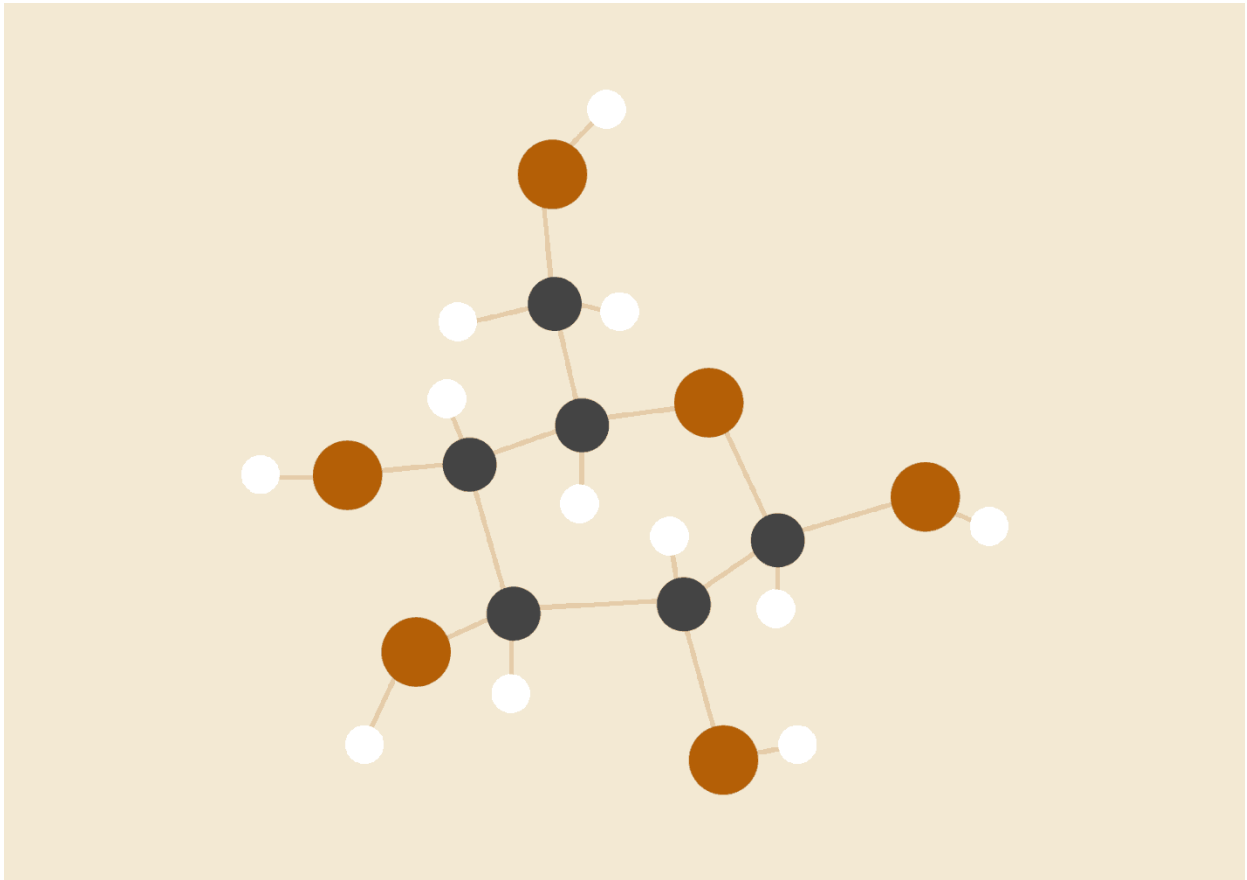


# ***RAPPORT PROJET M8***

- Analyse crise cardiaque –



**SBAI Rokia / BOUHCHOUGH Imane / MATTICHE Othmane /  
REGRAGUI Yasmine**

2020 – 2021

## SOMMAIRE :

<b>1. DESCRIPTION DES DONNÉES.....</b>	<b>3</b>
1.1 La Variable Output .....	3
1.2 La Variable Âge.....	3
1.3 La Variable Sex .....	4
1.4 La Variable CAA.....	4
1.5 Variable exng.....	5
1.6 La variable oldpeak.....	6
1.7 La variable slp.....	6
1.8 Variable Cp .....	7
1.9 Variable Trtbps.....	7
1.10 La variable chol .....	8
1.11 Variable Thalachh.....	8
1.12 La variable fbs .....	9
<b>2. REGRESSION .....</b>	<b>9</b>
2.1 Régression multiple avec toutes les variables.....	9
2.2 Régression multiple en ajoutant une à une les variables.....	9
2.3 Régression multiple en enlevant les points aberrants .....	10
<b>3.TEST STATISTIQUES.....</b>	<b>11</b>
3.1 Test du Chi 2 .....	11
3.2 Test de Student : Pente de régression.....	12
<b>CONCLUSION.....</b>	<b>13</b>
<b>REFERENCES .....</b>	<b>14</b>

