A statistical learning model to predict heart disease predisposition through physiological parameters and habits

Andrea TRESOLDI Lorenzo LEONI

University of Bergamo, Department of Management, Information and Production Engineering

June 12, 2022

Abstract

According to the Center for Disease Control and Prevention (CDC), heart disease is one of the leading causes of death for people in the U.S. The aim of this study is to predict heart disease predisposition through physiological parameters and habits (e.g. BMI, diabetes, sleep time and smoking) by using statistical learning methods.

Keywords: heart disease, statistical learning methods, logistic regression, undersampling.

1 Dataset description

Originally, the dataset[1] comes from the CDC which conducts annual telephone surveys to gather data on the health status of U.S. residents. This one includes 319 795 instances collected only in 2020. The columns, instead, are questions asked to candidates about their personal key indicators and their health status, such as “Do you have serious difficulty walking or climbing stairs?” or “Have you smoked at least 100 cigarettes in your entire life?”. From the original dataset only the most significant 18 variables have been selected or rather the variables which it’s thought to be directly correlated with the hearth disease’s insurgence: In detail, they are present both qualitative and quantitative variables (table 1 for more information about basic statistics):

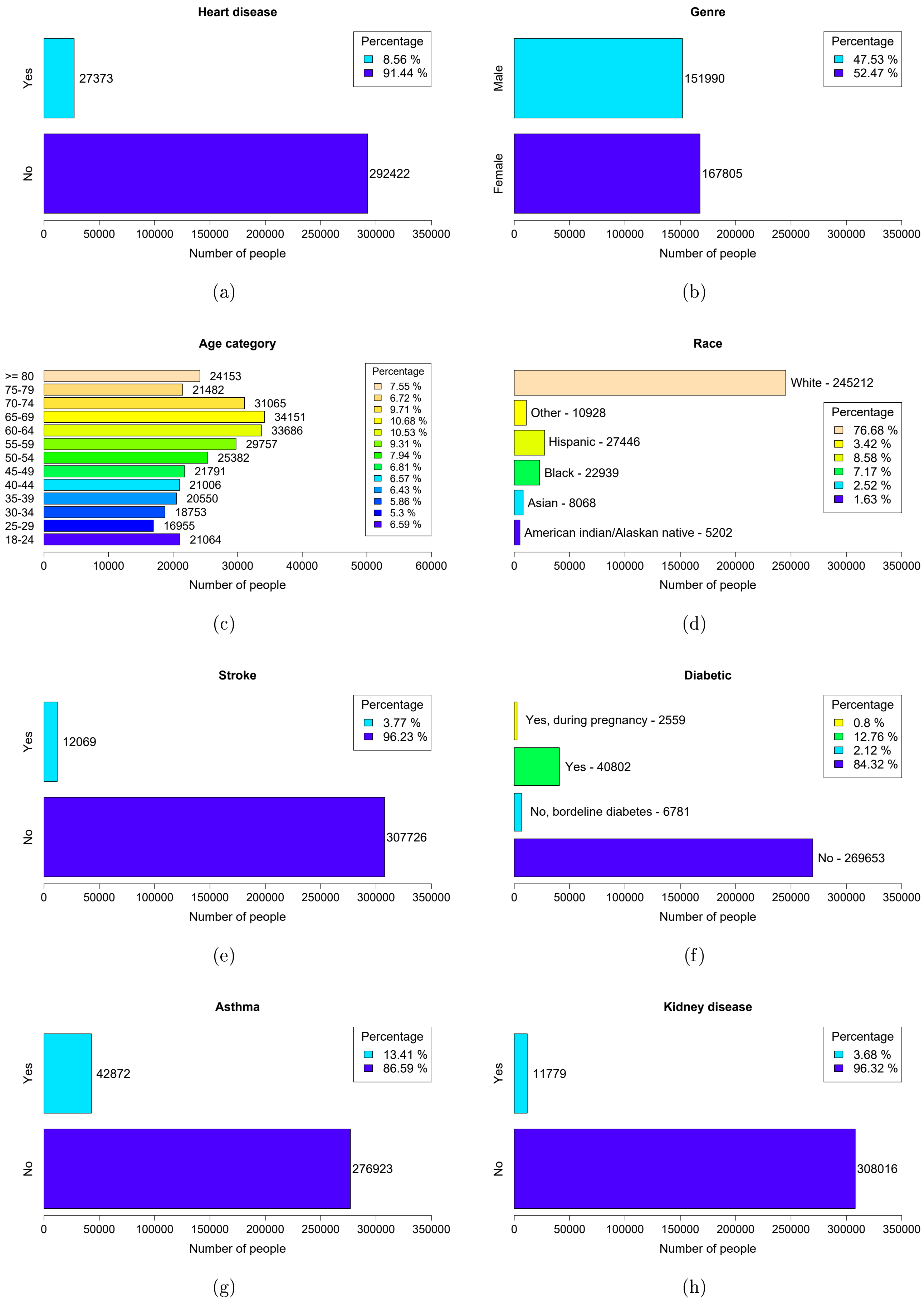
- **Heart disease**: it’s a categorical binary variable which expresses if the respondent has ever reported having coronary heart disease (CHD) or myocardial infarction (MI) (figure 1(a)).
- **Genre**: it’s a categorical binary variable which expresses the genre of the candidate (figure 1(b)).
- **Age category**: it’s a categorical variable (13 categories) which expresses the age range of the interviewee (figure 1(c)).
- **Race**: it’s a categorical variable (6 categories) which expresses the race/ethnicity of the respondent (figure 1(d)).

