UNCOVERING LIFESTYLES PATTERNS: A MACHINE LEARNING APPROACH

Lumia Giulia, Bongiovanni Sabrina Anna, Cortina Katia,
Tranchina Salvatore, Dolce Rosolino

ABSTRACT

In order to reveal the clustering structure underlying the data, several clustering methods were implemented and compared. The choice fell on the model-based approach, leading to three clusters. Moreover, for the purpose of assigning new persons to the clusters previously identified, a predictor was developed.

A comparison was carried out among the random forest and two main versions of the gradient boosting algorithms: the most accurate predictor was the basic gradient boosting algorithm, capable of correctly predicting more than 91% of the test set observations.

OUTLINE

1. INTRODUCTION

This report is aimed to show the results of a clustering and predictive analysis carried out in order to reveal the possible presence of groups of people sharing some features with respect to their socio-demographical status, lifestyle and interests. According to this general purpose a survey was conducted. The collected data refers to 512 individuals, whereas the features are:

- **Sex;**
- **Age**;
- **Weight;**
- **Height**;
- **Marital status;**
- **Sexual orientation;**
- **Occupational status;**
- **Own, Father and mother educational level;**
- **Desiring/actually having children;**

SOCIO-DEMOGRAPHICAL FEATURES

- **Physical activity status and frequency**;
- **Free time amount;**
- **Social networks usage;**
- **Diet;**
- **Meat, fish, legumes, vegetables and fruits consumption;**
- **Coffee, junk food and alcohol consumption;**
- **Smoking status**;
- **Haircut;**
- **Transportation habits;**
- **Average sleeping hours;**

LIFESTYLE

- **Mental, physical, social, financial and career wellness;**
- **Prevalent mood in the last month;**
- **Interest with respect to movies, tv-series, videogames, romances, chess, gambling, essay and theatre;**
- **Interest with respect to volunteering activities;**
- **Agreement with respect to the following ideas: polygamy, Greta Thunberg, Andrew Tate;**

GENERAL WELLNESS & INTERESTS

2. DATA PREPROCESSING

The first step of the analysis aims to make data suitable for the subsequent clustering and classification steps. Moreover, some variables were recoded and removed.
In fact, the survey questionnaire was not specifically created for this analysis, thus, according to the study aim, some features are not meaningful. In particular, the excluded variables are:

- Sex;
- Haircut;
- Interest with respect to volunteering activities;
- Agreement with respect to the following ideas: polygamy, Greta Thunberg, Andrew Tate.

In order to concentrate the information, weight and height were summarized, leading to a unique BMI (Body Mass Index) variable. This quantity refers to a well-known index, that allow to assess, as instance, whether a person is underweight or obese.
The weight value considered alone, in fact, may be misleading. As instance, given a weight of 60 kg a 1.75 meters woman can be considered with normal weight while a 1.50 meters woman would be classified as overweight.

TABLE 2.1: BMI CONDITIONS

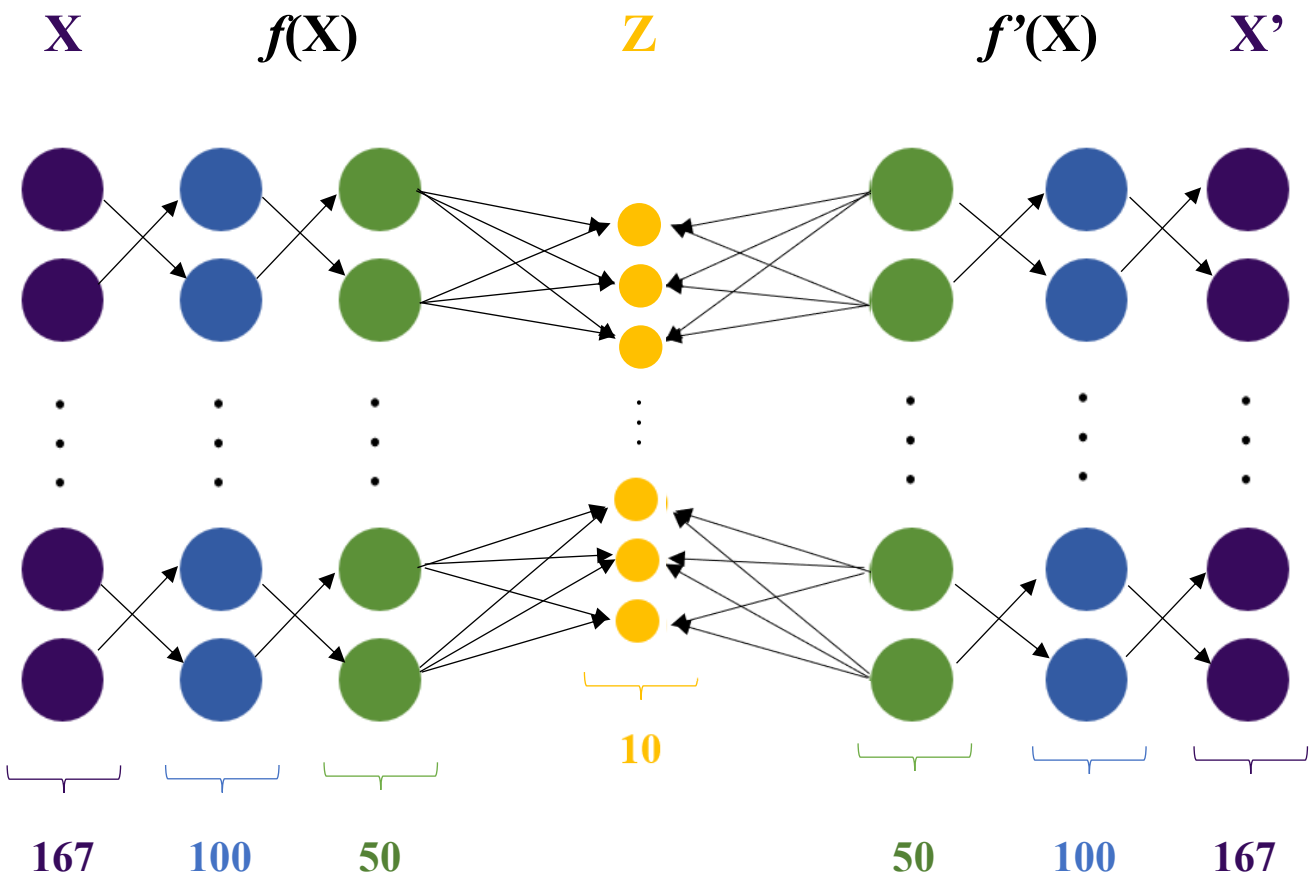
CLASS	BMI
Underweight	<18,5
Normal weight	18,5 ≤ BMI < 25
Overweight	25 ≤ BMI < 30
Obesity grade I	30 ≤ BMI < 35
Obesity grade II	35 ≤ BMI < 40
Obesity grade II	≥ 40

On the basis of the nature of the methods and algorithms that will be used in the following analysis steps, some variables were pre-processed. In particular:

- STANDARDIZATION
  Numerical variables, i.e. age and BMI index, were standardized in order to overcome the effect of the different scales;
- ONE-HOT ENCODING
  Since most of the variables have a categorical nature, an encoding was necessary. Given a variable taking *n* different values, *n* different binary columns were created. The resulting dataset is made up of 167 columns.
- DIMENSION REDUCTION
  The drawback of the one-hot encoding method concerns the fact the original dataset dimensions increase significantly. There are, however, some dimensions reduction methods, able to concentrate the information given by the original variables, using lower dimensions. A comparison between PCA (Principal Components Analysis) and Autoencoders was made. Since the variables are of mixed typology the chosen method was the second one: autoencoders that are based on a neural network. Given a dataset having *n* variables, the resulting neural network will have *n* neurons both in the first and

in the last layer and a number of hidden layers and neurons in them chosen on the base of a grid search approach. The *n* features are efficiently represented via an encoder function (and a decoder one) learnt by minimizing a loss function between the original information and the recoded one. In this way it was possible to represent this information in a lower dimension, given by the number of neurons in the central hidden layer.
As a result, the following neural network was obtained (PLOT 2.1), with a MSE between the original variables and the ones obtained by the means of the recoding function of about 0.097.

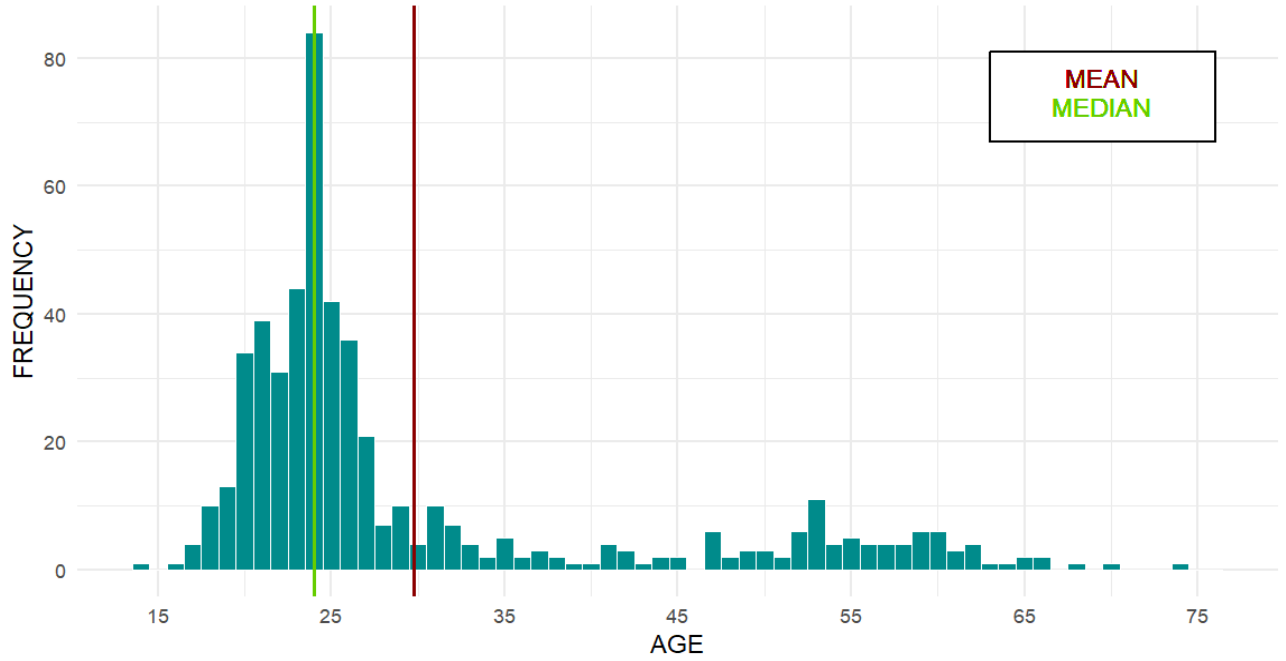
PLOT 2.1: UNDERCOMPLETE AUTOENCODER WITH FIVE FULLY CONNECTED HIDDEN LAYERS



3. EXPLORATORY ANALYSIS

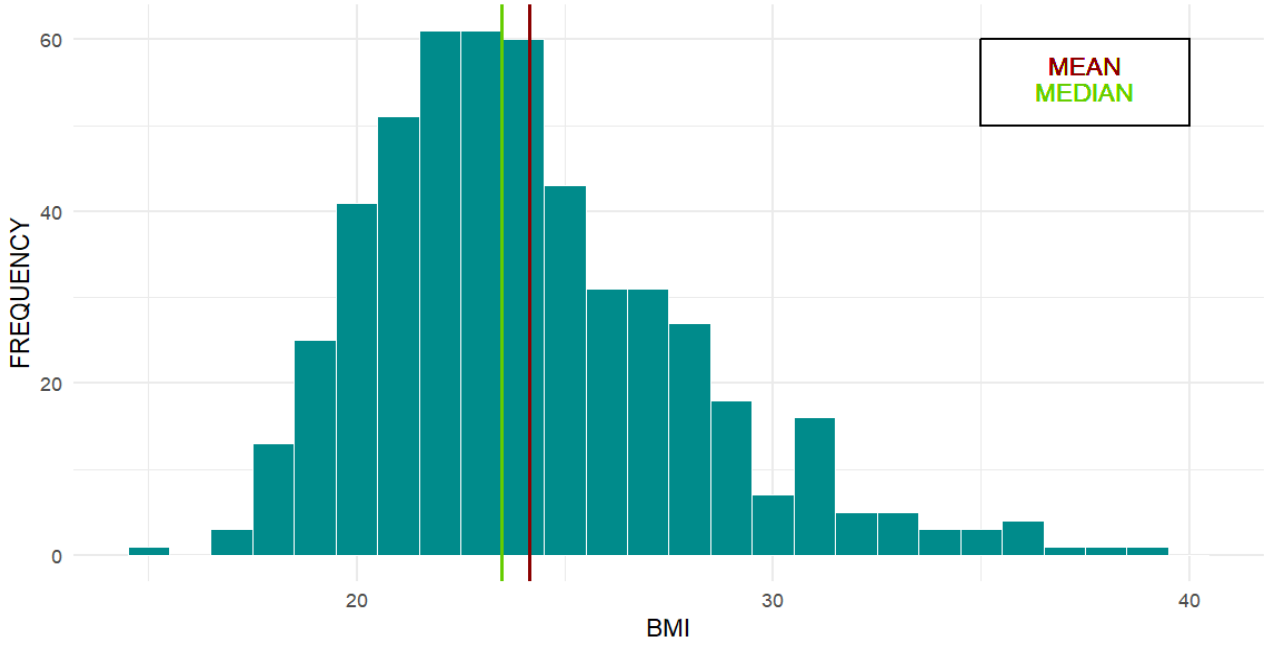
Before implementing the clustering and classification phases an overall description of the participants was considered necessary. Having this purpose, some marginal and conditional distributions were graphically represented.

