Predicting occurrence of Diabetes in PIMA Indian Women

Abhinav Srinivasan, Edward Rao and Sathish Sakthivelan

August 4, 2021

Introduction

The data set we've chosen is a diabetes-related data set that was acquired from https://data.world/data-society/PIMA-indians-diabetes-database. The PIMA Indian Diabetes Data set, originally from the National Institute of Diabetes and Digestive and Kidney Diseases, contains information of 768 women under the age of 21 from a population of PIMA Indian heritage near Phoenix, Arizona, USA. The objective of the data set is to diagnostically predict whether or not a 21 year old female has diabetes based on the diagnostic measurements found in the data set.

Background

Diabetes is a chronic condition in which the blood sugar levels are much higher than normal for a prolonged duration of time after carbohydrate consumption. The root cause of this chronic condition is that the body do not produce Insulin or develops a resistance to insulin, a hormone that breaks carbohydrates into glucose. Diabetes is categorically classified into two types: Type 1 and Type 2.

- Type 1 represents the condition where the pancreatic beta cells do not produce any insulin needed and requires the patient to inject insulin manually.
- Type 2 represents the condition where the insulin develops resistance to insulin sensitivity this leading to increased Blood Glucose level in the body.

Diabetes when diagnosed in women is even more harder than that of on men as it can affect both the mother and their unborn children during pregnancy. Also, Diabetic women have higher risk of having a heart attack, miscarriages. The unborn babies could be born with birth defects. Creating a statistical learning model will help us predict whether a particular PIMA Indian female under 21 years old will develop diabetes based on the predictors provided. If we are able to predict the likely occurrence of diabetes before its onset, We could very likely reduce the risk of diabetes in these females and design a health plan such that they could mitigate those risks and extend their diabetes-free life.

Data

The overview of the sample data shows that we have 9 columns of which one seems to be the Response variable, while others seems to be Predictor variables.

Table 1: Sample of PIMA Indian Diabetes data set

Pregnancies	Glucose	BP	Skin	Insulin	BMI	DPF	Age	Outcome
6	148	72	35	0	34	0.63	50	1
1	85	66	29	0	27	0.35	31	0
8	183	64	0	0	23	0.67	32	1
1	89	66	23	94	28	0.17	21	0
0	137	40	35	168	43	2.29	33	1
5	116	74	0	0	26	0.20	30	0

Data Description

The below table provides the detailed look at all the variables present in the data set. The description and the data type of each variable is presented along with the variable's names. Based on the available information, we can clearly see that as we expected, there is one Response variable and the eight predictor variables.

Table 2: Variable Description of PIMA Indian Diabetes data set

Variable Name	variable Type	Variable Description
Pregnancies	integer	Number of times a participant is pregnant
Glucose	integer	Plasma glucose concentration a 2hr in an oral glucose tolerance test
BP	integer	It consists of Diastolic blood pressure (when blood exerts into arteries between heart)(mmHg)
Skin	integer	Triceps skinfold thickness (mm). It concluded by the collagen content
Insulin	integer	2-hour serum insulin (µU/ml)
BMI	numeric	Body Mass Index in kg/m^2
DPF	numeric	Pedigree Diabetes Function: The function that represents how likely they are to get the disease by extrapolating from their ancestor's history. An appealing attributed used in diabetes prognosis
Age	integer	Age of participants
Outcome	integer	Diabetes class variable, '1' represent the patient is diabetic and '0' represent patient is not diabetic

Exploratory Data Analysis

Data Introduction

The data set consists of 768 entries and 9 variables with all predictors are consisted entirely of continuous variables. As seen below, the data set appeared to be clean with all rows are complete, with no missing values and consistent inputs, essentially suggesting that only a minimal data preprocessing may be needed.

Table 3: Basic Statistics in Raw Count

rows	768
columns	9
discrete_columns	0
continuous_columns	9
all_missing_columns	0
total_missing_values	0
complete_rows	768
total_observations	6912
memory_usage	36600

Missing Values

Upon further investigation of the data, it was noted that, though we have all the rows filled and no missing values, the data in certain columns represent abnormal values for those biological measurements. The biological measurements like skin thickness, Glucose, BP, Insulin, BMI cannot be zero, while Pregnancies and Outcome can be 0. Thus, we can conclude that the missing values in our data set is represented by zeros.

By creating a subset with the columns containing abnormal values and quantifying the missing values, We found that we have a staggeringly higher observation of missing data in our data set.

[1] 376

Table 4: Null Values by Predictor Variables

	x
Glucose	5
BP	35
Skin	227
Insulin	374
BMI	11
DPF	0
Age	0

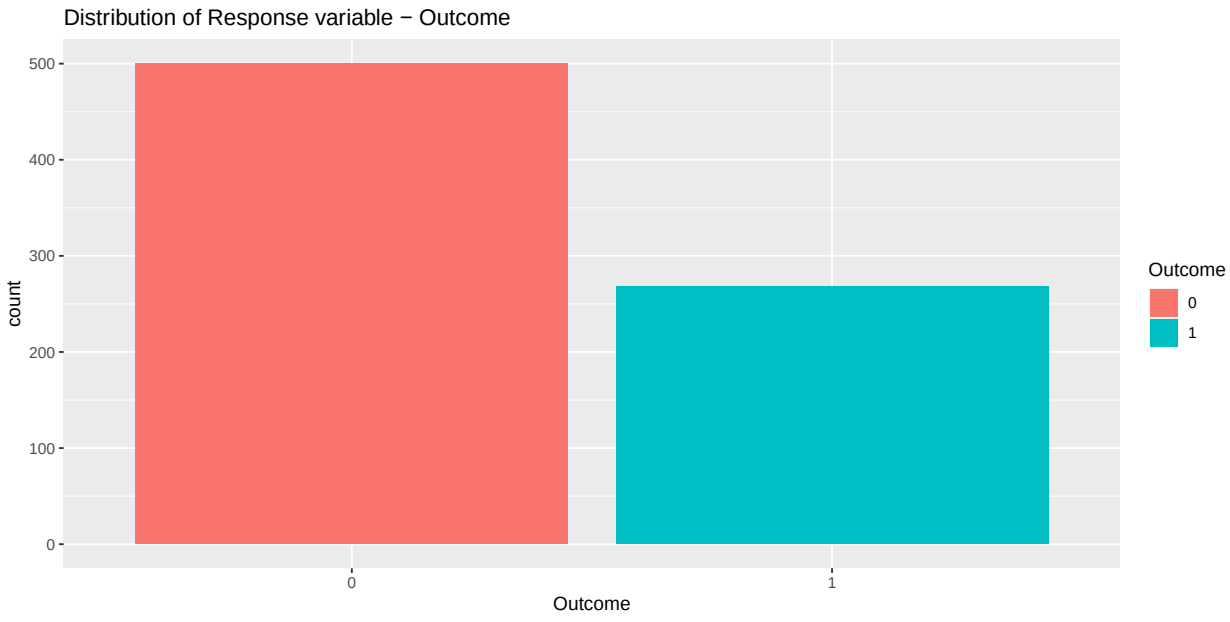
Multivariate Imputation by Chained Equations

Due a higher % of missing values in the data, eliminating the missing values will lead to a significant data loss. This loss will in turn result in a poor prediction model. We have chosen to perform a Multivariate Imputation by Chained Equations to impute the null values in the data set. For the MICE imputation we decided to consider m = 30 to impute the missing values using Weighted predictive mean matching method.

Once the MICE imputation is completed we reran the initial check to validate if all the null values have be handled, which we were able to confirm.

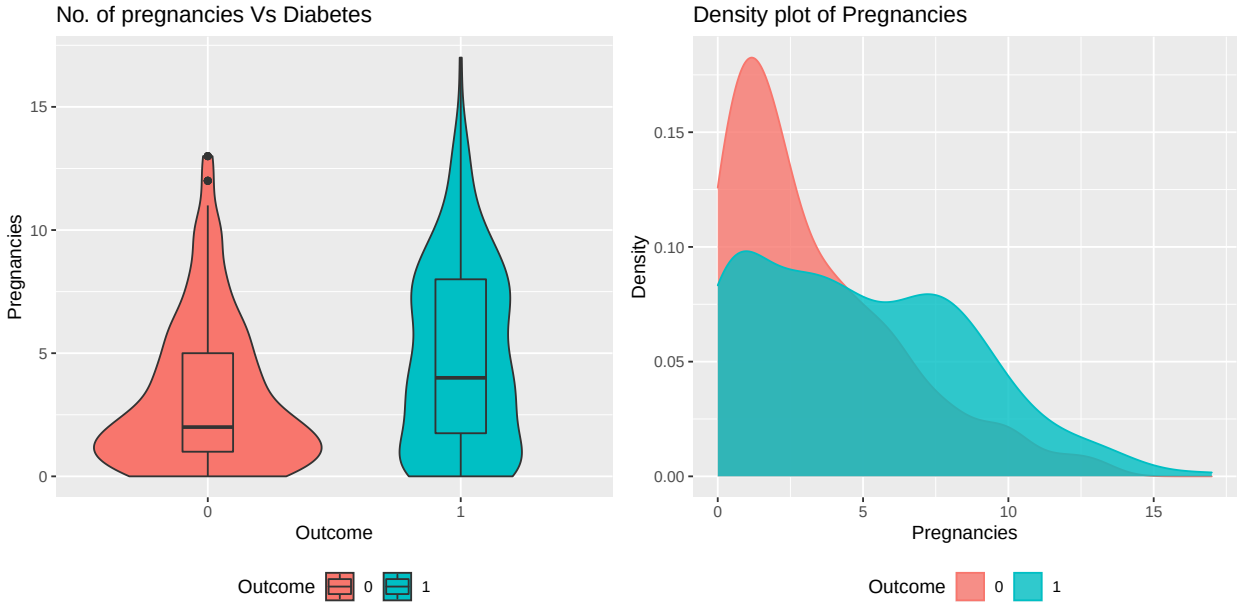
Variables

Response Variable - Outcome



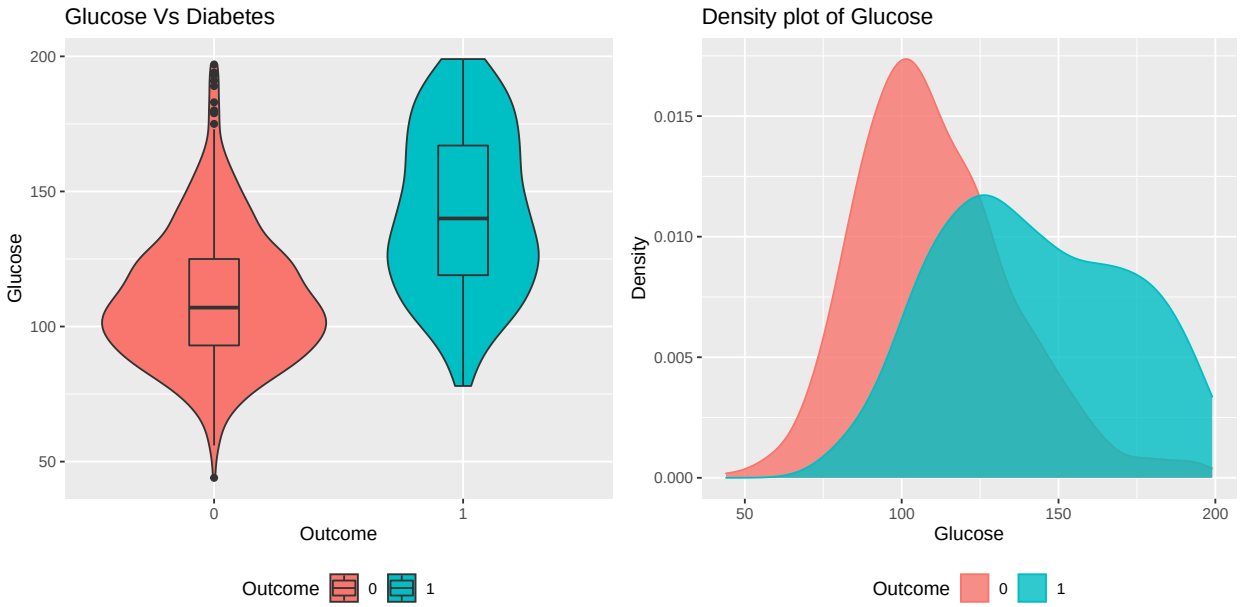
The 0 and 1 are binary interpretations of the negative and positive outcomes of diabetes being present. Out of 768 records, we find that 268 women were positively diagnosed as Diabetic and 500 women were tested as non-diabetic.

Predictor Variables - Pregnancies



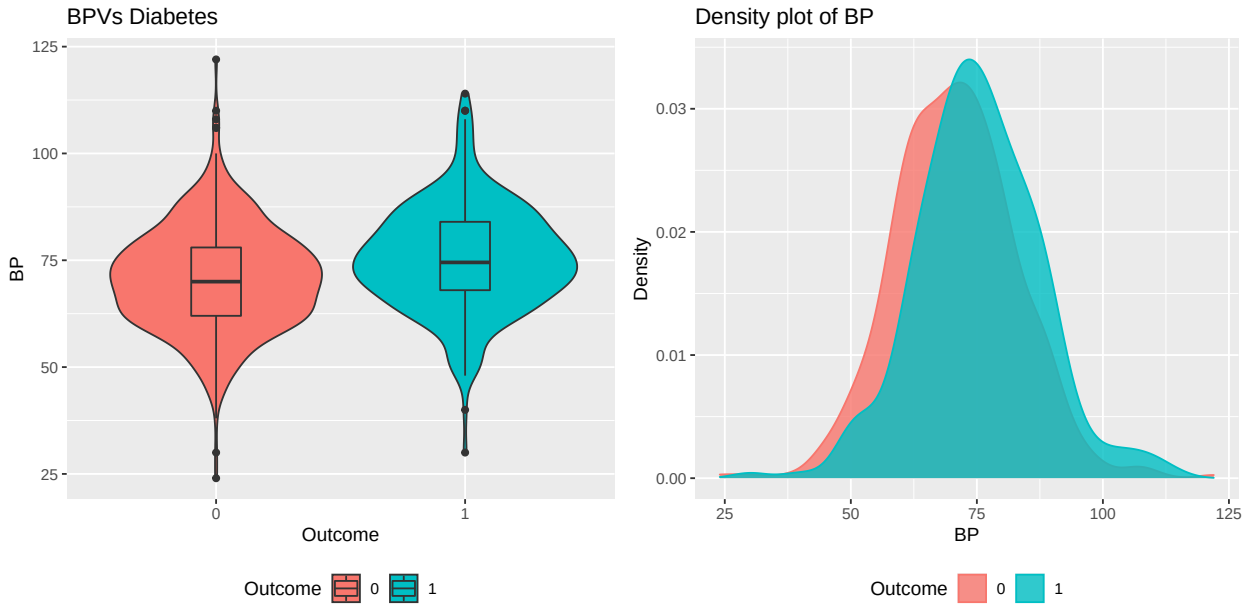
Academically we expect that there is direct correlation between pregnancy and Diabetes. Many women are diagnosed with diabetes for the first time during their pregnancies, called as Gestational Diabetes. Whereas, from the figures above, we do not seem to have a clear relationship between Diabetes and Pregancy.

Predictor Variables - Glucose



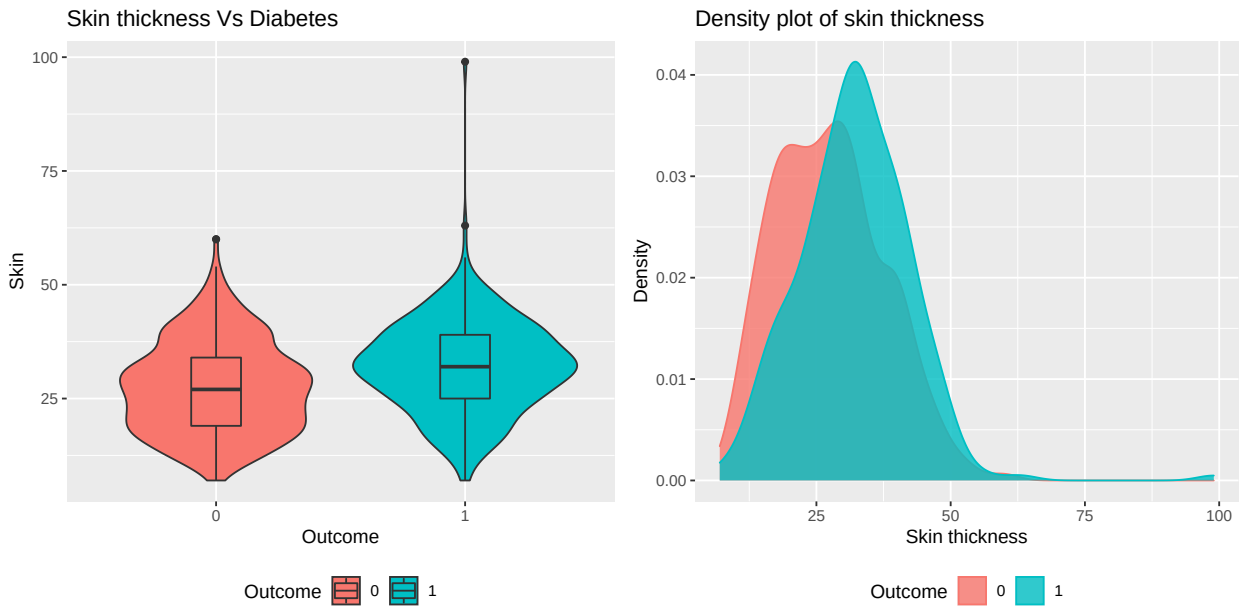
Glucose and prolonged duration of high sugar levels are primary indicators of diabetes. from the figure above we see that there is a clear difference in glucose level tested in the women who are diabetic and those who are non-diabetic. The density plot shows that there are substantially higher number of individuals who have higher levels of glucose are reported as diabetic. Another interesting thing to note is the normal bell-curve distribution for glucose plot density in individuals without diabetes. Though the density plots of both outcomes are overlapping, we can assume that Glucose to be one of the good predictor for the response.

