HEART DISEASE

To learn which tests matter in detecting heart disease, three Machine Learning Algorithms (logistic, SVM and decision trees) were utilized. Logistic Regression were found to be the most effective with 88 % accuracy. The most important attributes are chest pain type, fluoroscopy results, thallium heart scan, ECG + treadmill test and blood cholesterol test. This method can save 4 % of the cost, greatly simplify and save a lot time during the process of heart disease detection.

Predictive Modelling Group 17

Rohit Saluja
Antti Korhonen
Rainer Klein

Table of Contents

Background and motivation 2

Goals 2

1. Identify heart disease predictors 2
2. Suggest the optimal test and their order 2
3. Identify unnecessary tests 2

Solution Process 2

1. Project methodology 2
2. Data exploration 3

Original Data 3

Costs 4

3. Data preparation 5

Feature extraction 5

Feature scaling 5

Data split 6

4. Modelling 6

Logistic Regression 6

Support Vector Machines (SVM) 8

Decision Trees 9

Results 11

1. Comparison of regression models 11
2. Effective attributes 11
3. Economic effects 12

Conclusion 13

References 13

BACKGROUND AND MOTIVATION

The data used in this study has been obtained Kaggle [1]. It is originally from the UC Irvine Machine Learning Repository [2], which in turn received it from the Cleveland Clinic Foundation in 1988. The data set consists of 303 instances of 14 health related attributes. One of the attributes is the existence of heart disease which is tried to be predicted based on the other 13 attributes. The data had no NaN values. We also did not find strong correlation between the different features.

The motivation behind this study is to find possible cost and time savings related to caring of possible hearth disease patients. Even small percentual savings and optimizations in the working process can lead to significant economical savings and even more importantly can improve health care quality by easing possibly long waiting times for the treatments. The savings can be achieved if some of the tests which are conducted normally, is identified to have only a minor effect to accuracy and therefore can be left out.

GOALS

1. Identify heart disease predictors

The first aim of the study is to pinpoint the attributes that most accurately detect the presence of heart disease.

2. Suggest the optimal test and their order

Having identified these potent attributes, the optimal set of tests can be suggested. Starting with the most effective predictor, the hospital could have a rather reliable result quickly and cheaply. We could then continue with additional tests if necessary.

3. Identify unnecessary tests

The final aim of the study is to identify tests which have no or little predictive power so time and money can be saved be stopping these tests. To this end, test costs provided with the dataset are utilized.

SOLUTION PROCESS

1. Project methodology

After finding the interesting dataset and research questions we started to familiarize with the data and confirm that it is suitable for answering to our questions. The process was started with exploratory data analysis, visualization and preprocessing. This guarantees sufficient basic understanding of the data set so that real machine learning models can be built to derive further insights from the data. This workflow is presented in the figure 1.

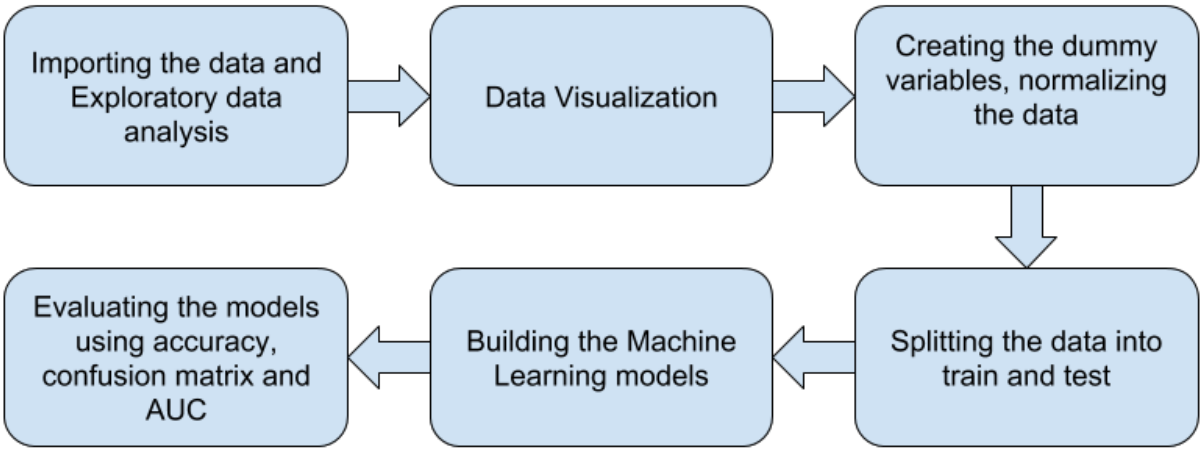

Figure 1: Project methodology.

