développement des maladies coronariennes. Les variables incluaient des facteurs comme la tension artérielle systolique, la consommation de tabac et d'alcool, le taux de LDL dans le sang, l'adiposité et l'IMC, les comportements de type A ainsi que l'âge et l'historique familial de maladies coronariennes. L'ensemble de données avait deux classes, soit la présence ou l'absence de maladies coronariennes : les instances correspondaient, pour chacune, aux données recueillies chez un individu. La vérification des données a permis de conclure que les variables présentaient une proportion d'individus sujets au développement de maladies coronariennes, de par leurs habitudes de vie (grande consommation d'alcool ou de tabac), ou encore de par leurs caractéristiques inhérentes (tension artérielle systolique élevée, taux de LDL élevé, grande adiposité, grand IMC, etc.). Au niveau du prétraitement, il a été possible d'ajuster le nombre d'instances de l'ensemble d'entraînement avec la méthode SMOTE, et de normaliser et d'appliquer une PCA sur l'ensemble d'entraînement puis sur l'ensemble de tests, en fonction de ce qui avait été appliqué sur l'ensemble d'entraînement.

Considérant que la variable de la classe étudiée étant dépendante et qualitative et que les autres variables étaient indépendantes et, dans la majorité, quantitatives, il a été possible de sélectionner la régression logistique comme test statistique. Il s'agissait au préalable de vérifier la colinéarité des variables ainsi que la corrélation bisérienne ponctuelle. Le test de la multicolinéarité a permis de vérifier que la plupart des VIFs étaient très grands, ce qui soulignait une grande corrélation entre les variables et pointait vers l'utilisation d'une régression logistique univariée, et ce, afin que chacune des influences des variables indépendantes sur la variable dépendante soit distinguable. La corrélation bisérienne ponctuelle permit de confirmer une corrélation légèrement positive et significative entre les variables indépendantes et la variable dépendante, sauf pour l'alcool. Les régressions logistiques effectuées par la suite avaient toutes des pentes positives, et les modèles expliquaient une bonne partie de la variabilité des résultats (entre 63 % et 70 %) en plus d'être significatifs (sauf pour l'alcool, ce qui était attendu). Le tabac est ainsi attendu comme étant une variable importante dans la catégorisation, en fonction des résultats de la régression.

Il s'agissait ensuite de produire une étude comparative de classification. Au niveau de la méthodologie, la valeur de gini et le paramètre meilleur de la stratégie de coupure ont été sélectionnés, sans limites de la profondeur de l'arbre. Pour le reste des paramètres, il a fallu effectuer une étude comparative en, tout d'abord, contrôlant les paramètres par la méthode de *pruning* pour sélectionner le nombre d'instances minimum par nœud et par feuille et, ensuite, en recherchant les meilleures combinaisons possibles de paramètres externes au modèle. En ce qui a trait aux analyses et aux interprétations de l'étude comparative de classification, nous avons d'abord présenté un seul arbre de décision en fonction des paramètres utilisés, ainsi qu'une matrice de confusion. La figure 19, c'est-à-dire la représentation graphique de la PCA, a permis de visualiser que certaines instances semblaient bien se distinguer, ce qui a nécessairement un impact sur la classification. La visualisation de l'arbre de classification a été possible à la figure 20. Finalement, la figure 21 permettait de visualiser les performances du modèle qui, somme toute, semble être relativement performant dans le cadre de cette génération.

Une étude comparative sur le nombre d'instances par nœud et par feuille, selon les paramètres inscrits au tableau 9, a été effectuée. Un nombre d'instances minimum par nœud très bas engendrait le développement d'un arbre très complexe, et rapprochait le modèle du surentraînement. Lorsque le minsplit augmentait, et ce, jusqu'à un minimum de 90 instances, les performances du modèle augmentaient ; après 90 instances, les performances étaient en diminution. Il a été décidé de poursuivre l'étude avec un nombre minimum d'instances par nœud et par feuille de 60 et de 30, de manière respective. Il a été décidé de poursuivre les analyses principalement avec le taux de classement général suite à la visualisation des métriques de performance de l'*accuracy* et de l'indice Kappa. Il a été possible de déterminer que les arbres de décision sont plus sensibles à certains paramètres que les machines à vecteur de support, ou que les arbres de décision ont tendance à varier davantage d'une itération à l'autre. La normalisation par MinMaxScaller(), quant à elle, augmentait les performances du modèle. La validation croisée était plus stable, au niveau des performances, que les séparations uniques entrent 10 et 90 %. Il a aussi été possible de déterminer que l'arbre serait possiblement capable de classer, de façon plus adéquate, seulement l'une des deux classes après un ratio de test de 70 %.

