Università degli Studi di Milano

Data Science and Economics

Supervised learning problem: smokers classification based on body signals

Student:
Michele Bartesaghi
Registration number: T16334

ACADEMIC YEAR 2021-2022

Contents

1 Abstract **2**

2 The dataset **3**
2.1 Pre-processing 5
2.2 Exploratory data analysis 6
2.3 Correlation 10

3 Classification **11**
3.1 Decision tree 11
3.2 Random forest 15
3.3 K-NN 17
3.4 Logistic regression 19

4 Conclusions **22**

Bibliography **23**

Chapter 1

Abstract

The aim of this project is to classify smokers on the basis of body signals, namely physical individual features, also called *bio-signals*. In order to do so, a dataset which is a collection of medical records regarding innumerable people's biological characteristics has been downloaded. Moreover, it is interesting to look at the problem the other way around and investigate how impacting smoking is on the human body, should this be the case.

After a brief pre-processing, a study of the different variables is carried out, in an attempt to deeply understand the nature of the data and what they communicate as they are.

Ultimately, a few supervised machine learning techniques are used to achieve the objective stated before, namely a decision tree, a random forest, a k-nn and a logistic regression. All these models achieve a reasonable result in terms of accuracy, but the random forest proves to be the best one.

Chapter 2

The dataset

The dataset presents 55692 records, each one corresponding to a different individual, and 27 variables, which are both quantitative and qualitative. Observe that dealing with mixed-type data is not as easy as focusing on numeric only features, but it represents a valuable challenge, as in real life people have to deal with various data. The observations were collected in 2020 as part of the annual health check carried out by the National Health Insurance Corporation, in South Korea.

The response variable is called *smoking* and it is a dichotomic vector, each entry of which belongs to $\{0, 1\}$, where 0 means non-smoker, whilst 1 stands for smoker. Here are the other variables, followed by a concise description:

- ID,
- gender,
- age,
- height(cm),
- weight(kg),
- waist(cm),
- eyesight(left),
- eyesight(right),
- hearing(left),
- hearing(right),
- systolic,
- relaxation,
- fasting blood sugar,
- Cholesterol,
- triglyceride,
- HDL,
- LDL,
- hemoglobin,
- Urine protein,
- serum creatinine,
- AST,
- ALT,

- Gtp,
- oral,
- dental caries,
- tartar.

The majority of the variable names are self-explicative, but it might be of use to specify them and give a further explanation about the ones which may cause some doubts. *ID* is a unique identification code, while *gender* is a binary vector with a distinction between "M" (male) and "F" (female). *Age* is measured in 5 years units, while *height(cm)* and *weight(kg)* are measured in 5 centimetres and 5 kilograms units respectively. Analogously, *waist(cm)* is measured in 5 centimetres units. Moreover, both *eyesight(left)* and *eyesight(right)* are continuous, varying in $[0.1, 9.90]$, whereas *hearing(left)* and *hearing(right)* can only assume values in $\{1, 2\}$, leaving space to some speculation about their actual meaning. Probably they stand for a low or high hearing power respectively. *Systolic* and *relaxation* (diastolic) are measurements of the two different blood pressure types and *fasting blood sugar* is the level of glucose in one's blood (measured in mg/dL). Also *Cholesterol* is measured in mg/dL, as well as *triglyceride*, *HDL* and *LDL*, which are the notorious "good" and "bad" cholesterol respectively. *Hemoglobin*, which is responsible for carrying the oxygen in the blood, is measured in g/dL, whilst *Urine protein* stands for the amount of protein expelled with urine (measured in mg), categorised into 6 levels. *GTP* is called gamma-glutamyl transpeptidase and its level rises in presence of liver diseases, therefore, together with *ALT*, it is a useful parameter to assess the condition of the liver. Furthermore, *AST* plays a similar role, but it is also found in other organs like the brain, the pancreas, the heart or the kidneys. They are all measured in U/L (units per litre). *Oral* stands for the oral examination status and it assumes one value only, while *serum creatinine* is a waste product of the normal wear and tear of muscles (measured in mg/dL). Ultimately, *dental caries* varies in $\{0, 1\}$, where 1 indicates the presence of dental caries, and *tartar* varies in $\{Y, N\}$, where "N" obviously stands for the absence of tartar.

2.1 Pre-processing

First and foremost, it is possible to observe that in this dataset there are no missing values, which are frequently difficult to handle. Secondly, the dataset is inspected to a deeper level to understand whether some variables should be removed. A common rule of thumb suggest to get rid of variables that can only assume one possible value, as well as categorical variables with too many levels. For this reason, the variable *oral* can be excluded from the analysis from now on, as it presents only and solely entries equal to 1. Moreover, the column *ID* is rather useless in terms of the objective of this work, because no relevant information is added by the individual's unique identifier at all. Finally, the numeric features representing different levels (or categories) are all transformed into factors; they are:

- gender,
- hearing(left) and hearing(right),
- dental caries,
- tartar,
- Urine protein,
- smoking.

The next step that needs to be taken is to analyse each variable individually, in order to identify potential outliers. It is clear from this plot (2.1) that the majority of the variables presents potential outliers that could hinder the performance of the algorithms. However, a closer look to the summary of these box plots reveals that the number of potential outliers is rather large in almost every feature. For this reason, these points do not appear to be incorrectly entered nor measured data and they do not seem to affect the assumptions required to fit the models that will be used later, therefore they will not be considered as outliers for the moment.