2. Data exploration

Original Data

There are relatively few attributes and the data exploration can be achieved with histograms. Below, in Table 1, these attributes are presented.

Table 1: Attributes

#	ATTRIBUTE	TYPE	TEST RESULT	TEST
1	age	int	Age in years	Questionnaire
2	sex	bin	Male/female	Questionnaire
3	cp	int	Chest pain type	Questionnaire
4	trestbps	int	Resting blood pressure	Blood pressure
5	chol	int	Blood cholesterol level	Blood test
6	fbs	bin	Blood sugar level > 12mg/dl	Blood test
7	restecg	int	ECG results	ECG
8	thalach	int	Maximum heart rate	Thallium heart scan
9	exang	bin	Exercise induced angina (yes/no)	Teadmill + ECG
10	oldpeak	real	ST depression induced by exercise	Teadmill + ECG
11	slope	int	Slope of the peak exercise ST segment (-/0/+)	Teadmill + ECG
12	ca	int	No. of major vessels colored (0/1/2/3)	Flouroscopy
13	thal	int	Heart tissue damage type (none/type 1/type 2)	Thallium heart scan

In each chart presented in the figures 2-5, two distributions for that attribute are plotted – one for the case where heart disease was detected (target>0) and one for the case where no heart disease was detected (target=0). After checking the set of plots produced, it was noticed that following

four appear to show significant correlation: cp (Figure 2), ca (Figure 5), exang (Figure 4) and slope (Figure 3). These will be further researched in the modelling part of the study.

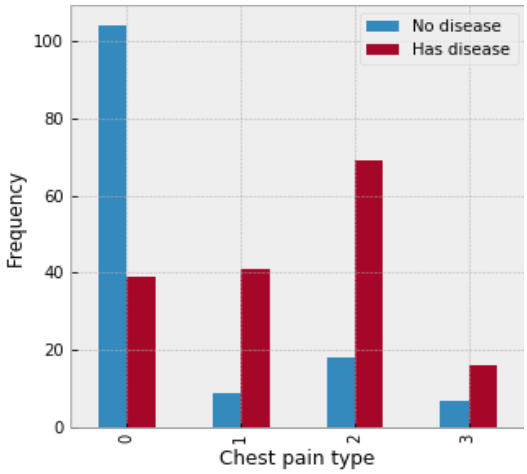

Figure 2: Chest pain type (cp) vs heart disease. Chest pain types: 0 Typical angina, 1 Atypical angina, 2 Non-anginal pain and 3 Asymptotic

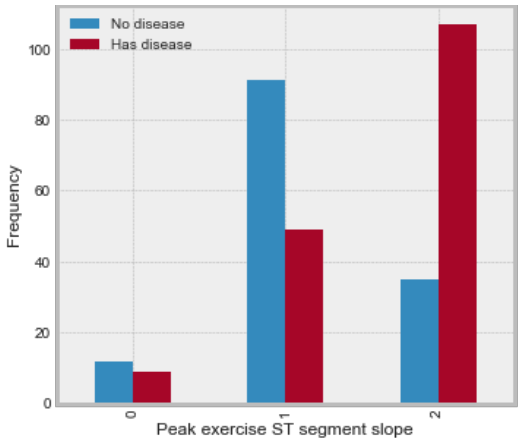

Figure 3: Peak exercise ST segment slope (slope) vs heart disease. Slope categories: 0 Up, 1 Flat and 2 Down

