Report Analysis Home Assignment 5

1. Input data

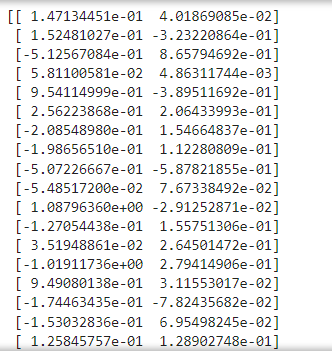


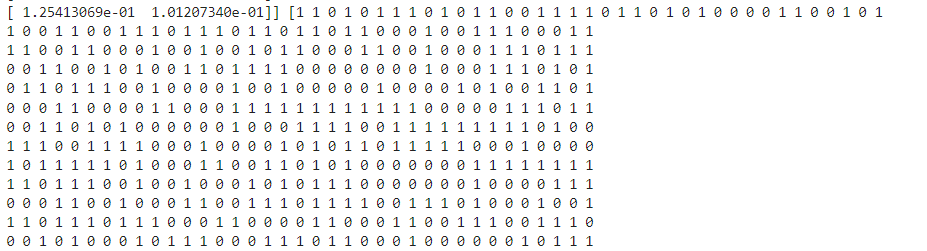
In the initial stage, the sample data is given as input for the cluster named x\_blobs and labels are given as y\_blobs. Here is the following output with input codes for first step (a).

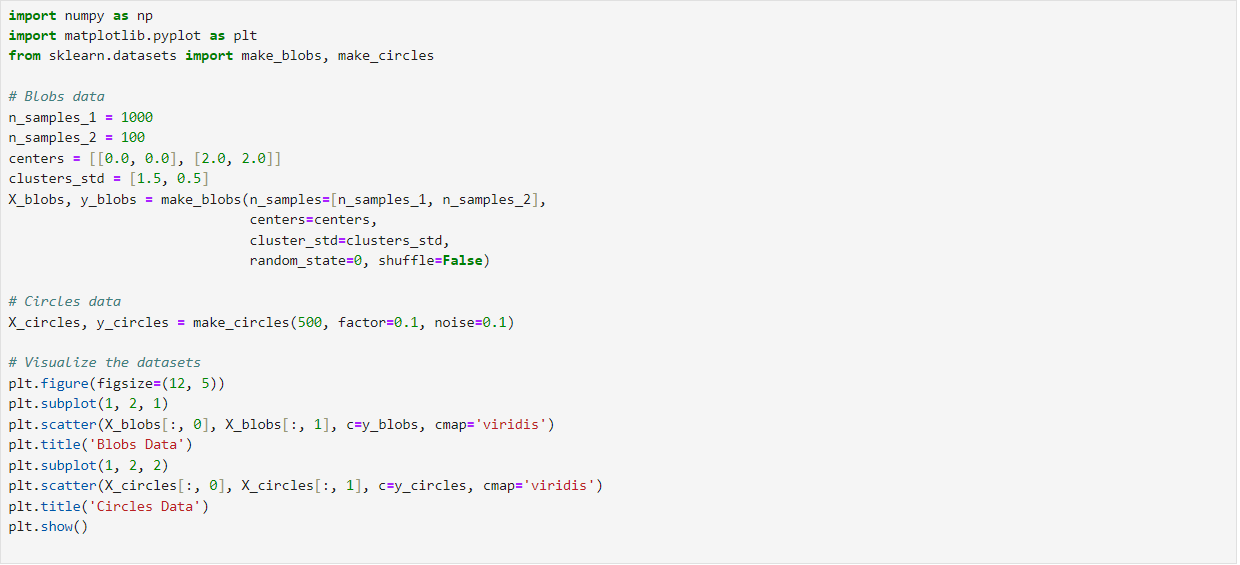
1. Input data another dataset for the task.

