Knowledge Discovery and Data Mining of Biomedical Data with Decision Trees

Marilyn K. McClelland

mmcclell@nccu.edu

School of Business, North Carolina Central University
Durham, NC 27701, USA

Abstract

Traditional knowledge discovery processes are applied in a biomedical population study to develop a data warehouse of diverse clinical, phenotypic, psychosocial, and genetic data associated with hypertension. Experiences of an informationist as a member of a biomedical research team are shared. Issues encountered with missing data and data analysis is discussed. The use of decision trees which can accommodate missing data is explored. SAS Enterprise Miner 4.2 is used to develop classification and predictive decision trees for hypertension in African Americans. Preliminary knowledge discovery through the use of decision trees suggests psychosocial components such as anger as well as traditional metabolic syndrome components such as waist size are important factors in hypertension of healthy, community dwelling, African Americans. Advantages and limitations of decision trees are discussed. More broadly, the biomedical research team continues to benefit from the knowledge management infrastructure implemented as part of the knowledge discovery and databases (KDD) process described here.

Keywords: KDD, data mining, biomedical informatics, decision tree, knowledge management

1. INTRODUCTION

A few years ago a university colleague asked if I could help with the data from a recently funded population study. The study is the first of its kind at our university and no one had planned for the management of the large volume of data being generated. Here I describe the application of traditional knowledge discovery approaches to understand the context of the data, prepare and preprocess the data for analysis, as well as preliminary analysis and knowledge discovery with this rich set of biomedical data.

The study is motivated by the disproportionate occurrence of hypertension among African-Americans, a well known, though little understood health disparity. Predisposition to cardiovascular disease is recognized as multifaceted where genetics, environmental influences, and psychosocial factors interact. A gene-environment cardiovascular study as part of P20 MD 00175, Excellence in Partnerships for Community Outreach and Research on Health Disparities and Training (EXPORT) is underway at North Carolina Central University (NCCU). Thus far, the study has enrolled 176 subjects of African descent and collected over 1000 data points per subject utilizing 15 intake devices. Data collected includes psychosocial, clinical, phenotypic, and genetic data. This study tests the hypothesis that the increased incidence of cardiovascular disease in the African American population is the result of gene-environment interactions, in particular, factors such as psychosocial stress of institutionalized racism differentially impact African Americans with a susceptible cardiovascular genotype and result in elevated cardiovascular reactivity to stress.

Knowledge discovery and database (KDD) processes are utilized for data management. Here I report experiences with the use of knowledge discovery and data mining techniques for pattern identification in the data. I share insights from utilizing decision trees for knowledge discovery with a diverse biomedical database with psychosocial, clinical, phenotypic, and genetic data of African-American subjects.
Preliminary analysis reported here examines resting systolic blood pressure (SBP). Medical guidelines recognize SBP over 140 as hypertensive. SBP of 120 to 139 is considered pre-hypertensive. SBP below 120 is considered normal systolic blood pressure.

2. OUR STUDY

Study Participants
Subjects studied include 176 community dwelling participants (63 males, 113 females) self reported African Americans.

Eligibility criteria for entry:
- age of 18 to 65 years
- must attend or work at NCCU or live in the surrounding Durham, Wake, Orange county regions
- must be without any cardiovascular disease (self report)
- not on any hypertensive medication (self report)

Informed consent was obtained for each participant in accordance to NCCU Institutional Review Board and federal guidelines.

The study protocol is illustrated in Figure 1. Broad categories for the data collected include psychosocial instruments, clinical, reactivity, demographic, and genetic.

Psychosocial Measures
- Anxiety: State-Trait Anxiety Inventory (STAI) (Spielberger, 1983)
- Perceived racism: Perceived Racism Scale (PRS) (McNeilly, Anderson, Armstead, Clark, Corbett, Robinson, Pieper, & Lepisto, 1996)
- Anger expression: Spielberger Anger Expression (SAE) (Spielberger, Johnson, Russell, Crane, Jacobs, & Worden, 1985)
- Depression: Beck Depression Index (BDI) (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961)
- Hostility: Cook Medley Hostility scale (CMH) (Cook & Medley, 1954)
- John Henryism: John Henry Active Coping (JHAC) (James, Hartnett, & Kalsbeek, 1983)