## INTRODUCTION

Nous avons derrière nous des centaines d'années de recherches et de découvertes concernant le corps humain. Aussi complexe que fascinant soit-il, ce dernier est un ensemble, parfaitement organisé d'organes. Et malgré tous les progrès médicaux que le monde a connus jusqu'à présent, quelques énigmes persistent à rester irrésolubles. Le cancer est une maladie incurable, le sida, le diabète ... Ce que nous allons aborder n'est pas une maladie incurable mais presque. Les arrêts cardiaques sont des crises qui surgissent soudainement, sans raison diraient certains, d'autres sont persuadés du contraire. Alors, peut-on prédire une crise cardiaque? C'est ce que nous allons voir lors de l'étude d'un jeu de données de patients.

Nous avons travaillé sur des données concernant des patients, ces données regroupent 303 observations et 14 variables. Cependant, pour éviter toute sorte de confusion, nous avons préféré en supprimer 2, par manque d'informations sur ces dernières. Nous nous retrouvons alors avec 12 variables : **age, sex, caa, exng, oldpeak, slp, cp, trtbps, chol, thalachh, fbs et output**. Nous verrons par la suite plus en détail la description de ces variables.

**L'étude complète et bien détaillée du projet se trouve dans le fichier Python .ipynb joint au rapport.**

La variable output est la variable prépondérante, elle montre si un patient est plus susceptible ou non d'avoir une crise cardiaque. C'est la variable qui nous intéressera le plus, vu qu'on étudiera l'impact de chaque variable sur cette variable output, afin de savoir si on peut prédire une crise cardiaque chez un patient. Ci-dessous un aperçu des données:

	age	sex	cp	trtbps	chol	fbs	thalachh	exng	oldpeak	slp	caa	output
0	63	1	3	145	233	1	150	0	2.3	0	0	1
1	37	1	2	130	250	0	187	0	3.5	0	0	1
2	41	0	1	130	204	0	172	0	1.4	2	0	1
3	56	1	1	120	236	0	178	0	0.8	2	0	1
4	57	0	0	120	354	0	163	1	0.6	2	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...
298	57	0	0	140	241	0	123	1	0.2	1	0	0
299	45	1	3	110	264	0	132	0	1.2	1	0	0
300	68	1	0	144	193	1	141	0	3.4	1	2	0
301	57	1	0	130	131	0	115	1	1.2	1	1	0
302	57	0	1	130	236	0	174	0	0.0	1	1	0

*Extraction des 10 premières lignes des données sur les patients*

Le tableau ci-dessous représente les moyennes, médians, variances et écart-types de toutes nos variables quantitatives, après quelques calculs, nous avons obtenu :

	Moyenne	Médian	Variance	Ecart Type
Age	54.366337	55.0	82.484558	9.082101
Sex				
Caa		0.0		
Exng				
Oldpeak	1.039604	0.8	1.348095	1.161075
Slp		1.0		
Cp		1.0		
Trtbps	131.623762	130.0		
Chol	246.264026	240.0		
Thalach	149.646865	153.0		
Fbs				

## 1. DESCRIPTION DES DONNÉES

### 1.1 La Variable Output

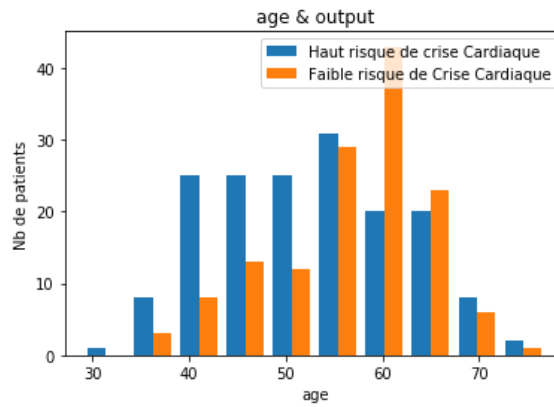
Lors de notre étude, on expliquera la variable output. Dans les données, quand output = 1, cela veut dire que le patient a un très haut risque de crise cardiaque, et l'output = 0 veut dire le contraire. Sur Python, nous avons calculé le nombre de patients qui ont un output à 1 et ceux qui ont un output à 0. On a donc obtenu le résultat suivant :

Haut risque de crise cardiaque	Faible risque de crise cardiaque
165	138

### 1.2 La Variable Âge

Cela n'arrive qu'aux plus vieux ! Est-ce toujours vrai ? On essaiera de répondre à cette question en étudiant cette variable.

