Assignment 2: Evaluating Risk Factors in Health Care *

Emanuel Tomé †, João Carvalho ‡

University of Porto, 15^{th} May 2021

1 Introduction

The Framingham Heart Study is an association under the National Heart, Lung and Blood Institute (USA), born in 1948 (Massachusetts), whose goal is to study "the common factor or characteristics that contribute to cardiovascular disease (CVD)" [1], over generations. For this study, there were selected 5209 healthy men and women, in other words, who had not developed yet apparently, any cardiovascular symptoms, nor had suffered hearth attacks or strokes. The careful monitoring of its population over the time, it was crucial thus enabling the identification of major CVD risk factors and their effects towards other variables.

Having said that, the purpose of this assignment is to create predictive models based on supervised learning classification techniques in order to predict the 10 years risk of CVD using demographic, behavioural and medical history data of the patients. This work is divided in six sections: introduction, which is the present section; the exploratory data analysis, where a brief description of the dataset set as well of its descriptive statistics are presented; Logistic Regression, where all the details of the procedures to fit a Logistic Regression is described; State of the Art model: Decision Trees, where the details of fitting a decision tree to the data are also described; Feature Importance, where Permutation Feature Importance is applied in order to know the most relevant risk factors; and finally the Conclusions, where the main achievements are pointed out.

2 Exploratory data analysis

The Framingham Heart study dataset has 16 variables and 4240 instances. It is composed of seven binary variables (`male`, `currentSmoker`, `BPMeds`, `prevalentStroke`, `prevalentHyp`, `diabetes` and `TenYearCHD`), five ratio-scaled and continuous numerical variables (`totChol`, `sysBP`, `diaBP`, `BMI` and `glucose`), three ratio-scale and discrete numerical variables (`age`, `cigsPerDay` and `heartRate`) and one ordinal variable (`education`). The meaning of each variable is as follows [2]: `male`: 1 if the gender is male, 0 if it is female; `Age`: Age of the patient; `Current Smoker`: whether or not the patient is a current smoker; `Cigs Per Day`: the number of cigarettes that the person smoked on average in one day; `BP Meds`: whether or not the patient was on blood pressure medication; `Prevalent Stroke`: whether or not the patient had previously had a stroke; `Prevalent Hyp`: whether or not the patient was hypertensive; `Diabetes`: whether or not the patient had diabetes; `Tot Chol`: total cholesterol level; `Sys BP`: systolic blood pressure; `Dia BP`: diastolic blood pressure; `BMI`: Body Mass Index; `Heart Rate`: heart rate; `Glucose`: glucose level; `TenYearCHD`: 10 year risk of coronary heart disease CHD (1, means "Yes", 0 means "No").

This dataset comprises a classification problem where the target variable is `TenearCHD`. In other words, the objective is predict the patient's 10 year risk coronary heart disease given behavioural and medical history of the patient.

In Table 1 are presented the descriptive statistics of all variables. The first thing that can be noticed is that there are some variables with missing values, namely `education` (2.48% of missing values), `cigsPerDay` (0.68%), `BPMeds` (1.25%), `totChol` (1.18%), `BMI` (0.45%), `heartRate` (0.02%) and `glucose` (9.15%). We decide to remove any row with missing values, ending with a dataset with 3658 patients. It can also be noticed from Table 1 that any of the variables change sign, being all of them positive. The range of variation of the variables is considerably different, therefore standardisation will be applied to the non-binary variables before fitting the models. For instance, while the variable `totChol` has a range of variation between 107 and 696, the variable `BMI` has a range of variation between 44 and 143.

*This work was submitted on the framework of the course Data-Driven Decision Making

†Emanuel Tomé is a student at the Faculty of Sciences of the University of Porto. Currently, he is enrolled in the 1^{st} year of the Master's in Data Science (e-mail: up200702634@edu.fc.up.pt).

‡João Carvalho is a student at the Faculty of Sciences of the University of Porto. Currently, he is enrolled in the 1^{st} year of the Specialisation on Computational Statistics and Data Analysis (e-mail: up 202009073@edu.fc.up.pt).

Since the Minimum and Maximum values of some variables are far from the first and third quartiles, respectively, those variables seem to have outliers. In fact, this is confirmed by the variables plots and box-plots presented in Appendix A (sections A.1 and A.2, respectively), namely variables `totChol`, `sysBP`, `diaBP`, `BMI`, *heartRate* and `glucose`. However, since the outliers may be related with the patients that it is intended to identify, all of them were kept.

Table 1: Descriptive statistics of the variables.

	count	mean	std	min	0.25	0.5	0.75	max
male	4240	0.429	0.495	0	0	0	1	1
age	4240	49.58	8.573	32	42	49	56	70
education	4135	1.979	1.02	1	1	2	3	4
currentSmoker	4240	0.494	0.5	0	0	0	1	1
cigsPerDay	4211	9.006	11.922	0	0	0	20	70
BPMeds	4187	0.03	0.17	0	0	0	0	1
prevalentStroke	4240	0.006	0.077	0	0	0	0	1
prevalentHyp	4240	0.311	0.463	0	0	0	1	1
diabetes	4240	0.026	0.158	0	0	0	0	1
totChol	4190	236.7	44.591	107	206	234	263	696
sysBP	4240	132.355	22.033	83.5	117	128	144	295
diaBP	4240	82.898	11.91	48	75	82	90	142.5
BMI	4221	25.801	4.08	15.54	23.07	25.4	28.04	56.8
heartRate	4239	75.879	12.025	44	68	75	83	143
glucose	3852	81.964	23.954	40	71	78	87	394
TenYearCHD	4240	0.152	0.359	0	0	0	0	1

This dataset is highly unbalanced, where the target variable `TenYearCHD` has about 15% of values from class 1 and the remaining 85% from class 0. This is an important issue and should be considering when fitting and evaluating the classifiers. Note that the patients that were removed due to missing values did not change this percentage of each class in the target variable. Finaly, in Figure 39 (Appendix A.3) are presented a a grid of the pairwise relationships of the numeric variables, where it can be noticed that the two classes of the target variables are not clearly separated by any of this variables.

3 Logistic Regression

Logistic Regression is a linear model for classification and assumes a linear decision boundary. Since the the linear assumption is very strong, the logistic regression is unlikely to overfit and therefore it tends to create a very simplistic model that has high bias and low variance errors. Since the data being analysed is unbalanced, in order to fit the model a grid search procedure with a Stratified 5-fold cross-validation strategy was used in order to aid in the selection of the hyperparameters of the Logistic Regression. The evaluation metric used was the f1-score and the following hyperparameters were considered in the grid search:

- C - Inverse of the regularisation parameter - between 0.1 and 20, by steps of 0.1;
- fit_intercept - true or false;
- class_weight - None or Balanced (the balanced mode adjust weights inversely proportional to class frequencies);

The evaluation metric f1-macro was also considered and similar results were obtained. The output hyperparameters of the grid search are the following: $C = 1.9$, fit_intercept: true and class_weight 'balanced'.

We then fit the best model obtained by the grid search to the train dataset and obtained the classification report and confusion matrix presented in Tables 2 and 3, respectively. This results were obtained using standardising the non-binary variables before fitting the model. Similar results were obtained when we did not standardised the variables (see Appendix A.2). However, we kept this model in order to be able to analyse the importance of each variable in section 5 and compare that king of analysis with the Permutation feature importance.

Considering now the obtained metrics presented in Tables 2 and 3, poor metrics were obtained especially to class 1. One can also notice that 355 patients out of 1098 are misclassified. From those, the more important misclassified may be the 54 from class 1 that was classified as class 0 since the ones wrongly classified as class 1 may be submitted to complementary medical exams.

Table 2: Classification Report - Logistic Regression.

