Python for Data Analysis
Obesity Dataset

AUBRY Corentin

BAROUX Alexandre

CORE IBO1

Principales étapes

- Découverte du dataset
- Analyse des données
- Pré processing
- Machine learning
- API

1 - Découverte du dataset

Le dataset et son contexte

Le Dataset utilisé s'intitule "Estimation of obesity levels based on eating habits and physical condition Data Set".

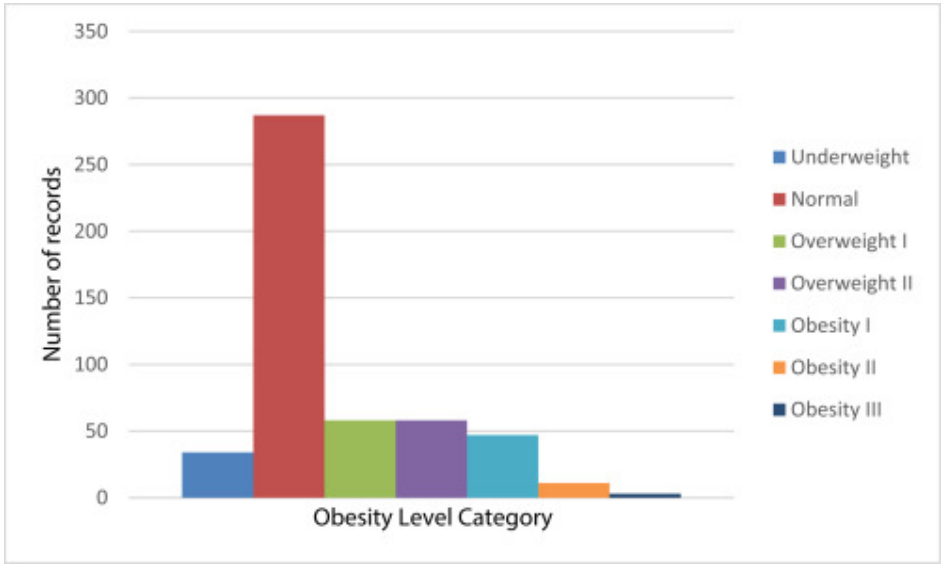
Les auteurs de ce dataset sont :

- Fabio Mendoza Palechor
- Alexis de la Hoz Manotas

Ils ont créé le dataset dans le but de pouvoir utiliser le dataset pour générer des outils de calcul intelligents pour identifier le niveau d'obésité d'un individu et pour construire des systèmes de recommandations qui surveillent les niveaux d'obésité.

La construction du dataset

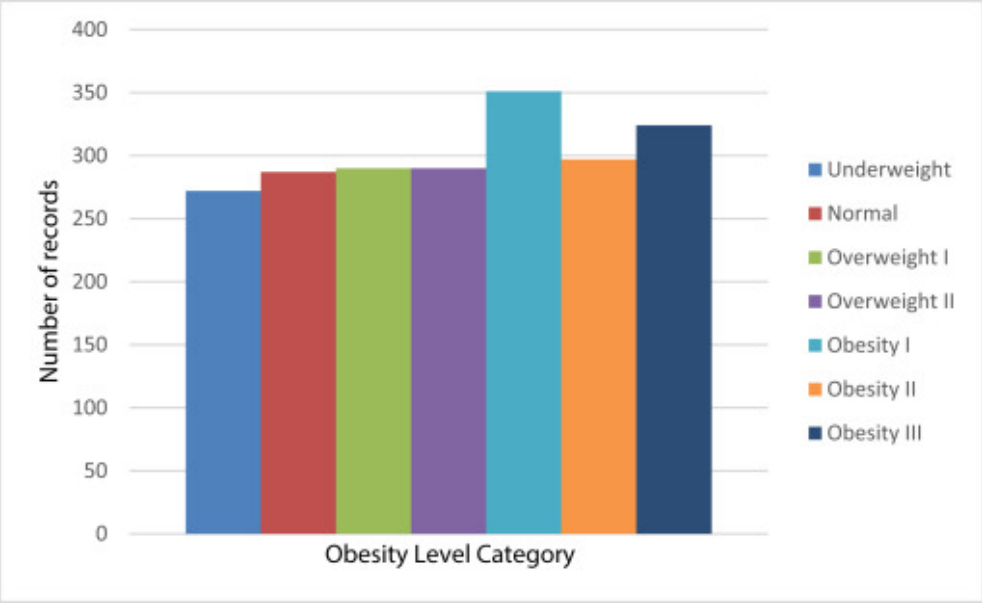
Ce dataset résulte d'une enquête en ligne sur la corpulence de la population du Mexique, du Pérou et de la Colombie. En premier lieu les données utilisés proviennent d'une enquête en ligne. Le but étant de récupérer la corpulence d'un individu par calcul de son IMC et un ensemble de données. L'enquête a duré 30 jours et a récupéré 485 profils de personnes entre 14 et 61 ans.



La répartition des individus suivant leurs corpulences a donnée la répartition visible sur le graphique.
Cette répartition n'est pas équilibré, certaines corpulences ne sont presque pas représenté.

La construction du dataset

Les premières données donnant lieu à une répartition déséquilibré selon la colonne cible cela aura des conséquences sur la qualité du modèle de machine learning qui serait entraîné dessus. Pour éviter cela les auteurs ont rééquilibré la répartition grâce à des données synthétiques généré en utilisant l'outil Weka et le filtre SMOTE proposé par d'autres chercheurs dans cet article : “SMOTE: synthetic minority over-sampling technique”



Les données synthétiques ont été rajoutées en se basant sur celles de l’enquête. Ces données synthétiques représentent 77% du dataset finale. La répartition finale est visible sur ce graphique

Objectif : prédire la corpulence d'un individu en se basant sur les caractéristiques suivantes :

- son sexe
- son âge
- sa taille
- sa masse
- ses antécédents familiaux avec le surpoids
- sa consommation de nourriture calorique
- sa consommation de légumes
- son nombre de repas quotidiens
- ses habitudes de grignotage
- son tabagisme
- sa consommation d'eau
- est ce que l'individu surveille son nombre de calories
- sa pratique d'activités physiques
- son temps d'utilisation d'objets électroniques
- sa consommation d'alcool
- son moyen de transport usuel

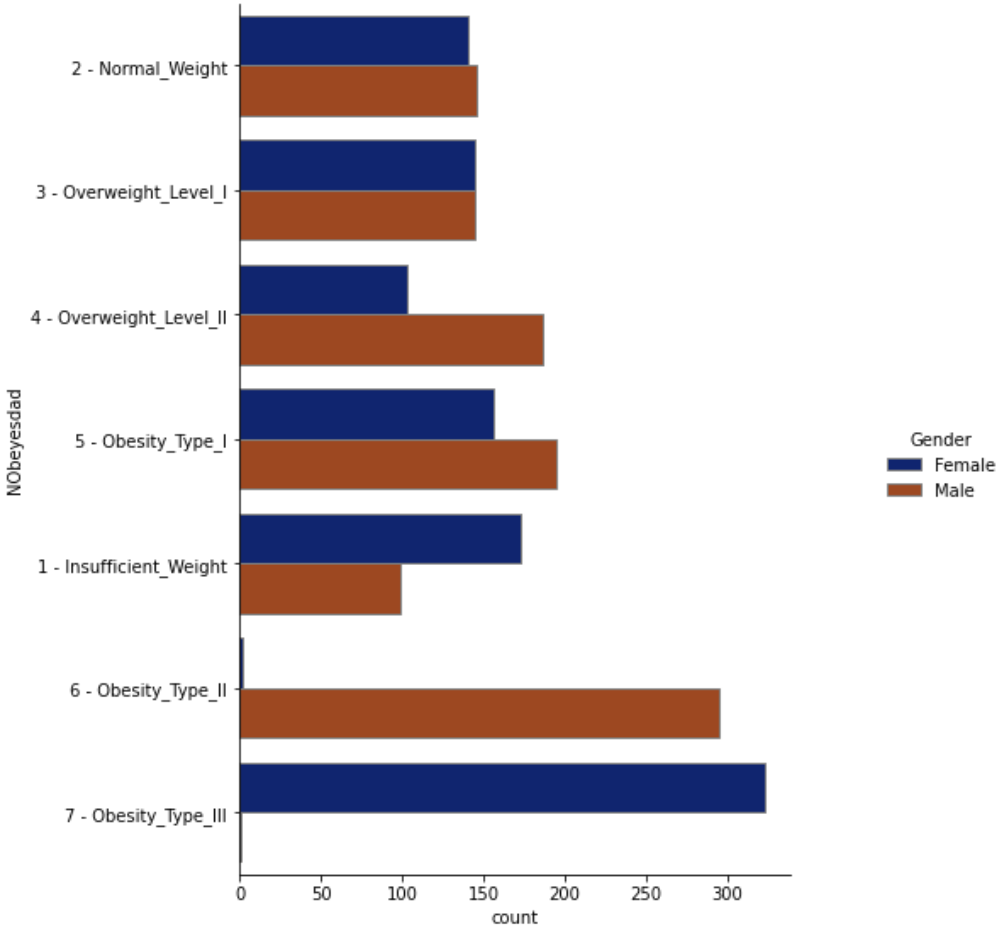
2 - Analyse des données

Comment avons nous analysé les données

Nous avons décidé d'analyser chacune des variables présentes dans le dataset. Le but de cette analyse a été de vérifier la structure des différentes variables, de connaître la répartition des individus suivant chaque variable et enfin d'essayer de classifier les colonnes en fonction de leurs impacts sur la corpulence d'un individu.

