Prediction of heart failure

by Robert Nowak

Heart failure is a condition that develops when your heart doesn’t pump enough blood for your body’s needs. This can happen if your heart can’t fill up with enough blood. It can also happen when your heart is too weak to pump properly. The term "heart failure" does not mean that your heart has stopped. However, heart failure is a serious condition that needs medical care.

Every year in Poland there are about 70 thousand heart attacks. Every fifth patient dies within 12 months of the heart attack.

Aim of this project is to develop algorithm which will help identify risk group vulnerable to heart failures based on medical exams.

Algorithm was developed in python 3. The code is attached and written in JuPyther notebook.

Exploratory data analysis (EDA)

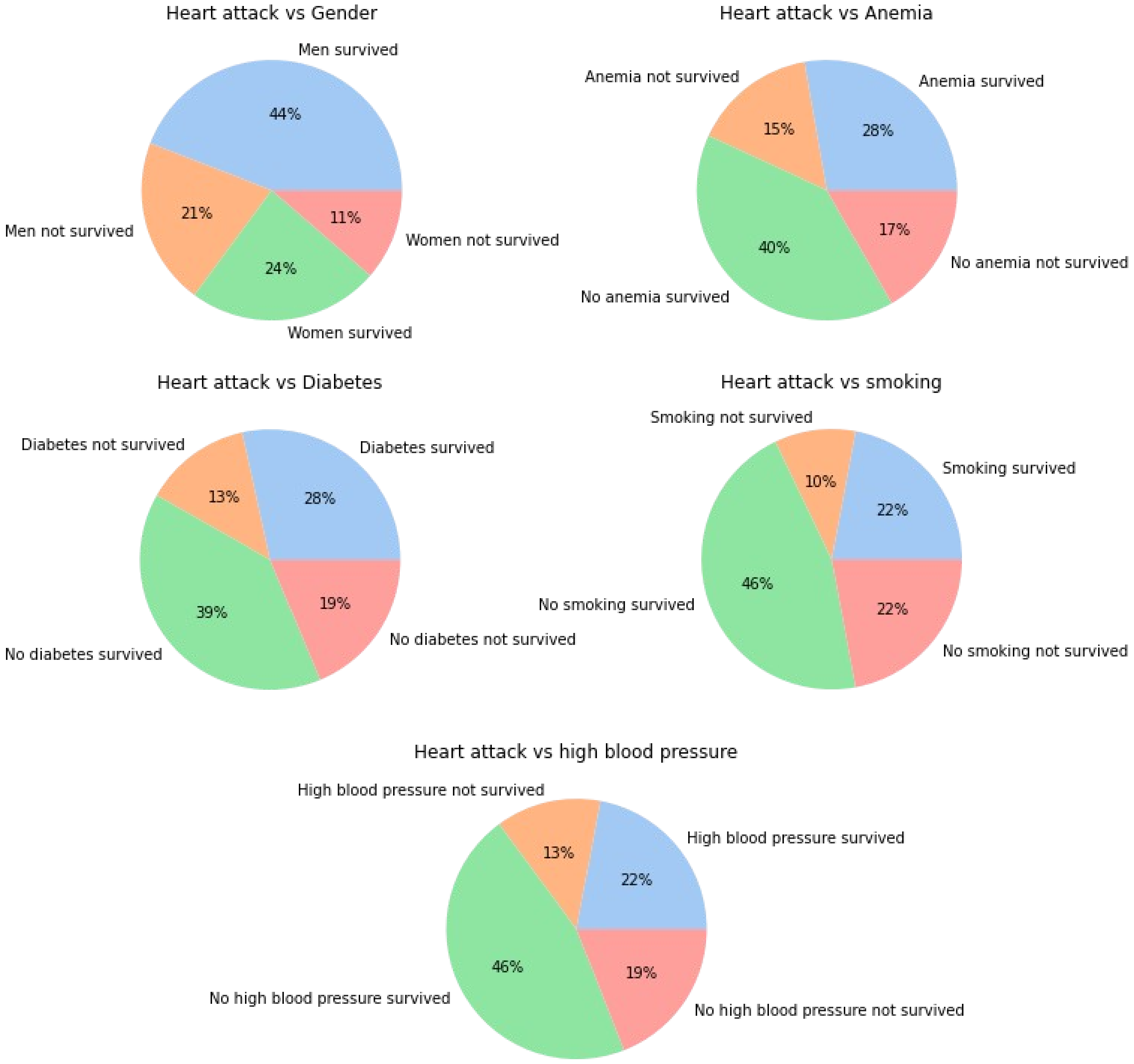
Database was downloaded from Kaggle, direct link:
https://www.kaggle.com/andrewmvd/heart-failure-clinical-data

This database consist of 299 rows and 13 columns:
- Age (Integer)
- anemia - decrease of red blood cells or hemoglobin (Boolean)
- creatinine phosphokinase - level of the CPK enzyme in the blood (mcg/L)
- diabetes - if the patient has diabetes (Boolean)
- ejection fraction - percentage of blood leaving the heart at each contraction (percentage)
- high blood pressure - if the patient has hypertension (Boolean)
- platelets - platelets in the blood (kiloplatelets/mL)
- serum creatinine - level of serum creatinine in the blood (mg/dL)
- serum sodium - level of serum sodium in the blood (mEq/L)
- sex - woman or man (binary)
- smoking - if the patient smokes or not (Boolean)
- time - follow-up period (days)
- DEATH_EVENT - if the patient deceased during the follow-up period (Boolean)

All boolean predictors were converted to integer with 0 for false and 1 for true and binary sex column was converted to integer with 0 for woman and 1 for man by the author.

