Projet GoMycode : Heart Diseases (Comparaison entre les Technique d'apprentissage supervisé)

Introduction :

Il s'agit d'un dataset Heart Diseases de la plateforme UCI contenant un ensemble de données **303** tuples (rows) et de **14** attributs.

Chaque rows décrit un état d'un patient.

Le dataset contient un attribut cible (Target) cet attribut est binaire et signifie **0 : Patient non_atteint de maladie cardiaque.** **1 : Patient atteint de maladie cardiaque.**

Explication des attributs :

Chaque aatributs présent dans le dataset représente les symptôme clinique (symptôme perceptible) d'un patient ou biologique (bilan) ou bien le résultat d'un ECG effectué sur le patient.

Attributs :

1- Age : l'âge de la personne en années
2- Sexe : 1 : homme 0 : femme.
3- Cp : type de douleur thoracique contient 4 valeurs (Valeur 0: asymptomatique,Valeur 1: angor atypique, Valeur 2: douleur non angineuse, Valeur 3: angor typique)
4- Trestbps : tension artérielle au repos de la personne.
5- Chol : mesure de cholesterole
6- Fbs : glycémie á jeun de la personne.
7- Restecg : résultats électrocardiographiques au repos (0 : montre une hypotrophie ventriculaire gauche probable, 1 : normale, 2 : présentant une anomalie de l'onde ST-T).
8- Thalach : fréquence cardiaque maximale atteinte par la personne.
9- Exang : angor induite par l'exercice (1 : oui, 0 : non).
10- Oldpeak : dépression ST induite par l'exercice par rapport au repos
11- Slope(pente) : la pente du pic de l'exercice segment ST 0 : descente 1 :plat 2 : ascendant.
12- Ca : nombre de vaisseaux principaux
13- Thal : trouble sanguin 0 : supprimé 1 pas de flux dans la partie du cœur 2 : flux normal 3 : flux observé mais pas normal.

Data Visualization

1- En premier lieu on commence par importer les données qui sont sous format csv.
2- On commence á visualiser les 5 premières ligne de notre dataset grâce á la fonction head()

	age	sex	cp	trestbps	chol	fbs	...	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	...	0	2.3	0	0	1	1
1	37	1	2	130	250	0	...	0	3.5	0	0	2	1
2	41	0	1	130	204	0	...	0	1.4	2	0	2	1
3	56	1	1	120	236	0	...	0	0.8	2	0	2	1
4	57	0	0	120	354	0	...	1	0.6	2	0	2	1

3- Vérifier les plages de valeurs pour chaque attributs grâce á la fonction unique()
Par exemple pour l'attribut âge il y'a 41 valeurs possibles et pour sexe il ya 2 valeurs (0 : homme 1 : femme)

```
age          41
sex           2
cp            4
trestbps     49
chol        152
fbs           2
restecg       3
thalach      91
exang         2
oldpeak      40
```

4- On va résumer pour chaque attributs les valeurs count, mean, standard deviation, min, max valeurs dans un tableau récapitulatif → fonction describe()

	age	sex	cp	...	ca	thal	target
count	303.000000	303.000000	303.000000	...	303.000000	303.000000	303.000000
mean	54.366337	0.683168	0.966997	...	0.729373	2.313531	0.544554
std	9.082101	0.466011	1.032052	...	1.022606	0.612277	0.498835
min	29.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000
25%	47.500000	0.000000	0.000000	...	0.000000	2.000000	0.000000
50%	55.000000	1.000000	1.000000	...	0.000000	2.000000	1.000000
75%	61.000000	1.000000	2.000000	...	1.000000	3.000000	1.000000
max	77.000000	1.000000	3.000000	...	4.000000	3.000000	1.000000

5- On vérifie ensuite si on n'a pas de missing values (valeurs manquante) grâce á la fonction isnull().sum() et isnull().values.any() → pour cette dernière elle retourne vrai ou faux (False si tts les attributs ne possèdent pas de valeurs manquantes comme dans notre cas)

