Practical Data Science
(Decision Tree & Naïve Bayes)

Write R Scripts or use R to perform any mathematical operations while solving the following problems.

Problem 1: Applying CART, C4.5 and NaiveBayes Algorithms

Given the following training data with 4 categorical variables and 1 target variable.

RID	age	income	student	credit_rating	Class: buys_computer
1	<=30	high	no	fair	no
2	<=30	high	no	excellent	no
3	31 . . . 40	high	no	fair	yes
4	>40	medium	no	fair	yes
5	>40	low	yes	fair	yes
6	>40	low	yes	excellent	no
7	31 . . . 40	low	yes	excellent	yes
8	<=30	medium	no	fair	no
9	<=30	low	yes	fair	yes
10	>40	medium	yes	fair	yes
11	<=30	medium	yes	excellent	yes
12	31 . . . 40	medium	no	excellent	yes
13	31 . . . 40	high	yes	fair	yes
14	>40	medium	no	excellent	no

Do the following:

a. Build a decision tree using CART algorithm manually without any pre and post pruning.

b. Predict the class of following test observation using the tree you constructed in part-a:

   age<=30, income=medium, student=yes, credit_rating=fair

c. Build a decision tree using C4.5 algorithm manually without any pre and post pruning.

d. Predict the class of following test observation using the tree you constructed in part-d:

   age<=30, income=medium, student=yes, credit_rating=fair

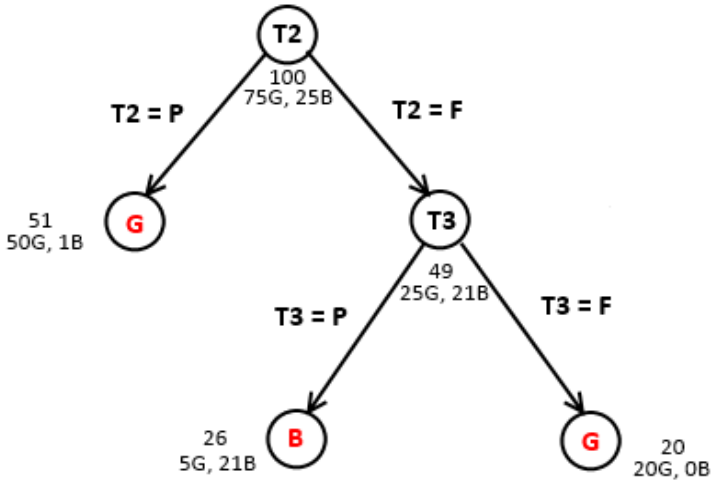
e. Build a probabilistic model using Naïve Bayes algorithm manually without considering any parameters.

f. Predict the class of following test observation using the tree you constructed in part-d:
age<=30, income=medium, student=yes, credit_rating=fair

Problem 2: Applying Cost-complexity Pruning

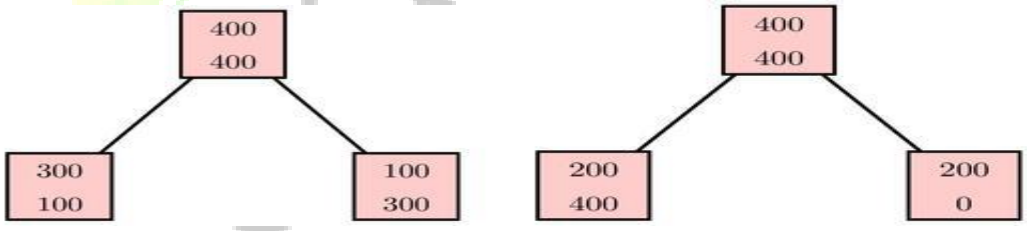
Given the following tree, apply the cost complexity pruning discussed in class for cp values of 0, 1/20, 1/10, 1/8, 1/3, 1. Do the following for each of cp value separately:

a. Compute the pruned and unpruned cost at every internal node.
b. Find out the pruned tree.



Problem 3: Impurity vs Misclassification Rate for tree growth

Find the misclassification rate of following subtrees independently. Which measure do find useful to grow the tree: misclassification rate or impurity?



Practical Data Science
(Decision Tree & Naïve Bayes)

Problem 4: Applying Naïve Bayes Algorithm on Continuous & Categorical Data

Given the training data in the table below (*Tennis* data), predict the class of the following new example using Naïve Bayes classification: outlook=overcast, temperature=60, humidity=62, windy=false. Assume Gaussian distribution for numerical attributes and use Laplace's Correction factor while estimating likelihoods.

outlook	temperature	humidity	windy	play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	86	false	yes
rainy	70	96	false	yes
rainy	68	80	false	yes
rainy	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rainy	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rainy	71	91	true	no