Machine Learning M – 2023/24
FINAL PROJECT

Estimation of Obesity Levels based on Eating Habits and Physical Condition with Machine Learning

Author: Aníbal Hernando Novo

Abstract

In this project my objective is to analyze the healthy and physical variables given in the dataset[1] to estimate the obesity levels of each person. The data includes eating habits and physical condition of individuals from the countries of Mexico, Peru and Colombia.

The analysis is done throw different types of clustering methods like K-means, Hierarchical Clustering and Dbscan.

In the second part of the project, I use some classification models like K-Nearest Neighbors, Support Vector Machine, Decision Tree and Random Forest to make future predictions about the obesity level of a new person based on the obtained variables after the data collection. As a result, most of the models obtain high accuracy confirming their good performance.

Contents

1. Introduction

The dataset is from UC Irvine ML Repository, is about health and medicine area, has 2111 instances and 17 features. Each instance refers to some questions asked to some people in Mexico, Peru and Colombia. The variable target "obesity level" is known as IMC^2 (Corporal Mass Index) and is obtain from the next formula:

$$IMC = weight(kg) \ / \ height(m)^2$$

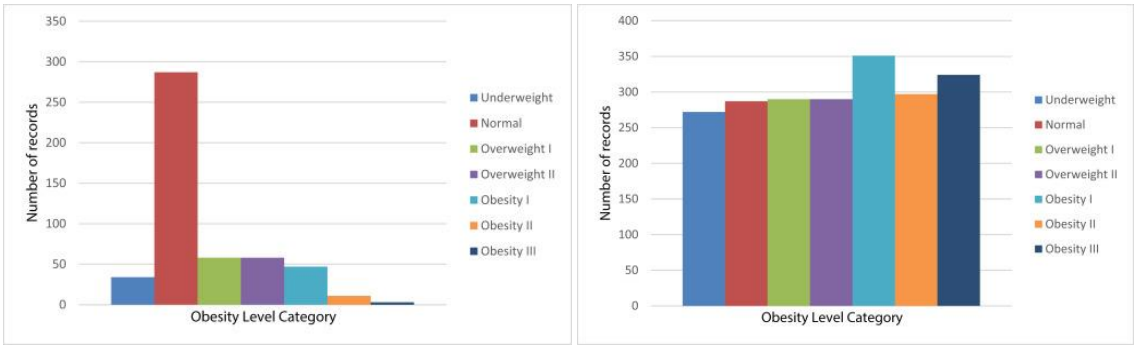
It is known that eating habits and physical condition are very important to have a good healthy. Nowadays, we have different types to classify people based in their obesity levels:

- Insufficient Weight: IMC <18,5
- Normal Weight: IMC 18,5 - 24,9
- Overweight Level I: IMC 25 -29.9
- Overweight Level II: IMC 30-34.9
- Obesity Level I: IMC 35-39,9
- Obesity Level II (Morbid): IMC 40-49,9
- Obesity Level III (Extreme): IMC >50

After the labeling process was finished, the categories of obesity levels were unbalanced (Fig.1.a), and this presented a learning problem for the data analysis methods.

Normal class had most of the instances because this data (23%) was collected directly from users through a survey in a web platform. To solve the imbalanced data problem, they have generated

data synthetically (77%) using the Weka tool and the SMOTE[3] filter. The class distribution is in Fig.1.b.



(a) (b)

Figure 1. Imbalanced data (a), Balanced data (b)

1.1. Objectives

In this project I want to differentiate between two types of Machine Learning (ML) techniques to analyze the data.

The first one is Clustering, this is aimed to do an analysis when you don't have a variable 'label' and create the clusters without prior information (Unsupervised Learning). The objective is to create clusters where the values between clusters are different from each other (heterogeneous) and their internal values are similar (homogeneous). In other words, I want to see if the difference between people that has good health and physical habits is represented in the obesity levels.

In this type of analysis, I have done some methods like K-means (iterative distance-based), Hierarchical clustering, EM-cluster (probabilistic model-based) and Dbscan (density-based).

The second one is Classification, which objective is classifying people based in some healthy and physical features in different obesity levels (Supervised Learning) with a 'label' variable that indicates the level. The goal is getting the best model that classifies people obesity level based in the obtained data, and then classify with high accuracy any instance in the correct level. I use methods like K-Nearest Neighbors, Support Vector Machines (SVM), Decision Trees and Random Forest (RF).

