Estimation of obesity levels based on eating habits and physical condition - Data Set

https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+

Carla CAUCHE – janvier 2021

Sommaire

- Contexte
- Data
- Préparation des données
- Modélisation

Contexte

Tenants et aboutissement

- Ce dataset à été créer pour but d'**estimer le niveau d'obésité** des populations de Colombie, Pérou et Mexique à partir de leurs habitudes alimentaires et condition physique.
- Il a été créer par 2 chercheurs de l'Universidad de la Costa en **Colombie**. C'est en partie pour cela que la population d'Amérique Latine est ciblée.
- Il peut également avoir pour but d'**identifier l'impact** de certains facteurs sur l'apparition de problème d'obésité.

Problème

- En 2017, **60%** des sud-américain étaient en surpoids.
- En 2019, ces chiffres augmentent à **70%** pour les Mexicains.

- L'obésité peut présenter un vrai danger et pouvoir le prévoir pourrait aider à le prévenir.

- Nous allons donc pour la suite chercher à **prédire l'indice d'obésité** à partir de plusieurs facteurs.

Data

Les données

- Le dataset contient 2111 informations sur 17 variables
- 23% des données ont été collecté directement à travers un sondage et les autres **77% ont été généré synthétiquement**.
- Il faut prendre en compte qu'une majorité des données sont synthétique et peuvent donc ne pas représenter au mieux la population.

Les variables

Attribut	Type	Signification
Gender	{"Female","Male"}	
Age	int	
Height	mètre	
Weight	kg	
Family history w/ overweight	{"yes","no"}	
Frequent consumption of high caloric food (FAVC)	{"yes","no"}	
Frequency of consumption of vegetables (FCVC)	1, 2, 3	1:Never; 2:Sometimes; 3:Always
Number of main meals (NCP)	1, 2, 3, 4	
Consumption of food between meals (CAEC)	{"No","Sometimes","Frequently","Always"}	
Smoke	{"yes","no"}	
Consumption of water daily (CH2O)	1, 2, 3	1:Less than a liter; 2:Between 1 and 2L; 3:More than 2L
Calories consumption monitoring (SCC)	{"yes","no"}	
Physical activity frequency (FAF)	0, 1, 2, 3	0:None; 1:1 or 2 days; 2:2 or 4 days; 3:4 or 5 days
Time using technology devices (TUE)	0, 1, 2	0:0-2 hours; 1:3-5 hours; 2:More than 5 hours
Consumption of alcohol (CALC)	{"No","Sometimes","Frequently","Always"}	
Transportation used (MTRANS)	{"Automobile","Bike","Motorbike","Public_Transportation","Walking"}	
NObesity	{"Insufficient Weight", "Normal Weight", "Overweight Level I", "Overweight Level II", "Obesity Type I", "Obesity Type II", "Obesity Type III"}	

- Seuls les variables *Age*, *Height* et *Weight* sont continues.
- Les autres variables sont discrètes et qualitatives.

Cible

- NObesity est déterminé à partir de l'IMC = $\frac{Poids}{Taille^2}$

- Les variables *Height* et *Weight* ne sont donc pas intéressantes. Si elles sont connues, une prédiction n'est pas nécessaire.

NObesity	IMC
Insufficient Weight	< 18.5
Normal Weight	18.5 - 24.9
Overweight	25 - 29.9
Obesity Type I	30 - 34.9
Obesity Type II	35 - 39.9
Obesity Type III	> 40

Préparation des données

Traitement

- Les données générées synthétiquement ne sont pas sous la bonne forme et doivent donc être modifiées.
- Par exemple FAF devrait avoir comme valeur 0, 1, 2 ou 3 et TUE devrait avoir 0, 1 ou 2 mais ce n'est pas le cas.
- Voici donc les changements effectués: pour FCVC, NCP, CH2O, FAF et TUE

$0 < x \leq 1$	1
$1 < x \leq 2$	2
$2 < x \leq 3$	3
$3 < x \leq 4$	4

FAF	TUE
0.000000	1.000000
3.000000	0.000000
2.000000	1.000000
2.000000	0.000000
0.000000	0.000000
...	...
1.676269	0.906247
1.341390	0.599270
1.414209	0.646288
1.139107	0.586035
1.026452	0.714137

