CHRIST

(DEEMED TO BE UNIVERSITY)

BANGALORE · INDIA

MCA472B: Machine Learning

CAT3: Component 2:

By,

Athira T.P

2147244

4 MCA B

Write analysis of all the algorithms applied on the dataset chosen, visualization and interpretation with respect to the algorithms are mandatory.

Dataset Chosen: Diabetes Dataset

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

The datasets consist of several medical predictor variables and one target variable (Outcome) Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age etc and the class variable is (Outcome)

- Pregnancies: Number of times pregnant
- Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- BloodPressure: Diastolic blood pressure (mm Hg)
- SkinThickness: Triceps skin fold thickness (mm)
- Insulin: 2-Hour serum insulin (mu U/ml)

• BMI: Body mass index (weight in kg/(height in m)^2)

• DiabetesPedigreeFunction: Diabetes pedigree function • Age: Age (years) • Outcome: Class variable (0 or 1)

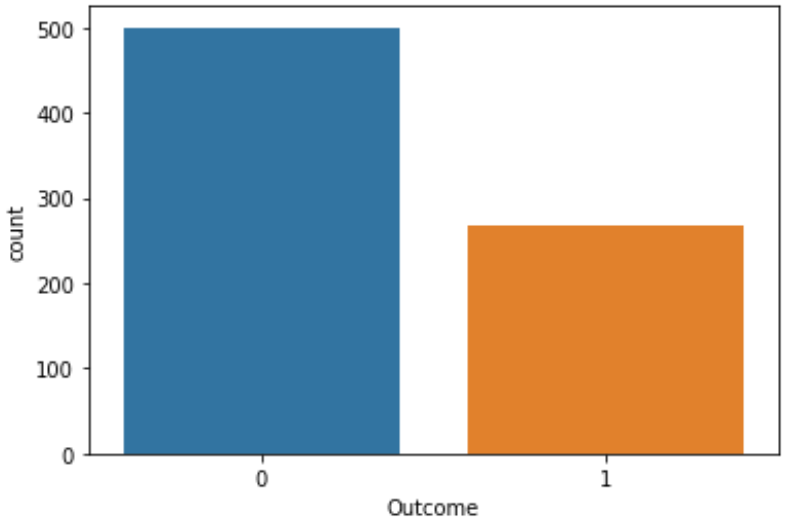
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

PIMA DIABETES DATASET

LAB1: EDA AND VISUALIZATION:

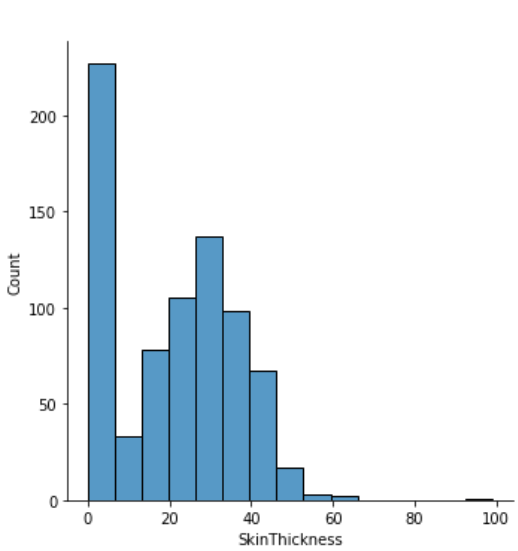
Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations. It is major step to fine-tune the given data set(s) in a different form of analysis to understand the insights of the key characteristics of various

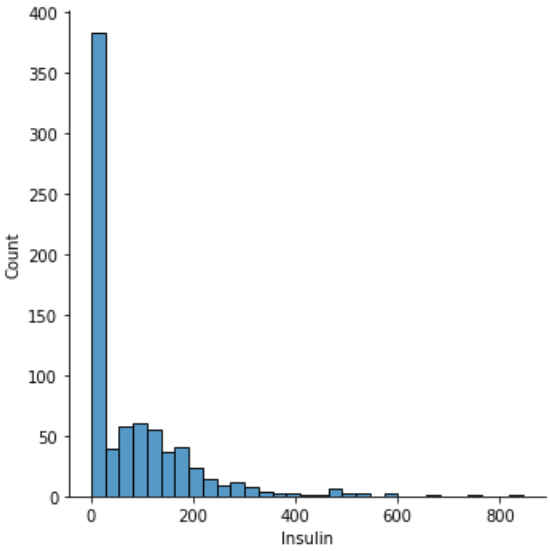
entities of the data set like column(s), row(s) by applying Pandas, NumPy, Statistical Methods, and Data visualization packages. Exploratory Data Analysis is a data analytics process to understand the data in depth and learn the different data characteristics, often with visual means. This allows you to get a better feel of your data and find useful patterns in it.

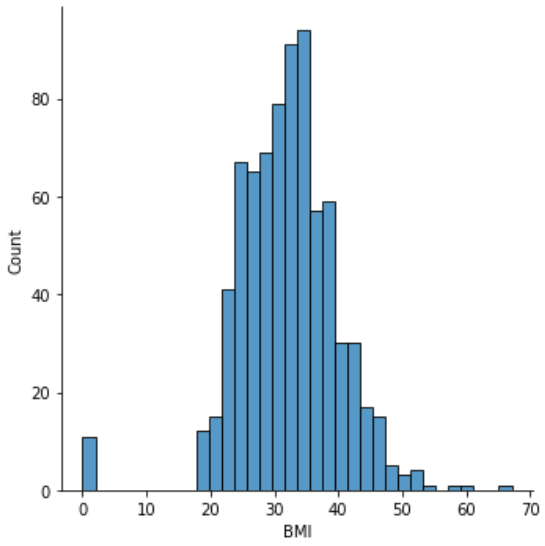


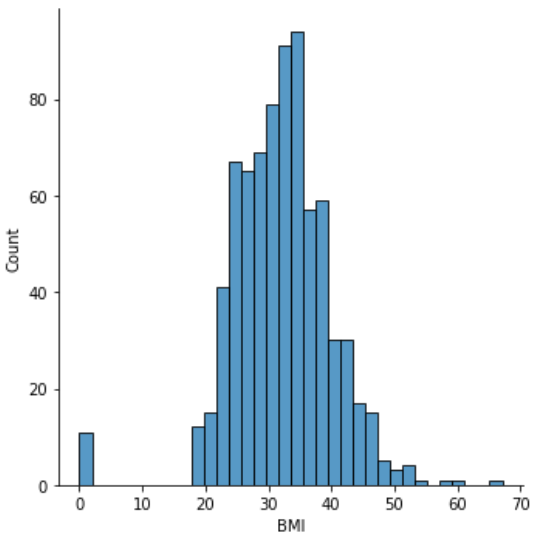
Interpretation:

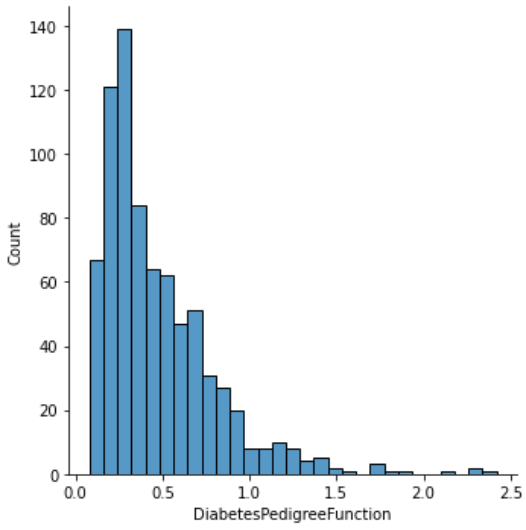
Using countplot we can observe that with that target variable (Outcome) having categories [Diabetic and Non-Diabetic] there is a slight imbalance in the dataset.

