Assignment- 3

Q.1)

Another variable of class integer is created called res which represents 1 if crime rate for the suburb is above median and 0 otherwise. Classification is done for those two classes using LDA, logistic regression and knn. Data is divided in test and train set in the ratio of 1:1(253 rows in test and 253 to fit the model i.e. train set). The fitted model is used to predict results of the test set data.

Correlation between the variables is calculated against crime rate. Now classification is done for the best 5,10 and all the variables.

For 5 variables,

Logistic regression for the test set predicts that out of 253 suburbs in boston,124 has crime rate below median and for remaining 129 has error above median. Error in logistic regression is very less that is 1.18%.

LDA for the test set predicts that out of 253 suburbs in boston,165 has crime rate below median and for remaining 88 has error above median. While for LDA error is the maximum which is around 16.6%.

knn performed with k=5 to 14. Knn error is least for k=7 which is around 9.5%.

For 10 variables,

Logistic regression for the test set predicts that out of 253 suburbs in boston,124 has crime rate below median and for remaining 129 has error above median but has a little bit more error than for the 5 variable classification. Error in logistic regression is very less as compared to LDA and knn, that is 1.97%.

LDA for the test set predicts that out of 253 suburbs in boston,156 has crime rate below median and for remaining 97 has error above median. While for LDA error is the maximum which is around 14.6%.

knn performed with k=5 to 14. Knn error is least for k=5 which is around 9.5%

For all the variables,

Logistic regression for the test set predicts that out of 253 suburbs in boston,119 has crime rate below median and for remaining 134 has error above median but has a little bit more error than for the 5 and 10 variable classification. Error in logistic regression is very less as compared to LDA and knn, that is 4.74%.

LDA for the test set predicts that out of 253 suburbs in boston,150 has crime rate below median and for remaining 103 has error above median. While for LDA error is again maximum following the trend for 5 and 10 variables, which is around 16.2%.

knn performed with k=5 to 14. Knn error is least for k=5 which is around 8.6%

So we can conclude that error with classification using logistic regression in increases with the no of predictors. For most related minimum predictors error is minimum but it fits the model better than other methods for two classes.

LDA error is the maximum for any number of variables as compared to LR and knn. It doesn't follow any fix trend as that of LR. But roughly we can conclude that error decreases as no of related variables are increased.

Knn error varies with no of neighbors values. It gives high accuracy with less no of neighbors. As k value increases error goes on increasing. Error for least NN is still less than LDA but higher than LR.

So, to infer the problem, logistic regression is the best method for classification.

Q.2)

a)

