# Multivariate Analysis of a Breast Cancer Dataset

MVA - Final Project

*Carles Garriga Estrade i Balbina Virgili Rocosa*

*06/27/2018*

# Contents

**1 Introduction** **3**

**2 The Dataset** **3**

**3 The Pre-process of Data** **4**
3.1 Missing values . . . . . 4
3.2 Feature selection . . . . . 4
3.3 Outlier detection . . . . . 5
3.4 Feature extraction . . . . . 5

**4 The Protocol of Validation** **7**

**5 Data visualization** **7**
5.1 Factorial analysis on Mixed Data . . . . . 7
5.2 Clustering . . . . . 11

**6 Predictive methods** **13**
6.1 Generalized linear model . . . . . 13
6.2 Decision Trees . . . . . 14
6.3 Random Forest . . . . . 15

**7 Conclusions** **16**

# 1 Introduction

Breast cancer is known to be the most occurred cancer for women worldwide and one of the most common cancer that causes death. During the past decade, breast cancer has been deeply studied and thanks to it, an increase on the prognosis rate and a decrease on death rate have been achieved. Nevertheless, further research is still needed to achieve full understanding of its mechanism and corresponding efficient treatment. It is known to happen because most cancer treatments tend to be too general, since the doctors normally give patients of different characteristics similar suggestions. But the usage of statistical methods to understand cancer pretend to improve the prediction of cancer and easily identify significant genes to be able to apply a more specific treatment to each patient.

The main objective of this project is to successfully develop a multivariate analysis from a chosen dataset, which contains information about breast cancer of many different patients. More concretely, we want to, finally, be able to create a good model to predict a possible eventual death of patients that suffer breast cancer.

# 2 The Dataset

The chosen dataset is called *NKI Breast Cancer* and its data is collected by NKI (Netherlands Cancer Institute). The dataset contains information of 272 breast cancer patients and 1570 attributes for each of them, which include 3 patient related attributes, 13 clinical attributes and 1554 gene attributes. More detailed, the attributes recorded for each patient are the following ones:

- *Patient*: [String] categorical patient identifier.
- *Id*: [Integer] numerical patient identifier.
- *Age*: [Integer] patient age.
- **Eventdeath**: [Boolean] whether the patient died of breast cancer or not. This is the value that we are interested in predicting for each individual.
- *Survival*: [Float] survival time.
- *Timerecurrence*: [Float] recurrence time. It is generally the same as survival time when the patient did not die from breast cancer and, in other cases, it is smaller.
- *Chemo*: [Boolean] whether the patient has received a chemotherapy or not.
- *Hormonal*: [Boolean] whether the patient has received hormonal therapy or not.
- *Amputation*: [Boolean] whether forequarter amputation has been used as a treatment or not.
- *Histtype*: [Factor] histological type, which refers to the growth pattern of the tumors.
- *Diam*: [Integer] diameter of the tumor size.
- *Posnodes*: [Integer] number of positive nodes.
- *Grade*: [Factor] three levels indicating cancer grade, which is an indicator of how quickly a tumor is likely to grow and spread.

- *Angioinv*: [Factor] three levels indicating the extent to which the cancer has invaded blood vessels.
- *Lymphinfil*: [Factor] three levels indicating the level of lymphocytic infiltration.
- *Barcode*: Clinical patient identifier.
- **Gene attributes**: [1554 Float]

As we can realize, **our dataset has more variables than individuals.**

# 3 The Pre-process of Data

Data is likely to contain many errors, be incomplete or inconsistent. That is why data preprocessing is needed to deal with those issues to prepare the data for further analysis / processing. Once the dataset has been loaded, the first step that must be done is the pre-processing of the data. For it, the following techniques have been studied and applied to our chosen dataset.

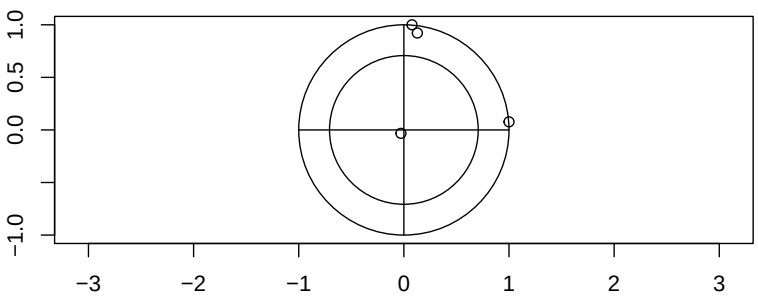
## 3.1 Missing values

As we already know, missing values are represented with the missing code NA by default. That is why, its evaluation has been performed by searching for NA values on the loaded dataset. With the results retrieved, we can confirm that there are no missing values on the given data so no further treatment is needed to be done to be able to compute them.

## 3.2 Feature selection

Afterwards, we have analyzed the different attributes given for each individual. At first glance, we have realized that Patient, Id and Barcode do not contribute with any useful information for the further analysis, as all of them are just unique identifiers of each patient of the dataset. Therefore, all of them are removed from the dataset.

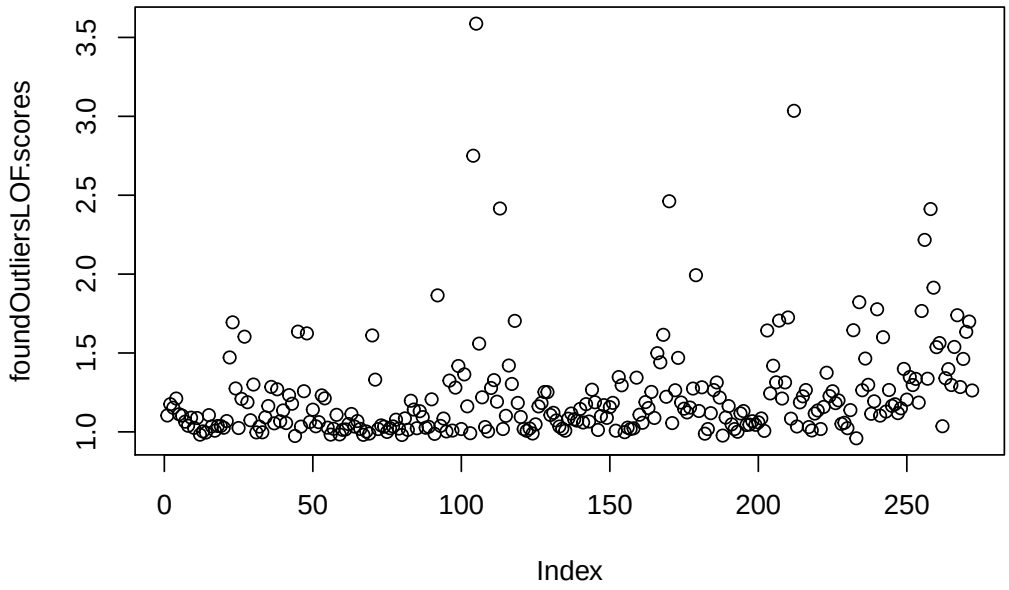
Furthermore, we have performed a correlation analysis for the numerical clinical variables (excluding the gene ones) in order to avoid co-linear variables. That is why we have computed the correlation between variables and we have identified the ones that are highly correlated ($> 0.9$). With the results retrieved, we can see that survival and timerecurrence are highly correlated, and we have decided to delete survival to not produce multicollinearity in an independent variable.