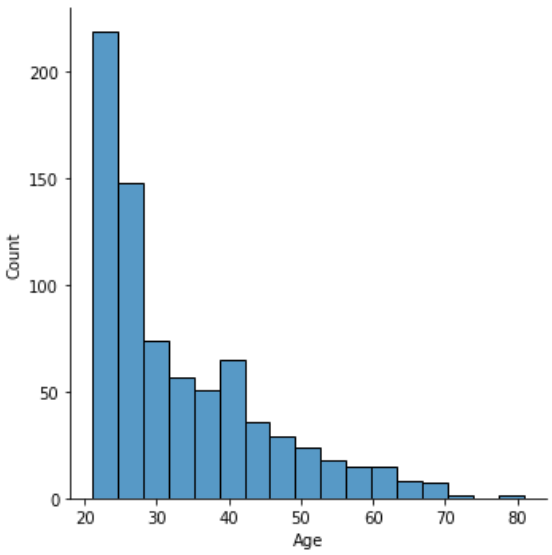

Count
SkinThickness


Count
Insulin


Count
BMI

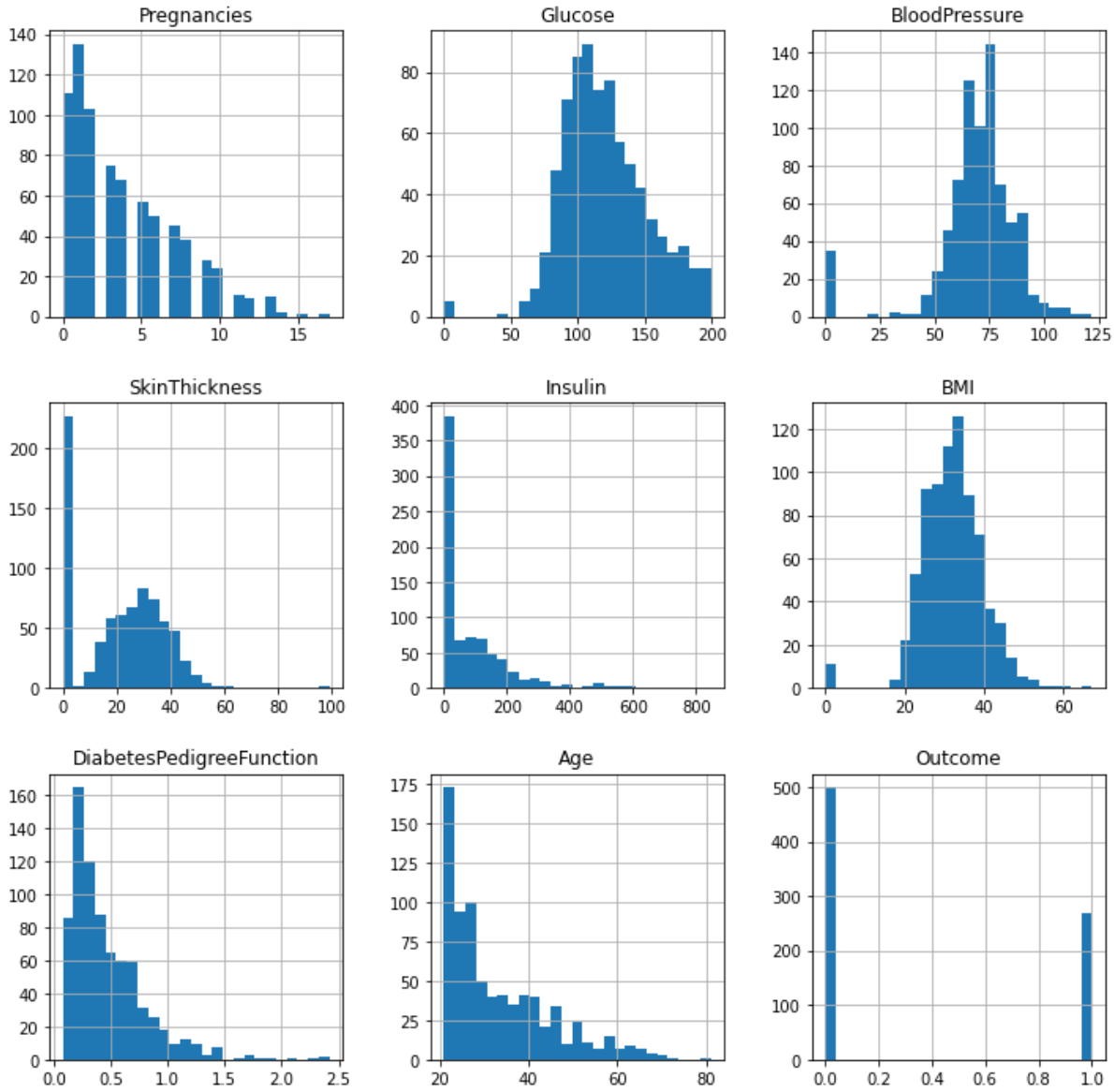

Count
BMI


Count
DiabetesPedigreeFunction


Count
Age

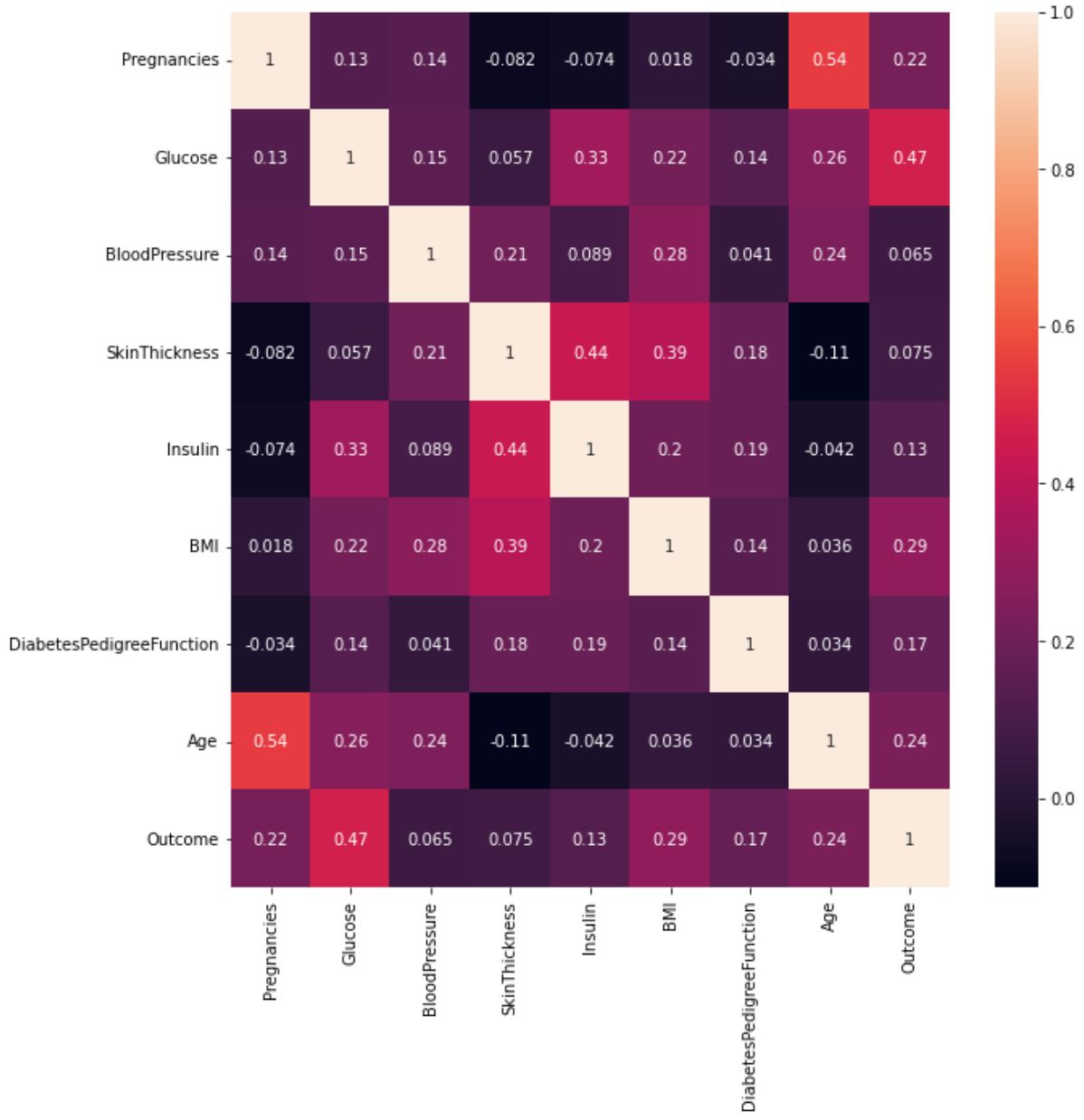
Interpretation:

Using the distribution plot we can clearly observe all the features detailly regarding the features skewness, skewness can be observed in each plots as most of the features are either rightly skewed or leftly skewed. BMI column values are normally distributed across the plot.



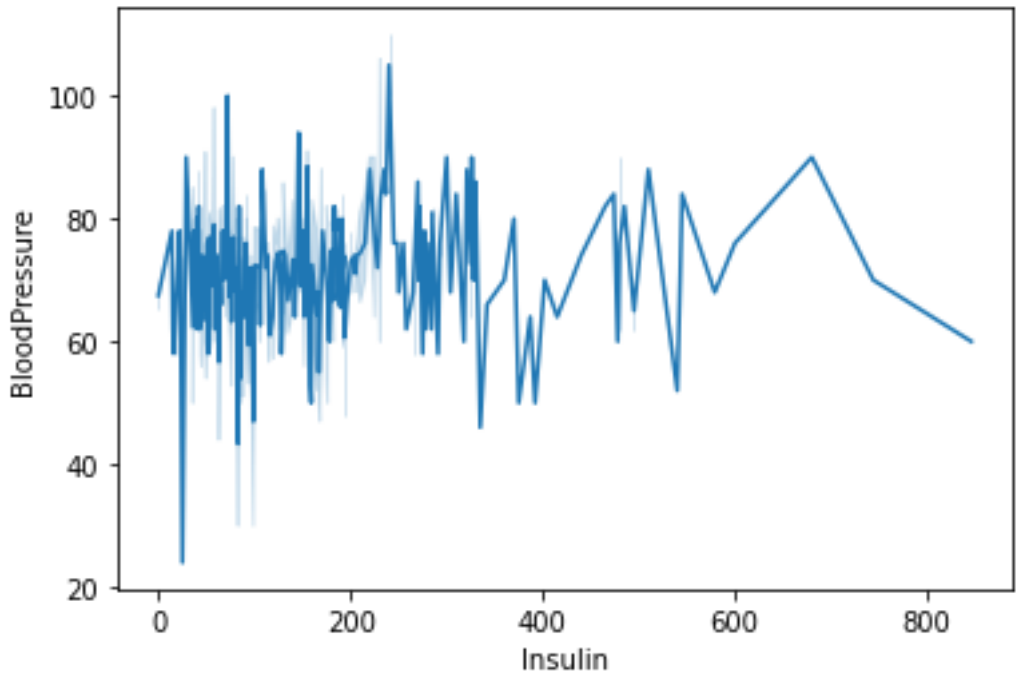
Interpretation:

With the help of histogram we can understand that Mean is slightly more than the median for most of the features. So it is right skewed most of the features.