PLOT 3.1: AGE DISTRIBUTION

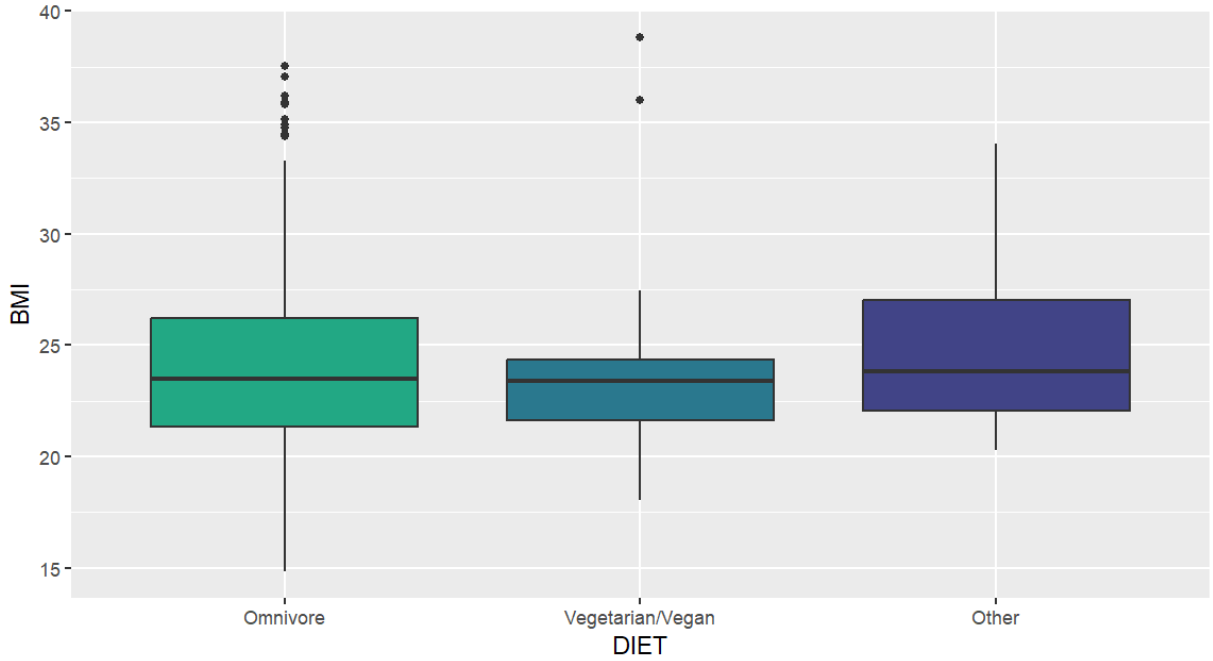


PLOT 3.1 shows the empirical distribution of the participants age while PLOT 3.2 displays their BMI index. Survey participants range in age from 14 to 74 years old.
On average there are young, normal weight individuals, with an age of about 30 years and a BMI of about 24 points. Both distributions are positively skewed: half of the participants has at most 24 years and at least a 23.5 BMI points.

PLOT 3.2: BMI DISTRIBUTION

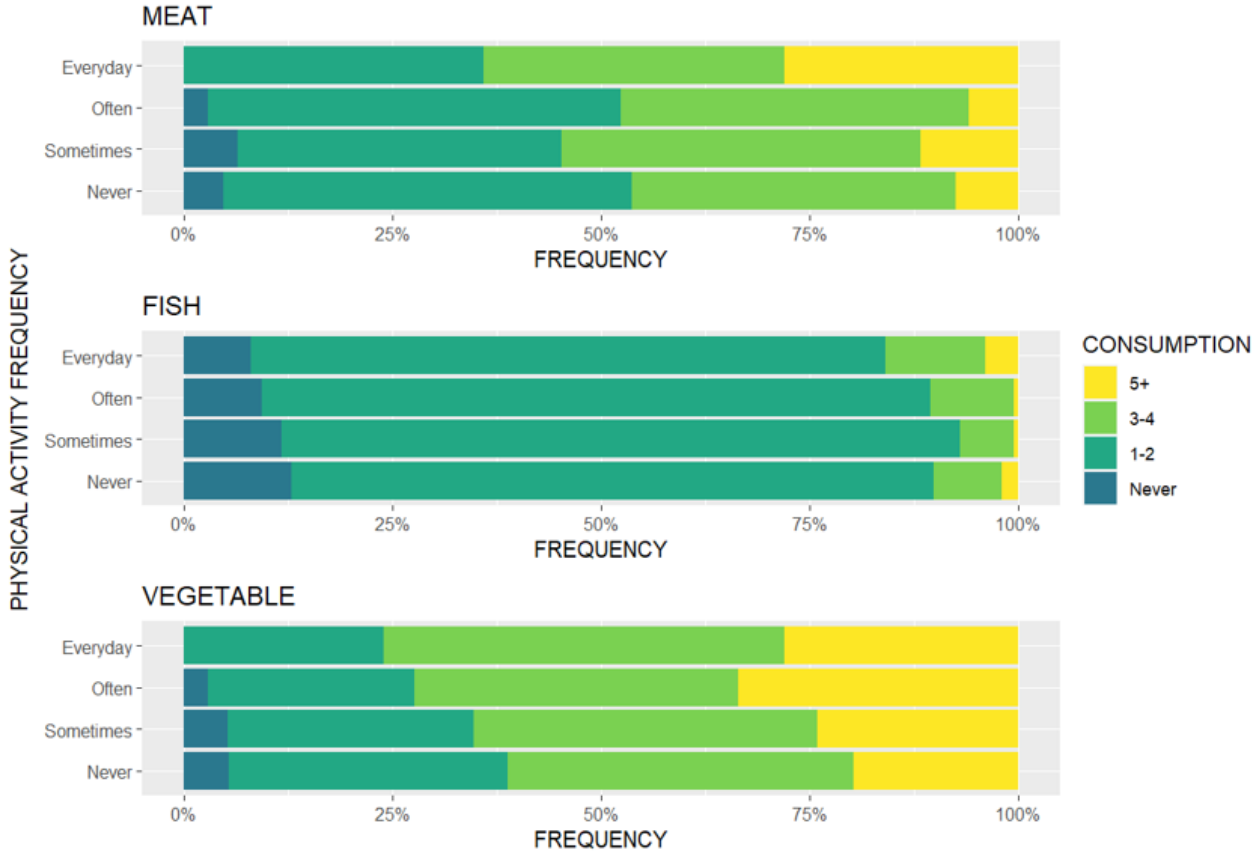


PLOT 3.3: CONDITIONAL BMI DISTRIBUTION GIVEN THE DIET TYPE



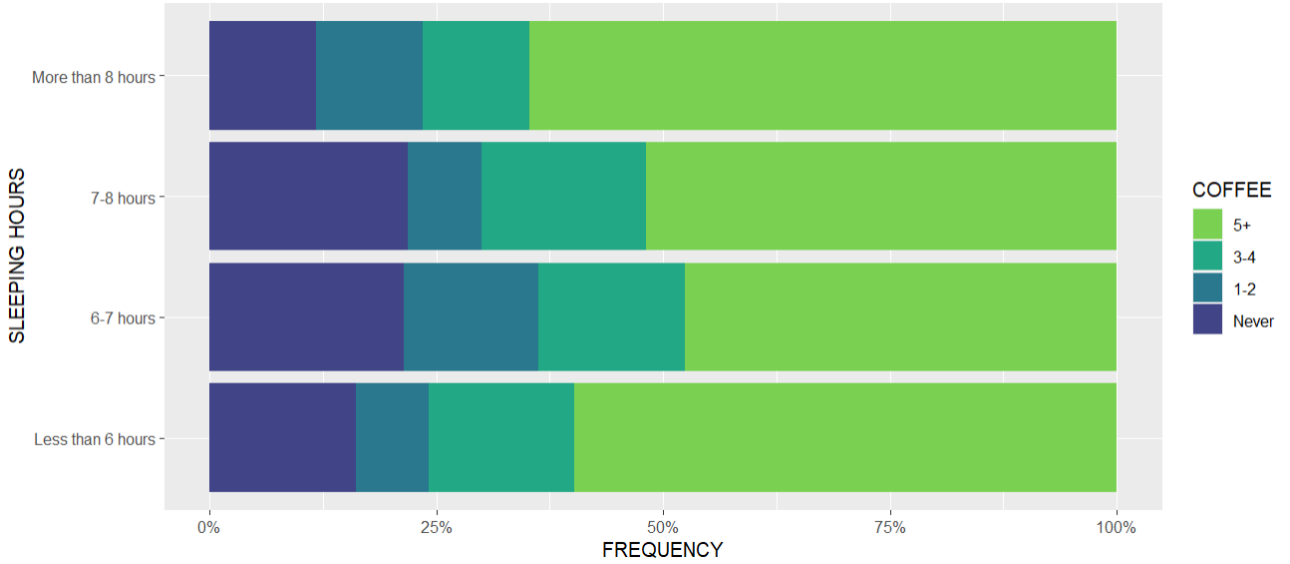
The median BMI is found out to be about 23 for all diet types (PLOT 3.3). However, vegetarians and vegans are almost all normal weight, since more than 3/4 of them has a BMI value below 25, with the only exception given by two individuals with a BMI between 35 and 40. People following an omnivore diet show greater variability in terms of BMI with respect to the other diets. In fact, there are several underweight and obese individuals. On the other hand, people following another dietary regimen, i.e. flexitarians, fishetarians and fruitarians, are mostly normal weight and overweight.

PLOT 3.4: CONDITIONAL MEAT, FISH AND VEGETABLES CONSUMPTION DISTRIBUTIONS GIVEN P.A.



