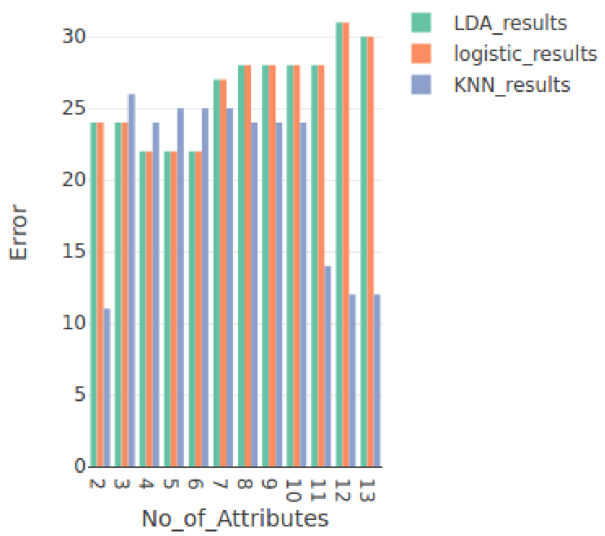
Sai Kiran Putta

Problem 1)

- Doing sanity check, sum(is.na()) gives us 0 missing values in the data.
- Convert the feature crim as categorical. This is done by checking if each observation is less or greater than the mean of feature.
- Generate train and test datasets randomly.
- Different subsets of features are generated by forward subset selection. After the subsets are selected, those features are used to fit Logistic Regression, LDA and KNN models.
- For KNN classification, for a given feature set k values 3,5,7,9 are used and the minimum error of those is considered as the final Leastsquares value for that feature subset.
- The results are as follows.

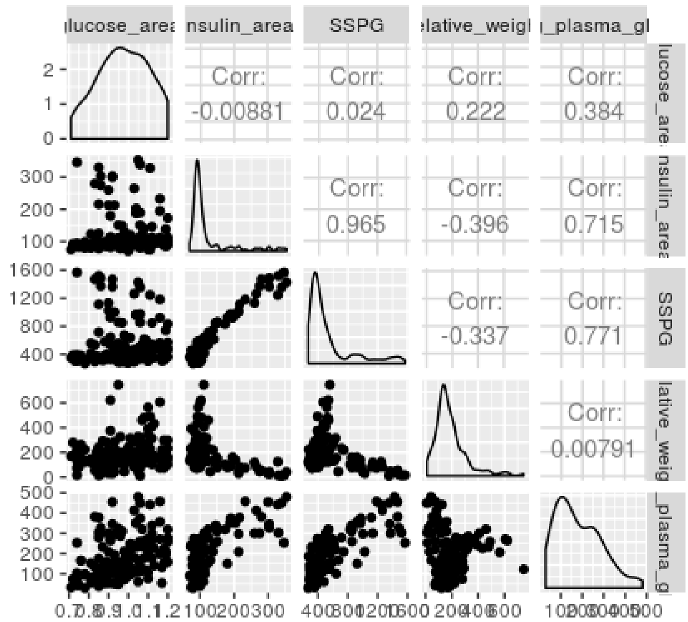


Inference: For all the subset of features, LDA and Logistic are performing the same way. KNN is almost equal or doing much better than the rest as it can be seen from above figure.

Problem –2)

