

REPUBLIQUE DU CAMEROUN

Paix-Travail-Patrie

\*\*\*\*\*\*\*\*\*\*

UNIVERSITE DE YAOUNDE I

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

ECOLE NATIONALE SUPERIEUR POLYTECHNIQUE DE YAOUNDE

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

DEPARTEMENT DU GENIE INFORMATIQUE

REPUBLIC OF CAMEROON

Peace-Work-Fatherland

\*\*\*\*\*\*\*\*\*\*\*\*

UNIVERSITY OF YAOUNDE I

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

NATIONAL ADVANCED SCHOOL

OF ENGINEERING OF YAOUNDE

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

DEPARTEMENT OF COMPUTER SCIENCE ENGINEERING

**INTELLIGENCE ARTIFICIELLE ET APPLICATIONS NUMERIQUES**

Rapport de Projet :

Détection des maladies cardiaques

Etudiant

NGUAZONG AUREL 20P001

Sous la supervision : M. Bitha

Année académique :

2023/2024

Table des matières

[INTRODUCTION 3](#_Toc167130622)

[I. Méthodologie de conception 5](#_Toc167130623)

[1. Importation et exploitation de la dataset 5](#_Toc167130624)

[2. Visualisation des données 6](#_Toc167130625)

[3. Feature Processing 8](#_Toc167130626)

[4. Séparation des données test et données d’entraînement 9](#_Toc167130627)

[5. Modèle d’entrainement 9](#_Toc167130628)

[II. Remarque et commentaire 13](#_Toc167130629)

# INTRODUCTION

Intitulé du projet :

Les maladies cardiaques, également appelées maladies cardiovasculaires, sont un terme général utilisé pour désigner les maladies et affections affectant le cœur et le système circulatoire. C’est une cause majeure de handicap partout dans le monde. Le cœur étant l’un des organes les plus vitaux du corps, ses maladies affectent également d’autres organes et certaines parties du corps. Il existe plusieurs types et formes de maladies cardiaques. Les plus courants provoquent un rétrécissement ou un blocage des artères coronaires, un dysfonctionnement des valvules cardiaques, une hypertrophie de la taille du cœur et plusieurs autres conduisant à une insuffisance cardiaque et à une crise cardiaque.

Faits marquants selon l'OMS (Organisation Mondiale de la Santé) :

Les maladies cardiovasculaires (MCV) sont la principale cause de décès dans le monde.

On estime que 17,9 millions de personnes sont mortes de maladies cardiovasculaires en 2019, ce qui représente 32 % de tous les décès dans le monde. Parmi ces décès, 85 % étaient dus à une crise cardiaque ou à un accident vasculaire cérébral.

Plus des trois quarts des décès dus à des maladies cardiovasculaires surviennent dans des pays à revenu faible ou intermédiaire.

Sur les 17 millions de décès prématurés (moins de 70 ans) dus à des maladies non transmissibles en 2019, 38 % étaient dus à des maladies cardiovasculaires.

La plupart des maladies cardiovasculaires peuvent être évitées en s’attaquant aux facteurs de risque comportementaux tels que le tabagisme, une mauvaise alimentation et l’obésité, l’inactivité physique et la consommation nocive d’alcool.

Il est important de détecter les maladies cardiovasculaires le plus tôt possible afin que la prise en charge avec des conseils et des médicaments puisse commencer.

Le jeu de données Analyse\_cardiaque comporte en son sein des données reparties dans les colonnes suivantes :