Data is complete, without missing or wrong type values. There are no duplicated records.

Distribution of the categorical parameters



Heart attack vs Gender

Heart attack vs Anemia

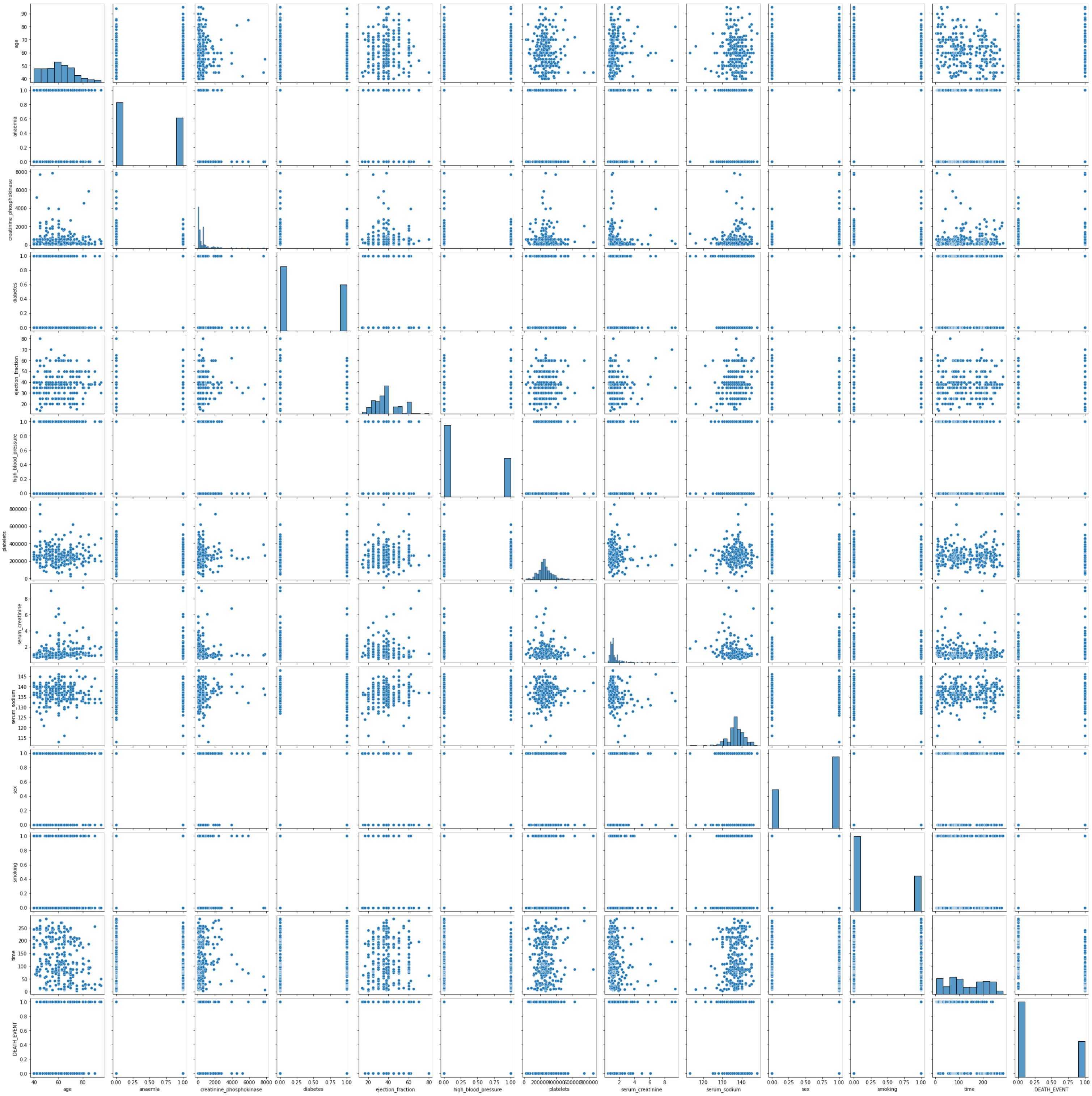
Heart attack vs Diabetes

Heart attack vs smoking

Heart attack vs high blood pressure

Patients who did not die and do not have given attribute make up approx 40% of the database. Given that the overall distribution of DEATH_EVENT is 68% survieved to 32% died, the algorithm might better predict patients from low risk group. The database might be imbalanced.