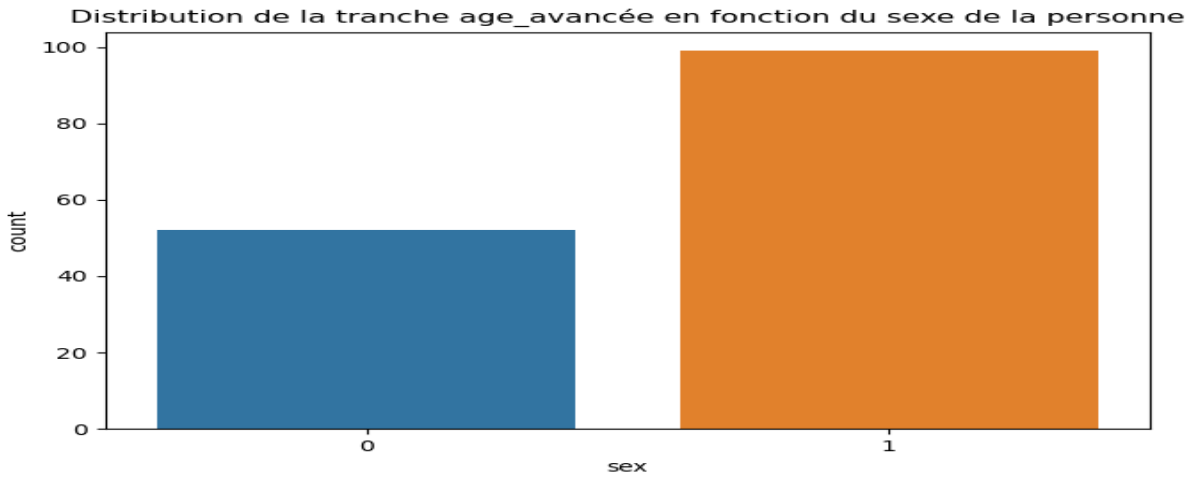
```
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64
False
```

6- Matrice de corrélation permet de voir les corrélations entre toutes les variables. En quelques secondes, on peut voir si un attribut est positivement ou négativement corrélé avec notre prédicteur (cible), comme par exemple l'attribut CP (douleur thoracique est + corrélé avec le target ce qui est logique car plus on a une douleur thoracique plus on est sujet á une maladie cardiaque)

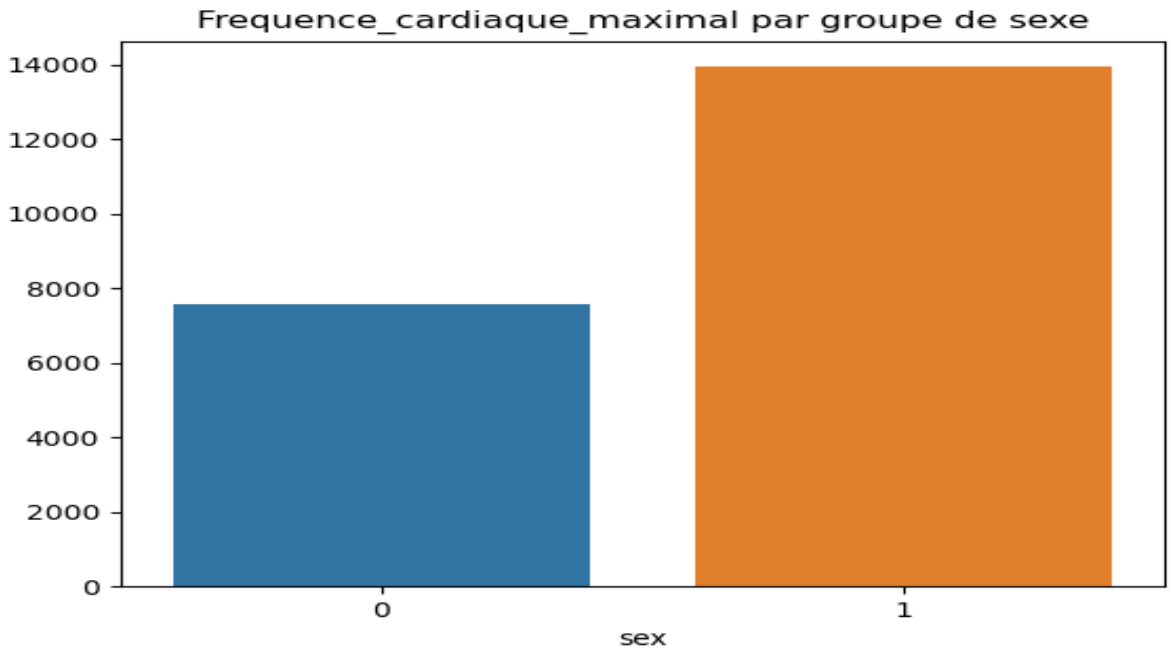
7- L'attribut Age :
J'ai diviser les patients en 3 catégorie :

