CHRIST

(DEEMED TO BE UNIVERSITY)

BANGALORE · INDIA

MCA472B: Machine Learning

CAT3: Component 2:

By,

Nithya S

2147255

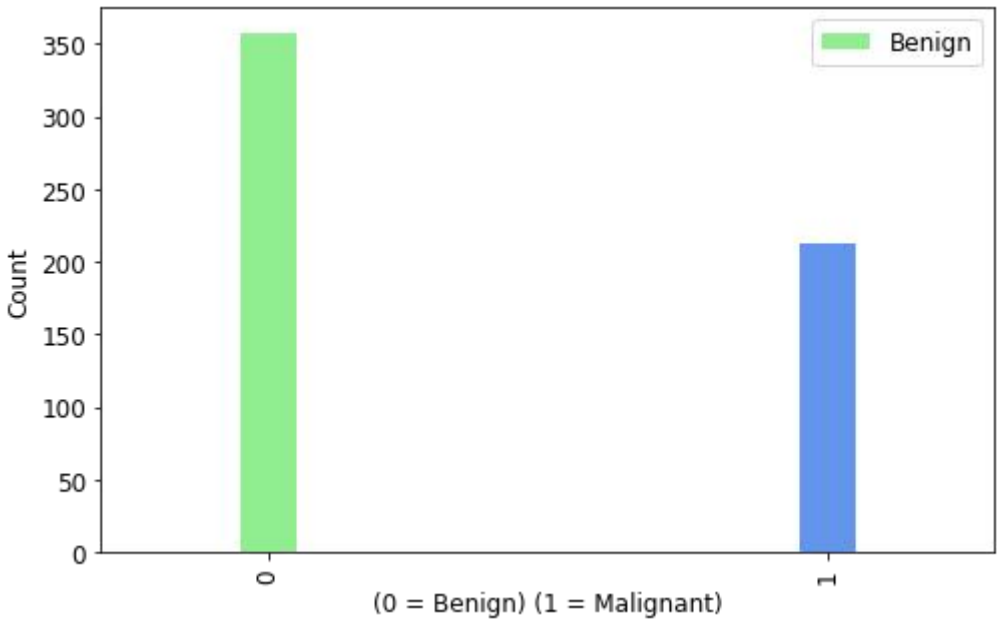
4 MCA B

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

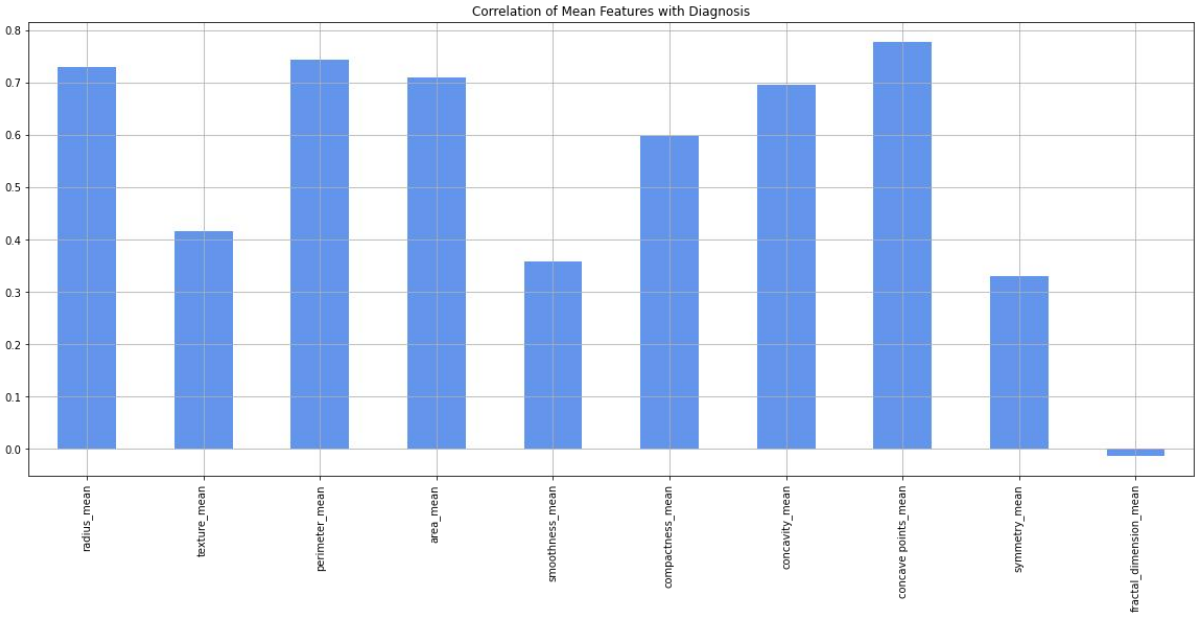
EDA AND VISUALIZATION:

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.

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.

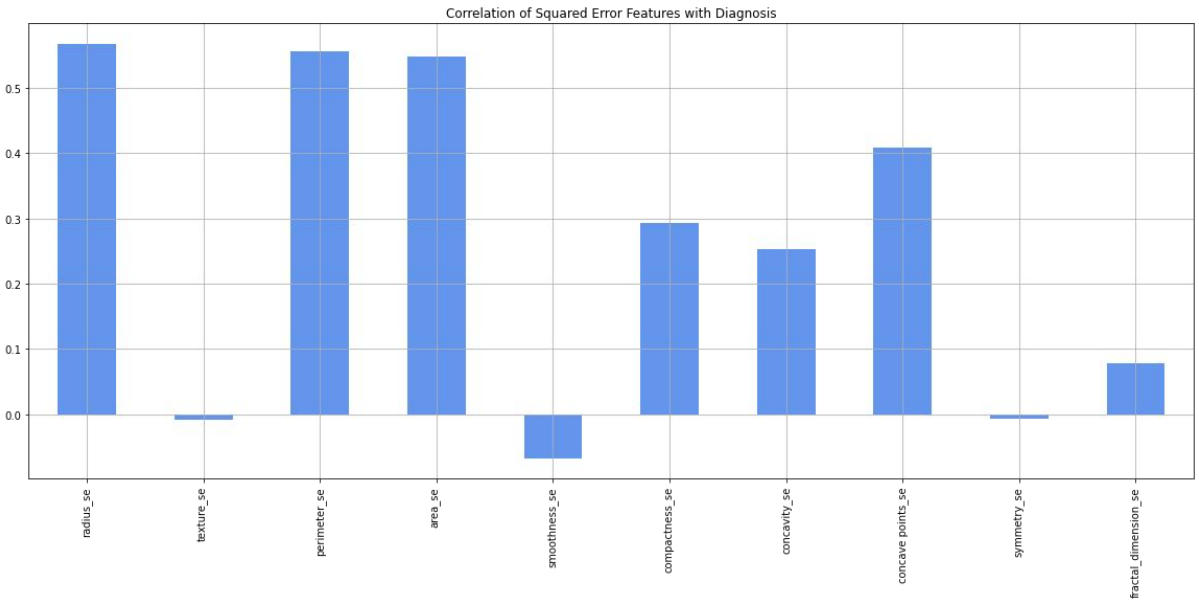


Observation: *We have 357 malignant cases and 212 benign cases so our dataset is* Imbalanced, we can use various re-sampling algorithms like under-sampling, over-sampling.



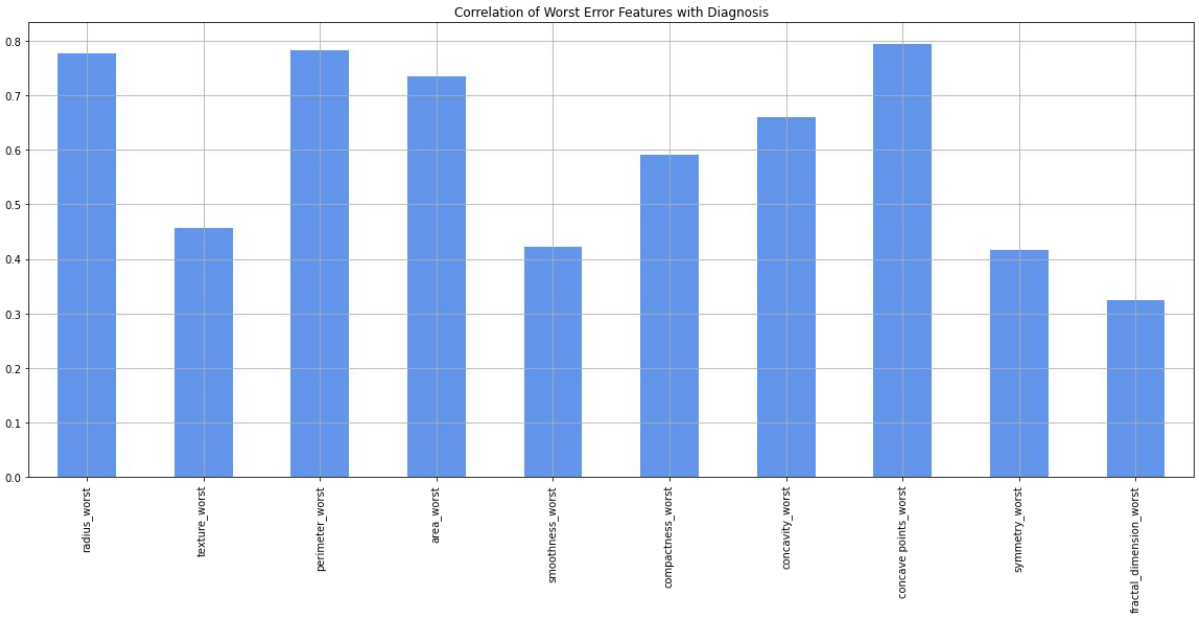
Observations:

- *fractal_dimension_mean least correlated with the target variable.*
- *All other mean features have a significant correlation with the target variable.*



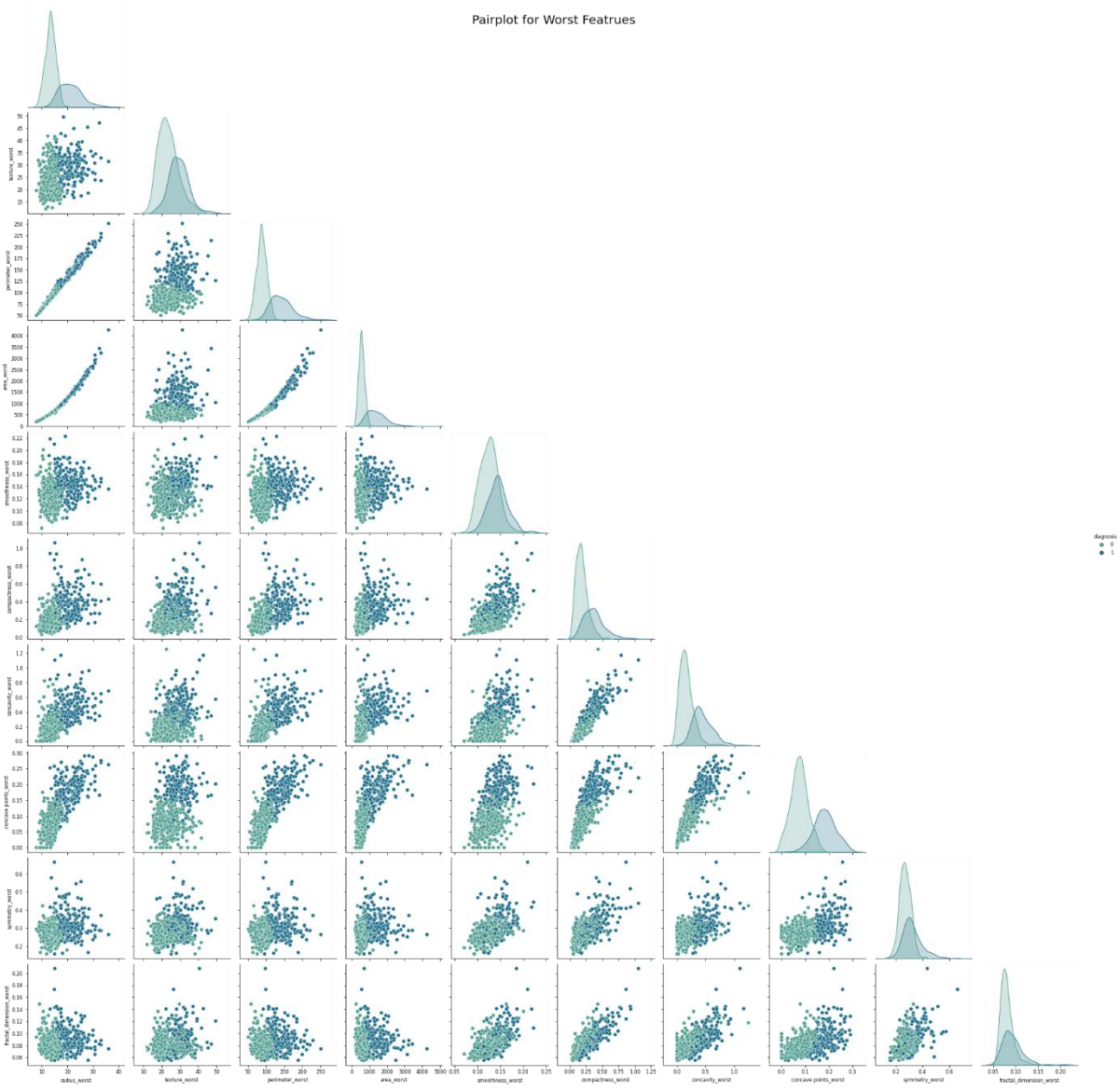
Observations:

- *texture_se, smoothness_se, symmetry_se, and fractal_dimension_se are least correlated with the target variable.*
- *All other squared error features have a significant correlation with the target variable.*

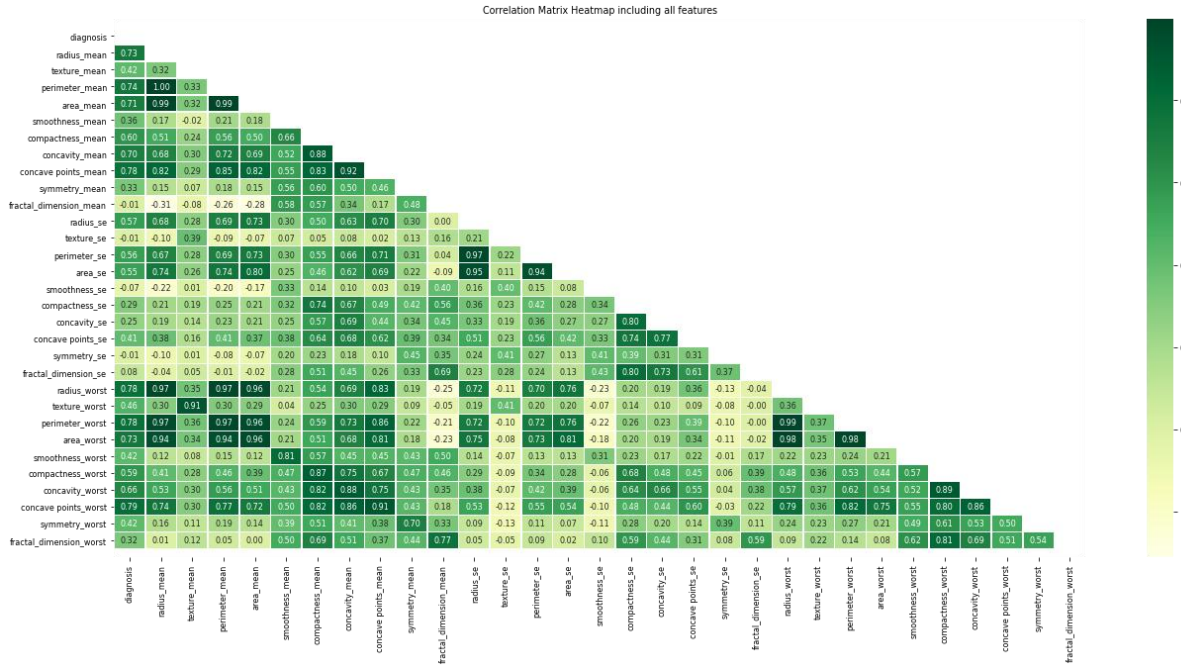


Observation:

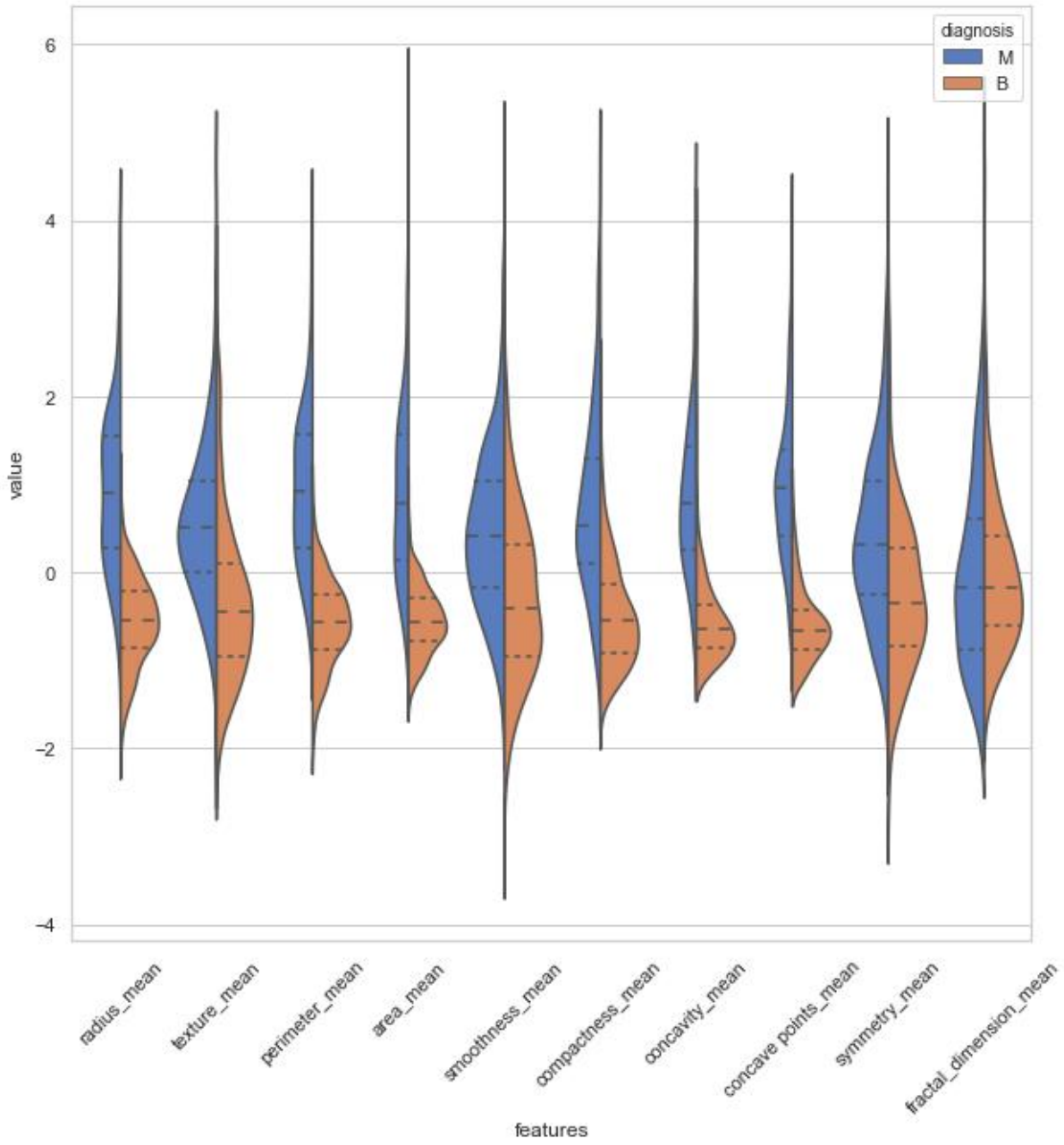
- *All worst features have a significant correlation with the target variable.*



Observations: Almost perfectly linear patterns between the radius, perimeter, and area attributes are hinting at the presence of multicollinearity between these variables. Another set of variables that possibly imply multicollinearity are the concavity, concave_points, and compactness.

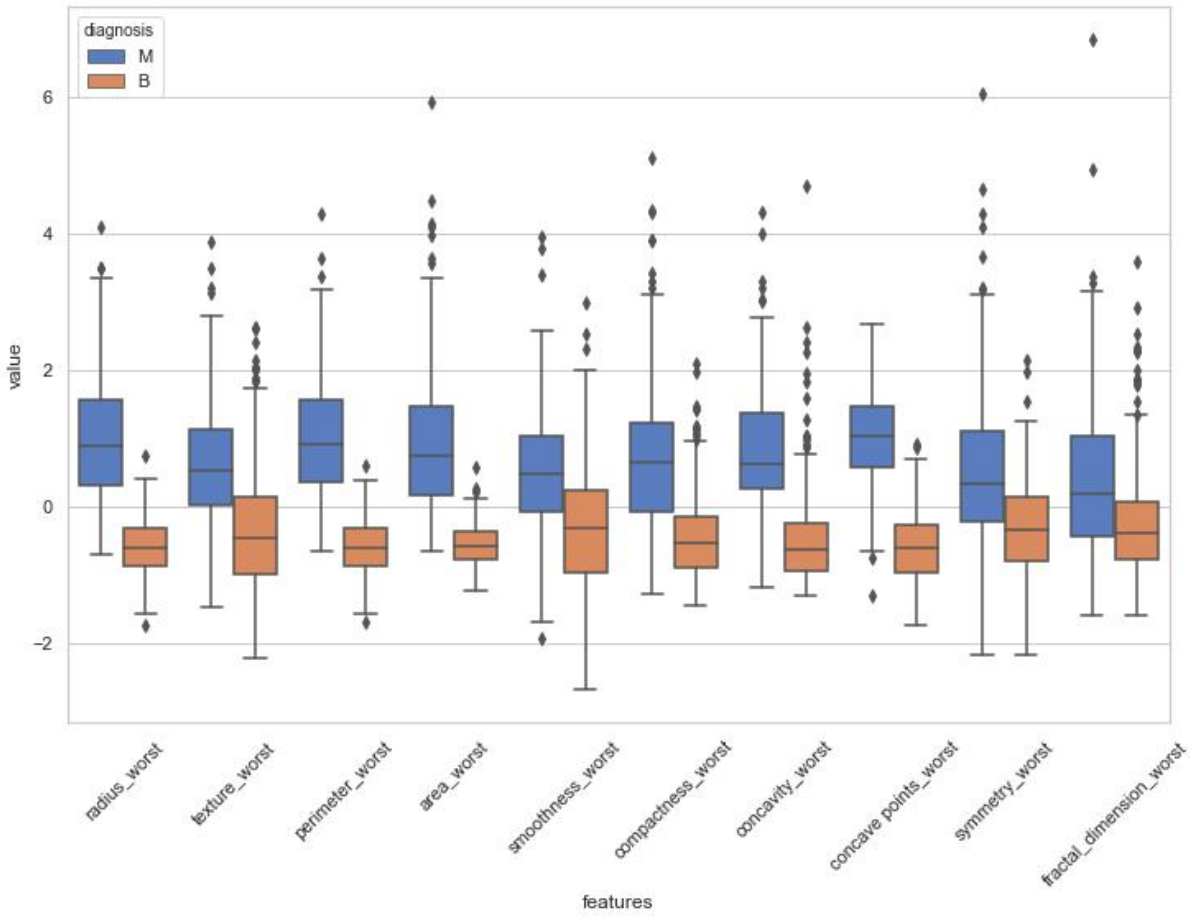


Observations: We can verify multicollinearity between some variables. This is because the three columns essentially contain the same information, which is the physical size of the observation (the cell). Therefore, we should only pick one of the three columns when we go into further analysis.



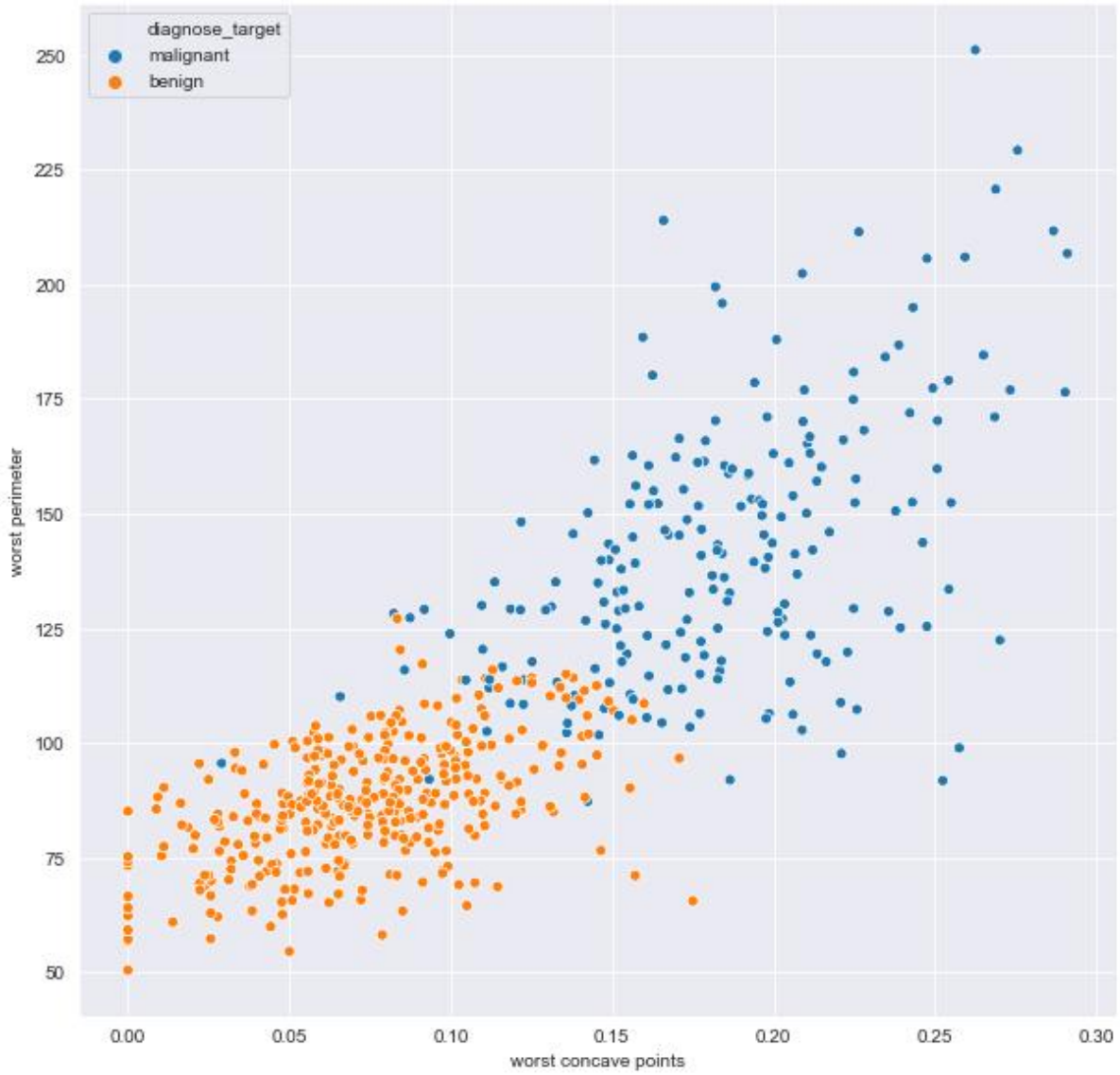
Observations:

A violin plot is more informative than a plain box plot. While a box plot only shows summary statistics such as mean/median and interquartile ranges, the violin plot shows the full distribution of the data in breast cancer dataset.



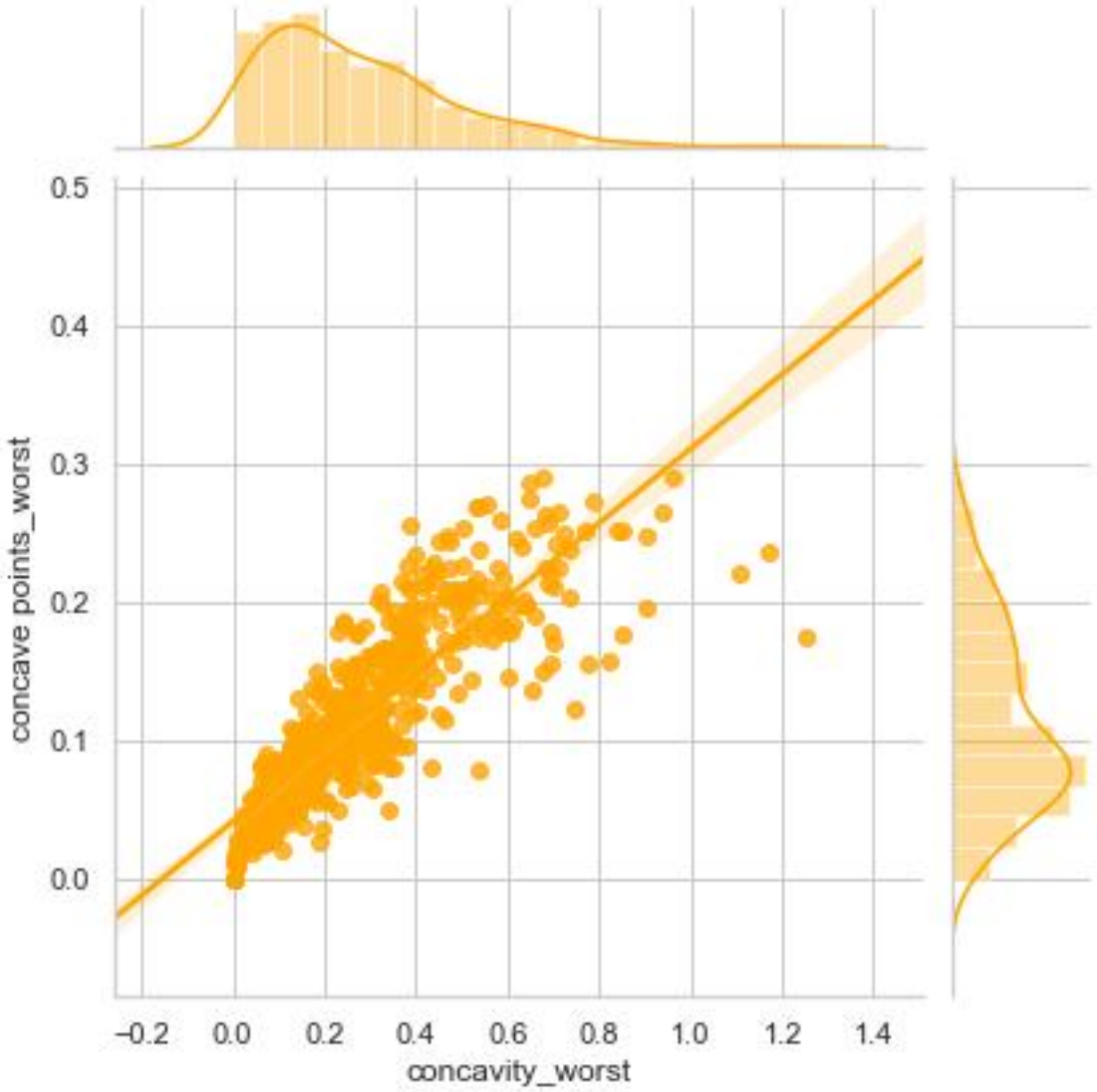
Observations:

Most of the features has outliers in the dataset.



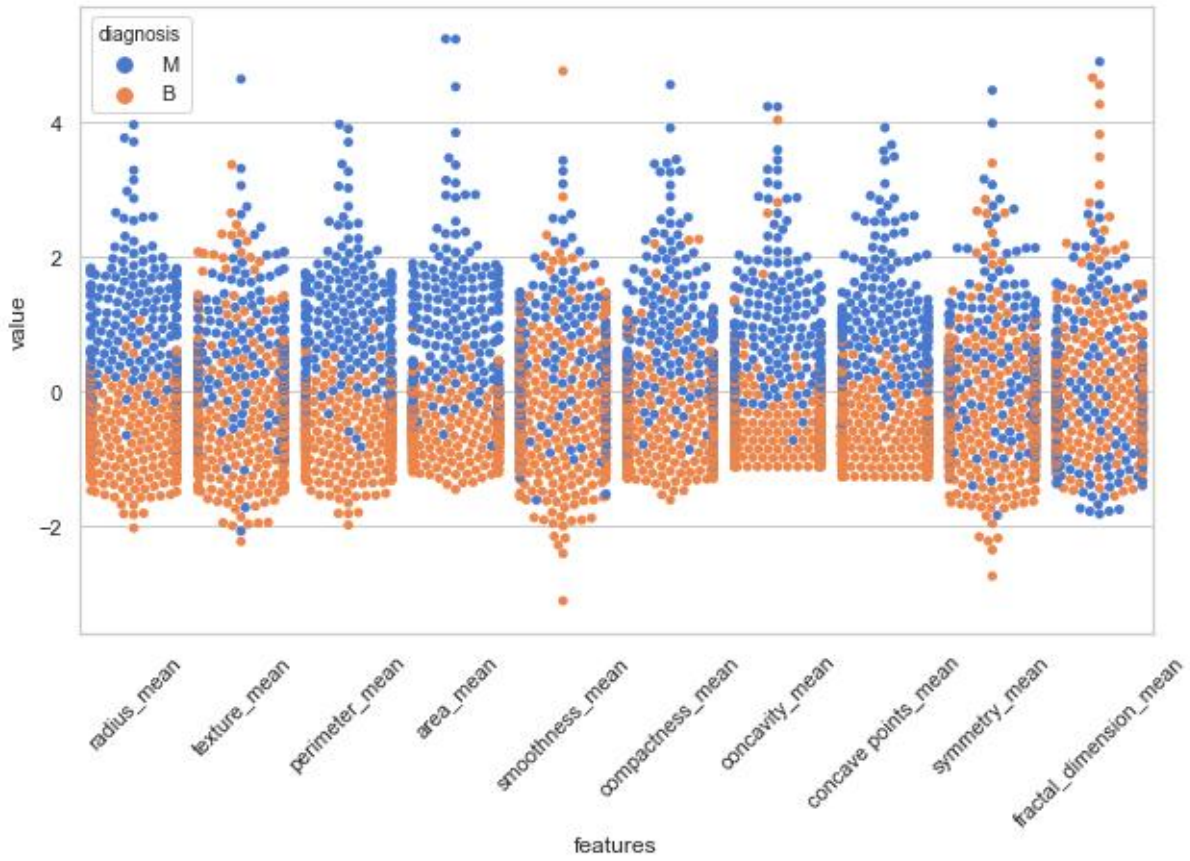
Observations:

This data visualization help to understand the worst concave points and the perimeter and the relationship with the diagnosis target (malignant / benign).