[1] the dataset and some information about it comes from this link: `https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease`.

- **Stroke**: it's a categorical binary variable which expresses if the candidate has ever had a stroke (figure 1(e)).
- **Diabetic**: it's a categorical variable (4 categories) which expresses if the interviewee has ever had diabetes (figure 1(f)).
- **Asthma**: it's a categorical binary variable which expresses if the respondent has ever had asthma (figure 1(g)).
- **Kidney disease**: it's a categorical binary variable which expresses if the candidate has ever had kidney disease, not including kidney stones, bladder infection or incontinence (figure 1(h)).
- **Skin cancer**: it's a categorical binary variable which expresses if the interviewee has ever had skin cancer (figure 2(a)).
- **Smoking**: it's a categorical binary variable which expresses if the respondent has smoked at least 100 cigarettes in his entire life (figure 2(b)).
- **Alcohol drinking**: it's a categorical binary variable which expresses if the candidate is a heavy drinker, in details adult men having more than 14 drinks per week and adult women having more than 7 drinks per week (figure 2(c)).
- **Difficulty walking**: it's a categorical binary variable which expresses if the interviewee has serious difficulty walking or climbing stairs (figure 2(d)).
- **Physical activity**: it's a categorical binary variable which expresses if the respondent has done physical activity or exercise during the past 30 days other than his regular job (figure 2(e)).
- **General health**: it's a categorical variable (5 categories) which expresses how is the candidate's general health condition (figure 2(f)).
- **Body Mass Index (BMI)**: it's a real positive number defined as the body mass (in kg) divided from the square of body height (in m) (figure 2(g)).
- **Physical health**: it's an integer positive variable which expresses how many days during the past 30 days the physical health of the interviewee was not good (figure 2(h)).
- **Mental health**: it's an integer positive variable which expresses how many days during the past 30 days the mental health of the respondent was not good (figure 3(a)).
- **Sleep time**: it's an integer positive variable which expresses on average how many hours of sleep the candidate gets in a 24-hours period (figure 3(b)).

2 Scientific questions

The main goal of this scientific study is searching, by using a statistical approach, a correlation between a set of personal key indicators about habits and health of Americans (the regressors) and their predisposition to develop a heart disease (the target variable). In detail, it's wanted to build a statistical learning model in order to reach different but complementary aims:



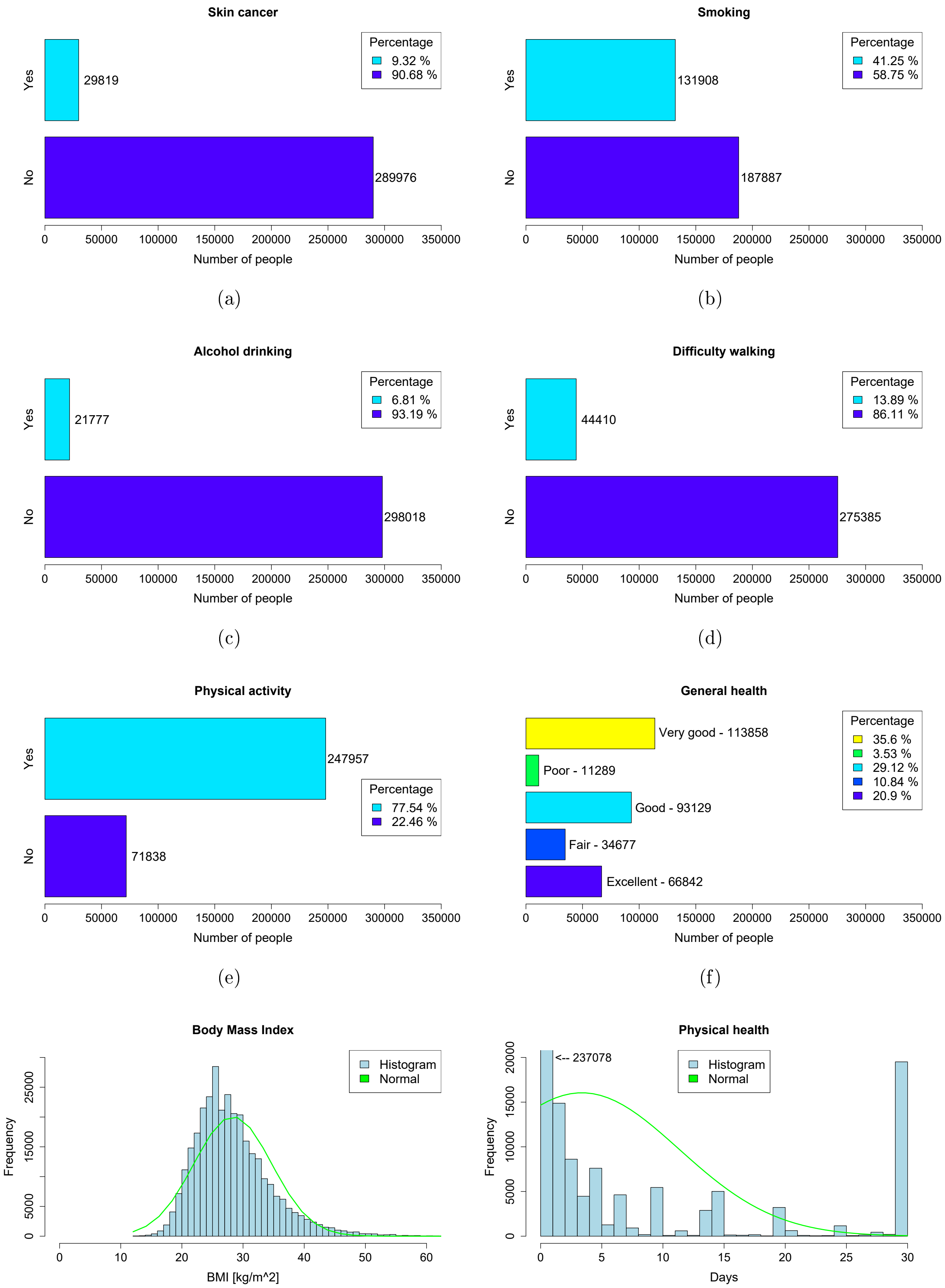
(a) (b)

(c) (d)

(e) (f)

(g) (h)

Figure 1: barplots of the categorical variables.



