Knowledge Discovery in Behavioral Data

Use of Decision Trees to Predict levels of Alcohol Problems

Mark Brosnan

for Lutz Hamel, CSC 499: University of Rhode Island

Finding The Data

- I needed to find data to use for my Data Mining work with Dr. Hamel.
- Psychology is my second major, and I have never read a paper that used data mining.
- The data sets collected in psychology research are large and fairly complicated.
- The CPRC is constantly collecting data.
- Dr. Mark Wood was nice enough to allow me access to raw data he collected in 2002 on the URI campus.

The Data Continued...

- In an effort to remain as unbiased as possible I did not read Dr. Wood's paper resulting from his analysis of the data.
- On first examination I saw that the data was typical of that found in Psychology studies, wide but not deep.
- The study had 425 subjects and over 1,200 pieces of data collected for each of them.
- I would have reduce the size of my domain, current data mining algorithms work much better with tables that have few columns (variables) and many cases (examples).

The Data Continued...

- As I examined the data I realized it was part of a longitudinal study and that the data was collected in three “waves.”
  - Wave 1 was collected before the subjects entered college and consisted largely of background information.
  - Wave 2 was collected during the subjects’ freshman year. Data for many measures was collected at this time.
  - Wave 3 was collected during the subjects’ sophomore year and asked the same questions as wave 2. Due to attrition, however, wave 3 contained fewer subjects.

The Data Continued...

- By exploring the data collected in only one wave I would be able to reduce the number of columns.

- I chose to examine wave 2. This wave had all of the data I would need and it had more subjects than wave 3.

- Wave 2 contained 440 columns and only 384 subjects. I would have to further narrow the scope of my exploration.

- I needed a systematic approach. Dr. Hamel suggested I consider using CRISP.

CRISP

- CRISP (CRoss Industry Standard Process) http://www.crisp-dm.org/index.htm
- There are 6 main steps to the CRISP process
  1. Understand the domain
  2. Understand the data
  3. Prepare the data
  4. Build the predictive model
  5. Evaluate the model
  6. Use the model.

CRISP and the Data

- CRISP encourages the user constantly analyze the quality of the results at each step and to loop back to previous steps if it is found that a different tactic would produce better results.
- I had started to look at the data "blind" but this would not work. I needed to better understand my domain.
- It was time to read Dr. Wood's paper.

Understand the Domain

- I found that the metric of the consequences of alcohol use was a measure called the YAAPST (Young Adults Alcohol Problems Screening Test). This test was administered to all subjects in wave 2.
- A subject's score on the YAAPST was the best available predictor of negative alcohol induced experiences. The goal of this research is to find the factors that contribute to students' alcohol problems, and ultimately to develop a program reducing the frequency of those problems.
- Dr. Wood used the YAAPST as his dependent variable, I chose to follow suit.

Understand the Domain

- The data consisted of questions from many measures. The subject's answers to those questions were then used to calculate a resultant score for each measure. By using scores for each measure instead of using every question I would be able to reduce the number of columns to 43.
- Furthermore, Dr. Wood developed a path model that he theorized would explain the variance in students YAAPST scores.
- His model was based on prior research investigating the cause of alcohol problems. Of the 43 possible "sub scores" Dr. Wood selected 10 independent variables to explain students alcohol problems.

Understand the Domain

- The final 10 measures (column label):
  1. Social lubrication outcome expectancy (EQ_SEW2)
  2. Tension reduction outcome expectancy (EQ_TRW2)
  3. Impulsively - sensation seeking (IMPSSW2)
  4. Negative affect (NEGAFFW2)
  5. Alcohol offers (ALCOFFW2)
  6. Perceived peer drinking environment (SOMODW2)
  7. Enhancement drinking motives (DMENHW2)
  8. Coping drinking motives (DMCOPEW2)
  9. Social reinforcement drinking motive (DMSOCW2)
  10. Alcohol use (AQW2_RE)
- This brought my total number of columns to 11, a manageable number.

Prepare the Data

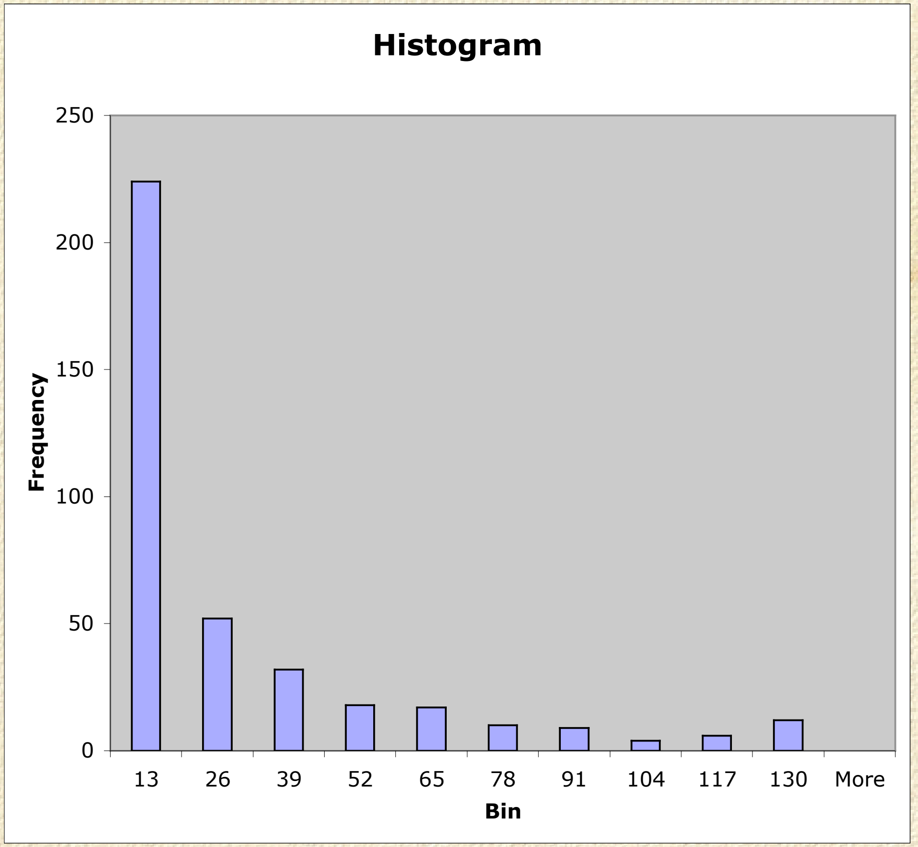
- All 11 variables consisted of continuous data. This does not usually lend itself to decision trees. The data mining tool I chose allowed the use of continuous independent variables but I would have to map the dependent variable into fixed categories.
- The process by which I chose my categories was not short and involved several iterations of the CRISP process.
- The scores on the YAAPST ranged from 0 to 256. I started by simply binning that data into 10 equal parts each with a range of 25.

