Tools & Techniques Lab Project

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

CardiPredict: A Machine Learning Solution to predict Heart Disease using SVM Algorithm

Abstract

This mini-project by sixth-semester computer science and engineering students at Kalinga Institute of Industrial Technology explores the use of Support Vector Machine (SVM) to predict heart disease based on patient attributes. Early detection is crucial in reducing mortality rates, but continuous monitoring by clinicians is not always possible. The use of machine learning techniques can aid in making decisions about lifestyle changes in high-risk patients, reducing complications, and improving prognosis.

Introduction

- Heart Disease causes around 12 million deaths globally every year.
- Early detection of heart disease is crucial for making informed decisions about lifestyle changes and reducing the risk of complications.
- This project aims to use machine learning algorithms to analyze patient data.
- The goal is to predict the likelihood of future heart disease using these algorithms.

Objectives

The main objective of developing this project are:

- To develop a machine learning model to predict the future possibility of heart disease by implementing Support Vector Machine(SVM).
- To determine significant risk factors based on medical dataset which may lead to heart disease.
- To analyze feature selection methods and understand their working principle.

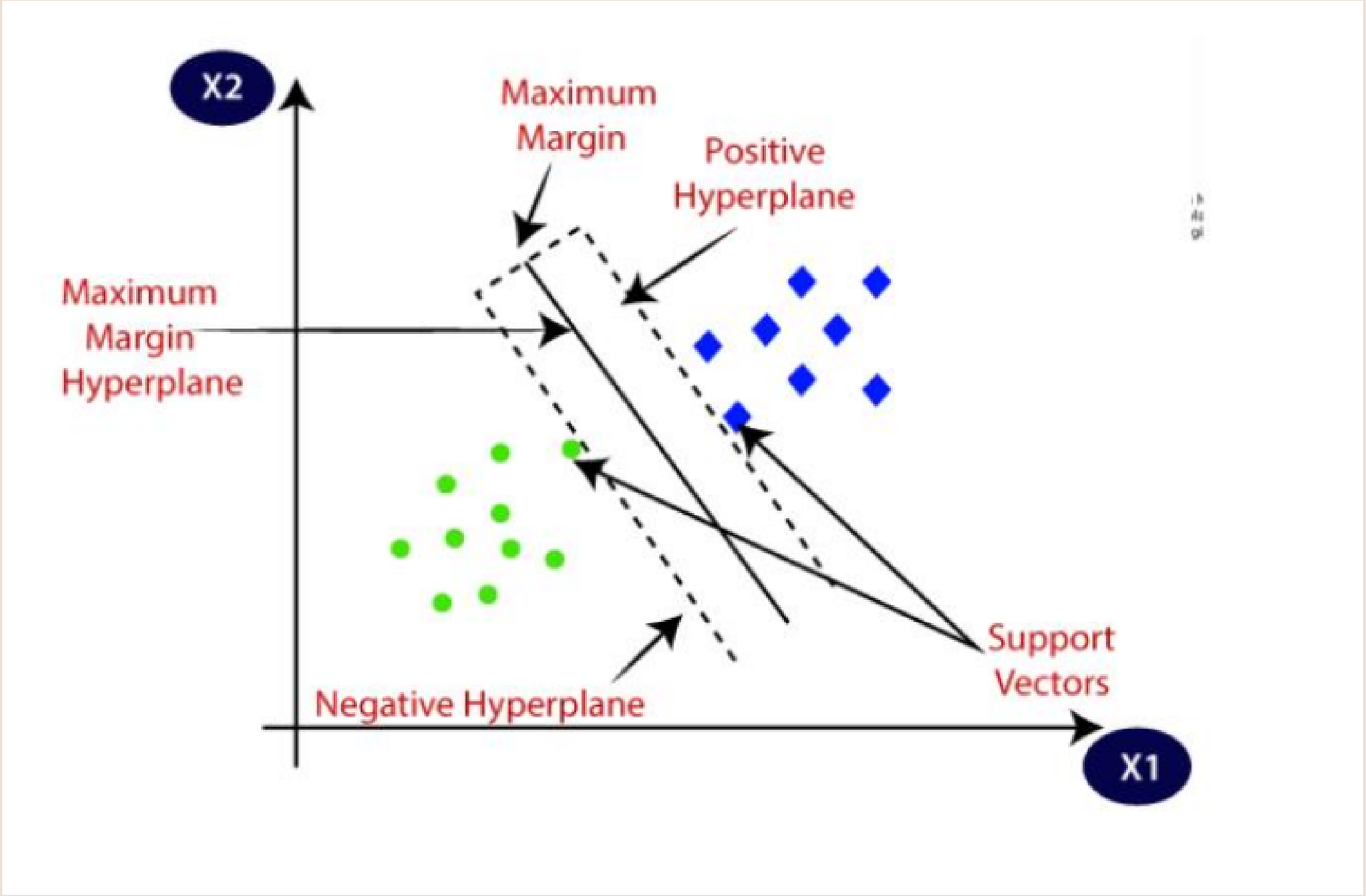
DATASETS

- The dataset is part of an ongoing cardiovascular study on residents of Framingham, Massachusetts.
- The dataset contains over 4000 records and 14 attributes.
- The attributes include age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiograph results, maximum heart rate, exercise-induced angina, ST depression induced by exercise, slope of the peak exercise, number of major vessels, and target ranging from 0 to 2, where 0 indicates the absence of heart disease.
- The dataset is in CSV format and is prepared as a data frame using the Pandas library in Python.

