# TWO DATA MINING APPLICATIONS FOR PREDICTING PRE-DIABETES

A Dissertation
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Guangjing You

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Industrial Engineering & Management

November 2015

Fargo, North Dakota

North Dakota State University
Graduate School

**Title**

Two data mining applications for predicting pre-diabetes

**By**

Guangjing You

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

**MASTER OF SCIENCE**

SUPERVISORY COMMITTEE:

Kambiz Farahmand
Chair

Jing Shi

Yarong Yang

Approved:

11/20/2015
Date

Om Prakash Yadav
Department Chair

# ABSTRACT

In this study, the performance of Logistic Regression and Decision Tree modeling is compared by using SAS Enterprise Miner for predicting pre-diabetes in US population by using several of the common factors from the type 2 diabetes screening criteria. From 17 variables of NHANES' three sets of dataset, a total of 13 risk factors were selected as predictors of pre-diabetes. A comparison of two data mining methodology showed that Decision Tree has a higher ROC index than Logistic Regression modeling. All ROC indexes for two models were greater than 77% indicating both methods present a good prediction for pre-diabetes. The predictive accuracy of the two models was greater than 72% on the whole dataset. Decision tree modeling also resulted in higher accuracy and sensitivity values than Logistic Regression modeling. Taken as a whole, the results of comparison indicated Decision Tree modeling is a better indicator to predict pre-diabetes.

***Keywords*：*Diabetes, pre-diabetes, logistic regression, decision tree, risk factor.***

## TABLE OF CONTENTS

# LIST OF TABLES

## LIST OF FIGURES

## LIST OF ABBREVIATIONS

| | |
|---|---|
| AACE | American Association of Clinical Endocrinologists Medical |
| ADA | American Diabetes Association |
| ANNs | Artificial Neural Networks |
| AOC | Area of Curve |
| ARIC | Atherosclerosis Risk in Communities |
| AUC | Area Under Curve |
| AUSDIAB | Australian Diabetes Obesity and Lifestyle Study |
| BLR | Binary Logistic Regression |
| BMI | Body Mass Index |
| BPSYS | Systolic Blood Pressure |
| CART/CRT | Classification and Regression Tree |
| CDC | Centers for Disease Control |
| CHAIN | Chi-Square Automatic Interaction Detection |
| CMI | Comorbidity Index |
| CMS | Centers for Medicate and Medicaid Services |
| CS | Chi-Square |
| CURES | Chennai Urban Rural Epidemiology Study |
| CVD/CV | Cardiovascular disease |
| DBP | Diastolic Blood Pressure |
| DT | Decision Tree |
| ER | Emergency Department Visits |
| ESRD | End-Stage Renal Disease |

| | |
|---|---|
| FBS | Fasting Blood Glucose |
| FFA | Free Fatty Acid |
| FP | False Positive |
| FPG | Fasting Plasma Glucose |
| FN | False Negative |
| GDM | Gestational Diabetes Mellitus |
| GLU | Blood Glucose |
| GT | Gamma-glutamyl Transferase |
| HbA$_{1C}$ | Glycated Hemoglobin |
| HDL | High-Density Lipoprotein |
| ICMR-INDIAB | Indian Council of Medical Research-Indian Diabetes |
| ICSI | Institute for Clinical Systems Improvement |
| IDF | International Diabetes Federation |
| IDRS | Indian Diabetes Risk Score |
| ID3 | Iterative Dichotomister |
| IFG | Impaired Fasting Glucose |
| IGT | Impaired Glucose Tolerance |
| HIS | Indian Health Service |
| K-NN | K-nearest Neighbor |
| LDL | Low Density Lipoprotein |
| LR | Logistic Regression |
| MLR | Multinomial Logistic Regression |
| NASH | Non-alcoholic Steatohepatitis |

| | |
|---|---|
| NHANES | National Health and Nutrition Examination Survey |
| NIDDM | Non-Insulin-Dependent Diabetes Mellitus |
| OGTT | Oral Glucose Tolerance Test |
| OV | Office Visits |
| PCOS | Polycystic Ovarian Syndrome |
| PDM | Pre- Diabetes Mellitus |
| PIDD | Pima Indian Diabetes Dataset |
| PLS-DA | Partial Least Square Discriminant Analysis |
| PLS-LDA | Partial Least Square-Linear Discriminant Analysis |
| PNN | Prototype Nearest Neighbor |
| ROC | Receiver Operating Characteristic |
| SAHS | San Antonio Heart Study |
| SVM | Support Vector Machines |
| TG | Triglyceride |
| TN | True Negative |
| TP | True Positive |
| TLGS | Tehran Lipid and Glucose Study |
| T2DM | Type 2 Diabetes Mellitus |
| VA/DoD | The Department of Veterans Affairs and the Department of defense |
| UCI | University of California |
| UCL | University College London |
| USPSTF | U.S. Preventive Services Task Force |
| WC | Waist Circumference |

| | |
|---|---|
| WEKA | Waikato Environment for Knowledge Analysis |
| WHO | World Health Organization |
| WHR | Waist-hip Ratio |

## INTRODUCTION

Diabetes is the fastest growing chronic disease in the world. In the United States, according to the Centers for Disease Control and Prevention (CDC) Diabetes report (2014), there were more than twenty-nine million people or 9.3% the U.S. population who had diabetes in 2012. From which, twenty-one million were diagnosed, and 8.1 million with diabetes were undiagnosed. Diabetes is a common chronic disease, which occurs when the pancreas does not produce enough insulin, or when the body cannot effectively use the insulin it produces. This leads to an increased concentration of glucose in the blood.

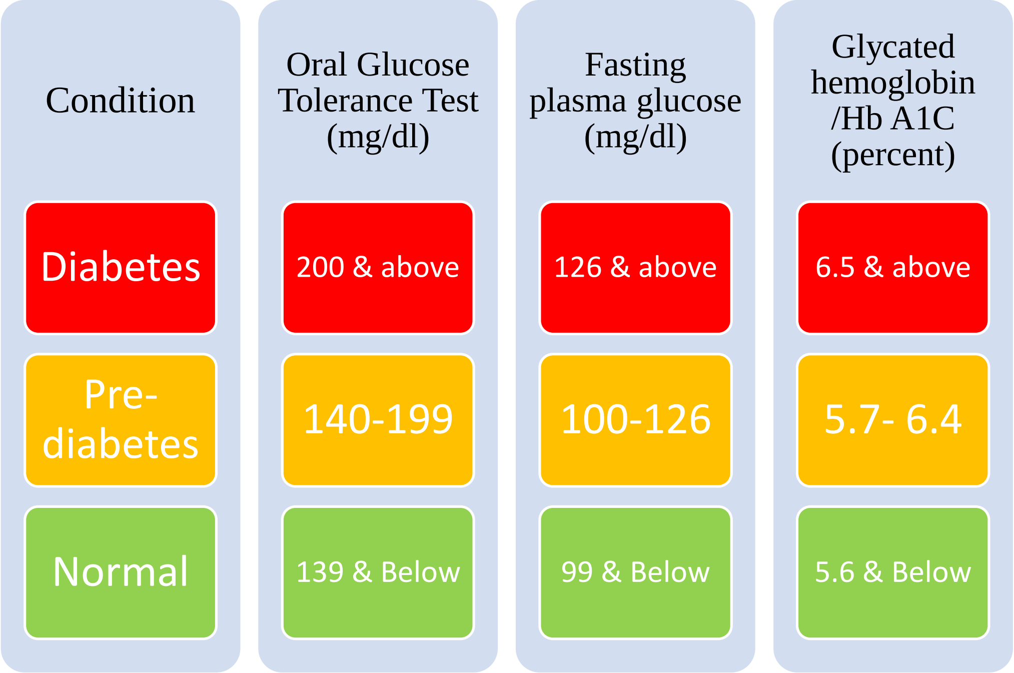
There are two main types of diabetes:

**Type 1 diabetes mellitus**: when most or all insulin producing beta cells in the pancreas have been destroyed, so there is a severe lack of insulin in the body.

