

CPM252 Foundations and Programming for Data Analytics Semester 1, 2020/2021

ASSIGNMENT (20%)

(Individual Work)

**NG WEI YI**

**(147630)**

**CPM 252**

**Abstract**

According to a report published by the World Health Organization (WHO), cardiovascular diseases (CVDs), also known by the term as "heart diseases" are the number one cause of death globally. Cardiovascular diseases (CVDs) such as [hypertensive heart disease](https://en.wikipedia.org/wiki/Hypertensive_heart_disease) and cerebrovascular disease have taken an estimated of 17.9 million lives every year.

Therefore, the experiment in this project is carried out to predict the odds of a patient getting heart disease using a suitable machine learning algorithm and also to reduce the number of heart disease patients in the near future.

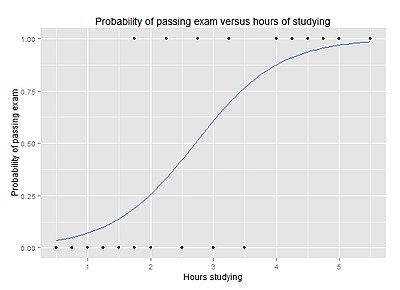
**Introduction**

The term "heart disease" is often used interchangeably with the term "cardiovascular disease", it describes a range of conditions that affect a person’s heart. Cardiovascular disease (CVD) is a class of diseases that involve the [heart](https://en.wikipedia.org/wiki/Heart) or [blood vessels](https://en.wikipedia.org/wiki/Blood_vessel). The diseases under the heart disease umbrella include [coronary artery diseases](https://en.wikipedia.org/wiki/Coronary_artery_disease) (CAD) such as [angina](https://en.wikipedia.org/wiki/Angina_pectoris) and [myocardial infarction](https://en.wikipedia.org/wiki/Myocardial_infarction) (commonly known as a heart attack). Other CVDs include [stroke](https://en.wikipedia.org/wiki/Stroke), [heart failure](https://en.wikipedia.org/wiki/Heart_failure) and [hypertensive heart disease](https://en.wikipedia.org/wiki/Hypertensive_heart_disease).

Heart disease is also one of the most common diseases among everyone and it has affected not just the elderly, but youth as well, both male and female. Hence, prediction of cardiovascular disease is regarded as one of the most important subjects in the section of clinical data analysis. Machine learning (ML) proves to be effective in assisting in making decisions and predictions from the large quantity of data produced by the healthcare industry.

Hence, two machine learning algorithms are chosen to perform an experiment to predict the odds of getting heart disease based on the data set provided. The two machine learning algorithms are Logistic Regression and Support Vector Machine (SVM).

Logistic regression is a statistical model and it is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Logistic regression is a predictive analysis and it is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. Logistic Regression is highly prone to outliers since it tries to maximize the conditional likelihood of the training data.



A graph of a logistic regression curve showing the probability of passing exam versus hours of studying

Support Vector Machine (SVM) is a [supervised learning](https://en.wikipedia.org/wiki/Supervised_learning) model with associated learning [algorithms](https://en.wikipedia.org/wiki/Algorithm) that analyse data used for [classification](https://en.wikipedia.org/wiki/Statistical_classification) and [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis). In the SVM algorithm, each data item is plotted as a point in n-dimensional space (where n is number of features) with the value of each feature being the value of a particular coordinate. Then, classification is performed by finding the hyper-plane that differentiates the two classes. Support Vectors are samples on the margin.

A hyperplane (Indicated by the dotted line in the diagram below) in a [high-](https://en.wikipedia.org/wiki/High-dimensional_space) or infinite-dimensional space can be used for [classification](https://en.wikipedia.org/wiki/Statistical_classification), [regression](https://en.wikipedia.org/wiki/Regression_analysis) or other tasks like detecting outliers. The best hyperplane is the one that has the largest separation or [margin](https://en.wikipedia.org/wiki/Margin_(machine_learning)) between the two classes. If the distance from the hyperplane to the nearest data point on each side is maximized, it is known as the [maximum-margin hyperplane](https://en.wikipedia.org/wiki/Maximum-margin_hyperplane).



**Objective**

The objective of this experiment is to choose the best machine learning algorithm to build a champion model to predict whether a patient has heart disease or not based on the existing data such as chest pain type, cholesterol level, patient’s age and gender.

**Justification of The Machine Learning Algorithms Selected**

Based on the data provided, patients can be categorized into two classes, with or without heart disease. Hence, two supervised machine learning algorithms are chosen to classify the patients since there are known inputs and outputs in the dataset. The two supervised machine learning algorithms are Logistic Regression and Support Vector Machine (SVM).

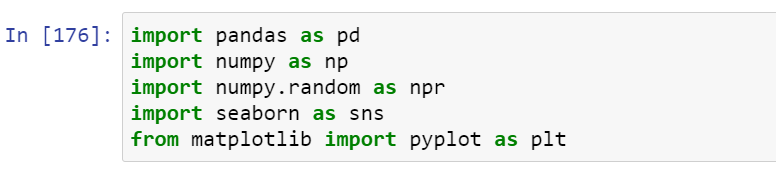
Both machine learning algorithms are also called supervised classification machine learning algorithms, which means both algorithms can be used to separate the data into two known classes or categories only since the final outcomes of the predictions will be either having heart disease or do not have heart disease (dichotomous).

Hence, unsupervised machine learning algorithms such as Hierarchical Clustering and K-Means Clustering cannot be used because unsupervised machine learning algorithms are suitable to be used where there are only known inputs in the dataset, the outputs are unknown. Furthermore, supervised regression machine learning algorithms such as Linear Regression and Regression Trees cannot be used because the final outcomes are categorical instead of continuous.

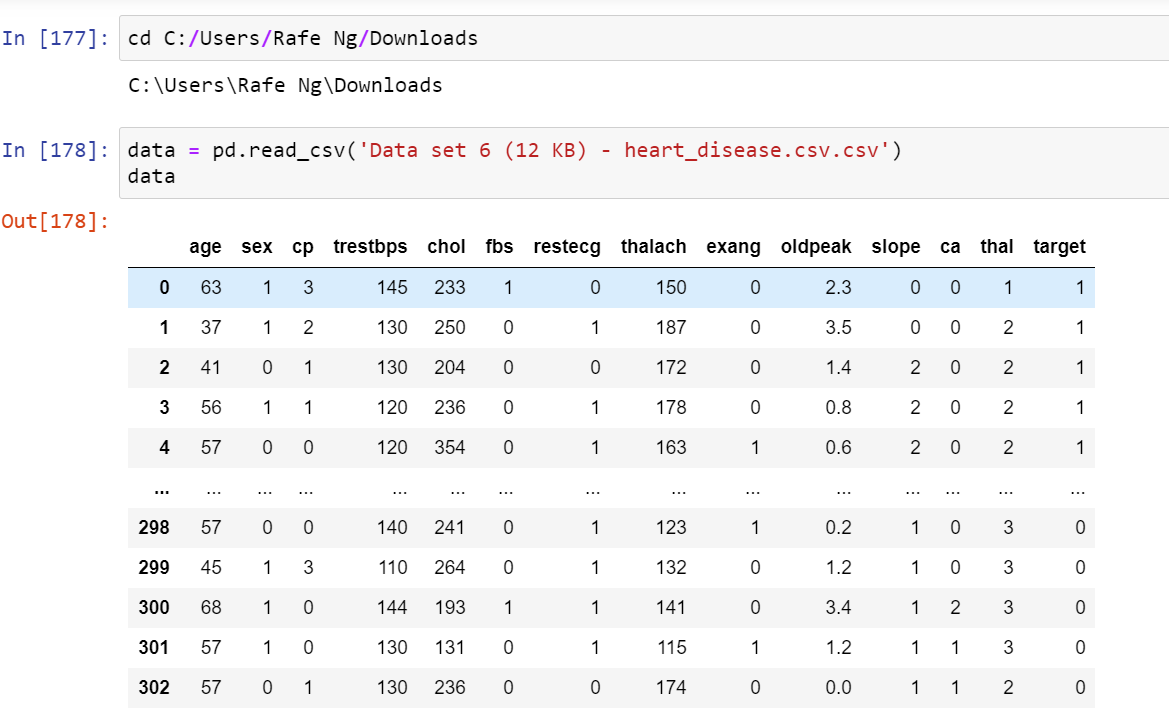
**Steps To Build The Machine Learning Models**

***Step 1: Import and Reading The Dataset***

Modules such as pandas, numpy, numpy.random, seaborn and matplotlib are imported into Juypter Notebook.

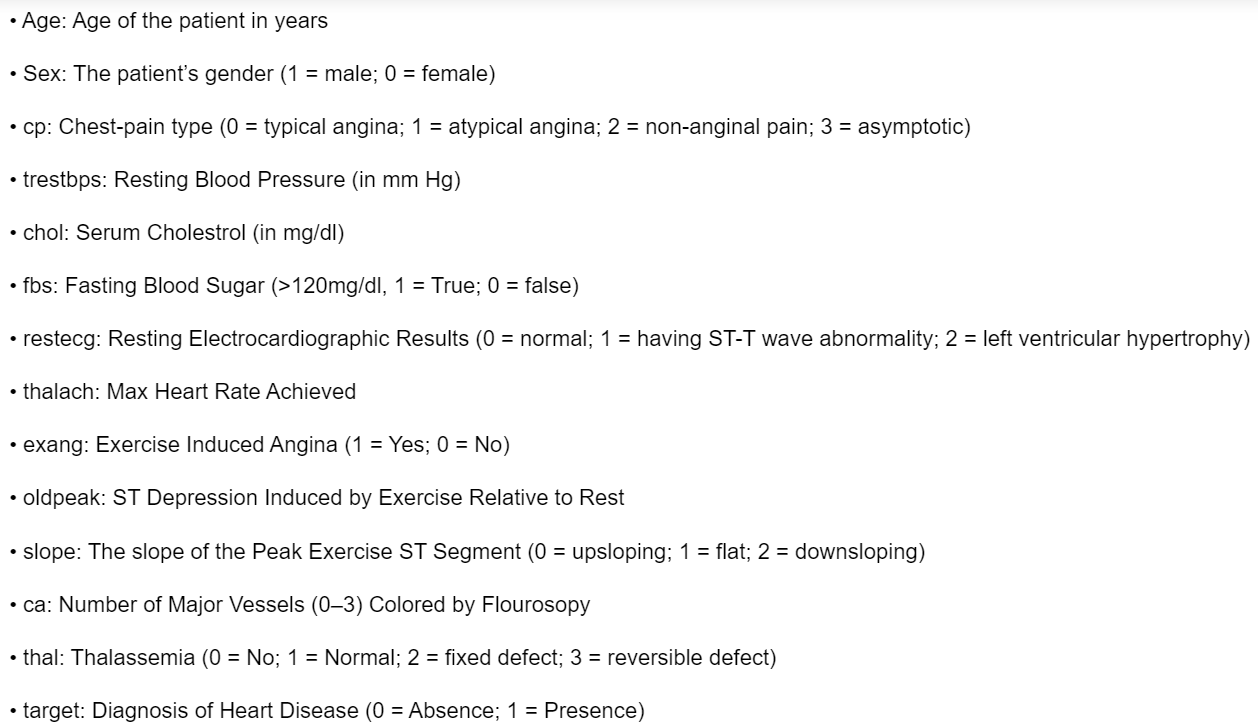


The file directory (absolute path) is stated in order to access the dataset from a specific file (in this case, the file is Downloads).



