



In this issue:

## Benefits of Knowledge Discovery Process for Biomedical Population Study

**Marilyn K. McClelland**

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

---

**Recommended Citation:** McClelland (2010). Benefits of Knowledge Discovery Process for Biomedical Population Study. *Journal of Information Systems Applied Research*, 3 (2). <http://jisar.org/3/2/>. ISSN: 1946-1836. (A preliminary version appears in *The Proceedings of CONISAR 2008*: §2325. ISSN: 0000-0000.)

This issue is on the Internet at <http://jisar.org/3/2/>

The **Journal of Information Systems Applied Research** (JISAR) is a peer-reviewed academic journal published by the Education Special Interest Group (EDSIG) of the Association of Information Technology Professionals (AITP, Chicago, Illinois). • ISSN: 1946-1836. • First issue: 1 Dec 2008. • Title: Journal of Information Systems Applied Research. Variants: JISAR. • Physical format: online. • Publishing frequency: irregular; as each article is approved, it is published immediately and constitutes a complete separate issue of the current volume. • Single issue price: free. • Subscription address: [subscribe@jisar.org](mailto:subscribe@jisar.org). • Subscription price: free. • Electronic access: <http://jisar.org/> • Contact person: Don Colton ([editor@jisar.org](mailto:editor@jisar.org))

### 2010 AITP Education Special Interest Group Board of Directors

Don Colton Brigham Young Univ Hawaii EDSIG President 2007-2008	Thomas N. Janicki Univ NC Wilmington EDSIG President 2009-2010	Alan R. Peslak Penn State Vice President 2010	
Scott Hunsinger Appalachian State Membership 2010	Michael A. Smith High Point Univ Secretary 2010	Brenda McAleer U Maine Augusta Treasurer 2010	George S. Nezelek Grand Valley State Director 2009-2010
Patricia Sendall Merrimack College Director 2009-2010	Li-Jen Shannon Sam Houston State Director 2009-2010	Michael Battig St Michael's College Director 2010-2011	Mary Lind North Carolina A&T Director 2010-2011
Albert L. Harris Appalachian St JISE Editor ret.	S. E. Kruck James Madison U JISE Editor	Wendy Ceccucci Quinnipiac University Conferences Chair 2010	Kevin Jetton Texas State FITE Liaison 2010

### Journal of Information Systems Applied Research Editors

Don Colton Professor BYU Hawaii Editor	Scott Hunsinger Assistant Professor Appalachian State Associate Editor	Alan R. Peslak Associate Professor Penn State Associate Editor	Thomas N. Janicki Associate Professor UNC Wilmington Associate Editor
---	---	---	--

This paper was selected for inclusion in the journal based on blind reviews from three or more peers placing it in the 30% acceptance rate category for papers submitted to CONISAR 2008.

EDSIG activities include the publication of JISAR and ISEDJ, the organization and execution of the annual CONISAR and ISECON conferences held each fall, the publication of the Journal of Information Systems Education (JISE), and the designation and honoring of an IS Educator of the Year. • The Foundation for Information Technology Education has been the key sponsor of ISECON over the years. • The Association for Information Technology Professionals (AITP) provides the corporate umbrella under which EDSIG operates.

© Copyright 2010 EDSIG. In the spirit of academic freedom, permission is granted to make and distribute unlimited copies of this issue in its PDF or printed form, so long as the entire document is presented, and it is not modified in any substantial way.

# Benefits of Knowledge Discovery Process for Biomedical Population Study

Marilyn K. McClelland  
mmcclell@nccu.edu

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

## Abstract

Knowledge discovery and database (KDD) processes and best practices are utilized to build a knowledge management infrastructure for the first biomedical population study conducted at North Carolina Central University. The infrastructure includes data dictionaries, process documentation, document repository and 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. 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 process described here.

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

## 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)
- Self-esteem: Interpersonal Support Evaluation List (ISEL) (Cohen & Hoberman, 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,... )

### 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 meas-

ured at 1-minute intervals throughout the procedure including the recovery period. A psychological stressor was given after baseline 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, Piatetsky-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, *Step 1. Understand the context*, 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. Its necessary to have height and weight in the same scale before calculating Body Mass Index ( BMI ) and to insure 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 hemolysis 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 hemolysis. 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 insure 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 psychosocials, 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 CP and AR. PROC TTEST reported the significance of the difference by sex of the baseline clinical measures.

PROC GLM with the SNK option calculated Student-Newman-Kuhl's test to assess significant differences in blood pressure readings during baseline, stress, and recovery periods as a repeated measures ANOVA. PROC CORR calculated correlations between psychosocial and clinical measures of blood pressure response to acute stressors.

Use of regression models for the biomedical data mart are challenged by missing data, lack of well defined hypotheses, and lack of linearity in relationships. We evaluated other models including hierarchical clustering and decision trees. Missing data again proved problematic for clustering. As with regression models, we either have to use a mean replacement or delete the entire attribute for clustering.

