

GRADO EN INGENIERÍA EN TECNOLOGÍAS INDUSTRIALES

Classification Project

Authors:

Manuel Trabado de la Cruz

Alfonso Sanz Giner

Ignacio Ortiz de Zúñiga

Madrid

18/02/2021

Contents

[Introduction 3](#_Toc64578625)

[Exploratory analysis 3](#_Toc64578626)

[Factors levels 3](#_Toc64578627)

[Missing 3](#_Toc64578628)

[Outliers 3](#_Toc64578629)

[Analysis 4](#_Toc64578630)

[Regression model 5](#_Toc64578631)

[Multilayer Perceptron (MLP) 6](#_Toc64578632)

[KNN model 7](#_Toc64578633)

[Decision Trees 8](#_Toc64578634)

[XGBoost 9](#_Toc64578635)

[Random forest 10](#_Toc64578636)

[Support Vector Machines (SVM) 10](#_Toc64578637)

[Conclusion 11](#_Toc64578638)

# Introduction

This assignment focuses on a classification problem of a set of given information about, at first, relevant information to build distinct models to predict the output via different methods.

# Exploratory analysis

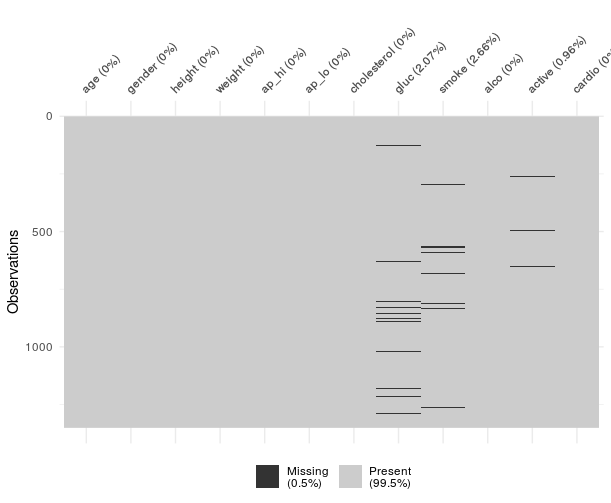
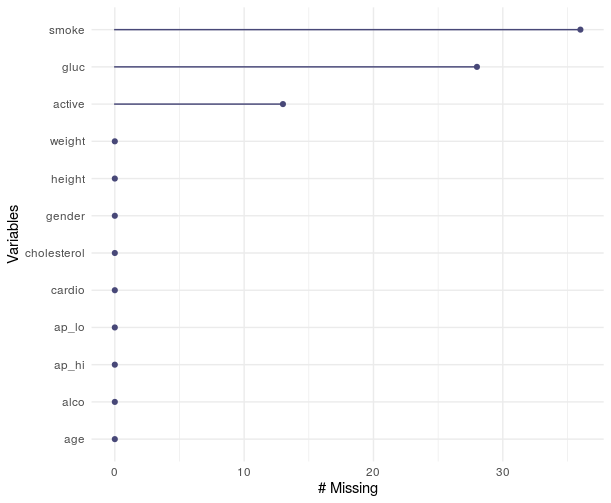
The dataset to analyse amounts to 1351 observations of 12 variables, of which only one is an output, the presence of cardiovascular disease, while the other are input information to use for the classification.

## Factors levels

The dataframe read form the csv is in a formatless table and therefore, factors and levels need to be determined for categorical variables such as the blood pressure, cholesterol, glucose, and the binary variables: gender, and smoking, drinking and physical activity.