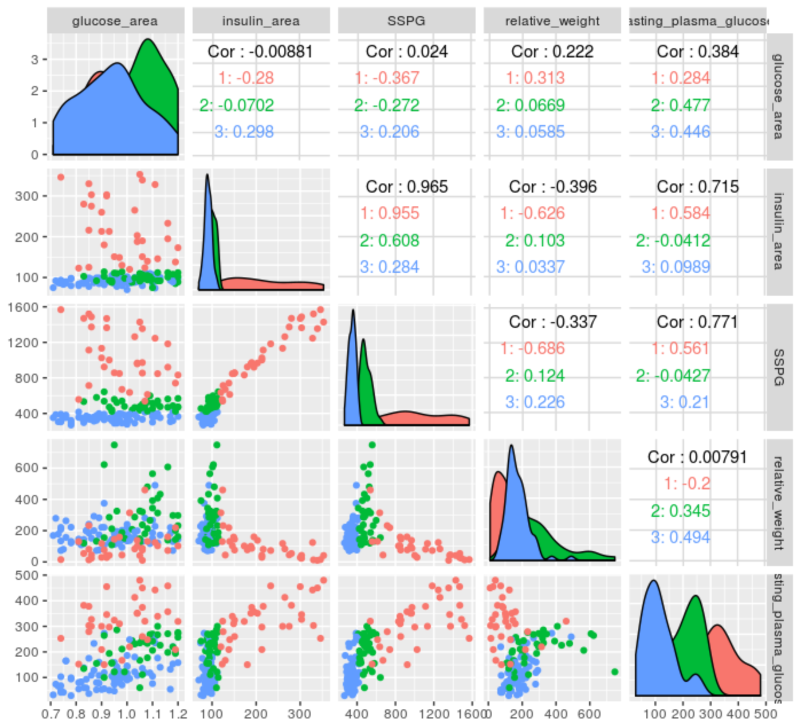
- Doing a sanity check on the data by, sum(is.na()) we get that there are no missing values in the data.
- Plotting a pair plot we get the following.

Plotting the pairplots for the complete feature set gives the following.



Inference: Almost all the features are normally distributed with right tails.

Plotting the pair plots with different classes seperately.



Inference: Looking at the feature like SSPG it is evidenent that distribution of individual classes is different. Looking at the scatter plots combination like insulin_area and SSPG it is evident that the classes are separable.

Looking at the results of models we are correct to infer that classes are separable.

b)
lda_error = 10
qda_error = 3
Error from qda is less. Hence it is better for this data.

c)
LDA prediction – 3
QDA prediction – 2

Problem – 4)
The errors from the models are as follows.
```
7.28474411546929
0.937178917181123
0.956253813731321
0.953445283156601
```

The error for 2nd degree polynomial is the minimum. The result is expected since the degree of the polynomial was 2 when we created the data.

Doing a summary of the model object we see that the coefficients of the models are low for degree – 1 and degree – 2 which agrees with the cross validation results.
It is significant for degree – 1 but for degree – 2 it is very low indicating how well the model fits the data.

3)

a) let's prove this using induction.

let $k=1$ i.e., there is only one class in the dataset.

That is, let n = total no. of rows in the dataset.

if $k=1 \Rightarrow$ probability of that class being chosen $= \frac{n}{n} = 1$

let $k=2$, i.e., there are two classes in the dataset.

The Complete probability of any of the class being chosen is,

let k be the no. of observations from class 1, i.e., $(n-k)$ are the observations from class 2.

Sum of posterior probabilities $= \frac{k}{n} + \frac{(n-k)}{n}$

$$= \frac{k+n-k}{n} = 1$$

let $k=(z-1)$ i.e., there are $(z-1)$ classes.

Sum of posterior probabilities is given by.

$$\frac{1}{n}\sum_{i=1}^{z-1} k_i + (n-k_i) = 1.$$

Hence by induction, Sum of posterior probabilities of classes is equal to one.

Scanned by CamScanner

5) Let's begin with,

$$P(x) = \frac{e^{(\beta_0+\beta_1 x)}}{1+e^{\beta_0+\beta_1 x}}$$

let $Y(x) = e^{\beta_0+\beta_1 x}$.

Then $P(x) = \dfrac{Y(x)}{1+Y(x)}$

So, $p(x)\left(1+Y(x)\right) = Y(x)$.

$$\Rightarrow P(x) = Y(x) - P(x)Y(x)$$
$$= \left(1-P(x)\right)Y(x)$$

Dividing by $1-P(x)$ by both sides we get.

$$\frac{p(x)}{1-P(x)} = Y(x)$$

$$\therefore \frac{P(x)}{1-P(x)} = e^{(\beta_0+\beta_1 x)}$$