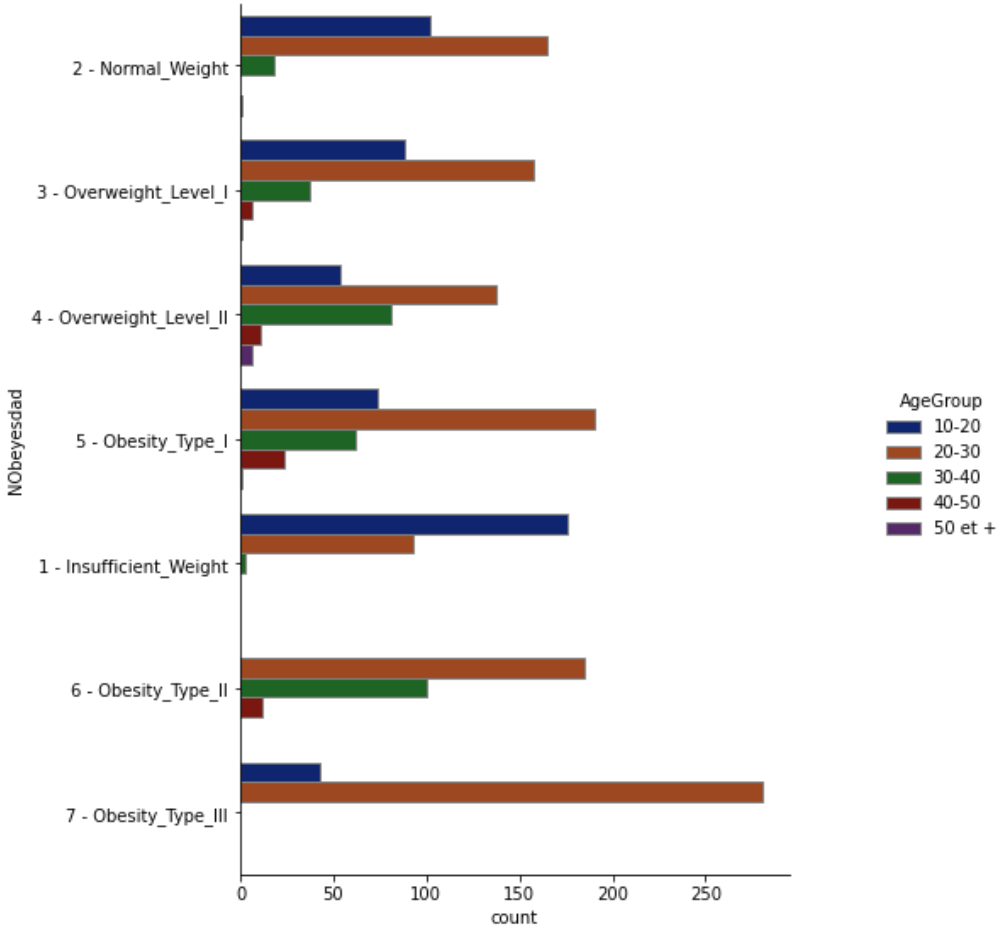
Cette étude avait également pour but de mieux comprendre les tenants et aboutissants des données synthétiques. On s'est particulièrement demandé si ces données pouvaient être génératrices de biais statistiques.

Analyse des données : le sexe



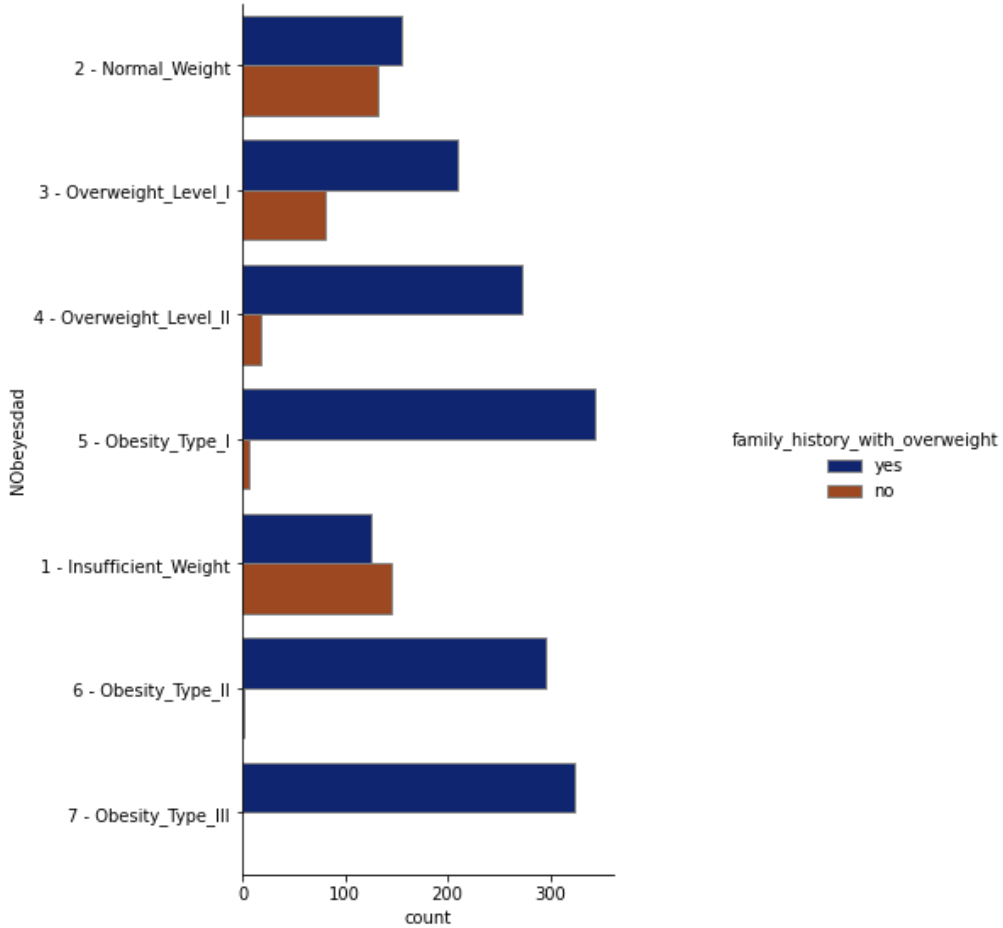
La différence de poids entre les hommes et les femmes semble peu significative, à l'exception de l'obésité de type 2 et 3, qui semblent être exclusives à un seul sexe.

Analyse des données : l'âge



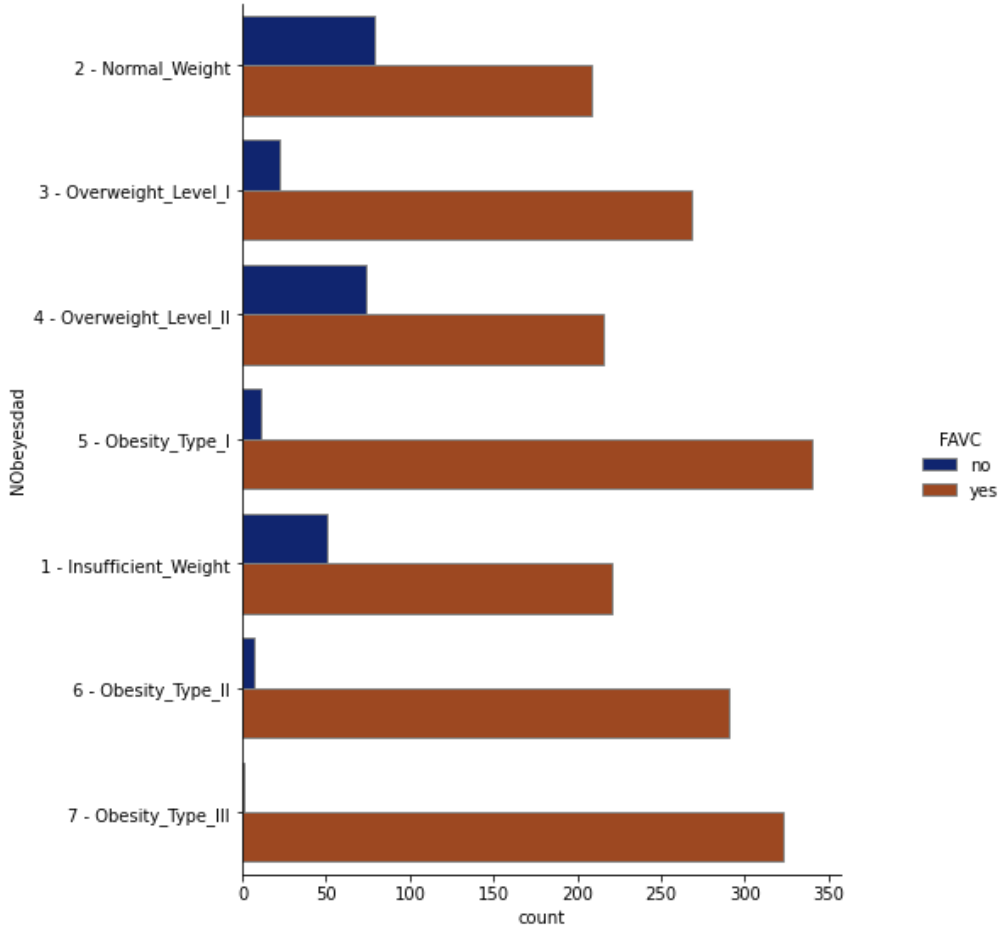
Les données de cette variable sont très déséquilibrées, en effet les 20-30 ans sont majoritaires dans chaque catégorie, à l'exception de la maigreur, qui semble davantage concerner les 10-20ans.
il semble donc difficile d'établir un lien entre l'âge et la corpulence à partir de ces données