Decision trees are a palatable approach since missing data can be accommodated. Decision trees also provide intuitive views of the data with a well understood paradigm of a branching tree that work well in presenting the models to the content area experts. Decision trees can handle high dimensional data like the multi-faceted biomedical data here. Decision trees are appropriate for hierarchical data. SAS Enterprise Miner 4.2 provided an easy implementation of decision trees since the data was already in a SAS data set. Appendix II provides an overview of Enterprise Miner 4.2 for decision trees.

A key question for decision tree is to determine the target attribute which becomes the root of the tree. One of our research questions is to explore significant factors contributing to baseline blood pressure for African Americans. Here we utilize a decision tree as a vehicle for exploration.

**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 decision 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 graph 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.

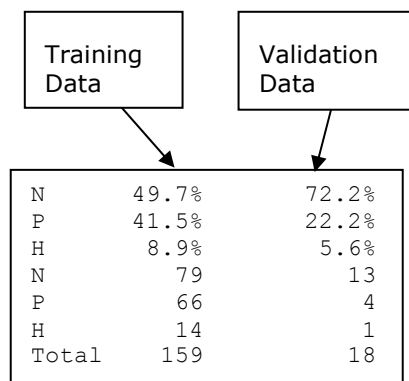
**Predictive Decision Tree for Resting Blood Pressure Diagnosis as Target:** A new variable was added to the data set with the value 'N' for normal when average baseline systolic blood pressure is below 120; value is 'P' for pre-hypertensive when average systolic blood pressure ranges from 120

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.



	Training Data	Validation Data
N	49.7%	72.2%
P	41.5%	22.2%
H	8.9%	5.6%
N	79	13
P	66	4
H	14	1
Total	159	18

Figure 5: Trunk of Prediction Tree

The tree is shown in figure 6. Enterprise Miner generated trees have color coding of the nodes to reflect low values (yellow) and high values (red) of the target. Waist\_CAVG is the average waist circumference in centimeters for the subject. Two waist measures are recorded for the subject. If one measure differs by more than a half a centimeter, a third measurement is taken. Waist circumference provides the best partition for the next branch of the tree. The left branch for waist below 88.08 cm. The interpretation is with a waist below 88.08 cm, 61.5% will be normal SBP, 31.7% will be pre-hypertensive and 6.7% hypertensive. On the right hand side, with a waist circumference above

88.08 cm, only 27.3% are normal SBP, 60% are pre-hypertensive, and 12.7% are hypertensive. So having a large waist circumference cuts the likelihood of normal SBP from 61.5% to 27.3 % or just about in half. Likewise, the likelihood of pre-hypertensive goes from 31.7% to 60%, or almost double. Hypertensive also almost doubles from 6.7% to 12.7%.

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 elevate 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 (Breault, Goodall, & Fos, 2002) (Dolce, Quintieri, Serra, Lagani, & Pignolo, 2008) (Guan, Huang, Guo, Wang, & Zhou, 2008) (Ho, Park, Jang, & Jee, 2007, Zhu, Li, Zhao, Cui, Zhao, Chang, Feng, & Wei, 2008) (Goel, Misra, Kondal, Pandey, Vikram, Wasir, Dhingra, & Luthra, 2008).

I plan to investigate techniques useful in other biomedical research like boosting (Deconinck, Zhang, Petit, Dubus, Ijjaali, Coomans, & Vander Heyden, 2008) (Elith, Leathwick, & Hastie, 2008), and two stage models where decision trees identify or select variables for subsequent use with a regression model (Ferguson, Siddique, & Karri-

son, 2008, Kohrt, Olshen, Bermas, Goodson, Wood, Henry, Rouse, Bailey, Philben, Dirbas, Dunn, Johnson, Wapnir, Carlson, Stockdale, Hansen, & Jeffrey, 2008). I am also interested evaluating the performance of random forests classifiers and support vector machines for the dataset as many have shown these to be better classifiers (Statnikov, Wang, & Aliferis, 2008).

#### **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. 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. Valuable contributions were made by Dr. Mildred Pointer, Dr. Richard Bukowski, Ms. Sadiqa Yancey, Ms. Ranim Abou-chacra, Ms. Patricia Petrusi, Dr. Jonathan Livingston, Dr.