(a)

(b)

(c)

(d)

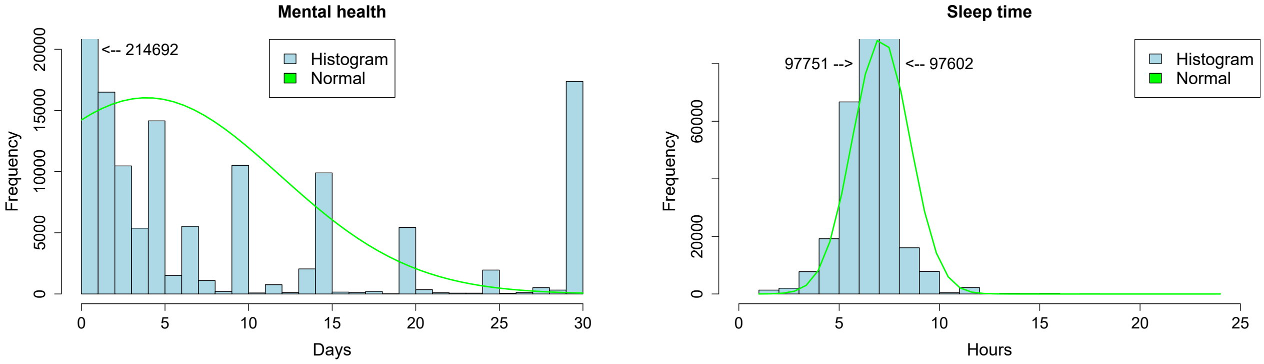
(e)

(f)

(g)

(h)

Figure 2: barplots of the categorical variables and histograms of the quantitative variables.



(a)

(b)

Figure 3: histograms of the quantitative variables.

	Mean	*Std*	1^{st} *qu.*	*Median*	3^{rd} *qu.*	*Kurtosis*	*Skewness*	*JB p-value*
BMI	28.33	6.36	24.03	27.34	31.42	3.89	1.33	~ 0
Physical health	3.37	7.95	0	0	2	5.53	2.60	~ 0
Mental health	3.90	7.96	0	0	3	4.40	2.33	~ 0
Sleep time	7.10	1.44	6	7	8	7.85	0.68	~ 0

Table 1: main statistics concerning the quantitative variables.

- **prediction**: given a new set of values about key indicators regarding a new person, it's wanted to predict if this one is exposed to the risk in developing heart disease. Are the considered regressors sufficient to define a statistical learning model which is able to predict correctly the target variable or it's necessary taking into account other physiological parameters?

- **inference**: it's desired to understand which personal key indicators are really associated with the risk in developing heart disease, what is the relationship between the target variable and each regressor and to comprehend which is the best model and related complexity that explains the correlation.

3 Methodology

The model used to achieve the aforementioned goals is the *logistic regression*. It's wanted to estimate a model that, given a set of regressors, provides the probability Pr (figure 4) a person can develop heart disease. Fixed a classification threshold th, it's possible to use estimated probabilities to classify the new instances into 2 categories: the candidate is predisposed to heart disease (*Yes* class) or not (*No* class). To build the starting model have been taken into account all the remaining 17 variables; as many of this regressors are categorical, it has been necessary to convert these one in dummy variables in order to consider them into a logistic model.

$$Pr(Y = Yes|X) = \frac{e^{\beta_0+\beta_1\cdot X_1+...+\beta_{17}\cdot X_{17}}}{1+e^{\beta_0+\beta_1\cdot X_1+...+\beta_{17}\cdot X_{17}}}$$

Figure 4: equation of the logistic regression where Y is the target variable *heart disease* and X are all the remaining 17 regressors.

3.1 Data selection

As it's possible to see in figure 1(a) the dataset is characterized by an unbalance on the target variable *heart disease*. By estimating the logistic model using the full dataset it's noted its poor capability in predicting the minority class. In order to reach the balancing between the two classes, during the entire scientific study it has been used an approach named *undersampling*. This technique allows to build a balanced dataset: all the 27 373 instances of the minority class are maintained into the new dataset, while the instances of the majority class are randomly reduced since the balancing is reached, that is 27 373 instances both *Yes* and *No* class[2].

