Instituto Superior Técnico - UL

# MECD - Multivariate Analysis

## Project Report


| *Authors:* | *IST ID:* |
|---|---|
| Catarina Costa | 86582 |
| Inês Tavares | 86593 |
| Catarina Oliveira | 86963 |
| Gonçalo Mestre | 87005 |

*Professora Rosário de Oliveira*


January 2020

# Contents

# 1 Introduction

## 1.1 Problem Description

Most organizations race for a reduction on expenses and an increase on productivity and customer satisfaction. In order to achieve that, the organizations should manage the factors that affect their performance. One factor that has a huge effect on an organization's performance is the human resources. Employees with low performance hinder the progress of a business, causing a vital loss for organizations. The absenteeism is considered to be one of the factors that compromise this progress.

Understanding the causes of absenteeism may help to empower an organization with the tools to gain maximum profit.

## 1.2 Objectives

With that being said, it was decided to study a dataset with records of absenteeism at work from July 2007 to July 2010 at a courier company in Brazil. This dataset was created by PhD students working on their thesis and can be found on [1]. The main goal of this project is to give an answer to: "How can we predict the time of absence of a worker, by knowing the time of the year, people's habits and other given information?".

Basically, the aim of this work is to discover the factors and causes of employees' absence and how these can affect the time of their absence. By finding which factors are more correlated with absenteeism, the organization gains a highly competitive advantage tool that could be used to address the consequences of the employees' absence and help to improve the process of recruitment and crisis management.

# 2 Data Preliminary Analysis

The chosen dataset contains 21 attributes, which are: *ID*, *Reason for absence*, *Month of absence*, *Day of the Week*, *Seasons*, *Transportation Expense*, *Distance from Residence to Work*, *Service Time*, *Age*, *Work load Average/day*, *Hit Target*, *Disciplinary failure*, *Education*, *Son*, *Social Drinker*, *Social Smoker*, *Pet*, *Weight*, *Height*, *Body Mass Index* and *Absenteeism time in hours*.

## 2.1 Initial Modifications

It was decided to initially remove the variables *Height*, *Weight*, *Body Mass Index* and *ID*, because they don't seem to be relevant for this study.

Some of the variables had too long names, so they were changed to be easier to display the plots regarding them. The variables *Distance from Residence to Work*, *Absenteeism time in hours*, *Month of absence* and *Day of the week* were renamed to *Work Distance*, *Absenteeism*, *Month* and *Week day*, respectively.

In order to avoid missing values, some categorical and quantitative variables initially assumed the value 0, being *Absenteeism* one of them. However, since this is the target variable, it does not make much sense having rows in which this variable's information is missing. Therefore, all the rows in which the variable *Absenteeism* takes the value 0 were removed. By doing this, all the missing values for the remaining variables were also eliminated. To avoid any computation problems, the values of the variable *Work load average/day* were rounded, so that they had no decimal places.

## 2.2 Variable's changes

The attribute *Reason for Absence* can assume one of 28 values according to what disease or other reason is representing (the explanation of each one of those values can be found in the ***Appendix***). To simplify the data analysis, this variable was divided in 8 classes, by joining all the possible diseases into category 1: Disease (1), Patient follow-up (2), Medical Consultation (3), Blood Donation (4), Laboratory Examination (5), Unjustified Absence (6), Physiotherapy (7) and Dental Consultation (8).

*Day of the Week* can take one of 5 values, where 2 refers to Monday, 3 to Tuesday and so on. Similarly, the variables *Seasons* and *Education* can take one of 4 values, for each season: Summer (1), Autumn (2), Winter (3) and Spring (4); and to level of education: high school (1), graduate (2), postgraduate (3), and master and doctor (4). There are also three binary variables: *Disciplinary failure*, *Social Drinker* and *Social Smoker* (yes=1 and no=0).

Given this, since these variables are not quantitative, as well as *Month of absence* (can take one of 12 values according to the month it represents: January (1),...,December(12)), they were converted into categorical or binary ones, using the command ***as.factor*** in *R*.

Although the dataset had the target attribute in quantitative value (hours of absenteeism), for a more comprehensive analysis of the results, it was decided that it was best that absenteeism hours were divided into classes: number of hours lower than 8 (1); number of hours equal to 8 (2); number of hours between 8 and 40 (3); number of hours greater than 40 (4).

After this, this variable was also turned into a categorical one.

## 2.3 Outliers

In order to observe the distribution of the numerical data and analyse the existence of potential outliers, the box plots of the quantitative variables were computed, using *R*, these can be found on Figure 1.

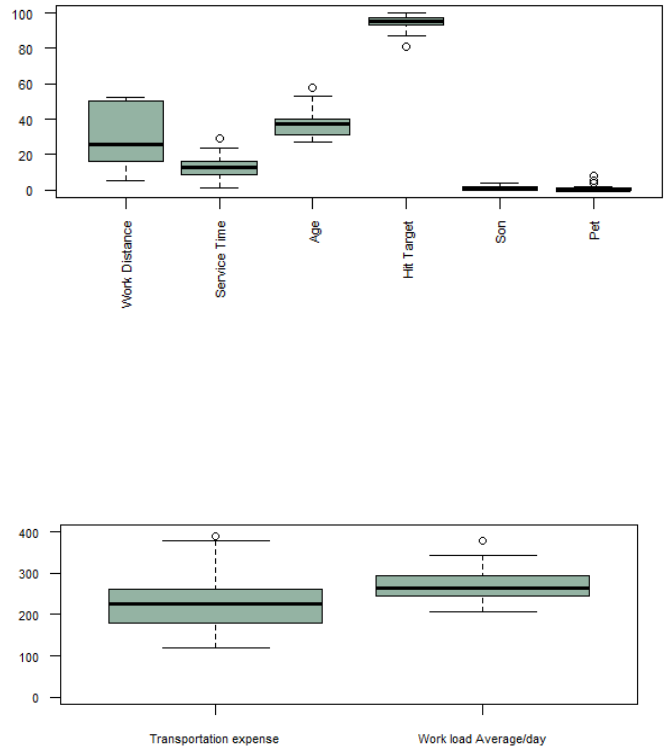


Figure 1: Boxplots of *Work Distance*, *Service Time*, *Age*, *Hit Target*, *Son*, *Pet*, *Work Load Average/day* and *Transportation Expense*

Notice that it was computed the boxplot of *Work Load Average/Day* and *Transportation Expense* in separate for reasons of scalability, since they took much higher values than the other variables. When observing the boxplots, it is concluded that most of them have outliers. It is also possible to see that, in most cases, the data is skewed. In case of a normal distribution, the mean and the median are approximately closer. This can lead one to think that most of the variables do not follow a normal distribution. In order to obtain more trustworthy results, the outliers obtained for each variable were identified, considering the values for each class of absenteeism in separate. The rows found with outliers for 3 or more variables were 207 and 209 (found in class where the number of hours of absenteeism is equal to 8). Therefore, these two rows were removed from the dataset.