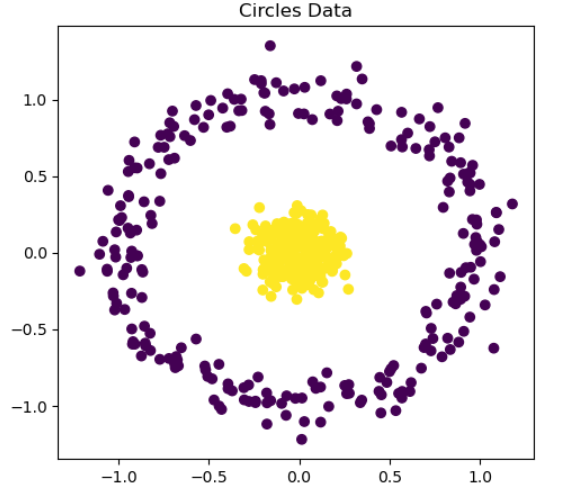
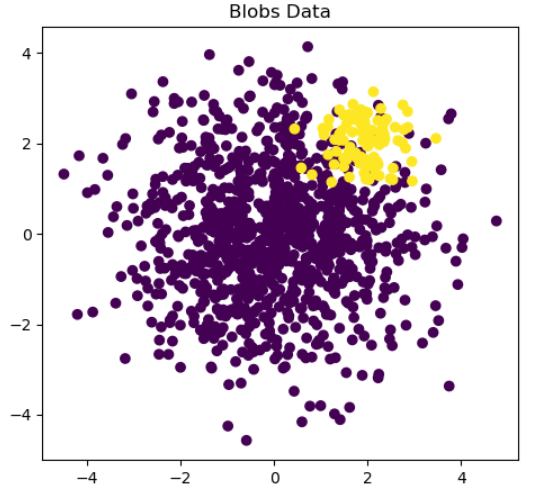


Output of this data set is.

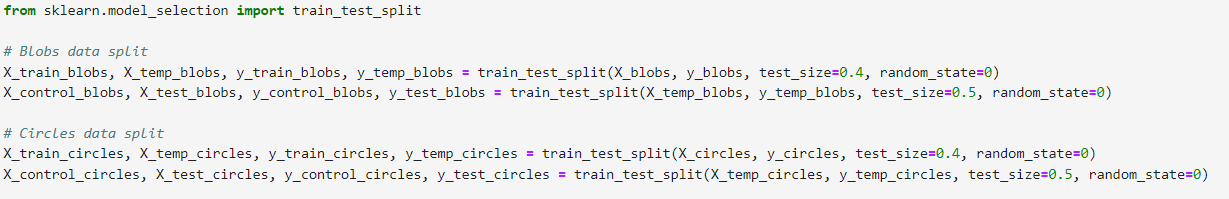




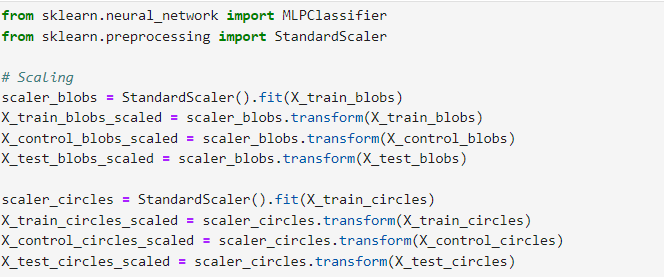
1. With the above input dataset here is the combined code in addition to that I created a graph to visualize both datasets for proper understanding. 



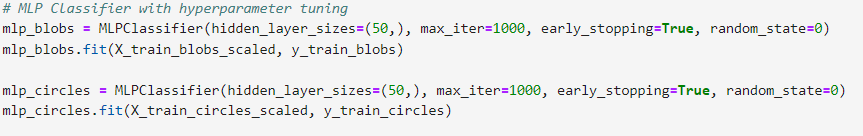
1. In the next step, Divide data for the training, control, and test samples. Control sample used for the hyperparameters setting. Test sample use for the final evaluation of the model quality. Here is the input code.



