

**Département d’informatique et d’ingénierie**

# **INF6333-01 Éléments d'intelligence artificielle appliquée**

**Rapport finale**

**Titre : Détection de maladie des Cardiovasculaires**

**Mamadou Bamba Diop**

**Joël Barthelemy**

**Université du Québec en Outaouais**

1. **INTRODUCTION**

Les maladies cardiaques et cardiovasculaires jouent depuis des années un rôle dévastateur dans la vie de beaucoup de personnes. Malgré la baisse des taux de décès au cours des dernières années ; Les maladies cardiovasculaire restent l'une des principales causes de décès et d'invalidité dans le monde. Selon les statistiques, une estimation de 17 millions de décès annuel est attribuable aux maladies cardiovasculaires soit le tiers de tous les décès et plus fréquemment dans les pays industrialisés.

La source de ces données statistiques provient, le plus souvent au moment de l'examen médical où différentes données sont collectées. Toutefois, le traitement de ces données peut être complexes parce que, les données sont parfois manquantes, incomplètes, en double ou composées de valeurs aberrantes. Lorsque les données entrantes ne sont pas correctement nettoyées, le traitement devient problématique dans le cas où nous voulons classifier les patients en bonne santé ou atteints de maladies cardiovasculaires. Notre travail vise à classifier certaines catégories de personnes les plus susceptibles de développer une maladie cardiovasculaire.

Une telle tâche d’apprentissage automatique peut jouer un rôle essentiel dans la prédiction de la présence/absence de maladies cardiaques et plus encore. En effet, si cette information est prévue à l'avance, elle peut apporter des informations vitales. Notamment aux experts de la santé qui pourront décider plus rapidement et ensuite adapter leur diagnostic et leur traitement en fonction de chaque catégorie de patient.

Poussés par ce besoin critique, nous avons décidé de prévoir le risque de maladie du cardiaque en utilisant des méthodes comparatives cohérentes. en appliquant les mêmes techniques de prétraitement au même ensemble de données et en comparant ensuite différents algorithmes de classification. Re-échantillonnage est appliqué pour le jeu de données d'entraînement, conduire les données de prévision d'entraînement pour voir les risques de surajustement, afin de fixer le modèle et de rechercher la meilleure précision.

1. **Objectif**

L'objectif principal de ce projet consiste à classer un patient sur la base d'un ensemble de données. La classification se fera en fonction des valeurs des 12 attributs fournis par le jeu de données qui compte 69303 instances qui provient du datasets kaggle. Toutes les valeurs des attributs ont été collectées au moment de l'examen médical.

Ce projet vise à trouver le modèle le plus approprié à la problématique. Pour y arriver, quelques visualisations de données seront effectuées pour donner un meilleur sens aux données. À partir des attributs catégoriciels nous créerons des graphes représentant les incidences dans le but d’apercevoir les catégories impliquées. Nous pourrons ainsi déterminer si le risque est négligeable ou important. On se servira de variables continues pour obtenir le sens des distributions et des cibles. Nous créerons ensuite de nouveaux attributs et supprimerons certains d'entre eux.

Cette exploration nous permettra d'établir si des valeurs aberrantes et des données manquantes sont présentes ou non. Et seront traitées soit sur la moyenne, si les données sont manquantes. Ou normaliser en calculant le z-Score si des valeurs aberrantes existent. Nous procéderons ensuite à un échantillonnage sans remplacement des données pour avoir une idée de l'organisation générale des données. Par la suite, diviser les attributs en données de test et données d'entraînement.

L’étape d’entraînement constitue une étape cruciale pour le projet. Pour ce faire, on utilisera deux modèles. Le modèle de l’arbre de décision pour la classification, et le modèle de la régression logistique. Ces deux modèles seront utilisés pour prédire les données d’entraînement et les données de test.

Ces deux modèles feront l'objet d'une évaluation afin de déterminer lequel sera le plus performant. En se basant sur leur précision, leur taux d’erreur moyenne, le recall, le f1-score et le support. Par la suite, un graphique représentant la précision des tests et de l'entraînement sera présenté pour montrer le surajustement du modèle testé. Une optimisation sera réalisée dans le modèle en obtenant les meilleures performances grâce à de nouveaux paramètres.

1. **Exploration et prétraitement des données**

Dans cette section, nous présenterons les différentes étapes dans le processus de prétraitement des données et les décrirons.

**Présentation de l’ensemble des données**

L’ensemble de données est composé de 69303 instances avec 12 attributs caractéristiques.

Nous étudierons les caractéristiques de l’attribut cible cardio, qui nous permettront d’avoir les prédictions et de classer les risques par catégories d’âge, par indice masse corporelle, par genre et par taux de glucose.

Les différents attributs de l’ensemble de données sont les suivants:

•âge (en jours converti en années: 0 à 70 ans).

•taille(cm)

•poids (float(kg))

•Genre(code catégoriciel): valeur 1: femme 2: homme

•ap\_hi (pression artérielle systolique )

•ap\_lo (pression artérielle diastolique)

•cholesterol 1:normal, 2: dessus de la normale, 3: bien au dessus de la normale

•glucose 1: normal, 2: dessus de la normale, 3: bien au dessus de la normale

•fumeur :(binaire)

•preneur d’alcohol : (binaire )

•Présence ou absence de maladie cardiovasculaire: cardio

Tous ces attributs sont impliqués dans l’évaluation du risque de détection de maladie cardiaque chez le patient.

1. **Exploration des données**

L'exploration de données a été la première phase du projet. Il s'agit d'une étape cruciale qui nous a permis de rendre les données plus significatives. En particulier, en faisant la distinction entre les types de données, c'est-à-dire les variables continues et catégoriques (voir la figure 4). Un meilleur traitement est assuré dans la deuxième phase puisque nous connaissons préalablement les types de valeurs que nous devrons traiter. Mais plus intéressant encore, cette distinction nous permet de nous identifier à l'aide de graphes représentant l'incidence cardiovasculaire avec des variables catégoriciels en fonction de l'attribut cible (cardio). Ceci est illustré par la figure 1, qui représente un exemple d'incident cardiaque par genre.

