1. Introduction

The main objective of this study is to build a model that can predict the heart disease occurrence, based on a combination of features (risk factors) describing the disease. Different machine learning classification techniques will be implemented and compared upon standard performance metric such as accuracy. The dataset used for this study was taken from UCI machine learning repository, titled "**Heart Disease Data Set**". The stages of the study are listed as follows

- Dataset structure & description
- Analyse, identify patterns, and explore the data
- Data preparation
- Modelling and predicting with Machine Learning
- Evaluation
- Conclusion

2. Dataset Structure and Description

Heart Disease Data Set[1] is used as a data set in this study. Cleverand database has 303 sample in this dataset. This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. The independent variable that needs to be predicted, 'diagnosis', determines whether a person is healthy or suffer from heart disease. Experiments with the Cleveland database have concentrated on endeavours to distinguish disease presence (values 1, 2, 3, 4) from absence (value 0). There are several missing attribute values, distinguished with symbol '?'. The header row is missing in this dataset, so the column names have to be inserted manually. Its attributes and descriptions are given below;

Number	Feature	Detail
1	Age	Age in years
2	Sex	1 for male; 0 for female
3	Chest pain type	Value1: typical angina. Value2: atypical angina. Value3: non-anginal pain. Vlaue4: asymptomatic
4	Resting blood pressure	In mm hg on admission to the hospital
5	Serum cholesterol	In mg/dl
6	Fasting blood sugar > 120 mg/dl	1 for true; 0 for false
7	Resting electrocardiographic results	Value0: normal. Value1: having ST-T wave abnormality (T-wave inversions and/or ST elevation or depression of > 0.05 mV). Value2: showing probable or definite left ventricular hypertrophy by Estes's criteria

8	Maximum heart rate achieved	centered
9	Exercise-induced angina	1 for yes; 0 for no
10	ST depression induced by exercise relative to rest	In mm Hg on admission to the hospital
11	Number of major vessels	(0-3) colored by fluoroscopy
12	The slope of the peak exercise ST segment	Value1: upsloping. Value2: flat. Value3: downsloping
13	Thallium heart scan	3 for normal; 6 for fixed defect; 7 for reversible defect
14	diagnosis	Value0: no disease. Value1: heart disease

Table1: Feature Details

Sex and *chest_pain* are categorical features. *Fasting_blood_sugar, fasting_blood_sugar, electrocardiographic, induced_angina, slope, no_of_vessels, thal, diagnosis* are ordinal features. *Age, blood_pressure, serum_cholestoral, max_heart_rate, ST_depression* are continuous feature which taking values between any two points or between the minimum or maximum values.

3. Analyse, identify Patterns, and Explore the Data

Diagnosis data is ordinal data. In order to make classification and have 2 Ground truth (sick or not), we should group them in 2 labels as shown below;

Diagnosis	Count
0	164
1	139

Table2: Diagnosis Count

Thus, the distribution of patients is as follows;

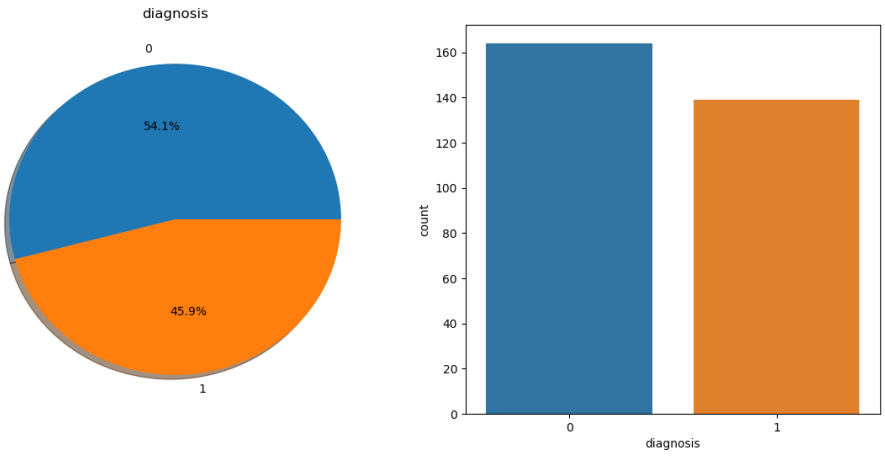

Figure1: Disease Distribution

There is missing data in 2 columns. I explored saving these data rows by replacing missing data with the feature's mode (since these were categorical features). However this made the model predict worse. I ended up dropping the rows with missing data, loosing 6 of the 303 data points.