Label	precision	recall	f1-score	support
0	0.93	0.69	0.79	931
1	0.28	0.69	0.40	167
accuracy			0.69	1098
macro avg	0.61	0.69	0.60	1098
weighted avg	0.83	0.69	0.73	1098

Table 3: Confusion matrix - Logistic Regression.

		Predicted label	
		0	1
True label	0	639	292
	1	51	116

4 State of the art model: Decision Trees

Decision Trees is a flowchart similar to roots of plants, where in each node there is a statistical test on an attribute of the given data, and in each branch the outcome of the test. The ground nodes of a certain diagram are known as the leaf nodes. When the response variable y is categorical/binary, there is a classification type of problem, if y is continuous, there is a regression type of problem. Decision Trees happen to be flexible, capable of handle high dimension problems, and usually select the relevant variables. Due to the computational advances, it is very used nowadays.

Following the same procedure as in Logistic Regression, and having non-binary standardized variables, it was used a grid search method with 5 cross-validation splits strategy and f1_macro as its evaluation metric, with the aim of setting the hyperparameters for the model, those being:

- criterion - The function to measure the quality of a split - Gini or Entropy;
- class_weight - Weights associated with a class - Balanced or None;
- ccp_alpha - Complexity parameter used for Minimal Cost-Complexity Pruning - range between 0 and 10, by 0.0025.

The output hyperparameters of the grid search are: criterion: Gini, class_weight: Balanced, ccp_alpha = 0.0075. The following plot represents the Decision Tree diagram.

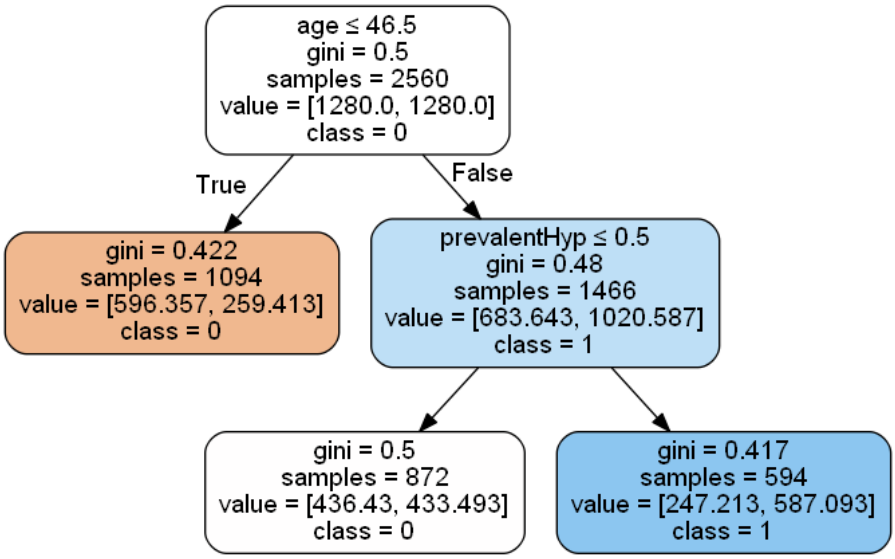

Figure 1: Decision Tree Diagram.

Given these hyperparameters, the model is obtained and a Classification Report and a Confusion Matrix in generated in Tables 4 and 5, respectively.

Table 4: Classification Report - Decision Tree.

Label	precision	recall	f1-score	support
0	0.88	0.78	083	916
1	0.30	0.47	0.37	182
accuracy			0.73	1098
macro avg	0.59	0.63	0.60	1098
weighted avg	0.79	0.73	0.75	1098

Table 5: Confusion matrix - Decision Tree.

		Predicted label	
		0	1
True label	0	719	197
	1	97	85

Regarding the results in Table 5, there were 87 subjects being put wrongly in class 1, in other words, healthy people classified as risky. However, the reall issue here comes when there are 136 subjects misclassified as class 0 i.e. being considered healthy when in fact, they are risky.

5 Feature Importance

Feature importance scores are important when building a predictive model since they allow to have insight into the data, insight into the model and is the basis for dimensionality reduction and feature extraction. Therefore, Feature importance is useful for better understanding of the data, for better understanding of the model and to reduce the number of input features. In this work, we start to look to the Logistic Regression coefficients and the applying Permutation Feature Importance (PFI). PFI is a model inspection technique and is useful to break the relationships between the features and the target. Although the PFI is a model agnostic technique, it does not reflect to the intrinsic predictive value of a feature by itself but how important this feature us for a particular model [3].

5.1 Feature Importance using the Logistic Regression coefficients

In Figure 2 are presented the obtained coefficients of the Logistic Regression model. From the analysis of that Figure, the majority of the variables are important, with exception of variables `diaBP`, `currentSmoker` and `heartRate` since their coefficients are small. These results are considerable different from the ones obtained using Permutaion Feature Importance, as will be seen in the next section.

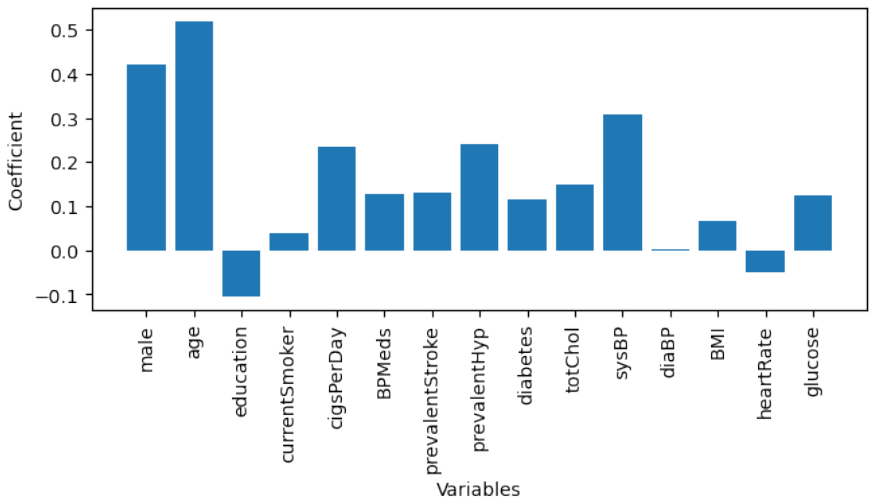

Figure 2: Coefficients of the Logistic Regression model obtained in section 3.

5.2 Permutation Feature Importance

In Figures 3, 4 and 5 are presented the obtained coefficients of the PFI using the model trained in section 3 for all dataset, for the training dataset and for the test dataset, respectively. It can be noticed that the variables `age`, `totChol`, `sysBP` have high coefficients in the three datasets. There are also more differences between the test dataset and the other two, as it was expected. For the text dataset the variables `male`, `glucose`, `heartRate` and `cigsPerDay` has also importance. Note that positive values of this coefficients indicate that the variable is important to predict class 1 and negative values indicate that the variable is important to predict the class 0 [4]. It is also interesting to note that some variables, such as `education`, have different signals for the training and test datasets.

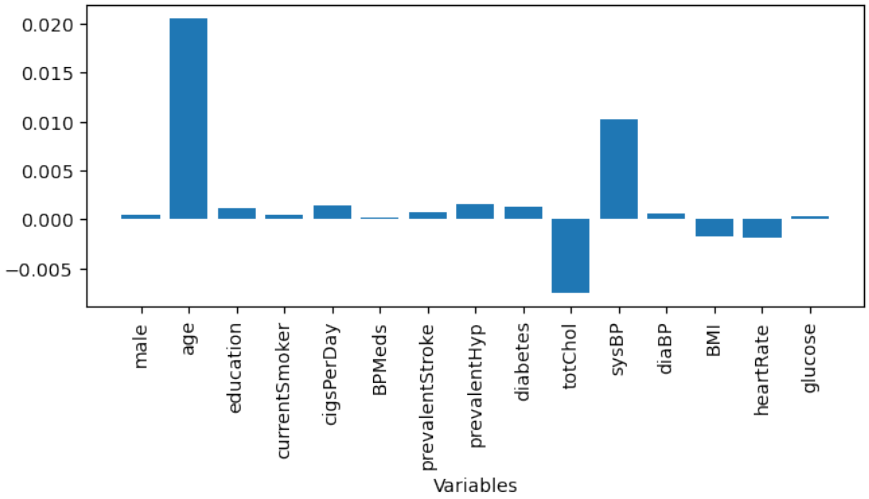

Figure 3: Feature Importance Permutation with all dataset.

