Rendu Machine learning

Groupe:

Nael Mezdari

Amine Baha Eddine

Sayed Drissi

Hugo Sinprasith

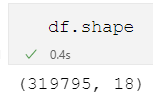
[Lien du Repo](https://github.com/nlmz94/machineLearningHeartDisease/tree/test_branch) (branche “test\_branch”)

Comment on a abordé le problème

Tout d’abord, nous avons fait un bilan des compétences de chacun. Ainsi, nous avons pu découvrir les affinités de chacun pour travailler efficacement.

Ensuite, nous avons analysé ensemble le jeu de données pour définir les bases et pour s’assurer qu’on parte tous sur la même longueur d’onde.

Enfin, nous avons chacun de notre côté essayé de manipuler le jeu de données avec Jupyter, Pandas, etc… car même si nous savions ce qu’on voulait réaliser, nous avions besoin de pratiquer pour au moins avoir les bases.

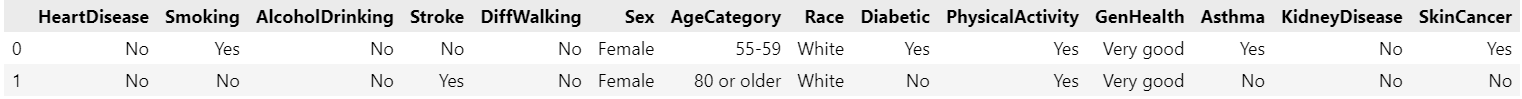
Exploration du dataset

Forme : 319795 lignes et 18 colonnes

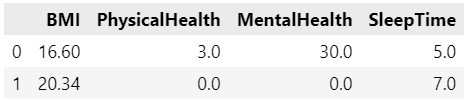
Types des différentes colonnes (float64 ou object)

Deux types de données :

* Variables de catégories (14)



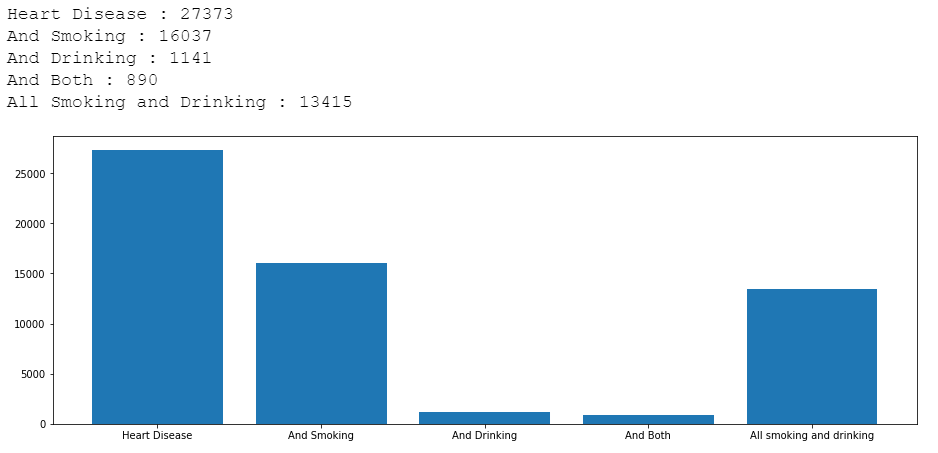
* Variables discrètes (4)



Aucune valeur nulle.

