Decision Tree Computer Assignment

1. Train and test a decision tree classifier on the Fisher Iris dataset.

a. Load the csv files using the python library `pandas`.

b. Use the Decision Tree Classifier from the python library `sklearn` (scikit learn) to train a decision tree on the dataset.

c. Use `entropy` (i.e. information gain) as the `criterion` instead of the default `gini`.

d. Write your own function to calculate the classification accuracy of the decision tree. You may not use any library here. The function should have the input data and labels, and the classifier, as input arguments, so that the same function can be used to calculate training accuracy as well as testing accuracy.

e. Report the training and testing accuracy of the classifier using your function.

f. Now vary `max_depth` and `min_samples_leaf` from their default values so that the classifier can generalise better. Your aim in changing these parameters is to reduce overfitting, so change the parameters accordingly.

g. Train the new decision tree and report its training and testing accuracy using your function. The new accuracy may or may not be lower than the previous accuracy.

h. Also train a decision tree classifier using `gini` as the `criterion`, first with default parameter values, then with the changed values to reduce overfitting. Report the training and testing accuracies in both cases.

i. Write a brief README.txt file mentioning how to run your code on the command line, including any command-line arguments you may have used.

Details about the Fisher Iris Dataset:

1. Each instance has 4 continuous-valued attributes: sepal length, sepal width, petal length and petal width of a flower in centimetres.

2. Using these, each flower is to be classified as one of 3 classes: Iris setosa, Iris versicolor, and Iris virginica.

3. The training set `iris_train.csv` has 120 examples and the testing set `iris_test.csv` has 30 examples.