Paper 4057-2019
A Comparison of Algorithmic and Epidemiologic Approaches for Modeling Predictors of Type II Diabetes

Wendy Ballew, Rebekah Fallin, Sharon Pearcey, Audrey Whittle, Kennesaw State University

ABSTRACT

Algorithmic variable reduction techniques have the potential to reduce the intensive pre-processing required for regression. This paper employs algorithmic techniques as a variable selection method and an alternative to regression for identifying salient factors. The first approach utilized random forest as a substitute for standard pre-processing and variable reduction, however this method generated a theoretically invalid set of predictors and was not explored further. Subsequently, a set of logistic regression and decision tree models are compared for their efficacy in identifying predictors of Type II diabetes. The optimal decision tree model (a Chi-Squared model with default Enterprise Miner® options) performed comparably to the logistic regression model in terms of Average Squared Error. These findings suggest that data mining methods can successfully be used to supplement traditional epidemiologic research methods in future research.

INTRODUCTION

Modern data mining techniques can often reduce the need for data cleaning prior to analysis. The volume and diversity within healthcare data provide an ideal environment for exploring data mining, and the adoption of data mining methods in the healthcare industry has increased rapidly in recent years. Although many studies have examined the value of algorithmic methods to detect and identify factors associated with heart disease, there is a paucity of research into their usefulness in identifying risk factors for other diseases. Epidemiologic research using the traditional statistical approach often requires time-intensive data cleaning prior to analysis. These approaches also have additional limitations, such as the assumptions necessary for regression: normality, small or no collinearity, homoscedasticity, and the need to impute missing data.

To explore the viability of data mining methods into health research, we selected a health issue where data is validated and easy to obtain; there is significant literature demonstrating an association between the variables and the outcome; there is a relatively limited amount of research into the application of data mining techniques to the issue; and there is a longstanding traditional statistical approach. To this end, we chose the NHIS 2017 data on Type II diabetes.

BACKGROUND

Over 30 million Americans (9.4%) have diabetes (CDC, 2017). The costs, both financial and humanitarian, are enormous. Diabetes can lead to devastating health consequences such as heart disease, nerve damage, kidney damage, vision loss, and bacterial and fungal infections (Mayo Clinic, 2018). Furthermore, it was the seventh leading cause of death in the US in 2015 (CDC, 2017).

The CDC (2017) estimates that 95% of the diagnosed and undiagnosed cases of diabetes in the US are Type II. Type II diabetes occurs when the body becomes insulin resistant or the body cannot make enough insulin to keep up with increased needs (Mayo Clinic, 2018). Although this is the main cause of diabetes, there are many factors that can increase one’s risk. Some of the risk factors include being overweight, having more fat distributed in the abdomen, being more than 45 years old, having close family members with diabetes, having a sedentary lifestyle, and being non-Hispanic Black or Hispanic (CDC, 2017; Mayo Clinic, 2018). Many of these factors are associated with modifiable lifestyle choices. Understanding the relationship among behavior and diabetes is imperative in the hopes of decreasing the incidence of this devastating disease.

DATA MINING IN HEALTH DATA

Although the causes and risk factors associated with diabetes have been investigated extensively, there are few studies that have used data mining procedures to look at this disorder. Algorithmic methods are not plagued by the same limitations as traditional regression modeling. Additionally, there are many national health data sets with hundreds of variables available to provide information from diverse groups of individuals. Currently, data mining is underused in the investigation of health issues in large data sets. This project uses SAS® Enterprise Miner (EM) to investigate risk factors of Type II diabetes using a random forest as a variable reduction technique and decision trees to build a predictive model.

PROBLEM

The purpose of this study was two-fold: first, to address whether data mining techniques can aid in the variable reduction and selection process for health data using a random forest; second, to compare the models which result from regression and data mining. We replicated prior research on diabetes risk factors using a theoretically and clinically validated subset of variables while also creating a separate subset of variables selected using the random forest method.

DATA

The data used for analysis was the 2017 National Health Interview Survey (NHIS). The NHIS is a nationally representative household survey managed by the National Center for Health Statistics, a subset of the CDC. The survey collects information on a wide variety of health status indicators and lifestyle factors of its subjects, and is a face-to-face survey conducted yearly by members of the U.S. Census Bureau. Households are chosen using a multi-stage area probabilistic design, which allows the sample to be representative without being cumbersome. We chose to use only the 2017 survey; therefore, our exploration is conducted on cross-sectional data and risk cannot be estimated.