Prepare the Data

- I found that outliers were effecting my results. Again I turned to Dr. Wood's paper.
- Like Dr. Wood, I adjusted scores for "far outliers" to 1 value greater than the greatest non-far-outlier. This reduced my range to 0 -126. My 10 bins now each had a range of 13. (The YAAPST scores consisted of only whole numbers so I could not use the more accurate 12.6 bin size.)
- Using these bins I could build models that nicely explained the training cases but I was getting poor predictive power with my test cases

Prepare the Data

- I needed to take another look at the data.
- Looking at the histogram and the confusion matrices resulting from my decision trees I knew that I needed more subjects in each bin.

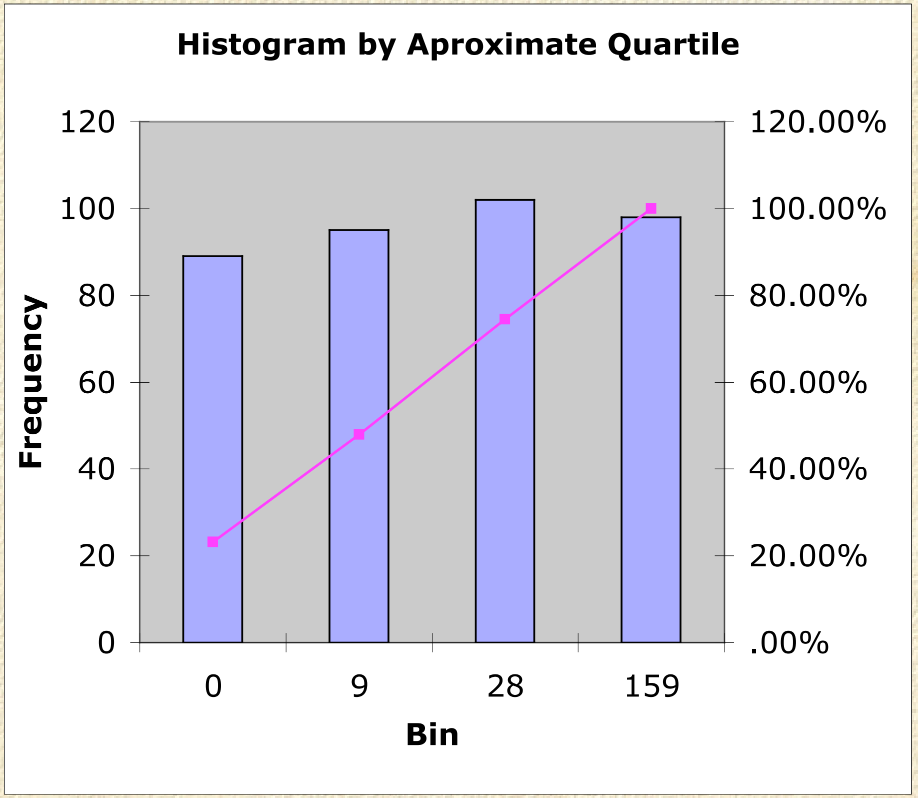


Prepare the Data

- Through several more iterations of the CRISP process I realized that simple equal binning of the data would not work.
- I considered using means and standard deviation to determine my bins but quickly realized that that would not be appropriate for the highly skewed data.
- I chose to use quartiles and bin my data into 4 categories.

Prepare the Data

- By calculating the quartiles I developed a better sense of exactly how skewed the data really was.
- The scores of the first three quartiles combined ranged from 0-28. The fourth quartile scores ranged from 29-126.
- The quartiles are not perfect because while the data is continuous it consists of only whole numbers.
- The first three quartiles account for 74.48% of the subjects.

