Project 1 - Analysis Methods to predict the onset of diabetes in PIMA Indians

Santanu Mukherjee

Introduction

The PIMA Indians dataset is a selection of data form a larger database and originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether a patient has diabetes, based on certain diagnostic measurements that is included in the dataset. All patients here are females and at least 21 years old of PIMA Indian heritage.

The datasets consists of several medical predictor variables and one target variable, **Outcome**. Predictor variables includes the number of pregnancies the patient has had, Glucose level, Blood Pressure data, Skin Thickness, their BMI, insulin level, and age. The PIMA .csv file was downloaded and saved in a data frame. This dataset has 768 observations (rows) and 9 variables (columns). The response variable is "Outcome" and the other 8 are predictor variables.

Objectives

1. Independent data analysis to understand the nature of data
2. Data pre-processing for model analysis
3. Modelling the data using one or two methods and predict the onset of diabetes
4. Conclusion regarding the results related to prediction accuracy and errors for the models used.

Data Structures

As mentioned earlier, the **PIMA** data is in a .csv file and we have pulled the data and stored in a data frame. The response variable is *Outcome* and it has a value of 0 or 1.
The variables $Diabetes$ and BMI are of $numeric$ type. The other variables are all of type int.

Methods

Initially, exploratory data analysis (EDA) has been performed to understand the nature of data. In this process, several charts are used to see the relationship between the predictors and the response. In the entire dataset, the outcome of 500/768 (65%) of the patients are negative, meaning they are predicted not to have diabetes.

The EDA also showed that there is no significant correlation between predictors and so there is no reason to ignore any predictor while performing modelling.

The other significant area where data pre-processing was required was to identify, analyze and perform imputation of missing values. It has been found out that there are in total 763 missing values. The process of imputation was carried out to take care of the missing values and the method that was used for the imputation was **pmm (predictive mean matching)**.

Statistical Analysis

The below diagram shows the correlation between the predictors and it is evident that there is no STRONG correlation between the predictors. This signifies that there is no need to drop any variable while modelling the data.

