Ecrit par : Fillot Romain, Sabatier Audric

Rapport Modélisation

Analyse et explication de notre travail

2024

Contexte du projet

Notre projet vise à mettre en pratique nos connaissances en matière d'apprentissage automatique. L'objectif principal est de construire un modèle aussi précis que possible pour prédire le taux d'obésité, traduit par l'Indice de Masse Corporelle (IMC). Cette problématique revêt une importance croissante dans le contexte actuel de préoccupation mondiale concernant la santé publique et le bien-être individuel.

Plan

Données

Cette partie a pour but de décrire nos données ainsi que d'évoquer les problèmes rencontrés lors de notre exploration, ainsi que la façon dont nous les avons résolus. Nous vous présenterons également quelques statistiques.

Modèles et apprentissage

Ici, nous évoquerons nos choix de modèles en les présentant et en donnant des détails sur les paramètres utilisés. Nous expliquerons également notre méthode d'entraînement et de tests.

Métriques et décision

Cette partie a pour but d'évaluer et de critiquer la performance de nos modèles. Nous examinerons ces indicateurs pour donner un avis sur nos choix.

Conclusion

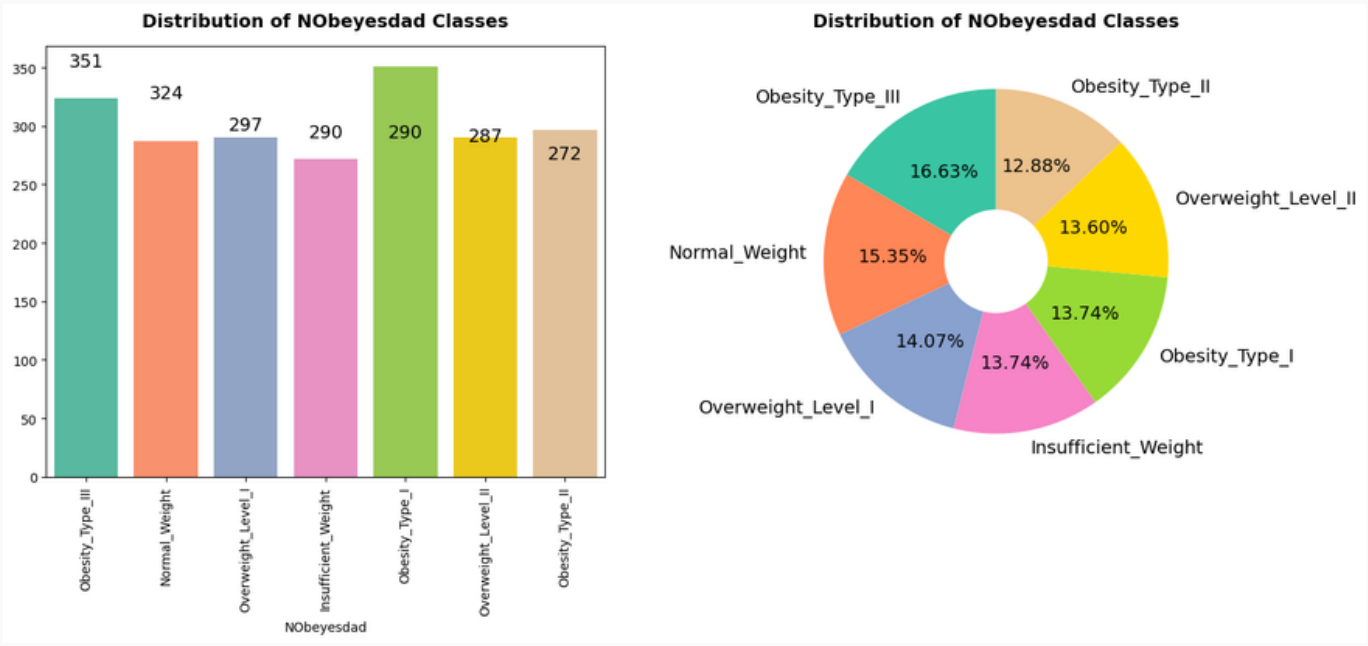
Nous vous proposerons une conclusion synthétisant notre travail en présentant des pistes d'amélioration futures.