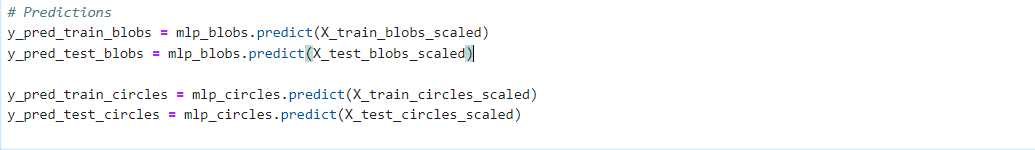
1. Create NN models with the usage of MLPClassifier. Data should be scaled. The same scaling method should be applied to the training, control, and test data set. For MLPClassifier try to find the best value of the hyperparameters.



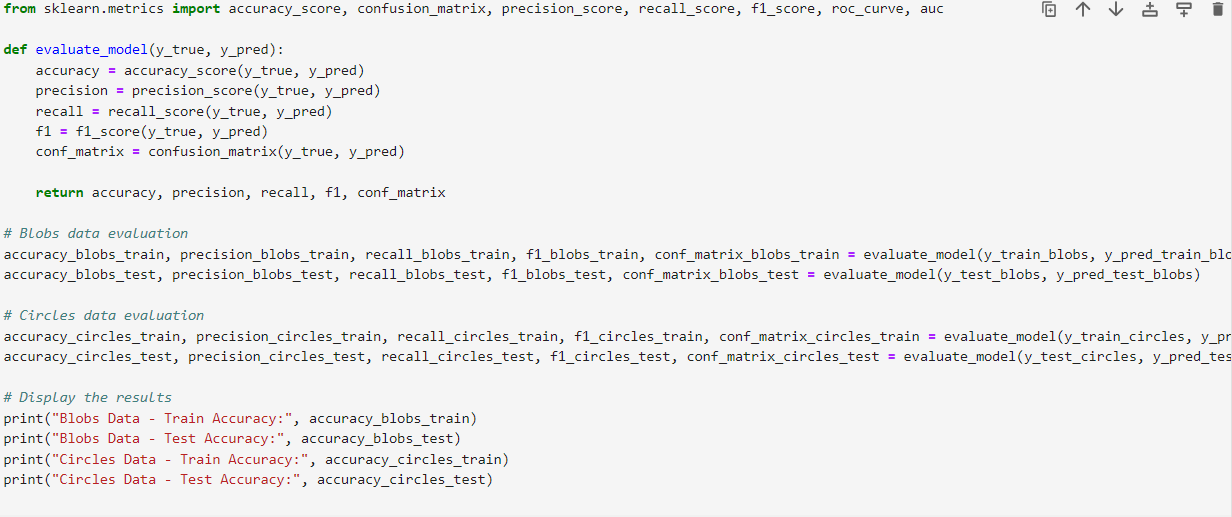
1. Create different architectures of the NN by variation of the hidden\_layer\_sizes. Use early\_stopping=True.Show lost functions at the first and last iterations for each variant.

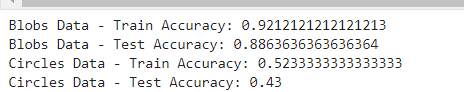


1. Make predictions based on the NN models.



1. Evaluate the quality of the NN models based on accuracy, confusion matrix, precision, recall, F1 score, ROC curve, AUC. Make conclusions.





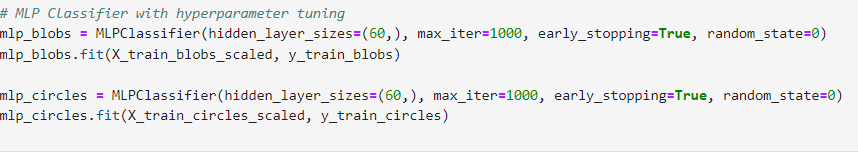
Analysis table for different hidden layers and selecting best value for our model for both test and train factor.

