Ensemble Implementation

Junaid Mahmud

Data Import

Data was uploaded to the drive and then imported

Data Preprocessing

The following preprocessing was done.

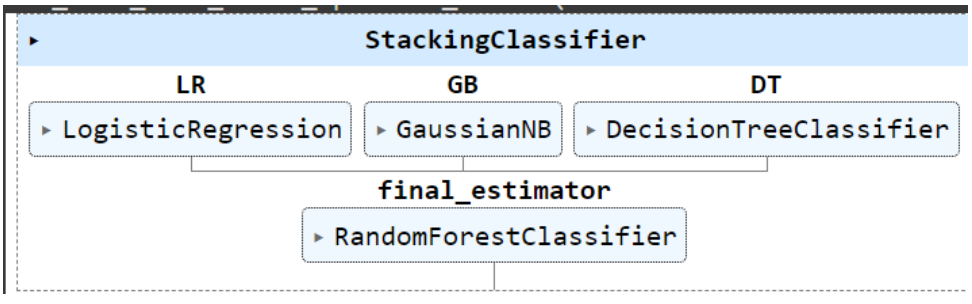
1. Data Encoding of the Categorical Variables.
2. Data Shuffling
3. Scaling
    a. Standard Scaling
    b. Min Max Scaling
4. No Null Value Found
5. No Outlier Found

Ensemble Building

The following estimators were used:

1. Logistic Regression
2. Gaussian NB
3. Decision Tree

The final Model was the **Random Forest Classifier**



Ensemble Method: Stacking Classifier

A stacking Classifier was used to stack the models

Experimentation Details

The experiment was done on the data as follows:

1. Data with no scaling
2. Standard Scaled Data
3. Min Max Scaled Data

The **F1 Score** of the ensemble along with the estimators are as follows

	Ensemble	Logistic Regression	Gaussian NB	Decision Tree	Random Forest Classifier
No Scaling	91%	66%	60%	91%	95%
Standard Scaler	94%	87%	54%	91%	95%
Min Max Scaler	92%	71%	55%	92%	94%

Result Discussion

- The ensemble performed significantly better than Logistic Regression and Gaussian NB.
- The performance of the Decision Tree was very close to the ensemble for unscaled data and mixed scaled data. However, for standard scaled data, the ensemble outperformed.

It appears that, after standard scaling, the model had the best performance which is an F1 score of 94%. But this still is less than the final model Random Forest Classifier (F1 - 95%). In other scaling methods, the f1 score of the stack was substantially lower than the Random Forest Classifier.

The reason assumed behind the stacked ensemble underperforming compared to the Random Forest Classifier is, that the ensembled model acquires both the strengths and weaknesses of the estimators. In this experience, the two estimators (LR, GB) were significantly weak, which made the ensemble perform slightly lower than the Random Forest Classifier. If we use strong estimators in the future, we can have a better performance for this.