## 2.4 Variables Correlation

The next step was to analyse the correlation between all the quantitative variables. For that, it was firstly computed a correlogram, which is shown on Figure 2.

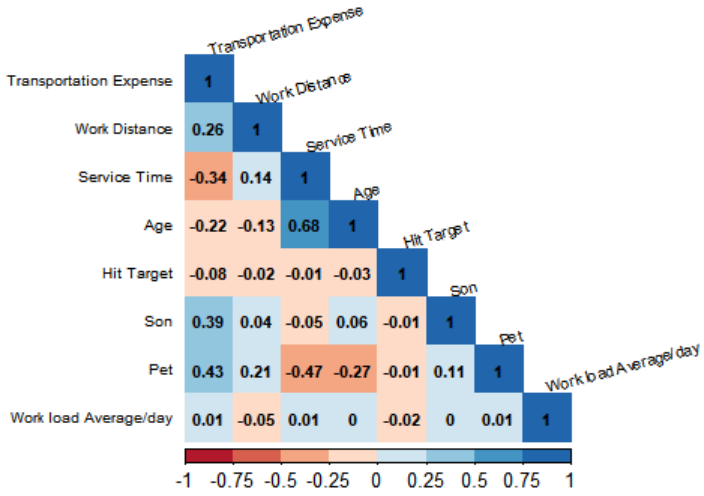


Figure 2: Correlation between the quantitative variables

By the observation of the correlogram, it is not possible to see any relevant correlation between the variables, except for *Age* and *Service Time*, which seem to have meaningful and positive correlation (0.68).

The Mutual Information (normalized) between all the quantitative variables was also computed, to find some dependence between the variables that was not evident when using the Pearson's correlation.

The heatmap in Figure 3 shows the Mutual Information values between all the variables, including the categorical and the binary ones. The values greater than 0.6 appear in grey, as the mutual information between that pair of variables is already considerably high.

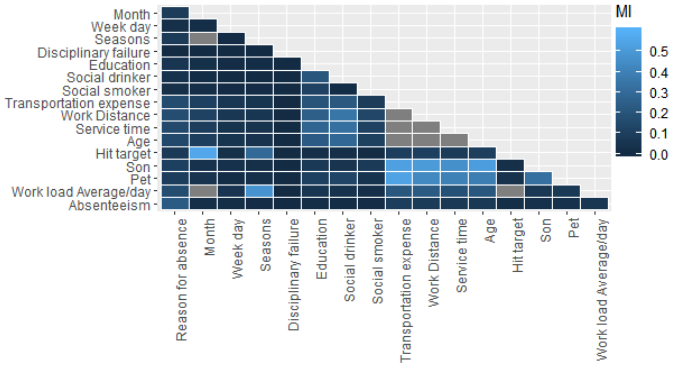


Figure 3: Mutual Information's heatmap

A value for mutual information greater than 0.8 was obtained for the following pairs of variables: *Transportation Expense* and *Work Distance* (0.854); *Transportation Expense* and *Age* (0.835); *Service Time* and *Work Distance* (0.867); *Age* and *Work Distance* (0.870); *Age* and *Service Time* (0.830). A high value of 0.789 was also obtained for the variables *Transportation Expense* and *Service Time*.

As it can be noticed, there is an evident dependence between the variables *Age*, *Work Distance*, *Service Time* and *Transportation Expense*, since the mutual information between *Age* and the remaining three variables is so

high, as well as the mutual information between them.

Since Pearson's correlation values obtained for the variables *Age* and *Service Time* was also considerable high (0.68), it seems reasonable to remove the variable *Service Time*. It was also decided to remove the variables *Transportation Expense* and *Work Distance*, since they are also dependent from each other and from a third variable, *Age*, which remains in the dataset.

The bar chart in Figure 4 shows the proportion of each absenteeism class by *Age*.

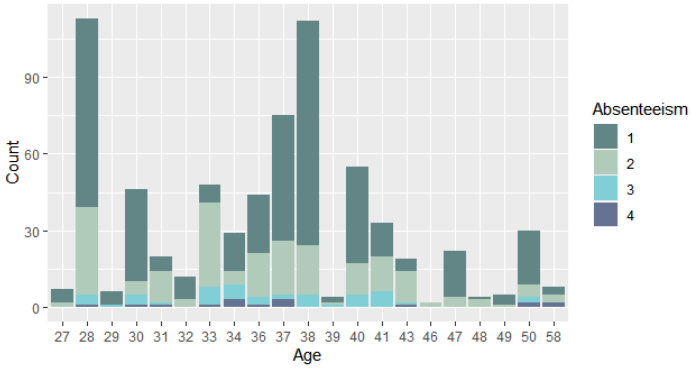


Figure 4: Proportion of absenteeism classes by age

As it can be seen, workers with 28 and 38 years old are the ones with more absences at work (mostly less than 8 hours).

After this, it was used the Shapiro-Wilks test to confirm the non normality of the variables. All the p-values obtained were lower than $10^{-14}$, which is much lower than 0.05 (significance level). Therefore, comes the conclusion that the samples deviate from normality, as it was expected.

Regarding the categorical and binary variables, was used the command ***table*** in *R*, to see the number of occurrences of each category.

By observing the results obtained for the variable *Reason for absence*, it is possible to see that the most commonly used reasons for absence are having a certain Disease (38%), Medical Consultation (21%) and Dental Consultation (16%).

In case of *Month*, *Week Day* and *Seasons*, the number of observations for each category is very similar.

By analysing the occurrences for the variable *Education*, it is possible to see that 82% of the work absences are from people with only high school level of education. All the remaining categories have similar values. The bar chart regarding the proportion of each level of education by class of *Absenteeism* is shown in Figure 5.

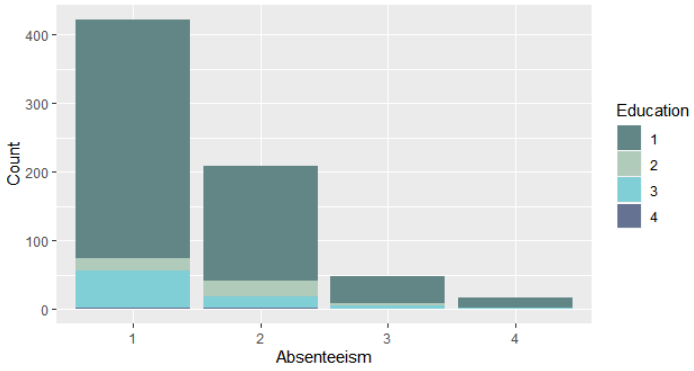


Figure 5: Proportion of each level of Education by Absenteeism's class

While the results obtained for the variable *Social Drinker* are not very relevant, because the proportion of absences from people who regularly drink alcohol and don't drink are similar, the results obtained for *Social Smoker* show that 93% of the work absences are from people that do not smoke. A bar chart considering the proportion of workers that are social drinkers, social smokers, both of them or none of them by class of absenteeism was computed, and is shown in Figure 6.