## Missing

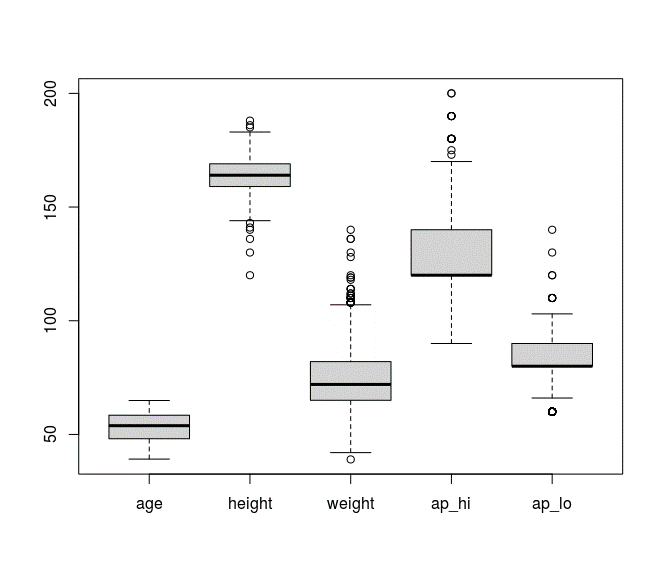
There are 77 missing values in total, 28 of the glucose, 36 of smoking and 13 of physical activity, all of them are exclusive, meaning there are not any observation with 2 NA values.



These observations mean only a 5.7% of the data and consequently removing them would not significantly affect the results, especially while using cross validation. Nevertheless, another more data-efficient involves filling the values with the mean (median as they are factors) of the respective variable or computing a KNN model to predict its most possible value.

## Outliers

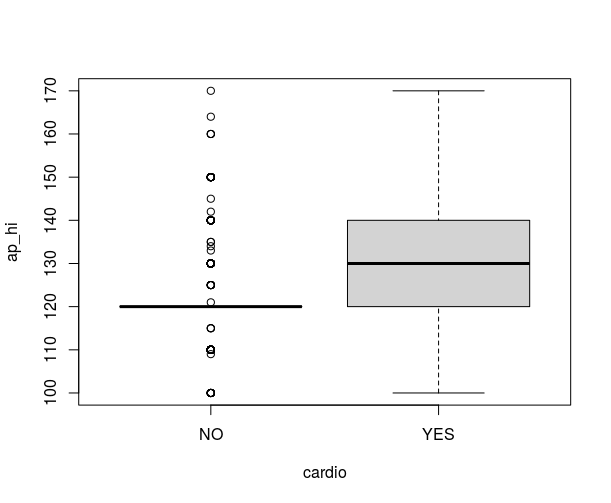
The second most noticeable defect when analysing the data is the presence of -1 values in the numeric variables age and height, with 37 and 31 cases respectively for a total of 68 or 5% of the total. Since the missing values were already predicted there 68 cases can be removed without a significant impact.

After doing so the outliers can be obtained individually for each numeric variable with boxplots

Finding out that there are no outliers in age, but there are some observations under 141 cm in height and above 185 cm (9 in total), 30 cases of a weight higher than 108 kg (1 below 39) and another 79 outliers in pa\_hi and ap\_lo.

Of all these we will only remove the outliers of lower height since the rest could be significant attributes to include in our model even as extreme cases.

Another option for detecting outliers is representing the numeric variables group by the output (cardio) but although these removes the height outliers it also arrives to the conclusion that the majority of the NO results are at ap\_hi=120, but we cannot remove the rest as they are an important amount of data.

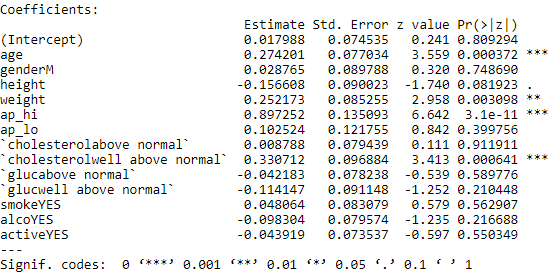


## Analysis

The correlation and distribution analysis shows the highest relation to the output “cardio” in the variables age, ap\_hi, cholesterol. Therefore, these variables are going to be the more relevant and used in the models.

