UNCOVERING CARDIOVASCULAR DISEASES RISK FACTORS: A MACHINE LEARNING APPROACH

Lumia Giulia, Bongiovanni Sabrina Anna

ABSTRACT

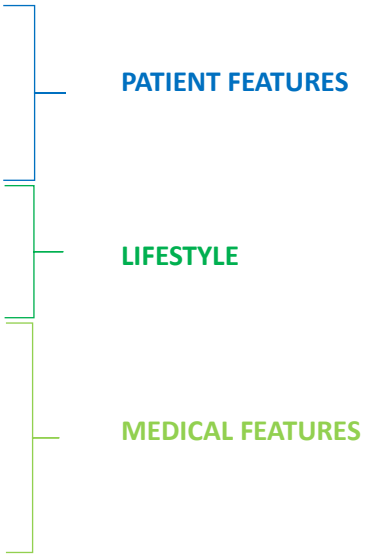
The cardiovascular diseases risk factors were explored by the mean of various machine learning techniques. As a result, it was possible to obtain several diagnostical tools, all with about 70% accuracy, capable of predicting whether a patient has a cardiovascular disease on the base of its characteristics. Irrespective of the complexity of the technique, performances are all similar, meaning that there may be some lack of information to explore.

INTRODUCTION

This report is aimed to show the results of an analysis carried out in order to produce some assessments about the probability of having a cardiovascular disease, taking into account some risk factors.

The available data refer to:

- **Age**;
- **Gender**;
- **Weight**;
- **Height**;
- **Physical activity**;
- **Alcohol intake**;
- **Smoking**;
- **Systolic blood pressure**;
- **Diastolic blood pressure**;
- **Cholesterol**;
- **Glucose**;
- **Cardiovascular disease** (presence or absence).



The analysis is developed in six stages:

1. Data preprocessing;
2. Exploratory analysis;
3. Single decision tree;
4. Random forest;
5. Gradient boosted trees: basic version;
6. Gradient boosted trees: stochastic version.

1. DATA PREPROCESSING

The first step of the analysis aims to make data suitable for the subsequent model adaptation. Moreover, some variables were transformed in order to became more meaningful, as instance on base of some existing criteria:

- **Blood pressure.**
  Since all the available data refer to adults (the age ranges between 29 and 65 years), it was possible to summarize the information about the systolic and diastolic blood pressure according to the American college of cardiology guidelines:

BLOOD PRESSURE	SYSTOLIC		DIASTOLIC
normal	<120	AND	<80
elevated	[120-130)	AND	<80
hypertension stage 1	[130-140)	OR	[80-90)
hypertension stage 2	>=140	OR	>=90
hypertensive Crisis	>180	AND/OR	>120

- **BMI index.**
  The information about the weight of a person was summarised taking into account also his/her height and sex. The weight value considered alone, in fact, doesn't allow to assess whether a person is underweight or obese.
  As instance, given a weight of 60 kg a 1.75 meters woman can be considered with normal weight while a 1.50 meters woman would be classified as overweight.
  On the base of this well-known index patient can be classified as:

BMI	MAN	WOMAN
underweight	<20,1	< 18,7
normal	[20,1-25,1)	[18,7-23,9)
overweight	[25,1-30)	[23,9-28,7)
obese S1	[30-35,1)	[28,7-35,1)
obese S2	[35,1-40]	[35,1-40]
obese S3	>40	>40

Moreover, since data contained explicit errors and medical anomalies, some corrections were needed:
- the maximum value of height was set to 2.4 meters;
- the minimum value of BMI is 3.47 while the max is 298,67;
- there can't exist cases in which the systolic blood pressure is lower than the diastolic;
- blood pressure can't be negative.