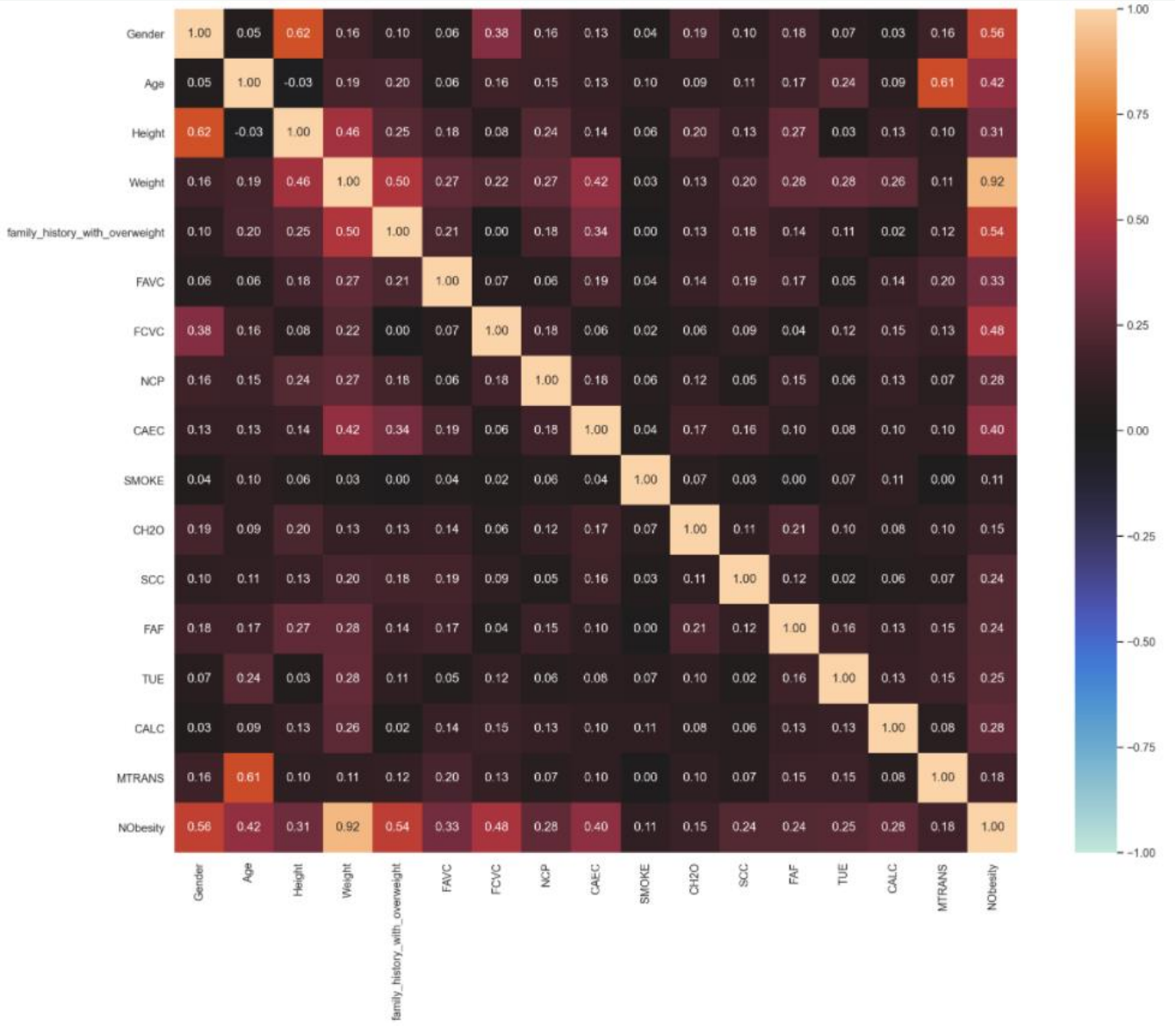
Encoding variables qualitatives

Attribut	Encoding
Gender	0:"Female",1:"Male"
Family history w/ overweight	1:"yes",0:"no"
Frequent consumption of high caloric food (FAVC)	1:"yes",0:"no"
Consumption of food between meals (CAEC)	0:"No",1:"Sometimes",2:"Frequently",3:"Always"
Smoke	1:"yes",0:"no"
Calories consumption monitoring (SCC)	1:"yes",0:"no"
Consumption of alcohol (CALC)	0:"No",1:"Sometimes",2:"Frequently",3:"Always"
Transportation used (MTRANS)	1:"Automobile",2:"Bike",3:"Motorbike",4:"Public_Transportation",5:"Walking"
NObesity	0:"Insufficient Weight", 1:"Normal Weight", 2:"Overweight Level I", 3:"Overweight Level II", 4:"Obesity Type I", 5:"Obesity Type II", 6:"Obesity Type III"

Association entre les variables

- Nous calculons l'associations entre les variables qualitatives en utilisant la méthode Cramer's V.

