DATA MINING

- Please provide succinct answers to the questions below.
- You should submit an electronic pdf or word file in blackboard.
- Please include the names of all team-members in your write up and in the name of the file.
- You should include all the R functions you use.

Problem 1. Consider the following data set:

record number	income	student	credit-rating	buys-computer
1	high	no	fair	no
2	high	no	excellent	no
3	low	no	excellent	yes
4	medium	no	fair	no
5	low	yes	fair	no
6	low	yes	excellent	yes
7	low	no	excellent	yes
8	medium	yes	fair	yes
9	low	yes	fair	no
10	medium	yes	fair	yes
11	medium	yes	excellent	yes
12	medium	no	excellent	no
13	high	yes	fair	no
14	medium	yes	excellent	yes

(a) Considering "buy-computer" as the target variable, which of the attributes would you select as the root in a decision tree that is constructed using the gain ratio impurity measure?

(b) For the same data set, suppose we decide to construct a decision tree using the Gini index impurity measure. Which attribute would be the best option to use as the root node?

(c) Use the Gini index impurity measure and construct the full decision tree for this data set.

(d) Using your decision tree, provide two decision rules that we can use to predict whether a student is going to buy computer or not. Justify your choice.

Problem 2. Given the dataset in the following table, use the Naïve Bayes classifier to classify the new point (T,F,1.0).

ID	a_1	a_2	a_3	True Class
1	T	T	5.0	Y
2	T	T	7.0	Y
3	T	F	8.0	N
4	F	F	3.0	Y
5	F	T	7.0	N
6	F	T	4.0	N
7	F	F	5.0	N
8	T	F	6.0	Y
9	F	T	1.0	N

Problem 3. This exercise relates to the College data set, which can be found in the file College.csv. It contains a number of variables for 777 different universities and colleges in the US. The variables are

- Private : Public/private indicator
- Apps : Number of applications received
- Accept : Number of applicants accepted
- Enroll : Number of new students enrolled
- Top10perc : New students from top 10% of high school class
- Top25perc : New students from top 25% of high school class
- F.Undergrad : Number of full-time undergraduates
- P.Undergrad : Number of part-time undergraduates
- Outstate : Out-of-state tuition
- Room.Board : Room and board costs
- Books : Estimated book costs
- Personal : Estimated personal spending
- PhD : Percent of faculty with Ph.D.'s
- Terminal : Percent of faculty with terminal degree
- S.F.Ratio : Student/faculty ratio
- perc.alumni : Percent of alumni who donate
- Expend : Instructional expenditure per student
- Grad.Rate : Graduation rate

(a) Read the data into R. Call the loaded data "college".

(b) Look at the data. You should notice that the first column is just the name of each university. We don't really want R to treat this as data. However, it may be handy to have these names for later. Try the following commands:

> rownames (college) → college [,1]

You should see that there is now a row.names column with the name of each university recorded. This means that R has given each row a name corresponding to the appropriate university. R will not try to perform calculations on the row names. However, we still need to eliminate the first column in the data where the names are stored. Write a code to eliminate the first column.

(c) Provide a summary statistics for numerical variables in the data set.

(d) Use the pairs() function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix A using A[,1:10].

(e) Use the plot() function to produce side-by-side boxplots of Outstate versus Private.

(f) Create a new qualitative variable, called Elite, by binning the Top10perc variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%. Follow the code below.

```
> Elite ← rep ("No",nrow(college))
> Elite[college$Top10perc > 50] ← "Yes"
> Elite ← as.factor(Elite)
> college ← data.frame(college,Elite)
```

i. Explain each line of the above code.

ii. Use the summary() function to see how many elite universities there are. Now use the plot() function to produce side-by-side boxplots of Outstate versus Elite.

(g) Use the hist() function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command par(mfrow=c(2,2)) useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to this function will divide the screen in other ways.

Problem 4. This exercise involves the "Auto" data set.

(a) Remove the missing values from this data set.

(b) Which of the predictors are quantitative, and which are qualitative? How do you check this information?

(c) What is the range of each quantitative predictor?

(d) What is the mean and standard deviation of each quantitative predictor?

(e) Remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

(f) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

(g) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.

Problem 5. Download the data set salary-class.csv. This data set (drawn from census data) will be used to predict whether a person has income more or less than $50K (this is 1990 data). The fields are as follows:

AGE: Age of person
EMPLOYER: area of employment (government, private, etc.)
DEGREE: Highest academic degree
MSTATUS: Marital Status
JOBTYPE: type of job (clerical, cleaners, etc.)
SEX: male or female
C-GAIN: capital gain claimed on taxes last year

C-LOSS: capital loss claimed on taxes last year
HOURS: average hours per week worked
COUNTRY: country of origin
INCOME:$\leq$ 50K or $>$ 50K

Use R to answer the questions below.

(a) Import the data set. Use a partition node to divide the data into 60% train, 40% test.

(b) Create the default C&R decision tree. How many leaves are in the tree?

(c) What are the major predictors of INCOME? Justify your choice. How can you get this information from the software?

(d) Give three rules that describe who is likely to have an INCOME $>$ 50K and who is likely to have an income $\leq$ 50K. These rules should be relevant (support at least 5% in the training sample) and strong (either confidence more than 75% "$>$ 50K" or 90% "$\leq$ 50K"). If there are no three rules that meet these criteria, give the three best rules you can.

(e) Create two more C&R trees. The first is just like the default tree except you do not "prune tree to avoid overtting" (you need to let the model to grow to its full depth). The other does prune, but you require 500 records in a parent branch and 100 records in a child branch. You can also play with the complexity parameter. How do the three trees differ (briefly). Which seems most accurate on the training data? Which seems most accurate on the test data?