

MET Bhujbal Knowledge City, Nashik

Department of Computer Engineering

Academic Year : 2021-22

**Data Mining and Warehousing**

**Mini-Project Report**

On

**Heart Disease Risk Prediction**

Submitted by:

Tanvi A. Nirmal (B-15)

Sakshi S. Thorat (B-17)

Guided by: Prof. Vishal Patil

DEPARTMENT OF COMPUTER ENGINEERING

Academic Year 2021-22

|  |  |
| --- | --- |
| **Contents** | **Page No.** |
| 1. Problem Statement | …1 |
| 2. Introduction | …2 |
| 3. Dataset | …3 |
| 4. Classification Models | …5 |
| 5. Result | …7 |
| 6. Conclusion | …9 |

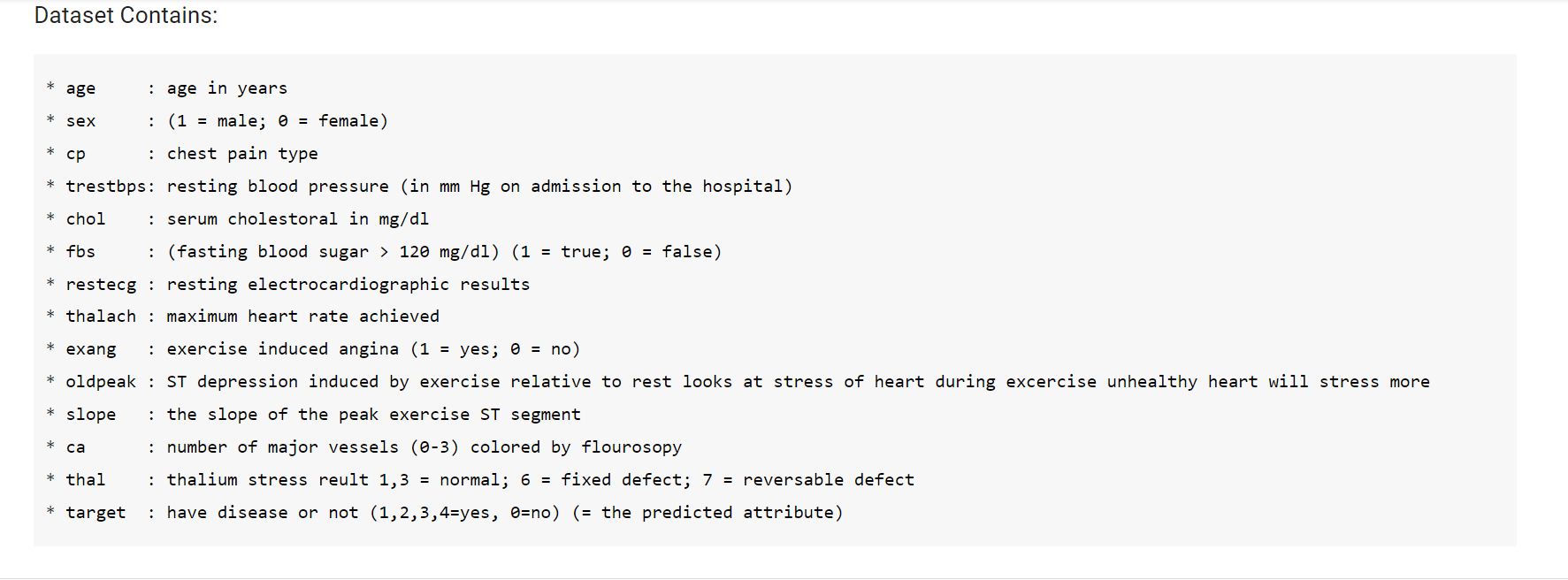
**Problem Statement:**

Consider a labelled dataset belonging to an application domain. Apply suitable data preprocessing steps such as handling of null values, data reduction, and discretization. For prediction of class labels of given data instances, build classifier models using different techniques (minimum 3), analyze the confusion matrix and compare these models. Also apply cross validation while preparing the training and testing datasets.