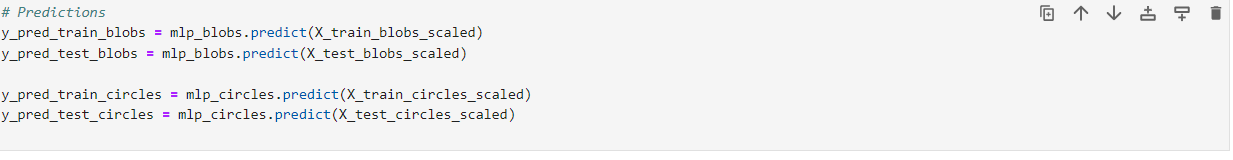
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Hidden Layer Size | Blobs Data - Train Accuracy | Blobs Data - Test Accuracy | Circles Data - Train Accuracy | Circles Data - Test Accuracy |
| 50 | 0.9212 | 0.8864 | 0.5233 | 0.4300 |
| 60 | 0.9227 | 0.8864 | 0.936 | 0.91 |
| 70 | 0.9212 | 0.8864 | 0.6200 | 0.6700 |
| 80 | 0.9212 | 0.8864 | 0.6200 | 0.6700 |

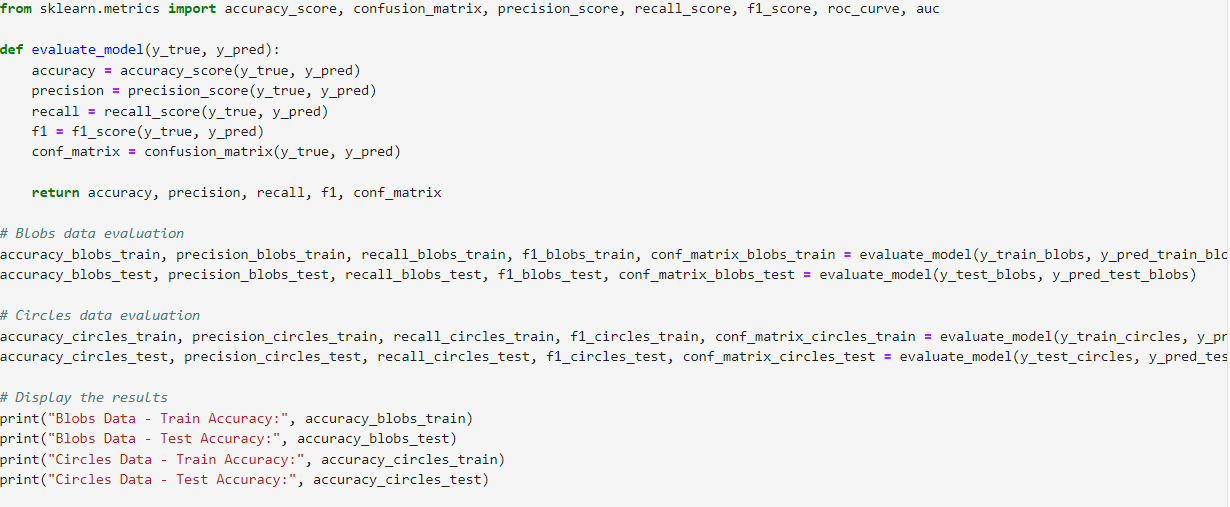
Conclusion

As it is clear that with hidden layer size 60 gives better accuracy for both the data set. Moreover, it is equally noticed that the hidden layer for dataset blobs actually does not create a huge impact in most of the cases it is actually at the same time circle data is highly impacted by this moreover for circle data more we are increasing the hidden layer sizes the more accuracy for train and test is reducing.

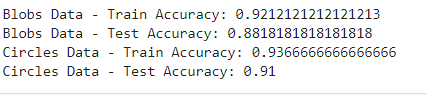
Finally input code is.