METHODS AND ALGORITHMS USED

Support Vector Machine

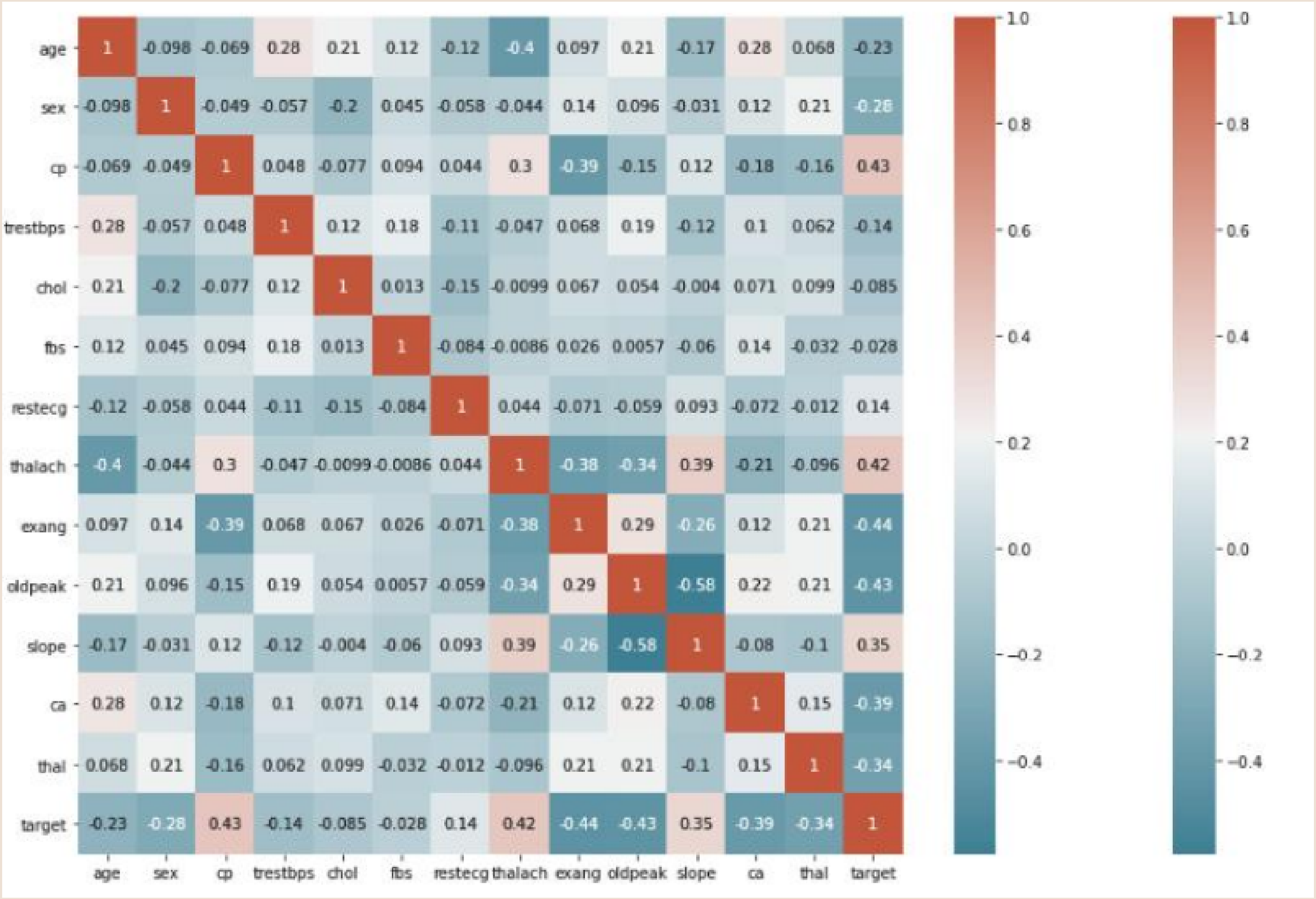
- The goal of Support Vector Machine (SVM) is to create the best line or decision boundary that can separate n-dimensional space into classes.
- The purpose of the decision boundary is to classify new data points into the correct category in the future.
- The decision boundary is called a hyperplane, and SVM selects extreme points/vectors that help in creating the hyperplane.
- The extreme cases are called support vectors, and hence the algorithm is named Support Vector Machine.
- SVM uses support vectors to classify data points into different categories.
- The diagram provided shows two different categories classified using a decision boundary or hyperplane.


X2
Maximum
Margin
Positive
Hyperplane
Maximum
Margin
Hyperplane
Support
Vectors
Negative Hyperplane
X1