Quelques éléments pertinents ont été remarqués vers la fin du projet, et ce en ce qui concerne les limites et les faiblesses de cette étude. En ayant les ressources temporelles disponibles, ces limites auraient bien entendu été adressées. En premier lieu, les résultats de classification ont une variance relativement élevée. En effet, de nombreuses itérations de classification produisent des métriques de performance variant d'environ 15%, et ce en conservant exactement les mêmes paramètres. Une tentative de résolution de ce problème a été effectuée en contrôlant toutes les fonctions comprenant un élément aléatoire avec un état aléatoire spécifique (*random_state*), mais en vain, cela n'a pas réglé la problématique. En outre, le principe de la validation croisée aurait pu être appliqué différemment. C'est-à-dire que la recherche de paramètres optimaux aurait pu s'effectuer uniquement sur un ensemble d'entraînement unique dans lequel une validation croisée serait appliquée. Une fois le paramétrage complété, un test final sur l'ensemble de test, jusqu'à lors intouché, serait conduit. De cette manière, l'ensemble de test n'aurait jamais influencé la démarche de modélisation, rendant la méthodologie plus représentative d'une problématique réelle. Par ailleurs, comme expliquée dans la section de la méthodologie, la recherche de paramètres s'est effectuée en deux temps, soit pour le *pruning* et ensuite pour les paramètres externes à l'arbre. Cependant, cette technique aurait pu être améliorée et être fusionnée pour vérifier l'ensemble des combinaisons possibles. Certaines fonctions déjà existantes performent ce type de tâches, telle *sklearn.model_selection.GridSearchCV*. Finalement, la variable *famhist*, variable indépendante binaire, aurait peut-être dû être manipulée autrement. En effet, dans le cadre de cette étude, nous avons traité la variable *famhist* comme les autres variables indépendantes quantitatives. Cependant, les ingénieries de caractéristiques appliquées à ces variables (changement d'échelle et réduction de dimensionnalité) ne s'appliquaient peut-être pas de la même manière à une variable binaire. Bien que ces éléments exposent les faiblesses du travail, il est clair que cet exercice de réflexion est nécessaire et contribue à l'amélioration de la méthodologie pour de futurs travaux.

6. CONCLUSION GÉNÉRALE

De manière générale, les données étaient déjà très propres, ce qui a été apprécié. En effet, ces dernières avaient déjà été utilisées dans le cadre d'études et d'analyses préalables, ce qui a probablement eu une influence sur leur propreté. Les différentes méthodes apprises dans le cadre du projet ainsi que les connaissances acquises en sont des précieuses, puisqu'elles serviront effectivement dans un avenir universitaire, mais aussi dans nos emplois futurs. Ces outils sont indispensables aux connaissances des scientifiques des données et il est apprécié que cela nous soit montré rapidement dans notre parcours universitaire. En ce qui a trait aux personnes ayant suivi le cours sans base relative au sujet de la programmation, la pente d'apprentissage a été assez rude. Toutefois, comme mentionnées, ces connaissances sont importantes et serviront probablement dans un avenir rapproché, et ce, même hors du cadre de la formation de scientifique des données. Le projet était d'envergure, mais l'ouverture du professeur fasse à l'allongement du temps fût la bienvenue.

L'ensemble du projet était agréable. Celui-ci nous a permis de tirer des conclusions claires concernant les risques de maladie cardiovasculaire ainsi que de corroborer les études sur le sujet (Fung *et al.* 2001; Hu 2003; Perdesen 2006; Srour *et al.* 2019). De plus, les études statistique et comparative nous on fait découvrir une pléiade de nouvelles fonctions dans python. Ce sont grâce à ces fonctions que notre projet a pu tirer des résultats concluants. En suivant les résultats du projet, il est donc clair que pour réduire les risques de maladie cardiovasculaire, il est impératif de réduire les facteurs de risque étudiés (Fung *et al.* 2001; Hu 2003; Perdesen 2006; Srour *et al.* 2019). D'un point de vue sociétal, il est important de comprendre les causes sous-jacentes des maladies cardiovasculaires afin de les prévenir. Cela est d'autant plus important sachant que celle-ci est la première cause de mort dans le monde selon l'OMS (2015). Il s'agit d'un enjeu mondial qui pourrait sauver la vie de millions d'êtres humains dans le futur : pour se faire, nous devons donc prendre conscience des facteurs de risque afin de les réduire. C'est pourquoi notre projet avait une certaine utilité, puisqu'il a permis de corroborer les études modernes et la méthode de classification, si améliorée, pourrait être utilisée d'un point de vue de la catégorisation des patients des facteurs de risque cardiaques.

7. RÉFÉRENCES

Bramer M. 2016. Principles of data mining. Springer, 425 p.

Caruana R et Niculescu-Mizil A. 2006. An empirical comparison of supervised learning algorithms. Proceedings of the 23rd international conference on machine learning : 161-168.

Chawla NV, Bowyer KW, Hall LO et Kegelmeyer WP. 2002. SMOTE : synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16 : 321-357.

Enders FB. 2013. Collinearity. Consulté le 2020-12-13, https://www.britannica.com/topic/collinearity-statistics

Fawcett T. 2006. An introduction to ROC analysis. Pattern Recognition Letters, 27 : 861-874.