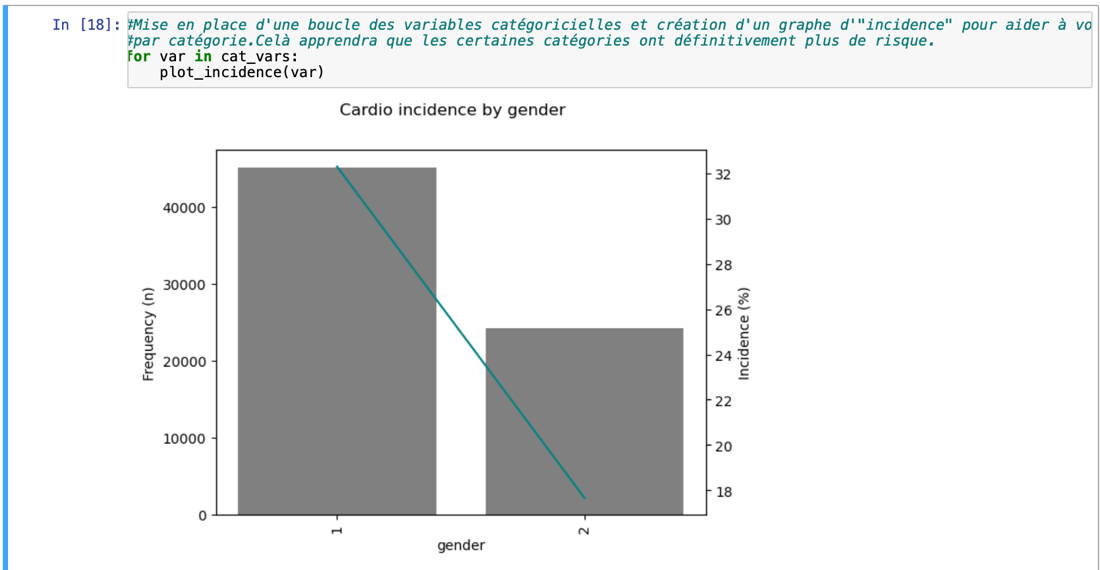


Figure: 1

Nous remarquons que le genre 2 représentant les hommes. Enregistre plus de risques, où sont plus confrontés à des maladies cardiovasculaires que le genre 1 représentant les femmes.

Le subplots est utilisé pour représenter les incidences. Ce graphique agit comme une bande utilitaire et aide à créer des dispositions communes de sous-parcelle.

La boîte a moustache sera utilisée pour trouver des valeurs aberrantes, ce qui est un moyen rapide d'afficher le profil des attributs. Ce diagramme est tiré de données continues qui sont plus susceptibles d'enregistrer des valeurs aberrantes. Autrement dit, des valeurs très différentes des autres valeurs de l'ensemble de données. La figure 2 montre des valeurs aberrantes dans l'attribut cible en fonction de la taille.

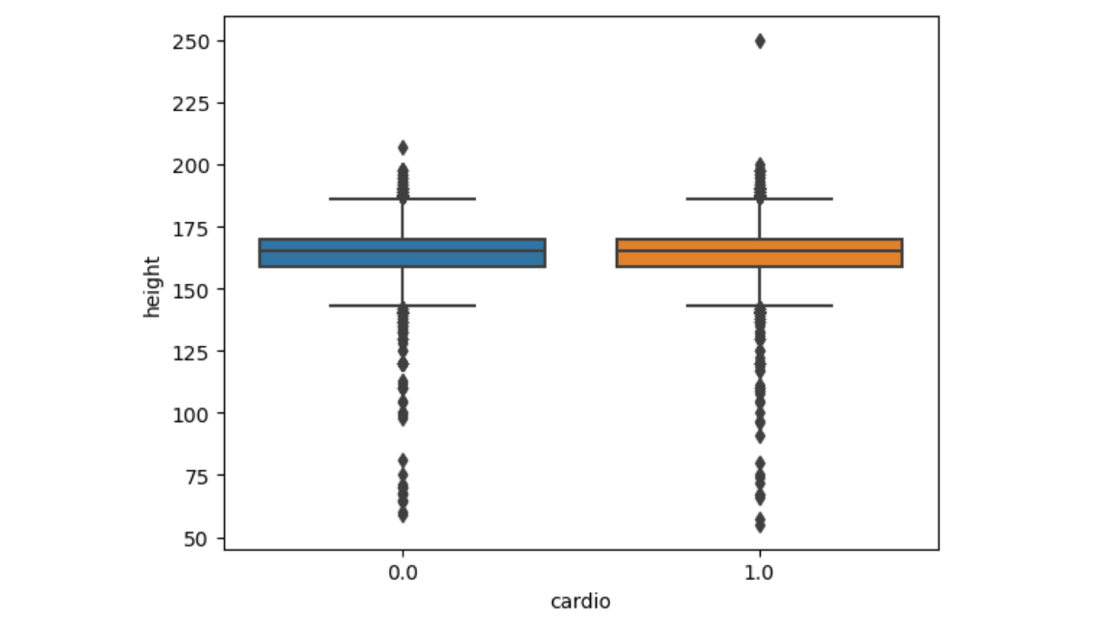


Figure: 2

Dans cette représentation, on note un point très éloigné des autres points. Ce point représente bien une donnée aberrante dont il a été nécessaire de traiter.

Nous avons pu vérifier l'existence ou non de données manquantes comme le montre la figure 3.

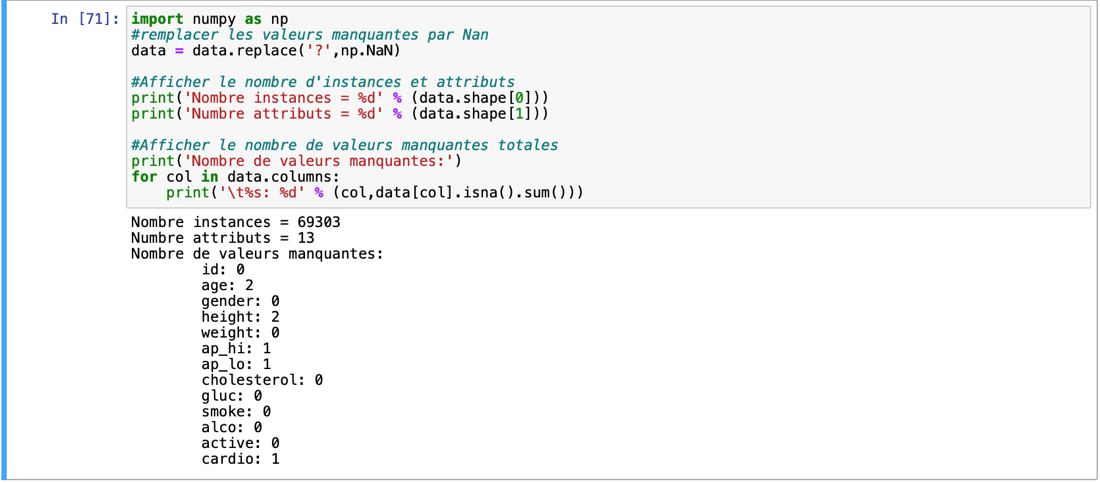
[[1]](#footnote-1)

Figure: 3

Les résultats ont montré la présence de 8 données manquantes dispersées au niveau âge, taille, ap\_hi, ap\_lo et cardio. Ces données manquantes seront plus tard traitées.