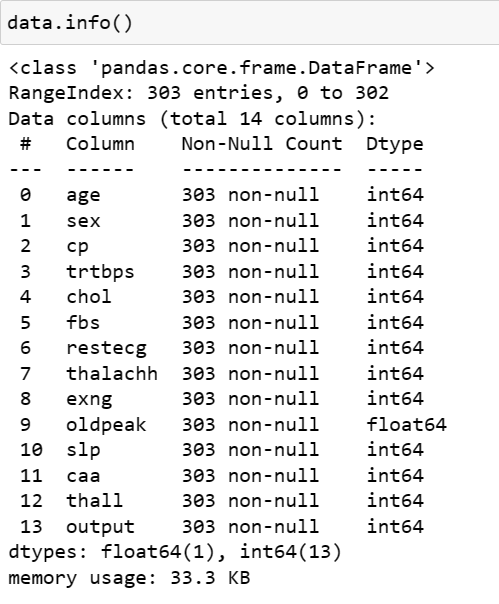
* Âge "age" : L'âge de l'individu.
* Sexe "sex" : Le sexe de l'individu.
* Type de douleur thoracique "cp" : Le type de douleur thoracique ressentie par l’individu.
* Pression artérielle au repos "trtbps" : tension artérielle au repos de l'individu (en mm Hg).
* Cholestérol "chol" : Les niveaux de cholestérol de l'individu (en mg/dL).
* Taux de sucre dans le sang à jeun "fbs" : Le taux de sucre dans le sang à jeun de l'individu (> 120 mg/dL est considéré comme élevé).
* Résultats électrocardiographiques au repos (ECG) "restecg" : Résultats de l'électrocardiogramme au repos.
* Fréquence cardiaque maximale atteinte "thalachh" : la fréquence cardiaque maximale atteinte par l'individu pendant l'exercice.
* Angine induite par l'exercice "exng" : Si la personne a souffert d'angine de poitrine pendant l'exercice.
* Dépression ST induite par l'exercice "oldpeak" : Dépression ST induite par l'exercice par rapport au repos.
* Pente du segment ST d'exercice de pointe "slp" : La pente du segment ST d'exercice de pointe.
* Nombre de vaisseaux sanguins majeurs colorés par fluoroscopie "saa" : Le nombre de vaisseaux sanguins majeurs colorés par fluoroscopie.
* Thalassémie "thall" : Un trouble sanguin ; différents types de thalassémie peuvent être représentés dans l’ensemble de données.
* Cible "Output" : Si l'individu présente le risque d'une crise cardiaque (généralement représenté sous forme binaire : 0 pour aucune crise cardiaque, 1 pour une crise cardiaque).

Vous êtes appelez à créer un modèle capable de prédire des risques d'accident cardiaque chez un patient en utilisant des méthodes d'apprentissage automatique.

# Méthodologie de conception

## Importation et exploitation de la dataset

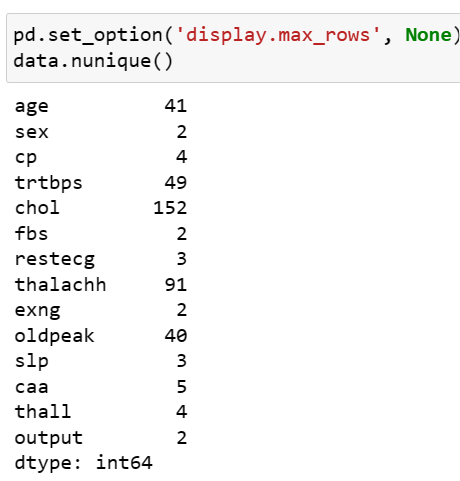
La dataset importée comporte 303 lignes et 14 colonnes de base et sont diversifiées selon les types de données car toutes les colonnes n’ont pas le même type de données.



Après avoir vérifié les données dupliquées et après les avoir supprimées, l’on obtient une dataset de 302 lignes et 14 colonnes.



Par la suite, l’on inspecte la dataset pour vérifier les valeurs différentes de chaque colonne.



L’on constate dans ce cas que chaque colonne a sont intervalle de valeurs différentes assez disproportionnées avec l’âge qui a 41 valeurs différentes ou le taux de cholestérol qui a 152 valeurs différentes tandis que le sexe a 2 valeurs différentes. Cela conduit au tableau de description :



N’ayant pas de minimum, maximum ou de moyenne équivalent, il va falloir réguler pour pouvoir entraîner le/les modèle(s).

## Visualisation des données

Pour commencer l’on sépare la dataset en 2 :

* Les variables numériques : celles dont leurs valeurs sont une représentation d’une mesure numérique.