I also plotted distributions. I cleaned the data so the data types aligned for the analysis portion.

The predictor column ('num') contained 5 categories (0-4) making it difficult to perform a classification. Also, the distribution of the predictor column showed extreme class imbalance. I found instructions that a '0' value indicates no presence of heart disease while 1-4 indicates possibility of heart disease. Therefore, I mapped boolean values to this column which changed the 1-4 values to '1'. Now this became a classification problem of identifying patient as not having heart disease (0 prediction) or having the presence of heart disease (1 prediction). Doing this also addressed the previous class imbalance problem so I no longer needed to explore options to address this such as stratification, oversampling, or synthetic minority oversampling technique (SMOTE).

Gender distribution and disease distribution by gender are given below.

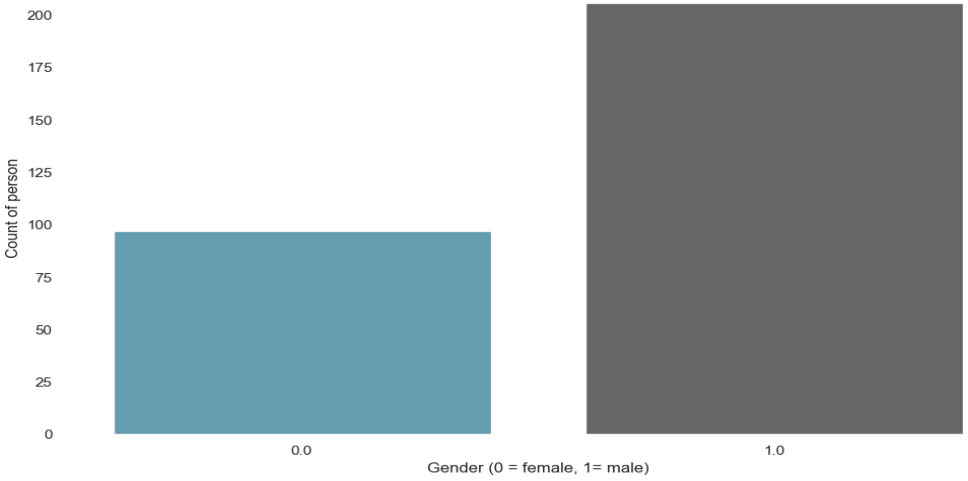

Figure2: Distribution of Gender

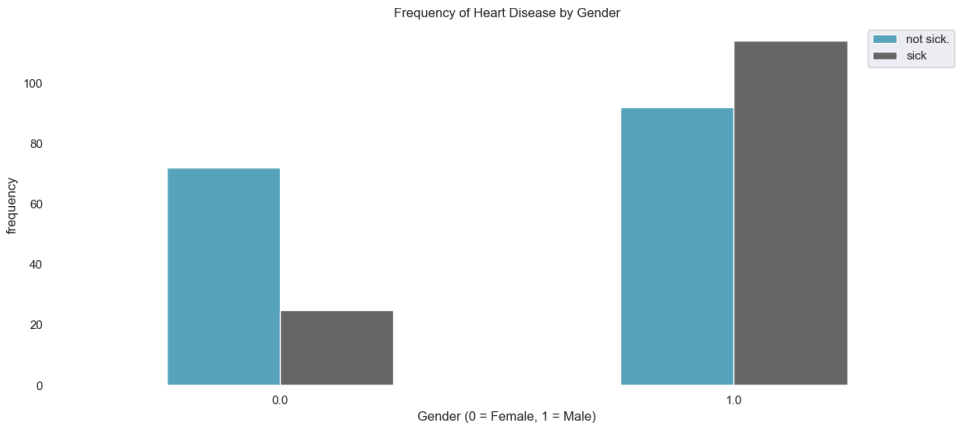

Figure3: Disease Distribution by Gender

Disease distribution by age is given below.

