Task 4 – Rohit Garg

1. Objective

To train a classifier on the dataset to predict the probability of breast cancer.

2. Data

There are 116 rows and 10 columns.

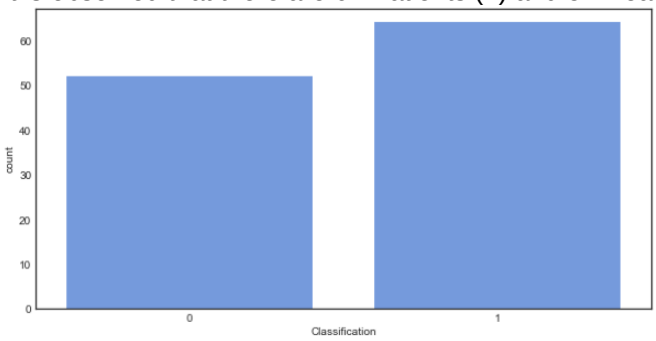
2.1. Independent

It is observed that there are no missing values in any of the independent variables

	count	min	mean	max	std
Adiponectin	116.0	1.7	10.2	38.0	6.8
Age	116.0	24.0	57.3	89.0	16.1
BMI	116.0	18.4	27.6	38.6	5.0
Glucose	116.0	60.0	97.8	201.0	22.5
HOMA	116.0	0.5	2.7	25.1	3.6
Insulin	116.0	2.4	10.0	58.5	10.1
Leptin	116.0	4.3	26.6	90.3	19.2
MCP.1	116.0	45.8	534.6	1,698.4	345.9
Resistin	116.0	3.2	14.7	82.1	12.4

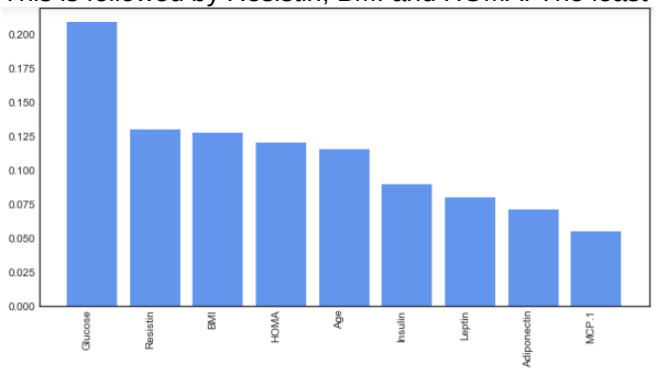
2.2. Dependent

It is observed that there are 64 Patients (1) and 52 Healthy controls (0)



2.3. Feature importance

It is observed that the Glucose has the highest contribution to predicting the breast cancer. This is followed by Resistin, BMI and HOMA. The least contribution is by MCP 1.



3. Model

Gradient boosting combines a set of weak learners and delivers improved prediction accuracy. The outcomes predicted correctly are given lower weight compared to the miss-classified outcomes. The hyper-parameters of this ensemble model can be divided into 3 categories:

- **Tree-Specific Parameters:** These affect each individual tree in the model.
- **Boosting Parameters:** These affect the boosting operation in the model.
- **Miscellaneous Parameters:** Other parameters for overall functioning.

The optimal hyper-parameters are determined using iterative process. The best hyper-parameters are:

- **Loss function:** deviance
- **Max depth:** 5 (maximum depth of the individual regression estimators)
- **Max features:** sqrt (number of features to consider when looking for the best split)
- **Min samples leaf:** 30 (minimum number of samples in the leaf node)
- **Number of estimators:** 20 (number of boosting stages to perform)

4. Performance

It is observed that the cut-off is at **0.50**. Cut-off is decided based on F1-Score.

- The model has an accuracy of **0.862** and F1 score of **0.879**
- The model has AUROC of **0.913**, Gini of **0.825** and KS of **0.675**

5. Conclusion

Based on the performance it can be concluded that the model performs a good job at classifying the breast cancer as Patients (1) or Healthy controls (0)