**Type 2 diabetes mellitus**: when the pancreas still produces insulin but body cannot use insulin properly.

Type 1 diabetes often happens in children and adolescents. However, type 2 diabetes is the most common form of diabetes. In adults, type 2 diabetes accounts for about 90% to 95% of all diagnosed cases of diabetes. Patients with type 2 diabetes require long-term health management plans (ADA, 2013). According to the statistics of CDC (2014), $ 245 million were used for the total costs and lost work and wages for people with diagnosed diabetes. From these numbers, one can see that type 2 diabetes has significant financial impact. In this study, the assumption is the type 2 diabetes if a particular kind of diabetes is not mentioned.

Data from the National Diabetes Statistics report (2014), 86 million American age greater and equal 20 years had pre-diabetes in 2012. It is mean that more than 1 out of 3 adults have pre-diabetes. A person with pre-diabetes who has a blood sugar level higher than normal, but not high enough for a diagnosis of diabetes. At this stage, patients may be considered to have pre-diabetes. Often, they have Impaired Glucose Tolerance (2-hour OGTT values between 140 and 199 mg/dl), IFG (FPG between 100 and 126 mg/dl), or an A1C of 5.7–6.4%. Blood test level was shown in figure 1 as follow:



| Condition | Oral Glucose Tolerance Test (mg/dl) | Fasting plasma glucose (mg/dl) | Glycated hemoglobin /Hb A1C (percent) |
|---|---|---|---|
| Diabetes | 200 & above | 126 & above | 6.5 & above |
| Pre-diabetes | 140-199 | 100-126 | 5.7- 6.4 |
| Normal | 139 & Below | 99 & Below | 5.6 & Below |

Figure 1. Diabetes and pre-diabetes diagnosis criteria (WHO, 2014)

These individuals are at higher risk for developing type 2 diabetes and other serious health problems, including heart disease, and stroke. Without lifestyle changes to improve their health, 15% to 30% of people with pre-diabetes will develop type 2 diabetes within five years. But, 9 out of 10 adults do not know who have pre-diabetes. Therefore, identifying individuals at high risk for pre-diabetes is an urgent need.

Effective diabetes screening could improve people’s quality of life and reduce the cost of health care system. Screening should be sequential, not a one-time event. However, when and how to screen asymptomatic individuals is a complex decision. In order to group the patients who have the same condition and make a screening schedule for same group. Based on these requirement, how to accurately predict and diagnose diabetes or pre-diabetes are vital for healthcare system.

The objective of this study is compare qualitative models in data mining for pre-diabetes. Data mining is the processing of analyzing large-scale data in order to descript, understand and predict trends in the data. This is the reason why data mining technologies were used to analyze the constantly increasing volumes of data for diabetes.

# DATA DESCRIPTION

## Data Collection

Data of National Health and Nutrition Examination Survey (NHANES) in website of Centers for Disease Control and Prevention (CDC) were released to the public in 2-year cycles, all participants were interviewed from 15 different country locations selected from a sampling frame that included all 50 states in U.S. and District of Columbia. Following this method, data was selected from 1999 to present. In this study, the recent three 2-year cycles data were chose and analyzed. From which, data of 2011-2012 was used and eliminated all participants with any of the "missing", "refused", and "don't know" among total 9756 participants. Each of data set, represent the two-year data release cycle number. In the original data, there are two dependent variables, first is "Ever doctors told you have pre-diabetes" and the other one is "Doctor told you that you have diabetes". These two variables were combined together, and if anyone was told "Borderline"-on the verge of diabetes, they will be considered the pre-diabetes patients. Ultimately, there were 4312 survey participants who had the integrated information about what was needed. NHANES 2007-2008 demographics data had a total 10,149 participants and 2009-2010 demographics data have total 10,537 participants. For 2007-2008 and 2009-2010 NHANES data, the same procedure was done as data of 2011-2012. The final total observations are 2985 and 3357. In final version, all diabetes patients had been deleted. Therefore, these data were applied to analyze which factors would cause pre-diabetes and to find how these factors predict pre-diabetes.

## Variables

To detect people with diabetes in early stage, based on the study of Tan et al (2014) comparison was made eight guidelines: 1. The American Diabetes Association (ADA, 2014); 2. American Association of Clinical Endocrinologists medical (AACE, 2013); 3. The World Health Organization (WHO, 2006); 4. The Indian Health Service (IHS, 2011); 5. Centers for Medicate and Medicaid Services (CMS); 6. The Department of Veterans Affairs and the Department of defense (VA/DoD, 2010); 7. The Institute for Clinical Systems Improvement (ICSI, 2012); 8. The International Diabetes Federation (IDF, 2006). All of them were utilized to care for patients with type 2 diabetes. Individuals might have some early signs of the disease but do not exactly meet the criteria for diagnosis. Several of these guidelines considered patients to be at high risk for undiagnosed type 2 diabetes if they had 1 or more of the following diabetes risk factors, For example, the ADA (2014) guidelines recommendations for screening for type 2 diabetes: a family history of diabetes (defined as diabetes in a parent, brother, or sister, or some combination thereof), hypertension, cardiovascular disease (myocardial infarction, heart failure, atrial fibrillation, stroke, peripheral vascular disease), lipid metabolism disorders, obesity (Body Mass Index, BMI ≥25 kg/$m^2$), age ≥ 45 years, and a history of gestational diabetes mellitus. Five comparison tables were made to see what differences are present between these guidelines. Risk factors will be considered based on these criteria to predict pre-diabetes. The one of five comparison tables of type 2 diabetes screening criteria shown in table 1.

Table 1. Type 2 diabetes screening criteria

| | ADA | AACE | WHO | IHS | CMS | VA/DoD | ICSI | IDF |
|---|---|---|---|---|---|---|---|---|
| Overweight (BMI ≥ 25 kg/$m^2$) | ★ | | ★ | ★ | | ★ | ★ | ★ |
| First-degree relative with diabetes | ★ | ★ | ★ | ★ | ★ | ★ | ★ | ★ |
| Women who delivered a baby weighing > 9 lb | ★ | ★ | ★ | ★ | ★ | ★ | ★ | ★ |
| Hypertension (>140/90 mmHg) | ★ | ★ | ★ | ★ | ★ | ★ | ★ | ★ |
| HDL cholesterol <35 mg/dl or triglyceride level >250 mg/dl<br>(*VA/DoD HDL cholesterol <40 mg/dl)* | ★ | ★ | ★ | ★ | ★ | ★ | ★ | ★ |
| IGT or IFG on previous testing/Pre-diabetes | ★ | ★ | ★ | | ★ | ★ | ★ | ★ |
| History of Gestational Diabetes Mellitus (GDM) | ★ | | ★ | ★ | | ★ | ★ | ★ |
| Polycystic ovarian syndrome (PCOS) | ★ | ★ | ★ | ★ | | ★ | ★ | ★ |
| Acanthosis Nigricans | ★ | | ★ | ★ | | ★ | ★ | ★ |
| Other clinical conditions associated with insulin resistance (e.g., Severe obesity, PCOS, Acanthosis Nigricans) | | | ★ | | | | ★ | |
| History of CVD | ★ | ★ | | ★ | | | | ★ |
| High-risk race/ethnicity | ★ | ★ | ★ | | | ★ | ★ | ★ |
| Physical inactivity | ★ | ★ | | | | ★ | | ★ |
| Age ≥45 years<br>(WHO Age≥35 years<br>CMS Age ≥65 years) | ★ | | ★ | | ★ | ★ | | ★ |
| History of Vascular Disease | | | ★ | | | ★ | | |
| Antipsychotic therapy for schizophrenia or severe bipolar disease | | ★ | | | | ★ | | |
| Abdominal obesity | | | | | | ★ | | |

Table 1. Type 2 diabetes screening criteria (continued)

| | ADA | AACE | WHO | IHS | CMS | VA/DoD | ICSI | IDF |
|---|---|---|---|---|---|---|---|---|
| Non-alcoholic steatohepatitis (NASH) | | | | | | ★ | | |
| Dyslipidemia | | | | | | ★ | | |
| Cardiovascular Risk Factors | | | | | | | ★ | |
| A1C ≥ 5.7% | ★ | | | | | | | ★ |
| Those with prediabetes should be tested annually | ★ | | | | | ★<br>Repeat screening every 1-3 year | | ★ |
| If results are normal, repeat test every 3 years | ★ | | ★ | | | | | ★ |

In the present study, the following dependent variable coding is used: ‘No Pre-diabetes’=0, and ‘Pre-diabetes’=1. According to the data from the National Health and Nutrition Examination Survey. 17 variables were chose, including gender (male/female), age, race/ethnicity (six levels), served active duty in US armed forces (yes/no), born of U.S. (yes/no), citizenship status (yes/no), education level (five levels), marital status (six levels), total number of people in the family (six levels), annual family income (fourteen levels), ever told you have health risk for diabetes (yes/no), smoked (two levels), physical activity (three levels), high cholesterol level (yes/no), hypertension status (three levels), diet (five levels), and BMI. Age and body mass index are continuous variables, while the other 15 factors are categorical variables. These two continuous variables are explain below:

1) **Age** is the most important factor for the risk of type 2 diabetes, because the incidence of diabetes increases steeply with age of the fifty articles total in table 3, there are forty-one studies (82%) mentioned age as predictor in their regression model. Effective pre-

diabetes screening can reduce the incidence of diabetes and cost of treatment. For instance, Chung et al. (2014) reconsidered the age thresholds of screening using cross-sectional analysis of a nationally representative sample from the National Health and Nutrition Examination Survey, 2007–2010. This study examined the optimal age for opportunistic universal screening, compared to different screening, methods recommended by the U.S. Preventive Services Task Force (USPSTF) and American Diabetes Association (ADA) guidelines.