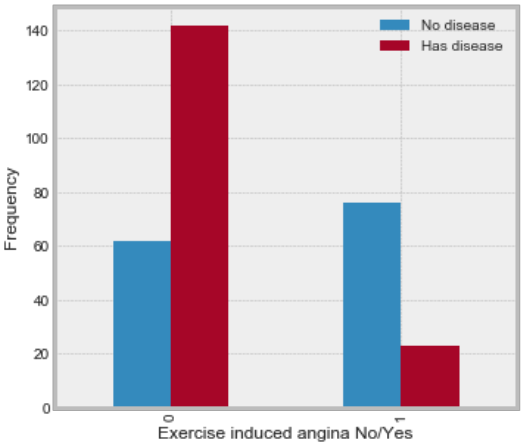

Figure 4: Exercise induced angina (exang) vs heart disease

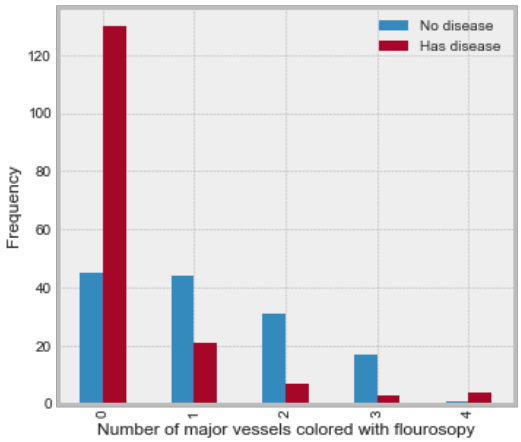

Figure 5: No. of major vessels colored by fluoroscopy (ca) vs heart disease

Costs

Table 2 displays the cost of the test required to attain each attribute value. It was originally presented in 1985 Canadian Dollars, which is roughly equal to 2019 Euros (1.02 € to be exact) [3] [4]. The cost information is from the Ontario Health Insurance Program's fee schedule [5]. The costs in the *Cost* column are for individual tests, considered in isolation. When tests are performed in groups, there may be discounts, due to shared common costs. Groups of tests with common costs are identified in the *Test Group* column. Marginal costs with discounts are in the *Test 2* column. As it can be seen from the table, some of the tests have to be done individually.

For example, the classic treadmill test including ECG provides values for 3 attributes – *exang*, *oldpeak* and *slope*. Once the test has been carried out to get a value for slope, gaining values for also *oldpeak* and *exang* has a positive but nearly zero marginal cost.

Table 2: Test costs

	ATTRIBUTE	COST	TEST GROUP	COST 2
1	age	1	-	1
2	sex	1	-	1
3	cp	1	-	1
4	trestbps	1	-	1
5	chol	7	A	5
6	fbs	5	A	3
7	restecg	16	-	16
8	thalach	103	B	1
9	exang	87	C	1
10	oldpeak	87	C	1
11	slope	87	C	1
12	ca	101	-	101
13	thal	103	B	1

3. Data preparation

Feature extraction

The data exploration conducted showed that the data set includes also categorical attributes (*sex*, *cp*, *restecg*, *slope*, *thal*). In order to fit these variables into a regression model, new so called dummy variables have to be done for those, for example, variable *sex* is converted to two variables *sex_male* and *sex_female* which are having values 0 (false) and 1 (true) [6]. Pandas library for Python has ready-made function *get_dummies* which was used to conduct the division into dummy variables.

Feature scaling

We normalized our data using feature scaling from sklearn preprocessing, *quantile_transform* function. The scaling has to be done because some continuous values in the dataset were very large in magnitude and it is preferable to bring those values to the same scale range with other values. This confirms that the models will not give more weight to features only because of their bigger magnitude. However, also the original dataset without scaling is kept so in each further modelling step, the effect of normalization can be assessed by comparing the results achieved with unnormalized data.

Data split

Normally in data science, the original data set is divided into two parts, as known as training and testing parts. This allows that the same dataset can be used for both training and testing of a model in a way that model performance is evaluated with new unseen (for the model) data. The split was conducted in a way that the training set has 70 % and the test set has 30 % of the original data set. The 70/30 ratio was chosen as it (along with 80/20) has commonly been shown to produce good results and to allow the model have enough data for training. The sklearn library in Python has a ready-made function *train_test_split* which is used to conduct the split. From the figure 6 below, it can be seen that the split was successful since the proportion of patients which have and does not have hearth disease does not change significantly.

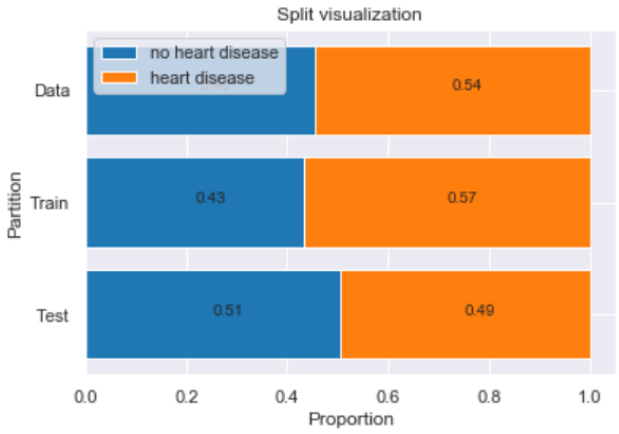

Figure 6: Proportions of patients having hearth disease in original, training and testing datasets.

4. Modelling

We have chosen three supervised-learning algorithms, namely Logistic Regression, Support Vector Machines and Decision Trees, to build the heart disease prediction model. All three models were built using both normal (unnormalized) and normalized datasets. The performance of the models was compared with confusion matrices and accuracy scores. This allowed us to choose the best predictive model out of six possibilities (three different algorithms with both unnormalized and normalized data).