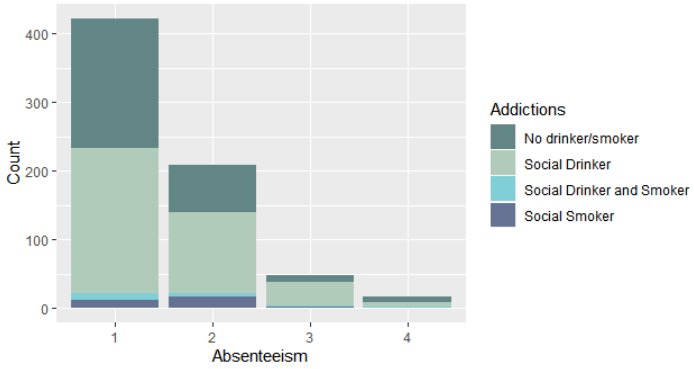


Figure 6: Proportion of social smokers/drinkers and non drinkers and smokers by Absenteeism's class

It can be easily observed in Figure 6 that a considerably high proportion of absences is from non smokers and non drinkers (39%).

Considering the target variable, *Absenteeism*, it is possible to notice that the absenteeism time is mostly lower than 8 hours (61%), or equal to 8 hours (30%), which corresponds to one day of work.

For *Disciplinary Failure*, it was obtained that 100% of the work absences are from people without disciplinary failures. So, although this variable was taken into consideration, it was found it had no obvious part on the target variable. Therefore, it was removed from the dataset.

The outputs obtained are displayed in the ***Appendix***.

The values of mutual information regarding the categorical and binary variables (also shown in Figure 3), suggest that there aren't any relevant dependencies be-

tween them or even between them and the quantitative variables, except in the case of *Work load Average/day* and *Month* and *Work load Average/day* and *Hit Target*. However, even these values are lower than 0.80. Therefore, all the remaining categorical and binary variables were kept.

After this preliminary analysis, the dataset contains 13 variables, which are *Reason for Absence*, *Week Day*, *Month*, *Seasons*, *Education*, *Social Drinker*, *Social Smoker*, *Age*, *Hit Target*, *Son*, *Pet*, *Work load average/day* and *Absenteeism* (target). It initially had 740 rows, which were reduced to 694.

# 3 Data Classification

For the classification, the goal is to predict the classes of observations and in this case the values of *Absenteeism* correspond to the classes.

This part starts with dimensionality reduction, using Principal Component Analysis (PCA) for the quantitative variables and Multidimentional Scaling (MDS)[2] for the categorical ones.

## 3.1 Dimensionality reduction

In PCA, was used standardized variables because some of them have small values comparing to others and all should have the same importance, for example the number of *Pet* or *Son* take small values when compared with *Hit target* or *Work load average/day*, but this variables are also important and must be relevant for this analysis. The function used in *R* was *prcomp*. The obtained results are presented in Figure 7.

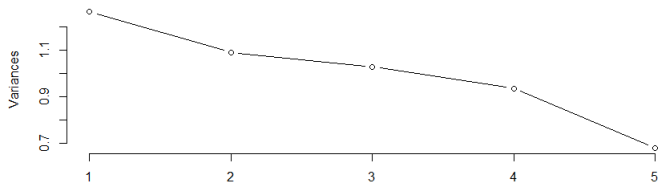


Figure 7: Relation between variance and principal components

The threshold used to define how many principal components should be kept was $\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{71} \lambda_j} \geq 80\%$, where $k$ is the small number of principal components that satisfies the condition and $\lambda_i$ is the eigenvalue for principal component $i$.

In this case, 4 principal components were kept and they explain 86.37% of the variability of the dataset.

| | PC1 | PC2 | PC3 | PC4 |
|---|---|---|---|---|
| Age | -0.67677819 | 0.13953744 | -0.3482909 | -0.07943722 |
| Hit target | 0.05996430 | -0.59364322 | -0.3420561 | 0.71932955 |
| Son | 0.14552560 | 0.49310069 | -0.7699853 | 0.07888582 |
| Pet | 0.71883886 | 0.06247886 | -0.1549568 | -0.17075880 |
| Work load Average/day | 0.02164617 | 0.61730473 | 0.3805287 | 0.66398283 |

Figure 8: Explanation of principal components

First principal component explains most of the variability and it gives more importance to *Age* and *Pet* comparing to the other variables. Additionally, *Age* has a contrast with the remaining quantitative variables.

In MDS, categorical variables were transformed into binary, using the one hot encoding technique [3] (function *onehot*), so the distance between all variables could be measured in the same way, with *Hamming distance* [4]. This distance is very useful for binary variables. The function used to calculate this distance was *hamming.distance*.

Then, MDS was applied, using function *cmdscale*, and the obtained output is displayed in Figure 9.

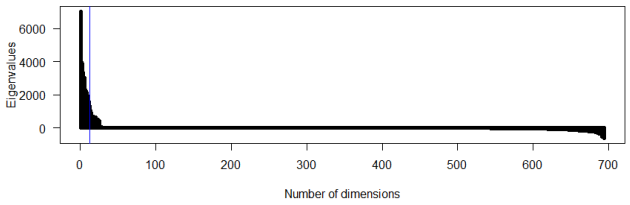


Figure 9: Relation between number of dimensions and eigenvalues in MDS. Blue line corresponds to $x = 12$, number of dimensions kept

The *Goodness of fit* corresponds to 0.805, when 12 dimensions are used. More dimensions add small information so they are not necessary.

The dataset used for classification was composed with the scores of the first 4 principal components, the scores of the first 12 dimensions of multidimensional scaling and *Absenteeism*. Although the number of dimensions was smaller in the beginning, the complexity of the information now is not higher, because it was used MDS after one hot encoding, which means that the complexity of each variable was reduced.

