Assignment 1: Decision tree learning for cancer diagnosis

In this mini-project, you will implement a decision-tree algorithm and apply it to breast cancer diagnosis. For each patient, an image of a fine needle aspirate (FNA) of a breast mass was taken, and nine features in the image potentially correlated with breast cancer were extracted. Your task is to develop a decision tree algorithm, learn from data, and predict for new patients whether they have breast cancer. Dataset can be downloaded from U.C. Irvine Machine Learning Repository.

1. Collect the data set from my website. Each patient is represented by one line, with columns separated by commas: the first one is the identifier number, the last is the class (benign or malignant), the rest are attribute values, which are integers ranging from 1 to 10. The attributes are (in case you are curious): Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses. (Note that the UCI document page specifies a different number of attributes, because it refers to a set of several related datasets. For detailed information of the dataset that we use here, see this document.)
2. Implement the ID3 decision tree learner, as described in Chapter 3 of Mitchell. You may program in C/C++, Java. Your program should assume input in the above format.
3. Implement both *misclassification impurity* and *information gain* for evaluation criterion. Also, implement split stopping using chi-square test.
4. Divide the data set randomly between training (80%) and testing (20%) sets. Use your algorithm to train a decision tree classifier and report accuracy on test. Run the same experiment 100 times. Then calculate average test performances (accuracy, precision, recall, f-measure, g-mean).
5. Compare performances by varying the evaluation criteria. Make a table as follows:

Evaluation Criteria	Accuracy	Precision	Recall	F-measure	G-mean
misclassification impurity					
information gain					

6. Answer the following:
    a. Which evaluation criterion and confidence level work well? Why?
    b. Do you see evidence of overfitting in some experiments? Explain.