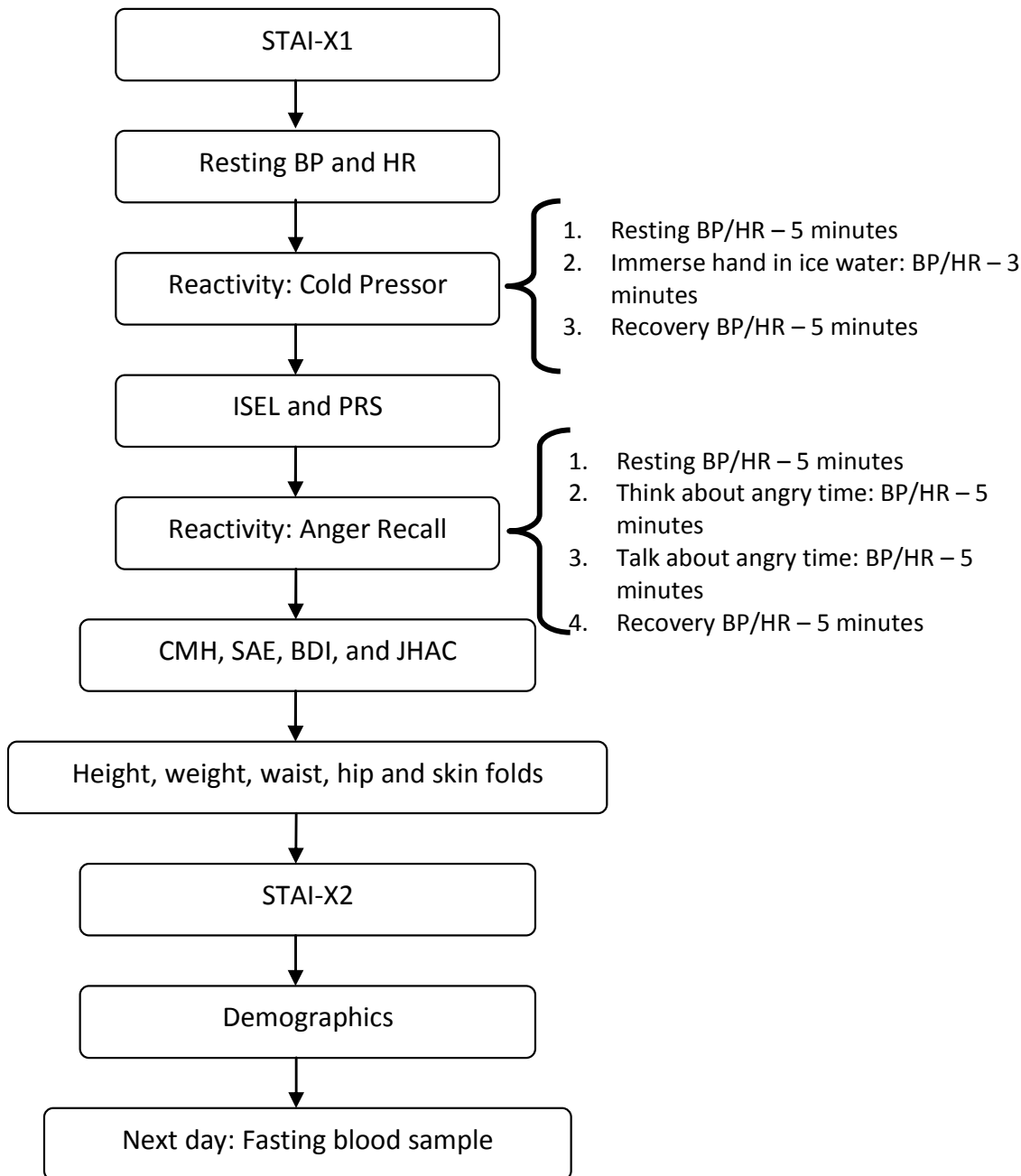
Redford Williams, and Dr. Elizabeth Hauser. SAS Institute provided training in the use of SAS Enterprise Miner for predictive modeling.