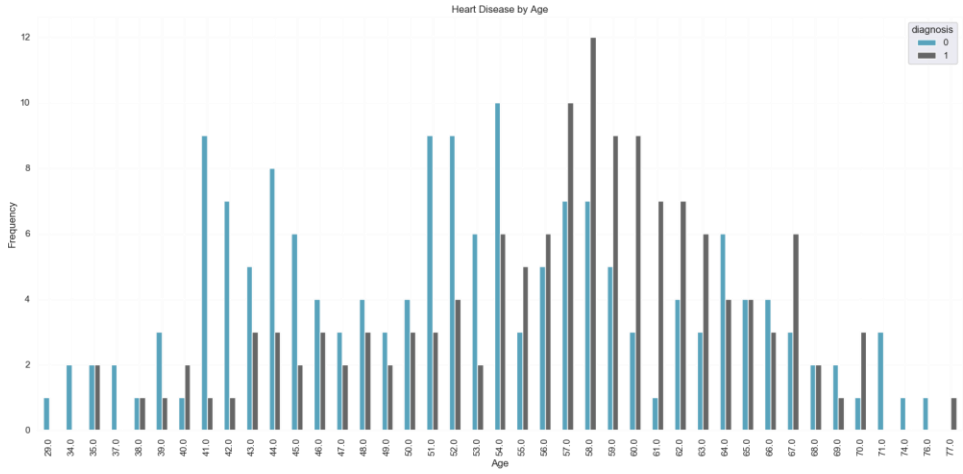

Figure4: Disease Distribution by Age

Disease Frequency by Slope Variable is given below.

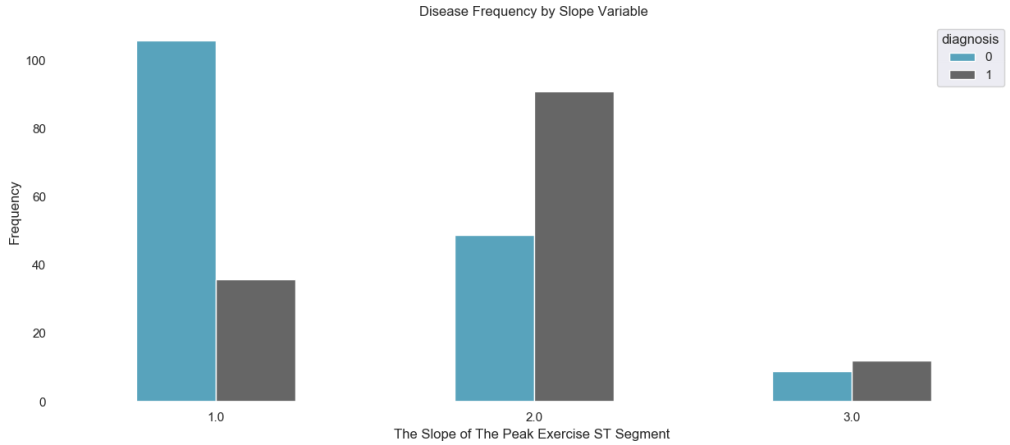

Figure5: The Slope of The Peak Exercise ST Segment

Frequency of Heart Disease by Fasting Blood Sugar is given below.

