Heart Disease Project Report.

Introduction and Objective

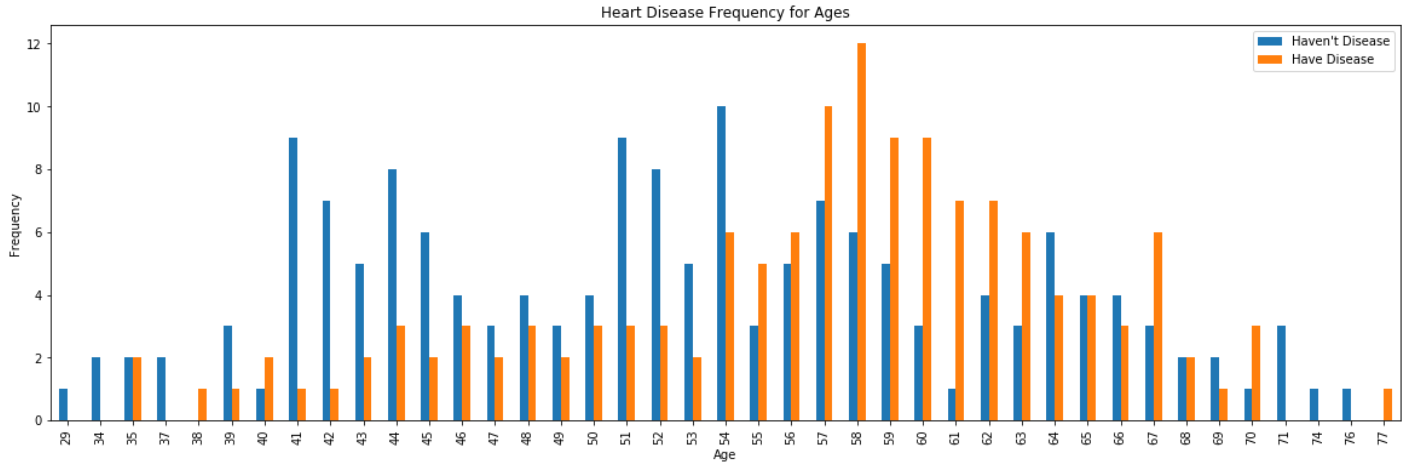
The dataset, Cleveland Heart Disease Dataset from the UCI, was compiled by Robert Detrano, M.D., Ph.D using data from the V.A. Medical Centre, Long Beach and Cleveland Clinic Foundation. The data-set can be found on the UCI Machine Learning Repository

The objective of this project is to predict the presence of heart disease in individuals by using different machine learning algorithms, given a variety of features related to the disease. The dataset contains 13 features/explanatory variables that are believed to be correlated with heart disease in some capacity. Individuals who are believed to be at risk of heart disease (through underlying conditions) can use this tool to spot how likely they are to suffer from heart disease. Individuals who are disease averse can use this tool to identify what steps they can take to reduce their likelihood of getting the disease.

The project will consist of the following tasks:

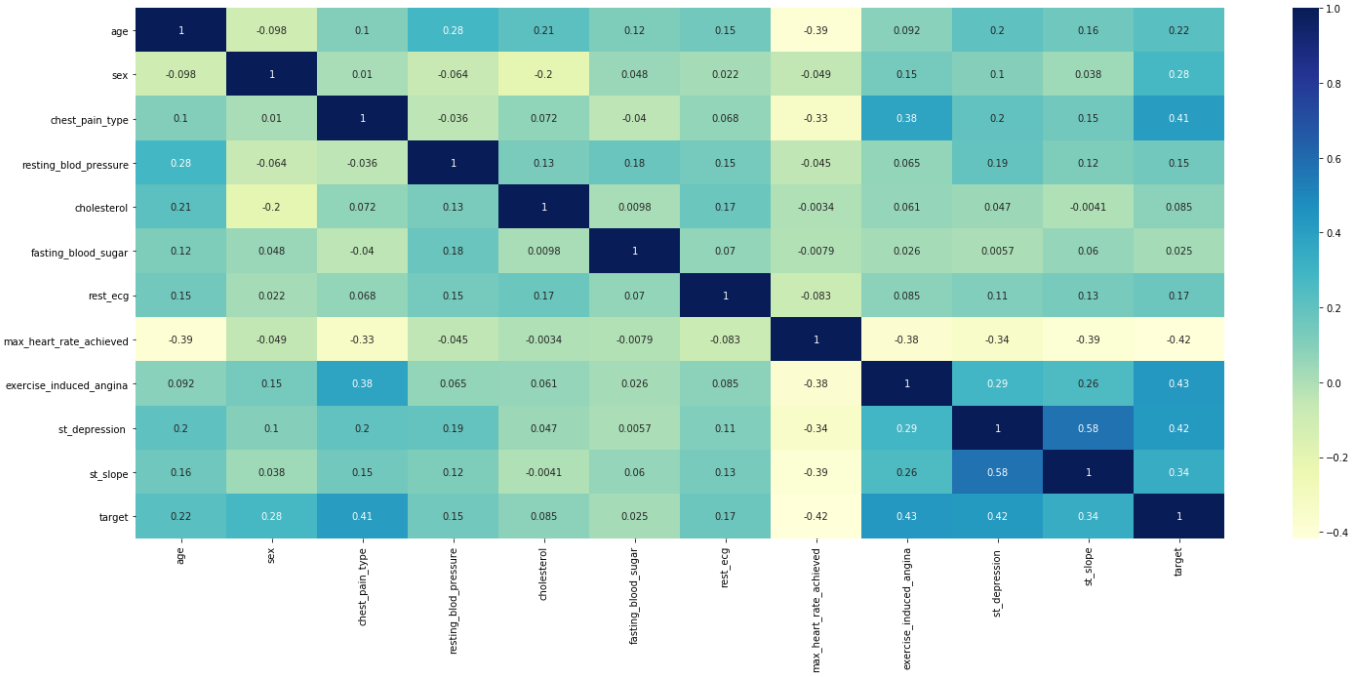
- Datas Analysis - To conceptualise the data, deal with missing values and remove extraneous columns to prep the data for machine learning.
- Descriptive and Inferential Statistics - To explore the data and make generalisations about which features are most closely linked to heart disease
- Machine Learning models - Logistic regression, K-nearest neighbours, classifier, support vector machine, decision tree classifier, random forest classifier and XGBoost classifier will be used to determine training accuracy.
- Model Tuning - Hyperparameter tuning will be used to yield a lower error.

Data Analysis



The plot above gives an overview of the heart disease dataset. The plot illustrates the posited trend that heart disease presence/likelihood increases with age. The peak at the middle is closer to the middle due to the distribution of participants surveyed (there were less participants aged 65 and above than between 55-65).

2 of the 13 columns contained missing values. Averages could not be taken for these columns because they were categorical, hence the rows with missing values were deleted.



The correlation matrix above illustrates the correlation between each feature/ explanatory variable related to our target variable, presence of heart disease. The values show that none of the features are significantly correlated with each other.

Hypothesis Test 2

• H0: $\rho = 0$

• H1: $\rho \neq 0$

In words:

• The null hypothesis states that there is no significant relationship between cholesterol and heart disease.

• The alternative hypothesis states that there is a significant relationsip between cholesterol and heart disease.

In [33]:
```
corr_coeff, p_val = spearmanr(heart_disease.chol, heart_disease.target)
print('Spearman\'s Correlation Coefficient: ' + str(corr_coeff))
print('p-value: ' + str(p_val))
```

```
Spearman's Correlation Coefficient: 0.11565400290797957
p-value: 0.046433924233820936
```

p-value is less than 0.05 so we reject the null hypothesis and accept the alternate

Hypothesis Test 3

• H0: $\rho = 0$

• H1: $\rho \neq 0$

In words:

• The null hypothesis states that there is no significant relationship between resting blood pressure and heart disease.

• The alternative hypothesis states that there is a significant relationsip between resting blood pressure and heart disease.

In [36]:
```
corr_coeff, p_val = spearmanr(heart_disease.trestbps, heart_disease.target)
print('Spearman\'s Correlation Coefficient: ' + str(corr_coeff))
print('p-value: ' + str(p_val))
```

```
Spearman's Correlation Coefficient: 0.13174028158972423
p-value: 0.023165003218545328
```

p-value is less than 0.05 so we reject null and accept alternate hypothesis.

The Spearman's coefficient (as shown above), however, shows that both cholesterol and resting blood pressure are significantly correlated with the presence of heart disease in individuals.