EXPLORATORY DATA ANALYSIS

Correlations

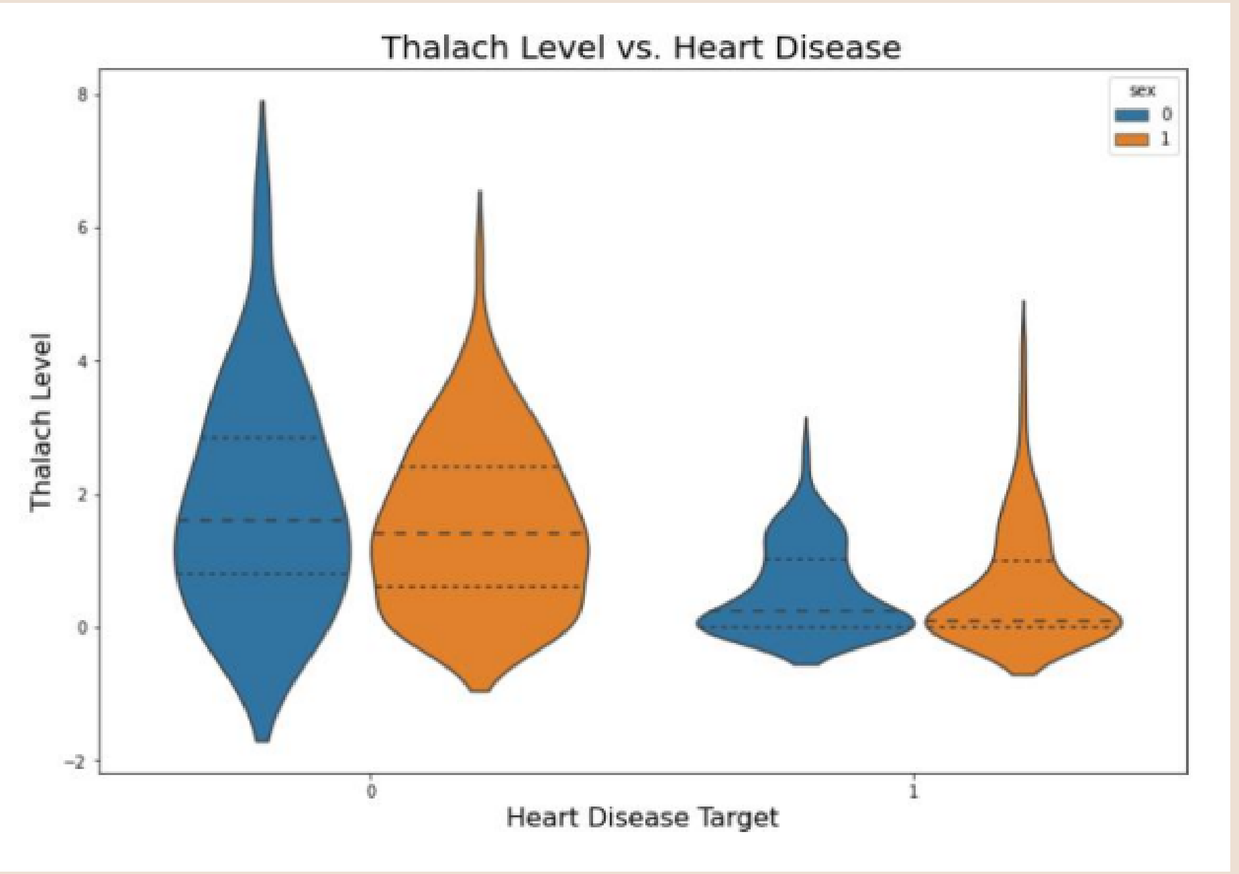
- A correlation matrix is a table that displays correlation coefficients for different variables, allowing for the identification and visualization of patterns in data.
- In the given dataset, there is a positive correlation between chest pain (cp) and the target variable, which makes sense since a greater amount of chest pain increases the chance of heart disease.
- There is a negative correlation between exercise-induced angina and the target variable, which also makes sense because narrowed arteries can slow down blood flow during exercise.
- Pair plots are another way to visualize correlations between variables, but in this case, only continuous features were used to avoid clutter.
- ST segment depression can contribute to heart disease, with low ST depression indicating greater risk.
- The "slope" hue in the plots refers to the peak exercise ST segment, with both positive and negative heart disease patients exhibiting similar distributions across the three slope categories.