### 3.3 Outlier detection

Another factor that we need to take into account are the outliers of our data. In order to detect and remove them, two methods can be used. As our dataset has more columns than individuals, we are not allowed to implement the first implemented method that uses the Mahalanobis distance and the principal components.

Then, a second method, based on the local outlier factor, is applied. It provides a score based on the distance of a point to its k-nearest neighbors (density of the neighborhood) compared with the density of the neighbors of the first point (distances of the k-nearest neighbors of the neighbor). The resulting score value for each point determines if the individual would be considered an outlier or not.




With the results retrieved, we can see that most of the individuals lie between 0 and 2, and there are just a few individuals that have an score value calculated above 2, which we could detect as outliers. But the individual with the higher score calculated is 3.5, so we think that the difference is not high enough to be considered as outliers.

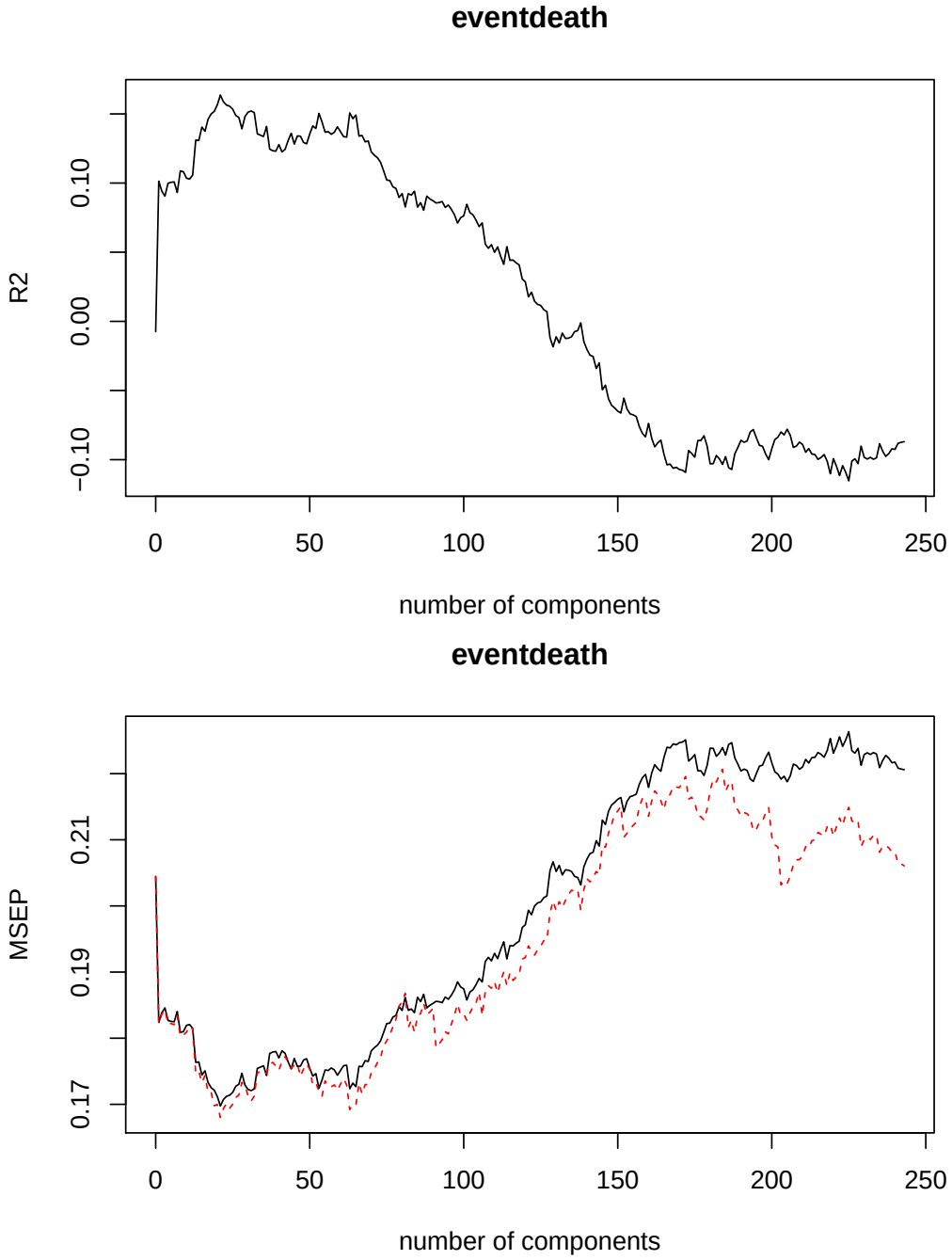
### 3.4 Feature extraction

Finally, as having a large number of variables makes the data quite difficult to treat, we have reduced the dimensionality of the data, by substituting all the gene attributes for its most important principal components.

To do it, we have performed a principal components regression (PCR) to all gene attributes (variables 13 to 1566 of the dataset). The basic idea behind PCR is to calculate the principal components of the desired data and then use some of these components as predictors in a linear regression model fitted using the typical least squares procedure. The linear model defined has been the eventdeath as response variable taking into consideration all the other variables of the training set.

For being able to reduce the dimensionality with PCR, we need to decide which number of components is the best one, by obtaining the minimum number of components that keep most of the variability

of the original variables and the ones that maximizes the R2 and minimizes the MSEP the most. With the results obtained, we can see that the number of components that we need to define are the first 21.




We finally join the first 12 variables of the data (regarding clinical attributes) with the new calculated ones, which have reduced the dimensionality of the gene attributes. This way, we have been able to reduce the dimensionality of the whole data from 1566 to 33. Therefore, now, **our dataset has more individuals than variables** and it will be easier to treat and extract latent information than from the initial dataset.

