Diagnosing Heart Disease

Using Supervised ML Classification Techniques

Soheila Rahmani

Overview

According to the Centers for Disease Control and Prevention (CDC), Heart disease is the leading cause of death for men, women, and people of most racial and ethnic groups in the United States. One person dies every 36 seconds in the United States from cardiovascular disease. About 655,000 Americans die from heart disease each year—that's 1 in every 4 deaths. Heart disease costs the United States about $219 billion each year from 2014 to 2015. This includes the cost of health care services, medicines, and lost productivity due to death.

Goal : Classifying whether a person is suffering from heart disease or not, applying 3 common Machine Learning techniques, Logistic Regression, Random Forest Classifier, K-Nearest Neighbor Classifier.

Dataset

age - age in years
sex - (1 = male; 0 = female)
cp - chest pain type
value1: typical angina
value2: atypical angina
value3: non-anginal pain
value4: asymptomatic
trestbps - resting blood pressure (in mm Hg on admission to the hospital)
chol - serum cholesterol in mg/dl
fbs - (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
restecg - resting electrocardiographic results
value0: normal
value1: having ST-T wave abnormality(T wave inversions and/or ST elevation and depression of >005 mV)
value2: showing probable or definite left ventricular hypertrophy by Estes criteria.

In this machine learning project, the dataset have been collected from Kaggle (https://www.kaggle.com/sid321axn/heart-statlog-cleveland-hungary-final) with consisting of 1190 records of patients from US, UK, Switzerland and Hungary with 11 features and 1 target variable.

The classification goal is to predict whether the patient has a risk of future coronary heart disease (CHD) or not.

thalach - maximum heart rate achieved in beats per minutes(bpm)
exang - exercise induced angina (1 = yes; 0 = no)
oldpeak - ST depression induced by exercise relative to rest
slope - the slope of the peak exercise ST segment
diagnosis - have disease or not (1 = yes, 0 = no)