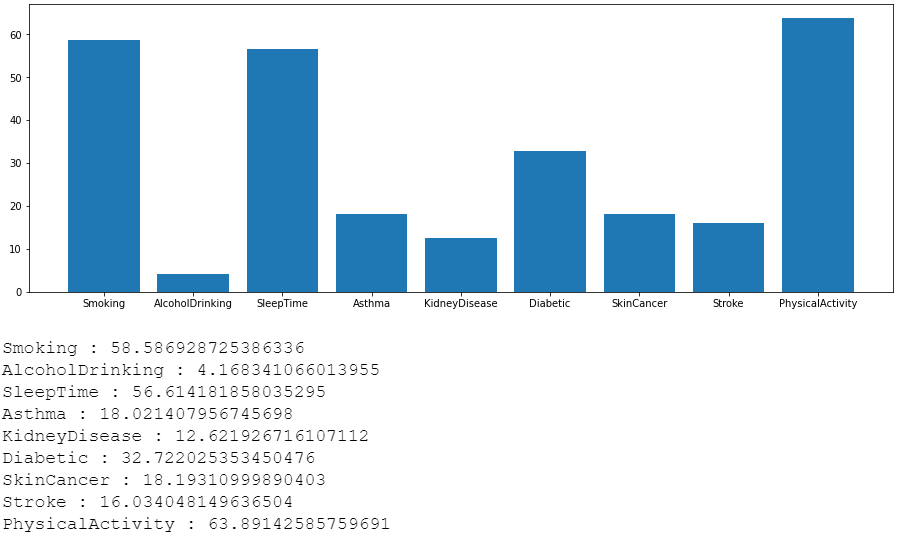
Remarques supplémentaires :

* 27373 personnes sur les 319795 qui souffrent de maladies cardiaques.
* Sur ces 27373 personnes, 16037 d'entre elles fument, 1141 boivent de l'alcool, 890 fument et boivent
* Aussi, sur les 27373 personnes, 15497 ont un temps de sommeil inférieur aux 8 heures recommandées



Etude de chaque colonne (sur les personnes qui souffrent de maladies cardiaques) :

* la majorité des sujets sont entre 23 et 33 BMI
* 58.5 % des sujets fument, 41.5% ne fument pas
* 4.1% de sujets boivent de l'alcool
* 56.6% de sleeptime < 8 heures
* 18% qui font de l’asthme
* 12% qui ont une maladie des reins
* 32% qui ont le diabète
* 18% qui ont un cancer de la peau
* 16% qui ont déjà eu une crise cardiaque
* 63.9% font une activité physique régulière