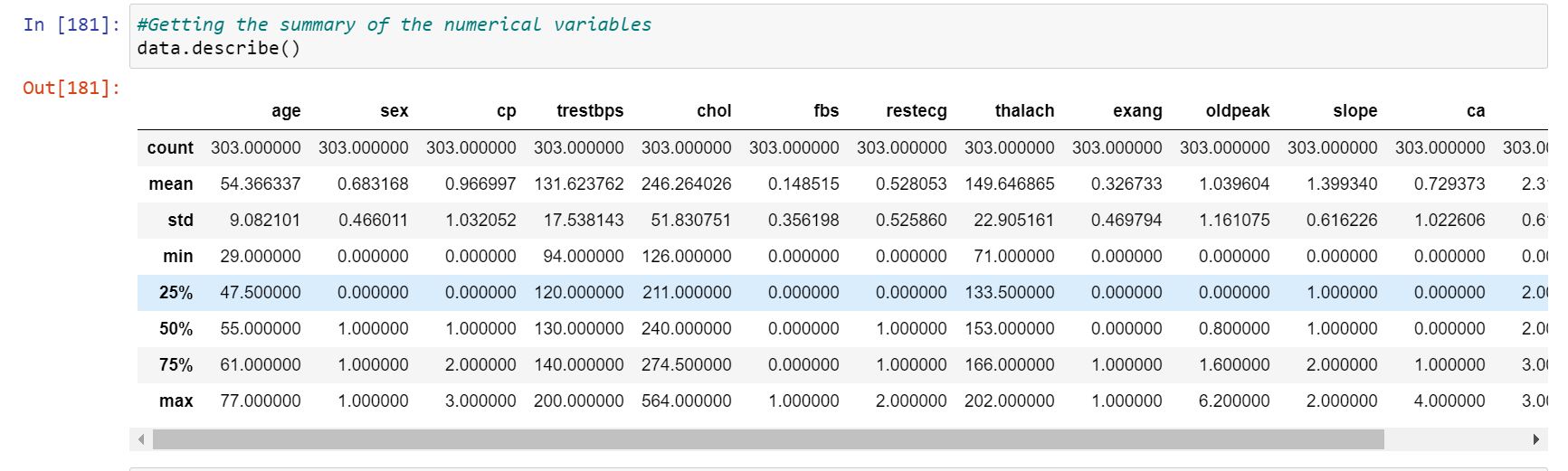
***Step 2: Data Description***

The meaning of some of the column headers are not obvious, hence, this step is taken to state the meaning of the unobvious columns header. The meanings are stated as below:

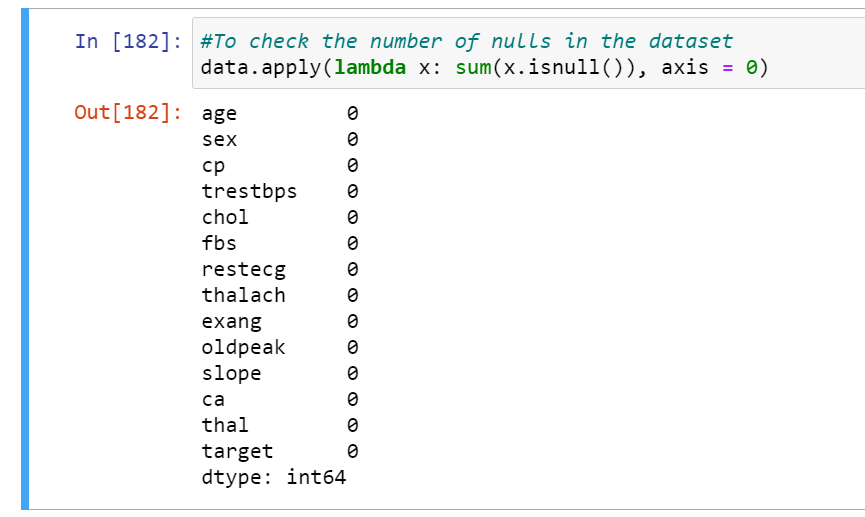


Sources: <https://www.kaggle.com/ronitf/heart-disease-uci> & <https://archive.ics.uci.edu/ml/datasets/heart+disease>

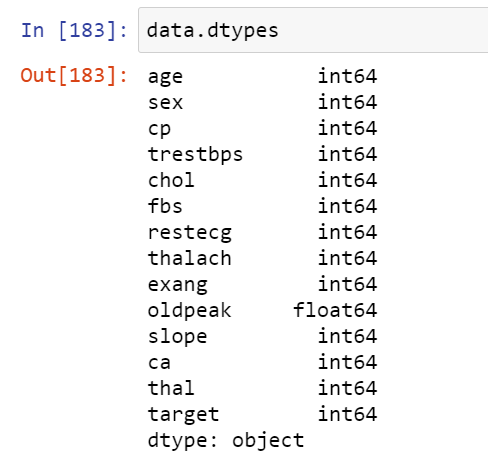
Then, the summary of all numerical values are stated:



The number of nulls in the dataset is checked:



Data type object (describes how the bytes in the fixed-size block of memory corresponding to an array item should be interpreted) is stated (in this case, it is int64 and float64):



***Step 3: Data Analysis***

*Feature Selection*

Feature Selection is the process where those features which contribute most to the prediction variable or output in which we are interested in will be automatically or manually selected.

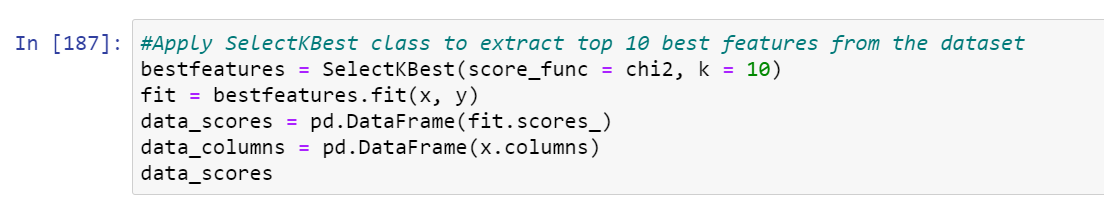
Accuracy of the models will decrease if they have irrelevant features in the data and this makes the models learn based on irrelevant features.

*1. Univariate Selection*

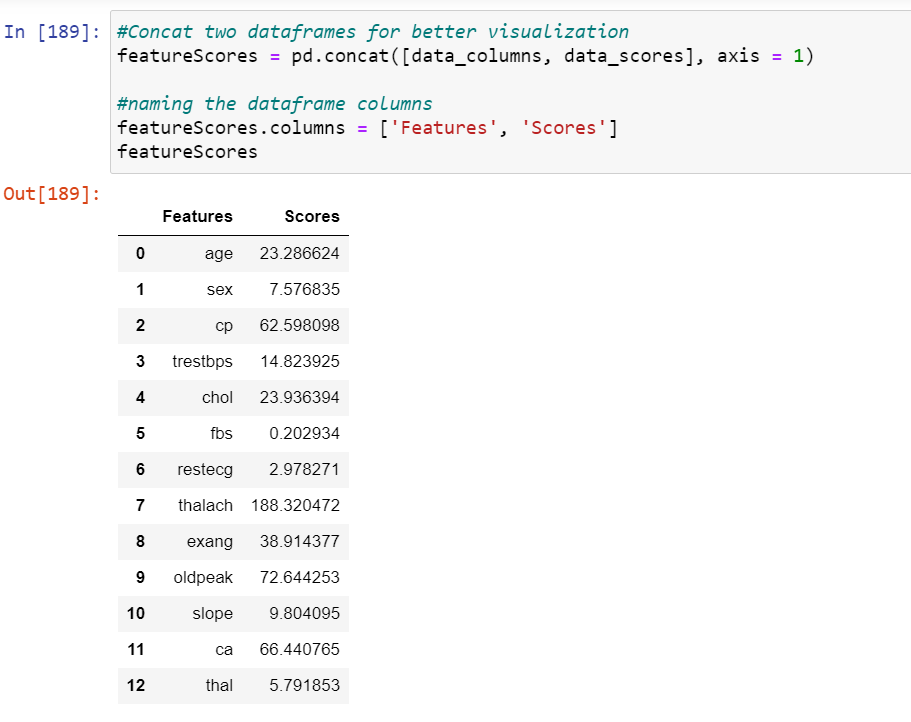
Univariate selection works by selecting the best features that have the strongest relationship with the output variable based on univariate statistical tests. It can be seen as a pre-processing step to an estimator.



Extract the top 10 features from the dataset:



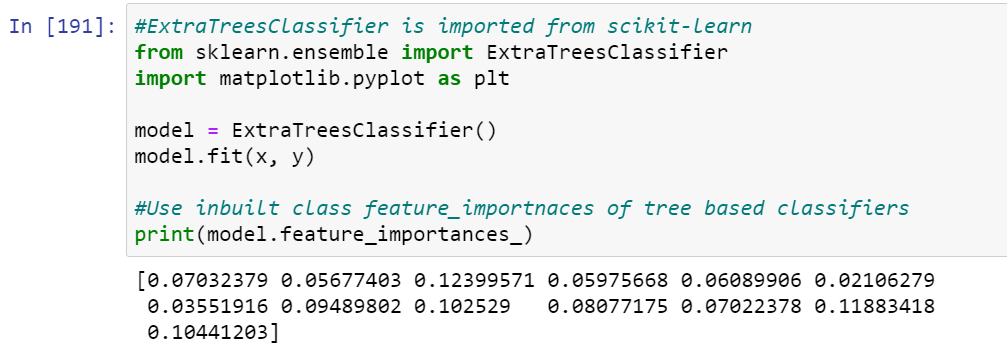
The top 10 features from the dataset and their scores:



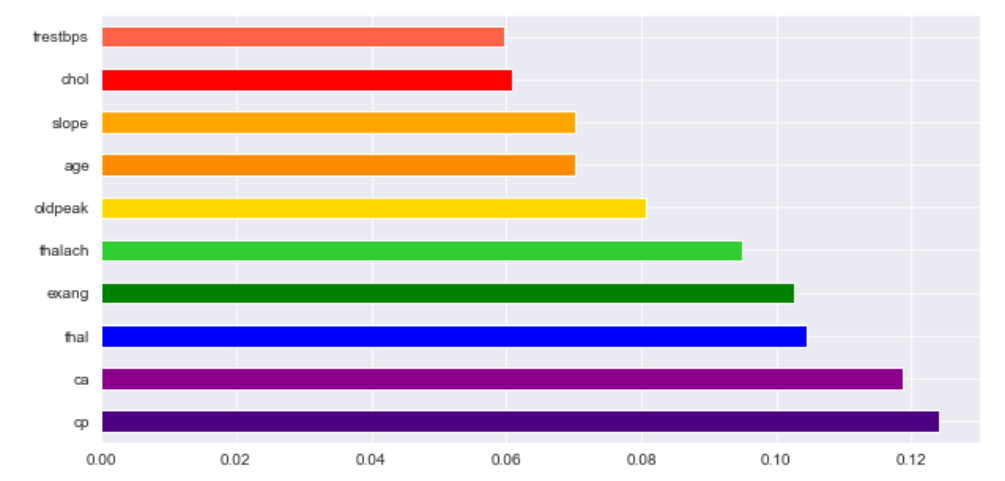
*2. Feature Importance*

Feature importance gives a score for each feature of the data. Features that are more important or relevant towards the output variable, will have higher scores compared to those features which are less important or relevant.