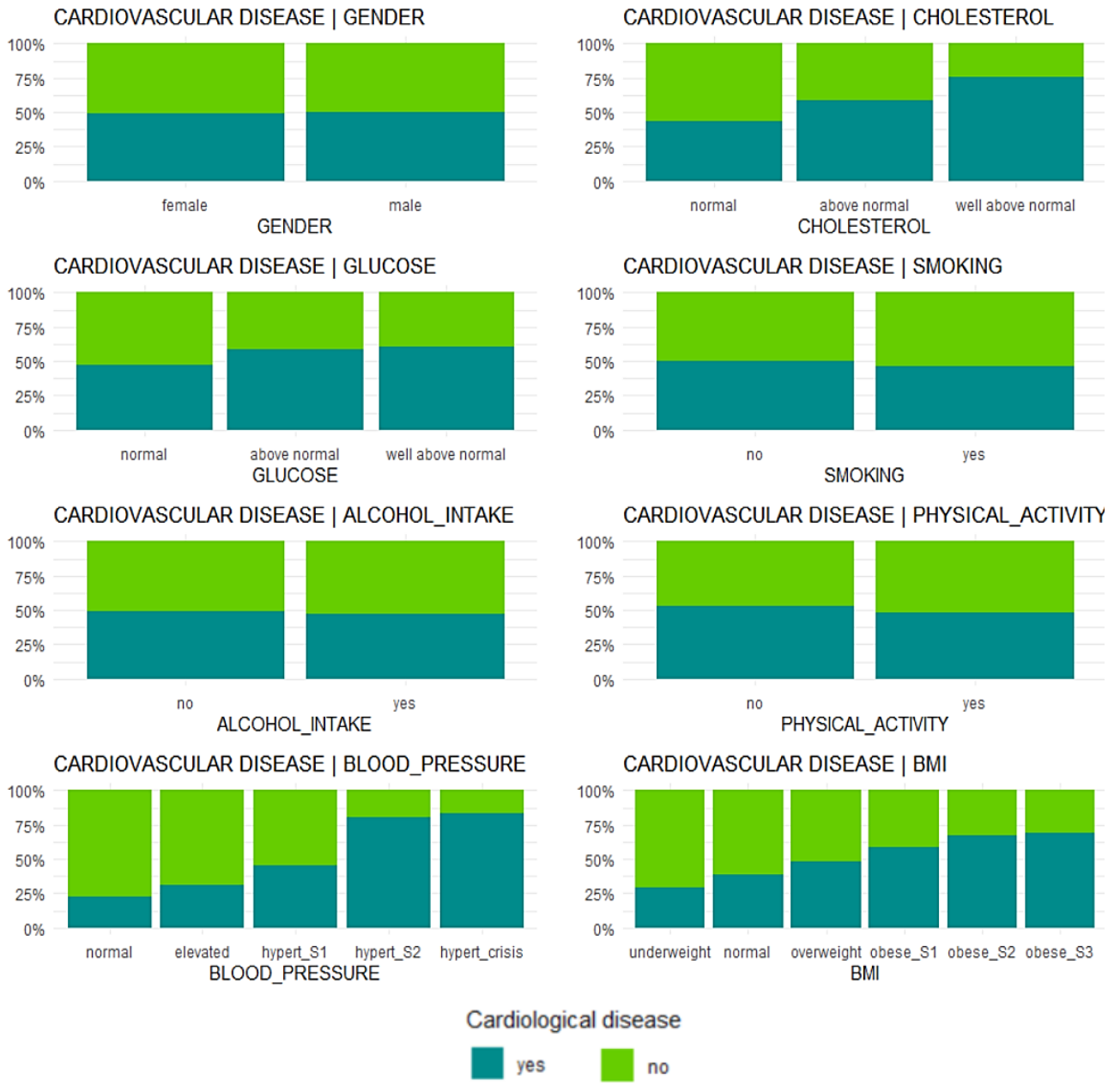
2. EXPLORATORY ANALYSIS

The main aim of the analysis is to develop a diagnostic tool with respect to the presence of cardiovascular diseases.
The original dataset was randomly partitioned in:
- **training set** (80% of the patients), used in order to train the ML classificators;
- **test set** (20% of the patients), used in order to test the performance of the obtained tools.

The partition was carried out stratifying with respect to the response variable to classify (cardiovascular disease), in order to maintain the original balance between the two categories (presence or absence of the disease).

With an exploratory purpose it may be useful to highlight the discriminant power of each of the variables available, with respect to the presence of a cardiovascular disease.

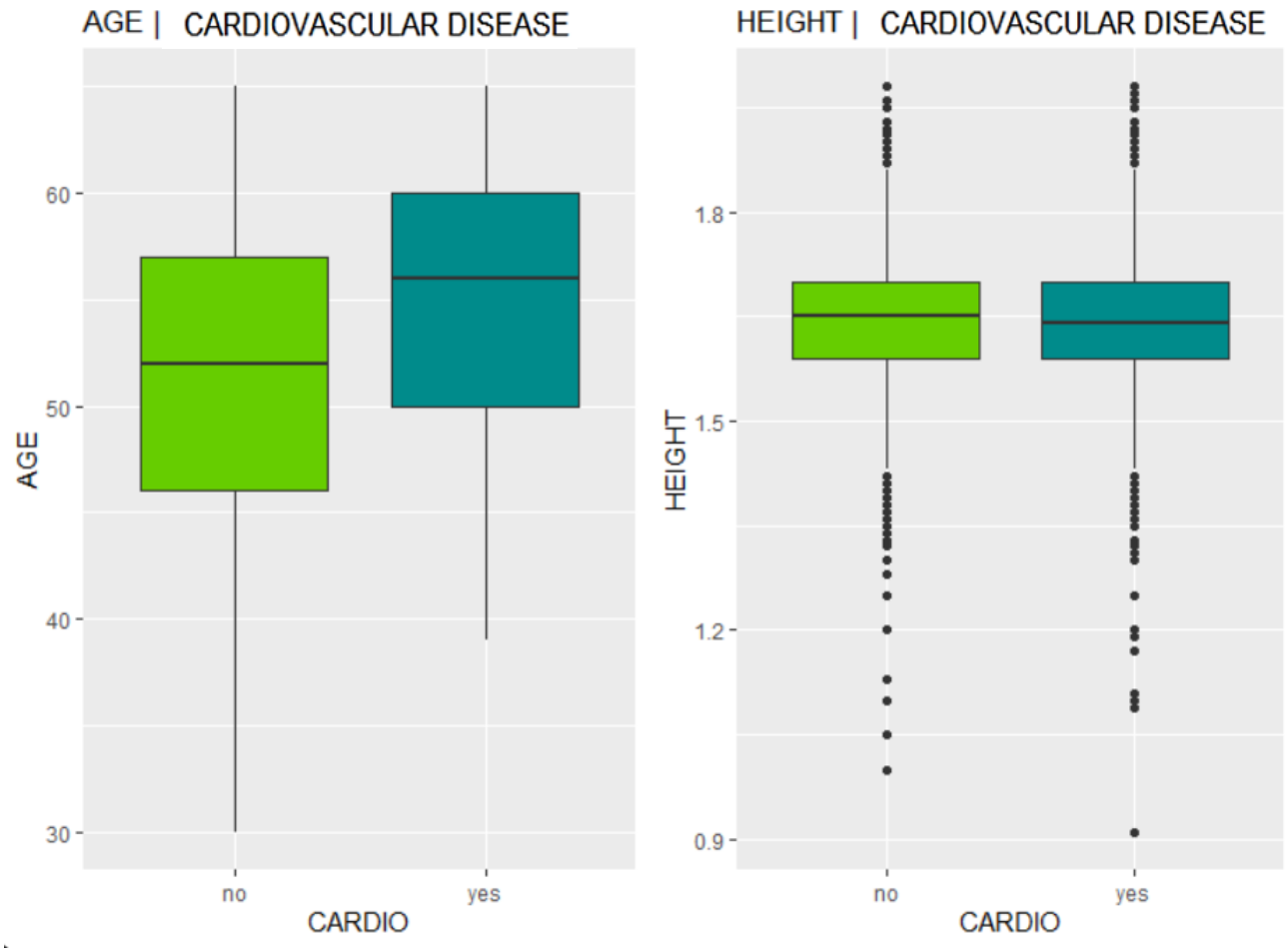
PLOT 2.1: DISCRIMINANT POWER OF EACH CATEGORICAL VARIABLE



