# Task 1 – Rohit Garg

## Objective

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

## Data

There are 1000 rows and 12 columns.

### Independent

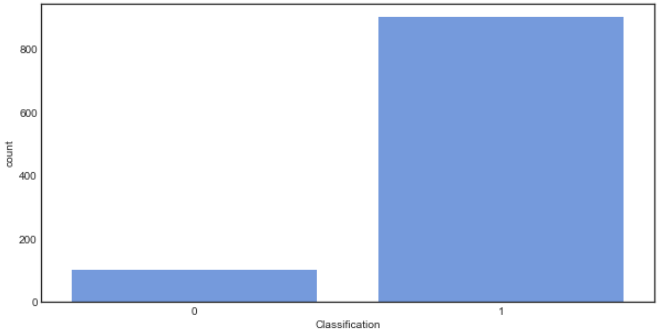
It is observed that there are no missing values in any of the independent variables

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **count** | **min** | **mean** | **max** | **std** |
| **AGE** | 1,000.0 | 20.0 | 53.5 | 79.0 | 8.8 |
| **Urea** | 1,000.0 | 0.5 | 5.1 | 38.9 | 2.9 |
| **Cr** | 1,000.0 | 6.0 | 68.9 | 800.0 | 60.0 |
| **HbA1c** | 1,000.0 | 0.9 | 8.3 | 16.0 | 2.5 |
| **Chol** | 1,000.0 | 0.0 | 4.9 | 10.3 | 1.3 |
| **TG** | 1,000.0 | 0.3 | 2.3 | 13.8 | 1.4 |
| **HDL** | 1,000.0 | 0.2 | 1.2 | 9.9 | 0.7 |
| **LDL** | 1,000.0 | 0.3 | 2.6 | 9.9 | 1.1 |
| **VLDL** | 1,000.0 | 0.1 | 1.9 | 35.0 | 3.7 |
| **BMI** | 1,000.0 | 19.0 | 29.6 | 47.8 | 5.0 |
| **Sex** | 1,000.0 | 0.0 | 0.6 | 1.0 | 0.5 |

### Dependent

It is observed that there are 898 Patients (1) - Diabetic or Predict-Diabetic.

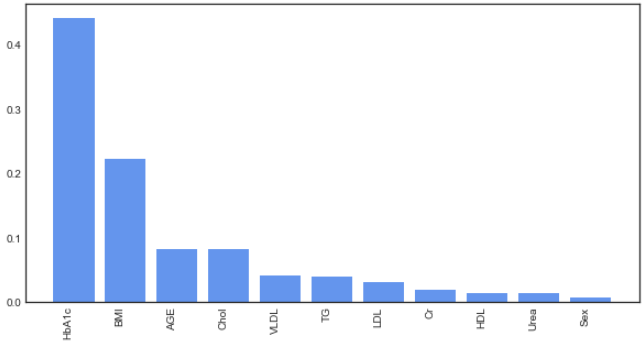
There are 102 Healthy (0) - Non-Diabetic



### Feature importance

It is observed that the HBA1C has the highest contribution to predicting the diabetes.

This is followed by BMI. The least contribution is by Sex.



## 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:** 10 (minimum number of samples in the leaf node)
* **Number of estimators:** 50 (number of boosting stages to perform)

## Performance

It is observed that the cut-off is at **0.15**. Cut-off is decided based on F1-Score.

* The model has an accuracy of **1.00** and F1 score of **1.00**
* The model has AUROC of **1.00**, Gini of **1.00** and KS of **0.98**

## Conclusion

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