Predictor Variables - BP



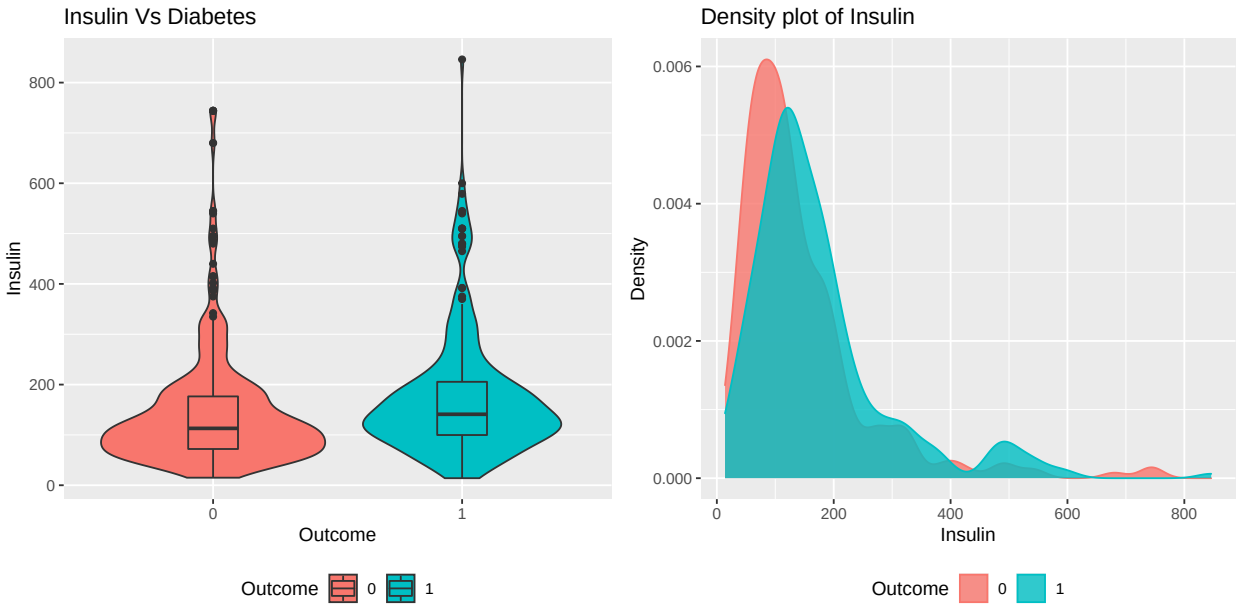
Academically, Over time diabetes will damage the arteries and thus causing high blood pressure. Diabetes as such is a cause for high blood pressure and not the otherway around. Our data seems to implicate the same. There seems to be no clear relationship between diabetes and BP, with diabetic patients having only a slightly higher overall BP level than non-diabetic patients. The Density plot of both outcome are almost identical and overlapping each other suggesting that BP may not be a good predictor for the response.

Predictor Variables - Skin



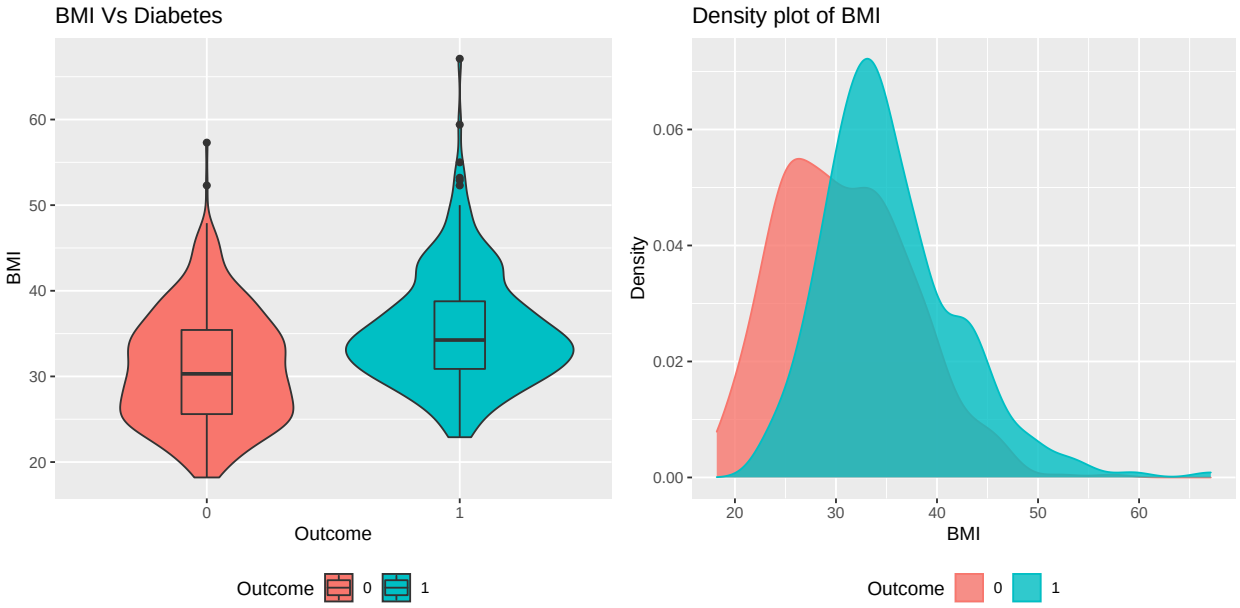
Observations on skin thickness seem to suggest that there seems to be no clear relationship between diabetes and skin thickness. Based on density plot, we see that there might be some relation between them. The range of skin thickness in diabetic patients is thicker than those of non-diabetics, although the concentration remains the same. This could imply skin thickness being an effect and not a cause of diabetes.

Predictor Variables - Insulin



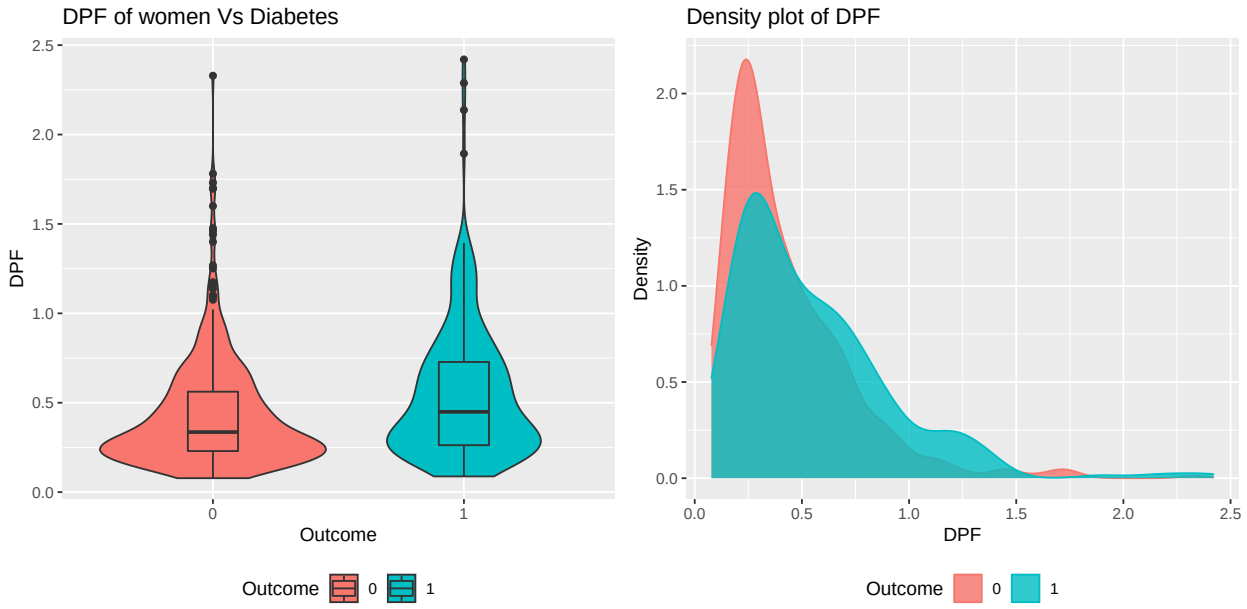
Observations on Insulin also seem to suggest that there seems to be no clear relationship between diabetes and serum insulin levels. Insulin levels are also relatively standard, and although the peaks for both diabetic and non-diabetic subjects are relatively low, the patients with diabetes seem to lean towards slightly higher insulin levels. With the insulin being resistant, not all insulin is used up in the breaking down of the insulin hence that could explain the sligthtly higher insulin levels in diabetic women.

Predictor Variables - BMI



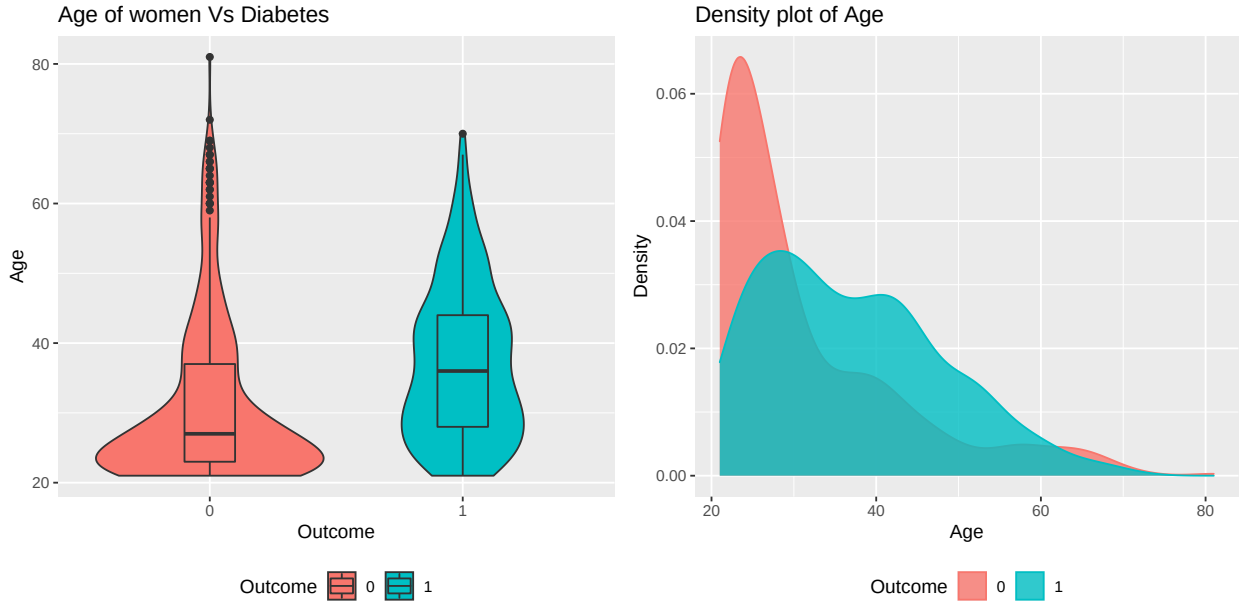
The BMI range for a non-diabetic is about 20-60 where as BMI of a Diabetic is about 25-70. This suggests that the BMI reveals a noticeably higher peak for diabetic patients overall, and indicates that there is a clear relationship between diabetes and BMI that needs to be explored. Density plot makes it

Predictor Variables - DPF



The Diabetes Pedigree Function, indicated by DPF, is basically a calculated value of diabetic history in relatives to indicate likelihood of diabetes. Interestingly, this diagram supports the measure and its intended function but also demonstrates that many non-diabetic patients also have a higher DPF, which could imply that the measure is not perfect, or simply that diabetes in kin does not translate to diabetes in an individual.

Predictor Variables - Age



Age seems to be a very telling indicator of diabetes, and suggests that diabetes becomes more prominent in an individual as they advance in their age. Age is also associate with other previous factors such as Pregnancy and glucose levels suggesting that there might be a possibility for Bi-Variate Associations

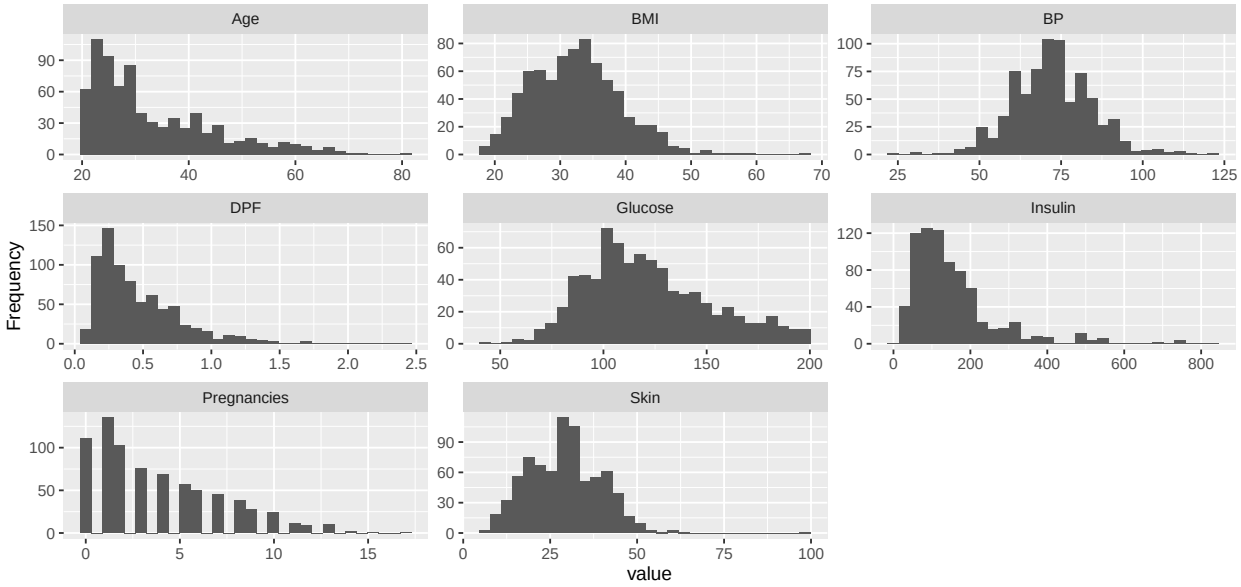
Univariate Analysis

Descriptive Statistics

Table 5: Descriptive Statistics

	Pregnancies	Glucose	BP	Skin	Insulin	BMI	DPF	Age
Mean	3.85	121.56	72.35	29.10	154.17	32.47	0.47	33.24
Std.Dev	3.37	30.51	12.38	10.43	117.25	6.90	0.33	11.76
Min	0.00	44.00	24.00	7.00	14.00	18.20	0.08	21.00
Q1	1.00	99.00	64.00	21.00	79.00	27.50	0.24	24.00
Median	3.00	117.00	72.00	29.00	122.00	32.35	0.37	29.00
Q3	6.00	140.50	80.00	36.00	185.00	36.60	0.63	41.00
Max	17.00	199.00	122.00	99.00	846.00	67.10	2.42	81.00
MAD	2.97	29.65	11.86	10.38	78.58	6.89	0.25	10.38
IQR	5.00	41.25	16.00	15.00	106.00	9.10	0.38	17.00
CV	0.88	0.25	0.17	0.36	0.76	0.21	0.70	0.35
Skewness	0.90	0.53	0.13	0.55	2.22	0.58	1.91	1.13
SE.Skewness	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09
Kurtosis	0.14	-0.28	0.81	1.88	6.57	0.85	5.53	0.62
N.Valid	768.00	768.00	768.00	768.00	768.00	768.00	768.00	768.00
Pct.Valid	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Histograms



The above histograms present a general spread of all the predictor variables and their frequency spread. Visually we find that the Age, DPF, Pregnancies, Insulin are left skewed. BP, BMI and Skin seems to be normal. Glucose seems to be right skewed. To be sure of these assumptions we need to proceed with Shaprio-Wilk's normality test and Q-Q plots to validate our assumptions.

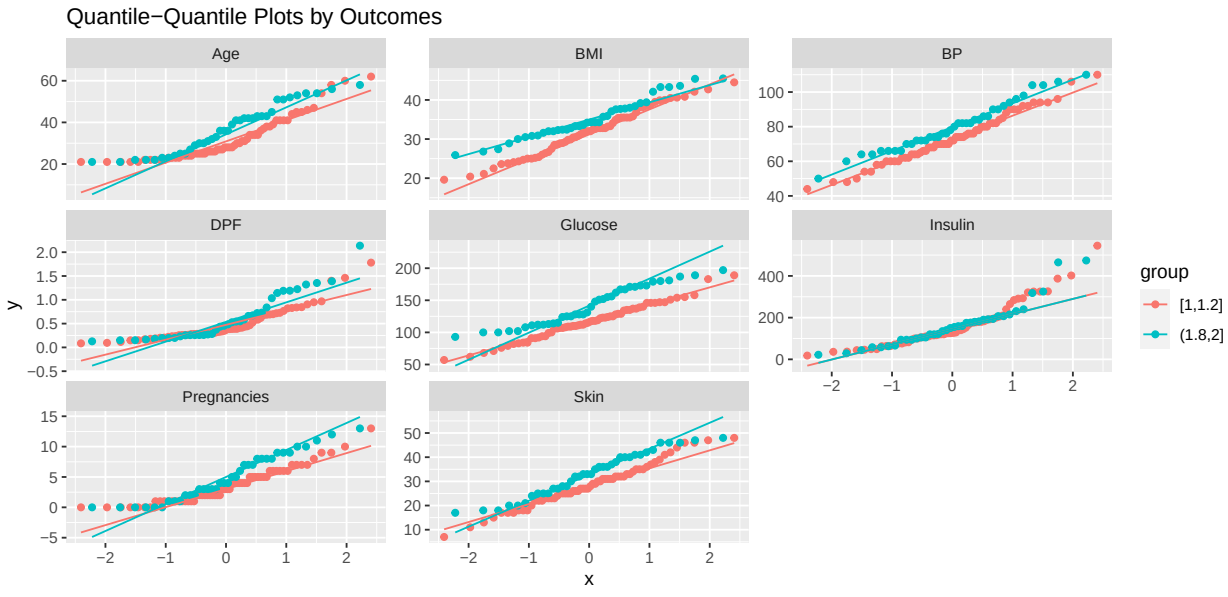
Shapiro-Wilk's normality test

Table 6: Shapiro-Wilk's normality test for PIMA Indian Diabetes data set

Outcome	variable	statistic	p
1	Age	0.80	0.000
1	BMI	0.98	0.000
1	BP	0.99	0.002
1	DPF	0.83	0.000
1	Glucose	0.97	0.000
1	Insulin	0.77	0.000
1	Pregnancies	0.88	0.000
1	Skin	0.98	0.000
2	Age	0.95	0.000
2	BMI	0.95	0.000
2	BP	0.99	0.024
2	DPF	0.86	0.000
2	Glucose	0.97	0.000
2	Insulin	0.83	0.000
2	Pregnancies	0.94	0.000
2	Skin	0.95	0.000

Shaprio Wilk's test for normality allows us to identify whether the data is normally distributed. From the p-values in the above table it is clear that all predictor variables, except for BP, have a p-value lower than 0.001 suggesting that the sample deviates from normality.