So, the current dimensions of the data, after performing the whole preprocessing, are the following

ones. Note that the number of individuals remains the same.

```
##  [1] "age"           "eventdeath"     "timerecurrence" "chemo"
##  [5] "hormonal"      "amputation"     "histtype"       "diam"
##  [9] "posnodes"      "grade"          "angioinv"       "lymphinfil"
## [13] "Comp 1"        "Comp 2"         "Comp 3"         "Comp 4"
## [17] "Comp 5"        "Comp 6"         "Comp 7"         "Comp 8"
## [21] "Comp 9"        "Comp 10"        "Comp 11"        "Comp 12"
## [25] "Comp 13"       "Comp 14"        "Comp 15"        "Comp 16"
## [29] "Comp 17"       "Comp 18"        "Comp 19"        "Comp 20"
## [33] "Comp 21"
```

# 4 The Protocol of Validation

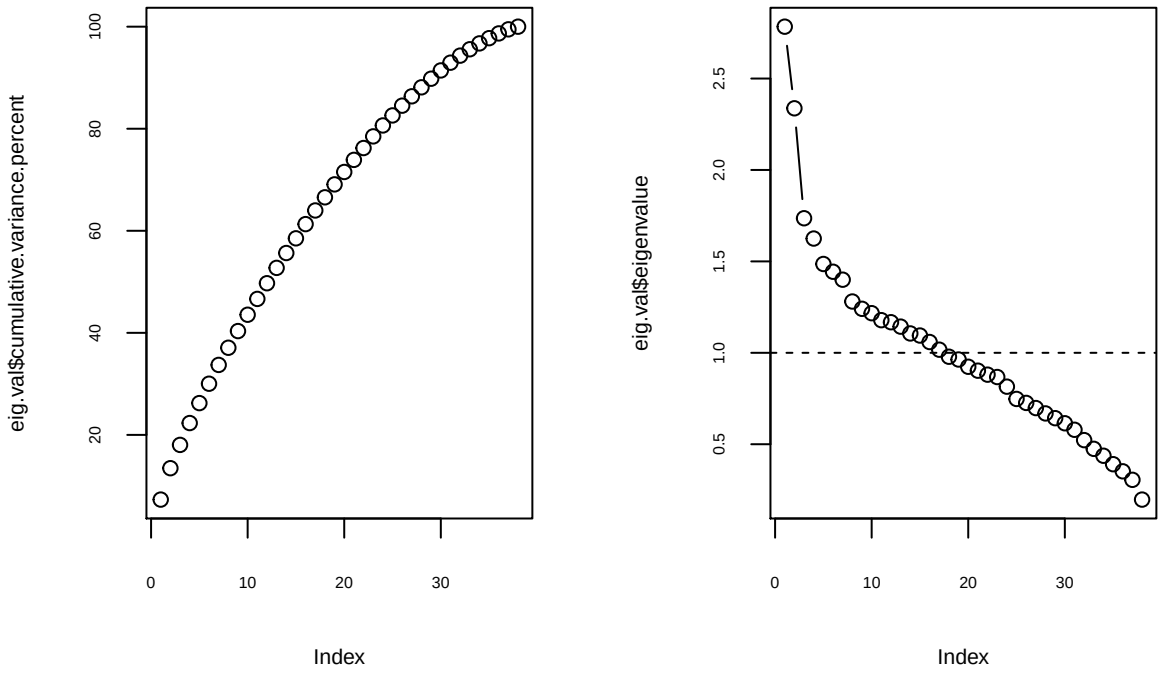
As we have already stated before, the total number of individuals that our dataset has is 272. Even it is not a very large number, we have preferred to apply the Holdout method for selecting a validation dataset because we have few patients that differ much to the others and we think that Leave-One-Out method would perform worse on our data.

So, using the Holdout method, we have separated the data into testing set (20% of data) and training set (80% of data remaining).

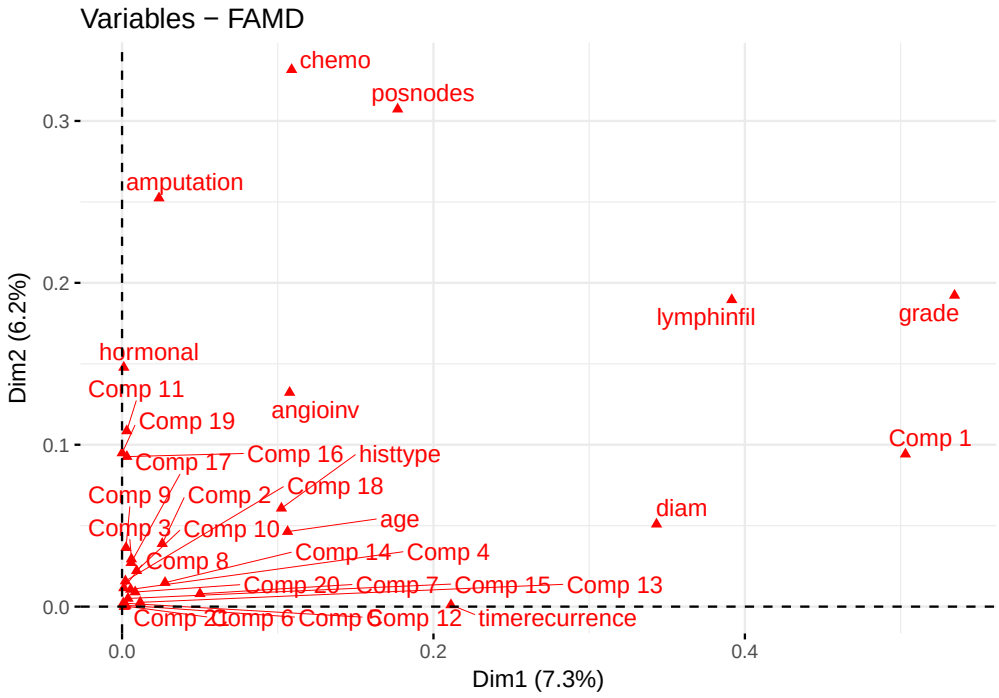
# 5 Data visualization

## 5.1 Factorial analysis on Mixed Data

After developing the whole preprocessing and defining the protocol of validation, our newly merged dataset contains both qualitative and quantitative data. This introduces us the problem of having to compute both the principal component analysis (for quantitative variables) and multiple correspondence analysis (for qualitative data). However, factoMineR provides a principal component method that allows to have mixed type data: Factor Analysis for Mixed type Data (or FAMD). As a matter fact, this method can be interpreted as a PCA on the quantitative data and MCA on the qualitative one. During this analysis, the training set of the data are used as active data, while the test individuals and the evendeath variable are taken as supplementary.