## 3.2 Cross Validation Technique

The dataset was split into train (70%) and test (30%).

```
> table(scores$class)
  1   2   3   4
425 206  48  15
```

```
> table(train$class)
  1   2   3   4
298 142  35  11
```

Figure 10: Number of observations per class in all dataset (left) and in training data (right)

*k-fold* cross validation [5] was the technique used in this study, because it calculates the average of the $k$ recorded errors and tries to find the best model for different parts of the training data, so the values obtained

are more accurate. This was applied only to training data, using functions *trainControl* - for *k-fold* cross validation technique - and *train* - to train the model with the training data, choosing a certain method. After the best model is created, testing data was used to evaluate the performance of the model, using function *predict*. This strategy was considered because the model should make tests on data it has never seen before. The need to test on never seen data comes from the fact that when a classifier is trained it always gets an amount(should be kept small) of bias regarding the training data.

There were used some supervised learning methods for classification, such as linear, nonlinear and ensemble algorithms.

In order to compare models, it was considered accuracy and F1-score values and analysed the confusion matrix. The accuracy (output of the function *confusionMatrix*) gives the proportion of correct classifications, but since this dataset is imbalance, this is not a good measure when used alone as explained on [6]. Through the confusion matrix, it is possible to understand how the predictor works in each class. And finally, F1-score is the weighted average of precision and recall, which is a good parameter for this study. The last one was calculated using function *F1_Score*.

## 3.3 Supervised Learning methods applied

For the **linear algorithms**, was used Fisher's Discriminant Analysis or Linear Discriminant Analysis (LDA) [7] which is a technique that assumes that the predictor variables are normally distributed and have identical covariance matrices. This algorithm was considered because it aims to maximize the distance between the mean of each class and minimize the spreading within the class itself. One of the inputs for function *train* was *method="lda"*.

| | Reference | | | |
|---|---|---|---|---|
| Prediction | 1 | 2 | 3 | 4 |
| 1 | 102 | 30 | 3 | 0 |
| 2 | 23 | 31 | 9 | 4 |
| 3 | 1 | 3 | 1 | 0 |
| 4 | 1 | 0 | 0 | 0 |

| | Class: 1 | Class: 2 | Class: 3 | Class: 4 |
|---|---|---|---|---|
| Sensitivity | 0.8031 | 0.4844 | 0.076923 | 0.000000 |
| Specificity | 0.5926 | 0.7500 | 0.979487 | 0.995098 |
| Pos Pred Value | 0.7556 | 0.4627 | 0.200000 | 0.000000 |
| Neg Pred Value | 0.6575 | 0.7660 | 0.940887 | 0.980676 |
| Prevalence | 0.6106 | 0.3077 | 0.062500 | 0.019231 |
| Detection Rate | 0.4904 | 0.1490 | 0.004808 | 0.000000 |
| Detection Prevalence | 0.6490 | 0.3221 | 0.024038 | 0.004808 |
| Balanced Accuracy | 0.6979 | 0.6172 | 0.528205 | 0.497549 |

Figure 11: Confusion matrix and Statistics by class using LDA

By observing Figure 22, it can be concluded that the classifier does not identify correctly any observation of class 4 (lowest probability) and in the most cases it classifies an observation as class 1 (highest value of prevalence). Balanced Accuracy has good values for classes 1 and 2 and not so good ones for the classes with lower prior probability.

For the **nonlinear algorithms**, Quadratic Discriminant Analysis (QDA) [7] was considered because there is no assumption that the covariance matrix of classes is the same, but it did not work because some classes have few observations ($n_k \leq p$, where $n_k$ is the number of observations in class $k$ and $p$ is the number of features).

Also Regularized Discriminant Analysis (RDA) [7] was implemented because it can lead to an improvement of the discriminant analysis, it improves misclassication risk when the number of observations per class is lower than the number of features, which happens with this dataset. So this method can be an alternative to QDA. It was used *method="rda"*.

| | Reference | | | |
|---|---|---|---|---|
| Prediction | 1 | 2 | 3 | 4 |
| 1 | 100 | 31 | 3 | 0 |
| 2 | 27 | 31 | 9 | 4 |
| 3 | 0 | 2 | 1 | 0 |
| 4 | 0 | 0 | 0 | 0 |

| | Class: 1 | Class: 2 | Class: 3 | Class: 4 |
|---|---|---|---|---|
| Sensitivity | 0.7874 | 0.4844 | 0.076923 | 0.00000 |
| Specificity | 0.5802 | 0.7222 | 0.989744 | 1.00000 |
| Pos Pred Value | 0.7463 | 0.4366 | 0.333333 | NaN |
| Neg Pred Value | 0.6351 | 0.7591 | 0.941463 | 0.98077 |
| Prevalence | 0.6106 | 0.3077 | 0.062500 | 0.01923 |
| Detection Rate | 0.4808 | 0.1490 | 0.004808 | 0.00000 |
| Detection Prevalence | 0.6442 | 0.3413 | 0.014423 | 0.00000 |
| Balanced Accuracy | 0.6838 | 0.6033 | 0.533333 | 0.50000 |

Figure 12: Confusion matrix and Statistics by class using RDA

The obtained results are very similar with the previous one, but in this case the model does not classify any point into class 4. So, although the accuracy values are the same, F1-score is lower in this case.

Support Vector Machine [8] finds a hyperplane in N-dimensional space that separates the different classes, with the maximum distance between data points from different classes. Since it can be non-linear, it can separate different classes in a more embracing way. Soft-margin technique was used so the algorithm could make some mistakes in classification and keep margin as wide as possible to allow that other points could still be correctly classified. It was chosen radial kernel because it was not possible to represent the data points, there is no prior knowledge about their distribution [9]. This method was considered because it creates classification areas, which can be used to classify new data points in a more generic way. The input used was *method="svmRadial"*.

| | Reference | | | |
|---|---|---|---|---|
| Prediction | 1 | 2 | 3 | 4 |
| 1 | 105 | 32 | 3 | 0 |
| 2 | 22 | 32 | 10 | 4 |
| 3 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 |

| | Class: 1 | Class: 2 | Class: 3 | Class: 4 |
|---|---|---|---|---|
| Sensitivity | 0.8268 | 0.5000 | 0.0000 | 0.00000 |
| Specificity | 0.5679 | 0.7500 | 1.0000 | 1.00000 |
| Pos Pred Value | 0.7500 | 0.4706 | NaN | NaN |
| Neg Pred Value | 0.6765 | 0.7714 | 0.9375 | 0.98077 |
| Prevalence | 0.6106 | 0.3077 | 0.0625 | 0.01923 |
| Detection Rate | 0.5048 | 0.1538 | 0.0000 | 0.00000 |
| Detection Prevalence | 0.6731 | 0.3269 | 0.0000 | 0.00000 |
| Balanced Accuracy | 0.6973 | 0.6250 | 0.5000 | 0.50000 |

Figure 13: Confusion matrix and Statistics by class using SVM

This method does not classify any observation into classes 3 and 4, the ones with lowest probabilities, and that is not what is intended, so this is not a good algorithm is this case, even the F1-score is high.