2) **Body Mass Index (BMI)** is another important factor. The equation of BMI =Weight (kg)/Height $(\mathrm{m})^2$= 703*Weight (lb)/Height $(\mathrm{inch})^2$, so the BMI index will increase with weight. Tayek's (2002) showed, weight loss alone will not cure the type 2 diabetes, but it could reduce the incidence of type 2 diabetes. As mentioned, following the six guidelines, adults should be evaluated for type 2 diabetes if they are overweight (BMI ≥ 25 kg/m2) and have one or more of the factors list in table 1. In addition, waist circumference is another form of BMI could be a factor. All of the fifty articles total in table 3, there are twenty-five papers (50%) mentioned BMI as predictor in model, and fifteen studies (30%) mentioned waist circumference as factor in their regression model.

Details of the 17 variables' are summarized in table 2.

Table 2. Explanatory variable explanation

| **Factor** | **Variable** | **Variable assignment rules** |
|---|---|---|
| Age | $x_1$ | 20 years of age or older |
| Body mass index | $x_2$ | Body mass index calculated by the weight in kilograms divided by the square of the height in meters |
| Gender | $x_3$ | Gender of patient (Male=**1**; Female=**2**) |
| Race | $x_4$ | Race/ethnicity (Mexican American=**1**; Other Hispanic=**2**; Non-Hispanic White=**3**; Non-Hispanic Black=**4**; Other Race-Including Multi-Racial=**5**) |
| US armed forces | $x_5$ | Served active duty in US armed forces (Yes=**1**; No=**2**) |
| Born of U.S. | $x_6$ | Born of U.S. (Born in 50 US States or Washington, DC=**1**; Others=**2**) |
| Citizenship status | $x_7$ | Citizenship status (Citizen by birth or naturalization=**1**; Not a citizen if the US=**2**) |
| Education level | $x_8$ | Education level (Less than 9th grade=**1**; 9-11th grade (Includes 12th grade with no diploma)=**2**; High school graduate/GED or equivalent=**3**; Some college or AA degree=**4**; College graduate or above=**5**) |
| Marital status | $x_9$ | Marital status (Married=**1**; Widowed=**2**; Divorced=**3**; Separated=**4**; Never married=**5**; Living with partner=**6**) |
| Total number of people in family | $x_{10}$ | Total number of people in family (1 People=**1**; 2 People=**2**; 3 People=**3**; 4 People=**4**; 5 People=**5**; 6 People=**6**; 7 or more people in the family=**7**) |
| Annual family income | $x_{11}$ | Annual family income ($ 0 to $ 4,999=**1**; $ 5,000 to $ 9,999=**2**; $10,000 to $14,999=**3**; $15,000 to $19,999=**4**; $20,000 to $24,999=**5**; $25,000 to $34,999=**6**; $35,000 to $44,999=**7**; $45,000 to $54,999=**8**; $55,000 to $64,999=**9**; $65,000 to $74,999=**10**; $20,000 and Over=**12**; Under $20,000=**13**; $75,000 to $99,999=**14**; $100,000 and Over=**15**) |
| Ever told you have health risk for diabetes | $x_{12}$ | Ever been told by a doctor or other health professional that you have health risk for diabetes (Yes=**1**; No=**2**) |
| Smoked | $x_{13}$ | Smoked at least 100 cigarettes in life (Yes=**1**; No=**2**) |
| Physical activity | $x_{14}$ | Physically active moderate recreational activities ( Less than 3 days Moderate activities a week=**1**; More than 3 days Moderate activities a week=**2**) |
| High cholesterol level | $x_{15}$ | Doctor told you have High cholesterol level (Yes=**1**; No=**2**) |
| Hypertension status | $x_{16}$ | Hypertension status (High Blood Pressure=**1**; Borderline Hypertension=**2**; No=**3**) |
| Diet | $x_{17}$ | How healthy is the diet (Excellent=**1**; Very Good=**2**; Good=**3**; Fair=**4**; Poor=**5**) |

Age ($\boldsymbol{x_1}$), gender ($\boldsymbol{x_3}$), race ($\boldsymbol{x_4}$), served in US armed forces ($\boldsymbol{x_5}$), born of U.S. ($\boldsymbol{x_6}$), citizenship status ($\boldsymbol{x_7}$), education level ($\boldsymbol{x_8}$), marital status ($\boldsymbol{x_9}$), total number of people in family ($\boldsymbol{x_{10}}$) and annual family income ($\boldsymbol{x_{11}}$) were selected from demographics data of NHANES. The remaining seven variables are from questionnaire data of NHANES. Due to the use of logistic regression and decision tree both continuous and categorical variables be processed. Therefore, it is not necessary to change the form of these variables.

## Regression Methods Literature Review

### Logistic Regression

There are several papers focusing on the selection of factors for type 2 diabetes or pre-diabetes. In primary care clinical sciences, logistic regression has been used to investigate the factors of diabetes. Bonora et al. (2004) investigated 1,000 random white people of Bruneck, Italy between the ages of 40 to 79 years. They used logistic regression modeling to suggest age, body mass index (BMI), hypertension, dyslipidemia, IFG (Impaired fasting glucose) and IGT (Impaired glucose tolerance) are significant. Meng et al. (2013) compared the three data mining models for predicting diabetes or pre-diabetes using 12 risk factors: logistic regression, artificial neural networks (ANNs) and decision tree. These data total included 1487 participants from two communities in Guangzhou, China. There are twelve variables in logistic regression: age, family history of diabetes, marital status, educational level, work stress, duration of sleep, physical activity, gender, eating fish, drinking coffee, preference for salty food, and BMI. Meng et al. (2013) also used chi-square test to choose the risk factors; the result is same with logistic regression selection but the importance of variables are different.

Borrell et al. (2007) selected 7,231 U.S. adults aged 20 years or older without diabetes and not pregnant. Logistic regression modeling was used to predict the probability of the individuals having undiagnosed diabetes, using the factors were: age, sex, ethnicity, family history of diabetes, self-reported hypertension, hypercholesterolaemia, and periodontal disease. Gray et al. (2010) used the data on 6,186 subjects aged 40-75 years from UK. Age, ethnicity, sex, first-degree family history of diabetes, antihypertensive therapy or history of hypertension, waist circumference, and BMI were included in final logistic regression model. Similar in method, Griffin et al. (2000) based on 1,077 British individuals aged 40-64 years without diabetes. Age, sex, prescribed antihypertensive medication, prescribed steroids, BMI, family history of diabetes, and smoking status were calculated to the diabetes risk score. Schmidt et al. (2005) produced risk functions for detecting incident diabetes on a randomly selected half of the sample using logistic regression models. Factors were considered includes: age, sex, ethnicity, parental history of diabetes, use of medication for hypertension, height, various measures of obesity (waist, weight, BMI, waist-to-hip ratio, each investigated one at a time), systolic blood pressure, fasting glucose, HDL cholesterol, triglycerides, and fasting insulin.