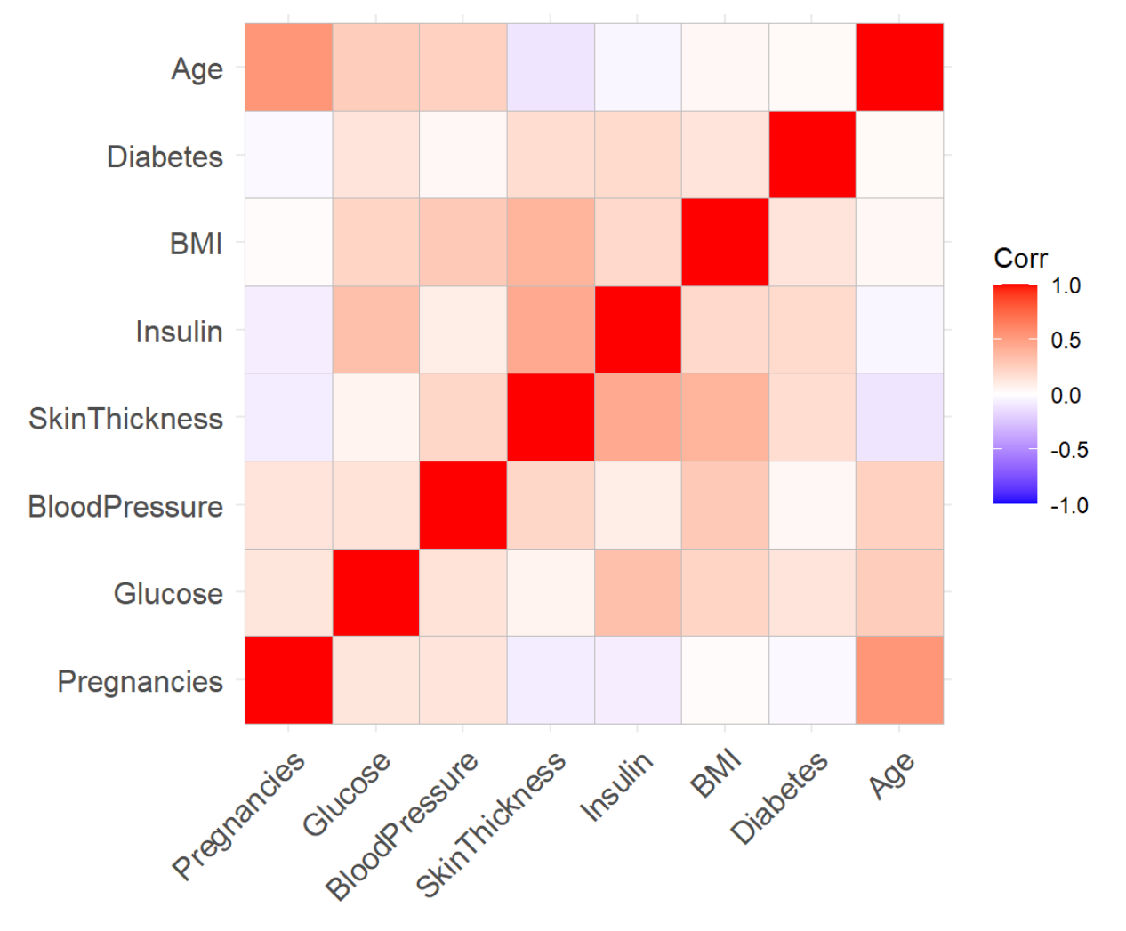

Figure 1: Correlation between Predictors

Data pre-processing

The below diagram shows that there are missing data points in the dataset. So, pre-processing of the data is very important. Data imputation is used to help solve for the missing values. There are 763 missing data points in this $PIMA$ dataset.