Naive Bayes Classifier [10] usually is a good method that assumes that numerical variables have normal distribution. It also assumes that all features are independent (strong assumption). It returns the probability of each observation belongs to a certain class, using Naives Bayes Theorem, and in the end it classifies the observation into the class with higher probability. It was used *method="nb"*.

```
          Reference
Prediction   1   2   3   4
         1 107  20   4   1
         2  17  28   4   2
         3   1  14   5   1
         4   2   2   0   0
```

```
                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity            0.8425   0.4375  0.38462  0.00000
Specificity            0.6914   0.8403  0.91795  0.98039
Pos Pred Value         0.8106   0.5490  0.23810  0.00000
Neg Pred Value         0.7368   0.7707  0.95722  0.98039
Prevalence             0.6106   0.3077  0.06250  0.01923
Detection Rate         0.5144   0.1346  0.02404  0.00000
Detection Prevalence   0.6346   0.2452  0.10096  0.01923
Balanced Accuracy      0.7669   0.6389  0.65128  0.49020
```

Figure 14: Confusion matrix and Statistics by class using Naive Baeyes Classifier

Although the strong assumption this method has, that is highly unlikely to be satisfied with a high number of variables, it works fine with this dataset. It does not classify correctly any observation from class 4, but the values of Balanced Accuracy are good.

Neural Networks [11] learn to recognize correlations between certain relevant features and optimal results, creating connections between feature signals and what those features represent. This algorithm was chosen because of its efficiency with variables that may seem not correlated. The input for function *train* was *method="nnet"*.

```
          Reference
Prediction   1   2   3   4
         1 101  27   3   0
         2  24  28   9   3
         3   2   6   1   0
         4   0   3   0   1
```

```
                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity            0.7953   0.4375 0.076923 0.250000
Specificity            0.6296   0.7500 0.958974 0.985294
Pos Pred Value         0.7710   0.4375 0.111111 0.250000
Neg Pred Value         0.6623   0.7500 0.939698 0.985294
Prevalence             0.6106   0.3077 0.062500 0.019231
Detection Rate         0.4856   0.1346 0.004808 0.004808
Detection Prevalence   0.6298   0.3077 0.043269 0.019231
Balanced Accuracy      0.7125   0.5938 0.517949 0.617647
```

Figure 15: Confusion matrix and Statistics by class using Neural Networks

This algorithm works fine with this dataset, it caught one observation from class 4, which is as improvement considering the previous methods. The accuracy and F1-scores are good.

K-Nearest Neighbour assumes that similar observations exist in close proximity, that means closer observations are more similar than distant ones. This was chosen because of its versatility and simplicity. In order to find the best $k$ for the model, it was tested with different values of $k$ to find the one with highest accuracy. The results obtained using *method="knn"* can be found in Figure 16.

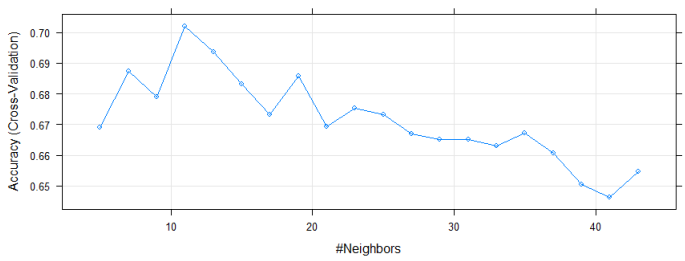


Figure 16: Relation between $k$ values and accuracy of the respectively model

This algorithm was implemented considering $k = 11$.

```
          Reference
Prediction   1   2   3   4
         1 113  39   5   0
         2  14  20   7   3
         3   0   5   1   1
         4   0   0   0   0
```

```
                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity            0.8898  0.31250 0.076923  0.00000
Specificity            0.4568  0.83333 0.969231  1.00000
Pos Pred Value         0.7197  0.45455 0.142857      NaN
Neg Pred Value         0.7255  0.73171 0.940299  0.98077
Prevalence             0.6106  0.30769 0.062500  0.01923
Detection Rate         0.5433  0.09615 0.004808  0.00000
Detection Prevalence   0.7548  0.21154 0.033654  0.00000
Balanced Accuracy      0.6733  0.57292 0.523077  0.50000
```

Figure 17: Confusion matrix and Statistics by class using KNN

This method does not classify correctly any observation of class 4, in fact it does not classify any observation into this class. Class 2 has worst values and consequently, class 1 also has them, because more than half of the observations from class 2 are classified as class 1.

And, finally, for the **ensemble algorithms**, was used Random Forest [12], which is a set of decision trees that operate together, each individual tree in the random forest returns a class and the one that has more votes becomes the final prediction. This method was used because it avoids overfitting, using different trees to classify an observation. The function used was *rf*.

```
          Reference
Prediction   1   2   3   4
         1 111  29   6   0
         2  16  29   4   2
         3   0   3   3   1
         4   0   3   0   1
```

```
                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity            0.8740   0.4531  0.23077  0.250000
Specificity            0.5679   0.8472  0.97949  0.985294
Pos Pred Value         0.7603   0.5686  0.42857  0.250000
Neg Pred Value         0.7419   0.7771  0.95025  0.985294
Prevalence             0.6106   0.3077  0.06250  0.019231
Detection Rate         0.5337   0.1394  0.01442  0.004808
Detection Prevalence   0.7019   0.2452  0.03365  0.019231
Balanced Accuracy      0.7210   0.6502  0.60513  0.617647
```

Figure 18: Confusion matrix and Statistics by class using Random Forest

This method chooses all classes and it detects observations of class 1, 2 and 3 better than Neural Networks classifier, the overall accuracy is also better and the F1-score is similar.

The accuracy and F1-score values can be found in Table 3.

| | **Accuracy** | **F1-score** |
|---|---|---|
| **LDA** | 0.6442 | 0.7786 |
| **RDA** | 0.6442 | 0.7663 |
| **SVM** | 0.6587 | 0.7865 |
| **Naive Bayes** | 0.6731 | 0.8263 |
| **Neural Networks** | 0.6587 | 0.7953 |
| **Random Forest** | 0.6923 | 0.8132 |
| **KNN** | 0.6442 | 0.7958 |

Table 1: Accuracy and F1-score of each method

The best algorithm for this dataset, considering accuracy, F1-score values and the respective confusion matrix, is the Random Forest, because this method works better with imbalanced datasets, which is this case, because the decision trees that compose the Random Forest work by learning a hierarchy of if/else questions and this can force all classes to be addressed, instead of only addressing the most common classes.

# 4 Data Clustering

To perform the cluster analysis of this dataset it was used the pre-processed data and then were tried two different approaches.