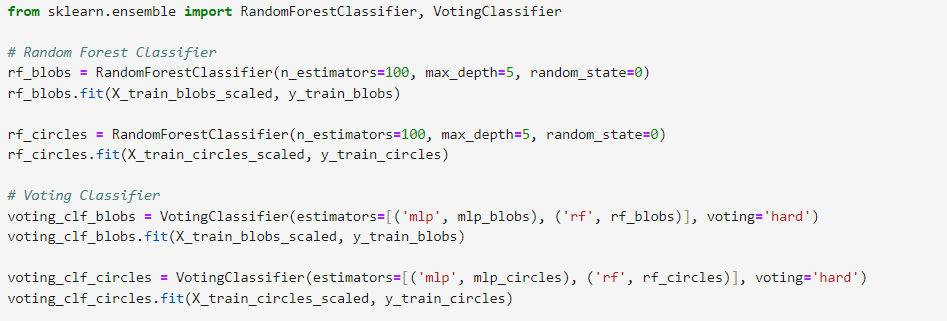


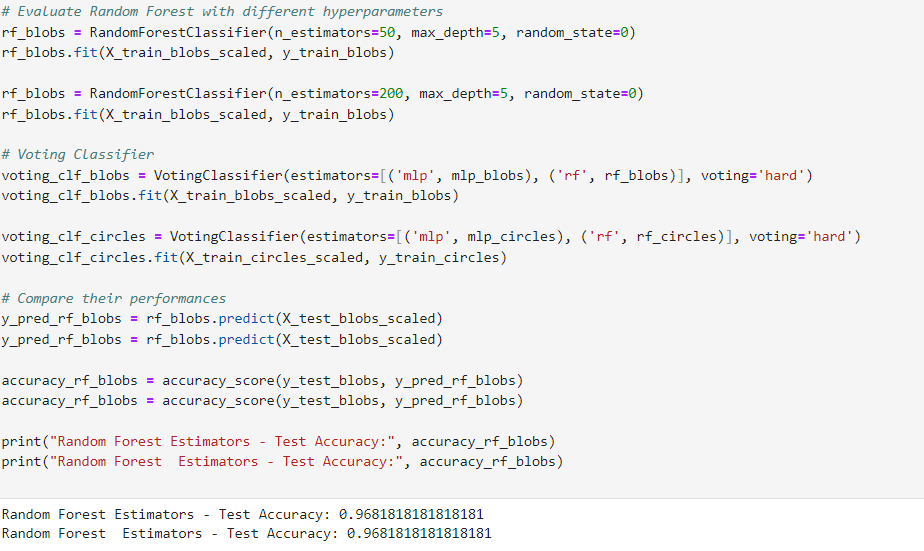


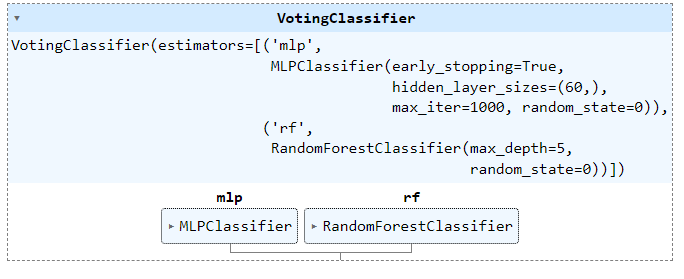
And final output is:-



1. Create Ensembles of models with the usage of the Random Forest Classifier. Consider different values of max\_depth of max\_features bootstrap,n\_estimators Additionally use VotingClassifier to consider different voting and weights.







Here is the the analysis table where a comparison is made by changing the different values of n\_eatimators and max\_depth and trying to find the best model for RandomForestClassifier.

|  |  |
| --- | --- |
| Model Configuration | Random Forest - Test Accuracy |
| n\_estimators=200, max\_depth=5 | 0.9682 |
| n\_estimators=200, max\_depth=6 | 0.9591 |
| n\_estimators=50, max\_depth=5 | 0.9636 |
| n\_estimators=100, max\_depth=5 | 0.9682 |
| n\_estimators=100, max\_depth=6 | 0.9591 |

As we can see the best n\_estimators=100, max\_depth=5 so further models have been created using this below are the input code and output.