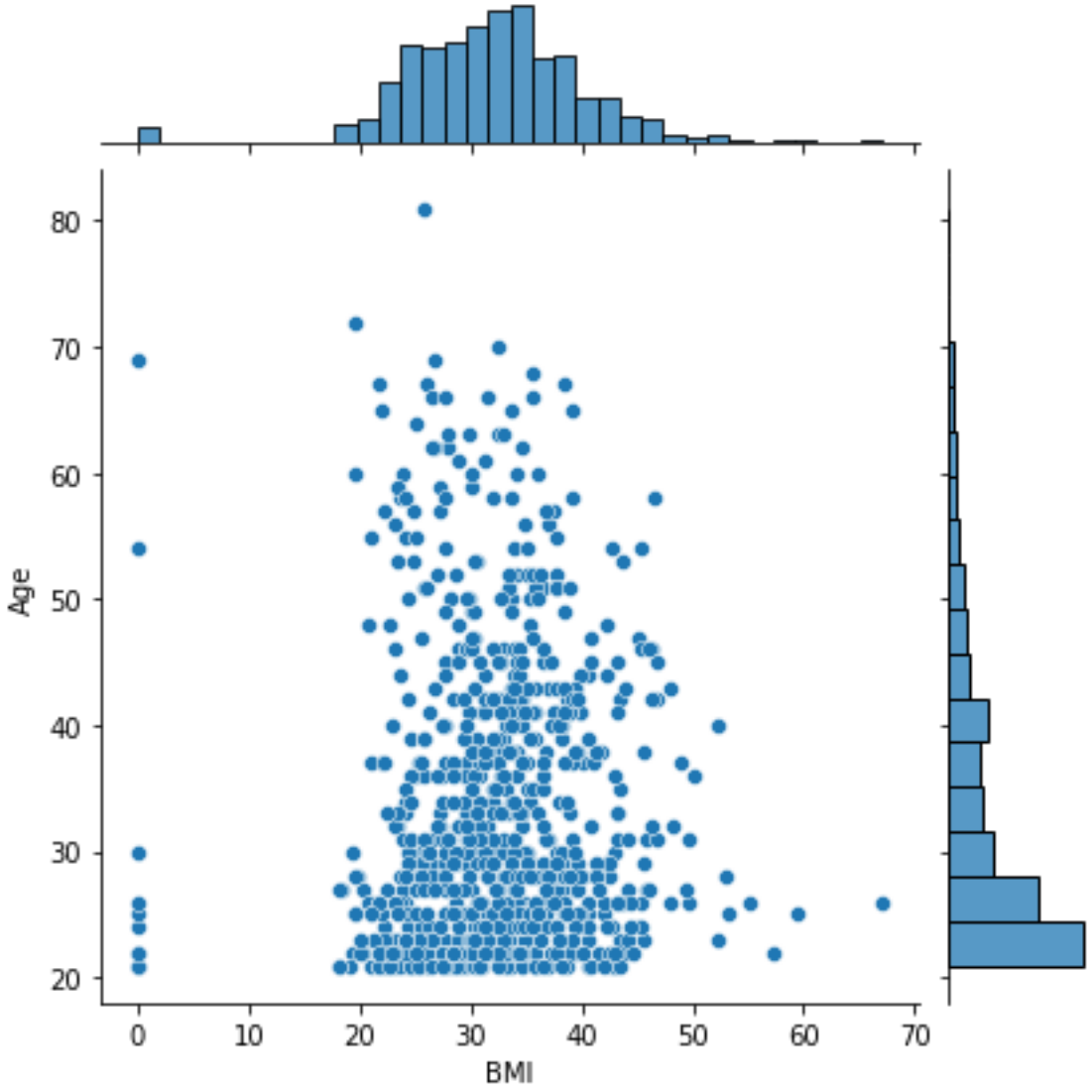
<u>Interpretation</u>:

With the help of plotting heatmap we can clearly observe that Correlation Matrix reveal that bloodpressure and insulin are positive correlated whereas bmi and cholestral are positively correlated.



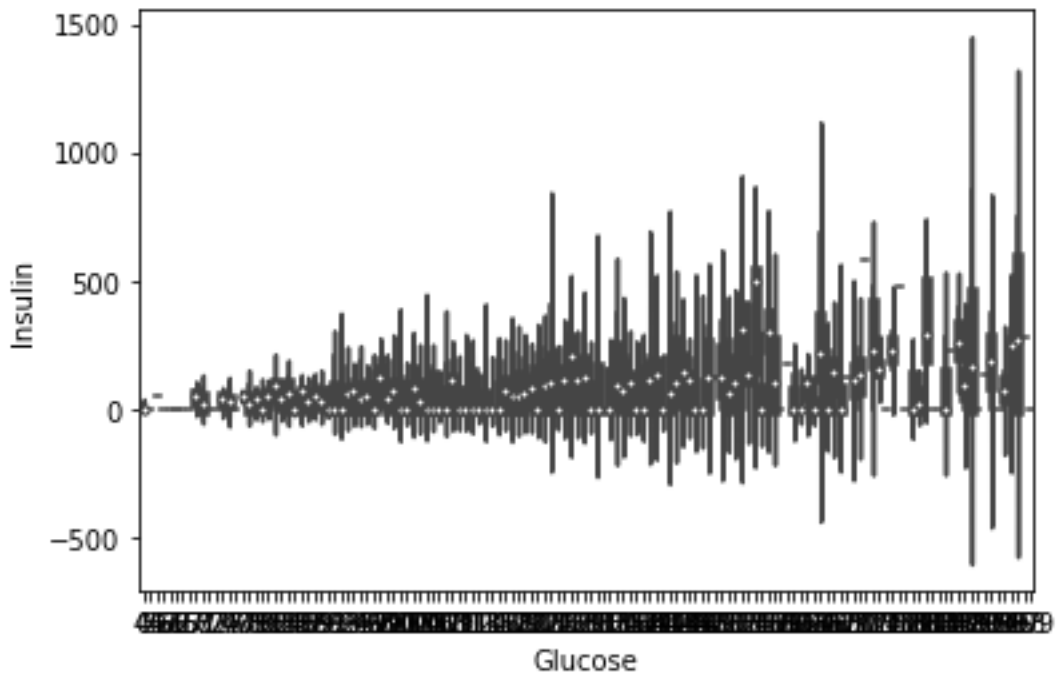
Interpretation:

With the help of line plot changes and trends over different changes in blood pressure and insulin regard to effect of diabetes, it is also helpful to show small changes that are difficult to measure in other graphs.



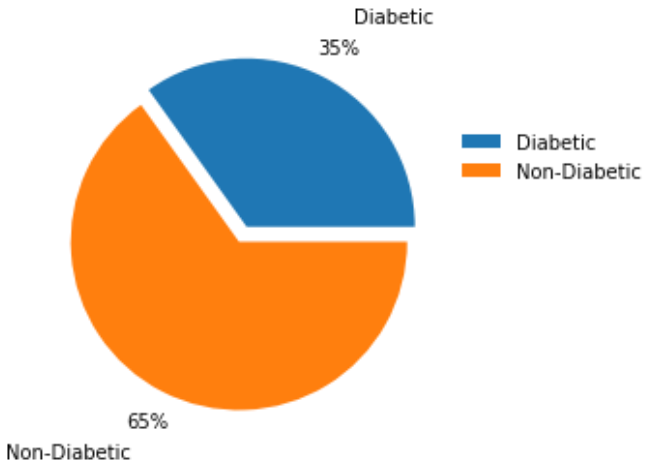
Interpretation:

With the help of plotting joint plot displays relationship between the variables as well as the univariate graph of diabetes outcome with respect to Age and BMI.



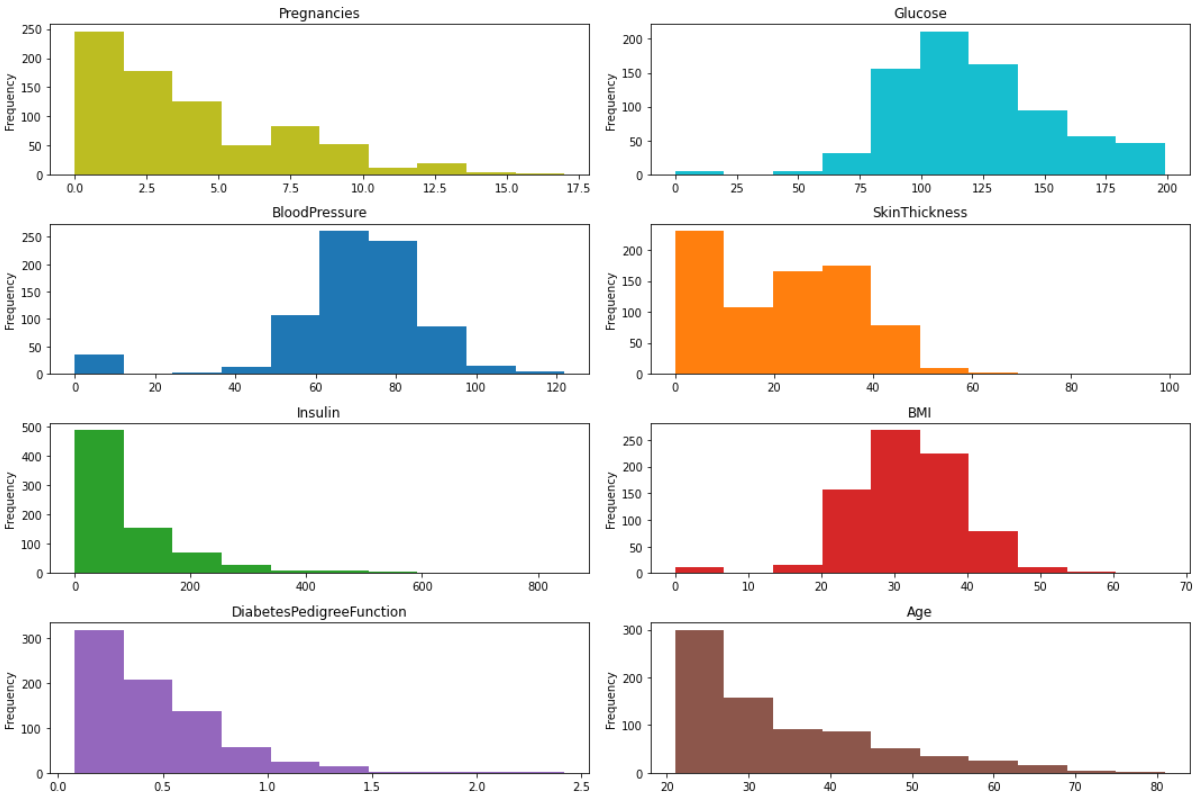
Interpretation:

With the help of violin plot displays distribution of insulin and glucose across the samples in the dataset. Mainly used to observe their numerical distribution of how insulin and glucose is related.