2. Preprocessing

2.1. Variables

The dataset has 17 variables and there aren't missing values, this helps us to do the analysis. The data is balanced so I don't have problems like:

- Biased Predictions: model could become biased towards the majority class, leading to poor performance on the minority class to unseen data.
- Misleading Accuracy: accuracy is not a good metric for imbalanced datasets because the model can achieve high accuracy predicting the majority class all the time.

Now, I am going to explain the variables.

- *"Gender"* (Categorical):
  - Male: 1
  - Female: 0
- *"Age"* (Continuous): in years
- *"Height"* (Continuous): in meters
- *"Weight"* (Continuous): in kilograms
- *"FHWO (Family history with overweight)"* (Binary):
  Has a family member suffered or suffers from overweight?
  - Yes: 1
  - No: 0
- "*FAVC"* (Binary):
  Frequent consumption of high caloric food?
  - Yes: 1
  - No: 0
- *"FCVC"* (Integer)
  Frequency of consumption of vegetables?
  - Never: 1
  - Sometimes: 2
  - Always: 3
- *"NCP"* (Continuous)
  How many main meals do you have daily?
  - Between 1 y 2: 1
  - Three: 2
  - More than three: 3
- *"CAEC"* (Categorical)
  Do you eat any food between meals?
  - No: 0
  - Sometimes: 1
  - Frequently: 2
  - Always: 3
- *"SMOKE"* (Binary):
  Do you smoke?
  - Yes: 1
  - No: 0
- *"CH20"* (Continuous):
  How much water do you drink daily?
  - Less than a liter: 1
  - Between 1 and 2 liters: 2
  - More than 2 liters: 3
- *"SCC"* (binary):
  Do you monitor the calories you eat daily?
  - Yes: 1
  - No: 0
- *"FAF"* (Continuous):
  How often do you have physical activity?
  - I do not have: 0
  - 1-2 days: 1
  - 2-4 days: 2
  - 4-5 days: 3
- *"TUE"* (Integer):
  How much time do you use technological devices such as cell phone, videogames, television, computer and others?
  - 0–2 hours: 0
  - 3–5 hours: 1
  - More than 5 hours: 2
- *"CALC"* (Categorical):
  How often do you drink alcohol?
  - I do not drink: 0
  - Sometimes: 1
  - Frequently: 2
  - Always: 3
- *"MTRANS"* (Categorical):
  Which transportation do you usually use?
  - Walking: 0
  - Bike: 1
  - Motorbike:2
  - Public Transport: 3
  - Automobile:4
- *"NObesydad"* (Categorical):
  Obesity Level
  - Insufficient Weight
  - Normal Weight
  - Overweight Level I
  - Overweight Level II
  - Obesity Type I
  - Obesity Type II
  - Obesity Type

2.2. Distribution of Variables

After using SMOTE the data has been balanced:

Insufficient Weight	Normal Weight	Obesity Type I	Obesity Type II	Obesity Type III	Overweight Level I	Overweight Level II
272	287	351	297	324	290	290

I going to use distance-based clustering methods, so I think that it is also important to evaluate methods such as normalization. The variables have different ranges which would lead to different weights in the various algorithms. For this reason, I scaled the variables to see their distribution. The next plot shows the scaled distribution of variables.