Clinical Measures:
- Resting blood pressure (systolic and diastolic)
- Resting mean arterial pressure
- Resting heart rate
- Insulin
- Glucose
- Homeostasis model assessment (HOMA)
- Urinalysis
- Cholesterol (HDL/LDL)
- Triglycerides
- Height
- Weight
- Body Mass Index
- Skin fold measures
- Body circumferences (i.e. waist, thigh, arm)

Cardiovascular Reactivity Measures:
- Psychological stressor (Anger recall)
- Physiological stressor (hand immersion in ice water)

Cardiovascular reactivity has been investigated as an indicator for a predisposition to hypertension and the development of cardiovascular disease. Both psychological and physiological stressors have been used to assess vascular reactivity and both stressors are believed to produce the same qualitative vascular reactivity response.

The cardiovascular measures of systolic and diastolic blood pressures as well as heart rate were assessed during each of the two types of stressors. Following baseline blood pressure and heart rate measurements (GE Dinamap model Pro 100) participants were administered the cold pressor test consisting of submersion of the hand in ice cold water for three minutes followed by a five-minute recovery period. Blood pressure was measured at 1-minute intervals throughout the procedure including the recovery period. A psychological stressor was given after base-
line blood pressures were returned to baseline. Two mental stressors were used in the study. The Stroup test or color naming was used for the first 25 participants. However, we observed no significant change in blood pressure in the first 25 participants and consequently, used anger recall for the subsequent recruited participants. Anger recall stress consisted of 5 minutes of contemplating an event that evoked anger, 5 minutes talking about the event, and 5 minutes recovery period. Blood pressure was measured at 1-minute intervals throughout the procedure. Individuals were classified as low and high responders using the median split approach.

**Demographic Measures**
- Birth date
- Sex
- Race
- Job title
- Student status
- Income
- Years of education
- Mother’s education
- Mother’s job title
- Father’s years of education
- Father’s job title
- Family medical history
- Physical activity level

**Genetic Data:**
To date we’ve genotyped three SNPs (single nucleotide polymorphisms). The researchers plan additional high throughput screening of multiple snps on genes associated with pathways related to cardiovascular reactivity and hypertension.

**3. KNOWLEDGE DISCOVERY PROCESS**
The knowledge discovery process illustrated in figure 2 is a simplification of the knowledge discovery and databases (KDD) process described by Fayyad, Piatetsky-Shapiro, and Smyth (Frayyad, Piatetksy-Shapiro, & Smyth, 1996). The researcher had no prior experience with biomedical research or KDD for biomedical research, but was aware that others have recognized the potential benefits of scientific discovery from the application of KDD in the biomedical domain (Jiawei, Russ, Vipin, Heikki, & Daryl, 2002).

**Step 1: Understand the Context**
The start of the KDD process required that I learn not only this particular project but also biomedical research generally. I started the process by reading the funded research proposal. Next steps included interviews with the research scientists, study coordinator and research technician as well as course work in genetics, bioinformatics, and biomedical data laws, regulations and guidelines. I mentored graduate students to employ use case methodology as one part of the documentation of the user’s relationship with the data. The knowledge gained through interviews with the scientists was also documented in data dictionaries and code books for all data elements. The data dictionaries included established naming conventions for the attributes, and documented domains, and valid ranges. For example, the questions on the John Henry psychosocial instrument are scored with either a 1 or 0 whereas the Cook-Medley instrument questions are scored a 1, 2, 3 or 4. Another example is documenting whether height is recorded with a metric scale or in feet and inches. It is necessary to have height and weight in the same scale before calculating Body Mass Index (BMI) and to assure the use of the appropriate BMI formula.

An overarching database design was developed that allowed each data dictionary to represent a table in the database. Thus, the development of the data dictionaries also considered that ultimately, the data would be stored in a database that supports analysis across the data sets.

Much of the data was not in an electronic format and existed as paper forms completed by the subject and stored in file folders. In some cases, some of the data had already been entered into a spreadsheet but often numeric and text data was entered in the same cell. The interviews with the scientist probed for how the data would ultimately be analyzed or interpreted in order to gain insight for developing a data structure that would have the flexibility to support the desired analysis. For example, a spreadsheet of the blood analysis included occasional comments of hemolyze in the
same cell with a number. We assembled the research team to discuss likely analysis of the blood data and the impact of the hemolyze. The author learned the hemolyze observed and recorded by the technician indicated degradation in the blood sample. Thus we added two fields to the blood analysis data dictionary and data table; one to indicate the specimen integrity and a comment field for the technician to record the observation of the blood hemolyze. The cell was edited to only have the numeric entry for the blood work. The example described here also illustrates capturing the knowledge of the biomedical technician relating to interpretations of one piece of data that will be useful in knowledge extraction from the aggregation of the individual data sets.

