MACHINE LEARNING ASSIGNMENT 1

CHENNURI PRATEEK
16110042

Q1(a)
Ans: if data is also added as an attribute/feature to the existing dataset, each value of the date is absolutely unique . Therefore choosing date as the root node wouldn't be giving us the maximum information gain. Moreover, if the depth increases with date as the root node, the model will lead to absolutely overfitting the data.
As discussed in the class, Outlook would be chosen as the root node since among the available feature this particular feature gives us the maximum information gain.

(b)
Handling Missing values in Decision Trees:
Whenever we encounter a missing value of a feature having continuous values, we generally replace it by the mean of that particular feature values. However, this is the case when we have a regressive output

When we have a classifying output, we look at the output (‘y’) of the row having the missing feature value. Then we collect all the data (‘sub_data’) having (y) as the output.

We now take the mean of the values of the feature (which has a value missing) and place the mean at the point where the value is missing.

The code for all the questions can be found using this link:
https://gist.github.com/chennuri91/c1342ee2b54eeb7441c43d7b53edca98#file-ml_assignment1_16110042-ipynb