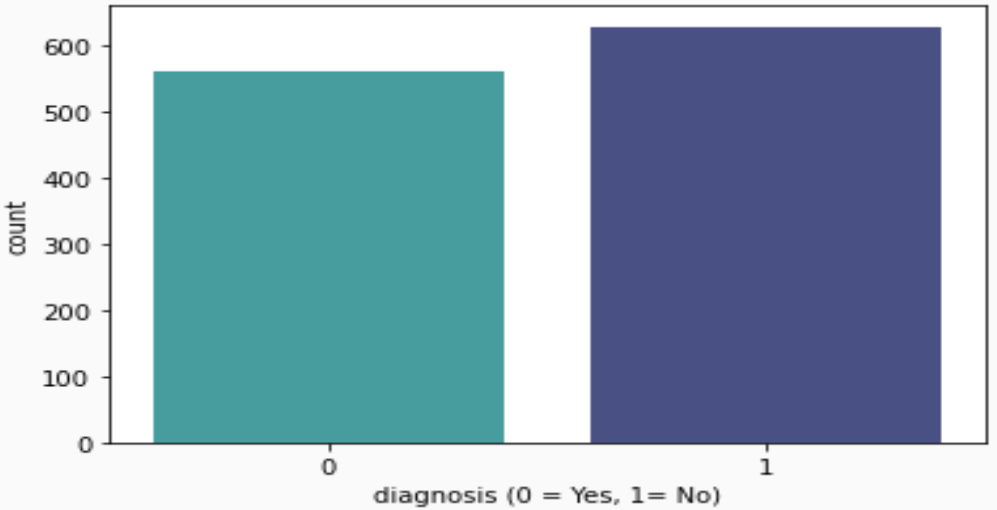
Data Visualization

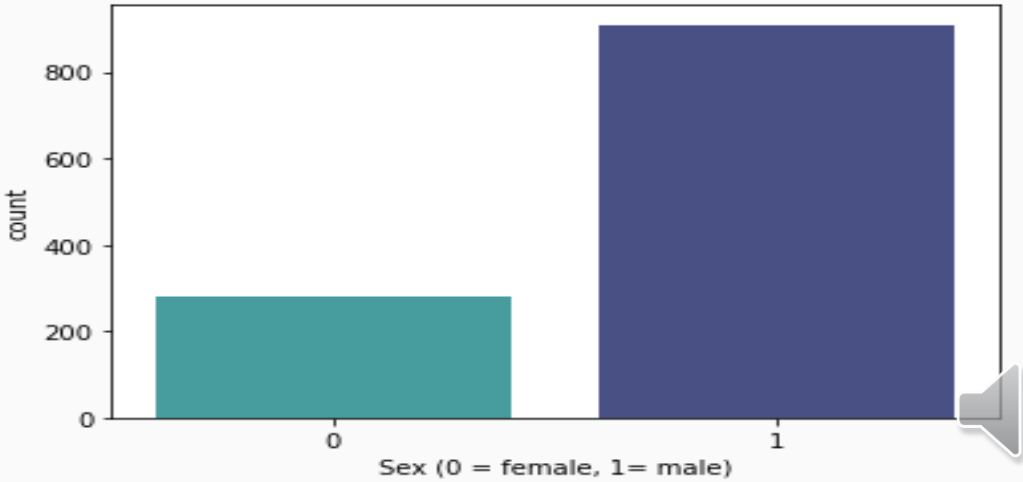
Percentage of Patients Haven't Heart Disease: 47.54%