# Regression model

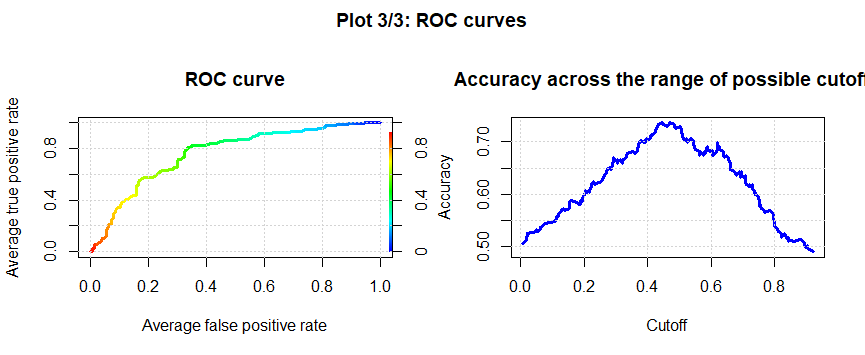
The first model involves a simple linear regression of all input variables to determine which are significant and which can be removed from the model:



The results of this model shows that the most significant attributes are age, ap\_hi, cholesterol and less but still relevant weight. This model also has an accuracy of 0.7 and kappa of 0.41.

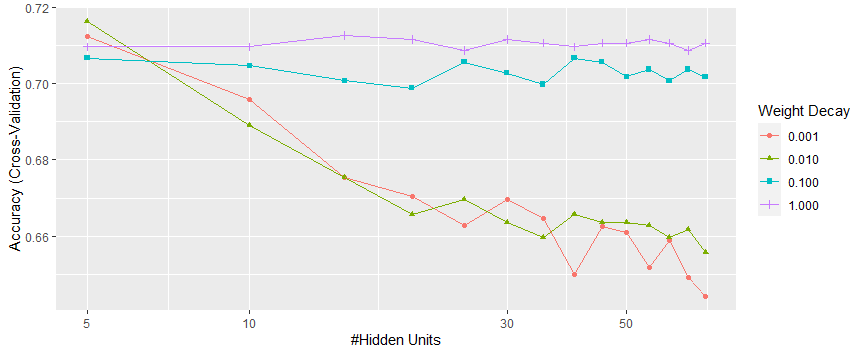
The iteration process requires the removal of the less significant variables (with the higher p value) until finding a simpler and accurate model.

Simplifying the model to use only the 3 variables mentioned before we arrive to a model with an accuracy of 0.73 and kappa of 0.46 while having an acceptable performance on the test predictions:

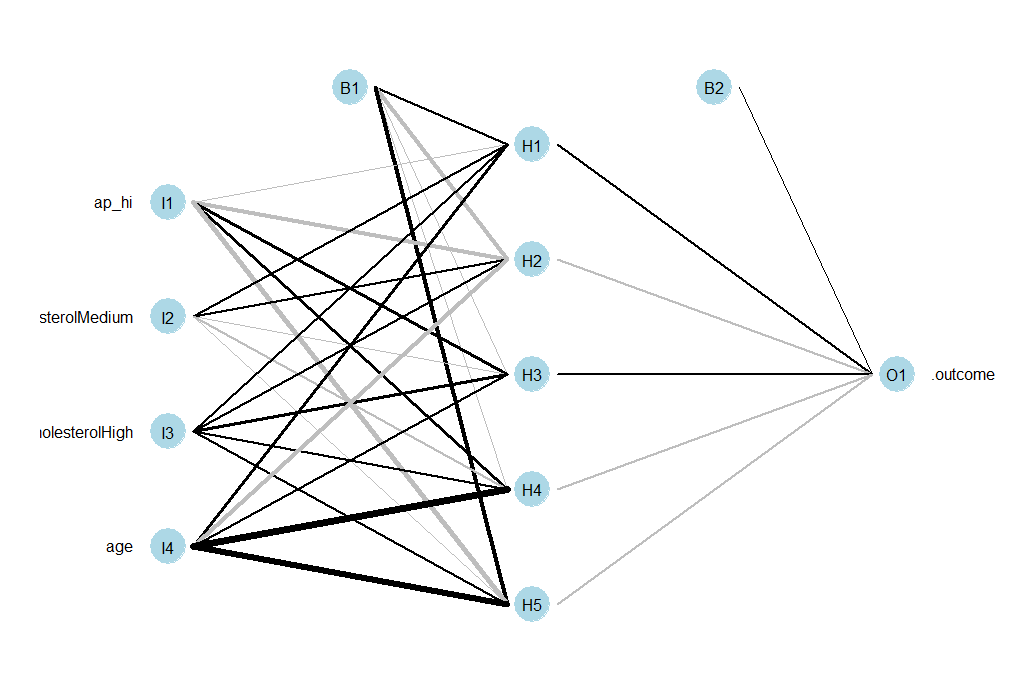


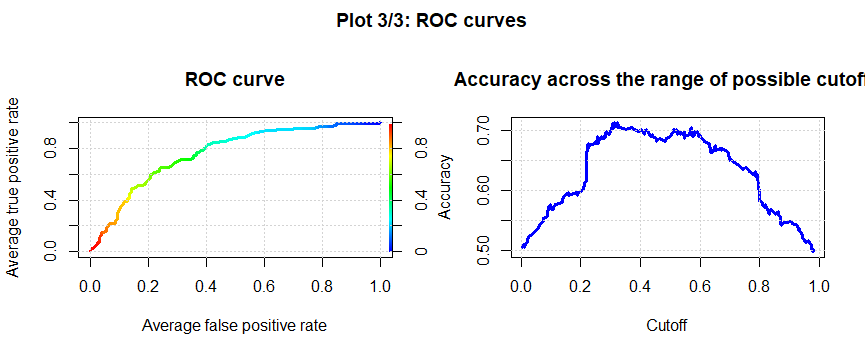
# Multilayer Perceptron (MLP)

The second model attempted is the Multilayer Perceptron method which requires to find the optimal number of neurons and decay. For this, we train a sequence between 5 and 70, and decay of 0.001, 0.01, 0.1 and 1.



The optimum is achieved at 5 neurons and a decay of 0.01, to accomplish a reasonable equilibrium between complexity and a cross validation accuracy (0.72).

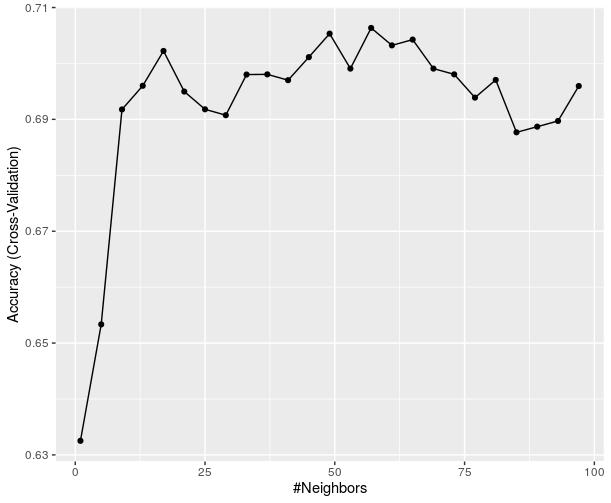




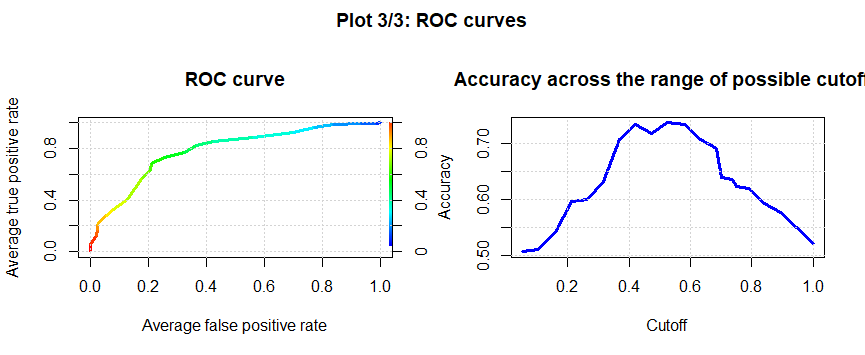
# KNN model

The next model will be using the k-nearest neighbours’ method and the same 4 relevant inputs. With an optimum value of k of 19 it is achieved an accuracy of 0.74 and kappa of 0.47.