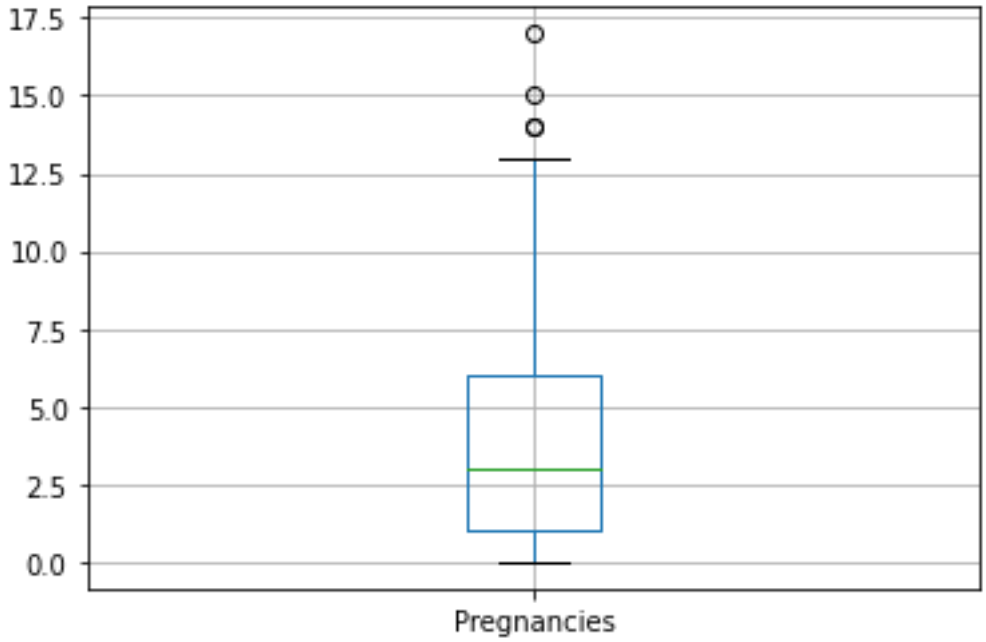
Interpretation:

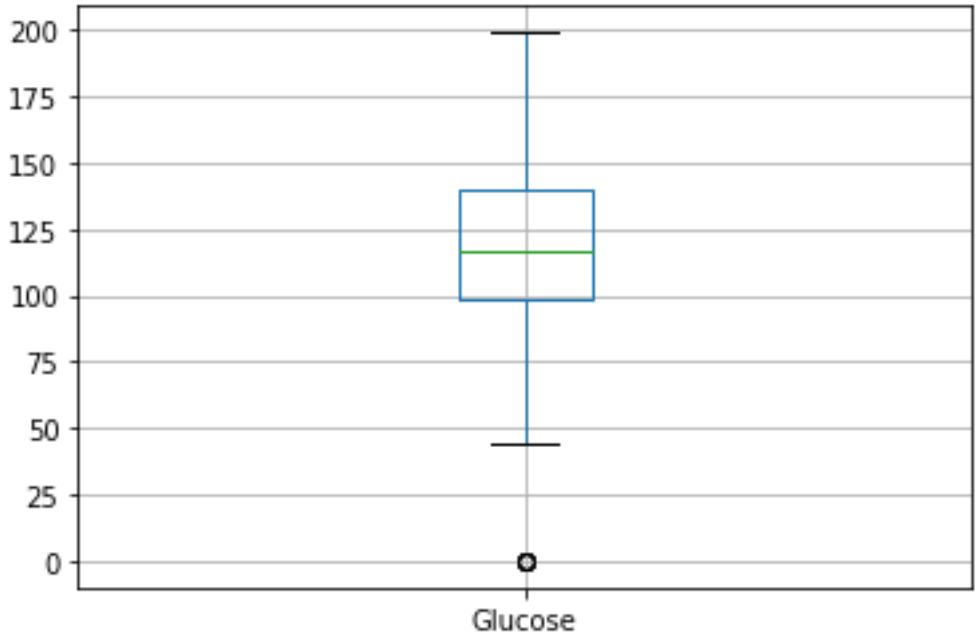
With the help of pie chart we can clearly observe the class imbalance a slight imbalance in the dataset with 65% being Non diabetic and 35% being diabetic.

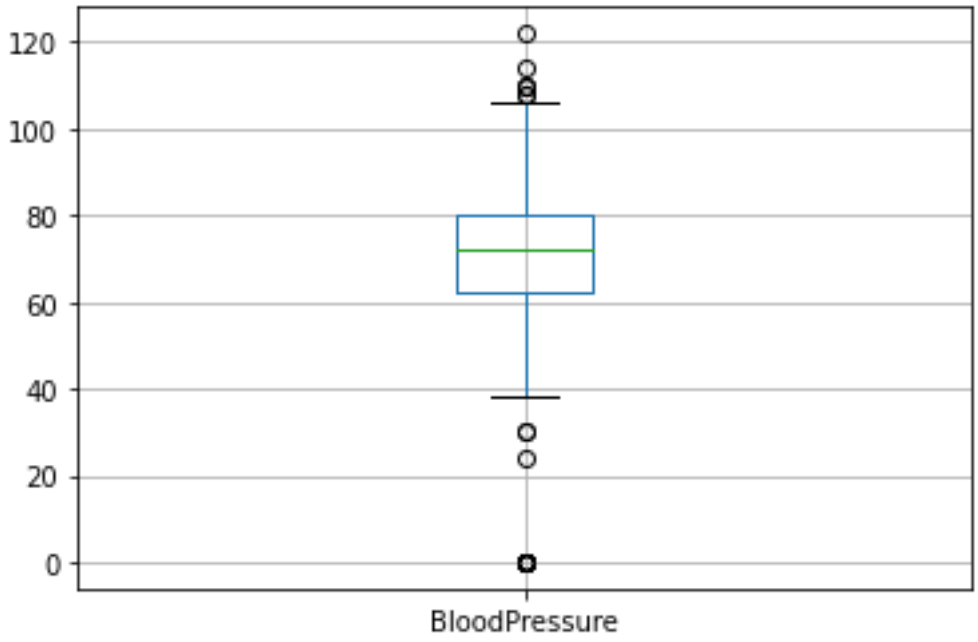


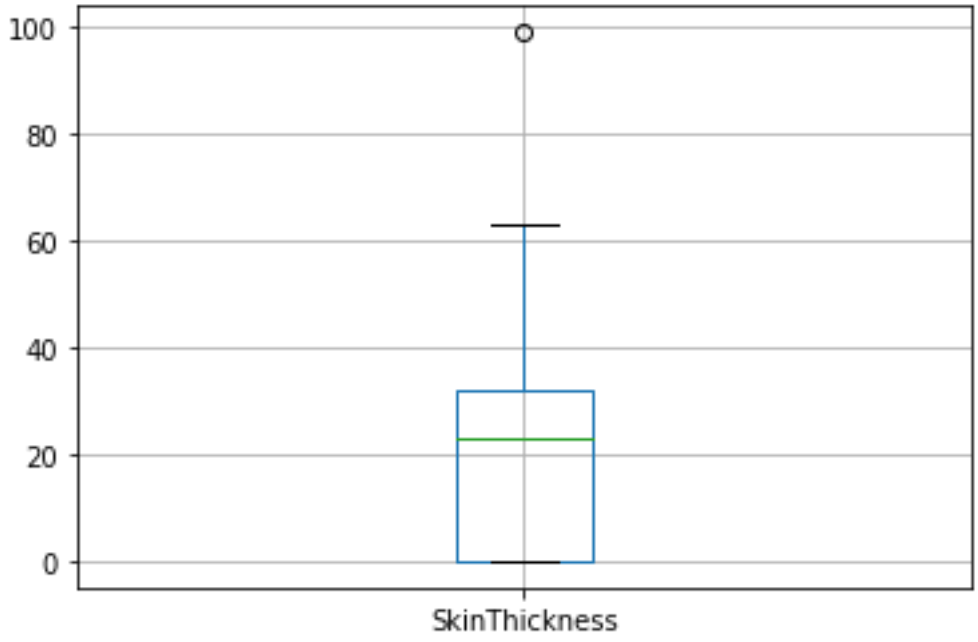
Interpretation:

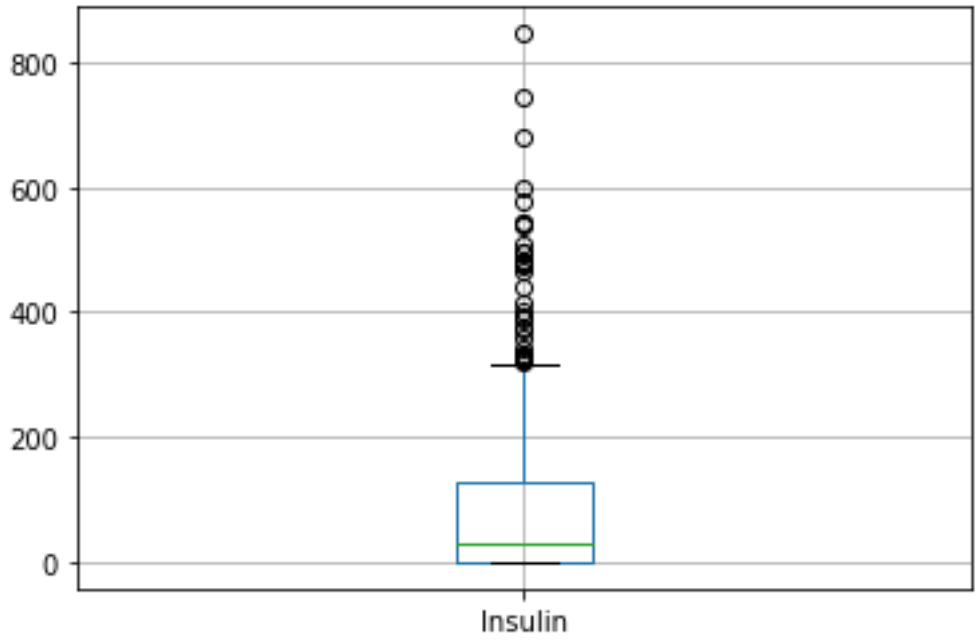
With the help of subplots various layouts of subplots of each feature is plotted individually showing the information of the measure of skewness, all the features detailly regarding the features skewness, skewness can be observed in each plots as most of the features are either rightly skewed or leftly skewed. BMI column values are normally distributed across the plot.