PLOT 2.1 shows the conditional distribution of the presence of cardiovascular diseases with respect to each of the categorical variables. Some considerations can be made:

- Variables like gender, smoking, alcohol and physical activity, taken alone, doesn't seem to have any discriminant power with respect to the cardiological disease.
  In fact, as an example, knowing that a person is a female doesn't facilitate the classification because not only the risk of having a disease is the same for both male and female, but also the two categories are equally likely;
- Variables like cholesterol, BMI index and blood pressure seem to have a relevant discriminant power. As instance, a person with a well above normal level of cholesterol is far more likely to have a cardiovascular disease than a person with a normal level.
  It is also possible to notice that the risk of having a cardiovascular disease increases as BMI index, cholesterol level and blood pressure increase.

PLOT 2.2: DISCRIMINANT POWER OF EACH QUANTITATIVE VARIABLE



PLOT 2.2 shows conditional boxplots of both age and height with respect to the presence of a cardiovascular disease. It is possible to notice that while height does not seem to have any discriminant power, age may help in the classification of sick people. As instance if a person is 1.7 meters both categories (presence or absence of the disease) are equally likely.
Knowing instead that a person is older makes the presence of the disease more likely.

3. SINGLE DECISION TREE

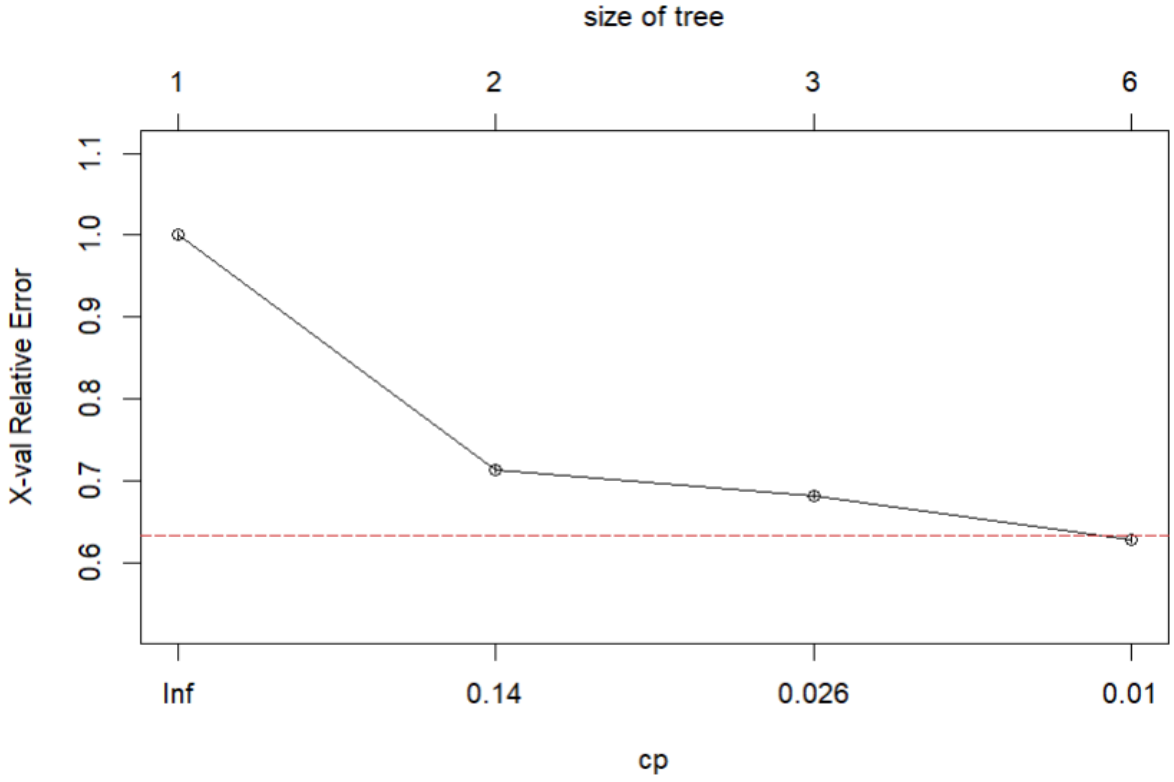
Decision trees are a machine learning algorithm that can be used for both classification and regression. The classification is carried out according to some splitting rules, by answering a series of questions regarding the value of variables chosen in order to maximize the reduction in the heterogeneity Gini index.
The objective is to partition the initial set of observations in order to obtain sets that are "pure" with respect to the values of the variable to predict, i.e. cardiovascular disease.

3.1 PRUNING AND OPTIMAL DECISION TREE

Usually the implementation of a decision tree consists of growing a complex tree, that is a tree with an elevated number of leaves and then prune it, in order to remove leaves that are less useful, obtaining a tree that is a trade-off between accuracy and complexity.
One of the main pruning techniques of the maximal tree identifies as potential optimal trees those having a cross-validated misclassification rate below a threshold computed as follows:

THRESHOLD = MINIMUM CV MISCLASSIFICATION RATE + ITS STANDARD ERROR

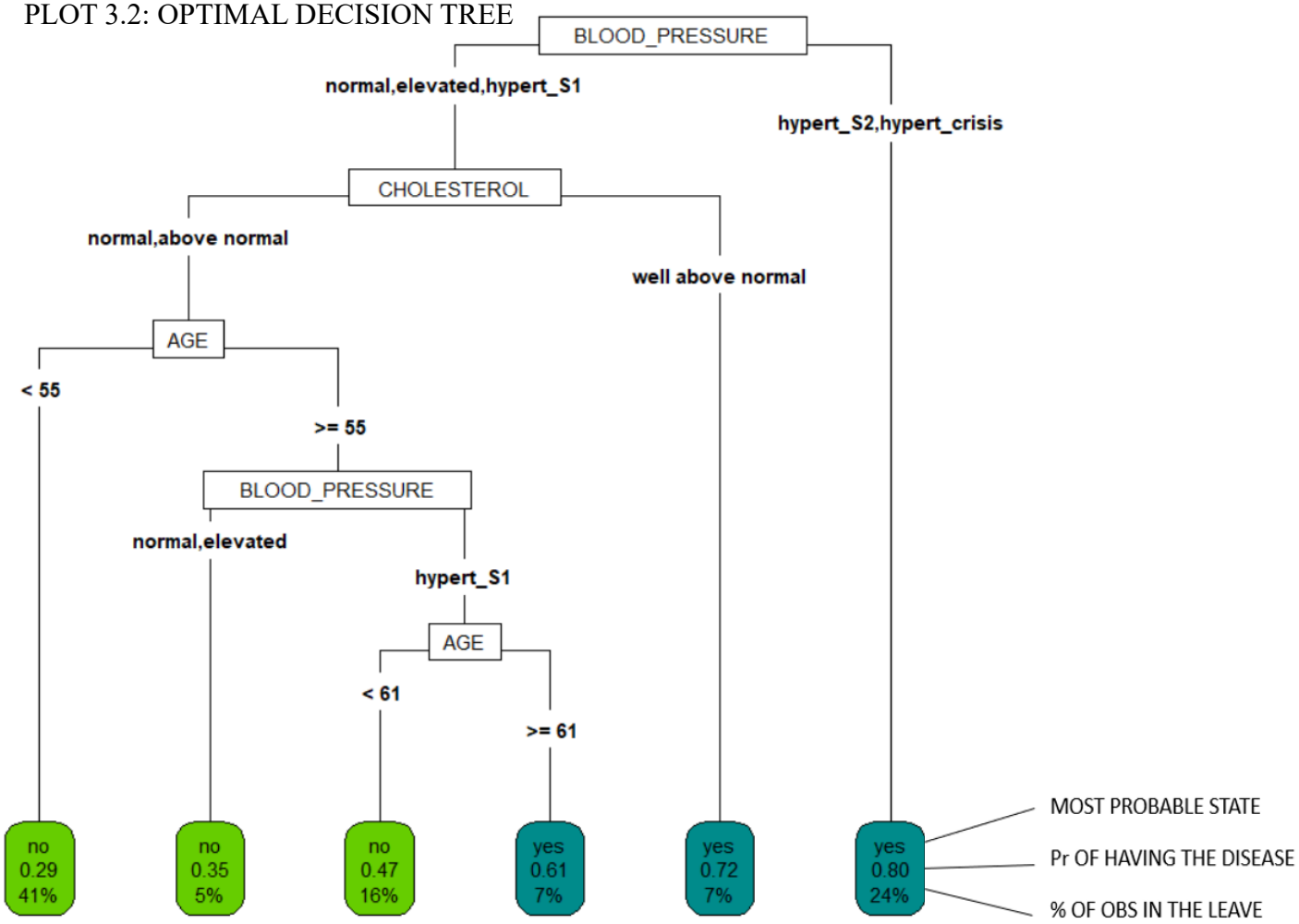
PLOT 3.1: MISCLASSIFICATION RATE AS A FUNCTION OF THE TREE COMPLEXITY



PLOT 3.1 shows that, in the analysed case, the maximal tree corresponds to the optimal tree with six leaves. One of the main advantage of decision trees is that the output can be represented graphically with some key elements:

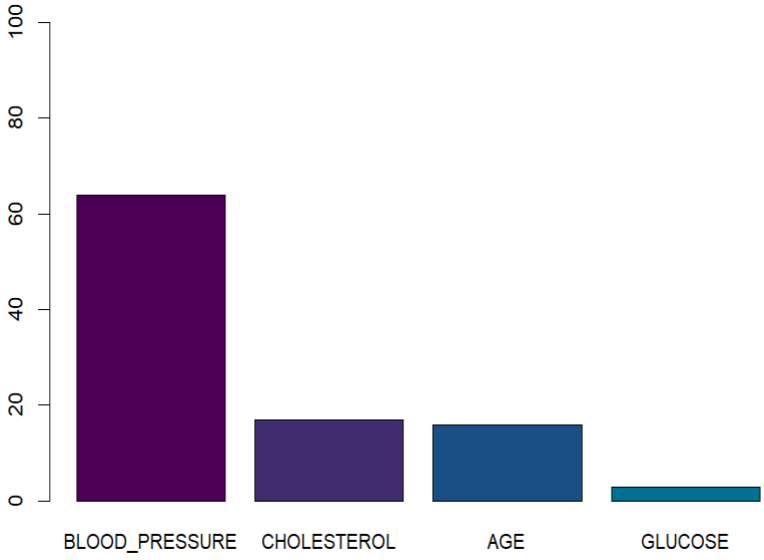
- Nodes, that are sets of observations to be partitioned;
- Branches, that are the conditions that allow partitions;
- Leaves, that are the final nodes, made up of "pure" nodes with respect to the presence of a cardiovascular disease.

PLOT 3.2: OPTIMAL DECISION TREE



PLOT 3.2 shows the selected tree. Considering the first splitting rule it is possible to notice that, as an example, patients with a more severe hypertensive status (second stage and crisis) will have 0.80 probability of having a cardiovascular disease while younger patients (<55 years) with a less severe hypertensive status, a lower cholesterol level will be more likely to be healthy.

PLOT 3.3: DECISION TREE VARIABLE IMPORTANCE



Blood pressure, cholesterol, age and glucose are the most important variables, i.e. those that determine the highest variation in the heterogeneity index when removed. The results partially overlaps those obtained in the exploratory analysis, when the discriminant power was assessed.

