Database Systems

HW 8: *A Simple Decision Tree Classifier:*

We are monitoring the financial health of 143 currently operating companies. In the prior few years, 107 companies have gone bankrupt, and we are trying to spot companies that might be at risk of going bankrupt in the near future. We are looking to identify currently operating companies that share characteristics with companies that have already gone bankrupt.

Over time, we have developed a scorecard for the financial risk of each company based on six factors in a company's performance. These are stored in the database file you will load.

The scorecard contains the following information[1].
Attribute Information: (P=Positive, A=Average, N=negative, B=Bankruptcy, NB=Non-Bankruptcy)

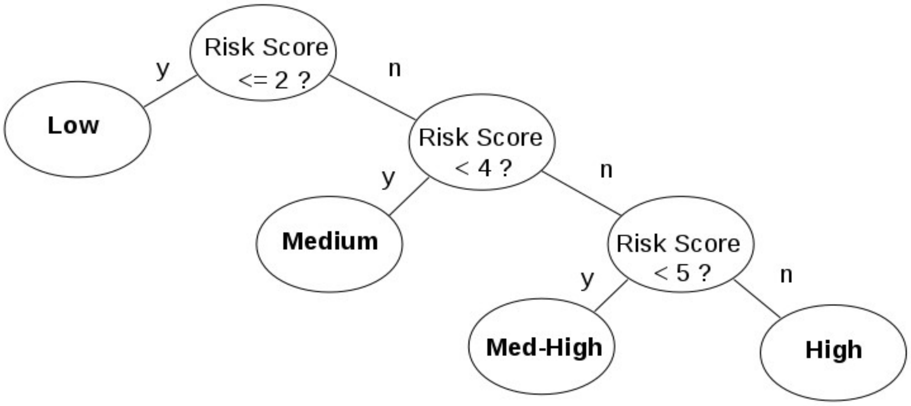
Field:	Possible Values:
1. Company ID	Integer.
2. Industrial Risk:	{P ,A, N}
3. Management Risk:	{P, A, N}
4. Financial Flexibility:	{P, A, N}
5. Credibility:	{P, A, N}
6. Competitiveness:	{P, A, N}
7. Operating Risk:	{P, A, N}
8. Class:	{B, NB}

There will be *four* classifications. We will group each company into a risk-level group, based on how many of the company's risk factors are scored as 'N', or below-average (negative) on that metric.

The groups (risk levels) will be defined by their 'N' score as follows:

<= 2	Low
< 4	Medium
< 5	Medium-High
> 5	High

These classifications correspond to the leaf nodes of a a Decision Tree, which looks like this[2]:



Your Assignment:

Download the .csv data file, create a table for the data, and load the data from the .csv file into your table.
The meaning of the data is explained in the file "info.txt".
Generate a "risk score" for each company, by adding 1 point for each 'N' seen in the company's six rating fields.
Using a Decision Tree approach (as in the diagram above), classify each company into one of the four groups, based on their risk score.

Report the number of companies at each risk level from the *bankrupt* group.

Report the number of companies at each risk level from the *non-bankrupt* group.

Make a report of currently operating companies that are at a risk level of 'Medium' or higher.
We will have to monitor these, to make sure we don't get burned if they go bankrupt.

Hints:

To count up the number of 'N' ratings a company has, you may want to look at each column, and say CASE WHEN column='N" THEN 1 ELSE 0 END.
This will have to be repeated across all 6 columns and added up to generate the score for each company (row).

Once you have a score for each company, you may have to employ *nested* CASE statements to be able to traverse down the decision tree.
For example:

```
CASE WHEN condition1 THEN class_1
     WHEN condition2 THEN class_2
     ELSE class_3
END
```

Submit:

One .sql file with all the commands you used to load your .csv data and execute your queries.

1. Data Source: Martin.A, Uthayakumar.j and Nadarajan, Dr.V.Prasanna Venkatesan February 2014

2. Here we are using only one factor to classify which group each company belongs to. Often we will change from one factor to another at each step in the tree in order to classify.