There were 32,617 households interviewed for the 2017 NHIS, totaling 78,543 persons surveyed. Due to the nature of the survey, only a sample of adults answered the full set of health-related questions, including those pertaining to diabetes. Our final sample size consisted of 26,742 civilian adults. When the Household, Family, and Adult data files were combined, there was a total of 1304 variables. Many of these variables affected the responses of the fields that follow them, as sections of the survey were either answered or skipped depending on the respondents' answers to previous questions. It is impractical to examine the full set of variables for a traditional analysis due to multi-collinearity and interaction. However, for variable reduction using a random forest, we retained the full set of available variables under the assumption that tree models are unaffected by the above issues.

For our target variable, we chose to explore the presence of Type II diabetes versus no diabetes. This was not included in the original dataset due to the flow of survey questions. The outcome variable was created using two original variables: a flag for the presence of diabetes, and a categorical indicator of the type of diabetes. The few respondents who had 'Type I diabetes', 'Prediabetes', or 'Refused to Answer' were treated as missing for our analysis (Table 1).

Table 1: Distribution of Target Variable

dib_2	Frequency	Percent
Has Type II Diabetes	1150	4.30
Does not have Type II Diabetes	24949	93.30
Missing/Type I/Prediabetes	643	2.40

DATA CLEANING/VALIDATION

Before beginning our 'traditional' logistic regression analysis, we attempted to reduce dimensionality and collinearity by pre-selecting a subset of indicators. Since our goal was to compare model effectiveness rather than identify new possible predictors of diabetes, we chose inputs which have been associated with diabetes in existing literature. Our final selection included 30 variables consisting of demographic, behavioral, and health factors. Since our dataset is a survey, all the included variables contained the levels 'Refused', 'Not Ascertained', and 'Don't Know'. As these levels contained very few records, they were binned with the missing category for all inputs. A full list of variables used for this analysis is included in Table A 1 in the Appendix.

After selecting the subset of variables, data exploration consisted of Proc Freq in SAS and Graph Explore in EM to determine if any transformation, replacement, binning, or imputation was needed. In general, all the variables were already suitable for analysis, most likely since the dataset is collected and maintained by the CDC. We found only one variable which needed transformation (BMI) and one variable with too many levels which needed binning (EDUC1- Highest level of education completed). For the transformation of BMI, we explored two possible transformations: 1) the 'Best' transformation as chosen by EM, and 2) the epidemiology industry standard Log transformation. In preparation for regression, all variables were imputed in EM using tree and median methods. Finally, in another attempt to reduce dimensionality and multi-collinearity, two sets of related variables were combined into single binary indicators. These were 'Have you been told in the last year to/Are you currently participating in a weight loss program?' and 'Have you been told in the last year to/Are you currently reducing fat and calories?'

For our data mining approach, we used a random forest method to pare down the full list of variables, using a positive Out of Bag GINI (OOB GINI) as the selection criteria. Upon reviewing the model, we were left with 91 possible predictors. To maintain consistency among the logistic regression models, these subgroups were also binarized by grouping the 'Refused', 'Not Ascertained', and 'Don't Know' levels together with the missing category.

ANALYSIS

Initially, the random forest method seemed to capture useful variables for the tree and logistic regression models; further inspection revealed that the output provided spurious correlations with the dependent variable (Table A 2 in the Appendix). Although a few variables selected by this algorithmic method (i.e., "Imputed_Ever been told you have hypertension" and "BMI") are likely related to modifiable health behaviors that could lead to Type II diabetes, most of the other variables selected by this method were either not related to or were more likely a consequence rather than a cause of Type II diabetes. For example, "total height in inches" and "taking a low dose aspirin on your own" do not have face validity and are likely not instrumental in preventing Type II diabetes. More importantly, other variables selected by this method were likely a consequence of having Type II diabetes (i.e, "Imputed_Seen/talked to foot doctor, past 12 m" and "Use any adaptive devices such as magnifiers"). Given this, we did not continue with further analyses of the models using the variables selected by this method. The literature-validated variables, however, were precursors to diabetes. Therefore, we were able to continue with our second objective of comparing decision trees with logistic regression models.

LOGISTIC REGRESSION MODELS

Using a data partition node in Enterprise Miner, the dataset was split into two parts for the regression models: 70% of the data was used for model training while 30% was used for model validation. Four separate regression analyses were then performed: a full regression and three variants of stepwise regressions (see Diagram, Figure A 2 in Appendix). The stepwise selection criterion for entry and elimination were the defaults used by Enterprise Miner; both the entry and removal criterion were set to 0.15 for all stepwise models.

