BUSINESS DATA MINING (IDS 572)

HOMEWORK 2
DUE DATE: WEDNESDAY, SEPTEMBER 14 AT 03:00 PM

- 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.
- One submission is sufficient for the entire group.
- Please include any R codes you use to answer the questions.

Problem 1. Download Pima Indian Diabets data set from blackboard. This dataset is to be used to predict a restul of a diabetic test (class value 1 is interpreted as "tested positive for diabetes").

(a) Load the data and check the attributes of the data. How many variables are in this data set?

(b) Choose the first 80% of the data for training and the remaining 20% data for testing.

(c) Use "rpart" function to create a tree using the training data . What is the accuracy of your model based on training data?

(d) Plot your decision tree. How many leaves are in your tree? Are these leaves pure?

(e) Provide two strongest If-Then rules from this decision tree. Please explain why these rules are chosen.

(f) Apply the decision tree on test data and report your prediction. What is the accuracy of your model in the test data?

(g) Do parts (c), (e), and (f) for a "ctree" function as well. Are there any significant differences between these two dicision trees?

Notice that you will have to clean data first (e.g. remove outlier, handle missing values, ...). Carefully report what you do to clean data.

Problem 2. Please answer this question without using software. Consider the dataset in the table below for classifying whether a type of food is appealing or not.

Appealing	Temperature	Taste	Size
No	Hot	Salty	Small
No	Cold	Sweet	Large
No	Cold	Sweet	Large
Yes	Cold	Sour	Small
Yes	Hot	Sour	Small
No	Hot	Salty	Large
Yes	Hot	Sour	Large
Yes	Cold	Sweet	Small
Yes	Cold	Sweet	Small
No	Hot	Salty	Large

(a) What is the information gain associated with choosing Temperature as root? What about Taste and Size? Briefly write the steps of your calculations.

(b) Draw a full decision tree for the dataset by choosing appropriate features for splitting. Justify each split.

Problem 3. Please briefly justify your answer to the following questions in one or two sentences.

(a) Consider two decision trees that always yield the same class label, given the same test sample. Do both the trees need to have the same number of nodes? You can justify your answer with an example.

(b) Describe the purpose of separating the data into training and testing data.

(c) Which problem do we try to address when using pruning? How do we do that? Please explain.