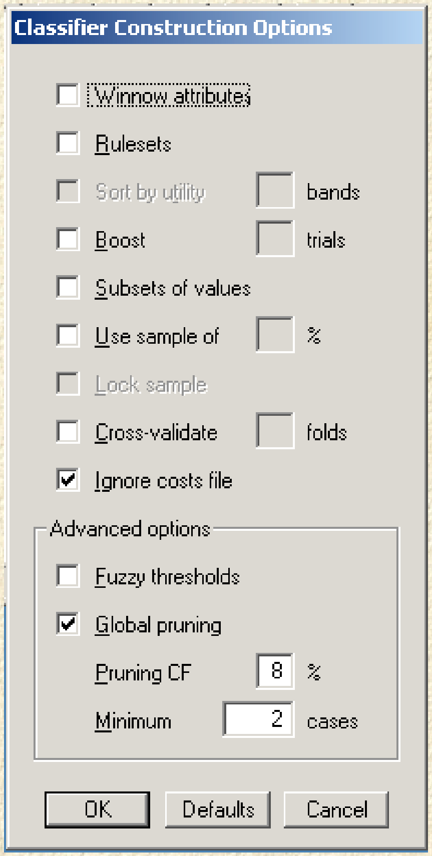


Prepare the Data

- I finally realized that for my experiment using decision trees and this data I should convert the continuous YAAPST scores into binary data with a score of 0 given to all subjects scoring from 0-28 and a score of 1 given to all subjects scoring above 28.
- I would be building models to predict wether or not a subject would score in the 4th quartile for the occurrence of negative alcohol related consequences.

Build the Predictive Model

- This is the window used to set the parameters of any decision trees built using the C5.0 data mining tool.
- For now we will ignore the costs file. This is a file that assigns weighted values to various subject misclassifications.
- We are interested in tuning the model using the global pruning options.
- A higher value in the "Pruning CF" box will allow more complex trees to be developed. The more complex the tree the more likely that the tree has over-fit the data. This reduces the generalizability of the model.
- The number in the "Minimum" box indicates the minimum number of cases that can be contained in any one leaf of the tree.
- These factors combine to reduce tree complexity. The trick is to find the most accurate, simple, and generalizable tree.



Classifier Construction Options

☐ Winnow attributes

☐ Rulesets

☐ Sort by utility ☐ bands

☐ Boost ☐ trials

☐ Subsets of values

☐ Use sample of ☐ %

☐ Lock sample

☐ Cross-validate ☐ folds

☑ Ignore costs file

Advanced options

☐ Fuzzy thresholds

☑ Global pruning

Pruning CF 8 %

Minimum 2 cases

Build the Predictive Model

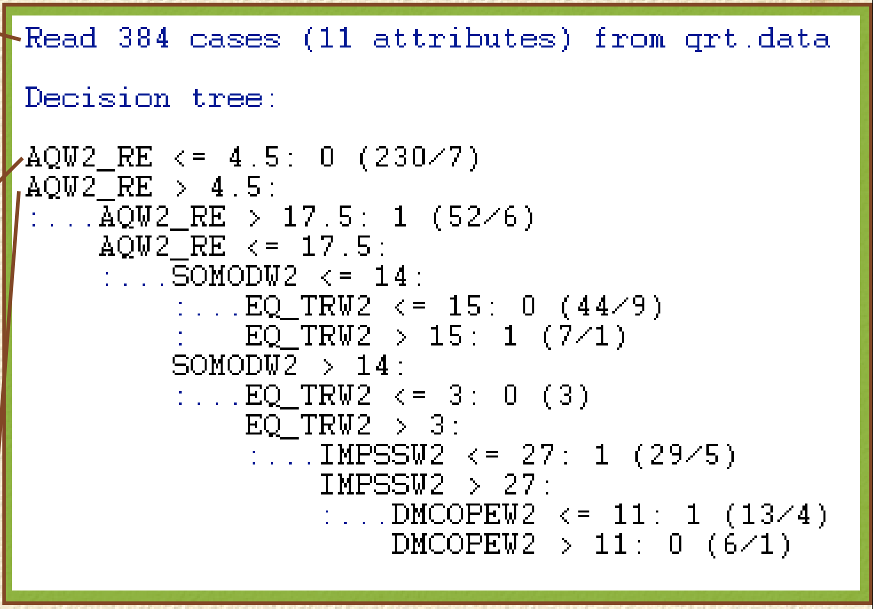
This is the decision tree resulting from the settings displayed on the previous slide.

One of the benefits of using decision tree algorithms is that the results are fairly easy to understand. This tree is no different.

The first line states 384 cases were used to develop this tree, each case had 11 attributes (independent variables), and which text file contains the data.

The first split in the tree is on the AQW2_RE attribute, if the subject's score is <= 4.5 then the model assigns them to class 0. Of the 384 cases examined 230 followed this branch, 7 of them were misclassified.

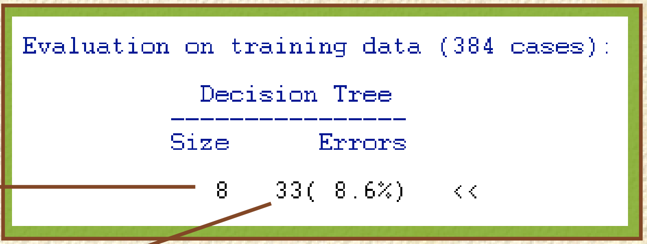
If the AQW2_RE attribute score is > 4.5 than the case is sent down the other branch for further analysis. This continues until all cases have been classified.



```
Read 384 cases (11 attributes) from qrt.data

Decision tree:

AQW2_RE <= 4.5: 0 (230/7)
AQW2_RE > 4.5:
:...AQW2_RE > 17.5: 1 (52/6)
    AQW2_RE <= 17.5:
    :...SOMODW2 <= 14:
        :...EQ_TRW2 <= 15: 0 (44/9)
        :   EQ_TRW2 > 15: 1 (7/1)
        SOMODW2 > 14:
        :...EQ_TRW2 <= 3: 0 (3)
            EQ_TRW2 > 3:
            :...IMPSSW2 <= 27: 1 (29/5)
                IMPSSW2 > 27:
                :...DMCOPEW2 <= 11: 1 (13/4)
                    DMCOPEW2 > 11: 0 (6/1)
```

Evaluate the Model

This is the basic information C5.0 gives us in order to evaluate our decision trees.