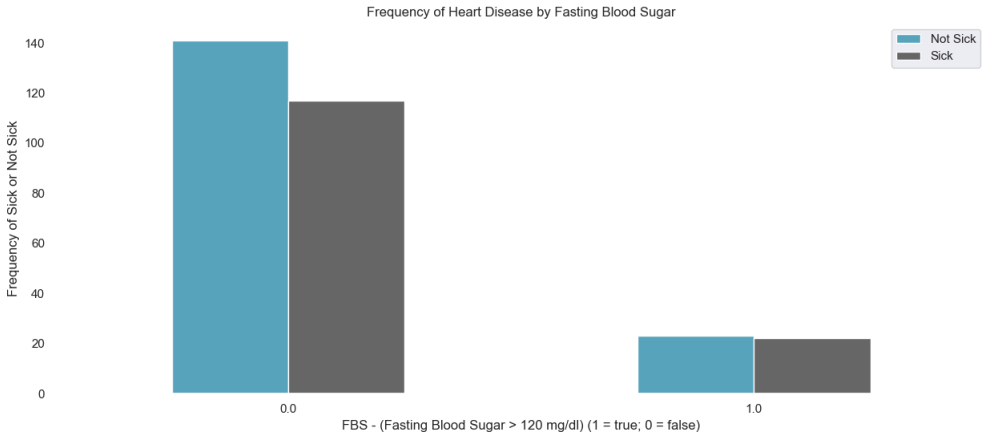

Figure6: Frequency of Heart Disease by Fasting Blood Sugar

Heart Disease Frequency by Chest Pain Type is given below.

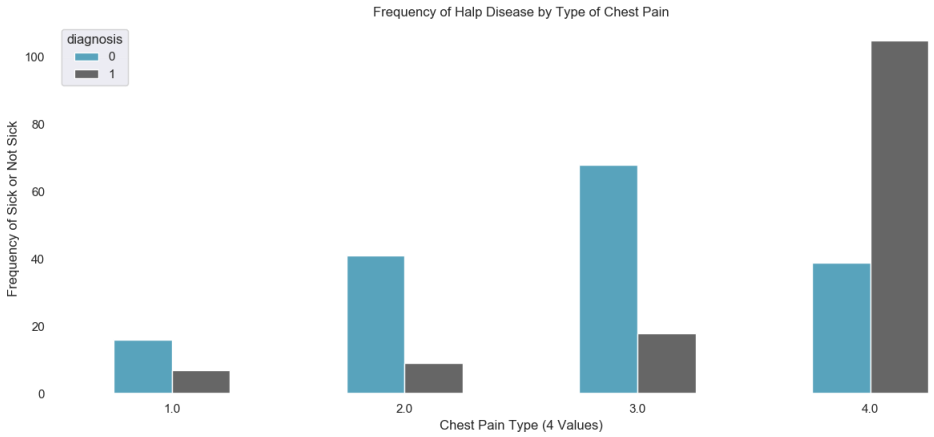

Figure7: Heart Disease Frequency by Chest Pain Type

Findings;

- Men are much more prone to get a heart disease than women.
- The higher number of vessels detected through fluoroscopy, the higher risk of disease.
- While soft chest pain may be a bad symptom of approaching problems with heart (especially in case of men), strong pain is a serious warning!
- Risk of getting heart disease might be even 3x higher for someone who experienced exercise-induced angina.
- The flat slope (value=2) and downslope (value=3) of the peak exercise indicates a high risk of getting disease

Almost even distribution suggests *fasting blood sugar* to be a very weak feature for our prediction, therefore it could be excluded from our model. It is not likely that leaving this variable will improve our model accuracy, yet it shouldn't make it worse too. I decide to keep this variable as it is and confirm my hypothesis through checking feature importance of a few models.

4. Data Preparation

There are 6 missing values in total in the two columns of the Dataset.

Column	Count of Missing value
Age	0
sex	0
chest_pain	0
blood_pressure	0
serum_cholestoral	0
fasting_blood_sugar	0
electrocardiographic	0
max_heart_rate	0
induced_angina	0

ST_depression	0
slope	0
no_of_vessels	4
thal	2
diagnosis	0

Table2: Missing Values in the Dataset

Both columns containing missing values are categorical. In such case, mode (most frequently occurring value in a given vector) is usually used for filling 'nans'.

There are 303 samples in total in the dataset. I use 70% of them for training and 30% for testing. Accordingly, the distribution is as follows

- train_set_x shape: (212, 13)
- train_set_y shape: (212,)
- test_set_x shape: (91, 13)
- test_set_y shape: (91,)

Data needs to be normalized or standardized before applying to machine learning algorithms. For feature selection;

firstly, standard deviation is used in the *classifications.py* file.

Then, PCA(Principal Component Analysis) and standard deviation are used in the *classifications_PCA.py* file.

