**Heart Disease Prediction**

Adi Revach

Almog Yakovian

**Abstract**

One of the leading causes of death is heart disease.

In many situations the patient does not know that he is at risk until he has a heart attack, so we would like to predict if the patient is at increased risk for heart disease and thus help prevent heart attack or stroke.

In this study our recommendation for those findings is a step-by-step process of examinations to arrive at an accurate diagnosis and yet prevent patients from performing any unnecessary tests.

**Introduction**

We would like to predict which patients will be at high risk of developing heart disease based on various medical data.

As mentioned, patients who have experienced a heart attack may suffer irreversible damage to the heart muscle, which requires

essential modification in longevity, as they are at increased risk for further heart attacks and await long rehabilitation.

However, it is a reversible process if it is treated in time.

The database includes the following attributes for diagnosing high-risk patients: cholesterol, age, sex, blood pressure, diabetes, ECG test results, fluoroscopy, cardiac mapping test results.

The last column represents a final analysis using an X-ray examination and is therefore a factor of comparison.

The process we would like to offer is gradual when in the first stage the patient will undergo a set of routine tests for initial diagnosis and appropriate treatment.

If one of the tests raises suspicion, the patient will perform more comprehensive tests and he will refer for fluoroscopy

and cardiac mapping tests.

**Methodology:**

**Pre-Processing:**

The data set includes 270 rows that represent the patient and 14 columns that represent the attributes.

We classified each column into a suitable variable type:

In order to convert nominal information to binary we used the method one-hot.

We have converted the Target column from nominal to binary because there are 2 classifications:

1: Presence - There is heart disease

0: Absence - The patient is healthy

**Data Transformation:**

Normalization - since the data range is positive, we chose to perform a min-max normalization.

**Data Reduction:**

The data set includes columns that represent medical examination.

Our goal is to check if there is a correlation? between certain medical examinations in order to reduce costs without compromising the quality of the analysis.

1. Correlation test for the numerical variables:

It can be seen that there is no correlation between the numerical variables.

2. Correlation test between the presence of heart disease and type 2 diabetes.

Although the result showed that there is no correlation, we decided to keep this column because this examination is cheap and important for the patient's general health.

3. Anova test - We focused on the columns representing the tests ECG, fluoroscopy (Number of vessels fluoroscopy, and cardiac mapping (Thallium):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Cholesterol | ST depression | Exercise angina | Number of vessels fluoroscopy | Slope of ST |
| Thallium | ~0 | ~0 | ~0 | <0.05 |  |
| EKG |  |  |  | <0.05 | >0.05 |

\* for p-value > 0.05 we reject the 0 hypothesis and there is no significant relationship between the variables

4. Chi Square test for the nominal variables:

|  |  |  |
| --- | --- | --- |
| Slope of ST | Chest pain type |  |
| 28.3 | 32.2 | Thallium |

\* for  > critical value we reject the 0 hypothesis and there is significant relationship between the variables

Depending on the results, as an initial step, we decided to drop the fluoroscopy column and the Thallium column.

In the second stage, we will recommend performing the cardiac mapping test and fluoroscopy, in case the initial test raises the suspicion of the presence of the disease.

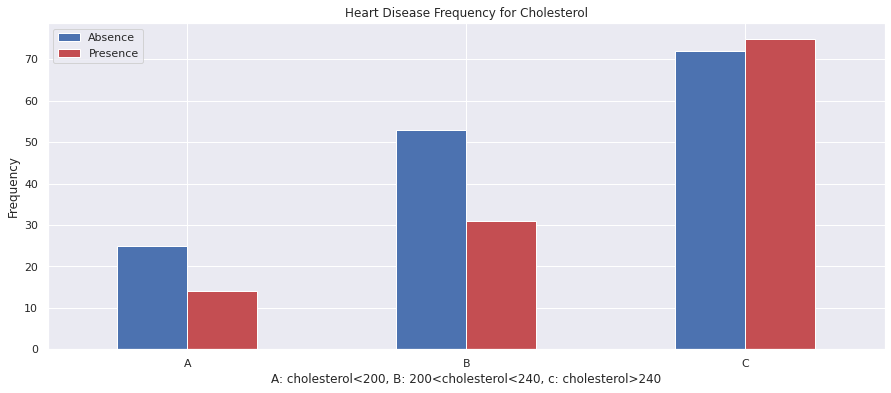
The following tests are tests in which we received a high correlation between them and the heart mapping test and fluoroscopy, so we will focus on them in order to decide on a follow-up test:

* EKG
* Exercise angina
* high cholesterol
* ST depression
* Slope of ST

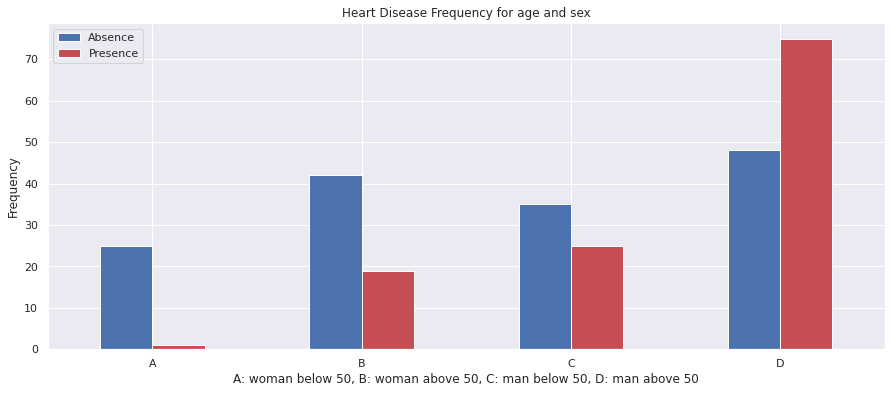
In addition, to get general insights on the dataset we visualized the data:

1. We examined the relationship between cholesterol and the presence of heart disease, according to which it can be seen that the higher the cholesterol

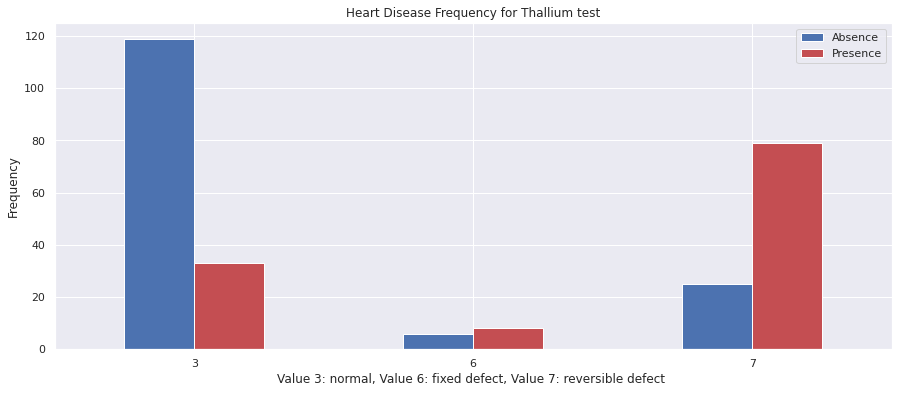
Higher, there is a higher chance of heart disease presence.



1. We examined the relationship between age, sex and and heart disease, and found that men have more heart disease From women, especially men over the age of 50.



1. We examined the association between the presence of heart disease and a thallium test, and found that there was a direct relationship between the test to disease.

****

**Algorithms**

\* The process is shown for both stages.

We have selected 3 comparison algorithms that are suitable for are classification problem:

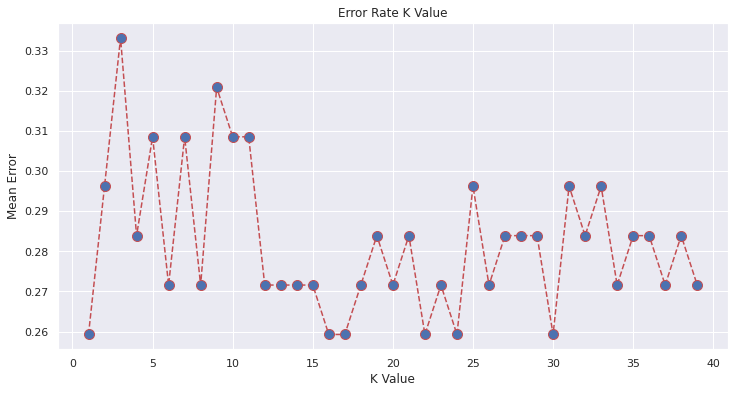
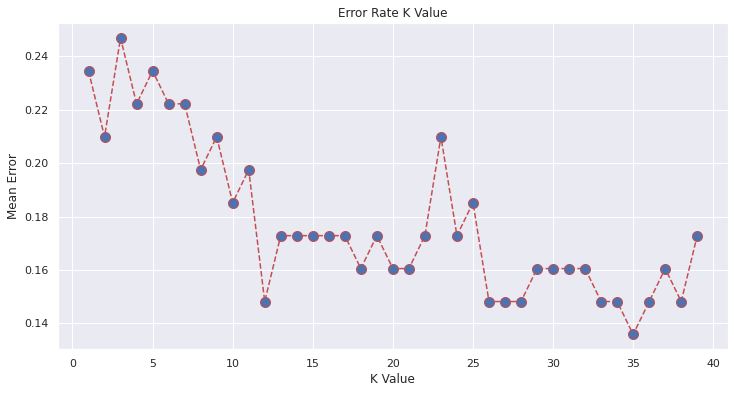
1. K- Nearest Neighbors