Percentage of Patients Have Heart Disease: 52.46%

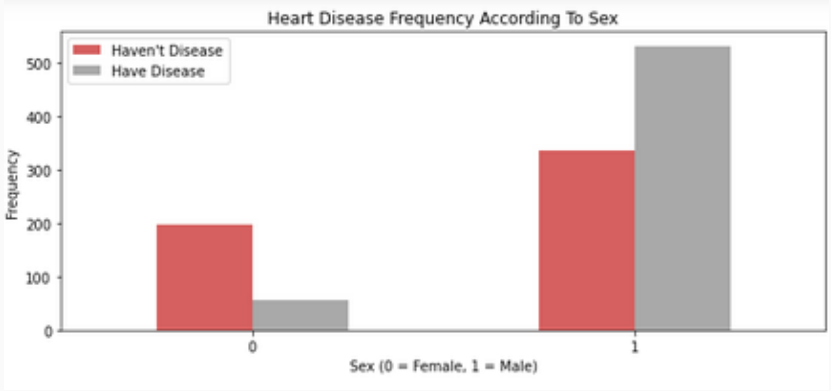
Percentage of Female Patients: 22.52%

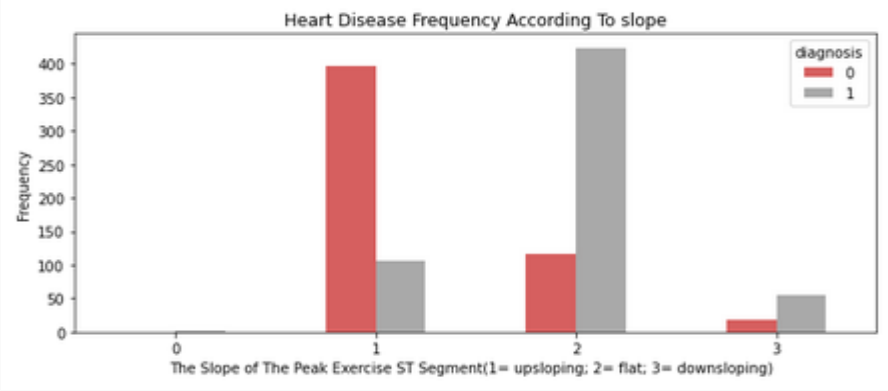
Percentage of Male Patients: 77.48%

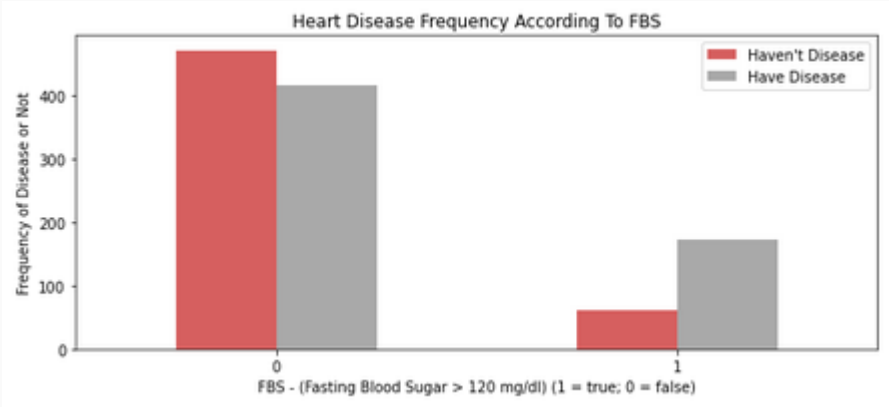


