Python for data analysis

Project

Dataset:

Le dataset n’a aucune valeurs manquantes. Il y 17 attributs, et 2111 entrées.

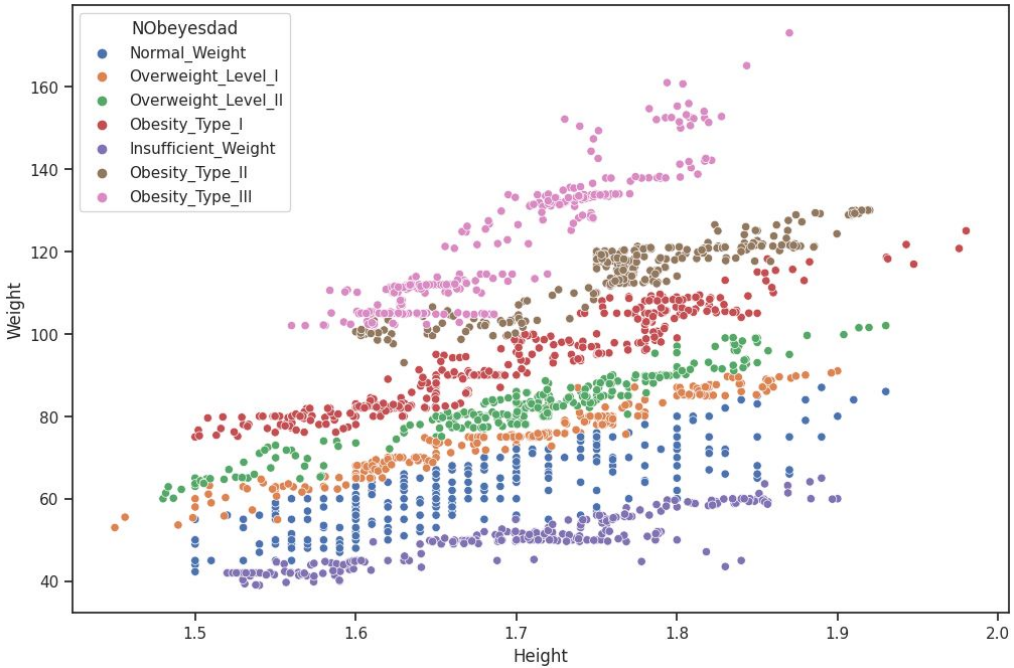
Le dataset parle de l’obésité et de la relation entre l'âge (Age), le poids (Weight), la taille(Height), ainsi que d’autres questions posées au patient, pour en déduire son type d’obésité, qui peut varier:
Poids Insuffisant, Poids Normale,
Surpoids de niveau I, Surpoids de niveau II, Surpoids de niveau III,
Obésité de Type I, Obésité de Type II, Obésité de Type III

Les questions posées au patient

Le genre : (Homme/Femme)
Antécédent d'obésité dans la famille : (Oui/Non)
"FAVC" : Consommation fréquente de nourriture hautement calorique (Oui/Non)
"FCVC" : Fréquence de consommation de légumes (Jamais/Parfois/Toujours)
"NCP" : Nombre d'aliments de base consommé chaque jours (1/2/3/>4)
"CAEC" : fréquence d'alimentation hor repas (Jamais/Parfois/Souvent/Toujours)
"Smoke" : Fume (Oui/Non)
"CH2O" : La quantité d'eau ingéré chaque jours (<1L, [1L: 2L], >2L)
"SCC" : Surveillance des calories quotidiennes: (Oui/Non)
"FAF" : La fréquence d'activité physique par jour: (0/[1:2],[2:4],[4:5])
"TUE" : temps devant les écrans ([0:2heures], [3:5heures]

A noter:

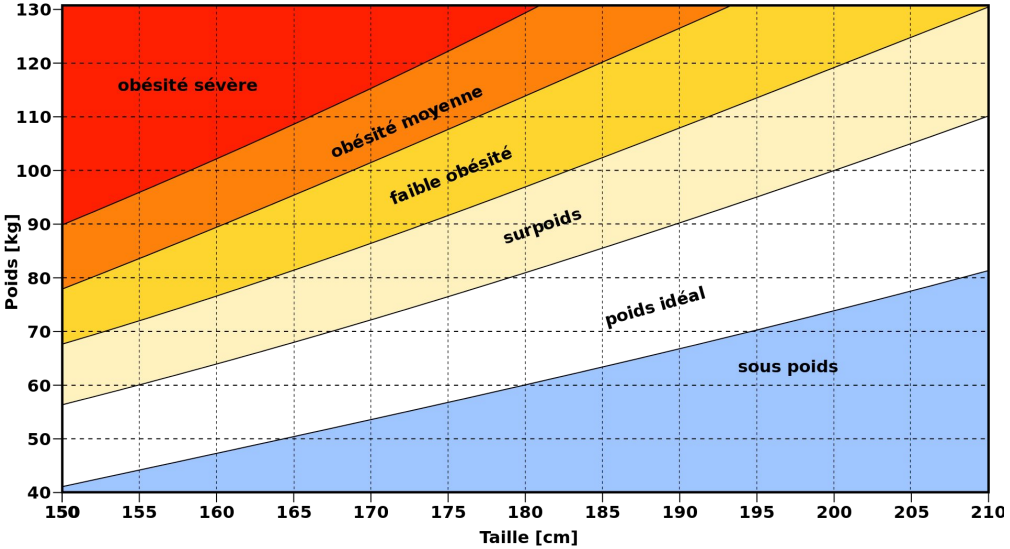
Certaines données de ce dataset ont été générées automatiquement par ordinateur pour augmenter artificiellement la taille du dataset de base, qui devait être trop petit.



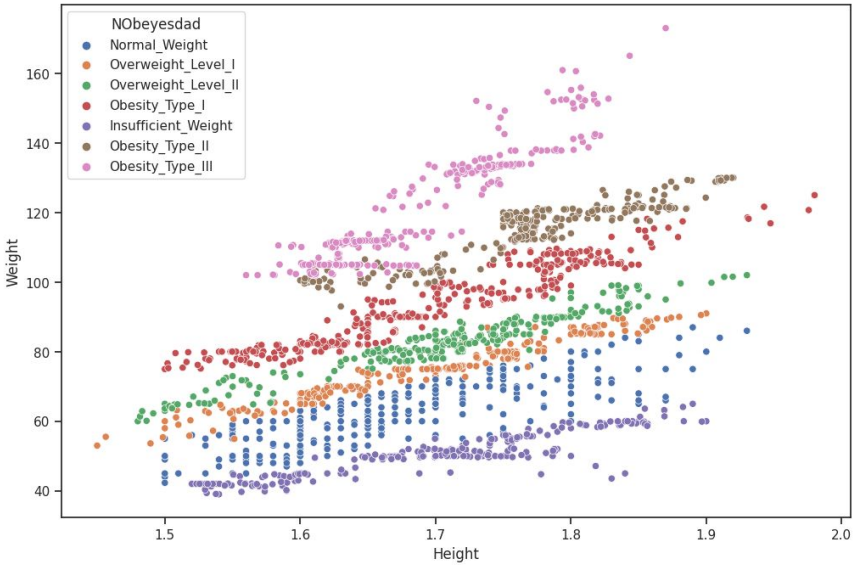
On peut voir que les données bleues (Poids normales), semblent ne pouvoir prendre que certaine plage de valeurs. On peut le voir dans les rayures verticales que forme le nuage de points bleu.

Ajout de nouvelles variables

Nous avons voulu voir quels résultats nous obtiendrons en utilisant l'IMC, qui ne s'appuie que sur la taille et le poids: IMC = Poids/Taille²