**Step 2: Preprocess the Data**

**Step 2. Preprocess** the data includes tasks to detect data entry errors as well as tasks to prepare the data for use in modeling and analysis. Two primary objectives exist for preprocessing the data: one is to assure the integrity of the data and avoid what we used to call ‘garbage in – garbage out’ through data cleaning. The other is reflected in another well known saying – ‘he could not see the forest for the trees’. In other words, strive to balance having sufficiently rich data to reflect all the relevant dimensions yet not flooding the analysis with so many attributes that the data noise prevents the recognition of patterns.

Data were analyzed with SAS 9.1.3 for Windows (SAS Institute, 2006). Preliminary data analysis cleaned, verified, and validated data before proceeding with further analysis. Univariate statistics for psychosocial measures, blood, BMI, BP, reactivity, and SNPs were examined to identify extreme values. Extreme values were checked for data entry errors and compared against the source data forms. Domain and range values were established for each data source. SAS was used to identify any data values outside the expected range and to generate an exception report. Here again, data identified as out of the expected range or domain was manually examined and checked against the data source for any needed corrections. Graphs of the data were examined for outliers and unusual patterns that warrant more scrutiny. SAS calculated Cronbach’s alpha for each psychosocial scoring protocol (all instruments are .6 or higher).

In addition to data cleaning, data preparation also includes 1) scrutinizing the data for relevance 2) checking for data redundancy, 3) data transformation, and 4) feature reduction. (Han & Kamber, 2006). The output of Step 2 is the selection of an appropriate subset of data as an input to Step 3. **Modeling and analysis.**

Although the diagram of the knowledge discovery process in figure 2 shows a linear flow; there may be cycles back through the steps. For example, correlations between variables identified in Step 3 could result in a return to Step 2 for removal of one of the variables from the relevant variable list for modeling as illustrated in Figure 3. In our correlation analysis, we found baseline systolic blood pressure to be highly correlated with baseline diastolic blood pressure. In selecting variables for analysis we use either baseline systolic blood pressure or baseline diastolic blood pressure but not both.

A data mart for further analysis is built from the database. Data transformations / reductions included scoring the psychosocial instruments as validated in prior studies, averaging multiple measurements, and calculations like Body Mass Index and HOMA. For example, each individual has between 15 to 30 measures of resting blood pressure and heart rate in the data base. The data mart contains the average resting blood pressure and standard deviation rather than the multiple measures.

Generalizing the attribute measurement to a higher level concept can lead to data reduction. Decision trees provide useful insights into possible breakpoints for transforming an attribute from interval data to a categorical attribute. Here each subject has 10 to 30 baseline systolic blood pressure measurements. One way to generalize the baseline blood pressure is to average all the baseline readings then categorize the subject as having normal blood pressure ( below 120 SBP), pre-hypertensive ( 120-139 SBP), or hypertensive ( above 140 SBP).

**Step 3. Modeling and analysis**

PROC MEANS was used to calculate the average and standard error for the baseline clinical measures and the reactivity during
Training data / validation data: Here, 90% of the data was randomly selected to use as a training data set for the decision tree. The remaining 10% of the data was allocated for validation of the decision tree after the tree was constructed with the training data set. SAS Enterprise Miner allows selection of a new seed value for the random value generator.

Classification decisions trees have an interval variable as the target. Predictive decision trees have a categorical variable as the target or root (Han & Kamber, 2006). Murthy provides a good description of decision tree induction and assessment (Murthy, 1998).

Classification Decision Tree for Average Resting Systolic Blood Pressure: Average Systolic Blood Pressure was declared as the target variable. Below is the list of the decision tree settings. Table 2 lists the interval variables along with univariate statistics for the variables. Table 3 provides assessment of the decision tree model. Figure 4 is a view of the decision tree that has been pruned to remove branches where the validation data indicates an over fitting of the model with the training data.

Interpretations of Classification Decision Tree: Figure 4 is a view of a decision tree with resting SBP as the root of the tree. In side each node, the left hand number is from the training data and the right side number is the average from the validation data.