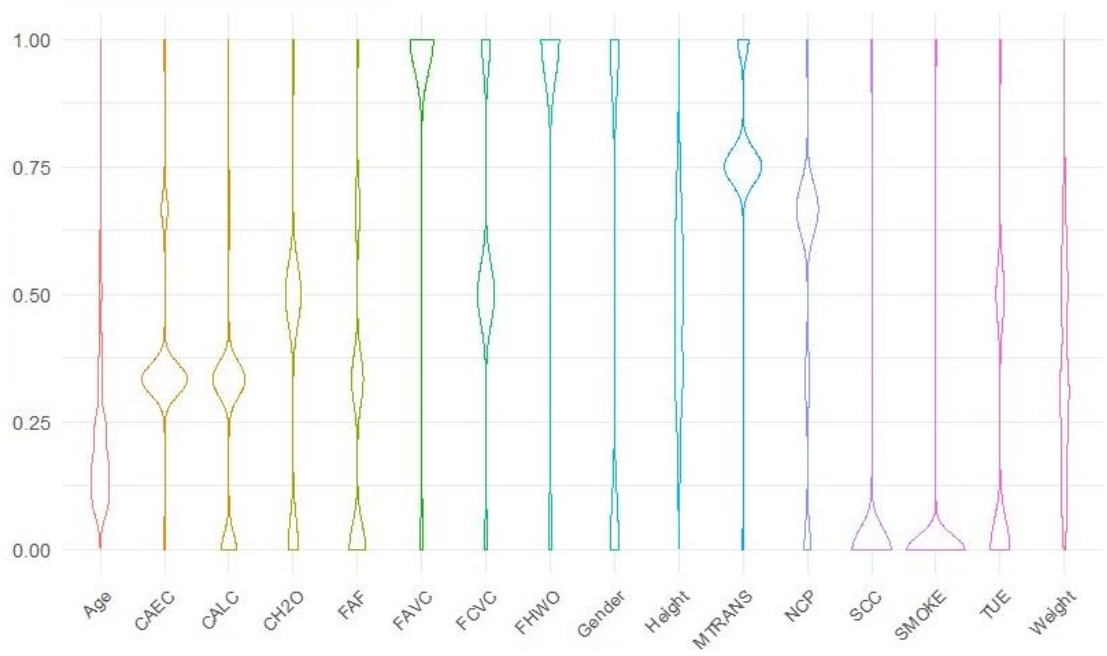


Figure 2. Distribution Normalization of the variables.

Most of the variables hasn't a normal distribution[4] because most variables are binary or discrete which means that we can't appreciate well the normality.

2.3. Random Forest - Importance of Variables

To do a better analysis, I analyze the importance of each variable, using the Random Forest method with Gini index[5].

In the context of a Random Forest model, the importance of a variable is measured by the reduction in the Gini index that the variable provides to the decision tree. Variables that contribute more to the reduction in the Gini index are considered more important.

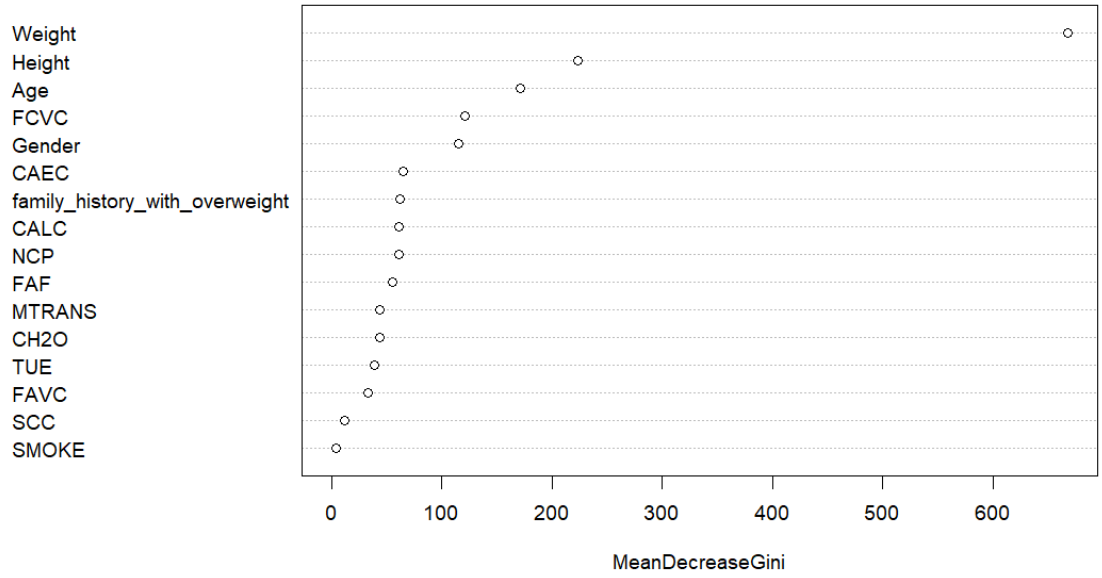


Figure 3. Variable Importance

We can see that Weight is the most influential by far, then Height, Age, FCVC and Gender has more influence than the others. In this case, "physical" variables are more important in the analysis. Although SMOKE and SCC have a very low importance, I am going to use all the variables, because there aren't too much, and I think there won't be efficient problems on training or the performance of the models.

3. Clustering

I ignore the variable 'label' to do clustering, then I will use it in supervised learning. Then I going to do a model-based clustering using EM algorithm for the parameter estimate to maximize verisimilitude. I standardize the data to make this clustering.

To calculate the number of clusters I use the Bayesian Information Criterion (BIC). In this case the clusters are ellipsoidal with 3 variables: Volume, Shape and Orientation of Axis.