A total of 562 participates in Kuwait agreed to be tested. These data were entered into a forward logistic regression modeling, some important factors were identified: age, waist circumference, blood pressure medication, sibling with diabetes (Al Khalaf et al, 2010). Al-Lawati et al. (2007) investigated 1,432 subjects without pregnant in Oman. Backward stepwise logistic regression modeling was used to obtain age, waist circumference, BMI, family history of diabetes, hypertension and coefficients of these factors. Based on 1,016 participants aged 55-74 years in the Netherlands, Baan et al. (1999) used stepwise logistic regression to obtain four factors: age, sex, use of antihypertensive medication, obesity (BMI ≥ 30).

## Multivariate Logistic Regression

In Statistics, multivariate analysis as a kind of statistical modeling that have 2 or more dependent or outcome variables (Van Belle, 2004), and multivariable analysis as a kind of statistical modeling in which there are multiple independent or response variables (Katz, 2005).

Lindström and Tuomilehto (2003) followed a random population sample of the ages of 35 to 64 years old with no antidiabetic drug treatment. The data from the Nation Population Register

Survey (N=4746) in 1987 and the FINRISK Studies Survey (N=4615) in 1992. Multivariate logistic regression modeling coefficient were used to develop a concise model, which assign a score for each of the following variable: age (45-54, 55-64), BMI, waist circumference, use of blood pressure medication, and history of high blood glucose. Gao et al. (2009) indicated that age, sex, BMI, waist circumference, family history of diabetes play the most significant role by applying a multivariate logistic regression modeling. They based on 3,094 Mauritian Indians between the ages of 20 to 65 years without diabetes during 11 years follow up. After a year later, Gao et al. (2010) used two years survey (2002, n=1986) and (2006, n=4336) from Chinese adults between the ages of 20 to 74 years. Age, waist circumference, family history of diabetes were significance predictors in the multivariate logistic regression model. According to 1,032 Egyptian subjects with no diabetes, the multivariate logistic regression equation included age, random plasma glucose, postprandial time, sex, BMI as predictors for undiagnosed diabetes (Tabaei and Herman, 2002). A total of 6,237 individuals in Canary Islands were applied to test the screening programs. The three predictors were: age, waist/height ratio, and family history of diabetes for men. For women, the four predictors were age, waist/height ratio, family history of diabetes, and gestational diabetes (Cabrera de León et al, 2008).

## Multiple Logistic Regression and Multivariable Logistic Regression

As early as 1999, Burke, et al (1999) followed participants in the San Antonio Heart Study (SAHS) for 7-8 years. The SAHS predicting model was created by using multiple logistic regression models. They used odds ratios for various factors to identify which factors will develop type-2 diabetes by estimating from these logistic regression models. Age, sex, ethnic group, neighborhood, and date of enrollment were significance predictors of diabetes. Carlsson et al. (2004) tested age, BMI, physical activity and alcohol consumption in the multiple logistic regression model, based on a prospective study of 11 years followed up of the incidence of diabetes in the Nord- Trondelag Health Survey. The result indicated that smoking influences the immune system in human diabetes. In the same way, Waki et al. (2005) investigated 12,913 men and 15,980 women aged 40-59 years old in the Japan Public Health Center-based study on cancer and cardiovascular disease. During 10 years of follow-up, their results show high alcohol consumption was positively associated with the incidence of diabetes in lean Japanese men (BMI$\leq$ 22 kg/$m^2$).

Chen et al. (2010) studied 6,060 Australians in a diabetes obesity and lifestyle study (AUSDIAB). These participants aged 25 years or older and did not have diagnosed diabetes by follow- up after 5 years. The final prediction model included nine factors: age, sex, ethnicity, smoking, parental history of diabetes, history of high blood glucose level, use of antihypertensive medications, physical inactivity and waist circumference. Stern et al. (2002) analyzed 5,158 participants between the ages of 25 to 64 years and not pregnant in San Antonio Heart Study. Multiple logistic regression modeling was used to indicate that age, sex, race, fasting plasma glucose (FPG), systolic blood pressure, HDL cholesterol, BMI, family history of diabetes were considered as factors. From NHANES (National Health and Nutrition Examination Survey), a total of 21,620 in U.S. aged 45 years or older were tested. Bang et al. (2004) used multiple logistic regression to determine age, sex, family history of diabetes, history of hypertension, obesity (BMI or waist circumference), and physical activity as participant characteristics associated with undiagnosed diabetes. Ko et al. (2010) derived 12,448 Hong Kong Chinese without diabetes. Age, sex, BMI, hypertension, dyslipidaemia, family history of diabetes, gestational diabetes were used to calculate the risk score of diabetes in the multiple logistic regression modeling. Among 26,001 subjects who derived from the Chennai Urban Rural Epidemiology Study (CURES) in India, age, waist circumference, physical activity, and family history of diabetes were identified as the four factors as used to develop the Indian Diabetes Risk Score (IDRS) based on multiple logistic regression analysis (Mohan et al, 2005). Ramachandran et al. (2005) tested 10,003 participants aged 20 years or older in India. Age, family history of diabetes, BMI, waist circumference, physical activity were identified as significant factors and applied for multiple logistic regression analysis.

A total of 2,364 Caucasian subjects were studied, who between the ages of 50 to 74 years old, who did not know if they had diabetes. Ruige et al. (1997) pointed out frequent thirst, pain during walking with need to slow down, shortness of breath when walking, age, sex, BMI, obesity (men), family history of diabetes, use of antihypertensive drugs, and reluctance to use bicycle for transportation were significant in backward stepwise multiple logistic regression model. Stepwise multiple logistic regression was applied on the optimal risk score for occurrence of Diabetes Mellitus among Hindustani Surinamese (n=336), African Surinamese (n=593), and Dutch (n=486). Age, BMI, waist circumference, resting heart rate, first-degree relative with diabetes, hypertension, history of CVD, ethnicity were included in the risk score (Bindraban et al, 2008). Glümer et al. (2004) studied in 6,784 individuals between the ages of 30 to 60 years in Denmark. Stepwise backward logistic regression were developed to calculate the diabetes risk score. Age, BMI, sex, known hypertension, physical activity, and family history of diabetes were included in final risk score.

A total of 429 Thai adults without diabetes were derived by stepwise multiple logistic regression to determine the risk equation. Age, BMI, and history of hypertension were significant in model (Keesukphan et al, 2007). Similarly, Pires de Sousa et al. (2009) based in a population of 1,224 subjects aged 35 years or older without known diabetes in Brazilian. They indicated age, BMI, and hypertension were significance factors to classify as type 2 diabetes patients by stepwise backward multiple logistic regression.

Aekplakorn et al. (2006) followed 2, 677 individuals between the ages of 35 to 55 years without diabetes in Thailand during 12 years. Multivariable logistic regression was used to indicate that age, BMI, waist circumference, hypertension, and family history of diabetes were significant predictive variables. The Atherosclerosis Risk in Communities (ARIC) study recruited 15,792 U.S. adults between the ages of 45 to 64 years as subjects. Schmidt et al. (2005) constructed a multivariable logistic regression modeling to indicate age, waist circumference, height, hypertension, family history of diabetes, ethnicity, HDL cholesterol, triglycerides, and fasting glucose were significant. Chaturvedi et al. (2008) studied in 4,044 individuals between the ages of 35 to 64 years in India. Age, blood pressure, waist circumference, and family history of diabetes were significant with $p<0.05$ in multivariable logistic regression analysis. Based on 1,549 participants from Rancho Bernardo Study in U.S.A., multivariable logistic regression were performed to indicate sex, age, triglycerides, and fasting plasma glucose as predictors (Kanaya et al, 2005).

### Cox Proportional hazards models

According to Perry et al. (1995), there are 7,735 middle aged (40 to 59 years old) British men who were selected at random from 24 towns in England Wales, and Scotland between 1978 and 1980. Cox's proportional hazards models were used to assess which factors could develop noninsulin dependent diabetes. Seven variables were selected by proportional hazards regression: age, BMI, blood pressure, triglycerides, high density lipoprotein (HDL) cholesterol, heart rate, and uric acid. Kawakami et al. (1997) investigated a cohort of 2,312 male employees who worked at electrical company in Japan. They used analysis to indicate that the age an individual starts smoking and the number of cigarettes smoked per day are two important factors that increase the non-insulin-dependent diabetes mellitus (NIDDM) incidence over 8 years (1984-1992). Sugimori et al. (1998) contained similar methods during a 16 year epidemiologic study with a cohort of 1,851 males and 722 females from Tokyo, Japan. Age, fasting blood glucose (FBS), family history, hypertension, smoking, and body mass index (BMI) were significant factors for diabetes in males, whereas age, FBS, drinking, not eating breakfast, and hypertension were significant factors for diabetes in females. Manson et al. (2000) studied 21,068 American male physicians aged 40 to 84 years over 12 years in the Physicians' Health Study, who were initially free of diagnosed diabetes mellitus, cardiovascular disease, and cancer. Heir result indicated cigarette smoking was an independent and modifiable determinant of type 2 diabetes mellitus by using proportional hazards regression models.