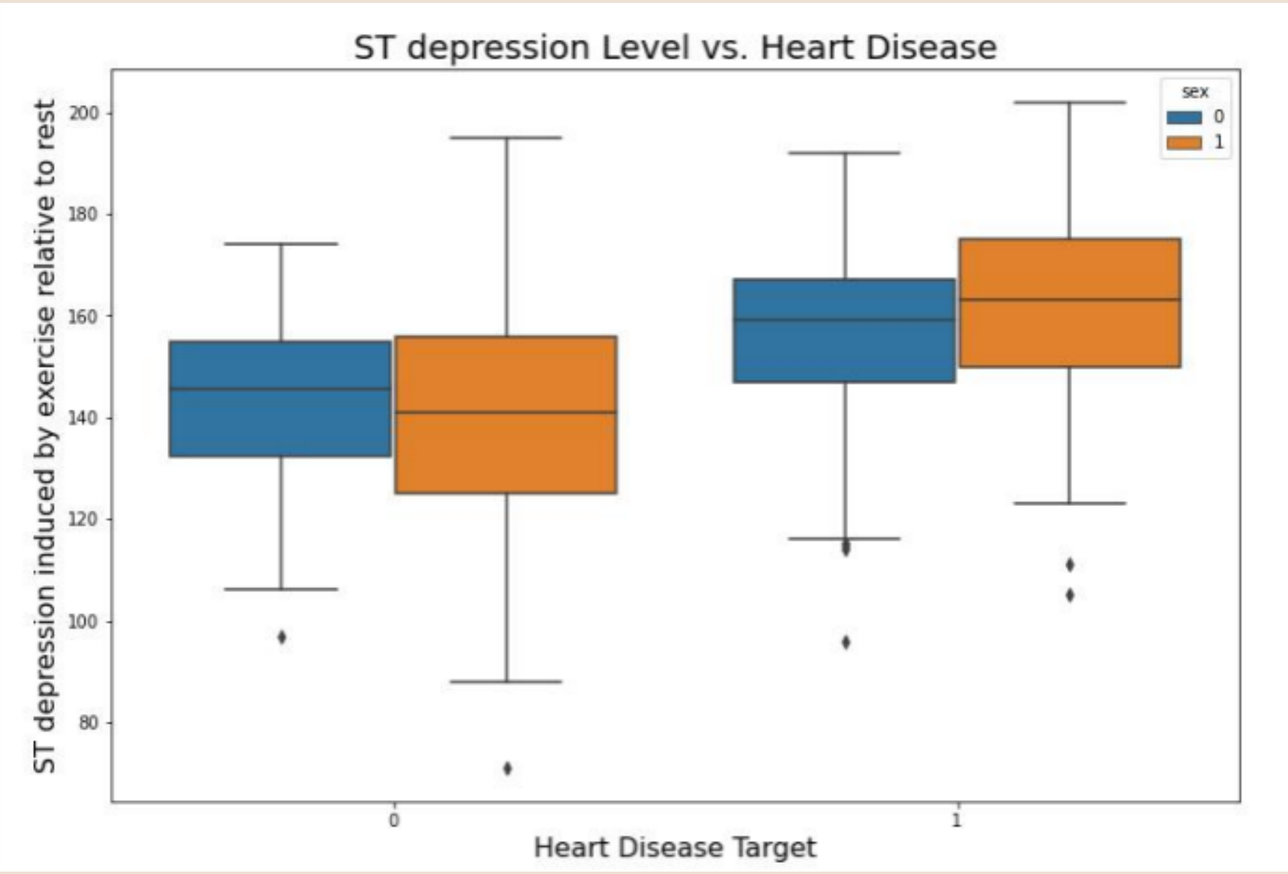

age 1 -0.098 -0.069 0.28 0.21 0.12 -0.12 -0.4 0.097 0.21 -0.17 0.28 0.068 -0.23
sex -0.098 1 -0.049 -0.057 -0.2 0.045 -0.058 -0.044 0.14 0.096 -0.031 0.12 0.21 -0.28
cp -0.069 -0.049 1 0.048 -0.077 0.094 0.044 0.3 -0.39 -0.15 0.12 -0.18 -0.16 0.43
trestbps 0.28 -0.057 0.048 1 0.12 0.18 -0.11 -0.047 0.068 0.19 -0.12 0.1 0.062 -0.14
chol 0.21 -0.2 -0.077 0.12 1 0.013 -0.15 -0.0099 0.067 0.054 -0.004 0.071 0.099 -0.085
fbs 0.12 0.045 0.094 0.18 0.013 1 -0.084 -0.0086 0.026 0.0057 -0.06 0.14 -0.032 -0.028
restecg -0.12 -0.058 0.044 -0.11 -0.15 -0.084 1 0.044 -0.071 -0.059 0.093 -0.072 -0.012 0.14