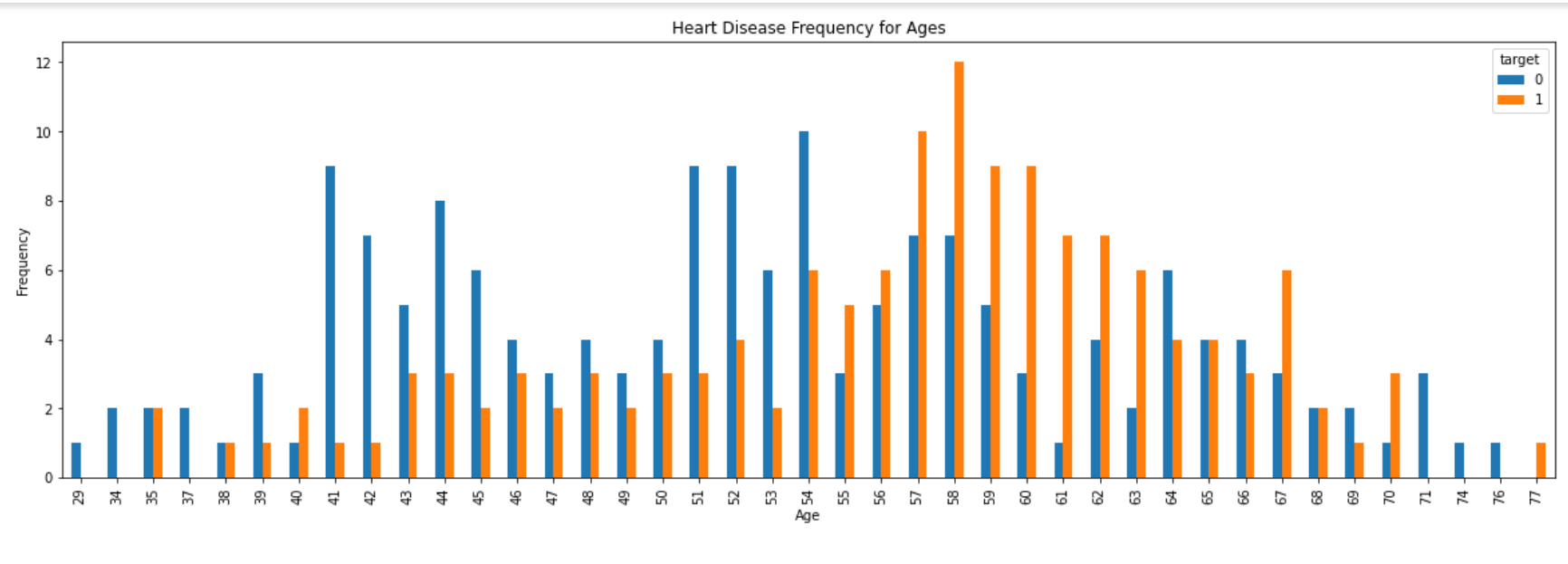
**Introduction:**

It is a well known fact that Heart Diseases are currently the leading cause of death across the globe. The development of a computational system that can predict the presence of heart diseases in patients will significantly reduce the mortality rates and substantially reduce the costs of healthcare. Machine learning is used across many spheres around the world. Especially it is gaining more popularity in the healthcare industry. Machine learning can play an essential role in predicting presence or absence of a critical disease, for an instance, Heart disease, etc. Such information predicted well in advance, can provide important insights to the doctors, who can then carry out the treatments of the patients accordingly and efficiently. In addition to that, heart disease prediction is carried out using different approaches such as Support Vector Classifier, Logistic regression, Naïve Bayes Classifier, K-Nearest Neighbor and Random Forest Classifier.