3.1. Probabilistic Model-Based Clustering - EM Cluster

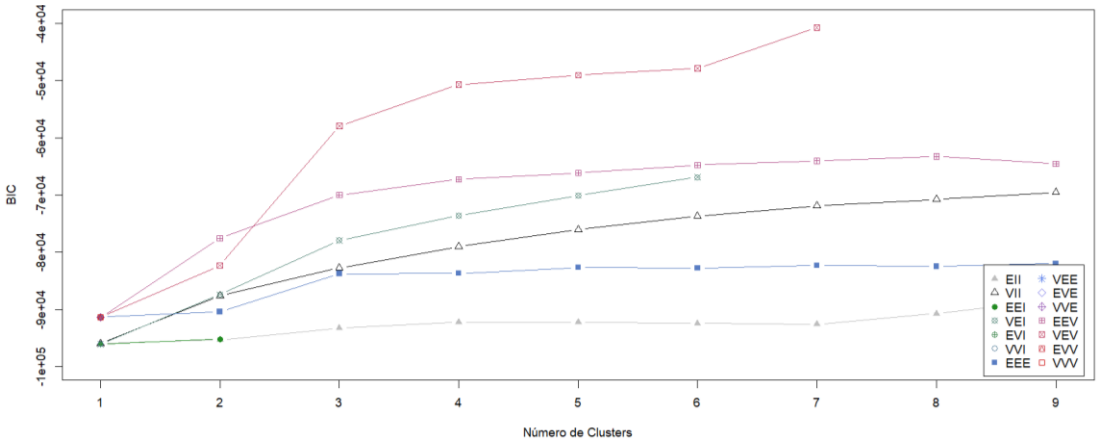


Figure 4. BIC valor criteria

We can see that the maximum BIC[6] is reached in 7 clusters with VEV model, this means:

$$\sum_k = \lambda_k D_k A D_k^T$$

In other words, Distribution: ellipsoidal, Volume: variable, Shape: equal, and Orientation: variable.

Now I are going to analyze internal validation methods[7] to obtain the number of clusters with K-Means algorithm.

1. Sum of square (SSQ) distance to centroids: in this case I want to minimize the values to have better cluster quality. To do this, I make a function from k = 2:10, repeating the k-means 50 times and plotting the means.

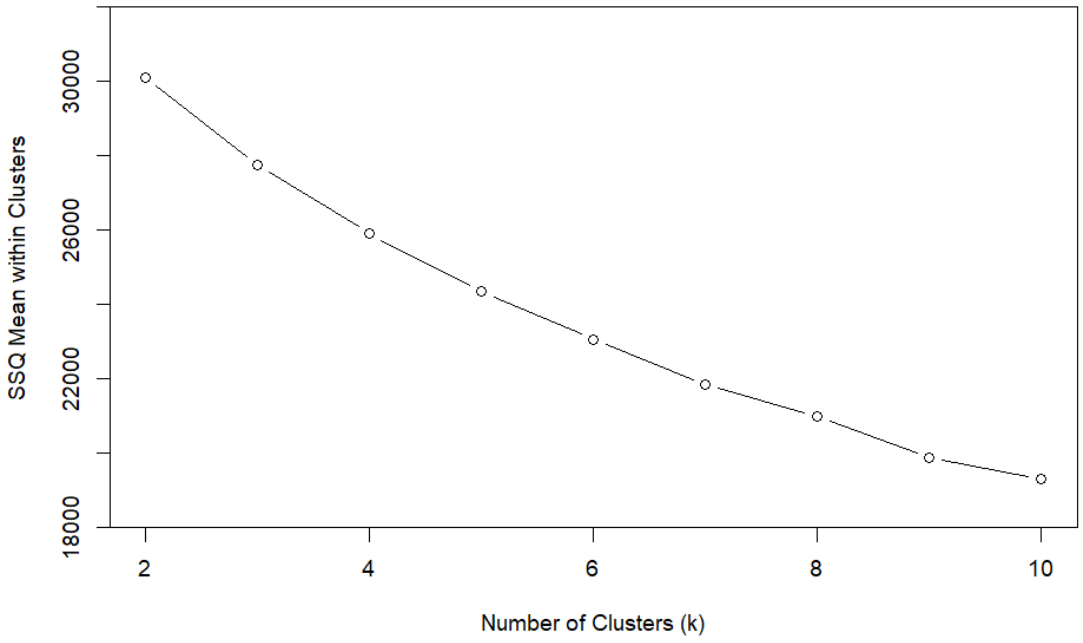


Figure 5. SSQ Plot

I don't appreciate clearly the elbow of the plot to conclude something, but approximately, the elbow could be between 6-8 clusters.

