Ahsanullah University of Science and Technology

Department of Computer Science and Engineering

Program: Bachelor of Science in Computer Science and Engineering

Course No: CSE 4108

Course Title: Artificial Intelligence Lab
Lab Group : C1

Project

Date of Submission: 6 / 3 / 2022

Submitted to:

Mr. Faisal Muhammad Shah
Associate Professor, Department of CSE, AUST.

Mr. Md. Siam Ansary
Lecturer, Department of CSE, AUST.

Submitted by,

Name: Ullash Bhattacharjee
Student ID: 180104103

Name: Bijoy Ranjan Mandol
Student ID: 170104083

Contents Page

Problem Description:

Cardio Vascular Disease is very common to world - wide. CVDs are the leading cause of death globally, taking an estimated 17.9 million lives each year. If a person who has high cholesterol, diabetes, high blood pressure, angina are more likely to have cardio vascular diseases like heart attack , stroke , heart failure in future. So, it will be better if we can predict the person will have cardio vascular diseases or not by using some attributes. We used six models in our project to predict the person will have CVDs or not and showed comparisons between them to know which one performed better.

Dataset Description:

The dataset contains the information of the people on different aspects. This dataset has 999 rows and 13 columns in total. There are 13 features in each row of the dataset. They are : Age , Gender , Exang(exercise induced angina) , weight , trestbps(resting blood pressure) , thalac(maximum heart rate) , cholesterol , fastingBS(Fasting Blood Sugar) , slope , chestpain , resting ecg, thal(thallasemia) and cardio.It is estimated by WHO that 82 percent of people who die of coronary vascular disease are 65 or older. Besides high cholesterol , diabetes, high blood pressure , having angina are more likely to have cardio vascular disease in future. Here the Cardio is the targeted feature in our model. The other twelve features are independent. We used 70 percent of the dataset to train our models and 30 percent of the dataset to test our models.

Description of the models:

We used six models to predict the result of our project. They are as follows:

Model 1 : K - Nearest Neighbor

K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the

dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

Model 2 : Logistic Regression

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.

Model 3 : Naïve Bayes Classifier

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

Model 4 : Support Vector Machine

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new

data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

Model 5 : Decision Tree Classifier

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of **a** dataset, branches represent the decision rules and each leaf node represents the outcome.

In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

Model 6 : Random Forest Classifier

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Result Table:

KNN , Logistic Regression, Naïve Bayes , SVC , Random Forest and Decision Tree , these six models are used to measure the results. To compare the results four performance metric scores are executed. They are the accuracy , precision, recall and f1 – score.

Algorithm	Accuracy	Precision	Recall	F1 – Score
KNN:	**0.77**	**0.73**	**0.84**	**0.78**
Logistic Regression:	**0.73**	**0.72**	**0.73**	**0.73**
Naïve Bayes:	**0.69**	**0.66**	**0.77**	**0.71**
SVC:	**0.77**	**0.75**	**0.81**	**0.78**
Random Forest:	**0.81**	**0.80**	**0.83**	**0.81**
Decision Tree:	**0.74**	**0.74**	**0.72**	**0.73**

From the table, we can see that Random Forest Classifier has the higher accuracy , precision , recall and f1-score between all of them. And the second one is Support Vector Machine.

Discussion:

Now let's see a graph of Actual vs Predict of our Model 1 , KNN model:

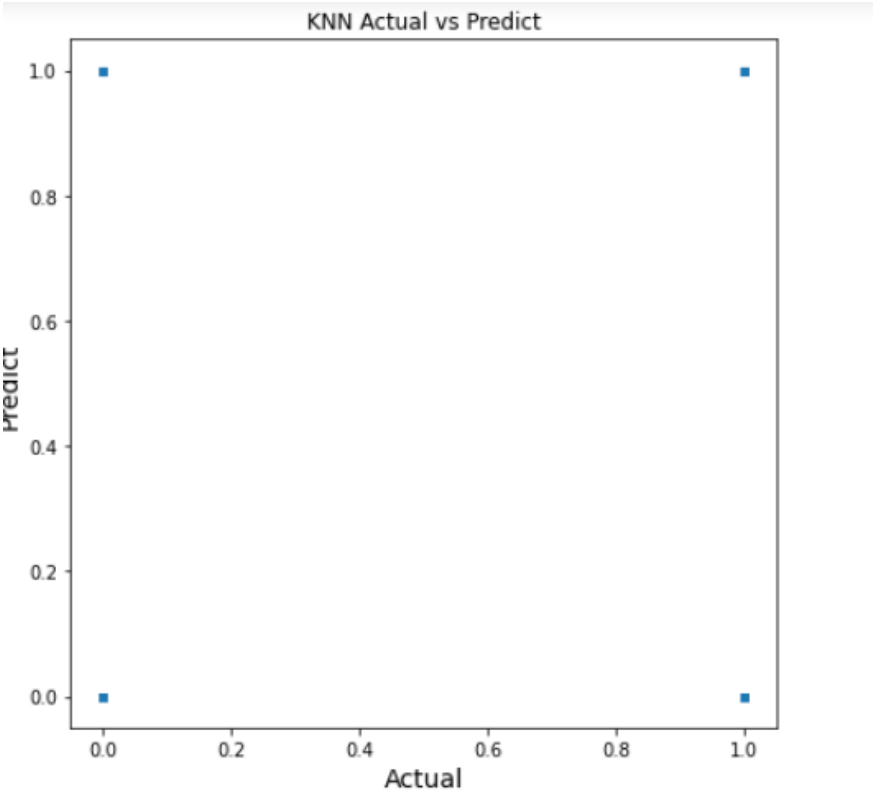


Figure1 : Actual VS Predict graph of model 1

A histogram of actual vs predict of our model KNN:

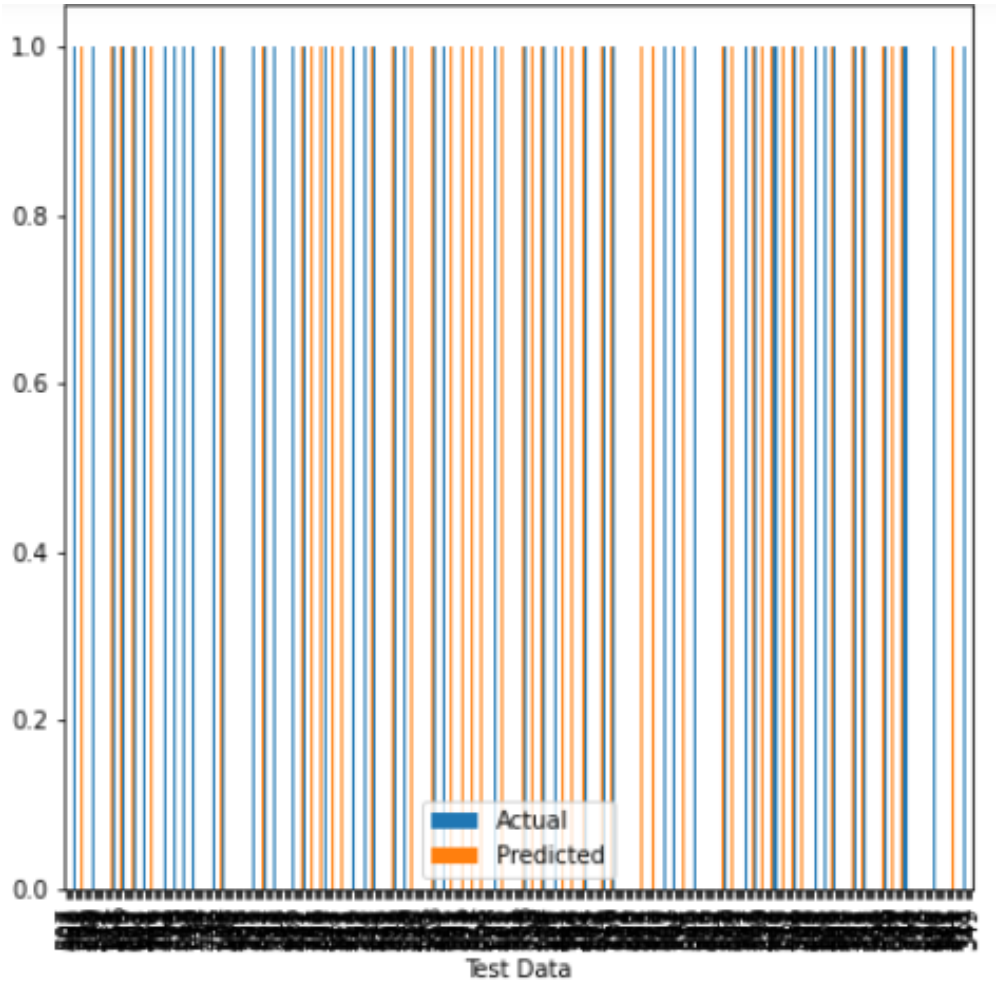
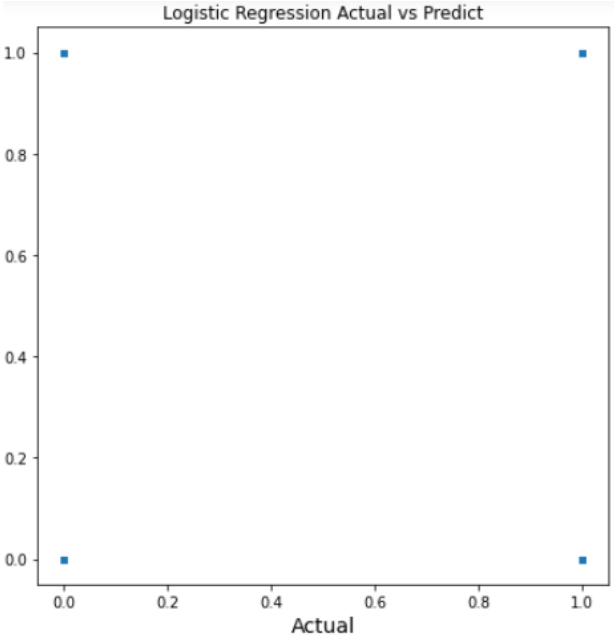
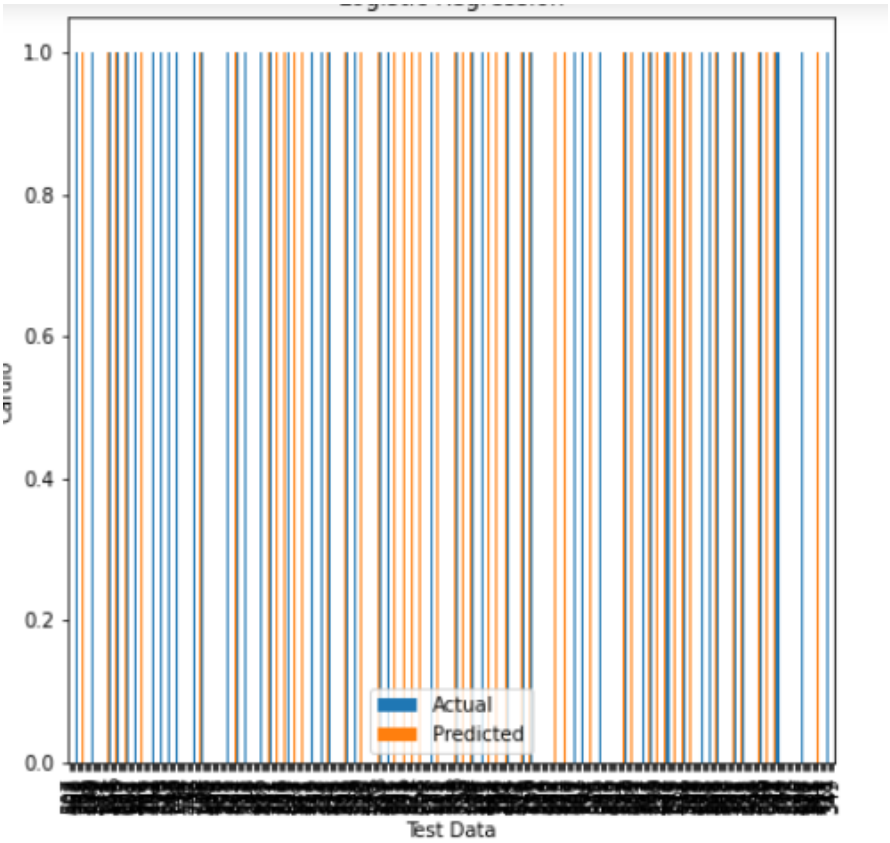