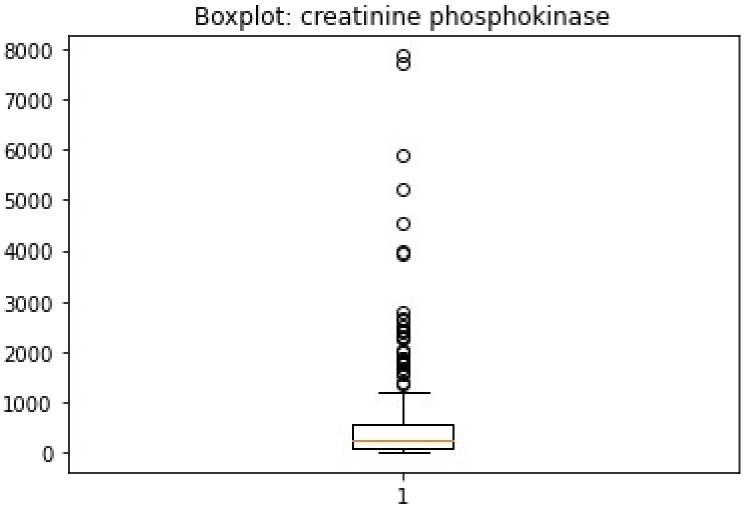
Correlations and histograms

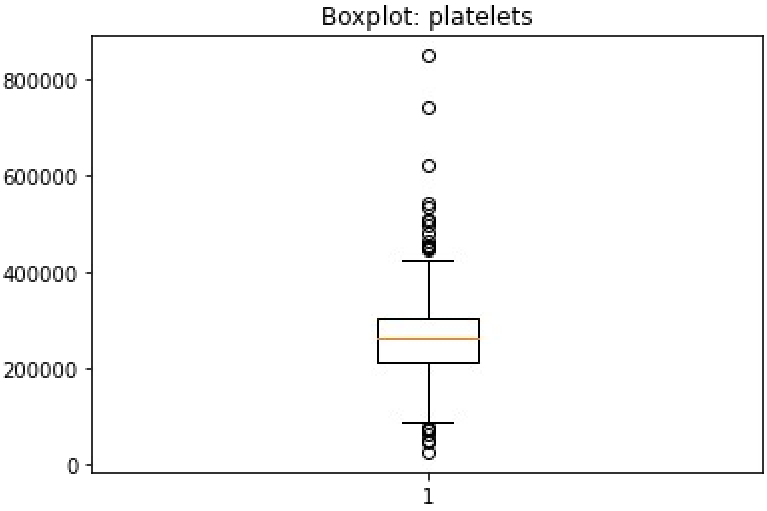


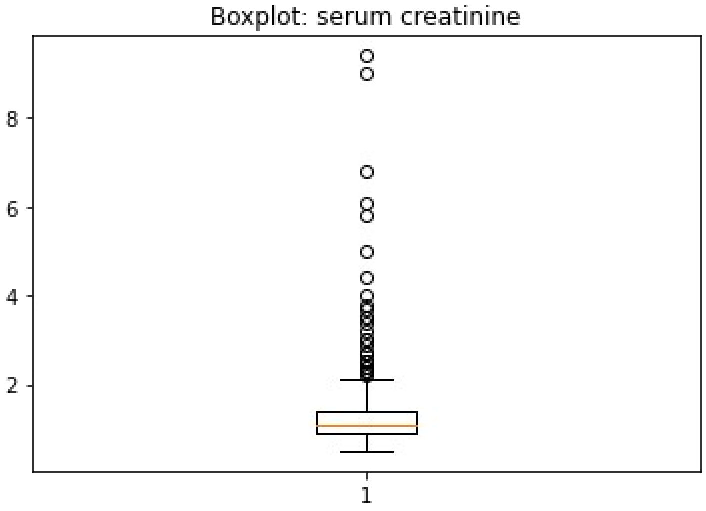
There are no high correlations in the database. Pairplot above proves how scattered the data is. It was also checked by calculating Pearson's and Spearman’s colinearity. For both tests the are no colerations with score higher than 0,7.

The histograms for each variable are in the diagonal of the pairplot.