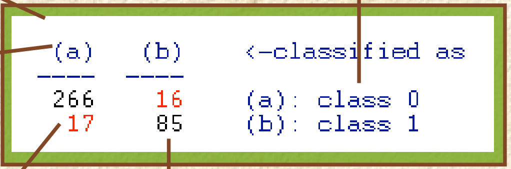
```
Evaluation on training data (384 cases):

          Decision Tree
        ----------------
        Size      Errors

           8   33( 8.6%)   <<
```

The "8" in the "size" column is an indication of the depth (complexity) of the tree.

This model resulted in 33 errors for an error rate of 8.6% or an overall accuracy rate of 91.4%.
This is a very accurate model, it is probably over-fit.

"a" is class 0 indicating a YAAPST score of 28 or lower, and "b" is class 1 or a YAAPST score greater than 28.

This table is called a confusion matrix. It displays exactly where predictions differed from the actual YAAPST bin.



```
  (a)   (b)    <-classified as
 ----  ----
  266    16    (a): class 0
   17    85    (b): class 1
```

The columns represent the classifications generated by the decision tree. In this case the tree classified 283 (266 +17) subjects as "a" and the remaining 101 subjects as class "b."

The numbers "17" and "16" are red, indicating errors.
The red "17" is in column "a" (predicted class 0) and the second row (actually class 1).
The red "16" is in column "b" (predicted class 1) and the first row (actually class 0.)

The black "266" and "85" indicate agreement between the model and the actual YAAPST bin.

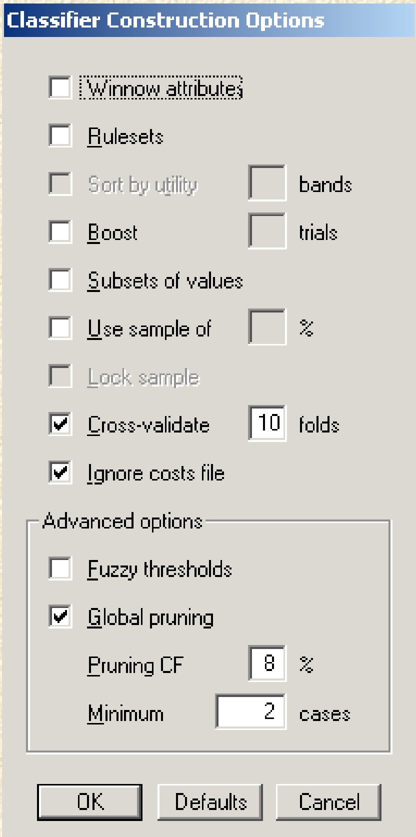
Re-evaluate / Rebuild the Model

- The CRISP process calls for constant re-evaluation of the model and for the miner to "tune" the parameters of the model until the optimal model is found.

- While "tuning" the model over fitting the training data must be avoided. A model that is over fit will very accurately model the training data but it will have poor generalizablity.

Re-evaluate / Rebuild the Model

- Here is a summary of a few of the techniques I used to tune each model.
- First I used K fold cross-validation to prevent over fitting. This option is selected in the window to the right.
- Cross validation divides the data into "k" folds or "test blocks." In this case I have chosen k=10, meaning that each test block will be 10% of the entire data set. Each block is of the same size and has roughly the same class distribution.
- For each test block to be analyzed, a decision tree is created using the remaining 90% of the data. That tree is then used to predict the category of each case in the test block. The % error of the resulting predictions is calculated.
- The same process is used for all 10 blocks. This allows all data in the set to be used for testing trees while maintaining a separation between test data and training data to ensure the tree can not over fit its test data.



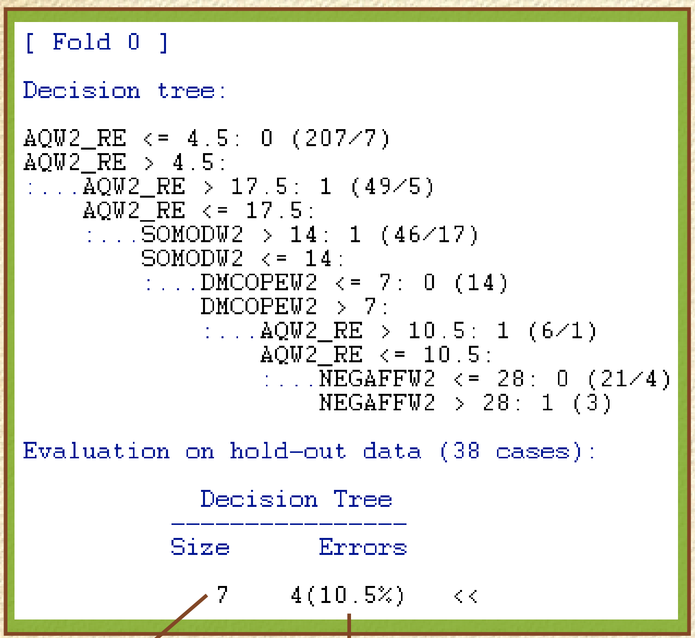
Classifier Construction Options

☐ Winnow attributes
☐ Rulesets
☐ Sort by utility ☐ bands
☐ Boost ☐ trials
☐ Subsets of values
☐ Use sample of ☐ %
☐ Lock sample
☑ Cross-validate 10 folds
☑ Ignore costs file

Advanced options
☐ Fuzzy thresholds
☑ Global pruning
Pruning CF 8 %
Minimum 2 cases

Re-evaluate / Rebuild the Model

This are the first two trees created using cross fold validation. We see each tree and we have much of the information we used to evaluate our previous tree.