Logistic Regression

The first algorithm that we used for heart disease prediction was Logistic Regression from the *sklearn.linear_model* package. As table 3 presents quite good accuracy was achieved with that.

Table 3: Accuracy comparison of Logistic Regression with Unnormalized and Normalized Data

Model	Accuracy
Logistic Regression with Unnormalized Data	87.90 %
Logistic Regression with Normalized Data	87.91 %

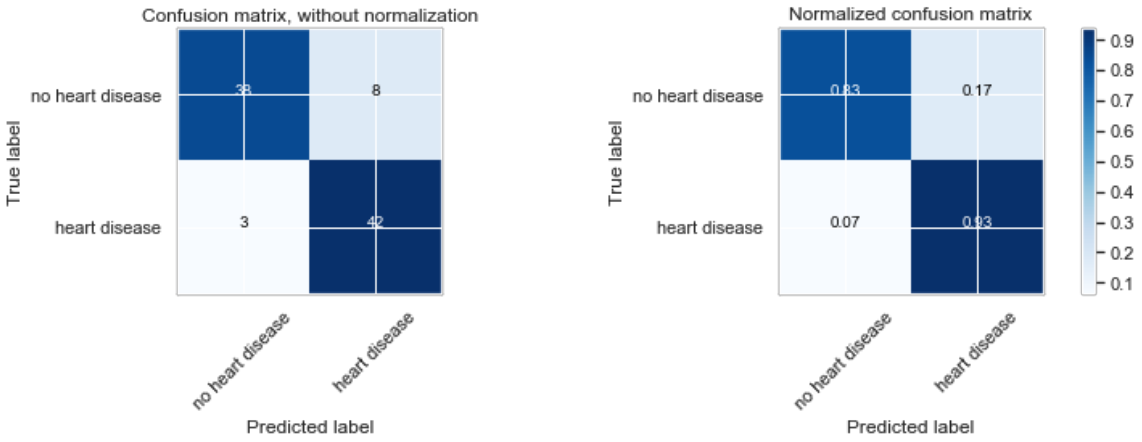

Figure 7: Confusion Matrix for Logistic Regression with Unnormalized Dataset

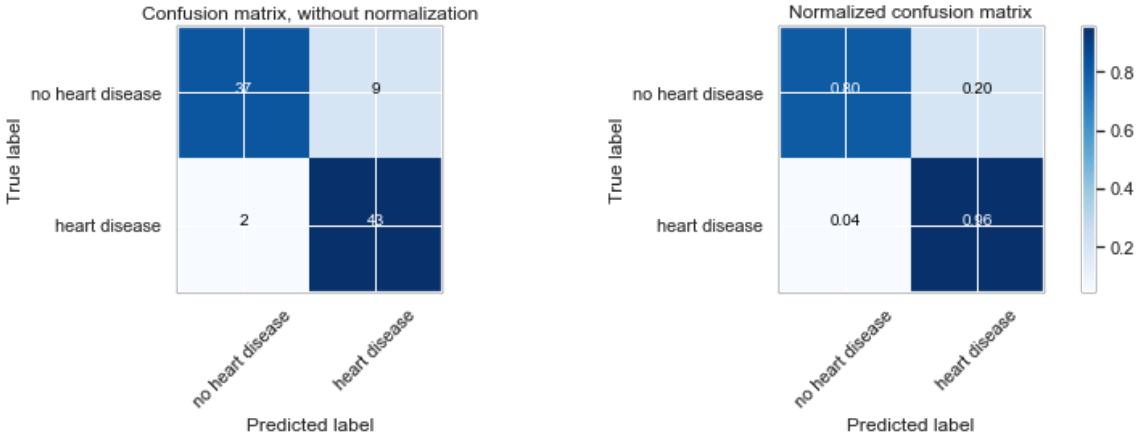

Figure 8: Confusion Matrix for Logistic Regression with Normalized Dataset

The main criteria used to evaluate the performance of our model was the number of times the model predicted someone did not have a heart disease, but they actually had it, as known as false negative result. The evaluation is done by this manner since only looking at accuracy as a measure of success of the model can often be misleading and especially in the field of health care false negative test results can be especially harmful and cause even fatal outcomes if patient having a heart disease is analyzed not to have it.

From the figures 7 it can be seen that the Logistic Regression model with unnormalized data predicted 7 % of hearth disease patients who actually have it are predicted wrongly not to have it. Corresponding value for the model with normalized data was 4 % as presented in the figure 8. According to these findings, it is visible that Logistic Regression model with normalized dataset has better performance on identifying heart disease.