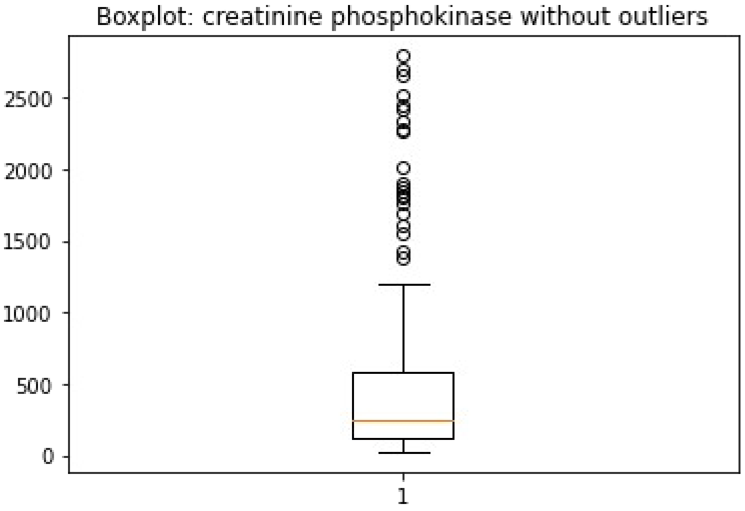
Outliers

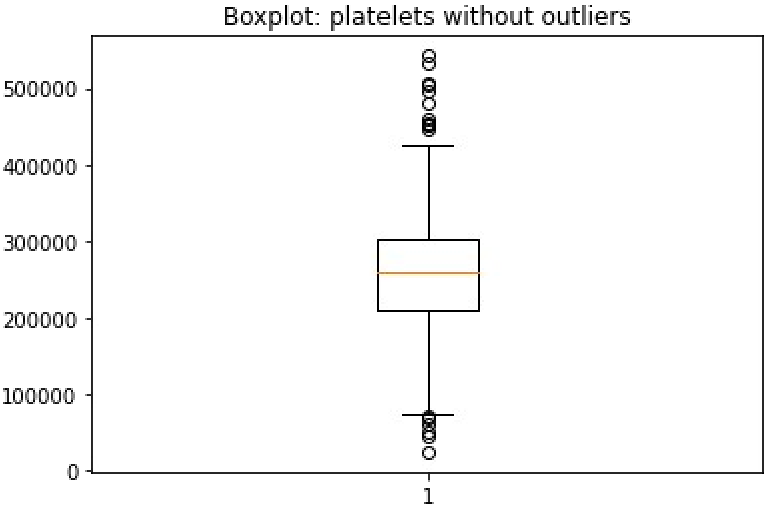


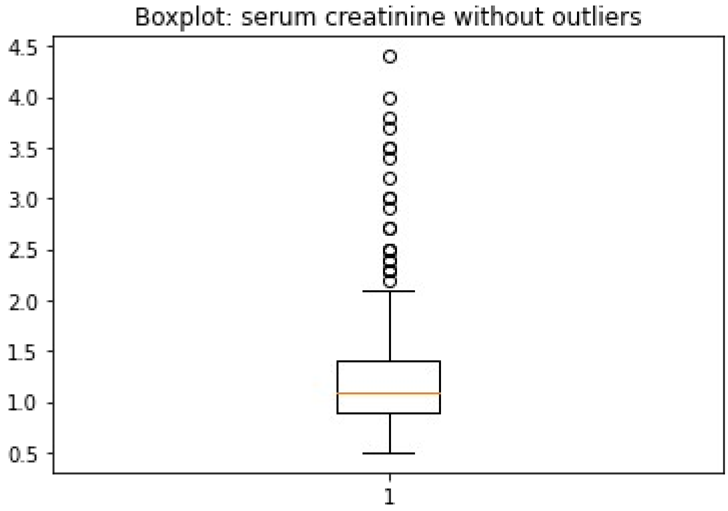




The only worrying boxplots are the one above, the rest doesn’t have any outliers or very few. Even though patients can have such extreme medical exams results influence of the outliers will be checked.
Any records with attribute higher or lower than 3 standard deviations from the median will be removed. Below are graphs when the outliers were removed.





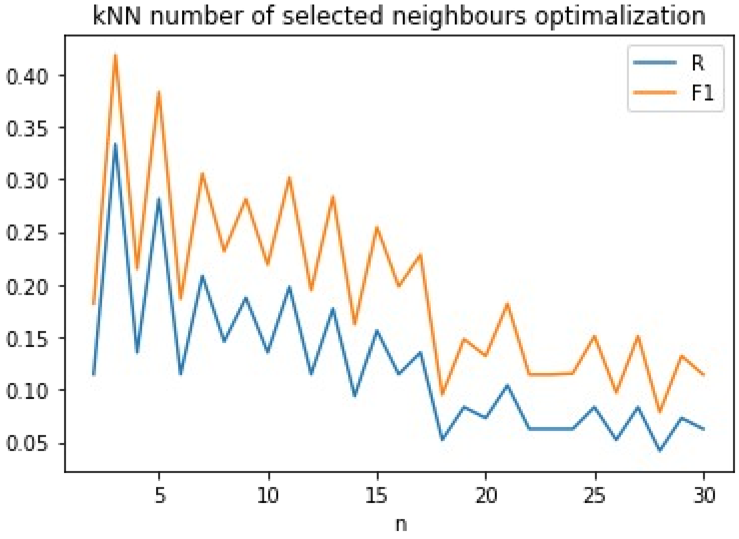


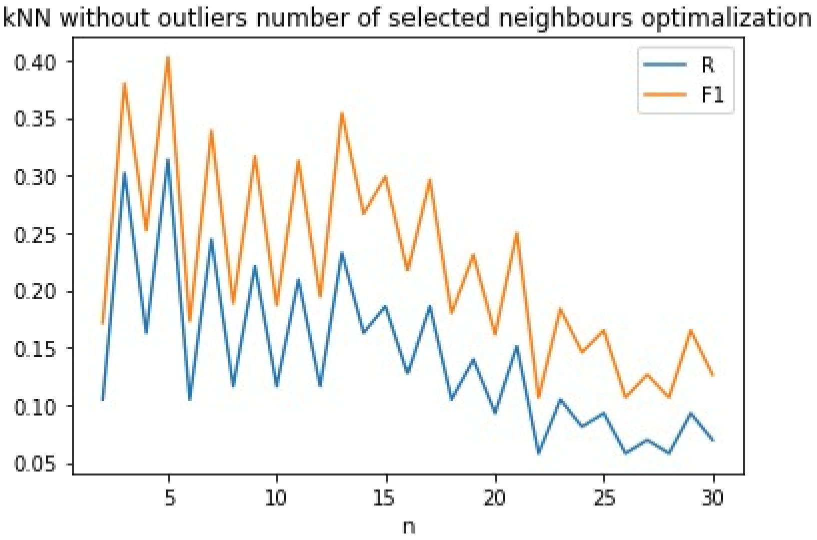
Modeling

Time variable has no sense in prediction of heart attack because it is follow-up time we cannot know before. Therefore it will be dropped from database.

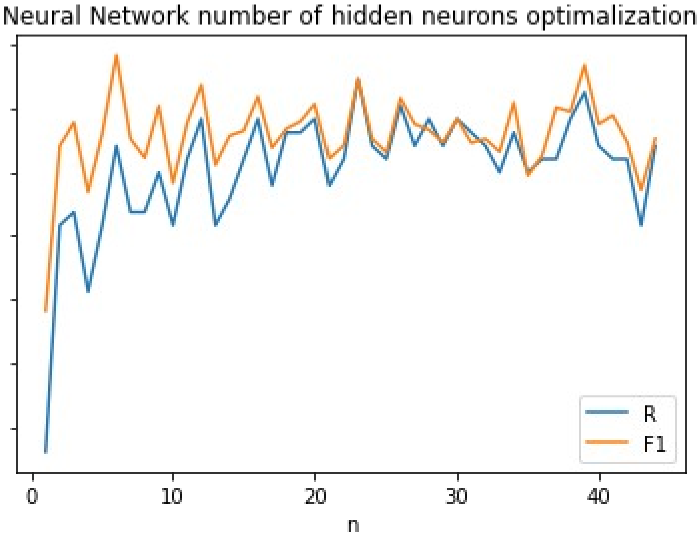
The most important is to correctly predict hearth attack for people from risk group. Wrong diagnosis could mean their death. Other way around if patient with low probability of heart attack were diagnosed incorrectly, to be in high risk group, nothing threatens his life.