2. Silhouette coefficient (Si): $\boldsymbol{Davg}$ is the average distance of $\boldsymbol{Xi}$ data points within the cluster of $\boldsymbol{Xi}$ and $\boldsymbol{Dmin}$ the minimum of these (average) distances, over the other clusters.

$$S_i = \frac{Dmin_i^{out} - Davg_i^{in}}{\max\{Dmin_i^{out}, Davg_i^{in}\}}.$$

In this case for all clusters from k = 2:10 exists negative values in silhouette so this is not valid, and I won't use it.

Moreover, in the next plot I have the average silhouette width for all k which is very low, so I can conclude that I could have overlapping clusters.

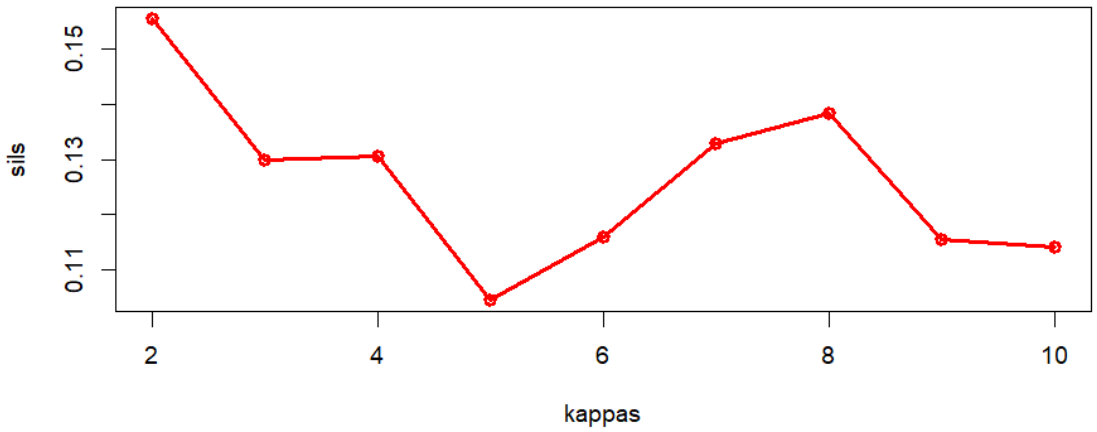


Figure 6. Average Silhouette Width

After analyzing the results of these methods, I have decided to do the k-means clustering with k=7.

3.2. Iterative Distance Based - K-means

Then I do, k-means with k=7. This method is based in EM algorithm and use the Euclidean distance[8] as measure of the distance between points and centroids.

I obtain that *between_SS / total_SS = 33.2%*, this indicates that about one-third of the total variance is explained by the clustering.

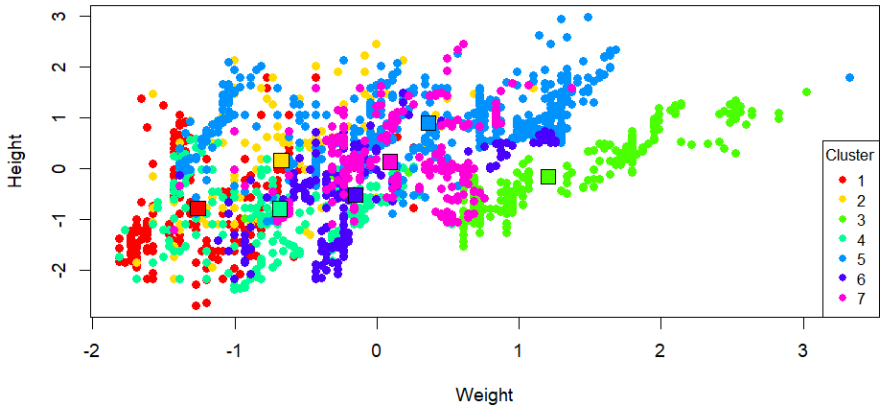


Figure 7. Clustering between Height and Weight

In this dispersion graph between the 2 principal variables based on importance (Weight and Height), so that clusters 1 and 2 are more distinguishable than the other clusters which are more concentrated between themselves. I prove it in the next plot.