Feature importance is an inbuilt class that comes with Tree Based Classifiers, Extra Tree Classifier will be used for extracting the top 10 features for the dataset.



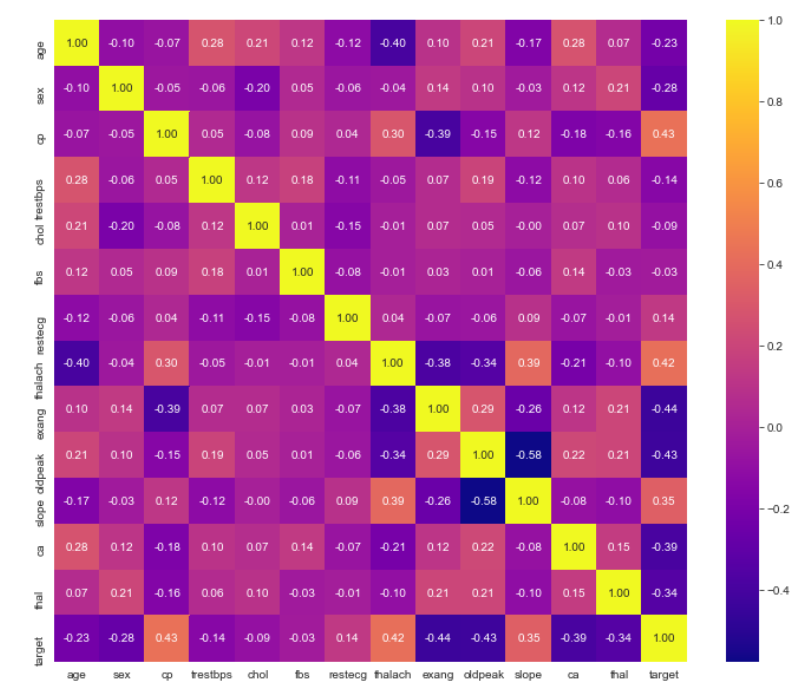
A graph of the top 10 features against their scores is plotted:



*3. Correlation Matrix with Heatmap*

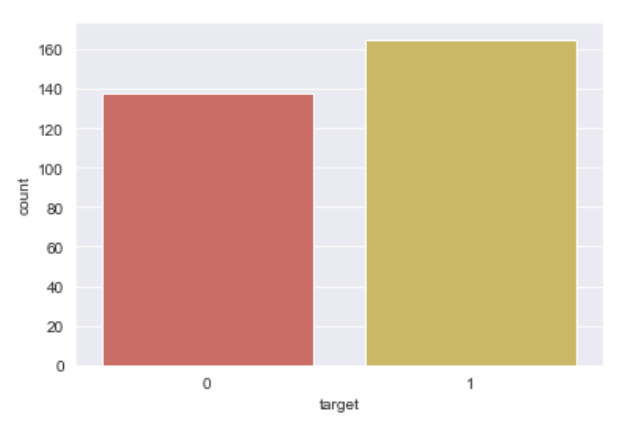
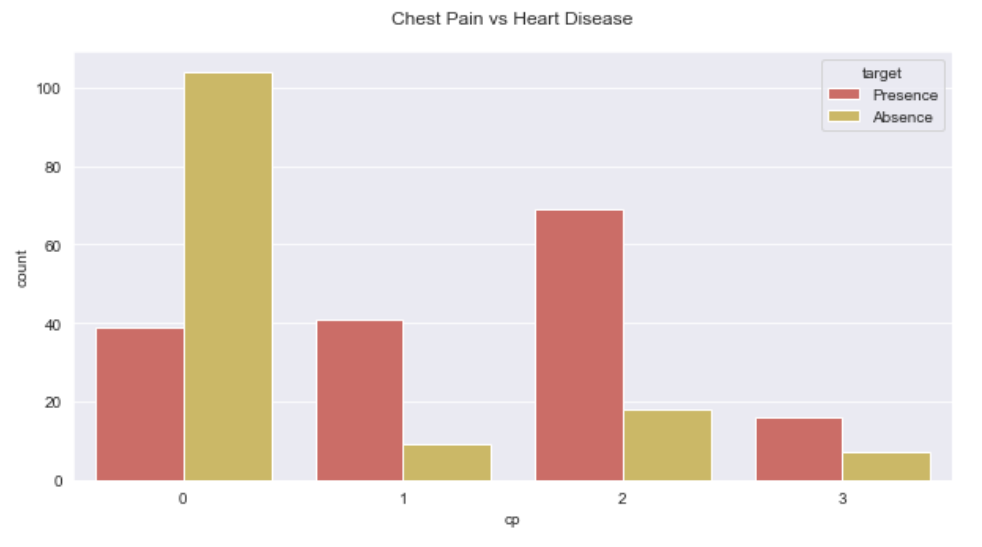
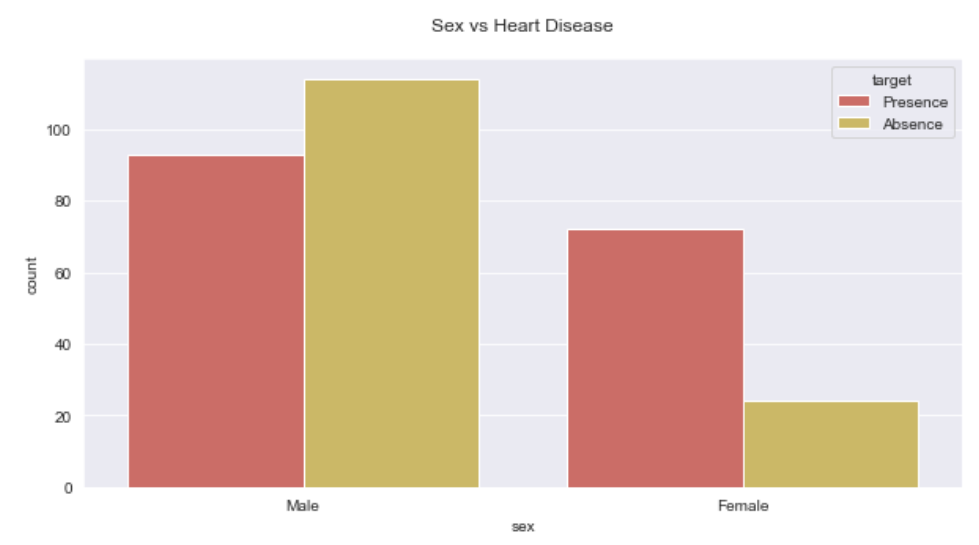
How the features are related to each other or the target variable can be stated using correlation. Correlation can be positive (increase in one value of feature increases the value of the target variable) or negative (increase in one value of feature decreases the value of the target variable).

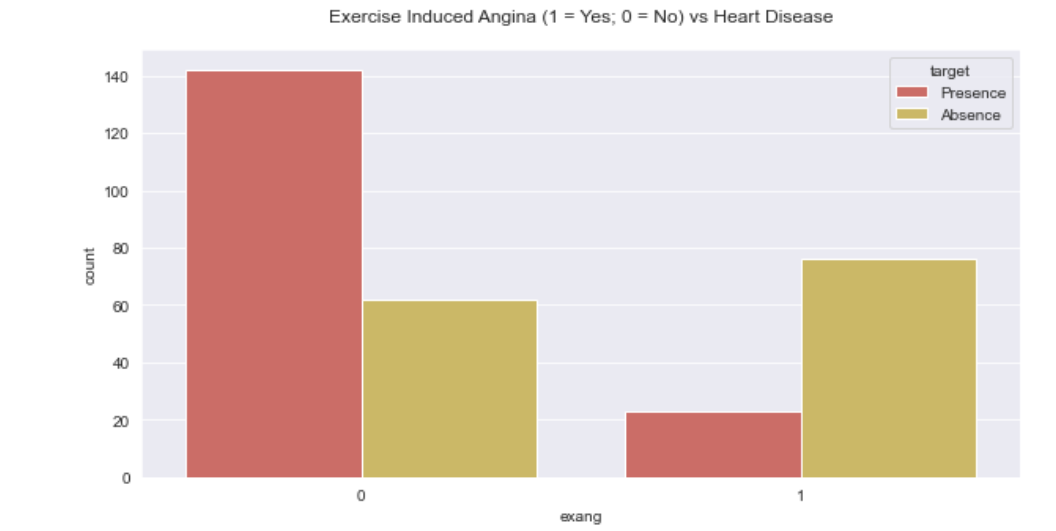
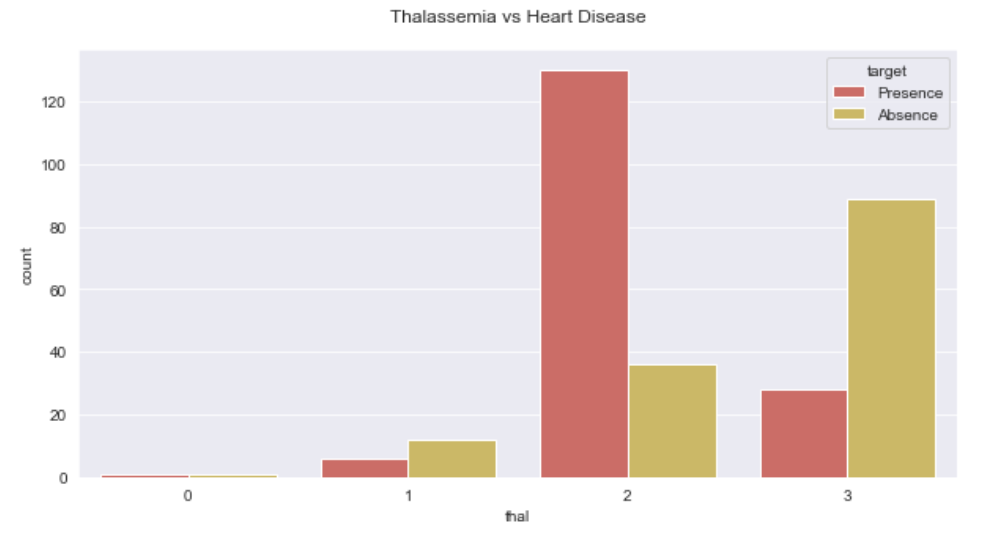
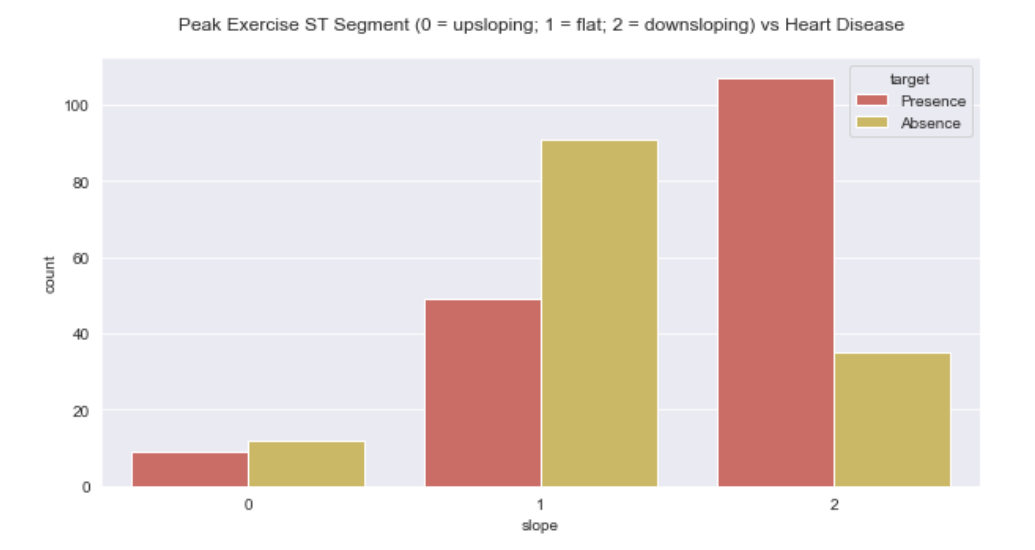
Heatmap makes it easy to identify which features are most related to the target variable, a heatmap of correlated features is plotted using the seaborn library.