Multivariable Cox proportional hazards regression was used to construct a model for predicting the incidence of diabetes over 10 years in Chinese people. Age, elevated fasting glucose, body mass index, white blood cell count, triacylglycerol, and HDL-cholesterol were found as predictors to create the model (Chien et al, 2009). Sun et al. (2009) followed 73,961 individuals between the ages of 36 to 74 years over a median 3.15 years in the Taiwan periodic health-check population, and derived risk functions using multivariate Cox regression. Factors eventually included: age, gender, education level, smoking status, BMI, waist circumference, hypertension, high FPG, and HDL cholesterol. Among 19,257 hypertensive patients in Anglo-Scandinavian, Gupta et al. (2008) used multivariable cox proportional hazards regression to indicate the significant of predictors: age, sex, fasting plasma glucose (FPG), BMI, serum triglycerides, HDL cholesterol, alcohol intake, and systolic blood pressure. Schulze et al. (2007) investigated 9,729 men and 15,438 women aged 35-65 years old in Germany. Age, waist circumference, height, history of hypertension, physical activity, smoking, consumption of red meat, whole-grain bread, coffee, and alcohol were significant in the model. Hipposley-Cox et al, (2009) used a similar model to investigate primary care health records 2,540,753 patients between the ages of 25 to 79 years from 19 of the Qsearch databases in England. Age, sex, body mass index, smoking status, family history of diabetes, social deprivation, treated hypertension, cardiovascular disease, and current use of corticosteroids were the significant factors. In Tuomilehto et al.'s (2010) paper, a total 1,429 individuals aged 40-70 years with BMI 25-40 kg/ $m^2$ were randomly recruited from nine countries (Canada, Germany, Austria, Norway, Denmark, Sweden, Finland, Israel and Spain). Acarbose treatment, gender, serum triglyceride level, waist circumference, fasting plasma glucose, height, history of cardiovascular disease (CVD) and hypertension were included in multivariable Cox proportional hazards regression model.

### Other methods

Based on 46,239 Chinese adults aged 20 years or older, Yang et al. (2010) indicated age (older), sex (male), a family history of diabetes, overweight, obesity, central obesity, increased heart rate, elevated systolic blood pressure, elevated serum triglyceride level, educational level below college, and urban residence were all significantly associated with an increased risk of diabetes by using multivariable multinomial logistic models. In addition, as above, all factors except sex (male) and urban residence were significantly associated with an increased risk of pre-diabetes. Between 1980 to 1996, the Nurses' Health Study followed 84,941 female nurses who were tested for dietary and lifestyle factors in relation to type 2 diabetes. They got results by using pooled logistic regression; overweight was the single most important predictor of diabetes, though lack of exercise, a poor diet, currently smoking, and abstinence from alcohol were significant to increase risk of diabetes (Hu et al. 2001). Based on 12, 729 American adults aged 45-64 years, Kahn et al. (2009) demonstrated age (55 years or older), diabetic status of parents, hypertension, race (black), smoking status, waist circumference, rapid pulse, and nonuse of alcohol were significant factors by applying Weibull proportional hazards regression.

For pre-diabetes and undiagnosed diabetes, Heikes et al. (2008) used the data from the Third National Health and Nutrition Examination Survey to build two logistic regression and classification tree analysis models. These two models were used to designate any individuals who have a high risk for 'undiagnosed diabetes or pre-diabetes', 'pre-diabetes', and 'neither undiagnosed diabetes or pre-diabetes'. In the estimated coefficients equations of logistic regression, only 8 variables (age, gender, weight, standing height, waist-to-hip ratio, BMI, and high blood pressure) meet the $p< 0.05$ significance level for entry into the model.

Balkau et al. (2008) investigated 1,863 men and 1,954 women aged 30-65 years old in France, and used logistic regression to test for interactions with sex. The result show the predictors were fasting glucose, waist circumference, smoking, and hypertension for men, whereas fasting glucose, BMI, hypertension, and diabetes in family for women. Kolberg et al. (2009) devised a model development process applying multiple statistical approaches to reduce the number of factors based on six biomarkers (adiponectin, C-reactive protein, ferritin, inter-leukin-2 receptor A, glucose, and insulin) from 6,600 Danes followed over 5 years. Based on 3,140 participants aged 54 years, Wilson et al. (2007) used two modelling methods (Cox proportional hazards model and multivariate logistic regression) to estimate the risk of type 2 diabetes. The results show parental history of diabetes, BMI, triglycerides, HDL cholesterol, fasting plasma glucose (FPG), and blood pressure were significant association with incidence of diabetes. Furthermore, Xie et al. (2010) used classification and regression tree models based on 15,540 Chinese adults aged 35-74 years. The significant predictors for type 2 diabetes for men were age and waist circumference and for women were age and waist/hip ratio. In addition, Woolthuis et al. (2009) used multiple statistical analysis methods ($x^2$ test and logistic regression) to perform and analyze the data based on 49,229 practice population in Netherlands. Among diagnostic models containing various factors, a model containing obesity alone was the best predictor of undiagnosed diabetes. All literature review sources are given below in table 3.

Table 3. Logistic regression for diabetes or pre-diabetes

| **Source** | **Country** | **Method for adjustment** | **Predictors in the model** |
|---|---|---|---|
| Griffin et al, 2000 | UK | Logistic Regression | Sex, prescribed antihypertensive medication, prescribed steroids, age, BMI, family history of diabetes, smoking status |
| Bonora et al, 2004 | Italy | Logistic Regression | Age, BMI, Hypertension, Dyslipidemia, IFG (Impaired fasting glucose) and IGT (Impaired glucose tolerance) |
| Schmidt et al, 2005 | USA | Logistic Regression | Age, ethnicity, parental history of diabetes, FPG, systolic blood pressure, waist circumference, height, HDL cholesterol, triglycerides |
| Borrell et al, 2007 | USA | Logistic Regression | Age, sex, ethnicity, family history of diabetes, self-reported hypertension, hypercholesterolaemia, periodontal disease |
| Gray et al, 2010 | UK | Logistic Regression | Age, ethnicity, sex, first-degree family history of diabetes, antihypertensive therapy or history of hypertension, waist circumference, BMI |
| Baan et al, 1999 | The Netherlands | Stepwise Logistic Regression | Age, sex, use of antihypertensive medication, obesity (BMI ≥ 30) |
| Al Khalaf et al, 2010 | Kuwait | Forward Stepwise Logistic Regression | Age, waist circumference, blood pressure medication, diabetes in sibling |
| Al-Lawati et al, 2007 | Oman | Backward Stepwise Logistic Regression | Age, waist circumference, BMI, family history of diabetes, hypertension |
| Burke et al, 1999 | USA | Multiple Logistic Regression | Age, sex, ethnic group, neighborhood, and date of enrollment |
| Stern et al, 2002 | USA | Multiple Logistic Regression | Age, sex, ethnicity, FPG, systolic blood pressure, HDL cholesterol, BMI, family history of diabetes |