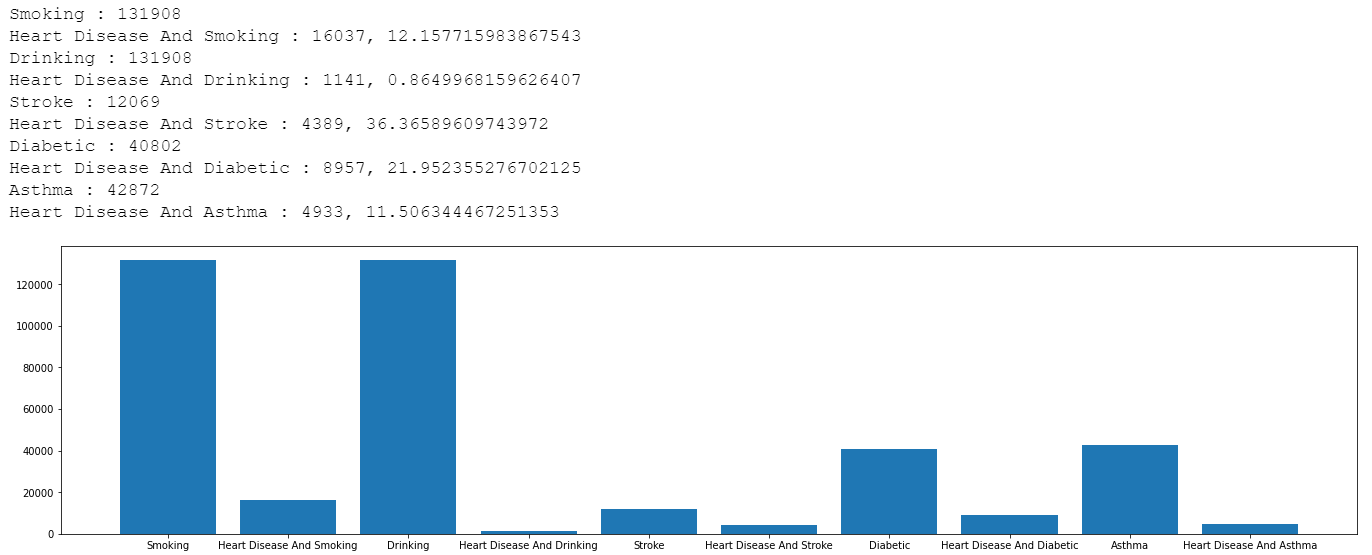


Combinaisons de données

Après l'étude initiale, on essaie de comparer les sujets qui sont malades et ceux qui ne le sont pas à travers plusieurs différentes caratéristiques.

On commence par récupérer le nombre total de personnes qui on “True” sur la caractéristique (par exemple Smoking) et on récupère le nombre de personnes qui souffrent de maladies cardiaques et de la caractéristique choisie aussi.

On compare les deux valeurs qu'on trouve :



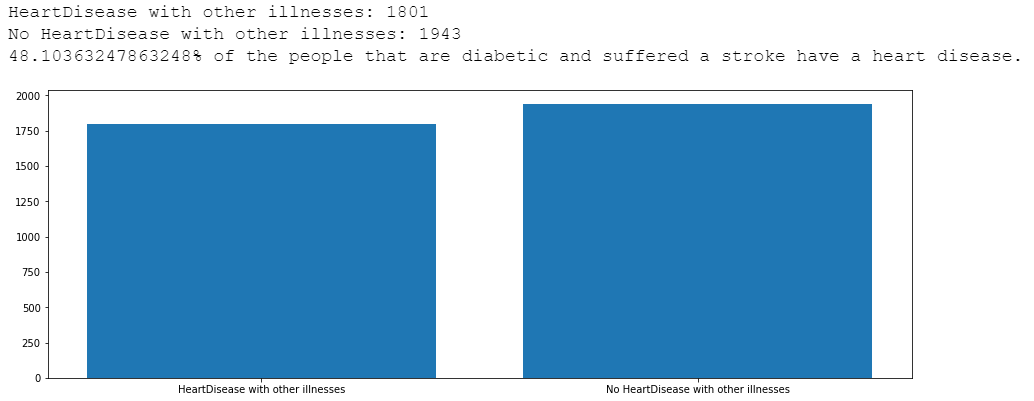
Théories qu’on peut émettre

En observant les données qu’on a,

On peut voir que 36% des personnes qui souffrent de maladies cardiaques ont subi une crise cardiaque.

On peut aussi voir que 21% des personnes qui souffrent de maladies cardiaques sont diabétiques.

On peut alors combiner les deux pour voir combien de personnes qui sont diabétiques et qui ont subi une crise cardiaque souffrent d’une maladie du coeur.

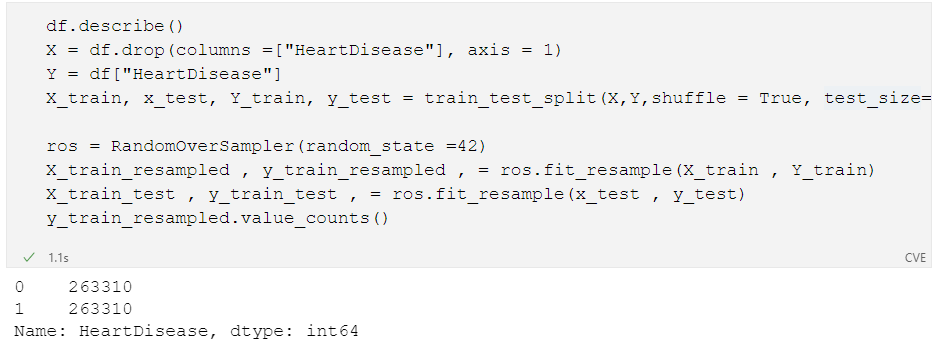


Résultat: 48% sont malades d’une maladie du cœur.

