Machine Learning

Assignment 1: Decision trees

1. Build a decision tree by taking as input a maximum depth and by randomly splitting the dataset as 80/20 split i.e., 80% for training and 20% for testing. Provide the accuracy by averaging over 10 random 80/20 splits. Consider that particular tree which provides the best test accuracy as the desired one. 30 marks
2. What is the best possible depth limit to be used for your dataset? Provide a plot explaining the same. 20 marks
3. Perform the pruning operation over the tree obtained in question 2 using a valid statistical test for comparison. 30 marks
4. Print the final decision tree obtained from question 3 following the hierarchical levels of data attributes as nodes of the tree. 10 marks
5. A brief report explaining the procedure and the results 10 marks

Dataset:

1. COVID-19 India statistics:

   It is also a time-series data containing the state-wise details of the COVID-19 statistics in India. The attributes are date, time, state/union territory, confirmed Indian and foreign nationals, cured, deaths and confirmed (Indian+foreign nationals). Target attribute is the number of deaths.
   Filenam: IndiaCOVIDStatistics.csv

Submission instructions:

1. Submit your codes by implementing them only in PYTHON. No other programming language is allowed. Do not use any library functions for building the decision tree.
2. Submit a README file which will contain the instructions on how to execute your code.
3. Submit a report briefly explaining the procedure and the results.

The source code, README file, and the report must be uploaded as a single compressed file (.tar.gz or .zip). The compressed file should be named as: {Group_Number}_ML_A1.zip or {Group_NUMBER}_ML_A1.tar.gz. Do not submit the data file within the compressed file.