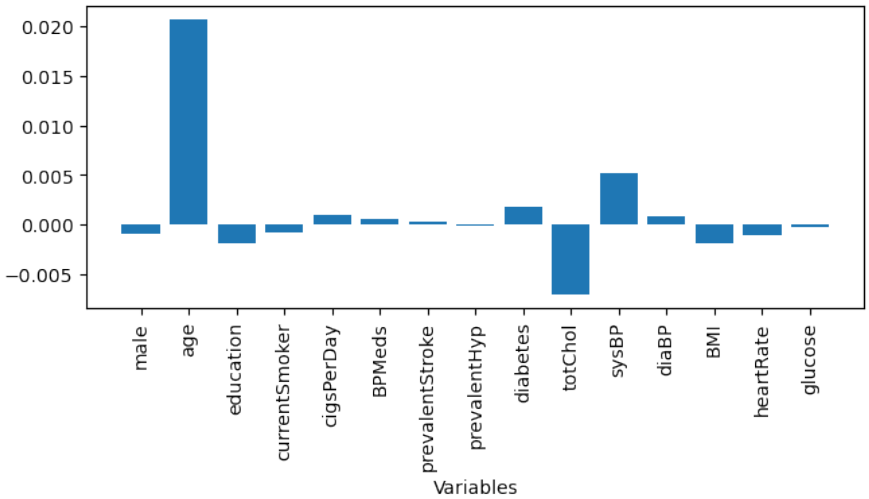

Figure 4: Feature Importance Permutation only with the train dataset.

Comparing to the risk factors pointed out by QRISK3 [5], which are age, systolic blood pressure, smoking status and ratio of total serum cholesterol to high-density lipoprotein cholesterol together with body mass index, ethnicity, measures of deprivation, family history, chronic kidney disease, rheumatoid arthritis, atrial fibrillation, diabetes mellitus, and antihypertensive treatment, we may conclude that these analysis identified the main CVD risk factors.

In Tables 8 and 9 (see Appendix C) are presented the Classification reports obtained from Logistic Regression using only the variables `male`, `age`, `cigsPerDay`, `totChol`, `sysBP`, `heartRate` and `glucose` as input variables. This model was obtained using the same procedure reported in section 3. However, the performance of the obtained model using only the referred variables is similar to the one obtained previously.

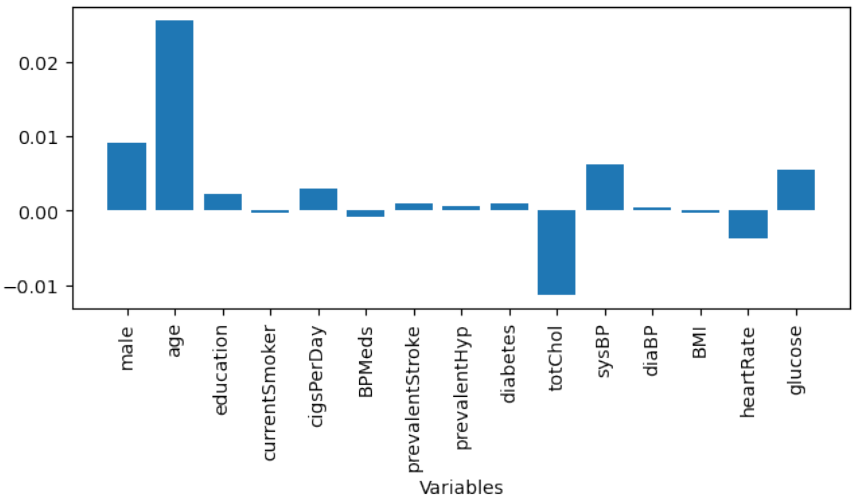

Figure 5: Feature Importance Permutation only with the test dataset.

5.3 Feature Importance using the Decision Trees

After applying the permutation technique it was predictable that there only two features in the model based on its own diagram. These features are `age` and `prevalentHyp` whose coefficients are -0.0143 and 0.0087, respectively. This means `age` is used for class 0 prediction, and `prevalentHyp` for class 1.

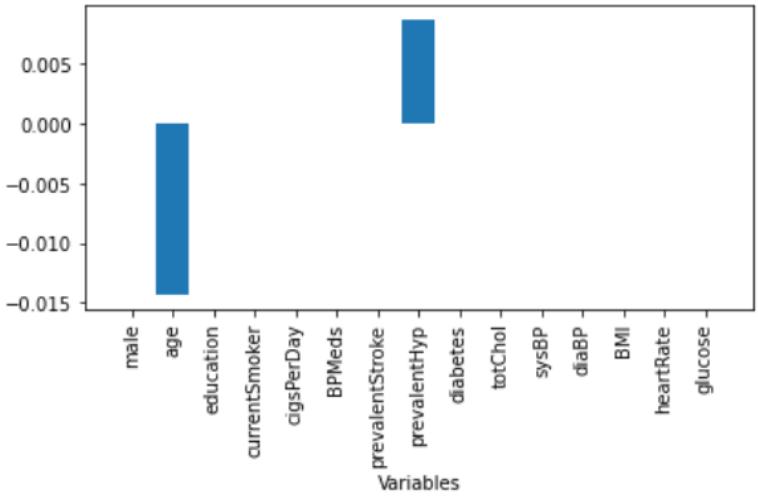

Figure 6: Feature Importance Permutation.

6 Conclusions

The knowledge of risk factors in Health Care is of extreme importance not just to the people in order to be able to avoid risk and unhealthy behaviours, but also for the governments in order to decide which are the best public health policies and strategies to take. In this work we applied machine learning on data collected in the Framingham study to analyse the cardiovascular disease risk.

First, we started by a exploratory data analysis, where the meaning of each variable of the dataset is described, as well as their descriptive statistics. It was found that the dataset is unbalanced, which means that there are much more subjects in one of the classes we intend to predict rather than the other. There are about 15% for class 1 and around 85% for class 0.

Regarding the fitted models, in order to determine the hyperparameters of the models (Logistic Regression and Decision Trees), we used a Stratified 5-fold cross-validation strategy varying some of the more important hyperparametrs. Regarding the obtained performances, based on the Confusion Matrix, the Logistic Regression model and the Decision Tree model behave similarly, misclassifying much more individuals who are healthy

(class 1) rather than the opposite situation, even though the Logistic Regression has evaluation metrics slightly better although it is a less expressive model than the Decision Trees.

With respect to feature importance, it is important in order to have insights into the data and models. Therefore, it allowed us to determined the risk factors of CVD. When applying Permutation Feature Importance in the Logistic Regression model, the variables `male`, `age`, `cigsPerDay`, `totChol`, `sysBP`, `heartRate` and `glucose` was highlighted as the ones more related with the target variable. This is coherent with the risk factors pointed out by QRISK3 [5]. Using those variables, we were able to reduce the number of features used maintaining a similar performance.Regarding the PFI for the Decision Tree model, since the model itself is using only 2 features, `age` and `prevalentHyp`, in this specific case, applying the permutation method is irrelevant. However, it is interesting to notice that age is the most important risk factor for both models. On other hand, the variable `prevalentHyp` is not so important in the Logistic Regression model as it is in the Decision Tree model.

References

[1] *About FHS*. URL: `https://framinghamheartstudy.org/fhs-about/`.