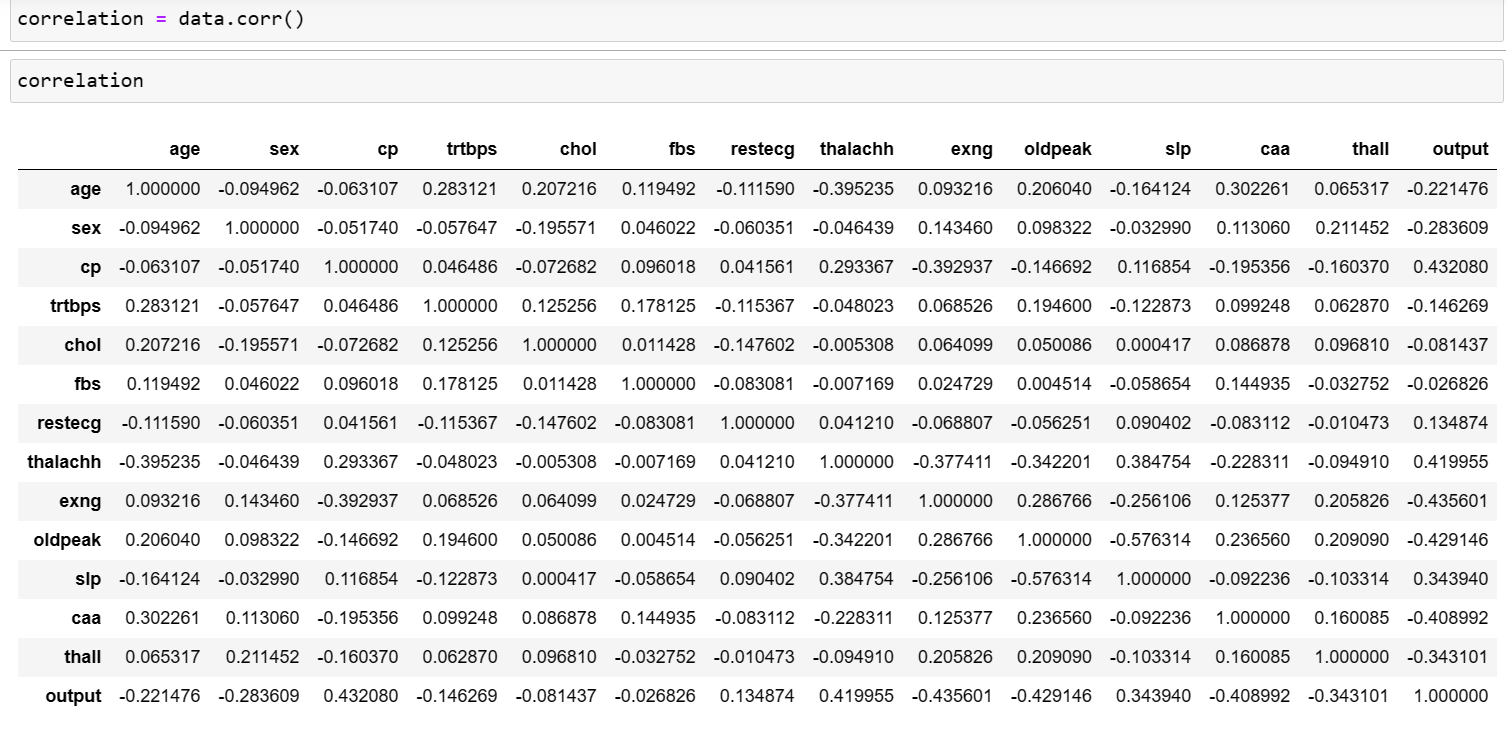


* Les variables catégoriques : celles qui représentes une catégorie labélisée par une valeur numérique.



Par la suite l’on visualise le tout pour compter les valeurs de chaque variable.

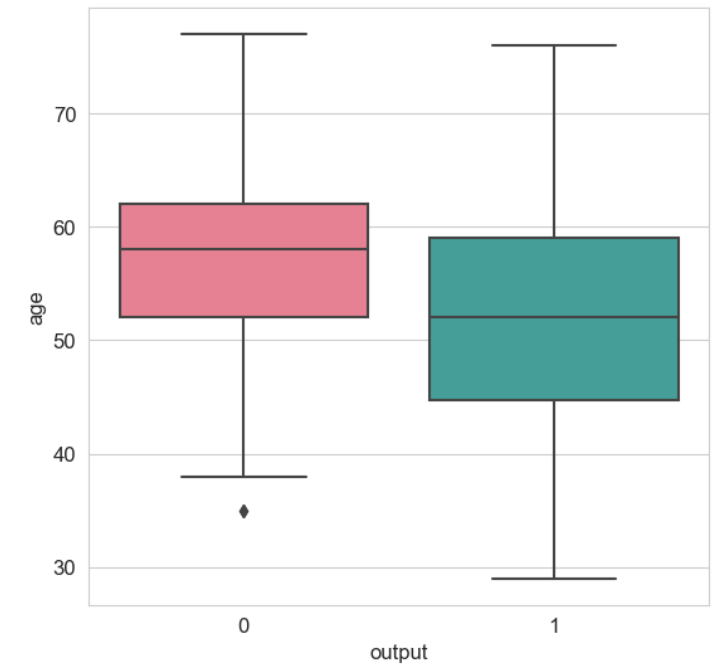
Afin d’avoir la corrélation entre les variables et leur impact entre elles, l’on doit concevoir la matrice de corrélation.



