REPORT on ASSIGNMENT 1

Afia Afrin

Student ID: 1105040

Section: A2

|  |  |  |  |
| --- | --- | --- | --- |
| Evaluation Criterion | Average Accuracy | Average Precision | Average Recall |
| Information Gain | 0.946494 | 0.950757 | 0.870196 |

1. Table: Following table shows average accuracy, precision and recall after running the program for 100 times:
2. Supervised learning is comprised of two phases:
3. Training phase
4. Testing phase

If we used all data for training, the program would show too much high accuracy while testing on the same dataset. Training data set is a very small portion of the entire distribution. So it helps to capture the distribution properly by separating some of the training data as testing data set. Furthermore, we would not be able to know if it over fits the training dataset, as all data have been used for training.

1. Training and testing data set have been chosen randomly. In some cases, the program has performed too much better in training set in comparison to testing set, which indicates over fitting. Usually this occurs due to noisy training data set, which generates a more specific tree instead of a smaller tree causing over fitting.