Table 3. Logistic regression for diabetes or pre-diabetes (continued)

| Source | Country | Method for adjustment | Predictors in the model |
|---|---|---|---|
| Carlsson, et al, 2004 | Germany | Multiple Logistic Regression | Age, BMI, physical activity and alcohol consumption |
| Mohan et al, 2005 | India | Multiple Logistic Regression | Age, abdominal obesity (waist circumference), physical activity, family history of diabetes |
| Waki et al, 2005 | Japan | Multiple Logistic Regression | High alcohol consumption |
| Ramachandran et al, 2005 | India | Multiple Logistic Regression | Age, family history of diabetes, BMI, waist circumference, physical activity |
| Bang et al, 2009 | USA | Multiple Logistic Regression | Age, sex, family history of diabetes, history of hypertension, obesity (BMI or waist circumference), physical activity |
| Chen et al, 2010 | Australia | Multiple Logistic Regression | Age, sex, ethnicity, parental history of diabetes, history of high blood glucose, use of antihypertensive medication, smoking status, physical activity, waist circumference |
| Ko et al, 2010 | Hong Kong | Multiple Logistic Regression | Age, sex, BMI, hypertension, dyslipidaemia, family history of diabetes, gestational diabetes |
| Ruige et al, 1997 | The Netherlands | Backward Stepwise Multiple Logistic Regression | Frequent thirst, pain during walking with need to slow down, shortness of breath when walking, age, sex, obesity (BMI), obesity (men), family history of diabetes, use of antihypertensive drugs, reluctance to use bicycle for transportation |
| Glümer et al, 2004 | Denmark | Stepwise Backward Multiple Logistic Regression | Age, BMI, sex, known hypertension, physical activity, family history of diabetes |
| Keesukphan et al, 2007 | Thailand | Stepwise Multiple Logistic Regression | Age, BMI, history of hypertension |
| Bindraban et al, 2008 | The Netherlands | Stepwise Multiple Logistic Regression | Age, BMI, waist circumference, resting heart rate, first-degree relative with diabetes, hypertension, history of CVD, ethnicity |
| Pires de Sousa et al, 2009 | Brazil | Stepwise Backward Multiple Logistic Regression | Age, BMI, hypertension |
| Aekplakorn et al, 2006 | Thailand | Multivariable Logistic Regression | Age, sex, BMI, abdominal obesity (waist circumference), hypertension, family history of diabetes. |
| Chaturvedi et al, 2008 | India | Multivariable Logistic Regression | Age, blood pressure, waist circumference, family history of diabetes |
| Kanaya et al, 2005 | USA | Multivariable Logistic Regression | Sex, age, triglycerides, FPG |
| Lindström et al, 2003 | Finland | Multivariate Logistic Regression | Age, BMI, waist circumference, use of blood pressure medication, history of high blood glucose, physical activity, daily consumption of vegetables |
| Cabrera de León et al, 2008 | Canary Islands | Multivariate Logistic Regression | Men: age, waist/height ratio, family history of diabetes<br>Women: age, waist/height ratio, family history of diabetes, gestational diabetes |

Table 3. Logistic regression for diabetes or pre-diabetes (continued)

| **Source** | **Country** | **Method for adjustment** | **Predictors in the model** |
|---|---|---|---|
| Gao et al, 2010 | China | Multivariate Logistic Regression | Age, waist circumference, family history of diabetes |
| Tabaei and Herman, 2002 | Egypt | Multivariate Logistic Regression | Age, random plasma glucose, postprandial time, sex, BMI |
| Perry et al, 1995 | UK | Cox Proportional Hazards Regression | Age, BMI, blood pressure, triglycerides, high density lipoprotein (HDL) cholesterol, heart rate, and uric acid |
| Kawakami et al, 1997 | Japan | Cox Proportional Hazards Regression | Younger age at starting smoking and the number of cigarettes smoked per day |
| Sugimori et al, 1998 | Japan | Cox Proportional Hazards Regression | Men: Fasting blood glucose (FBS), age, family history, hypertension, smoking, and body mass index (BMI)<br>Women: not eating breakfast, FBS, age, drinking, and hypertension |
| Manson, et al, 2000 | USA | Cox Proportional Hazards Regression | Cigarette smoking is an independent and modifiable determinant |
| Schulze et al, 2007 | Germany | Multivariate Cox Proportional Hazards Regression | Waist circumference, height, age, hypertension, intake of red meat, intake of whole-grain bread, coffee consumption, alcohol consumption, physical activity, former smoker, current heavy smoker (≥ 20 cigarettes/day) |
| Gupta et al, 2008 | UK, Ireland, Sweden, Denmark, Iceland, Norway, Finland | Multivariate Cox Proportional Hazards Regression | Age, sex, FPG, BMI, randomized group, triglycerides, systolic blood pressure, total cholesterol, use of non-coronary artery disease medication, HDL cholesterol, alcohol intake |
| Chien et al, 2009 | Taiwan | Cox Proportional Hazards Regression | Age, BMI, WBC count, and triacylglycerol, HDL cholesterol, FPG levels |
| Gao et al, 2009 | Mauritius | Cox Proportional Hazard Regression | Age, sex, BMI, waist circumference, family history of diabetes |
| Hippisley-Cox et al, 2009 | UK | Cox Proportional Hazards Regression | Age, BMI, family history of diabetes, smoking status, treated hypertension, current treatment with corticosteroids, diagnosis of CVD, social deprivation, ethnicity |
| Sun et al, 2009 | Taiwan | Multivariable Cox Proportional Hazard Regression | Sex, education level, age, current smoking status, BMI, waist circumference, family history of diabetes, hypertension, FPG |
| Tuomilehto et al, 2010 | Canada, Germany, Austria, Norway, Denmark, Sweden, Finland, Israel, Spain | Multivariable Cox Proportional Hazard Regression | Acarbose treatment, sex, serum triglyceride level, waist circumference, FPG, height, history of CVD, diagnosed hypertension |
| Kahn et al, 2009 | USA | Weibull Proportional Hazard Regression | Diabetic mother, diabetic father, hypertension, ethnicity, age, smoking status, waist circumference (sex), height (sex), resting pulse (sex), weight (sex) |

Table 3. Logistic regression for diabetes or pre-diabetes (continued)

| **Source** | **Country** | **Method for adjustment** | **Predictors in the model** |
|---|---|---|---|
| Balkau et al, 2008 | France | Sex-specific Logistic Regression | Men: waist circumference, smoking status, hypertension.<br>Women: waist circumference, family history of diabetes, hypertension. |
| Hu et al. 2001 | China | Pooled logistic regression | Overweight, lack of exercise, a poor diet, current smoking, and abstinence form alcohol |
| Wilson et al, 2007 | USA | Cox Proportional Hazards Regression Multivariate Logistic Regression | FPG, BMI, HDL cholesterol, parental history of diabetes, triglyceride level, blood pressure |
| Heikes et al, 2008 | USA | Logistic Regression and Classification Tree Analysis | Age, waist circumference, history of gestational diabetes, family history of diabetes, ethnicity, high blood pressure, weight, height, parental diabetes, exercise |
| Woolthuis et al, 2009 | Netherlands | $x^2$ test and logistic regression | Obesity |
| Xie et al, 2010 | China | Classification and Regression Tree | Men: age, waist circumference<br>Women: age, waist/hip ratio |
| Yang et al, 2010 | China | Multivariable, Multinomial, logistic models | Male sex, older age, a family history of diabetes, overweight, obesity, central obesity, increased heart rate, elevated systolic blood pressure, elevated serum triglyceride level, educational level below college, and urban residence |
| Meng et al, 2013 | China | Logistic Regression, Artificial Neural Networks (ANNs) and Decision Tree | Age, family history of diabetes, marital status, educational level, work stress, duration of sleep, physical activity, preference for salty food, gender, eating fish, drinking coffee, and body mass index |
| Kolberg et al, 2009 | Denmark | **U** (univariate logistic regression analyses),<br>**E** (exhaustive enumeration of small multivariate logistic models),<br>**H** (six different heuristic model-building methods, including forward, backward, and stepwise selection, Kruskal-Wallis, random forest, and Eigengene-based linear discriminant analysis with three different statistical learning algorithms, including logistic regression, linear discriminant analysis, and support vector machines), and<br>**B** (frequency of selection within 100 bootstrap replicates using the same basic heuristic model-building methods) | Adiponectin, C-reactive protein, ferritin, interleukin 2 receptor A, glucose, insulin |

## Decision Tree Literature Review

Decision tree-ID3, C4.5, and C5.0 Algorithm

The Fuzzy ID3 (Iterative Dichotomiser 3) algorithm as a precursor to the C4.5 is used by training on a dataset to produce a decision tree. Daveedu et al. (2012) indicated that a total of eight variables were used to obtain decision tree for individual clusters: age, BMI, number of times pregnant, plasma glucose concentration after 2 hours in an oral glucose tolerance test, diastolic blood pressure, triceps skin thickness, 2-hour serum insulin, and diabetes pedigree function.