The accuracy curve is not completely legible with multiple maximums but for keeping the model simple we selected the first maximum instead of the absolute maximum at ~60 neighbours.

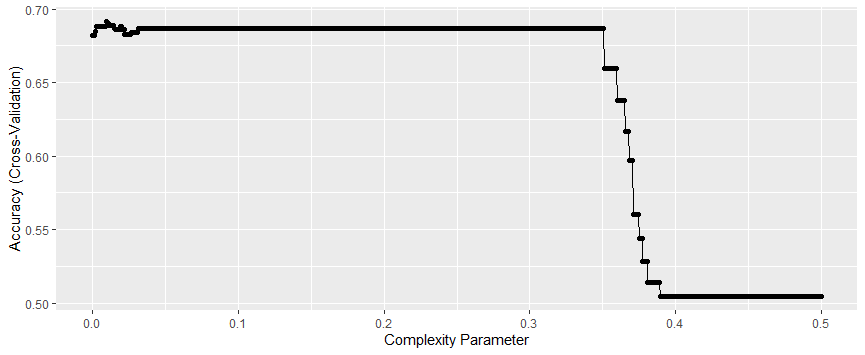


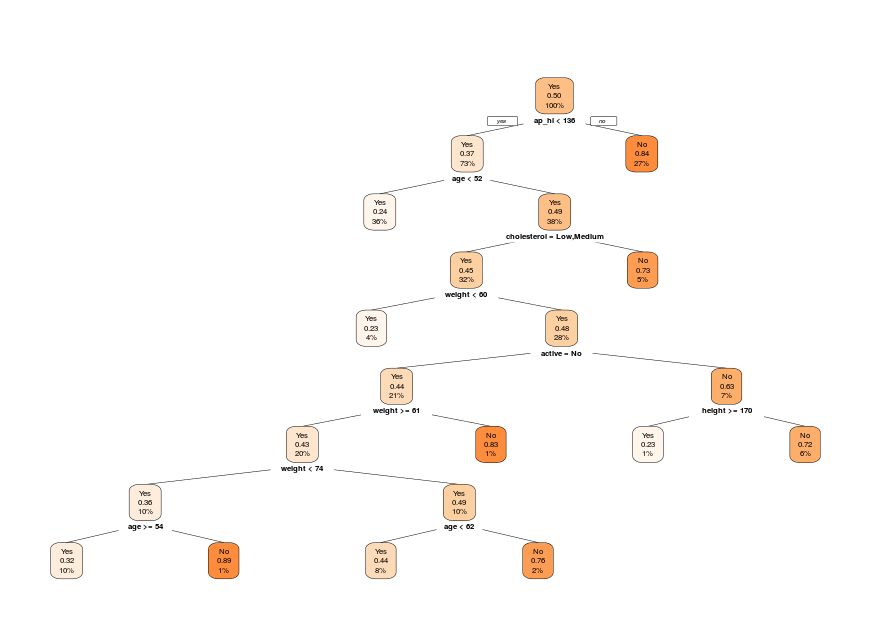
This model proves to be a most accurate method of predicting the cardio variable in comparison with MLP and regression.



# Decision Trees

For the Decision Tree it is required to select a complexity parameter. We plotted the graph of accuracy of the trees with all input variables for a sequence from 0 to 0.5 in complexity and the optimum selected was 0.05 with an accuracy of almost 71%.