Dans l'histogramme, on remarque que les moins de 55 ans ont plus de chance d'avoir une crise cardiaque. Cependant, à partir de 60 ans, on observe un pic (plus de 40 personnes sont exposées à un faible risque d'arrêt cardiaque, contre 20).



*Histogramme âge & output*

### 1.3 La Variable Sex

La variable sex est une variable qui admet 1 pour un homme et 0 pour une femme. Nous avons visualisé la variable sex en fonction de la variable output à travers un histogramme.

En analysant cet histogramme cela n'aura aucun sens de dire que les hommes ont plus de chance d'avoir une crise cardiaque que les femmes. Car d'après nos calculs, on a dans nos données 207 hommes et 96 femmes. Donc on ne peut pas tirer de conclusion hâtive de cet histogramme. C'est pour cette raison qu'on a décidé de calculer les rapports :

On a donc obtenu pour :

- Les hommes :  $= \frac{93}{114} = 0.8157894736842105$
- Les femmes :  $\frac{72}{24} = 3.0$

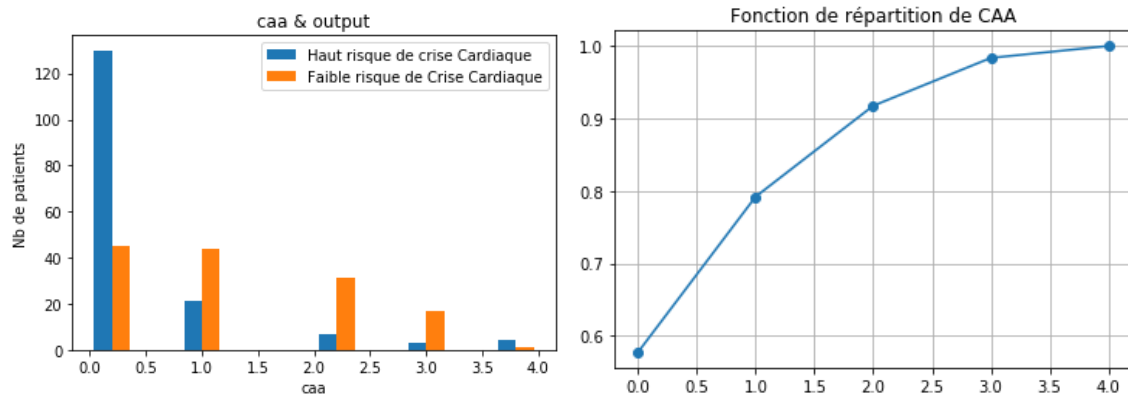
On conclut donc que : le rapport pour les femmes est supérieur à celui des hommes. Ce qui veut dire que les femmes sont plus susceptibles d'avoir des crises cardiaques que les hommes.

### 1.4 La Variable CAA

Cette variable fait référence au nombre des grands vaisseaux sanguins du cœur, c'est-à-dire les plus grosses artères et veines qui se fixent aux oreillettes et aux ventricules et transportent le sang vers et depuis le système circulatoire systémique et le système circulatoire pulmonaire. Les principaux vaisseaux sanguins connectés à votre cœur sont l'aorte, la veine cave supérieure, la veine cave inférieure, l'artère pulmonaire, les veines pulmonaires et les artères coronaires.

L'histogramme montre qu'un grand nombre de personnes qui ont un haut risque de crise cardiaque ont peu de grands vaisseaux sanguins. Plus de 120 de ces personnes là n'ont aucun

de ces vaisseaux cités précédemment. Ceci est très bien illustré aussi par les boîtes à moustaches. On obtient aussi une fonction de répartition croissante. (Voir fichier .ipynb)



*Histogramme et fonction de répartition caa & output*

## 1.5 Variable exng

La variable exng désigne les personnes qui ont eu une angine de poitrine , ou angor. C'est un symptôme cardiaque qui désigne la douleur ou l'inconfort thoracique causé par une interruption temporaire de l'apport de sang oxygéné au cœur. Généralement, l'angine de poitrine :