For a subject, we can classify the subject based on their characteristics and determine a likely resting systolic blood pressure. The first attribute for classification is the subject's age. The left branch of the tree is for subjects less than 42.5 years. The training data for subjects less than 42.5 years has 118 subjects with an average resting SBP of 117.2. The right branch of the tree is subjects greater than or equal to 42.5 years with a much higher SBP of 131.5.

On the left hand branch for younger subjects, the next attribute to branch is the average waist circumference where a larger waist has a higher resting SBP. The larger waist branches on sex with men having higher resting SBP than women.

Predictive Decision Tree for Resting Blood Pressure Diagnosis as Target: A new variable was added to the data set with the value ‘N’ for normal when average baseline systolic blood pressure is below 120; value is ‘P’ for pre-hypertensive when average systolic blood pressure ranges from 120 to 139 and ‘H’ for hypertensive when SBP is above 140.

After some experimentation, I returned to the use of the default parameter settings in SAS Enterprise Miner. Potts describes and
explains SAS decision tree parameter settings (Potts & Rothman, 2006). Default SAS Enterprise Decision tree parameters settings are shown in Table 2.

**Interpretations of Predictive Decision Tree:** A first node or trunk for the predictive tree for hypertension is Figure 5. The results of the training data set are on the left and the validation data set is on the right for each variable. The side by side display of result of the training and validation data facilitates review for fit and over-fit of the model to the data. For the training set, 49.7% have normal SBP (N), 41.5% are pre-hypertensive (P) and 8.8% are hypertensive (H). In the validation set, 72.2% are normal, 22.2% are pre-hypertensive, and 5.6% are hypertensive.

The next branch after a large waist is the triglyceride level. Here the number of hypertensive subjects is low so I only interpret the results for normal and pre-hypertensive. A larger sample size would be useful. High Triglycerides combined with a large waist, further increase the risk of pre-hypertension – whereas low triglycerides reduce the risk. High anger-in combined with high triglycerides and a large waist further elevates the risk of pre-hypertension to 86.7%.

**Concerns with Decision Trees:** The decision tree structure suggesting relationships between the variables changes with small tweaks in the data. For example, changing the seed used to split the data sets into training and validation produce different trees. One technique to minimize this effect is a cross validation 10 fold which splits the data into 10 sets and runs multiple regression trees to develop a composite set of relationships (Braga-Neto, Hashimoto, Dougherty, Nguyen, & Carroll, 2004). However, the sensitivity of the decision tree structure to the input parameters is a concern. Austin reports on using the same dataset yet the resulting decision trees are different in R and S-Plus (Austin, 2008).

Another concern is a lack of symmetry in the tree relationships. For example, anger control is a factor in the predictive hypertension tree with high triglycerides and a large waist. It would be difficult to decide how to incorporate the anger control parameter in a multiple regression model since the relationship is contingent on the values of other parameters.

**Advantages of Decision Trees:** The research scientists readily grasped the parameter relationships depicted by the decision trees. The decision trees provided an intuitive illustration of the parameters and were warmly received by the scientists especially in comparison to a technique like neural...
networks though neural networks often have better predictive ability.

**Step 4. Knowledge Discovery**

The application of the knowledge discovery process to the data collected in the cardiovascular study reported here has contributed to the understanding of the relationships between psychosocial factors and blood pressure in African Americans with analysis based on correlation, MANOVA, and regression models (Pointer, Livingston, Yancey, McClelland, & Bukoski, 2008). Although the decision tree analysis has been useful in understanding some of the patterns in the data, the concerns identified above have cautioned the author to proceed carefully in reporting the decision tree outcomes as the primary modeling tool for identifying patterns in the data. The study has received an additional five years of funding to increase the number of participants and to add a longitudinal component that will track blood pressure and other measures over time. The larger sample size and longitudinal data will allow the researchers to further investigate the relationships between psychosocial, clinical, genetic and environmental factors suggested by the preliminary data.

Several of the conferences for dissemination of hypertension and cardiovascular disease research are attended by a large percentage of physicians seeking insights into their clinical practice. As mentioned previously, the medical doctors and biomedical researchers on the research team found the data pattern represented by the decision trees appealing and intuitive. Decision trees or classification and regression trees (CART) have been useful in biomedical research to support clinical practice. As mentioned previously, the concerns described above led me to realize that decision trees would be THE analysis tool for classification or input variable selection in a two stage model. Additional research will investigate two stage models in addition to other classification or input variable selection methods.