The first approach using the pre-processed data without changing anything else, using the gower distance metric and different clustering methods, such as hierarchical, partitioning and density based methods.

The second approach was done by doing multidimensional scaling on the categorical variables and by doing a principal component analysis on the quantitative variables. After this it was used the euclidean distance metric and the clustering methods used in the first approach to cluster the data.

## 4.1 First approach - Gower Distance Metric

On this approach it was used the Gower distance metric as it is explained on the course slides [13]. Even though this is not the most common distance metric used for cluster analysis it was chosen because it is the choice when handling mixed type data.

This distance choice allows to handle mixed type data, by returning for a quantitative feature the partial dissimilarity as the ratio between the absolute difference of observations $i$ and $j$ and the maximum range observed from all observations and returning for a qualitative feature the partial dissimilarity as only if observations $i$ and $j$ have the same value, zero otherwhise. Further details on this and on how to handle mixed type data for cluster analysis can be found on [14].

After getting the dissimilarity matrix there were performed the hierarchical clustering methods, in which for all of them was first used the function *NbClust* from the package with the same name. This function was used to get the number of clusters for the corresponding algorithm that has the best silhouette width and the corresponding width.

After this, it was used *agnes* to compute the agglomerative hierarchical clustering for each of the four hierarchical methods used, *Single Linkage*, *Complete Linkage*, *Average Linkage* and *Ward's Method*. A further detailed explanation of the four methods can be found on [13].

For each method it was made the plot of the dendogram and then the partitions for that method were obtained by cutting the tree according to the output of *NbClust* and to the visualized dendogram.

After the hierarchical clustering methods were used it was applied a partitioning method, *Partitioning Around Medoids*, also explained on [13]. The choice of this method came from the will to try a different type of method which worked way different from hierarchical clustering and even though *kmeans* is a far more popular method, this one was chosen because it allows the dissimilarity matrix to be used, being more robust to outliers and noise.

As a last method on this approach it was used a density based method, *DBSCAN*, as recommended by the teacher, so another different way of thinking could be applied to the data, since only partitioning and hierarchical methods were being applied and the results were not that good. Further details on this method can be found on [15].

## 4.2 Second Approach - Dimensionality Reduction and Euclidean Distance Metric

On this second approach, first there were used dimensionality reduction techniques on the data and only then computed the dissimilarity matrix. The dimensionality reduction techniques were applied in the same way as in the *Data Classification*, so the same explanation used in Section 3 can be applied here.

After this it was used the euclidean distance to compute the dissimilarity matrix, this distance as the gower one, can be found on [13]. This was the chosen metric due to it's simplicity. It was made a choice to not standardize the reduced dimensionality data, because if there are predominant components that came from the *multidimensional scaling* or from the *principal component analysis*, they should remain with that dominance.

With the dissimilarity matrix computed there were applied the same methods to this matrix, as in the first approach.

Using just the data that resulted from the dimensionality reduction it was also applied model based clustering with the function *Mclust*. The models were estimated by the EM algorithm, and the optimal model is then selected according to the BIC criterion. The EM algorithm explanation can also be found on [13].

## 4.3 Results

To evaluate the quality of the obtained partitions, it was used the silhouette coefficient which is explained on [13].

When doing a preliminary analysis on the obtained results, both the partitions obtained with the *Single Linkage* result were excluded, since the result obtained for both is a really big cluster with almost all data and a really small one with just 1 or 2 observations. This is something expected, since this algorithm is really sensitive to outliers. Another partition removed is the one obtained with the *DBSCAN* algorithm on the second approach, since it is composed of one cluster containing the majority of the data and 58 clusters in which none of those has more than 4 elements, being this result pretty useless.

For the remaining partitions it was computed the Average Silhouette Width which can be found on Table 2.

| Method | Average Silhouette |
|---|---|
| **Gower-Complete** | 0.1530 |
| **Gower-Average** | 0.2342 |
| **Gower-Ward** | 0.1350 |
| **Gower-PAM** | 0.1381 |
| **Gower-DBSCAN** | -0.0613 |
| **Euclidean-Complete** | 0.1504 |
| **Euclidean-Average** | 0.2574 |
| **Euclidean-Ward** | 0.2038 |
| **Euclidean-PAM** | 0.1372 |
| **Euclidean-Mixture Models** | 0.1145 |

Table 2: Average Silhouette for each method

As it is possible to see there is only one method that results in an average silhouette width over 0.25. Having in mind that this partition is not the best result possible it was tried a different method which had not really good results.

This method starts by choosing a relevant cluster from one of the obtained partitions. This cluster needs to have a good average silhouette width. After it will be repeated the cluster analysis for the rest of the data. With data the same analysis is done, and this can be continued until there are only partitions with good values for the average silhouette width. Figure 19 illustrates this.

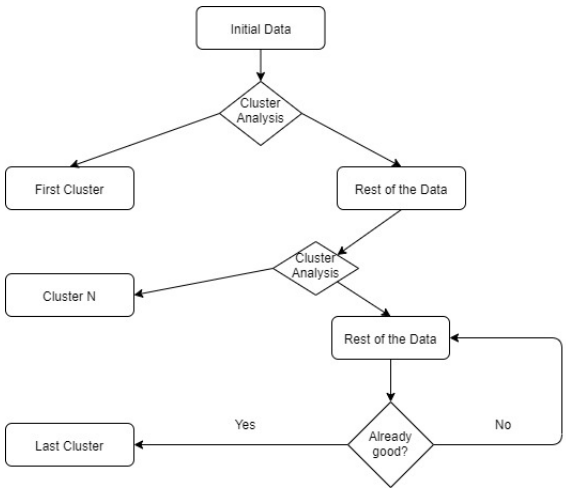


Figure 19: Fluxogram for the different method tried

With this method the results didn't have much better values neither did the partitions seemed more relevant, so it was chosen as the obtained partition the one produced with the Average Linkage algorithm for the second approach.