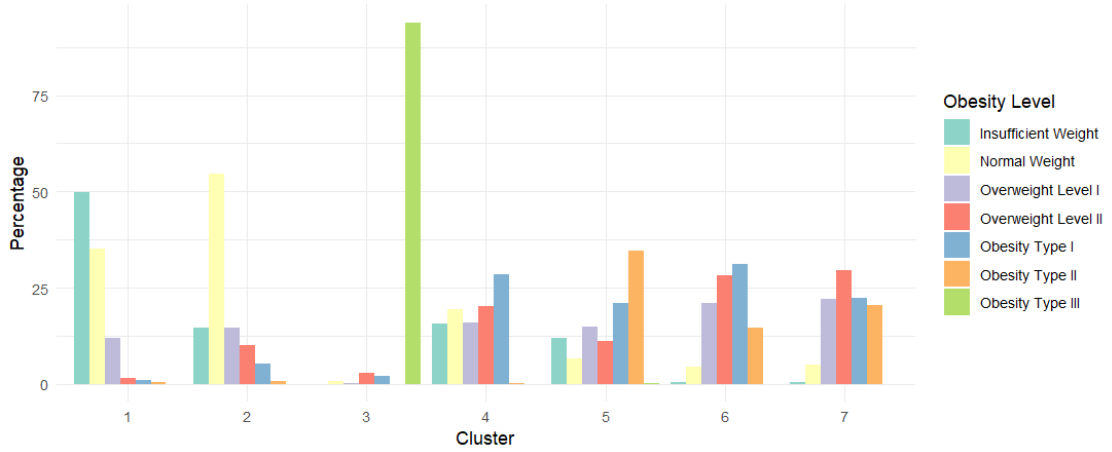


Figure 8. K-means clustering. Bar plot clustering for k=7

Exists a big difference between Obesity Type III and the rest of levels, where one cluster is composed by most of the instances. The rest of the clusters are very homogeneous with the difference of clusters 1 and 2 where we can observe a good classification of Insufficient and Normal Weight.

3.3. Hierarchical Clustering

To do hierarchical agglomerative clustering I use complete linkage which utilize Euclidean distance method for the distance matrix and calculate the distance between the maximum in all the possible pairs. I prune the dendrogram in 3 clusters, because there is a big gap for this number of clusters:

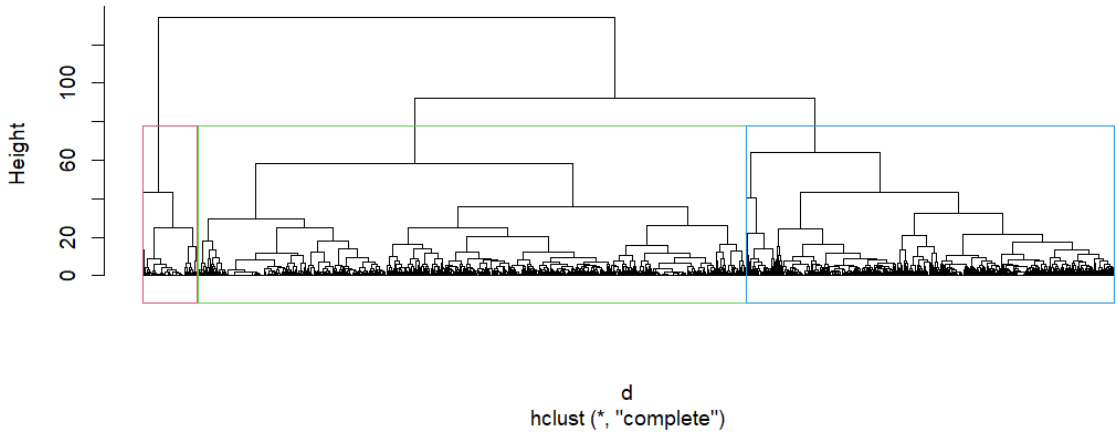


Figure 9. Cluster Dendrogram

The clusters are good defined, which implies good internal similarity of the groups found and diversity between them. The next plot shows the cluster distribution obtained by this unsupervised learning.