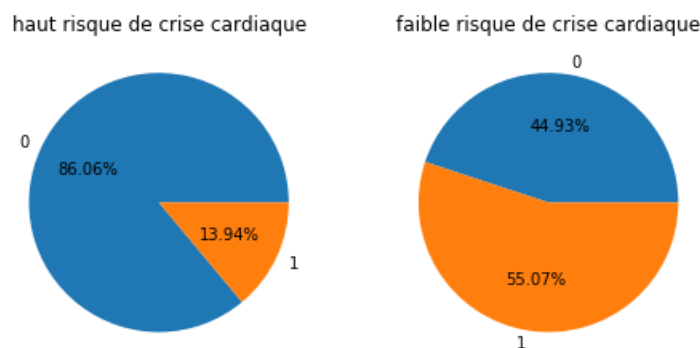
- disparaît au repos ou à la prise de médicaments;
- est le signe d'un risque accru de subir une crise ou un arrêt cardiaque.

Cette variable prend deux valeurs :

- La valeur 1 : Quand une personne a déjà fait une crise d'angor.
- La valeur 2: Quand une personne n'a jamais fait de crise d'angor.

On a tout d'abord divisé l'échantillon en deux catégories. Les personnes ayant eu une crise cardiaque et ceux qui n'en ont pas eu. On a ensuite étudié si ces personnes ont déjà eu une crise angor, nous avons donc obtenu deux camemberts.

On remarque que 86.06% des personnes qui ont un haut risque de faire une crise cardiaque n'ont jamais eu de crise d'angor. Alors que plus de la moitié (55.07%) des personnes qui ont un faible risque de crise cardiaque ont déjà eu une crise d'angor.



## 1.6 La variable oldpeak

La variable oldpeak est une variable qui est classé comme suit :

Rang	Risque
<1.5	Faible
1.5-4.2	Elevé
2.55 >	Très Élevée

L'histogramme et les boites à moustache (voir fichier .ipynb) nous montrent que 100 individus étaient exposés à un risque faible et que le reste se trouve entre 0.2 et 6. On remarque qu'entre 0 et 1.5 (risque faible) le nombre d'individus est plus élevé. Et on remarque aussi une légère densité entre 1.5 et 6 (risque élevé à très élevé).

## 1.7 La variable slp

Il existe deux types de crise cardiaque :

Infarctus du myocarde avec élévation du segment ST (STEMI)

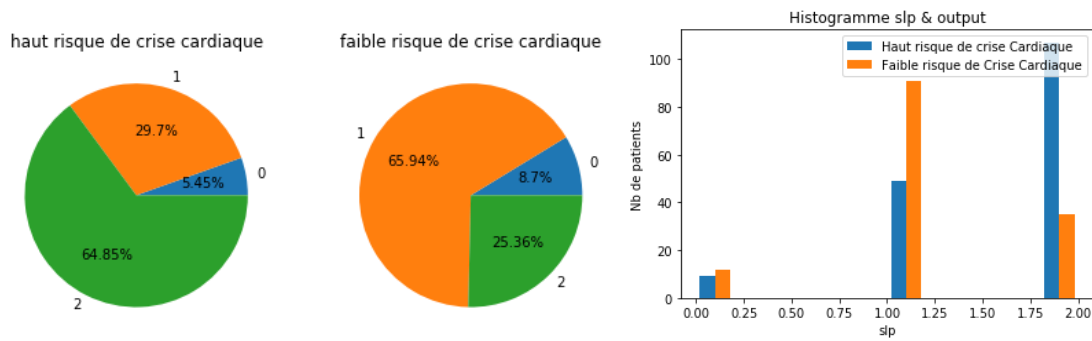
- Il se produit lorsque l'obstruction artérielle est totale, c'est-à-dire que la circulation sanguine dans une artère coronaire transportant le sang riche en oxygène vers le cœur est complètement interrompue.
- Puisque l'apport de sang oxygéné est interrompu, toute l'épaisseur du muscle cardiaque qui est alimentée par cette artère s'endommage, ce qui peut entraîner la mort de celui-ci.
- Cela se voit sur l'électrocardiogramme (ECG), qui affichera une élévation du segment ST correspondant au siège de la lésion à la paroi du cœur.

Infarctus du myocarde sans élévation du segment ST (NSTEMI)

- Il se produit lorsque l'obstruction artérielle est partielle, c'est-à-dire que la circulation sanguine dans une artère coronaire transportant le sang riche en oxygène vers le cœur n'est pas complètement interrompue.
- Seule une partie de l'épaisseur de la paroi du cœur s'endommage.
- L'ECG n'affichera aucune élévation du segment ST.

La variable slp représente ce segment ST et elle prend 3 valeurs différentes : (0 = ascendant; 1 = flat; 2 = descendant).