Observations:

With the help of plotting joint plot displays relationship between the variables as well as the univariate graph.



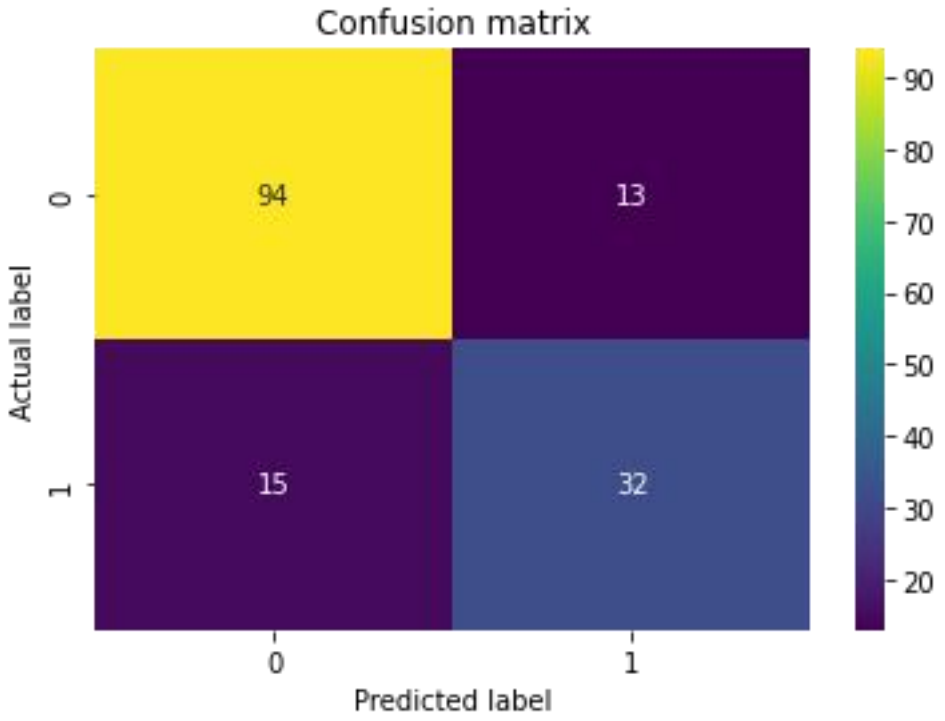
Observations

With the help of swarm plot a kind of scatter plot that is used for representing categorical values in the dataset

KNN (K NEAREST ALGORITHM)

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. 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.

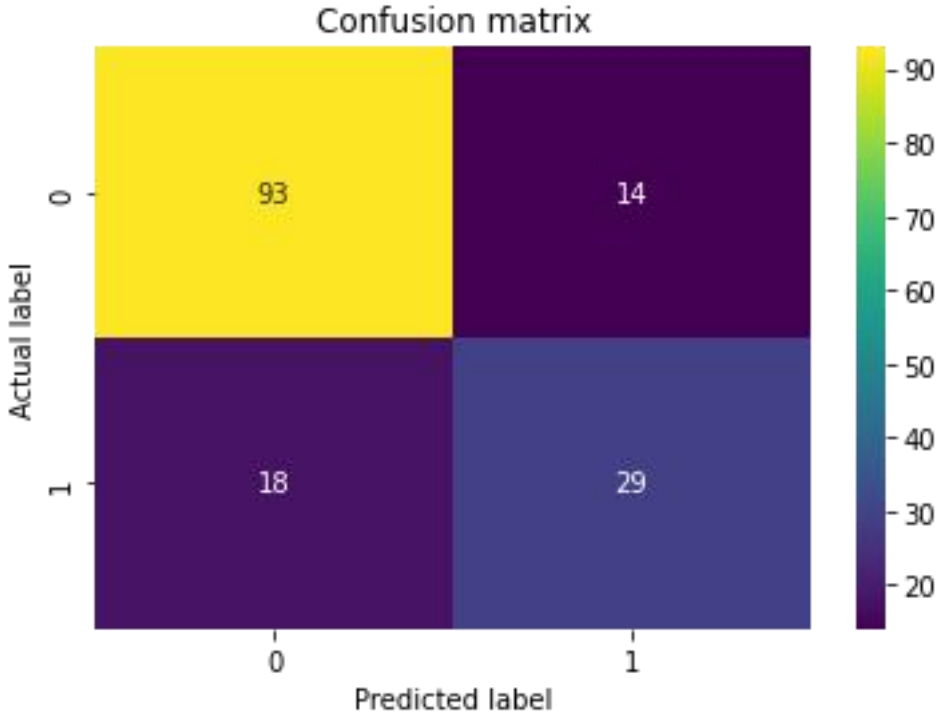


After performing the KNN algorithm on dataset we observe that we have got almost 81 % accuracy and inference from the confusion matrix that almost 126 values have been correctly predicted using knnn. Thus we can understand that almost 81% values of datapoints can be correctly predicted for dataset using Knn Algorithm.

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.

NAIVE 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.

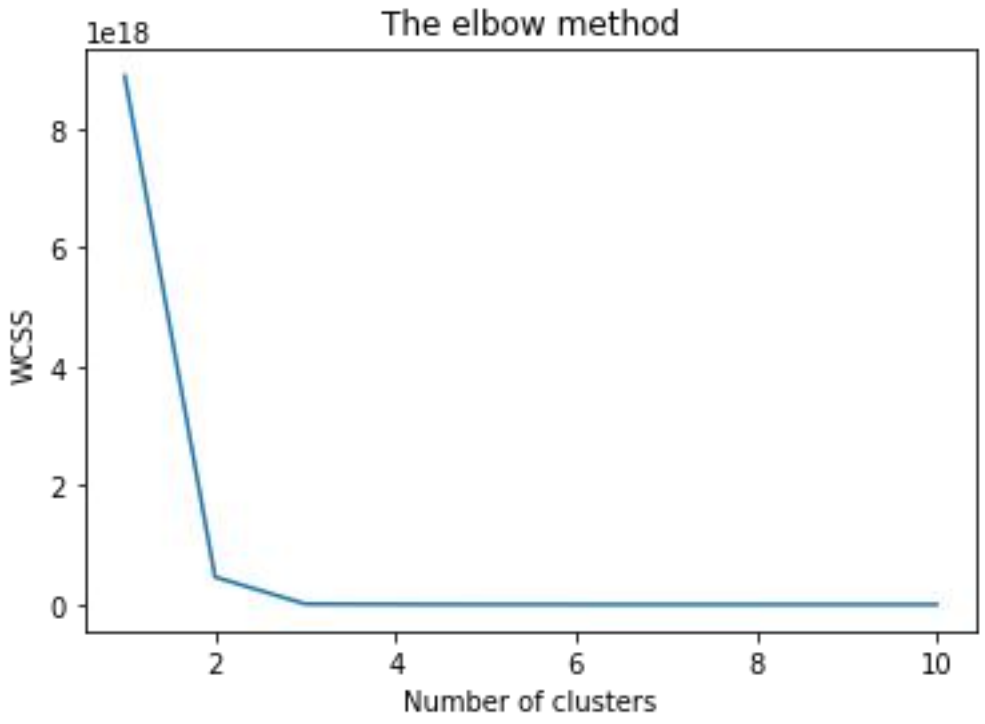


After performing the Naive bayes algorithm on dataset we observe that we have got almost 80 % accuracy and inference from the confusion matrix that almost 122 values have been correctly predicted using Naive Bayes.

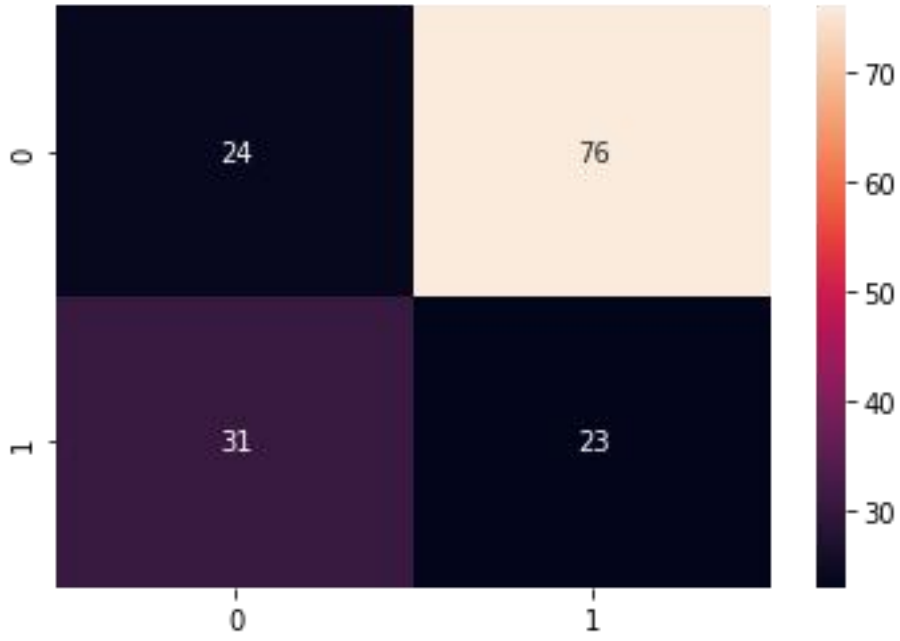
By comparatively knn and naive bayes we can conclude that knn is better predicted as its accuracy is 81% than naive bayes whose accuracy is just 79%.