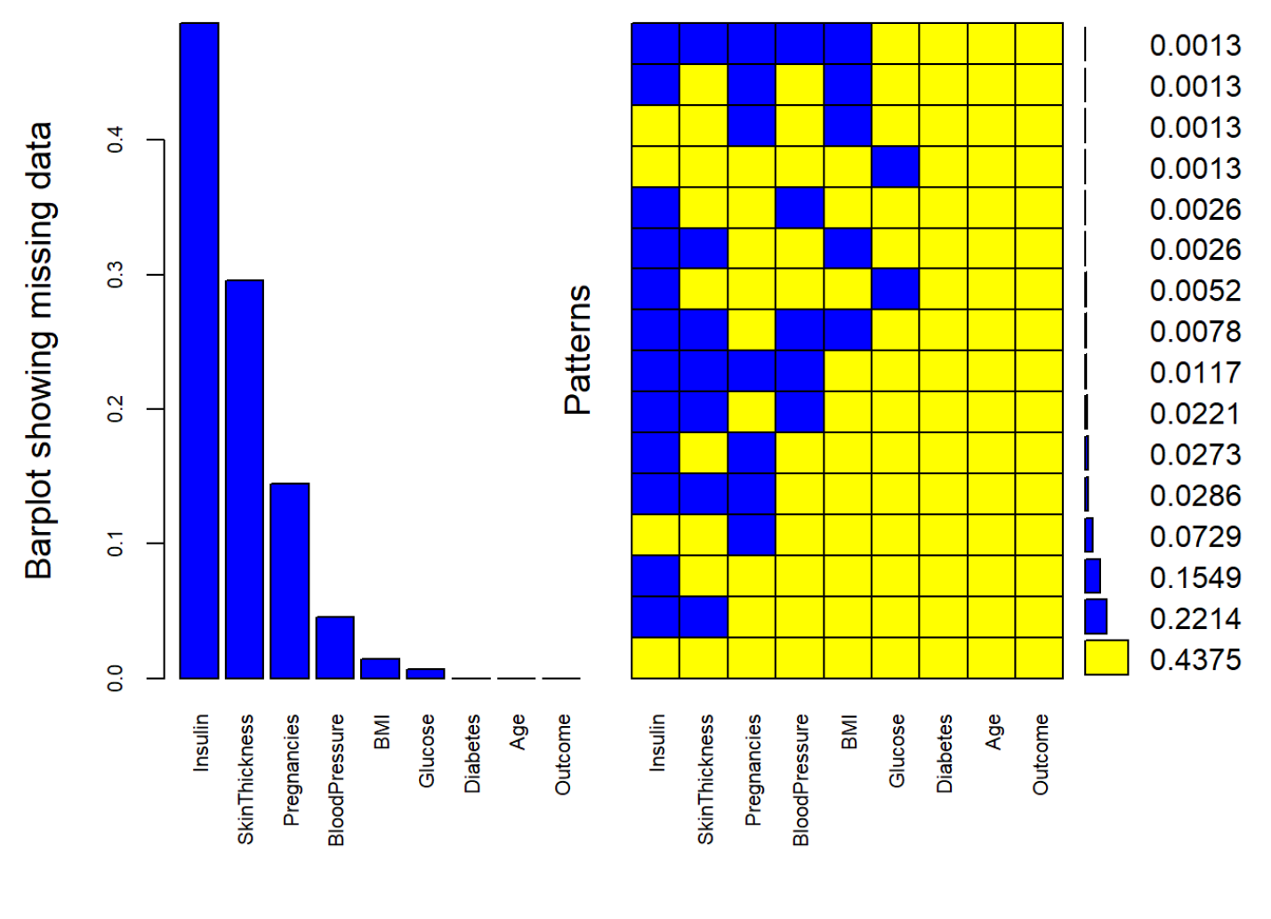

Figure 2: Missing Value Data and Patterns

Imputation of data using the pmm (predictive mean matching) method

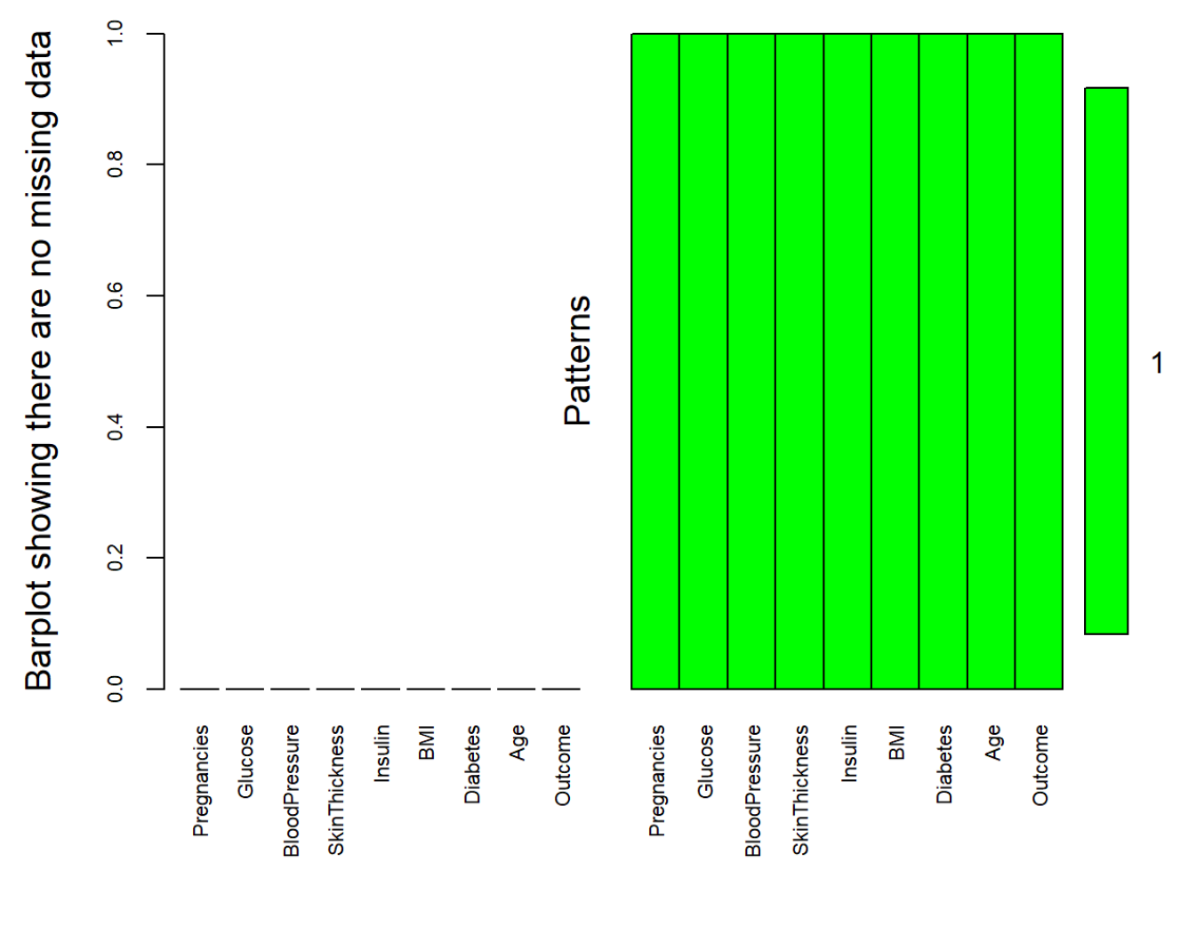

Figure 3: Data Imputation Complete - No Missing values

As you see here , there are no missing values in the data after imputation of the data.

Modelling of the data: Logistic Regression

Logistic Regression was performed with 80% (training data) and 20% (test data). The accuracy of the prediction of the onset of diabetes that this model gave was **0.844**