On pourrait conclure à travers les camemberts que les personnes à haut risque avaient majoritairement un slp descendant, et que celles qui ont un faible risque avaient en majorité un slp qui est plat (donc pas de signe d'infarctus).

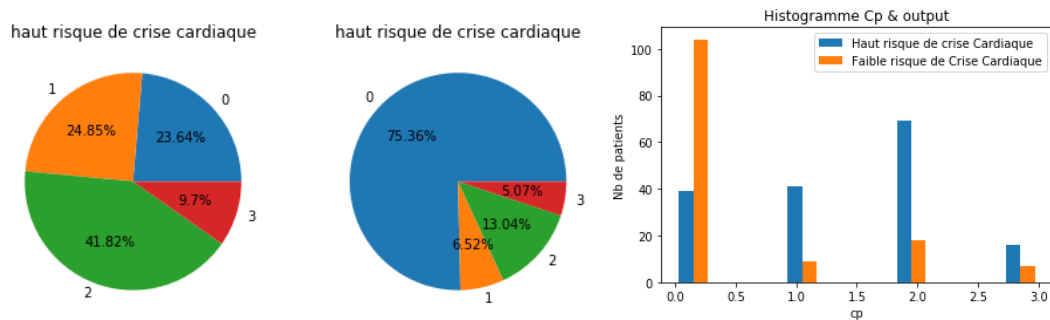


*Camemberts et histogramme slp & output*

## 1.8 Variable Cp

Cette variable représente le type de la douleur thoracique (Chest Pain). Cp = 1 représente une angine de poitrine typique, cp =2 est une angine de poitrine atypique, cp = 3 quand ce n'était pas une angine de poitrine et 0 représente les personnes qui sont symptomatiques.

On arrive à voir que la variable output a une relation avec le type de cp. On remarque que les personnes qui ont un haut risque de crise cardiaque sont celles qui ont eu des douleurs thoraciques de type 1 et 2. Et que la majorité des personnes asymptomatiques sont celles qui ont un faible risque de crise cardiaque.



*Histogramme et camemberts cp & output*

## 1.9 Variable Trtbps

Cette variable correspond à la tension artérielle au repos en mmHg. D'après nos recherches on a trouvé que la pression artérielle est donnée sous forme de 2 chiffres:

- **Pression systolique** : La pression lorsque votre cœur expulse le sang.
- **Pression diastolique** : La pression lorsque votre cœur se repose entre les battements.

Par exemple, si votre tension artérielle est 140/90mmHg, cela signifie que vous avez une pression systolique de 140mmHg et une pression diastolique de 90mmHg.

Voici comment se classifie cette variable :

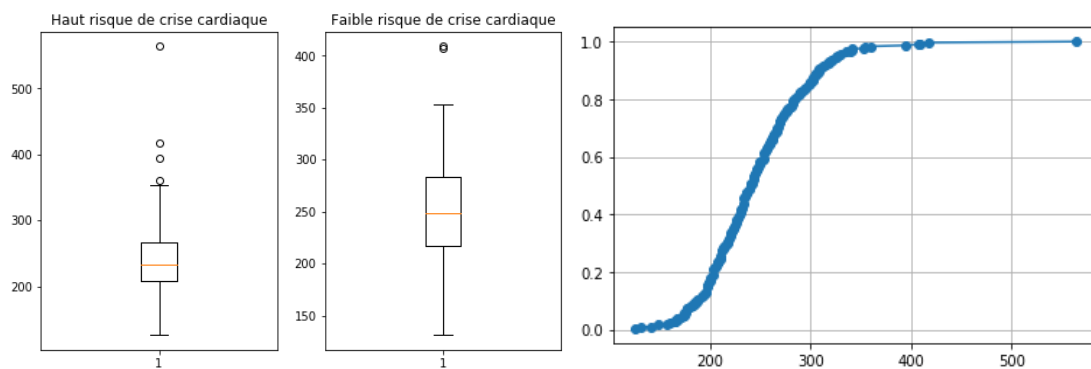


Pression artérielle <u>idéale</u>	Entre 90/60mmHg et 120/80mmHg
<u>Hypertension</u> artérielle	140/90mmHg ou plus
Pression artérielle <u>faible</u>	90/60mmHg ou moins.

On remarque que les moyennes de cette variable pour les patients à haut et à faible risque de crise cardiaque se rapprochent et sont toutes les deux supérieures à 140/90 mmHg. Ces conclusions nous ont mis un doute, mais on ne peut toujours rien conclure. Il faudrait faire la régression ainsi que des tests pour en tirer une conclusion plus exacte et voir si cette variable a une relation avec la variable output.