Avec bien entendu son graphe de corrélation.

L’on remarque que les variables qui ont un fort impact sur la sortie sont (cp, thalachh, slp). Ce seront les variables à considérer en priorité.

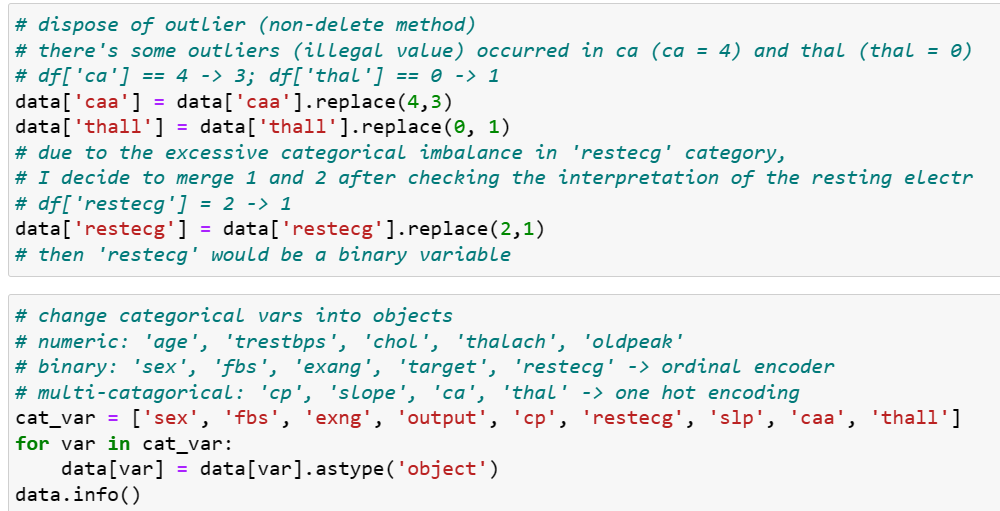
Concernant la visualisation des catégories des variables entre elle en analyse bivariée, l’on peut prendre pour exemple la relation entre l’âge et la sortie :