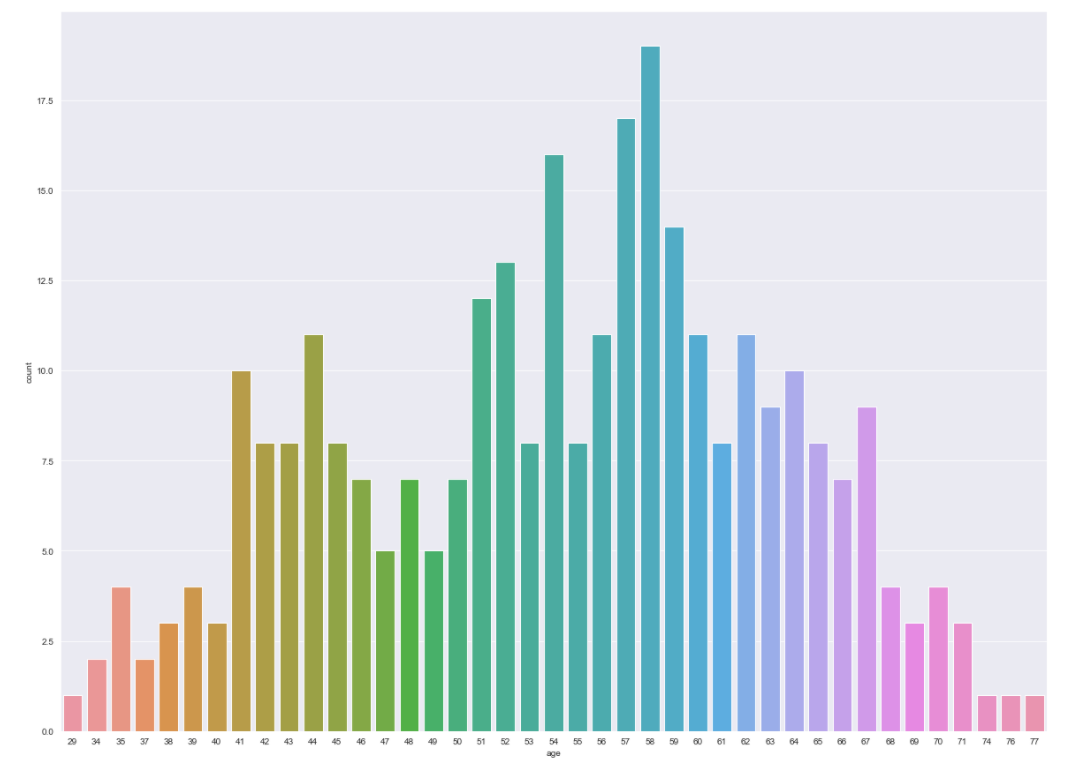
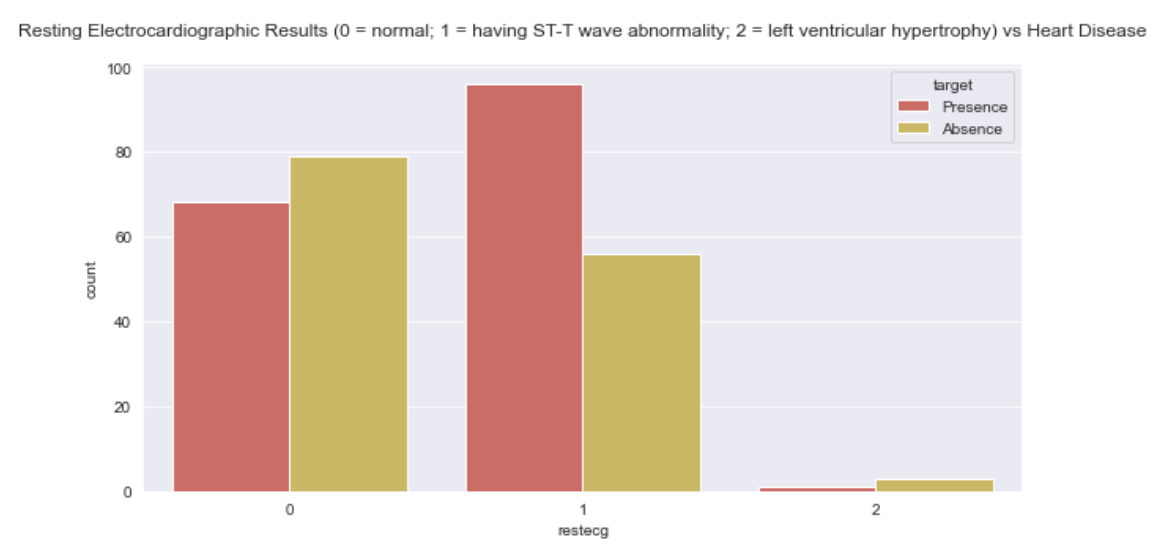
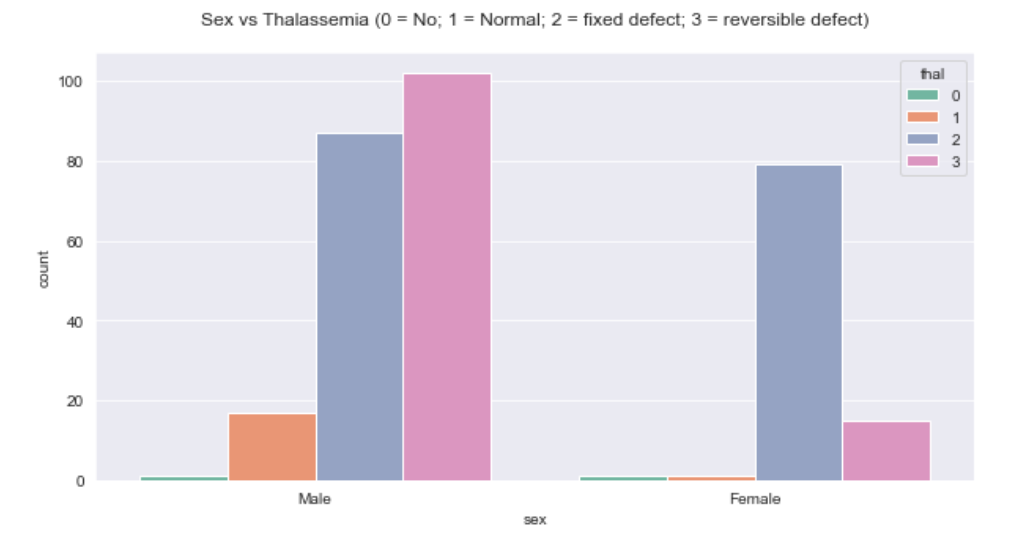
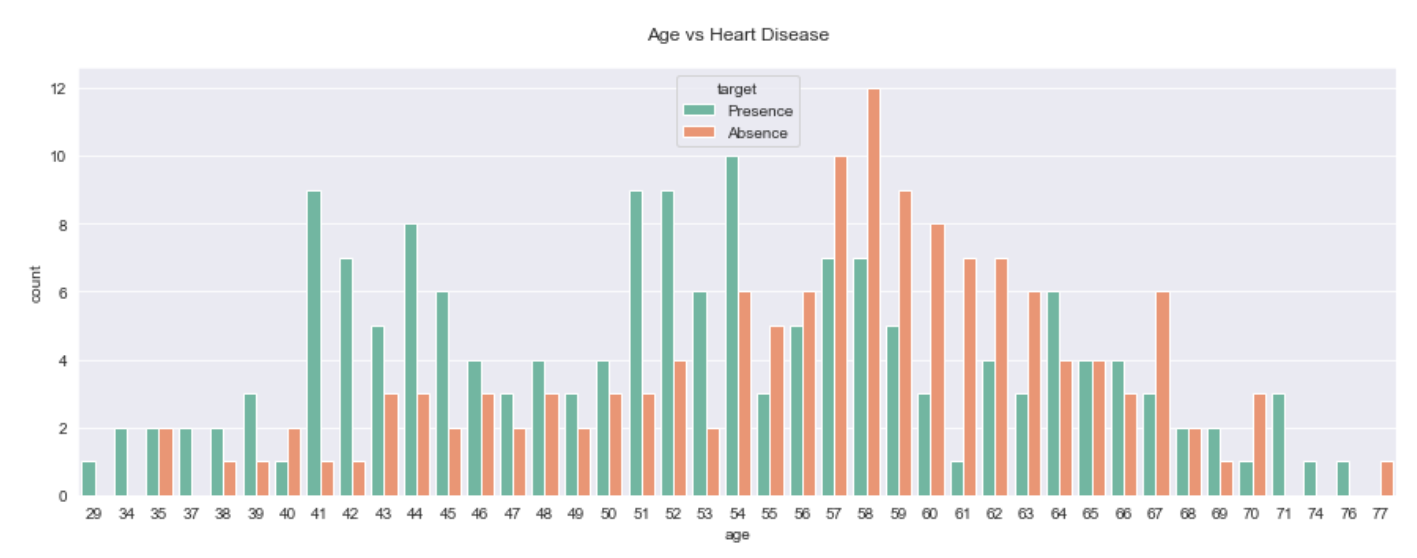
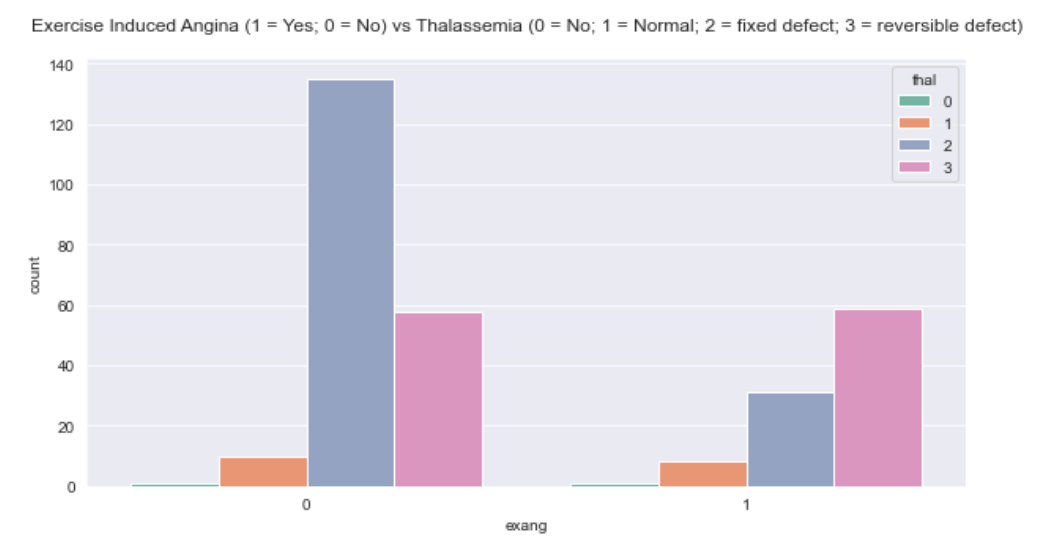