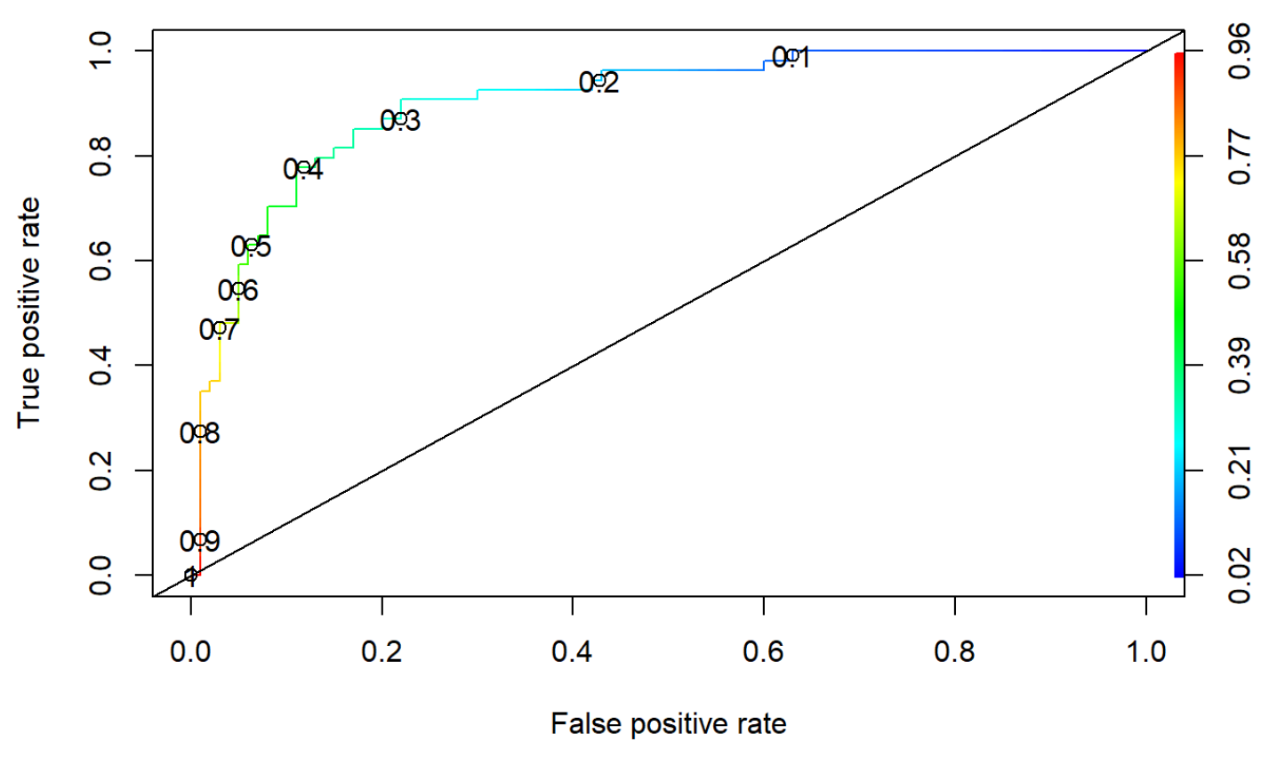

Figure 4: ROC Curve for Logistic Regression

SVM (Linear) & SVM (Radial Kernel)

Built SVM with Linear and Radial kernel on the same dataset (80% training and 20% test dataset).

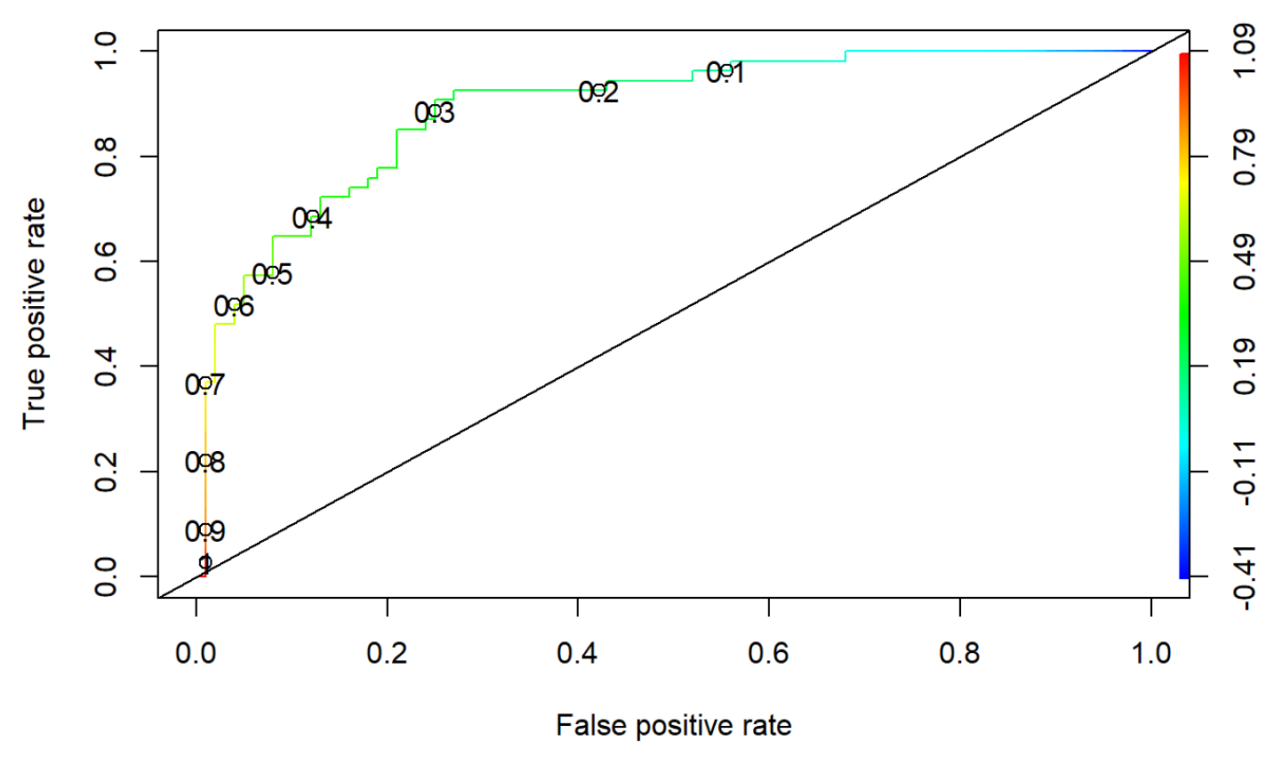

Figure 5: ROC Curve for SVM - Linear

Accuracy of the prediction of the onset of diabetes for SVM Linear is **0.811**.