Support Vector Machines (SVM)

The second algorithm that was used for heart disease prediction was the Support Vector Machines (SVM) from *sklearn.svm.svc*. Function *GridSearchCV* from *sklearn.grid_search* was used for the selection of the various hyper parameters like C, gamma and kernel. *GridSearchCV* uses cross validation (3- fold by default) and runs SVM on these folds of the dataset until it arrives with the combination of kernel, C and gamma with the best accuracy. *GridSearchCV* iterated through set of hyper parameters that were manually given to it in form of a dictionary and found the best possible combination. The best hyper parameters for the SVM model with the unnormalized dataset were C = 1, kernel = linear, gamma = auto and the best hyper parameters for the SVM model with the normalized dataset were also C = 1, kernel = linear, gamma = auto. The table 4 presents accuracy scores and figures 9 and 10 shows the confusion matrices for the SVM models.

Table 4: Accuracy comparison of SVM with Unnormalized and Normalized Data

Model	Accuracy
SVM with Unnormalized Data	87.91 %
SVM with Normalized Data	82.40 %

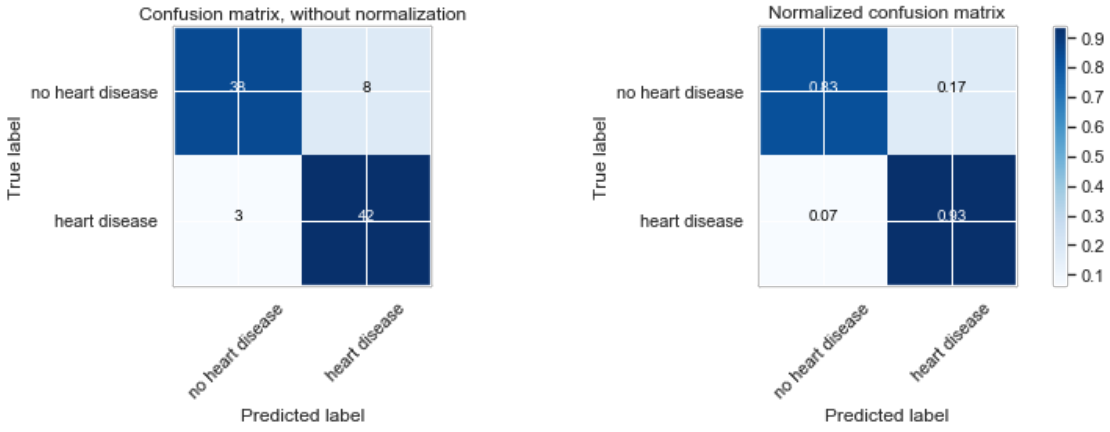

Figure 9: Confusion Matrix for SVM with Unnormalized Dataset

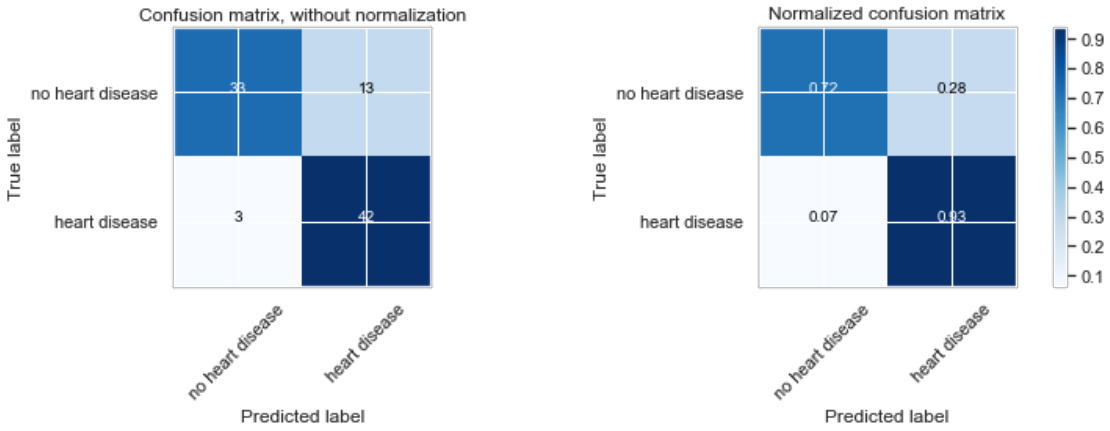

Figure 10: Confusion Matrix for SVM with Normalized Dataset

From the confusion matrices it can be seen that the SVM Model with both unnormalized and normalized data predicts incorrectly that 7 % of patients who have heart disease would not have it. In this case it is noticed that normalizing the dataset before conducting the model did not affect to the outcome. Also, it is noticed that the model is performing worse than the Linear Regression model.