3.2 Model selection

For the model selection has been used the *backward stepwise technique* based on z-test using undersampling. Iteratively, step-by-step, only the least significant regressor, or rather the variable with highest relatively p-value α, has been removed from the model. In order to avoid that the choice of the least significant regressor depends on the specific randomly undersampled dataset used to estimate the logistic model, every single step has been repeated 500 times (by using a different and randomly undersampled dataset every time) to remove the regressor that has shown to have its p-value α more than 1 % the most times respect to the other ones. This procedure has been applied iteratively until to obtain a model in which every variable was significant ($\alpha < 1\,\%$) at least 35 % of the times, or rather more than 150 times on 500 iterations.

3.3 Model validation

After obtaining the final model of logistic regression through the backward stepwise procedure, it has passed to its validation using a classification threshold *th* equal to 50 %. For this step it has been used a classic 70/30 validation approach employing a randomly undersampled dataset to build train and test data. The error's indexes have been taken into account to evaluate the model's performances are:

- train and test misclassification errors (MISE);
- train and test false-positive errors (FPE);
- train and test false-negative errors (FNE).

As it has been done for model selection, also for validation, in order to obtain classification results independent by the specific dataset used to create training and test data, the procedure has been repeated 1000 times, using a different randomly undersampled dataset for each iteration, to get an average value for every error's indexes. Moreover, to evaluate the sensibility of the model in prediction as the classification threshold changes, different values of the threshold *th* between 30 % and 70 % have been used to perform the validation.

[2] more information about random undersampling and other undersampling techniques at this link: https://www.mastersindatascience.org/learning/statistics-data-science/undersampling/

3.4 Outliers

After completing the study with the logistic regression using the full dataset, it has tried to improve model's performances removing any outliers from the data. The technique has been used for the removal plans to eliminate those instances which present outlier values about the four quantitative regressors (*BMI*, *sleep time*, *mental health* and *physical health*): it is enough that an instance has the value of a single regressor on the tail of its distribution to be removed from the dataset. In detail, a value of a regressor is an outlier if it's outside of the following range:

$$[Q_1 - \alpha \cdot IQR, Q_3 + \alpha \cdot IQR]$$

where Q_1 it's the first quartile of regressor's distribution, Q_3 the third quartile, $IQR = Q_3 - Q_1$ the interquartile range and finally k is a variable parameter that allows to regulate the width of the filter. This technique hasn't been applied to the qualitative variables because it's not significant. In order to analyse the model's performances without outliers it has been used the same approach (subsection 3.3) used to validate the logistic regression model with outliers. It's also important to note that the process of model selection without outliers has produced the same results (subsection 4.1) of model selection using the full dataset, therefore the model used for validation without outliers is the same one used for validation with outliers.

3.5 Other statistical learning approaches

So that to have a term of comparison about test classification errors of logistic regression, it has been tested also other models using the same validation approach (subsection 3.3), all the regressors and the entire dataset:

- LDA, QDA and naive Bayes;
- k-nearest neighbours;
- random forests.

4 Data analysis and result discussion

4.1 Model selection

Starting from considering all the 17 regressors, 38 including dummy variables, the selection's results have brought to build a model with only the 31 most significant variables, or rather a model containing only regressors that have resulted significant at least 35 % of the times (table 2). An unexpected behaviour affects *physical activity*: although it may seem a relevant regressor to explain the target variable *heart disease*, it has been selected to be removed. This could be due to the fact that *physical activity* refers only to the last 30 days before the interview instead of a longer period and mostly, being a categorical binary variable, it's not able to quantify the frequency. Another interesting aspect to note is that the average model's AIC remains constant during the iterations although the model's complexity decrease; this confirms the removed regressors are not meaningful to explain the target variable.

4.2 Model validation

4.2.1 Classification threshold *th* = 50%

As it's possible to see in table 3, model's performances are not influenced by the undersampling technique: in fact, having iterated the validation process 1000 times and having used a different

Step	*Number of regressors in the model*	*Least significant regressor*	*Not significant*	*Average model's AIC*
1	38	Age category (25-29)	500 times	54 038
2	37	Race (other)	500 times	54 019
3	36	Physical activity (Yes and No)	498 times	54 017
4	35	Diabetic (Yes, during pregnancy)	495 times	54 009
5	34	Race (White)	483 times	54 029
6	33	Race (Hispanic)	415 times	54 016
7	32	Diabetic (No, borderline diabetes)	348 times	54 038
8	31	/	/	54 040

Table 2: model selection steps based on z-test.