Analyse des données : antécédents familiaux



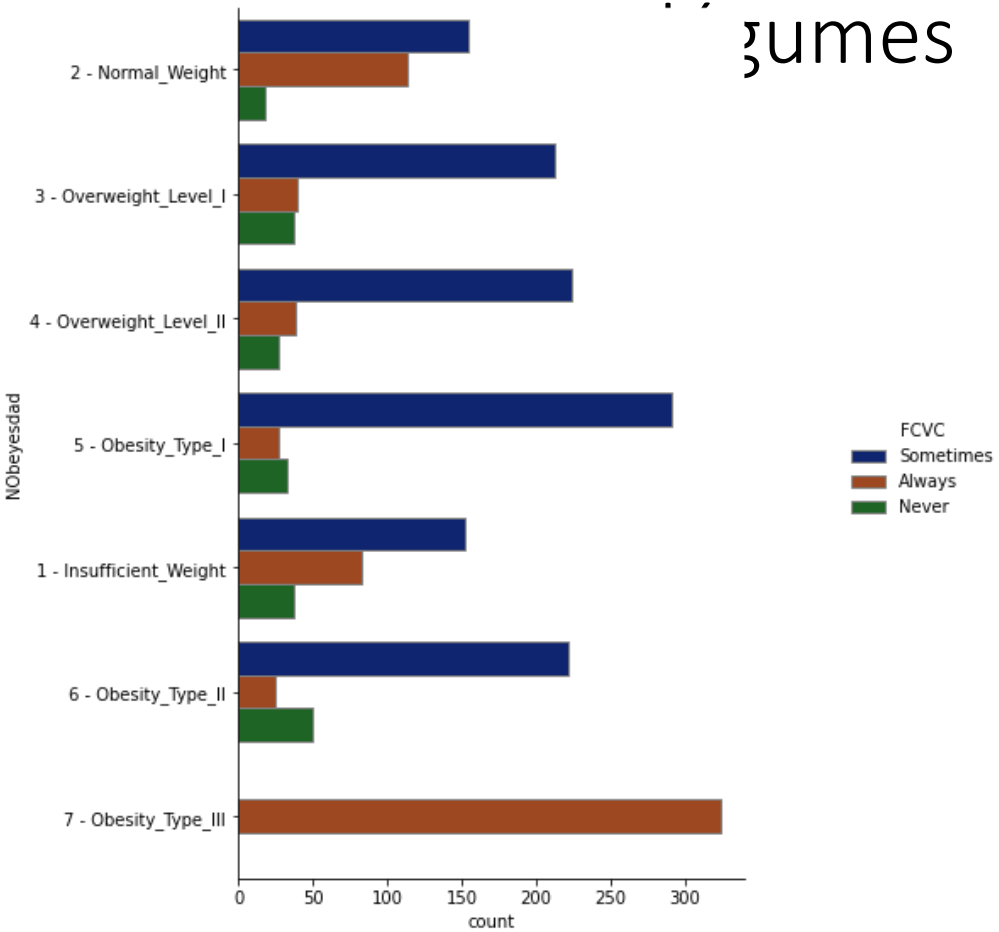
L'obésité semble ne toucher que les personnes ayant des antécédents familiaux avec le surpoids. De même, le surpoids touche essentiellement ce même type de personnes.
Cette variable semble donc avoir une influence très forte sur la corpulence d'un individu

Analyse des données : consommation de calories



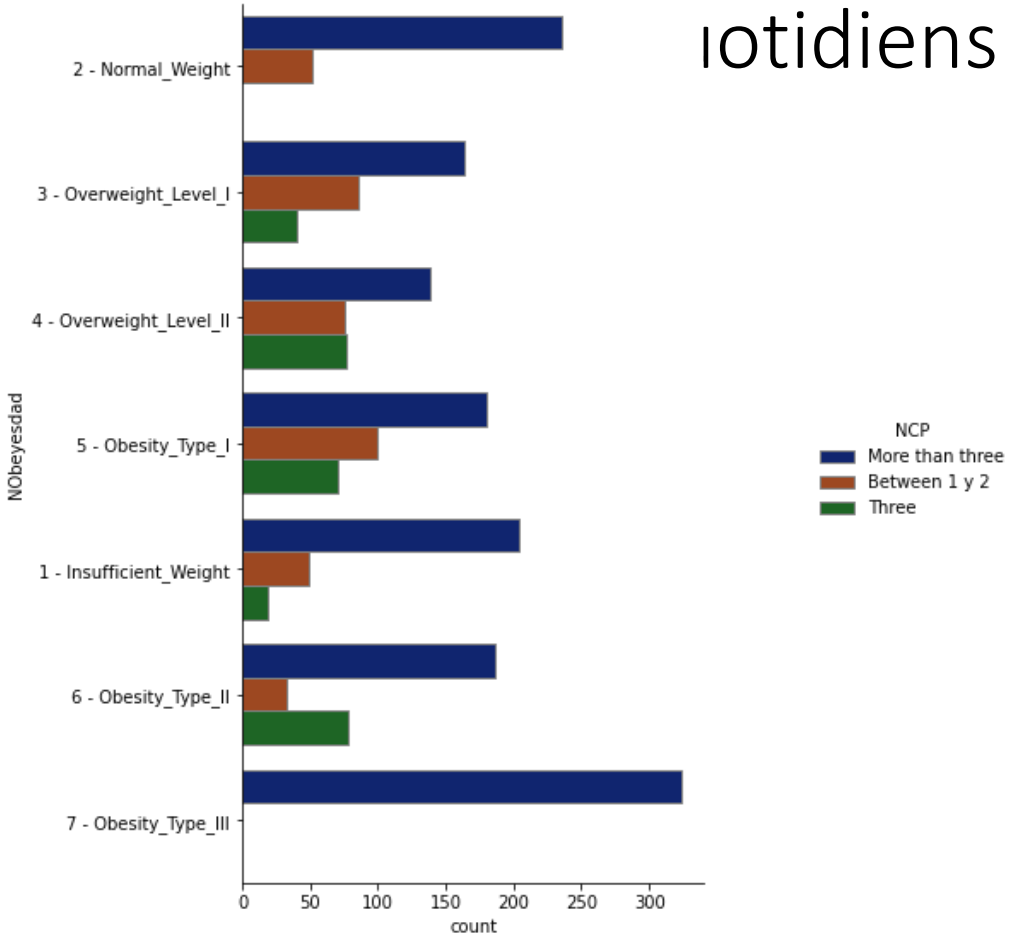
Malgré un fort déséquilibre des données, il semblerait que l'obésité concerne essentiellement des personnes consommant de la nourriture calorique.
il semble donc y avoir un lien entre la consommation de nourriture calorique et la corpulence