[2] Dileep. *Logistic regression To predict heart disease*. June 2019. URL: `https://www.kaggle.com/dileep070/heart-disease-prediction-using-logistic-regression`.

[3] *4.2. Permutation feature importance¶*. URL: `https://scikit-learn.org/stable/modules/permutation_importance.html`.

[4] Jason Brownlee. *How to Calculate Feature Importance With Python*. Aug. 2020. URL: `https://machinelearningmastery.com/calculate-feature-importance-with-python/`.

[5] *QRISK*. May 2021. URL: `https://en.wikipedia.org/wiki/QRISK`.

Appendixes

A Exploratory data analysis

A.1 Plots of variables

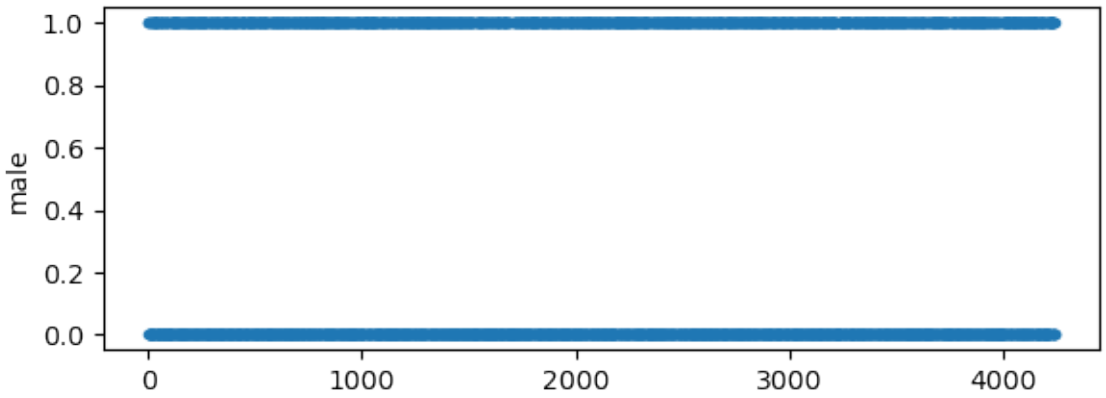

Figure 7: Male.

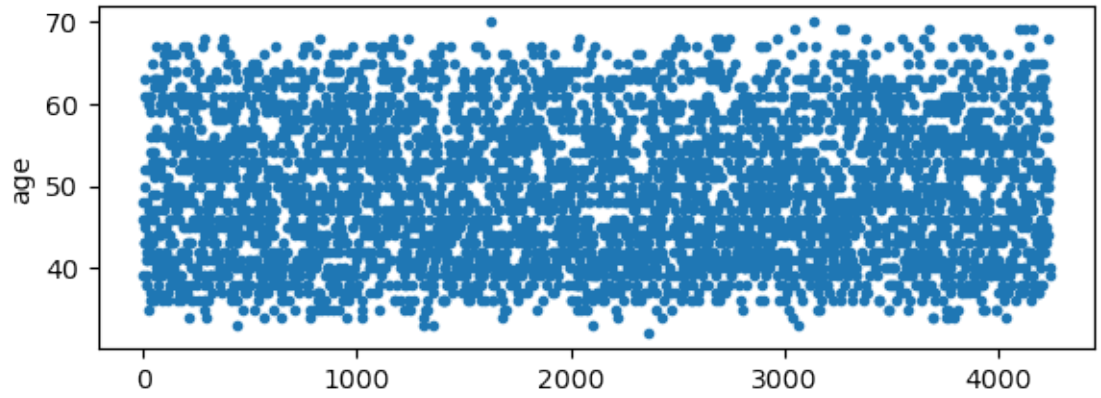

Figure 8: Age.

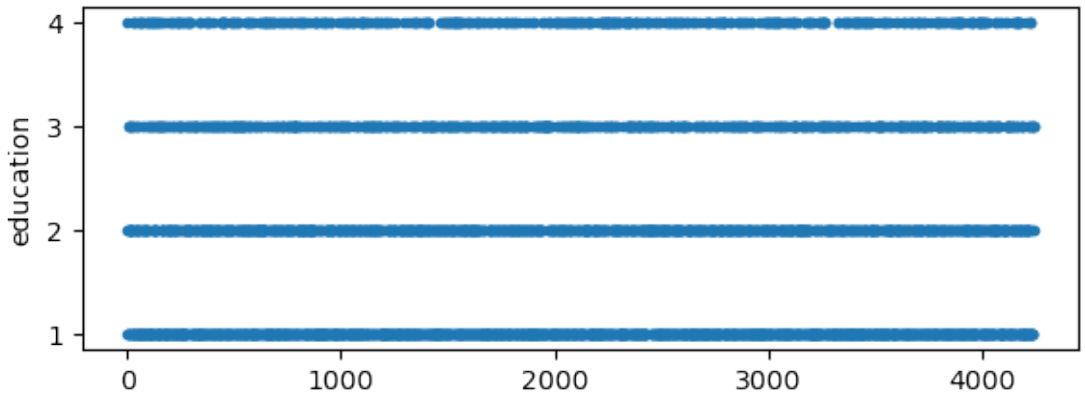

Figure 9: Education.

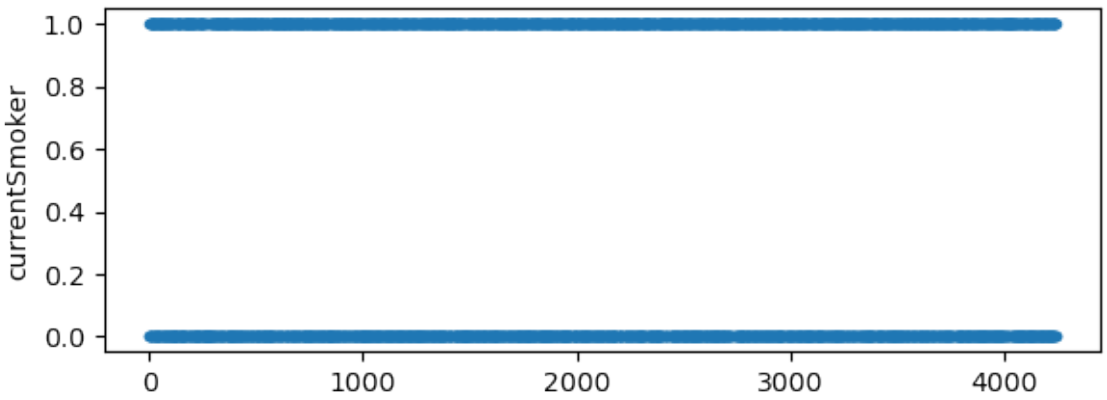

Figure 10: Current Smoker.

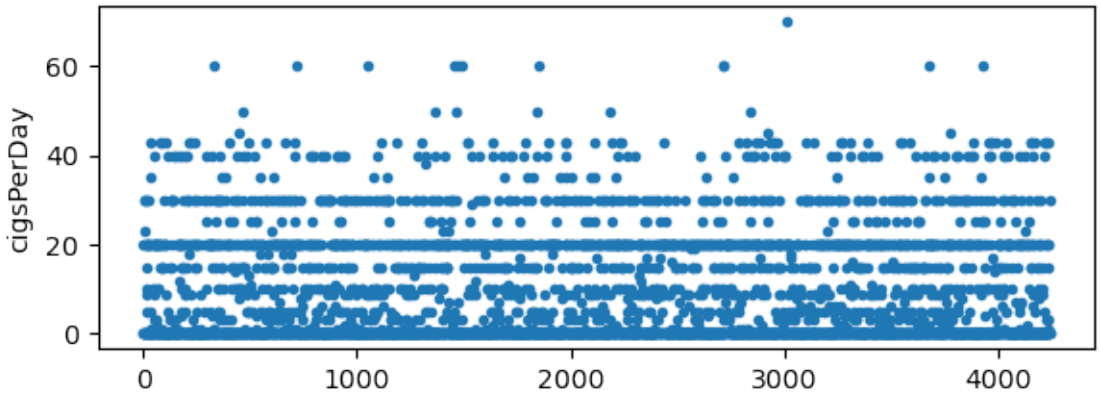

Figure 11: Cigarettes per Day.

