Assignment 1 Report
Decision Tree Learning for Cancer Diagnosis

Name: Fahmid Morshed Fahid

ID: 1105021

Evaluation Criterion	**Accuracy**	**Precision**	**Recall**	**Overfitting**
Information Gain	0.9276	0.9533	.8361	Moderately Overfit

Question: Why are you dividing the data set 80% into training and 20% into test data rather than using 100% data for training?

Answer:

If we had used all the data on training, then no new data would have been available for testing our decision tree, thus, we had to use some data from the training data set which would have imposed OVERFITTING (optimistic accuracy). This is because, the noises in the training data (if any) are already accounted while forming the Decision Tree, so feeding the subset of the training data will result more aligned with the correct result. But in real scenario, the new data will not mimic the training data accordingly and thus will have different kind of noises, so our algorithm will fall behind the expected accuracy. So, to avoid OVERFITTING, we separated the data set into two disconnected group such as TRANING SET (80%) and TEST SET (20%).

Question: Do you see evidence of overfitting in some experiments? Explain.

Answer:

Yes, some experiments show evidence of overfitting. This is because our decision tree is not absolutely optimal. To explain more thoroughly, the tree was created based on finite number of training data (about 600), and there are cases and patterns of noises that are not confined within these training data; theoretically we need all the combinations of attributes in the learning set to overcome the overfitting. As that is not realistic, there are cases where our decision tree performs well under the already familiar patterns and noises (if any) of the training data set (from which it learned the Decision Tree), and performs poorly in new separate test data set.