Analyse des données : consommation de légumes



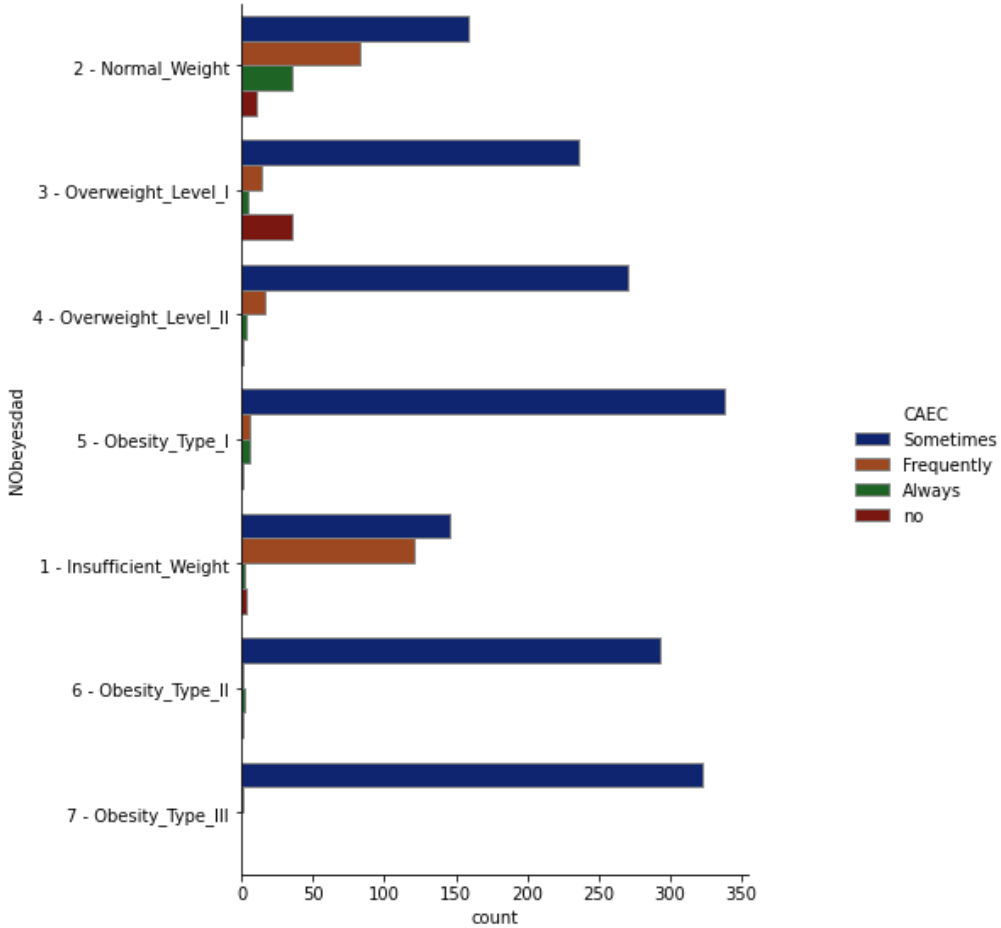
Il ne semble pas y avoir de liens entre la corpulence et la consommation de légumes, étant donné que les résultats sont les mêmes dans chaque catégorie, à l'exception de l'obésité de type 3, où étonnamment, tous les individus consomment des légumes à chaque repas. Ce résultat contre-intuitif est probablement dû au faible nombre de données

Analyse des données : nombre de repas quotidiens



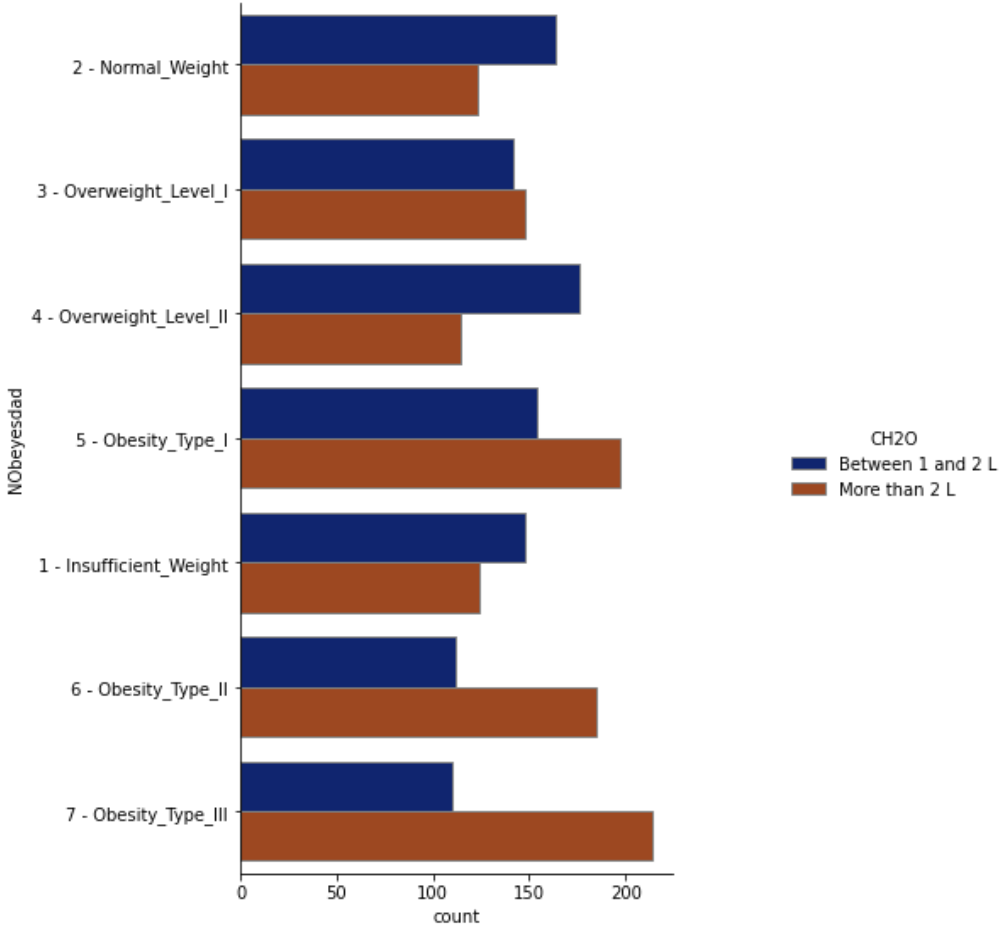
Cette variable est également très déséquilibrée, les personnes prenant plus de 3 repas par jour, sont surreprésentées, et majoritaires dans chaque catégorie.

Analyse des données : le grignotage



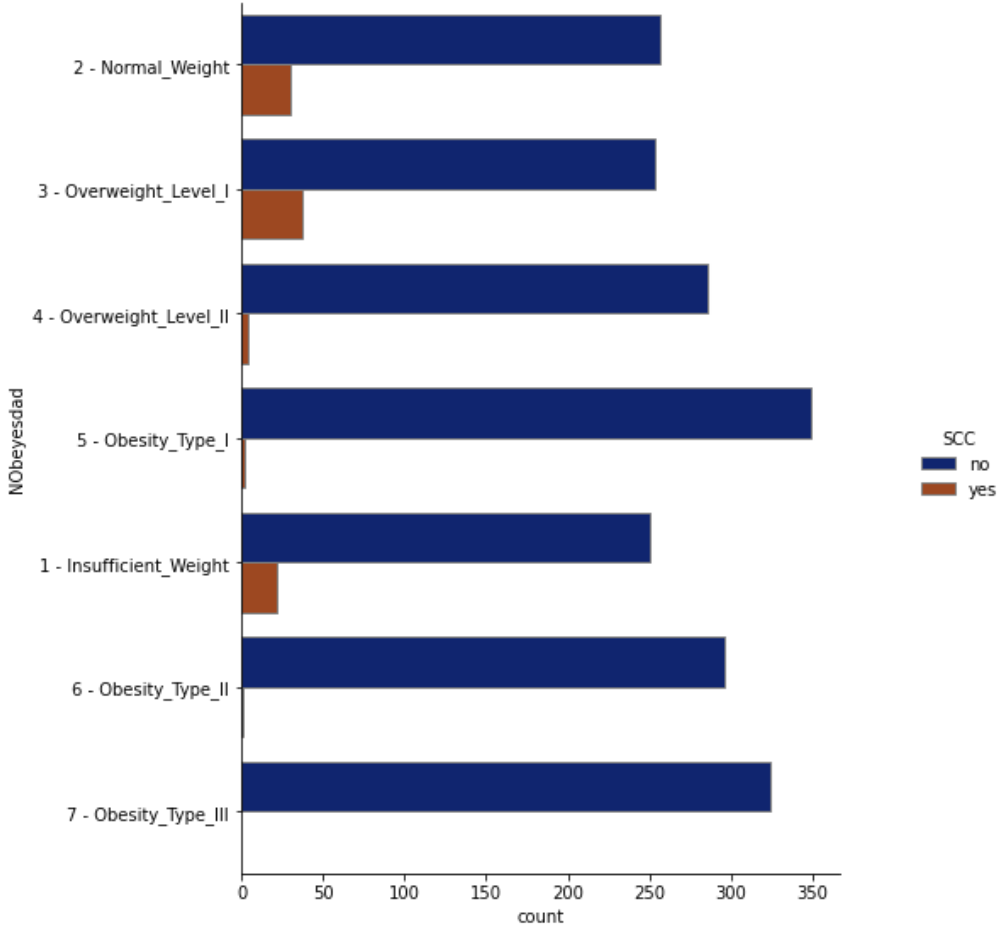
Malgré un fort déséquilibre des données, on peut tout de même constater que les personnes qui ne grignotent pas ne sont pas ou très peu concernées par l'obésité et le surpoids de type 2. Mais il semble cependant difficile de trouver d'autres liens entre le grignotage et la corpulence