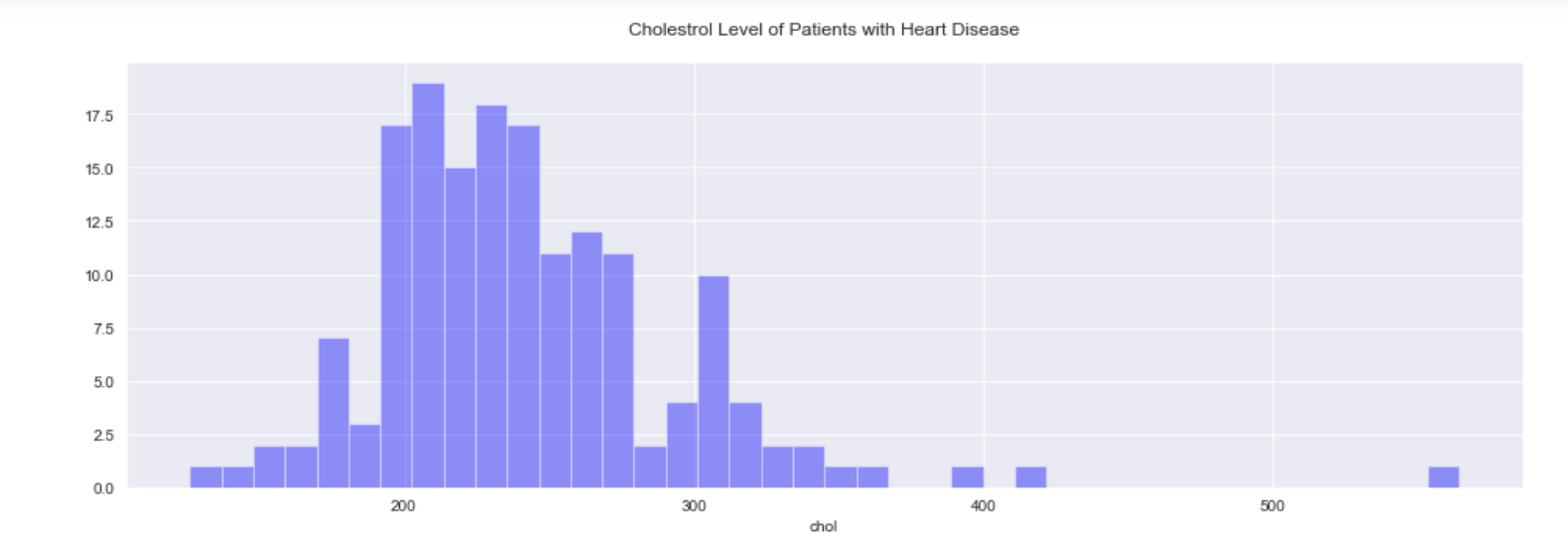
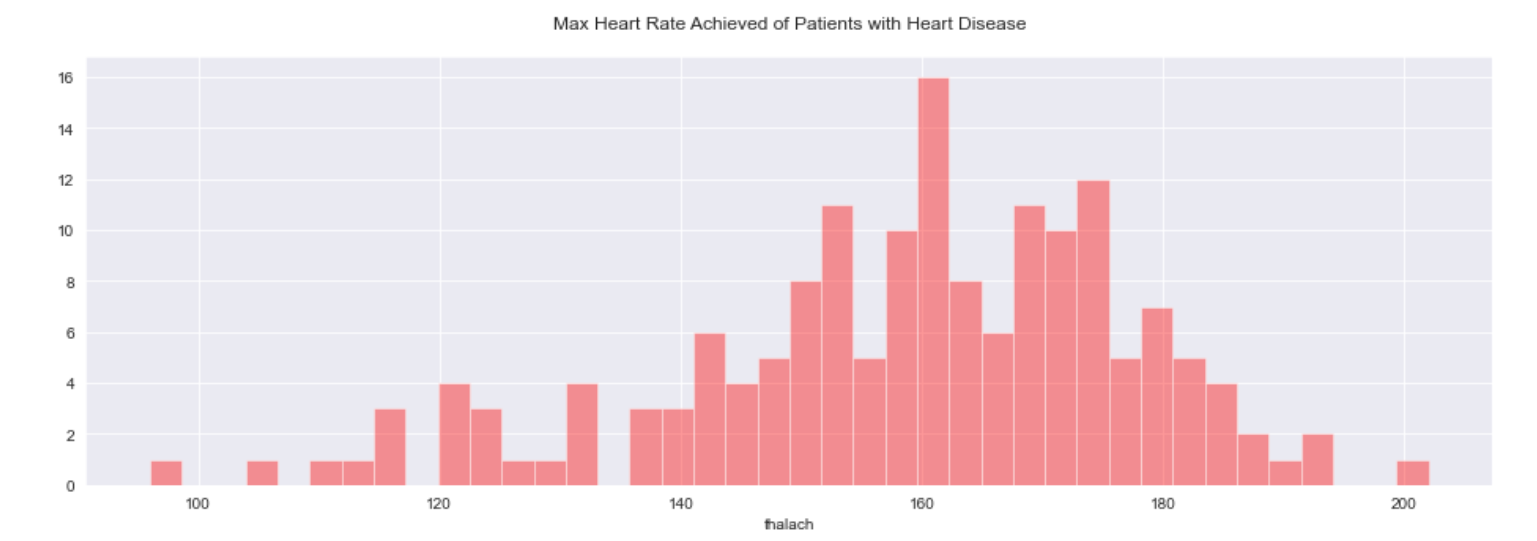
***Step 4: Data Visualization***

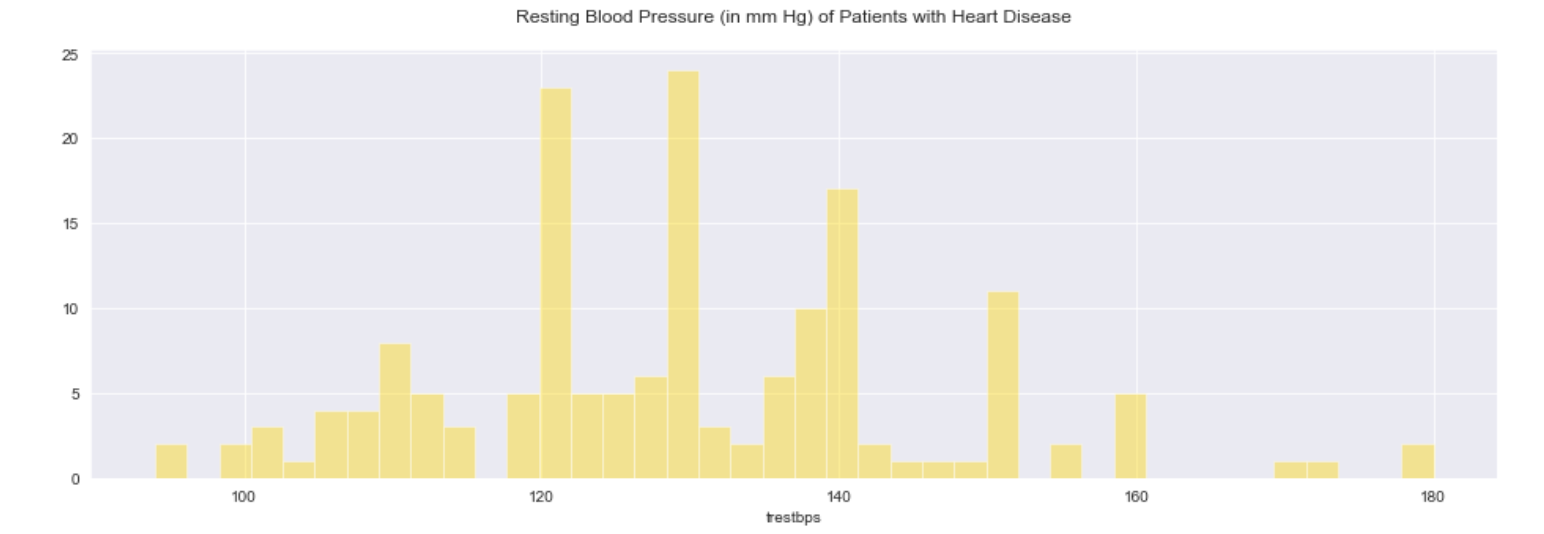
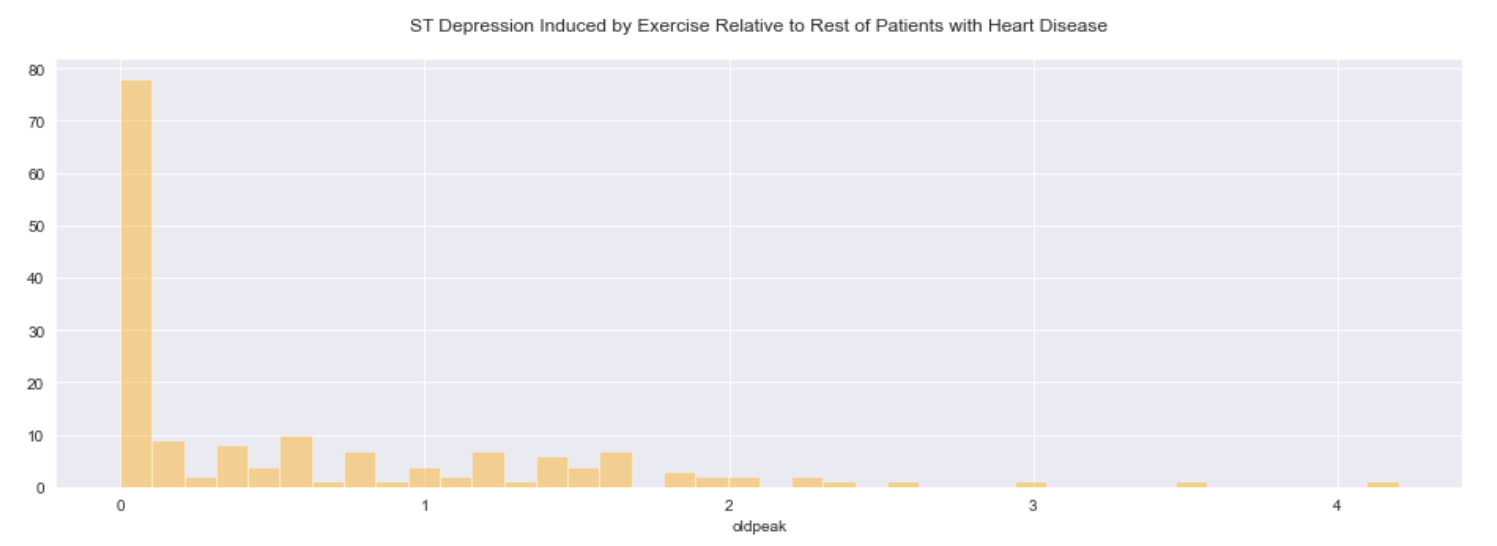
Libraries like Seaborn and Matplotlib are used to plot graphs in order to be better visualize the data.