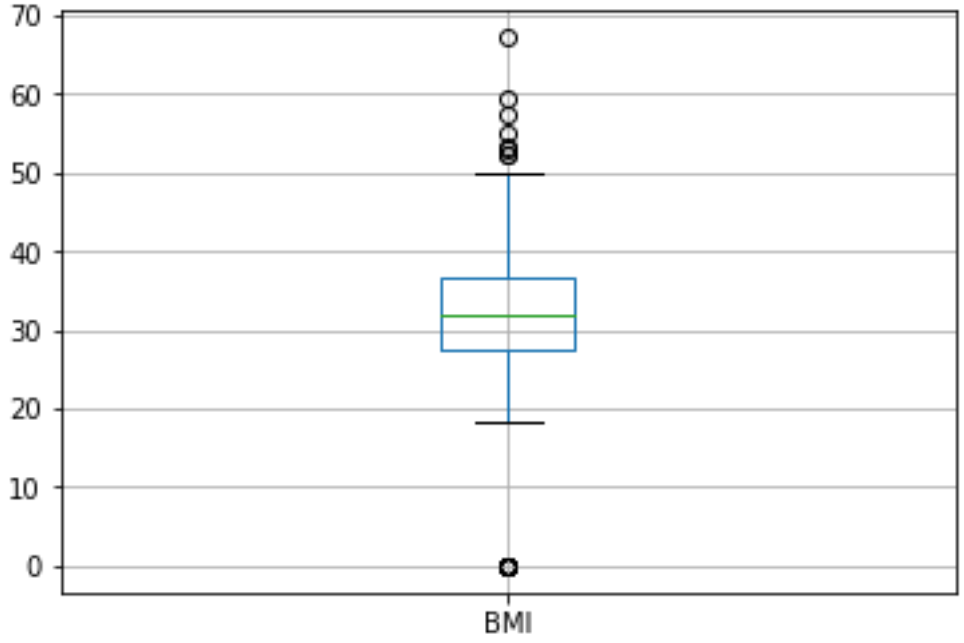

17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Pregnancies

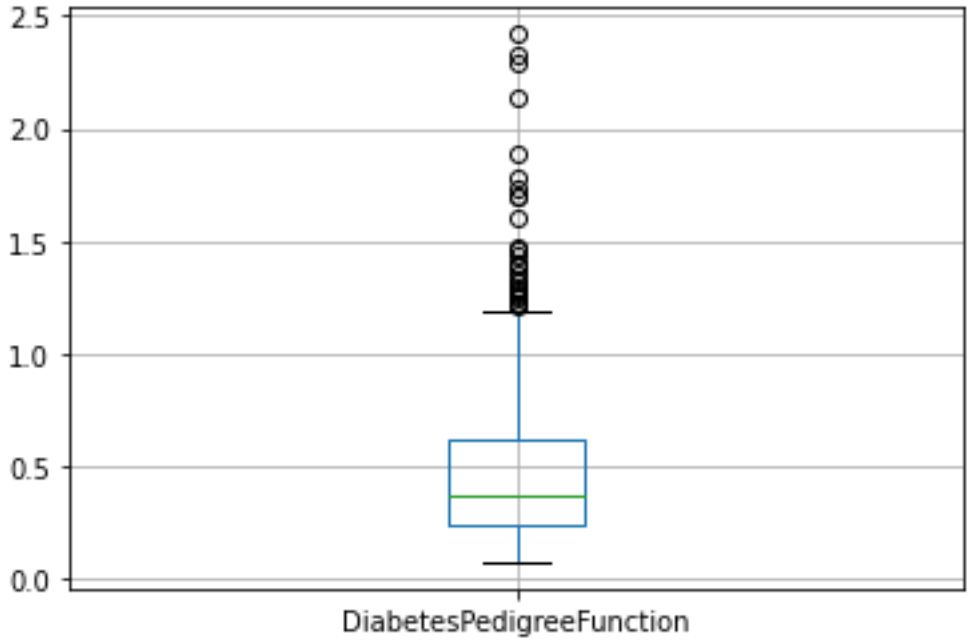

200
175
150
125
100
75
50
25
0
Glucose

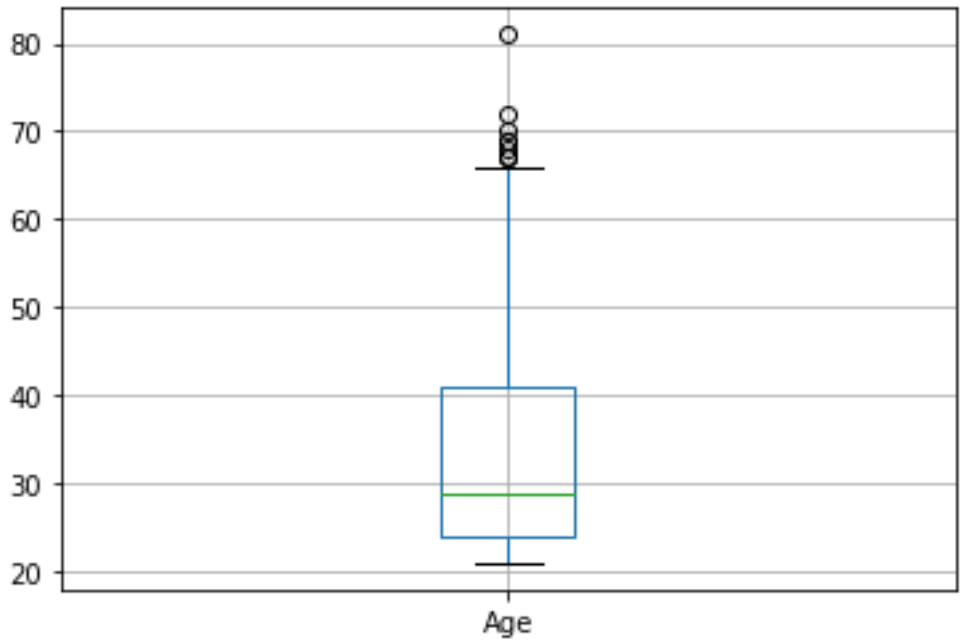

120
100
80
60
40
20
0
BloodPressure


100
80
60
40
20
0
SkinThickness


800
600
400
200
0
Insulin


70
60
50
40
30
20
10
0
BMI



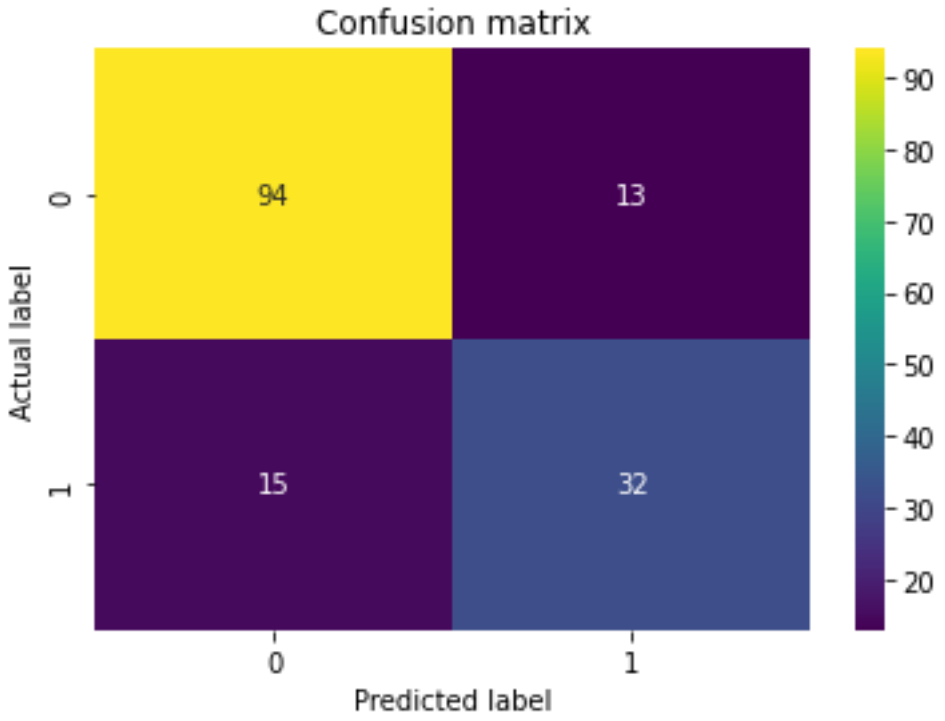


Interpretation:

Most of the features have Outliers especially BMI, Bloodpressure, Insulin, DiabetesPedigreeFunction.

LAB 2: KNN (K NEAREST ALGORITHM)

- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.
- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

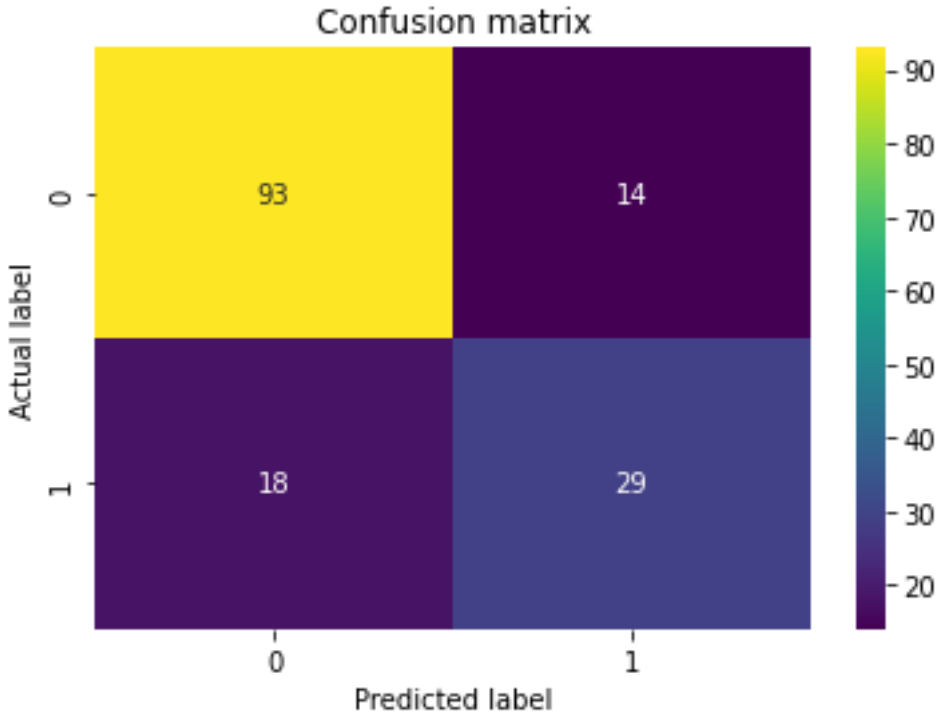


Interpretation:

- After performing the KNN algorithm on diabetes dataset we observe that we have got almost 81 % accuracy and inference from the confusion matrix that almost 126 values have been correctly predicted using knnn. Thus we can understand that almost 81% values of datapoints can be correctly predicted for diabetes dataset using Knn Algorithm. It is simple to implement.
- It is robust to the noisy training data
- Always needs to determine the value of K which may be complex some time.
- The computation cost is high because of calculating the distance between the data points for all the training samples.

LAB 3: NAÏVE BAYES ALGORITHM

- Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.
- It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.