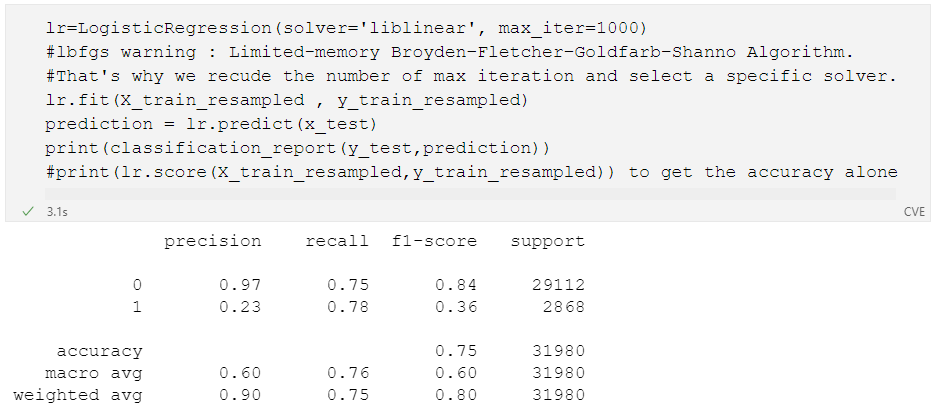
Etude à travers des modèles

Split des données

Pour appliquer nos différents modèles, on doit faire en sorte qu'on ai le meme nombre de lignes et de colonnes pour X(L1) et Y(L2)



Modèle 1: Logistic Regression



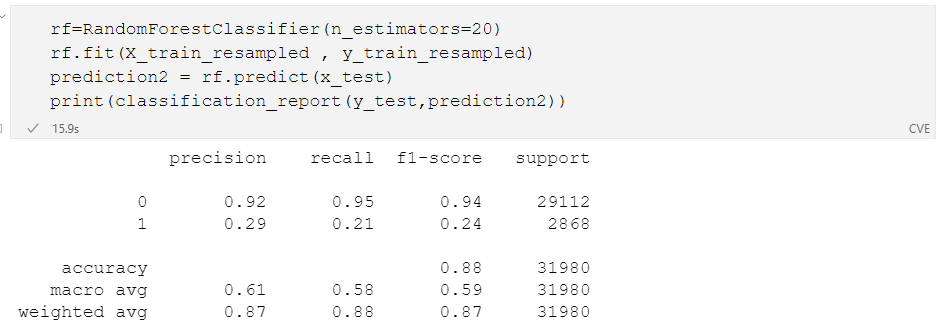
On remarque que notre taux de précision est de 76%.

Cette valeur est due au fait que ce modèle est plus précis avec des valeurs continues, alors que nous disposons principalement de valeurs discrètes.

Nos données comme "Smoking" et "Diabetic" sont plus importantes que d'autres valeurs pour déterminer si le sujet est malade d'une maladie de cœur. Hors ce modèle accorde la même importance à chaque feature.

Ce résultat est peut-être aussi dû à un surentraînement à cause de nos valeurs répétitives d'où l'importance de préparer nos données afin de l'éviter.

Modèle 2: Random Forest



On remarque que notre taux de précision est de 58%

Le taux de précision varie suivant le paramètre "n\_estimators".