2nd step:

The number of neighbors that will get the least errors is k=35:

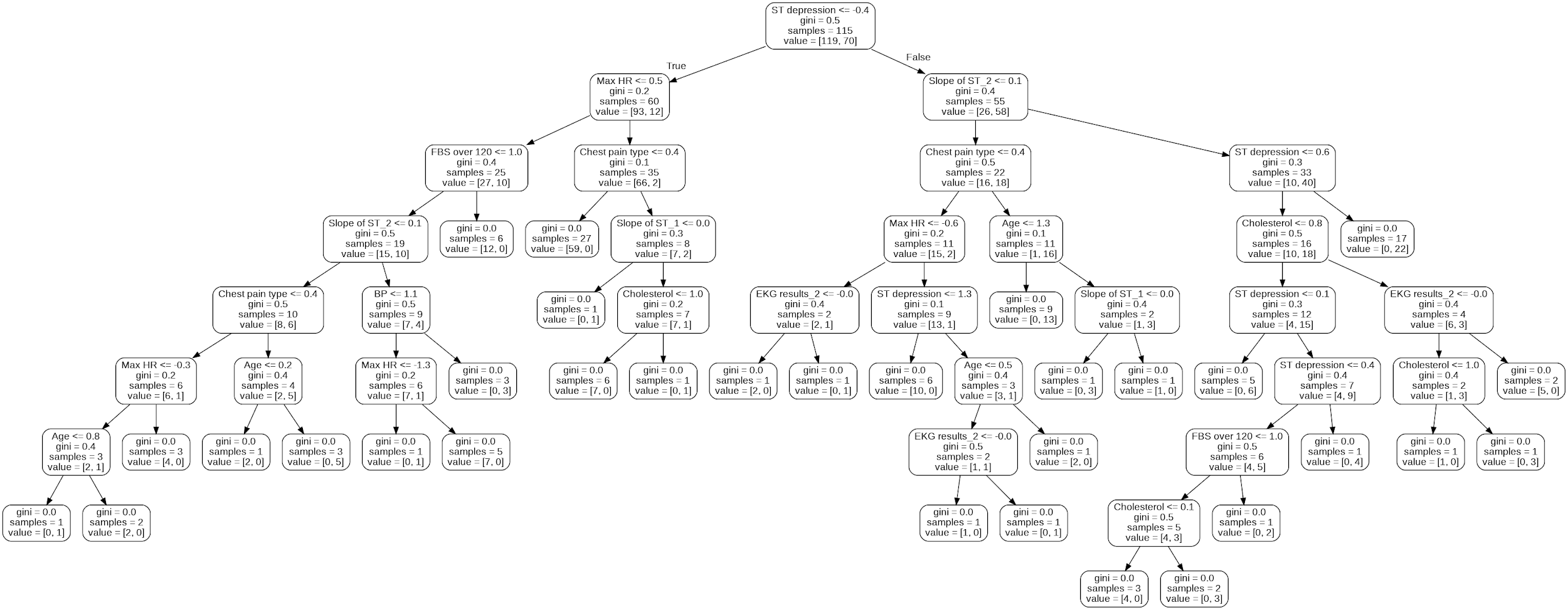
1st step:

The number of neighbors that will get the least errors is k=1:



1. SVM
2. Random Forest

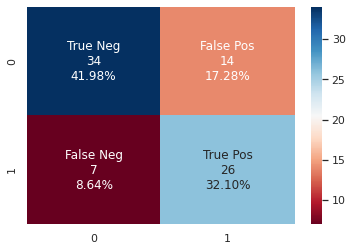
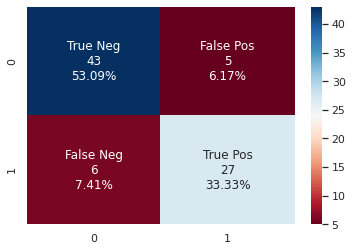
The following is a tree from the algorithm:



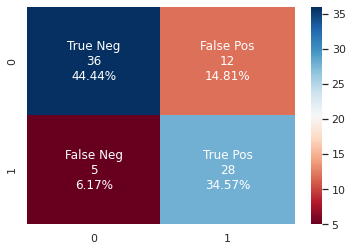
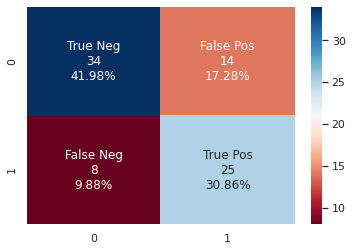
**Results:**

The following diagram shows the classification accuration between the algorithms we used.

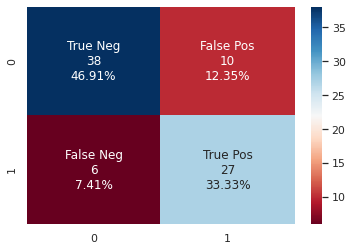
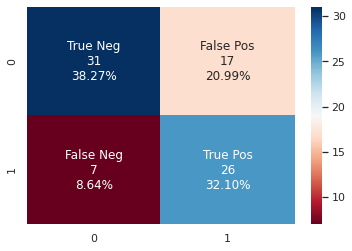
KNN confusion matrix:

1st step: 2nd step:

SVM confusion matrix:

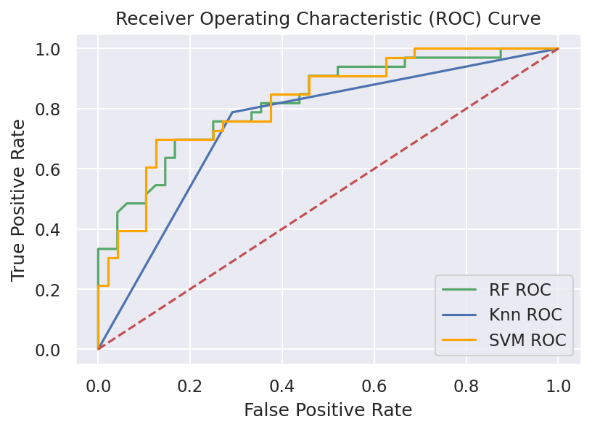
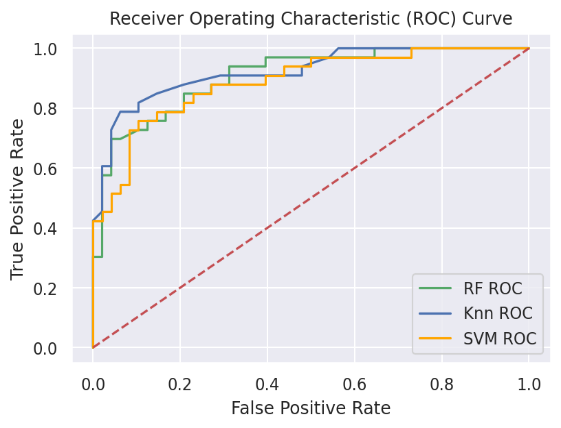
1st step: 2nd step:

Random Forest confusion matrix:

****1st step: 2nd step:

ROC graph:

1st step: 2nd step:

****

**Conclusions:**

We recommend a process that counts several steps to arrive at an accurate diagnosis and prevent patients from performing unnecessary medical examinations.

Since our classification problem deals with medical issues, our goal was to minimize the false negative index as possible.

In the first step, in each of the algorithms there are patients who were classified as healthy but in practice they are sick.

In the second step, the false negative index was reduced for each of the algorithms.

ROC diagram:

For the first step, we received that the AUC index of the SVM model is the highest.

For the second step, it can be seen that the area under the KNN curve is the largest .

In this situation, the KNN model gives better results compared to the three models we examined, both in terms of accuracy and both in terms of benefit and cost.

If we examine the economic-business aspect, the model will identify sick patients from the population with the highest level of accuracy, while conducting fewer tests in vain.

With that being said, this algorithm will get more information from the tests.

The preference is to perform tests “worthless” rather than to miss real sick patients.

According to the Recall index, for the first step, the recall index of the KNN model is the highest (0.788) and therefore we selected it.

For the second stage, the recall index of the SVM model is the highest (0.848) and therefore we chose it.

copy of final

correlation