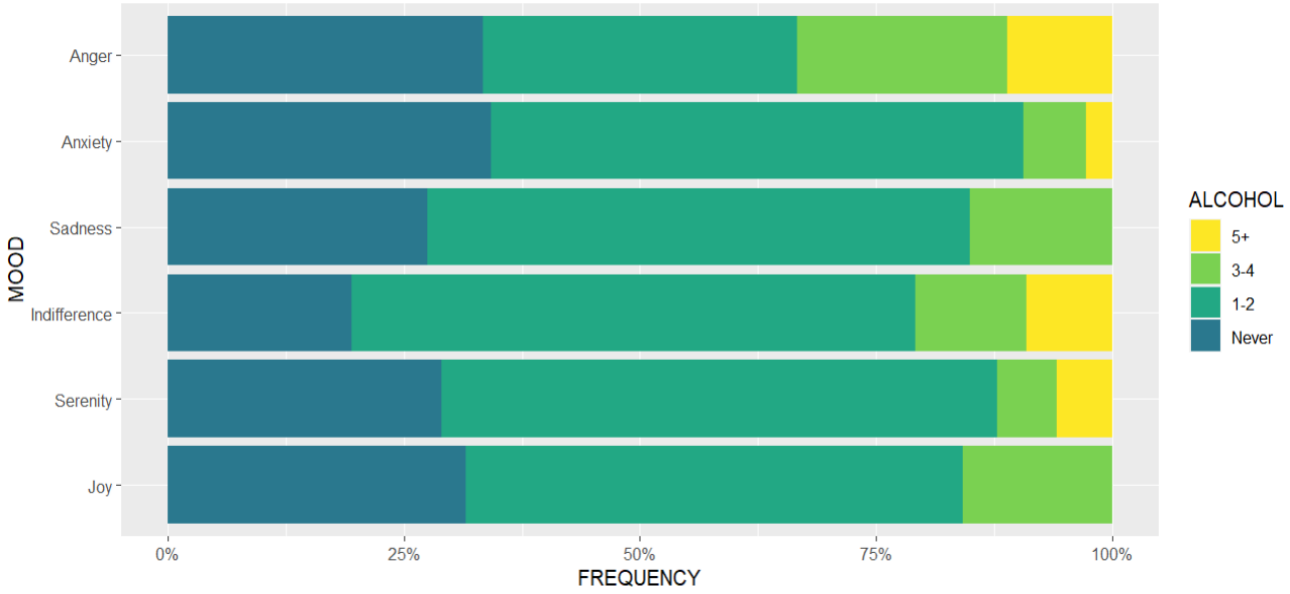
PLOT 3.4 shows that more than half of those who train daily, during the week consume meat at least 3 days. The vegetable consumption increases proportionally with the frequency of physical activity performed; specifically, anyone who trains everyday eats vegetables at least 1-2 days per week. Finally, the most frequent fish consumption is 1-2 days per week, regardless of the frequency of physical activity. However, the percentage of those who totally exclude fish from their diet increases as the frequency of activity decreases.

PLOT 3.5: CONDITIONAL COFFEE CONSUMPTION DISTRIBUTION GIVEN THE SLEEPING HOURS



PLOT 3.5 shows how the coffee consumption changes according to the hours of sleep.
It is possible to notice that the higher consumption (5 or more times weekly) is prevalent among those who sleep on average more than 8 hours or less than 6 hours.

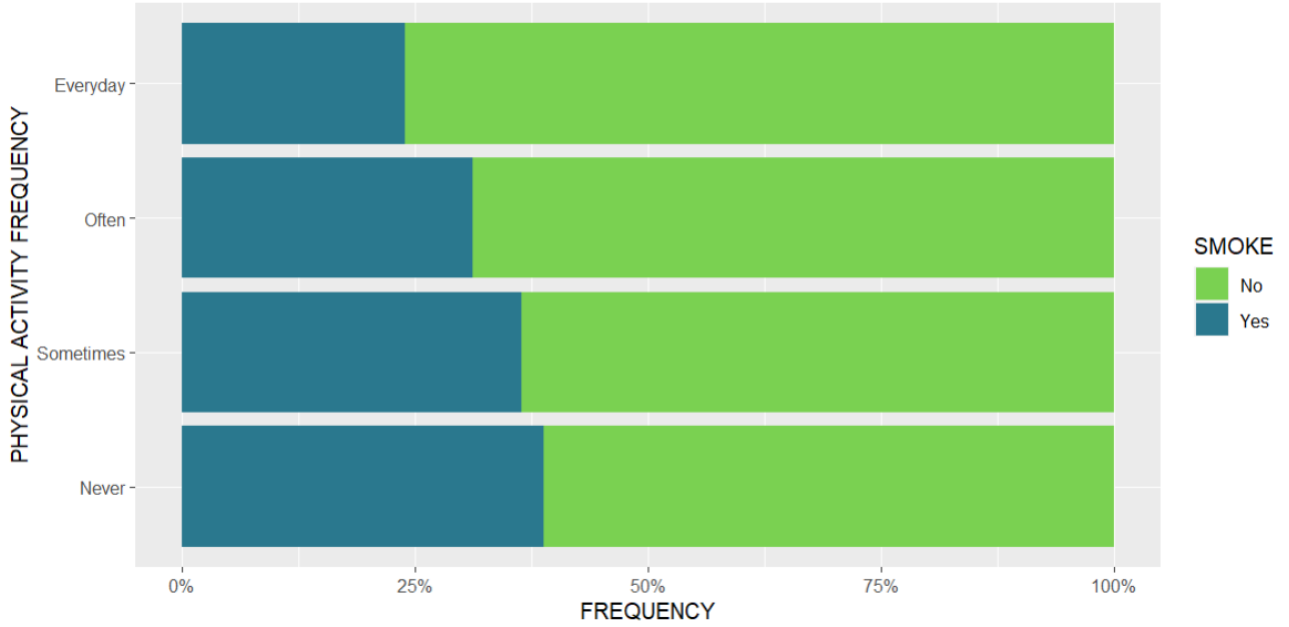
PLOT 3.6: CONDITIONAL ALCOHOL CONSUMPTION DISTRIBUTION GIVEN THE MOOD



From PLOT 3.6 emerges that the most predominant category of alcohol consumption is between 1 and 2 days a week, regardless of the prevalent mood in the past month. It is also

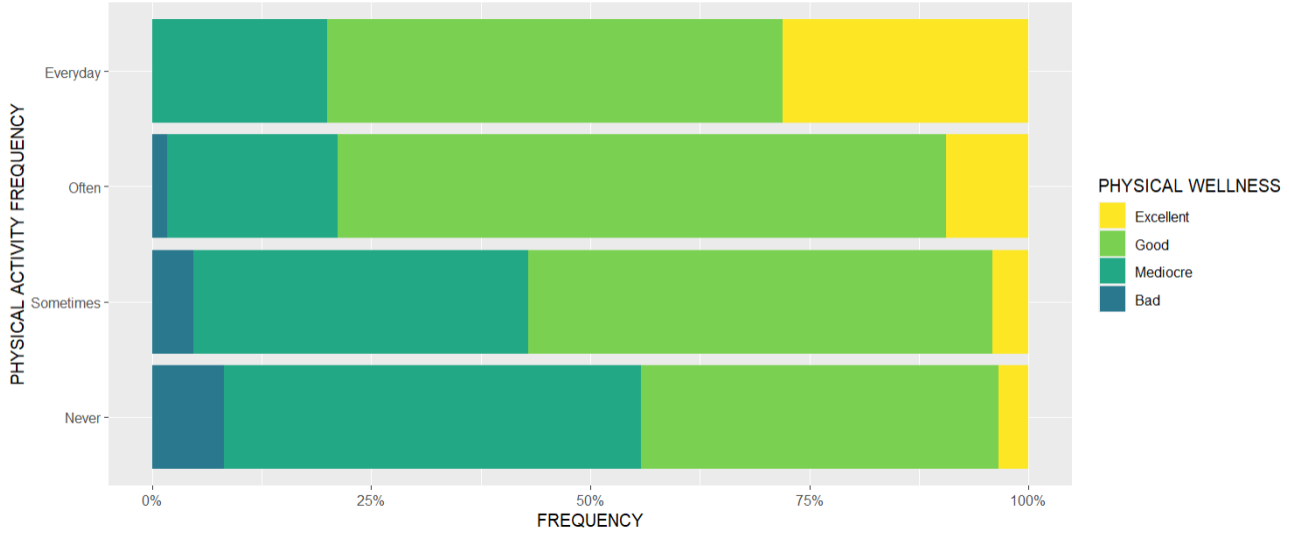
possible to notice that about 30% of the participants with a feeling of anger heavily consume alcohol, 3 or more times per weeks.

PLOT 3.7: CONDITIONAL SMOKE DISTRIBUTION GIVEN PHYSICAL ACTIVITY FREQUENCY



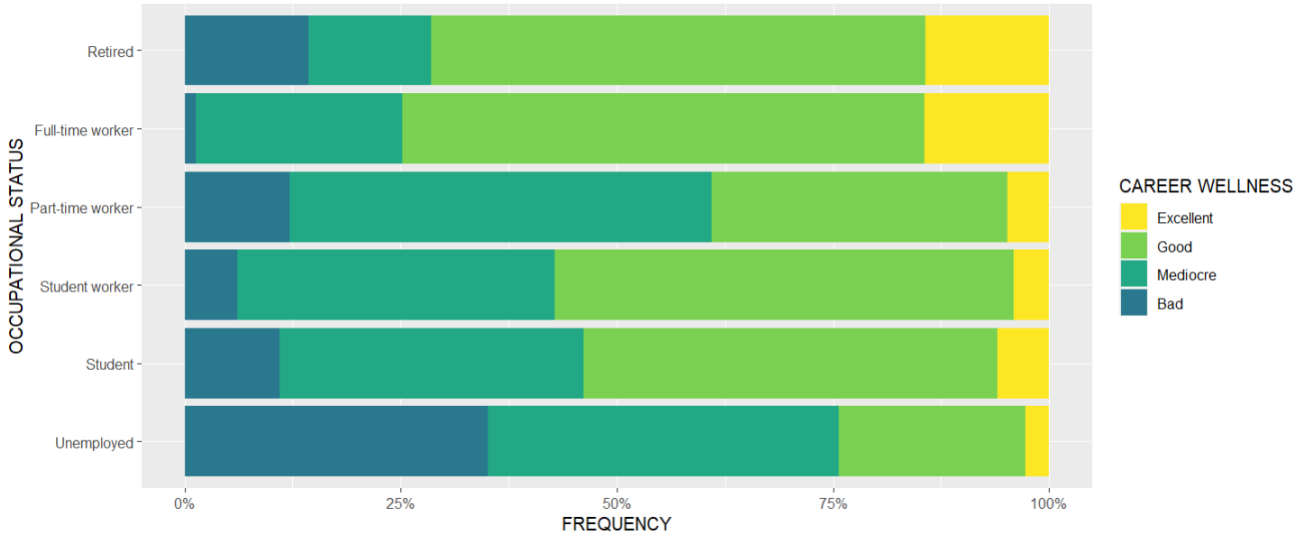
PLOT 3.7 shows how the smoking habit change when the physical activity one changes too. It appears that as the physical activity frequency increases, the proportion of smokers decreases. In particular, just a quarter of those who exercise daily reported being smokers.

PLOT 3.8: CONDITIONAL P.A. FREQUENCY DISTRIBUTION GIVEN THE PHYSICAL WELLNESS



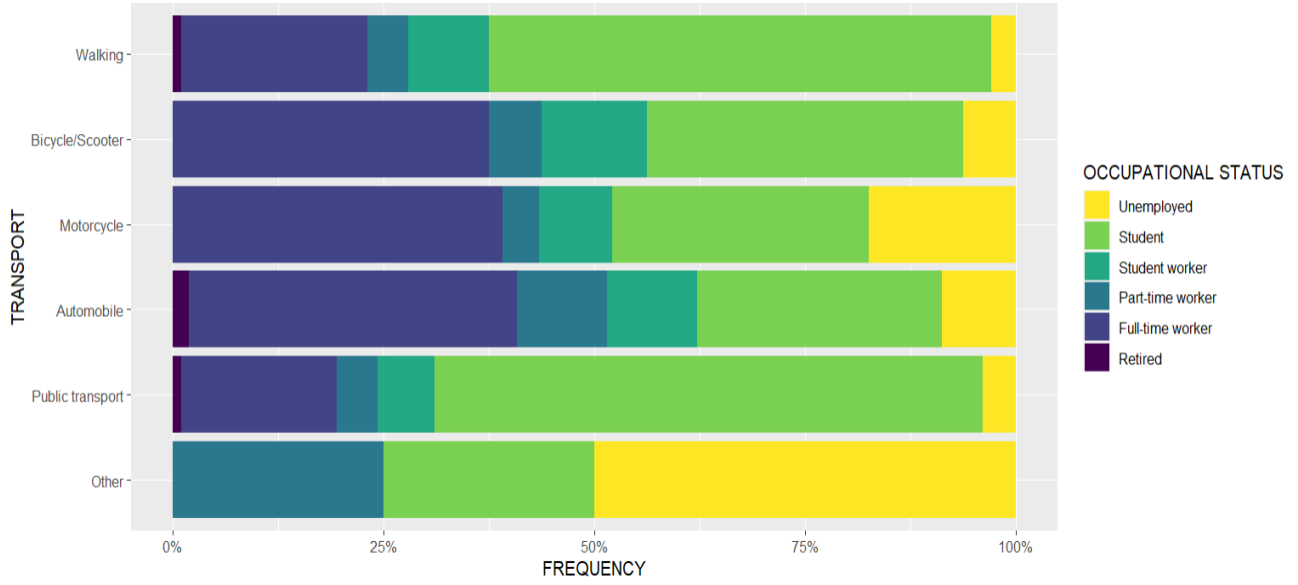
PLOT 3.8 shows how the physical wellness changes according to the weekly physical activity frequency. It is possible to notice that the physical wellness improves as the frequency increases. In fact, over 50% of individuals who do not engage in physical activity report a bad or mediocre physical wellness whereas all the participants who exercise on a daily basis report at least a mediocre wellness level.

