LAB Assignment 6
CSL 2050

By :- Akshat Jain B21CS005

1. We have used make_moons to Construct the dataset.
    - 1.1.1. We have Plotted the make_moons dataset of size 1000 with the red colour denoting one class and blue denoting another class.
    - 1.1.2. We then split the data into train and test and then trained a simple Decision tree Classifier and plotted it. After that, We have iteratively performed hyperparameter tuning for max depth and then plotted the decision boundary for its best iteration according to accuracy.
    - 1.1.3. We trained a BaggingClassifier from sklearn on the same dataset and plotted the decision boundary.
    - 1.1.4. We trained a RandomForest from sklearn on the same dataset and plotted the decision boundary. As we have plotted the decision boundary for all three classifiers, we can see subtle differences in all three decision boundaries. The accuracy of the classifiers varies as acc(Random Forest) > acc(Tuned DTC) > acc(Bagging Classifier)
    - 1.1.5. We vary the number of estimators for both the random forest classifier and Bagging Classifier. We then Find what the optimum value for the no of estimators is w.r.t the accuracy metric.
    - 1.2. Here we make the Bagging Classifier from Scratch.
        - 1.2.1. Here we have made a class called BaggingClassifierScratch with the fit and predict Functions which works like a normal Bagging classifier from sklearn.
        - 1.2.2. We then implement this From scratch classifier with n_estimators = 10
2. Boosting
    - 2.1. We train an AdaBoost Classifier with n_estimators = 100
    - 2.2. We train an XGboost Classifier n_estimators = 100 and subsample = 0.7
    - 2.3. We then print the accuracies for both the above models for training and test datasets
    - 2.4. Along with 2.5
    - 2.5. In 2.4 and 2.5 we have Plotted the Graphs for a range of values of num_leaves and Max_depth. In green, we have training Accuracy, and in

red, we have testing Accuracy. We find the maximum test accuracy is at num_leaves = 2 and max_depth = 1. We can also say reliably that somewhere just after num_leaves = five and max_depth = 5, the model starts overfitting because the testing accuracy reaches its second peak; however, the training accuracy increases until it reaches 1.

2.6. Some Good parameters to vary for combatting overfitting are:

2.6.1. Keeping num_leaves low

2.6.2. Keeping max_depth to avoid growing a deep tree

2.6.3. Use feature sub-sampling by set feature_fraction

These are some of the methods which can we used to combat overfitting.

2.7. The best performance is from the tuned LightGBM model. We find that acc(lgbm) > acc(adaboost) > acc(xgboost)

3. Voting Classifier

3.1. We train a Bayes classifier and print its accuracy

3.2. We then select DTC, Random Forest, Adaboost models along with the Bayes classifiers as the estimators in the Voting Classifier. We train for both ‘soft’ and ‘hard’ voters and see that the ‘hard’ classifier has better accuracy. We also find out that the accuracy of the Voting classifier is the best one we have achieved in this complete lab.