**4. CONCLUSIONS**

This was my first experience working with diverse biomedical data. In addition, I am part of a research team that did not have prior experience with large data sets. I am often in the role of educator on the research team explaining steps required to clean, validate, and analyze the data as well as the importance of documentation and backup; all elements of the KDD process. Over the years, team members have changed, and having a data and document repository is an important asset for new as well as old team members. The knowledge management infrastructure developed to store the data and all the associated documents continues to be an important component for research for the NCCU EXPORT Cardiovascular team as the research is ongoing and additional data is added to the repository frequently.

Here, I described preliminary experiences with using SAS miner 4.3 and the pros and cons of decision trees to identify patterns with diverse data. Initially, I thought decision trees would be THE analysis tool for classification and prediction. However, with further study and investigation, the concerns described above led me to realize that decision trees are better suited for feature selection or input variable selection in a two stage model. Additional research will investigate two stage models in addition to other classification algorithms.

**5. REFERENCES**


Han, Jiawei & Micheline Kamber. 2006. *Data Mining Concepts and Techniques*. second ed. San Francisco: Morgan Kaufman.


6. ACKNOWLEDGEMENTS

This work was supported by grants from the National Heart, Lung, and Blood Institute HL59868, the National Center on Minority Health and Health Disparity P20 MD00175 and the National Institute of Health / National Library of Medicine F38 LM008882.
Appendix I – Figures and Tables referenced in the paper

1. Resting BP/HR – 5 minutes
2. Immerse hand in ice water: BP/HR – 3 minutes
3. Recovery BP/HR – 5 minutes

Figure 1: Study Protocol

1. Understand the context
2. Preprocess the data
3. Modeling and analysis
4. Knowledge Discovery

Figure 2: KDD Process
Table 1 Clinical Profile of Subjects with significant Female/Male differences

<table>
<thead>
<tr>
<th>Variable</th>
<th>Total n</th>
<th>Overall $X$</th>
<th>Female $n$</th>
<th>Female $X$</th>
<th>Male $n$</th>
<th>Male $X$</th>
<th>P-value $^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>176</td>
<td>32.5 ± 1</td>
<td>113</td>
<td>32.5 ± 1</td>
<td>63</td>
<td>32.5 ± 1</td>
<td>0.9655</td>
</tr>
<tr>
<td>Systolic</td>
<td>175</td>
<td>121 ± 1</td>
<td>112</td>
<td>118.0 ± 1</td>
<td>63</td>
<td>126 ± 2</td>
<td>0.0006</td>
</tr>
<tr>
<td>Diastolic</td>
<td>175</td>
<td>74 ± 1</td>
<td>112</td>
<td>74 ± 1</td>
<td>63</td>
<td>74 ± 1</td>
<td>0.8393</td>
</tr>
<tr>
<td>MAP</td>
<td>175</td>
<td>91 ± 1</td>
<td>112</td>
<td>90 ± 1</td>
<td>63</td>
<td>93 ± 1</td>
<td>0.0880</td>
</tr>
<tr>
<td>BPM (HR)</td>
<td>175</td>
<td>69 ± 1</td>
<td>112</td>
<td>71 ± 1</td>
<td>63</td>
<td>66 ± 2</td>
<td>0.0880</td>
</tr>
<tr>
<td>HOMA</td>
<td>101</td>
<td>1.6 ± 0.2</td>
<td>63</td>
<td>2 ± 0.2</td>
<td>39</td>
<td>0.89 ± 0.2</td>
<td>0.0009</td>
</tr>
<tr>
<td>Glucose</td>
<td>118</td>
<td>71.8 ± 3</td>
<td>76</td>
<td>75.1 ± 4</td>
<td>42</td>
<td>65.7 ± 4</td>
<td>0.0918</td>
</tr>
<tr>
<td>Insulin</td>
<td>105</td>
<td>8.4 ± 0.7</td>
<td>66</td>
<td>9.9 ± 1</td>
<td>39</td>
<td>5.9 ± 1</td>
<td>0.0074</td>
</tr>
<tr>
<td>BMI</td>
<td>165</td>
<td>28.9 ± 0.5</td>
<td>107</td>
<td>29.3 ± 0.7</td>
<td>58</td>
<td>28.0 ± 0.8</td>
<td>0.2415</td>
</tr>
<tr>
<td>Triglycerides</td>
<td>116</td>
<td>88 ± 4</td>
<td>74</td>
<td>87 ± 5</td>
<td>42</td>
<td>89 ± 7</td>
<td>0.8704</td>
</tr>
<tr>
<td>Cholesterol</td>
<td>116</td>
<td>182 ± 4</td>
<td>74</td>
<td>185 ± 5</td>
<td>42</td>
<td>178 ± 6</td>
<td>0.3822</td>
</tr>
<tr>
<td>HDL</td>
<td>116</td>
<td>57 ± 2</td>
<td>74</td>
<td>58 ± 2</td>
<td>42</td>
<td>54 ± 2</td>
<td>0.1764</td>
</tr>
<tr>
<td>LDL</td>
<td>116</td>
<td>107 ± 4</td>
<td>74</td>
<td>109 ± 5</td>
<td>42</td>
<td>104 ± 6</td>
<td>0.4476</td>
</tr>
</tbody>
</table>