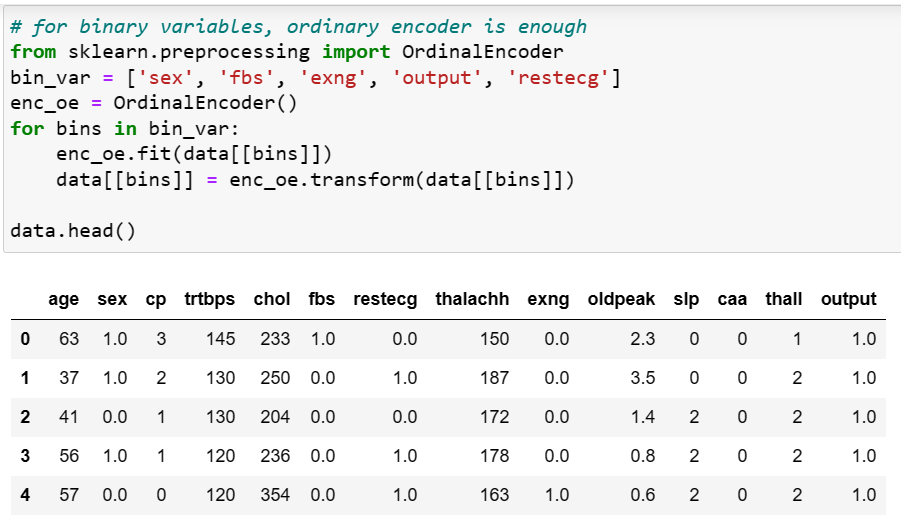


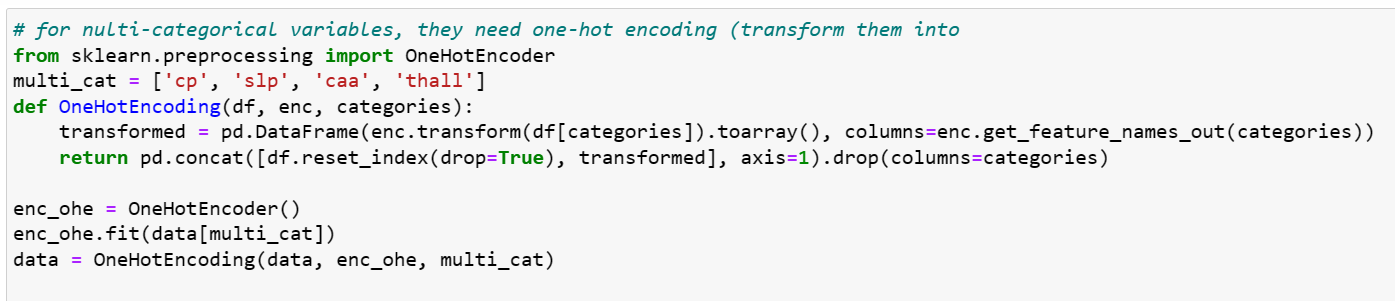
Cela s’interprète par le fait que ceux qui auront tendance à être malade sont plus dans la tranche des 45 à 59 ans et ceux non malades entre environ 52 à 62 ans. Ça n’a pas d’impact significatif dans l’analyse, c’est pour cette raison que sont taux d’impact dans la matrice de confusion plus haut était de -22%.

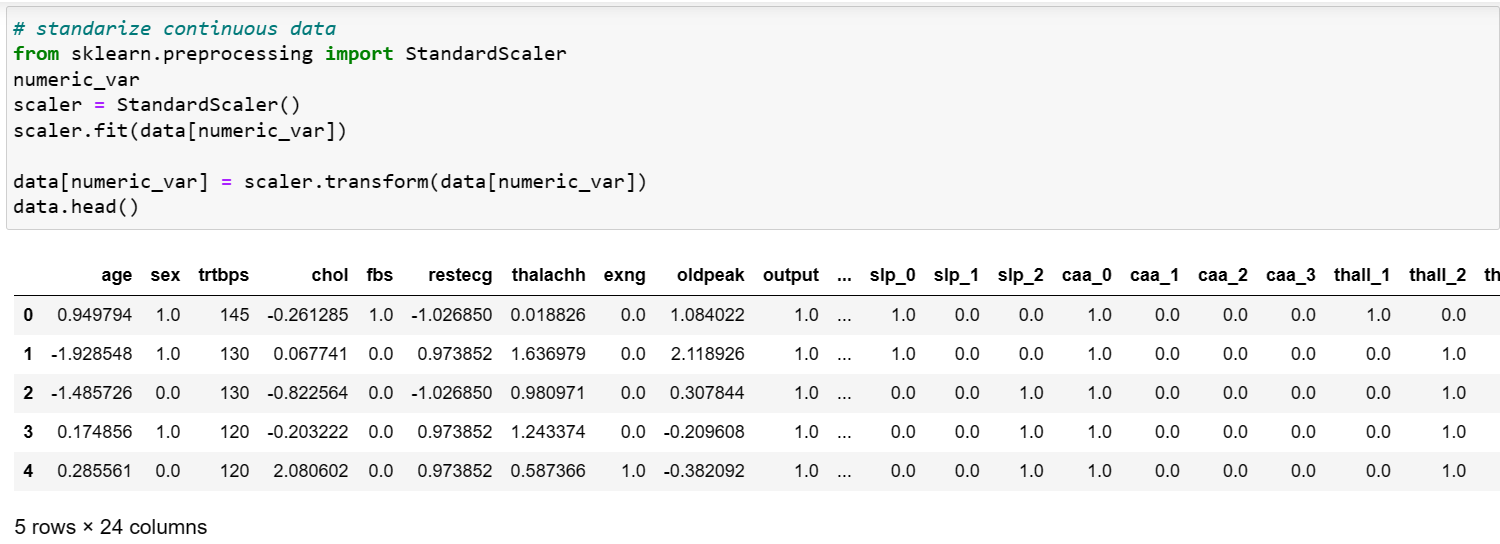
## Feature Processing

Cette partie est essentiellement conçu pour normaliser les données dans un même intervalle pour homogénéiser la dataset afin de ne pas avoir d’erreur dans l’entraînement.