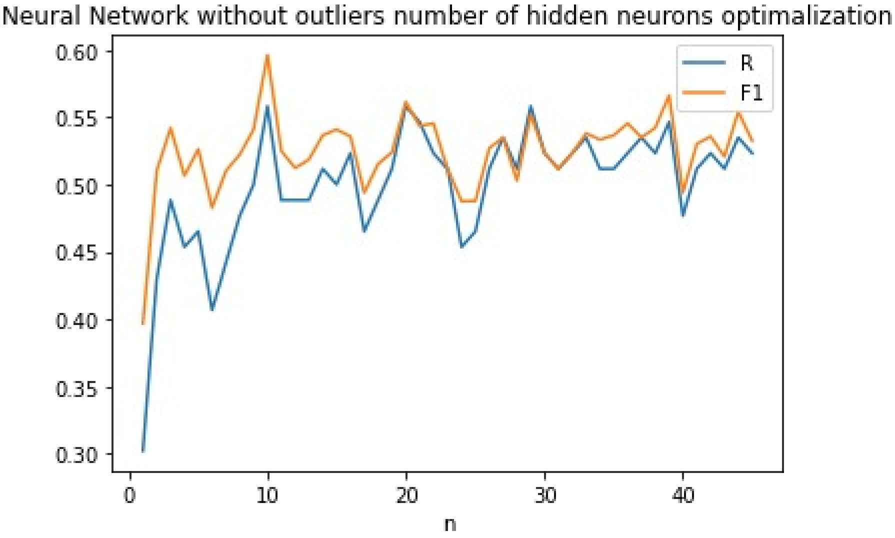
To estimate performance of selected models, recall and F1 score will be calculated. Recall – it describes how many patients with high risk were diagnosed correctly. F1 score describes overall performance of the algorithm. It is dependent on correct diagnosis of patients with high risk factor and on all other misdiagnosed.
Research will be developed by using k-Nearest Neighbor, Random Forest and Neural Network algorithms. Data will be standarized.




kNN without outliers number of selected neighbours optimalization

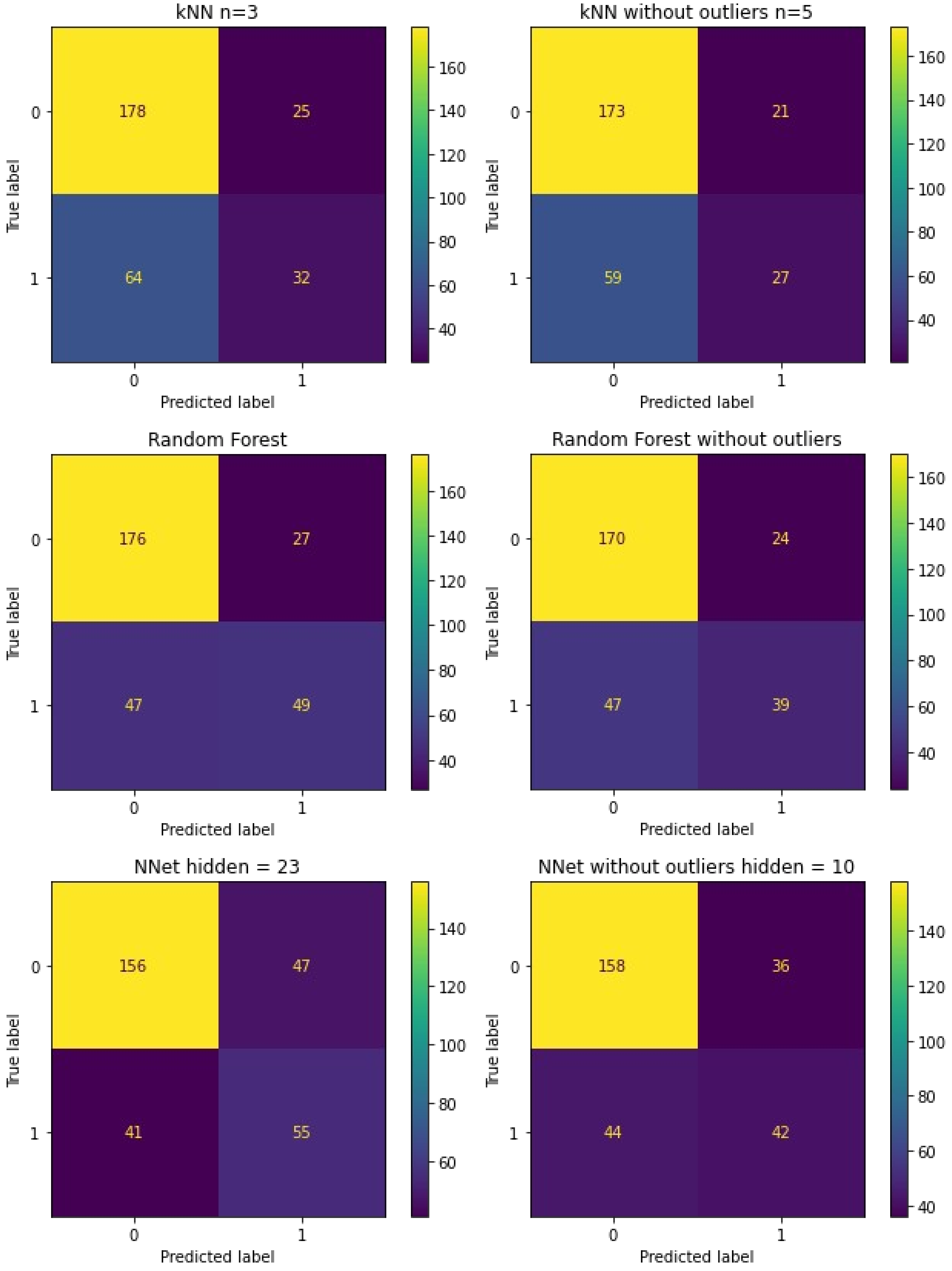
K-NN has best results are for n = 3 and for the case without outliers n = 5.