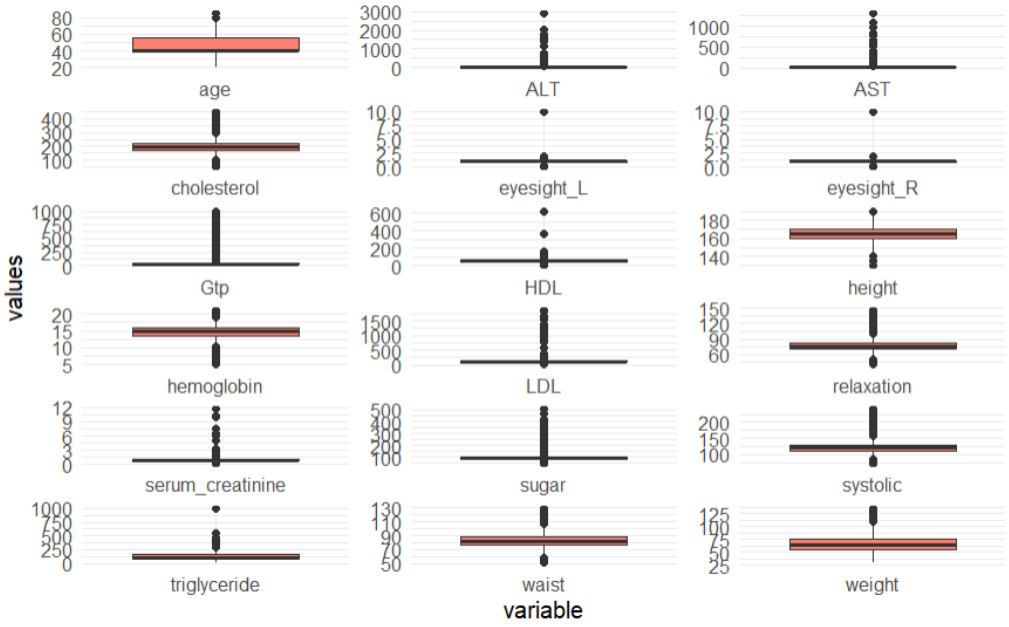

Figure 2.1: Box plots of the numeric variables.

2.2 Exploratory data analysis

Now it is possible to retrieve some information from the available data, in order to see whether they match common believes about the relationship between smoking and health, or they lead to unexpected findings.

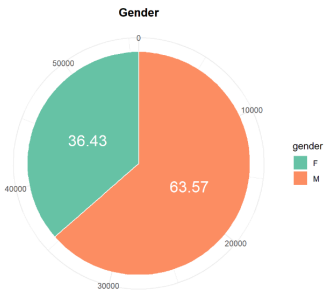

Figure 2.2: Gender distribution.

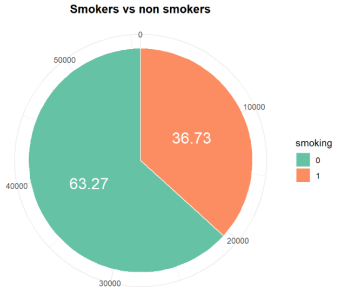

Figure 2.3: Smokers distribution.

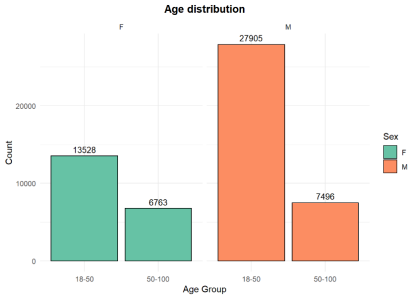

Figure 2.4: Age distribution according to the gender.

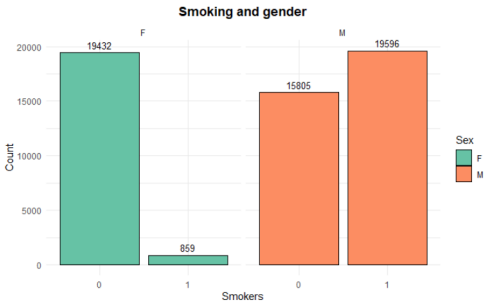

Figure 2.5: Smokers distribution according to the gender.

According to these plots, there is a larger number of male individuals and a prevalent percentage of non-smokers. Later results with and without a class balancing technique will be compared. It is also clear that most of the people are aged between 18 and 50 years old, where there are more male individuals. Moreover, it is incredibly evident how unbalanced the proportions of smokers between the two different genders are. In the table below there are people's average age, weight and height divided by gender and smoking status.

gender	smoking	avg_age	avg_height	avg_weight
F	0	49	156.00	56.00
F	1	46	157.00	56.00
M	0	42	170.00	71.00
M	1	41	170.00	72.00

Now a few graphs about more meaningful bio-signals will be reported, in order to determine whether smoking has a significant impact on the human body. Notice that all the categories designed to distinguish various "health statuses" are derived from online papers and websites, which will be eventually included in the bibliography.

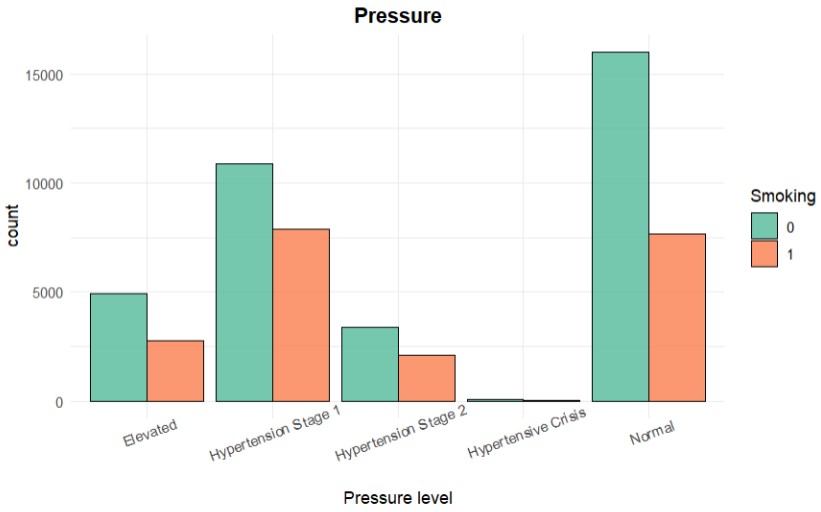

Figure 2.6: Blood pressure according to smoking status.