thalach -0.4 -0.044 0.3 -0.047 -0.0099 -0.0086 0.044 1 -0.38 -0.34 0.39 -0.21 -0.096 0.42
exang 0.097 0.14 -0.39 0.068 0.067 0.026 -0.071 -0.38 1 0.29 -0.26 0.12 0.21 -0.44
oldpeak 0.21 0.096 -0.15 0.19 0.054 0.0057 -0.059 -0.34 0.29 1 -0.58 0.22 0.21 -0.43
slope -0.17 -0.031 0.12 -0.12 -0.004 -0.06 0.093 0.39 -0.26 -0.58 1 -0.08 -0.1 0.35
ca 0.28 0.12 -0.18 0.1 0.071 0.14 -0.072 -0.21 0.12 0.22 -0.08 1 0.15 -0.39
thal 0.068 0.21 -0.16 0.062 0.099 -0.032 -0.012 -0.096 0.21 0.21 -0.1 0.15 1 -0.34
target -0.23 -0.28 0.43 -0.14 -0.085 -0.028 0.14 0.42 -0.44 -0.43 0.35 -0.39 -0.34 1
age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal target
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4

EXPLORATORY DATA ANALYSIS

Violin and Box-plots

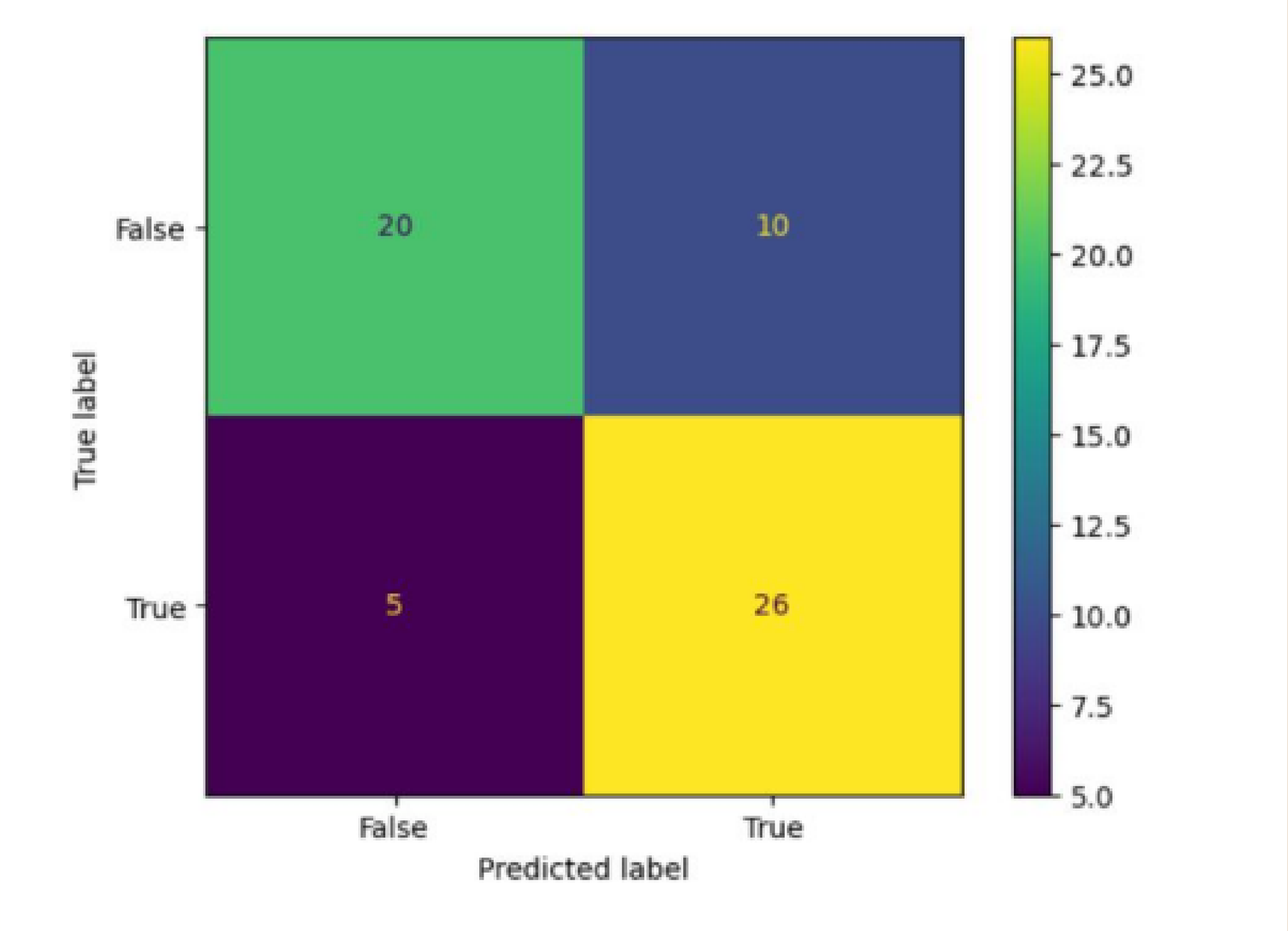
- Box and violin plots are useful for displaying the basic statistics and distribution of data, as well as identifying outliers.
- They can be used to compare the distribution of a variable across different categories.
- Positive and negative patients have different distributions of ST depression levels, with positive patients having a lower median and a greater concentration of data between 0 and 2, while negative patients have a median between 1 and 3.
- There are no major differences between male and female target outcomes, except that males have slightly larger ranges of ST depression.