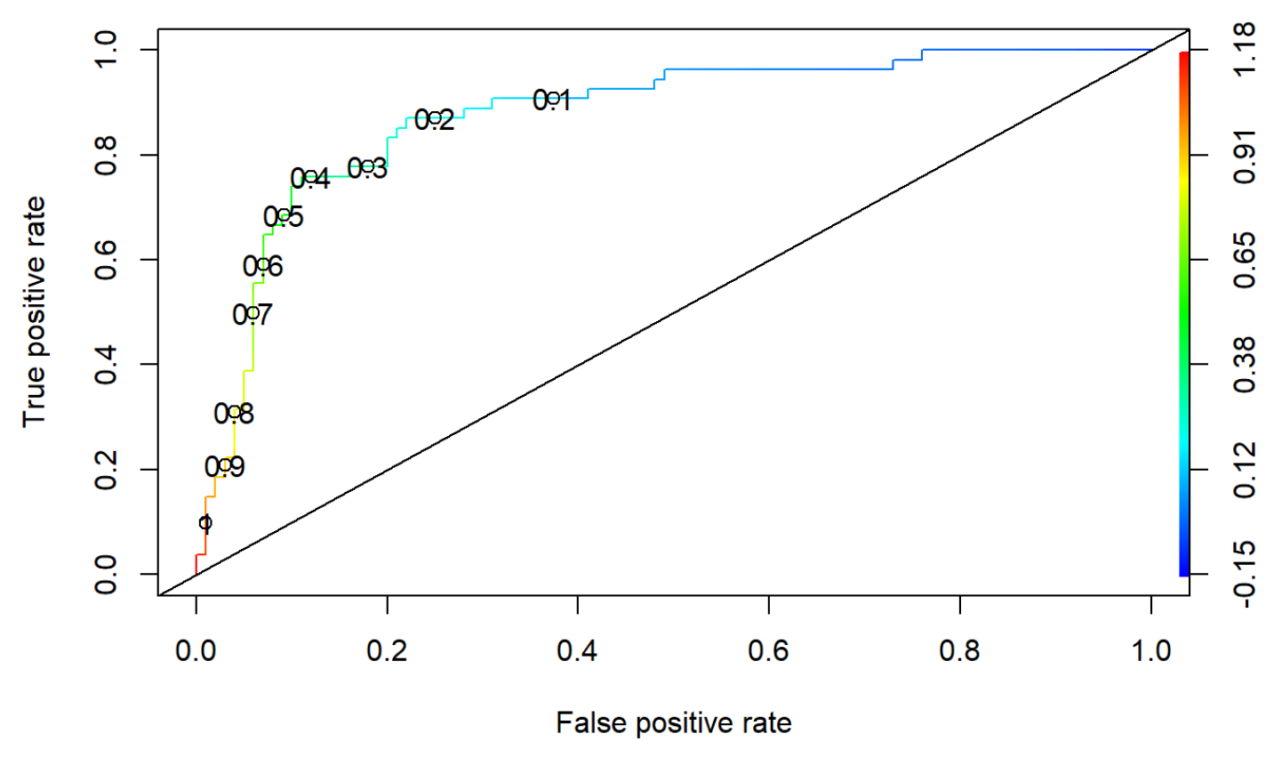

Figure 6: ROC Curve for SVM - Radial

Accuracy of the prediction of the onset of diabetes for SVM Radial is **0.824**.