QQ Plots by Outcomes

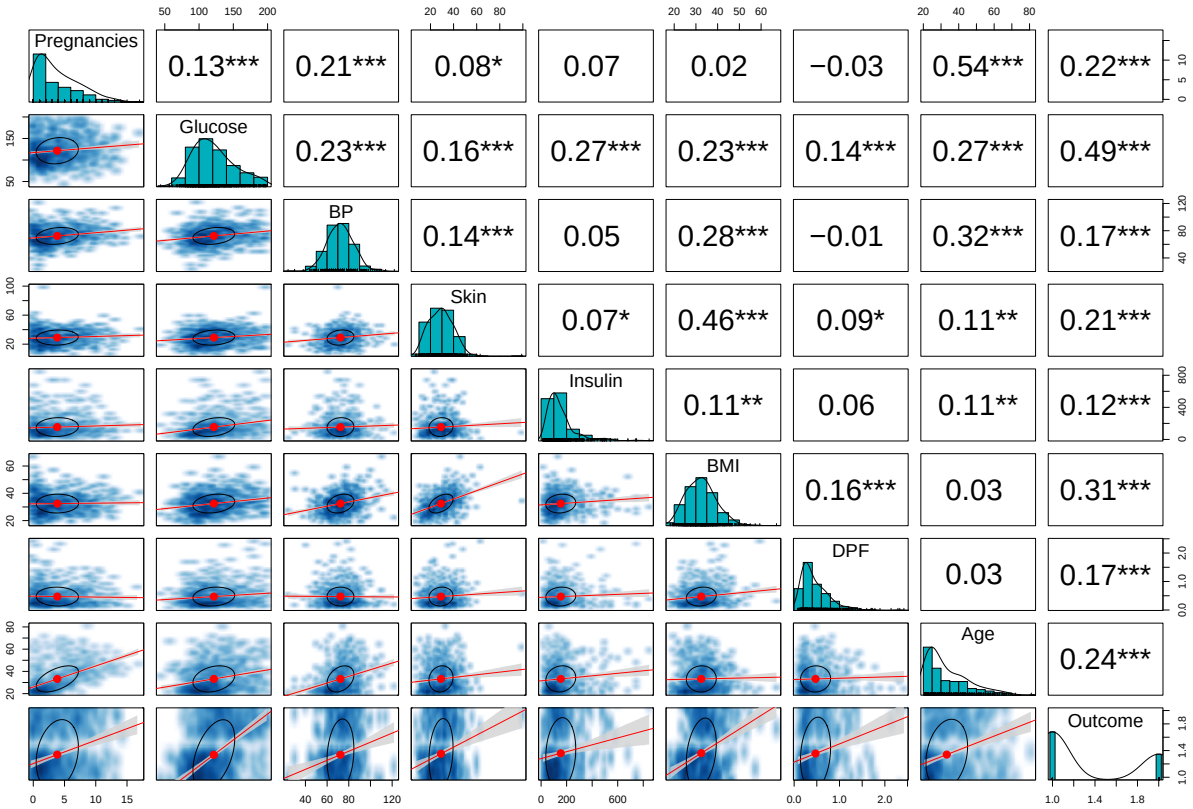


Based on the Q-Q plots, we can draw the following inferences for our data set

- The presence stair-step appearance in the Pregnancies and BP Q-Q plot suggests that data values are highly repeated

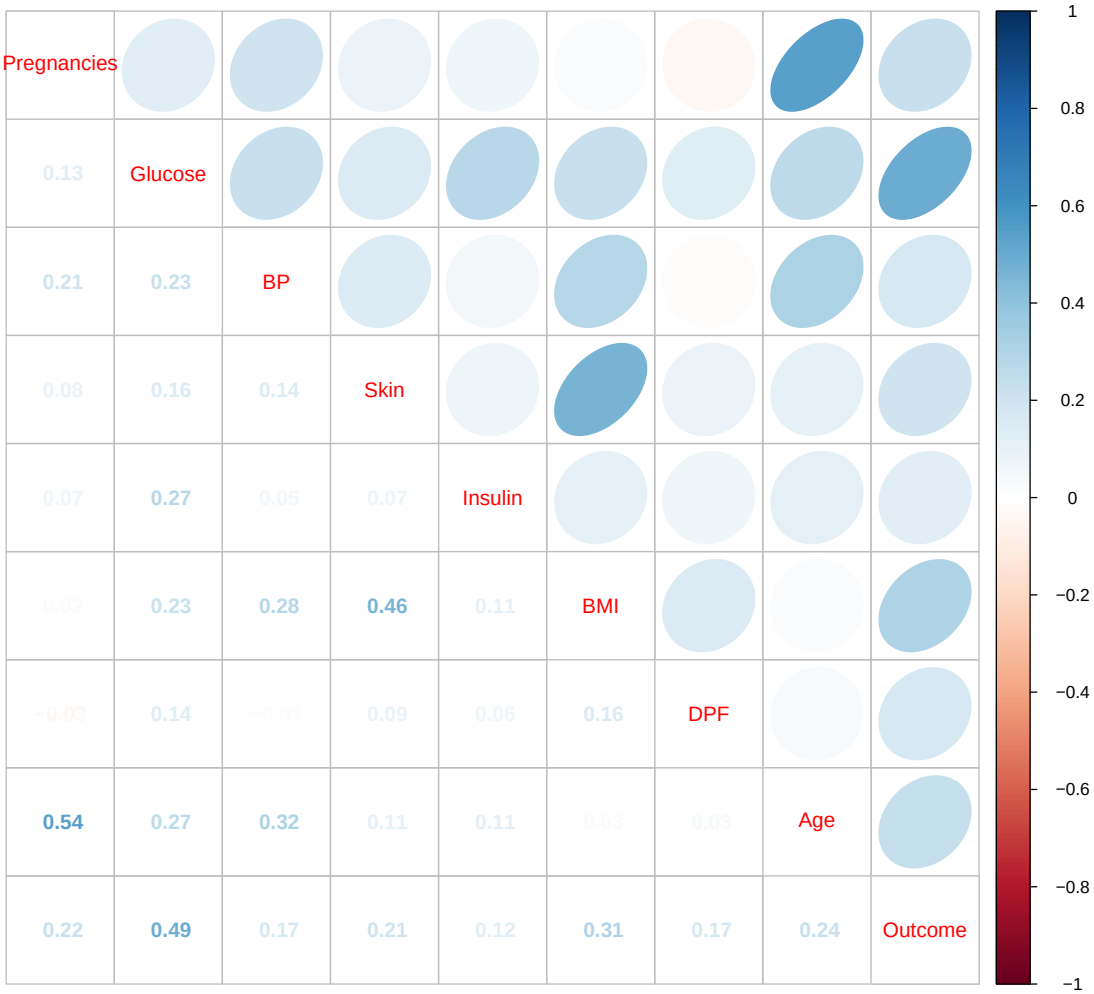
- Data distribution for Glucose, BMI and Skin are normal but there are a lot of Outliers.
- Data distribution for Age, DPF and Insulin shows data is skewed to the Right.
- We see that the Outliers are present in both the Diabetic and Non-diabetic data.
- The plot for non-diabetic is more skewed to the right for Age than for the diabetic.
- BP has no role in both Diabetic and Non-diabetic

Scatter Plot Matrix



Scatter matrix show the spread and concentration of each variable. We see that there are correlations that are identified (by stars) as significant.

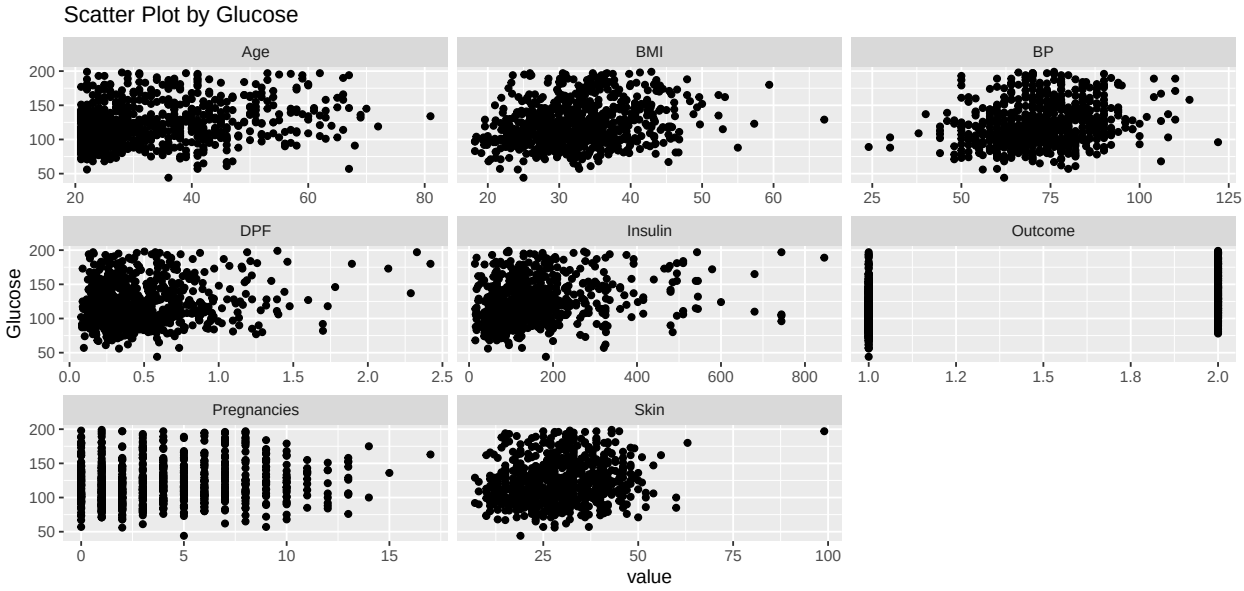
Correlation Analysis



We notice that the Glucose is highly correlated with the target variables, whereas BP has low correlation values. Based on the Scatter Plot matrix and Correlation Analysis, it is clear that there are possible Bi-variate relations that needs to be studied. the following seems to be predominant correlation

- BMI and Skin (0.465)
- Insulin and Glucose (0.30)
- Pregnancy and Age (0.544)
- BP and Age (0.318)
- BMI and Insulin (0.119)
- BMI and BP (0.2482)

Scatter Plots by Glucose

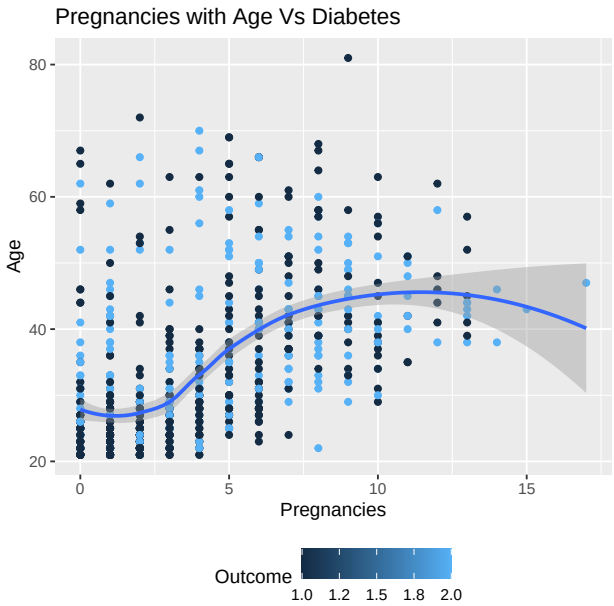


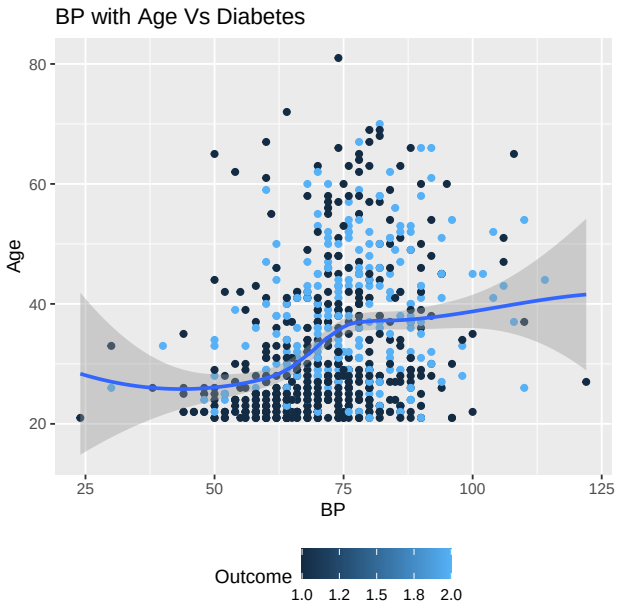
We do see that Glucose plays a major role in interacting with every variable. In other terms, the Blood Glucose level plays a major role in not only predicting the diabetes, it also affects/being affected by the other predictor variables.

Bi-variate Analysis

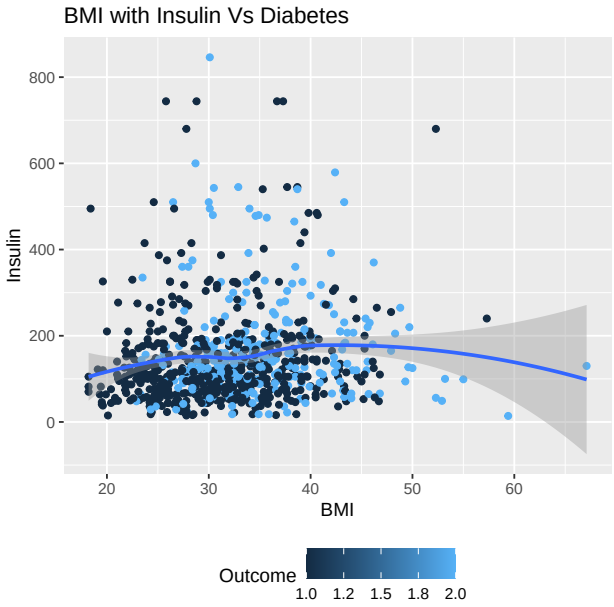
Now we perform Bi-Variate analysis to see how combination of two variables affect the occurrence of Diabetes:

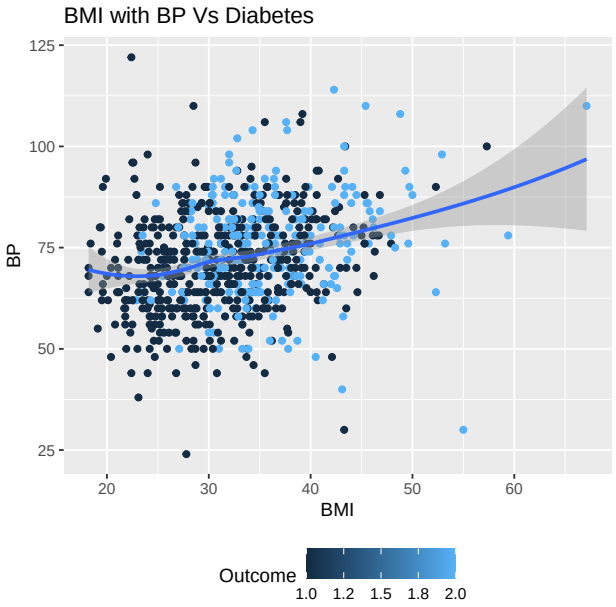
- Women with low BMI and skin thickness did not have Diabetes
- Non-diabetic women have lower levels of Glucose and Insulin compared to their Diabetic counterparts who recorded high levels of Glucose and a wide range of Insulin.





- No clear boundary can be established between the diabetic and non-diabetic women based on number of pregnancies and their age.
- Young women with normal BP did not have Diabetes

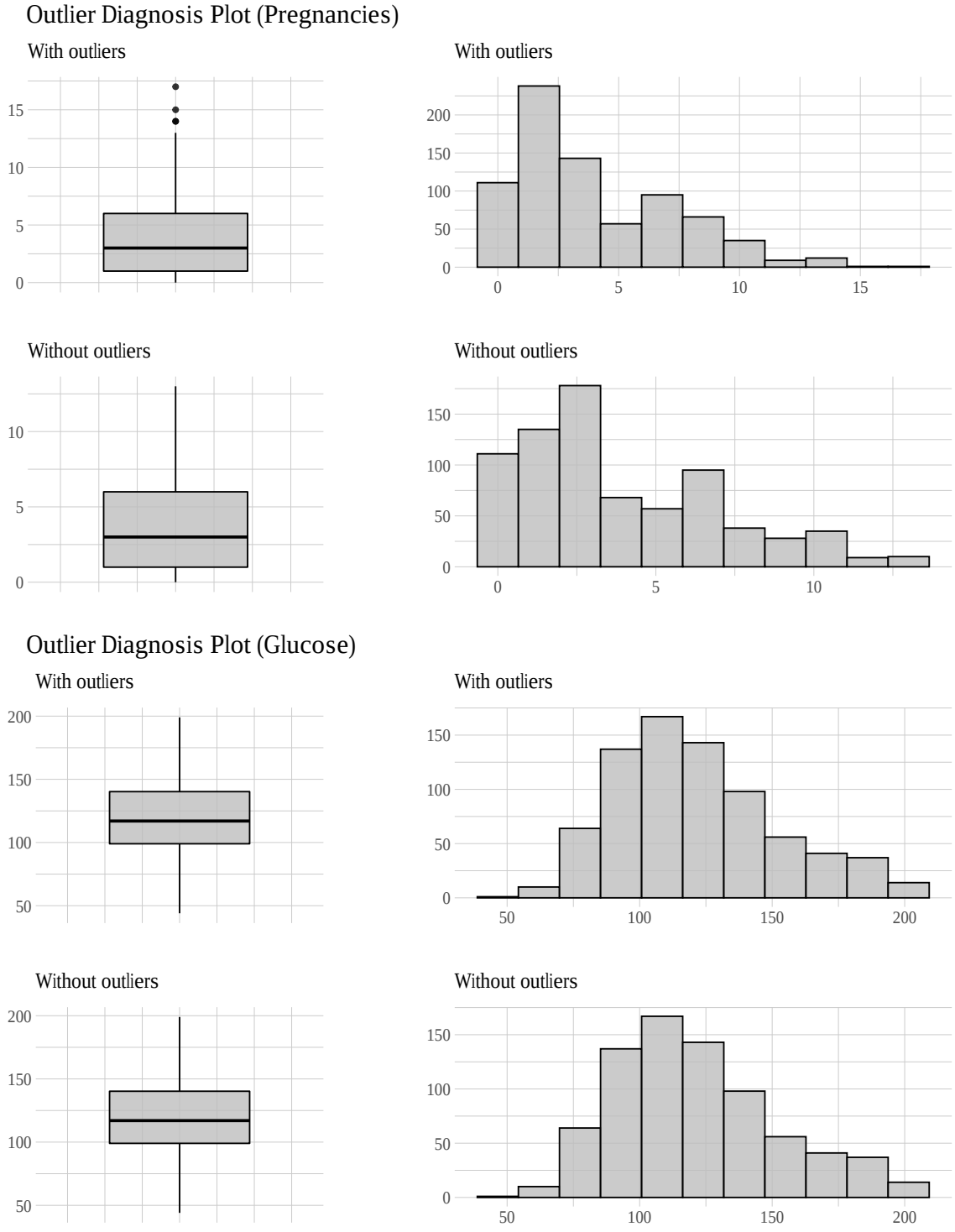




- Diabetic women can be differentiated from non-diabetic based on BMI and BP values
- Women with low BMI and Insulin content did not have Diabetes

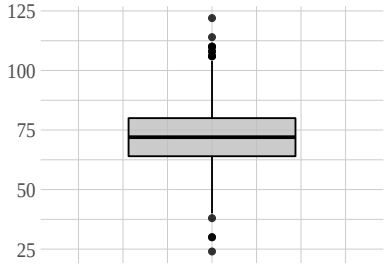
Outliers

The Violin plots, and other plots has constantly showed that there are a lot of outliers in our dataset and needs to be taken care of before algorithm training and testing.