Données

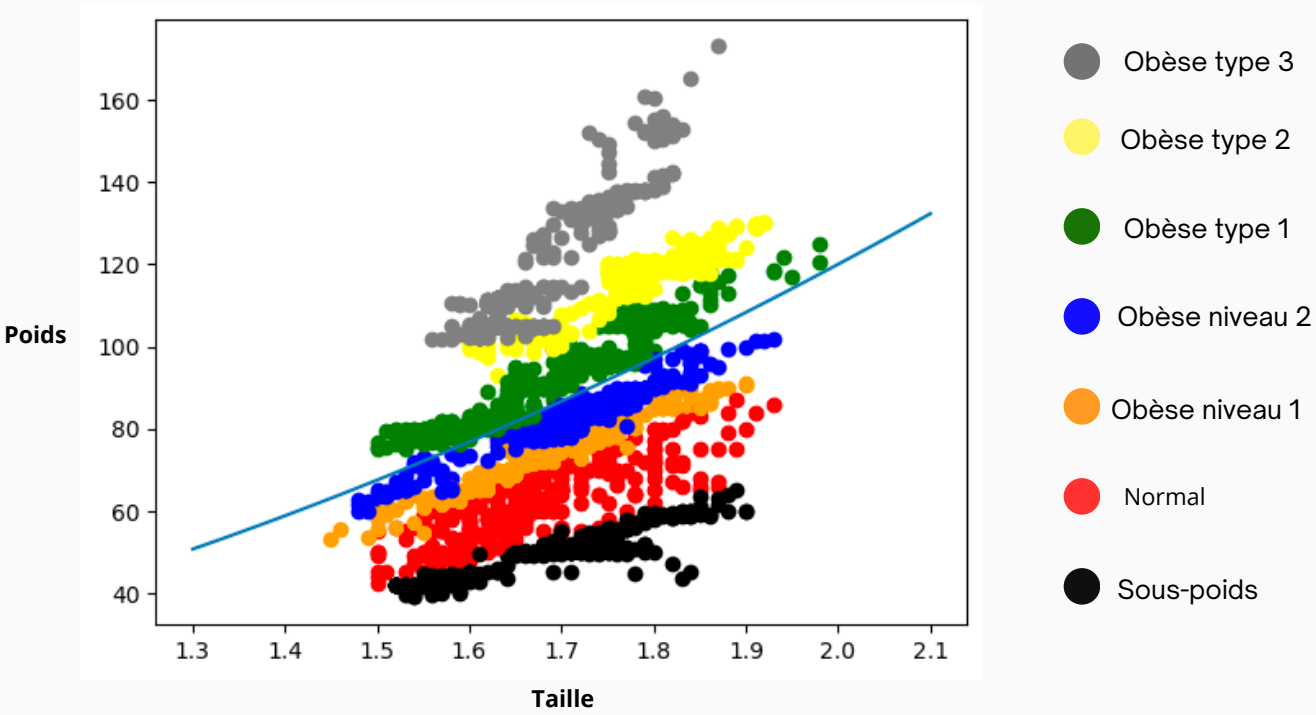
Nos données proviennent d'un sondage proposé à des Mexicains, Péruviens et Colombiens lors de cette étude durant laquelle, des questions ont étés posées à de multiples personnes. Nous voyons que nous avons deux grandes catégories de variables qui résultent de ces question : les habitudes ainsi que les caractéristiques des personnes.

Ci- dessous un récapitulatif des variables de l'étude :

Variable	Question (in survey)	value	signification
Gender	What is your gender?	Female / Male	Gender
Age	what is your age?	-	Age
Height	what is your height?	-	Height
Weight	what is your weight?	-	Weight
family_history_with_overweight	Has a family member suffered or suffers from overweight?	Yes/No	If member of family have overweight
FAVC	Do you eat high caloric food frequently?	Yes/No	If people eat high caloric food frequently
FCVC	Do you usually eat vegetables in your meals?	Never/Sometimes/Always	Frequency of vegetables eating
NCP	How many main meals do you have daily?	1-2/3/>3	Number a meals/day
CAEC	Do you eat any food between meals?	No/Sometimes/Frequently/sometimes	If poeple eat between meals
SMOKE	Do you smoke?	yes/no	If people smoke
CH2O	How much water do you drink daily?	<1L/1-2L/>2L	Quantity of water drinks/day
SCC	Do you monitor the calories you eat daily?	yes/no	If people count calories eats
FAF	How often do you have physical activity?	NOT/1-2 days/2-4days/4-5days	Number of days wuth physical activity/week
TUE	How much time do you use technological devices such as cell phone, videogames, television, computer and others?	0-2/3-5/>5	Hours/days with technological devices
CALC	how often do you drink alcohol?	NOT/sometimes/Frequently/always	Level of alcohol drinking
MTRANS	Which transportation do you usually use?	Automobile/Motorbike/Bike/Public Transportation/Walking	Transportation that people use