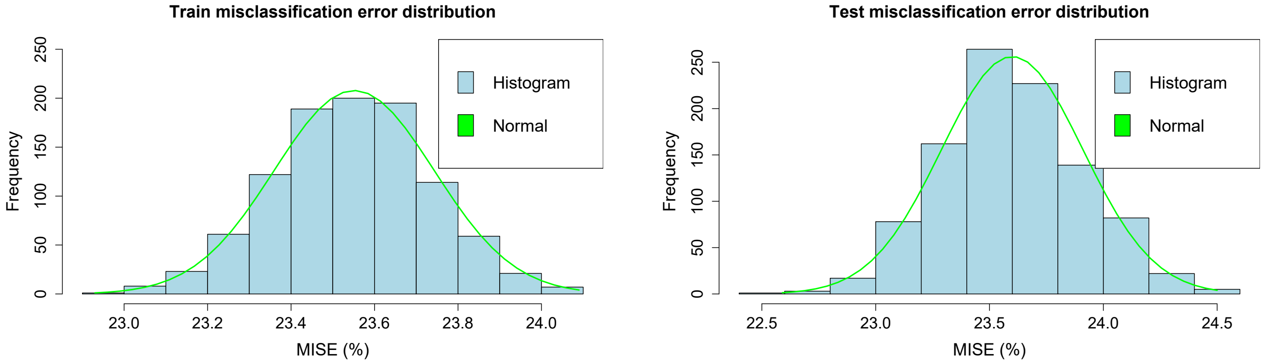
randomly balanced dataset in each iteration, it allowed to verify the values of train and test MISE, FPE and FNE are characterised by a low standard deviation. This highlights not only the independence of model's performances from the specific random dataset used to train the model, but also that it doesn't suffer from variance's problem. Another aspect it can be seen is that mean train and test classification errors are comparable to underline the fact the model has good test prediction capabilities and it doesn't suffer from overfitting. Finally, it's interesting to underline the normal distribution of train and test MISEs (figure 5), an assumption that is confirmed by the p-values of Jarque-Bera test greater than 5 %.

	Min.	*Mean*	*Max.*	*Std*	*JB p-value*
Train MISE	22.92	23.56	24.17	0.18	33
Test MISE	22.61	23.60	24.59	0.31	11
Train FPE	24.27	25.30	26.16	0.29	94
Test FPE	23.70	25.36	26.97	0.51	67
Train FNE	20.90	21.81	22.79	0.29	66
Test FNE	20.39	21.84	23.42	0.51	10

Table 3: statistics about classification errors (in %) resulting from 1000 iterations and classification threshold at 50 %.

4.2.2 Sensitivity to the classification threshold

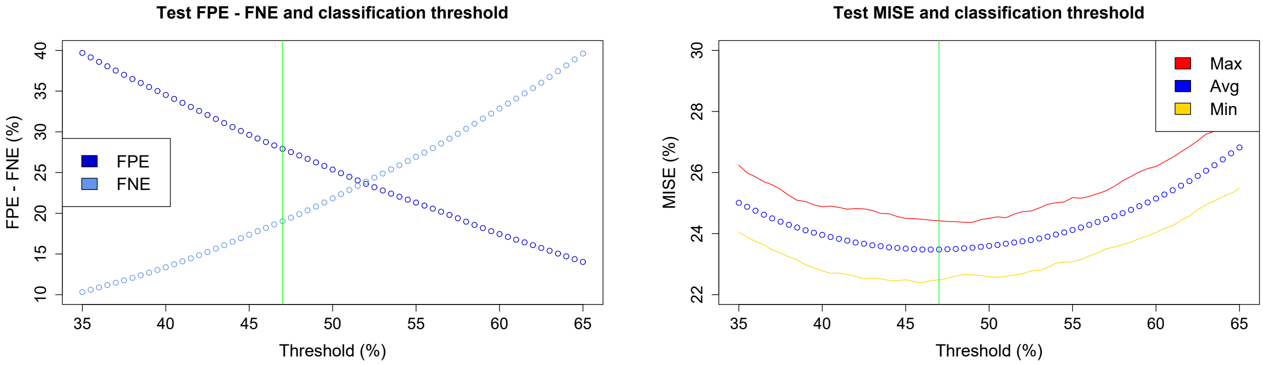
Table 3 shows that, using a classification threshold equal to 50 %, test mean FPE is greater than test mean FNE ($\overline{FPE}_{50} = 25.36\%$, $\overline{FNE}_{50} = 21.84\% \rightarrow \Delta_{50} = 5.52\%$), or rather the model performs worse in predicting *Yes* instances when the real class is *No* rather than another way