PLOT 3.9: CONDITIONAL CAREER WELLNESS DISTRIBUTION GIVEN THE OCCUPATIONAL STATUS



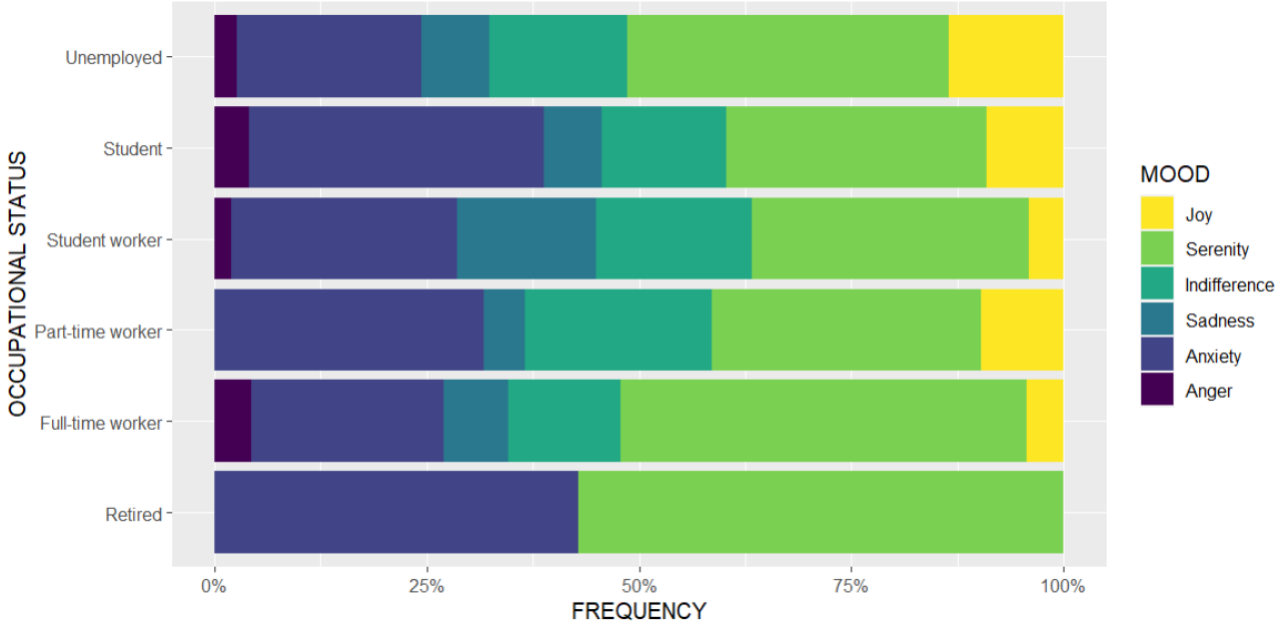
PLOT 3.9 shows that the participants having the best career wellness are full-time workers or retired. In fact, among them it is possible to find the greater proportion of the highest wellness levels (good and excellent). Conversely, more than 75% of the unemployed have a bad or mediocre career wellness, followed by part-time workers and students.

PLOT 3.10: CONDITIONAL TRANSPORTS DISTRIBUTION GIVEN THE OCCUPATIONAL STATUS



PLOT 3.10 highlights the difference among transportation preferences according to the occupational status. Among those who prefer either walking or taking public transports the great majority are students. On the other hand, those who use the automobile, the motorcycle or the bicycle/scooter as their main means of transport are mostly full-time workers.

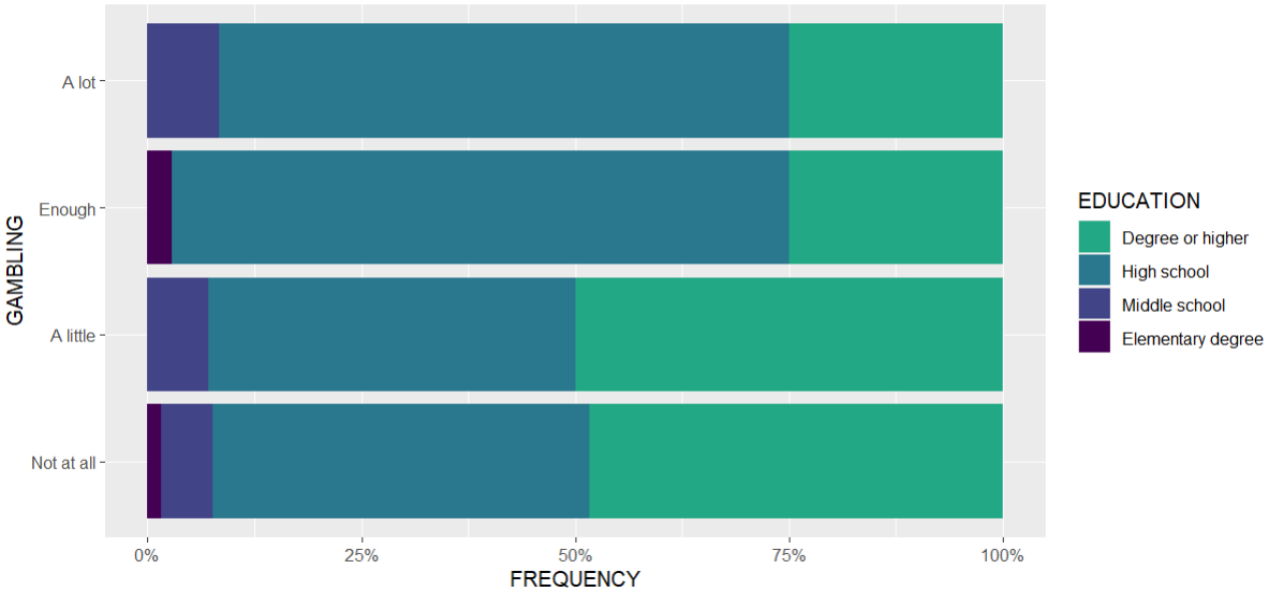
PLOT 3.11: CONDITIONAL MOOD DISTRIBUTION GIVEN THE OCCUPATIONAL STATUS



By examining the participants mood besides the employment status (PLOT 3.11), it emerges that more than the 50% of full-time workers and retired manifest positive feelings, like joy or serenity. In contrast, students, student-workers and part-time workers predominantly manifest negative feelings such as anxiety, sadness and indifference. However, the overall number of retired people is almost irrelevant since it is equal to 7. It worths mention that a relevant proportion of students felt during the last month a sense of anxiety.

From PLOT 3.12 emerges that those who are a little or not at all interested in gambling mostly hold a degree or a higher title. In contrast, the majority of those moderately or very interested in gambling have at most a high school diploma. It needs to be pointed out that there are only few participants with a middle or elementary educational level.

PLOT 3.12: CONDITIONAL GAMBLING DISTRIBUTION GIVEN THE EDUCATIONAL LEVEL



4. CLUSTERING ANALYSIS

One of the key objectives of the analysis was to explore the presence of natural groups within the questionnaire participants, i.e. groups of people who share common or similar values of the collected features, allowing a deeper understanding of their overall diversity or similarity. Three clustering methods were implemented and compared:

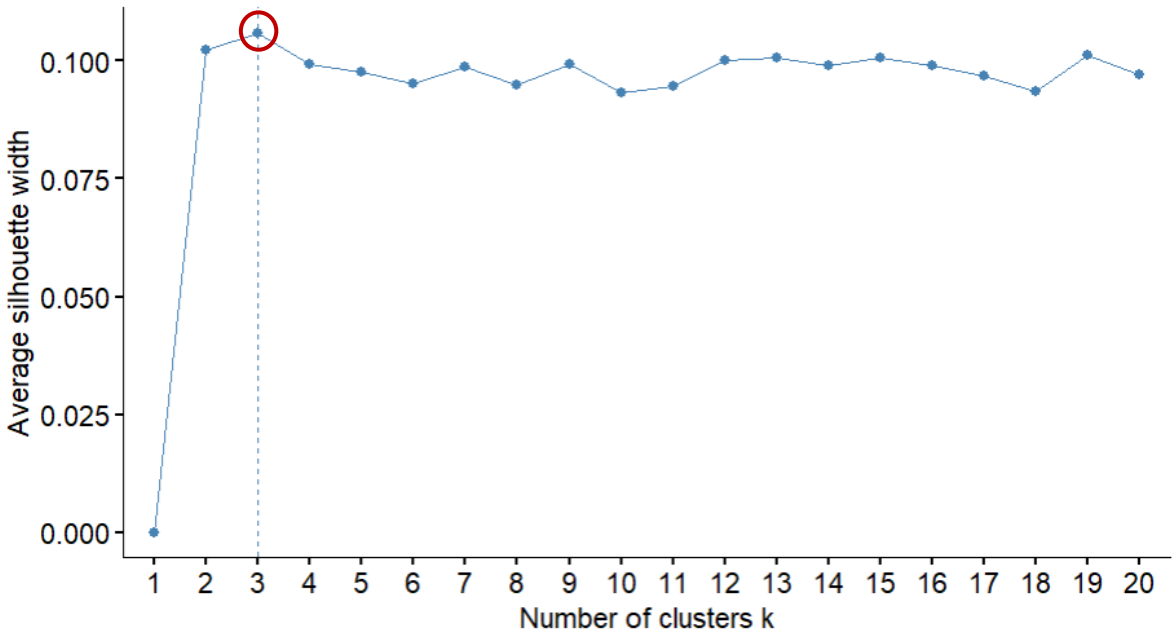
- K-means clustering;
- Agglomerative hierarchical clustering;
- Model-based clustering.

Since the original dataset has a high dimension due to the one-hot encoding previously performed, there may be some computational problems. As an example, the model-based approach is well known to have this kind of limitation that can be, however, usually addressed by the mean of a dimensionality reduction approach, such as the aforementioned PCA and encoders. As expected, the model-based algorithm required the use of the reduced dataset. In order to assure comparability, the clustering analysis was conducted on both the original and the reduced datasets obtained by the mean of autoencoders.