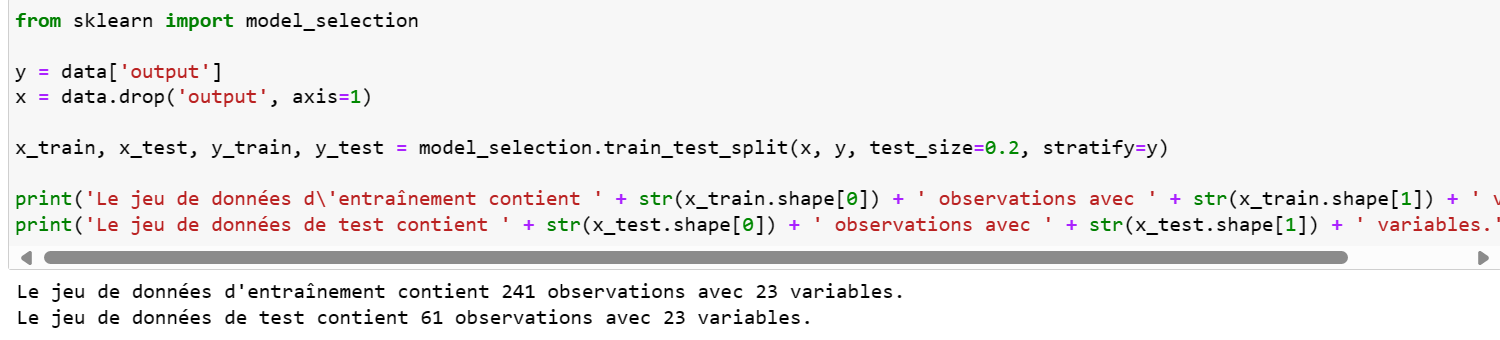






## Séparation des données test et données d’entraînement

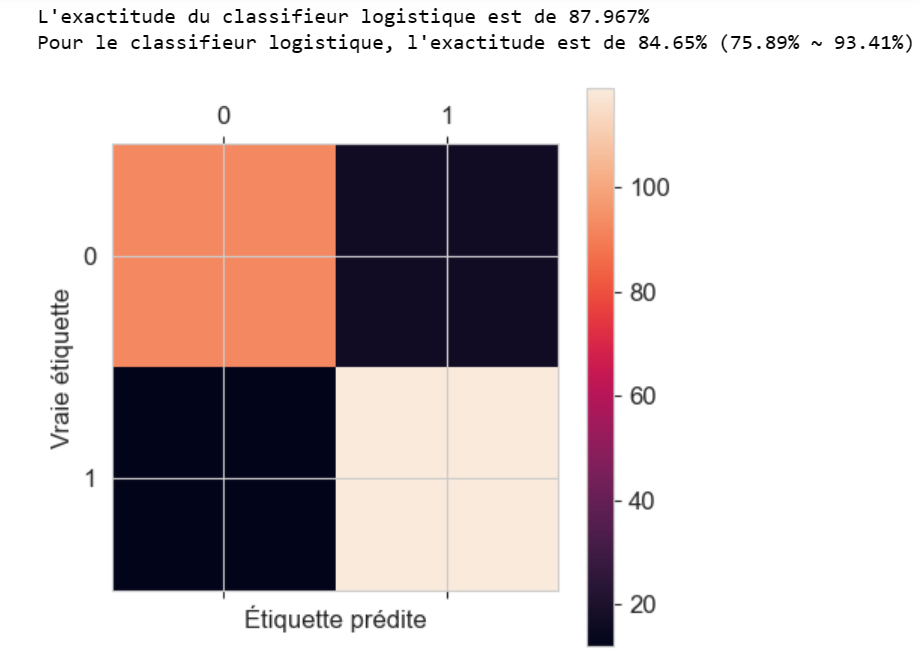
L’on sépare les données selon le ratio 8 : 2.