Nous voyons sur les figures ci-dessus que la répartition des différents types d'obésité est quasiment égale. Nous avons donc une très bonne répartition de la variable cible, nous n'avons donc pas besoin d'effectuer de sur/sous-échantillonnage.



Ci-dessus, la représentation de l'IMC des personnes : poids en kg/taille^2 (en m). Nous voyons donc que la représentation de notre variable cible est linéaire avec deux features (taille, poids).

Lors de l'exploration de nos données, nous avons identifié quelques problèmes, nécessitant ainsi une préparation de ces données pour les rendre exploitables.

Préparation

Nous avons rencontré des incohérences dans les données, ce qui a nécessité une série d'actions pour assurer leur qualité. Tout d'abord, nous avons identifié que les âges et les nombres de repas dans la journée étaient représentés de manière inconsistante, certains en nombres réels et d'autres en entiers. Pour homogénéiser cela, nous avons pris la décision d'arrondir tous les âges au nombre entier le plus proche.

Ensuite, afin de traiter les variables comportant des champs textuels, nous avons opté pour la discrétisation en associant un nombre entier à chaque valeur, par exemple en associant "bus" à 1 pour la variable "Moyen de transport".

Par ailleurs, la présence de lignes dupliquées a été remarquée. Étant donné la nature de notre échantillon et considérant qu'il est improbable que deux individus répondent de manière exactement identique, nous avons procédé à la suppression de ces valeurs en doublon. Cela a réduit notre ensemble de données de 2111 à 2087 lignes.

Enfin, selon le modèle sélectionné, nous avons également pris en compte la normalisation des données pour assurer une comparaison équitable entre les différentes caractéristiques.

Nos modèles

Linear SVC (Support Vector Classifier)

Pour choisir ce modèle, nous avons utilisé le "scikit-learn algorithm cheat-sheet", qui est un organigramme conçu pour donner aux utilisateurs un guide approximatif sur la façon d'aborder les problèmes concernant les estimateurs à essayer sur nos données.
Après avoir répondu aux questions,nous sommes arrivé au Linear SVC. Ce modèle est une variante du classificateur des machines à vecteurs de support (SVM - Support Vector Machine) utilisé pour des tâches de classification dans l'apprentissage automatique supervisé. Contrairement à la version classique des SVM, qui est souvent utilisée pour des tâches de classification non linéaire en utilisant des noyaux non linéaires, le Linear SVC est spécifiquement conçu pour les tâches de classification linéaire.

Le fonctionnement de création est le suivant :

- **Hyperplan de décision linéaire :** Le modèle Linear SVC cherche à trouver un hyperplan de décision linéaire qui sépare les données en classes dans un espace de caractéristiques. Cet hyperplan est une frontière de décision linéaire qui maximise la marge entre les classes.
- **Optimisation de la marge :** La marge est la distance entre l'hyperplan de décision et les exemples les plus proches de chaque classe, appelés vecteurs de support.

Paramètres :

- **mutli_class** : Ce paramètre définit la stratégie de classification multi-classe à utiliser dans de tels cas, soit ovr ou crammer_singer. **ovr** entraîne les classificateurs n_classes one-vs-rest, tandis que **crammer_singer** optimise un objectif commun sur toutes les classes

- **dual** : Sélectionnez l'algorithme pour résoudre le problème d'optimisation double ou primaire(**false**, **true** ou **auto**)

- **max_iter** : Le nombre maximum d'itérations à exécuter

- **C** : “la force de régularisation” qui peut avoir un impact significatif sur la convergence de LinearSVC