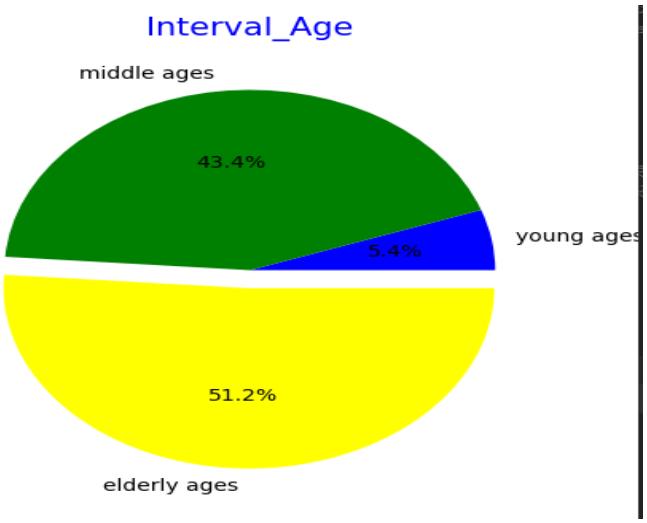
Personnes_Age_jeunes entre 29ans et inferieur á 40 → on n'a pas beaucoup de patients dans cette catégorie **16**
Personne_Age_moyen entre 40 et inf á 55 → on compte plus ou moins beaucoup de personne **128**
Personne_Age_Senior sup á 55 → on compte beaucoup de personne dans cette catégorie **151**

- **Distribution de l'interval Age_Senior (Pour les sénior on observe que les femmes sont plus nombreuses que les hommes)**



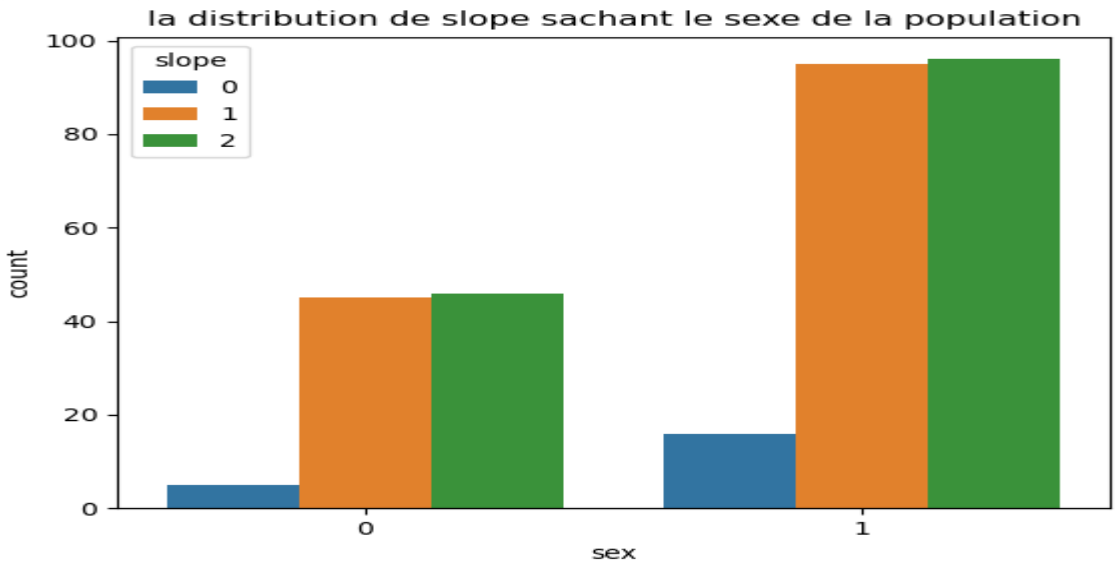
- **Fréquence cardiaque maximal on observe qu'elle est plus fréquente chez les femmes d'un âge sup á 55ans**



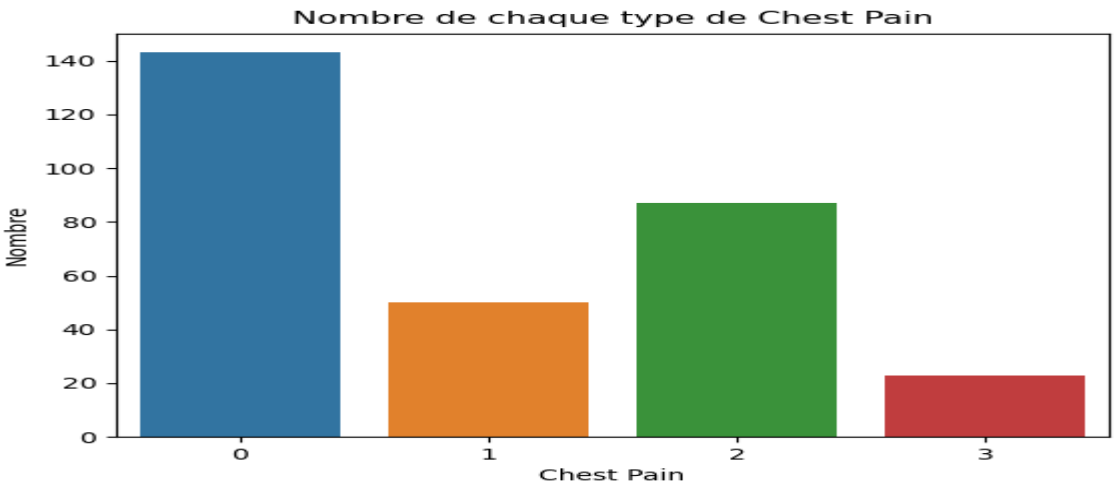
- 

8- L'attribut Sexe :