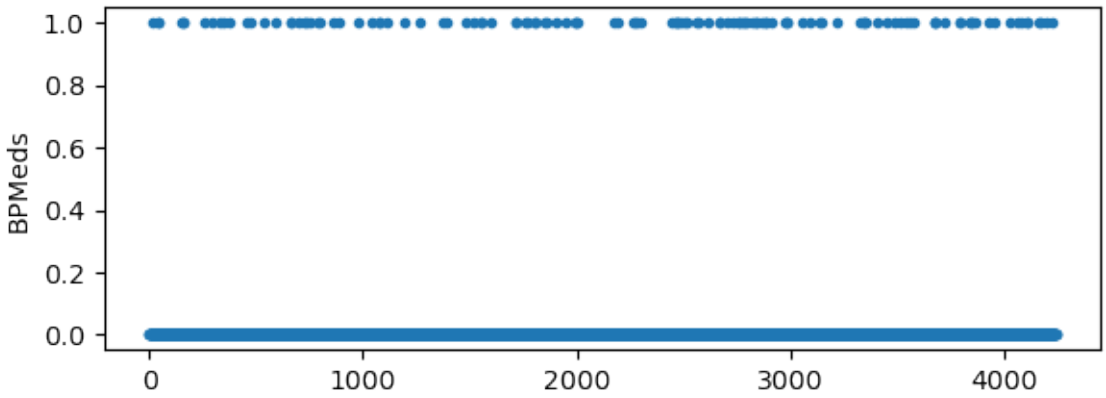

Figure 12: BPMeds.

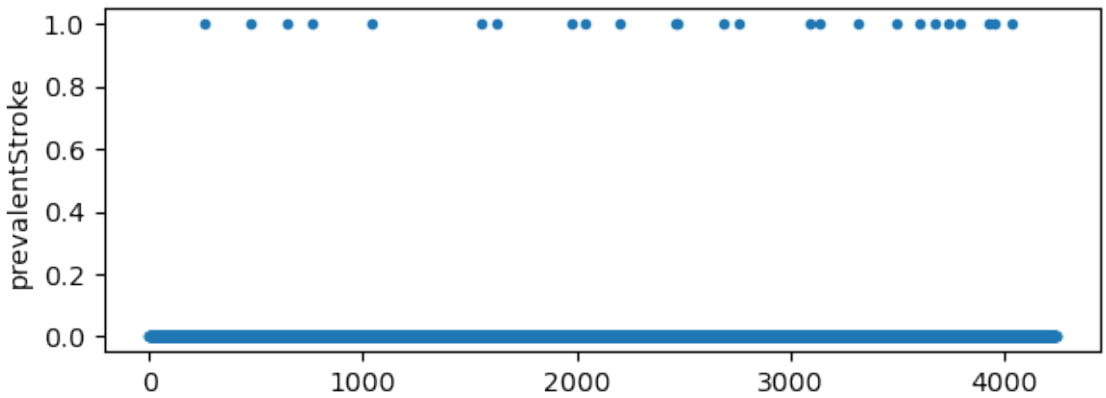

Figure 13: Pravalent Stroke.

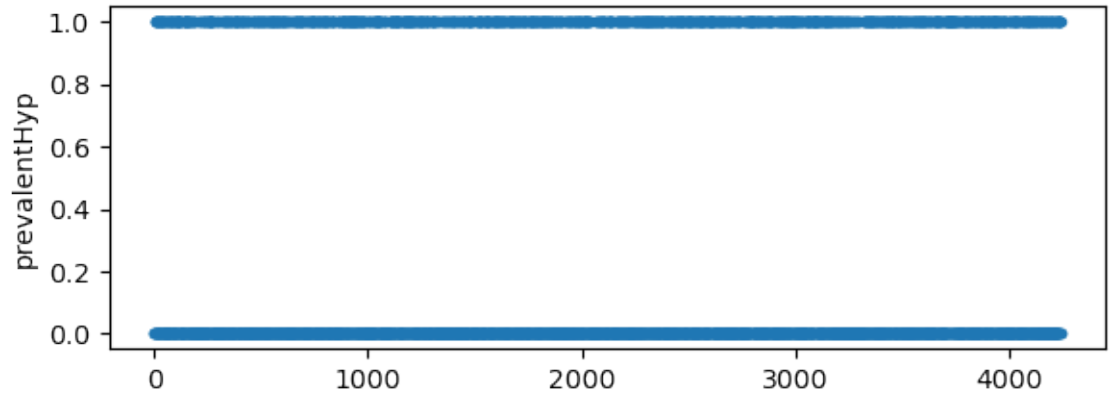

Figure 14: Prevalent Hypertension.

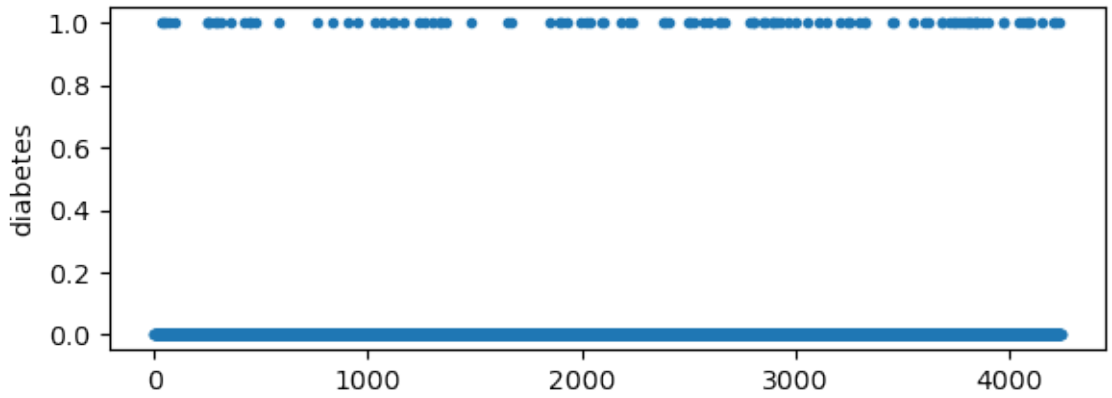

Figure 15: Diabetes.

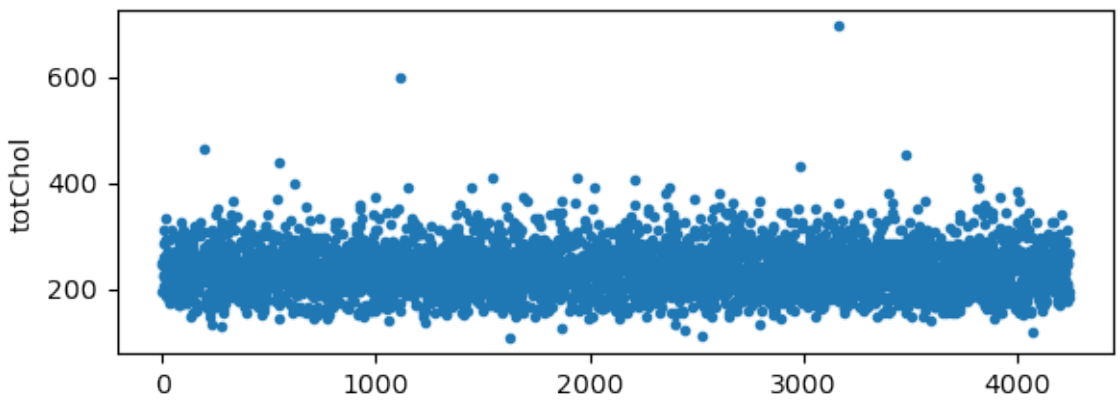

Figure 16: Total Cholesterol.