Analyzing the obtained partition and the summaries of each cluster shown on Figures 20 and 21 it is possible to verify that even though cluster 1 seems pretty general, cluster 2 has a specific set of characteristics.

```
Reason for absence     Month          Week day Seasons Education Social drinker Social smoker      Age
1      :247        3      : 78   2:146    1:158   1:563     0:279          0:648          Min.   :27.0
3      :147        2      : 71   3:136    2:179   2: 23     1:389          1: 20          1st Qu.:30.0
8      :111        7      : 64   4:142    3:161   3: 78                                   Median :37.0
7      : 68        10     : 62   5:114    4:170   4:  4                                   Mean   :36.1
2      : 34        11     : 54   6:130                                                    3rd Qu.:38.0
6      : 33        5      : 52                                                            Max.   :58.0
(Other): 28        (Other):287
  Hit target         Son              Pet             Work load Average/day  Absenteeism
Min.   : 81.00   Min.   :0.0000   Min.   :0.0000   Min.   :206.0          1:413
1st Qu.: 93.00   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:243.2          2:193
Median : 95.00   Median :1.0000   Median :0.0000   Median :264.0          3: 47
Mean   : 94.69   Mean   :0.9671   Mean   :0.7051   Mean   :270.9          4: 15
3rd Qu.: 97.00   3rd Qu.:2.0000   3rd Qu.:1.0000   3rd Qu.:285.0
Max.   :100.00   Max.   :4.0000   Max.   :8.0000   Max.   :379.0
```

Figure 20: Summary of the first cluster

```
Reason for absence     Month          Week day Seasons Education Social drinker Social smoker      Age
1      :15         3      :5    2:7      1: 6    1: 6      0:26           0: 0           Min.   :36.00
5      : 6         5      :5    3:5      2:10    2:20      1: 0           1:26           1st Qu.:40.00
2      : 2         8      :5    4:3      3: 6    3: 0                                   Median :40.00
3      : 2         1      :4    5:4      4: 4    4: 0                                   Mean   :40.92
8      : 1         11     :3    6:7                                                     3rd Qu.:40.00
4      : 0         2      :1                                                            Max.   :48.00
(Other): 0         (Other):3
  Hit target         Son              Pet             Work load Average/day  Absenteeism
Min.   :91.0     Min.   :1.000    Min.   :0.000    Min.   :206.0          1:12
1st Qu.:93.0     1st Qu.:2.000    1st Qu.:0.000    1st Qu.:239.5          2:13
Median :96.0     Median :2.000    Median :0.000    Median :250.0          3: 1
Mean   :95.5     Mean   :1.846    Mean   :1.077    Mean   :269.8          4: 0
3rd Qu.:98.0     3rd Qu.:2.000    3rd Qu.:0.000    3rd Qu.:312.0
Max.   :99.0     Max.   :2.000    Max.   :5.000    Max.   :343.0
```

Figure 21: Summary of the second cluster

Analysing the summary of cluster two it is possible to check that this cluster has the group of all entries in which the worker that missed is simultaneously "Social Smoker", but not "Social Drinker".

Another characteristic of this cluster is that the majority of the entries are from workers that have 40 years and that the range of ages is much more narrow than the original one, having entries only from workers that are aged between 36 and 48 years.

## 4.4 Comments on the Results an Recommendations

As it is possible to see from the results, these are not the best, the average silhouette width values are just in the limiar of the acceptable and the clusters are not the most relevant for the study in question. Even though that the separation identifies a clear type of person, it doesn't help identifying the number of hours of work the person will be absent for.

One motive that made this analysis not that good may be a common problem in machine learning, known as "the curse of high dimensionality". This problem refers to the fact that one higher dimensionality spaces, common distance metrics are not as a good and also that when one has a lot of dimensions, a lot of observations are needed too to cover a relevant amount of the input space, as explained on [16]. On this data set for the cluster analysis, there were 12 variables, many of them categorical, and only 694 observations.

Another thing that could help on this analysis would be increasing the number of observations, to do this there could be used methods of oversampling. This methods could help dealing with the problem of high dimensionality, but also could help dealing with the fact that this dataset is clearly imbalanced, which makes the clustering task harder.

As a last recommendation, as this dataset represents people behaviour(work absenteeism), there could also be

tried other more advanced forms of cluster analysis, like fuzzy clustering, in which a point can belong to one or more clusters at the same time having their membership in a particular cluster corresponding to some probability. More information on fuzzy clustering can be found on [17].

# 5 Data Classification, into the obtained Clusters

It was asked to repeat the data analysis (supervised methods) but now using the clustering results as the new real values for the response variable. With that being said, it was considered the clustering technique that got the best results and used their classification as the supposed real values for the diagnosis variable. At the first instance, it would be expected that the results would get much better given that the new observations are divided in accordance with their similarities (similar entries imply smaller distance which means a bigger probability of being in the same cluster).

## 5.1 Supervised Methods

It was performed the same analysis with cross-validation as the one made in section 3 and the final results obtained are summarized in the table shown bellow:

| | Accuracy | F1-score |
|---|---|---|
| **LDA** | 0.976 | 0.9874 |
| **RDA** | 1 | 1 |
| **SVM** | 1 | 1 |
| **Naive Bayes** | 1 | 1 |
| **Neural Networks** | 0.9904 | 0.995 |
| **Random Forest** | 1 | 1 |
| **KNN** | 0.9808 | 0.9901 |

Table 3: Accuracy and F1-score of each method

When observing the table, it can be conclude that opposed to what happened in the prior analysis now there are much better results, RDA, SVM, Naive Bayes and Random Forest present the next confusion matrix:

```
          Reference
Prediction   1   2
         1 201   0
         2   0   7
```

Figure 22: Confusion matrix for classification methods with accuracy and f1-score equal to 1

This, alongside the previous presented scores shows that these four classification methods accurately predicted the class of all observations that belong to test data.

## 5.2 Using different *classes*

The best classification method obtained in Section 3 was Random Forest, and it was also one of the best of section 5.1. With that given and since this method is one of the most recommended for imbalanced data sets (which is also the case on this analysis), the results of Random Forest method are the ones being presented in more detail in Figures 23 and 24.

```
                Accuracy : 0.6827
                  95% CI : (0.6148, 0.7453)
    No Information Rate : 0.6106
    P-Value [Acc > NIR] : 0.01865

                   Kappa : 0.37

 Mcnemar's Test P-Value : 0.13630

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity            0.8425   0.4844  0.23077 0.250000
Specificity            0.5926   0.8264  0.97949 0.980392
Pos Pred Value         0.7643   0.5536  0.42857 0.200000
Neg Pred Value         0.7059   0.7829  0.95025 0.985222
Prevalence             0.6106   0.3077  0.06250 0.019231
Detection Rate         0.5144   0.1490  0.01442 0.004808
Detection Prevalence   0.6731   0.2692  0.03365 0.024038
Balanced Accuracy      0.7176   0.6554  0.60513 0.615196
```

Figure 23: Overall statistics for RF using *Absenteeism* variable

```
                Accuracy : 1
                  95% CI : (0.9824, 1)
    No Information Rate : 0.9663
    P-Value [Acc > NIR] : 0.0008084

                   Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.9663
         Detection Rate : 0.9663
   Detection Prevalence : 0.9663
      Balanced Accuracy : 1.0000

       'Positive' Class : 1
```

Figure 24: Overall statistics for RF using clustering results

Observing both figures and having in consideration all obtained results using Absenteeism and clustering results as the diagnosis variable, one can conclude that with the second one there were obviously better outcomes, since it obtained greater scores on all performance evaluators.