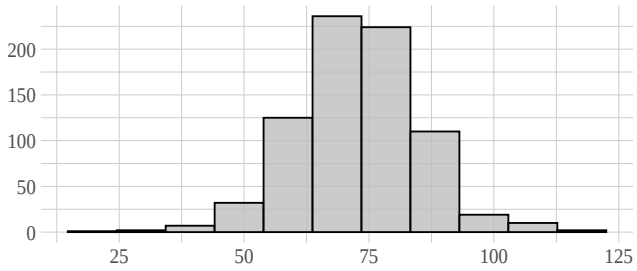


Outlier Diagnosis Plot (BP)

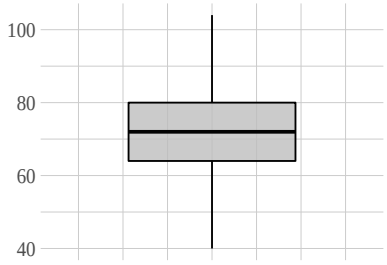
With outliers



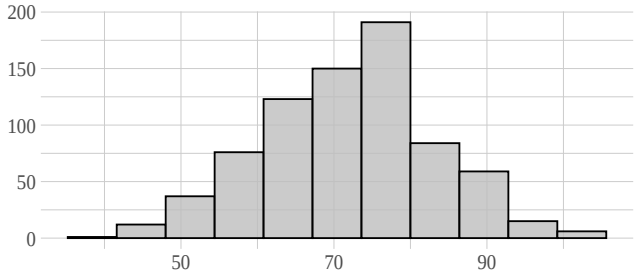
With outliers



Without outliers

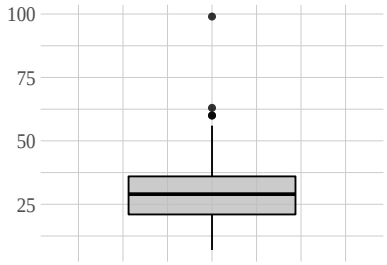


Without outliers

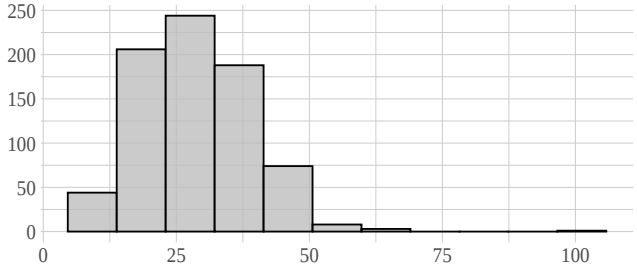


Outlier Diagnosis Plot (Skin)

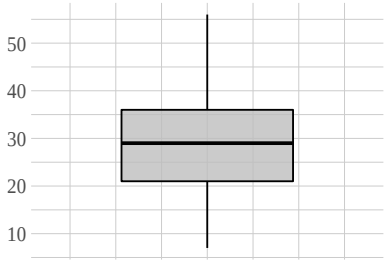
With outliers



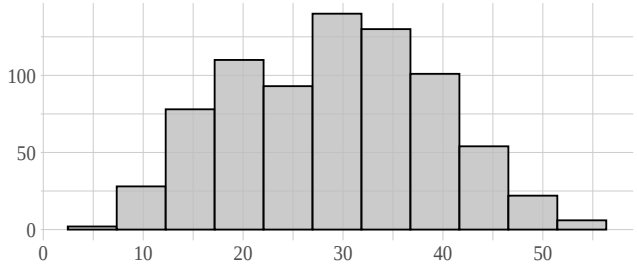
With outliers



Without outliers

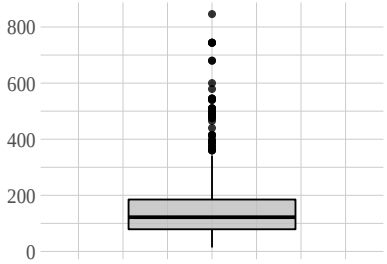


Without outliers

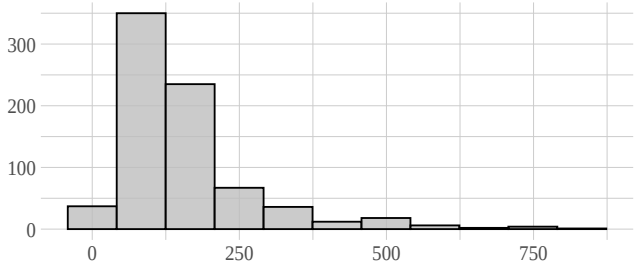


Outlier Diagnosis Plot (Insulin)

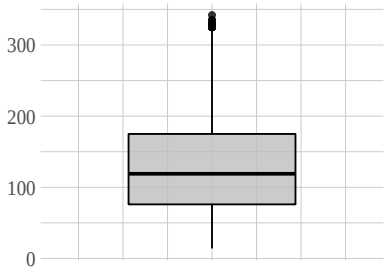
With outliers



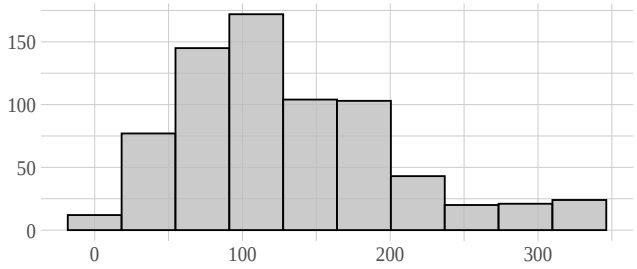
With outliers



Without outliers

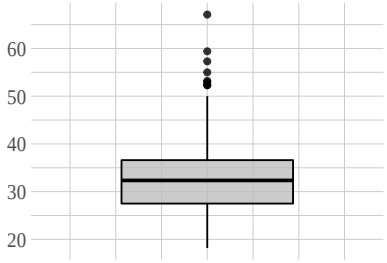


Without outliers

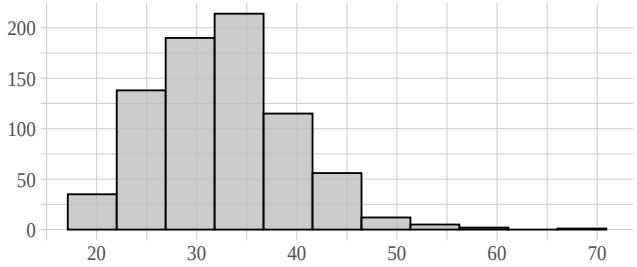


Outlier Diagnosis Plot (BMI)

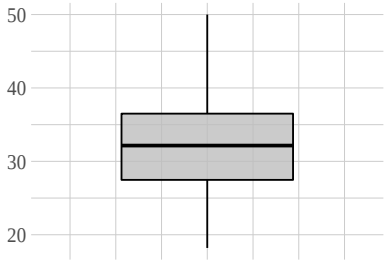
With outliers



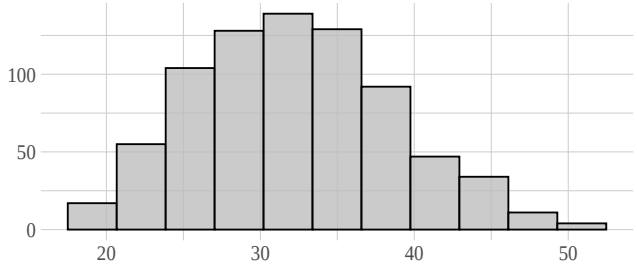
With outliers

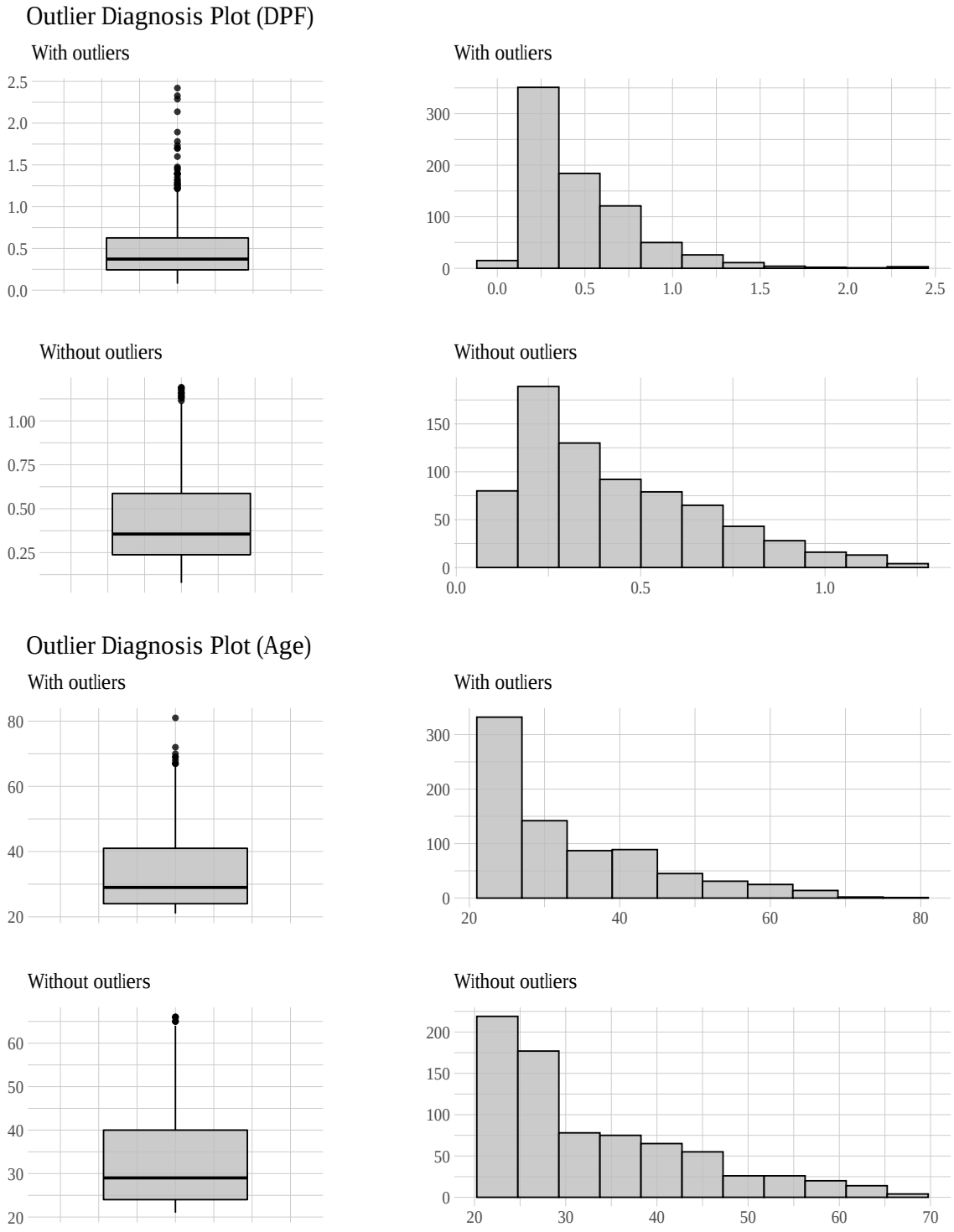


Without outliers



Without outliers





From the analysis on the outliers,

- We find that Outliers did not affect data on Pregnancies, Glucose, BP, DPF or Age.
- Once the Outliers were removed from the Skin thickness, Insulin and BMI data, the data seems to be normal.

We chose to use IQR to handle the outliers thus making the data training and testing more accurate.

Algorithm Testing

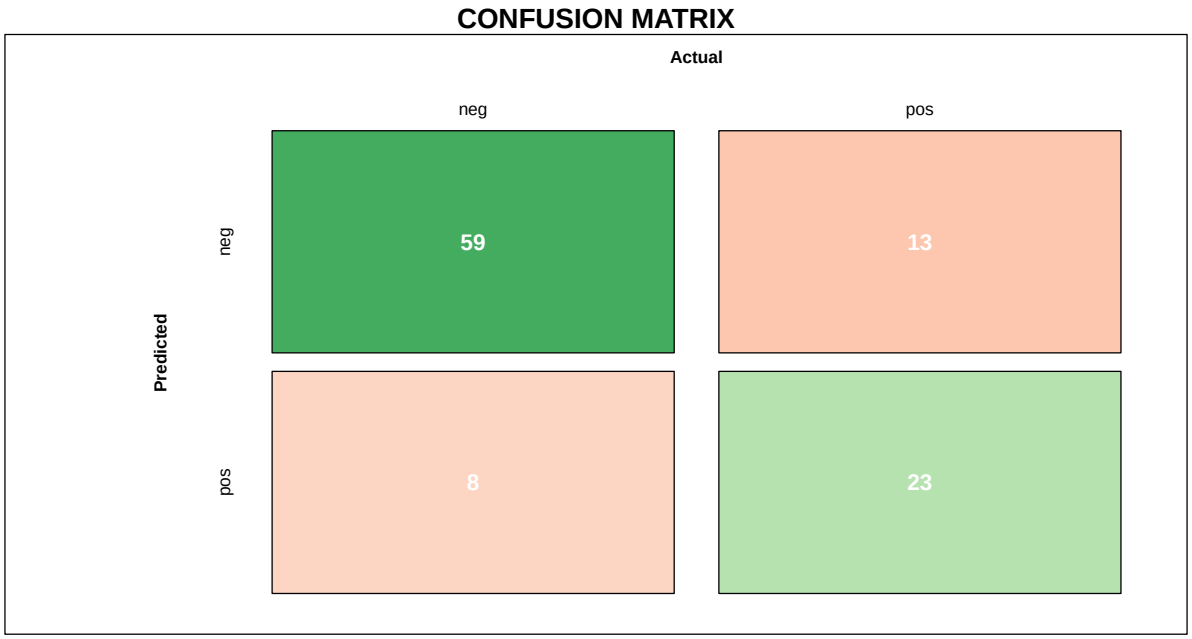
For our dataset, we decided to train, test and compare the following statistical learning methods.

- Logistic Regression
- KNN
- XGBOOST - eXtreme Gradient BOOSTing
- Support Vector Machines
- Random Forest
- Rpart CART - classification and Regression Decision Trees

In order to do the models, the dataset is split into two: Training and Test data.

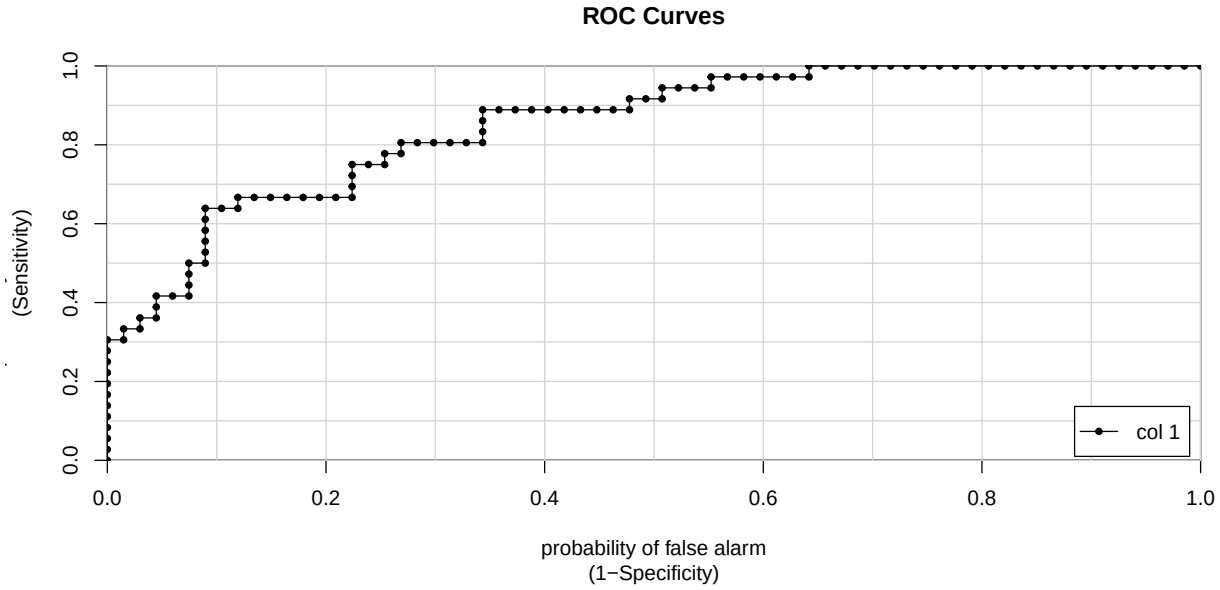
- 85% of the dataset (589 observations) will be used as Training set which we will use to train the model
- 15% of the dataset (103 observations) will be used as Testing set which we will use to test the trained model

Logistic Regression



DETAILS

Sensitivity	**Specificity**	**Precision**	**Recall**	**F1**
0.638889	0.880597	0.741935	0.638889	0.686567
AccuracyPValue	**McnemarPValue**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
0.00093	0.382733	0.741935	0.819444	0.759743
Accuracy	**Kappa**	**Prevalence**	**Detection Rate**	**Detection Prevalence**
0.796117	0.536732	0.349515	0.223301	0.300971

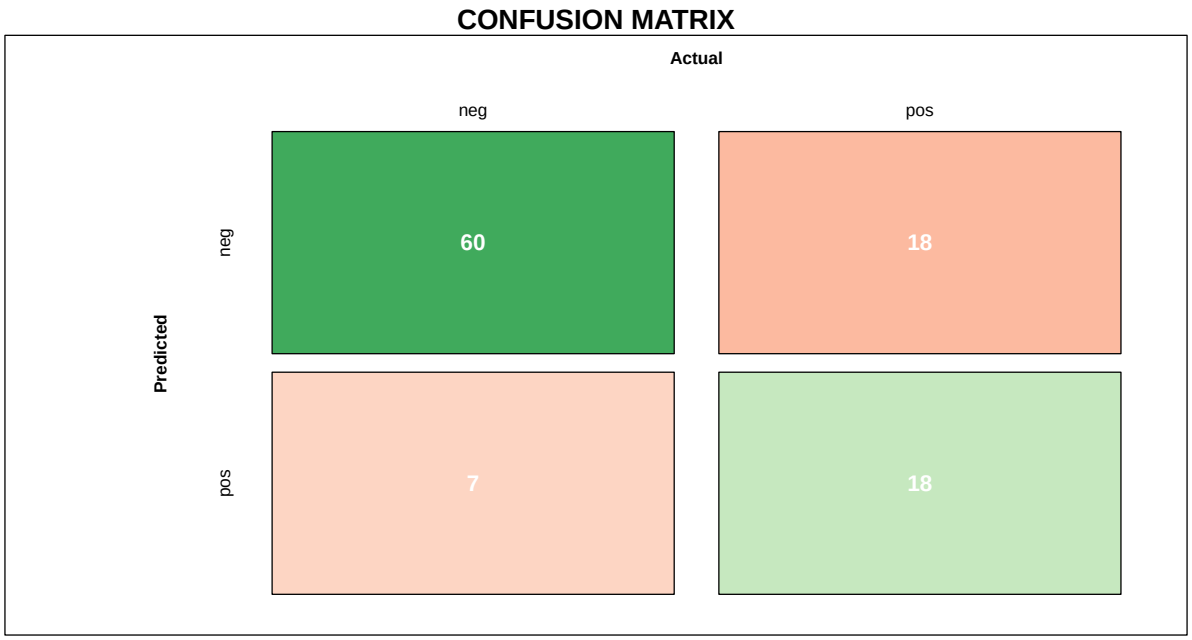


Based on the Logistic Regression model we see that:

- The ROC value is 0.833
- The accuracy is 0.8
- The area under the curve has a value of 0.85
- The F1 score is 0.69
- The Sensitivity is 0.64
- The Precision is 0.74

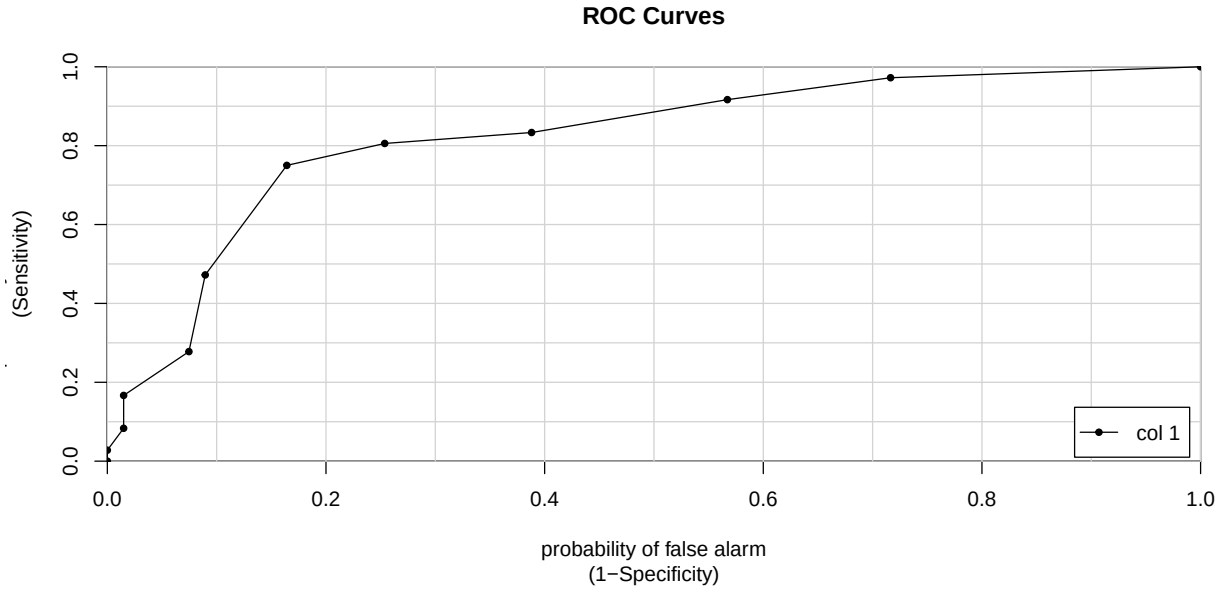
We get a 80% accuracy score on the test data. Our Precision for the model stands at 0.74. This indicates that 74% of the time our model classified the patients in a high risk category when they actually had a high risk of getting diabetes. The Recall/Sensitivity is 0.64, implying that 64% of the time people having actually having high risk were classified correctly by our model.

KNN



DETAILS

Sensitivity	**Specificity**	**Precision**	**Recall**	**F1**
0.5	0.895522	0.72	0.5	0.590164
AccuracyPValue	**McnemarPValue**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
0.013253	0.0455	0.72	0.769231	0.697761
Accuracy	**Kappa**	**Prevalence**	**Detection Rate**	**Detection Prevalence**
0.757282	0.425608	0.349515	0.174757	0.242718

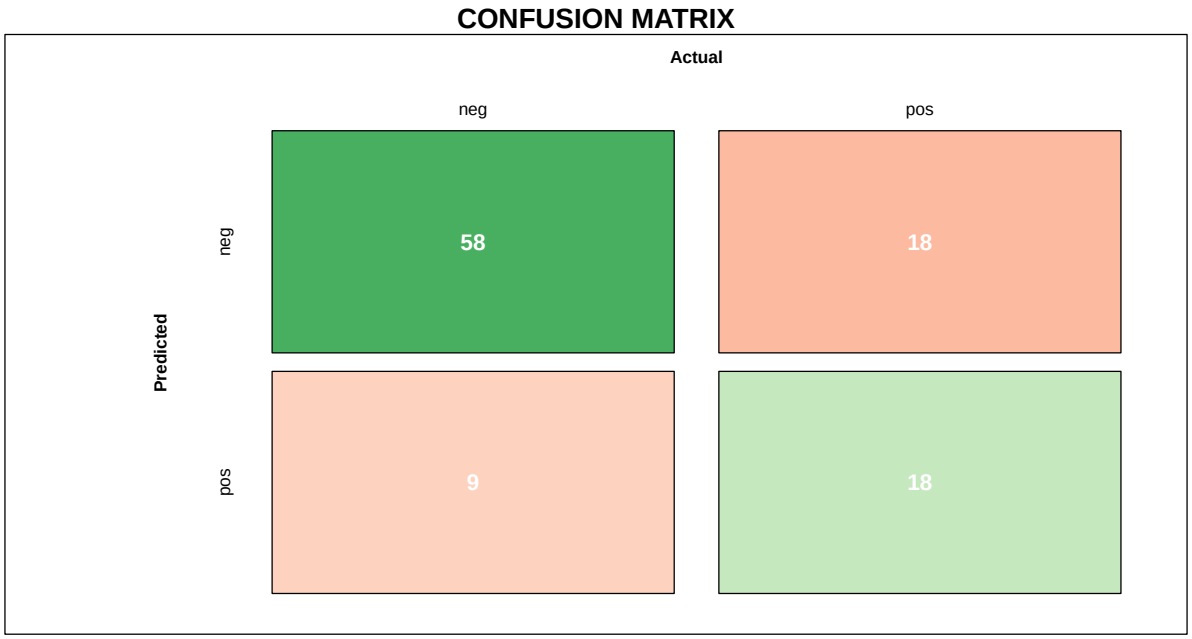


Based on the KNN model we see that:

- The accuracy is 0.76
- The area under the curve has a value of 0.82
- The F1 score is 0.59
- The Sensitivity is 0.5
- The Precision is 0.72

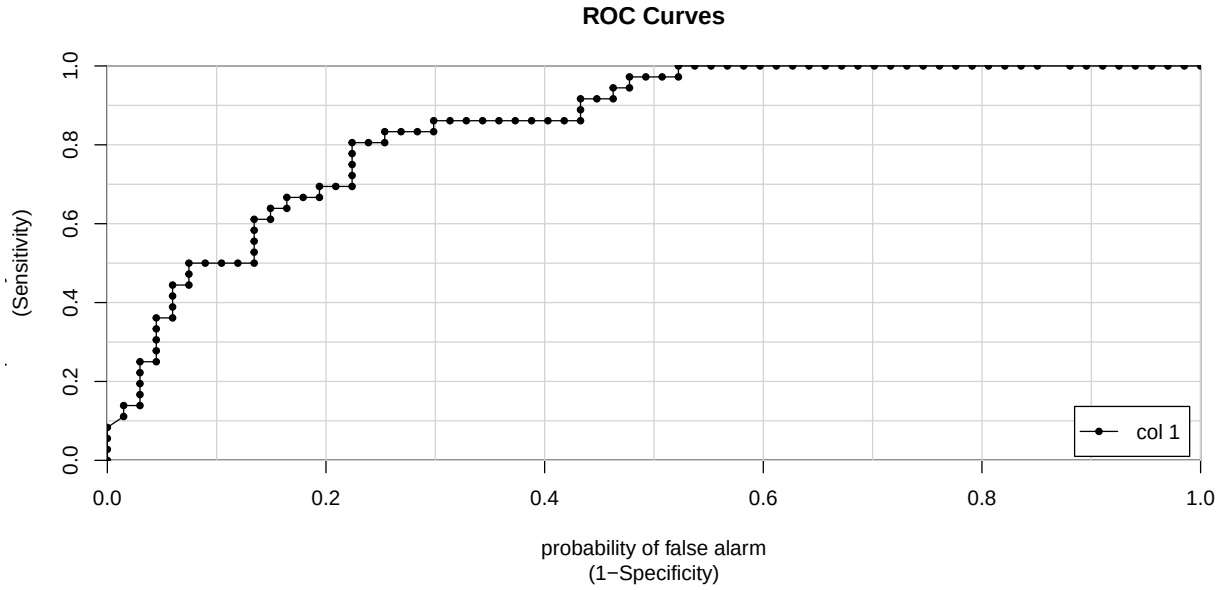
We get a 76% accuracy score on the test data. Our Precision for the model stands at 0.72. This indicates that 72% of the time our model classified the patients in a high risk category when they actually had a high risk of getting diabetes. The Recall/Sensitivity is 0.5, implying that 50% of the time people having actually having high risk were classified correctly by our model.

XGBOOST - eXtreme Gradient BOOSTing



DETAILS

Sensitivity	Specificity	Precision	Recall	F1
0.5	0.865672	0.666667	0.5	0.571429
AccuracyPValue	**McnemarPValue**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
0.037305	0.123658	0.666667	0.763158	0.682836
Accuracy	**Kappa**	**Prevalence**	**Detection Rate**	**Detection Prevalence**
0.737864	0.388119	0.349515	0.174757	0.262136

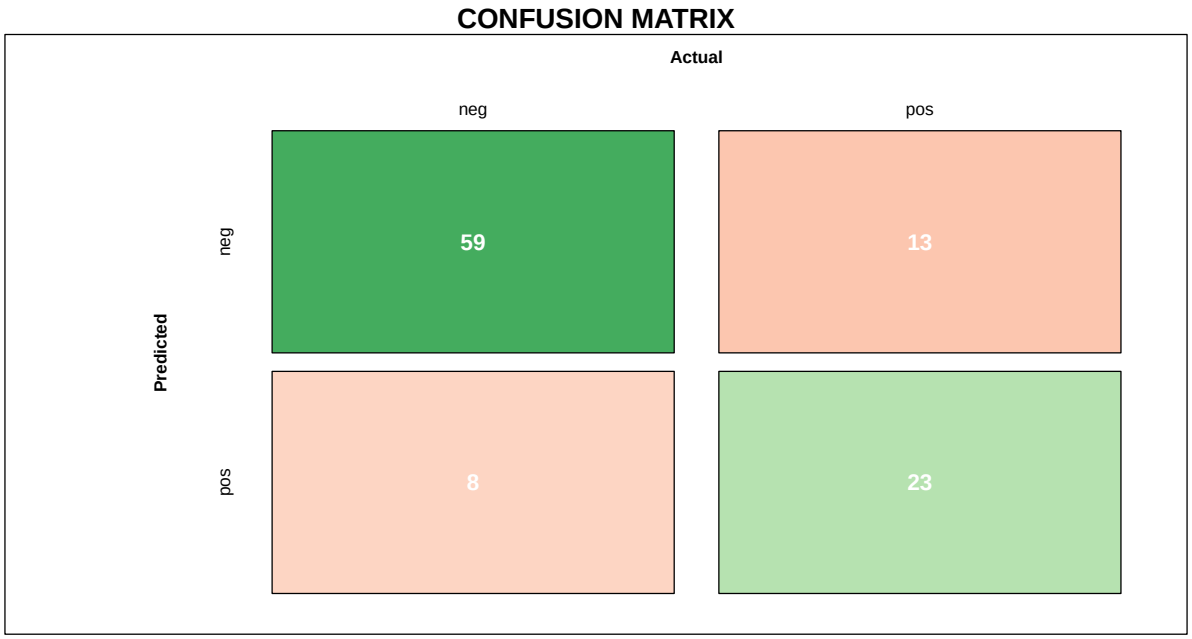


Based on the eXtreme Gradient BOOSTing model we see that:

- The ROC value is 0.813
- The accuracy is 0.74
- The area under the curve has a value of 0.85
- The F1 score is 0.57
- The Sensitivity is 0.5
- The Precision is 0.67

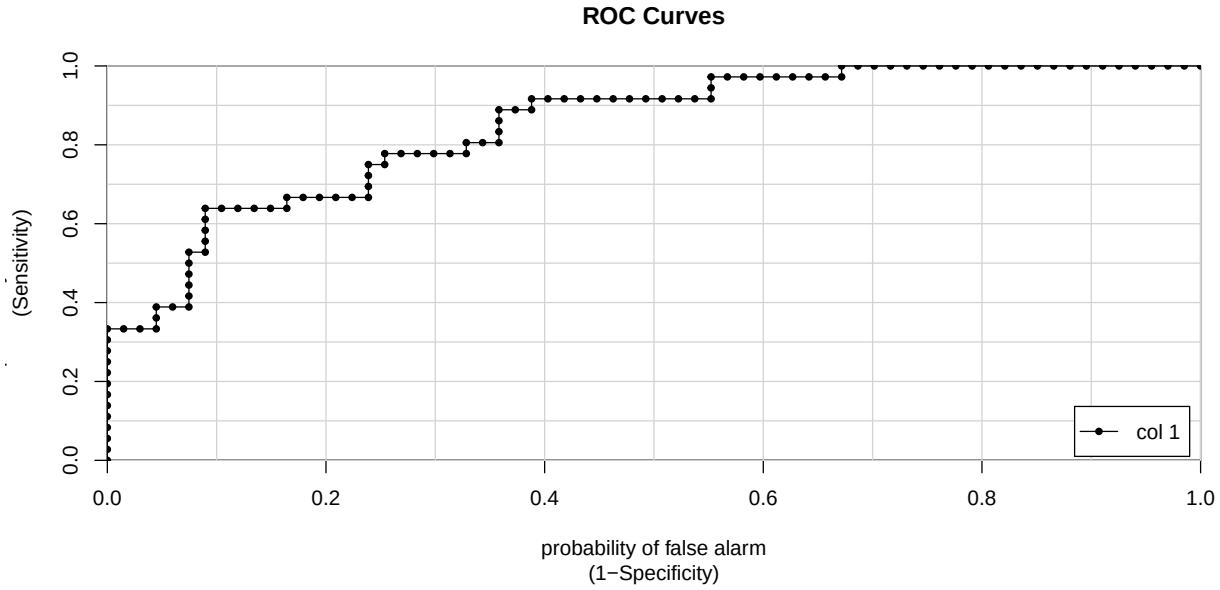
We get a 74% accuracy score on the test data. Our Precision for the model stands at 0.67. This indicates that 67% of the time our model classified the patients in a high risk category when they actually had a high risk of getting diabetes. The Recall/Sensitivity is 0.5, implying that 50% of the time people having actually having high risk were classified correctly by our model.

Support Vector Machines



DETAILS

Sensitivity	**Specificity**	**Precision**	**Recall**	**F1**
0.638889	0.880597	0.741935	0.638889	0.686567
AccuracyPValue	**McnemarPValue**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
0.00093	0.382733	0.741935	0.819444	0.759743
Accuracy	**Kappa**	**Prevalence**	**Detection Rate**	**Detection Prevalence**
0.796117	0.536732	0.349515	0.223301	0.300971

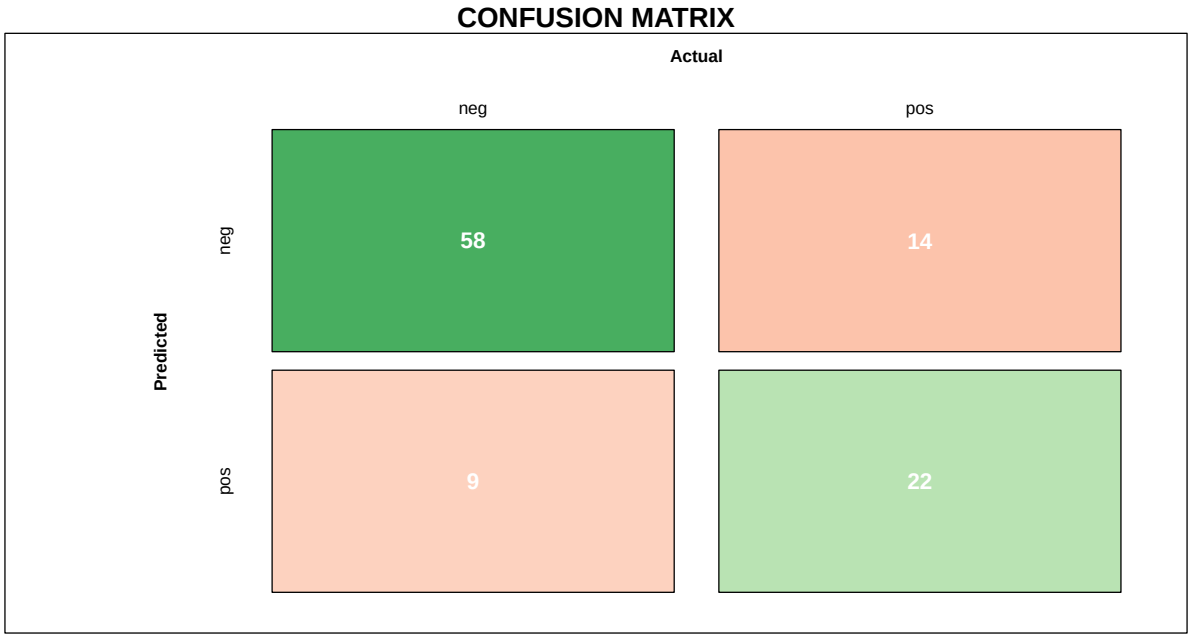


Based on the Support Vector Machines model we see that:

- The ROC value is 0.839
- The accuracy is 0.8
- The area under the curve has a value of 0.85
- The F1 score is 0.69
- The Sensitivity is 0.64
- The Precision is 0.74

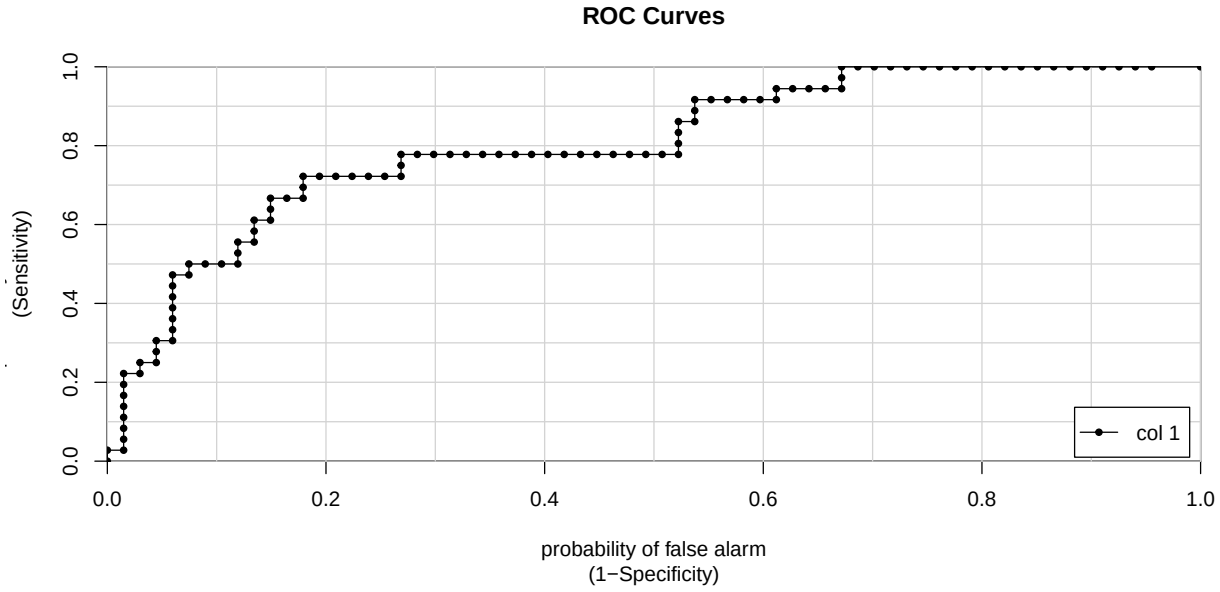
We get a 80% accuracy score on the test data. Our Precision for the model stands at 0.74. This indicates that 74% of the time our model classified the patients in a high risk category when they actually had a high risk of getting diabetes. The Recall/Sensitivity is 0.64, implying that 64% of the time people having actually having high risk were classified correctly by our model.