K MEANS CLUSTERING

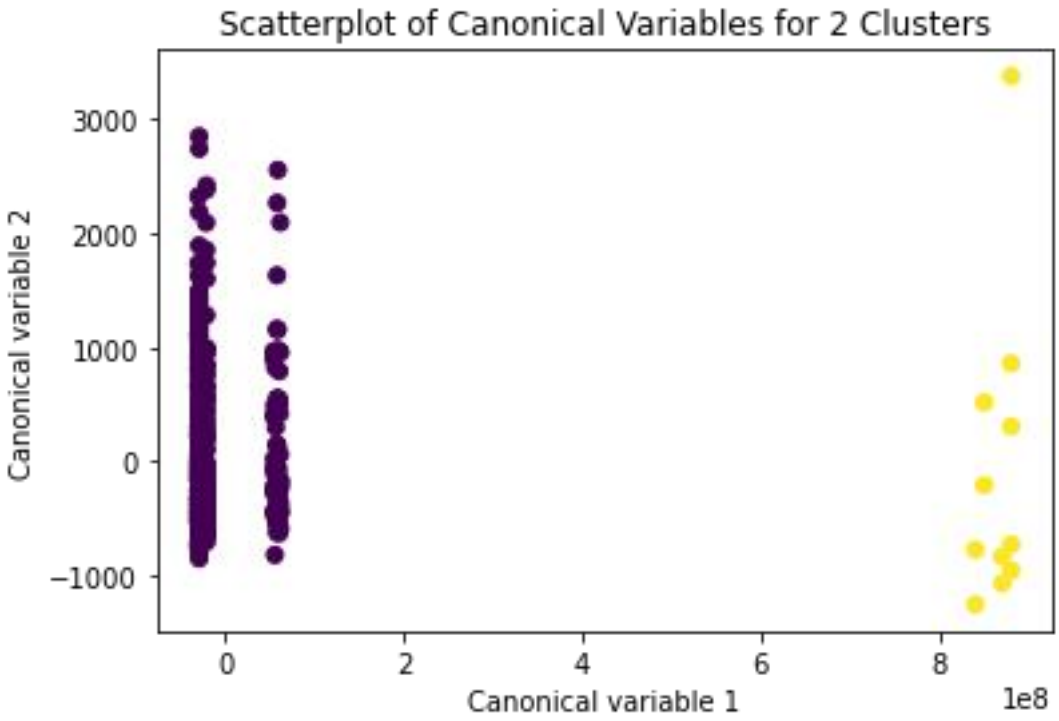
K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabelled dataset into different clusters. It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabelled dataset on its own without the need for any training.



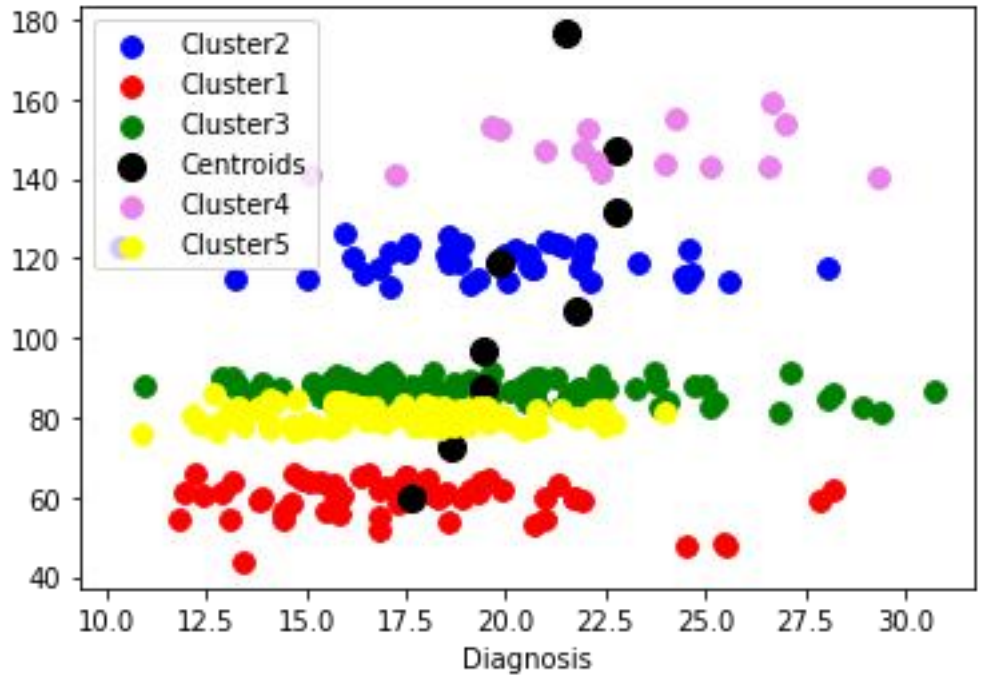
From the above elbow method we can determine the optimal number of clusters into which the data may be clustered here as 2



After performing the K means algorithm on dataset inference from the confusion matrix that almost 47 values have been correctly predicted.



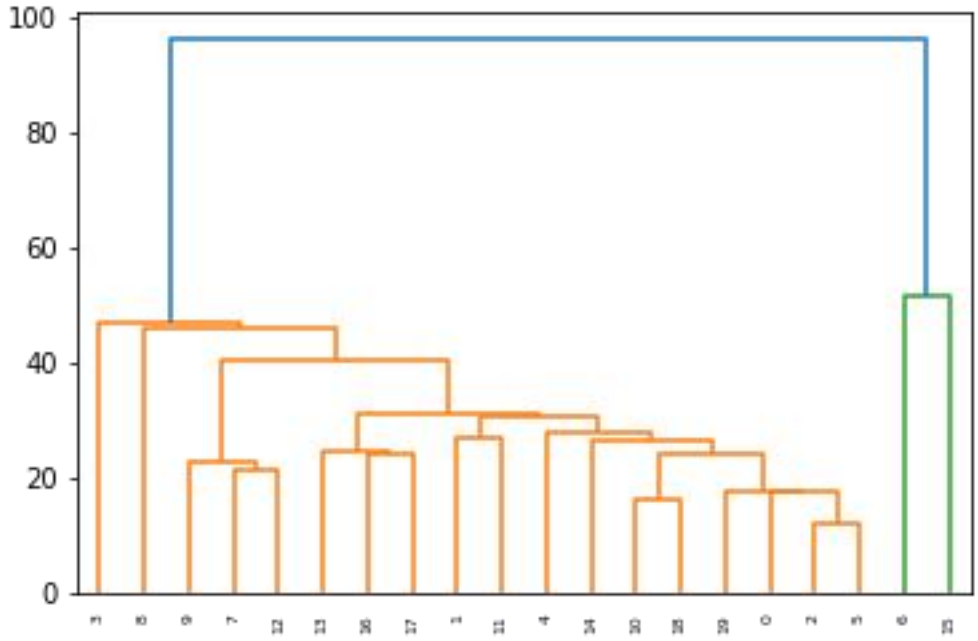
K-means algorithm can easily adjust to the changes. If there are any problems, adjusting the cluster segment will allow changes to easily occur on the algorithm.



The algorithm used is good at segmenting the 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.

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.



The hierarchy of clusters in the form of a tree for the dataset with this tree-shaped structure is dendrogram showing the clustering for the various features.

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.