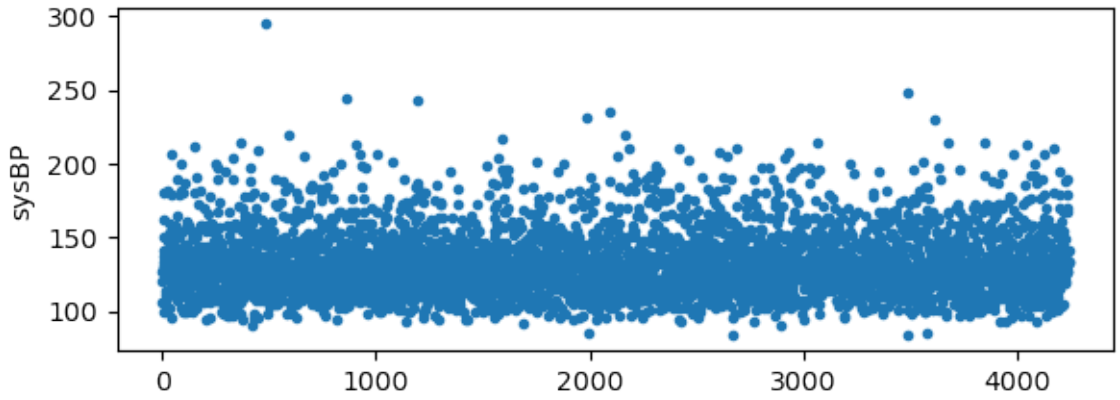

Figure 17: Systolic Blood Pressure.

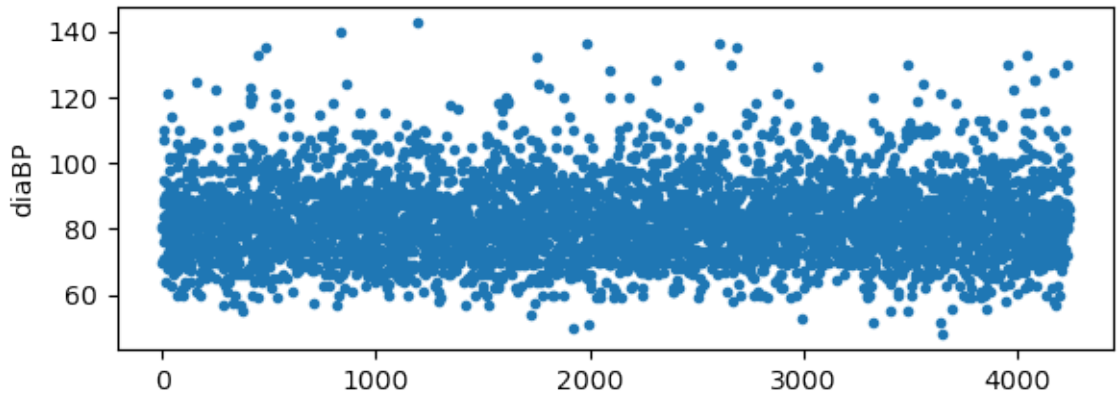

Figure 18: Dia Blood Pressure.

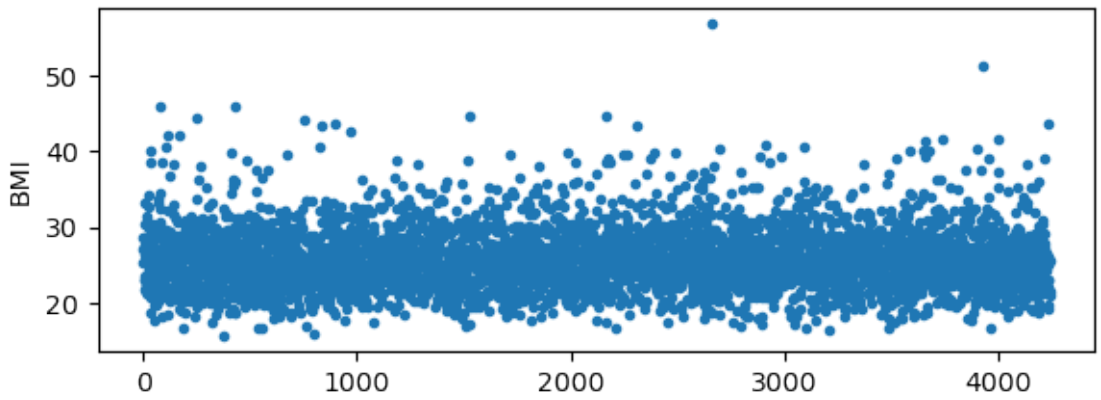

Figure 19: Body Mass Index.

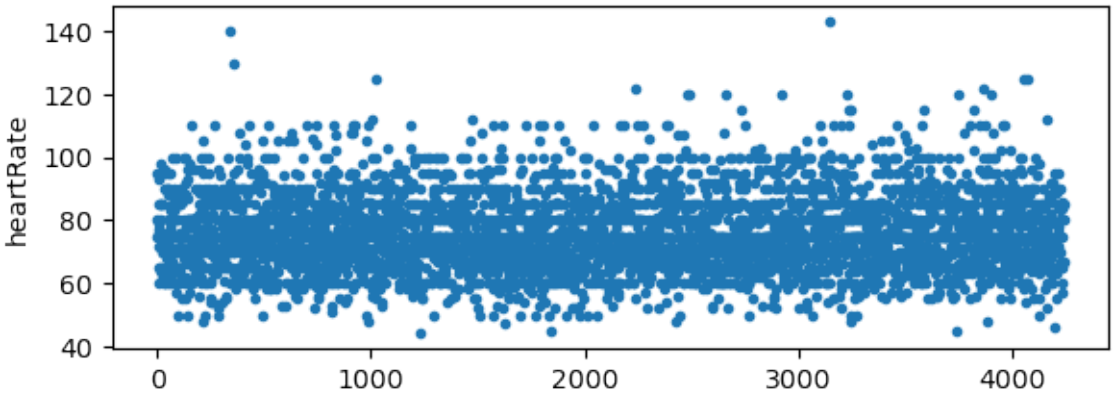

Figure 20: Heart Rate.

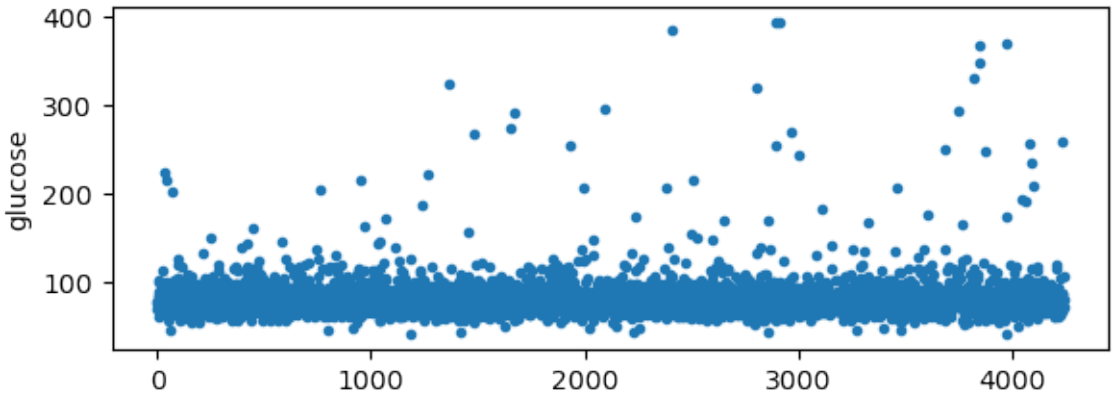

Figure 21: Glucose Levels.