Random Forest



DETAILS

Sensitivity	**Specificity**	**Precision**	**Recall**	**F1**
0.611111	0.865672	0.709677	0.611111	0.656716
AccuracyPValue	**McnemarPValue**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
0.003891	0.404248	0.709677	0.805556	0.738391
Accuracy	**Kappa**	**Prevalence**	**Detection Rate**	**Detection Prevalence**
0.776699	0.492611	0.349515	0.213592	0.300971

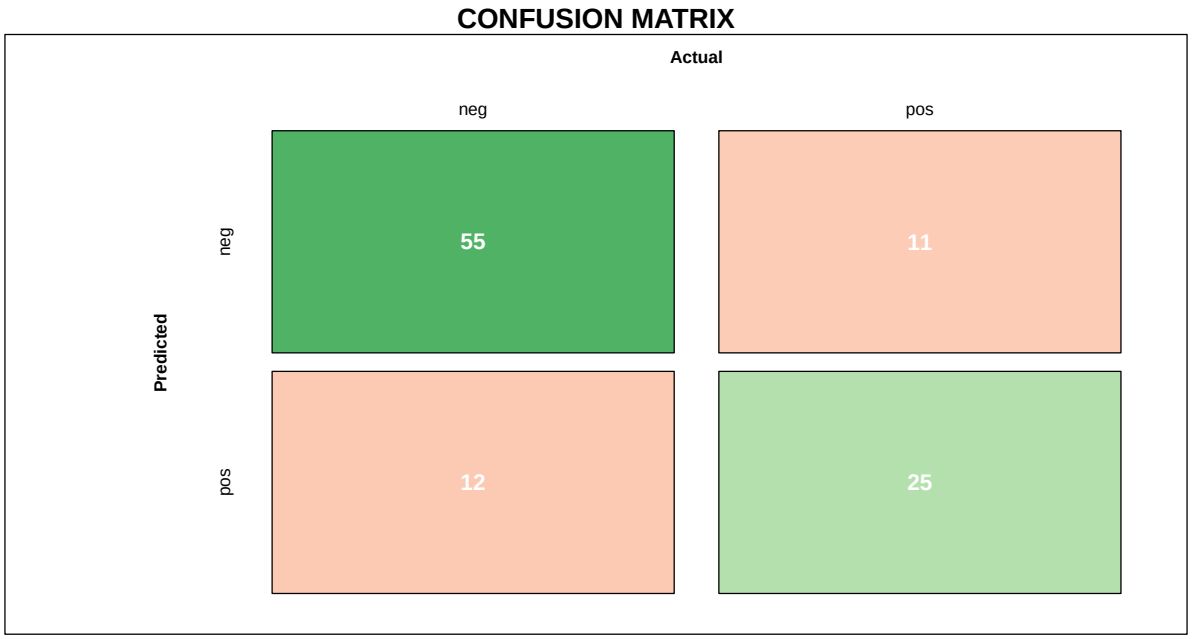


Based on the Random Forest model we see that:

- The accuracy is 0.78
- The area under the curve has a value of 0.81
- The F1 score is 0.66
- The Sensitivity is 0.61
- The Precision is 0.71

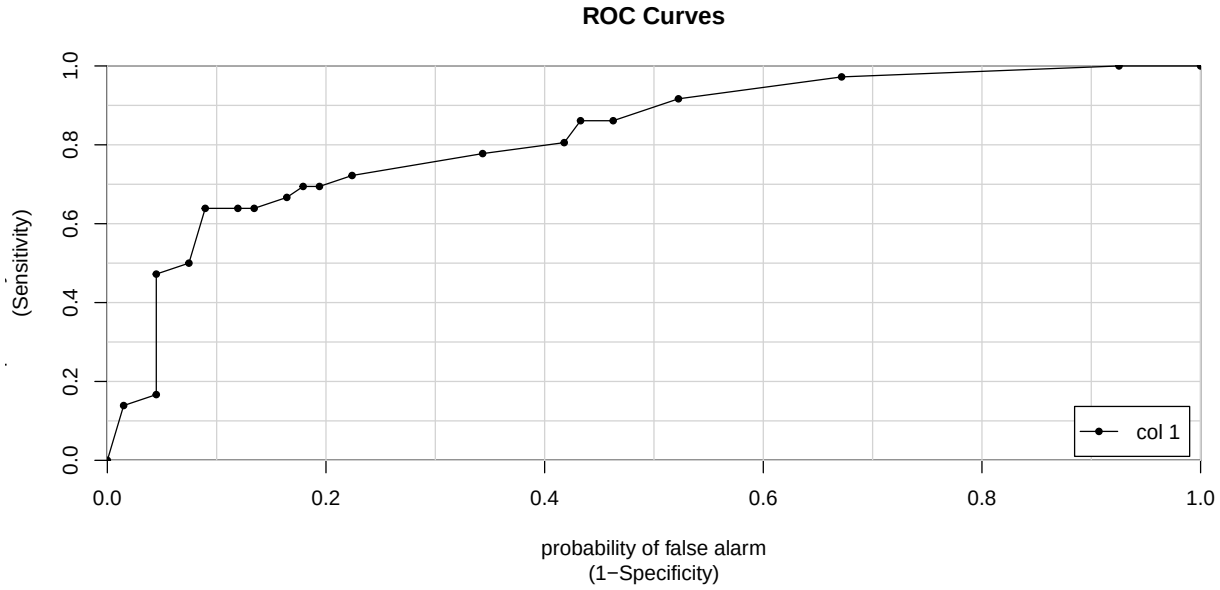
We get a 78% accuracy score on the test data. Our Precision for the model stands at 0.71. This indicates that 71% of the time our model classified the patients in a high risk category when they actually had a high risk of getting diabetes. The Recall/Sensitivity is 0.61, implying that 61% of the time people having actually having high risk were classified correctly by our model.

Rpart CART - classification and Regression Decision Trees



DETAILS

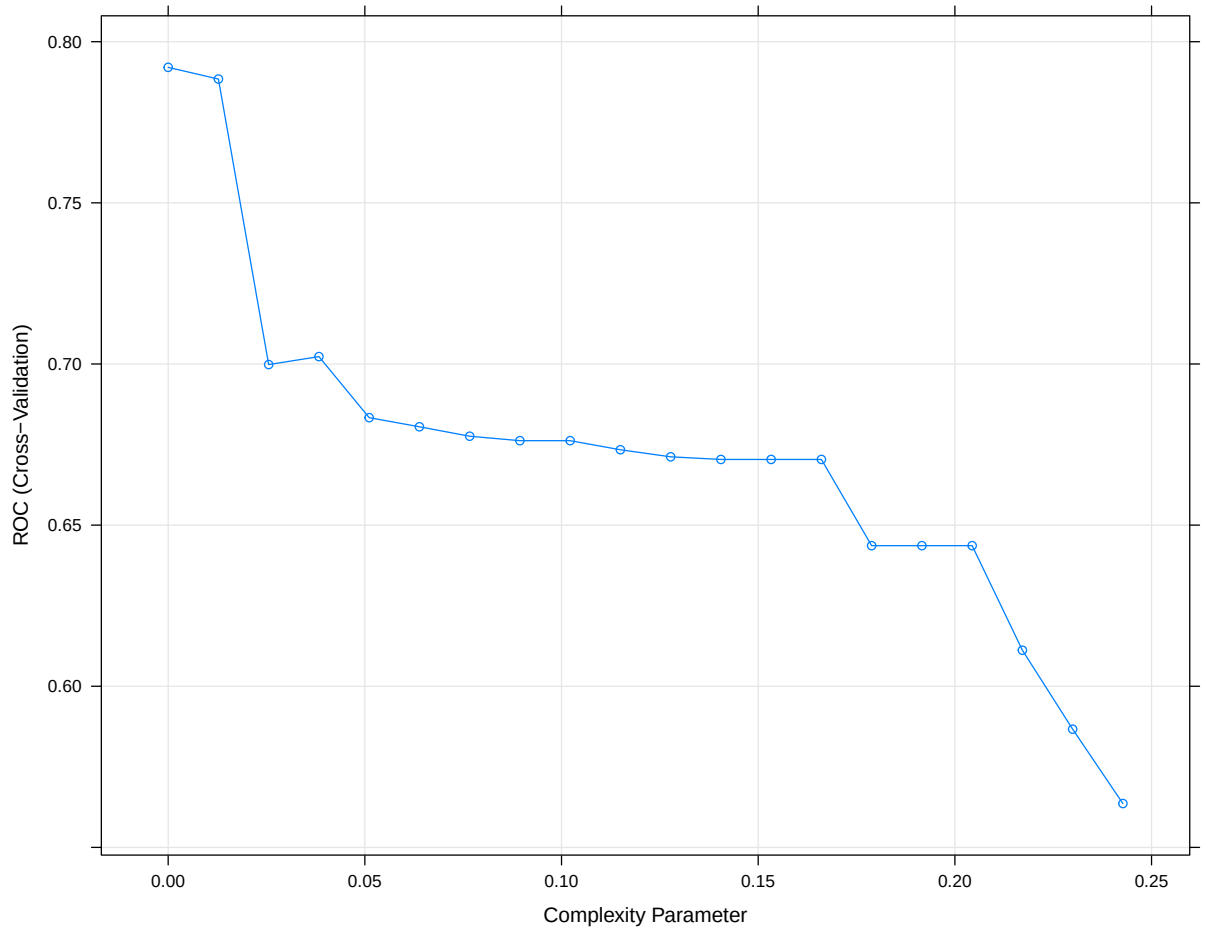
Sensitivity	**Specificity**	**Precision**	**Recall**	**F1**
0.694444	0.820896	0.675676	0.694444	0.684932
AccuracyPValue	**McnemarPValue**	**Pos Pred Value**	**Neg Pred Value**	**Balanced Accuracy**
0.003891	1	0.675676	0.833333	0.75767
Accuracy	**Kappa**	**Prevalence**	**Detection Rate**	**Detection Prevalence**
0.776699	0.512049	0.349515	0.242718	0.359223



Based on the Rpart CART classification and Regression Decision Trees model we see that:

- The accuracy is 0.78
- The area under the curve has a value of 0.82
- The F1 score is 0.68
- The Sensitivity is 0.69
- The Precision is 0.68

We get a 78% accuracy score on the test data. Our Precision for the model stands at 0.68. This indicates that 68% of the time our model classified the patients in a high risk category when they actually had a high risk of getting diabetes. The Recall/Sensitivity is 0.69, implying that 69% of the time people having actually having high risk were classified correctly by our model.



cp 1 0 n= 589

node), split, n, loss, yval, (yprob) * denotes terminal node

1) root 589 210 neg (0.650 0.350)
   2) Glucose< 1.3e+02 389 81 neg (0.792 0.208)
     4) Age< 28 212 18 neg (0.915 0.085)

     5) BMI< 31 118 2 neg (0.983 0.017) *
     6) BMI>=31 94 16 neg (0.830 0.170)
       18) BP>=54 87 11 neg (0.874 0.126)
       36) Glucose< 1.1e+02 52 3 neg (0.942 0.058) *

37) Glucose>=1.1e+02 35 8 neg (0.771 0.229)

38) BMI>=34 27 3 neg (0.889 0.111) *
39) BMI< 34 8 3 pos (0.375 0.625) *19) BP< 54 7 2 pos (0.286 0.714)*
7) Age>=28 177 63 neg (0.644 0.356)

8) BMI< 26 33 0 neg (1.000 0.000) *
9) BMI>=26 144 63 neg (0.562 0.438)
22) Glucose< 1e+02 40 7 neg (0.825 0.175) * 23) Glucose>=1e+02 104 48 pos (0.462 0.538)
46) DPF< 0.56 71 31 neg (0.563 0.437)

47) DPF>=0.43 12 2 neg (0.833 0.167) *
48) DPF< 0.43 59 29 neg (0.508 0.492)

49) BMI>=35 25 7 neg (0.720 0.280)
372) Insulin< 1.8e+02 16 2 neg (0.875 0.125) *373) Insulin>=1.8e+02 9 4 pos (0.444 0.556)*
50) BMI< 35 34 12 pos (0.353 0.647)
374) Age>=44 9 3 neg (0.667 0.333) *375) Age< 44 25 6 pos (0.240 0.760)*
750) DPF< 0.19 7 3 neg (0.571 0.429) 751) DPF>=0.19 18 2 pos (0.111 0.889) *
51) DPF>=0.56 33 8 pos (0.242 0.758) *
3) Glucose>=1.3e+02 200 75 pos (0.375 0.625)

10) BMI< 30 53 18 neg (0.660 0.340)

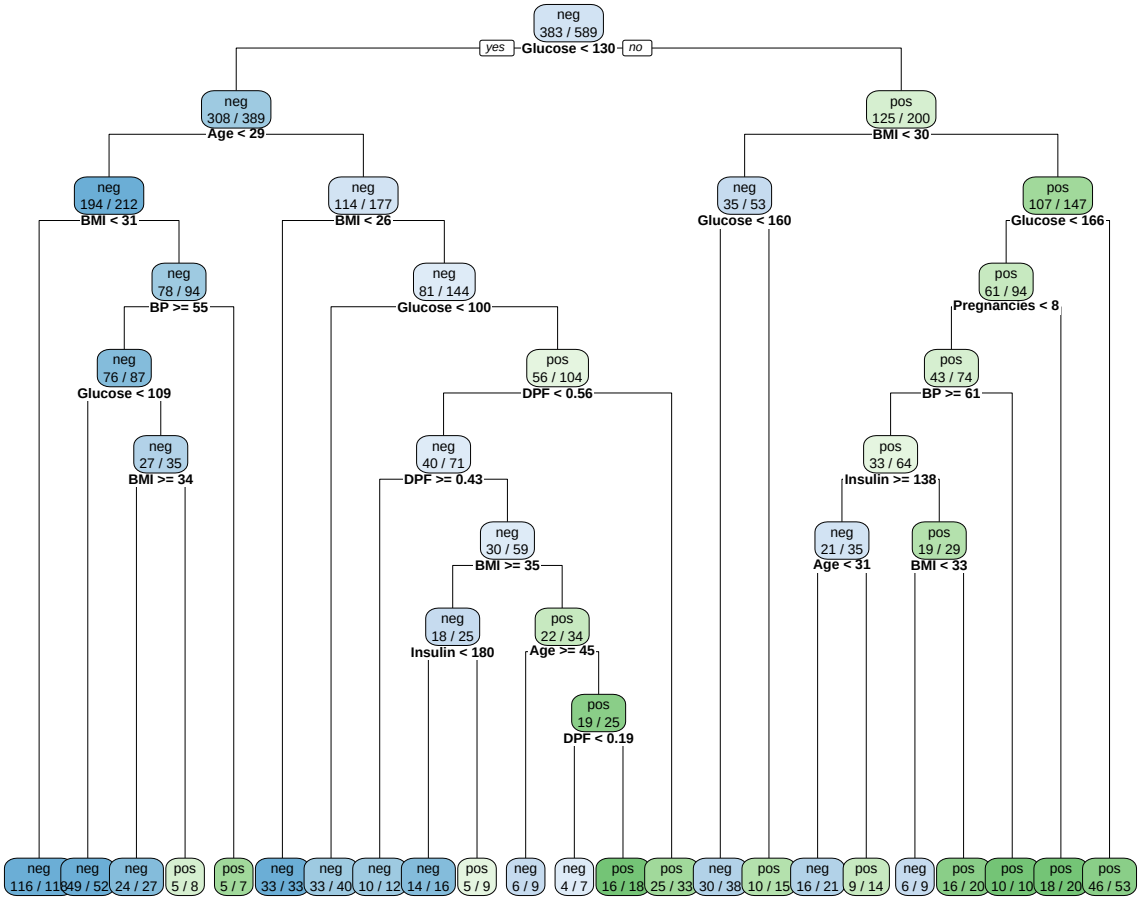
11) Glucose< 1.6e+02 38 8 neg (0.789 0.211) *
12) Glucose>=1.6e+02 15 5 pos (0.333 0.667) *
13) BMI>=30 147 40 pos (0.272 0.728)

14) Glucose< 1.7e+02 94 33 pos (0.351 0.649)
28) Pregnancies< 7.5 74 31 pos (0.419 0.581)
56) BP>=61 64 31 pos (0.484 0.516)

57) Insulin>=1.4e+02 35 14 neg (0.600 0.400)

58) Age< 31 21 5 neg (0.762 0.238) *
59) Age>=31 14 5 pos (0.357 0.643) *
60) Insulin< 1.4e+02 29 10 pos (0.345 0.655)