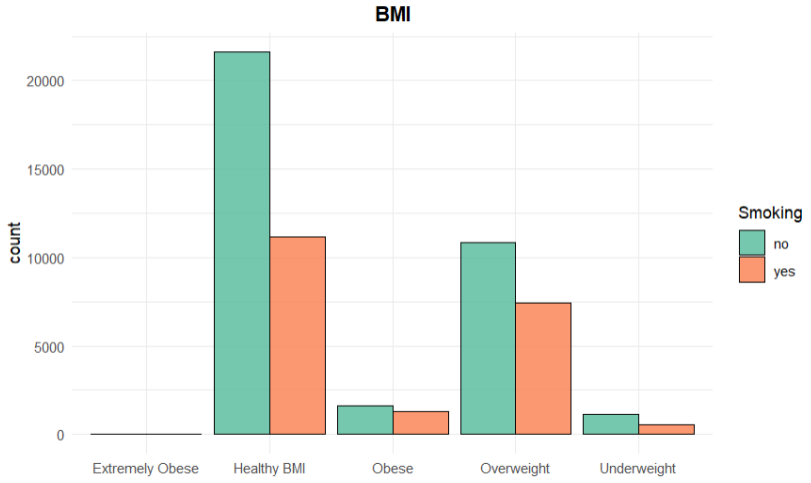

Figure 2.7: Body Mass Index according to smoking status.

Even though it is commonly believed that smokers are at a higher risk of heart diseases, nothing highly relevant can be said about these graphs. With regard to 2.6, it is safe to say that most of the non-smokers are in the normal category, meaning that a non-smoker is less likely to suffer from hypertension. On the other hand, a considerable number of smokers falls into the Hypertension Stage I category, with respect to the total of smokers observed in other categories. As far as the Body Mass Index (BMI) is concerned,

similar considerations can be made: most of the non-smokers have a healthy BMI, but also the majority of smokers belongs to the same category. Finally, observe that in the Extremely Obese category, we have an equal percentage of both smokers and non-smokers.

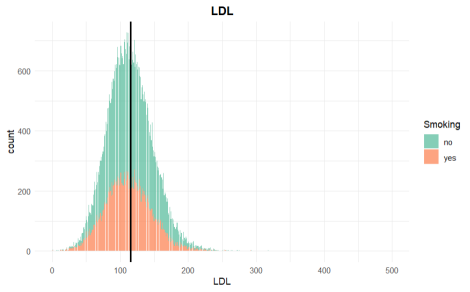

Figure 2.8: LDL levels. The black line represents the mean.

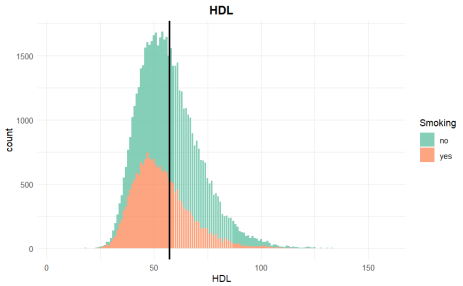

Figure 2.9: HDL levels. The black line represents the mean.

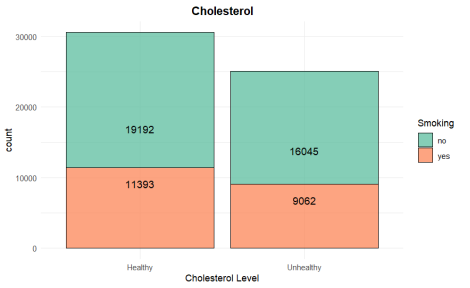

Figure 2.10: Total cholesterol levels

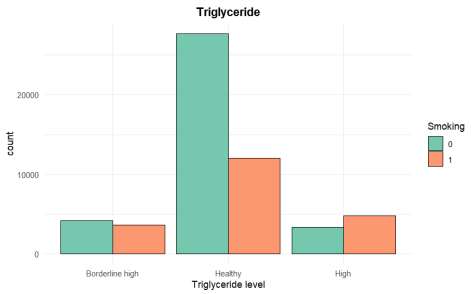

Figure 2.11: Triglycerides levels

Both the LDL and the HDL distributions are similar between non-smokers and smokers. However, among those people with a high concentration of triglycerides, smokers are prevalent, whilst almost every non-smoker has a healthy level of triglycerides. Ultimately, the table below shows that smokers seem to be more prone to suffer from liver diseases, according to what has been stated while describing the variables.

smoking	Alt_mean	Ast_mean	Gtp_mean
0	24.74	25.31	30.89
1	30.99	27.69	55.57

2.3 Correlation

In the last section of this chapter, the correlation between the continuous variables of the dataset is studied.

According to the graph (2.12), it is safe to state that the following variables are almost perfectly correlated:

- *weight* and *waist*,
- *systolic* and *relaxation*,
- *cholesterol* and *LDL*,
- *AST* and *ALT*.