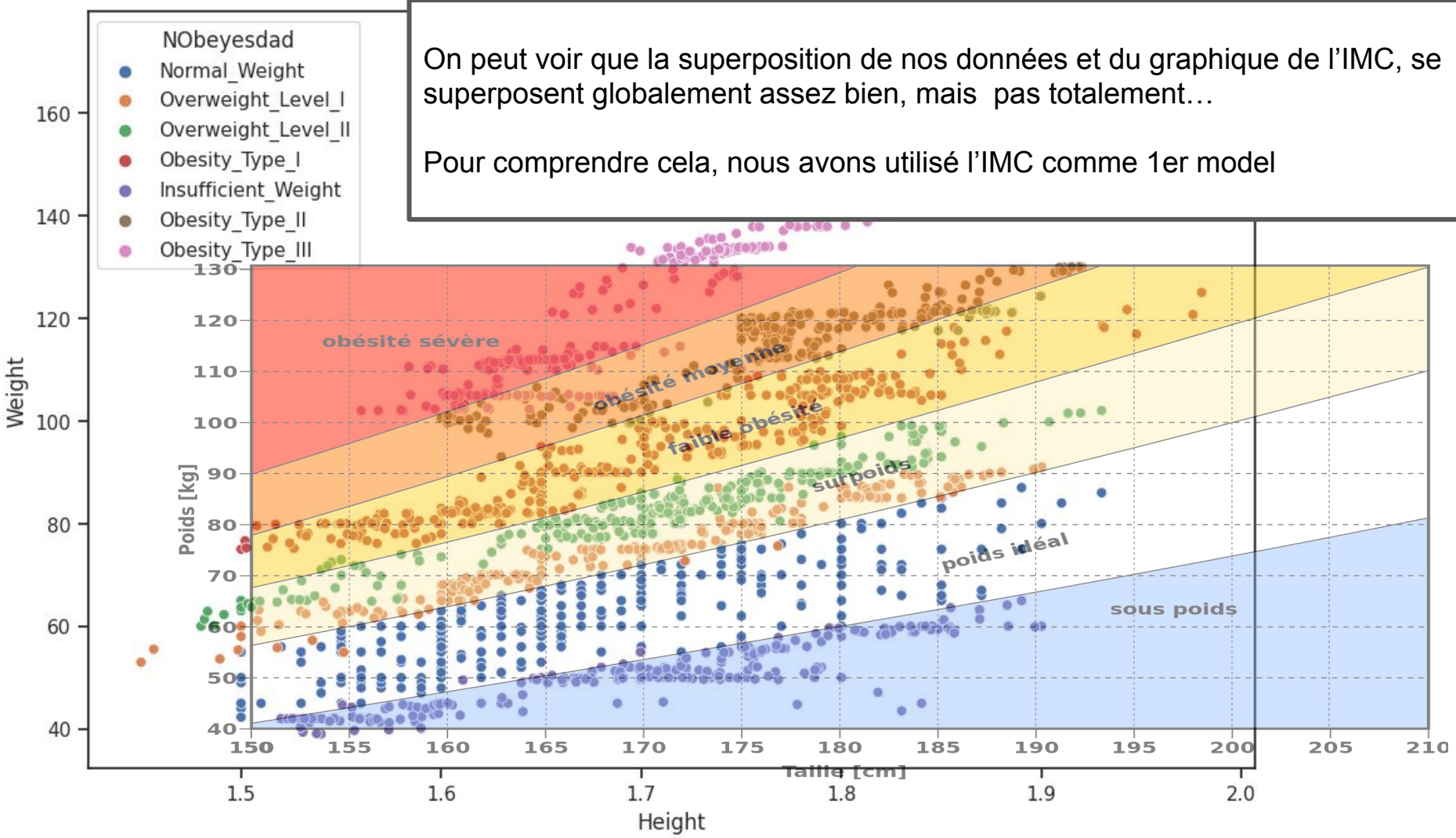
IMC - Wikipédia



Nos données

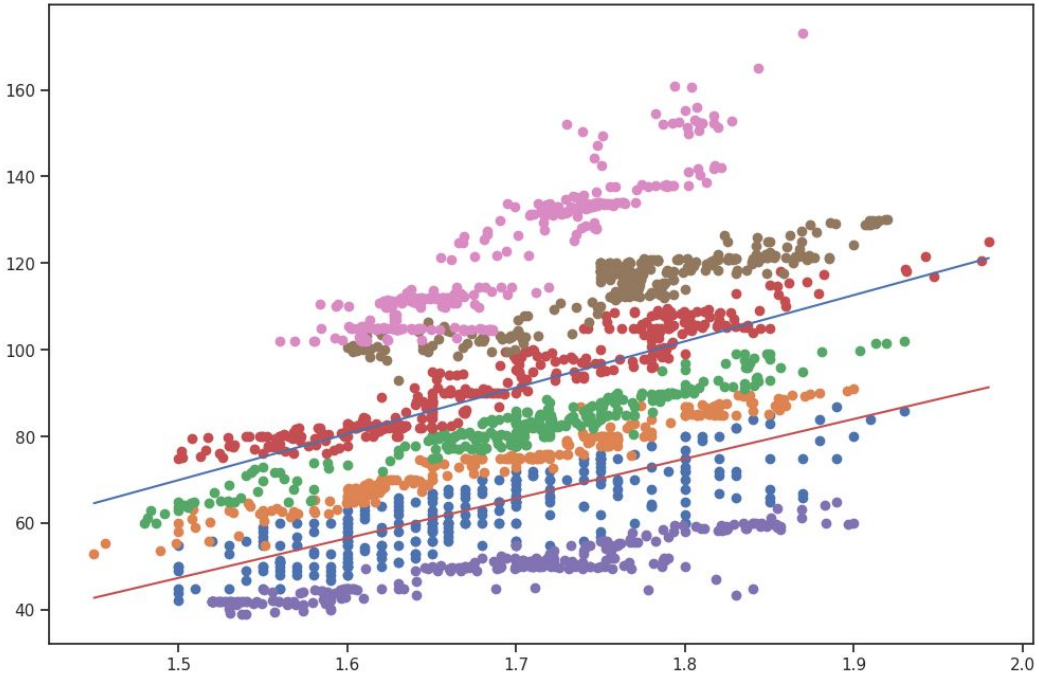
On peut voir que la superposition de nos données et du graphique de l'IMC, se superposent globalement assez bien, mais pas totalement…