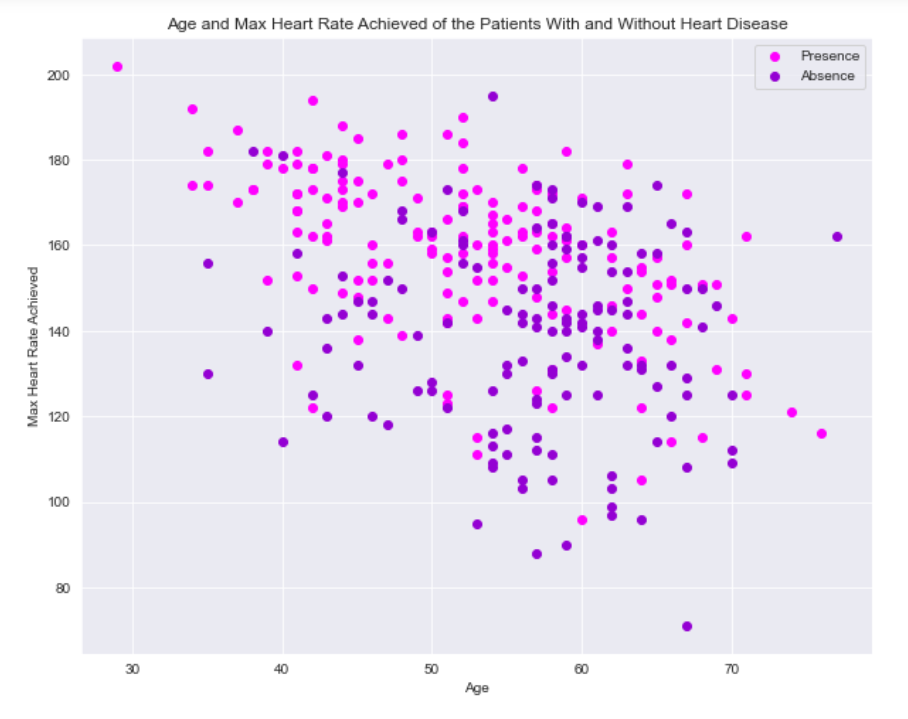
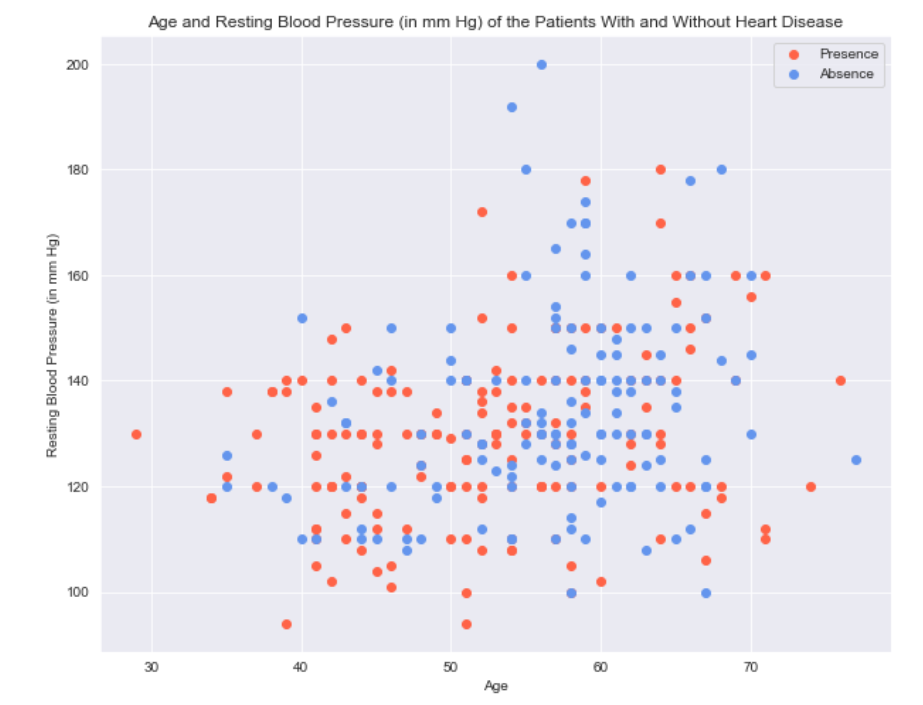
 

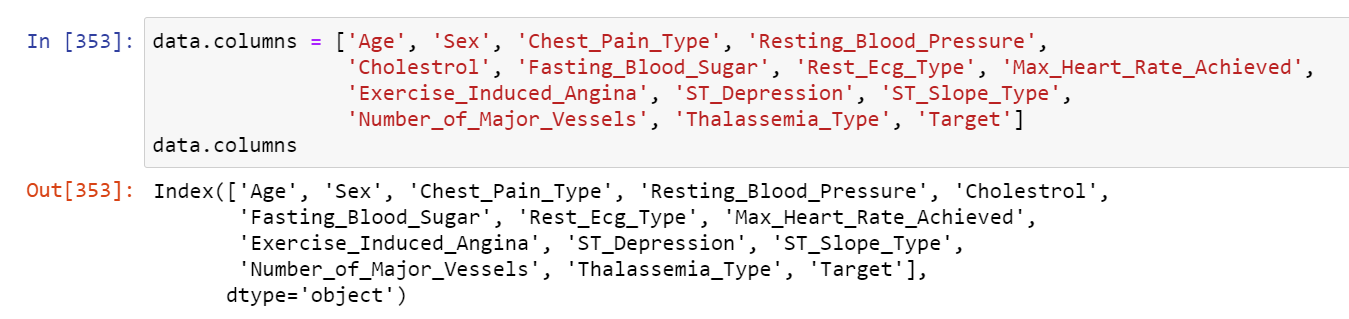
 



***Step 5: Data Pre-processing***

This step transforms raw data into an understandable and readable format in order to make it suitable for building and training a Machine Learning model.

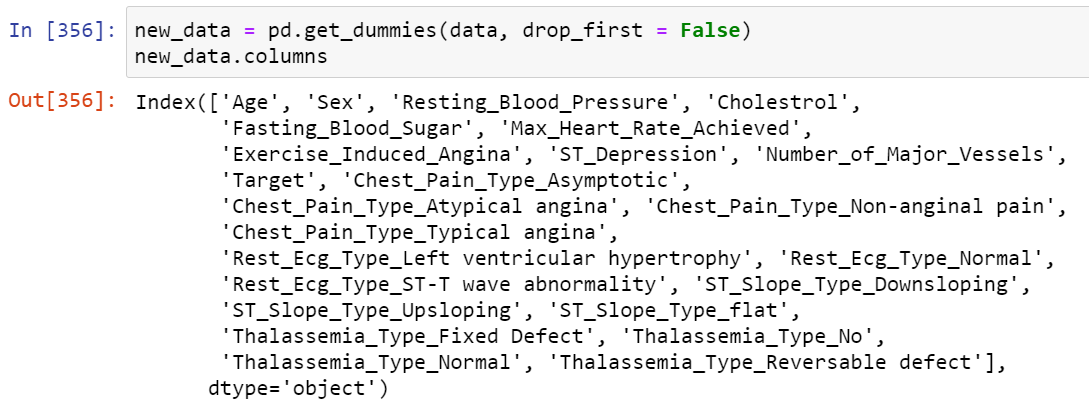
The unobvious column headers are changed into readable information:



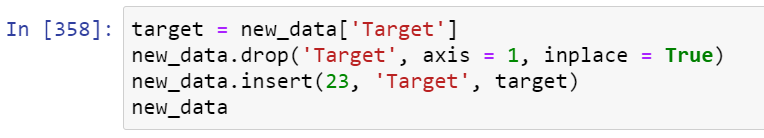
Generate the categorical columns values:



Dummy Encoding is then used, dummy variables take the values 0 and 1 to indicate the presence or absence of the specific categorical effect that will affect the outcome. Value 1 will represent presence while value 0 will represent absence of that particular variable.



The category ‘Target’ is removed then re-inserted in the last column:



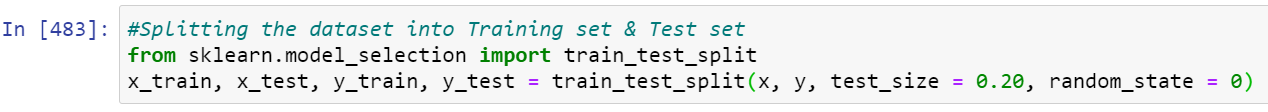
***Step 6: Model Evaluation***

In this step, the two supervised classification machine learning algorithms, Support Vector Machine (SVM) and Logistic Regression are used in the models.

