Introduction

In Estonia, cardiovascular diseases are the leading cause of death. Therefore for our project, we decided to predict the risk of heart disease prior to negative outcomes like myocardial infarctions (heart attacks) taking place. To achieve our goal, we used different data mining methods and trained several predictive models.

Data

The dataset used for this project was obtained from Kaggle. The data was collected by The Behavioral Risk Factor Surveillance System (BRFSS) which is the premier system of health-related telephone surveys in the United States. The survey is conducted annually and each year the dataset contains more than 300 columns.

Based on extensive research regarding heart disease risk factors, we selected a subset of features from BRFSS 2015, when 441,456 adults were interviewed. To gain a better understanding of the variable names and answers, we had to thoroughly study the 137-page BRFSS codebook report: https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf

Preprocessing

Out of the original 330 columns in the BRFSS 2015 dataset we decided to keep only 22. Many columns in the original dataset either contain repetitive information or consist mostly of blank values due to being specifying questions to a previously asked question.

We decided to only use columns containing data previously proven to be related to heart disease as proven by previous medical research.

The remaining columns included data such as: whether the respondent has experienced heart disease (either coronary heart disease or myocardial infarction), whether the respondent has high blood pressure, the respondents age, BMI and more.

All of the chosen columns included ordinal data. If a row had a missing value in any column, the row was dropped.

Model

We trained multiple different models using different algorithms and sampling methods. We used five different training algorithms: Decision Tree, Random Forest, K Neighbors, Gaussian Naive Bayes and AdaBoost.

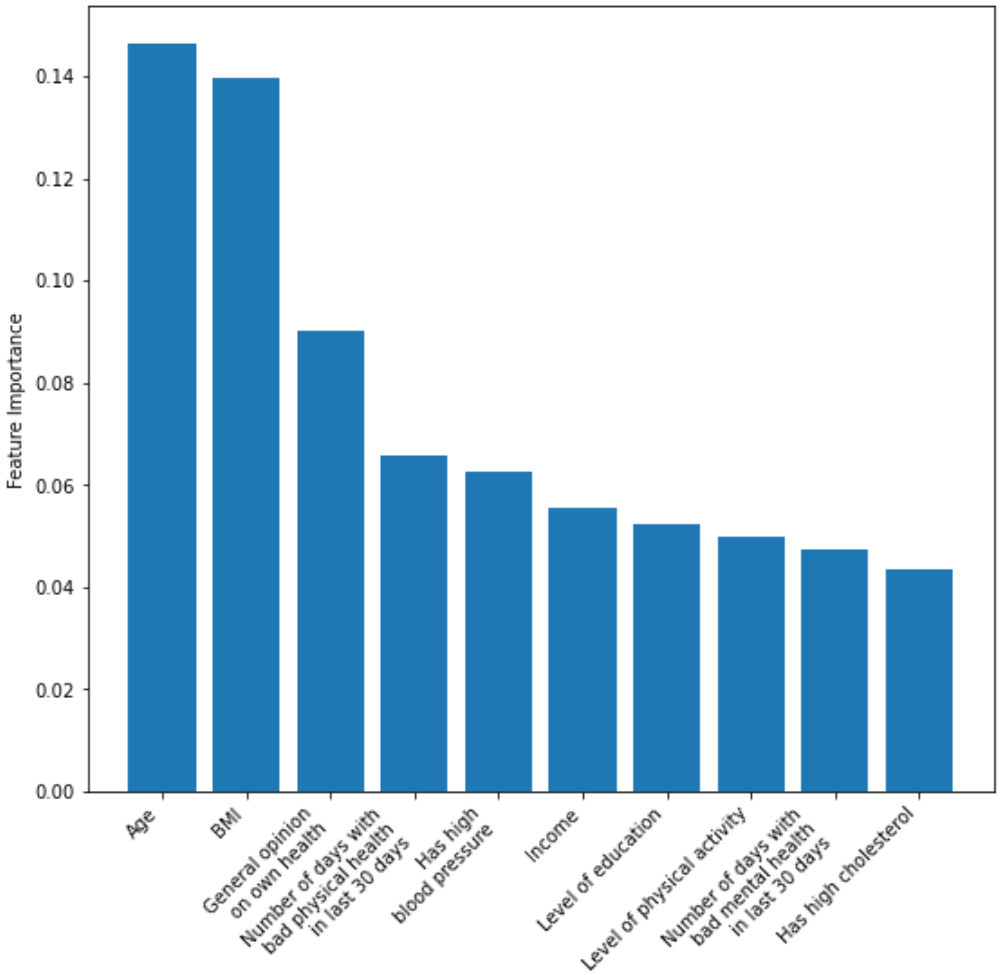
For every algorithm we used three different sampling methods: no sampling, undersampling and oversampling. In total we compared fifteen different models. The basis of comparison was the

F-scores of the models on the validation set. In general, models with unsampled data performed the worst and the models with oversampled data performed the best.

Results

Out of the trained models, the one with the best F-score on the validation set was a Random Forest classifier with oversampled data. On the validation data, the model achieved a F-score of 0.970. On the test data, the model achieved a F-score of 0.971, with the precision being 0.945 and accuracy 0.999. The 10 most important features and their importance are shown on graph 1.1.

Graph 1.1. Most important features and their importance



We also found features which are strongly correlated with heart disease (graph 1.2) by calculating pairwise correlations between all features and the target variable. The results of that activity mostly match those shown on graph 1.1.

Graph 1.2. Most important features according to absolute pairwise correlations