# REPORT ON ASSIGNMENT 1

Decision Tree Learning for Cancer Diagnosis

SUBMITTED BY:
SIAMUL KARIM KHAN
1105104

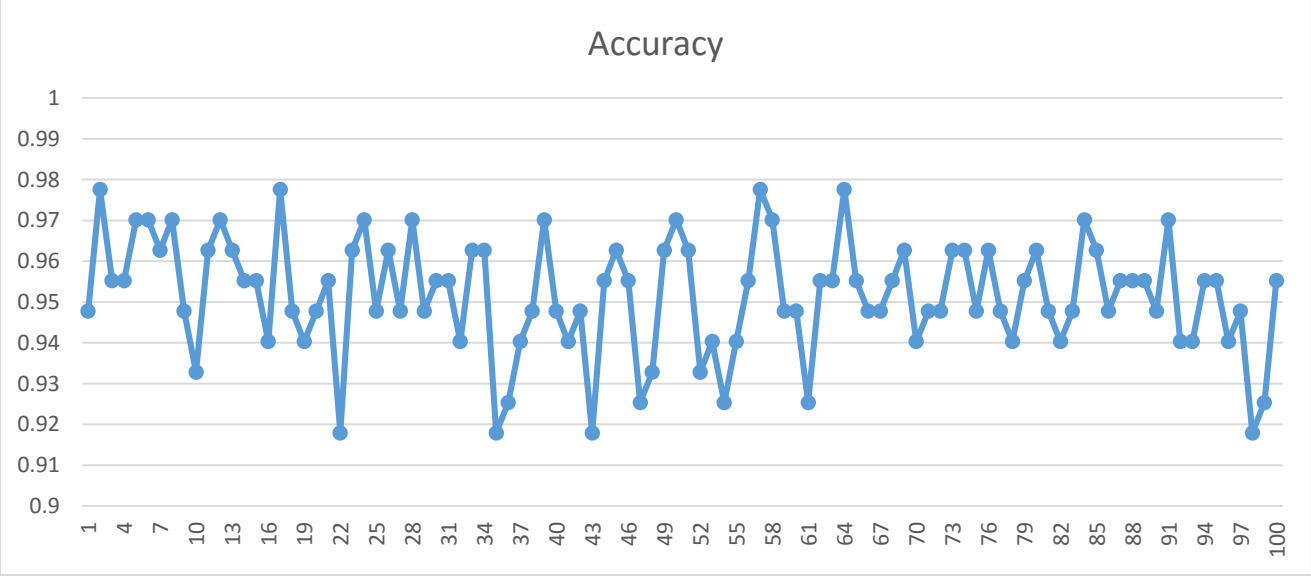
| Evaluation Criterion | Accuracy | Precision | Recall |
|---|---|---|---|
| Information Gain | 0.9516417910447769 | 0.9486077168567739 | 0.9081975204076883 |

Q. Why are you dividing the dataset 80% into training and 20% into test data rather than using 100% data for training?

A: If we use 100% data for training, we may over-fit the data by considering the random fluctuations in our data while fitting our model. So we separate the dataset into training and test set. The test set is unlikely to exhibit the same random fluctuations as the training set and so can be used to assess our model and examine whether we are over-fitting the false characteristics of the training set.

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

A:




In each of our iteration we are fitting a tree to a random 80% set of data. We can see that some trees perform better on their corresponding test sets. The tree that perform worse can be thought of as over-fitted as opposed to those that perform better on the test set.