**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 at 85% accuracy. The most important attributes are chest pain type, fluoroscopy results, thallium heart scan, ECG + treadmill test and blood cholesterol test. This method saves 4% of costs and greatly simplifies the process by identifying test which do not add value.

Predictive Modelling

Group 17

Rohit Saluja

Antti Korhonen

Rainer Klein

Table of Contents

[Background 3](#_Toc1391434)

[Goals 3](#_Toc1391435)

[1. Identify heart disease predictors 3](#_Toc1391436)

[2. Suggest the optimal test and their order 3](#_Toc1391437)

[3. Identify unnecessary tests 3](#_Toc1391438)

[Solution Process 3](#_Toc1391439)

[1. Data exploration 3](#_Toc1391440)

[Original Data 3](#_Toc1391441)

[Costs 5](#_Toc1391442)

[2. Data preparation 5](#_Toc1391443)

[Feature extraction 5](#_Toc1391444)

[Feature scaling 5](#_Toc1391445)

[Data split 5](#_Toc1391446)

[3. Modelling 6](#_Toc1391447)

[Logistic Regression 6](#_Toc1391448)

[Support Vector Machines (SVM) 7](#_Toc1391449)

[Decision Trees 8](#_Toc1391450)

[Results 9](#_Toc1391451)

[1. Comparison of regression models 9](#_Toc1391452)

[2. Effective attributes 10](#_Toc1391453)

[3. Economic effects 10](#_Toc1391454)

[References 11](#_Toc1391455)

# Background

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 attributes. One of the attributes is the existence of heart disease.

# Goals

## Identify heart disease predictors

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

## 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.

## Identify unnecessary tests

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

# Solution Process

## 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.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 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 |

Table 1: Attributes

In each chart, 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 five appear to show significant correlation: cp (Figure 1), ca (Figure 5), exang (Figure 3), slope (Figure 2) and thal (Figure 4).

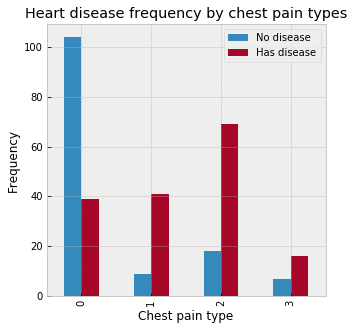


Figure 1: Chest pain type (cp) vs heart disease

|  |  |
| --- | --- |
| Chest pain types:   1. Typical angina 2. Atypical angina 3. Non-anginal pain 4. Asymptotic | Slope categories:   1. Up 2. Flat 3. Down |
| Thalassemia defect type:   1. Normal 2. Fixed defect 3. Reversible defect |  |

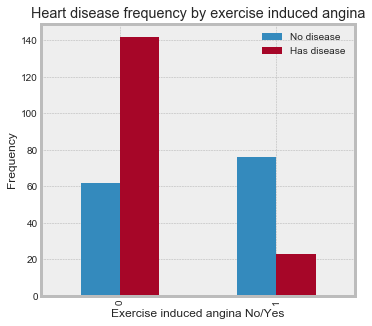


Figure 3: Exercise induced angina (exang) vs heart disease

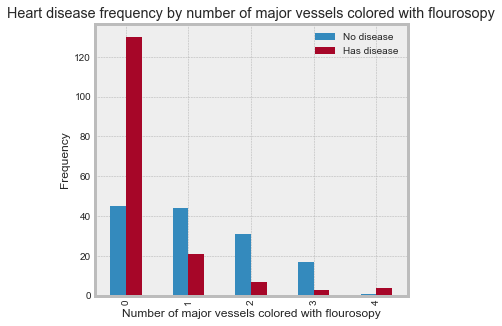


Figure 5: No. of major vessels colored by fluoroscopy (ca) vs heart disease

### Costs

Table 2 displays the cost of the test required to attain each attribute value. It was originally presented in 1985 Canadian Dollars, which as it turns out 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.

For example, the classic treadmill + ECG test 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

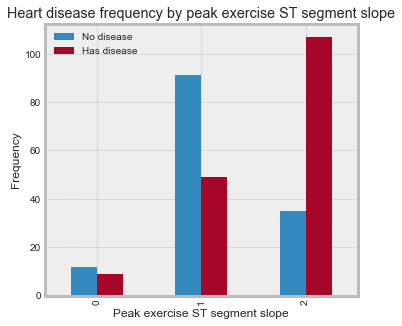


Figure 2: Peak exercise ST segment slope (slope) vs heart disease

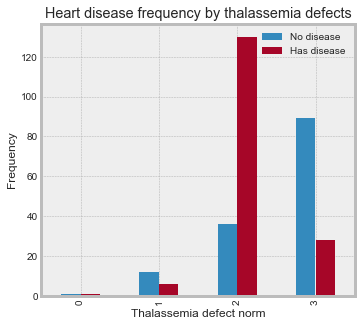


Figure 4: Thalassemia defect type (thal) vs heart disease

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 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 |

## Data preparation

### Feature extraction

From data exploration it became clear that some of the studied attributes (sex, cp, restecg, slope, thal) were categorical. Fitting a regression model on these requires the use of dummy variables [6]. To this end pandas get\_dummies function was utilized.