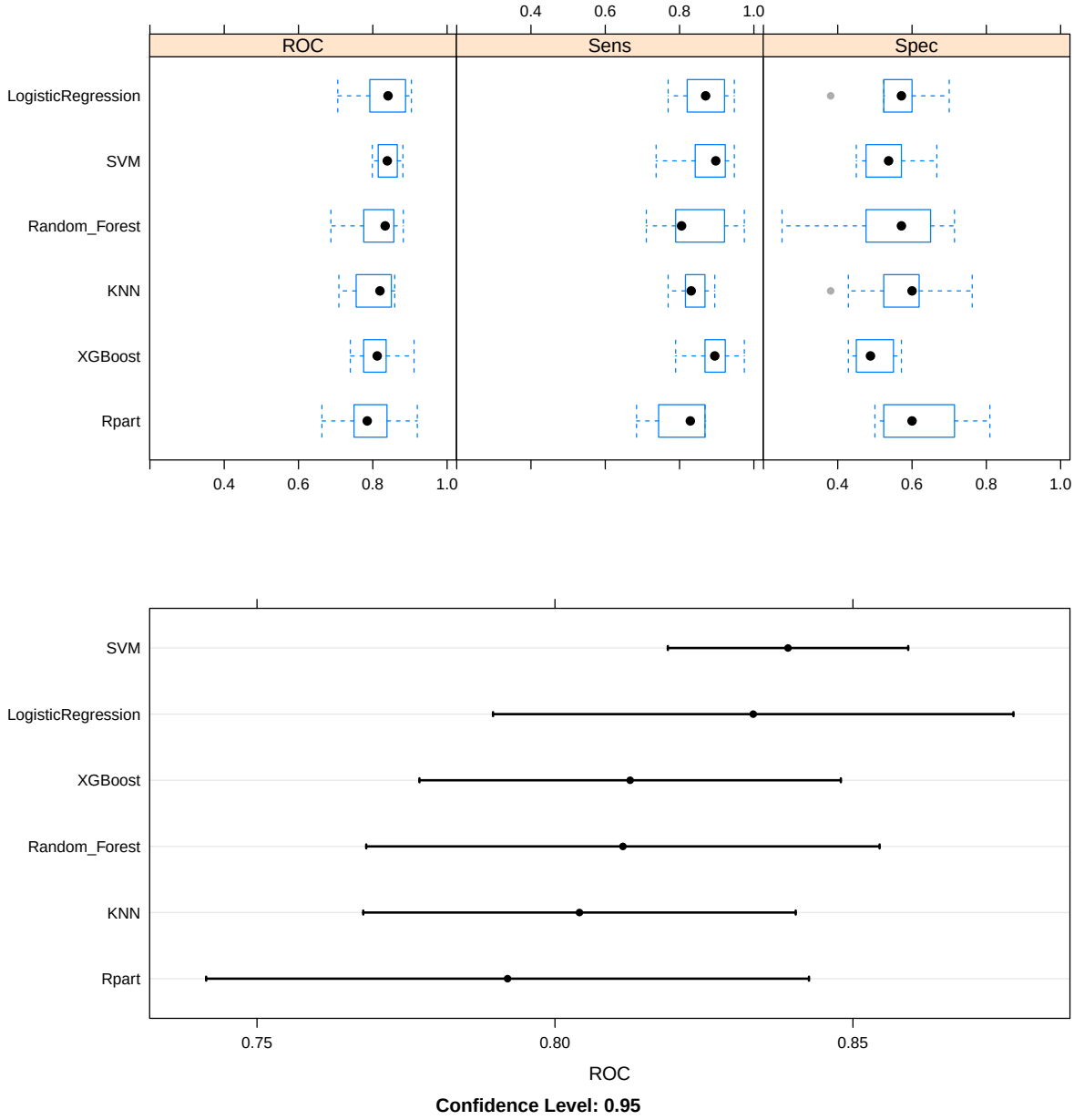
61) BMI< 33 9 3 neg (0.667 0.333) *
62) BMI>=33 20 4 pos (0.200 0.800) *
63) BP< 61 10 0 pos (0.000 1.000) *
29) Pregnancies>=7.5 20 2 pos (0.100 0.900) *
15) Glucose>=1.7e+02 53 7 pos (0.132 0.868) *



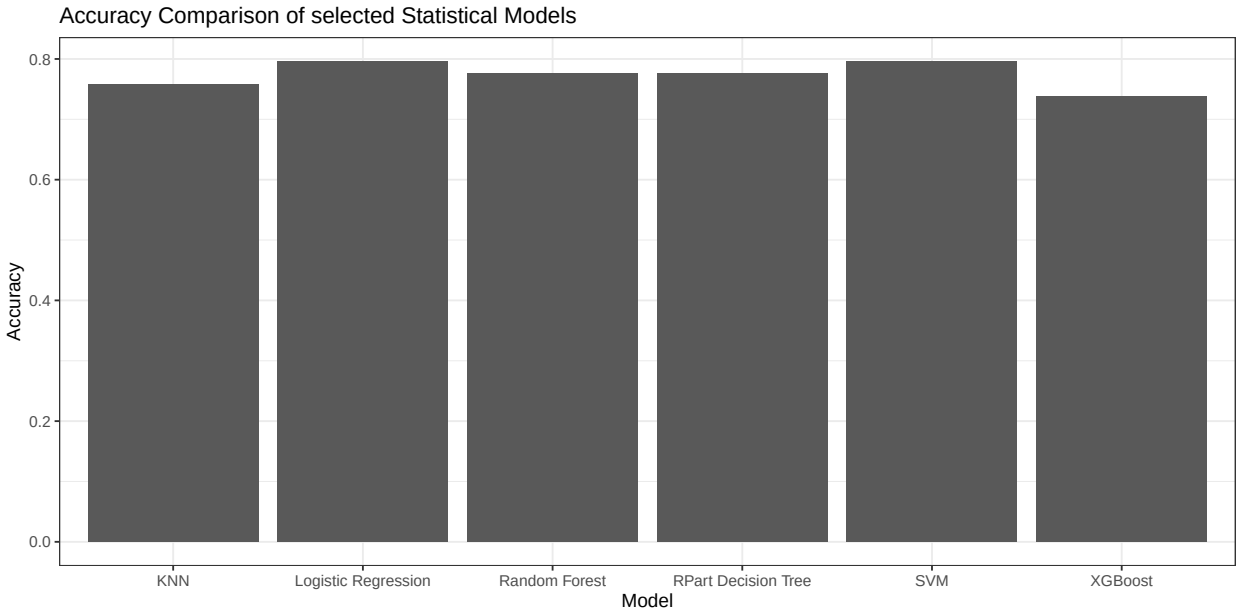
- We see from the ROC (Cross-Validation) vs Complexity parameter chart that as we hit a threshold of 0.2 on the model complexity, the ROC drastically drops from 0.70 to under 0.60.
- From the Decision Tree diagram, we see that Glucose values of more than 124 mg/dL play a major role in the predicting the probable outcome of Diabetes, more than any other factors and only followed by BMI and Age. Other attributes such as Diabetes Pedigree Function, Pregnancies, Blood Pressure, Skin Thickness and Insulin also contribute to the prediction.

Model Comparison

Now that we have trained and tested six models, lets compare the results from each model during both Training and Testing phase to see which performed better than the others.



Based on resampling fit results, we can see that Support Vector Machines model performed best in the training the data set among all of the six statistical learning models.



Core Algorithm

We obtain the highest accuracy from the Support Vector Machines model, with the score reaching 0.8. This implies that our model predicted classified correctly 80% of the times.

The Precision score stood at 0.74, implying our model correctly classified observations with high risk in the high risk category 74% of the times. The Recall stood at 0.64. We also have an F1 score of 0.69, which is defined as the harmonic mean of precision and recall and assigns equal weight to both metrics. However, for our analysis it is more important for the model to have low false negative cases, as it will be dangerous to classify high risk patients in low risk category. Thus, we will prioritize looking individually at both Precision and Recall rather than F1.

Based on all these findings, we have decided to select the Support Vector Machines model as our core model to proceed with tuning and improving the prediction.

Core Algorithm Parameter Tuning

The Support Vector Machine is a very useful classification technique. SVM methods can handle both linear and non-linear class boundaries. It can be used for both two-class and multi-class classification problems. In real life data, the separation boundary in most datasets is generally nonlinear. Technically, the SVM algorithm perform a non-linear classification using what is called the kernel trick. The most commonly used kernel transformations are polynomial kernel and radial kernel.

Now that we have selected our core algorithm to be Support Vector Machines, we will define the parameters to be tuned and the type of resampling next.

For our model we will be using 10-fold cross-validation resampling methods. We have also selected ROC as the performance metric for the resampling profile. We have normalized the variables to make their scale comparable before building the SVM classifier by setting the option preProcess = c("center","scale")

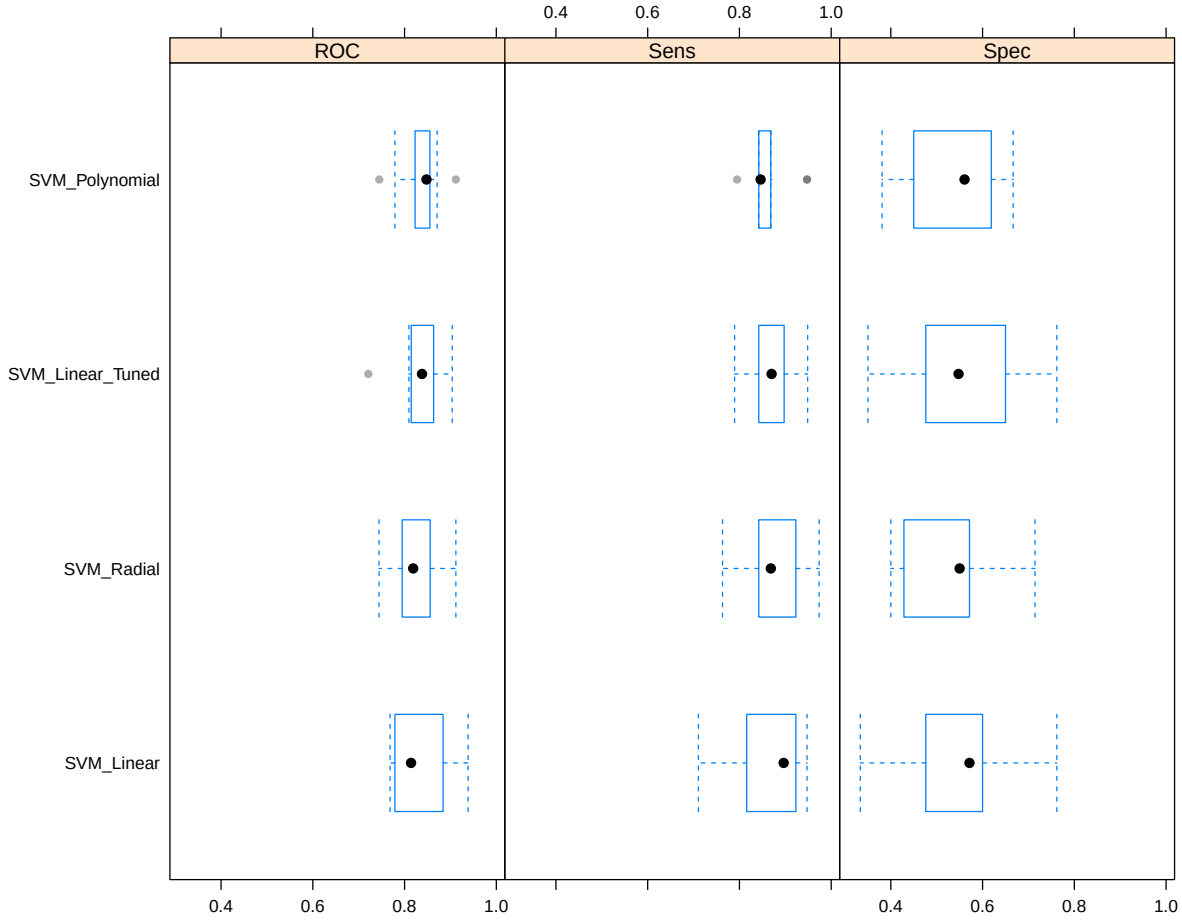
Next we are going to fit the data into the following models:

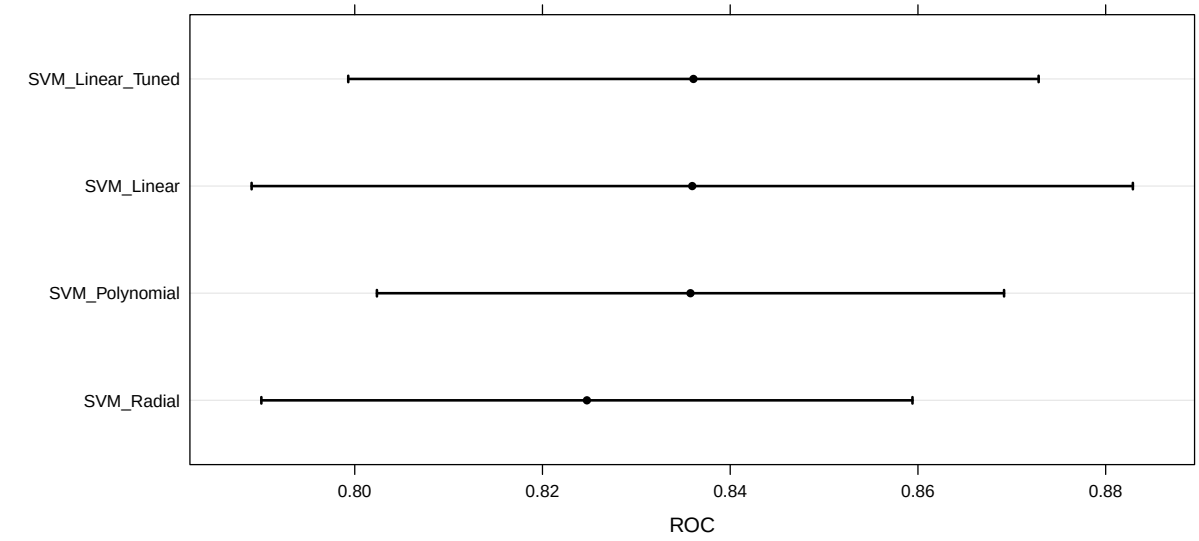
- SVM Linear 10-fold CV
- SVM Linear 10-fold CV Tuned
- SVM Radial kernel
- SVM Polynomial kernel

For each model, we have made slight adjustments in order to apply alterations to the data training and determine how they improve upon the base core SVM linear testing. For our parameter tuning, we include a parameter that allows us to set the particular values that the main default variables will be using. The radial Kernel function is used for non-linear decision boundary transformation, and we have altered the tuneLength in order to inform the algorithm to use (in our case) 10 different default values. Our polynomial kernel is of similar but otherwise defined in its arbitrary boundary sequencing. Given these models, we will now make statistical statements about their performance differences by collecting the resampling results using 'resamples'

```
Call:
summary.resamples(object = resamps)

Models: SVM_Linear, SVM_Linear_Tuned, SVM_Radial, SVM_Polynomial
Number of resamples: 10

ROC
                 Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
SVM_Linear       0.77    0.78   0.81 0.84    0.88 0.94    0
SVM_Linear_Tuned 0.72    0.82   0.84 0.84    0.86 0.90    0
SVM_Radial       0.74    0.80   0.82 0.82    0.85 0.91    0
SVM_Polynomial   0.74    0.82   0.85 0.84    0.85 0.91    0

Sens
                 Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
SVM_Linear       0.71    0.82   0.90 0.87    0.92 0.95    0
SVM_Linear_Tuned 0.79    0.85   0.87 0.87    0.90 0.95    0
SVM_Radial       0.76    0.84   0.87 0.87    0.91 0.97    0
SVM_Polynomial   0.79    0.84   0.85 0.86    0.87 0.95    0

Spec
                 Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
SVM_Linear       0.33    0.48   0.57 0.55    0.60 0.76    0
SVM_Linear_Tuned 0.35    0.48   0.55 0.55    0.64 0.76    0
SVM_Radial       0.40    0.45   0.55 0.53    0.57 0.71    0
SVM_Polynomial   0.38    0.45   0.56 0.54    0.61 0.67    0
```

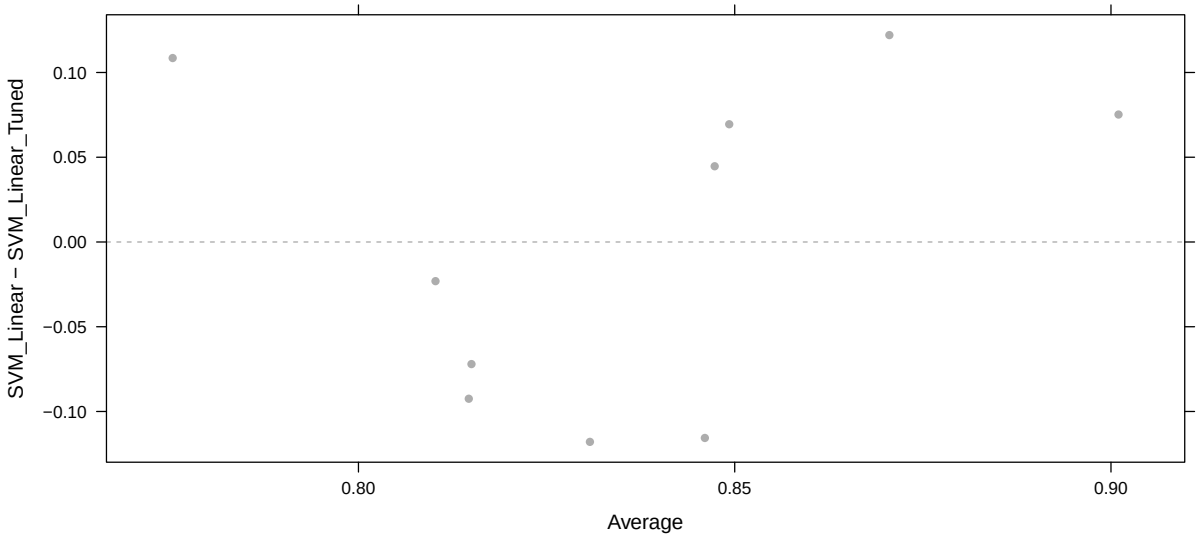
We observe the ROC, Sensitivity, and Specificity values for the different models above. Notice the zero values that will not be included. We can achive higher ROC, sensitivity and specificity but that comes at a cost. Now that we have the resampling results, we are going to create several lattice plot methods to visualize the resampling distributions: density plots, box-whisker plots and scatterplot matrices. In the below box-whisker plot, each of the SVM models is compared with their respective box-whsiker plots for ROC, Sensitivity and Specificity.