Moreover, a few variables appear to be rather highly correlated, namely *height* and *weight*, as well as *height* and *hemoglobin*. However not all the correlated variables will be removed a priori, because one can aright believe that they can still bring some additional information. In fact, only *waist* and *ALT* will be excluded from the analysis from now on. For example, it would be appropriate to remove *LDL* without excessive hesitation if cholesterol were the exact sum of *LDL* and *HDL*, which is not the case. Observe also that there is no significant negative correlation between variables. Finally, it might be required to evaluate the importance of these correlations again when considering the logistic regression approach.

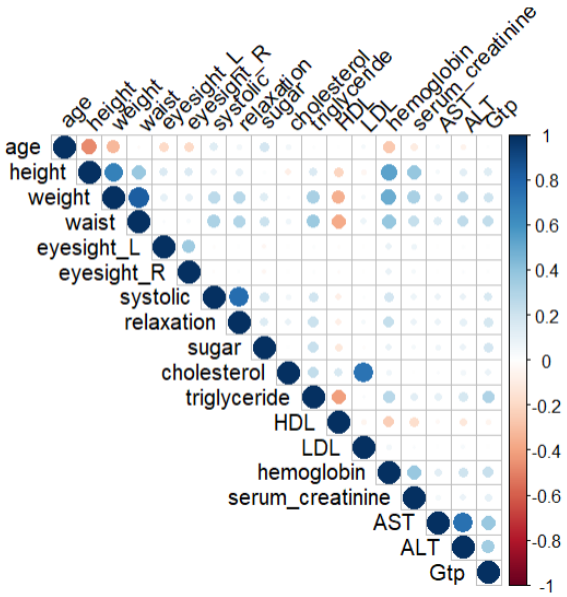

Figure 2.12: Correlation between continuous variables.

Chapter 3

Classification

In this chapter different machine learning techniques will be applied to seek to correctly classify a person either as a smoker or a non-smoker, given the bio-signals described above. Typically, the performance of a classifier $\widehat{C}(x)$ will be measured by the *misclassification error rate*

$$Err_{Te} = Avg_{i \in Te} \mathbb{1}[y_i \neq \widehat{C}(x_i)]$$

where Te is the test set. First of all the dataset is split into training set and test set, leaving 30% of the data to test the classifiers trained on the training set.

3.1 Decision tree

The first algorithm used to classify data points is a decision tree. This method is rather simple and really useful for its interpretability[1], hence it is widely used, notwithstanding it is easily outperformed by other learning algorithms in terms of prediction accuracy. Moreover, trees can easily handle all kind of qualitative predictors.

A tree is made up of *nodes* and *labels*. At a given internal node, the label $X_j < t_k$ indicates the left-hand branch originating from that split, while the right-hand bough corresponds to $X_j \geq t_k$[2]. The terminal nodes are also called *leaves* and the process of building a tree corresponds to a partition of the predictors space; in other words the set of possible values for $X_1, \ldots, X_p$ gets divided into J distinct (non-overlapping) regions, called $R_1, \ldots, R_J$. Then, for every observation that falls into the $j-th$ region, the same prediction is made, which is the most commonly occurring class of the

[1] Decision trees mirror human decision-making approaches.

[2] In case of qualitative variables the labels are $X_j = t_k$ or $X_j \neq t_k$.

response variable for the training observations in R_j. In order to maintain a great interpretability, the predictors space is divided into *boxes*, and the goal is to find $R_1, \ldots, R_J$ such that the classification error rate[3]

$$E = 1 - \max_k(\widehat{p}_{mk}),$$

is minimised, where $\widehat{p}_{mk}$ is the proportion of the training observations in the $m-th$ region that belong to the $k-th$ class. However, this quantity is often replaced with other effective measures. In practice, the *Gini index* defined as

$$G = \sum_{k=1}^{K} \widehat{p}_{mk}(1 - \widehat{p}_{mk})$$

is used as an evaluation of the total variance across the K classes, hence the aim is to minimise this measure, estimating the *purity* of a node: a small value of G indicates that a node contains predominantly observations from a single class. Due to the fact that it is impossible to consider every possible partition of the predictors space, at each step the best split is chosen, without looking ahead in order to pick a split that could lead to a better tree. The predictor X_j is selected together with the cut point s such that the split

$$\{X|X_j < s\} \quad \{X|X_j \geq s\}$$

leads to a reduction of the Gini index. The process is repeated, splitting one of the previously identified regions, until a stopping criterion is reached; for instance, one might want that every node has a Gini index smaller than a certain threshold and if it is not the case, that node needs further splitting.

The tree classifier below achieves an accuracy[4] of 74.6% on the test set. In order to achieve this result, a cross validation on both the maximum depth k and the complexity parameter have been performed, with the optimal k being 7 and the optimal cp being about 0.0012[5]. We can also observe that *gender*, *Gtp*, *triglyceride* and *age* are the most used variables to split the nodes.

[3]The fraction of the training observations in that region that do not belong to the most common class.

[4]Accuracy = (TP+TN)/(TP+TN+FP+FN)

[5]The complexity parameter is used to control the size of the decision tree. If the cost of adding another variable to the decision tree from the current node is above the value of cp, then the tree building does not continue.