Foerster M, Marques-Vidal P, Waeber G, Vollenweider P et Rodondi N. 2010. Association entre consomation d'alcool et facteurs de risque cardiovasculaire : une étude sur la population lausannoise. Revue Médical Suisse, 6 : 505-509.

Fortmann-Roe S. 2012. Understanding the bias-variance tradeoff. Consulté le 2020-12-14, http://scott.fortmann-roe.com/docs/BiasVariance.html

Fung TT, Rimm EB, Spiegelman D, Rifai N, Tofler GH, Willett WC et Hu FB. 2001. Association between dietary patterns and plasma biomarkers of obesity and cardiovascular disease risk. The American Journal of Clinical Nutrition, 73 : 61-67.

García S, Fernández A, Luengo J et Herrera F. 2009. A study of statistical techniques and performance measures for genetics-based machine learning : accuracy and interpretability. Soft Computing, 13 : 959-977.

GeeksforGeeks. 2020. Detecting multicollinearity with VIF - Python. Consulté le 2020-12-13, https://www.geeksforgeeks.org/detecting-multicollinearity-with-vif-python/

Gouvernement du Canada. 2011. Le tabagisme et la maladie cardiaque. Consulté le 2020-12-05, https://www.canada.ca/fr/sante-canada/services/preoccupations-liees-sante/tabagisme/legislation/etiquetage-produits-tabac/tabagisme-maladie-cardiaque.html

Hamdaoui Y. 2020. Cardiovascular disease. Consulté le 2020-11-28, https://www.kaggle.com/yassinehamdaoui1/cardiovascular-disease

Hu FB. 2003. Plant-based foods and prevention of cardiovascular disease : and overview. The American Journal of Clinical Nutrition, 78 : 544S-551S.

Institut de cardiologie de Montréal. 2020. Obésité Consulté le 2020-12-05, https://www.icm-mhi.org/fr/obesite

Li L. 2019. Principal component analysis for dimensionality reduction. Consulté le 2020-12-05, https://towardsdatascience.com/principal-component-analysis-for-dimensionality-reduction-115a3d157bad

OMS. 2015. Questions-réponses l'hypertension artérielle. Consulté le 2020-12-05, https://www.who.int/features/qa/82/fr/#:~:text=Plus%20la%20pression%20est%20%C3%A9lev%C3%A9e,se%20rel%C3%A2che%20(pression%20diastolique).

Patel N et Upadhyay S. 2012. Study of various decision tree pruning methods with their empirical comparison in weka. International Journal of Computer Applications, 60 : 20-25.

Pennsylvania State University. 2018. Introduction to Generalized Linear Models,. Consulté le 2020-12-14, https://online.stat.psu.edu/stat504/node/216/

Perdesen BK. 2006. The anti-inflmmatory effect of exercise : its role in diabetes and cardiovascular disease control. Essays in Biochemistry, 42 : 105-117.

Rossouw JE, Du Plessis JP, Benadé AJ, Jordaan PC, Kotzé JP, Jooste PL et Ferreira JJ. 1983. Coronary risk factor screening in three rural communities. The CORIS baseline study. South African medical journal, 64 : 430-436.

Rossouw JE, Jooste PL, Chalton DO, Jordaan ER, Langenhoven ML, Jordaan PCJ, Steyn M, Swanepoel ASP et Rossouw LJ. 1993. Community-based intervention : the Coronary Risk Factor Study (CORIS). International Journal of Epidemiology, 22 : 428-438.

Sanharawi ME et Naudet F. 2013. Comprendre la régression logistique. Journal français d'ophtalmologie, 36 : 710-715.

Scikit-learn. 2020a. sklearn.preprocessing.MinMaxScaler. Consulté le 2020-12-05, https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

Scikit-Learn. 2020b. sklearn.decomposition.PCA. Consulté le 2020-12-05, https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

SciPy. 2014. scipy.stats.pointbiserialr. Consulté le 2020-12-13, https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.pointbiserialr.html

Srour B, Fezeu LK, Kesse-Guyot E, Allès B, Méjean C, Andrianasolo RM, Chazelas E, Deschasaux M, Hercberg S, Galan P, Monteiro CA, Julia C et Touvier M. 2019. Ultra-processed food intake and risk of cardiovascular disease: prospective cohort study (NutriNet-Santé) The BMJ, 365 : I1451.

Statistique Canada. 2013. Taux de cholestérol chez les Canadiens, 2009 à 2011. Consulté le 2020-12-05, https://www150.statcan.gc.ca/n1/pub/82-625-x/2012001/article/11732-fra.htm

Statistique Canada. 2015. Obésité abdominale et facteurs de risque de maladie cardiovasculaire à l'intérieur des catégories d'indice de masse corporelle. Consulté le 2020-12-05, https://www150.statcan.gc.ca/n1/pub/82-003-x/2012002/article/11653-fra.htm

Sucky RN. 2020. Logistic regression model fitting and finding the correlation, P-value, Z score, confidence interval, and more. Consulté le 2020-12-14, https://towardsdatascience.com/logistic-regression-model-fitting-and-finding-the-correlation-p-value-z-score-confidence-8330fb86db19