Pour comprendre cela, nous avons utilisé l'IMC comme 1er model



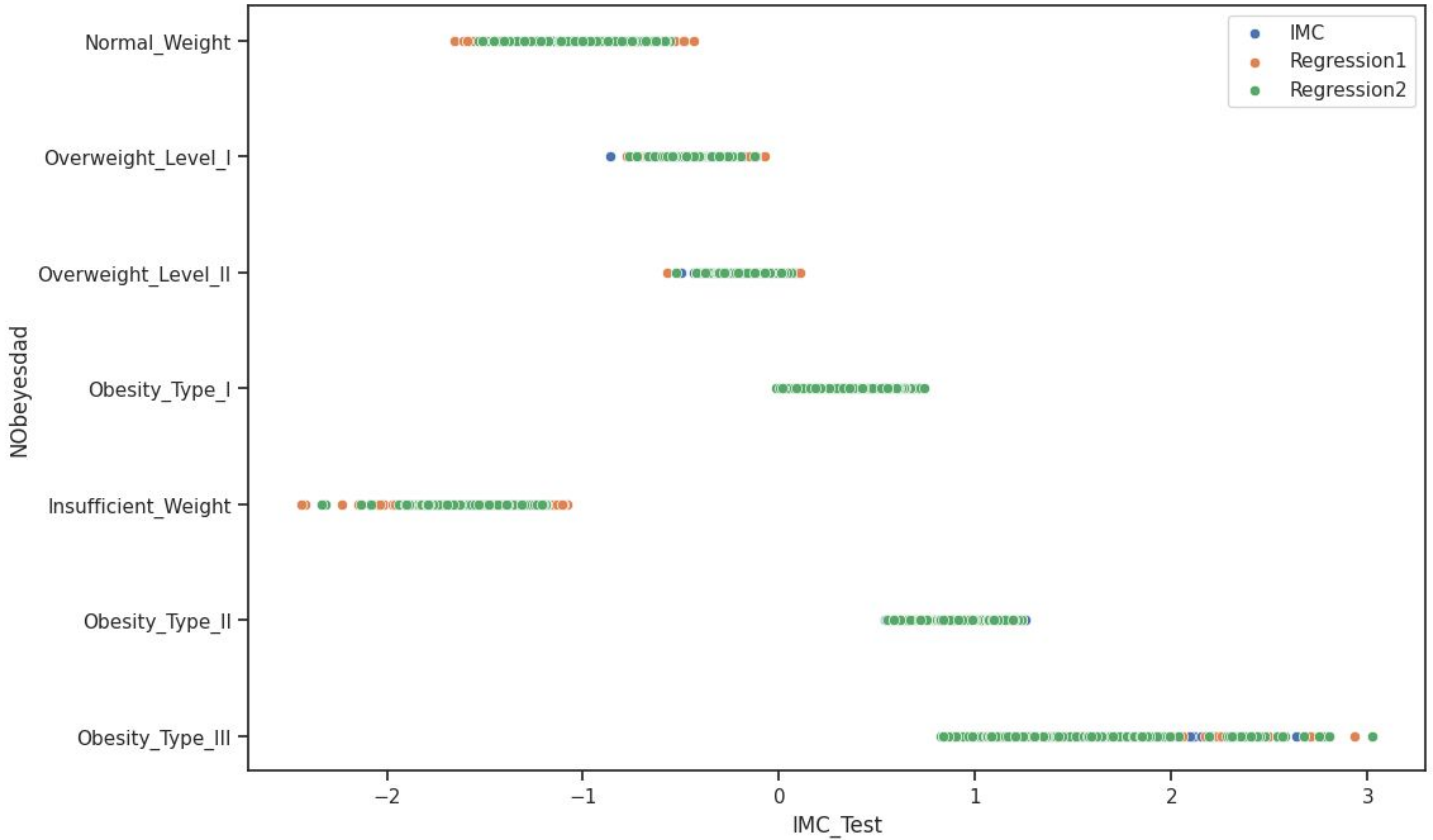
Régression linéaire

On peut constater que les différentes données semblent suivre une droite, donc on pourrait relier ces 2 variable de façon linéaire:

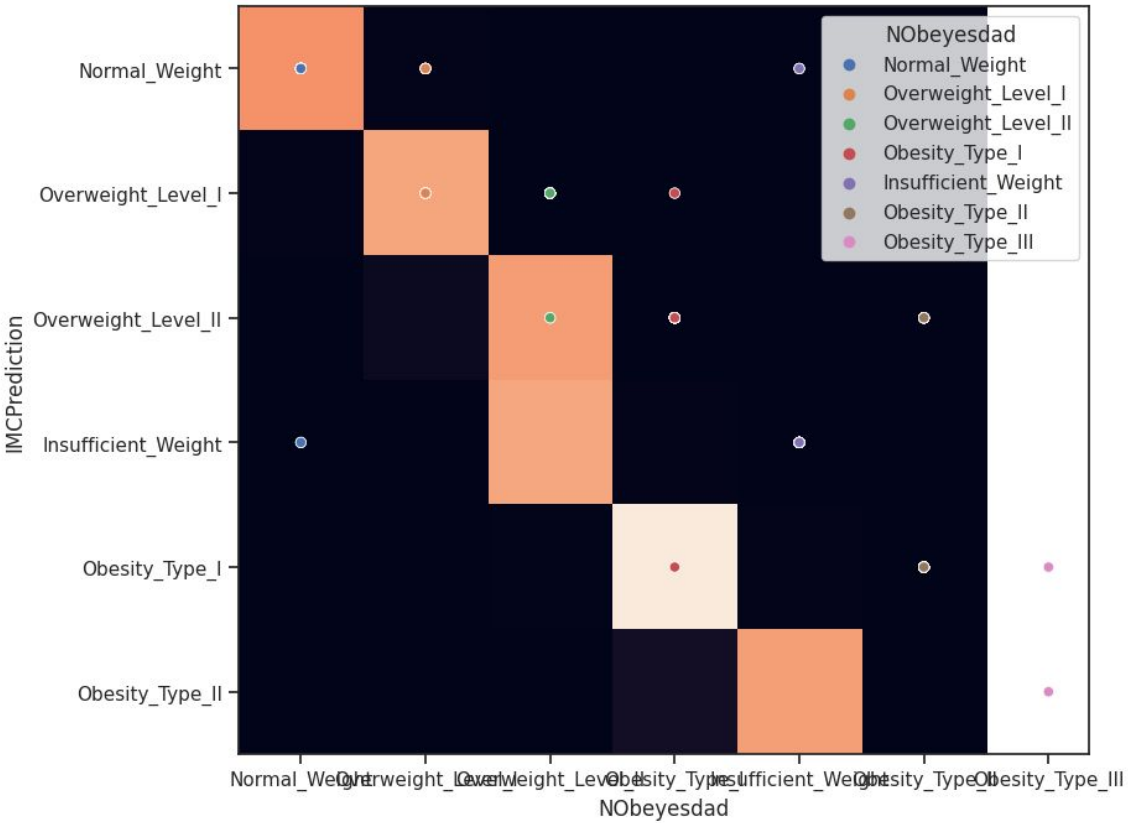


En faisant une régression linéaire de chacune des catégorie, puis en faisant la moyenne. On utilisera 2 façons de faire la moyenne: la moyenne des coefficients directeurs, et la moyenne de l'angle des pentes.

On peut voir que de ces 3 nouvelles variables, c'est l'IMC la plus représentative, comme nous le montre ce graphe:

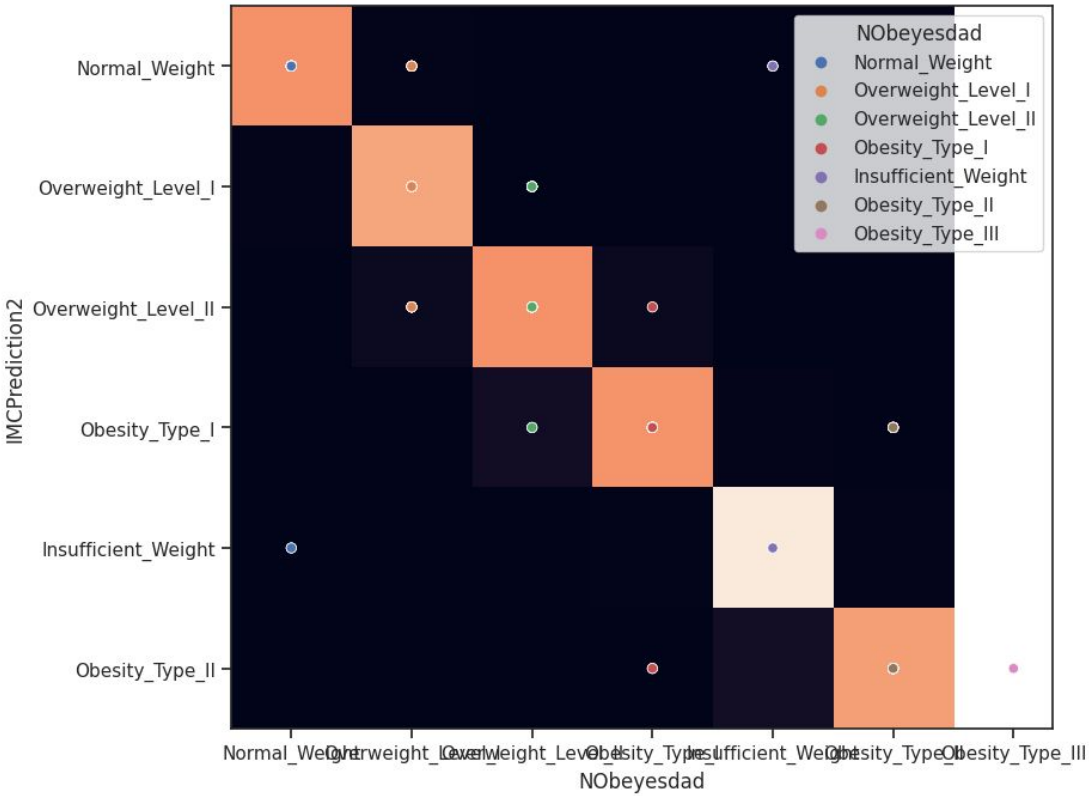


Prédiction avec l'IMC



Il y a 40% de bonnes réponses, ce qui prouve que l'IMC ne fait pas tout.