Il est essentiel de noter que la représentation de l’âge est rendue possible, grâce à la fonction lambda de python. Cette fonction permet de faire la conversion pour transformer l’âge en année au lieu en jour.

1. **Traitement des données**

* Données manquantes

A ce niveau, il a fallu faire face aux données manquantes. Les données manquantes sont des instances vides. Afin d'être en mesure de les traiter, deux solutions étaient possibles. La première solution proposée serait de les remplacer par la moyenne ou la médiane grâce à la librairie python numpy. La moyenne ou la médiane, ces données manquantes seront remplacées par la valeur la plus fréquente ou la moyenne de la série de valeurs de l'attribut. La Figure 5 montre le remplacement de l'ensemble des données manquantes par la médiane.

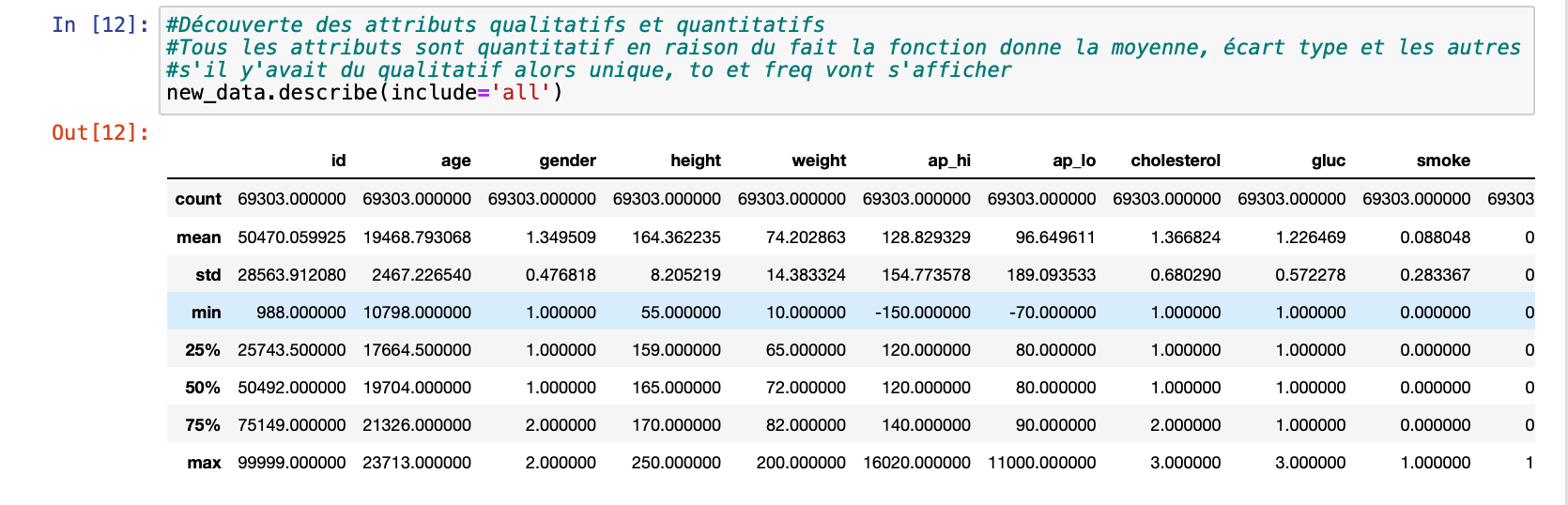


Figure: 4

La deuxième solution possible serait de supprimer totalement les instances contenant les données manquantes. La suppression de ses données n’etaient pas necessaire vue qu’elles representaient moins de 1% des données.