However, note that using Absenteeism as a response variable there are 4 possible levels of classification against 2 with the clustering results. Which means that with the last one it transforms into a binary classification, which facilitates the prediction. Also note that the train and test datas are so imbalanced that even if in section 5.1 all test data observations were predicted as class 1 the accuracy would still be higher than 0.9, which by itself shows the obtained scores can be a little misleading.

# 6 Conclusion

After the data's choice, the first problem to be faced was having a continuous diagnosis variable. Four classes were created based in information interest, which may not have been the best approach , because they turned out to be strongly imbalanced. For a new analysis it would be recommended to have in consideration the number of observations. That is, the separation should be in balanced classes, instead of interesting (but imbalanced) ones.

To deal with that imbalance one must be careful choosing it's performance metrics and the algorithms to solve this problem. Confusion Matrix and F1-Score can be good choices for this metrics and Random Forest is also a good algorithm for this kind of datasets.

One way to tackle this in a further work could be by doing oversampling on the data, principally on classes 3 and 4 which are clearly not well represented.

Although the obtained clusters on the unsupervised analysis are not that good nor useful, it is possible to see that this dataset still holds some relations that allow to group the work absenteeism in clusters, like the one found for absenteeism of social smokers that are not social drinkers.

Hereupon, with that clusters the obtained classifier had a good performance, but it may have been a too good one. It is possible that the obtained cluster generalized to much all data, which favored the prediction.

# References


[1] UCI. *Absenteeism at work Data Set*. URL: https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work.

[2] STHDA. *Multidimensional Scaling*. URL: http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/122-multidimensional-scaling-essentials-algorithms-and-r-code/.

[3] Hackernoon. *One Hot Encoding*. URL: https://hackernoon.com/what-is-one-hot-encoding-why-and-when-do-you-have-to-use-it-e3c6186d008f.

[4] ScienceDirect. *Hamming Distance*. URL: https://www.sciencedirect.com/topics/mathematics/hamming-distance.

[5] STHDA. *Regression Model Validation*. URL: http://www.sthda.com/english/articles/38-regression-model-validation/157-cross-validation-essentials-in-r/.

[6] Towards data science. *Dealing with Imbalanced Data*. URL: https://towardsdatascience.com/methods-for-dealing-with-imbalanced-data-5b761be45a18.

[7] STHDA. *Discriminant Analysis*. URL: http://www.sthda.com/english/articles/36-classification-methods-essentials/146-discriminant-analysis-essentials-in-r/#quadratic-discriminant-analysis---qda.

[8] Towards data science. *Support Vector Machine*. URL: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47.

[9] Data Flair. *SVM Kernel Functions*. URL: https://data-flair.training/blogs/svm-kernel-functions/.

[10] Monkey Learn. *A practical explanation of a Naive Bayes classifier*. URL: https://monkeylearn.com/blog/practical-explanation-naive-bayes-classifier/.

[11] pathmind. *A Beginner's Guide to Neural Networks and Deep Learning*. URL: https://pathmind.com/wiki/neural-network.

[12] Towards data science. *Random Forest*. URL: https://towardsdatascience.com/understanding-random-forest-58381e0602d2.

[13] Maria do Rosário Silva. *Cluster Analysis - Multivariate Analysis Course*. URL: https://fenix.tecnico.ulisboa.pt/downloadFile/282093452075835/Cluster%5C%20Analysis_v2.pdf.

[14] Towards data science. *Clustering on mixed type data*. URL: https://towardsdatascience.com/clustering-on-mixed-type-data-8bbd0a2569c3.

[15] STHDA. *DBSCAN*. URL: http://www.sthda.com/english/wiki/wiki.php?id_contents=7940.

[16] Pedro Domingos. *A Few Useful Things to Know About Machine Learning*. URL: https://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf.

[17] Science Direct. *Fuzzy Clustering*. URL: https://www.sciencedirect.com/topics/computer-science/fuzzy-clustering.

# Appendix

### Section 1

The variable *Reason for Absence* can initially assume 28 values, which represent:

1. Certain infectious and parasitic diseases
2. Neoplasms
3. Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
4. Endocrine, nutritional and metabolic diseases
5. Mental and behavioural disorders
6. Diseases of the nervous system
7. Diseases of the eye and adnexa
8. Diseases of the ear and mastoid process
9. Diseases of the circulatory system
10. Diseases of the respiratory system
11. Diseases of the digestive system
12. Diseases of the skin and subcutaneous tissue
13. Diseases of the musculoskeletal system and connective tissue
14. Diseases of the genitourinary system
15. Pregnancy, childbirth and the puerperium
16. Certain conditions originating in the perinatal period
17. Congenital malformations, deformations and chromosomal abnormalities
18. Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
19. Injury, poisoning and certain other consequences of external causes
20. External causes of morbidity and mortality
21. Factors influencing health status and contact with health services
22. Patient follow-up
23. Medical consultation
24. Blood donation
25. Laboratory examination
26. Unjustified absence
27. Physiotherapy
28. Dental consultation

The first 21 categories correspond to absences attested by the International Code of Diseases (ICD).

## Section 2

```
> #Categorical variables
> table(data$Absenteeism)

  1   2   3   4
425 206  48  15
> table(data$`Reason for absence`)

  1   2   3   4   5   6   7   8
262  36 149   3  31  33  68 112
> table(data$Month)

 0  1  2  3  4  5  6  7  8  9 10 11 12
 0 49 72 83 52 57 50 65 54 44 62 57 49
> table(data$`Week day`)

  2   3   4   5   6
153 141 145 118 137
> table(data$Seasons)

  1   2   3   4
164 189 167 174
```

```
> table(data$`Disciplinary failure`)

  0   1
694   0
> table(data$Education)

  1   2   3   4
569  43  78   4
> table(data$`Social drinker`)

  0   1
305 389
> table(data$`Social smoker`)

  0   1
648  46
```

```
> reasons_counts <- table(data$`Reason for absence`)
> reasons_counts / sum(reasons_counts)

          1           2           3           4           5           6           7           8
0.377521614 0.051873199 0.214697406 0.004322767 0.044668588 0.047550432 0.097982709 0.161383285
> education_counts <- table(data$Education)
> education_counts / sum(education_counts)

          1           2           3           4
0.819884726 0.061959654 0.112391931 0.005763689
> smoker_counts <- table(data$`Social smoker`)
> smoker_counts / sum(smoker_counts)

         0          1
0.93371758 0.06628242
> absenteeism_counts <- table(data$Absenteeism)
> absenteeism_counts / sum(absenteeism_counts)

         1          2          3          4
0.61239193 0.29682997 0.06916427 0.02161383
```