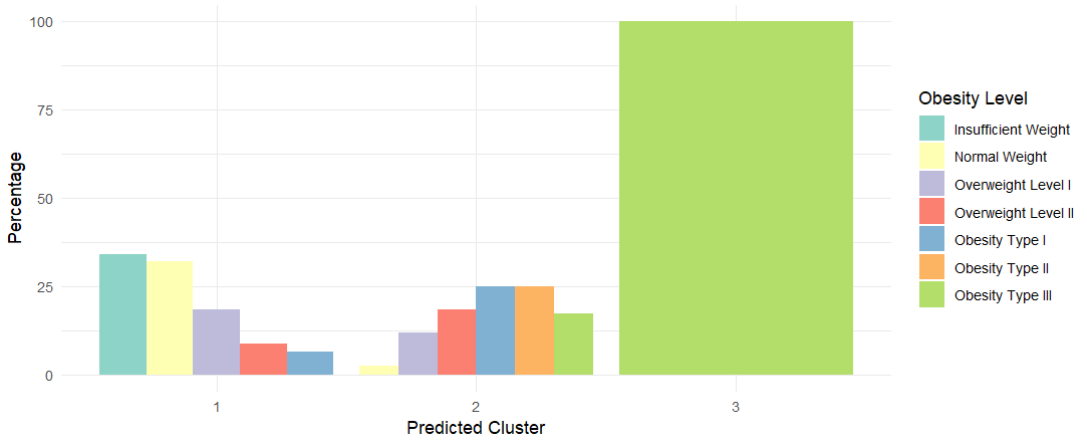


Figure 10. Hierarchical clustering

Exists a big difference between Obesity Type III like in the previous clustering. Is obvious the existence of huge contrast between people from Obesity Type III and other levels. About the rest levels, this clustering difference bad Overweight levels, but Insufficient Weight is good defined in the first cluster and Obesity Type II in the second one.

3.4. Dbscan

The last algorithm used is Dbscan, is based on the concept of instance density in the space. This method involves the division of points in different types of points, classified based on the number of points in a sphere of radius ε and contains at least τ data points.

In this method I have done a function to check which combination of these parameters maximize the quantity of the silhouette. This function compares the mean silhouette using Dbscan with $\varepsilon \in (0.3,0.9)$ and $\tau \in (5,30)$.

The best $\varepsilon = 0.49$ and $\tau = 26$ and the silhouette = 0.02. The problem is that the number of noise points = 1816 of 2111 instances, maybe because the data is sparse in a way that does not cluster well the data with these parameters.

4. Classification

To classify the instances, I have implemented 4 supervised learning algorithms, by the caret package[9], which are:

- K-nearest neighbors (KNN)
- Support Vector Machine (SVM)
- Decision Tree (DT)
- Random Forest (RF)

The dataset is divided into test (20% of instances) and training set (80% of instances). To optimize the hyperparameters of my models and check for overfitting, I implemented a k-fold Cross Validation technique on the training dataset, with k=5. This technique its higher computational cost but it provides a more robust estimate of the training and validation accuracy and the generalized error.

We have to remember which level is each number in the following tables:

0: Insufficient Weight, **1**: Normal Weight, **2**: Overweight Level I, **3**: Overweight Level II, **4**: Obesity Type I, **5**: Obesity Type II and **6**: Obesity Type III.

4.1. K-Nearest Neighbors (KNN)

The values of generalized error and empirical error are very similar, this means that the model has not overfitting. The accuracy is high, the model has good precision predicting new data.

Generalized Error	0.117
Empirical Error	0.143
Accuracy	0.883

		Reference						
Prediction		0	1	2	3	4	5	6
	0	52	7	1	0	0	0	0
	1	2	31	3	2	0	0	0
	2	0	11	47	1	0	0	0
	3	0	6	6	52	2	0	0
	4	0	2	1	2	66	1	0
	5	0	0	0	2	1	58	0
	6	0	0	0	0	1	0	64

Sensibility						
0	**1**	**2**	**3**	**4**	**5**	**6**
0.9630	0.5438	0.8103	0.8793	0.9429	0.9831	1.0000

In the sensibility table (measure of the proportion of positive cases that are good identify) I obtain the lowest value in the Normal Weight class. This means that almost half of Normal weight instances have been predicted bad, specifically most of them with Insufficient Weight and Overweight Level I.

4.2. Support Vector Machine (SVM)

I use kernel radial[10], because the data will be difficult modelling with linear separability. The hyperparameters are *C = 1* and *sigma = 0.04943*. The values of generalized error and empirical error are very similar, so the model has not overfitting. The accuracy is also very good, so the model makes good predictions.

Generalized Error	0.126
Empirical Error	0.129
Accuracy	0.874

		Reference						
		0	1	2	3	4	5	6
Prediction	0	45	3	1	0	0	0	0
	1	9	45	11	4	1	1	1
	2	0	6	41	5	0	0	0
	3	0	2	5	47	1	0	0
	4	0	1	0	2	68	0	0
	5	0	0	0	0	0	58	0
	6	0	0	0	0	0	0	63

Sensibility						
0	1	2	3	4	5	6
0.8333	0.7895	0.7069	0.8103	0.9714	0.9831	0.9844

The values in the sensibility table are relatively high, except Normal Weight and Overweight Level I in which exists a confusion between their predictions.

4.3. Decision Tree (DT)

The hyperparameter used is *cp*[11] *(complexity parameter) = 0.07411*, which is a threshold that determines the minimum amount of improvement in model fit necessary for an additional split to be performed on the tree.

The generalized error and empirical error are higher than previous algorithms, although I still approve the method because there isn't a significant difference between them. The accuracy is 0.55 so the model is not a good predictor.

