Task 3 – Rohit Garg

1. Objective

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

2. Data

There are 756 rows and 755 columns.

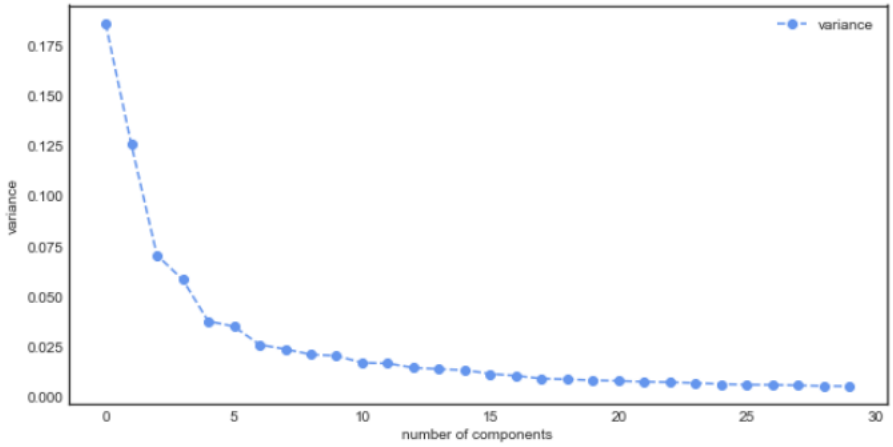
2.1. PCA

Dimension reduction is done by creating new variables (called components). It is done during pre-processing of data for predictive models or forecasting. The reasons for doing dimension reduction include:

- Multi-collinearity issues
- Computational issues of large number of predictors
- Noisy models due to over fitting

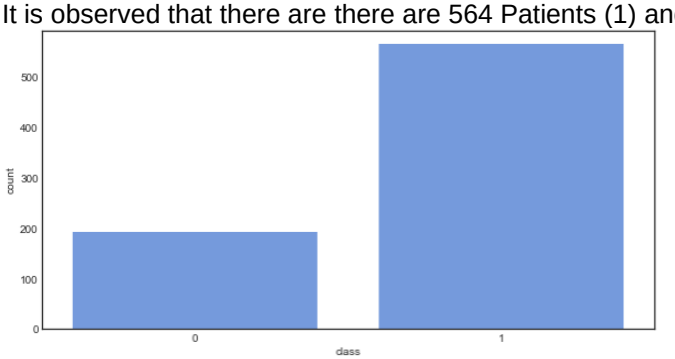
Principal Component Analysis (PCA) is an algorithm for dimension reduction. It linearly transforms the variables into components. Components can be determined by the percentage of variance explained. Before applying PCA the data needs to be scaled.

Original data has 754 variables. PCA is used to create 10 orthogonal components from these variables. These 10 components are considered for developing the model



2.2. Dependent

It is observed that there are there are 564 Patients (1) and 192 Healthy (0)



2.3. Independent

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

	count	min	mean	max	std
C1	756.00	-3.61	0.00	3.65	1.41
C2	756.00	-2.72	0.00	3.77	1.16
C3	756.00	-1.80	0.00	4.12	0.87
C4	756.00	-2.21	0.00	2.52	0.79
C5	756.00	-2.53	0.00	2.90	0.63
C6	756.00	-1.28	0.00	2.80	0.61
C7	756.00	-1.70	0.00	3.44	0.52
C8	756.00	-1.27	0.00	2.11	0.50
C9	756.00	-1.75	0.00	3.74	0.47
C10	756.00	-1.90	0.00	2.08	0.46

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:** exponential (Ada boost algorithm)
- **Max depth:** 10 (maximum depth of the individual regression estimators)
- **Max features:** auto (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:** 40 (number of boosting stages to perform)

4. Performance

It is observed that the cut-off is at **0.30**. Cut-off is decided based on F1-Score.

- The model has an accuracy of **1.0** and F1 score of **1.0**
- The model has AUROC of **1.0**, Gini of **1.0** and KS of **0.938**

5. Conclusion

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