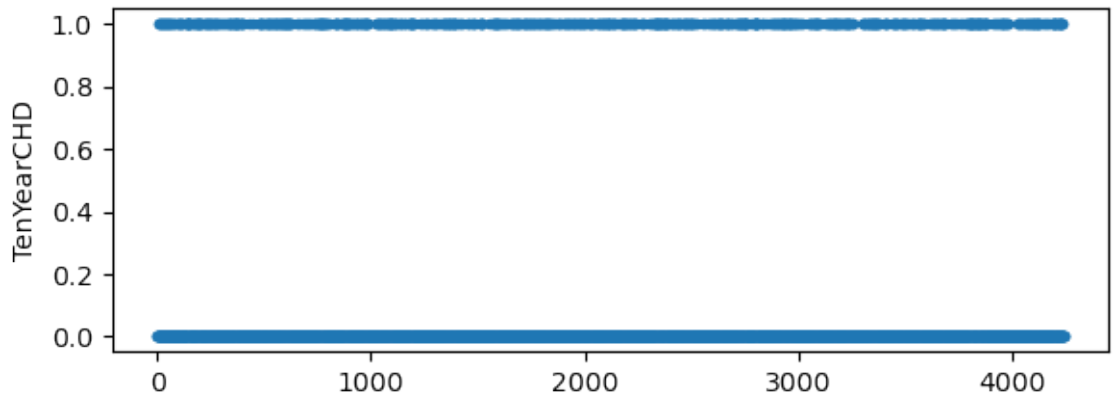

Figure 22: Ten Year CHD.

A.2 Box-plots

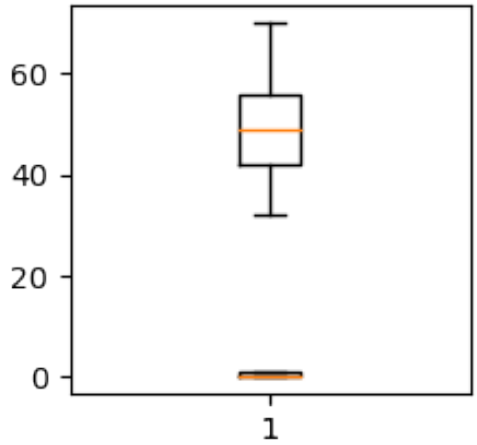

Figure 23: Box-plot of the variable Male.

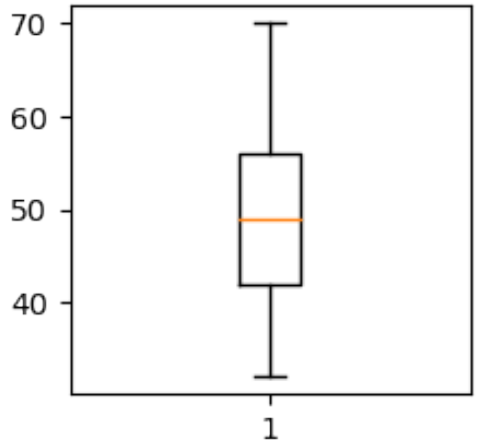

Figure 24: Box-plot of the variable Age.

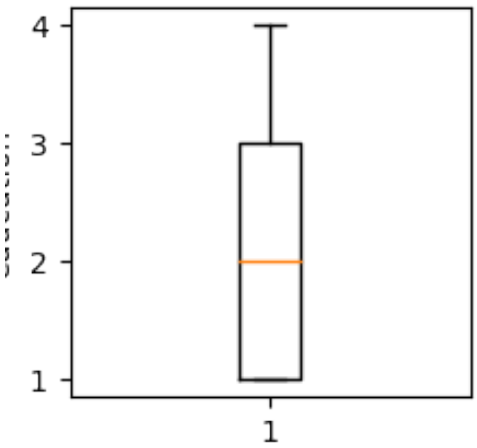

Figure 25: Box-plot of the variable Education.

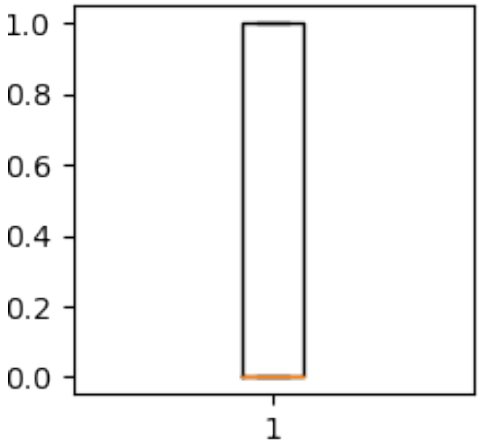

Figure 26: Box-plot of the variable Current Smoker.

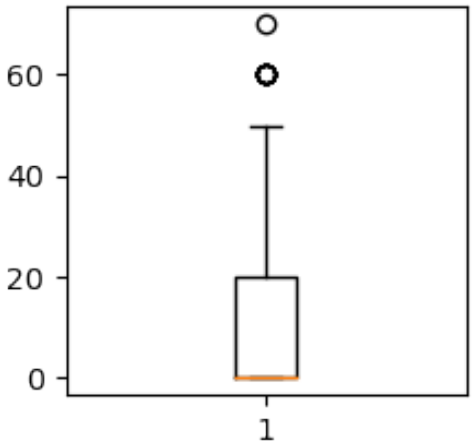

Figure 27: Box-plot of the variable Cigarettes per Day.

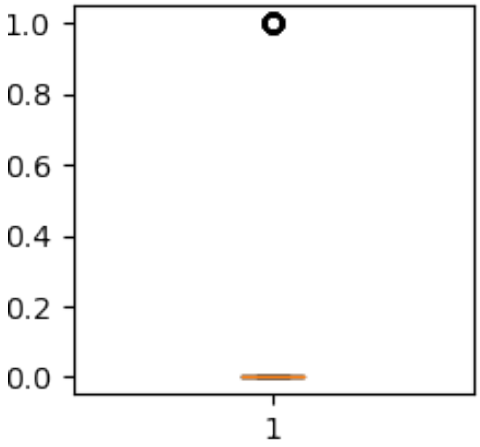

Figure 28: Box-plot of the variable BPMeds.

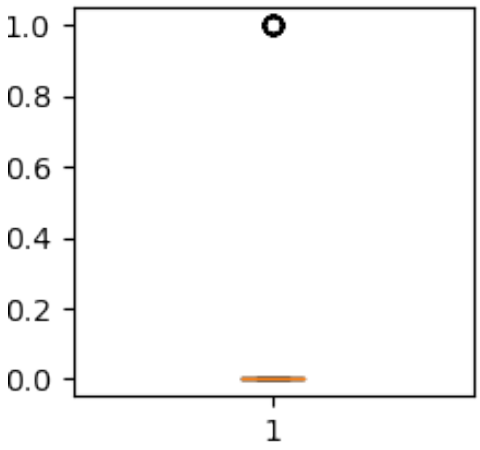

Figure 29: Box-plot of the variable Pravalent Stroke.

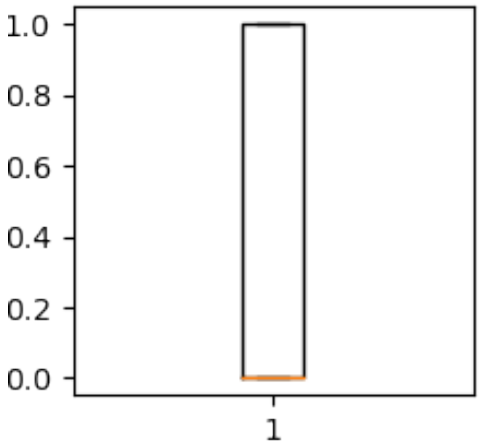

Figure 30: Box-plot of the variable Prevalent Hypertension.