And finally, LDA(Linear Discriminant Analysis) and standard deviation are used in the *classifications_LDA.py* file.

5. Modelling and Predicting with Machine Learning

In this study, the occurrence of heart disease is predicted with the highest accuracy. Several classification algorithms are used to achieve this. This section contains all the results from the study and presents the best performing according to the accuracy metric. Throughout the classification methods, k-nearest neighbor algorithm (k-NN), Decision Tree, Logistic Regression, Gaussian Naïve Bayes, Support Vector machine(SVM) and Linear SVM algorithms are used to solve supervised learning problems.

To run all these algorithms in a single function, a method as follows was developed.

*def train_model(X_train, y_train, X_test, y_test, classifier, **kwargs):*

Kwargs is name of sklearn library's algorithm.

Performance metrics and scores;

Accuracy: is the ratio between the number of correct predictions and total number of predications.

$acc = \frac{TP + TN}{TP + TN + FP + FN}$

Precision: is the ratio between the number of correct positives and the number of true positives plus the number of false positives.

$Precision (p) = \frac{TP}{TP + FP}$

Recall: is the ratio between the number of correct positives and the number of true positives plus the number of false negatives.

$recall = \frac{TP}{TP + FN}$

Problem characteristics in context of our case study:

TP = True positive (has heart disease). TN = True negative (has no heart disease). FP = False positive (has no heart disease) FN = False negative (has heart disease)

Scores of text results;

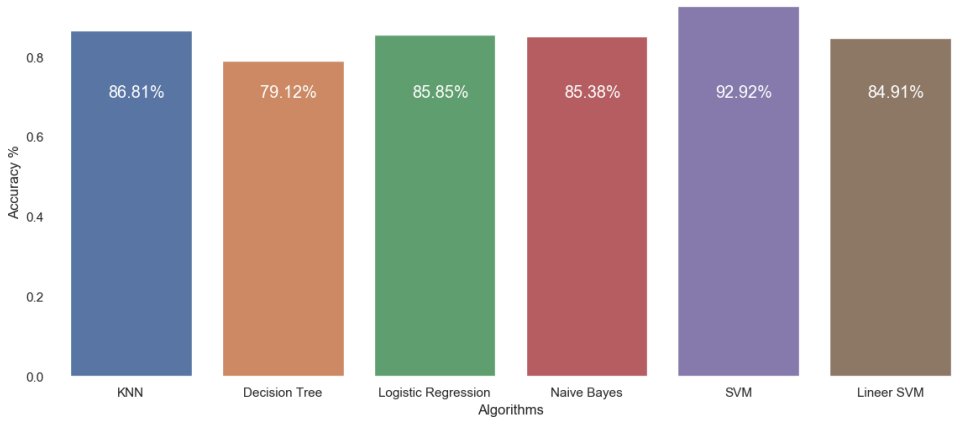

Figure8: Compare accuracy of Algorithms

Algorithm	Accuracy	Precision	Recall
KNN	0,868	0,842	0,82
Decision Tree	0,791	0,833	0,641

Logistic Regression	0,857	0,861	0,794
Gaussian Naïve Bayes	0,868	0,864	0,82
SVM	0,842	0,794	0,794
SVM Linear	**0,879**	0,889	0,82

Table3: Compare Performance Metrics of Algorithms

K-NN accuracy according to n_neigbors;

N_neigbor	**Accuracy**
1	0,747
2	0,791
3	0,832
4	0,846
5	**0,868**
6	**0,868**
7	**0,868**
8	0,857
9	0,857

Table4: Compare Accuracy of KNN Neigbors

Decision Tree accuracy according to max_depth;

Max Depth	**Accuracy**
1	0,747
2	0,725
3	0,769
4	0,78
5	0,78
6	**0,791**
7	0,758

Table5: Compare Accuracy of Decision Tree Depth

When statistical (manuel), PCA and LDA feature selection method's accuracy compared; the result of comparison can be observed in detail below.