- On remarque que les femmes sont plus nombreuses que les hommes (patients)
- Pour l'attribut ST on voit que les femmes sont plus susceptibles á une malade cardiaque car la pente du segment ST 0 est plus élevé chez les femmes



9- L'attribut ChestPain (douleur thoracique)



10- Filtring des données selon les patients pos ou neg

Nous remarquons que les patients atteint de la maladie cardiaque(pos) ont tendance á avoir une Fréquence cardiaque très élevés et un ST qui tend vers 0 ce qui prouve une dépression induite par un exercice par rapport au repos

```
(Positive Patients ST depression): 0.583030303030303
(Negative Patients ST depression): 1.5855072463768116
(Positive Patients thalach): 158.46666666666667
(Negative Patients thalach): 139.1014492753623
```

Modeling /Training

1- On commence par séparer les attributs (features) de l'attribut cible (target)
2- Ensuite on fait un split pour séparer les données test des données apprentissage
3- On procède ensuite á une normalisation scaling data
4- Nous procédons ensuite á la construction du modèle.

Matrix de confusion explication

24 est le nombre de vrais positifs dans nos données, tandis que 29 est le nombre de vrais négatifs.5 et 3 sont le nombre d'erreurs.

Il y a 5 erreurs de type 1 (Faux positifs) - Vous avez prédit un positif et c'est faux. Il y a 3 erreurs de type 2 (Faux négatifs) - Vous avez prédit un résultat négatif et c'est faux.

Par conséquent, si nous calculons la précision, c'est # Correct Predicted/ # Total. En d'autres termes, où TP, FN, FP et TN représentent le nombre de vrais positifs, de faux négatifs, de faux positifs et de vrais négatifs

(TP + TN)/(TP + TN + FP + FN). (24+29)/(24+29+5+3) = 0,87 = précision de 87%.

Features Importance
D'après le graphique de l'importance des caractéristiques ci-dessus, nous pouvons conclure que les 4 caractéristiques les plus importantes étaient le type de douleur thoracique (cp), la fréquence cardiaque maximale atteinte (thalach), le nombre de vaisseaux principaux (ca) et la dépression du ST induite par l'exercice par rapport au repos (oldpeak).

Prenons un scénario
C'est un homme de 20 ans, avec une douleur thoracique d'une valeur de 2 (angine atypique), et une tension artérielle au repos de 110.
Il a en outre un taux de cholestérol sérique de 230 mg/dl.
Il a une glycémie à jeun > 120 mg/dl.
Il a un résultat électrocardiographique au repos de 1.
La fréquence cardiaque maximale atteinte par le patient est de 140.
De plus, il souffre d'une angine de poitrine provoquée par l'exercice.
Sa dépression ST induite par l'exercice par rapport à la valeur au repos était de 2,2.
La pente du segment ST de pointe à l'effort est plate.

Il n'a pas de vaisseaux principaux colorés par fluoroscopie, et en outre, sa fréquence cardiaque maximale atteinte est un défaut réversible.

Sur la base de ces informations, on va voir si notre système peut classer notre patient ?

Conclusion

Avec l’algorithme Random forest on a atteint une précision d’apprentissage qui est supérieur á 80% comparait á Logistic reg qui est á 62 % et Le meilleure k des knn a obtenu un score de 75%
Sur les 13 caractéristiques que nous avons examinées, les 4 principales caractéristiques significatives qui nous ont permis de classer un diagnostic positif ou négatif sont le type de douleur thoracique (cp), la fréquence cardiaque maximale atteinte (thalach), le nombre de vaisseaux principaux (ca) et la dépression du ST induite par l'exercice par rapport au repos (oldpeak).
Notre algorithme d'apprentissage machine peut maintenant classer les patients atteints de maladies cardiaques. Nous pouvons maintenant diagnostiquer correctement les patients et leur apporter l'aide dont ils ont besoin pour se rétablir. En diagnostiquant ces caractéristiques à un stade précoce, nous pouvons éviter que des symptômes plus graves ne se manifestent plus tard.

Merci.