4. RANDOM FOREST

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

Random forests involve some hyperparameters that need to be tuned:

- Number of trees in the forest;
- Number of variables to consider at each split;
- Minimum size of each node in each decision tree;
- Sample fraction with respect to the original set of observations;
- Whether to sample observations with or without replacement.

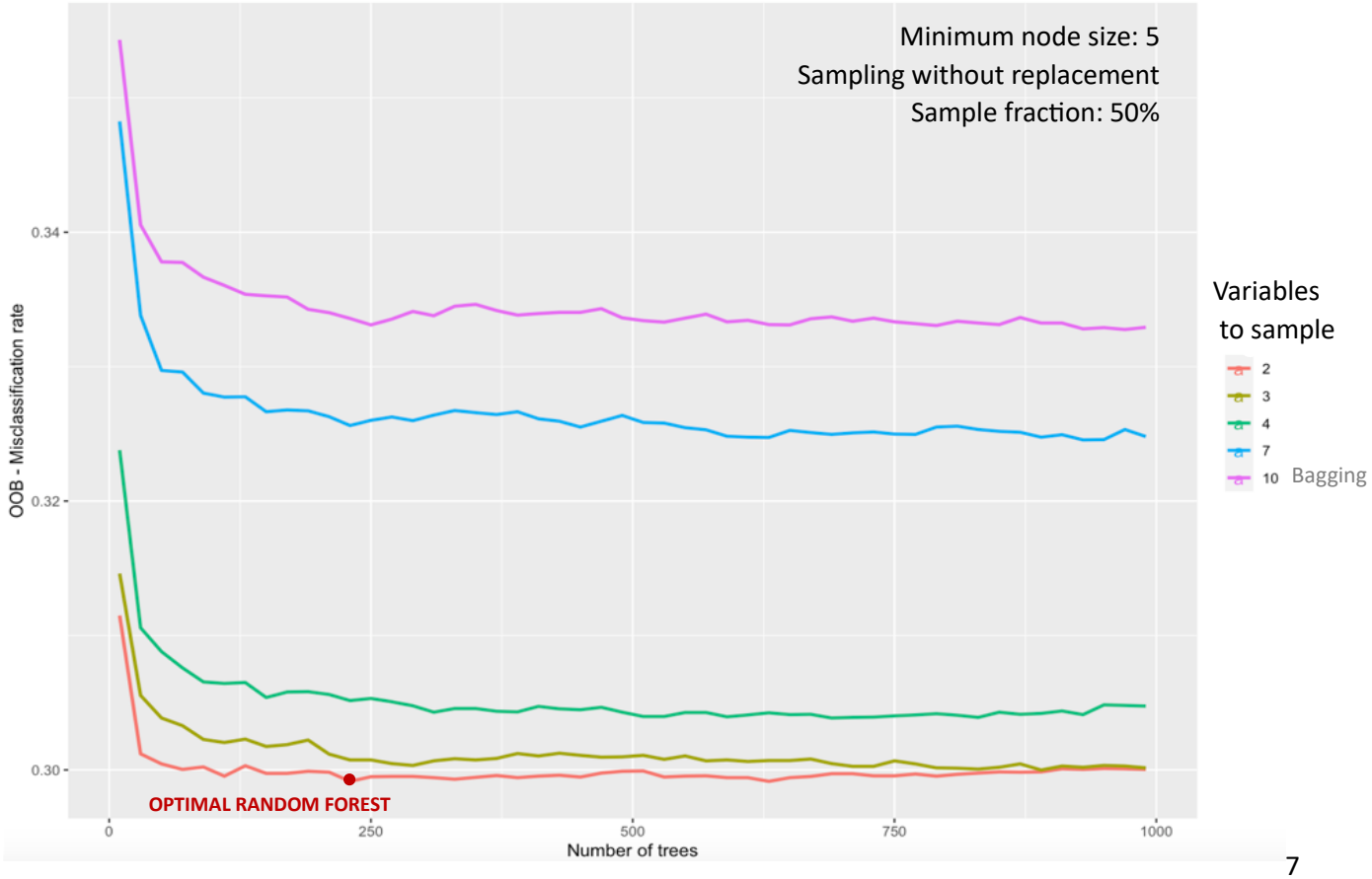
} They have the largest impact on results.

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

For computational reasons a two-steps tuning was implemented (final results are shown in TABLE 4.1):

- 1° step: minimum node size, sample fraction, sampling type;
- 2° step: given the hyperparameters previously tuned, the number of trees and of variables to sample at each split was tuned. PLOT 4.1 shows the performance of the forests when both the number of trees and of variables changes.