- Nous pouvons voir que *Weight* est fortement associé à NObesity, comme évoqué précédemment *Weight* est utilisé pour définir notre cible ce qui est donc normal.
- *SMOKE*, *MTRANS* et *CH2O* semblent avoir une très faible relation avec NObesity.

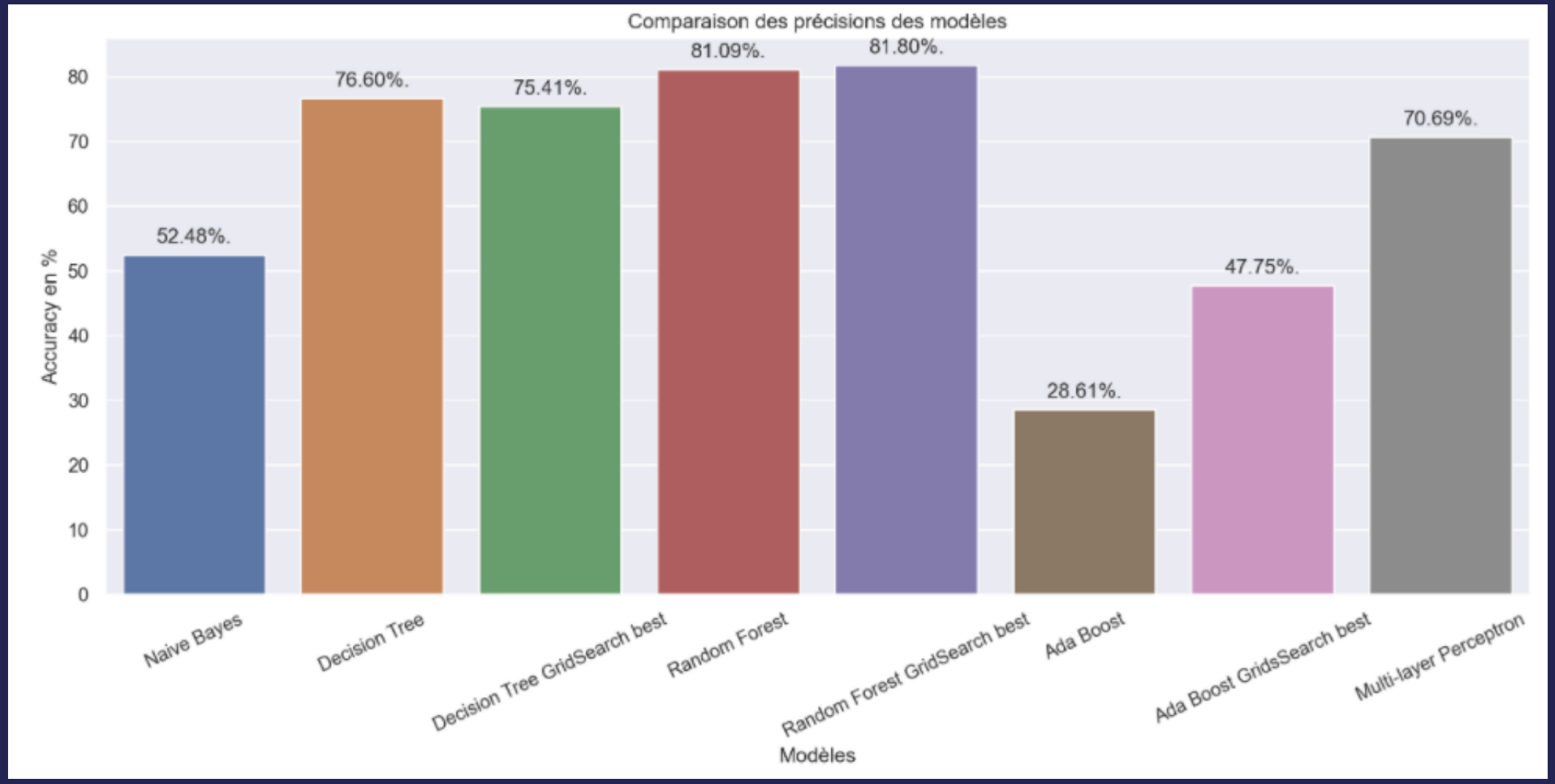


Modélisation

Classification

- Puisque la cible NObesity est discrète, nous avons un problème de classification.
- Voici donc quelques méthodes qui peuvent être intéressantes pour le traitement de notre problème :
  - Naive Bayes
  - Decision Tree
  - Random Forest
  - Ada Boost
  - Multi-layer Perceptron

Comparaison



Conclusion

- Ainsi, le modèle créé à partir de Random Forest est le plus performant des modèles testés.
- J'ai donc basé mon API sur celui-ci.