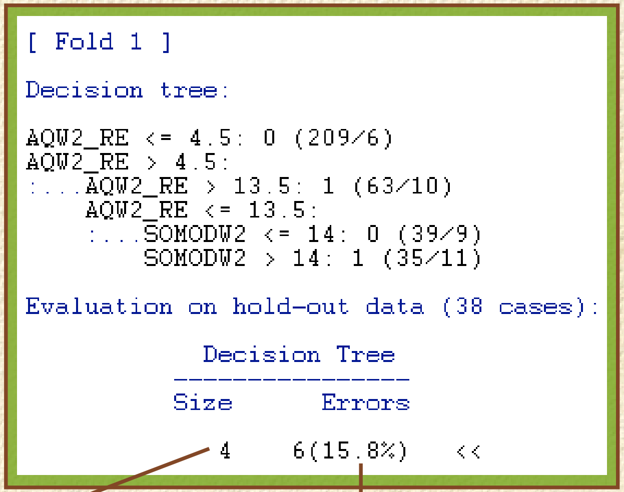


```
[ Fold 0 ]

Decision tree:

AQW2_RE <= 4.5: 0 (207/7)
AQW2_RE > 4.5:
:...AQW2_RE > 17.5: 1 (49/5)
    AQW2_RE <= 17.5:
    :...SOMODW2 > 14: 1 (46/17)
        SOMODW2 <= 14:
        :...DMCOPEW2 <= 7: 0 (14)
            DMCOPEW2 > 7:
            :...AQW2_RE > 10.5: 1 (6/1)
                AQW2_RE <= 10.5:
                :...NEGAFFW2 <= 28: 0 (21/4)
                    NEGAFFW2 > 28: 1 (3)

Evaluation on hold-out data (38 cases):

            Decision Tree
          ----------------
          Size      Errors

             7    4(10.5%)   <<
```

This model resulted in 4 errors for an error rate of 10.5% or an overall accuracy rate of 89.5%.

The "7" in the "size" column is an indication of the depth (complexity) of the tree.

You can see that this tree is quite different from that created in fold 0. This makes sense given that different data was used for training.



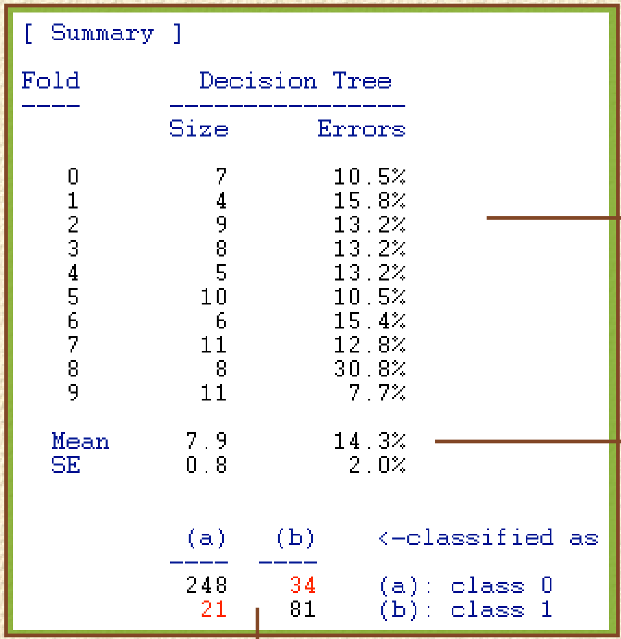
```
[ Fold 1 ]

Decision tree:

AQW2_RE <= 4.5: 0 (209/6)
AQW2_RE > 4.5:
:...AQW2_RE > 13.5: 1 (63/10)
    AQW2_RE <= 13.5:
    :...SOMODW2 <= 14: 0 (39/9)
        SOMODW2 > 14: 1 (35/11)

Evaluation on hold-out data (38 cases):

            Decision Tree
          ----------------
          Size      Errors

             4    6(15.8%)   <<
```

This model resulted in 6 errors for an error rate of 15.8% or an overall accuracy rate of 84.2%.

The "4" in the "size" column is an indication of the depth (complexity) of the tree.

Re-evaluate / Rebuild the Model

After all 10 trees are displayed with their individual evaluations a summary of all 10 trees is presented. This is the primary information used to evaluate the success of this itteration of the tuning process.



```
[ Summary ]

Fold        Decision Tree
----      ----------------
          Size      Errors

  0          7       10.5%
  1          4       15.8%
  2          9       13.2%
  3          8       13.2%
  4          5       13.2%
  5         10       10.5%
  6          6       15.4%
  7         11       12.8%
  8          8       30.8%
  9         11        7.7%

 Mean      7.9       14.3%
 SE        0.8        2.0%


           (a)   (b)    <-classified as
          ----  ----
           248    34    (a): class 0
            21    81    (b): class 1
```

This table shows each fold's tree size and error rate.

Here we see the mean and standard error for both tree size and error rate.

Here again we see a confusion matrix. It is important to understand what type of misclassifications the model is making. This matrix is a nice visual tool to aid that understanding.