PLOT 4.1: RANDOM FORESTS PERFORMANCE COMPARISON

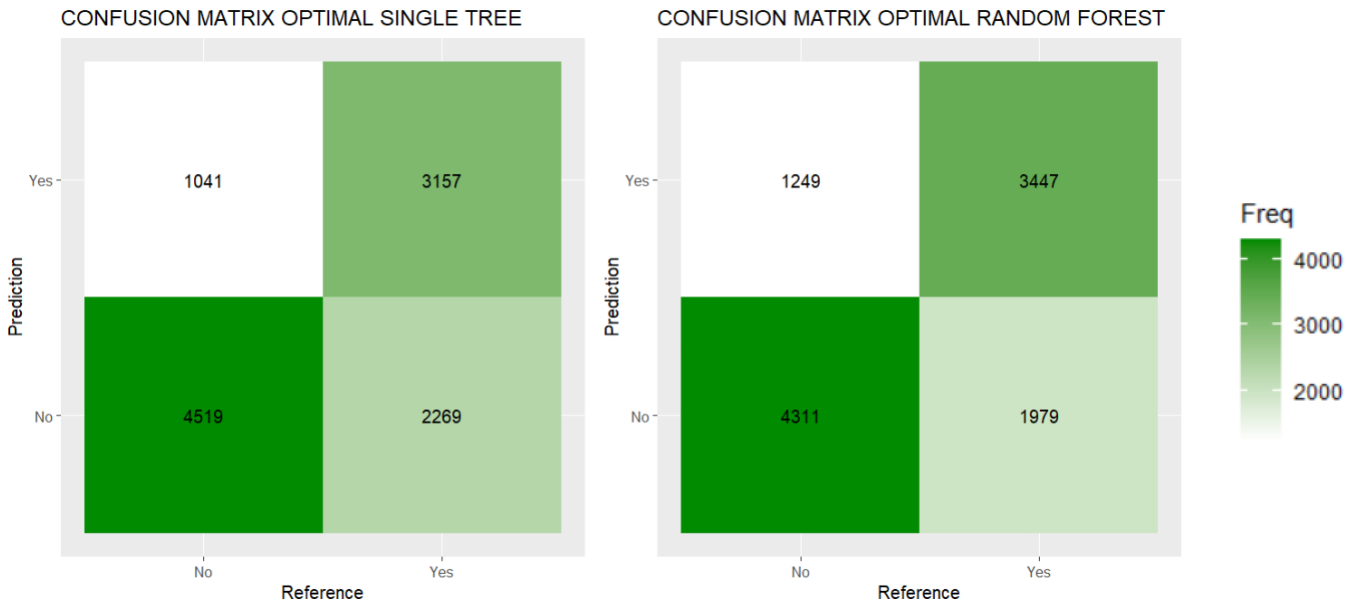


PLOT 4.1 highlights the performance of the random forests build with varying number of trees and of variables to sample at each split, given the hyperparameters tuned in the 1° step of the two-steps tuning. The performance was evaluated considering the proportion of out of bag observations wrongly classified, i.e. those observation not sampled in the building of the tree.
It is possible to notice that when the number of variables to sample increases, the performance decreases: bagging performs worse while forests whose trees are built sampling 2 variables at a time performs better. The optimal number of trees to include into the forest was about 230 (when the performance stabilizes).

TABLE 4.1: TUNED HYPERPARAMETERS

FEATURES OF THE SELECTED RANDOM FOREST	
MINIUM TREE NODE SIZE	5 observations
SAMPLING SCHEME	Without replacement
SAMPLE FRACTION	50% of the original sample size
NUMBER OF TREES	230
NUMBER OF VARIABLES TO SAMPLE	2

PLOT 4.2: CONFUSION MATRICES COMPARISON BETWEEN METHODS



PLOT 4.2 makes a comparison between confusion matrices of both the optimal decision tree and the random forest: they are representations able to compare the true health status with the one predicted by the model with respect to the test set observations.
Starting from these matrices it is possible to obtain some useful summary statistics (TABLE 4.2). Performances between the two methods are almost similar.
Using the more complex random forest algorithm, while both the proportion of correctly classified observations (accuracy) and the proportion of truly sick people identified (sensitivity) slightly increases, the proportion of healthy people identified by the model decreases (specificity).

TABLE 4.2: PERFORMANCE COMPARISON BETWEEN METHODS

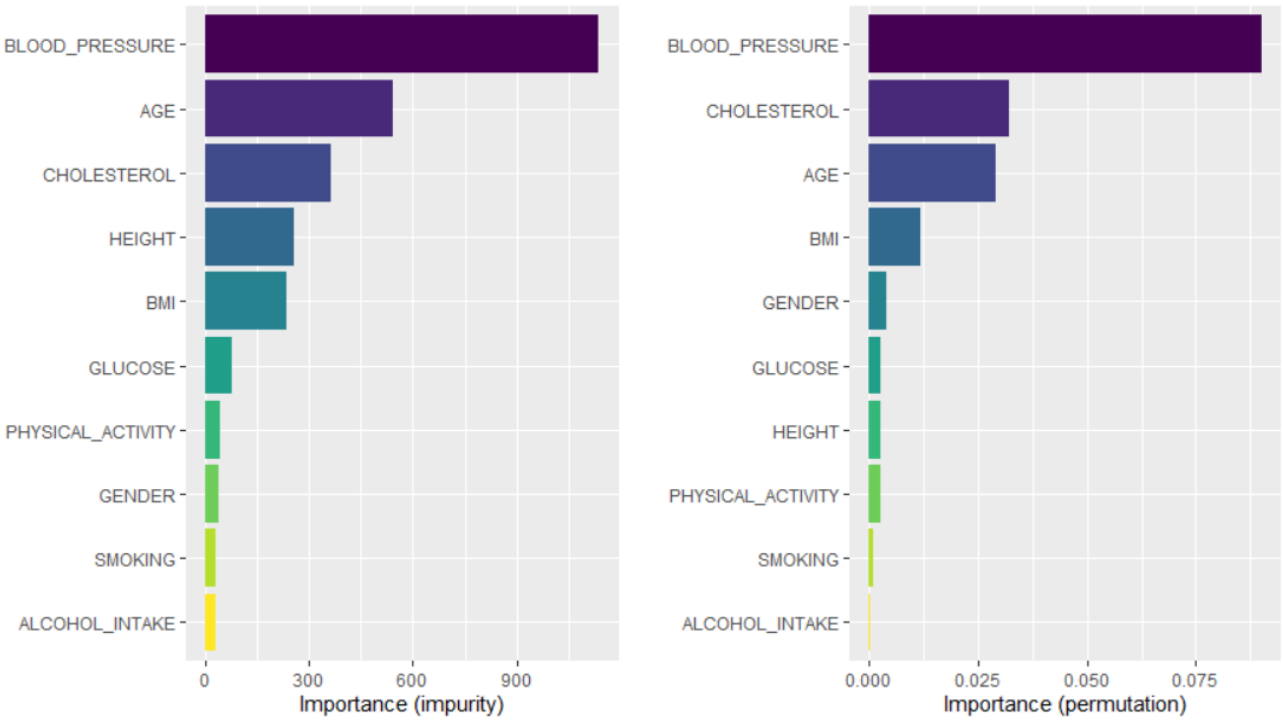
METRICS	DECISION TREE vs RANDOM FOREST	
ACCURACY	69.9%	70.6%
MISCLASSIFICATION RATE	30.1%	29.4%
SENSITIVITY	58.2%	63.5%
SPECIFICITY	81.3%	77.5%
+ PREDICTIVE VALUE	75.2%	73.4%
- PREDICTIVE VALUE	66.6%	68.5%

PLOT 4.3 shows the importance of each variable, computed using two different metrics:

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

Results are almost similar since the most important variables turn out to be those highlighted by the exploratory analysis, i.e. blood pressure, age, cholesterol while the less useful are smoking and alcohol intake.