Machine Learning

Before Tuning

	Model	Training Accuracy %	Testing Accuracy %
0	Logistic Regression	85.990338	87.777778
1	K-nearest neighbors	83.574879	88.888889
2	Support Vector Machine	92.753623	90.000000
3	Decision Tree Classifier	100.000000	70.000000
4	Random Forest Classifier	100.000000	87.777778
5	XGBoost Classifier	100.000000	81.111111

After Tuning

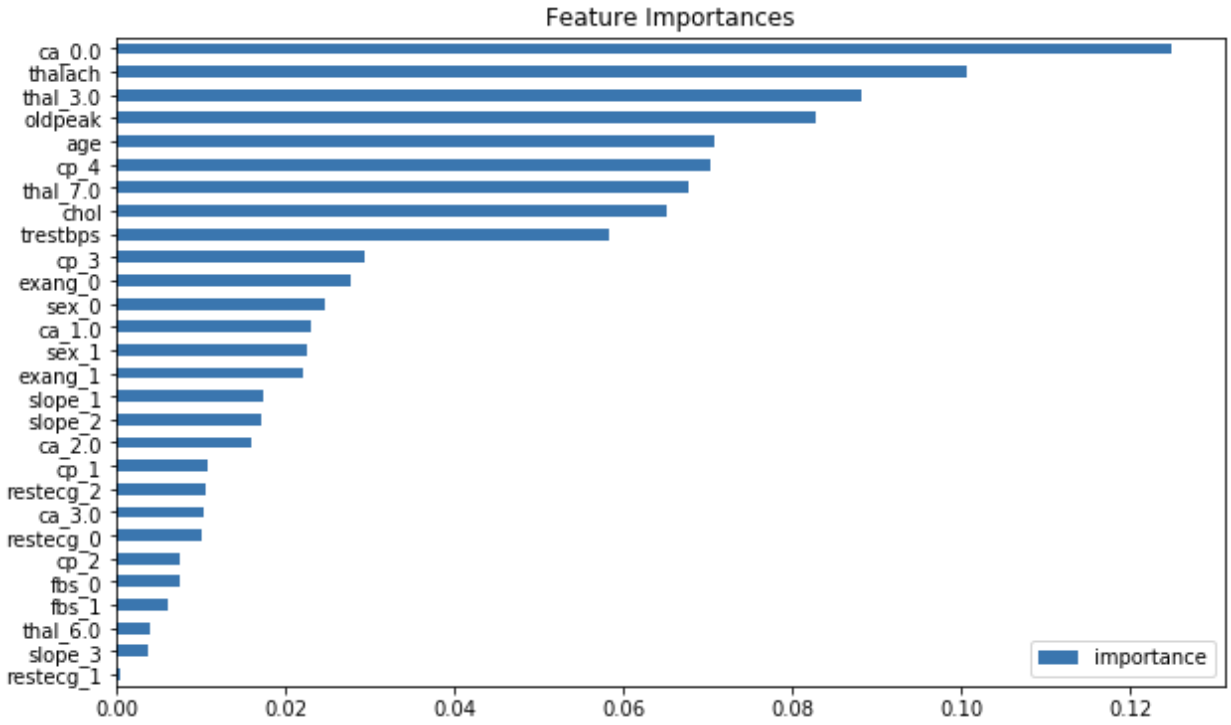
	Model	Training Accuracy %	Testing Accuracy %
0	Tuned Logistic Regression	85.990338	88.888889
1	Tuned K-nearest neighbors	82.125604	84.444444
2	Tuned Support Vector Machine	85.507246	90.000000
3	Tuned Decision Tree Classifier	85.990338	75.555556
4	Tuned Random Forest Classifier	98.067633	87.777778
5	Tuned XGBoost Classifier	85.507246	85.555556

This project's objective is to predict the presence of heart disease in individuals using machine learning classification techniques, 6 different techniques were used and evaluated against each other.

1. Logistic Regression
2. K-Nearest Neighbours Classifier
3. Support Vector machine
4. Decision Tree Classifier
5. Random Forest Classifier
6. XGBoost Classifier

The figures above show that the SVM and random forest models gave the best results.

Feature Importance for Random Forest and Conclusion



The figure above ranks the features of the random forest model in order of their importance. According to the model, 0 major blood vessels is the most important feature for heart disease classification.