According to the Indian diabetes dataset, Al and Asma (2011) used C4.5 in Waikato Environment for Knowledge Analysis (Weka) software. Age, diabetes pedigree function, BMI, 2-hour serum insulin, triceps skin fold thickness, diastolic blood pressure, plasma glucose concentration after 2 hours in an oral glucose tolerance test, and number of times pregnant were considered as predictors. They gave negative (non-diabetic) and positive (diabetic) as decision results, and the accuracy of the result was 78.18%. Luo et al. (2014) used C4.5 and multivariate logistic regression to selected variables from the data of 16,246 individuals aged 20 and older in Beijing, China. Combining the result of these two methods, nine factors were selected: age, diastolic blood pressure (DBP), high-density lipoprotein (HDL), waist, sex, cholesterol (CHOL), parental or sibling history, body mass index (BMI), and triglyceride (TG).

In a research paper presented by Vohra and Anshul (2014). A total of 206 individuals were collected from hospital records, the analyzed nine variables: age, number of times pregnant, fast glucose tolerance test, casual glucose tolerance test, diastolic blood pressure, serum insulin, triceps skin thickness, BMI, and diabetes pedigree function. They tested C4.5 and ID3 to determine which was more accurate to predict three levels (normal, pre-diabetes, and diabetes).

Comparing three data mining algorithm techniques (C4.5, Naïve Bayes, and IB1), Huang et al (2007) identified five predictors (age, diagnosis duration, insulin treatment, random blood glucose measurement, and diet treatment) as important factors for Ulster diabetes. Three popular decision tree algorithms (ID3, C4.5, and CART) were compared in the research paper presented by Lavanya and Usha (2011). Using data collected from the UCI Machine Learning Repository, they analyzed eight attributes for diabetes classifiers. The CART showed the best accuracy (99.45%) relative to C4.5 (96.24%) and ID3 (84.52%).

Based on Pima Indian Diabetes Dataset (PIDD), there are several studies comparing various data mining techniques. They all used on eight variables as factors: age, diabetes pedigree function, BMI, 2-Hour serum insulin, triceps skin fold thickness, diastolic blood pressure, plasma glucose concentration after 2 hours in an oral glucose tolerance test, and number of times pregnant. Karthikeyani et al, (2012) compared ten data mining algorithm methods: 1. C4.5; 2. Support vector machines (SVM); 3. K-nearest neighbor (K-NN); 4.Prototype nearest neighbor (PNN); 5. Binary Logistic Regression (BLR); 6. Multinomial logistic regression (MLR); 7. Classification and Regression Trees (CRT); 8. Chi-Square Classification and Regression Trees (CS-CRT); 9. Partial Least Square Discriminant Analysis (PLS-DA); 10. Partial Least squares-Linear Discriminant Analysis (PLS-LDA). Of these, C4.5, CRT, and CS-CRT were accepted as decision tree algorithms. Amatul et al. (2013) also compared ten algorithms in study, C4.5 and CS-CRT have the most accurate (86%). Total 768 patients were selected in same dataset, ten techniques: 1.F-score Feature Selection, k-means Clustering and SVM; 2. K-means algorithm; 3.Cascading K-means Clustering and K-Nearest Neighbor Classifier; 4. b-Colouring Technique in Clustering Analysis; 5. Feature Weighted Support Vector Machines and Modified Cuckoo Search; 6. Cascaded K-Means and Decision Tree C4.5;

7. Rough sets; 8. Prediction Model Discovery Using Rapid Miner; 9. Ensemble model (SVM, Discriminant analysis and Bayesian Network); 10. Neural Network and Fuzzy k-Nearest Neighbor Algorithm were applied to test the accuracy of the prediction.

Based on the same database and same factors, Visalatchi et al. (2014) compared five data mining algorithm methods, the most accurate was C4.5 (86%). Radha and B (2014) used Pima Indian Diabetes Dataset and Indian Council of Medical Research–Indian Diabetes (ICMR-INDIAB) study to compare five data mining methods: C4.5, SVM, *k*-NN, PNN, and Binary Logistic Regression (BLR). Karthikeyani et al. (2012) also compared C4.5, SVM, *k*-NN, PNN, Binary Logistic Regression (BLR), MLR, CRT, CS-CRT, PLS-DA, AND PLS-LDA by using the same eight variables. Similarly, Karegowda et al. (2012) make a hybrid model by using K-means clustering and decision tree C4.5 from the same database and same eight factors. The classification accuracy could be 93.33%.

C5.0 as an improved version of the C4.5 and ID3 algorithm was used by Toussi et al. (2009). They analyzed six variables (acute clinical symptoms, HbA1c, type of current treatment, body mass index (BMI), the existence of renal insufficiency, and being old (yes or no)) to explore who have the risk of diabetes. Meng et al. (2013) compared three data mining algorithms (logistic regression, artificial neural networks (ANNs) and decision tree-C5.0), according to twelve importance variables from the most to the least (age, educational level, family history of diabetes, marital status, preference for salty food, drinking coffee, duration of sleep, body mass index, work stress, eating fish, physical activity and gender), between these thee algorithms C5.0 was the best predictor and achieved a classification accuracy of 77.87%.

### Decision tree-Classification and Regression Trees (CART)

As early as in the 1990s, Barriga et al. (1996) used classification and regression tree (CART) to screen for impaired glucose tolerance and previously undiagnosed diabetes. Based on 583 Hispanic and 768 Non-Hispanic white observations, four variables (age, BMI, fasting glucose, and glycohemoglobin) were found to be the significant factors for identifying the different levels of diabetics. In a research paper presented by Herman et al. (1995), there were six total risk variables (age, sex, history of delivery of a macrosomic infant, obesity, sedentary lifestyle, and family history of diabetes) used indicators in CART models. CART was developed to identify individual who have a high risk for previously undiagnosed diabetes.

Breault et al. (2002) examined 30, 383 diabetes patients in New Orleans by using CART. In CART 4.0 version, HgbA1c >9.5 (0, 1) as target with ten predictors: age, sex, emergency department visits (ER), office visits (OV), comorbidity index (CMI), dyslipidemia, hypertension (HTN), cardiovascular disease (CV), retinopathy, and end-stage renal disease (ESRD). Miyaki et al. (2002) indicated CART was used to identify patients by two factors: macroangiopathy and microangiopathy. In the classification tree of macroangiopathy, five predictors were found: systolic blood pressure (BPSYS), triglyceride (TG), blood glucose (GLU), low density lipoprotein (LDL) and free fatty acid (FFA) at $p<0.02$ significance level. On the other hand, for microangiopathy group, there were five predictors that were significant at $p<0.02$: morbidity term, body mass index, high density lipoprotein (HDL), hemoglobin Alc (HBA1C), and blood glucose (GLU). Kavitha and Sarojamma (2012) developed a diabetes diagnostic system based on the CART method. Patients could login to the system and input their informations to check the status of diabetes. Age, fasting plasma glucose level (FPG), BMI, oral glucose tolerance test (OGTT), and A1c were five significant predictors in decision tree modeling.

In four common variables (age, waist-hip ratio (WHR), waist circumference (WC), and BMI) from Chinese adults between the ages of 35-74 years, Xie et al. (2010) separated men and women in different groups. For women, WHR and age as predictors were selected by CART, and for men, WC and age were selected to identify the diabetes risk levels. Sankaranaratanan and T (2014) used the dataset of 768 participants in UCL Machine Learning Repository which contained eight attributes. However, age, gender, and level value of glycated hemoglobin (HbA1C) were used to contribute the decision tree algorithm CART. Mochan and Ebell (2009) developed a risk assessment tool (CART and logistic regression) for detecting undiagnosed diabetes based on the study paper presented by Heikes et al (2008).

### Decision tree-other methods

In the Tehran lipid and glucose study (TLGS) database a total of 6,647 individuals aged older than 20 years were followed over 12 years. Classification by the decision tree was used to create a prediction model for identify the incidence of type 2 diabetes (Ramezankhani et al, 2014). In a research paper presented by Xin et al (2010), logistic regression and classification tree analysis were used to build a model to identify those who have a high risk of type 2 diabetes (T2DM) or pre-diabetes (PDM). In the classification tree analysis, waist-hip ratio (WHR), waist, hypertension, age and weight were found to be predictors in this model for detecting undiagnosed T2DM. Age, hypertension, waist-hip ratio (WHR), family history of diabetes and waist were found to be predictors in this model for detecting PDM and undiagnosed T2DM.