## 6. REFERENCES

- Austin, P. C. 2008. "R and S-PLUS produced different classification trees for predicting patient mortality". *J Clin Epidemiol*.
- Beck, A. T., C. H. Ward, M. Mendelson, J. Mock, & J. Erbaugh. 1961. "An inventory for measuring depression". *Arch Gen Psychiatry*, 4: 561-71.
- Braga-Neto, U., R. Hashimoto, E. R. Dougherty, D. V. Nguyen, & R. J. Carroll. 2004. "Is cross-validation better than resubstitution for ranking genes?" *Bioinformatics*, 20(2): 253-8.
- Breault, J. L., C. R. Goodall, & P. J. Fos. 2002. "Data mining a diabetic data warehouse". *Artif Intell Med*, 26(1-2): 37-54.
- Cohen, S. & HM Hoberman. 1983. "Positive events and Social Support as Buffers of Life Change Stress". *Journal of Applied Social Psychology*, 13: 99-125.
- Cook, W. & D. Medley. 1954. "Proposed hostility and pharisaic virtue scales for the MMPI". *Journal of Applied Psychology*, 38: 414-18.
- Deconinck, E., M. H. Zhang, F. Petitet, E. Dubus, I. Ijjaali, D. Coomans, & Y. Vander Heyden. 2008. "Boosted regression trees, multivariate adaptive regression splines and their two-step combinations with multiple linear regression or partial least squares to predict blood-brain barrier passage: a case study". *Anal Chim Acta*, 609(1): 13-23.
- Dolce, G., M. Quintieri, S. Serra, V. Lagani, & L. Pignolo. 2008. "Clinical signs and early prognosis in vegetative state: a decisional tree, data-mining study". *Brain Inj*, 22(7): 617-23.
- Elith, J., J. R. Leathwick, & T. Hastie. 2008. "A working guide to boosted regression trees". *J Anim Ecol*, 77(4): 802-13.
- Ferguson, M. K., J. Siddique, & T. Karrison. 2008. "Modeling major lung resection outcomes using classification trees and multiple imputation techniques". *Eur J Cardiothorac Surg*.
- Frayyad, U., G. Piatetsky-Shapiro, & P. Smyth. 1996. "From Data Mining to Knowledge Discovery in Databases". *AI Magazine*: 213-28.
- Goel, R., A. Misra, D. Kondal, R. M. Pandey, N. K. Vikram, J. S. Wasir, V. Dhingra, & K. Luthra. 2008. "Identification of insulin resistance in Asian Indian adolescents: classification and regression tree (CART) and logistic regression based classification rules". *Clin Endocrinol (Oxf)*.
- Guan, P., D. Huang, J. Guo, P. Wang, & B. Zhou. 2008. "Bacillary dysentery and meteorological factors in northeastern china: a historical review based on classification and regression trees". *Jpn J Infect Dis*, 61(5): 356-60.
- Han, Jiawei & Micheline Kamber. 2006. *Data Mining Concepts and Techniques*. second ed. San Francisco: Morgan Kaufman.
- Ho, S. H., H. Y. Park, Y. S. Jang, & S. H. Jee. 2007. "Risk stratification for LDL cholesterol using induction technique". *Stud Health Technol Inform*, 129(Pt 1): 571-5.
- James, S. A., S. A. Hartnett, & W. D. Kalsbeek. 1983. "John Henryism and blood pressure differences among black men". *J Behav Med*, 6(3): 259-78.
- Jiawei, Han, B. Altman Russ, Kumar Vipin, Mannila Heikki, & Pregibon Daryl. 2002. "Emerging scientific applications in data mining". *Commun. ACM*, 45(8): 54-58.
- Kohrt, H. E., R. A. Olshen, H. R. Bermas, W. H. Goodson, D. J. Wood, S. Henry, R. V. Rouse, L. Bailey, V. J. Philben, F. M. Dirbas, J. J. Dunn, D. L. Johnson, I. L. Wapnir, R. W. Carlson, F. E. Stockdale, N. M. Hansen, & S. S. Jeffrey. 2008. "New models and online calculator for predicting non-sentinel lymph node status in sentinel lymph node positive breast cancer patients". *BMC Cancer*, 8: 66.
- McNeilly, M. D., N. B. Anderson, C. A. Armstead, R. Clark, M. Corbett, E. L. Robinson, C. F. Pieper, & E. M. Lepisto. 1996. "The perceived racism scale: a multidimensional assessment of the experience of white racism among African Americans". *Ethn Dis*, 6(1-2): 154-66.
- Murthy, Sreerama K. 1998. "Automatic Construction of Decision Trees from Data: A

- Multi-Disciplinary Survey". *KDD Journal*, 2(4): 345-89.
- Pointer, M. A., J. N. Livingston, S. Yancey, M. K. McClelland, & R. D. Bukoski. 2008. "Psychosocial factors contribute to resting blood pressure in African Americans". *Ethn Dis*, 18(3): 289-93.
- Potts, William J.E. & Lorne Rothman. 2006. *Decision Tree Modeling Course Notes*. Cary, NC: SAS Institute, Inc.
- SAS Institute. 2006. *Base SAS 9.1.3 Procedures Guide*. 2 ed: SAS Institute.
- Spielberger, C. D. 1983. *Manual for the State-Trait Anxiety Inventory*. Palo Alto: Consulting Psychologists Press.
- Spielberger, C. D., E. H. Johnson, S. F. Russell, R.J. Crane, G. A. Jacobs, & T. J. Worden. 1985. The Experience and Expression of Anger: Construction and Validation of an Anger Expression Scale. In Chesney, M. A. & R. H. Rosenman, editors, *Anger and hostility in cardiovascular and behavioral medicine*. Washington DC: Hemisphere.
- Statnikov, A., L. Wang, & C. F. Aliferis. 2008. "A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification". *BMC Bioinformatics*, 9: 319.
- Zhu, H. L., Y. Li, Y. M. Zhao, H. Cui, Y. Zhao, X. H. Chang, J. Feng, & L. H. Wei. 2008. "[Classification tree analysis in serous ovarian adenocarcinoma patients for prognostic factors associated with three-year survival probability]". *Zhonghua Fu Chan Ke Za Zhi*, 43(3): 201-4.