Table 2 : Default SAS Miner 4.3 Tree parameters

1. Understand the context
2. Preprocess the data
3. Modeling and analysis
4. Knowledge Discovery

Figure 3

Splitting criterion: F Test
Significance Level: 0.2
Minimum number of observations in a leaf: 5
Observations required for a split search: 10
Maximum number of branches from a node: 2
Maximum depth of tree: 6
Splitting rules saved in each node: 5
Surrogate rules saved in each node: 0
Do not treat missing as an acceptable value
Model assessment measure: Average Square Error
Subtree: Best assessment value
Observations sufficient for split search: 159
Maximum tries in an exhaustive split search: 5000
P-value adjustment: KASS DEPTH
Apply KASS BEFORE choosing number of branches
<table>
<thead>
<tr>
<th>Variable</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Missing %</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLUCOSE</td>
<td>2.0</td>
<td>161.0</td>
<td>71.8</td>
<td>29.12</td>
<td>33.33</td>
<td>-0.489</td>
<td>1.084</td>
</tr>
<tr>
<td>CHOLESTEROL</td>
<td>94.0</td>
<td>327.0</td>
<td>182.3</td>
<td>42.71</td>
<td>34.46</td>
<td>0.932</td>
<td>1.146</td>
</tr>
<tr>
<td>TRIGLYCERIDES</td>
<td>29.0</td>
<td>216.0</td>
<td>88.1</td>
<td>41.33</td>
<td>34.46</td>
<td>1.191</td>
<td>1.033</td>
</tr>
<tr>
<td>HDL</td>
<td>27.0</td>
<td>118.0</td>
<td>56.8</td>
<td>15.87</td>
<td>34.46</td>
<td>0.690</td>
<td>1.191</td>
</tr>
<tr>
<td>LDL</td>
<td>11.0</td>
<td>221.0</td>
<td>107.0</td>
<td>39.17</td>
<td>34.46</td>
<td>0.705</td>
<td>0.553</td>
</tr>
<tr>
<td>INSULIN</td>
<td>0.00</td>
<td>42.5</td>
<td>8.5</td>
<td>7.50</td>
<td>40.68</td>
<td>1.948</td>
<td>4.926</td>
</tr>
<tr>
<td>AGE</td>
<td>18</td>
<td>65</td>
<td>32.5</td>
<td>12.4</td>
<td>0.565</td>
<td>0.524</td>
<td>-0.875</td>
</tr>
<tr>
<td>WAIST_CAVG</td>
<td>32.0</td>
<td>135.5</td>
<td>87.2</td>
<td>15.46</td>
<td>15.82</td>
<td>0.444</td>
<td>1.134</td>
</tr>
<tr>
<td>INCOME</td>
<td>1</td>
<td>20</td>
<td>4.85</td>
<td>4.78</td>
<td>26.55</td>
<td>1.578</td>
<td>2.164</td>
</tr>
<tr>
<td>EDUCATION</td>
<td>9</td>
<td>21</td>
<td>14.3</td>
<td>2.164</td>
<td>22.03</td>
<td>0.255</td>
<td>0.054</td>
</tr>
<tr>
<td>FATHER EDUCATION</td>
<td>3</td>
<td>21</td>
<td>12.7</td>
<td>3.53</td>
<td>31.07</td>
<td>0.348</td>
<td>0.192</td>
</tr>
<tr>
<td>MOTHER EDUCATION</td>
<td>4</td>
<td>20</td>
<td>12.9</td>
<td>3.19</td>
<td>27.12</td>
<td>0.044</td>
<td>0.233</td>
</tr>
<tr>
<td>RESTING SBP</td>
<td>93.7</td>
<td>170.1</td>
<td>121.0</td>
<td>14.2</td>
<td>0.565</td>
<td>0.672</td>
<td>0.789</td>
</tr>
<tr>
<td>RESTING DBP</td>
<td>56.2</td>
<td>102.2</td>
<td>73.4</td>
<td>9.32</td>
<td>0.565</td>
<td>1.012</td>
<td>0.757</td>
</tr>
<tr>
<td>RESTING HR</td>
<td>41.9</td>
<td>116.7</td>
<td>71.4</td>
<td>11.0</td>
<td>0.565</td>
<td>0.518</td>
<td>1.388</td>
</tr>