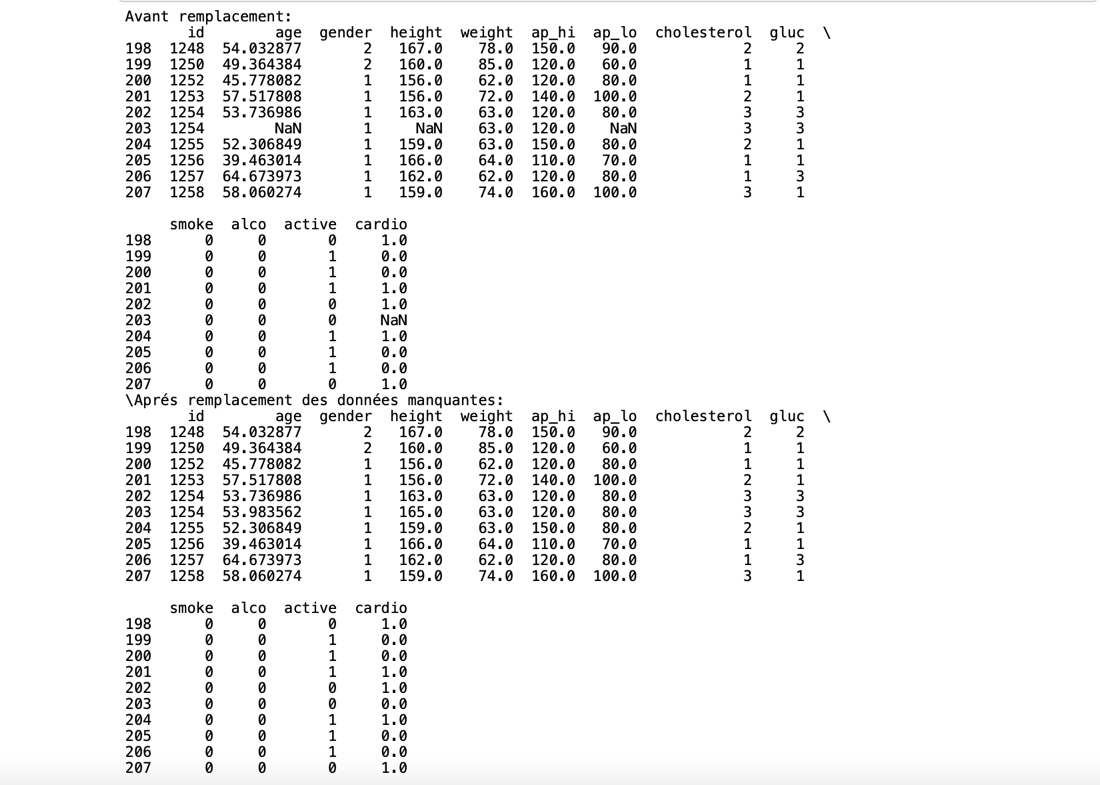


Figure: 5

* **Données aberrantes**

Les données aberrantes peuvent constituer un obstacle à la bonne performance d’un modèle. En effet, le traitement des données aberrantes est nécessaire dans cet ensemble de données en raison du fait que la plupart des données aberrantes constatées avec la boîte à moustache se localisent au niveau des attributs taille et poids. Nous avons décidé de les supprimer. Cela constitue une des raisons, mais la plus grande raison est due au fait que ces deux attributs ne sont pas trop informatifs pour le besoin de ce projet. Donc, obtenir un attribut qui peut les représenter est plus intéressant pour nous. Nous avons privilégié le BMI (indice de masse corporelle) qui permet d’estimer la corpulence d’une personne et est exprimé en kg/m2. Pour calculer le BMI, on divise le poids par la taille. Ainsi le BMI sera catégorisé également par la suite.

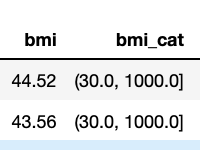


Figure: 6

Notre première approche pour traiter les données aberrantes a été de standardiser en calculant le z-score en divisant la moyenne de l’écart type. Les résultats sont donnés par la figure 7.

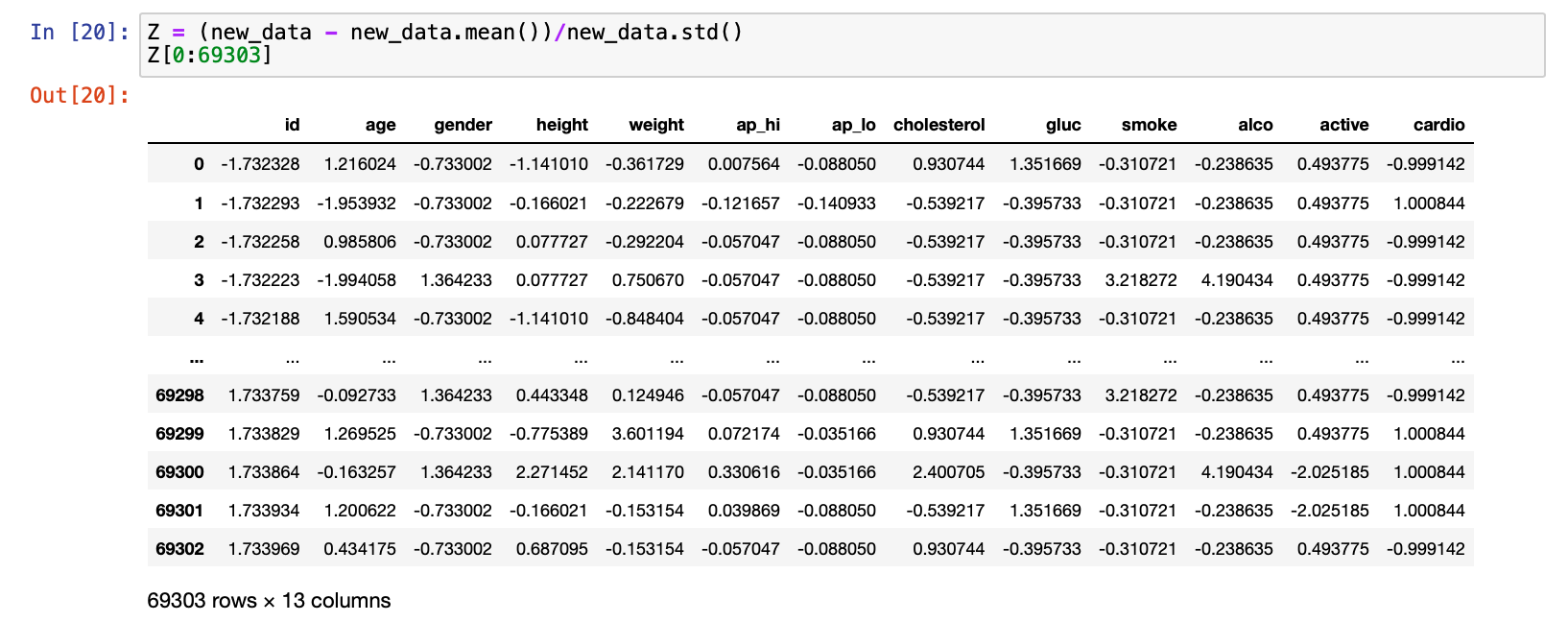


Figure: 7

À partir de ces résultats, nous avons établi la condition selon laquelle les cas où les valeurs sont supérieures à 3 sont considérés comme de grandes valeurs. Ce qui signifie des données aberrantes et celles inférieures à -3 sont considérées comme faible valeur. Cette solution a finalement éliminé aucune donnée aberrante due à la condition. Le but de ne pas perdre beaucoup de données importe la premiére solution a été conservée.

* **Echantillonage des données**

En raison du fait que nous avons des données multivariées, il était essentiel de découvrir en premier lieu la corrélation entre les attributs. Cette corrélation permet de distinguer les relations entre les attributs, comme on le voit à la figure 8.