Decision Trees

The third and final algorithm used to analyze the dataset was Decision Trees from the *sklearn.tree DecisionTreeClassifier*. During the analysis, different values for the *max_depth* parameter was tried in order to achieve best possible outcome. In the end optimum depth for the decision tree is three. The accuracy scores confusion matrices for Decision Tree models with unnormalized and normalized datasets are summarized in the table 5, figure 11 and figure 12.

Table 5: Accuracy comparison of Decision Trees with Unnormalized and Normalized Data

Model	Accuracy
Decision Tress with Unnormalized Data	83.51 %
Decision Tress with Normalized Data	71.42 %

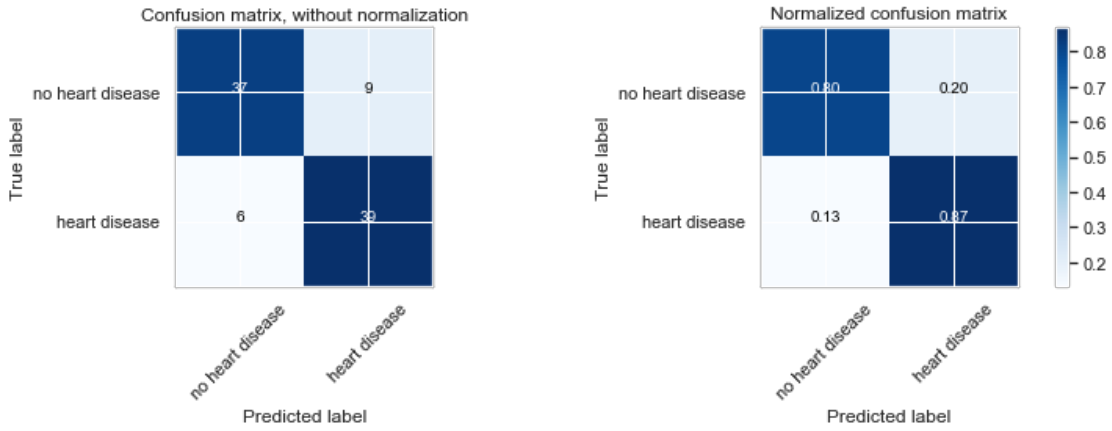

Figure 11: Confusion Matrix for Decision Trees with Unnormalized Dataset

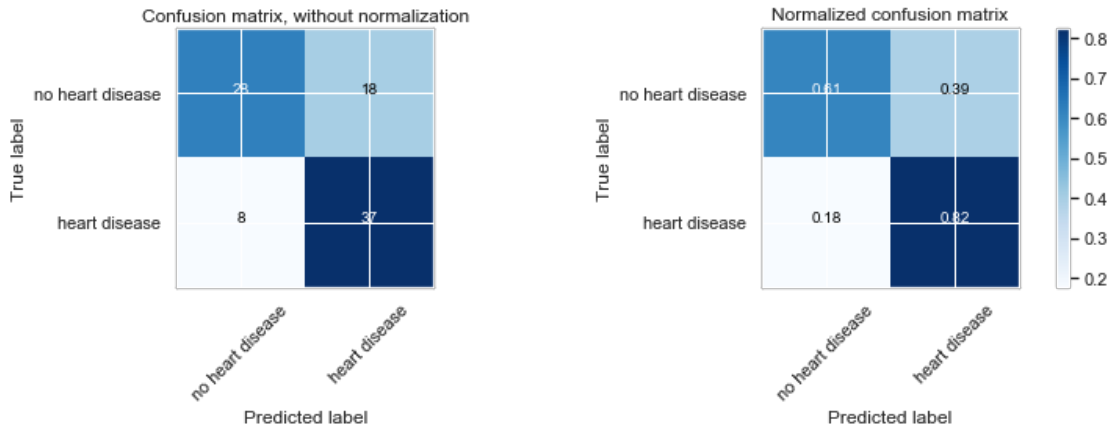

Figure 22: Confusion Matrix for Decision Trees with Normalized Dataset

Decision Tree model trained with unnormalized dataset achieved better accuracy than the decision tree that was trained with normalized dataset. Moreover, it can be observed from the confusion matrix that in 13 % of the cases that the Decision Tree Model trained with unnormalized data, predicts that patient did not have a heart disease when they had. The Decision Tree model trained with normalized data performed even worse and same score for it was even 18 %. Therefore, it can be concluded that Decision Tree model trained with unnormalized data set is a better choice from these two options.