<tr>
<td>BMI</td>
<td>17.2</td>
<td>49.7</td>
<td>28.9</td>
<td>6.65</td>
<td>7.345</td>
<td>0.829</td>
<td>0.290</td>
</tr>
<tr>
<td>ISEL</td>
<td>2</td>
<td>80</td>
<td>32.4</td>
<td>14.2</td>
<td>4.5</td>
<td>0.921</td>
<td>0.823</td>
</tr>
<tr>
<td>APPRAISAL</td>
<td>0</td>
<td>16</td>
<td>4.8</td>
<td>4.078</td>
<td>4.519</td>
<td>0.828</td>
<td>0.002</td>
</tr>
<tr>
<td>BELONGING</td>
<td>0</td>
<td>24</td>
<td>5.5</td>
<td>4.052</td>
<td>4.519</td>
<td>0.904</td>
<td>1.543</td>
</tr>
<tr>
<td>TANGIBLE</td>
<td>0</td>
<td>20</td>
<td>4.5</td>
<td>4.192</td>
<td>4.519</td>
<td>0.964</td>
<td>0.441</td>
</tr>
<tr>
<td>SELF ESTEEM</td>
<td>0</td>
<td>23</td>
<td>7.8</td>
<td>3.997</td>
<td>4.519</td>
<td>0.600</td>
<td>0.746</td>
</tr>
<tr>
<td>COOK HOSTILITY</td>
<td>1</td>
<td>18</td>
<td>9.4</td>
<td>3.36</td>
<td>1.695</td>
<td>-0.105</td>
<td>0.259</td>
</tr>
<tr>
<td>CYNICISM</td>
<td>0</td>
<td>10</td>
<td>6.3</td>
<td>2.205</td>
<td>1.695</td>
<td>-0.592</td>
<td>0.249</td>
</tr>
<tr>
<td>JHAC</td>
<td>4</td>
<td>60</td>
<td>49.7</td>
<td>7.11</td>
<td>1.695</td>
<td>-1.868</td>
<td>8.805</td>
</tr>
<tr>
<td>BDI</td>
<td>0</td>
<td>37</td>
<td>7.7</td>
<td>7.3</td>
<td>2.82</td>
<td>1.753</td>
<td>3.624</td>
</tr>
<tr>
<td>STAI-X1 ANXIETY</td>
<td>20</td>
<td>54</td>
<td>33.2</td>
<td>8.189</td>
<td>2.26</td>
<td>0.441</td>
<td>-0.801</td>
</tr>
<tr>
<td>STAI-X2 ANXIETY</td>
<td>21</td>
<td>54</td>
<td>37.5</td>
<td>8.081</td>
<td>74.58</td>
<td>0.272</td>
<td>-0.623</td>
</tr>
<tr>
<td>ANGER_IN</td>
<td>8</td>
<td>32</td>
<td>15.7</td>
<td>4.399</td>
<td>3.39</td>
<td>0.725</td>
<td>0.739</td>
</tr>
<tr>
<td>ANGER_OUT</td>
<td>8</td>
<td>31</td>
<td>14.8</td>
<td>4.20</td>
<td>3.39</td>
<td>1.268</td>
<td>2.155</td>
</tr>
<tr>
<td>SAE</td>
<td>28</td>
<td>65</td>
<td>46.9</td>
<td>6.5</td>
<td>3.39</td>
<td>-0.154</td>
<td>0.239</td>
</tr>
<tr>
<td>ANGER CONTROL</td>
<td>4</td>
<td>13</td>
<td>7.9</td>
<td>2.5</td>
<td>3.39</td>
<td>0.173</td>
<td>-0.954</td>
</tr>
<tr>
<td>PRS YEAR</td>
<td>0</td>
<td>151</td>
<td>40.1</td>
<td>25.2</td>
<td>1.69</td>
<td>0.999</td>
<td>2.069</td>
</tr>
<tr>
<td>PRS LIFE</td>
<td>0</td>
<td>157</td>
<td>51.1</td>
<td>28.9</td>
<td>1.69</td>
<td>0.693</td>
<td>0.874</td>
</tr>
<tr>
<td>PRS EMOTION</td>
<td>1</td>
<td>100</td>
<td>43.7</td>
<td>19.5</td>
<td>3.389</td>
<td>0.346</td>
<td>0.088</td>
</tr>
<tr>
<td>Variable</td>
<td>Min</td>
<td>Max</td>
<td>Mean</td>
<td>Std. Dev.</td>
<td>Missing %</td>
<td>Skewness</td>
<td>Kurtosis</td>
</tr>
<tr>
<td>----------------</td>
<td>-----</td>
<td>-----</td>
<td>------</td>
<td>-----------</td>
<td>-----------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>PRS BEHAVIOR</td>
<td>3</td>
<td>37</td>
<td>11.2</td>
<td>6.49</td>
<td>61.02</td>
<td>1.458</td>
<td>3.267</td>
</tr>
</tbody>
</table>