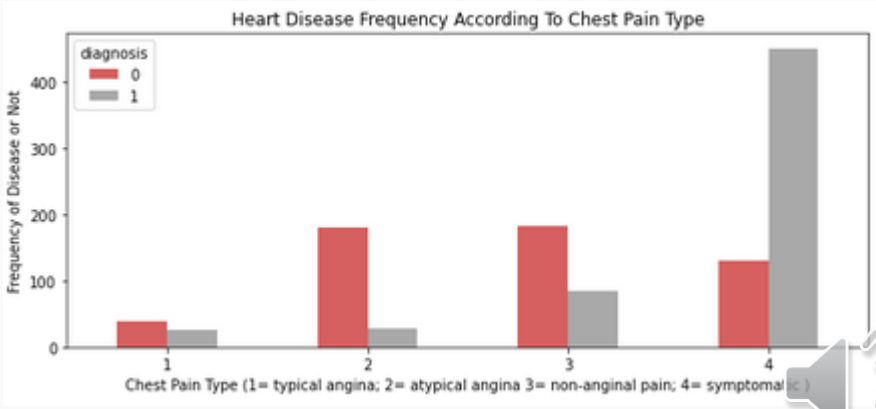


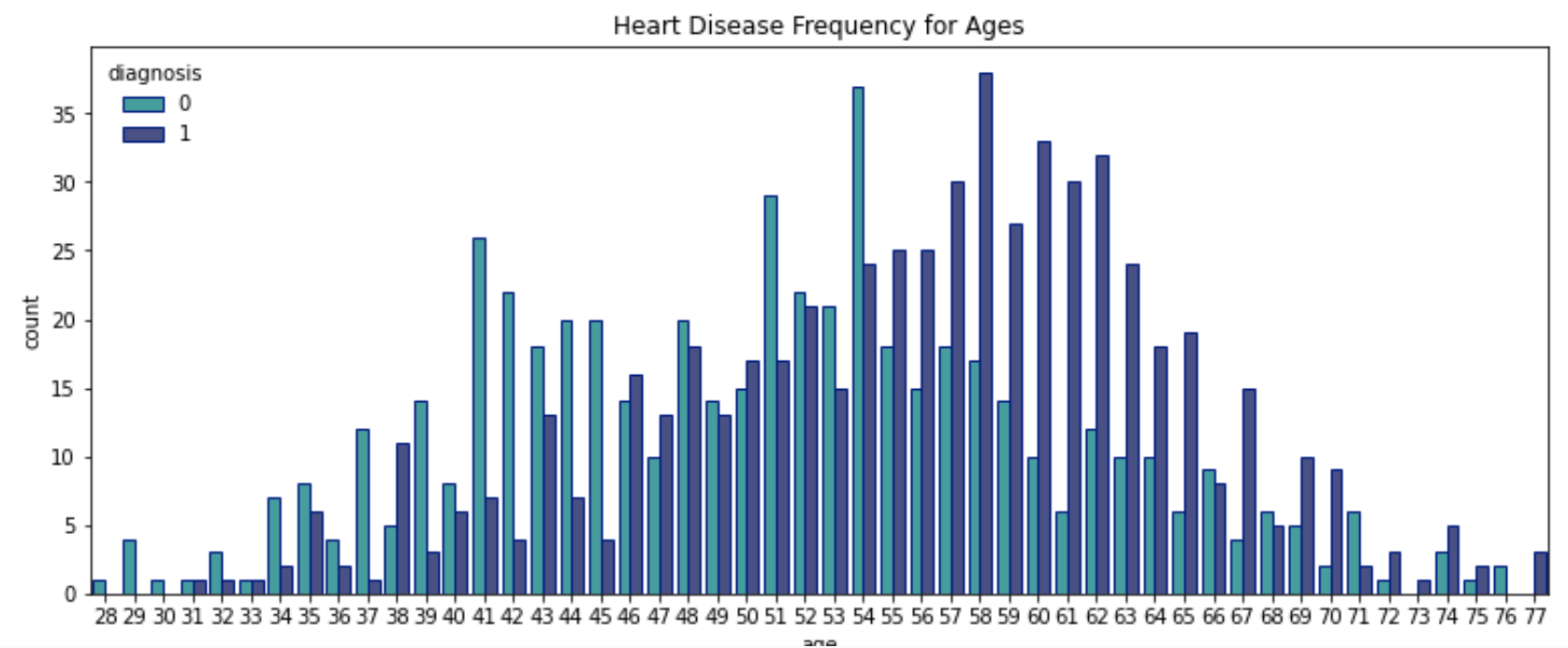
Data Visualization







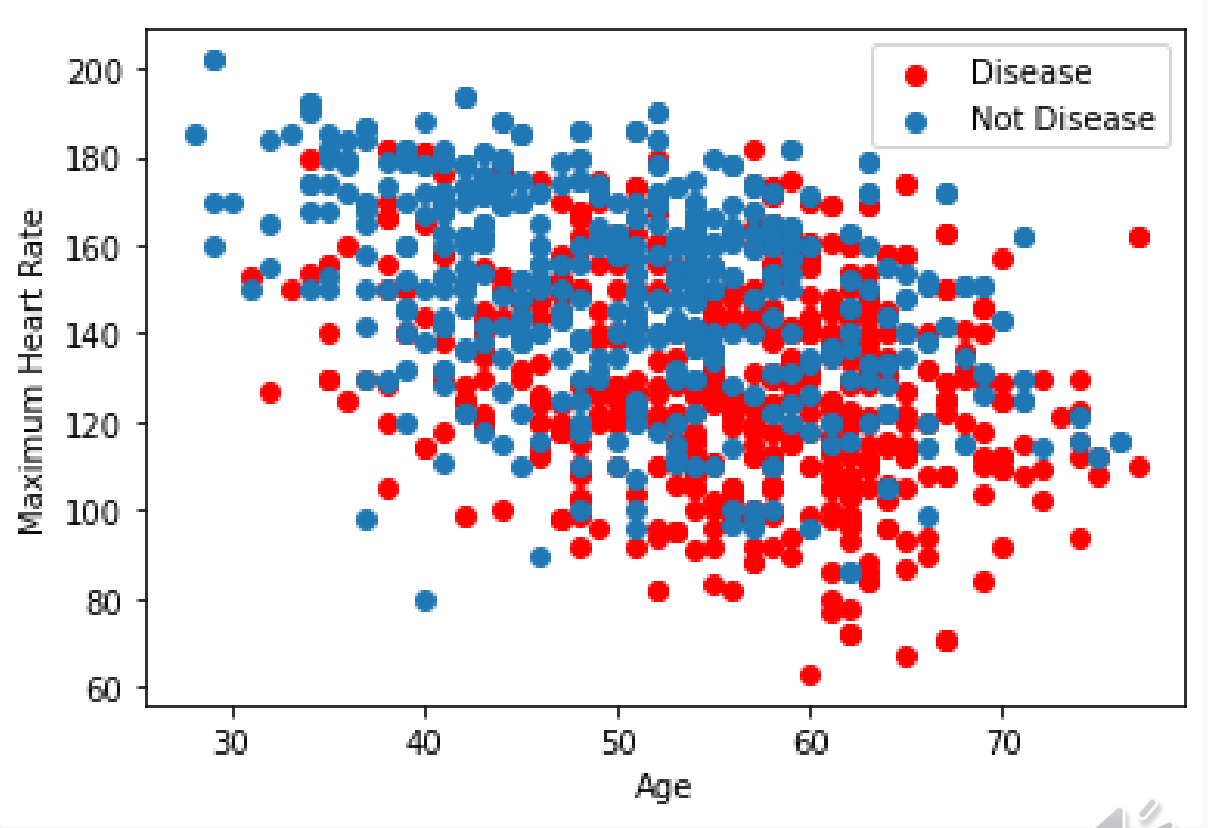




After age 58, the chance of heart disease decreases for both male and female. Age seems to have a positive correlation to the chance of heart disease.

Age and Max. Heart Rate Correlation with heart disease

Patients with the maximum heart rate between 80-160 and ages between 50-65 are mostly diagnosed with heart disease.