Generalized Error	0.452
Empirical Error	0.414
Accuracy	0.548

		Reference						
		0	1	2	3	4	5	6
Prediction	0	52	23	3	1	0	0	0
	1	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0
	3	2	34	55	56	50	1	0
	4	0	0	0	0	0	0	0
	5	0	0	0	1	20	58	0
	6	0	0	0	0	0	0	64

Sensibility						
0	1	2	3	4	5	6
0.9630	0.0000	0.0000	0.9655	0.0000	0.9831	1.0000

The sensibility in clusters 1, 2 and 4 are the values are bad classified. Most of these instances are predicted in cluster 3.

4.4. Random Forest (RF)

Now I'm going to calculate RF algorithm since it is composed of many decision trees. The hyperparameters that maximize the accuracy is $mtry^{12} = 9$ (number of features that are randomly selected as candidates in each split of a tree) and number of trees in our case 500.

The generalized error and empirical error are very similar too, so the model hasn't overfitting. The accuracy is the highest among the different methods used, so the model is a very good predictor.

Generalized Error	0.031
Empirical Error	0.039
Accuracy	0.969

		Reference						
Prediction		0	1	2	3	4	5	6
	0	52	2	0	0	0	0	0
	1	2	52	3	1	0	0	0
	2	0	1	55	0	0	0	0
	3	0	2	0	57	1	0	0
	4	0	0	0	0	69	1	0
	5	0	0	0	0	0	58	0
	6	0	0	0	0	0	0	64

Sensibility						
0	**1**	**2**	**3**	**4**	**5**	**6**
0.9630	0.9123	0.9483	0.9828	0.9857	0.9831	1.0000

The values in the sensibility table are high so RF is a good model. However, in other algorithms the sensibility of Normal and Overweight Level I is less than the rest.

5. Conclusion

To conclude, about clustering exists a big difference between Obesity Level Type III with the other ones. Moreover, it is appreciated a little contrast with Insufficient and Normal weights since the Overweight weights are very similar.

In classification, we achieve the best accuracy in Random Forest (96,9%). This could be because RF is a multiple Decision Tree evaluate with different data samples, this reduces the variance and mitigate the overfitting.

The sensibility in RF has lower values in Normal and Overweight Level I. In the other methods happen the same pattern. This could be because exist a little imbalanced in data, and the Overweight class has more instances than Normal one. Leading the model to misclassify the data into the major class, in this case Overweight.

We have seen that Machine Learning methods to classify new instances are more effective than casual assignment. In general, these algorithms have high accuracy, except Decision Tree, being a good way to assign people based in some healthy and physical features.

These techniques can be used in health centers or hospitals to check the patient's level of obesity and act accordingly. Early actions could prevent future diseases related to this problem that today cause deaths and worse living conditions. In my opinion, the focus should be on good healthy habits, leaving bad eating habits and lack of exercise.

6.Bibliography and references

1.Dataset: https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition

2.IMC and obesity levels: https://www.cdc.gov/healthyweight/spanish/assessing/bmi/adult_bmi/index.html

3.SMOTE filter: https://pubmed.ncbi.nlm.nih.gov/31467953/

4.Distribution of variables: https://es.quora.com/He-visto-que-una-distribuci%C3%B3n-de-probabilidad-generalmente-se-da-entre-0-y-1-es-esto-siempre-de-ese-modo-Por-qu%C3%A9-entre-0-y-1

5.Gini coefficient and Importance Variables: https://en.wikipedia.org/wiki/Gini_coefficient

6.Mclust and BIC information: https://cran.r-project.org/web/packages/mclust/mclust.pdf

7.Lesson 2024-05-09: https://elearning.unimib.it/course/view.php?id=51218

8.K-means explain in: https://www.datanovia.com/en/lessons/k-means-clustering-in-r-algorith-and-practical-examples/#k-means-algorithm

9.Caret package: https://www.machinelearningplus.com/machine-learning/caret-package/ and https://rebeccabarter.com/blog/2017-11-17-caret_tutorial

10.SVM in Lesson 2024-05-24: https://elearning.unimib.it/course/view.php?id=51218

11.Hyperparamter in Decision Tree: https://cran.r-project.org/web/packages/rpart/rpart.pdf

12.Hyperparameter in Random Forest: https://cran.r-project.org/web/packages/randomForest/index.html

---

Fig.1.a. and Fig.1.b.: https://pubmed.ncbi.nlm.nih.gov/31467953/

Figure 2 to 10: obtain in R by me.