**Heart Disease**

To learn which tests matter in detecting heart disease, three regression methods (logistic, SVM and decision trees) were utilized. Decision trees 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 2](#_Toc1303867)

[Goals 2](#_Toc1303868)

[1. Identify heart disease predictors 2](#_Toc1303869)

[2. Suggest the optimal test and their order 2](#_Toc1303870)

[3. Identify unnecessary tests 2](#_Toc1303871)

[Solution Process 2](#_Toc1303872)

[1. Data exploration 2](#_Toc1303873)

[Original Data 2](#_Toc1303874)

[Costs 2](#_Toc1303875)

[2. Data preparation 4](#_Toc1303876)

[Normalization 4](#_Toc1303877)

[Data split 5](#_Toc1303878)

[3. Modelling 5](#_Toc1303879)

[Logistic Regression 5](#_Toc1303880)

[SVM 5](#_Toc1303881)

[Decision trees 5](#_Toc1303882)

[Results 6](#_Toc1303883)

[1. Comparison of regression models 6](#_Toc1303884)

[2. Effective attributes 6](#_Toc1303885)

[3. Economic effects 6](#_Toc1303886)

[References 7](#_Toc1303887)

# 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.

# Goals

## Identify heart disease predictors

The dataset consists of 14 attributes, one of which is the presence of heart disease. 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 simply achieved with histograms. Below, in Table 1, these attributes are presented.

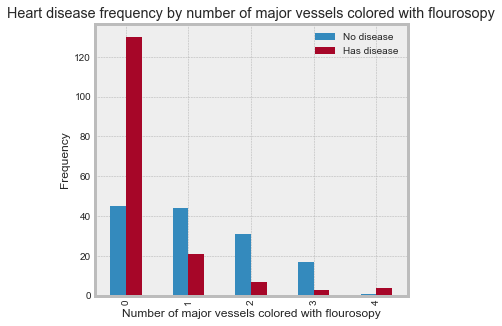
### Costs

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.

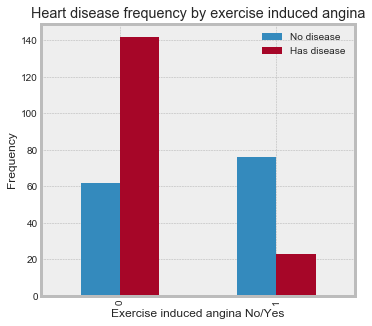
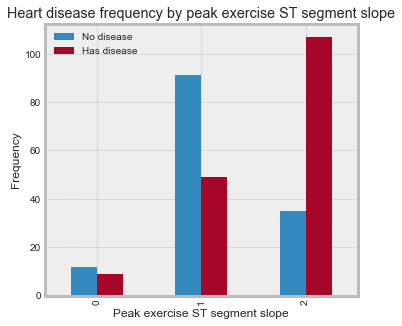
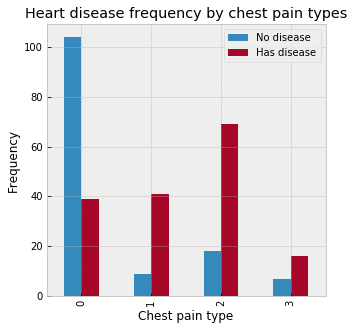
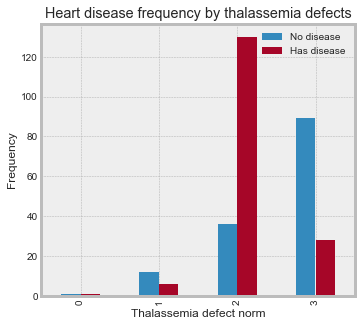
For example, the classic treadmill + ECG test provides values for 3 attributes – *exang*, *oldpeak* and *slope*. Once the test has been carried out to get a value for slope, gaining values for also *oldpeak* and *exang* has a positive but nearly zero marginal cost.

Table : 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 |

Table : 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 |



## Data preparation

### Normalization

### 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

### Logistic Regression

### SVM

### Decision trees

# Results

## Comparison of regression models

The most accurate

## 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]. |