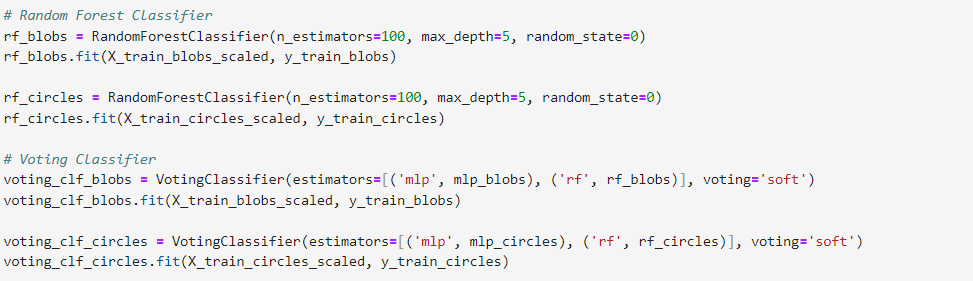
10. Create Ensembles Random forest

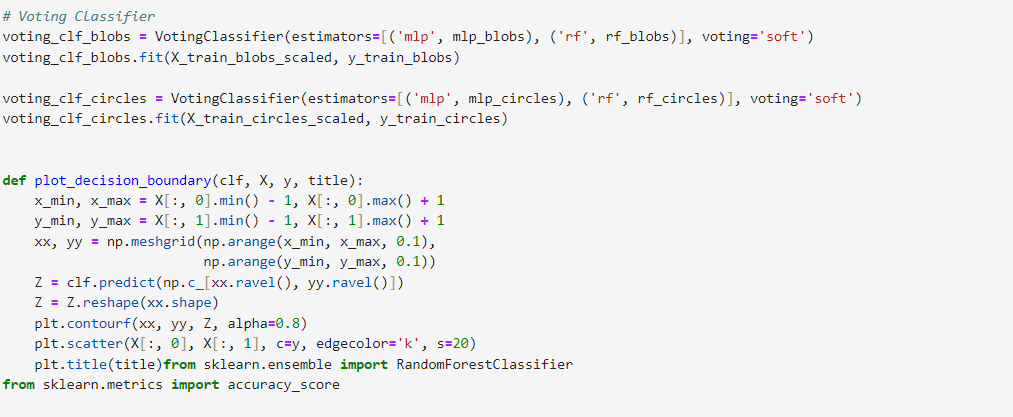
a. For base\_estimator use one/several models with tge basic parameters.

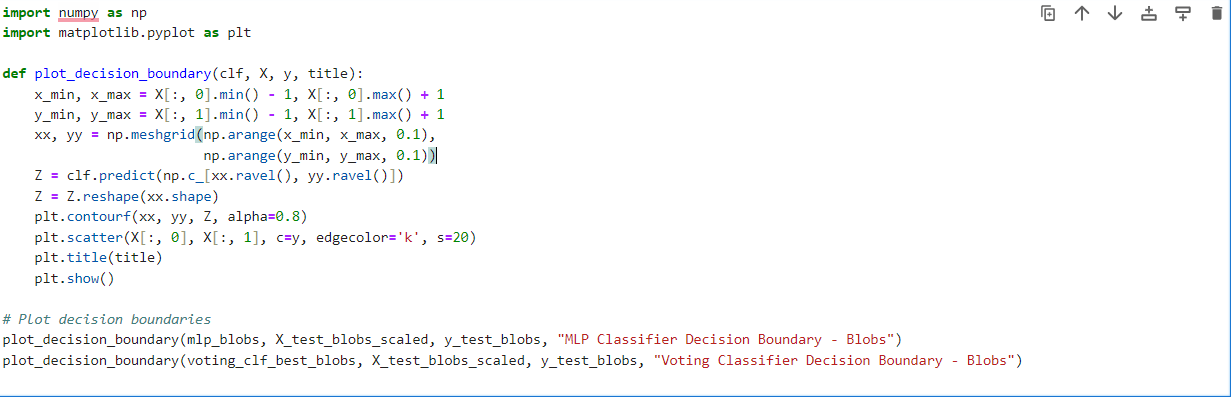
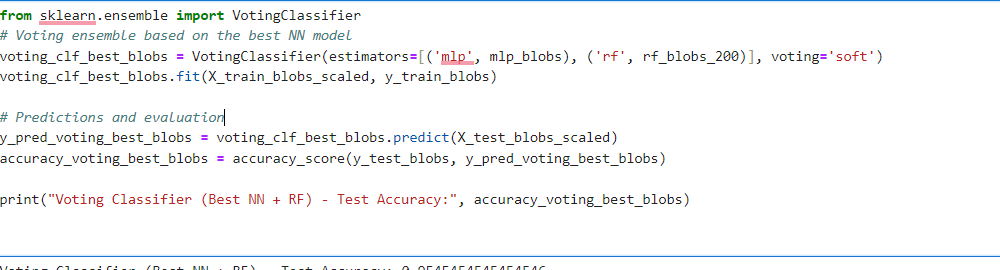
Compare different Ensembles which differs only with the hyperparameters.