PLOT 4.3: RANDOM FOREST VARIABLE IMPORTANCE



5. GRADIENT BOOSTED TREES: BASIC VERSION

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

Gradient boosted trees involve some hyperparameter to tune by the mean of a grid search algorithm:

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

The choice was based on the cross validated Bernoulli deviance loss function (TABLE 5.1) while the number of trees was that one starting from which the performance stabilises (PLOT 5.1).

PLOT 5.1: BASIC GRADIENT BOOSTED TREES ACCORDING TO THE NUMBER OF TREES

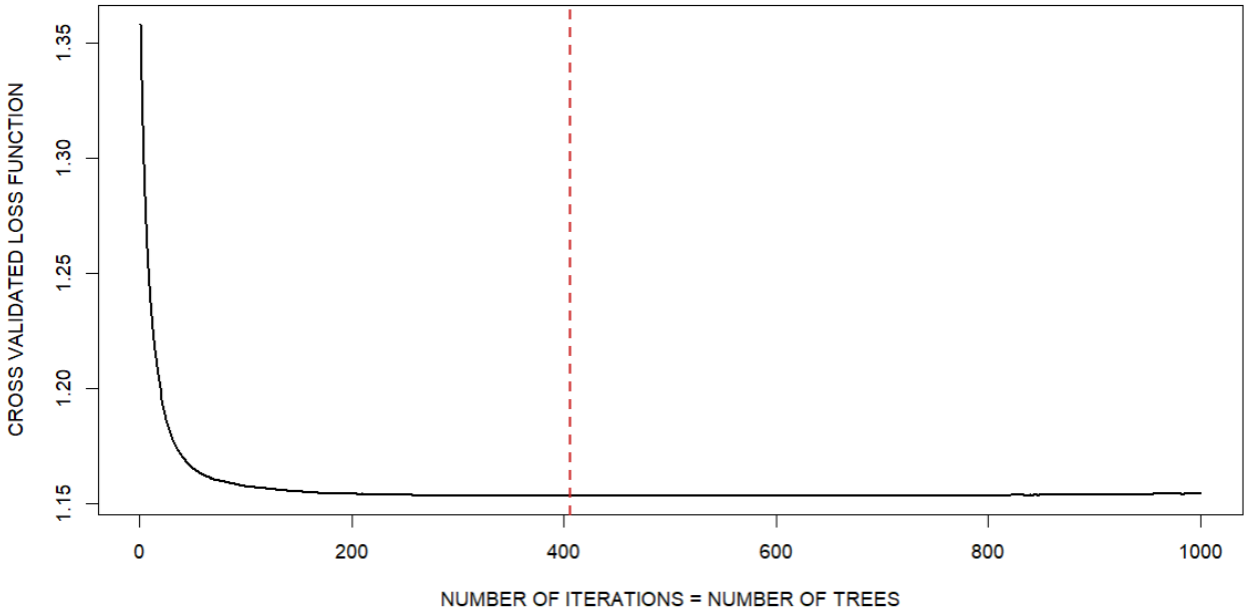


TABLE 5.1: TUNED HYPERPARAMETERS

FEATURES OF THE SELECTED RANDOM FOREST	
MINIUM TREE NODE SIZE	15 observations
TREE DEPTH	2
LEARNING RATE	0.1
MINIMUM NUMBER OF TREES NEEDED	406

6. GRADIENT BOOSTED TREES (STOCHASTIC VERSION)

The standard version of the Gradient Boosted Machine algorithm works by computing the gradient on the whole set of data. The following version consists of computing the gradient just on a random subsample of the original dataset, in order to obtain a method that is both faster and more robust with respect to the training data used. The hyperparameters that this method involves are the same of the previous one so the previously tuned values were used.
The additional hyperparameter consists of the sample rate, that is the amount of observation to sample at each iteration. It was tuned considering the accuracy value and was set equal to 0.25.