Results

1. Comparison of regression models

The best model for predicting the presence of the heart disease is **Logistic Regression model** that was trained with **normalized data**. Although the SVM model trained with unnormalized dataset has the same accuracy, the Logistic Regression model is a preferred choice because the number of cases in which it predicted that someone did not have a heart disease when they actually did was 4 % as compared to 7 % in SVM. This kind of difference is significant in the healthcare field and such difference could lead to loss of human lives.

The Receiver Operating Characteristic (ROC) curve is presented in the figure 13 and it can be seen that the results are quite similar for all models presented.

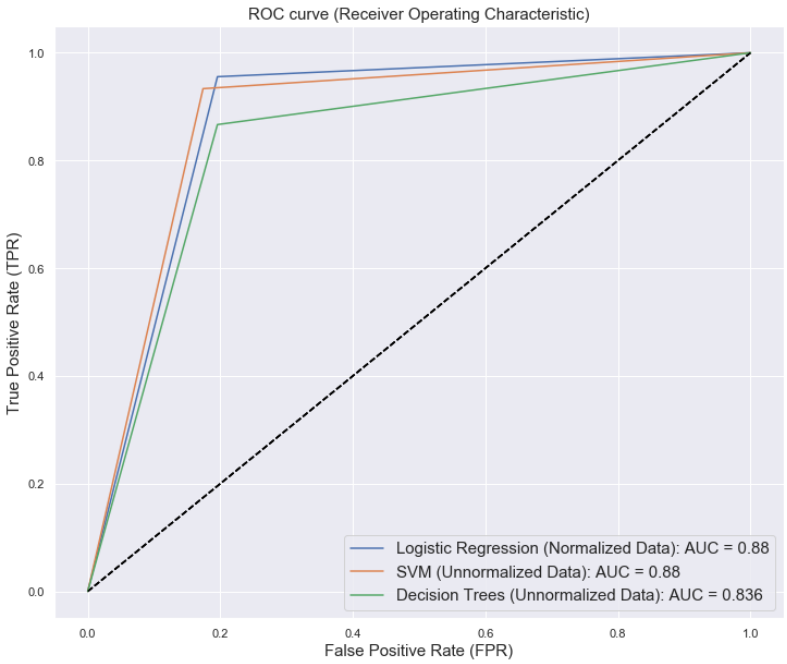

Figure 13: ROC curve

2. Effective attributes

As presented earlier, the Logistic Regression model trained with normalized data was the best performing model according to combinated accuracy score and confusion matrix analysis. Therefore, this model is chosen to be the one that is used to analyze which attributes are having the most significant effect in the model. These attributes are interesting since they are the ones that are the most accurate ones to predict if patient has heart disease.

Figure 14 presents importance of different variables. From there it is seen that the clearly most important feature is small value for *ca* (number of major vessels colored). Importance of this feature is circa 2.3 whereas the rest of the features have importance less than 1. There are still features indicating heart disease having absolute importance of at least 0.8: non-anginal chest pain (*cp*), fixed heart tissue defect (*thal*) and down slope (*slope*). To reflect to the goals of this study, the aforementioned attributes are the ones having the most importance. In total there are still 10 features having absolute importance over 0.5, but those are not expected to be most significant ones.