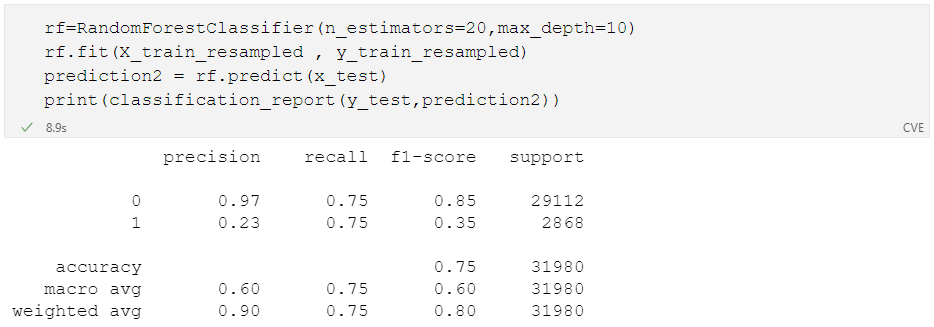
Ce paramètre correspond au nombre d'arbres qui va "prendre une décision" (choix binaire).

On constate qu'avec 20 arbres, le taux de précision est assez faible.

On en déduit que les différentes données n'ont pas la même importance sur le fait qu'une personne soit malade ou non.

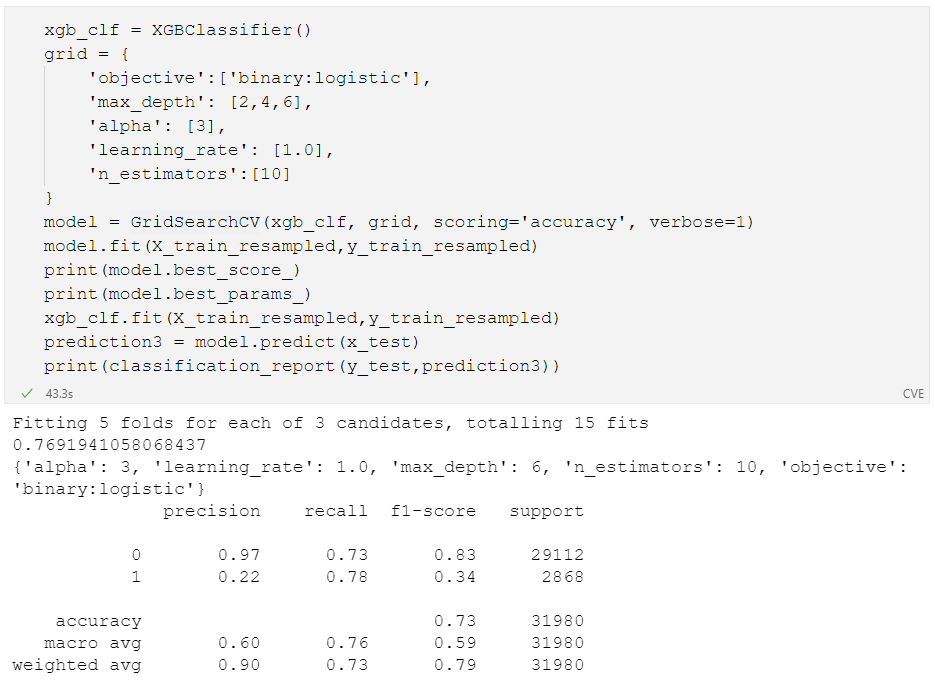
On suppose que la combinaison de certaines valeurs entre elles entraînera un meilleur taux de réussite.

Par exemple, une personne, dont les caractéristiques "smoking" et "Diabetic" sont plus élevées, est plus susceptible d'avoir une maladie du cœur.



Dans ce deuxième essai, on remarque que la combinaison des facteurs impacte grandement le taux de précision et la rapidité d'exécution.

Modèle 3: XGBoost



On remarque que notre taux de précision est de 76%

Ce modèle est très similaire au modèle Random Forest, sauf qu'il ne se charge pas d'optimiser le modèle lui-même.

On utilise donc GridSearchCV qui cette charge de trouver les paramètres optimaux pour le XGBoost.

Conclusion

On conclut que les modèles basés sur des arbres de décision sont plus efficaces et évitent le surapprentissage.

Grâce à ce data set, on a réussi à trouver les facteurs essentiels de la détection d’une maladie du coeur (le diabète et le fait de fumer).