Analyse des données : consommation d’eau



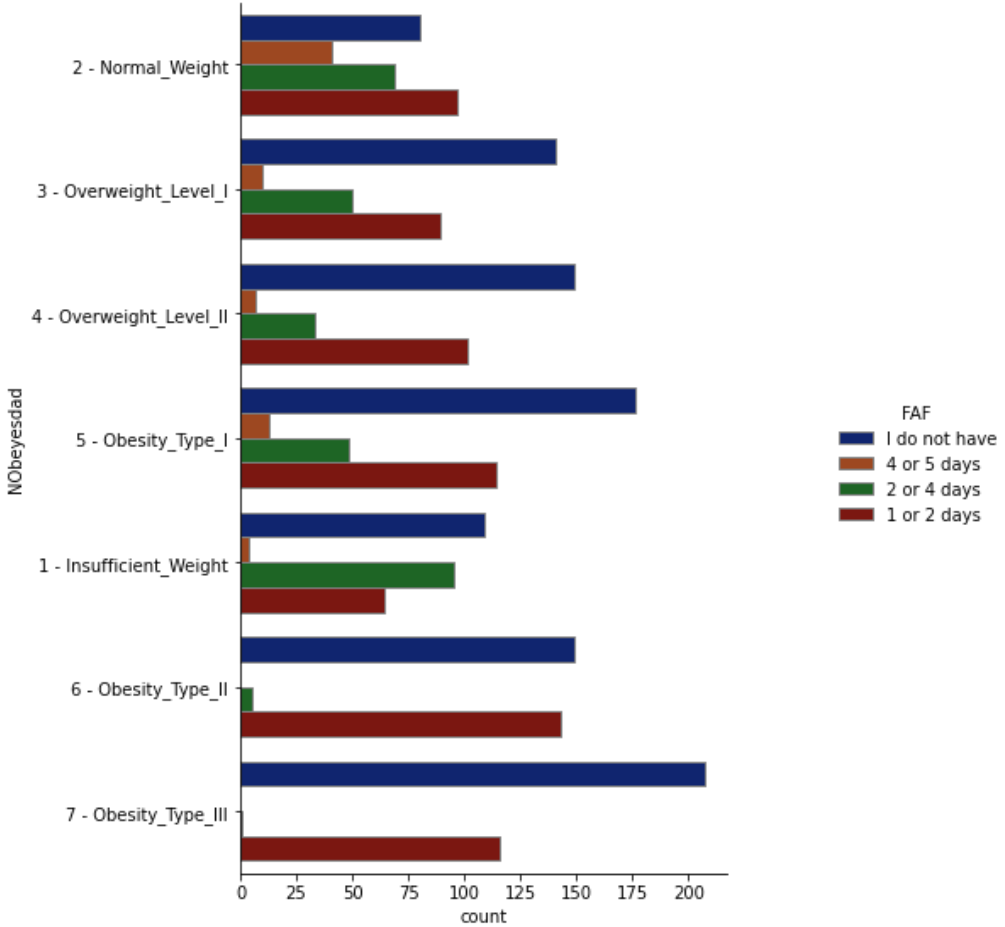
La consommation d’eau semble assez homogène dans chaque catégorie, à l’exception de l’obésité de type 2 et 3, où la consommation d’eau est beaucoup plus forte

Analyse des données : surveillance des calories



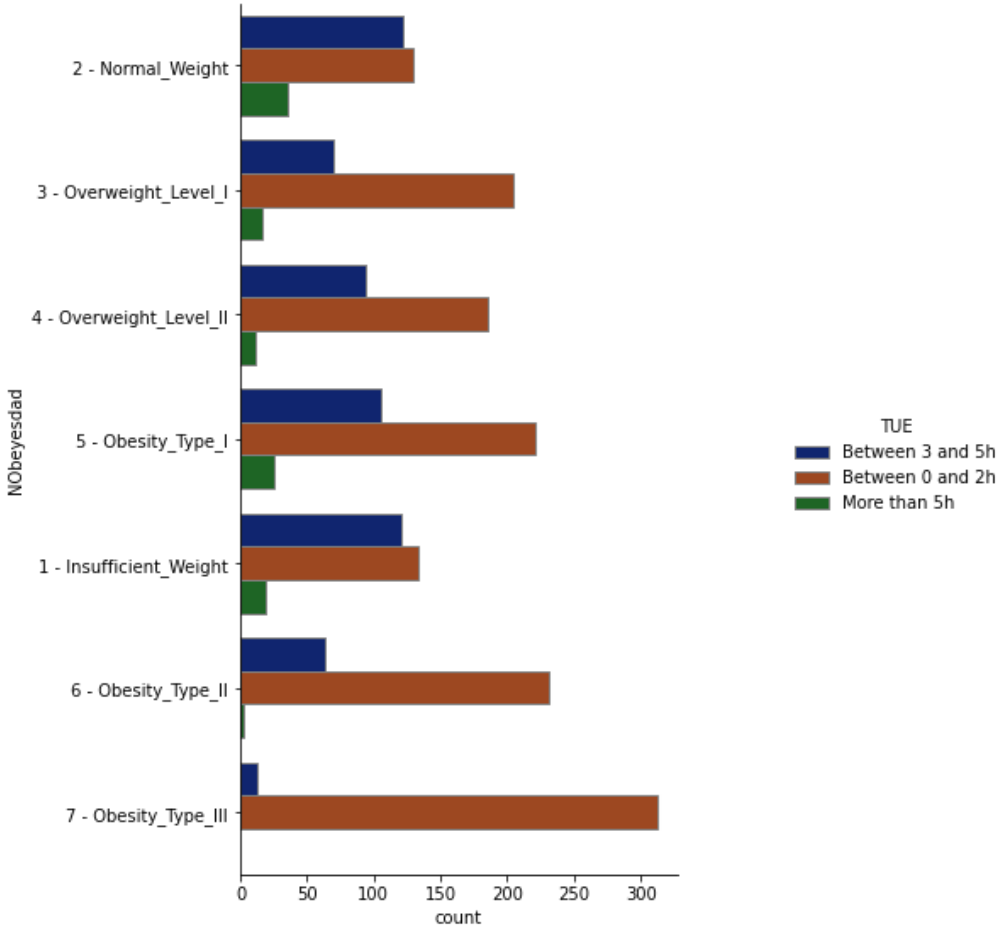
Bien que ces données soient très déséquilibrées, on constate que l'obésité et le surpoids de type 2 ne touchent pas ou très peu les personnes qui surveillent leurs calories. C'est donc un facteur important pour ce type de personne.

Analyse des données : activités physiques



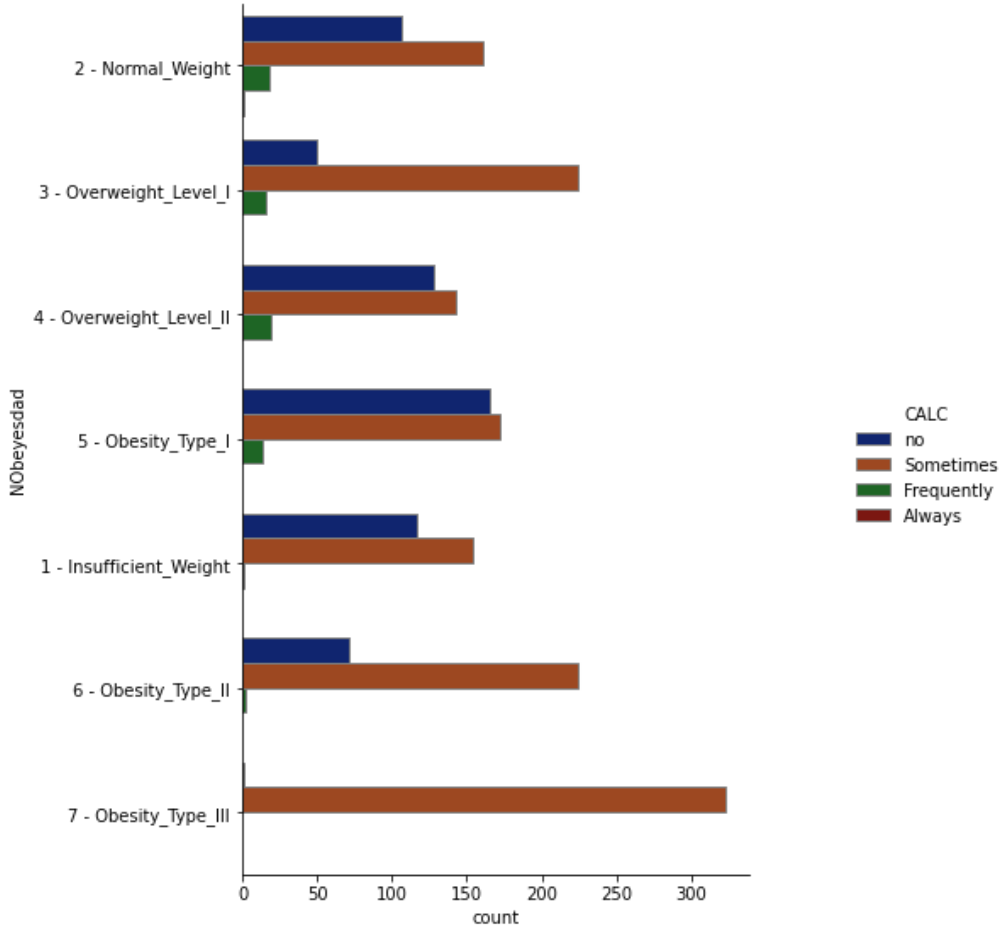
bien qu'une grande partie des individus ne pratiquent pas d'activités physiques, on remarque tout de même que celles qui en pratiquent beaucoup ont plus tendance à avoir une corpulence normale, et celles qui n'en pratiquent pas ont plus tendance à être obèses