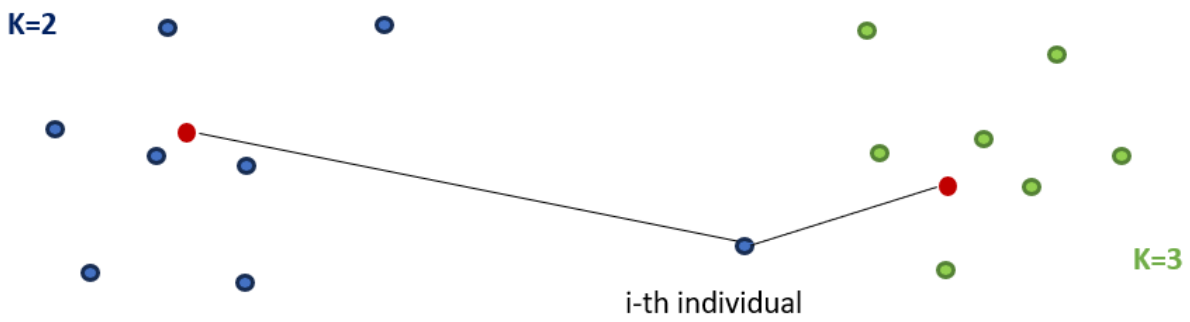
4.1 K-MEANS CLUSTERING

K-means clustering represents a frequently employed non-hierarchical clustering algorithm for the partitioning of observations into k groups (or clusters). The main objective of this algorithm is to obtain groups of individuals that present maximum heterogeneity among them and maximum homogeneity within them.
To achieve this goal, a desired number of clusters k needs to be initially set, by the mean of a grid search algorithm that works by maximizing a specific measure:
the silhouette measure, is a cluster validation approach that quantifies how well an observation is clustered, estimated by averaging the distance between clusters.
The silhouette measure increases as the clustering structure improves.

PLOT 4.1.1: SILHOUETTE METRIC ACCORDING TO THE NUMBER OF CLUSTERS *k*



PLOT 4.1.1 shows how the silhouette metric, computed on the base of the Gower distance (a measure able to handle mixed data types with outliers), changes according to the number of clusters *k*: the optimal value is equal to **three**.
The k-means algorithm works by randomly assign the n observations to three clusters. Then, *k* centroids are computed, besides a distance measure between each observation and each centroid. At each iteration a reassignment may be done according to the shortest distance previously computed and all the *k* centroids are updated. As instance, if the i-th person is assigned to the second cluster but the distance between it and the second centroid is greater than with respect to the third centroid, then the assignment will be changed.
An example of this situation is shown below:



The algorithm stops when for each individual the assigned cluster is the one with the nearer centroid. One of the main advantages of the K-means approach is its flexibility, due to its non-hierarchical nature. The main disadvantage refers to the high dependence on the initial cluster assignment and on the chosen *k* hyperparameter value.

4.2 HIERARCHICAL CLUSTERING

An alternative approach to the non-hierarchical k-means clustering concerns the hierarchical clustering methods. The main difference between the hierarchical and the non-hierarchical approaches regards the way in which each individual is assigned to a cluster:
in the hierarchical framework the assignment is permanent and the result is given by a hierarchy of nested clusters that can be easily visualised by the mean of a tree structure, called dendrogram. Clusters are obtained with successive divisions or fusions:

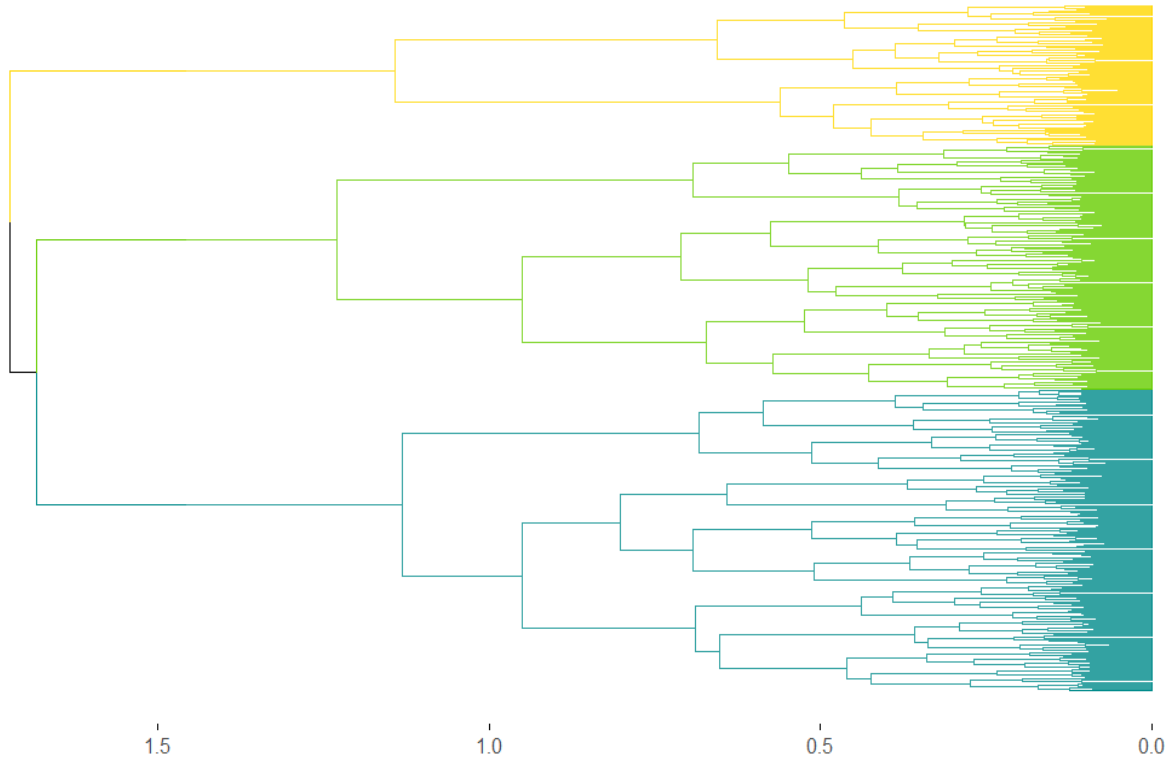
- DIVISIVE HIERARCHICAL CLUSTERING works in a top-down way.
  The algorithm starts by considering a unique cluster with all the individuals inside.
  At each iteration, the current cluster is split into two clusters that are considered most heterogeneous, on the base of the within-between deviance decomposition.
  The process is iterated until all observations make a cluster on their own.
- AGGLOMERATIVE HIERARCHIAL CLUSTERING works in a bottom-up way.
  Initially, the algorithm considers a cluster for each of the 512 individuals. At each iteration, most similar clusters are merged by augmenting the tolerated dissimilarity.
  The algorithm stops when all the individuals belong to the same cluster.
  Since clusters number and structure change iteratively, it is necessary to compute the chosen dissimilarity measure at each iteration, using different approaches:
  - **Single linkage**, works by considering as distance measure between two clusters the smallest one among the pairwise distances previously computed;
  - **Complete linkage**, works by considering as distance measure between two clusters the biggest one among the pairwise distances previously computed;
  - **Average linkage**, given two clusters all the pairwise distances between their members are computed. The average of these distances will be used as a distance/dissimilarity measure between the two clusters;
  - **Ward linkage**, at each iteration, works by merging the clusters having the smallest between-clusters distance.

TABLE 4.2.1: GRID SEARCH FOR THE LINKAGE CHOICE

LINKAGE METHODS	GOWER DISTANCE
Average	0,600
Single	0,304
Complete	0,776
Ward	0,934

The hierarchical clustering analysis was conducted by using an agglomerative approach: the linkage method was chosen with a grid search algorithm while the chosen distance measure is the Gower one. The Ward linkage method seems to perform better (TABLE 4.2.1) than the others.

PLOT 4.2.1: HIERARCHICAL CLUSTERING, WARD LINKAGE USING GOWER DISTANCE



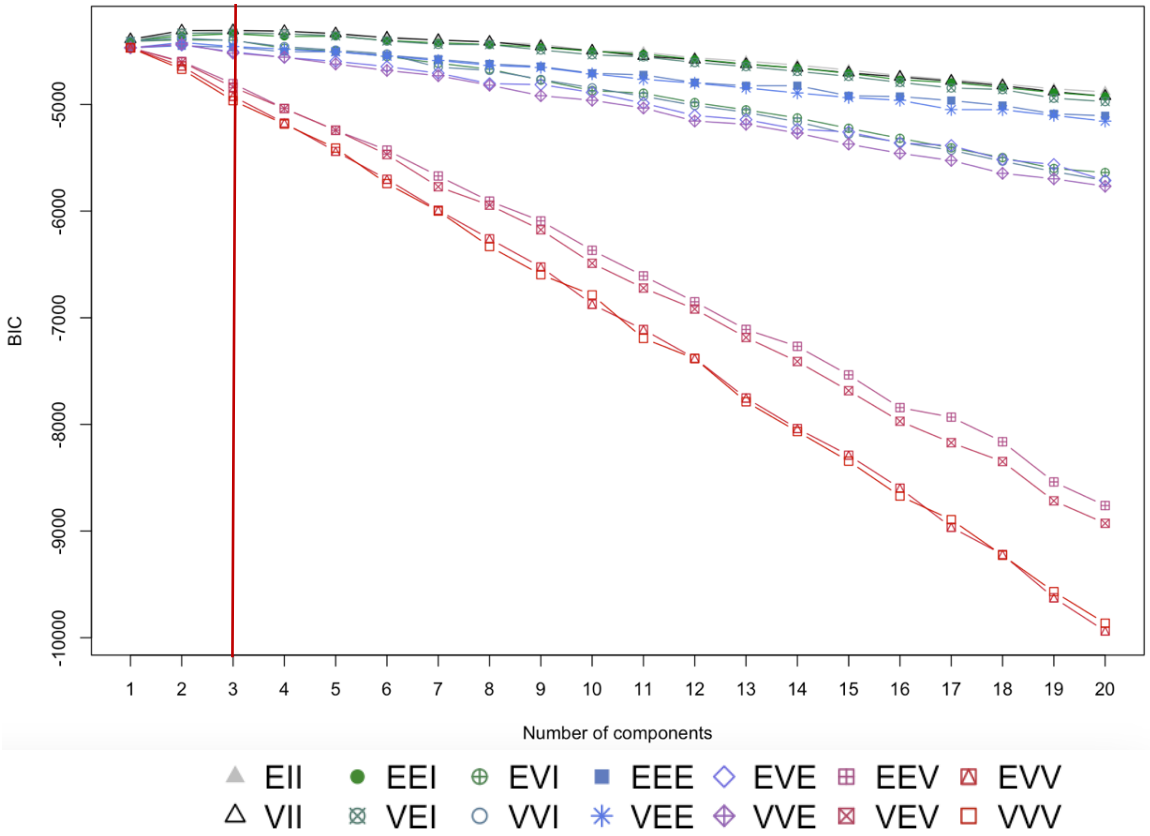
As aforementioned, one of the main advantages of hierarchical clustering methods is that they can be intuitively represented by the mean of a tree plot, also called dendrogram, highlighting their nested structure. This plot is useful not only from an aesthetic point of view, but also in order to choose the optimal number of clusters.
In fact, in contrast to non-hierarchical methods, hierarchical ones do not require to pre-specify the initial number of clusters (the *k* hyperparameter in the previously implemented k-means non-hierarchical approach): the optimal number of clusters is usually considered the one that require a large change in the distance measure to be modified (longer horizontal segments). In the current situation, the resulting dendrogram, shown in PLOT 4.2.1, emphasises the presence of **three clusters**.

4.3 MODEL-BASED CLUSTERING

The clustering methods previously implemented are based on a hard assignment that, in most real cases, turns out to be not really appropriate.
In fact, there may be some individuals whose cluster belonging is not really clear.
Model-based clustering allow to overcome this drawback, by the mean of a soft assignment.
Each individual is given a certain probability of belonging to each cluster:
the classification is done by considering the higher probability value.

In analogy with the k-means approach, the initial number of clusters *k*, is a hyperparameter that needs to be tuned by the mean of a grid search algorithm. The underlying assumption behind this method is that each k-th cluster can be considered a mixture of Gaussian distributions of order p (number of features in the considered dataset), having their own mean vector and variance-covariance matrix, thus leading to an overparameterization problem that can be overcome using the reduced dataset.
Acting this way, the model-based clustering considers *k* multivariate distributions of order 10 (number of neurons in the encoders neural network).