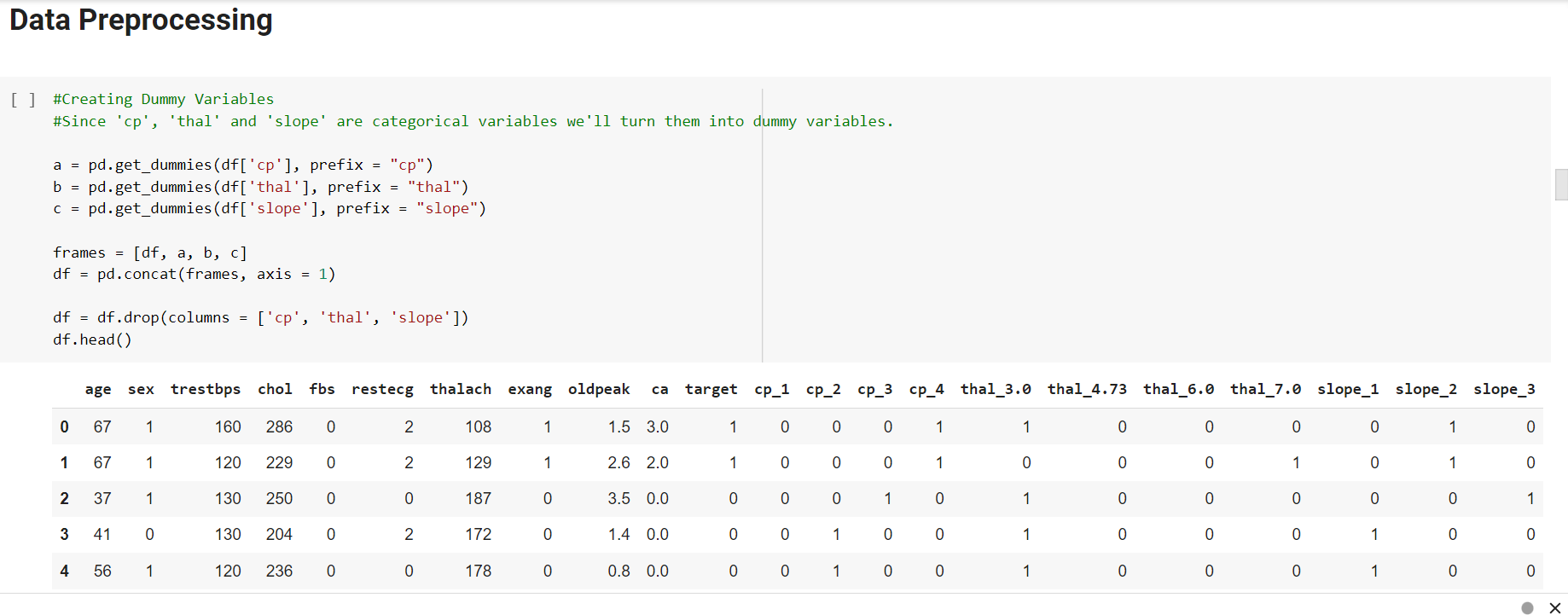
**Dataset:**



Data Exploration:



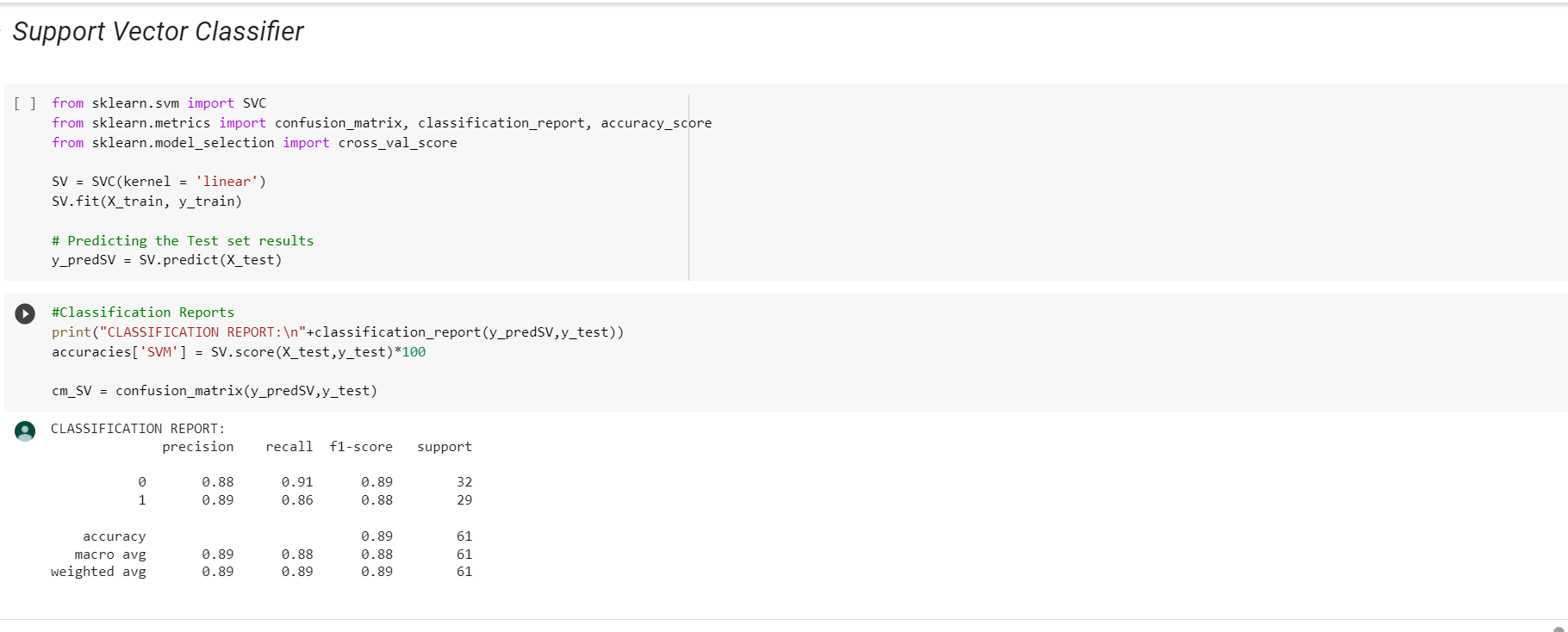
Data Preprocessing:





**Classification Model:**

We have trained the dataset using various classification models to predict the class label.