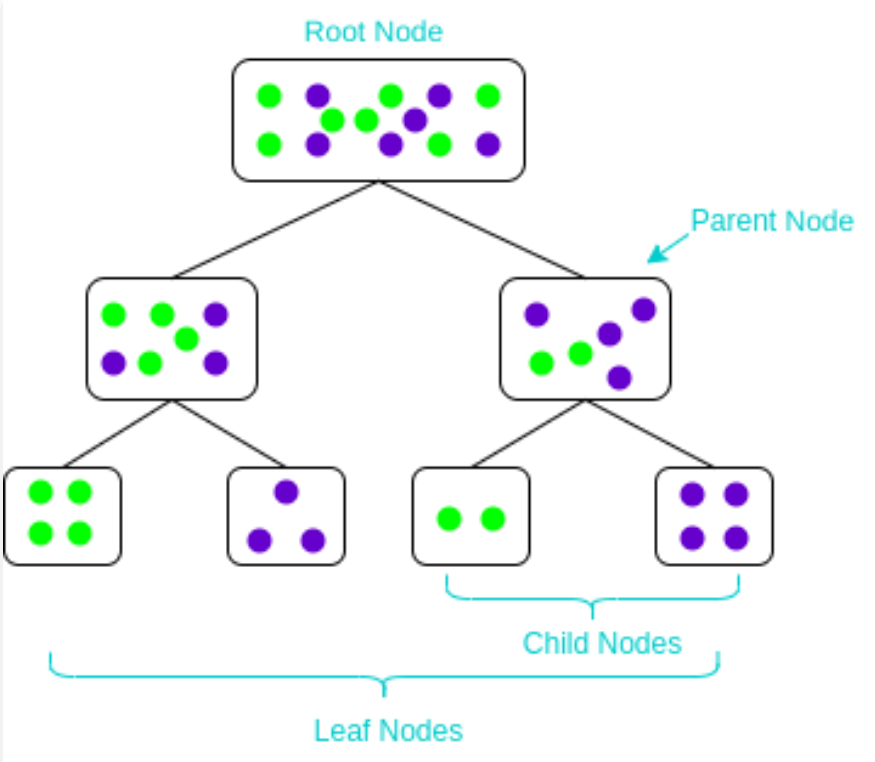
Nos modèles

Decision Tree Classifier

Nous avons décidé d'implémenter un modèle non linéaire pour tester cette approche. Ce modèle est basé sur un arbre de décision, qui est la méthode que nous connaissons le mieux suite à nos précédents cours.

Le fonctionnement de création est le suivant :

- Le root node représente toutes les données de bases
- Ensuite dans chaque noeud, on décide quelle variables scinde le mieux les données de manière homogène via l'entropie qui défini à chaque étape la variables la plus pertinente



Paramètres :

- **entropy** : permet de choisir l'entropie comme méthode de sélection des variables à chaque noeuds

- **random_state :** contrôle la génération de nombres pseudo-aléatoires il nous permet de garantir la reproductibilité des résultats lors de l'exécution du code -> utile pour le débogage

- **splitter (best) :** l'algorithme sélectionne la meilleure division à chaque nœud en maximisant la réduction de l'impureté (entropie)

Apprentissage
Entraînement

Pour l'entrainement nous avons séparé notre jeu de données en deux jeux de données :

Ainsi, nous avons essayé différents paramètres pour nos deux modèles afin de déterminer les plus intéressants. De plus, nous avons réalisé différentes combinaisons de caractéristiques afin d'évaluer leur importance. Pour le modèle **LinearSVC**, nous avons normalisé les données, car cela garantit que toutes les caractéristiques ont la même échelle, ce qui peut améliorer les performances du modèle. Pour le **DecisionTreeClassifier**, nous n'avons pas besoin de normaliser les données, car les divisions dans un arbre de décision sont basées sur les valeurs brutes des caractéristiques et non sur leurs échelles. Ainsi, la normalisation des données n'a pas d'impact direct sur la performance du modèle.

Le meilleur résultat pour les deux modèles est obtenu lorsque toutes les caractéristiques sont incluses. Bien entendu, certaines ont une plus grande importance, mais le fait de conserver uniquement celles qui sont les plus importantes n'améliore pas la performance globale du modèle.