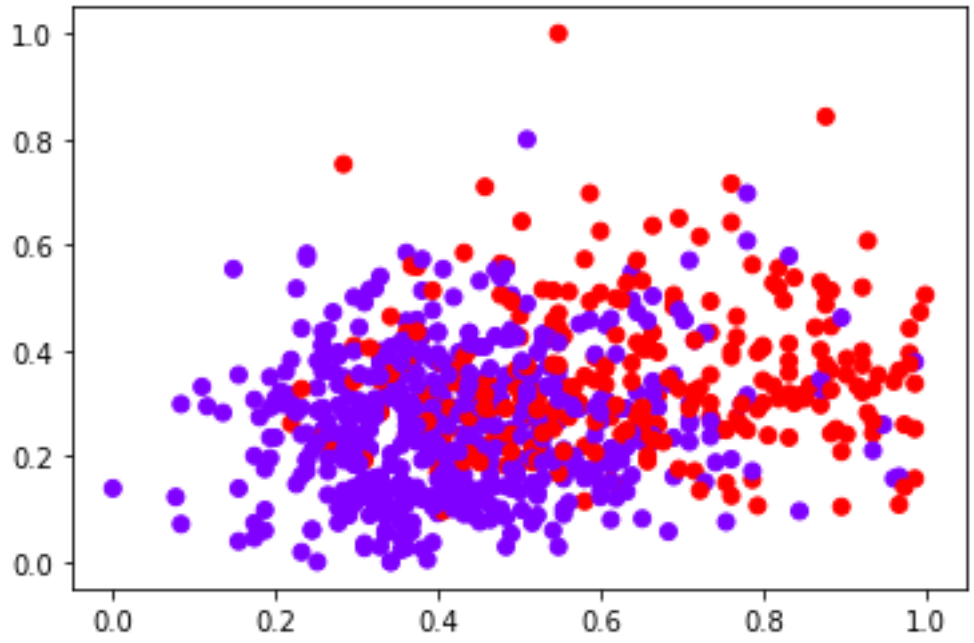
Interpretation:

After performing the Naive bayes algorithm on diabetes dataset we observe that we have got almost 79 % accuracy and inference from the confusion matrix that almost 122 values have been correctly predicted using Naive Bayes. Thus, we can understand that almost 79% values of datapoints can be correctly predicted for diabetes dataset using Naive bayes Algorithm. By comparatively understanding knn and naive bayes we can conclude that knn is better predicted as its accuracy is 81% than naive bayes whose accuracy is just 79%.

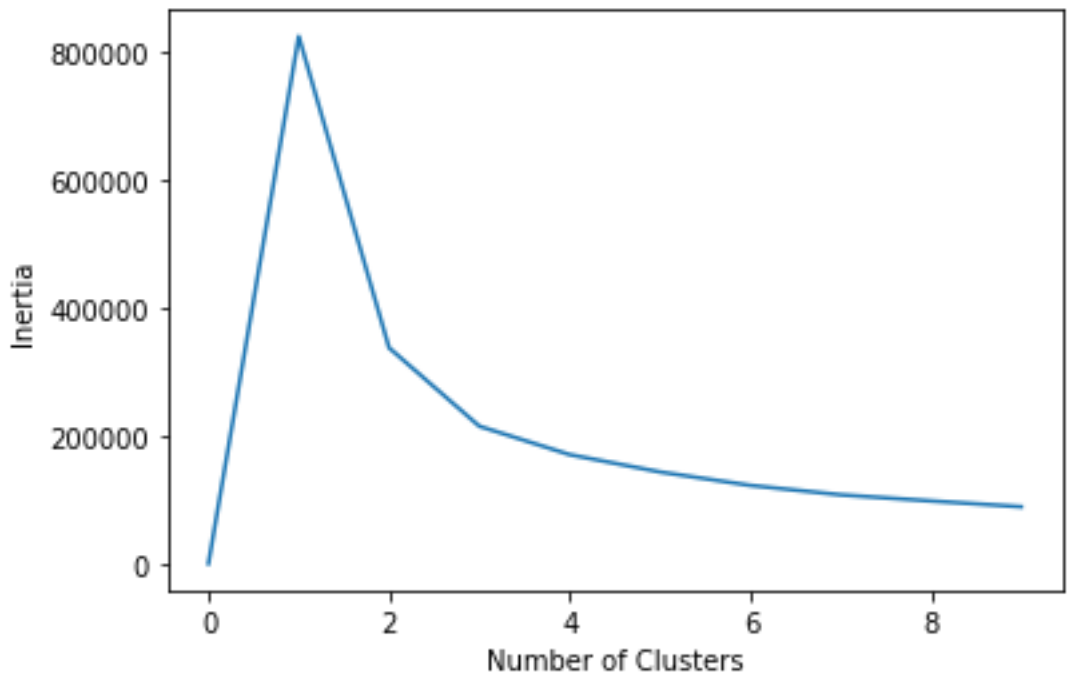
LAB 4: K MEANS CLUSTERING

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabelled dataset into different clusters. It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabelled dataset on its own without the need for any training.

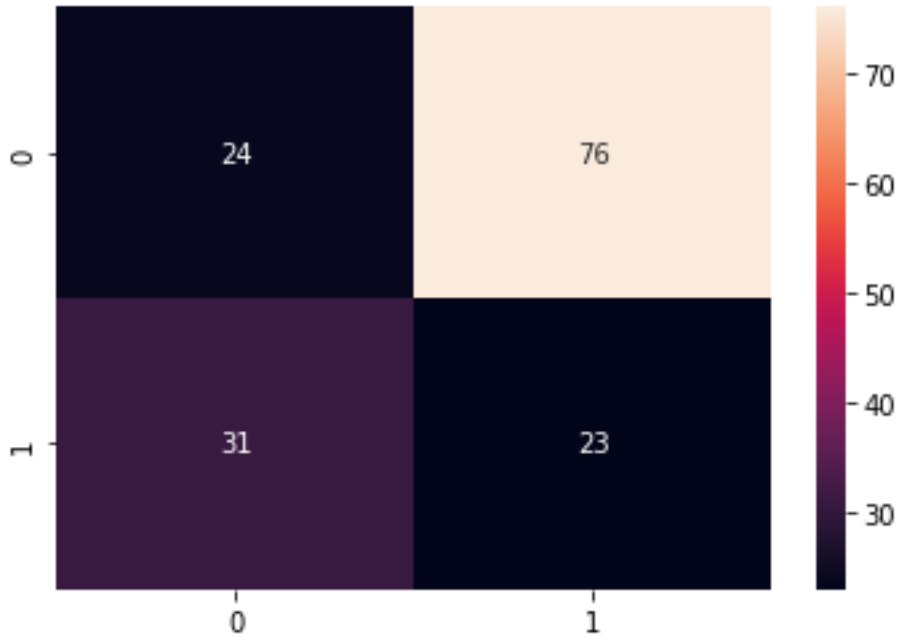
It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.



Interpretation: By assigning cluster as 2 we plotted the scatter plot with blood pressure and glucose showing the association between them.



Interpretation: From the above elbow method we can determine the optimal number of clusters into which the data may be clustered here as 2.

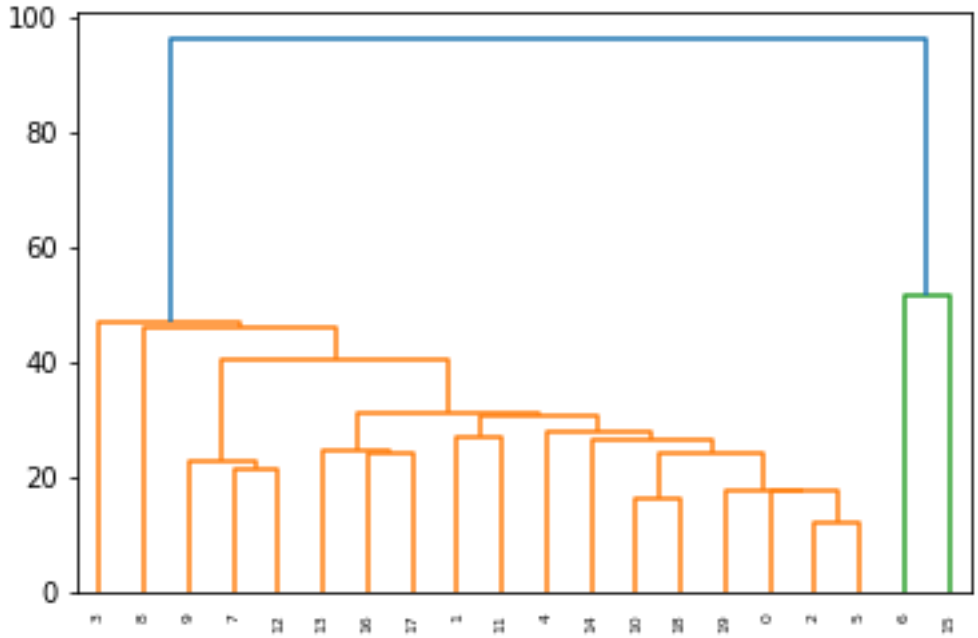


Interpretation:

After performing the K means algorithm on diabetes dataset inference from the confusion matrix that almost 47 values have been correctly predicted. K-means algorithm can easily adjust to the changes. If there are any problems, adjusting the cluster segment will allow changes to easily occur on the algorithm. The algorithm used is good at segmenting the diabetes data set. Efficiency depends on the shape of the clusters. K means works on minimizing Sum of squares of distances, hence it guarantees convergence. Computational cost is O(Knd), hence K means is fast and efficient.

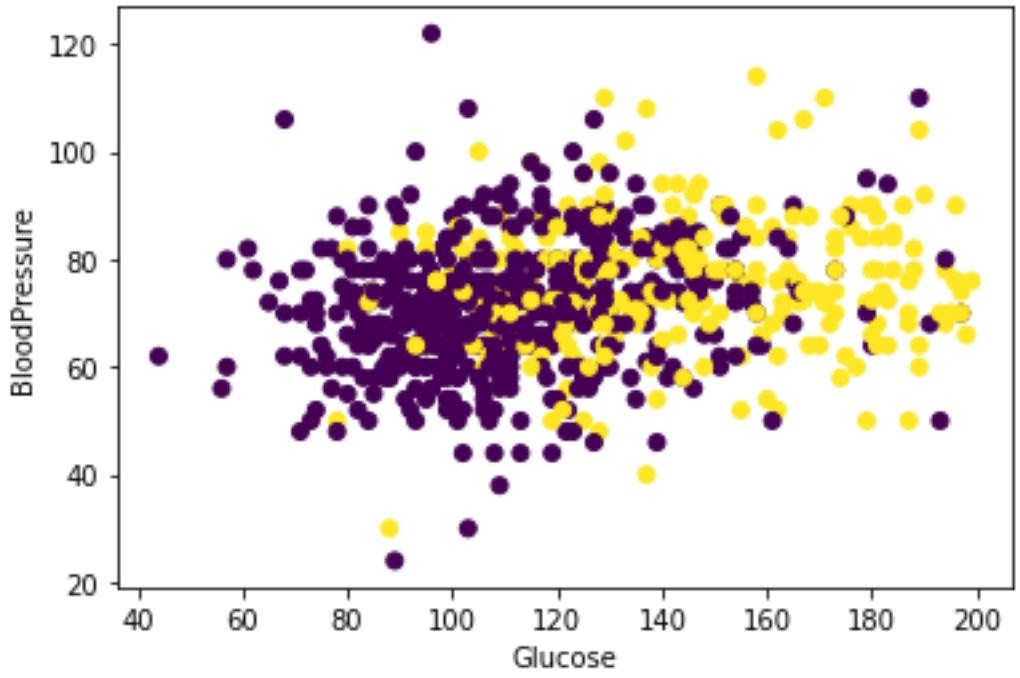
LAB 5: HIERARCHICAL CLUSTERING

Hierarchical clustering is another unsupervised machine learning algorithm, which is used to group the unlabelled datasets into a cluster and also known as hierarchical cluster analysis or HCA.