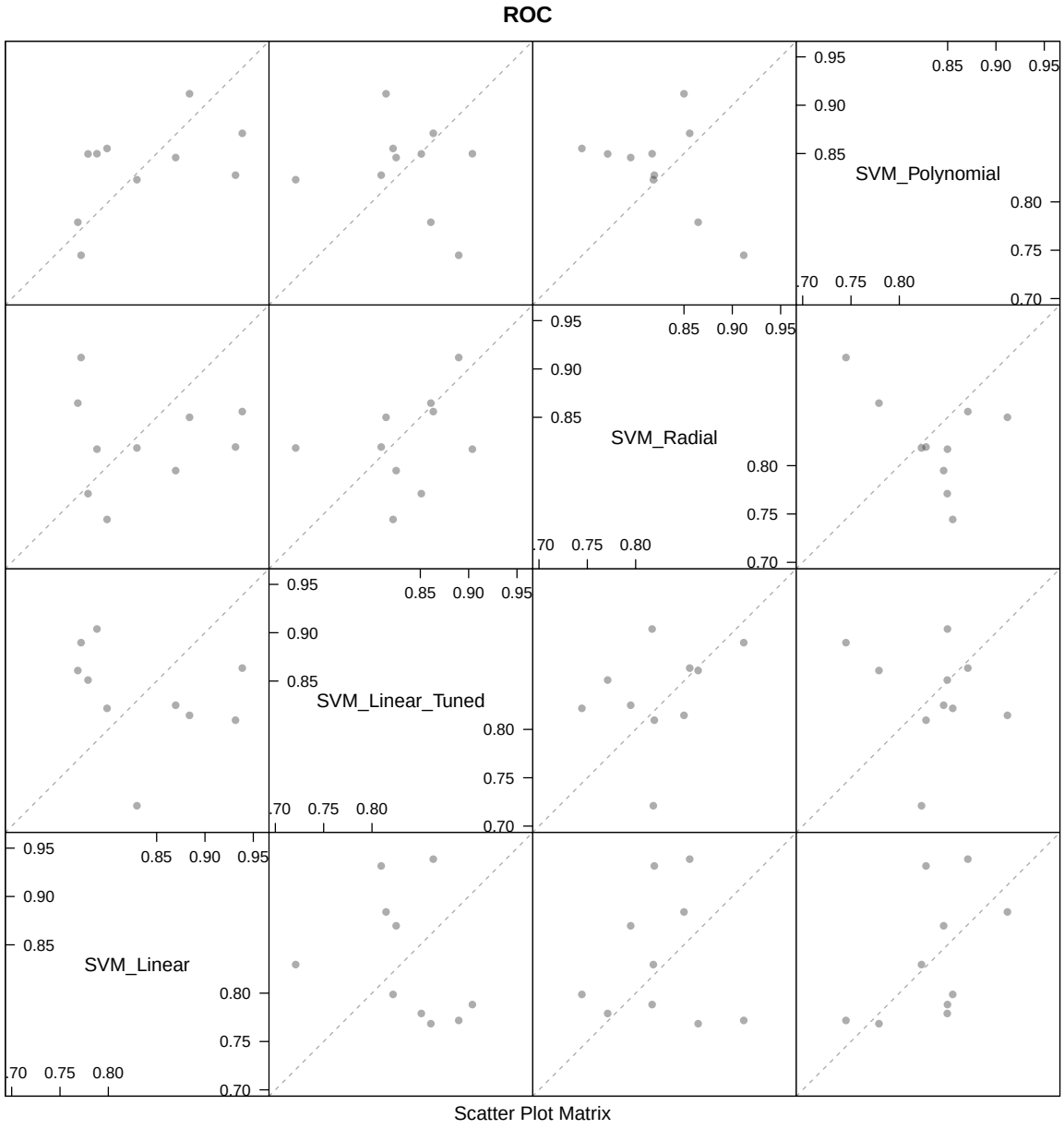




Confidence Level: 0.95

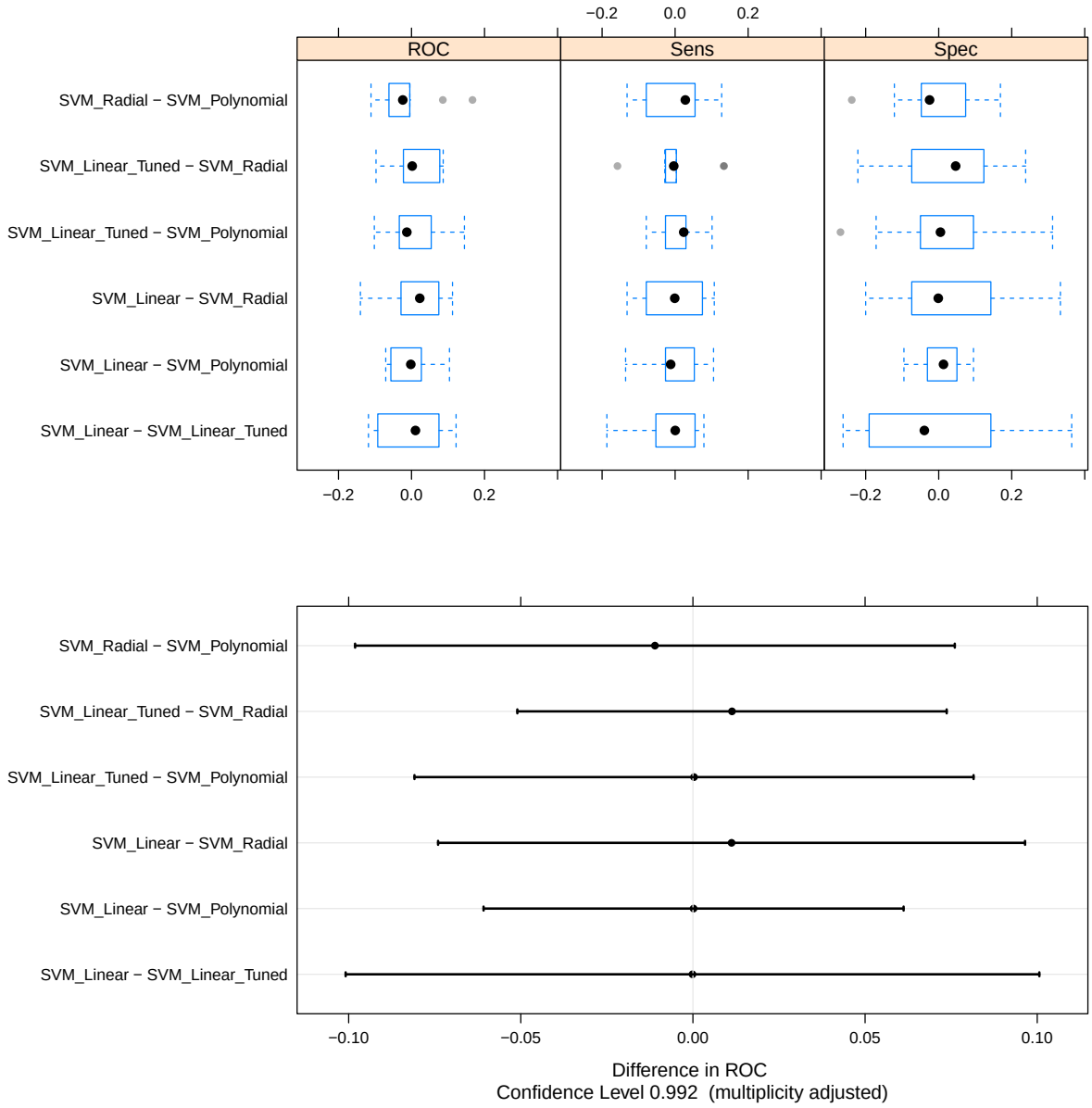
ROC





Above, we see in the dot-plot that the mean of SVM Polynomial ROC value is the highest. The confidence level for this plot has been set at 0.95. But as mentioned earlier, when Polynomial model is used, we see increased cost. we will fit the model to see what is the cost impact next. The subsequent scatterplot matrix is being utilized to compare each of the SVM model depending on their ROC values.

Since models are fit on the same versions of the training data, it makes more sense to make inferences on the differences between models. In this way we reduce any within-resample correlation that may exist. We can compute the differences, then use a simple t-test to evaluate the null hypothesis to conclude that there is no difference between models. Also, we see that the similarity between linear models and radial with a narrow spread in the ROC difference which is closer to zero.



We plot the box-whisker plot below to compare the differences along the same parameters for each SVM model. While the graphs look different, the values are very minute, with 0.2 being the maximum value difference in either direction. And finally, another dot plot is utilized to check and compute the differences. In this case, the confidence level has been readjusted to 0.992 and shows us a difference in the ROC values of each of the models that we have used. We see almost all the differences are higher than 0.05.

Model Fittings

Now that we have compared the ROC, Sensitivity and Specificity, let's look at the other metrics to understand how the parameter tuning and the other methods compare to each other.

Initially, we will check the best parameter tuning for each SVM model we are utilizing. After which, we will analyze the results that we have found. Going exclusively on the accuracy we can observe that SVM Linear and SVM Linear tuned models provide us with similar results. But if we consider the F1 score SVM Radical is higher than the rest of the SVM models.

A four-fold plot is used to provide better insights and check the accuracy of the different SVM models. The four-fold plot consists of the Positive and Negative values of Reference and Prediction.

SVM Linear CV and SVM Linear CV Parameter Tuned

Table 7: Best Tune Parameter SVM - Linear CV

C
0.1

We see that the for the best tune, the cost parameter C is the minimal at 0.1.

Both the base linear and tuned linear classifications have similar attributes and average means of values, with slight variations and divergences in scope and where the cost parameter of the "tuned" SVM is held tighter at a higher value of 0.5. There should be nearly no changes at all in resulting accuracy.

SVM Radial Kernel

Table 8: Best Tune Parameter SVM - Radial

sigma	C
0.12	0.25

The sigma value determines the "reach" of our training instances, and because of its low value the results of the classification is bound to be more linear. we can see that Radial is a bit pricey compared to Linear 10-fold model.

SVM Polynomial Kernel

```
line search fails -0.034 0.56 0.000011 -0.000000026 -0.0000000007 -0.00000000096 -0.000000000000000075lin
```

Table 9: Best Tune Parameter SVM - Polynomial

	degree	scale	C
39	3	0.01	1

While the outcome should be relatively similar, we can anticipate slight differences due to configuration differences in the models and the preprocessing for c and the tuning options. We see that for SVM Polynomial model the cost parameter is much higher at 1 and thus as discussed earlier this suggests that to attain similar or better performance, we will be spending higher cost

Parameter Tuning Results Comparison

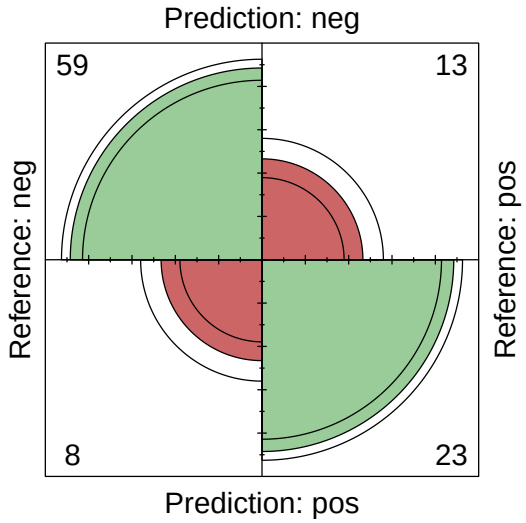
Table 10: Parameter Tuning Results Comparison

	Recall/Sensitivity	F1	Precision	AUC	Accuracy
results_support_vector_linear_cv	0.64	0.69	0.74	0.85	0.80
results_support_vector_linear_cv_tune	0.64	0.69	0.74	0.85	0.80
results_support_vector_radial	0.58	0.67	0.78	0.81	0.80
results_support_vector_polynomial	0.64	0.71	0.79	0.86	0.82

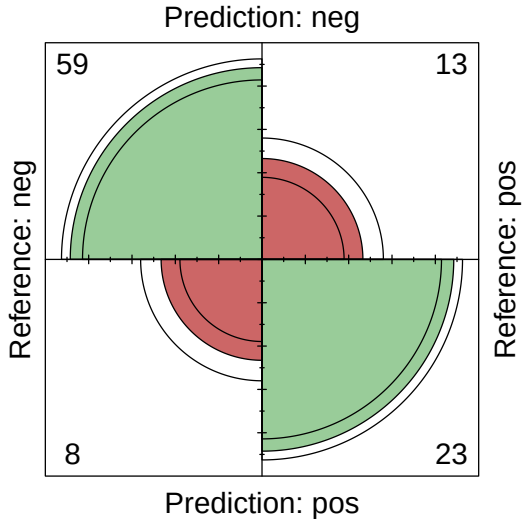
Looking at the results, we see how similar the results are and how this matches up with our prior predictions with parameter tuning. SVM radial appears to be the most different, likely as a result of the processing and exponential nature of radial. As expected, polynomial kernel yields a slightly higher accuracy along with the higher recall, F1, and precision values as linear methods, comparing the runtime cost, it is not cost effective to adapt SVM Polynomial Kernel Model.

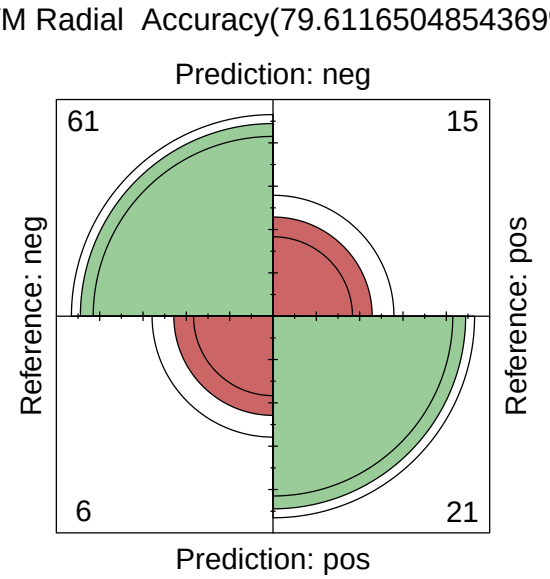
Accuracy Diagrams

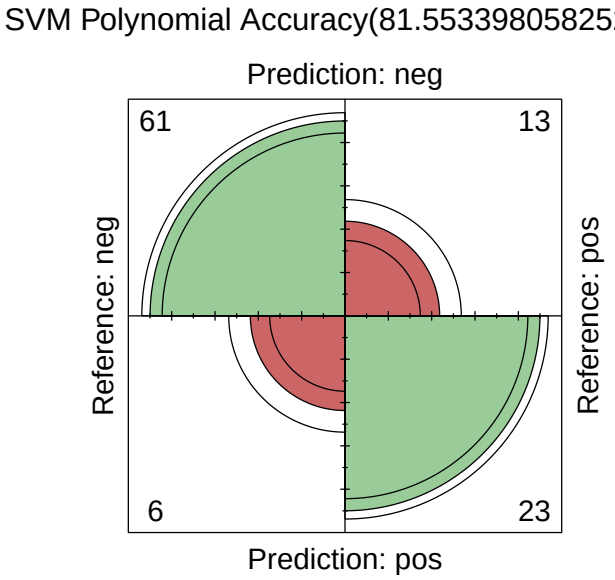
SVM Linear Accuracy(79.6116504854369%)



SVM Linear Tuned Accuracy(79.6116504854369%)







SVM Linear and SVM Linear Tuned models provide us with a similar accuracy of 79.61% and 79.61% while SVM Radical and SVM Polynomial has an accuracy of 79.61% and 81.55% respectively. The resulting diagrams will have a very similar appearance with only the False Negative value having slight changes. The single variable value appears in the False Negative sector, where radial actually yields a higher value of 15 comparatively to the 13 returned by the other kernels. The highest of the algorithm accuracies involves tuning in the kernel configuration. Considering all models have similar accuracies, precision and area under the curve, we can choose any of these model for predicting the Diabetic outcome. But cost efficient model will be Best Tuned SVM Linear 10-fold CV model.

Conclusion

The PIMA Indian Diabetes dataset which is a collection of 768 entries organized by 9 variables is used to predict the occurrence of diabetes in PIMA Indian women. This dataset consists of an aggregated compilation of various biometric information of females below 21 years old. Analyzing these parameters allow us to

discover trends and then eventually to predict the presence and likelihood of diabetes. We then fit the data into various algorithms, which included Logistic Regression, KNN, XGBOOST, SVM, Random Forest, and Rpart CART.

At an initial glance, the dataset appears to be simple, consisting of relatively few entries. But through our exploratory data analysis we understood the data better than before and realized that we had zero values, numerous outliers, etc. We used MICE imputation method to replace zero values with imputed values. Then we eliminated outliers through IQR method.

During our EDA, we performed numerous univariate analysis and a few bi-variate analyses to understand more about each of the variables involved in the dataset. We found that, each variable is interlinked and has a different effect on determining diabetes. Of all the factors involved, glucose was the one that played the greatest role when interacting with all the other variables. Meanwhile our data also showed that BMI (body mass index), BP (blood pressure), and skin thickness all played a role in diabetes, while other variables like Insulin or DPF (a predictor measure involving occurrence of diabetes in relatives) did not show a significant connection to diabetes. Bi-variate analysis of chosen variables to determine relation to each other and to diabetes were studies next and found that more combined factors played further stronger role in occurrence of diabetes.

Our next step was to decide which algorithm to be used for fitting, training, and testing for better prediction of diabetes with high accuracy. In order to begin testing, we first split off 85% of the dataset (589 observations) as training data to run through the algorithms and train the model and then used the remaining 15% of the dataset (103 observations) as our testing data that would deliver our results. We chose these values in part to provide sufficient training data for the algorithms to increase accuracy as well as help in performance when compiling in R, despite our relatively smaller database size.

Once we trained various models, we compared the precision, recall/sensitivity, F1 score, and accuracy for each of the models to determine which model performed better than the other. Upon validating, we found that the Support Vector Machines model performed better than other models with an accuracy of 80% and better Recall and AUC values. Logistic regression came as the closest second.

Now that we had our core model identified, we proceeded forward to tune the SVM model to obtain the better accuracy and recall results. We used four different approaches: the linear kernel, tuned linear, radial kernel, and polynomial kernel approaches. Out of the four, while polynomial's accuracy score was the highest, it also came at a higher runtime cost. Whereas, Linear 10-Fold, Linear 10-Fold best tuned and Radial all produced identical accuracy at much lower runtime cost. Thus, if we had to choose one of the four models, we will choose SVM Linear 10-Fold model for prediction as it has the best accuracy at the lowest cost.

To wrap things up, we can make a few brief comments on how the process of interacting closely with this dataset has allowed us to gain a better understanding of a condition that is so prevalent not just in the targeted focus of our dataset but also in the entire world. With the lessons learned from the Coronavirus outbreak and subsequent response, it would be prudent to use those experiences and pay attention to diseases and health conditions that can be just as debilitating, if not as widespread or infectious. As mentioned above, our findings can be used to help in understanding the link between diabetes and some of its causes, and perhaps translate this knowledge into other fields like heart problems or other BMI/blood pressure related diseases.