Re-evaluate / Rebuild the Model

This is a table showing the summarized results of 11 iterations of model tuning. The 2 settings manipulated were "pruning cf" and "minimum." Both of these directly effect the complexity of the resultant decision tree model. We are trying to find the most simple and accurate tree. The rows that are hi-lighted green have the lowest error rates.

The items in **blue** have been changed from the previous tree settings								Misclassifications	
winnow	boost	cross-Val	costs	pruning cf	minimum	size	error	actually 0	actually 1
no	no	10	ignore	8	2	7.9	14.3%	34	21
no	no	10	ignore	**10**	2	7.4	13.5%	27	25
no	no	10	ignore	**15**	2	8.8	14.6%	31	25
no	no	10	ignore	**20**	2	11.1	14.3%	25	30
no	no	10	ignore	**25**	2	11.1	14.6	33	23
no	no	10	ignore	25	**8**	4.5	15.1%	35	23
no	**no**	**10**	**ignore**	**25**	**6**	**5.0**	**13.8%**	**34**	**19**
no	no	10	ignore	25	**4**	7.3	13.8%	31	22
no	no	10	ignore	**10**	4	6.5	14.0%	26	28
no	no	10	ignore	10	**6**	5.8	14.8%	30	27
no	no	10	ignore	**15**	6	6.3	14.3%	27	28

The settings for the 3 hi-lighted rows are quite different but the error rates are similar. How to decide which settings to choose deds on your domain. For now it makes sense to choose the tree with the smallest mean size. This is the 2nd **bold** green row, with a pruning cf of 25 and minimum support per leaf of 6 cases. One thing to understand is that the selection of folds is random so building multiple trees with the same cross fold settings can result in different results.

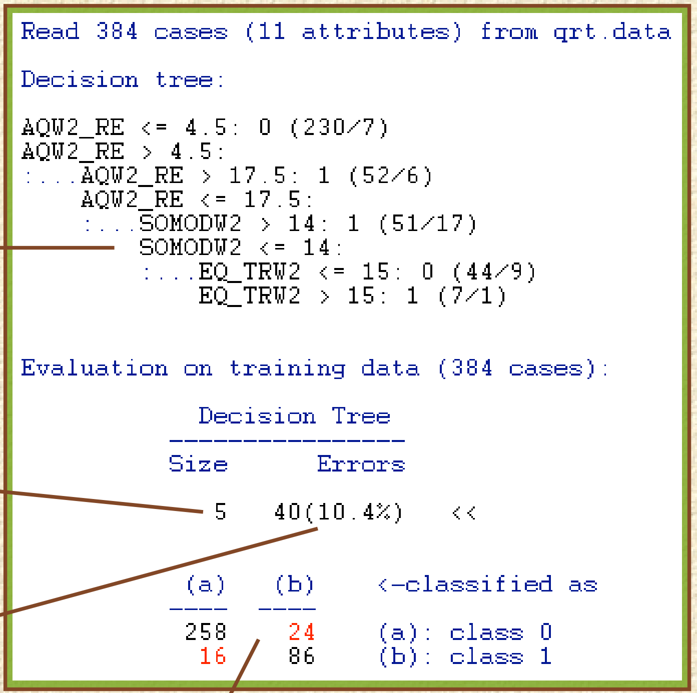
Re-evaluate / Rebuild the Model

Now we use the settings found to be best using the k-fold cross validation technique. We train the tree on the entire data set. Given that the settings were developed using cross validation the resultant tree should not be over fit to the data. This is the resultant tree.

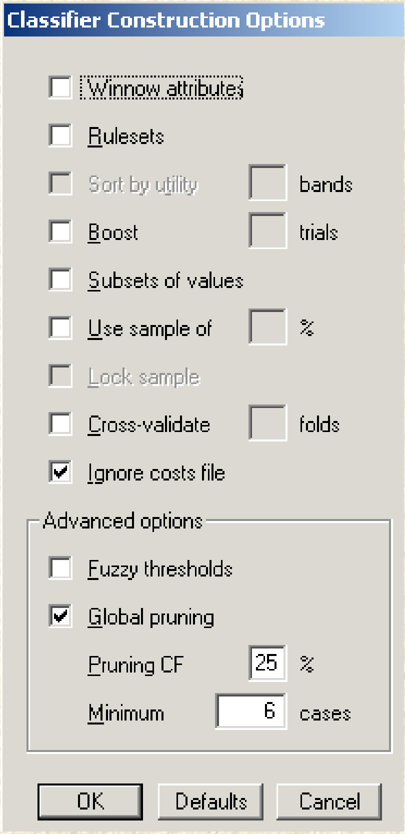
We have a tree that is easy to understand. It is also interesting to note that of the 11 attributes available the tree only uses 3.
AQW2_RE
SOMODW2
EQ_TRW2

size = 5 (a simple tree)

error rate = 10.4% for an overall accuracy of 89.6%.



```
Read 384 cases (11 attributes) from qrt.data

Decision tree:

AQW2_RE <= 4.5: 0 (230/7)
AQW2_RE > 4.5:
:...AQW2_RE > 17.5: 1 (52/6)
    AQW2_RE <= 17.5:
    :...SOMODW2 > 14: 1 (51/17)
        SOMODW2 <= 14:
        :...EQ_TRW2 <= 15: 0 (44/9)
            EQ_TRW2 > 15: 1 (7/1)


Evaluation on training data (384 cases):

            Decision Tree
          ----------------
          Size      Errors

             5   40(10.4%)   <<


           (a)   (b)    <-classified as
          ----  ----
           258    24    (a): class 0
            16    86    (b): class 1
```