Métriques choisies

Nous avons décidé de choisir trois métriques pour évaluer la performance de nos modèles :

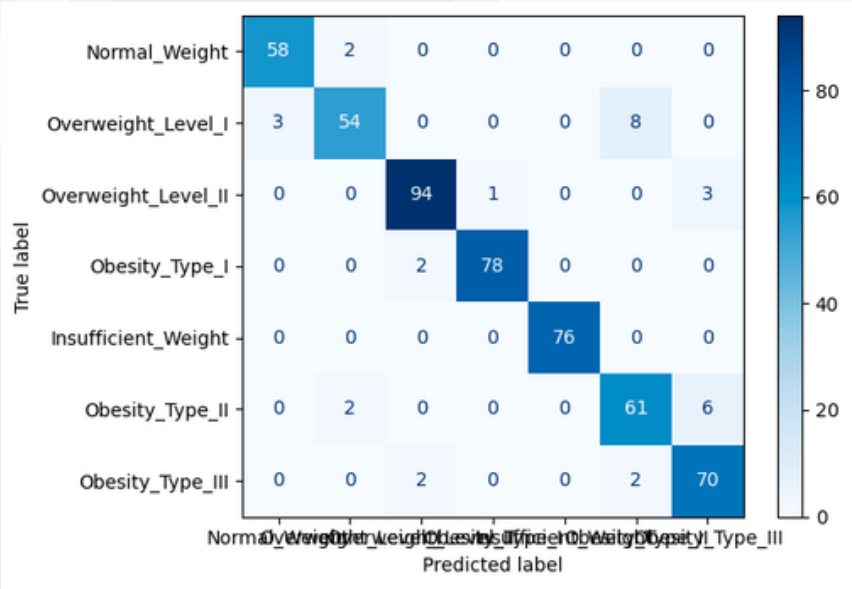
Accuracy **F1-score** **Matrice de confusion**

C'est métrique sont très souvent utilisées pour mesurer la performance d'un modèle de classification. Ce sont les plus facile à obtenir et à comprendre.

Voici les différents résultats que l'on obtient avec nos deux modèles :

DecisionTreeClassifier

Matrice de confusion



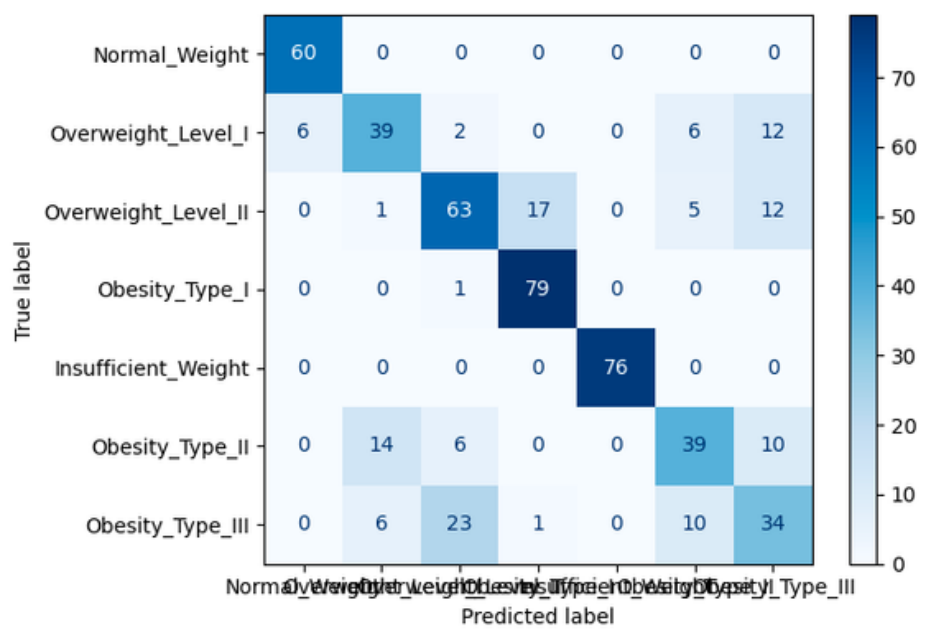
...	precision	recall	f1-score	support
Insufficient_Weight	0.95	0.97	0.96	60
Normal_Weight	0.93	0.83	0.88	65
Obesity_Type_I	0.96	0.96	0.96	98
Obesity_Type_II	0.99	0.97	0.98	80
Obesity_Type_III	1.00	1.00	1.00	76
Overweight_Level_I	0.86	0.88	0.87	69
Overweight_Level_II	0.89	0.95	0.92	74
accuracy			0.94	522
macro avg	0.94	0.94	0.94	522
weighted avg	0.94	0.94	0.94	522