### Feature scaling

We normalized our data using feature scaling from sklearn preprocessing, quantile\_transform function because some continuous values were very large in magnitude and we wanted to bring every value to the same scale range so that the models don’t give more weight to features because they are bigger in magnitude. In each further modelling step the effect of normalization is assessed by comparing the results with the model trained on unnormalized data.

### Data split

The data was split into a training and a test set with proportions 70/30. 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 enough data for training.The train\_test\_split function from sklearn was utilized.

## 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 normal (unnormalized) dataset and the normalized dataset, the confusion matrix and accuracy was compared in the end to choose the best predictive model.

### Logistic Regression

The first algorithm that we used for heart disease prediction was Logistic Regression from the sklearn.linear\_model package.

Table : 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 % |

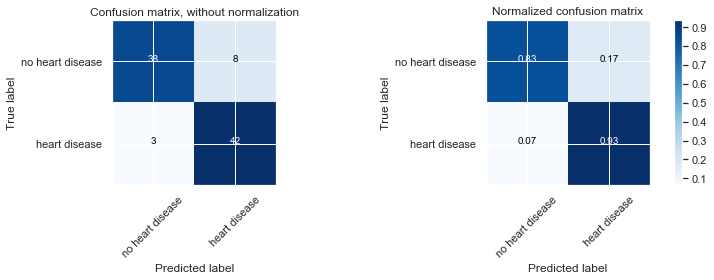


Figure : Confusion Matrix for Logistic Regression with Unnormalized Dataset

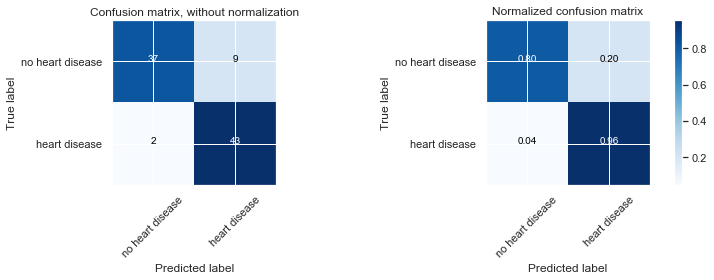


Figure : Confusion Matrix for Logistic Regression with Normalized Dataset

The main criteria that we used to evaluate the correctness of our model was the number of times the model predicted someone did not have a heart disease, but they actually had it as we are well aware that only looking at accuracy as a measure of success for the model can often be misleading. So, in healthcare where lives of people can be put into jeopardy by the slightest miscalculation, the lower this number (percentage), the more useful and better the model.

As it can be observed from the confusion matrix that in 7% cases the Logistic Regression Model with Unnormalized data predicts that someone did not have a heart disease when they actually did. This is has happened in 4 % of the cases for the Logistic Regression Model with Normalized Data. So, it is clearly visible that Logistic Regression model with Normalized Data set beats the Logistic Regression model with Unnormalized Dataset.

### Support Vector Machines (SVM)

The second algorithm that was used for heart disease prediction was the Support Vector Machines (SVM) from sklearn.svm.svc. 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 data set were C = 1, kernel = linear, gamma = auto and the best hyper parameters for the SVM model with the normalized data set were also C = 1, kernel = linear, gamma = auto. The results for both the SVM models have been summarized below.