Classifier Construction Options

☐ Winnow attributes
☐ Rulesets
☐ Sort by utility ☐ bands
☐ Boost ☐ trials
☐ Subsets of values
☐ Use sample of ☐ %
☐ Lock sample
☐ Cross-validate ☐ folds
☑ Ignore costs file

Advanced options
☐ Fuzzy thresholds
☑ Global pruning
Pruning CF 25 %
Minimum 6 cases

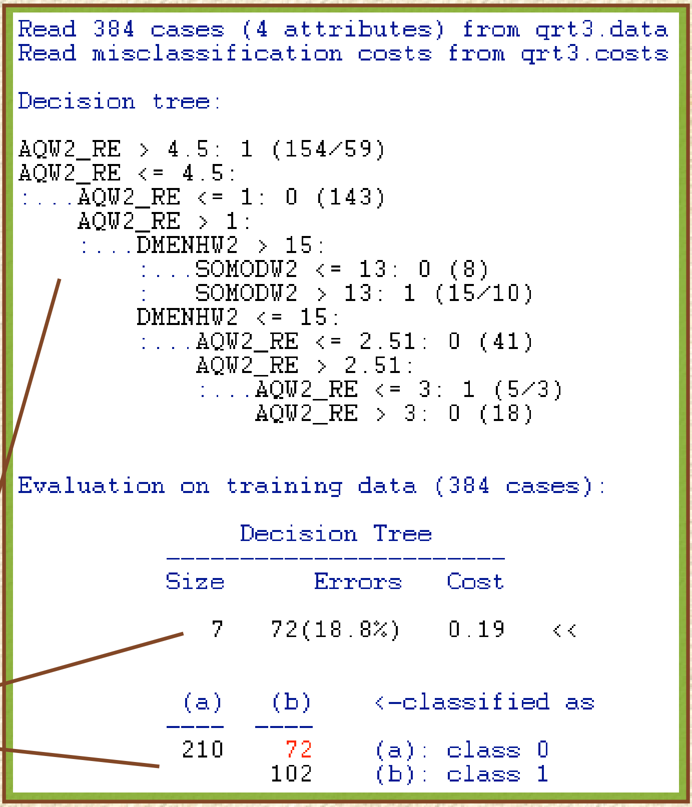
The confusion matrix shows that we are misclassifying class 0 subjects as class 1 slightly more than the opposite. The miner must determine if this is ok. For now this is fine.

Re-evaluate / Rebuild the Model

Given that the previous tree only used 3 attributes, I wanted to understand the predictive power of every variable. I chose to isolate each of the 10 independent variables by building trees using only 1 variable at a time. This proved interesting. Only 3 variables were able to support trees of ANY kind. It is also interesting to note that 1 of those variables was not the used in the previous tree, DMENHW2. EQ_TRW2 had no predictive power on it's own. I chose to train a new tree using only the 3 attributes that had power on their own.

After using cross validation and several other techniques that are beyond the scope of this presentation the following tree resulted. This tree has one other difference that has not yet been shown. The costs file was NOT ignored. Basically the cost file told the tree that misclassifying a person who is actually class 1 is 4 times more costly than misclassifying a class 0 person as class 1. In this case it is more important to catch everyone who has problems than it is to accidently tag those who are fine. The factor of 4 is somewhat arbitrary. That factor should be set by a person with much greater domain knowledge than I currently have.

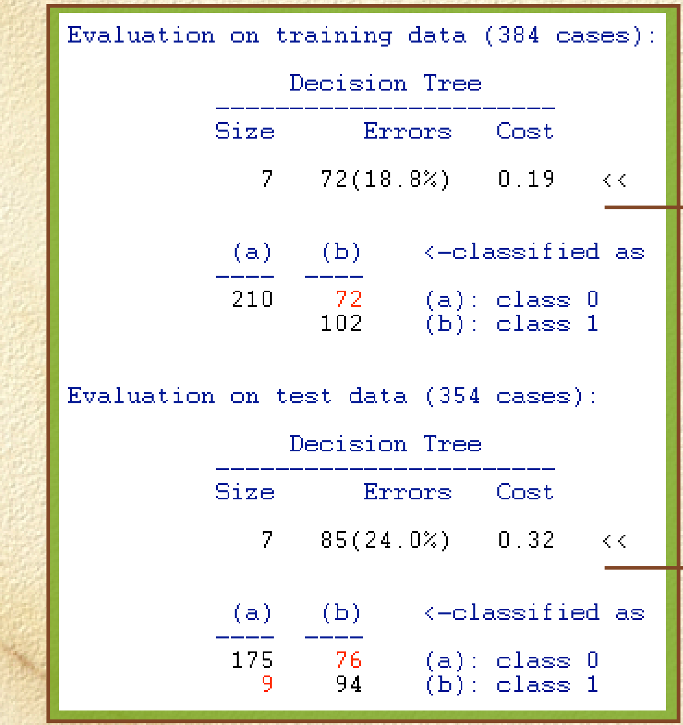
The size is small, the error is 18.8%, but note that no class 1 subjects were missed. Notice that AQW2_RE was by far the most used attribute.



```
Read 384 cases (4 attributes) from qrt3.data
Read misclassification costs from qrt3.costs

Decision tree:

AQW2_RE > 4.5: 1 (154/59)
AQW2_RE <= 4.5:
:...AQW2_RE <= 1: 0 (143)
    AQW2_RE > 1:
    :...DMENHW2 > 15:
        :...SOMODW2 <= 13: 0 (8)
        :   SOMODW2 > 13: 1 (15/10)
        DMENHW2 <= 15:
        :...AQW2_RE <= 2.51: 0 (41)
            AQW2_RE > 2.51:
            :...AQW2_RE <= 3: 1 (5/3)
                AQW2_RE > 3: 0 (18)


Evaluation on training data (384 cases):

               Decision Tree
          -----------------------
          Size      Errors   Cost

             7   72(18.8%)   0.19   <<


           (a)   (b)    <-classified as
          ----  ----
           210    72    (a): class 0
                 102    (b): class 1
```