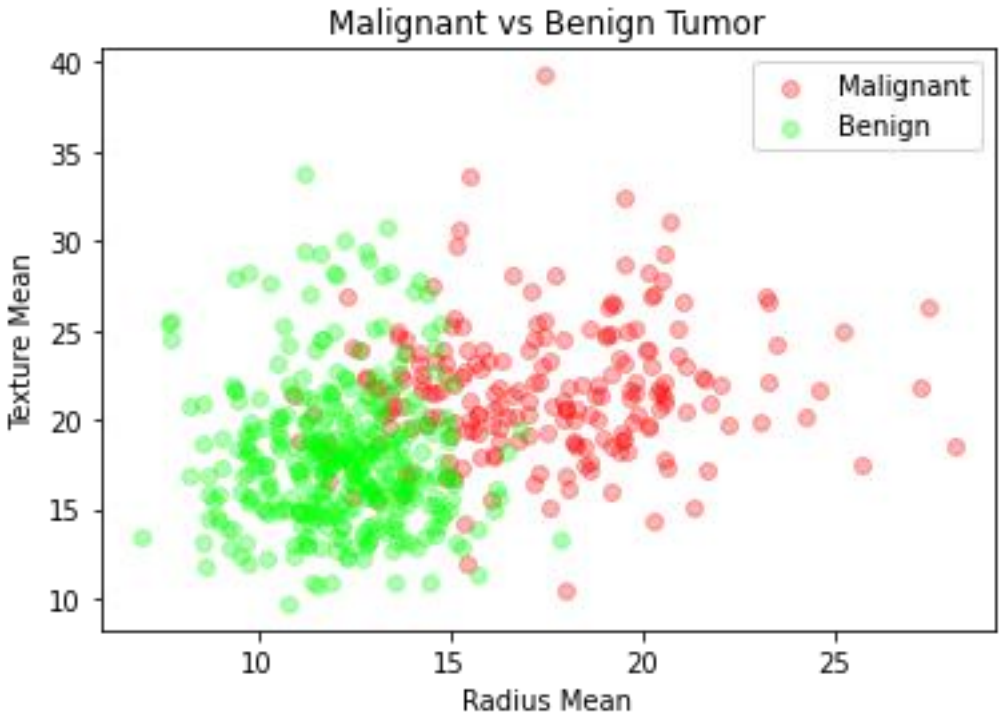
After performing the Hierarchical clustering algorithm on we get 65 % accuracy which is less compared to both KNN and Naïve Bayes which is more than 83%.

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.

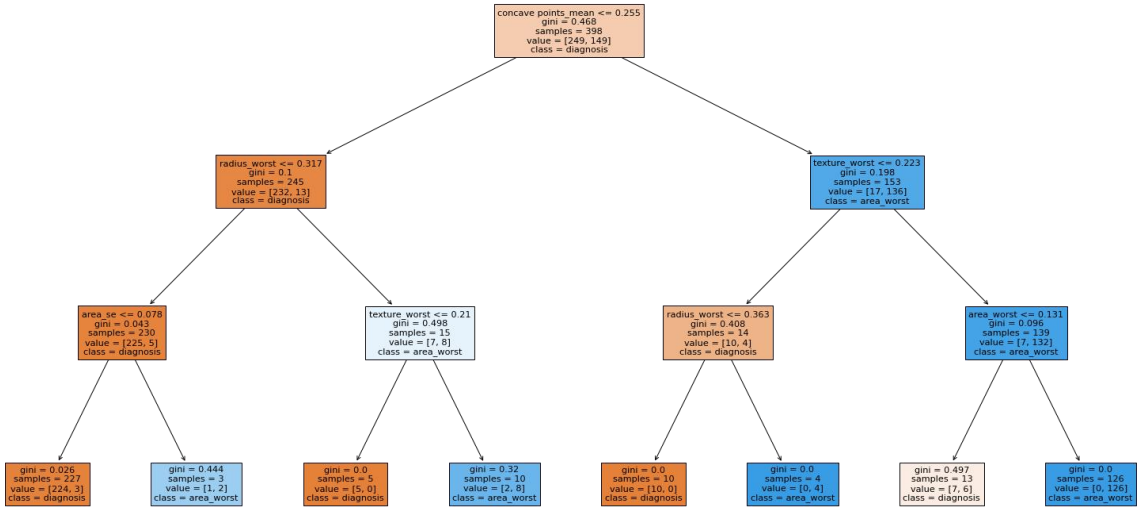
DECISION TREE

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.



IA 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.

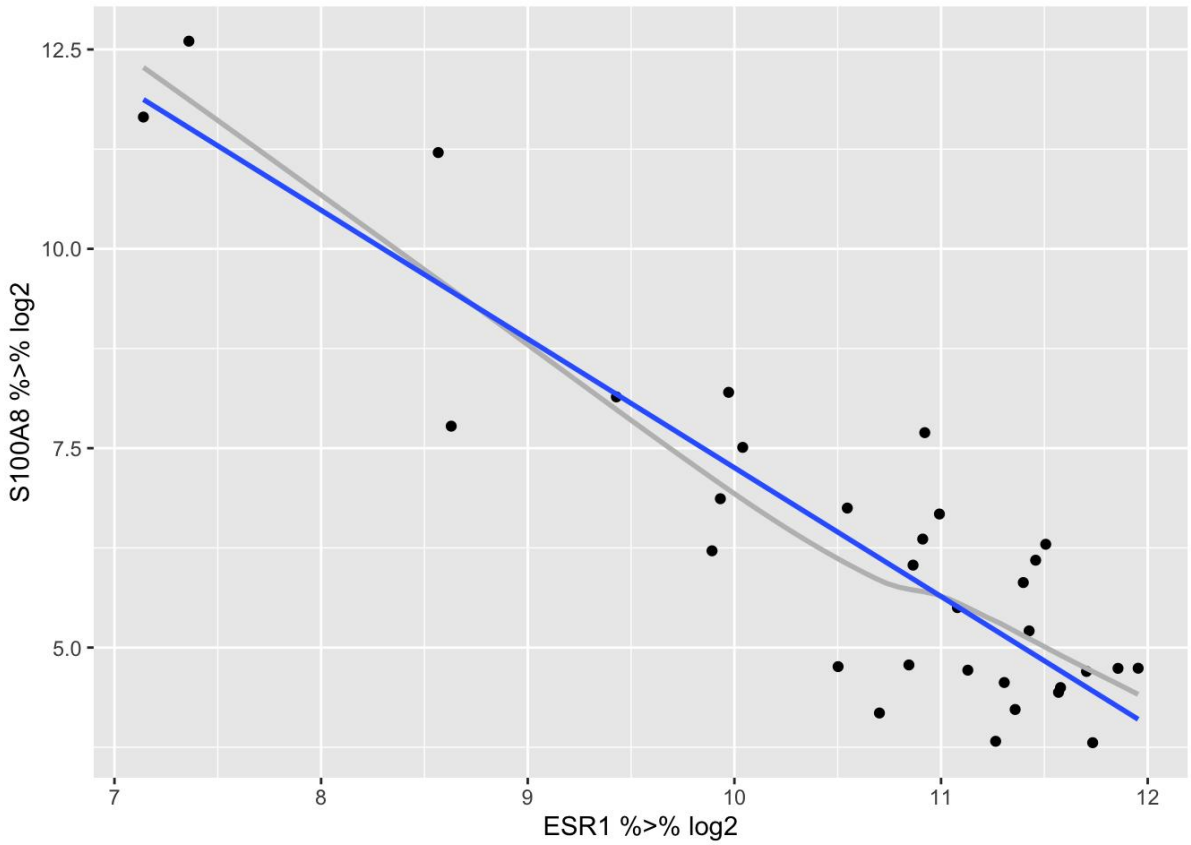


After performing the Decision Tree algorithm on dataset we get 93 % accuracy which is more compared Hierarchical clustering. It can be very useful for solving decision-related problems.

There is less requirement of data cleaning compared to other algorithms. But causes overfitting issues due to computational complexity.

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.