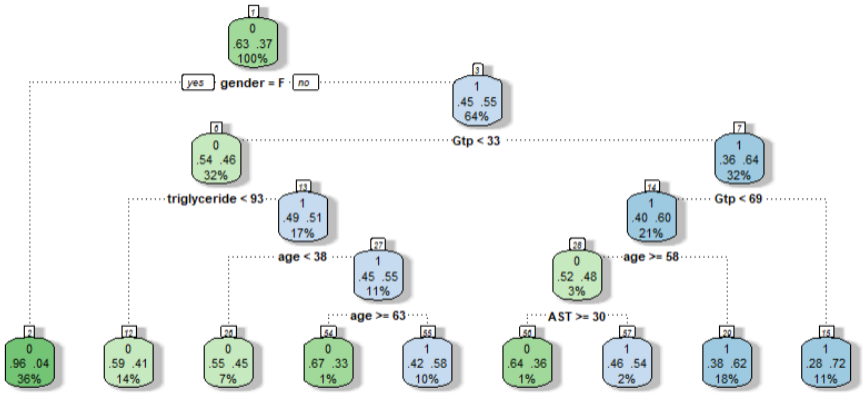

Figure 3.1: Decision tree with *"maxdepth = 5"* for interpretability purposes, achieving 73.9% of accuracy. According to figure 2.5, it makes sense that *gender* is highly relevant in the discrimination between smokers and non-smokers.

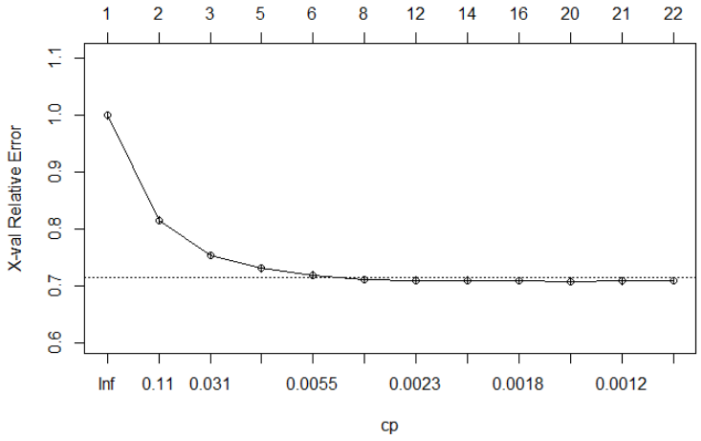

Figure 3.2: Complexity parameter tuning

In 3.4 one can see the confusion matrix of the tree above, where sensitivity is the *true positive rate* (TPR =TP/(TP+FN)) and specificity is the *true negative rate* (TNR=TN/(TN+FP)). In this particular case, sensitivity is the probability that an individual is classified as a non-smoker, given that he is not a smoker, and specificity the probability that an individual is correctly classified as a smoker.

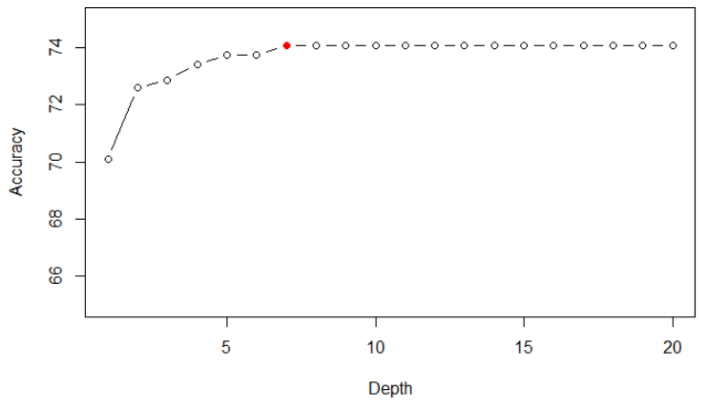

Figure 3.3: Maximum depth tuning. Going deeper than

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 8128 1791
         1 2443 4345

               Accuracy : 0.7466
                 95% CI : (0.7399, 0.7532)
    No Information Rate : 0.6327
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4666

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.7689
            Specificity : 0.7081
         Pos Pred Value : 0.8194
         Neg Pred Value : 0.6401
             Prevalence : 0.6327
         Detection Rate : 0.4865
   Detection Prevalence : 0.5937
      Balanced Accuracy : 0.7385
```

Figure 3.4: Decision tree confusion matrix with the unbalanced dataset.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 4627  383
         1 3762 7935

               Accuracy : 0.7519
                 95% CI : (0.7453, 0.7584)
    No Information Rate : 0.5021
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5046

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.5516
            Specificity : 0.9540
         Pos Pred Value : 0.9236
         Neg Pred Value : 0.6784
             Prevalence : 0.5021
         Detection Rate : 0.2769
   Detection Prevalence : 0.2999
      Balanced Accuracy : 0.7528
```

Figure 3.5: Decision tree confusion matrix with the balanced dataset.

Observe that in general there is a sort of trade-off between sensitivity and specificity and in this particular case one would be interested in increasing the specificity, meaning that the tree is more capable of guessing whether a smoker does actually smoke, which is the objective of this work. Finally, notice that a tree is biased by class imbalance. In figure 3.5, one can see how balancing the dataset[6] allows to increase the accuracy by 2 percentage points, as well as improve significantly the specificity, at the cost of some sensitivity loss.

[6]Meaning there is almost an equal number of smokers and non-smokers in the response variable.

3.2 Random forest

Random forest is an ensemble method that aggregates more decision trees in order to improve the classification performance. This technique builds the decision trees on bootstrapped training samples, that are samples repeatedly taken from the training set, and for each split it considers only a random selection of m out of p predictors (typically $m \approx \sqrt{p}$). After training our model on each of the B bootstrapped training sets in order to get a predictor $\widehat{f}^{*b}(x)$ (classifier trained on the $b-th$ training set), a record of the class predicted by each of the B trees is kept and a majority vote is taken in order to classify the new point[7].

It is possible to estimate the test error of such a model. Trees are repeatedly fit to bootstrapped subsets of the original observations and each tree makes use of around two-thirds of them. The remaining one-third of the observations is called *out-of-bag* (OOB) set of observations, hence the label of the $i-th$ point is predicted using the trees that observation was OOB for ($B/3$ predictions from which the majority vote is taken).