### 1.10 La variable chol

Cette variable représente le taux de cholestérol chez les patients en mg/dl.

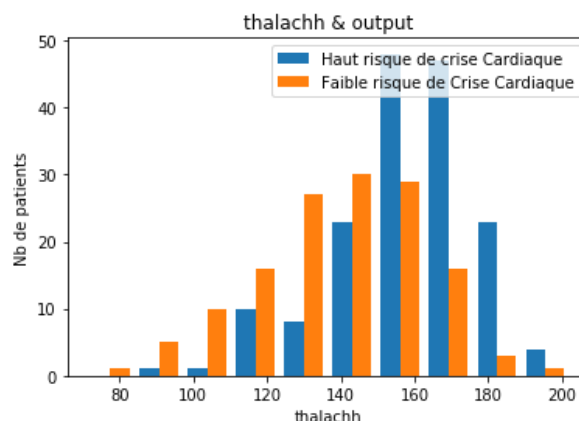


*Boîtes à moustache et fonction de répartition chol & output*

### 1.11 Variable Thalachh

Cette variable représente la fréquence cardiaque maximale atteinte.

On en conclut que plus thalachh augmente, plus le nombre de personnes à haut risque de crise cardiaque augmente. La tendance s'inverse pour les personnes à faible risque. On pourrait dire que la fréquence cardiaque maximale atteinte chez les personnes à haut risque est plus élevée. Dans ces figures, on voit clairement une relation entre thalachh et output.



### 1.12 La variable fbs

Cette variable fait référence au taux de sucre dans le sang du patient à jeun. Si le taux est  $>120$  mg/dl alors , fbs = 1 sinon fbs= 0.

On remarque que les patients à haut risque sont à peu près autant susceptibles que ceux à faible risque d'avoir un fbs $>120$ mg/dl.

Afin d'avoir une vision un peu plus claire de ces résultats, on a calculé les ratio pour obtenir :

- fbs $>120$ mg/dl(fbs=1):  $\frac{\text{Haut risque crise cardiaque}}{\text{Faible risque Crise cardiaque}} = 1.0455$
- fbs $<120$ mg/dl(fbs=0) :  $\frac{\text{Haut risque crise cardiaque}}{\text{Faible risque Crise cardiaque}} = 1.2241$

On ne remarque pas de grande différence entre les deux ratios, cela affirme la conclusion qu'on a fait précédemment. On pourrait dire que les personnes à haut risque sont un peu plus susceptibles d'avoir un fbs  $>120$  mg/dl.

## 2. REGRESSION

### 2.1 Régression multiple avec toutes les variables

On cherche à savoir si on peut prévoir le risque de faire une crise cardiaque (variable output) en fonction de plusieurs variables qui nous indiquent l'âge , le sexe , la douleur thoracique (cp) , la fréquence cardiaque maximale atteinte (thalachh) ... Nous avons donc fait la régressions de toutes les variables en fonction de output .Nous avons suivi le modèle suivant :

$$\text{output} = a_0 + a_1 \cdot \text{age} + a_2 \cdot \text{sex} + a_3 \cdot \text{caa} + a_4 \cdot \text{exng} + a_5 \cdot \text{oldpeak} + a_6 \cdot \text{slp} + a_7 \cdot \text{cp} + a_8 \cdot \text{trtbps} + a_9 \cdot \text{chol} + a_{10} \cdot \text{thalachh} + a_{11} \cdot \text{fbs}$$

Sous forme matricielle, cela nous donne :

$$y = X \cdot a \text{ avec } y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \text{ et } a = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \text{ et } X = \begin{pmatrix} 1 & \cdots & fbs_1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & fbs_n \end{pmatrix}$$

On a obtenu un  $R^2 = 0.496885718703031$  qui n'est pas satisfaisant vu que ce dernier est très éloigné de 1. Nous allons donc faire une régression en ajoutant les variables une à une.

