Machine Learning
Assignment 1

Name: Girish Chandar G
Roll No: 16110057
Department: Electrical Engineering

Q1

(a) The attribute '*Date*' gets chosen for root node since the information gain for the '*Date*' attribute will be the maximum as each day has a unique date and every data can be exactly classified with this attribute, i.e., there is a injective mapping from '*Date*' attribute of data to the target attribute. Choosing this attribute is not a good choice as it leads to overfitting(high variance).

(b) One method suggested in Tom Mitchell Chapter 3.7.4 is that we assign the missing value of an attribute, the most common value of that attribute in that particular node. Therefore by following this method for selection of root node the missing value in D3 will be assigned either 'Sunny' or 'Rain'(since both have same number of maximum occurrence in Outlook attribute).

    Another method that can be followed is assigning probability of each possible value of the respective attribute to the missing value. This implies the missing value of D3 of attribute Outlook will be assigned: (3/13) for Overcast,(5/13) for Sunny,(5/13) for Rain and this will be further propagated down the tree for classification.

All the remaining questions have been done in a single jupyter notebook and labeled and headings given accordingly

https://gist.github.com/girish1511/614c285df8b9a8195cee01df1c5202ac#file-assignment1-ipynb

References:

[1] *Machine Learning*, Tom Mitchell, McGraw Hill, 1997