Thalach Level vs. Heart Disease
sex
0
1
Thalach Level
8
6
4
2
0
−2
0
1
Heart Disease Target


ST depression Level vs. Heart Disease
sex
0
1
ST depression induced by exercise relative to rest
200
180
160
140
120
100
80
0
1
Heart Disease Target

Evaluation Metrics

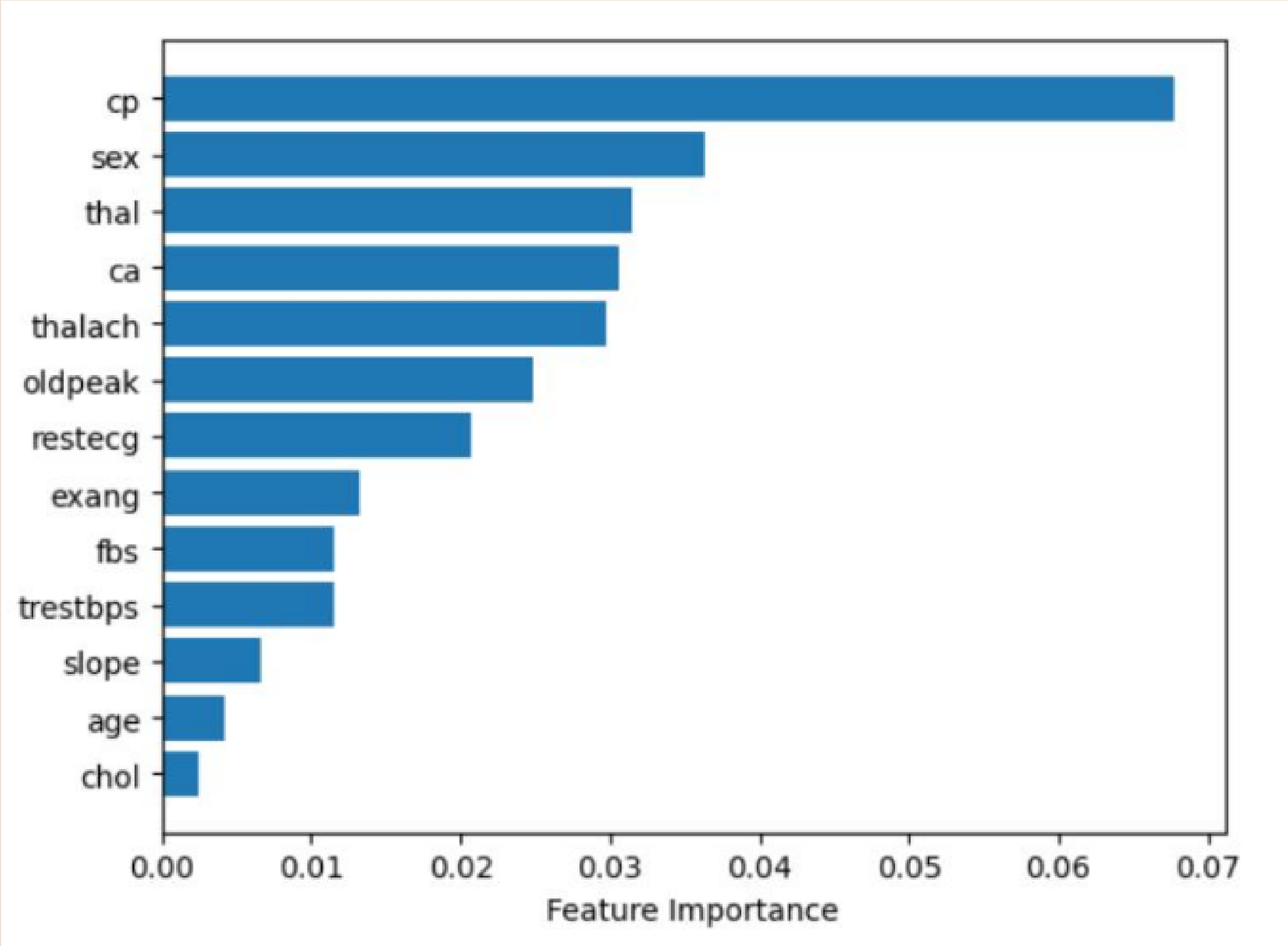
Confusion Matrix



Evaluation Metrics

Feature Importance

Feature Importance provides a score that indicates how helpful each feature was in our model. The higher the Feature Score, the more that feature is used to make key decisions & thus the more important it is



PREDICTIONS

Scenario:

- A patient develops cardiac symptoms & you input his vitals into the Machine Learning Algorithm.
- He is a 20 year old male, with a chest pain value of 2 (atypical angina), with resting blood pressure of 110.
- In addition he has a serum cholestoral of 230 mg/dl.
- He is fasting blood sugar > 120 mg/dl.
- He has a resting electrocardiographic result of 1.
- The patients maximum heart rate achieved is 140.
- Also, he was exercise induced angina.
- His ST depression induced by exercise relative to rest value was 2.2.
- The slope of the peak exercise ST segment is flat.
- He has no major vessels colored by fluoroscopy, and in addition his maximum heart rate achieved is a reversable defect.
- Based on this information, can you classify this patient with Heart Disease?

```
print(model3.predict(sc.transform([[20,1,2,110,230,1,1,140,1,2.2,2,0,2]])))
```

```
[1]
```

Yes! Our machine learning algorithm has classified this patient with Heart Disease. Now we can properly diagnose him, & get him the help he needs to recover. By diagnosing him early, we may prevent worse symptoms from arising later.

CONCLUSION

- The SVM algorithm has an accuracy of 75%, which is considered good but we need to be careful of over-fitting.
- Among the 13 features examined, the top 4 significant features that helped in classifying between positive and negative diagnosis were chest pain type (cp), gender (sex), maximum heart rate achieved (thal), and number of major vessels (ca).
- With the machine learning algorithm, we can now classify patients with heart disease and diagnose them properly. Early detection of these features may prevent worse symptoms from arising later.