PLOT 6.1: CONFUSION MATRICES COMPARISON BETWEEN GBM VERSIONS

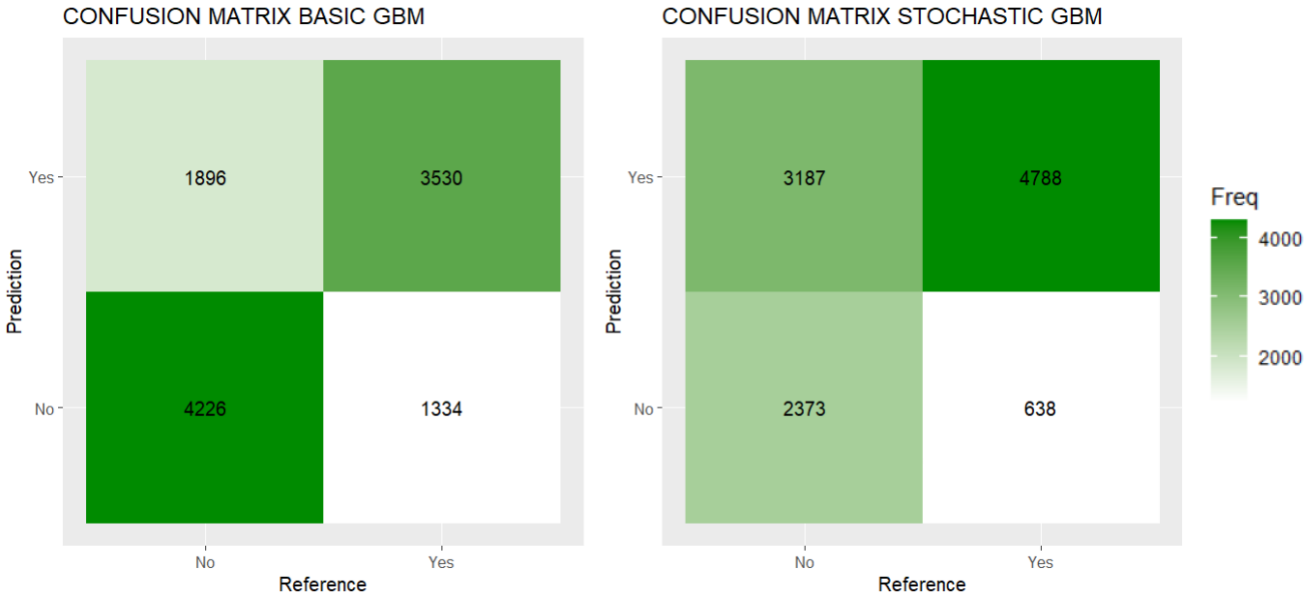
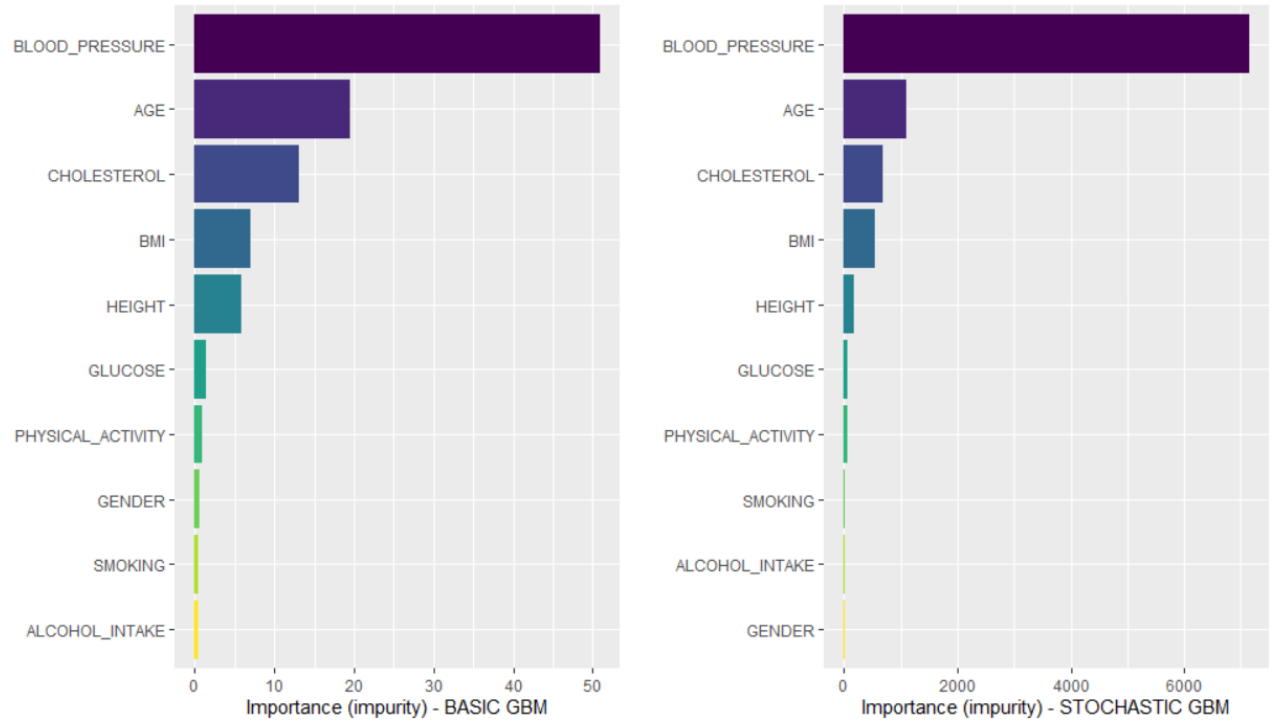


TABLE 6.1: PERFORMANCE COMPARISON BETWEEN GBM VERSIONS

METRICS	BASIC GBM vs STOCHASTIC GBM	
ACCURACY	70.6%	65.2%
MISCLASSIFICATION RATE	29.4%	34.8%
SENSITIVITY	72.6%	88.2%
SPECIFICITY	69.0%	42.7%
+ PREDICTIVE VALUE	65.1%	60.0%
- PREDICTIVE VALUE	76.0%	78.8%

TABLE 6.1 highlights summary performance statistics about the two GBM versions implemented, computed starting from the confusion matrices plotted in PLOT 6.1.
It is possible to notice that even if the accuracy decreases by 5 percentage points (from 70.6% to 65.2%), the sensitivity value increases significantly (from 72.6% to 88.2%):
stochastic GBM model may be preferred whether the main objective is to predict well the presence of the disease rather than the absence of it (specificity decreases).

PLOT 6.2: COMPARISON BETWEEN VARIABLE IMPORTANCE IN BOTH GBM VERSION



PLOT 6.2 shows the variable importance computed on the base of the change in the Gini index (impurity measure) when the variable is removed from the model.
The most important variables correspond to those highlighted by the other methods previously used: blood pressure, age, cholesterol and BMI.

CONCLUSION

The most accurate methods are random forest and basic GBM, even if the differences are not really relevant. There are however sufficient elements to suppose that some important variables were omitted, since some observations can’t be classified correctly by any method.