Table 2 below illustrates the validation statistics for the regression models using variables chosen by the literature, sorted by Average Square Error. The model containing Optimized BMI had the lowest ASE of 0.03292 (see Table A 3 in Appendix for significant variables). Odds ratio estimates are reported in Table A 4 in the Appendix. However, these estimates may be misleading due to the multicollinearity among the predictors.

Table 2: Regression Statistics for Validation Dataset - Chosen by Literature

Model	ASE	Misclassification Rate	ROC Index
Stepwise Regression with Optimized BMI	0.032920	0.042327	0.919
Stepwise Regression with Log-Transformed BMI	0.033319	0.041049	0.914
Stepwise BMI	0.033496	0.042455	0.913
Full Regression	0.034427	0.044373	0.901

The model with the lowest misclassification rate was the Log-Transformed BMI regression. The ROC index mirrored the results of the regression with Optimized BMI, but the ROC indexes are unusually high, likely due to the controlled nature of the data selection. The model selection graph (

Figure A 1) does not appear to be overfitted and the validation ASE closely mirrors the pattern of the training dataset.

DECISION TREE MODELS

After data preparation, we proceeded to create four decision tree models: Maximum Tree, Chi-Squared Default, Gini, and 3 Branch Chi-Squared trees. The Maximum Tree model was one which allowed the leaves to continue to grow until an optimal model was reached; the Chi-Squared Default tree had all default EM settings except for the splitting criteria, which was changed to Chi-Squared; the Gini tree also used default settings with GINI used as splitting criteria; and the 3 Branch Chi-Squared tree had an additional allowance of 3 branches per split (Table 3).

While the models were all structured differently, they had quantitatively similar performances, as shown in the following tables:

Table 3: Number of Leaves and Depth of Decision Tree Models

Model	Leaves	Depth
Chi-Squared Default	5	5
Maximum	32	7
GINI	7	7
Chi-Squared 3 Branch	7	6

The Maximum Tree was the best overall, outperforming all models in terms of ASE and ROC Index (Table 4). The Chi-Squared Default tree was the best in terms of Misclassification rate and second best performing in terms of both ASE and GINI coefficient. The GINI tree had the lowest misclassification rate and the second highest ROC Index. The Chi-Squared 3 Branch tree had the best GINI coefficient, the worst ASE, and the worst ROC Index.

Table 4: Validation Dataset Classifiers - Chosen by Literature

Model	ASE	Misclassification Rate	Gini Coefficient
Maximum	0.034232	0.045269	0.823
Chi-Squared Default	0.034671	0.042583	0.616
GINI	0.034997	0.042711	0.620
Chi-Squared 3 Branch	0.035727	0.044118	0.613

After comparing these statistics, we decided that ASE and Misclassification rate were the two metrics most important for the purpose of the analysis. In addition, interpretability of the model was also important. Therefore, the optimal tree chosen based on these criteria was the Chi-Squared Default tree shown in Figure A 3 in the Appendix.
Variables HYPMDEV2 (Now taking prescribed medicine for high blood pressure), DIBREL (Blood relative ever had diabetes), LOSEWT (Told to reduce calories or participate in weight loss program), and CHLMDEV2 (Ever prescribed medicine to lower cholesterol) were chosen as the most significant variables in this model. The purest leaf was shown to be the *No or Missing* leaf stemming from variable HYPMDEV2.

GENERALIZATION

A comparison of all models (Table A 5 in the Appendix) indicates that the regression models performed better than the tree models overall. Though all models performed well and were robust, the performance of the regression models may be inflated due to lack of consideration for the survey design. Tree models provide a different analysis approach to the logistic regression currently employed by healthcare researchers with the benefits of requiring fewer assumptions and less preparation. The analyses presented in this paper provide tentative evidence that data mining methods are a viable alternative to regression for classification of health data.

SUGGESTIONS FOR FUTURE

There are some issues in using the NHIS for our exploration; because the survey has a complex design and is not a simple random sample, we risk producing incorrect standard error estimates using traditional regression techniques. In general, the standard errors will be too small, so we are more likely to exclude important variables from our final model. While this is a concern, we decided to continue with our analysis since any important predictors that were identified could be assumed to be valid.

In future analyses, this issue with survey design can be overcome by accounting for the weights of each subject and utilizing Proc Survey Select in SAS. This will allow for accurate estimates of standard errors and may result in regression models with more significant predictors. Additionally, future research could explore the possibility of using random forest as a variable reduction method for health data after first pre-processing to eliminate invalid inputs. This design could improve efficiency while quickly identifying potential significant predictors of the target disease.