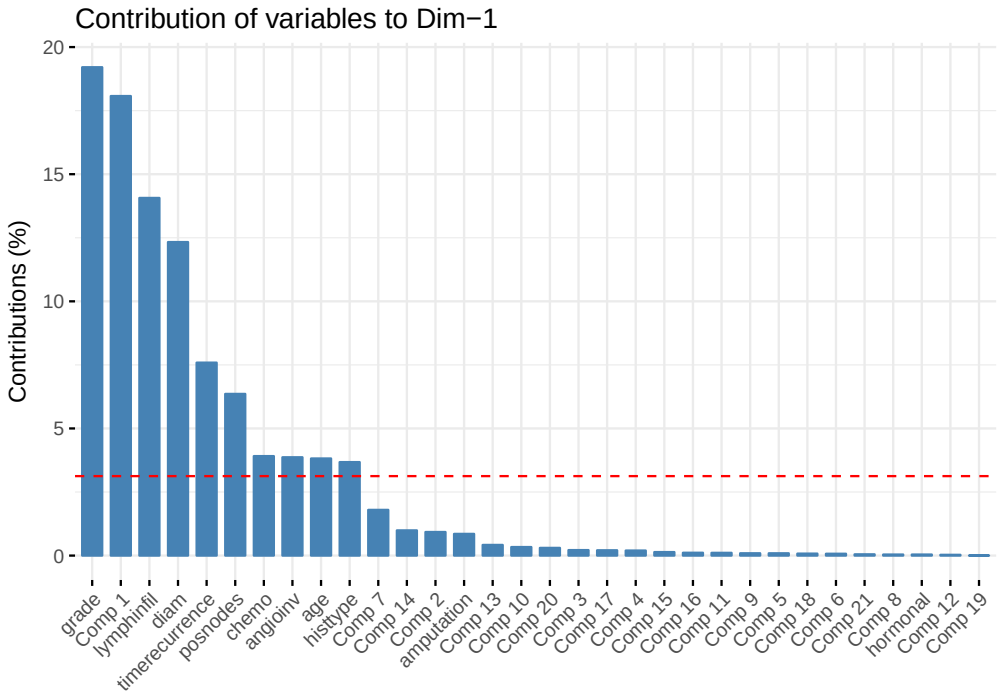



After obtaining the eigenvalues, we can compute the cumulative percentage of variance that each dimension represents (up to). Unfortunately, as we can see on the graphics retrieved, the cumulative variance percentages are represented as an "almost" linear distribution, which trades of as having similar percentage of variance represented per each dimension. And we are able to see that, taking into account only the first factorial plane, only the 13% of total variance is represented. So many dimensions should be kept in order to explain the significant percentage of the total inertia ($> 90\%$).

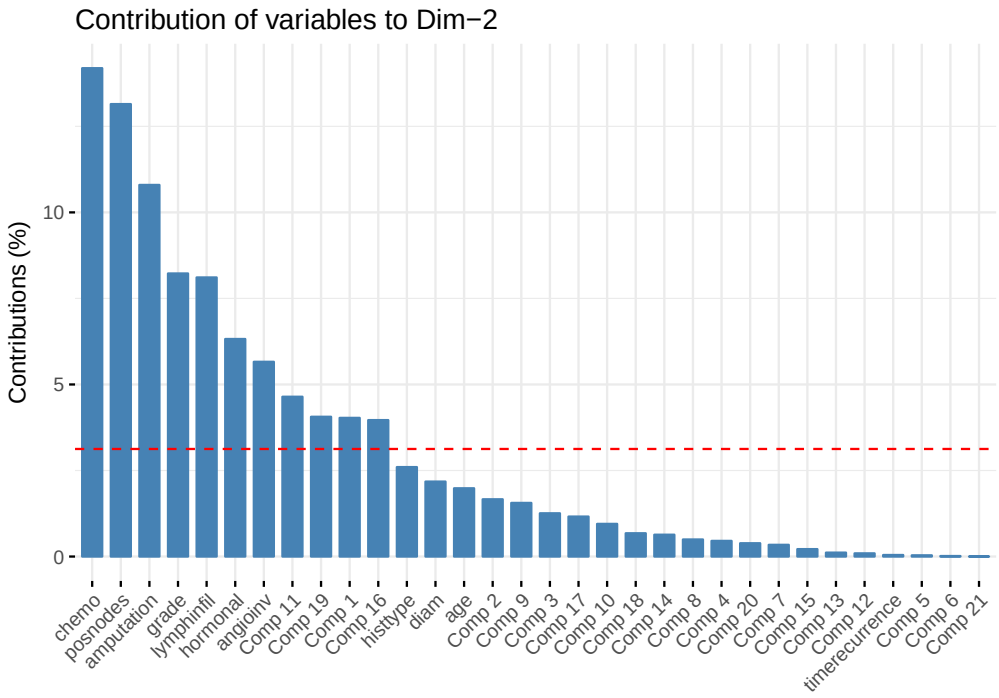



Moreover, plotting the variables in the first factorial plane allows us to distinguish several hidden groups:

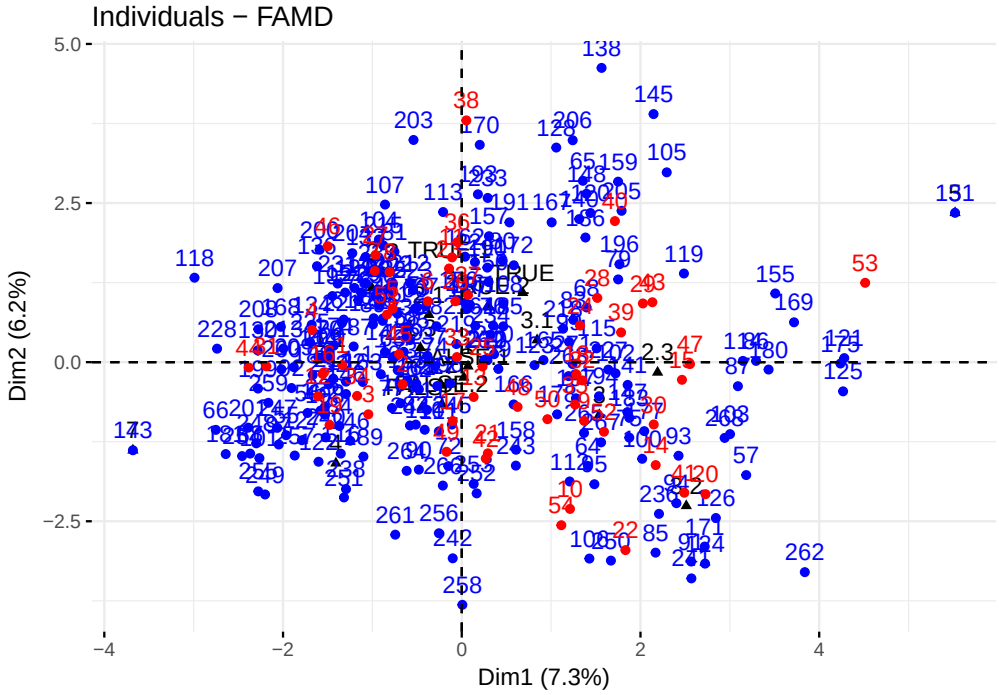
- A first group, that is based on the components *Comp x* extracted by the previous PCR execution. They don't represent much variance as the rest of factors in the first two dimensions. *hormonal*, *histtype* and *age* could also be relatable to those components.
- A second group based on *angionv*, being related in nearly the same measure to both first and second dimensions.
- A third group based on *grade*, *lymphinfil* and *diam*, being the most related to the first dimension. An hidden relationship could be explained, where the level of lymphocytic infiltration could determine the general grade of the tumor, also involving the firs component and the diameter of the tumor.
- Finally a last group based on *chemo* and *posnodes*, being the most related to the second dimension, could be related. As higher is the number of tumor nodes, more chances that a chemotherapy should be done. *amputation* could also be realatable to this group.




From the dimension point of view, *grade* , *Comp 1* and *lymphinfil* are the ones that contribute the most to the first dimension, followed by nearly the rest of the categorical factors. On the other side, the rest of the components *Comp x* are the one that contribute the less as well as hormonal.




For the second dimension, *chemo* , *posnodes* and *amputation* are the one that contribute the most to the first dimension, followed by nearly the rest of the categorical factors. On the other side (and the same as before) some components *Comp x* as well as time recurrence are the ones that contribute the less.



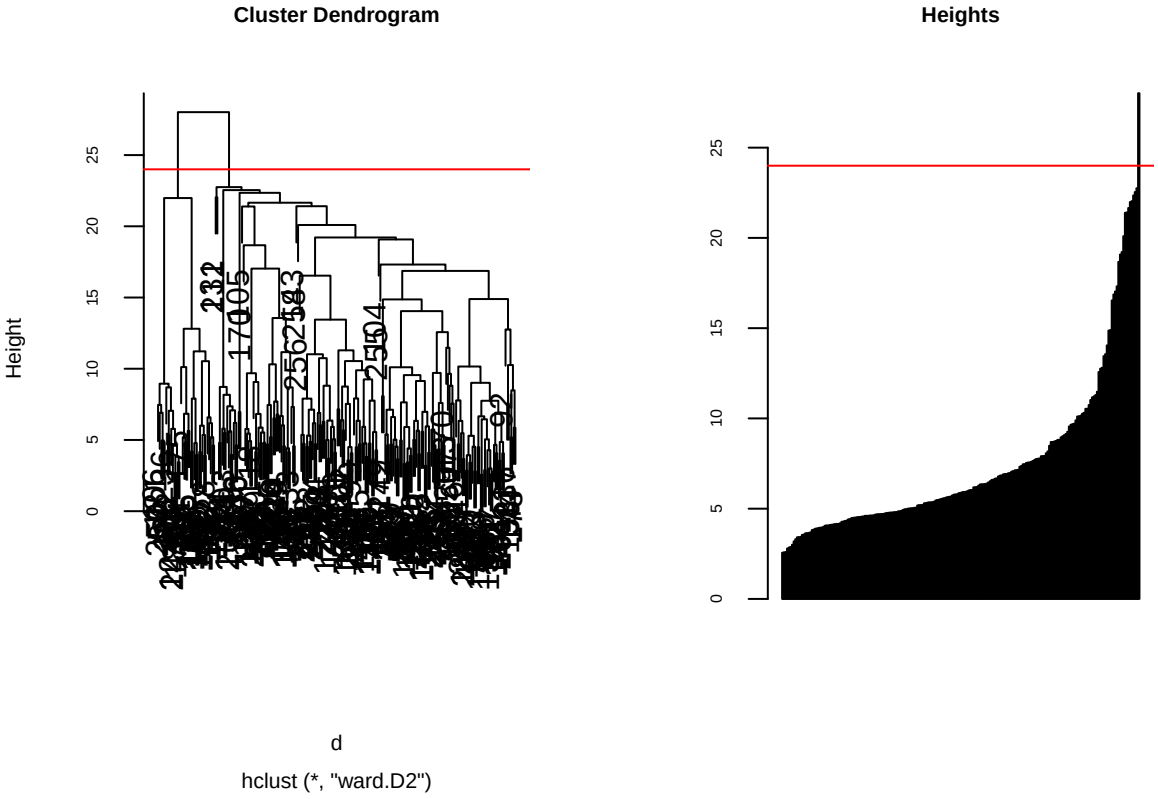

For the plot of individuals, we can observe a cloud with density near the center of the axis with a shape similar to an ellipsoid. However, several individuals remain outside of it and thus, being clearly entirely related to the first or second dimension (eg. 151, 53, 262 etc). Supplementary individuals (shown in red) could be considered inside the ellipsoid and they seem to don't provide

any drastic difference regarding the rest of the individuals in the first factorial plane.

However, we need to keep in mind that the first factorial plane, due to its low percentage of variance (~12%), cannot be used to represent the majority of variance explained by the dataset.

## 5.2 Clustering

Once having explored the hidden factors, we want to see if there is any available split for our individuals into clusters, in order to be able to characterize them. Our approach will be based on a consolidation clustering, where, first, an inital hierachical clustering is applied using the training set.