**Appendix I** – Figures and Tables referenced in the paper



**Figure 1: Study Protocol**

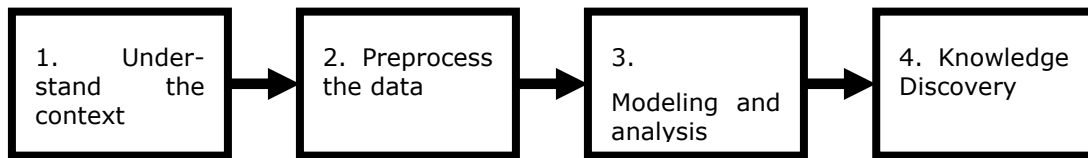


Figure 2 : KDD Process

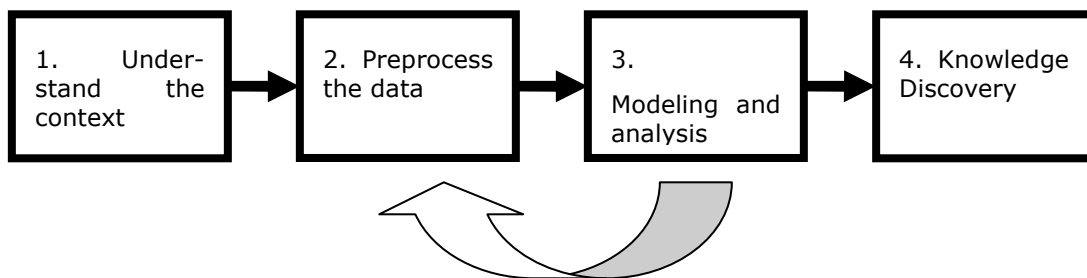


Figure 3

	Total n	Overall X	Female n	Female X	Male n	Male X	P-value <sup>a</sup>
Age	176	32.5 ± 1	113	32.5 ± 1	63	32.5 ± 1	0.9655
Systolic	175	121 ± 1	112	118.0 ± 1	63	126 ± 2	0.0006
Diastolic	175	74 ± 1	112	74 ± 1	63	74 ± 1	0.8393
MAP	175	91 ± 1	112	90 ± 1	63	93 ± 1	0.0880
BPM ( HR)	175	69 ± 1	112	71 ± 1	63	66 ± 2	0.0040
HOMA	101	1.6 ± 0.2	63	2 ± 0.2	39	0.89 ± 0.2	0.0009
Glucose	118	71.8 ± 3	76	75.1 ± 4	42	65.7 ± 4	0.0918
Insulin	105	8.4 ± 0.7	66	9.9 ± 1	39	5.9 ± 1	0.0074
BMI	165	28.9 ± 0.5	107	29.3 ± 0.7	58	28.0 ± 0.8	0.2415
Triglycerides	116	88 ± 4	74	87 ± 5	42	89 ± 7	0.8704
Cholesterol	116	182 ± 4	74	185 ± 5	42	178 ± 6	0.3822
HDL	116	57 ± 2	74	58 ± 2	42	54 ± 2	0.1764
LDL	116	107 ± 4	74	109 ± 5	42	104 ± 6	0.4476

Table 1 Clinical Profile of Subjects with significant Female/Male differences

**Appendix II – SAS Enterprise Miner 4.2 Overview**

Documentation for SAS Enterprise Miner can be found at:

<http://support.sas.com/documentation/onlinedoc/miner/>

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 2 : Default SAS Miner 4.3 Tree parameters**