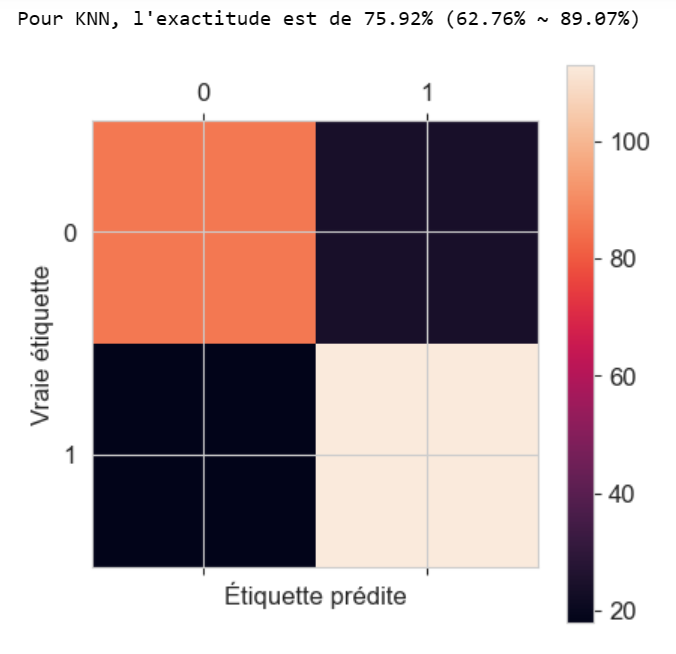
## Modèle d’entrainement

Concernant ce travail, l’on a testé plusieurs modèles différents pour essayer d’en tirer le meilleur possible. Dans ce cas, les résultats par modèles sont les suivant :

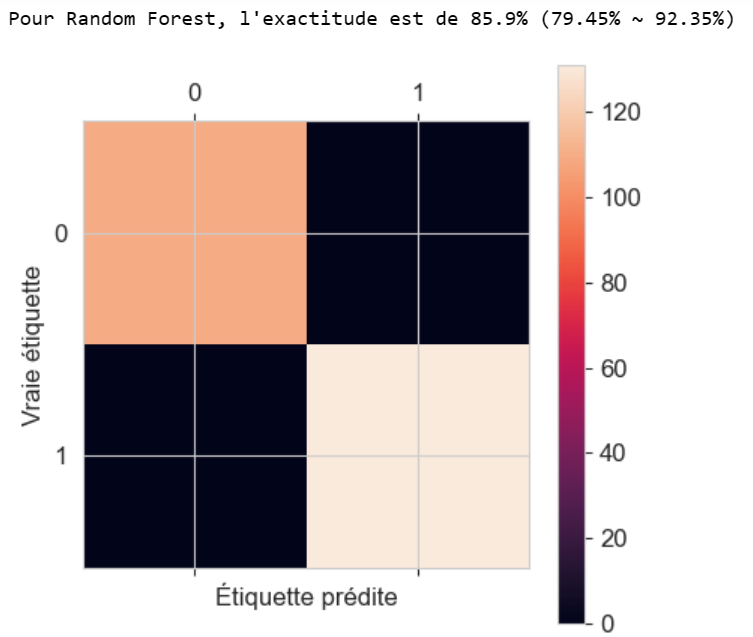
* Régression logistique :