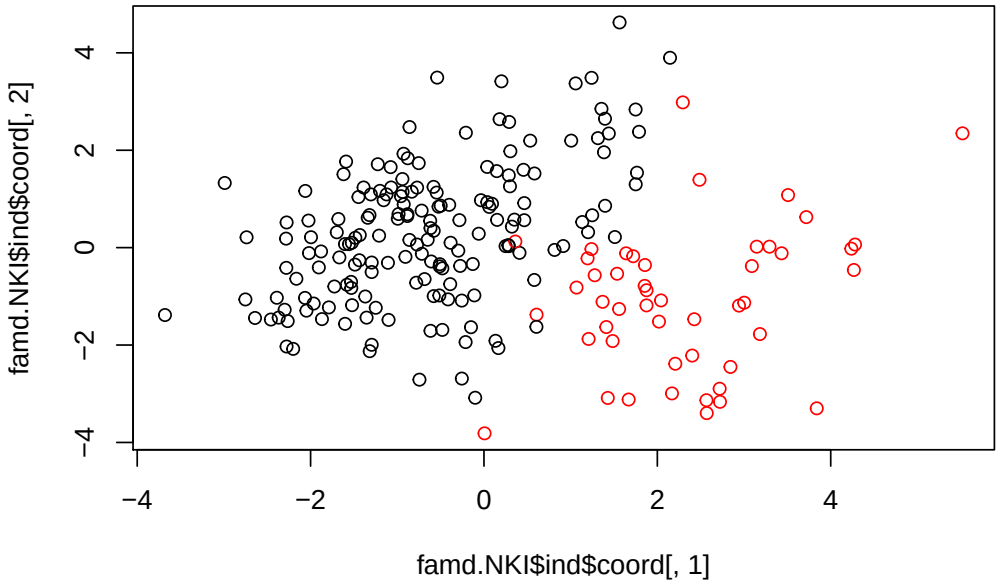



The resulting dendogram shows a uneven distribution of splits of individuals, which translates having more difficulties finding a suitable cut. An initial cut could be materialized into 2 classes. To guarantee that this cut is feasible, the depth of the number of clusters that this cut is going apply is revised (in the barplot of heights). Furthermore, the largest depth in the barplot of heights ends up being the cut that we want to apply. Other number of clusters could have been choosed but we decided to ensure not to have much numerous and complex clusters.

After deciding the number of clusters, one only iteration of k-means is executed. Hovewer, consolidation is applied using the centers calculated from the FAMD individuals coordinates.

The quality of the cut gives us a low value of 4.7365755%. This low value can be a consequence of the decision of splitting the populations in only two clusters of individuals.

After having executed kmeans, we recompute the quality of the clustering: 5.2792541%. Although there has been improved with regards to hierarchical clustering, we still consider it a low value.




Displaying the first factoral plane using the clusters as colors, allows us to see the splitting between individuals. However, due to the low variance percentage given, we still rely on the rest of dimensions in order to have a major variance explained.

Finally, the characteristics of each cluster are computed using the *catdes* method.

```
##
## Link between the cluster variable and the categorical variables (chi-square test)
## =================================================================================
##      p.value df
##
## Description of each cluster by the categories
## =============================================
## NULL
##
## Link between the cluster variable and the quantitative variables
## ================================================================
##               Eta2    P-value
## Comp.5 0.02259929 0.02645463
##
## Description of each cluster by quantitative variables
## =====================================================
## $`1`
##           v.test Mean in category Overall mean sd in category Overall sd
## Comp.5 -2.214508       -0.8222369   -0.2459172       6.604557   7.214727
##           p.value
## Comp.5 0.02679385
##
## $`2`
##          v.test Mean in category Overall mean sd in category Overall sd
## Comp.5 2.214508         1.795215   -0.2459172       8.750626   7.214727
##           p.value
```

```
## Comp.5 0.02679385
```

The first cluster can be identified by having:

- Quantitative variables:
  - Lower value on the fifth component *Comp 5* than the average.

The second cluster can be identified by having:

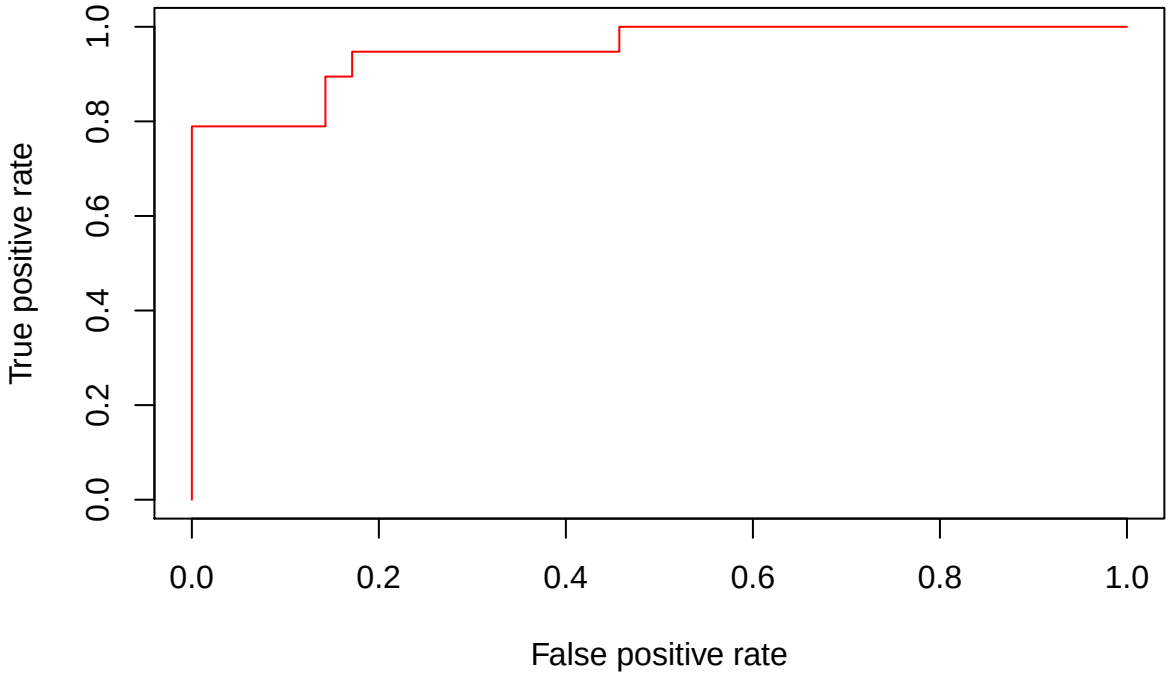
- Quantitative variables:
  - Higher value on the fifth component *Comp 5* than the average.

# 6 Predictive methods

Finally, we have developed three different models to predict a possible eventual death of patients that suffer breast cancer. Thanks to having a separated test set, we have 3 models, which have been previously trained with the training set, to test against unseen data.