Algorithm	**Accuracy of Statistical**	**Accuracy of PCA**	**Accuracy of LDA**
KNN	0,868	0,862	0,824
Decision Tree	0,791	0,813	0,78
Logistic Regression	0,857	0,824	0,868
Gaussian Naïve Bayes	0,868	0,857	0,868

SVM	0,842	0,835	0,868
SVM Linear	**0,879**	0,857	0,868

Table6: Compare Accuracy of Feature Selection Methods

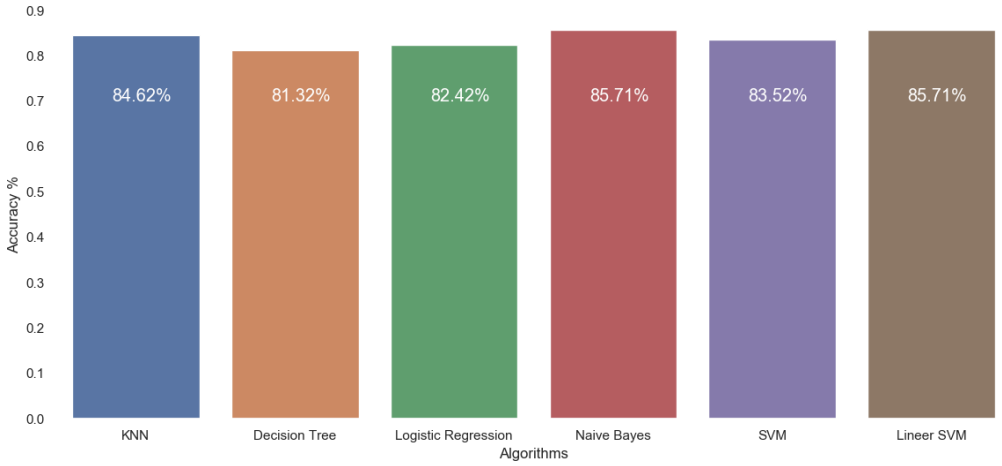

Figure9: Compare accuracy of Algorithms with PCA Feature Selection Method

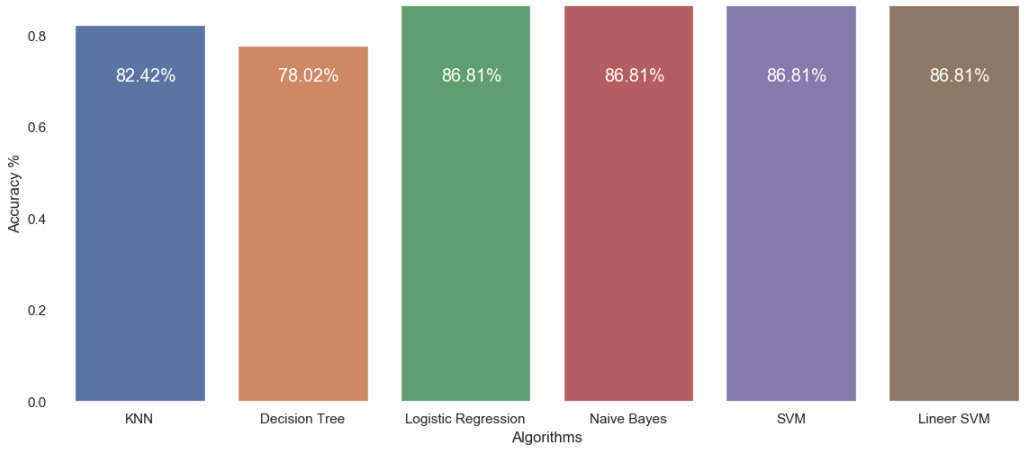

Figure10: Compare accuracy of Algorithms with LDA Feature Selection Method

6. Evaluation

First of all, when we examine Table 3, it can be seen that the success of SVM linear is the highest with 0.879. This may be due to the large number of features and the fact that SVM linear has proven itself in machine learning methods. It is known that the success of Decision Tree regression applications is low in the prediction of continuous values and in

complex data sets. In this study, it showed the lowest success with an accuracy value of 0.791.

In the KNN algorithm, the model was trained 9 times according to the neigbors and the success rates were specified in Table4. According to these success rates, the accuracy values of 5, 6, 7 neigbors were determined as the highest. The reason for this can be interpreted as the distance of the point to be predicted from the other points is optimal in 5, 6 and 7.