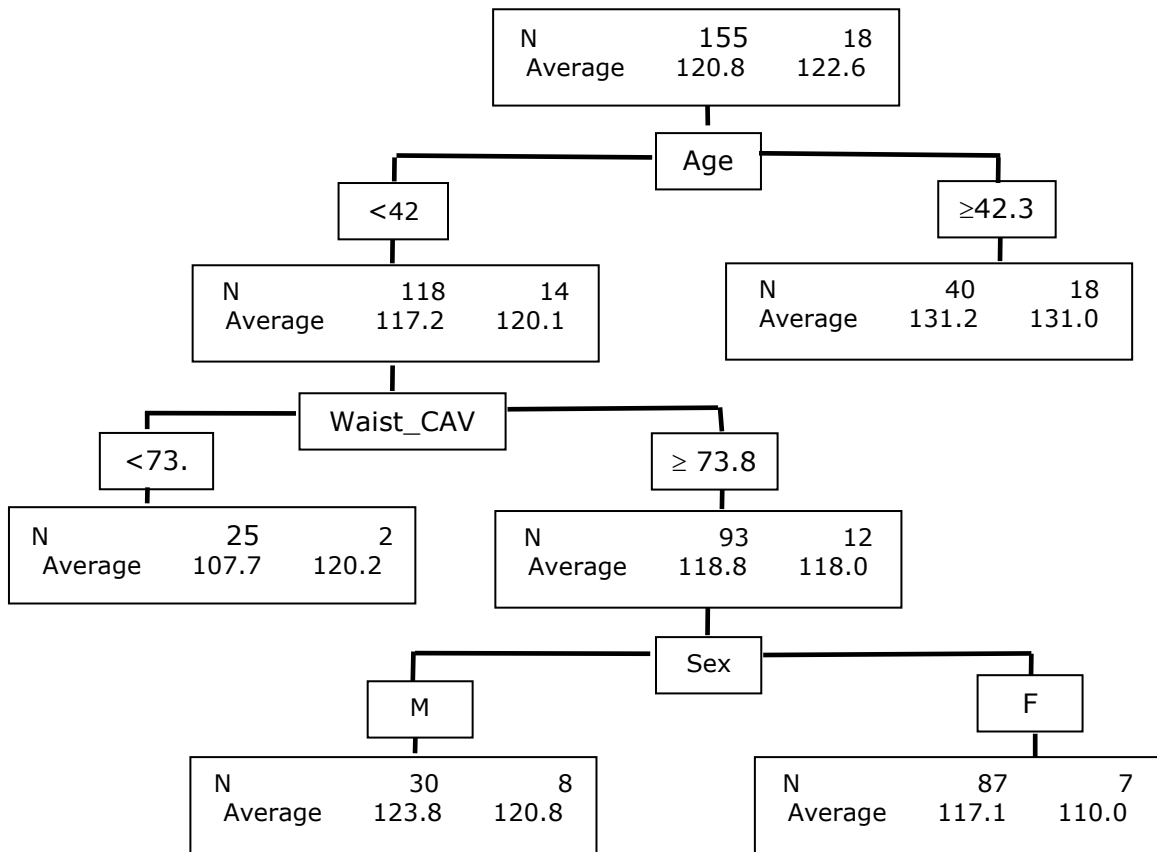
Variable	Min	Max	Mean	Std. Dev.	Missing %	Skewness	Kurtosis
GLUCOSE	2.0	161.0	71.8	29.12	33.33	-0.489	1.084
CHOLESTEROL	94.0	327.0	182.3	42.71	34.46	0.932	1.146
TRIGLYCERIDES	29.0	216.0	88.1	41.33	34.46	1.191	1.033
HDL	27.0	118.0	56.8	15.87	34.46	0.690	1.191
LDL	11.0	221.0	107.0	39.17	34.46	0.705	0.553
INSULIN	0.00	42.5	8.5	7.50	40.68	1.948	4.926
AGE	18	65	32.5	12.4	0.565	0.524	-0.875
WAIST_CAVG	32.0	135.5	87.2	15.46	15.82	0.444	1.134
INCOME	1	20	4.85	4.78	26.55	1.578	2.164
EDUCATION	9	21	14.3	2.164	22.03	0.255	0.054
FATHER EDUCATION	3	21	12.7	3.53	31.07	0.348	0.192
MOTHER EDUCATION	4	20	12.9	3.19	27.12	0.044	0.233
RESTING SBP	93.7	170.1	121.0	14.2	0.565	0.672	0.789
RESTING DBP	56.2	102.2	73.4	9.32	0.565	1.012	0.757
RESTING HR	41.9	116.7	71.4	11.0	0.565	0.518	1.388
BMI	17.2	49.7	28.9	6.65	7.345	0.829	0.290
ISEL	2	80	32.4	14.2	4.5	0.921	0.823

Variable	Min	Max	Mean	Std. Dev.	Missing %	Skewness	Kurtosis
APPRAISAL	0	16	4.8	4.078	4.519	0.828	0.002
BELONGING	0	24	5.5	4.052	4.519	0.904	1.543
TANGIBLE	0	20	4.5	4.192	4.519	0.964	0.441
SELF ESTEEM	0	23	7.8	3.997	4.519	0.600	0.746
COOK HOSTILITY	1	18	9.4	3.36	1.695	-0.105	0.259
CYNICISM	0	10	6.3	2.205	1.695	-0.592	0.249
JHAC	4	60	49.7	7.11	1.695	-1.868	8.805
BDI	0	37	7.7	7.3	2.82	1.753	3.624
STAI-X1 ANXIETY	20	54	33.2	8.189	2.26	0.441	-0.801
STAI-X2 ANXIETY	21	54	37.5	8.081	74.58	0.272	-0.623
ANGER_IN	8	32	15.7	4.399	3.39	0.725	0.739
ANGER_OUT	8	31	14.8	4.20	3.39	1.268	2.155
SAE	28	65	46.9	6.5	3.39	-0.154	0.239
ANGER CONTROL	4	13	7.9	2.5	3.39	0.173	-0.954
PRS YEAR	0	151	40.1	25.2	1.69	0.999	2.069
PRS LIFE	0	157	51.1	28.9	1.69	0.693	0.874
PRS EMOTION	1	100	43.7	19.5	3.389	0.346	0.088
PRS BEHAVIOR	3	37	11.2	6.49	61.02	1.458	3.267