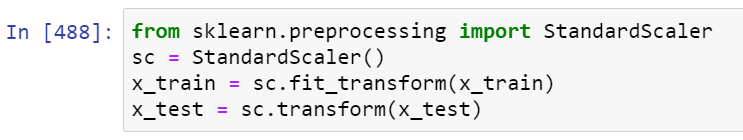
The category ‘Target’ is separated from the dataset:



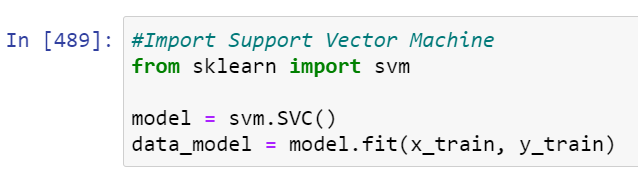
The dataset is then split into two sets, training set and train set with an 80:20 ratio. The training size is 80% while the test size is 20% of the whole data :



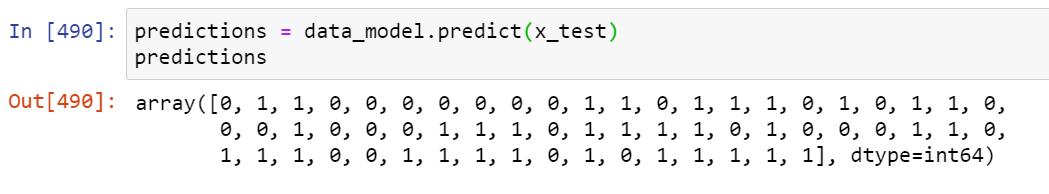
The features are then standardized by removing the mean and scaling to unit variance through importing StandardScaler from sklearn.preprocessing:



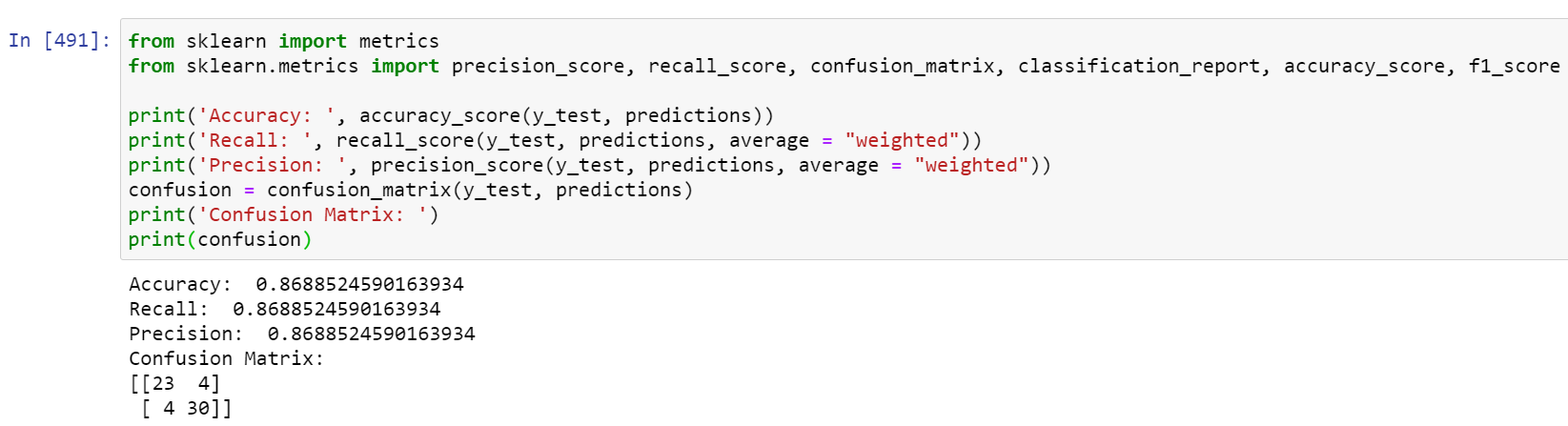
The machine learning algorithm, Support Vector Machine (SVM) is then imported from sklearn and it is fitted into the model:



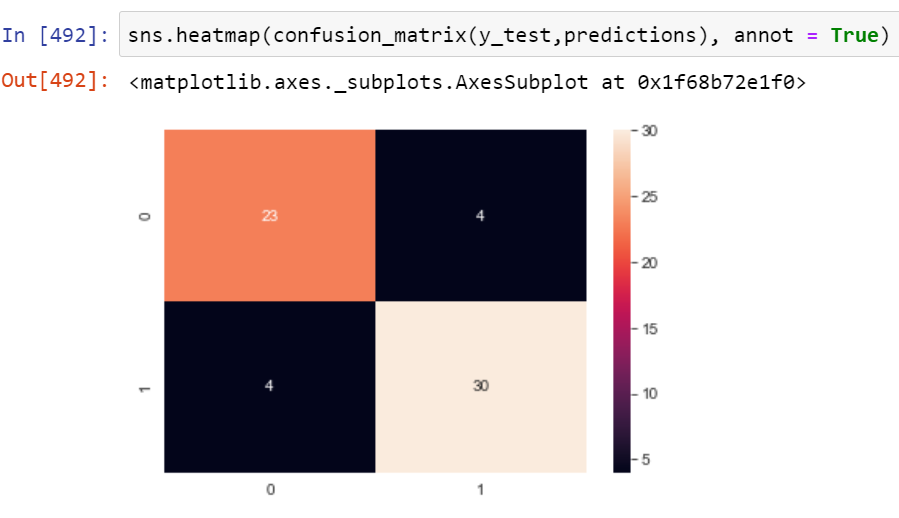
Once done importing and fitting into the model, predictions are then to be made using this model:



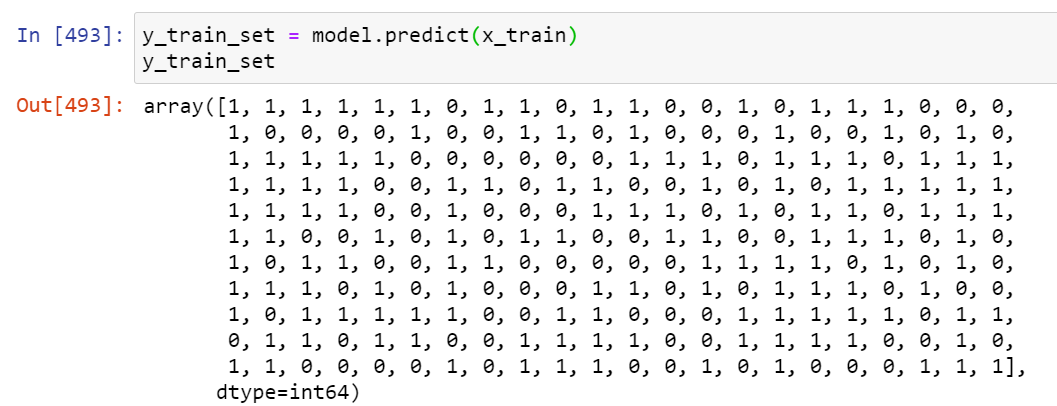
The accuracy score, recall score, precision score and the confusion matrix of this model is then shown:



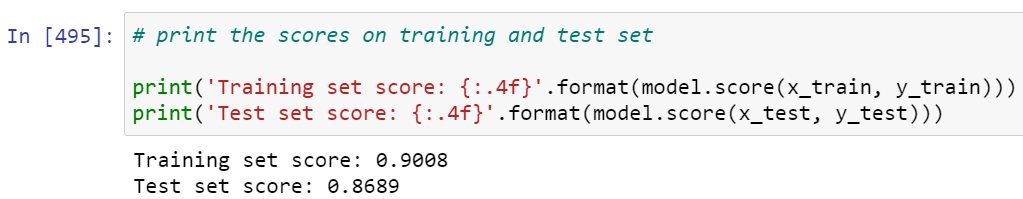
The heatmap of the confusion matrix is then shown as well:



Overfitting and underfitting issues are then checked:

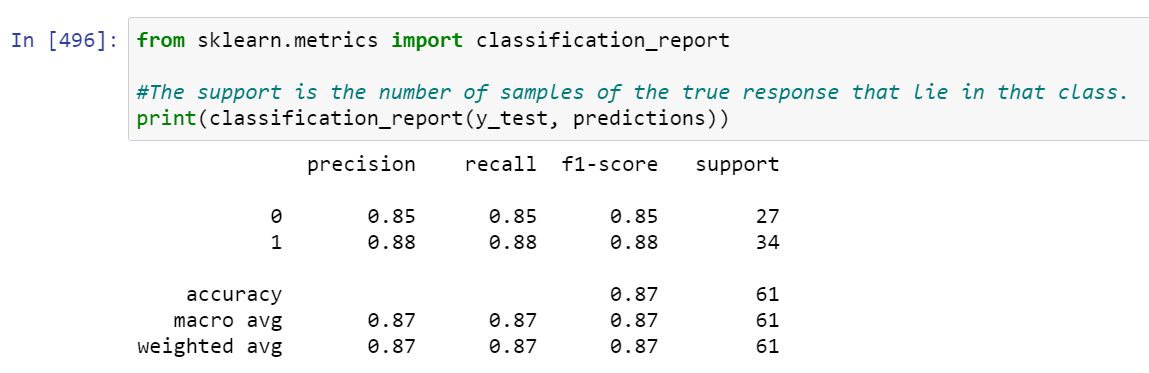


The training set and test set scores are shown to determine whether the model is overfitting or underfitting:

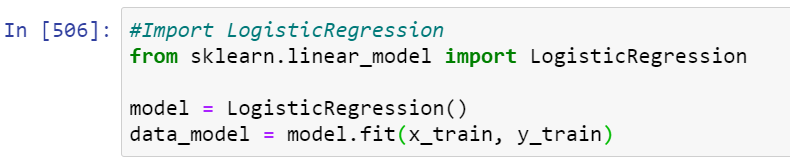


The training set accuracy is 0.9008 while the test set accuracy score is 0.8689. The two values are comparable so there is no overfitting issue.

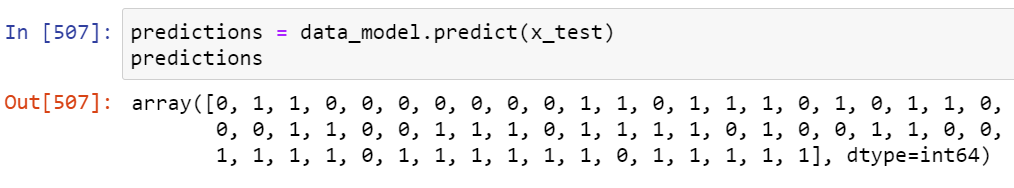
Classification report is then used to display the precision, recall, f1 and support scores of the model.