Analyse des données : temps d'écrans



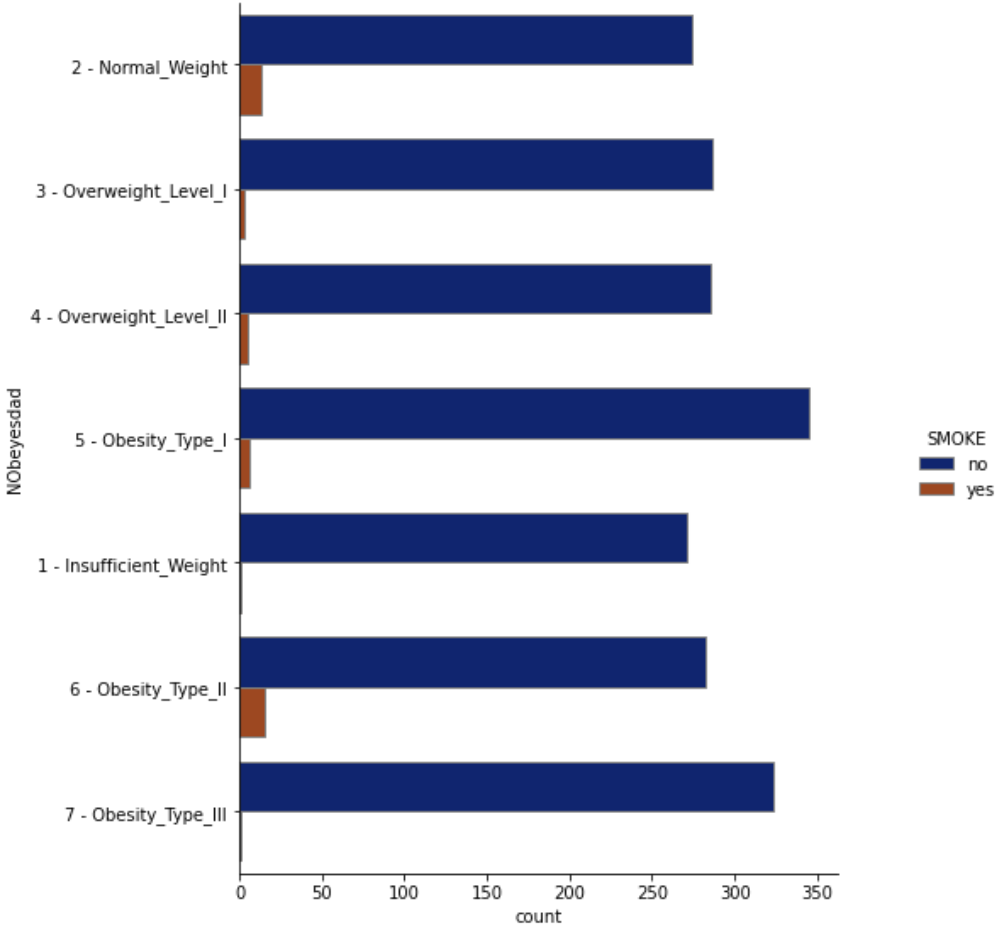
D'après les observations, la plupart des individus (et notamment les obèses) passent très peu de temps devant les écrans, ce qui semble contre-intuitif aujourd'hui. Cependant, la répartition n'est pas homogène.

Analyse des données : consommation d’alcool



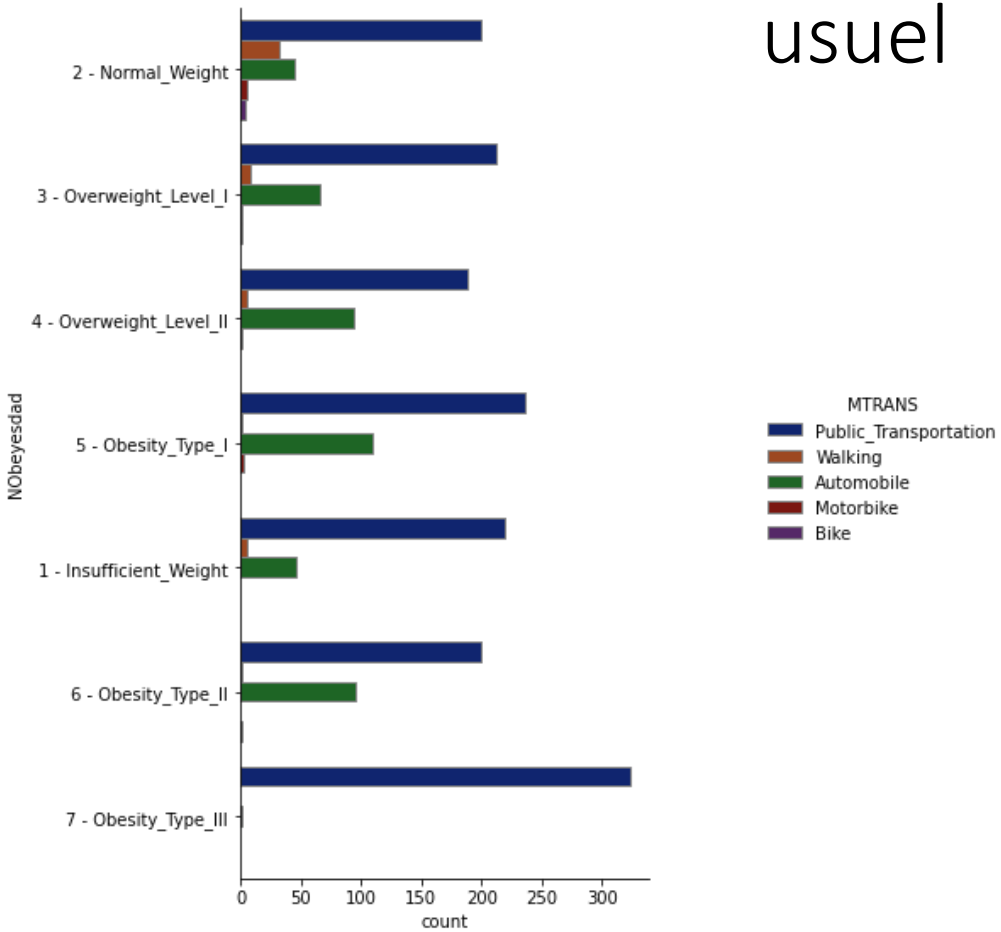
On constate que pour chaque catégorie, la plupart des personnes boivent de temps en temps, et qu’une seule personne a répondu boire tout le temps. Il semble donc difficile de dégager un lien entre la consommation d'alcool et la corpulence.

Analyse des données : le tabagisme



D’après les observations, la très grande majorité des individus ne fument pas, il est donc impossible de dégager un lien entre le tabagisme et la corpulence.

Analyse des données : le moyen de transports usuel



Les transports en commun sont les 2 principaux moyens de transport, cependant, la plupart des personnes se déplaçant à pied ou à vélo, ont une corpulence normale. Le moyen de transport semble donc avoir une influence sur la corpulence.

Conclusion de l’analyse des données

Lors de notre analyse, nous nous sommes rendu compte que pour certaines variables la répartition des individus n’était pas équilibrée. Cela étant, le dataset privilégie l’équilibre de la répartition sur la colonne cible dans l'intérêt de la construction de modèle informatique.

Nous avons remarqué certaines “bizarrerie” dans le dataset pour certaines corpulences notamment celle qui été le moins représentée dans le dataset de 485 profils issus de l'enquête. En effet il y a pour nous des biais statistiques qui ont été amplifiés par les données synthétiques notamment pour la corpulence “Obésité de type 3”. Les individus de cette corpulence sont très semblables. Il semble y avoir eu un effet “clonage” par le manque d’individus réels ayant servis de modèle. On entend par là que les biais statistiques des données réelles ont été amplifiés par “l’effet clonage”. (se référer au notebook pour plus de détails)

Analyse des données : synthèse

Nous avons obtenu une certaine classification des variables en fonction de leur influence sur la corpulence des individus. Cette influence à été déterminé par l'observation des données réalisés.

Critères les plus pertinents:

- Antécédents familiaux
- Consommation de nourriture calorique
- Surveillance du nombre de calories ingérées
- Pratique d'une activité physique
- Moyen de transport usuel

Critères moyennement pertinents:

- Sexe
- Nombre de repas quotidiens
- Grignotage entre les repas
- Consommation d'eau

Critères peu pertinents:

- Age
- Consommation de légumes
- Temps d'utilisation d'objets électroniques
- Consommation d'alcool
- Fumeur

Cette classification nous à permis de créer 3 datasets chacun avec un certain nombre de colonne. Nous avons ainsi pu tester nos modèles de machine learning par la suite suivant 3 datasets. L'idée étant de savoir si certaines colonnes pouvait être inutile.