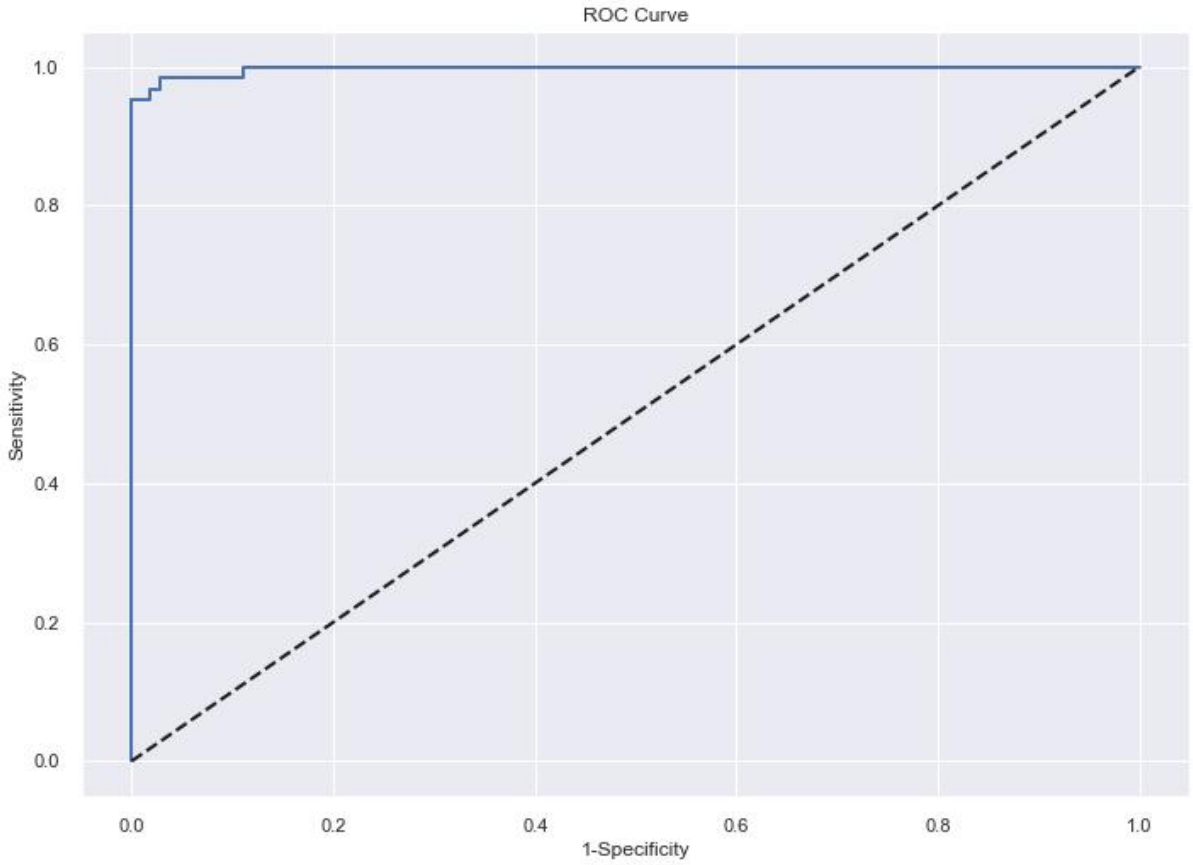


There is an extremely significant negative association between the S100A8 gene expression and that of ESR1 (p<<0.001p<<0.001). A patient with an ESR1 expression that is 2 times the expression of that of another patient will on average have an S100A8 expression that is 3.06 times lower (95% CI [2.48,3.79]).

Logistic Regression

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.

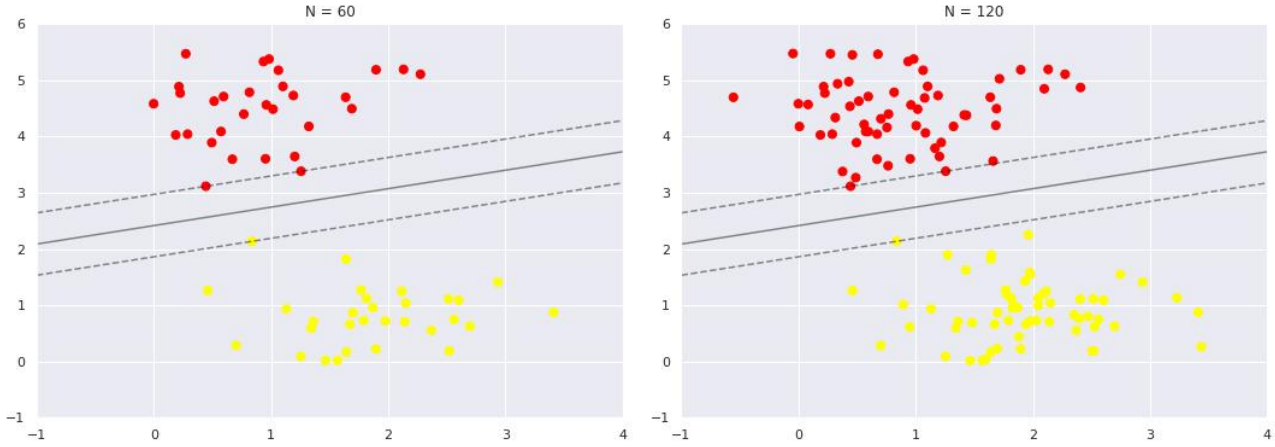


Accuracy = 0.988
Precision = 0.984
Recall = 0.984
F1_score = 0.984

With a cross validation of 5 folds and a threshold > 0.53 and a recall of 98%, following is the performance score of the Logistic Regression model with accuracy as the hyper parameter.

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:

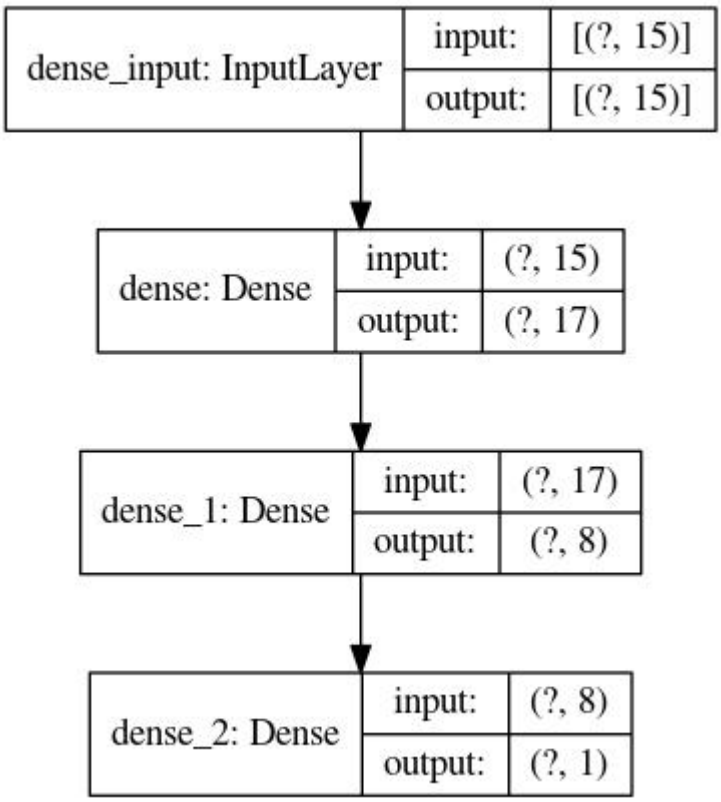
It creates the best line or decision boundary that can segregate n-dimensional space into 2 classes so that we can easily put the new data point in the correct category of either malignant or not.

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.

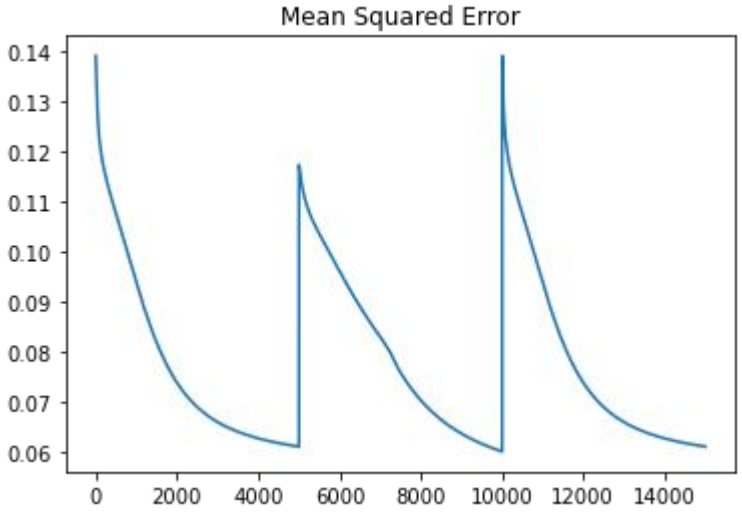
MLP uses backpropogation for training the network. MLP is a deep learning method.



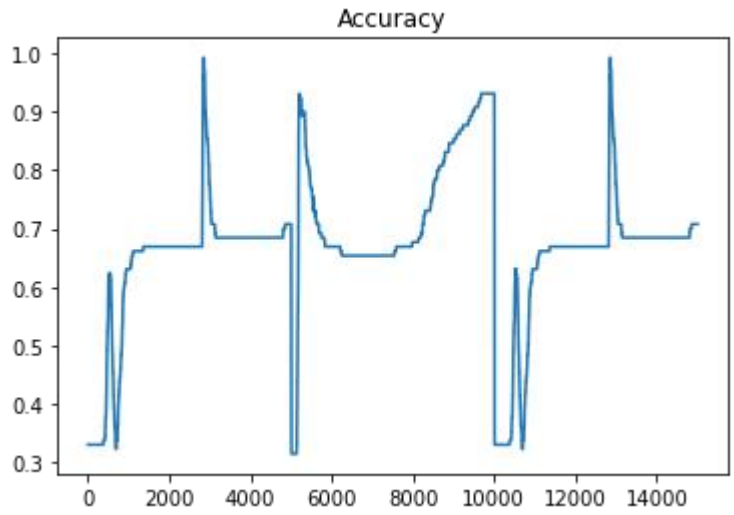
Can be applied for complex non-linear problems, with more large input data with better and quick predictions after training. The proper functioning of the model depends on the quality of the training. Computation can be difficult and time consuming.

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.



Optimises neural networks by propagating the error or loss into a backward direction by calculating the mean squared error.



Loss is calculated for each node and updates its weights accordingly in order to minimize the loss using gradient descent thus giving overall accuracy of 80%.