Interpretation:

The hierarchy of clusters in the form of a tree for the diabetes dataset with this tree-shaped structure is dendrogram showing the clustering for the features from BMI, Blood pressure, insulin and age.



Interpretation:

By assigning clusters we plotted the scatter plot with blood pressure and glucose showing the association between them, this plot shows high correlation of feature Blood pressure with Glucose.

After performing the Hierarchical clustering algorithm on diabetes dataset we get 65 % accuracy which is less compared to both KNN and Naïve Bayes which is more than 83%.

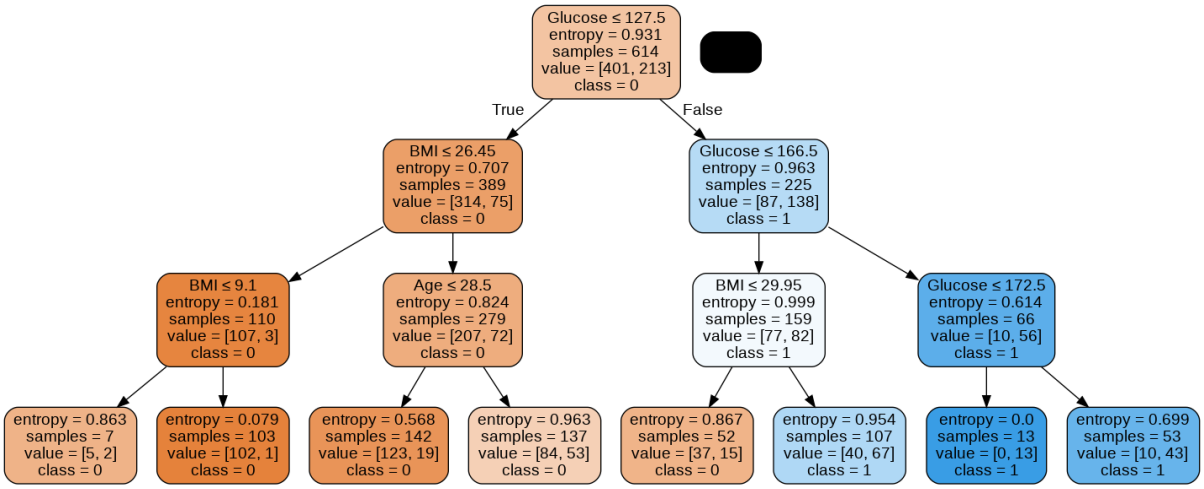
Hierarchical clustering outputs a hierarchy a structure that is more informative than the unstructured set of flat clusters returned by k-means. Therefore, it is easier to decide on the number of clusters by looking at the dendrogram.

Once the instances have been assigned to a cluster, they can no longer be moved around. Initial seeds have a strong impact on the final results and very sensitive to outliers.

LAB 6: DECISION TREE

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

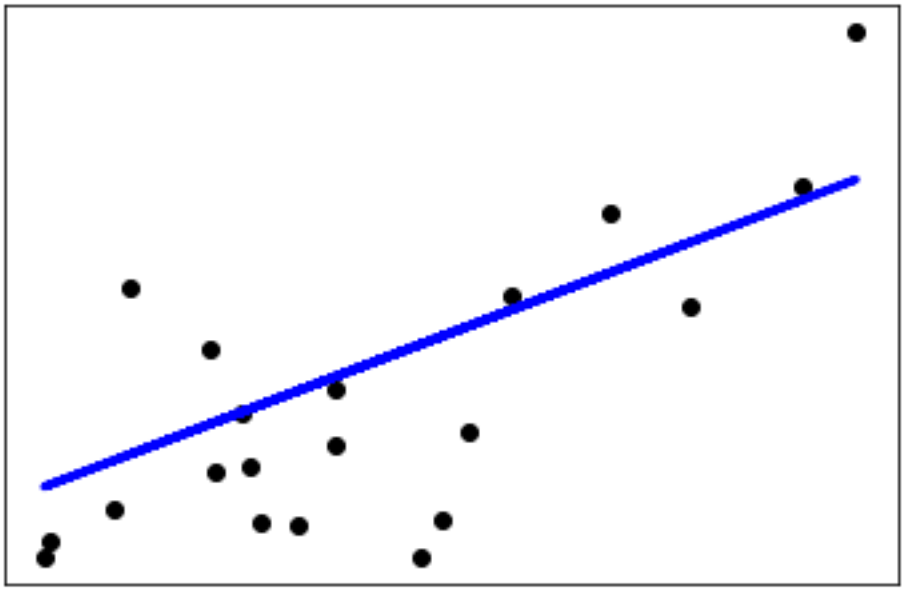


Interpretation:

- After performing the Decision Tree algorithm on diabetes dataset we get 79 % accuracy which is more compared Hierarchical clustering. It can be very useful for solving decision-related problems.
- It helps to think about all the possible outcomes for a problem.
- There is less requirement of data cleaning compared to other algorithms. But causes overfitting issues due to computational complexity.

LAB 6: B LINEAR REGRESSION

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis.



Interpretation:

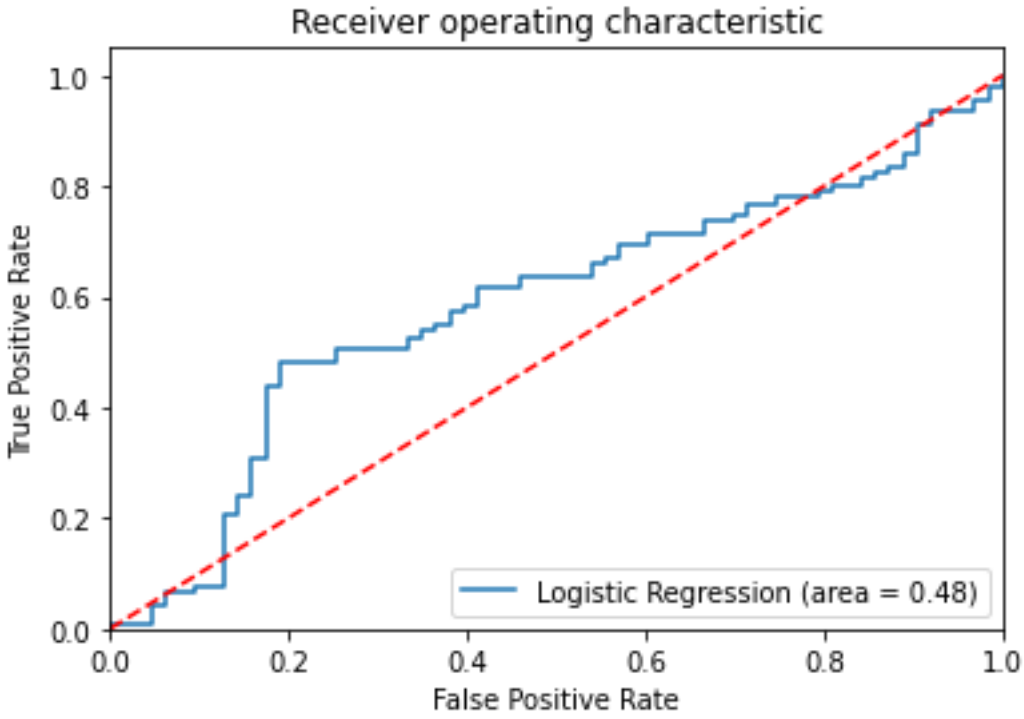
Linear regression shows the linear relationship between blood pressure and BMI which means it finds how the value of the dependent variable is changing according to the value of the independent variable. Here with the increase in blood pressure the level of BMI is also increasing from the observation with the accuracy percentage as 76%.

LAB 7: Logistic Regression

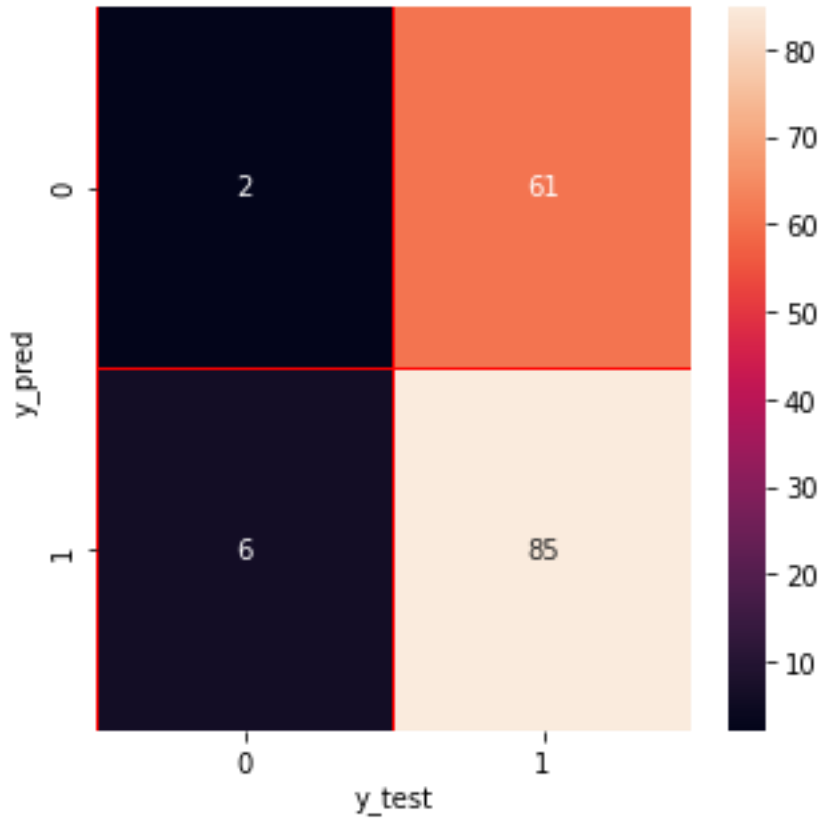
Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning

technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.