3- Pré processing

Pré processing : reformatage des variables

- pour l'analyse des données, les variables de type float ont été converties en int afin de réduire le nombre de valeurs possibles lors des group by
- pour l'analyse des données, les âges ont été groupés en tranches d'âges, et certaines valeurs renommées pour un affichage croissant (exemple no; sometimes; frequently; always)
- Pour entraîner les modèles de machine learning, il était indispensable de convertir toutes les variables de type string en un type numérique.
  - Les Yes/No ont donc été changés en 1/0
  - Les variables de type never/sometimes/always ont été changées en 0/1/2

Pré processing : préparation des datasets

- Les 2 subsets définis plus hauts sont créés
- les datasets sont séparés de la variable à prédire.
- les sets sont ensuite splittés en train et test set
- les données sont ensuite centrées et réduites

4-Machine learning

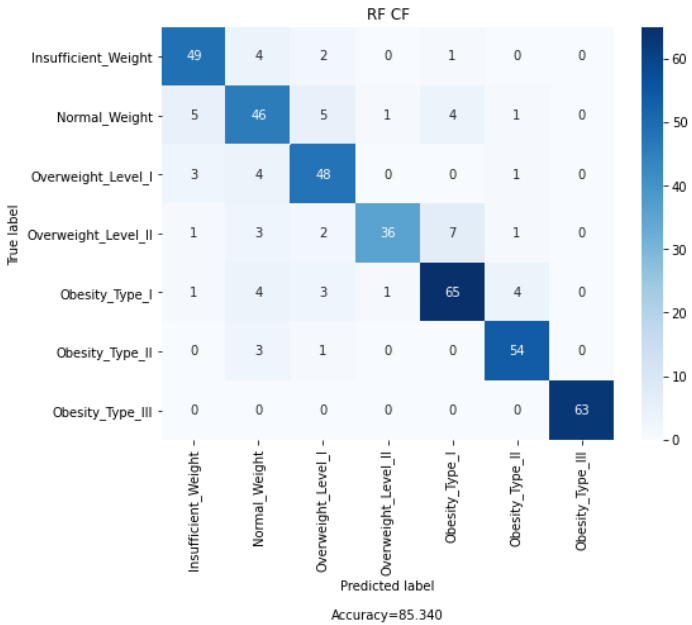
Machine learning

- En ne gardant que les critères les plus pertinents, aucun modèle n'a pu dépasser les 45% de précision moyenne
- En y ajoutant les critères moyennement pertinents, aucun modèle n'a pu dépasser les 65% de précision moyenne
- Finalement seul le dataset complet a pu donner des résultats intéressants et diversifiés.

Nous présenterons donc juste ici les modèles avec le dataset complet. Pour plus de détail se référer au notebook.

Machine learning : Random Forest

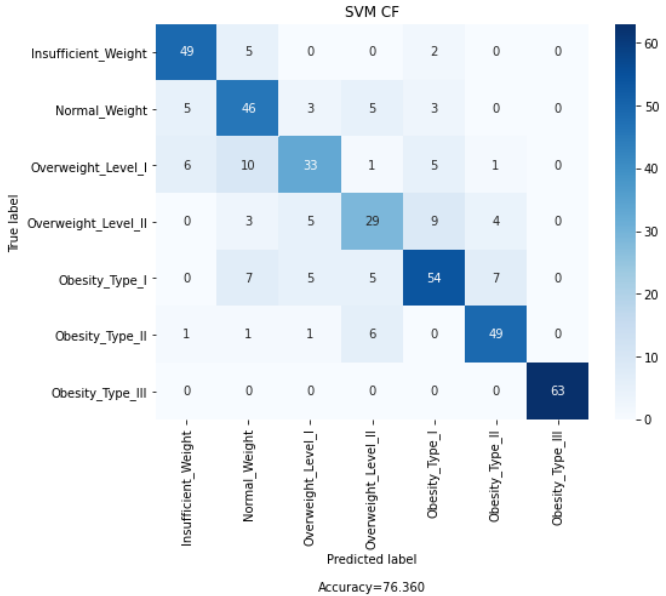
	precision	recall	f1-score	support
1 - Insufficient_Weight	0.83	0.88	0.85	56
2 - Normal_Weight	0.72	0.74	0.73	62
3 - Overweight_Level_I	0.79	0.86	0.82	56
4 - Overweight_Level_II	0.95	0.72	0.82	50
5 - Obesity_Type_I	0.84	0.83	0.84	78
6 - Obesity_Type_II	0.89	0.93	0.91	58
7 - Obesity_Type_III	1.00	1.00	1.00	63
accuracy			0.85	423
macro avg	0.86	0.85	0.85	423
weighted avg	0.86	0.85	0.85	423



Avec 86% de précision, le modèle Random forest est globalement très efficace, surtout pour l'obésité morbide. Il est néanmoins plus faible pour le poids normal étant donné la forte hétérogénéité des données. Cependant, il ne faut pas oublier que les données concernant l'obésité morbide sont souvent presque identiques, Le modèle risque donc de ne pas fonctionner pour un individu dont les caractéristiques divergent par rapport à la population du dataset. Cette catégorie est donc la moins pertinente pour évaluer l'efficacité d'un modèle

Machine learning : SVM

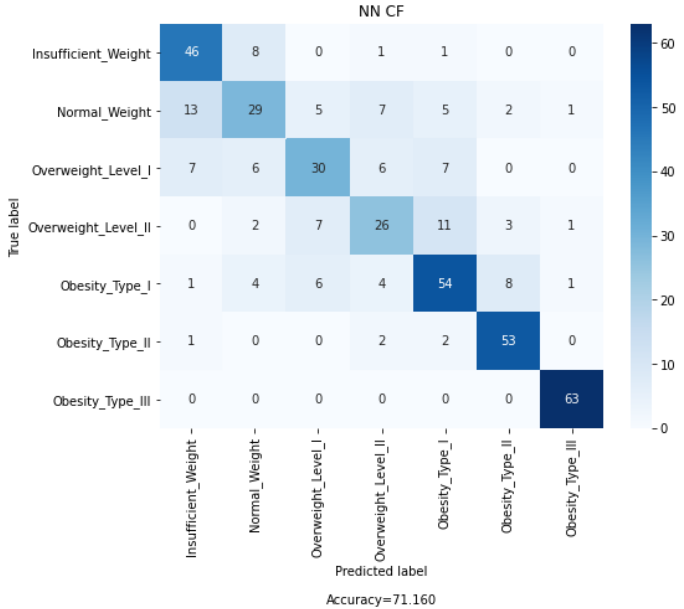
	precision	recall	f1-score	support
1 - Insufficient_Weight	0.80	0.88	0.84	56
2 - Normal_Weight	0.64	0.74	0.69	62
3 - Overweight_Level_I	0.70	0.59	0.64	56
4 - Overweight_Level_II	0.63	0.58	0.60	50
5 - Obesity_Type_I	0.74	0.69	0.72	78
6 - Obesity_Type_II	0.80	0.84	0.82	58
7 - Obesity_Type_III	1.00	1.00	1.00	63
accuracy			0.76	423
macro avg	0.76	0.76	0.76	423
weighted avg	0.76	0.76	0.76	423



Avec 76 % de précision, le modèle SVM semble assez efficace, mais il l'est surtout pour prédire l'obésité et la maigreur.
En dehors de l'obésité morbide, ce modèle semble beaucoup moins efficace que le précédent.

Machine learning : Neural Network

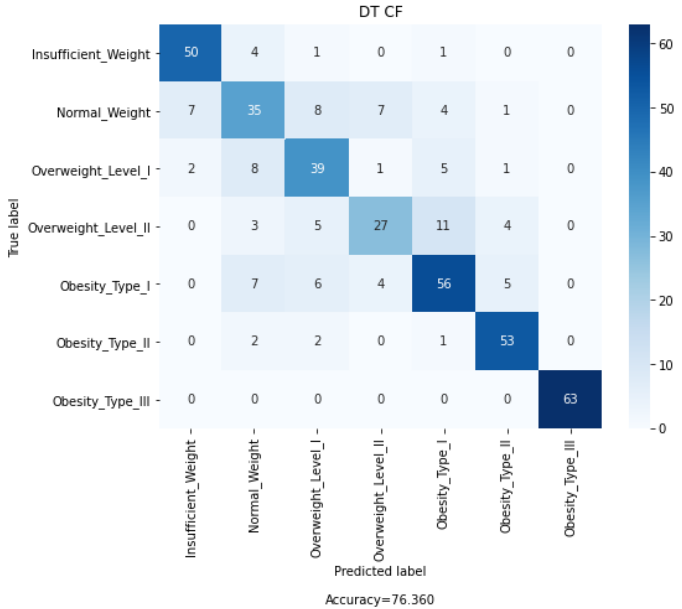
	precision	recall	f1-score	support
1 - Insufficient_Weight	0.68	0.82	0.74	56
2 - Normal_Weight	0.59	0.47	0.52	62
3 - Overweight_Level_I	0.62	0.54	0.58	56
4 - Overweight_Level_II	0.57	0.52	0.54	50
5 - Obesity_Type_I	0.68	0.69	0.68	78
6 - Obesity_Type_II	0.80	0.91	0.85	58
7 - Obesity_Type_III	0.95	1.00	0.98	63
accuracy			0.71	423
macro avg	0.70	0.71	0.70	423
weighted avg	0.70	0.71	0.70	423