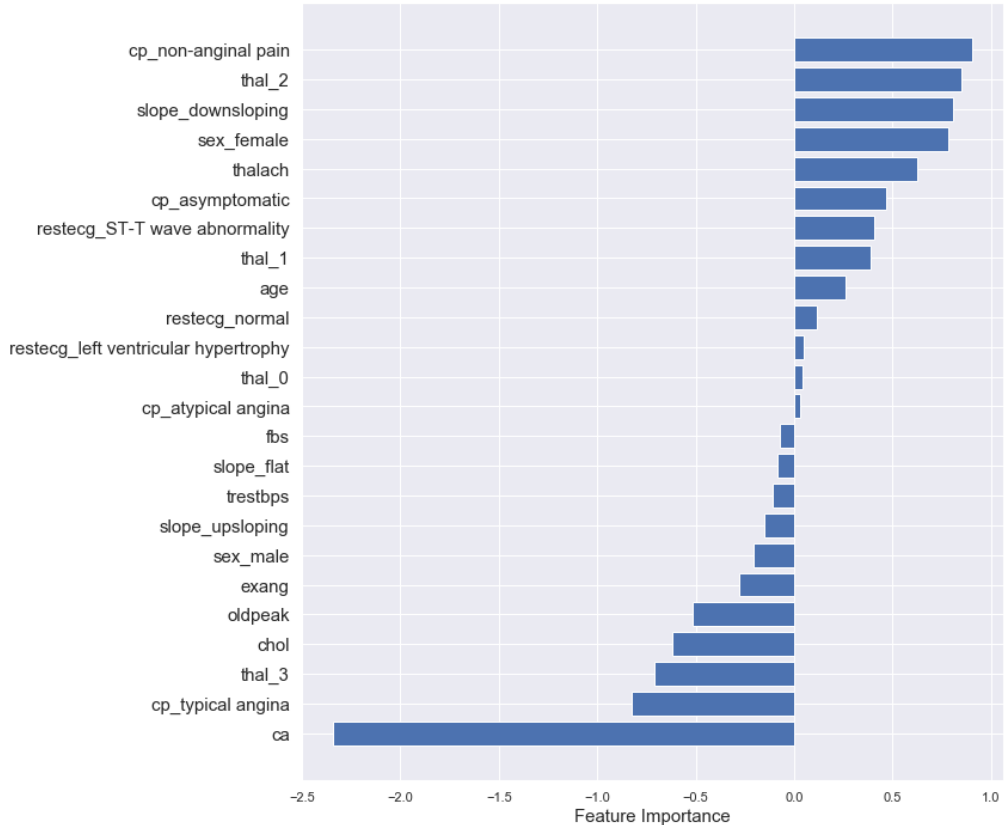

Figure 14: Importance of the variables.

There are still 14 features having smaller absolute importance than 0.5. From these *fbs*, *tresbps* and *exang* should be highlighted since they are not artificially created dummy variables but original variables. Because of the small importance of these three variables it can be questioned if they are needed in the hear disease identification process.

One of the goals of this study was to identify optimal set of tests in optimal order. Decision Tree model could have been used to conduct this but since it did not have enough good accuracy, we define the order by using costs and importance of information certain tests are providing. In addition to *age* and *sex* which are easy and straight forward to "test" we recommend conducting tests for all features which are identified having importance more than 0.8. If these tests result in negative result further test should be conducted to be sure that patient is not diagnosed not to have heart disease if that is not really the case.

3. Economic effects

Earlier we presented three attributes which are having only a minor importance for identifying heart disease. The costs of these tests are *fbs*:3; *tresbps*:1 and *exang*: 1, where for *fbs* and for *exang* it is expected that other tests belonging to same category are conducted since they are having more importance for the prediction. However, *fbs* is tested with *chol* which is not one of the most important features and has only limited importance and therefore it could be further studied and

discussed if it could also be left out. Without leaving *chol* out the total saving potential would be 5 equaling 3.7 % of total cost of all tests (134). By leaving out *chol* the saving would increase quite significantly to 9.0 %. It is good to remember that in addition to cost savings leaving out some of the tests would ease pressure that healthcare personnel are having and could also increase customer satisfaction. However, the dataset does not offer data to evaluate the possible amount of time savings and therefore we are not in a place to evaluate exact values for that and it could be further studied.

CONCLUSION

The study succeeded in identifying most important features to identify heart disease and from these *ca* was the most important one. This allowed that optimal set of tests and order of tests conducted could be evaluated and it is recommended to conduct the most tests providing information about most important features first and after that confirm the result with other tests if needed. Also, the study found that *fbs*, *tresbps* and *exang* tests are not providing much of importance for the test and therefore they could be considered to be left out in addition with *chol*. However, it has to be remembered that this study is based on one relatively small dataset and that in healthcare whole picture should be evaluated. For example, when evaluating whole healthcare system, it could be beneficial that even more tests (with higher cots) are conducted for the patient if it reduces need for extra visits in healthcare and costs to society related to that (like sick leave). Therefore, we recommend using our study as a base for further research and before leaving any of the tests out we recommend conducting more specific study to that particular test to better evaluate all of the effects. This is crucial to guarantee that no fatal mistakes are done.

REFERENCES

[1] Kaggle Inc, "Heart Disease UCI," 17 2 2019. [Online]. Available: https://www.kaggle.com/ronitf/heart-disease-uci.

[2] UC Irvine, "Heart Disease Data Set," 16 2 2019. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Heart+Disease.

[3] Official Data Foundation, "Canada Inflation Calculator," 17 2 2019. [Online]. Available: http://www.in2013dollars.com/ca/inflation/1995?amount=100.

[4] Currency Rate, "1.54 CAD Canadian Dollar to EUR Euro," 17 2 2019. [Online]. Available: https://cad.currencyrate.today/eur/1.54.

[5] UCI, "Heart disease," Costs, [Online]. Available: https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/costs/. [Accessed 17 2 2019].

[6] Princeton University Library, "Working with dummy variables," [Online]. Available: https://dss.princeton.edu/online_help/analysis/dummy_variables.htm. [Accessed 17 2 2019].