Observe that all these strategies to achieve a better model come at the cost of decreased interpretability of the results, in fact some people refer to random forest as a "*black box*".

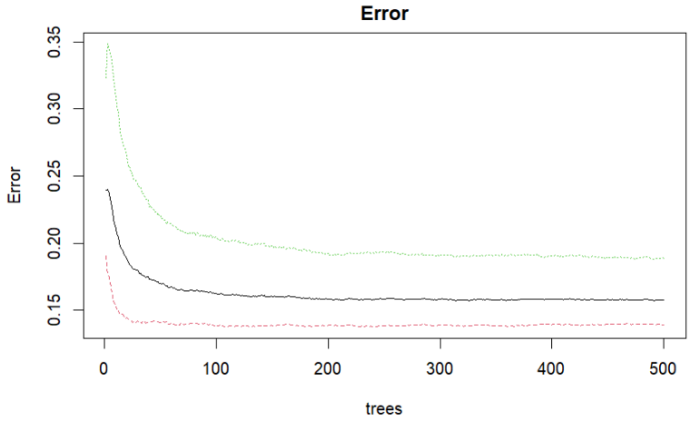

Figure 3.6: Random forest estimated test error, as a function of the number of trees.

In this graph one can observe that the estimated test error is stable after approximately sixty trees and each coloured line corresponds to a different class error: the green one regards the smokers class, while the red one corresponds to the non-smokers class, in fact the specificity is lower than

[7]The overall prediction is the most occurring class among the B predictions.

sensitivity. On top of this, the error rate is much lower with respect to the one achieved by the single decision tree showed before. As a matter of fact, with the random forest approach the accuracy of the model is about 83%, with both sensitivity and specificity increased by almost 10 percentage points. Moreover, in figure 3.9 one can see the total amount by which the Gini index is decreased by splits over a given predictor, averaged over all B trees. To put it another way, that plot gives an insight into the importance of the different variables when it comes to building the trees.

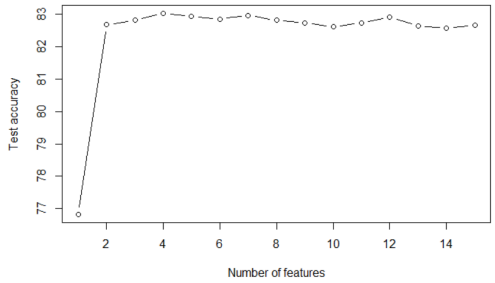

Figure 3.7: Test accuracy as a function of the number of features used at each split

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 8962 1253
         1 1609 4883

               Accuracy : 0.8287
                 95% CI : (0.8229, 0.8344)
    No Information Rate : 0.6327
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6358

