Part 1 – Tree Classifier

The programs for Part 1 are implemented and stored in the folder HW3\_Part1.

The files in the folder contain the following prefix

DTC – Decision Tree Classifier

BC – Bagging Classifier

RFC – Random Forest Classifier

GBC – Gradient Boosting Classifier

The other portion of the filename follows the naming convention of the provided datasets.

The program does not take any command line arguments and can be run in the following manner:

> python .\BC\_c300\_d100.py

Kindly change the TRAIN\_PATH, VALIDATION\_PATH and TEST\_PATH at line number 22, 23 and 24 respectively for the file to be run.

Kindly provide the complete path up to the filename in both the mentioned paths.

For questions 1, 2, 3 and 4, the best parameter settings, accuracy and F1 scores have been tabulated in the file ‘Report.xlsx’ in the first tab ‘Part1’

Question 5.

* Gradient Boosting Classifier yields the best overall generalization accuracy and F1 score. Bagging and Random Forest Classifiers train each tree independently on different subsets of the training dataset and then combines the prediction from all models. Gradient boosting on the other hand is an iterative approach, where it trains one tree at a time and each new tree corrects the mistakes made by the previously trained model. These weak trees are then combined to form a single strong model. However, in many cases it was also found that Random Forest performed better than Gradient Boosting and Bagging Classifier performed better than Random Forest and this can possibly be due to overfitting of the data by Gradient Boosting and Random Forest respectively. Also randomly selected subsets in Random Forest makes the model more robust and less prone to noise in the data.
* The accuracy and F1 score increase for each of the four models as the size of the training data is increased.
* As seen from the obtained results, the accuracy and F1 score increase for each of the four classifiers as the number of clauses are increased.

Question 6

* Random Forest Classifier yielded the best overall generalization accuracy. Random Forest Classifier performs better on real world data as it is noisy and contains many missing values. Random Forest is also harder to overfit as compared to Gradient Boosting.