classification report

Métriques choisies

LinearCSV

Matrice de confusion



classification report

...	precision	recall	f1-score	support
Insufficient_Weight	0.91	1.00	0.95	60
Normal_Weight	0.65	0.60	0.62	65
Obesity_Type_I	0.66	0.64	0.65	98
Obesity_Type_II	0.81	0.99	0.89	80
Obesity_Type_III	1.00	1.00	1.00	76
Overweight_Level_I	0.65	0.57	0.60	69
Overweight_Level_II	0.50	0.46	0.48	74
accuracy			0.75	522
macro avg	0.74	0.75	0.74	522
weighted avg	0.74	0.75	0.74	522

Décision

Synthèse des résultats

Il est facile de remarquer que le DecisionTreeClassifier est plus performant que le LinearSVC que ce soit niveau accuracy, recall et f1-score.

Nous pouvons en déduire plusieurs choses :

- Il semble y avoir une faible, voire aucune, relation linéaire entre les données. À l'exception de la taille et du poids, les caractéristiques de nos données ne présentent pas de tendances linéaires claires, ce qui explique les difficultés rencontrées par le modèle LinearSVC.
- Les modèles d'arbres de décision, y compris le DecisionTreeClassifier, sont généralement plus robustes au bruit dans les données que les modèles linéaires comme le LinearSVC. Nous avons du arrondir quelques données, ce qui à peut-être créées des valeurs aberrantes.

En outre, pour obtenir une précision de 0,74 avec le LinearSVC, nous avons dû expérimenter avec de nombreux paramètres différents et avons rencontré plusieurs problèmes de non-convergence. Pour résoudre ces problèmes, nous avons dû augmenter le nombre maximal d'itérations, ce qui a naturellement prolongé le temps d'exécution du programme. En revanche, pour le DecisionTreeClassifier, nous avons ajusté quelques paramètres et obtenu rapidement un score dépassant 0,90.

L’accuracy finale étant de 0,94 nous pouvons conclure que notre modèle est plutôt très performant et que nous pouvons en rester là.

Conclusion

Nous observons avec satisfaction qu'un de nos modèles présente un taux de prédiction extrêmement élevé, dépassant les 90%. Cette précision est un indicateur solide de notre capacité à prédire efficacement le niveau d'obésité d'une personne. Cependant, il convient d'aborder cette affirmation avec nuance, car elle repose sur l'utilisation de l'Indice de Masse Corporelle (IMC) pour évaluer l'obésité. Nous remettons en question la pertinence de l'IMC dans notre contexte, car il peut conduire à des résultats erronés, notamment en classant comme obèse des individus athlétiques qui ont une masse musculaire élevée. Cette limitation soulève la question de la validité de l'IMC comme mesure fiable de l'obésité. Nous sommes conscients de la complexité de cette problématique et de la nécessité d'explorer des alternatives plus précises pour évaluer le taux d'obésité.

Amélioration possible dans le futur

Nous pourrions intégrer la validation croisée (cross-validation) afin de diviser les données en plusieurs sous-ensembles, permettant ainsi une évaluation plus robuste du modèle. Avec un faible nombre de données comme nous avons (~2100) elle pourrait permettre d'utiliser toutes les données pour l'entraînement et l'évaluation, améliorant ainsi la fiabilité des performances estimées du modèle. De plus, en effectuant plusieurs itérations de formation et d'évaluation sur différents sous-ensembles des données, la validation croisée réduit la variance des performances du modèle. Nous pensons donc avoir de meilleurs résultats grâce à son utilisation. Il serait aussi intéressant d'essayer plusieurs autres modèles réputés de classification à savoir le Naive Bayes ou encore le Kernel Approximation qui utilisent d'autres méthodes de classification.

Merci pour votre attention

Fillot Romain, Sabatier Audric