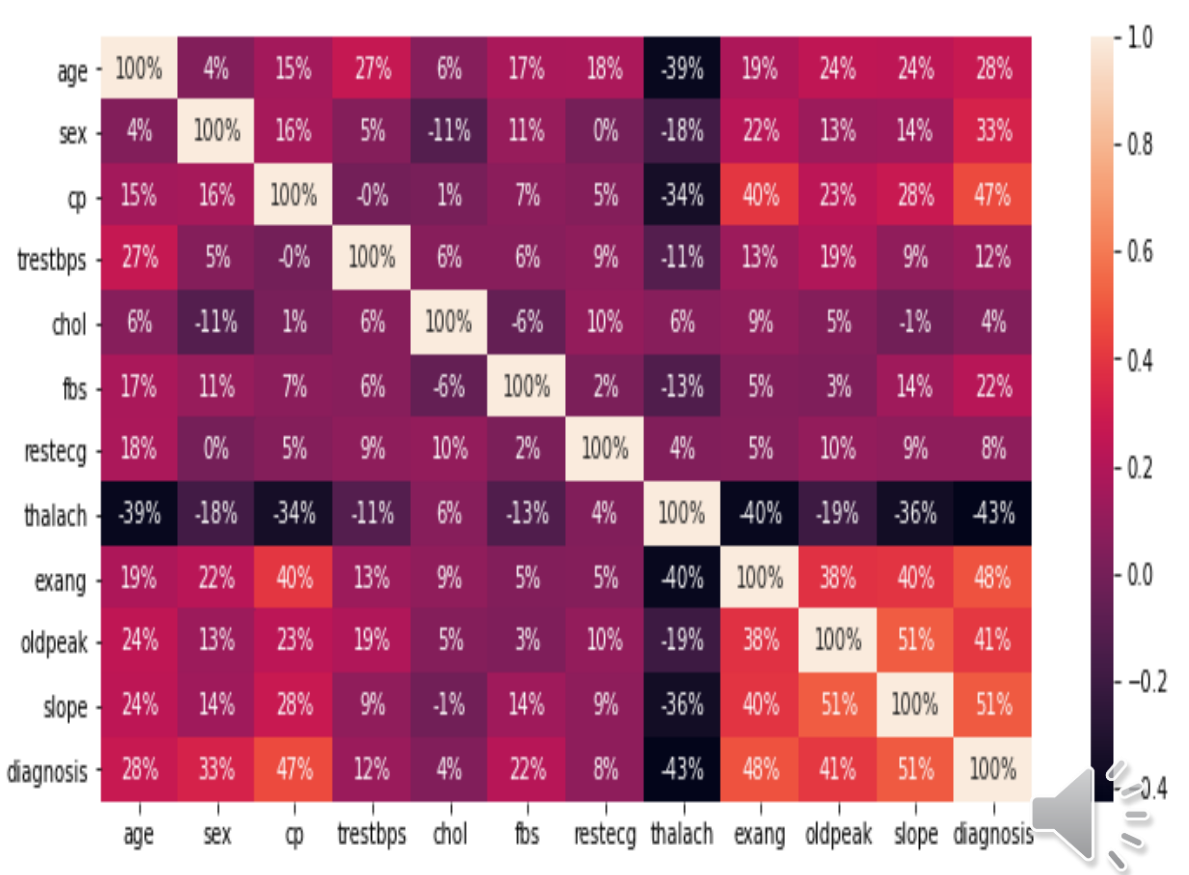


Seaborn Correlation Heatmap

Slope - Highest correlation
51%

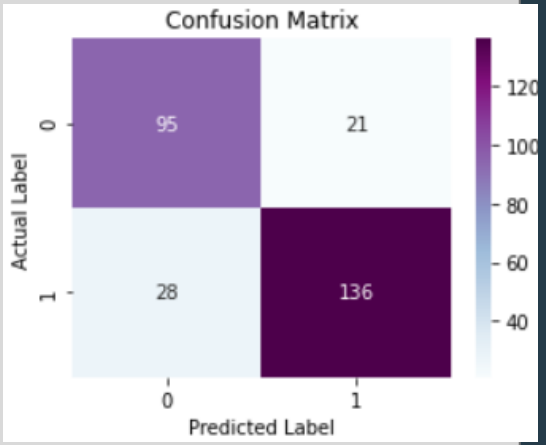
Exang - Second highest correlation
48%

Thalach - Lowest correlation
-43%



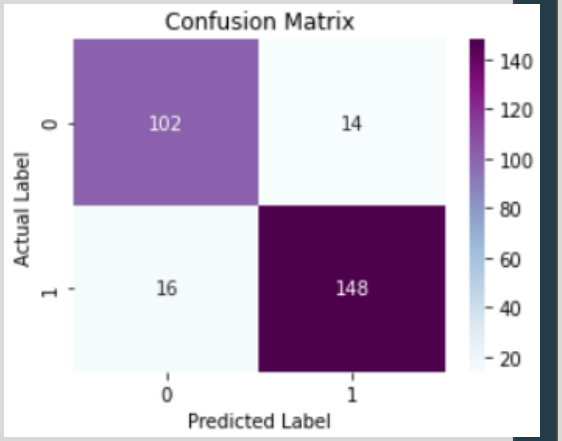
Result

To validate our data with the three different models, we used the features *age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, and slope with* 75% Training data, 25% Testing



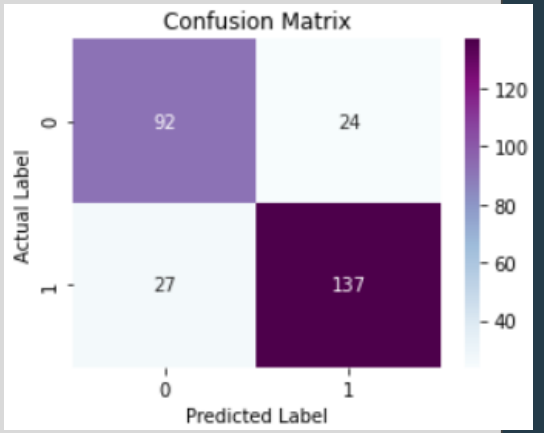
Logistic Regression Accuracy score

0.825



Random Forest Classifier Accuracy score

0.889



K-NN Classifier Accuracy score

0.817

Conclusion

Based on the results of this study, it was found that the highest risk of heart diseases falls under the age range 58 and younger and more likely in men than women.
By comparing the performances of the models we are able to see that Random Forest Classifier, with 89%, has the highest accuracy than the other two.

For the next study, I would like to research more about other risk factors that may cause heart disease, to contribute as adding factors to the dataset. I am also looking into finding other Machine Learning techniques with more accuracy which may help to improve the prediction results.