CONCLUSION

The results indicate that random forest is an ineffective replacement to established methods of pre-processing and variable reduction in health research when used alone. However, it could be useful in combination with a theory-based approach. The models produced by decision trees are comparable to logistic regression models. The disadvantage in using decision trees is that odds and risk ratios cannot be estimated for the important predictors. We believe that in future research, traditional and algorithmic techniques could be combined strategically to produce quality results efficiently.

REFERENCES

Centers for Disease Control and Prevention. 2018. “Diabetes home.” Accessed December 4, 2018. https://www.cdc.gov/diabetes/basics/type2.html.

Centers for Disease Control and Prevention. 2017. “National Diabetes Statistics Report, 2017.” Accessed January 22, 2019. https://www.cdc.gov/diabetes/data/statistics/statistics-report.html.

MayoClinic. 2018. “Type 2 diabetes: Symptoms and causes.” Accessed December 4, 2018. https://www.mayoclinic.org/diseases-conditions/type-2-diabetes/symptoms-causes/syc-20351193.

National Center for Health Statistics. 2017. “National Health Interview Survey." Public-use data file and documentation. Accessed January 22, 2019. https://www.cdc.gov/nchs/nhis/data-questionnaires-documentation.htm.

ACKNOWLEDGMENTS

The authors of this paper would like to acknowledge our faculty advisor, Dr. Sherry Ni.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Audrey Whittle
Team Lead
awhittl4@students.kennesaw.edu

Sherry Ni
Faculty Advisor
sni@kennesaw.edu

DISCLAIMER

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

APPENDIX

Table A 1: Variables Used in Literature-Based Analyses

Description	Name
Freq drank alcohol: Days per week	ALK12MWK
Body Mass Index (BMI)	BMI
CDC standard for legal marital status	CDCMSTAT
Ever told you had high cholesterol	CHLEV
Ever prescribed medicine to lower cholesterol	CHLMDEV2
Currently increasing physical activity	DBHVPAN
Told to increase physical activity, past 12 m	DBHVPAY
Presence or absence of Type II Diabetes (created)	dib_2
Blood relative ever had diabetes	DIBREL
Highest level of school completed	EDUC1
Total earnings last year	ERNYR_P
Obtaining affordable coverage	FCOVCONF
Any functional limitation, all conditions	FLA1AR
Education of adult with highest education in family	FM_EDUC1
Family type	FM_TYPE
Could not afford to eat balanced meals	FSBALANC
Worried food would run out before got money to buy more	FSRUNOUT
Geographic place of birth recode	GEOBRTH
Amount family spent for medical care	HCSPFYR
How long since last had health coverage	HILAST2
Ever been told you have hypertension	HYPEV
Ever prescribed medicine for high blood pressure	HYPMDEV2
Now taking prescribed medicine for high blood pressure	HYPMED2
Total combined family income (grouped)	INCGRP5
Been told to reduce fat/calories / Participating in weight loss program (created)	losewt
Currently reducing fat/calories / Participating in weight loss program (created)	losewt_c
Health Insurance coverage status	NOTCOV
Reported health status	PHSTAT
Sex	SEX
Ever smoked 100 cigarettes	SMKEV
Smoke freq: every day/some days/not at all	SMKNOW
Ever been told you had a stroke	STREV

Figure A 1: Model Selection from Optimal BMI

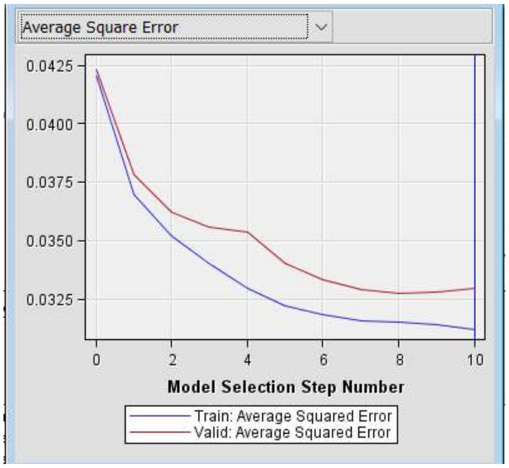


Table A 2: Full Regression Important Predictors (Chosen from Literature)