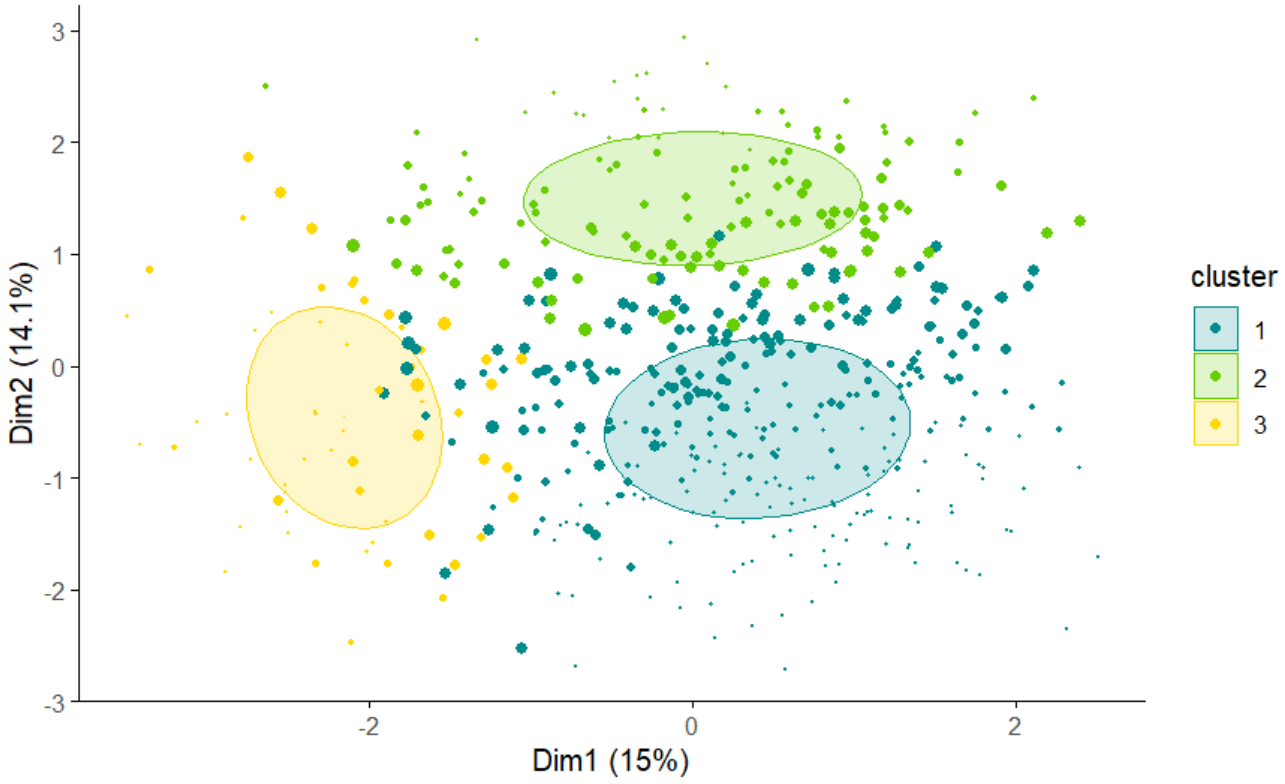
PLOT 4.3.1: BIC VALUE WITH VARYING NUMBER OF CLUSTERS AND CONSTRAINTS



The grid search algorithm works by computing the BIC value for each model, having a different number of clusters and features of the variance-covariance matrix.
The chosen mixture model has three clusters, i.e. three multivariate normal distribution with varying size and equal shape, made up of the 10 encoders previously selected.

Since the encoders lead to a reduced numerical representations of the original set of variables, it is possible to perform a PCA, in order to graphically represent the results.
PLOT 4.3.2 shows not only the detected clusters but also an important information:
the assignment uncertainty. The level of uncertainty is represented with a larger point size.
It is possible to notice that the individuals having larger uncertainty are those that lie in the region among the three clusters.

PLOT 4.3.2: UNCERTAINTY IN CLUSTER ASSIGNMENT



Given the difficulty in classifying certain observations, i.e. those in the area among the three clusters, it may be helpful to identify and characterize these individuals.
The first 10 participants whose assignment is most uncertain present an uncertainty value between 0.57 and 0.49.
These observations regard unmarried 36-year-old individuals with a mediocre level of physical, financial and mental wellness, who do not practice any specific physical activity and exercise only occasionally.
This suggests that there may be a problem in the data, probably due to the inappropriateness of some questions in the questionnaire.
For instance, individuals who reported not engaging in any sports, but occasionally exercising, may sometimes participate in activities such as football.

4.4 CLUSTERS PROFILING

In order to choose the best clustering method a first attempt was made according to the interpretability of the detected clusters. Each method was able to detect three different clusters having different dimensions.
The profiling was carried out by considering the mean or the modal values of the most important variables, i.e. those having the greater discriminant power with respect to the clustering belonging. For all the implemented methods, the most important variables are:

- Age;
- Marital status;
- Physical activity status;
- Physical activity frequency;
- Physical wellness;
- Mental wellness;
- Financial wellness;
- Career wellness;
- Occupational status;
- Educational level.

K-MEANS CLUSTERING

The k-means approach led to three clusters having the dimensions showed in TABLE 4.4.1.
On the base of the most discriminant features, the related profiles are the following:

- CLUSTER 1: young, unmarried students, having at least a degree and with an overall good wellness (physical, financial, mental and career wellness);
- CLUSTER 2: young, unmarried students, having a high-school title and with an overall mediocre wellness (physical, financial, mental and career wellness);
- CLUSTER 3: married, full-time workers adults, with a high-school title, and an overall good wellness (physical, financial, mental and career wellness).

TABLE 4.4.1: K-MEANS CLUSTERS DIMENSIONS

K-MEANS CLUSTERING	FREQUENCIES
Cluster 1	197
Cluster 2	184
Cluster 3	131

HIERARCHICAL CLUSTERING

The hierarchical clustering approach led to three clusters having the dimensions showed in TABLE 4.4.2.
On the base of the most discriminant features, the related profiles are the following:

- CLUSTER 1: young, unmarried students, having at least a degree and with an overall good wellness (physical, financial and mental wellness), who train frequently;

- CLUSTER 2: young, unmarried students, having a high-school title and with an overall mediocre wellness (physical, financial and mental), who train rarely;
- CLUSTER 3: married, full-time workers adults, having at least a degree and good levels of financial and mental wellness and a mediocre physical wellness, who never train.

TABLE 4.4.2: HIERARCHICAL CLUSTERS DIMENSIONS

HIERARCHICAL CLUSTERING	FREQUENCIES
Cluster 1	225
Cluster 2	182
Cluster 3	105

MODEL-BASED CLUSTERING

The model-based clustering approach led to three clusters having the dimensions showed in TABLE 4.4.3. The related profiles are the following:

- CLUSTER 1: young, unmarried students, having an overall good wellness (physical, social, financial and mental wellness), who train frequently and drink alcohol 1-2 times weekly;
- CLUSTER 2: young, unmarried students, having an overall mediocre wellness (physical, social, financial and mental), who never train and drink alcohol 1-2 times weekly;
- CLUSTER 3: married, full-time workers adults, having a good physical, social, financial and mental wellness, who never train or drink alcohol.

TABLE 4.4.3: MODEL-BASED CLUSTERS DIMENSIONS

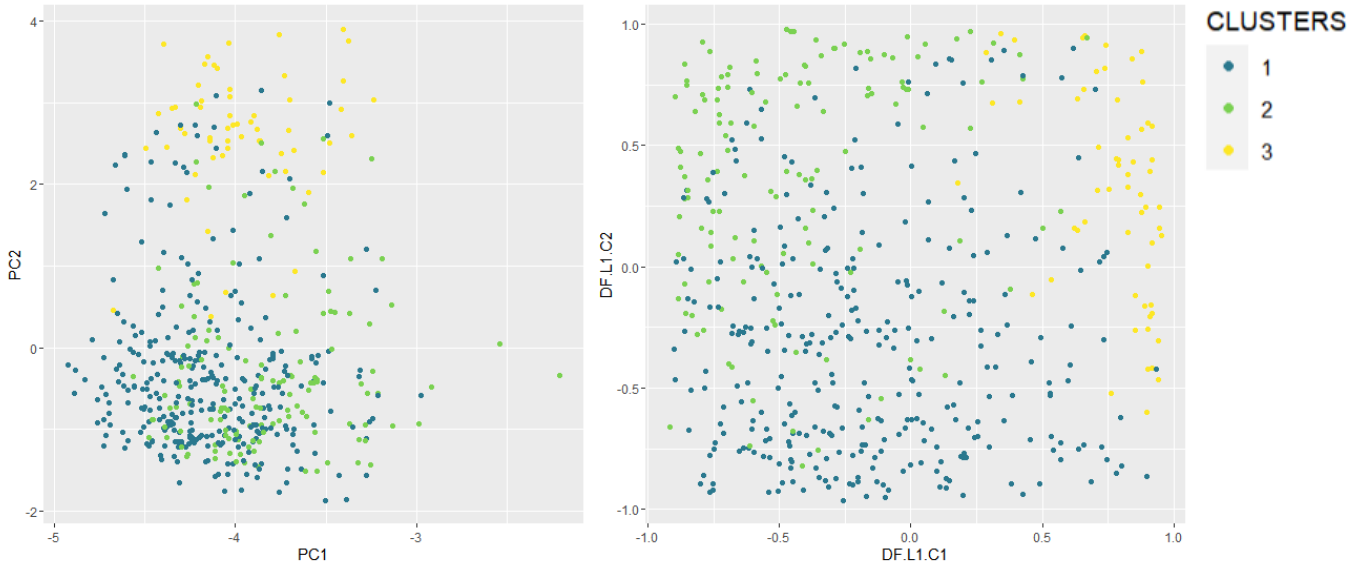
MODEL-BASED CLUSTERING	FREQUENCIES
Cluster 1	317
Cluster 2	136
Cluster 3	59

Since the obtained profiles are almost similar, the interpretability method cannot be used when choosing which clustering method adopt.
Instead, methods properties were taken into account:
given its flexibility, besides the possibility to detect uncertain observations, the choice fell on the model-based approach.

4.5 CLUSTERS VISUALIZATION IN 2D: PCA VS AUTOENCODERS

The choice about which dimensionality reduction method adopt fell on the autoencoders method for several reasons. The Principal Components Analysis, in fact, is more recommended when having quantitative variables. Nevertheless, the one-hot encoding allowed a numerical representation, but the original nature of these categorical variables remained unchanged. A more feasible method is given by the autoencoders, able to handle variables of mixed nature.
Moreover, in contrast to PCA, autoencoders are capable of catching complex and non-linear signals in data.

PLOT 4.5.1: 2D CLUSTERING REPRESENTATION USING DIFFERENT REDUCTION METHODS



At equal dimension reduction, PLOT 4.5.1 shows how the clustering structure changes according to the used dimensionality reduction method. The first plot shows just two of the first three principal components that it should have been used. Whereas the second plot show just two out of the ten components of the encoders hidden layer.
Despite being randomly picked, the two encoders seem to better represent the clustered structure in data, even if some overlapping individuals are present.
On the other hand, PCA tend to overlap more the clusters visualization, in particular with respect to the first and the second clusters.

5. CLASSIFICATION ANALYSIS

Having assigned each individual to a certain cluster, it may be useful developing a machine learning tool capable of assigning new future persons to a group, to a certain profile. The original dataset was randomly partitioned in:

- **training set** (80% of the participants), used in order to train the classificators;
- **test set** (20% of the participants), used in order to test the classificators performance.

The partition was carried out stratifying with respect to the response variable to classify (the cluster of belonging), in order to maintain the original dimension of each cluster (TABLE 4.4.3).

The following classification methods are all based on the simple decision tree structure and consist in different ways to combine trees together:

- random forest;
- basic gradient boosting;
- stochastic gradient boosting.

5.1 RANDOM FOREST

Random forest is an ensemble method that consists of assembling the decision of a certain number of simple decision trees, built on bootstrapped samples.
A particular feature of random forests is that at each split of each tree, the variable to be used can be chosen from a subset of the original set of variables, in order to highlight the discriminant power of those variables that may be masked by other ones.

Random forests involve some hyperparameters that need to be tuned:

- Number of trees in the forest;
- Number of variables to consider at each split;

} They have the largest impact on results.