SVM - Radial Kernel with tuning

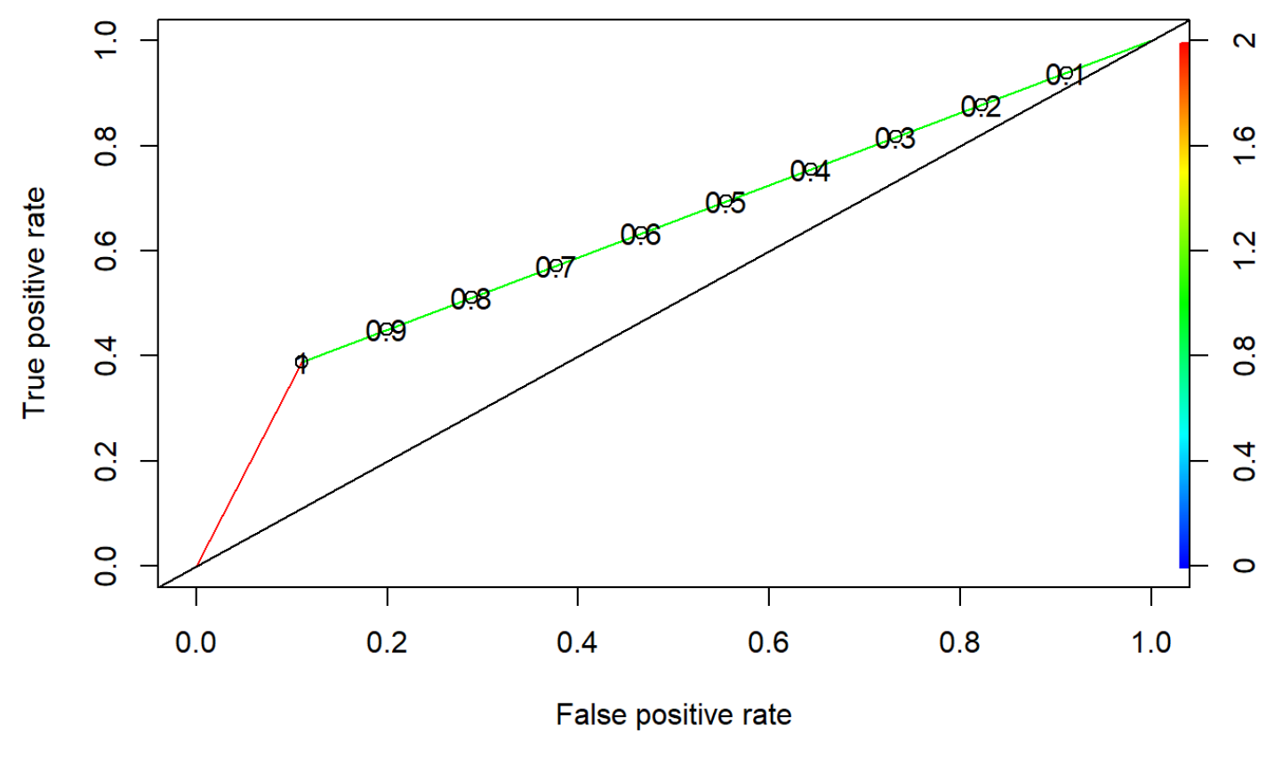

Figure 7: ROC Curve for SVM - Radial Tuning

Accuracy of the prediction of the onset of diabetes for SVM Radial is **0.694**.

Results

The below table provides a comparison of the results of the different parameters that the models are being evaluated for.

Table 1: Comparison of Model Statistics

	Accuracy	Root MSE	Sensitivity	Specificity
Logistic Regression	0.844	1.781	0.830	0.833
SVM - Linear	0.811	0.367	0.807	0.800
SVM - Radial	0.824	0.354	0.845	0.841
SVM-Radial Tune	0.694	0.433	0.726	0.594

Conclusion

In conclusion, it can be said that based on the current data split for this problem, Logistic Regression is a better model from Accuracy of prediction standpoint. But if we consider a model's performance based on lowest Root MSE, sensitivity (true positive rate) and specificity (true negative rate), then SVM- Radial (kernel) is the better model.

Out of the 4 models, SVM Radial Tune is the worst performing model.

One needs to keep in mind that as and when the data mix changes for training and test data, there can be different models that are suitable based on the data. In other words, there is no reason to assume that there exits an absolute correct model. In majority of the cases, a model's performance depends on the business needs for that organization and it is of utmost importance to understand the business value of any model generated.

References

Source : PIMA Indian dataset (https://www.kaggle.com/datasets/uciml/pima
indians-diabetes-database?
select=diabetes.csv)