Neural Network number of hidden neurons optimalization

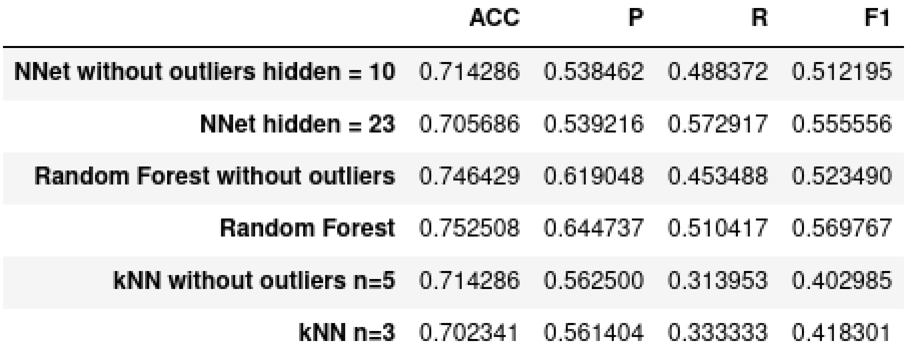

Neural Network without outliers number of hidden neurons optimalization

Due to high computational problems optimalization was only performed for one hidden layer.
Neural Network gives best results for 23 hidden neurons and for case without outliers 10.

Confusion matrices:



Results:



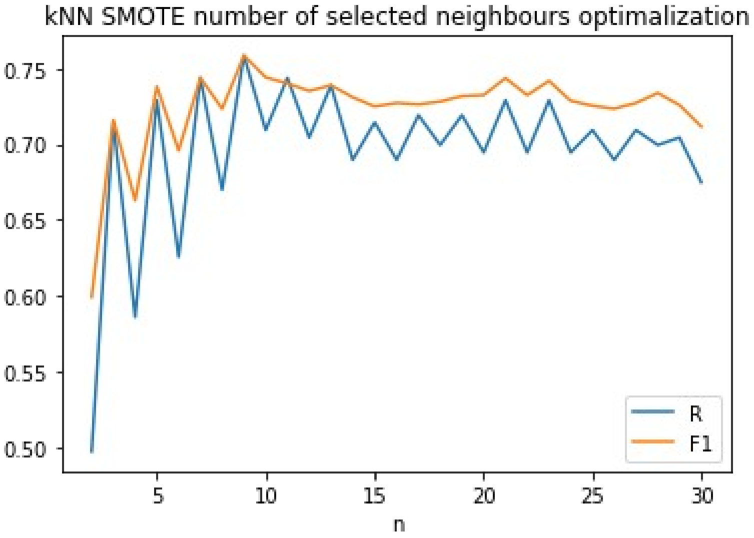
	ACC	P	R	F1
NNet without outliers hidden = 10	0.714286	0.538462	0.488372	0.512195
NNet hidden = 23	0.705686	0.539216	0.572917	0.555556
Random Forest without outliers	0.746429	0.619048	0.453488	0.523490
Random Forest	0.752508	0.644737	0.510417	0.569767
kNN without outliers n=5	0.714286	0.562500	0.313953	0.402985
kNN n=3	0.702341	0.561404	0.333333	0.418301

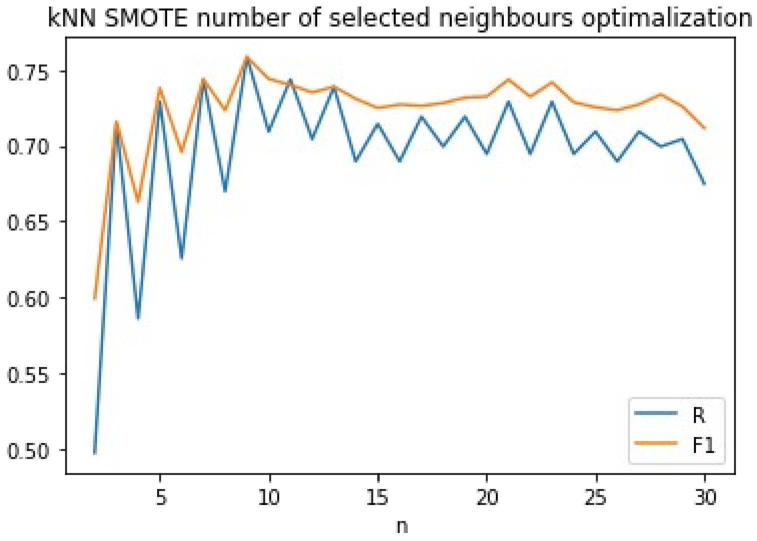
Data with removed outliers gave worse results. It means even though removed data looked like outlier it contained valuable information about this prediction.

