Indian Institute of Technology Indore
Discipline of Computer Science & Engineering
CS 403/603 Machine Learning
Assignment 2 -Decision Tree Learning

Some general instructions:

- Plagiarism in any form will not be tolerated.
- You are allowed to do any number of submissions before the deadline. However, only the last submission will be stored and used for evaluation.
- Submission of the assignment should be made using Google Assignments platform only.

In this assignment, you will be required to implement the Decision Tree algorithm from scratch using both (a) Information Gain and (b) Gini Index, to decide on the splitting attribute.

DataSet:
https://archive.ics.uci.edu/ml/machine-learning-databases/00267/data_banknote_authentication.txt (It is a txt-file but by changing the txt suffix to csv you can use the Pandas library's read_csv()-function.)
The data is about the ***authentication of banknotes***. If you like you can read more about the data here https://archive.ics.uci.edu/ml/datasets/banknote+authentication#

Attribute Information:

1. variance of Wavelet Transformed image (continuous)
2. skewness of Wavelet Transformed image (continuous)
3. curtosis of Wavelet Transformed image (continuous)
4. entropy of image (continuous)
5. *class* (integer)

In our data the classes are represented by "0" for **real** and "1" for **fake**.
Divide the dataset into 80:20 Train:Test ratio and perform the following task

Task:
(A) Train your decision tree classifier on the train-data (where you will use “class”), using the impurity measure:
a. Information Gain
b. Gini Index

Test your model on test-data (where the “class” label is unseen).
After prediction, report the individual accuracies on the test data obtained using (a) and (b).
Note that the “class” field should not be used in the classification process.

For both cases, write your program and print out the decision tree in format of your choice(but include a description in **description.txt**)

(B) Repeat the experiment using the decision tree algorithm implemented in scikit learn , using both Information Gain and Gini index. Report the accuracies on test data.

(**Deliverables)** Your report should contain :

1. The decision tree
2. The value of Information Gain and Gini Index of the root node using :
    a. Your model
    b. scikit learn
3. The labels generated on the test data and accuracy on the test data using :
    a. Your model
    b. scikit learn
