MACHINE INTELLIGENCE AND EXPERT SYSTEMS
AUTUMN SEMESTER - 2021
PROGRAMMING ASSIGNMENT-1 (DECISION TREE)

Q1. Build a decision tree from scratch using only Python, numpy and pandas, and calculate the accuracy of this tree on the *Fischer Iris* Dataset.

I. Load the csv data using the pandas library.
II. Define a new class called DecisionTree with methods **calc_accuracy**(data, labels, ...), **calc_info_gain**(data, attribute_name, ...) along with any other helper methods that you feel are necessary. The criterion for the decision tree will be **information gain** at every node.
III. The **calc_accuracy** function will compute the classification accuracy of the decision tree. The function should have the input data and labels as input arguments, so that the same function can be used to calculate training accuracy as well as testing accuracy.
IV. You are allowed to control the extent of the training process. It can be limited either by the number of iterations, a minimum amount of information gain necessary to make a split on the basis of some attribute, a combination of both or something else entirely. Unique and interesting ideas here would be rewarded with bonus points.
V. Report the training and testing accuracy of the model.
VI. Now implement a new method within the DecisionTree class called **prune_tree**() which will attempt to prune the tree in order to mitigate the false predictions due to overfitting of data and improve accuracy. The input parameters that decide whether a particular subtree should be pruned to a leaf or not are flexible and need to be decided by the student. Similar to (IV), the challenge is to come up with the right set of input parameters and algorithm that would lead to a maximum increase in the accuracy post-pruning.
VII. Write a README.txt file explaining how to run your code on the command line, including any command-line arguments you may have used. Also explain how you decided to structure your building and training process (into different functions and/or classes), how and when the decision to split the tree is made (through information gain's calculation and comparison) and at what point does the training stop (max iterations reached / gain in information is too less / both / something else?). Also provide the rationale for the prune_tree() method's implementation in your code.
VIII. The evaluation will be done on the basis of the following parameters (in decreasing order of their importance/weight) :
    A. program's structure and readability (include comments wherever necessary)
    B. the answer to (VII) (smart, logical and efficient answers will be rewarded)
    C. accuracy of the decision tree
    D. the efficiency of the training process' implementation (avoid repetitive calculations).

NOTE: Plagiarism will be dealt with severely. Any attempts to copy the code from another student or the internet will incur heavy penalties. The idea of this assignment is to test a student's programming skills and understanding of decision trees. Therefore, all the implementation decisions regarding the code's structure and training process of the model must be yours.

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.

Link to the data:
https://drive.google.com/drive/folders/1EnBi3nzEnrdjaLS421iNaU5_dWU77mO8?usp=sharing