## 6.1 Generalized linear model

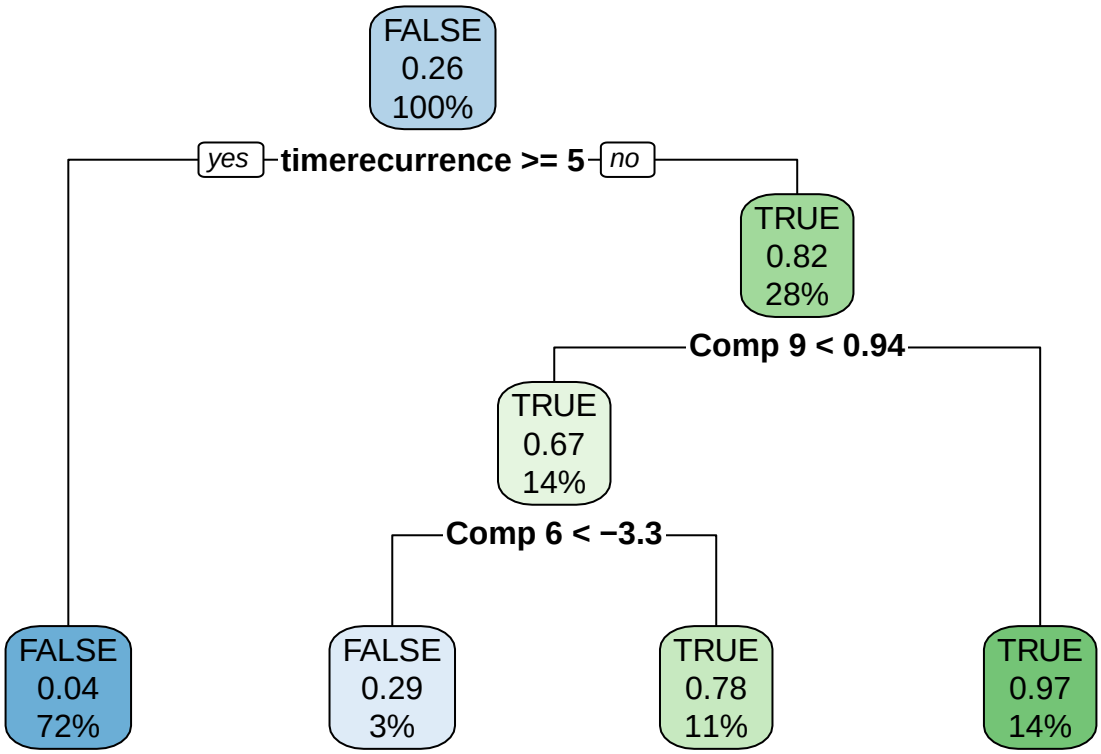
The first method used is a linear regression, via a generalized linear model with binomial family in order to predict binary outcomes. To do so, the glm function has been used by defining the appropriate formula for predicting the eventdeath variable taking into consideration all the other variables of the training set.



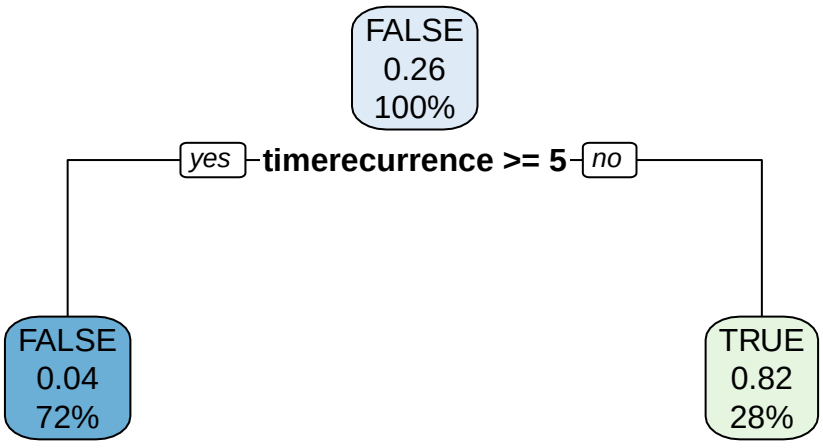

After setting a threshold for predicting the binary class (given probabilities), the accuracy achieved is 90.7407407 %.

### 6.2 Decision Trees

Once our data of test and training has been defined, we are able to obtain the decision tree to predict the variable eventdeath on the training data. To do so, the rpart function has been used by defining the appropriate formula for predicting the eventdeath variable taking into consideration all the other variables of the training set and also defining the complexity parameter and the number of cross-validations for our model. The visual representation of the cross-validation results obtained for our calculated decision tree are showed below.




We know that we obtain the optimal tree by pruning the maximal one up to the minimal cross-validation error. To decide the cutoff value for taking the decision more precisely, we have calculated the minimum error of our decision tree model.