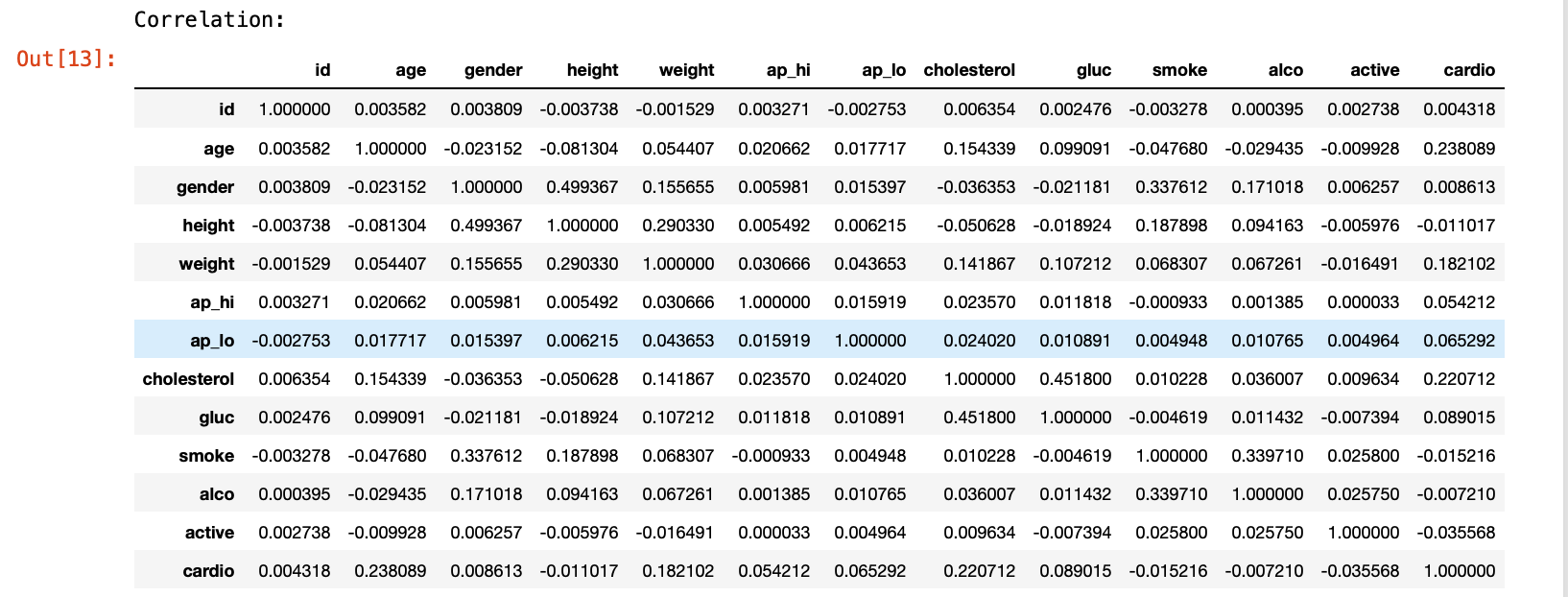


Figure: 8

Nous avons ainsi pu déterminer le type d'échantillonnage que nous allons faire pour voir comment les données sont organisées. Par conséquent, le type d'échantillonnage finalement choisi était celui sans remplacement afin de toujours conserver les données. Un échantillonnage avec n = 3 a été choisi ce qui a abouti au résultat de la figure 9. Un paramètre de random\_state a été ajouté pour sélectionner 1% des données et afficher l’échantillon sélectionné.

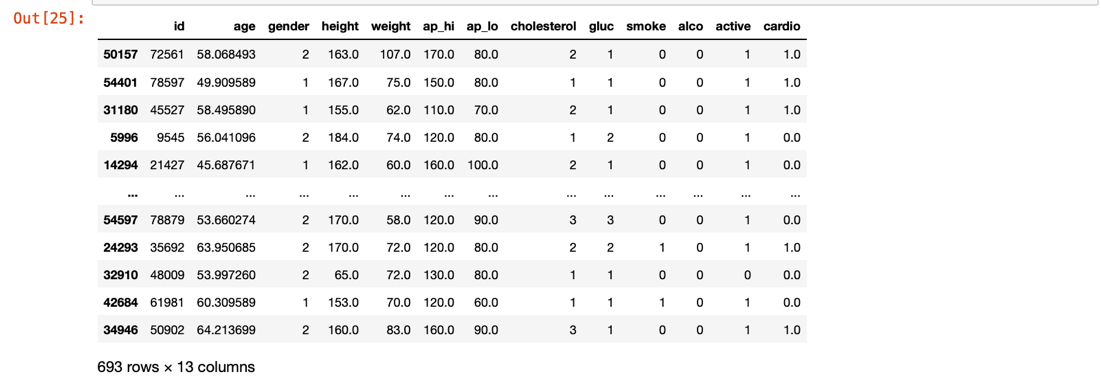


Figure: 9

Comme résultat, 693 lignes ont été affectées par l’échantillonnage.

* **Conversion d’attribut continu**

L'objectif du projet est de classifier les patients atteints de maladies cardiovasculaires selon leur âge et d'autres critères. L'attribut âge, cependant, est un attribut continu qui devra être transformé si les patients doivent être classés en fonction de l'âge. Pour ce faire, il fallait ajouter un attribut age\_cat qui catégorise les différents âges de l'ensemble de données. En effet, il s'obtient en dupliquant l'attribut âge et en le supprimant ensuite. La fonction qcut de Python reçoit les paramètres d'âge qu'elle duplique et regroupe des valeurs similaires en vue d'une catégorisation numérique.

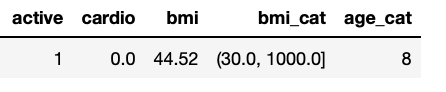


Figure: 10

Ainsi, comme résulat, age\_cat reçoit et catégorise les ages au nombre de 8.

* **Division des données de test et d’entrainement**

Dans le but de préparer les données d’entraînement et de test, celles-ci doivent être séparées afin de répartir les pourcentages. Ainsi dans notre cas, en raison des données réelles nous avons separé les données autour de 90% pour l’entrainement et 10% pour le test.



Figure: 11

Ainsi, comme résulats 62372 lignes ont été entrainés et 6931 lignes sont testé sur les 10 attributs.

* **Mise en place d’un pipeline**

La mise en place d’un pipeline a pour objectif de transformer les variables catégoricielles en numérique grâce à la variable hot encoder et utilisation de la variable scale pour les attributs continus. C’est une étape obligatoire depuis que l’apprentissage machine exige des données numériques. C’est important dans la classification, car si le modèle de classification utilise des données catégoricielles sans transformation, une erreur sera générée. Cela est causé par le fait que l’algorithme ne peut pas travailler avec des valeurs non numériques. En outre, les données peuvent se produire à différentes échelles, ce qui mène à des prévisions biaisées ou faussées en termes d'erreurs de classification et de taux d'exactitude.