Interpretation: From the above graph we can obtain the sigmoid function / S curve graph from the diabetes dataset determining the best cut off value for predicting whether a new datapoint of person getting diabetes or not is a "failure" (0) or a "success" (1).

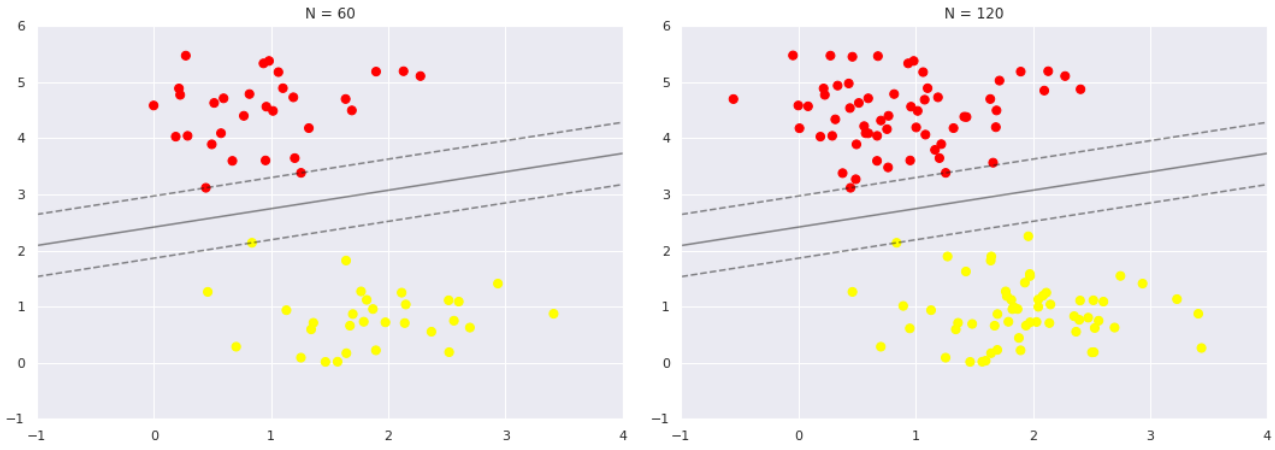


Interpretation:

- After performing the Logistic Regression algorithm on diabetes dataset we observe that we have got almost 56 % accuracy and inference from the confusion matrix that almost 88 values have been correctly predicted. Comparatively with all the algorithm logistic regression gives only 56% accuracy hence it is not more preferred to use logistic regression for this data set.

LAB 8: SUPPORT VECTOR MACHINE

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.



Interpretation:

It creates the best line or decision boundary that can segregate n-dimensional space into 2 classes so that we can easily put the new data point in the correct category of either diabetic or not. This best decision boundary hyperplane is plotted with the help of SVM.

Leftly we see the model and the support vectors for 60 training points rightly we have doubled the number of training points, but the model has not changed Three support vectors from the left panel are still the support vectors from the right panel. This

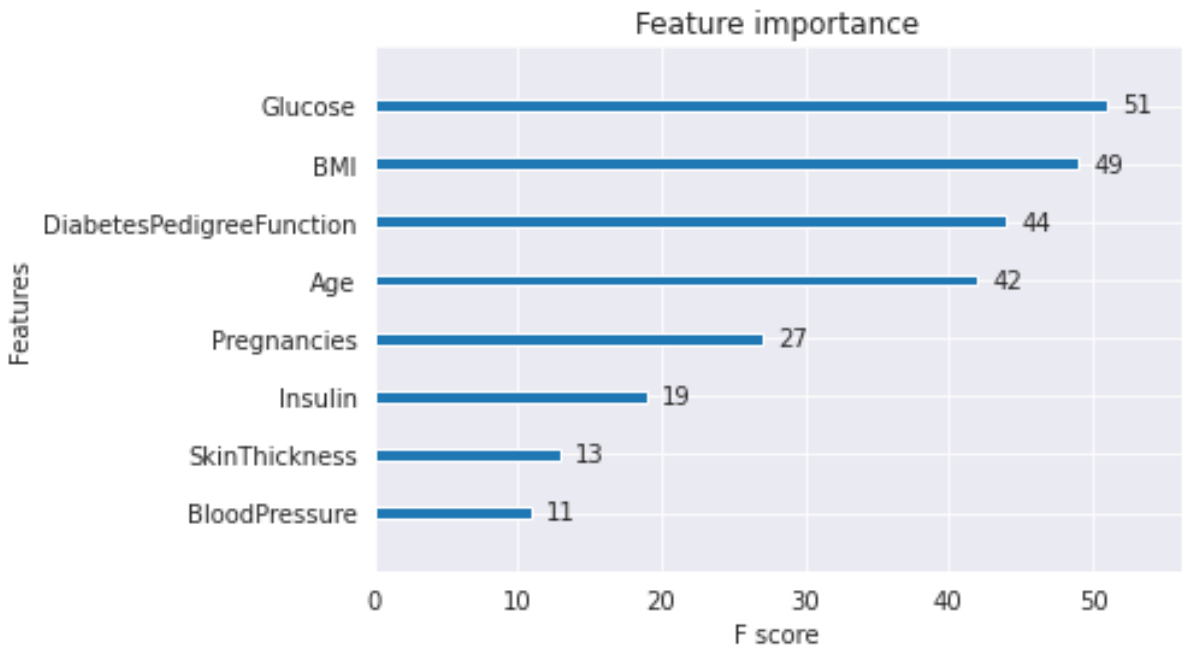
insensitivity to the exact behaviour of distant points is one of the strengths of the SVM model.

LAB 9: MLP (MULTI LAYER PERCEPTRON)

A multilayer perceptron (MLP) is a feedforward artificial neural network that generates a set of outputs from a set of inputs. An MLP is characterized by several layers of input nodes connected as a directed graph between the input and output layers.

MLP uses backpropagations for training the network. MLP is a deep learning method.

A multilayer perceptron is a neural network connecting multiple layers in a directed graph, which means that the signal path through the nodes only goes one way. Each node, apart from the input nodes, has a nonlinear activation function.



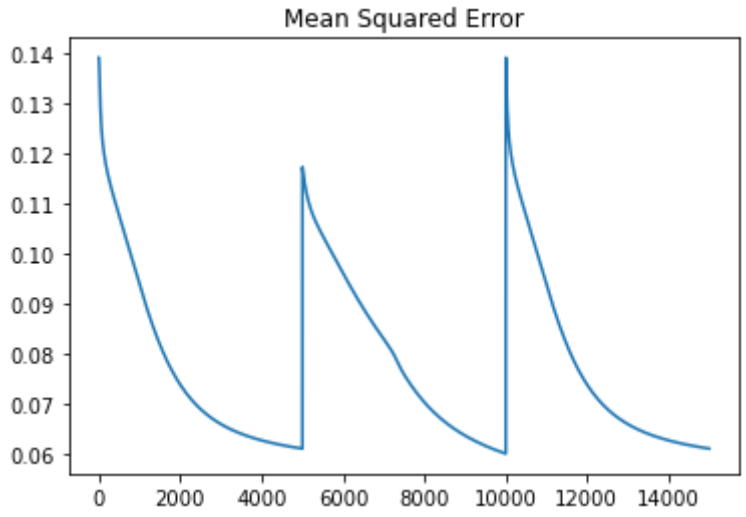
Interpretation:

- First three analysis mark BloodPressure, SkinThickness and Insulin as the least important features. Note that these features have very low correlation with Outcome. RFE marks SkinThickness as the least important.
- Three activation functions is used for hidden layers along with the displaying of the weights.
- Filter method: Calculating a metric like correlation coefficient between each feature and output separately as we did above. In this method all features are evaluated independently. Embedded methods: Methods like logistic regression or linear regression learn the coefficients that multiply each feature. Wrapper methods: Basically you have

an estimator and you train this estimator with the subsets of features. The subset giving the best score is selected and other features are eliminated.

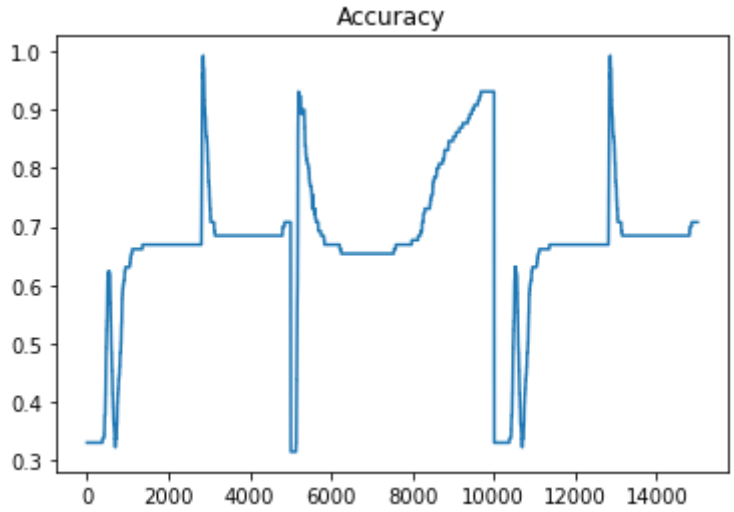
LAB 10: Back Propagation neural network

The core concept of BPN is to backpropagate or spread the error from units of output layer to internal hidden layers in order to tune the weights to ensure lower error rates. It is considered a practice of fine-tuning the weights of neural networks in each iteration. Proper tuning of the weights will make a sure minimum loss and this will make a more robust, and generalizable trained neural network.



Backpropagation neural network method is used to optimize neural networks by propagating the error or loss into a backward

direction by calculating the mean squared error for the diabetes dataset.



From the accuracy plotted graph we can understand that it finds loss for each node and updates its weights accordingly in order to minimize the loss using gradient descent thus giving overall accuracy of 80%.