**Table 3 – Interval Variables**

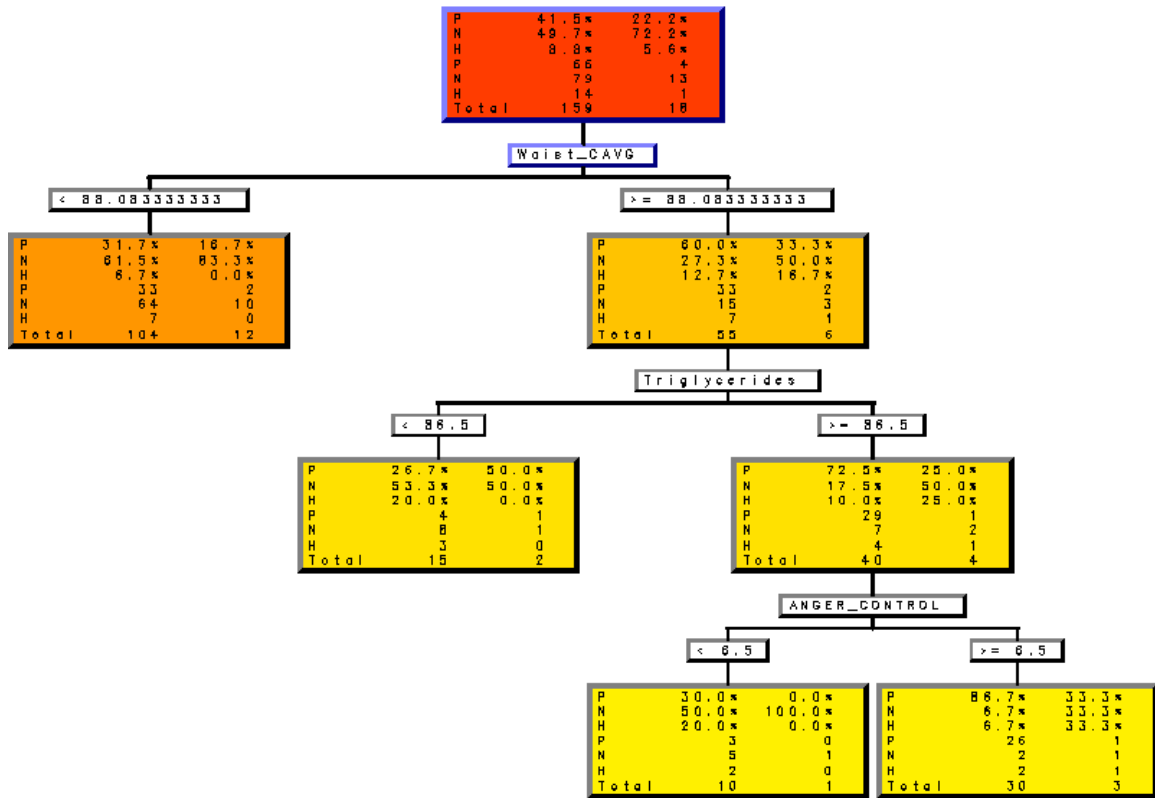
Fit Statistic	Training	Validation
Average Squared Error	140.72	162.26
Sum of Squared Errors	22233.26	2920.64
Root Average Squared Error	11.86	12.74
Maximum Absolute Error	38.55	29.17
Divisor for ASE	158.00	18.00
Total Degrees of Freedom	158.00	.
Number of Estimated Weights	4.00	.
Sum of Frequencies	158.00	18.00
Sum Case Weights * Frequencies	158.00	18.00