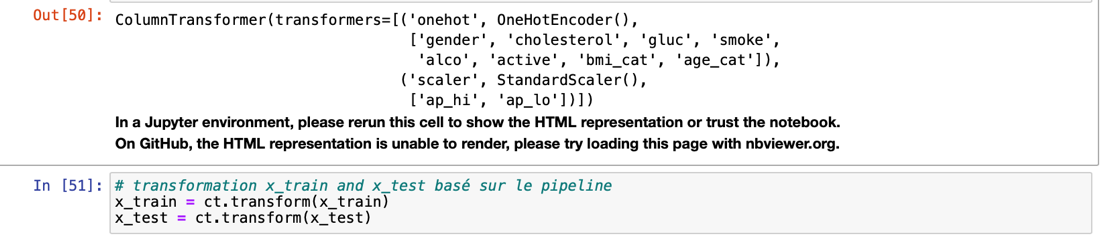


Figure: 12

1. **Entrainement et évaluation des modéles**
2. **Arbres de décisions**

Le modèle d’arbre de décision a été entraîné sur la base de deux paramètres qui sont la profondeur maximum max\_depth= 10 et l’état générateur random\_state= 42. Nous avons entrainés les modèles en se basant sur les données d’entraînement. Après avoir fini d’entraîner le modèle avec la méthode d’arbre de décision, nous avons fait des prédictions sur les données de test et mesuré la précision sur les données de test avec le MSE, l’erreur quadratique moyenne. Le MSE est une mesure qui caractérise la précision de l’estimateur. De cette façon, nous saurons que la précision est bonne quand le MSE est proche de la valeur 0. Une bonne estimation du MSE a été obtenue comme le montre la figure 13.

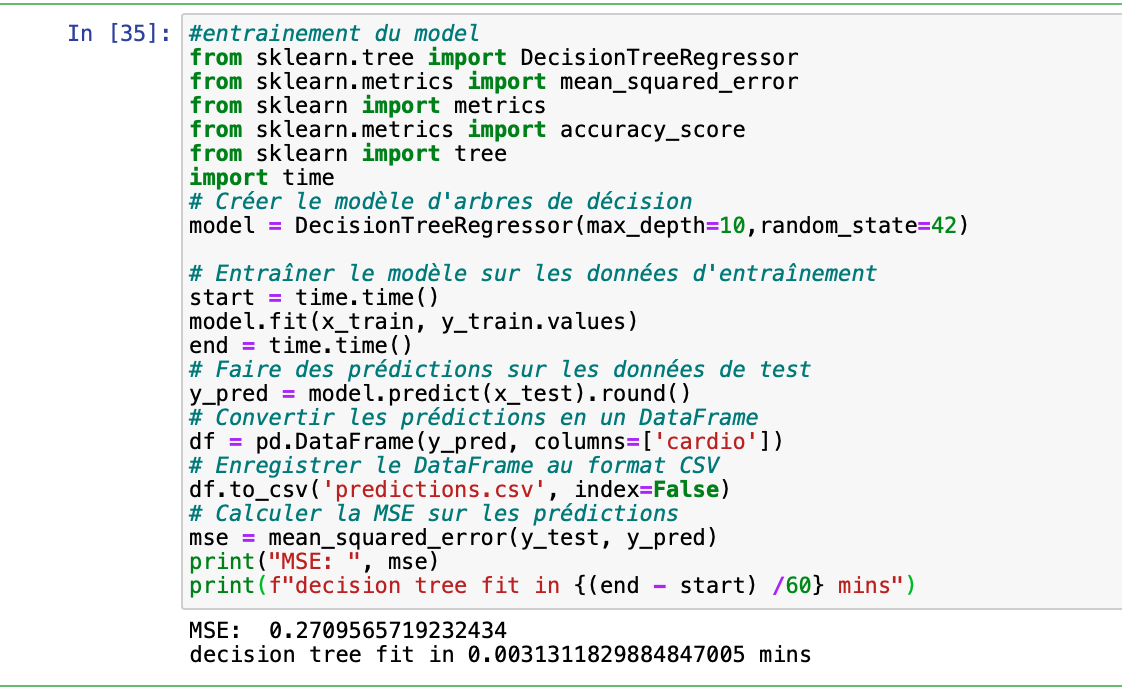


Figure: 13

Le modéle a réussit à s’ajuster sur un temps de 0.0031 minutes avec une précision estimée à plus de 70 % comme le montre la figure 14.

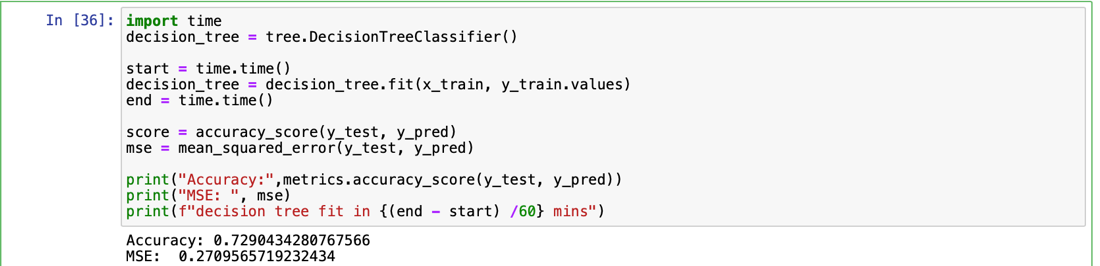


Figure: 14

De façon globale, nous obtenons les performances suivantes pour l’arbre de décision

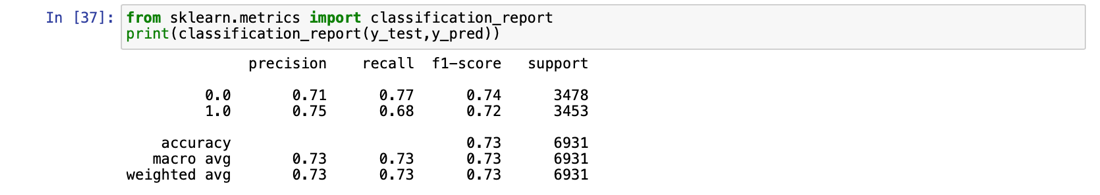


Figure: 15

Nous pouvons donc conclure que le modèle enregistre de bonnes performances sur les données de test. Sur la base de ces performances, nous voulions évaluer la précision des données de test et d'entraînement et effectuer la représentation graphique de la précision. La raison pour laquelle cette représentation est nécessaire est que celà nous permettra de distinguer, laquelle des données d’entrainements et de test est plus susceptible à un surajustement (overfitting). Pour effectuer cela un paramètre profondeur maximum max\_deph est établie compris de 2 à 50 qui va s’incrémenter dans l’entraînement du modèle. Le résultat est présenté par la figure 16.