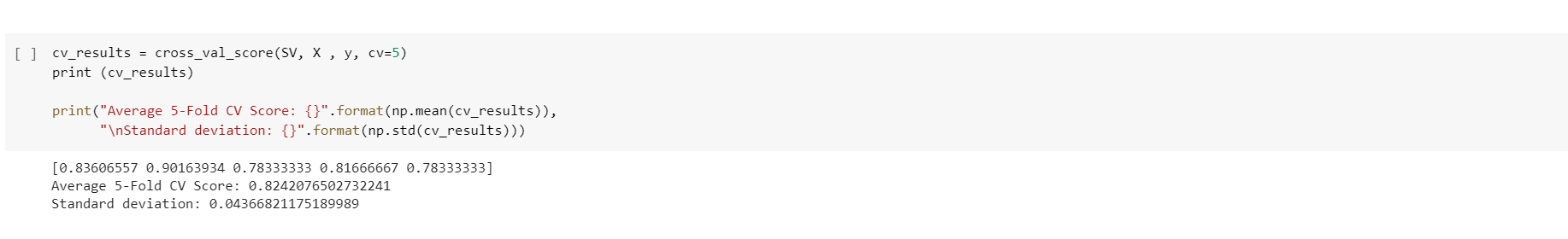


Fig.: Support Vector Classification Model

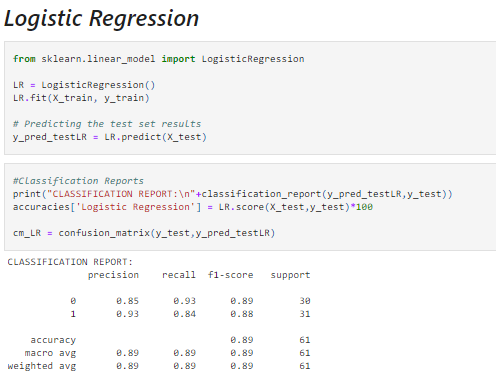
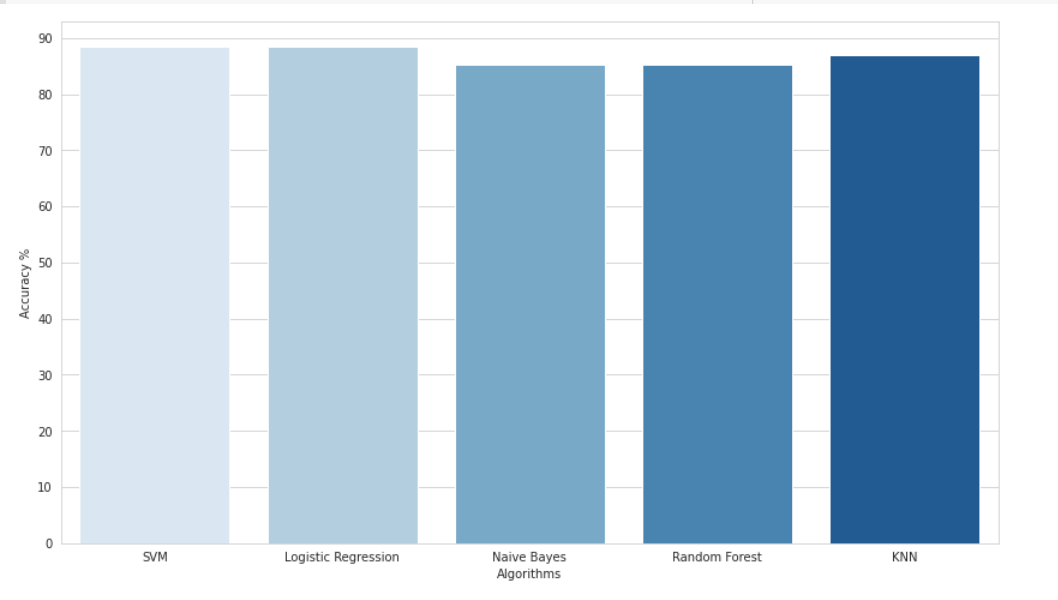


Fig.: Logistic Regression Classification Model

Similarly, we have used other classifiers: Naive Bayes, Random Forest Classifier and K-Nearest Neighbors Method on the dataset.

