Lab Report

Name: Devyani Gorkar
Roll no:B20ME027

Q1

Read the data

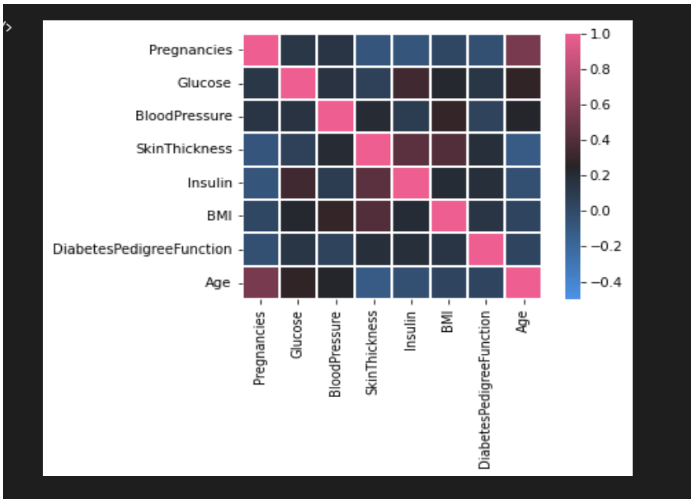
Q2

Split the data into 70:30 ratio using train test split.

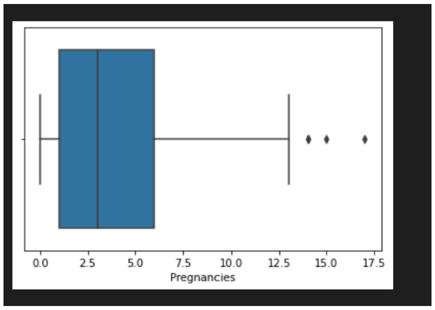
Q3.-

Preprocess the data and perform classification using the Naive Bayes classifier with inbuilt libraries.

Using seaborn library heat map we can check whether the features are independent or not.For applying bayes theorem it is important that the features of the dataset are independent.

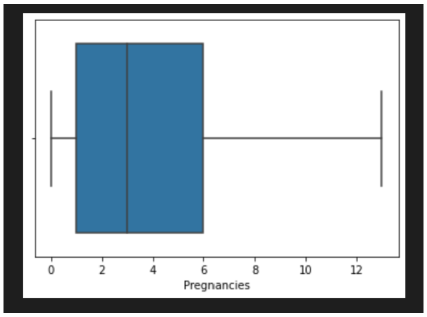


So it can be seen that dependencies is maximum across the diagonal and there are no other pink cells.This shows that the features are only dependent on themselves and hence naive bayes can used as a model.



On applying box plot to each feature ,outliers are observed in each feature which should be removed. Using the interquartile method for each feature we find the outliers and replace them with the mean of the data.

So after removing the outliers the box plot obtained is.

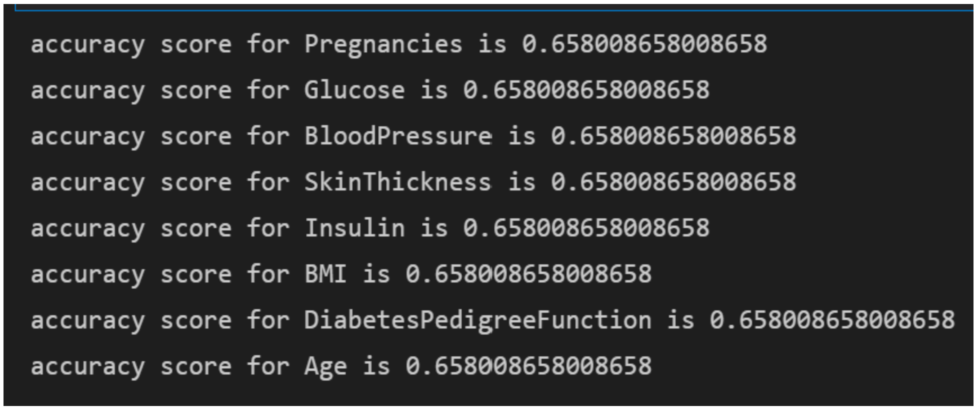


Hence the data is preprocessed as there are no outliers, no null values and no string values. Now we use the MultinomialNaiveBayes class of the Scikit learn library we do prediction.Accuracy obtained using library function is `0.670995670995671.`

Q-4

For binning KBinsDiscretizer from sklearn.preprocessing is used.Initially the number of bins kept are 10 and strategy for binning used is ‘uniform’.

Now a class called NB is created.Alpha is a constant kept to avoid the zero probability so that the result is not affected abruptly because of one zero.Prior is a list consisting of the prior probabilities of the two classes.count is a numpy array which stores the count of each value corresponding to a particular feature and the two classes.We defined a method prior_probability in the class Naive bayes .This function returns a list containing prior probabilities of each class.The next method is likelihood which takes X and y as input ,it returns a count matrix which an array of array.Each array represents the probabilities of a feature.The first row of each array stores probability P(x/w) when w=0 and next row stores probability when w=1.Then we predict output of each feature and accurac we get is.



```
accuracy score for Pregnancies is 0.658008658008658
accuracy score for Glucose is 0.658008658008658
accuracy score for BloodPressure is 0.658008658008658
accuracy score for SkinThickness is 0.658008658008658
accuracy score for Insulin is 0.658008658008658
accuracy score for BMI is 0.658008658008658
accuracy score for DiabetesPedigreeFunction is 0.658008658008658
accuracy score for Age is 0.658008658008658
```

On increasing the bin size the accuracy slightly increases because there is less loss of information as bin size is increased.

https://colab.research.google.com/drive/1IuKnK0ZdBFdwQkuI9T-pLPFFa3Wm29el?usp=sharing