 Mcnemar's Test P-Value : 3.227e-11

            Sensitivity : 0.8478
            Specificity : 0.7958
         Pos Pred Value : 0.8773
         Neg Pred Value : 0.7522
             Prevalence : 0.6327
         Detection Rate : 0.5364
   Detection Prevalence : 0.6114
      Balanced Accuracy : 0.8218
```

Figure 3.8: Random forest confusion matrix

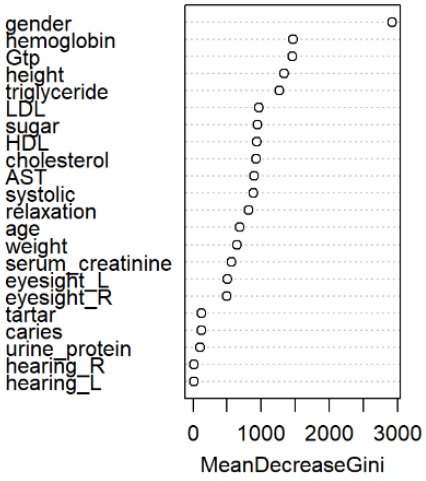

Figure 3.9: Variables importance plot

Ultimately, balancing the dataset causes the accuracy to decrease slightly (approximately 77.9%), but at the same time the model presents a higher specificity (approximately 87%), meaning that it is more efficient at classifying "true" smokers.

3.3 K-NN

At this point of the analysis, one could ask himself whether it would be a remarkable idea to fit a k-nearest neighbours (k-NN) model on the training set. It is fundamental to highlight that some variables shall be removed, should one try to apply the k-NN algorithm, as their values neither have a distance, nor can be compared with the numeric features values, on the basis of the euclidean distance. These are:

- gender,
- tartar,
- hearing(left),
- hearing(right),
- Urine protein,
- dental caries,

that have been all treated as factors up to this point. Then the remaining numeric features are scaled.

The idea behind the algorithm is the following: given a certain test point x_t, the classifier finds the k-nearest neighbours of that point in the given dataset and predict a class y_t, which is the majority vote of the classes of the k-nearest neighbours. It is quite obvious that this method completely relies on distance[8], therefore it is fundamental to scale the data. Notice that k-NN is quite fast at training[9], but it is rather demanding on memory. As k increases, the smoothness of the curve which represents the decision boundary increases as well. There are two extreme possible situations:

- $k = 1$: overfitting. The model achieves 100% accuracy on the training set, but it performs poorly on the test set;
- $k = m$: underfitting. Every point is classified according to the most common label in the dataset.

[8] typically the euclidean distance $d(x, y) = \sqrt{\sum(x_i - y_i)^2}$

[9] $\mathcal{O}(n \cdot d)$, where n is the number of distances and d the dimension of data points.

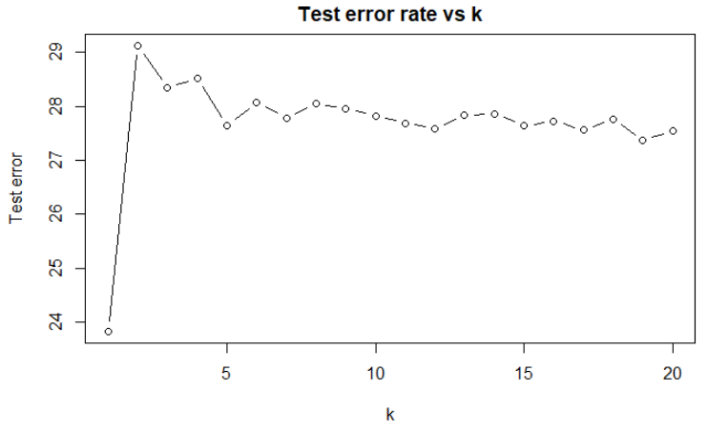

Figure 3.10: Test error rate as a function of k.

From this intuitive graph one can notice that $k = 1$ achieves the lowest error, but it would also lead to overfitting, hence $k = 19$ is used. However, with the nearly optimal k, this model achieves a 72.65 accuracy, getting close to the single decision tree performance. Observe that it is possible to see the behaviour of the k-NN on a subset of the data points. A Singular Value Decomposition (SVD) is performed and the data points are projected onto a 2-dimensional space. After taking a random subset of the projected observations the k-NN model is fitted and then the results are plotted. Every pink point with a triangle inside, as well as every blue point with a dot inside, is misclassified by the algorithm.

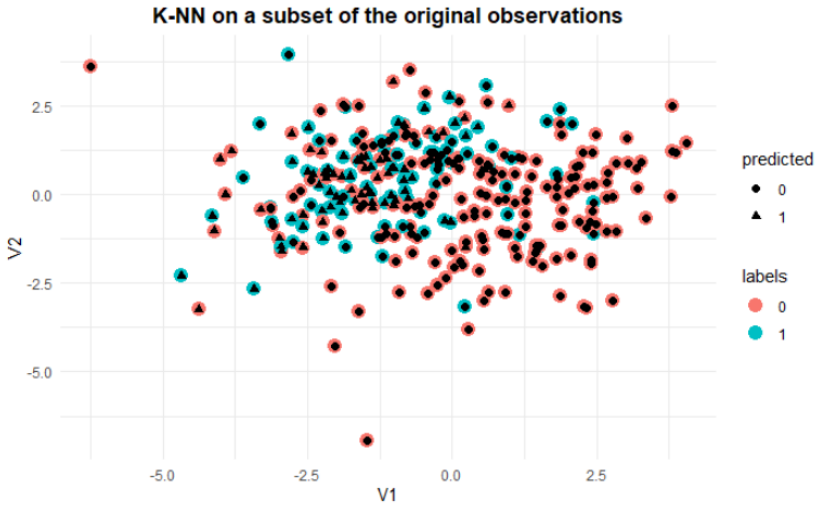

Figure 3.11: k-nn visualisation in the plane spanned by the two principal components.

Finally, notice how the two groups are overlapped. This is due also to the fact that the two principal components fail at describing the geometry of the data points at a sufficient extent.

In the k-NN approach, balancing the dataset does not seem to be beneficial in a significant way, therefore the results are omitted.

3.4 Logistic regression

The last algorithm used in an attempt to correctly classify smokers on the basis of a few bio-signals is the logistic regression. Observe often people are interested more in estimating the probabilities that a given point belongs to a certain category. In this case, in which the response variable is binary, it is possible to fit a linear regression model, but it might produce probabilities either smaller than zero or bigger than one, which is a contradiction. Let $p(X) = Pr(Y = 1|X)$. In general, the logistic regression uses the form

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}} \in (0, 1)$$

One can consider the *logit* transformation of $p(X)$, which is

$$log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

Then, the parameters are chosen as the ones that maximise the likelihood

$$\ell(\beta_0, \beta) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i))$$

and then they are used to calculate $\widehat{p}(X)$. Opposite to the classification methods that have been studied before, which are capable of handling skewed distributions and are robust with respect to outliers, for the logistic regression we need to verify that the dataset satisfies rather strong assumptions.

- The response variable is binary
- Observations are independent

These assumptions are both trivially satisfied, because the response variable is indeed binary and each observation is a record regarding a different individual, not related to the previous or following one by any means.

- Absence of multicollinearity among predictors