Description	Name
Total number of office visits, past 12 m	AHCNOYR2
Weight without shoes (pounds)	AWEIGHTP
Body Mass Index	BMI
Imputed_Weight problem causes difficulty with activity	IMP_AFLHCA18
Imputed_Fracture, bone/joint injury causes difficulty with activity	IMP_AFLHCA5
Imputed_Seen/talked to foot doctor, past 12 m	IMP_AHCSYR3
Imputed_Total height in inches	IMP_AHEIGHT
Impute_Freq drank alcohol past year: Time unit	IMP_ALC12MTP
Imputed_Cholesterol checked by doctor/nurse/health professional, past 12 m	IMP_APSCHCHK
Imputed_Doctor/health professional talked to you about diet, past 12 m	IMP_APSDIET
Taking low-dose aspirin on own	IMP_ASPONOWN
Use any adaptive devices such as magnifiers, talking materials	IMP_AVISDEV
Imputed_Ever been told you have hypertension	IMP_HYPEV

Table A 3: Optimal BMI Regression Important Predictors (Chosen from Literature)

Description	Name
Imputed_Ever been told you have hypertension	IMP_HYPEV
Imputed_Blood relative ever had diabetes	IMP_DIBREL
Imputed_Told to reduce calories or participate in weight loss program	IMP_LOSEWT
Imputed_Freq drank alcohol: Days per week	IMP_ALC12MWK
Imputed_Ever told you had high cholesterol	IMP_CHLEV
Transformed_Optimum_BMI	OPT_BMI
Imputed_Any functional limitation, all conditions	IMP_FLA1AR
Imputed_Told to increase physical activity, past 12 m	IMP_DVHPAY
Imputed_Cov stat as used in Health United States	IMP_NOTCOV
Imputed_Total earnings last year	IMP_ERNYR_P

Figure A 2: Enterprise Miner Diagram including Random Forest Variable Reduction and Analysis Using Literature Dataset