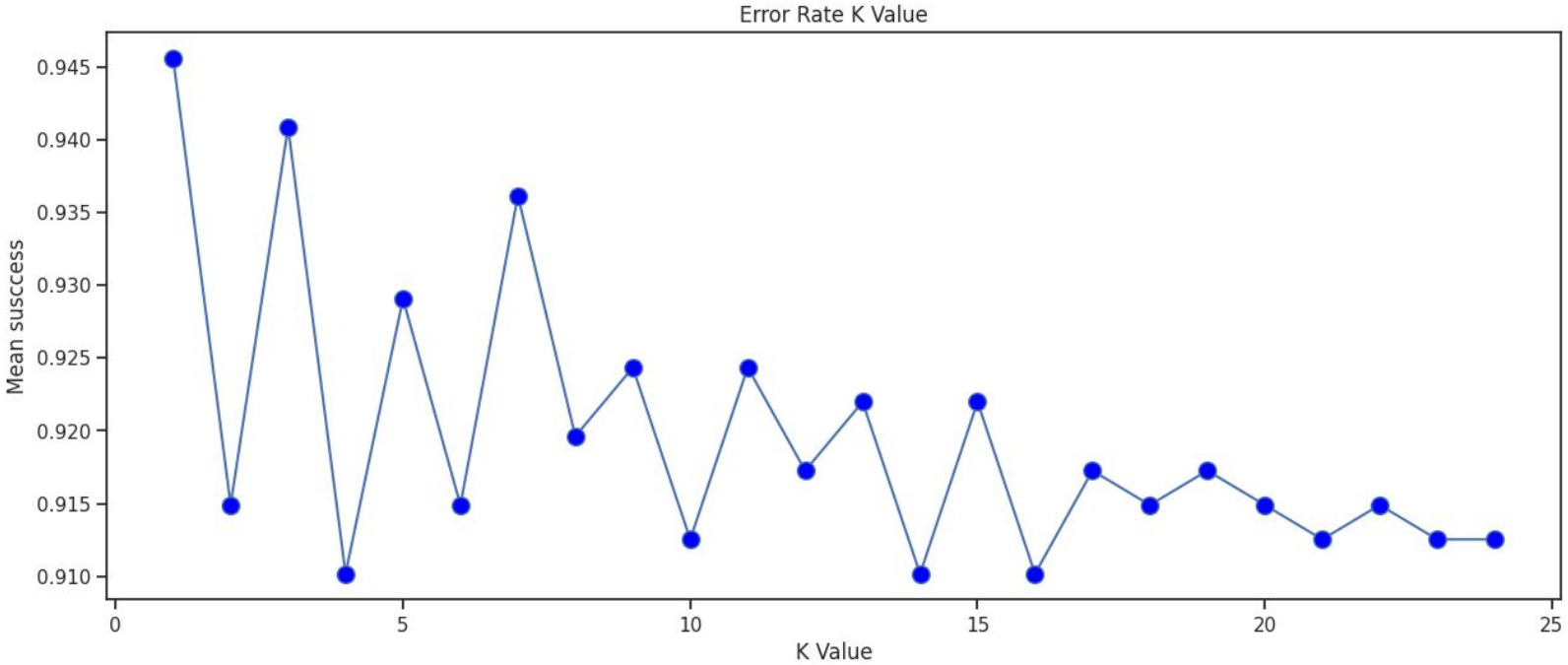
Prédiction avec l'IMC modifiée



En modifiant empiriquement les tranches de classification de l'IMC, on peut arriver à 81% de bonnes réponses, ce qui montre tout de même que le calcul de l'IMC peut être une nouvelle variable très pertinente.

KNN

Nous avons utilisé l'algorithme KNN, et l'on trouve 95% de bonne réponse, avec un k=1 :



RandomForest

Nous avons choisi d’utiliser l’algorithme RandomForest étant donné que celui-ci est généralement très pertinent en terme d’accuracy.

Nous avons utilisé sklearn pour mettre en place ce modèle, avec 50 arbres, et l’accuracy varie à chacun de nos test entre 0.96 et 0.99.

Nous sommes donc très satisfait du modèle utilisant RandomForest bien que le knn et la prédiction avec l’IMC ne soient pas mauvaises non plus.