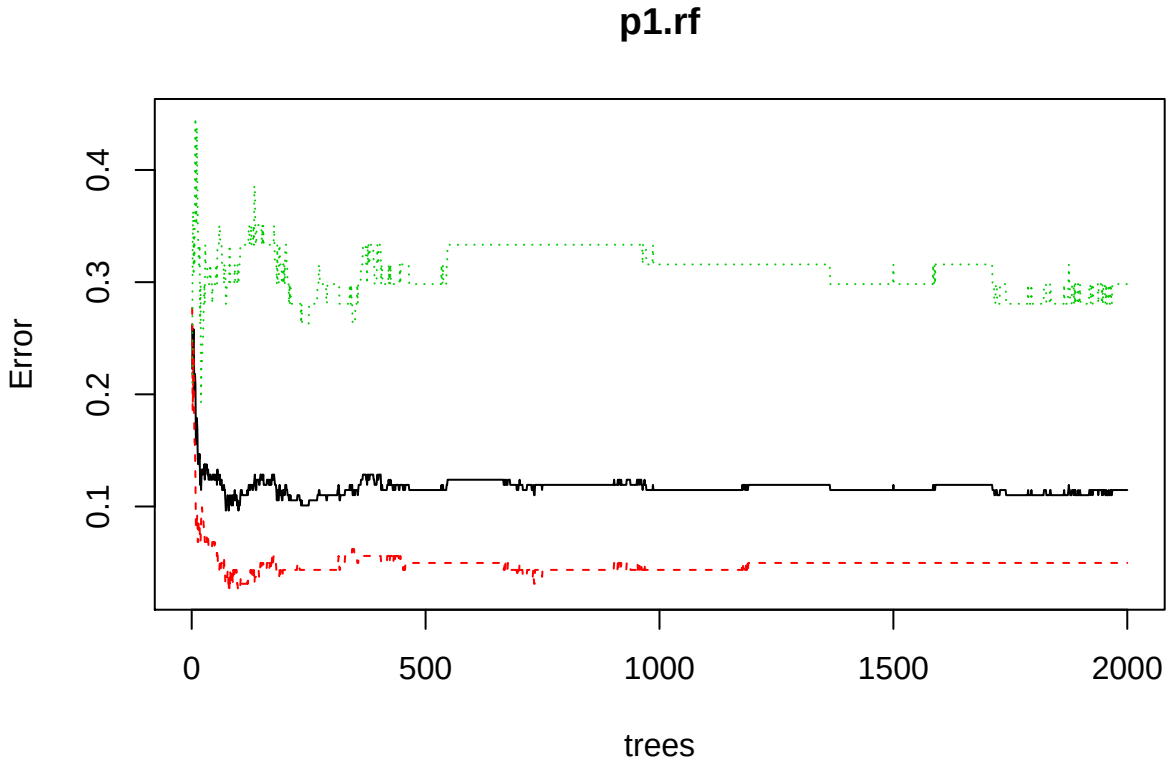



After the tree is prunned, we predict the test samples. The resulting accuraccy is 85.1851852%. We are surprised by the simplicity of the tree, only depending in one variable (among the 33 actual

ones) extracted from the training samples.

### 6.3 Random Forest

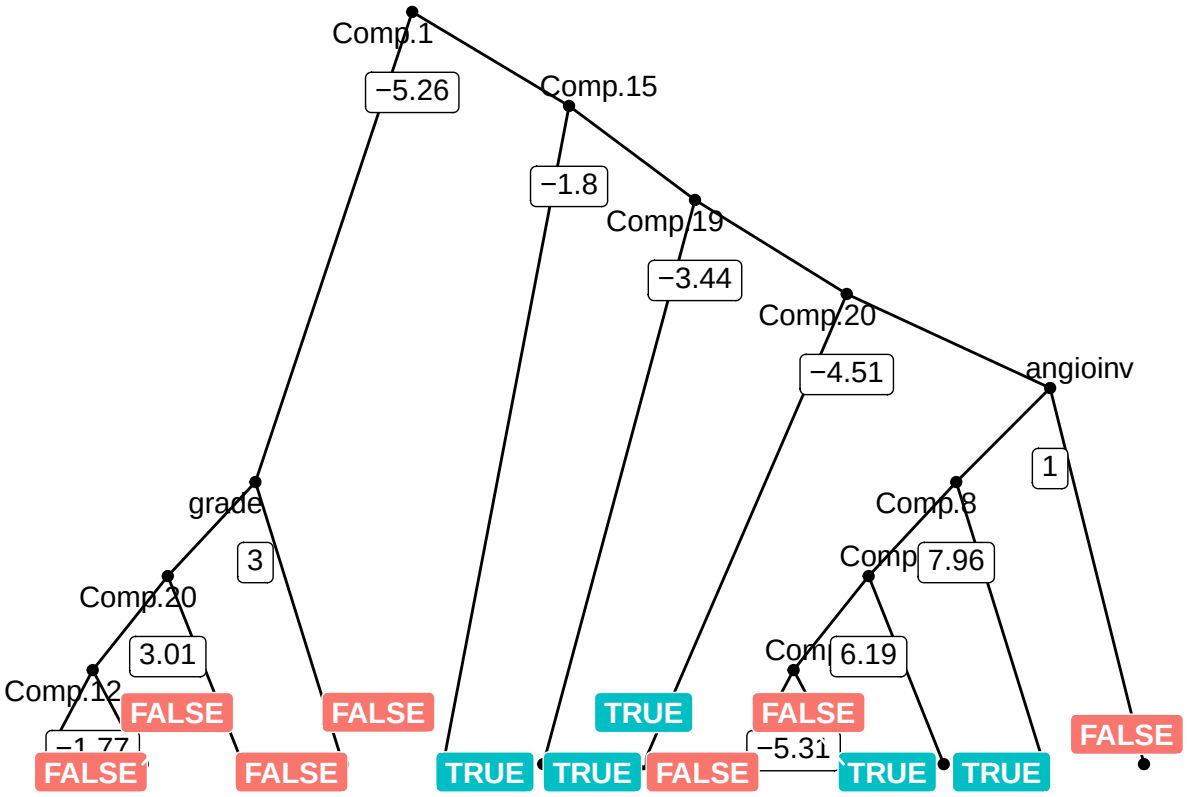
Finally, a random forest has been also developed.




In order to obtain the maximum possible accuracy, we have searched for the parameters that maximize the accuracy:

- ntree: Being the number of trees to create, this parameter is left with a high value (2000). The larger the value the maximum number of trees are created.
- nodesize: Minimum size of terminal nodes. Setting this number larger causes smaller trees to be grown (and thus take less time). The best value found is 20.
- mtry: Number of variables randomly sampled as candidates at each split. This parameter is left in default (since we do classification) being sqrt(#factors).

The maximum accuracy obtained for the the random forests is 83.3333333 which seems to be lower than the previous decision tree's 85.1851852 . However random forests have a higher computational time, no prunning to obtain the optimal solution is needed.




Taking a look at the tree that resulted the best test accuracy, we can observe that the complexity (in terms of tree size and depth) is quite higher compared to the decicion tree, which only a single timerecurrence decision is needed.

## 7 Conclusions

After finishing this project, the main conclusions that we would like to state are the following ones.

- We think that we have been able to perform a multiple variate analysis from a chosen dataset, by deeply analyzing and understanding the difficulties that the given data contained.
- We have realized the importance of performing a good preprocessing of the data to be able to obtain good results on the predictions.
- We have been able to apply a PCA to a dataset that contains more variables than individuals, realizing that the maximum number of dimensions is set to the number of variables of the dataset.
- We have been able to analyze and perform a dimensionality reduction from a dataset with a huge number of variables.
- Having mixed type data has brought us the oportunity to use Factor Analysis on Mixed type Data, this way, we have been able to compare categorical with continuous data.
- High difficulty of a cluster based spliting in hierarchical clustering has forbidden us a higher partition regarding the number of clusters. The quality of the cuts have been remained low despite applying consolidation clustering due to the number of clusters.

- It is needed to define a protocol of validation before using prediction models taking into account the length of individuals data. Houldout method should be used when it is possible.
- Training set should be representative of the whole data, meaning that all test set should not differ much from the observations calculated on the training set.
- Comparing the accuracy of the predictors, the linear regression is the one providing better validation accuracy.
- Even having a higher computational time and complexity, random forest has proven to be the worst predictor.

To conclude, we would like to add that we are glad to have been able to apply many different concepts learnt during the semester on a real dataset.