SMOTE

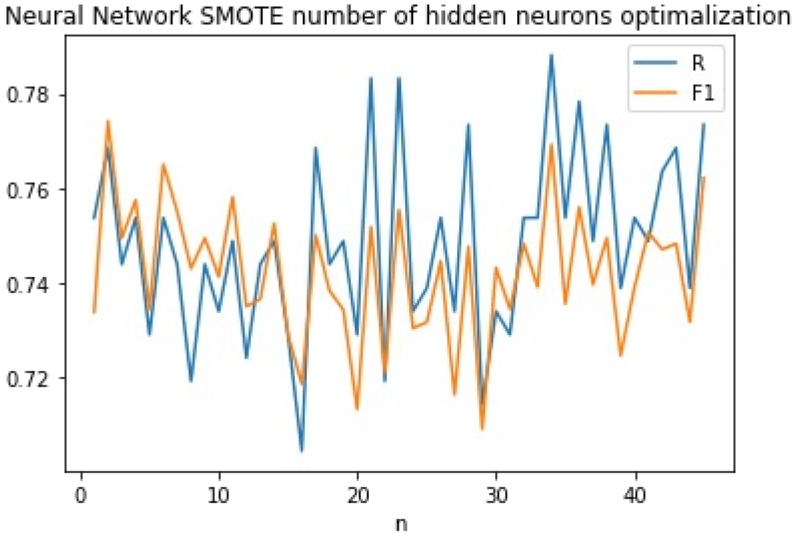
Results might be influenced by unbalanced database, 68% survived to 32% died because of heart attack. To overcome this problem Synthetic Minority Over-sampling Technique in short SMOTE will be used. This technique balance DEATH_EVENT column by oversampling records of patients who had heart failure. It will be performed on database with outliers since it gave better results then the one without them.

SMOTE generated 107 additional records. DEATH_EVENT is completely balanced 50% to 50% with 203 records for each case.



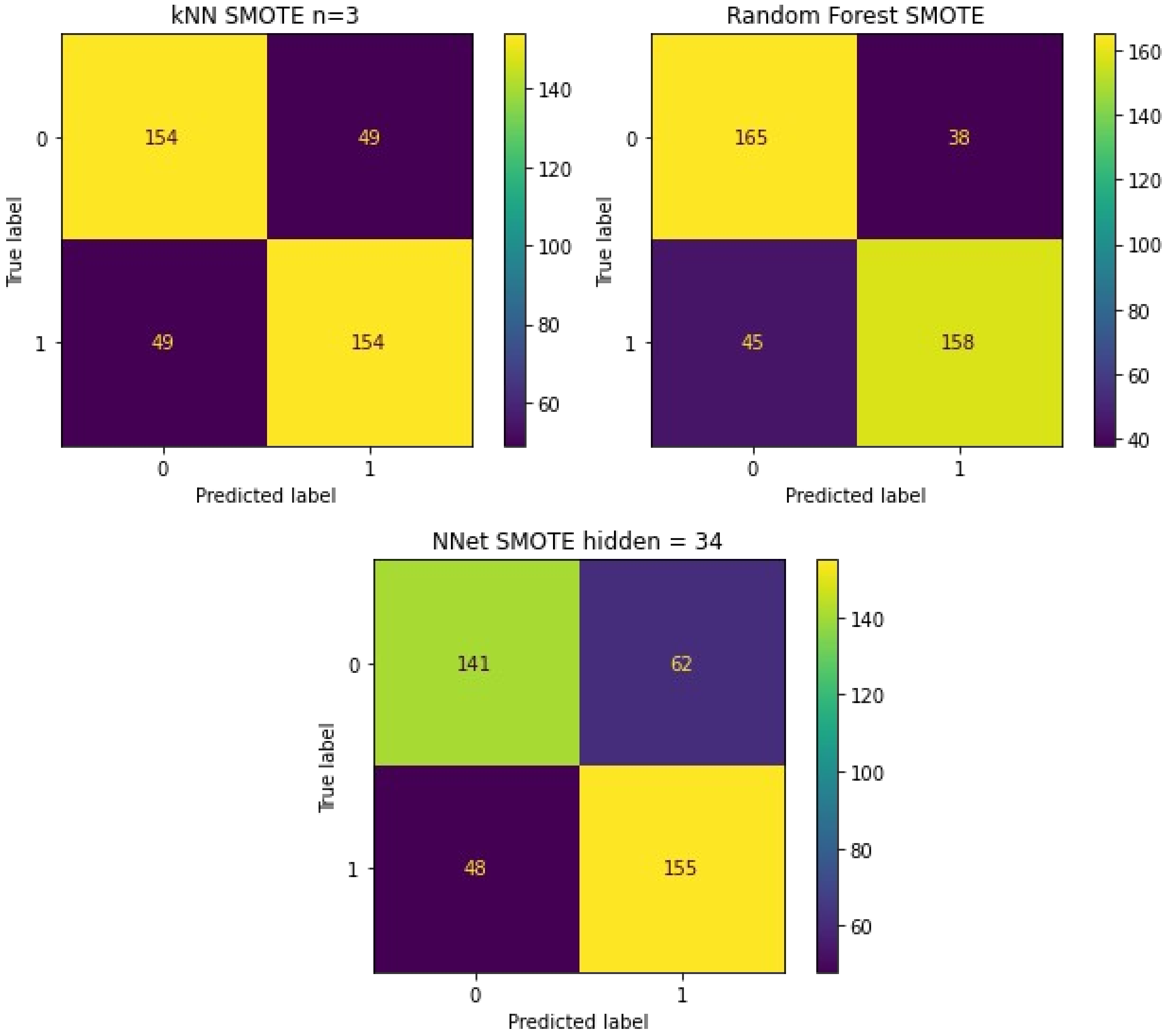


The best score for k-NN is for n=9.