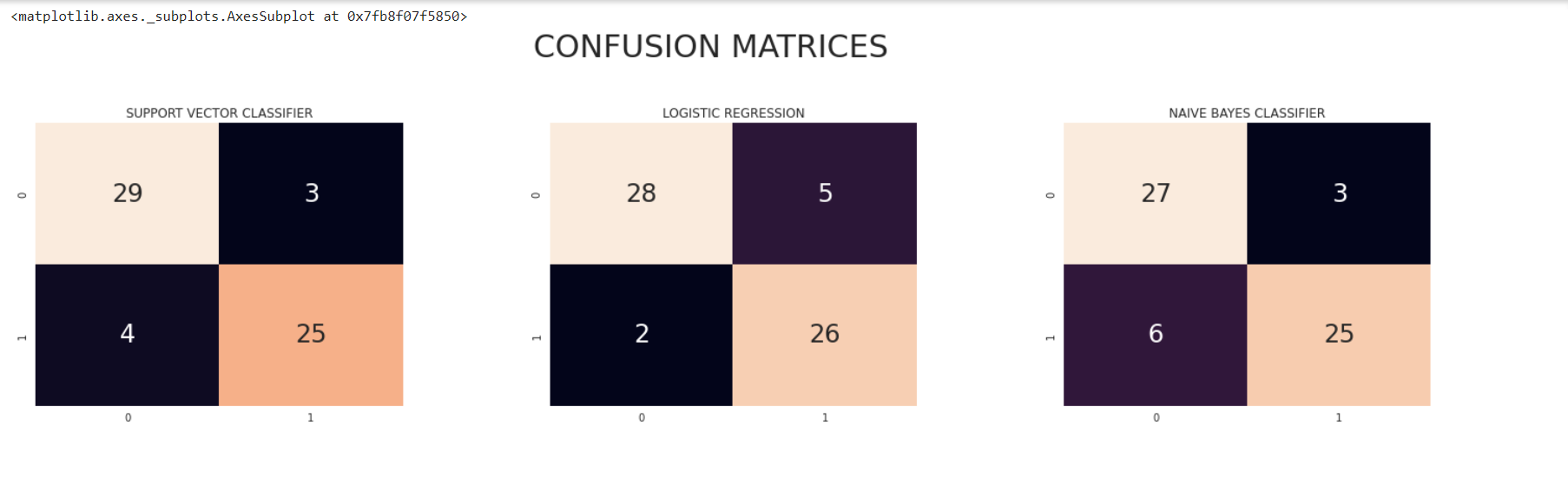
**Result:**

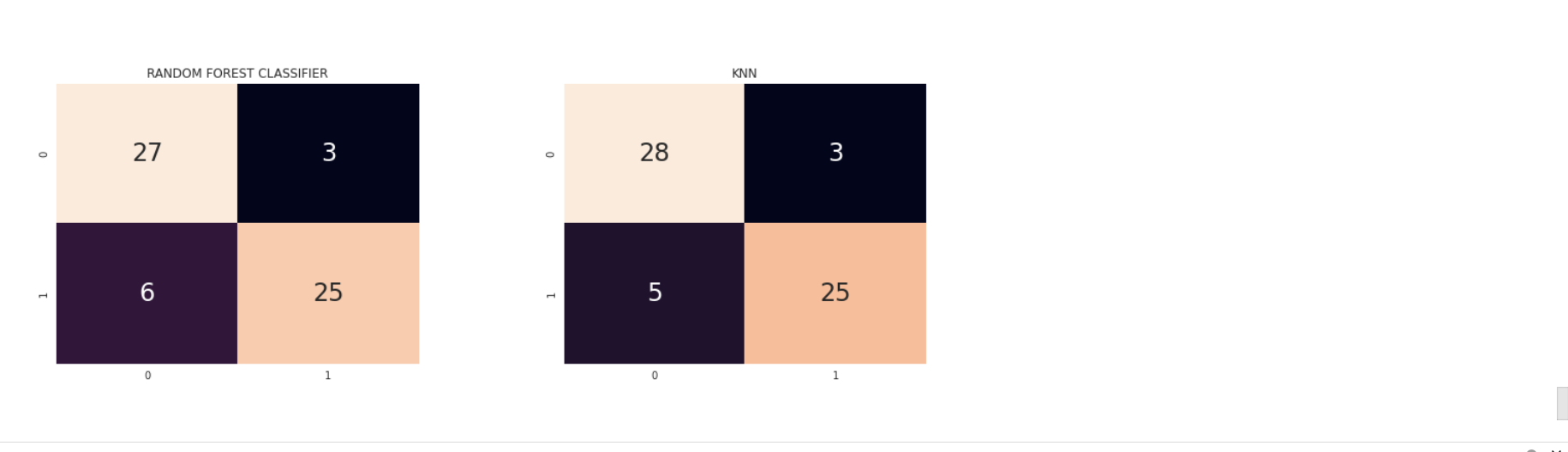
Comparison of the test accuracies of the above models is as follows:



The test accuracy for all the above models is greater than 85%.The highest being of Support Vector Machine and Logistic Regression, which is around 89%.

Comparison of Confusion Matrices:





**Conclusion:**

Thus we have successfully analyzed the Heart Disease dataset and performed data preprocessing steps. We have experimented with various classification models such as Support Vector Classifier, Logistic regression, Naive Bayes and Random Forest Classifier. We have then used these models to test perform prediction on the test dataset.