Table 3 – Interval Variables

<table>
<thead>
<tr>
<th>Fit Statistic</th>
<th>Training</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Squared Error</td>
<td>140.72</td>
<td>162.26</td>
</tr>
<tr>
<td>Sum of Squared Errors</td>
<td>22233.26</td>
<td>2920.64</td>
</tr>
<tr>
<td>Root Average Squared Error</td>
<td>11.86</td>
<td>12.74</td>
</tr>
<tr>
<td>Maximum Absolute Error</td>
<td>38.55</td>
<td>29.17</td>
</tr>
<tr>
<td>Divisor for ASE</td>
<td>158.00</td>
<td>18.00</td>
</tr>
<tr>
<td>Total Degrees of Freedom</td>
<td>158.00</td>
<td></td>
</tr>
<tr>
<td>Number of Estimated Weights</td>
<td>4.00</td>
<td></td>
</tr>
<tr>
<td>Sum of Frequencies</td>
<td>158.00</td>
<td>18.00</td>
</tr>
<tr>
<td>Sum Case Weights * Frequencies</td>
<td>158.00</td>
<td>18.00</td>
</tr>
</tbody>
</table>

Table 4 – Model Assessment
Figure 4: Systolic Blood Pressure Classification Decision Tree

---

Legend
P pre-hypertensive
N normal blood pressure
H hypertensive

Figure 6: Predictive Hypertension Decision Tree
Appendix Description of Enterprise Miner 4.2

Enterprise Miner 4.2 is a part of Base SAS and can be accessed through the Base SAS GUI. SAS also has a second generation of Enterprise Miner, currently 5.3, that requires additional licensing.

SAS Enterprise Miner has a drag and drop workflow as illustrated in Figure A-1. Users select and item from the left hand menu and drag it onto the right hand side palette. Data flows between objects are made by clicking the right hand corner of an object and dragging to the next object in the work flow. Right clicking an object reveals a menu for configuration and execution of the task.

Figure A-1
A view of the configuration options for the tree model are shown below in Figures A-2.

Results summary from the decision tree model includes tree ring, color coding for the target values, and a plot of the mean square error by the number of leaves for the training data and the validation data. This chart facilitates selecting a tree size that balances model fit and overfitting the data.

Enlarged view of plot of training data and validation data average square error for the number of leaves in the tree.
Tree ring plot illustrating mouse over feature to show the attribute values for the tree segment.

Competing splits chart by right clicking a node in the tree.
Example of tree diagram is shown below. Left hand values are the training data while right hand values show the validation data. The tree is color coded where light colors are low values and dark colors are high values of the target.