b. Create graphs for dependencies of n\_estimators for Ensembles and

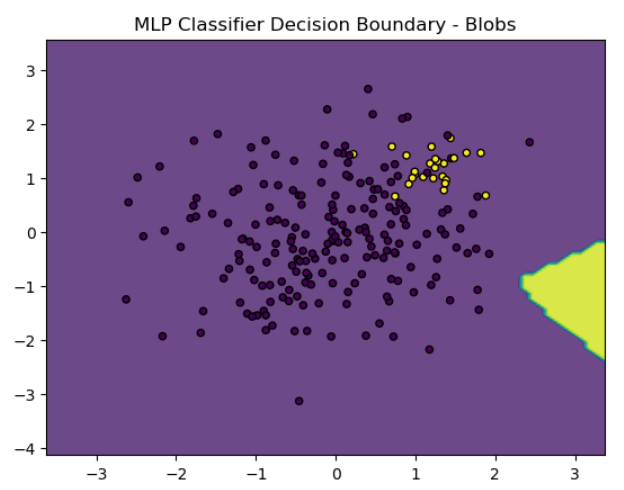
individual models.

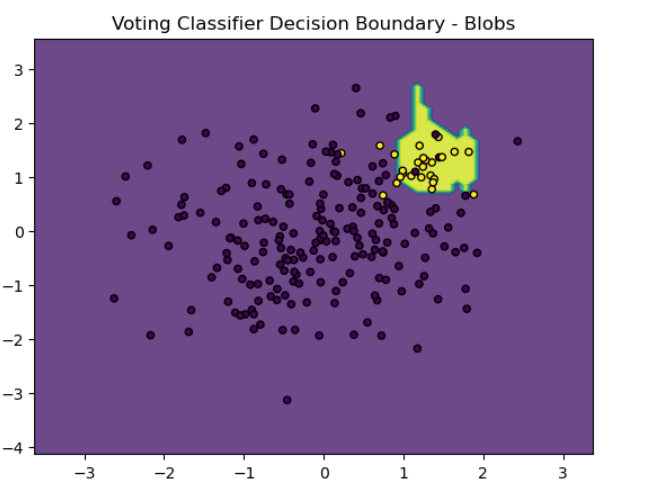






Changing the different values of the base parameter analysis of different graphs for decision boundary.