Figure2 : Histogram of Actual VS Predict

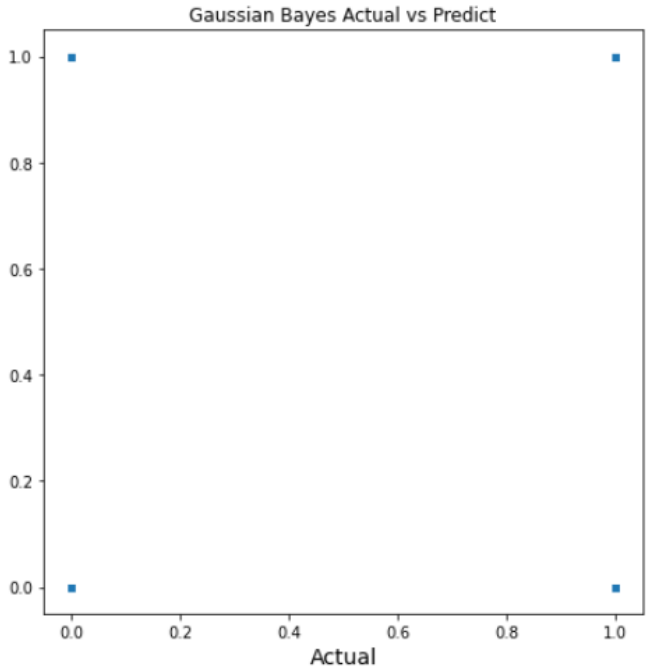
Actual vs Predict of our Model 2 , Logistic Regression model:



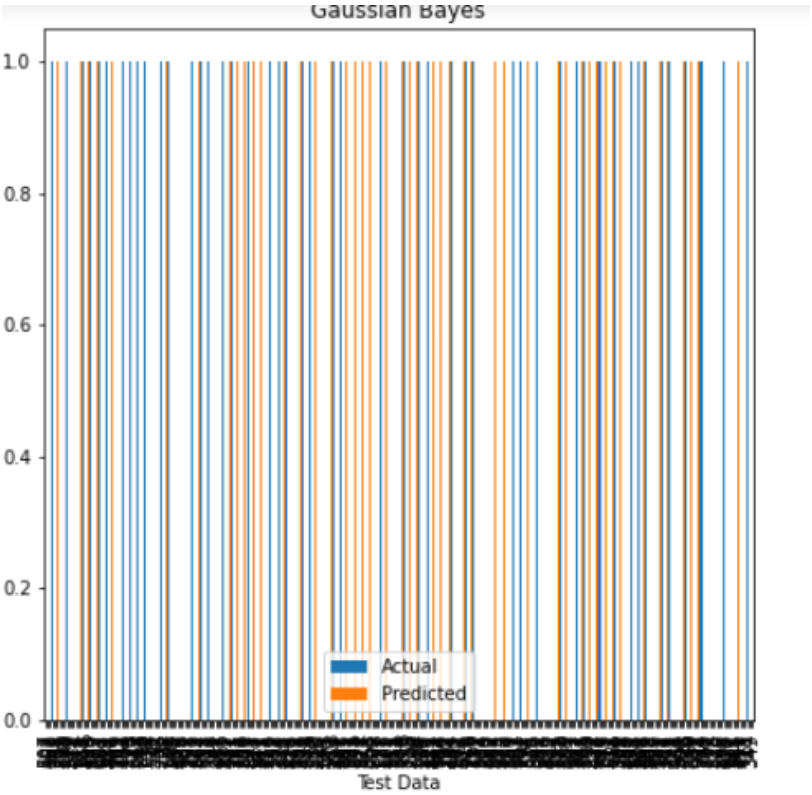
A histogram of actual vs predict of our model Logistic Regression:



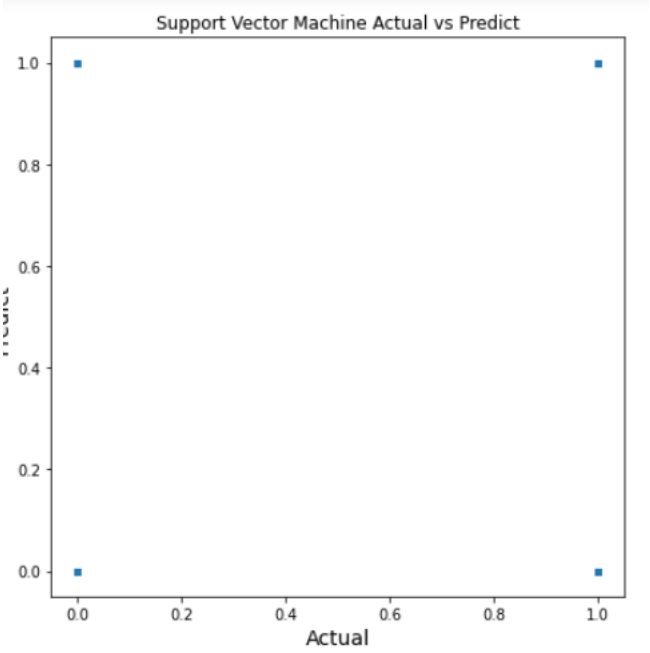
Actual vs Predict of our Model 3 , Gaussian Bayes model:



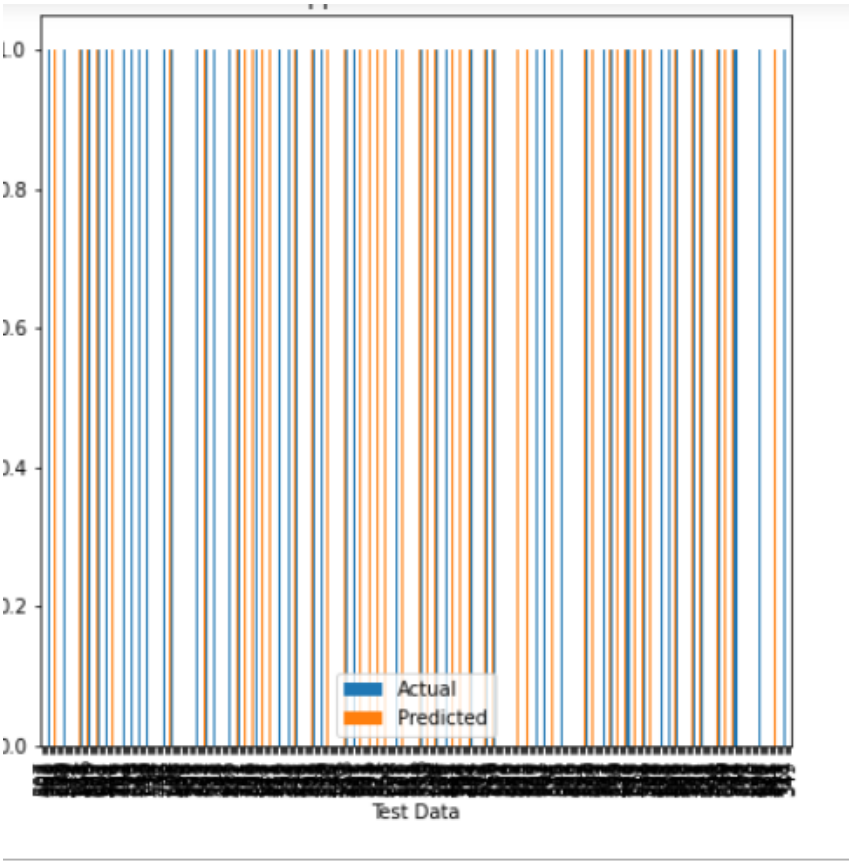
A histogram of actual vs predict of our model Gaussian Bayes:



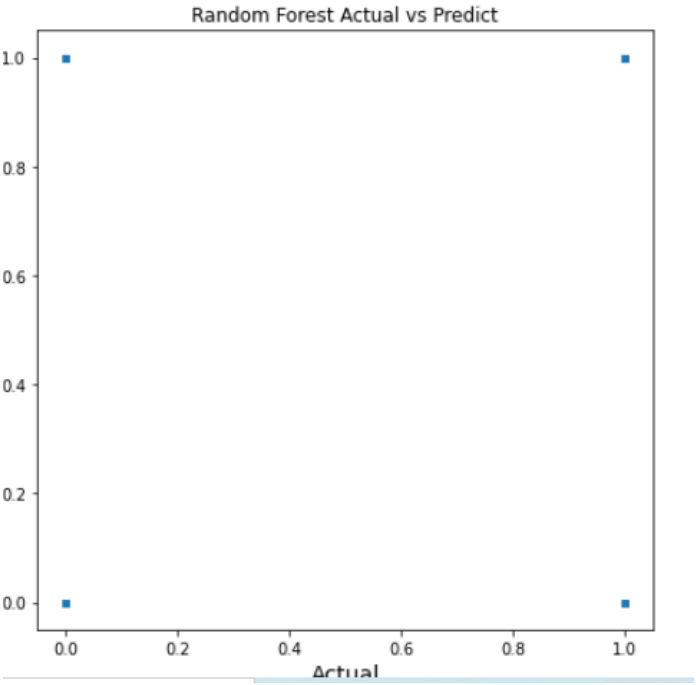
Actual vs Predict of our Model 4 , Support Vector Machine model:



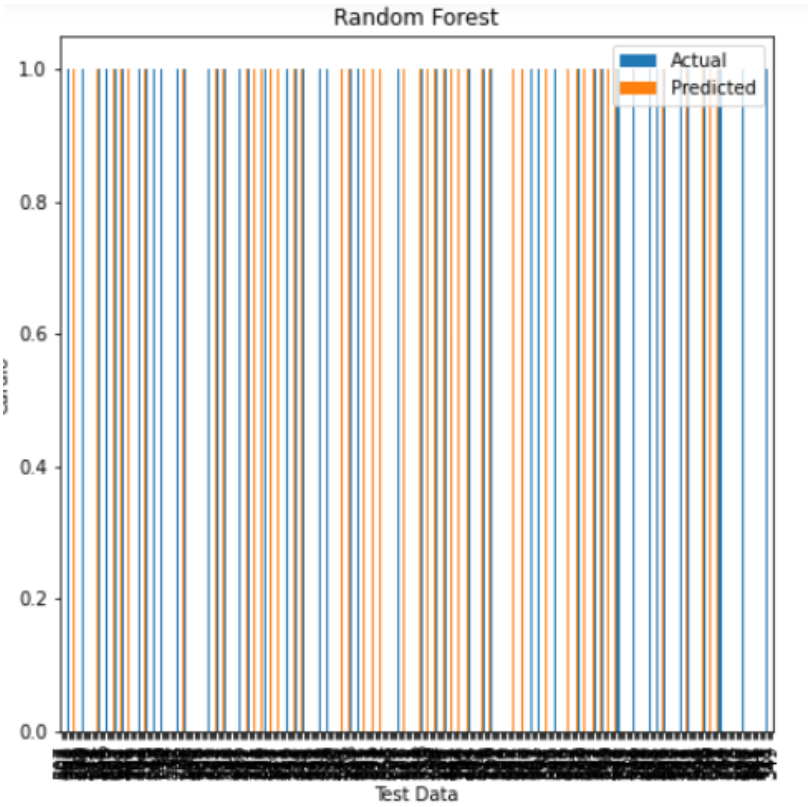
A histogram of actual vs predict of our model SVC:



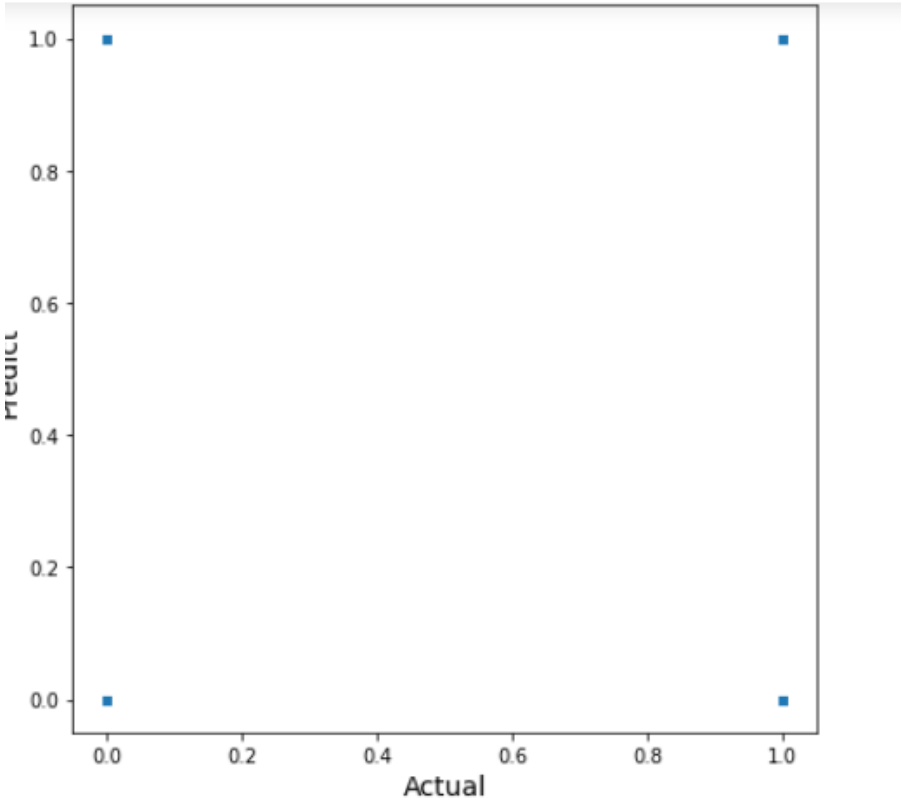
Actual vs Predict of our Model 5 , Random Forest model:



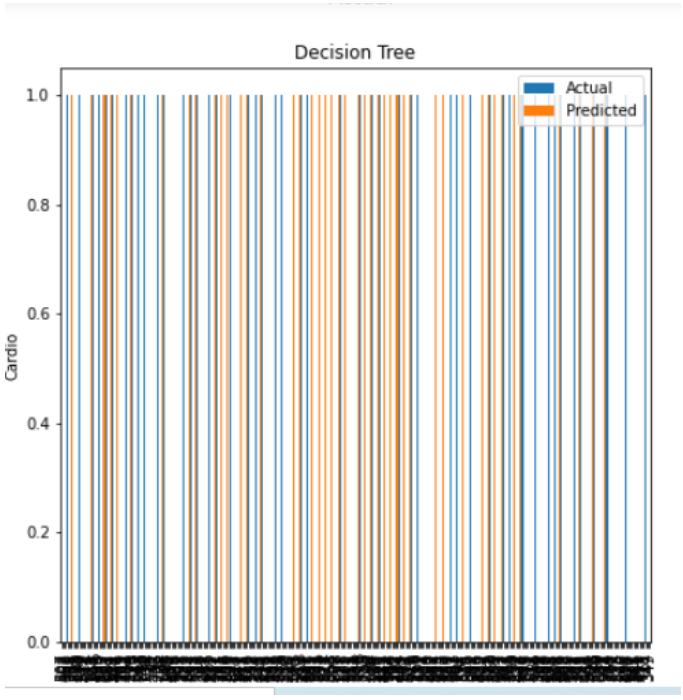
A histogram of actual vs predict of our model Random Forest:



Actual vs Predict of our Model 5 , Decision Tree model:



A histogram of actual vs predict of our model Decision Tree:



From all of the graphs we can see that our fifth model Random Forest performed better than all other models.
Random forest classifier has the accuracy of 81 percent that is the maximum of all others. Besides all the other performance metric like precision , recall , f1 – score random forest classifier is the highest. The second highest one is SVC and KNN with the accuracy of 77 percent. Decision Tree Classifier has the accuracy of 74 percent.
Between DT and Random Forest models , Random Forest performed better because Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset whereas Decision tree generates only one tree. A small change in the data can lead to a large change in the structure of the optimal decision tree. The greater number of trees in the random forest leads to higher accuracy and prevents the problem of overfitting.

SVM model works also relatively better because there is a clear margin of separation between classes. And it is also more effective in high dimensional spaces.

The KNN model also works well same as SVM. Because our dataset is roughly balanced. However with large data , the prediction stage might be slow.

The logistic Regression Model works average with the accuracy of 73 percent. Main limitation of logistic regression is the assumption of linearity between the dependent variable and independent variables. In the real world , the data is linearly separable.

The worst predicted model in our project is Naïve Bayes Classifier. We got the accuracy only 69 percent which is very less among all. We used Gaussian Bayes to predict our model because this classifier is employed when the predictor values are continuous and are expected to follow a Gaussian distribution. Naïve Bayes Classifier is basically used for text classification problem that includes a high dimensional training dataset.
So , finally after realizing all the graphs and result table values we can say that our model 5 Random Forest performed better than the all other models.

Benchmarking:

ID	**Project Work Percentage**
180104103	**65 %**

The first dataset is collected by ID , found in the documentation of the dataset. And also this ID made data 501 to 632 data in our dataset.
The Random Forest , Support Vector , Decision Tree and K Nearest Neighbor Model is developed by this ID. The Data Preprocessing Part and measurement of the performance metrics is also done by this ID.
Report and Documentation of the dataset is done by this ID.

170104083	**35 %**

The last dataset is collected by ID , found in the documentation of the dataset. And also this ID made data 305 to 500 data in our dataset.
Besides the logistic regression model , naïve bayes classifer model is developed by this ID.
Documentation of the dataset is done by this ID.