Likewise, in the Decision Tree algorithm, the model was trained 7 times according to the depths and the accuracy was specified in Table5. Among these accuracies, the 6th depth showed the highest success. However, the reason for this cannot be understood without simulating the tree.

The features that will be input for the model were determined by statistical(manual), PCA and LDA methods. The statistical method gives the most successful result in the comparison whose results are specified in Table6. The number of components is given as 5 in PCA and 1 in LDA. ***ValueError: n_components cannot be larger than min(n_features, n_classes -1)*** error is received when component parameter is 2 or higher in LDA. However, the total number of features is 14 and this number should not cause any problems. It was created as a result of analyzing the data set specified in the manual feature selection 3. *Analyse, identify Patterns, and Explore the Data* section. For example, the Fasting Blood Sugar feature was not used when training the model because this feature does not seem to have a significant effect on the outcome of being sick or not.

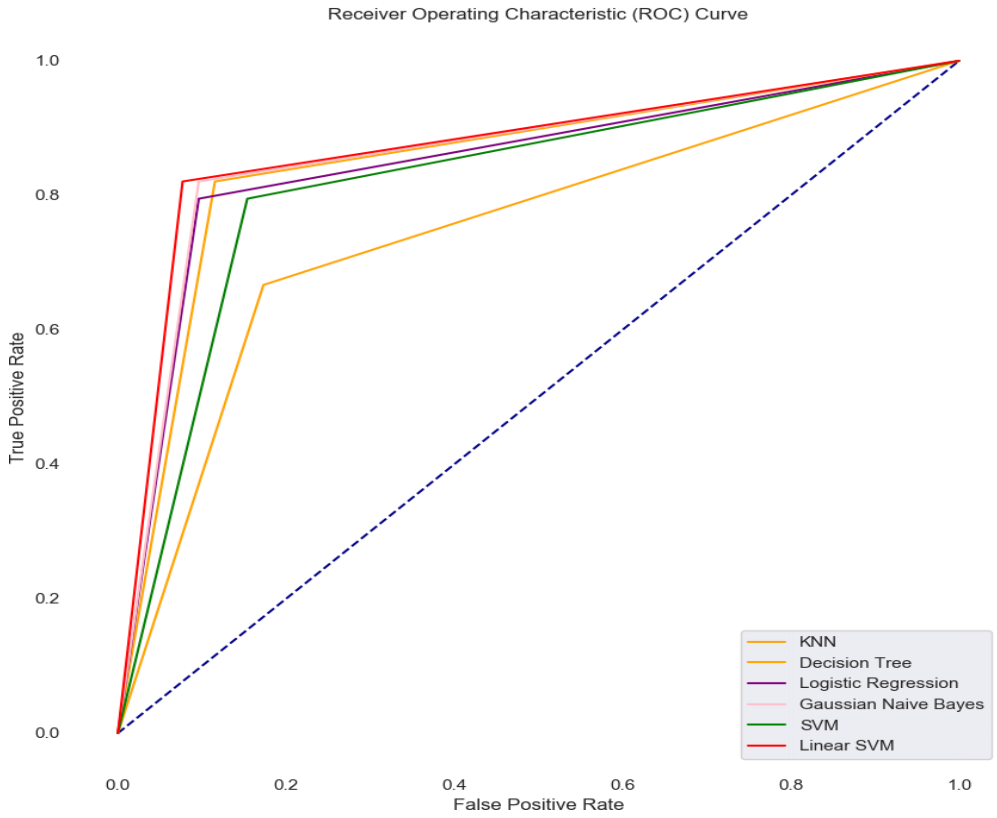

Figure11: ROC Curve

As a result, the Linear SVM algorithm achieved the highest accuracy with the statistical feature selection method, as indicated in Figure 11.

7. References

Aha, D. W. (n.d.). *Heart Disease Data Set*. UCI Machine Learning Repository: Heart disease dataset. Retrieved November 28, 2021, from https://archive.ics.uci.edu/ml/datasets/heart+disease.