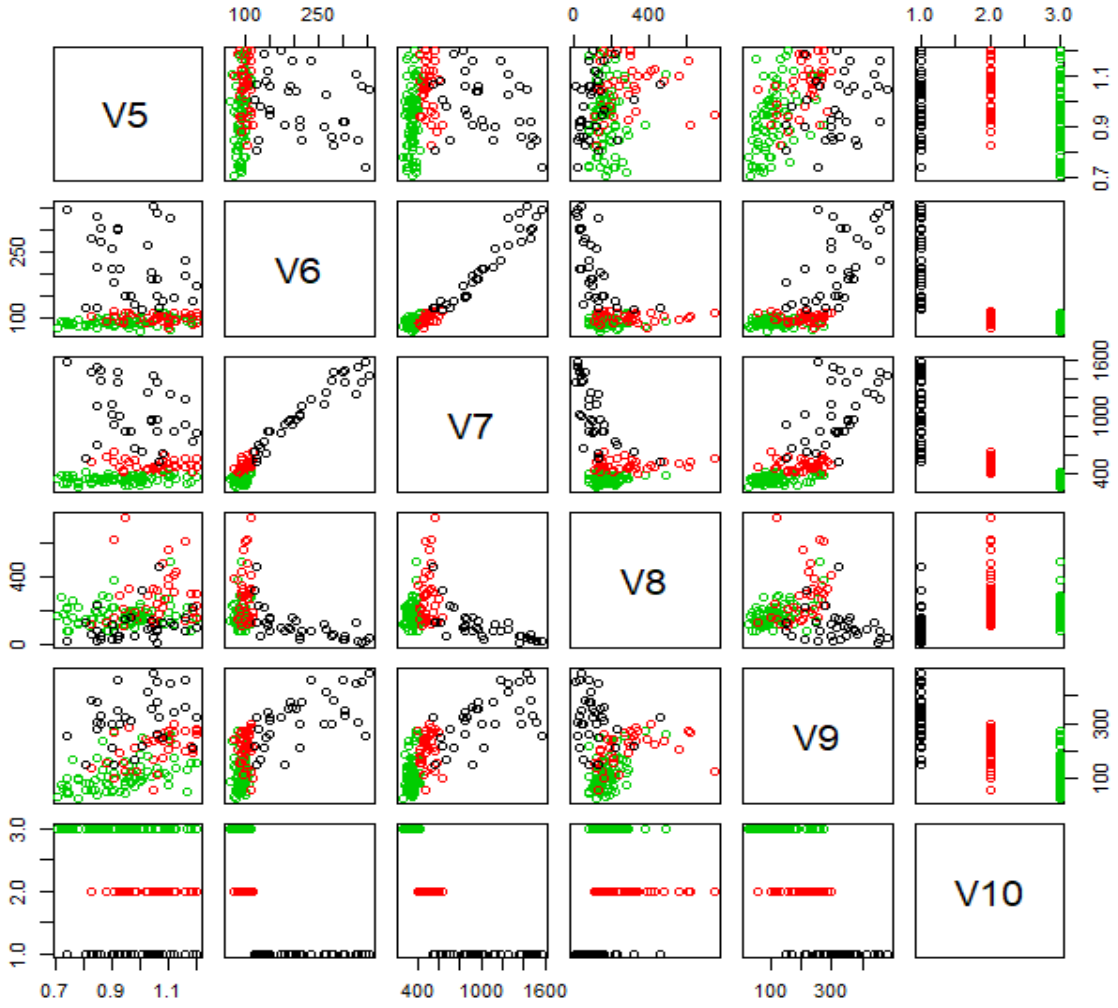


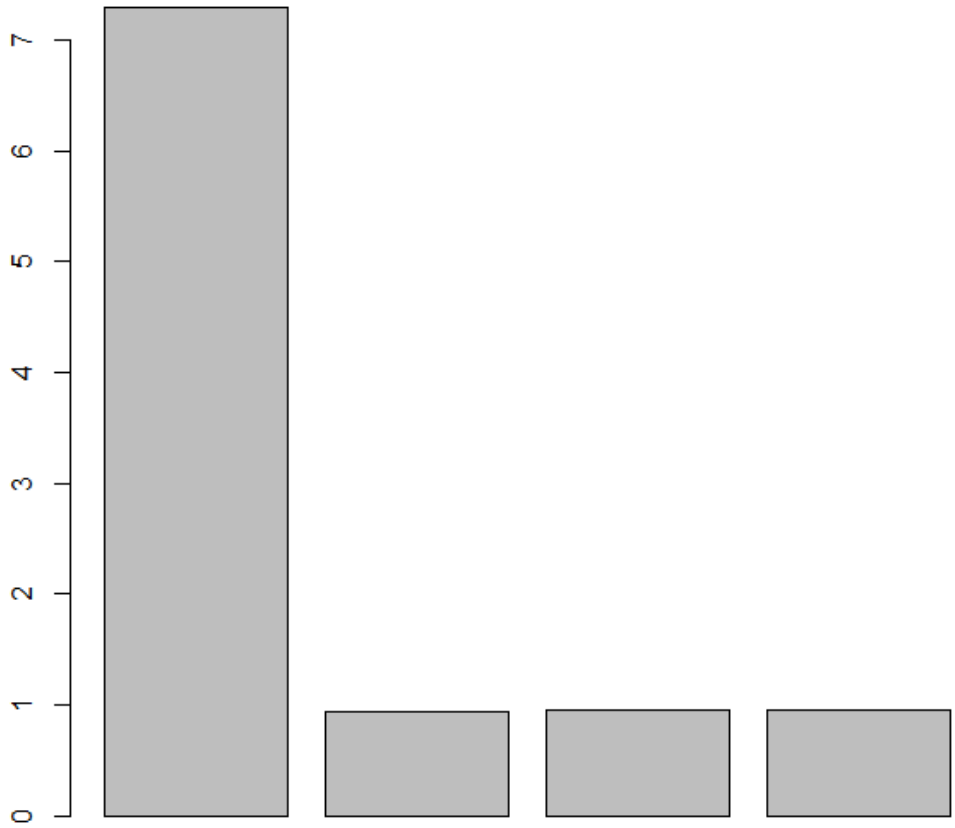
Figure above shows pairwise scatterplot for all the variables. From above graph we can interpret that, variables insulin and SSPG are highly correlated with respect to the classification. Then again fasting plasma glucose and insulin area are correlated.

b)When model is fitted with dataframe containing 5 variables and class variable, LDA error is around 16% while QDA error is 12%. We can conclude that for multiclass classification, QDA classifies data with more accuracy than LDA.

c)

Both LDA and QDA agrees on the data. Both classifies the given data in class 2.

Q.4)



The model with 2 degree polynomial equation has the least LOOCV value.

In LOOCV, we leave one row and fits data with other all rows. Now the left row acts as test data on which we test the model, calculate the error. The process is repeated for all n rows. Average of errors is called as LOOCV (Leave Out One Cross Validation Error).

The expected least error was for model with polynomial equation with degree 4. The error for degree 2,3,4 are comparable and fit model precisely.

Above graph shows comparison of errors for all the 4 models.

Table below shows coefficient values we got by fitting model using OLS for different equations

	B_o	B_1	B_2	B_3	B_4
Y = β0 + β1X + ε	-1.6254	0.6925			
Y = β0 + β1X + β2X^2+ ε	-1.550	6.189	-23.948		
Y = β0 + β1X + β2X^2 + β3X^3+ ε	-1.5500	6.1888	-23.9483	0.2641	
Y = β0 + β1X + β2X^2 + β3X^3 +β4X^4 + ε	-1.5500	6.1888	-23.9483	0.2641	1.2571

The p-values show that the linear and quadratic terms are statistically significants and that the cubic and 4th degree terms are not statistically significants. That is why error value for quadratic equation is least which is also observed in LOOCV error.