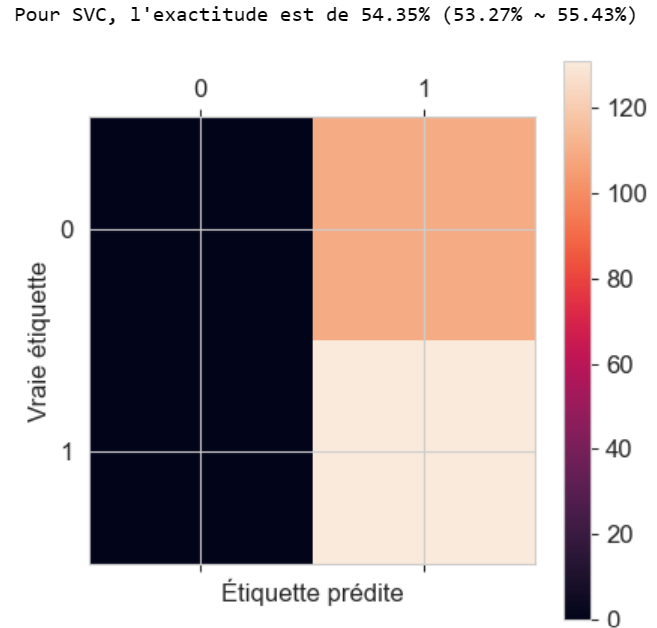
* Classification KNN



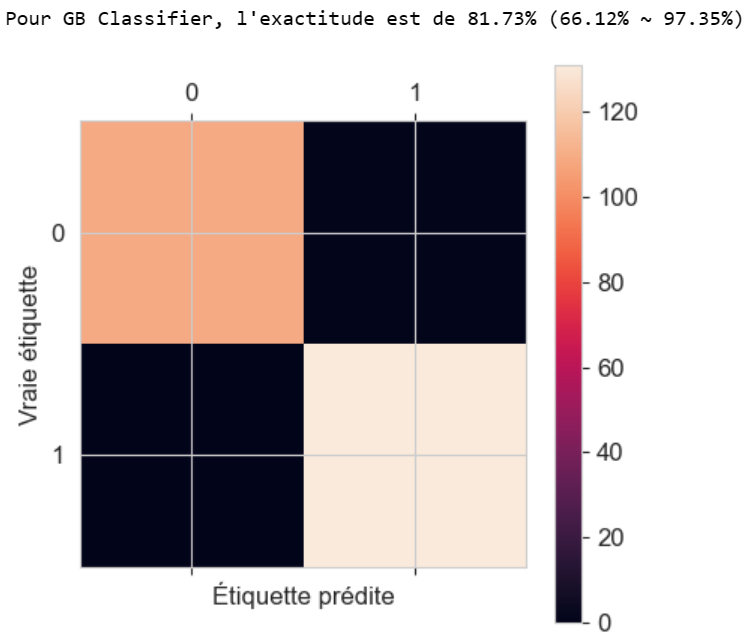
* Ramdom Forest



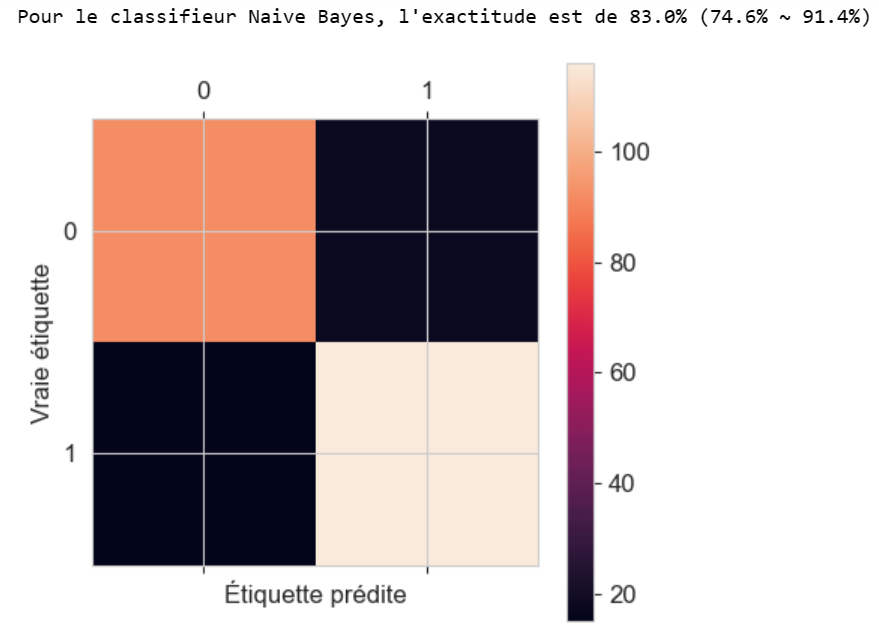
* Classification SVC



* Classification GB



* Classifieur Naive Baiye



L’on remarque pour ce cas que chaque modèle présente des caractéristiques différentes à un point où chacun présente des précisions plus ou moins importante mais l’on retient que le meilleur modèle est celui de la régression logistique avec 87,967%.

# Remarque et commentaire

Avec comme modèle prépondérant celui de la régression logistique, l’on conclut que celui-ci est le meilleur modèle pour l’entraînement. L’on pourrait optimiser les hyperparamètres pour arriver à d’autres précision mais cela ne ferait que nuire au performance si n’est pas bien maîtiser.