**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 an 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.

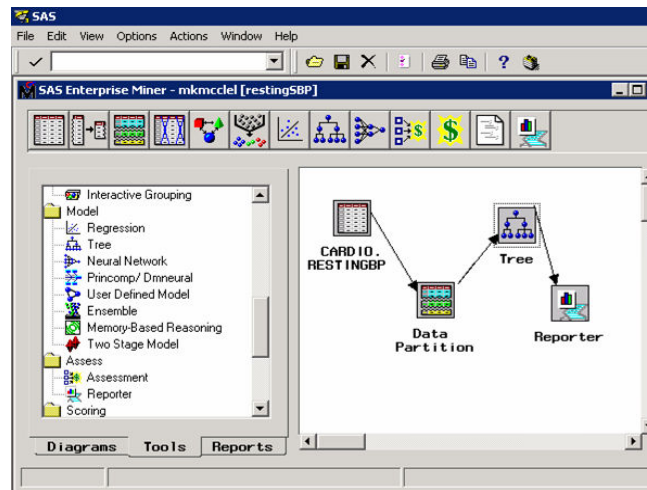


Figure A-1

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

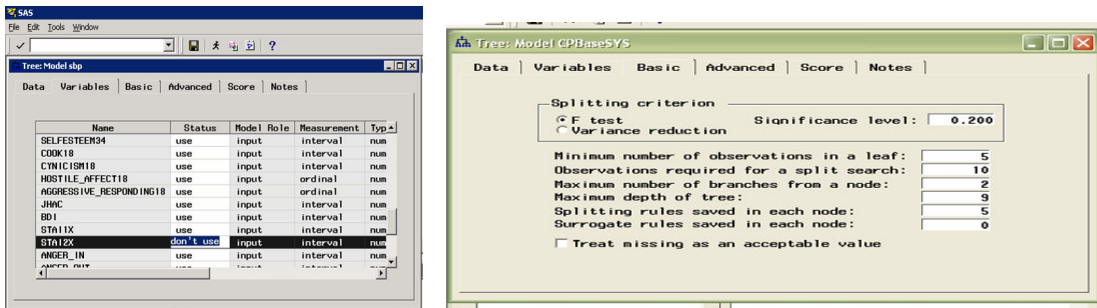
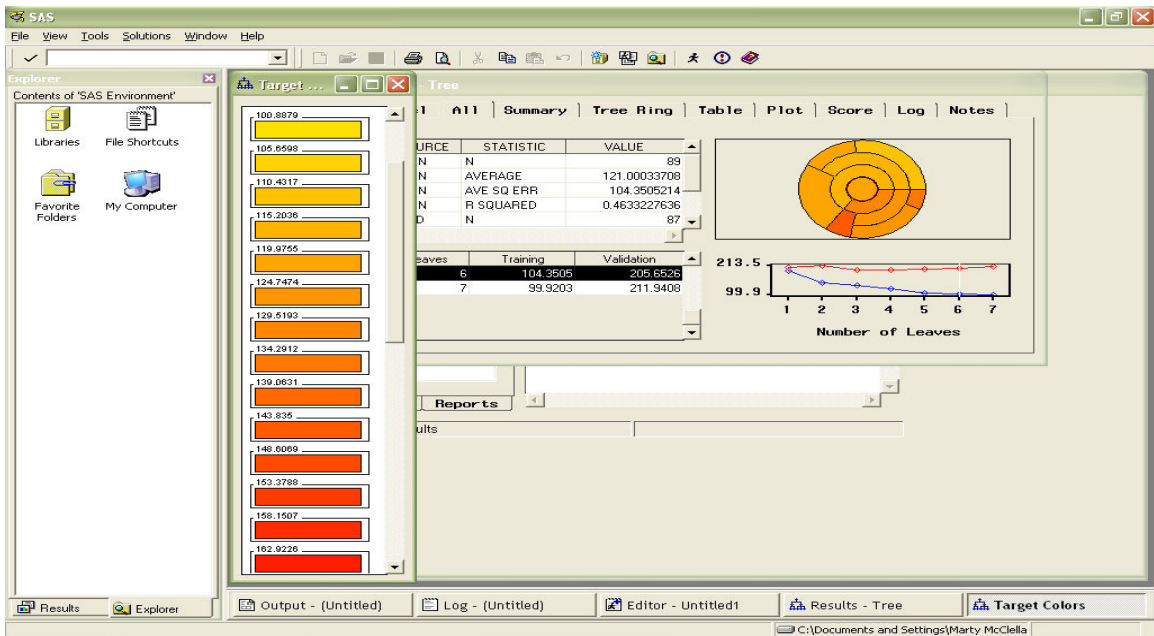
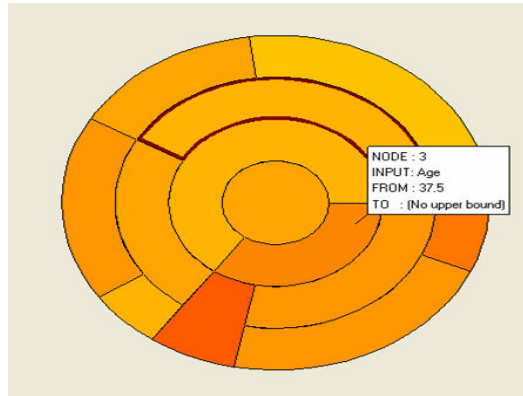


Figure A-2

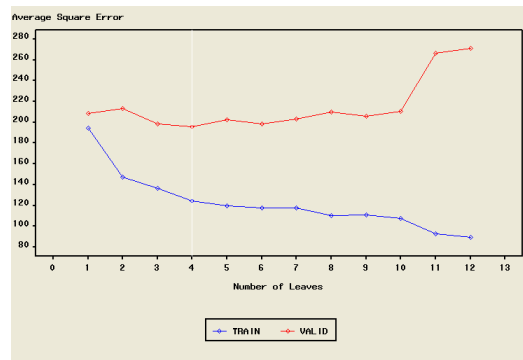
Results summary from the decision tree model includes tree ring, color coding for the target values, and a plot of the mean square error by the number of leaves for the training data and the validation data. This chart facilitates selecting a tree size that balances model fit and overfitting the data.



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.

Variable	Logworth	Groups	Label
WAIST_CAVG	4.054	2	Waist_CAVG
WAIST	3.617	2	
HOSTILE_AFFECT18	2.599	2	
BMI	1.727	2	
SEX	1.457	2	Sex

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.