(a) (b)

Figure 5: histograms of train (a) and test (b) misclassification errors resulting from 1000 iterations and classification threshold *th* at 50 %.

around. It's possible to change this behaviour by changing the threshold's value; in detail, as the threshold increases, FNE increases while FPE decreases (figure 6(a)). Finally, it's possible to affirm that threshold's value which guarantees the minimum test MISE is 47 % (figure 6(b)), despite of a bias' increase between FNE and FPE ($\overline{FPE}_{47} = 27.91\%$, $\overline{FNE}_{47} = 19.04\% \rightarrow \Delta_{47} = 8.87\%$) respect to the base case ($th = 50\,\%$).



(a) (b)

Figure 6: false-positive, false-negative (a) and misclassification (b) test errors as the classification threshold *th* changes resulting from 1000 iterations.

4.3 Outliers

As it possible to see in table 4, the mean values of test misclassification error have not improved compared to the statistics obtained without the outliers' elimination (table 3). Probably, by analysing the achieved results, it can be said that this removal technique is not the most suitable for the application domain of this study; it would be more correct using a different removal criteria depending on the meaning of the considered regressor.

4.4 Other statistical learning approaches

The performances on test data of LDA (table 5), QDA (table 6), naive Bayes (table 7), K-NN (table 8) and random forests (table 9) are shown below. As happened with outliers' removal, also with these statistical learning approaches no significant overall improvements have been

	Min.	*Mean*	*Max.*	*Std*	Num. of removed instances
$\alpha = 1$	22.54	23.88	25.54	0.45	116 488
$\alpha = 3$	22.60	23.81	25	0.39	68 117
$\alpha = 5$	22.69	23.76	24.90	0.37	52 009
$\alpha = 7$	22.33	23.67	24.73	0.35	39 921
$\alpha = 9$	22.46	23.66	24.89	0.35	37 199
$\alpha = 11$	22.83	23.75	24.88	0.34	21 580
$\alpha = 13$	22.78	23.74	25.03	0.32	20 159
$\alpha = 15$	22.47	23.62	24.48	0.30	0

Table 4: statistics resulting from 1000 iterations about test misclassification error (in %) without outliers as the number of instances removed changes.

achieved regarding all classification errors compared to the results obtained with the logistic regression.

Mean test MISE	*Mean test FPE*	*Mean test FNE*
23.66	25.68	21.64

Table 5: test classification errors (in %) of LDA model.

Mean test MISE	*Mean test FPE*	*Mean test FNE*
26.80	41.32	12.26

Table 6: test classification errors (in %) of QDA model.

Mean test MISE	*Mean test FPE*	*Mean test FNE*
27.79	20.29	35.29

Table 7: test classification errors (in %) of naive Bayes model.

Mean test MISE	*Mean test FPE*	*Mean test FNE*
24.88	28.79	20.97

Table 8: test classification errors (in %) of k-nearest neighbours non-parametric model with $k = 30$ and using normalized data.

Mean test MISE	*Mean test FPE*	*Mean test FNE*
23.98	28.20	19.76

Table 9: test misclassification errors (in %) of random forests non-parametric model with 300 trees for each of the 100 forests.

5 Conclusions

At the end of this study it can affirm that, also using different statistical learning approaches, it's not possible to obtain a test misclassification error below the 23 % threshold. Probably, in order to predict better if a person is exposed to the risk in developing heart disease, it would be appropriate taking into account also other significant regressors linked to the disease's onset as blood pressure, cholesterol's level, oxygen saturation and hereditary factors.