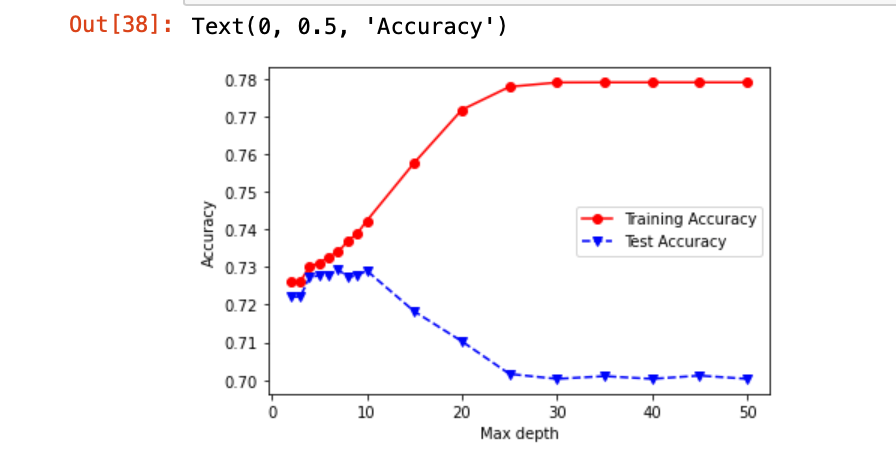


Figure :16

Nous pouvons constater que la précision de l'entraînement continuera d'évoluer avec l'augmentation de la profondeur maximale. Le modèle est de plus en plus complexe. D'autre part, la précision du test avait initialement évolué jusqu'à une évolution de 5, puis a enregistré une chute considérable en raison du surajustement.

Avant d'essayer un autre modèle, nous sommes aussi tentés d'essayer d'optimiser le modèle d'arbre de décision. Ceci a été fait par la sélection de nouveaux paramètres pour le modèle. Cette fois-ci, nous avons choisi un critère nommé entropy dans le modèle d’arbre de décision. L'entropie est une mesure informationnelle qui indique le désordre de caractéristiques avec l'attribut cible. La division optimale est choisie par la caractéristique avec moins d'entropie. Un autre paramétre max\_depth sera utilisé avec valeur 3. Les résultats sont illustrés par la figure 17.

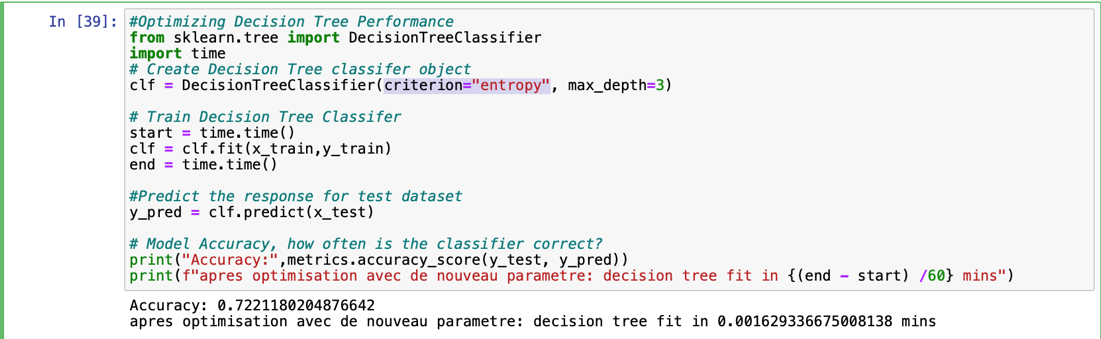


Figure: 17

Ainsi, après optimisation, la précision n'a pas beaucoup changé, on constate une petite difference.

1. **La régression logistique**

Le second modèle que nous avons été en mesure de tester est la régression logistique. Ce modèle a été entraîné tout comme l’arbre de decision, exceptés les paramètres qui changent. En effet, comme il s'agit d'une technique de classification, ce modèle est aussi capable de générer les données avec une bonne précision, comme l'indique la figure 18.

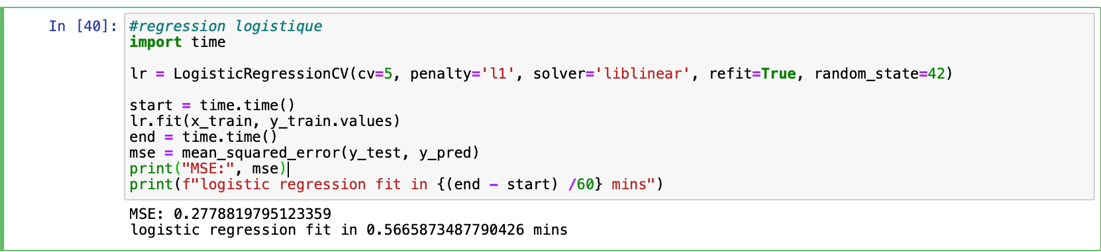


Figure: 18

Toutefois, nous remarquons une performance de rajustement plus longue en termes de temps que l'arbre de decision. Soit un temps de 0,56 minutes. La précision de l'erreur demeure la même que dans l'arbre de décision.

Les résultats de précision sont illustrés par la figure ci-dessous.

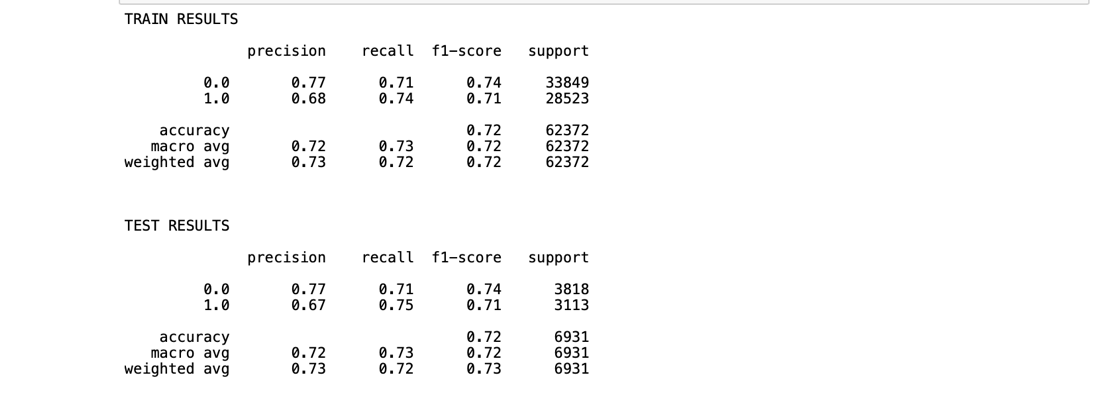


Figure: 19

En termes de précision, nous constatons une précision plus faible de l'arbre de décision dans les données de test. Soit 0.77 pour la régression logistique contre 0.71 pour l’arbre de décision. Les résultats des données d’entraînement sont bonnes mais l’arbre de décision remporte toujours avec une meilleure précision.