Table : Accuracy comparison of SVM with Unnormalized and Normalized Data

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

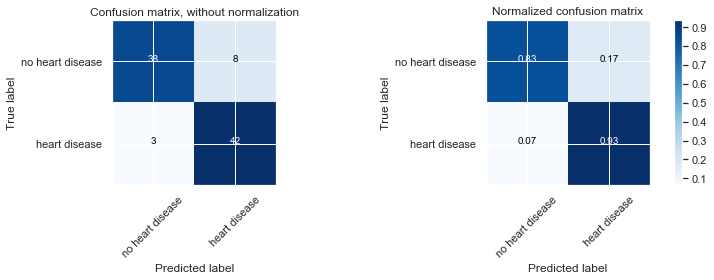


Figure : Confusion Matrix for SVM with Unnormalized Dataset

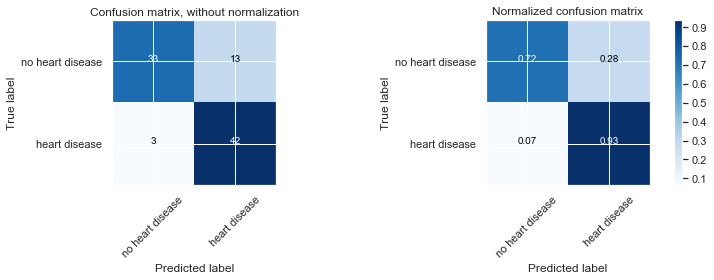


Figure : Confusion Matrix for SVM with Normalized Dataset

It is visible from the confusion matrix that in 7% cases the SVM Model with Unnormalized data predicts that someone did not have a heart disease when they actually did. This is has happened in 7 % of the cases for the SVM Model with Normalized Data as well. Keeping this information in mind we feel that no model is particularity a better choice than the other one.

### Decision Trees

The third and final algorithm that we tried was decision trees from the sklearn.tree DecisionTreeClassifier. We tried to tweak the maximum depth of the decision tree with the max\_depth parameter and during the process we found out that max\_depth = 3 was giving the best result and the accuracy was decreasing as we were increasing it. The results for decision trees trained with unnormalized and normalized dataset have been summarized below.

Table : 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 : Confusion Matrix for Decision Trees with Unnormalized Dataset

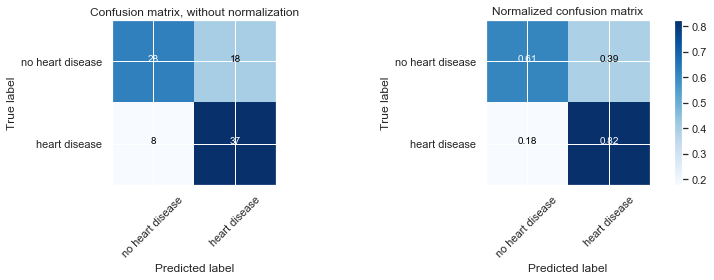


Figure : Confusion Matrix for Decision Trees with Normalized Dataset

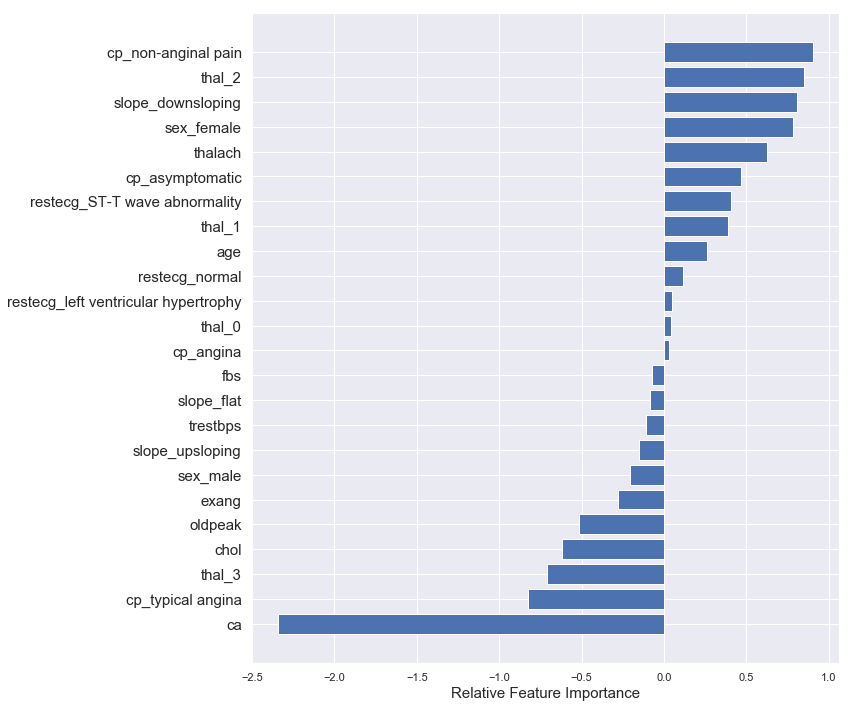
Decision tree that was trained with unnormalized data set had a better accuracy than the decision tree that was trained with Normalized dataset. Moreover, it can be observed from the confusion matrix that in 13% cases the Decision tree Model that was trained with Unnormalized data predicts that someone did not have a heart disease when they actually did, and this is 18 % in the case of the decision tree that was trained with normalized data. So, it can be concluded that Decision Tree model trained with unnormalized data set is a better choice.

### Results

## 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 no 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 and in medical world such an error could lead to loss of human life.  The AUC for the Logistic Regression model (normalized data) is also comparable to SVM (Unnormalized Data) and better. | /var/folders/w9/svvppq4d39x334g0p5ttlh9h0000gn/T/com.microsoft.Word/Content.MSO/C358E880.tmp |
|  | Figure : ROC curve |

## Effective attributes



## Economic effects

# 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]. |