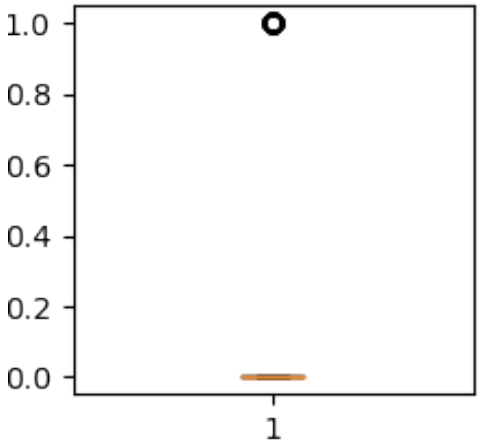

Figure 31: Box-plot of the variable Diabetes.

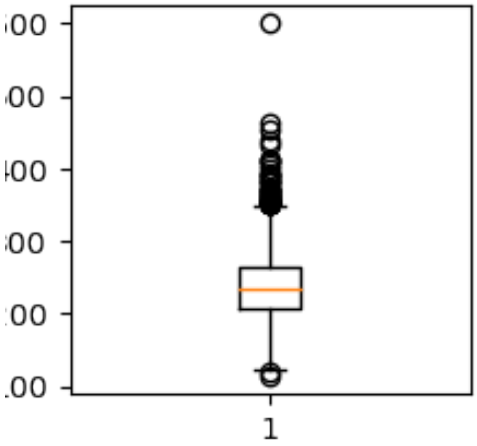

Figure 32: Total Cholesterol.

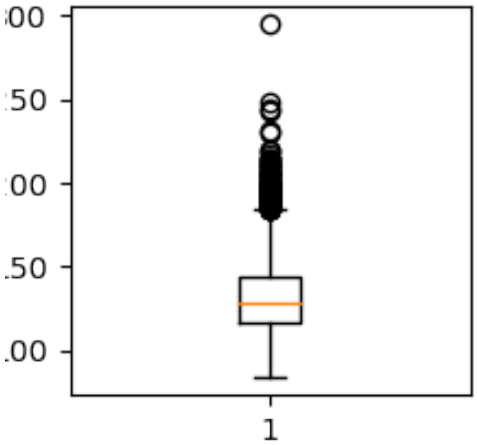

Figure 33: Systolic Blood Pressure.

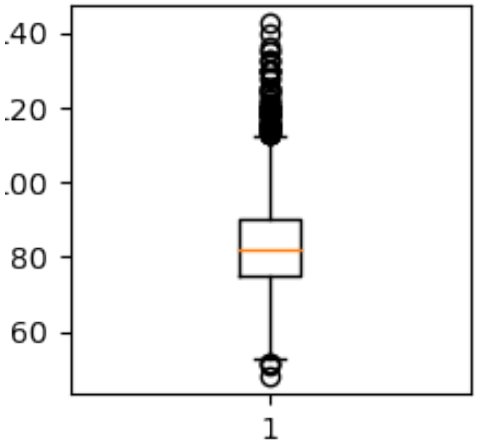

Figure 34: Dia Blood Pressure.

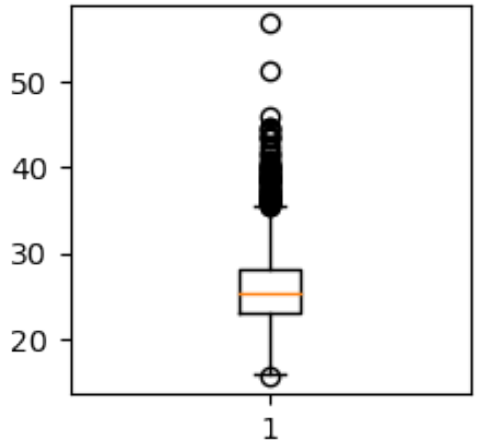

Figure 35: Box-plot of the variable Body Mass Index.

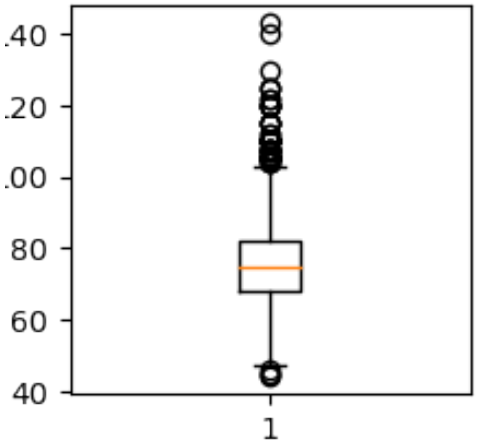

Figure 36: Box-plot of the variable Heart Rate.

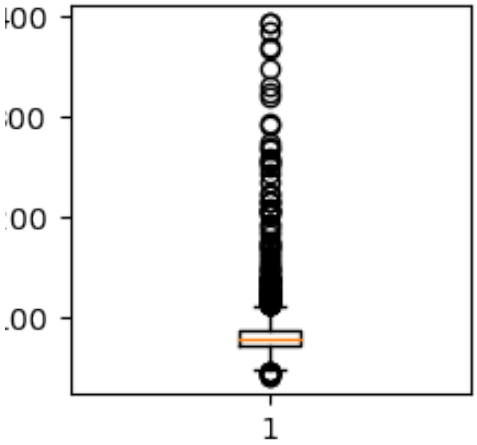

Figure 37: Box-plot of the variable Glucose Levels.

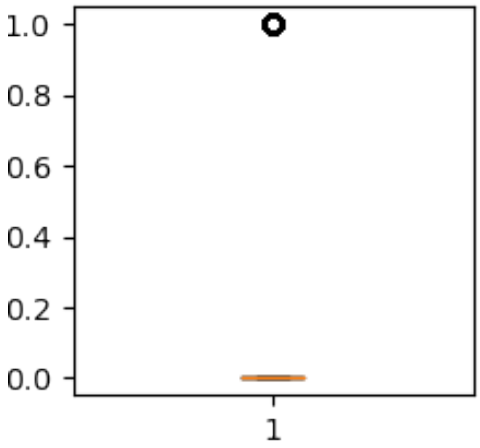

Figure 38: Box-plot of the variable Ten Year CHD.

A.3 Grid of pairwise relationships of numeric variables

Figure 39: Grid of pairwise relationships of the numeric variables.

B Logistic Regression

B.1 Obtained model and metrics without data standardisation

In this section are presented the classification report (Table 6) as well as the obtained confusion matrix (Table 7) for the obtained model without standardising the non-binary variables. The chosen hyperparameters by the grid search are the following: $C = 19.1$, fit_intercept: true and class_weight 'balanced'.

Table 6: Classification Report for the Logistic Regression model without standardising the non-binary variables.

Label	precision	recall	f1-score	support
0	0.91	0.60	0.73	931
1	0.24	0.69	0.35	167
accuracy			0.61	1098
macro avg	0.58	0.64	0.54	1098
weighted avg	0.81	0.61	0.67	1098

Table 7: Confusion matrix for the Logistic Regression model without standardising the non-binary variables.

		Predicted label	
		0	1
True label	0	559	372
	1	52	115

C Permutation Feature Importance

In Tables 8 and 9 are presented the Classification reports obtained from Logistic Regression using only the variables `male`, `age`, `cigsPerDay`, `totChol`, `sysBP`, `heartRate` and `glucose` as input variables.

Table 8: Classification Report - Logistic Regression - only using variables from PFI.

Label	precision	recall	f1-score	support
0	0.93	0.68	0.79	931
1	0.28	0.70	0.40	167
accuracy			0.69	1098
macro avg	0.61	0.69	0.60	1098
weighted avg	0.83	0.69	0.73	1098

Table 9: Confusion matrix - Logistic Regression - only variables from PFI.

		Predicted label	
		0	1
True label	0	637	294
	1	50	117