1. **Interprétation du modéle**

On a procédé à l'interprétation du modèle en générant un graphique des attributs catégoriciels de l'ensemble de données. Ce graphique a été obtenu en fixant l’objectif d’obtenir le ratio odds. Il s'agit d'une mesure statistique, utilisée en épidémiologie, qui exprime le niveau de dépendance entre les variables aléatoires.

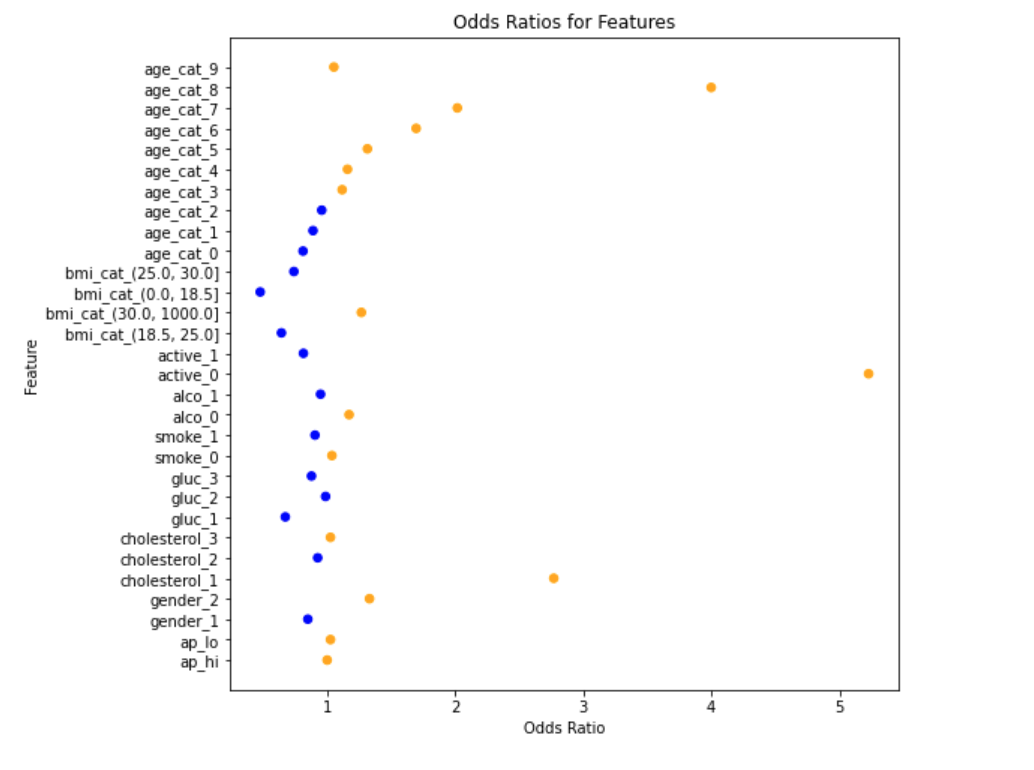
[[2]](#footnote-2)

Figure: 20

Selon le tableau et les catégories d'attributs, on peut prédire le risque de maladie cardiovasculaire dans les deux catégories suivantes :

* Catégorie d’âge: Les points orange augmentent au fur et à mesure de l'augmentation de la catégorie. Il en résulte une augmentation du risque de maladie cardiovasculaire avec l'âge. Car le risque le plus important est noté au niveau age\_cat\_8 en raison des points très distants.
* La catégorie du genre: On note également que le risque est également grand pour le genre 2 représentant le genre masculain.

Par conséquent, nous pouvons conclure que le risque de développer une maladie cardiovasculaire chez les hommes et en fonction de leur âge. Plus, l'homme est âgé, plus le risque de maladie cardiovasculaire augmente.

**CONCLUSION**

Sur la base des connaissances acquises dans le cadre de ce projet, les deux principaux facteurs à considérer pour la détection des maladies cardiovasculaires sont l'âge et le sexe. Grâce aux connaissances que nous avons pu en tirer. On peut conclure que plus, un homme est âgé, plus il risque de développer une maladie cardio-vasculaire.

Le recours à des intelligences artificielles a permis à la science de faire un bond en avant. L'application des concepts tirés du cours d’intelligence artificielle appliquée, nous a permis de questionner les données en réalisant les différentes étapes d'exploration, de prétraitement, d'entraînement, d'évaluation des modèles et l'interprétation des données. Il s'agit d'étapes cruciales dans l'interrogation des données dans l'apprentissage automatique. Pour les besoins de ce projet, nous estimons que le modèle d'arbre de décisions constituent le meilleur choix. Bien que le modèle de régression ait donné des résultats presque similaires, sa précision étant plus faible que celle de l'arbre de décision dans les données de test. Soit 0.77 pour la régression logistique contre 0.71 pour l’arbre de décision.

**REFERENCES:**

Heart-Stroke-Prediction-using-ML. [En ligne; accédé le 2 Mars 2023]. URL: <https://github.com/VipulGajbhiye/Heart-Stroke-Prediction-using-ML#algorithms>]

ML Pipelines [En ligne ; accédé le 2 Mars 20231.] URL : [ <https://spark.apache.org/docs/latest/ml-pipeline.html#main-concepts-in-pipelines>

Détection de maladie des Cardiovasculaires.[En ligne; accédé le 2 Mars 2023]. URL:

<https://www.kaggle.com/code/nattiv/logistic-regression-to-detect-cvd>

Exploration de données. [En ligne; accédé le 2 Mars 2023]. URL:

<http://www.cse.msu.edu/~ptan/dmbook/tutorials/tutorial3/tutorial3.html>

Data Processing. [En ligne; accédé le 2 Mars 2023]. URL:

<https://www.cse.msu.edu/~ptan/dmbook/tutorials/tutorial4/tutorial4.html>

Decision Tree. [En ligne; accédé le 2 Mars 2023]. URL:

<https://quantdare.com/decision-trees-gini-vs-entropy/>

Classification. [En ligne; accédé le 2 Mars 2023]. URL:

<https://www.cse.msu.edu/~ptan/dmbook/tutorials/tutorial6/tutorial6.html>

Effect of Correlated Attributes. [En ligne; accédé le 2 Mars 2023]. URL:

<https://www.cse.msu.edu/~ptan/dmbook/tutorials/tutorial5/tutorial5.html#5.3-Effect-of-Correlated-Attributes>

Détection de maladie des Cardiovasculaires.[En ligne; accédé le 2 Mars 2023]. URL:

<https://www.inspq.qc.ca/pdf/publications/590-MaladiesCoeursVasculairesCerebrales.pdf>

1. présence de données manquantes [↑](#footnote-ref-1)
2. graphique ratio odds [↑](#footnote-ref-2)