### 2.2 Régression multiple en ajoutant une à une les variables

En ajoutant une à une les variables, nous obtenons le tableau suivant :

Variables	Modèles	R2
Age	$a_0 + a_1.age$	0.05082261
Sex	$a_0 + a_1.sex$	0.07892536
Cp	$a_0 + a_1.cp$	0.18818093
Trtbps	$a_0 + a_1.trtbps$	0.02100503
Chol	$a_0 + a_1.chol$	0.00726571
Fbs	$a_0 + a_1.fbs$	0.00078656
Thalachh	$a_0 + a_1.thalachh$	0.17786542
Exng	$a_0 + a_1.exng$	0.1907567
Oldpeak	$a_0 + a_1.oldpeak$	0.18549905
Slp	$a_0 + a_1.slp$	0.11963095
Caa	$a_0 + a_1.caa$	0.15344769

On constate que la variable exng est la variable avec le plus grand R2 suivie par cp. Nous allons faire la régression linéaire de ces deux variables uniquement. Nous allons suivre le modèle suivant :

$$output = a_0 + a_1.cp + a_2.exng$$

Sous forme matricielle, cela nous donne :

$$y = X.a \text{ Avec } y = \begin{pmatrix} y_0 \\ y_1 \\ y_2 \end{pmatrix} \text{ et } a = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} \text{ et } X = \begin{matrix} 1 & cp_0 & exng_0 \\ 1 & cp_1 & exng_1 \\ 1 & cp_2 & exng_2 \end{matrix}$$

On obtient un  $R^2 = 0.022$ .

On remarque que même après avoir calculer le  $R^2$  relatif aux deux plus grandes variables, il reste assez petit. Ce qui nous pousse à dire qu'il y'a des valeurs aberrantes à retirer du modèle.

### 2.3 Régression multiple en enlevant les points aberrants

Après avoir détecté les points aberrants du modèle, grâce au script python, on trouve alors qu'on a 23 points à enlever. On les enlève et après avoir fait la régression sur le nouveau modèle, notre  $R^2$  devient  $R^2 = 0.607$ . Ce qui est une bonne amélioration, donc la qualité de ce modèle est bien meilleur que celle du modèle précédent.

### 3.TESTS STATISTIQUES

#### 3.1 Test du Chi 2

Nous avons testé l'indépendance de nos variables qualitatives avec la variable output afin de voir la relation d'indépendance entre elles.

Les hypothèses sont  $\begin{cases} H_0 = \text{indépendance} \\ H_1 = \text{dépendance} \end{cases}$

On a ensuite calculé la p-valeur, si cette dernière est inférieure à alpha (0.05) on rejette  $H_0$  et on dit que les deux variables sont dépendantes. Sinon, on garde  $H_0$  et on dit que les deux variables sont indépendantes.

Nous avons obtenu les tableaux de contingences et les p-valeurs suivants:

Variable Sex :

	Output=1	Output=0
Hommes	93	114
Femmes	72	24

p-valeur = 1.8767776216941503e-06

On rejette l'hypothèse  $H_0$ , sex et output ne sont pas indépendants.

Variable Exng :

	Output=1	Output=0
exng=1	23	76
exng=0	142	62

p-valeur = 7.45440933123568e-14

$H_0$  est rejetée, exng et output ne sont pas indépendants.

Variable Slp :

	Output=1	Output=0
slp0	9	12
slp1	49	91
slp2	107	35

p-valeur = 0.0029974669363164934

$H_0$  est rejetée, slp et output ne sont pas indépendants

Variable fbs :

	Output=1	Output=0
fbs=1	23	22
fbs=0	43	116

p-valeur = 0.7444281114149577

$H_0$  est gardée, fbs et output sont indépendants.

Variable Cp :

	Output=1	Output=0
cp0	39	104
cp1	41	9
cp2	69	18
cp3	16	7

p\_value = 1.3343043373050064e-17

$H_0$  est rejetée, cp et output ne sont pas indépendants

Variable Caa :

	Output=1	Output=0
caa0	130	45
caa1	21	44
caa2	7	31
caa3	3	17
caa4	4	1

p-valeur :2.712470211959332e-15

$H_0$  est rejetée, caa et output ne sont pas indépendants

Ce test prouve que toutes les variables qualitatives sont liées à output, mis à part fbs.

### 3.2 Test de Student : Pente de régression

Pour s'assurer de l'exactitude de nos coefficients et prouver que cela n'est pas du au fait du hasard, nous allons effectuer un test de Student. Nous avons donc fait 13 tests, autrement dit chaque variable avec la variable output.

Les hypothèses sont :  $\begin{cases} H0 = \text{indépendance} & a = 0 \\ H1 = \text{dépendance} & a \neq 0 \end{cases}$

Comme vu précédemment, nous avons déjà calculé la variance estimée, la variance calculée et les deux coefficients a et b, d'où :  $t = \frac{a}{\sqrt{\frac{\sigma^2}{Sx^2}}}$

Le nombre de degrés de liberté étant  $n - 2 = 301$ , la p-valeur se calcule avec la fonction `stats.t.pdf(t,ddl)` de python. On décide de garder  $H0$  ( $a = 0$ ) si la p-valeur est supérieure à 0,05 ( $pval \geq 0,05$ ), sinon on rejette l'hypothèse nulle.