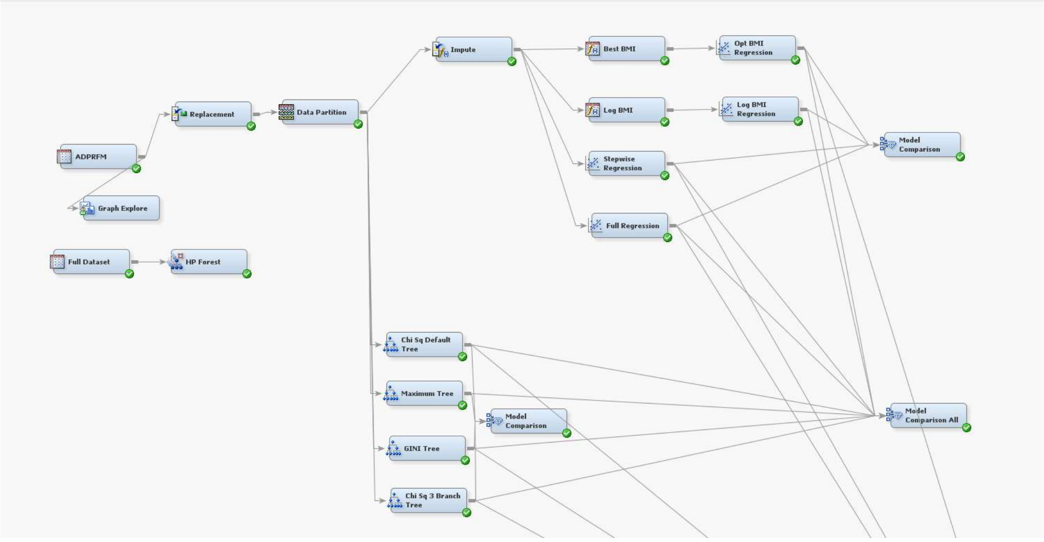


Figure A 3: Chi-Squared Default Tree

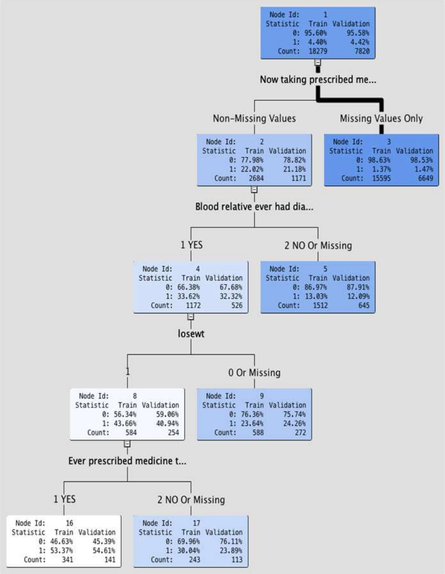
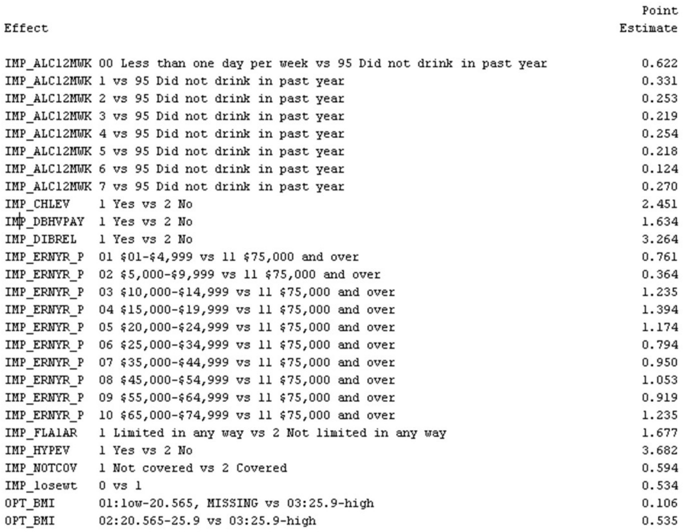


Table A 4: Odds Ratio Estimates for Optimal BMI Stepwise Regression



```
                                                                                Point
Effect                                                                       Estimate

IMP_ALC12MWK 00 Less than one day per week vs 95 Did not drink in past year      0.622
IMP_ALC12MWK 1 vs 95 Did not drink in past year                                  0.331
IMP_ALC12MWK 2 vs 95 Did not drink in past year                                  0.253
IMP_ALC12MWK 3 vs 95 Did not drink in past year                                  0.219
IMP_ALC12MWK 4 vs 95 Did not drink in past year                                  0.254
IMP_ALC12MWK 5 vs 95 Did not drink in past year                                  0.218
IMP_ALC12MWK 6 vs 95 Did not drink in past year                                  0.124
IMP_ALC12MWK 7 vs 95 Did not drink in past year                                  0.270
IMP_CHLEV    1 Yes vs 2 No                                                       2.451
IMP_DBHVPAY  1 Yes vs 2 No                                                       1.634
IMP_DIBREL   1 Yes vs 2 No                                                       3.264
IMP_ERNYR_P  01 $01-$4,999 vs 11 $75,000 and over                                0.761
IMP_ERNYR_P  02 $5,000-$9,999 vs 11 $75,000 and over                             0.364
IMP_ERNYR_P  03 $10,000-$14,999 vs 11 $75,000 and over                           1.235
IMP_ERNYR_P  04 $15,000-$19,999 vs 11 $75,000 and over                           1.394
IMP_ERNYR_P  05 $20,000-$24,999 vs 11 $75,000 and over                           1.174
IMP_ERNYR_P  06 $25,000-$34,999 vs 11 $75,000 and over                           0.794
IMP_ERNYR_P  07 $35,000-$44,999 vs 11 $75,000 and over                           0.950
IMP_ERNYR_P  08 $45,000-$54,999 vs 11 $75,000 and over                           1.053
IMP_ERNYR_P  09 $55,000-$64,999 vs 11 $75,000 and over                           0.919
IMP_ERNYR_P  10 $65,000-$74,999 vs 11 $75,000 and over                           1.235
IMP_FLA1AR   1 Limited in any way vs 2 Not limited in any way                    1.677
IMP_HYPEV    1 Yes vs 2 No                                                       3.682
IMP_NOTCOV   1 Not covered vs 2 Covered                                          0.594
IMP_losewt   0 vs 1                                                              0.534
OPT_BMI      01:low-20.565, MISSING vs 03:25.9-high                              0.106
OPT_BMI      02:20.565-25.9 vs 03:25.9-high                                      0.535
```

Table A 5: Model Comparison Validation Statistics (Chosen from Literature)

Model	ASE	Misclassification
Optimal BMI Regression	0.032972	0.042839
Log BMI Regression	0.033406	0.042327
Stepwise Regression	0.033603	0.043478
Maximum Tree	0.034319	0.046931
Full Regression	0.03456	0.04399
Chi-Squared Default Tree	0.034671	0.042583
GINI Tree	0.034997	0.042711
Chi-Squared 3 Branch Tree	0.035727	0.044118