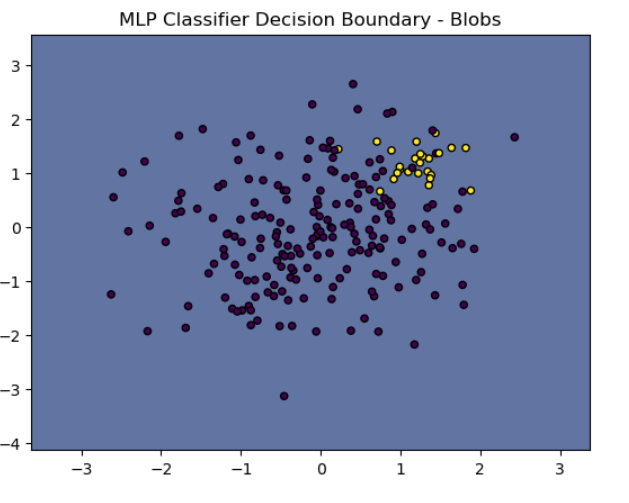


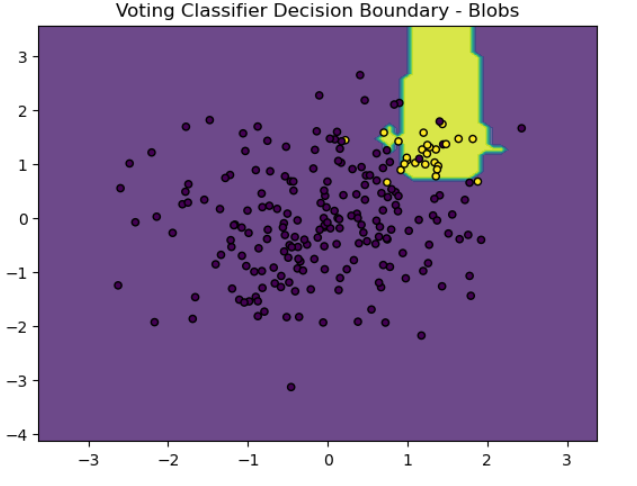


Analysis of the graph the MLP classifier has non-linear decision boundary, which might indicates good performance but also a risk of overfitting.

The voting classifier combines the decisions of multiple classifiers, resulting in a potentially more balanced decision boundary that could generalize better for the data which is unseen.

2. Another Graph for Analysis.



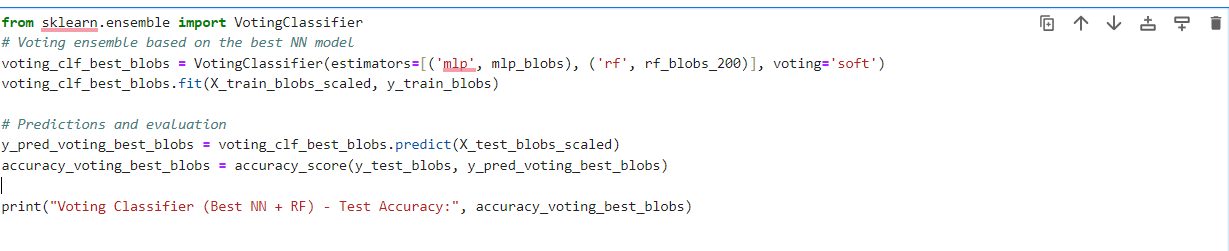


After making certain changes in the hyperparameter we could achieve better result and making more clear results.

1. Create ensemble voting based on the best NN model from the previous stages.

Here is the input code followed by the output.

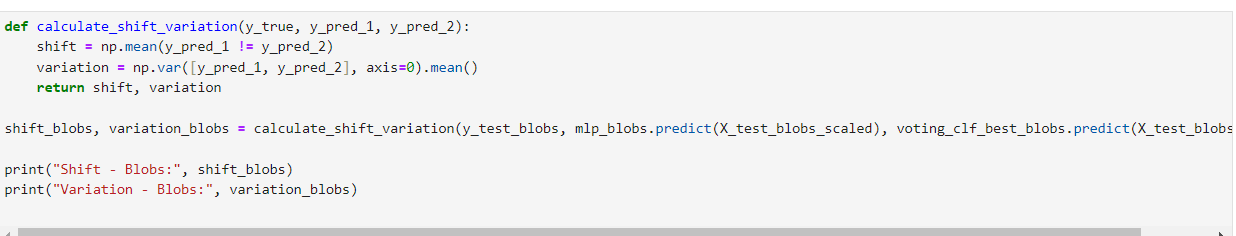






1. Calculate the shift and variation for the basic model and Ensemble.

Here is the input followed by the output.





1. Analysis and conclusion drawn on the basis of shift and variation.

High Accuracy

The Voting Classifier, combining the best Neural Network (NN) and Random Forest (RF), achieved a high test accuracy of 0.9681, indicating strong performance in classifying the blobs dataset.

Low Variation

A variation of 0.02952 indicates low variability in the model’s predictions, implying that the model is consistent in its decision making processs different subsets of the data.

Shift

A shift of 0.1182 suggests that the model’s predictions have some deviation when compared to the actual distribution. This could be due to slight overfitting or inherent variability in the data.