Nous avons obtenu des p-valeur presque toutes inférieures à alpha, qui dans notre cas est égale à 5%, on peut donc conclure que notre coefficient est correct, il y a 5 chances sur 100 qu'il soit dû au hasard, ce qui nous semble relativement bas.

Ajoutant à cela, que la connaissance des variables chol (quantitative) et fbs (qualitative) ne permet pas de prévoir les valeurs de output. Ce qui confirme le test de Chi2 fait précédemment.

## CONCLUSION

Ce projet a été une réelle opportunité de mettre en pratique ce que nous avons vu en cours tout au long du semestre. Nous avons pu traiter des données scientifiques et réelles et les analyser grâce aux méthodes et à nos connaissances. Cette analyse a porté sur 3 parties importantes : la description des 12 variables avec lesquelles nous avons travaillé, la partie de la régression multiple que nous avons effectué de 3 manières différentes (toutes les variables, en ajoutant les variables une à une puis en enlevant les points aberrants) et enfin une partie sur les tests statistiques.

D'après nos tests statistiques, nous pouvons enfin répondre à la question posée au tout début : toutes nos variables peuvent influencer le fait d'avoir une crise cardiaque, à l'exception de chol et fbs, qui ne dépendent pas de output.

De plus, réaliser ce projet nous a non seulement permis d'accroître nos connaissances personnelles et nos connaissances en tant que futurs ingénieurs, il nous a également permis d'établir une bonne cohésion de groupe. Etant donné qu'au début nous avions tous des idées différentes. Cependant, nous avons su mettre à profit la créativité et les atouts de chacun afin de mener à bien ce projet. Nous avons aussi pu développer des aspects de notre personnalité tels que la prise d'initiative et le respect des délais.

Toutefois, nous avons rencontré quelques difficultés lors de la réalisation de ce projet du fait qu'il fallait trier les données, savoir ce qui nous était utiles ou non. Cependant, la motivation de chacun nous a permis de surmonter les difficultés et cela nous a donné une idée concrète sur nos futurs projets dans les années à venir.

## REFERENCES

- <https://www.nhs.uk/common-health-questions/lifestyle/what-is-blood-pressure/#:~:text=As%20a%20general%20guide%3A,be%2090%2F60mmHg%20or%20lower>
- [https://www.researchgate.net/figure/CLASSIFICATION-OF-OLD-PEAK\\_tbl2\\_44260568](https://www.researchgate.net/figure/CLASSIFICATION-OF-OLD-PEAK_tbl2_44260568)
- <https://www.coeuretavc.ca/maladies-du-coeur/problemes-de-sante/crise-cardiaque>
- [https://fr.wikipedia.org/wiki/Angine\\_de\\_poitrine?fbclid=IwAR0jBNNNPXpLz0VEFA0WdcjRSIsTW\\_mbOgu5kLIDraEe47NS2s\\_LVyl\\_KNM](https://fr.wikipedia.org/wiki/Angine_de_poitrine?fbclid=IwAR0jBNNNPXpLz0VEFA0WdcjRSIsTW_mbOgu5kLIDraEe47NS2s_LVyl_KNM)
- [https://fr.wikipedia.org/wiki/Infarctus\\_du\\_myocarde?fbclid=IwAR2W-rUQecooAdqJ9qjl5y04\\_m-Mjtv5h-m4QmcWuExjjOpPOdb33ZJX\\_6w#:~:text=Le%20sympt%C3%B4me%20est%20la%20douleur,d'angine%20de%20poitrine](https://fr.wikipedia.org/wiki/Infarctus_du_myocarde?fbclid=IwAR2W-rUQecooAdqJ9qjl5y04_m-Mjtv5h-m4QmcWuExjjOpPOdb33ZJX_6w#:~:text=Le%20sympt%C3%B4me%20est%20la%20douleur,d'angine%20de%20poitrine)
- [https://fr.wikipedia.org/wiki/Pression\\_art%C3%A9rielle?fbclid=IwAR0OWz\\_tGaoj9zaL-bEysWuCGviEGQqMPmNRe4NCRxAJLcM3ifkGfZkwIPA](https://fr.wikipedia.org/wiki/Pression_art%C3%A9rielle?fbclid=IwAR0OWz_tGaoj9zaL-bEysWuCGviEGQqMPmNRe4NCRxAJLcM3ifkGfZkwIPA)