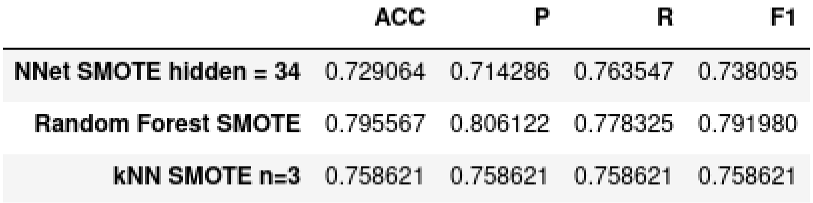


Neural Network has the best score with 34 neurons in hidden layer.

Confusion matrices:

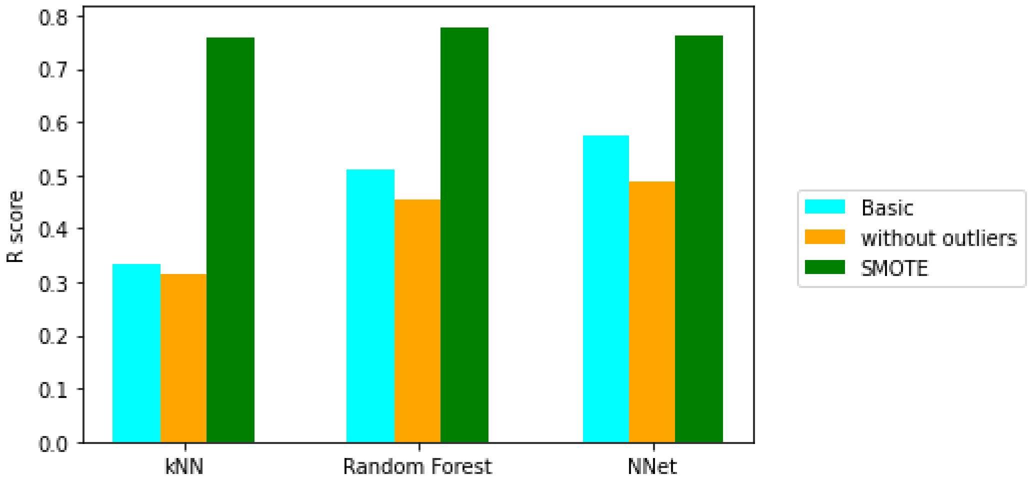


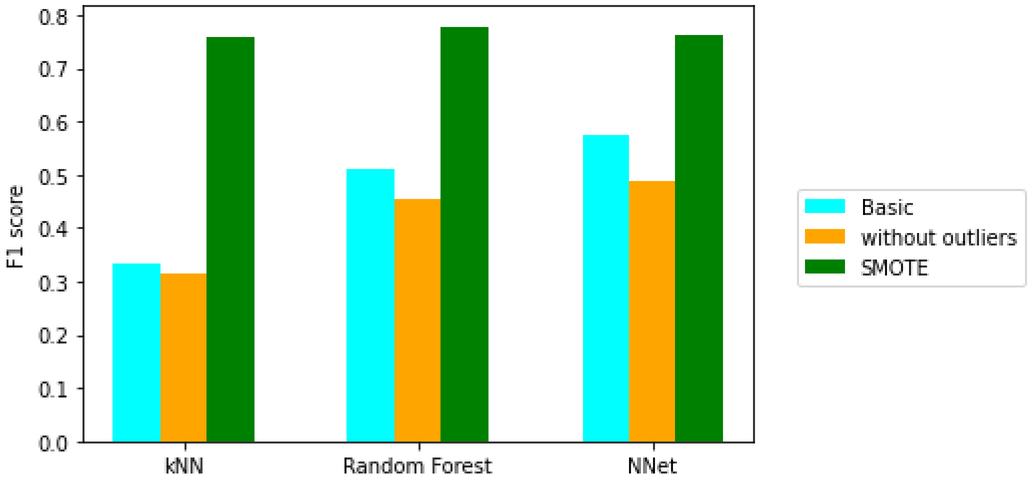
Results:



	ACC	P	R	F1
NNet SMOTE hidden = 34	0.729064	0.714286	0.763547	0.738095
Random Forest SMOTE	0.795567	0.806122	0.778325	0.791980
kNN SMOTE n=3	0.758621	0.758621	0.758621	0.758621

Conclusion





This research proved, not always, outliers should be removed. They might contain useful data and it is wise to check if in reality these variables can achieve these values.

Results for this database are terrible. Recall and F1 scores are the best for Neural Network with 23 neurons in hidden layer. Still they are terrible, R = 0.573, which mean's almost half of the patients from risk group were misdiagnosed.

SMOTE improved results. Now the best R and F1 score are for Random Forest with R = 0.778 and F1 = 0.792 but it means that 1 on 5 patients from risk group are misdiagnosed.

Bad results might indicate that predicting heart failure is more complex, it requires more data and/or more variables (results from medical exams). It is also possible that used models are not efficient for this topic. Further research is necessary.