We already checked the correlation among continuous variables before, and the same considerations hold. Remember that both *waist* and *ALT* have been removed from the dataset, due to their nearly perfect correlation with other variables. In addition, as for the logistic regression, a common rule of thumb states that any variable with a correlation coefficient greater than 0.7 with another variable, should be removed. This leads to the deletion of the *LDL* feature (correlation of 0.74 with Cholesterol). Moreover, the correlation among factor variables is calculated[10], and *hearing(left)* is perfectly correlated with *hearing(right)*, therefore the first one is excluded from the model. Ultimately, a new column *bmi* is created in order to ignore the *height(cm)* and *weight(cm)* features that are moderately correlated. The next step is to fit the model and calculate the Variance Inflation Factor, which underlines a collinearity problem with Urine protein. Once we have removed this last problematic feature from the data set, the next assumption can be examined.

- Absence of extremely influential outliers

The logistic regression model is more sensitive to outliers, compared with the methods fitted so far. Therefore, a more rigorous analysis of these atypical points is required. The logistic regression model is fitted and then the Cook's distance is computed for every observation in the data set. As the graph shows, there are a few potential outliers, that will be dropped.

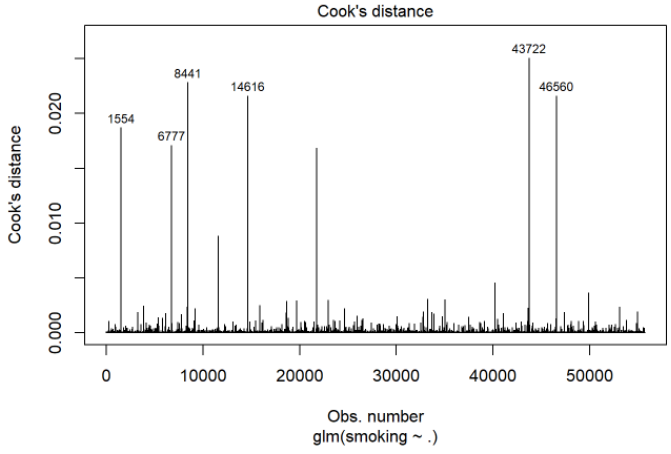

Figure 3.12: Cook's distance to identify potential outliers.

- Linear relationship between continuous explanatory variables and the logit of the response variable

[10]Using the cramerV() function, which is a measure of the association between two categorical variables.

One way to check this assumption is to fit the model and then plot the continuous independent variables against the probabilities of the model. It is rather safe to say that there is a fairly linear relationship, by looking at the figures below (3.13, 3.14).

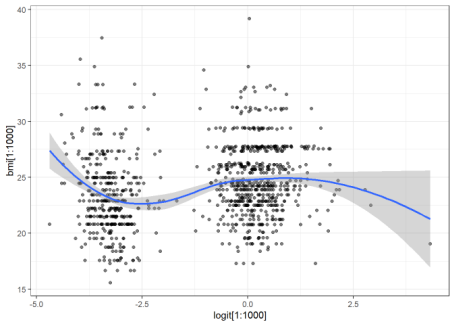

Figure 3.13: *bmi* vs log odds

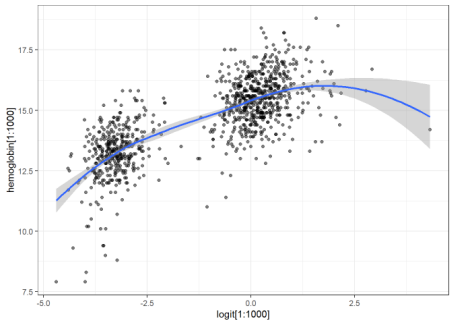

Figure 3.14: *hemoglobin* against log odds

Since all of the assumptions are now verified, it is possible to fit the model and classify the points in the test set. The logistic regression achieves a 74.68% accuracy, which is the second highest achieved so far[11]. It is crucial to look at the ROC (Receiver Operating Characteristic) curve, which is displaying both the sensitivity (along the y axis) and the specificity (1-specificity along the x axis) at the same time. In figure 3.15, also the AUC (Area Under the ROC Curve) is displayed. Notice that the higher the AUC, which varies between 0 and 1, the better the model. An AUC of 0.5 means that the model in no better than a random guessing made on the basis of a coin toss. Given the value of 0.828, it is safe to say that the model has a quite decent power of discrimination.

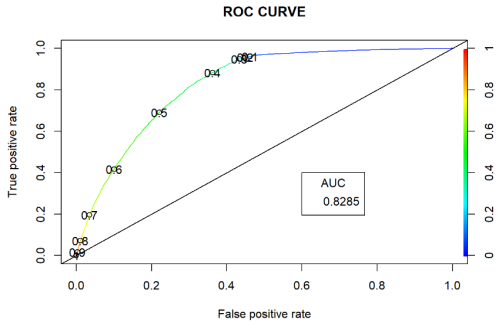

Figure 3.15: ROC-AUC.

[11] Also in the logistic regression case, balancing the dataset is not really effective.

Chapter 4

Conclusions

On the whole, this dataset did not need a sizeable amount of preprocessing. The data points revealed that a few of the common believes about the impact of smoking on the human body might be unfounded. Nevertheless, smoke does affect one's health to a great extent and should be avoided in any case. The decision tree achieved an accuracy of 74.66% in an attempt to classify a person as a smoker on the basis on a few bio-signals, and apparently the most relevant among the latter are both the Gtp and the AST levels, together with the triglycerides (apart from gender and age). This is in accordance with what has been observed at the beginning about the different average values of substances that are found in the liver between smokers and non-smoker. Moreover, balancing the dataset allowed the decision tree to achieve a slightly better accuracy of about 75%, together with a much higher precision in classifying smokers correctly. The random forest approach proved to be the best model, with about 8 rather outstanding additional percentage points of accuracy (82.8%) and an increase in both specificity and sensitivity. This has come at a cost for the interpretability, but the same variables highlighted before proved to be crucial in discriminating between the two classes. The k-NN approach has been chosen just for its highly intuitive nature, yet it performed surprisingly well, with an accuracy of 72.64 percentage points, achieved with k=19. Ultimately, a bit more of preprocessing prepared the way for a logistic regression model, which revealed a better accuracy (75.3%) with respect both to the decision tree classifier and the k-NN model, with a remarkable AUC = 0.82.

In summary, according to these results, about seven or eight people out of ten can be correctly identified as smokers, on the basis of essential human bio-signals. Needless to say, smoking tremendously affect our lives and it does not go unnoticed.

Bibliography

[1] *National Health Insurance Corporation_ Health Checkup Information*, accessed 10 June 2022, <https://www.data.go.kr/data/15007122/fileData.do>

[2] *Understanding Blood Pressure Readings*, accessed 10 June 2022, <https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings>

[3] *Normal Laboratory Values*, accessed 10 June 2022, <https://www.iapac.org/fact-sheet/normal-laboratory-values/>

[4] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*, Springer.