Avec 70% de précision, ce modèle est surtout efficace pour prédire l'obésité de type 2 et 3. En revanche il n'est pas très efficace pour le reste

Machine learning : Decision Tree

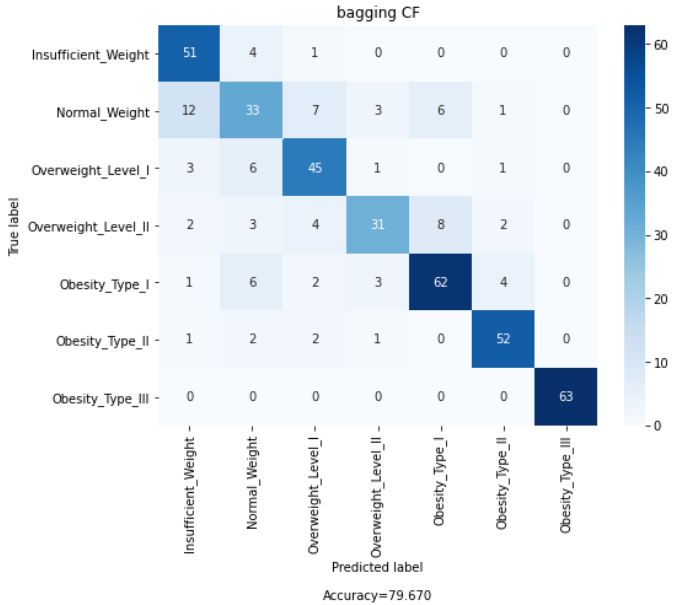
	precision	recall	f1-score	support
1 - Insufficient_Weight	0.85	0.89	0.87	56
2 - Normal_Weight	0.59	0.56	0.58	62
3 - Overweight_Level_I	0.64	0.70	0.67	56
4 - Overweight_Level_II	0.69	0.54	0.61	50
5 - Obesity_Type_I	0.72	0.72	0.72	78
6 - Obesity_Type_II	0.83	0.91	0.87	58
7 - Obesity_Type_III	1.00	1.00	1.00	63
accuracy			0.76	423
macro avg	0.76	0.76	0.76	423
weighted avg	0.76	0.76	0.76	423



On remarque que le modèle Decision tree est efficace pour prédire l'obésité ou la maigreur, mais l'est beaucoup moins pour le surpoids et la corpulence normale.

Machine learning : Bagging

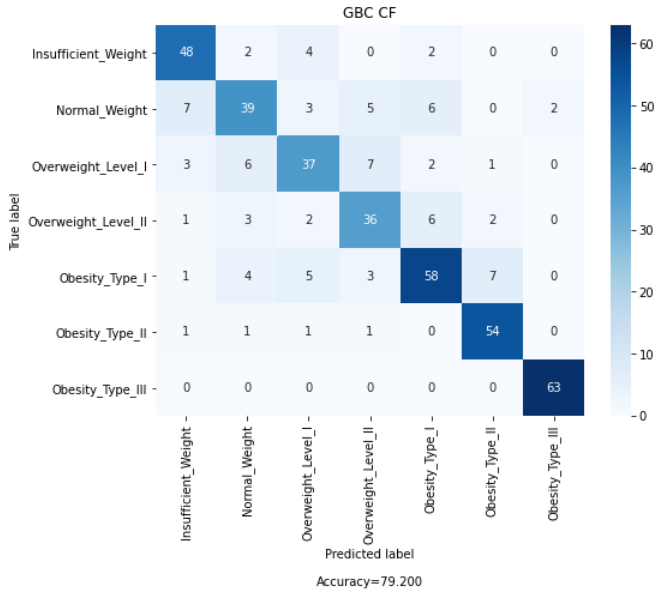
	precision	recall	f1-score	support
1 - Insufficient_Weight	0.73	0.91	0.81	56
2 - Normal_Weight	0.61	0.53	0.57	62
3 - Overweight_Level_I	0.74	0.80	0.77	56
4 - Overweight_Level_II	0.79	0.62	0.70	50
5 - Obesity_Type_I	0.82	0.79	0.81	78
6 - Obesity_Type_II	0.87	0.90	0.88	58
7 - Obesity_Type_III	1.00	1.00	1.00	63
accuracy			0.80	423
macro avg	0.79	0.79	0.79	423
weighted avg	0.80	0.80	0.79	423



Ce modèle est surtout efficace pou r prédire l'obésité, mais est très mo yen pour prédire le poids normal.

Machine learning : Gradient boosting

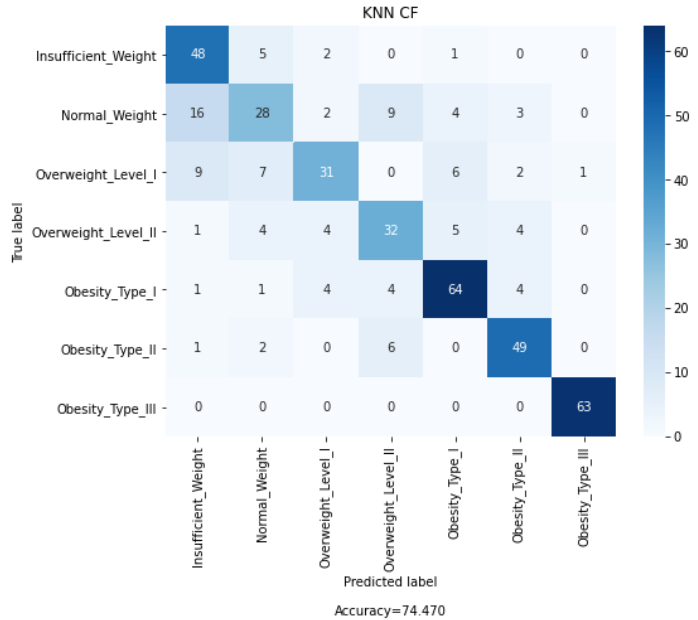
	precision	recall	f1-score	support
1 - Insufficient_Weight	0.79	0.86	0.82	56
2 - Normal_Weight	0.71	0.63	0.67	62
3 - Overweight_Level_I	0.71	0.66	0.69	56
4 - Overweight_Level_II	0.69	0.72	0.71	50
5 - Obesity_Type_I	0.78	0.74	0.76	78
6 - Obesity_Type_II	0.84	0.93	0.89	58
7 - Obesity_Type_III	0.97	1.00	0.98	63
accuracy			0.79	423
macro avg	0.79	0.79	0.79	423
weighted avg	0.79	0.79	0.79	423



Ce modèle prédit bien l'obésité et la maigreur, et arrive à prédire assez bien le poids normal et le surpoids.

Machine learning : KNN

	precision	recall	f1-score	support
1 - Insufficient_Weight	0.63	0.86	0.73	56
2 - Normal_Weight	0.60	0.45	0.51	62
3 - Overweight_Level_I	0.72	0.55	0.63	56
4 - Overweight_Level_II	0.63	0.64	0.63	50
5 - Obesity_Type_I	0.80	0.82	0.81	78
6 - Obesity_Type_II	0.79	0.84	0.82	58
7 - Obesity_Type_III	0.98	1.00	0.99	63
accuracy			0.74	423
macro avg	0.74	0.74	0.73	423
weighted avg	0.74	0.74	0.74	423



Le modèle KNN est surtout efficace pour prédire l'obésité, mais l'est beaucoup moins pour le reste.

Machine learning : synthèse

- Du fait de la grande homogénéité des données sur l'obésité morbide, tous les modèles sont très efficaces pour cette catégorie (entre 95 et 100% de précision)
- La plupart des modèles sont efficaces pour prédire l'obésité de type 1 et 2, ainsi que la maigreur
- En revanche, le surpoids et surtout la corpulence normale posent plus de problèmes, aucun modèle ne dépasse les 75% de précision pour ces catégories (à l'exception de Random forest)
- Random Forest est globalement le modèle le plus efficace, et l'est aussi pour la plupart des catégories
- Le réseau de neurones est le modèle le moins efficace, en plus d'être beaucoup plus lent à exécuter que les autres.

Retour sur les biais identifié

Comme nous l'avons évoqué dans la partie analyse des données, nous avons identifié que les individus de la catégorie "Obésité de type 3" sont très semblables et témoignent d'un effet clone provenant du manque de données réelles. Ce problème se répercute sur nos différents modèles de machine learning entraînés. Les résultats de classification pour cette corpulence sont excellents mais témoignent d'un surapprentissage. Les modèles reconnaissent surement les "clones" et malheureusement il y a fort à craindre que la reconnaissance de cette catégorie par les modèles fonctionne moins bien avec de nouvelles données de personnes appartenant à cette dite catégorie.

Ce problème est néanmoins cantonné à une seule corpulence.

5 - API

Création de l’API

Nous avons utilisé le modèle RandomForest préalablement entraîné pour créer une API prédisant la corpulence d’un individu suivant les 14 variables du dataset étudié.

Cette API est codée en python et utilise le micro framework open-source de développement web Flask. Cette API prend en entrée un tableau de données en JSON et renvoie en JSON la prédiction de la corpulence.

Concernant les informations pour l’utiliser et déployer l’API se référer au ReadMe partie “Comment utiliser l'API”.