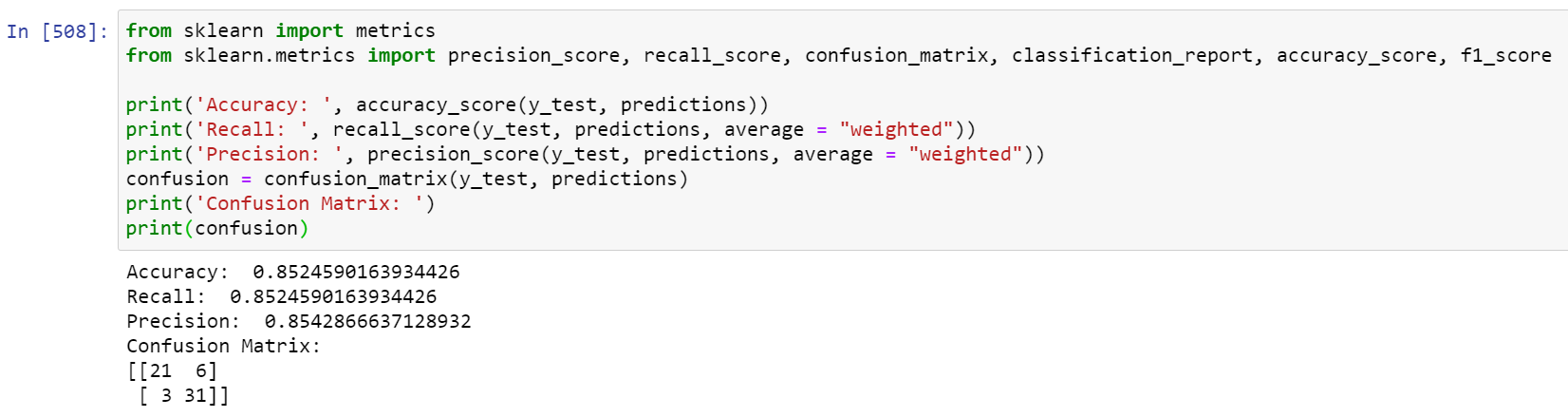
The same steps are repeated using another machine learning algorithm, Logistic Regression:



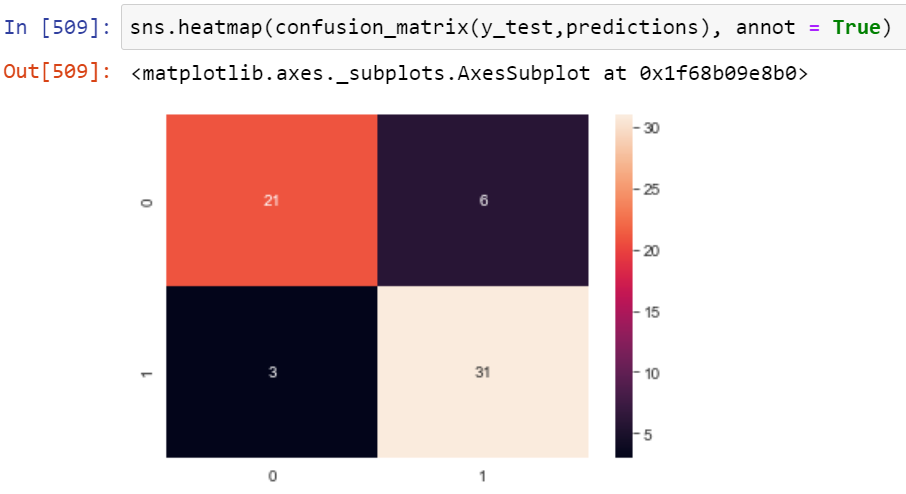
Once done importing and fitting into the model, predictions are then made using this model:



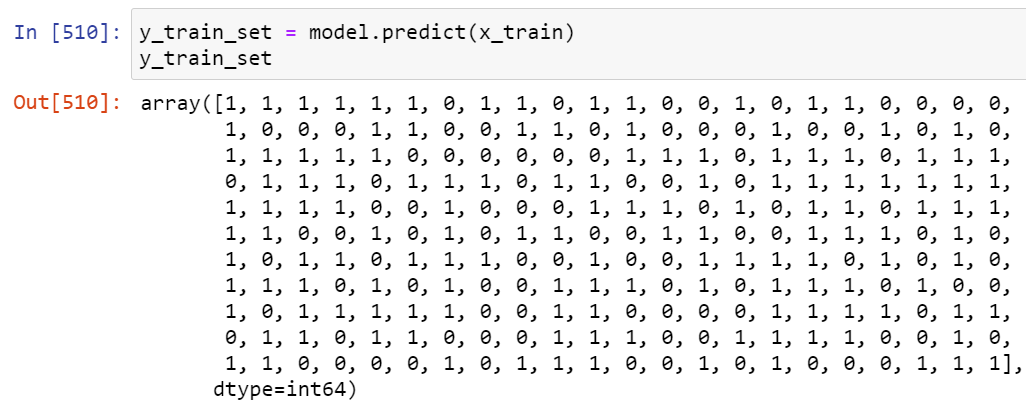
The accuracy score, recall score, precision score and the confusion matrix of this model is then shown:



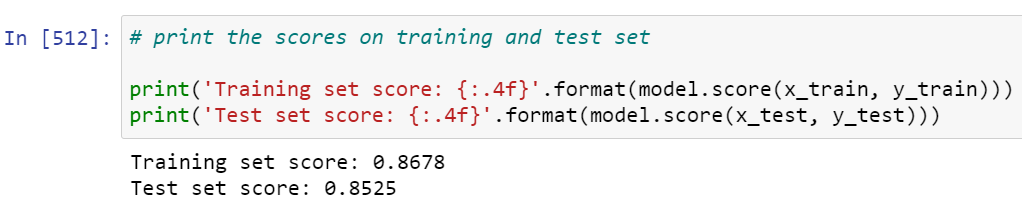
The heatmap of the confusion matrix is then shown as well:



Overfitting and underfitting issues are then checked:

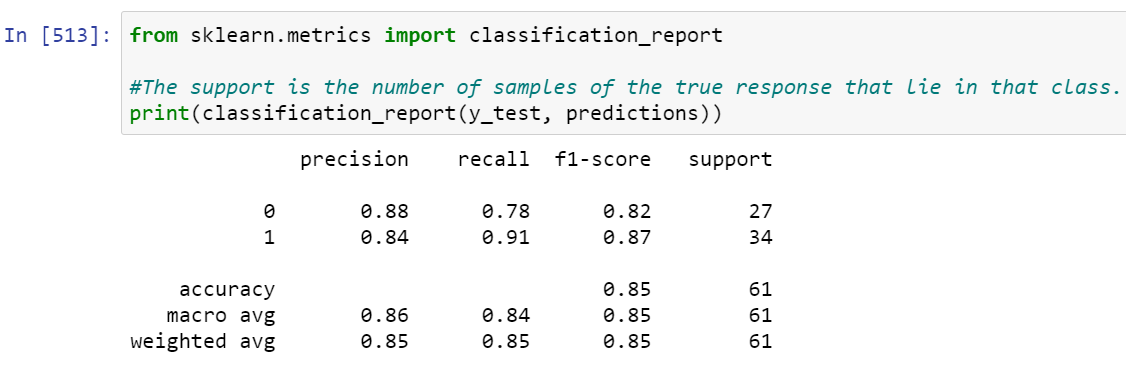


The training set and test set scores are shown to determine whether the model is overfitting or underfitting:



The training set accuracy is 0.8678 while the test set accuracy score is 0.8525. The two values are comparable so there is no overfitting issue.

Classification report is then used to display the precision, recall, f1 and support scores of the model.



Based on the test accuracy scores of both models, the model using Support Vector Machine (SVM) will likely be the champion model since its test accuracy score, 0.8689 is higher than the model using Logistic Regression, 0.8525.

**Comparison and Recommendation**

Although both Support Vector Machine (SVM) and Logistic Regression are supervised classification machine learning algorithm, where they are mostly used to solve classification problems like sorting data into categories. However, there are some key differences between these two algorithms.

For an example, SVM will try to find the best margin (distance from the hyperplane to the nearest data point on each side) to minimize the error on the data, whereas Logistic Regression maximizes the posterior class probability. Moreover, SVM is deterministic whereas Logistic Regression is probabilistic.

SVM can be used to solve both regression and classification problems whereas Logistic Regression can be used to solve classification problems. SVM is based on geometrical properties of the data while Logistic Regression is based on statistical approach.

Furthermore, SVM works well with unstructured and semi-structured data like text and images while Logistic Regression works with already identified independent variable.

Hence, Support Vector Machine (SVM) is the highly recommended machine learning algorithm to be used in a model and that model is more likely to be the champion model to accurately predict whether a patient has heart disease or not.

**Results and Discussion**

In this analysis, the test and training accuracy scores, recall score, precision score, F1 score and the confusion matrix of both models are obtained using Support Vector Machine (SVM) and Logistic Regression with the dataset split into two sets, training set and train set with an 80:20 ratio. The training size is 80% while the test size is 20% of the whole data.

A confusion matrix is an N X N matrix where N is the number of classes being predicted using a model. The accuracy is proportion of the total number of predictions being correct, precision is the proportion of the positive cases that identified being correct, recall is the proportion of the actual positive cases that identified being correct and F1 score is the harmonic mean of precision and recall values for a classification problem.

The training accuracy score obtained by using SVM in a model is 0.9008 whereas the test accuracy score obtained using this algorithm is 0.8689. Since both values are comparable, there is no overfitting issue.

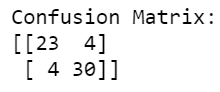
The training accuracy score obtained by using Logistic Regression in a model is 0.8678 whereas the test accuracy score obtained using this algorithm is 0.8525. Since both values are comparable, there is no overfitting issue.

The recall score obtained by using SVM is 0.8689 whereas the recall score obtained using Logistic Regression is 0.8525.

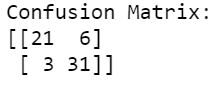
The precision score obtained by using SVM is 0.8689 whereas the precision score obtained using Logistic Regression is 0.8543.

The F1 score obtained by using SVM is 0.8689 whereas the F1 score obtained using Logistic Regression is 0.8512.

The confusion matrices obtained by using SVM and Logistic Regression are shown below:



Confusion matrix of the model using SVM



Confusion matrix of the model using Logistic Regression

**Concluding Remarks**

Heart disease is currently one of the most common diseases around the globe and has become one of the concerns of society today. Despite this, machine learning can be implemented to predict whether a patient has heart disease or not based on existing data since it is difficult to manually determine the odds of getting heart disease based on risk factors. Machine learning techniques are quite useful and accurate in a sense in predicting the output from existing data.

**Lessons Learned From The Project**

There are many lessons can be learned from this project, such as machine learning techniques are more efficient in predicting the odds of a patient getting heart disease compared to manually determine the odds based on risk factors. Machine learning techniques also reduce time and quite cost saving since not many manpower will be involved.

However, only with the right machine learning techniques, can an accurate result be produced. If the wrong machine learning techniques are being implemented such as choosing the wrong machine learning algorithm to make predictions, lives may be lost due to inaccurate results produced.

Hence, research must be done before using the model to make predictions in order to reduce errors. Once research is done and tests have been conducted producing an accurate result, the champion model can then be deployed.

Therefore, picking the right machine learning techniques, doing sufficient research before model deployment to reduce errors and solving the problems logically are also the important lessons that can be learned from this project.

**Conclusion**

Based on the results produced by this experiment, Support Vector Machine (SVM) is the more suitable machine learning algorithm to be used in a model to predict the odds of a patient getting heart disease, with a test accuracy score of 0.8689 compared to the test accuracy score of the model using Logistic Regression as the machine learning algorithm, which is 0.8525.

**References**

1. <https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118#:~:text=The%20term%20%22heart%20disease%22%20is,pain%20(angina)%20or%20stroke>.
2. <https://scikit-learn.org/stable/modules/feature_selection.html#univariate-feature-selection>
3. <https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e>
4. <https://en.wikipedia.org/wiki/Support_vector_machine>
5. https://en.wikipedia.org/wiki/Logistic\_regression
6. <https://www.statisticssolutions.com/what-is-logistic-regression>
7. <https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/>
8. <https://www.kaggle.com/ronitf/heart-disease-uci>
9. <https://archive.ics.uci.edu/ml/datasets/heart+disease>
10. <https://numpy.org/doc/stable/reference/arrays.dtypes.html>
11. <https://www.upgrad.com/blog/data-preprocessing-in-machine-learning/>
12. <https://www.geeksforgeeks.org/differentiate-between-support-vector-machine-and-logistic-regression/>
13. <https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/>
14. https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab\_1