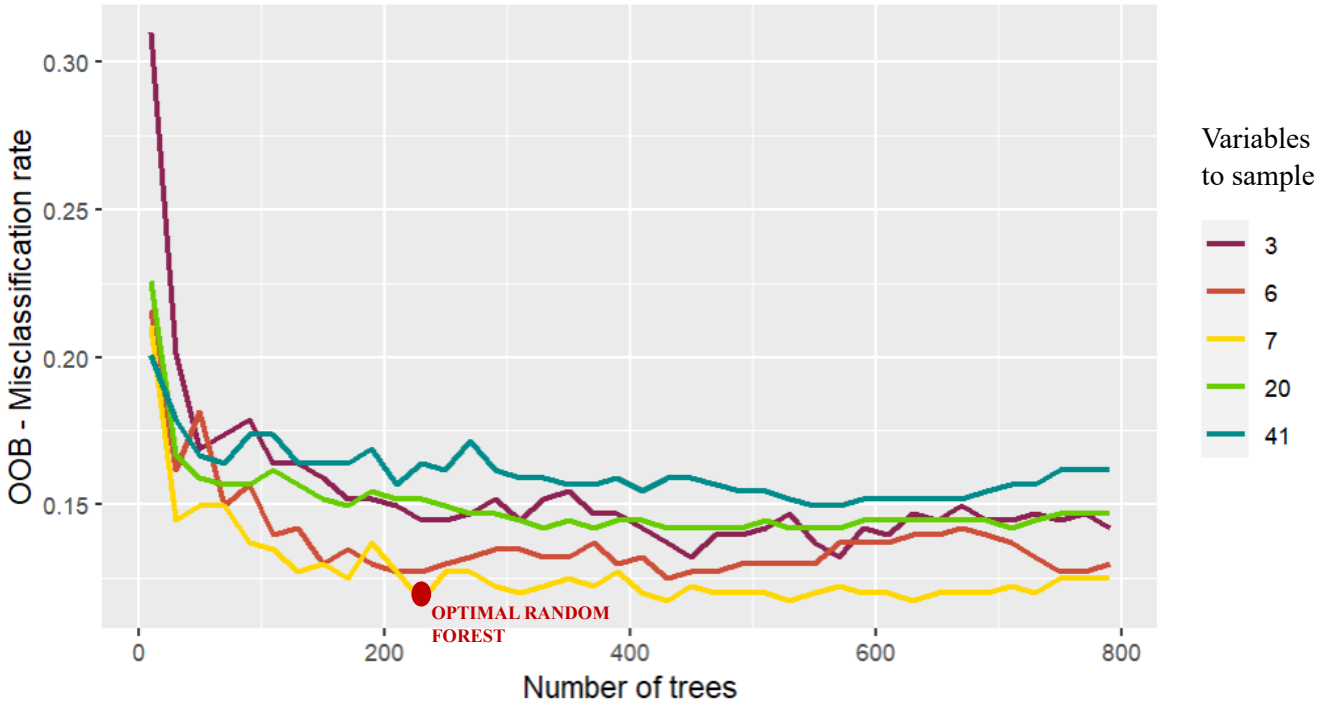
- Minimum size of each node in each decision tree;
- Sample fraction with respect to the original set of observations;
- Whether to sample observations with or without replacement.

Hyperparameters tuning was carried out by the mean of the grid search algorithm:
for each of all the possible hyperparameters combinations, a model was adapted.
The choice fell on the combination capable of minimizing the proportion of wrong prediction (misclassification rate) of the Out Of Bag observations (not sampled observations for each tree).

For computational reasons a two-steps tuning was implemented (final results are shown in TABLE 5.1.1):

- 1° step: minimum node size, sample fraction, sampling type;
- 2° step: given the hyperparameters previously tuned, the number of trees and of variables to sample at each split was tuned. PLOT 5.1.1 shows the performance of the forests when both the number of trees and of variables change.

PLOT 5.1.1: RANDOM FORESTS PERFORMANCE COMPARISON



It is possible to notice that a greater number of variables to be sampled does not always lead to better performances. The lowest out of bag misclassification rate is obtained by sampling 7 variables at a time. On the other hand, the optimal number of trees to be included in the forest was found to be around 230 with a misclassification rate (OOB) of about 0.118.

TABLE 5.1.1: TUNED HYPERPARAMETERS

FEATURES OF THE SELECTED RANDOM FOREST	
MINIMUM TREE NODE SIZE	5 observations
SAMPLING SCHEME	Without replacement
SAMPLE FRACTION	70% of the original sample size
NUMBER OF TREES	230
NUMBER OF VARIABLES TO SAMPLE	7

PLOT 5.1.2 displays the confusion matrix for the random forest classifier:
it consists in a representation able to compare the cluster of belonging with the one predicted, assigned by the random forest model with respect to the test set observations.
Starting from this kind of matrices it is possible to obtain some useful summary statistics that refer both to the overall and to the cluster-specific predictive performance (TABLE 5.1.2).

PLOT 5.1.2: RANDOM FOREST CONFUSION MATRIX

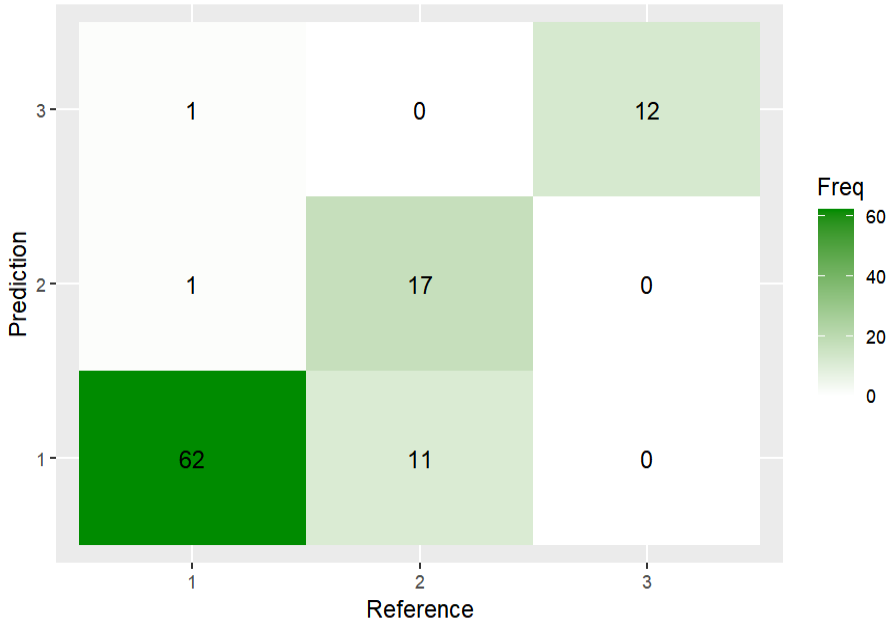


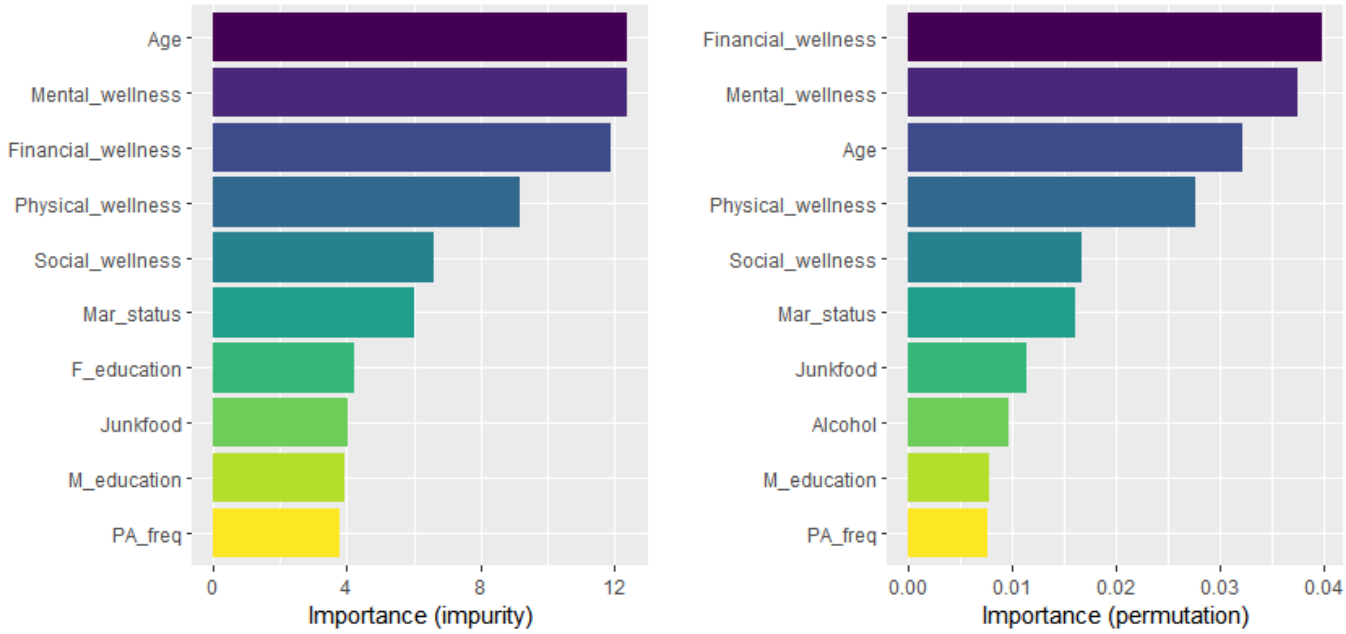
TABLE 5.1.2 shows that for the three identified clusters the proportion of correctly classified observations (accuracy) is equal to 0.875. Moreover, the values of specificity and sensitivity change across the clusters:

- Only the 60% of persons belonging to the second cluster were correctly assigned by the model (cluster-specific sensitivity equal to 60.7%);
- The optimal random forest was able to correctly assign all the individuals belonging to the third group (cluster-specific sensitivity equal to 100.0%). As a consequence, all those that weren't classified as belonging to this cluster, really didn't belong to it;
- Just ¾ of the participants assigned to the first cluster truly belong to it (cluster-specific specificity equal to 72.5%).

TABLE 5.1.2: RANDOM FOREST SUMMARY PERFORMANCE STATISTICS

METRICS	CLUSTER 1	CLUSTER 2	CLUSTER 3
ACCURACY	87.5%		
MISCLASSIFICATION RATE	12.5%		
SENSITIVITY	96.9%	**60.7%**	**100.0%**
SPECIFICITY	**72.5%**	98.7%	98.9%
+ PREDICTIVE VALUE	84.9%	94.4%	92.3%
- PREDICTIVE VALUE	93.5%	87.2%	**100.0%**

PLOT 5.1.3: RANDOM FOREST VARIABLES IMPORTANCE



PLOT 5.1.3 shows the importance of each variable, computed using two different metrics:

- the 1° metric considers the variation in the Gini index when the variable is removed from the model;
- the 2° metric considers the variation in the accuracy value when the levels of the variable are randomly permuted.

The results are almost similar. The most important variables, i.e. those having the greater discriminant power with respect to the clusters structure, are age, financial wellness, mental wellness, social wellness, physical wellness and marital status.