Re-evaluate / Rebuild the Model

The next step is to evaluate the model's performance on more data. Here I have used the model built using the 3 most powerful individual attributes found in wave 2 to predict the degree of alcohol problems found in wave 3.



```
Evaluation on training data (384 cases):

          Decision Tree
        -----------------------
        Size      Errors   Cost

           7   72(18.8%)   0.19   <<


         (a)   (b)    <-classified as
        ----  ----
         210    72    (a): class 0
               102    (b): class 1

Evaluation on test data (354 cases):

          Decision Tree
        -----------------------
        Size      Errors   Cost

           7   85(24.0%)   0.32   <<


         (a)   (b)    <-classified as
        ----  ----
         175    76    (a): class 0
           9    94    (b): class 1
```

The evaluation on the training data is the same as before

The test data is from wave 3 (collected from the same subjects 1 year later.) Wave 3 is not totally independent of wave 2 but the longitudinal prediction is of interest.

The error rate has increased to 24% but the rate of missing class 1 subjects is only 9.5%.

Additional evaluation of the Model

- Statistical measures of rater agreement are not currently used by data miners. Since rater agreement is, at its core, designed to compare classification techniques I believe that it should lend itself nicely to data mining.
- I explored two tests that measured the significance of the tendency of a model to over-rate or under-rate cases.
  - The McNemar Test of Marginal Homogeneity can be used on categories with more than one level. It measures the prediction tendency of each level of categorization, and returns a chi squared statistic indicating the presence of a bias for that level.
  - The Stuart-Maxwell Test provides similar information but rather for the model as a whole. It does not break down the bias by category level. For a binary case like mine both of these tests will return the same result.
  - There are many other measures that exist in the rater agreement bailiwick whose usefulness should be explored.

Conclusion

- The paper by Dr. Wood hypothesized 2 models each one a mechanism explaining the variance found in students' incidence of alcohol related problems. Both of his models fit the data reasonably well (model #2 fit significantly better) and they successfully explained approximately 70% of the variance in alcohol problems.
- Since I used the same variables and the same data it is interesting to compare our results.
- It is important to note that his work and mine have two quite different goals. He is attempting to explain variance and I am making predictions. Furthermore, I used bined data, he did not.

Conclusion continued...

- As I understand the analysis performed by Dr. Wood, there is no easy way to know anything about the 30% of cases that are not explained by his model. Given that 75% of the cases scored under 29 on the YAAPST it is possible that the model does not work well for those students who are in the 4th quartile. The opposite could also be true, those students with no problems might not be accounted for (1st quartile students all scored 0 on the YAAPST).
- The confusion matrix and measures of rater agreement allow the data miner to see exactly where the model is failing. This allows the miner to determine the significance of the errors and to tune the model accordingly.

Conclusion continued...

- Another significant difference between the two approaches is that path analysis requires a very specific hypothesis building causal links between independent variables. Some of the variables are then linked to the dependent variable. This technique yields insight into the motivations and inter-related nature of the problem, but it takes a great deal of domain knowledge to build.
- Data mining assumes nothing. It simply looks at everything and applies whichever variables are useful for prediction. In this case my models indicate that there is direct link between alcohol related problems and both the student's negative affect and their perceived drinking environment. These links existed, but were not direct, in Dr. Wood's hypothesis.

Conclusion continued...

- This work suggests that data mining can be a valuable tool to explore the many factors that contribute to the actions of an individual. Once significant variables are found, additional statical modeling techniques can be brought to bear. By combining many techniques a more complete understanding of social psychological domains can be achieved.
- Additional work is also currently being undertaken to develop models that will aid in the treatment of clients with addiction problems. These treatment programs use decision trees to provide feedback to patients while they work to overcome their addiction. Every possible avenue of support is helpful.
- Data mining has many potential uses in the social sciences and, as algorithms improve, it will become essential.

Works of interest

CRISP-DM Consortium and Special Interest Group. (n.d.). *Cross Industry Standard Process for Data Mining.* Retrieved March 15, 2003, from http://www.crisp-dm.org/index.htm

Hurlbut, S.C., & Sher, K.J. (1992). Assessing alcohol problems in college students. *College Health, 41*, 49-58.

Mitchell, T.M. (1997). Decision tree learning. In C.I. Liu, & A.B. Tucker (Ed.), *Machine Learning*. Boston: WCB/McGraw-Hill.

Read, J. P., Wood, M. D., Kahler, C.W., Maddock, J.E., & Palfai, T.P. (2003). Examining the role of drinking motives in college student alcohol use and problems. *Psycology of Addictive Behavior*s, 17, 13-23.

Uebersax, J. (2003). *Statistical methods for rater agreement.* Retrieved May 10, 2003, from http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm

Webley, P., & Lea, S. (2001). *Multivariate analysis II: Manifest variables analyses: Path analysis* (Topic 3). Retrieved April 2, 2003, from http://www.maths.ex.ac.uk/-wjk/psy6010/pathanal.html