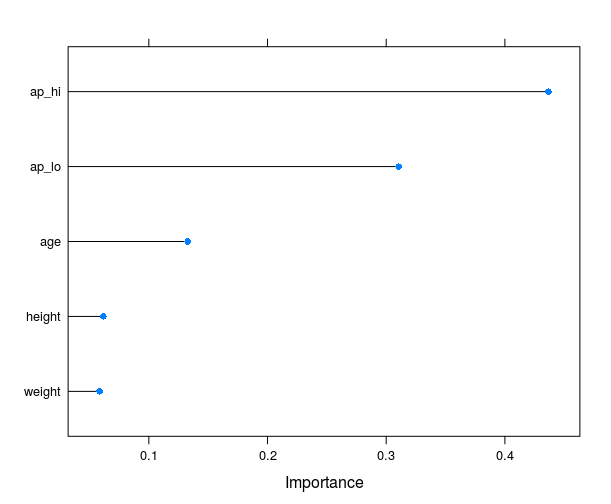




Although by manual testing and reducing the dimension of the problem by introducing only the age and ap\_hi variables we found an optimum of 0.0095 with an accuracy of 72% that has only one rule node: ap\_hi < 136. This tree may seem too simple, but it still achieves a respectful accuracy and shows how important ap\_hi is to our model.

# XGBoost

Using the XGBoost method once again we found out that the most important variables for this model (and the rest) are the blood pressure variables (ap\_hi and ap\_lo):



Even using these variables, the results of this model are not especially successful with an accuracy of 0.71 during training and 0.706 with the test data.

# Random forest

For the Random Forest method all input variables were utilized. After the model fit, the optimum value of predictors was set to 2 with an accuracy in cross validation of 0.725.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

In this model we could appreciate the effect of overfitting when comparing the values of accuracy in training (0.93) and test (0.74). This is due to the excessive complexity and overtraining in this model.

# Support Vector Machines (SVM)

For the SVM method the process to determine the optimum value of the hyperparameter C was the same as in the previous models, we iterated through several values of C and compare the accuracy obtained. The best value of C found was 0.01 with a cv accuracy of 0.72 and kappa 0.44.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

The final accuracy obtained in the confusion matrix was 0.7216, a value similar to the previous models.

# Conclusion

Gráfico

Descripción generada automáticamente

Of all the models used in this project the one with the highest mean accuracy and kappa values in cross validation is SVM. This may be due to the nature of the SVM model, that divides the input space in two regions allowing some rule violations in the border.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
|  |  | Accuracy (%) | |  |  | Area under ROC | |
|  |  | Trainning | Test |  |  | Trainning | Test |
|  | LogReg | 71,95 | 72,94 |  | LogReg | 0,7815 | 0,7588 |
|  | MLP | 72,93 | 69,41 |  | MLP | 0,8061 | 0,7615 |
|  | KNN | 73,5 | 73,72 |  | KNN | 0,8077 | 0,7808 |
|  | Decission tree | 69,89 | 72,16 |  | Decission tree | 0,7213 | 0,7362 |
|  | Random forest | 92,57 | 73,72 |  | Random forest | 0,9833 | 0,7742 |
|  | SVM | 72,04 | 72,16 |  | SVM | 0,7467 | 0,7244 |
|  | XGBoost | 71,75 | 70,59 |  | XGBoost | 0,806 | 0,7669 |

On the other hand, after analysing the results obtain in the confusion matrices the model with the higher accuracy in test is the KNN model. This model is able of generalize more than the other models because this model does not divide the space in two regions, but it compares a set of points near the point the model is evaluating. This process is good for this dataframe due to the low correlation between most of the variables.

Comparing the areas under the ROC curves of all the models we can see that the model with the biggest area in test is the KNN model.

Attending to the area under the ROC curve and the accuracy obtained in the test confusion matrix we have determined that the best model is the KNN.

Nevertheless, all the models analysed are similar in accuracy which determines that the best accuracy achievable is between 70-75%.

It has come to out attention that even a simple linear regression or a one node decision tree is capable of predicting the results with aprox. 70% accuracy. And it could be possible that with a bigger dataset we could have obtained a higher accuracy in the models.