5.2 GRADIENT BOOSTED TREES: BASIC VERSION

Gradient boosting is an ensemble machine learning technique used in regression and classification tasks that combines various weak learners, typically shallow trees. A weak model is one whose error rate is only slightly better than random guessing. Whereas random forests build an ensemble of independent trees, GBM build a sequence of trees in which each tree improves the previous one, so they are not independent. The key idea behind gradient boosting is to optimize an arbitrary differentiable loss function by the mean of the iterative gradient descent optimization algorithm:
at each iteration the previously solution is updated considering the gradient of the loss function computed with respect to the values predicted by the last solution (model), according to an hyperparameter called learning step.
Gradient boosted trees involve some hyperparameter to tune by the mean of a grid search algorithm:

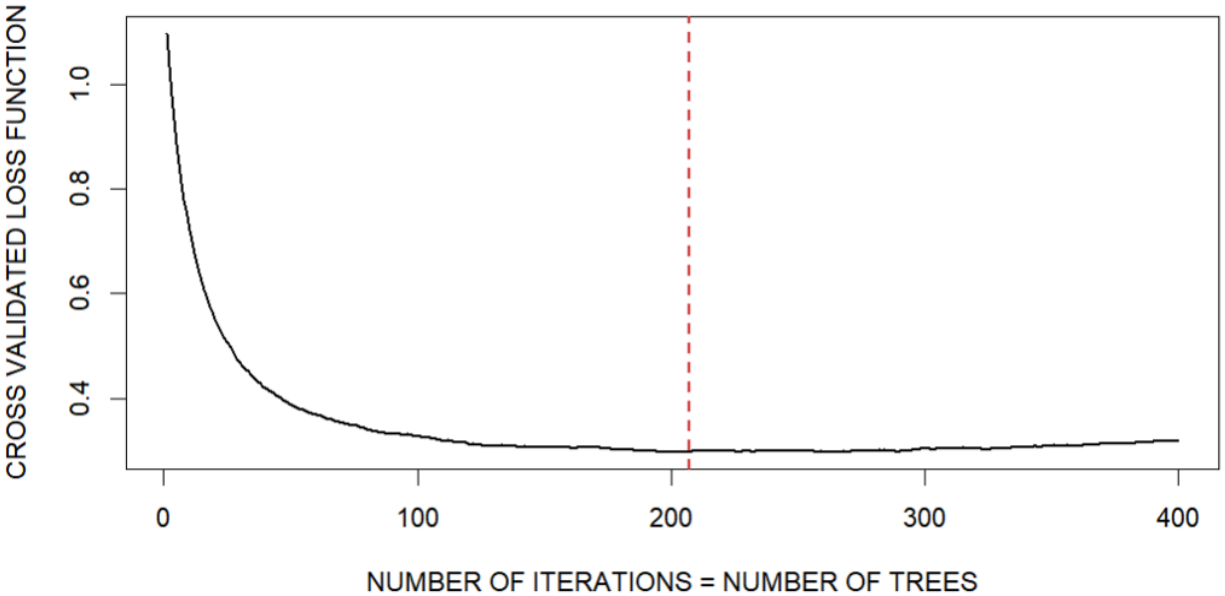
- learning rate (shrinkage), to control the impact of the gradient;
- interaction.depth, with respect to the maximum depth of each tree;
- n.minobsinnode, with respect to the minimum number of observation in each tree node;
- number of trees to combine.

The choice was based on the cross validated loss function (TABLE 5.2.1) while the number of trees was that one starting from which the performance stabilises (PLOT 5.2.1).

TABLE 5.2.1: TUNED HYPERPARAMETERS

FEATURES OF THE SELECTED BASIC GRADIENT BOOSTING	
MINIUM TREE NODE SIZE	15 observations
TREE DEPTH	2
LEARNING RATE	0.05
MINIMUM NUMBER OF TREES NEEDED	207

PLOT 5.2.1: BASIC GRADIENT BOOSTED TREES ACCORDING TO THE NUMBER OF TREES

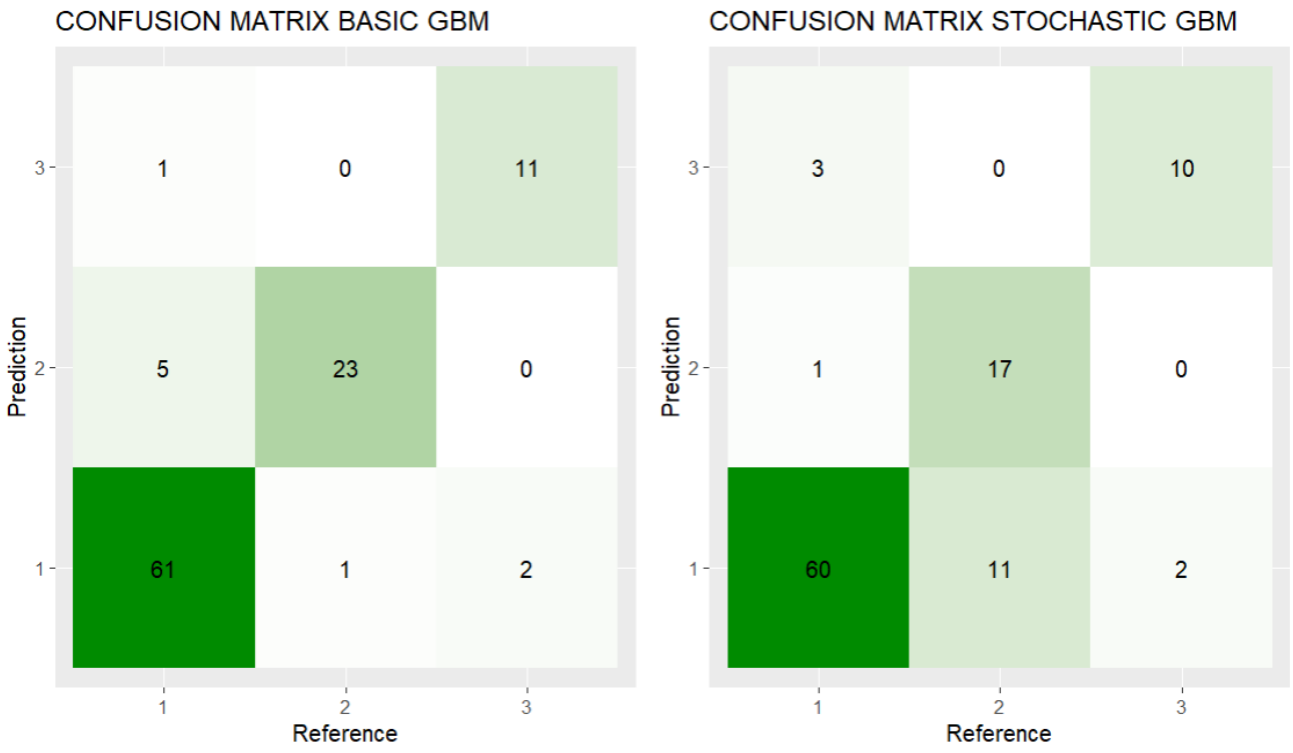


5.3. GRADIENT BOOSTED TREES: STOCHASTIC VERSION

The standard version of the Gradient Boosted Machine algorithm works by computing the gradient on the whole set of data. The following version consists of computing the gradient just on a random subsample of the original dataset, in order to obtain a method that is both faster and more robust with respect to the training data used.
The hyperparameters that this method involves are the same of the previous one, so the previously tuned values were used (TABLE 5.2.1). The additional hyperparameter consists of the sample rate, that is the amount of observation to sample at each iteration:
the chosen value, i.e. the one allowing to maximize the overall predictive accuracy, is equal to a 75% sample rate.

In order to make a comparison between the two implemented versions of the gradient boosting algorithm PLOT 5.3.1 shows the related confusion matrices while TABLE 5.3.1 and TABLE 5.3.2 highlight the derived performance statistics.

PLOT 5.3.1: CONFUSION MATRICES COMPARISON BETWEEN GBM VERSIONS



By looking at following tables it is possible to notice that:

- the basic gradient boosting is overall more capable of correctly assigning the true cluster, since the accuracy value is equal to 91.4% rather than 83.7% for the stochastic version;
- some results present few analogies. In fact, both models present the lowest sensitivity, positive predictive value and negative predictive value for the third, second and first cluster respectively, but the worst values refer to the stochastic GBM version;
- almost 100% of the individuals that truly do not belong to third cluster were not classified as belonging to the third cluster;

- overall the basic GBM performs better in classifying the clustering belonging, meaning that no additional stochasticity is needed in the classification process.

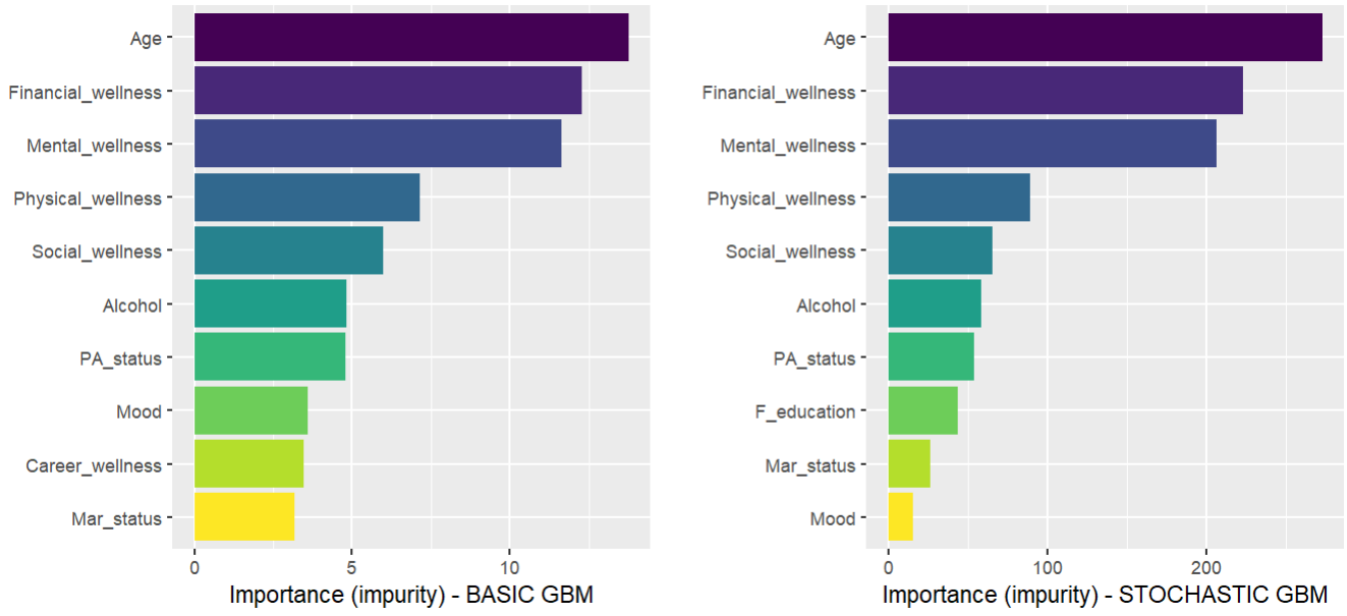
TABLE 5.3.1: BASIC GBM SUMMARY PERFORMANCE STATISTICS

METRICS	CLUSTER 1	CLUSTER 2	CLUSTER 3
ACCURACY	91.4%		
MISCLASSIFICATION RATE	8.6%		
SENSITIVITY	91.0%	95.8%	**84.6%**
SPECIFICITY	91.9%	93.8%	**98.9%**
+ PREDICTIVE VALUE	95.3%	**82.1%**	91.7%
- PREDICTIVE VALUE	**85.0%**	98.7%	97.8%

TABLE 5.3.2: STOCHASTIC GBM SUMMARY PERFORMANCE STATISTICS

METRICS	CLUSTER 1	CLUSTER 2	CLUSTER 3
ACCURACY	83.7%		
MISCLASSIFICATION RATE	16.3%		
SENSITIVITY	82.2%	94.4%	**76.9%**
SPECIFICITY	87.1%	87.2%	**97.8%**
+ PREDICTIVE VALUE	93.8%	**60.7%**	83.3%
- PREDICTIVE VALUE	**67.5%**	98.7%	96.7%

PLOT 5.3.2: COMPARISON BETWEEN VARIABLE IMPORTANCE IN BOTH GBM VERSION



PLOT 5.3.2 show the most important variables used in the GBM algorithms, in which the importance value is computed on the base of the change in the Gini index (impurity measure) when the variable is removed from the model. It is possible to notice that there are relevant similarities among the variables used by the two versions.

By making a comparison with respect to the random forest classification algorithm (PLOT 5.1.3) one can notice that age and the various types of wellness are always considered as important and discriminant features. Some differences regard the other variables, as an example the marital status, was considered having more importance in the random forest rather than in the two version of the GBM algorithm.

CONCLUSION

Among the 512 participants it was possible to detect the presence of three clusters:
one for the adults and two for the youngsters. There were, however, some middle-aged individuals that cannot be easily assigned to a cluster.
The most accurate clusters classifier, turned out to be the basic version of the gradient boosting machine, able to correctly predict more than the 90% of the test set individuals.