Report on Assignment 1:
Decision Tree Learning for Cancer Diagnosis

A. S. M. Ahsan-Ul-Haque

1205021

1. The performance measures for 5-fold cross validation is given below:

Cross Validation	1	2	3	4	5
True positive rate (sensitivity, recall, hit rate)	93.1034	87.2727%	76.0638%	69.6078%	75.6345%
True negative rate (specificity)	91.4365%	94.3243%	97.4063%	89.7281%	94.9853%
Positive predictive value (precision)	83.9378%	87.2727%	94.0789%	80.6818%	89.759%
Negative predictive value	96.5015%	94.3243%	88.2507%	82.7298%	87.027%
False positive rate (fall-out)	8.56354%	5.67568%	2.59366%	10.2719%	5.01475%
False negative rate (miss rate)	6.89655%	12.7273%	23.9362%	30.3922%	24.3655%
False discovery rate	16.0622%	12.7273%	5.92105%	19.3182%	10.241%
F1 score	88.2834%	87.2727%	84.1176%	74.7368%	82.0937%

2. (a) Why are you using cross validation? Do the dataset justify it?

Ans.

We are using cross validation for mainly two reasons:

i. The main purpose of machine learning is to perform well on new data. We cannot be sure about how good a classifier is (i.e. how well it predicts on data) based on training data alone. A classifier might be overfitting or underfitting and still perform well on the data it trains on. Cross validation provides us with new data with respect to the classifier.
ii. At the same time, the new data must be labeled (i.e. we must know for certain which class the new data comes from), otherwise we cannot determine recall, precision etc. Cross validation also helps with this because we already know the label of the new data.

2. (b) Besides accuracy, which of the criteria mentioned above should be used in cross validation for the given data set? Explain.

Ans.

Beside accuracy, precision and recall should also be used in cross validation. We can also use F1 Score alone because it takes both precision and recall into account.

Accuracy measures what portion of the data the classifier correctly classifies, i.e.

$$Accuracy = \frac{TP + FP}{P + N}$$

If the data is biased (i.e. the number of positive samples is much greater than the number of negative samples or vice-versa) then the accuracy measure is not very helpful. In real life, this case may occur frequently, for example: data samples of benign vs malignant cancer patients where the number of malignant cancer patients is very low.

Precision measures what portion of the predicted positive result is in fact positive, i.e.,

$$Precision = \frac{TP}{TP + FP}$$

But if the classifier predicts negative for all the samples, the precision will be very high (in fact 100%). This is where recall comes into play. Recall is defined as,

$$Recall = \frac{TP}{TP + FN}$$

So, recall is the measure of the portion of total positive data the classifier has correctly classified as positive. So, as in the previous example, if the classifier classifies every sample as negative, FN will be very high and Recall will be very low.

So, precision and recall check and balance each other.

The combined measurement of precision and recall can be addressed by their F1 Score (which is their harmonic mean), as defined below:

$$F1\ Score = \left[\frac{\frac{1}{Precision} + \frac{1}{Recall}}{2}\right]^{-1}$$

If either the precision or the recall is low, F1 score will be affected and become very low. So, F1 Score is the one measure that should be used to get a good idea about a classifier.