According to Taiwan’s National Health Insurance system, Liou et al (2008) applied three data mining techniques (logistic regression, neural network, and classification tree model) and compared their accuracy. Classification tree algorithm showed the best accuracy (99%), following by logistic regression (92%) and neural network (95%). Average days of drug dispensed, average medical expenditure per day, average dispensing service fees, average diagnosis fees, average drug cost per patient, average drug cost per patient per day, and average consultation and treatment were used to create the classification tree model.

Hische et al. (2010) analyzed metabolic syndrome from Berlin Potsdam Study (Mesy-Bepo) cohort and Dresden cohort by using a decision tree. Age and systolic blood pressure were used as the factors to build the decision tree. Bruno et al. (2014) adopted a decision tree to create a decision tree classifier, and they analyzed the dataset from Italian local health centers. Age, gender and some examination information (i.e. HDL cholesterol, gamma-glutamyl transferase (GT), venous blood, cardiovascular, general visit, glycated hemoglobin, and fundus oculi) were considered as nodes of the decision tree. The accuracy value was 98.6% by using the multiple level clustering strategy. Table 4 lists all literature review sources of decision tree modeling.

Table 4. Decision tree for diabetes or pre-diabetes

| Source | Country | Method(s) for adjustment | Risk predictors in the model |
|---|---|---|---|
| Daveedu et al, 2012 | Hybrid Classification System | ID3 | Number of times pregnant, Plasma glucose concentration a 2 hours in an oral glucose tolerance test, Diastolic blood pressure, Triceps skin thickness, 2-Hour serum insulin, BMI, Diabetes pedigree function, age |
| Al and Asma, 2011 | India | C4.5 | Number of times pregnant, Age, Diabetes pedigree function, BMI, Diastolic blood pressure, Plasma glucose concentration a 2 hours in an oral glucose tolerance |
| Luo et al, 2014 | China | Decision tree-C4.5 & Multivariate Logistic Regression | Age, diastolic blood pressure (DBP), high-density lipoprotein<br>(HDL), waist, sex, cholesterol (CHOL), parental or sibling history, body mass index (BMI), and triglyceride (TG) |
| Toussi et al, 2009 | France | C5.0 | Acute clinical symptoms, HbA1c, type of current treatment, body mass index (BMI) , the existence of renal insufficiency, and being old |
| Meng, et al, 2013 | China | Logistic Regression, Artificial Neural Networks (ANNs) and Decision Tree-C5.0 | Age, educational level, family history of diabetes, marital status, preference for salty food, drinking coffee, duration of sleep, body mass index, work stress, eating fish, physical activity, gender |

Table 4. Decision tree for diabetes or pre-diabetes (continued)

| Source | Country | Method(s) for adjustment | Risk predictors in the model |
|---|---|---|---|
| Vohra et al, 2014 | 206 instances | C4.5, and ID3 | Number of times pregnant, fast glucose tolerance test, casual glucose tolerance test, Diastolic blood pressure, serum insulin, Triceps skin thickness, BMI, Diabetes pedigree function, age |
| Huang et al, 2007 | Ulster | C4.5, Na ïve Bayes, and IB1 | Age, diagnosis duration, the need for insulin treatment, random blood glucose measurement and diet treatment |
| Lavanya and Usha, 2011 | UCI Machine Learning Repository | ID3, C4.5, and CART | Eight attributes |
| Karegowda et al, 2012 | India | K-means clustering and Decision tree C4.5 | Age, Diabetes pedigree function, BMI, 2-Hour serum insulin, Triceps skin fold thickness, Diastolic blood pressure, Plasma glucose concentration a 2 hours in an oral glucose tolerance test, Number of times pregnant |
| Karthikeyani et al, 2012 | India | C4.5, SVM, K-NN, PNN, BLR, MLR, CRT, CS-CRT, PLS-DA, PLS-LDA | Age, Diabetes pedigree function, BMI, 2-Hour serum insulin, Triceps skin fold thickness, Diastolic blood pressure, Plasma glucose concentration a 2 hours in an oral glucose tolerance test, Number of times pregnant |
| Amatul et al, 2013 | India | 1. F-score Feature Selection, k-means Clustering and SVM,<br>2. K-means algorithm,<br>3.Cascading K-means Clustering and K-Nearest Neighbor Classifier,<br>4. b-Colouring Technique in Clustering Analysis,<br>5. Feature Weighted Support Vector Machines and Modified Cuckoo Search,<br>6. Cascaded K-Means and Decision Tree C4.5, 7. Rough sets,<br>8. Prediction Model Discovery Using RapidMiner,<br>9. Ensemble model (SVM, Discriminant analysis and Bayesian Network)<br>10.Neural Network and Fuzzy k-Nearest Neighbor Algorithm | Age, Diabetes pedigree function, BMI, 2-Hour serum insulin, Triceps skin fold thickness, Diastolic blood pressure, Plasma glucose concentration a 2 hours in an oral glucose tolerance test, Number of times pregnant |
| Visalatchi et al, 2014 | India | C4.5, SVM, k-NN, Na ïve Bayes and Apriori | Age, Diabetes pedigree function, BMI, 2-Hour serum insulin, Triceps skin fold thickness, Diastolic blood pressure, Plasma glucose concentration a 2 hours in an oral glucose tolerance test, Number of times pregnant |
| Radha and B, 2014 | India | C4.5, SVM, *k*-NN, PNN, and Binary Logistic Regression (BLR) | Age, Diabetes pedigree function, BMI, 2-Hour serum insulin, Triceps skin fold thickness, Diastolic blood pressure, Plasma glucose concentration a 2 hours in an oral glucose tolerance test, Number of times pregnant |
| Herman et al (1995) | USA | CART | Age, sex, history of delivery of a macrosomic infant, obesity, sedentary lifestyle, and family history of diabetes |
| Barriga et al (1996) | USA | CART | Fasting glucose, age and BMI and glycohemoglobin |

Table 4. Decision tree for diabetes or pre-diabetes (continued)

| Source | Country | Method(s) for adjustment | Risk predictors in the model |
|---|---|---|---|
| Breault et al, 2002 | New Orleans | CART | HgbA1c >9.5 and 10 predictors: age, sex, emergency department visits, office visits, comorbidity index, dyslipidemia, hypertension, cardiovascular disease, retinopathy, end-stage renal disease |
| Miyaki et al, 2002 | Japan | CART | Macroangiopathy: systolic blood pressure (BPSYS), triglyceride (TG), blood glucose (GLU), low density lipoprotein (LDL) and free fatty acid (FFA)<br>Microangiopathy: morbidity term, body mass index, high density lipoprotein ( HDL), hemoglobin Alc (HBA1C) and Blood glucose (GLU) |
| Kavitha and Sarojamma, 2012 | 1025 individuals | CART | Age, BMI, OGTT, FPG and A1C |
| Sankaranaratanan and T, 2014 | UCI Machine Learning Repository | CART | Level value of Glycated Hemoglobin (HbA1C), age and gender |
| Heikes et al, 2008 | USA | CART and Logistic Regression | Age, waist circumference, history of gestational diabetes, parental diabetes, ethnicity, high blood pressure, weight, height, exercise more than peers |
| Xie et al, 2010 | China | CART and Multivariable Logistic Regression | Women: Waist-hip ratio (WHR), Age<br>Men: Waist circumference (WC), Age |
| Xin et al, 2010 | China | Classification tree analysis and logistic regression | T2DM: Waist-hip ratio (WHR), waist, hypertension, age and weight<br>PDM& T2DM: Hypertension, Waist-hip ratio (WHR), age, family history of diabetes and waist |
| Liou et al, 2008 | Taiwan | Logistic Regression, Neural Networks and Classification Tree | Average days of drug dispense, average medical expenditure per day, average dispensing service fees, average diagnosis fees, average drug cost per patient, average drug cost per patient per day, average consultation and treatment fees, average medical expenditure, and average amount claimed |
| Hische et al, 2010 | Germany | Decision tree | Age, systolic blood pressure |
| Ramezankhani et al, 2014 | Tehran Lipid and Glucose Study | Decision tree analysis | Fasting plasma glucose, body mass index, triglycerides, mean arterial blood pressure, family history of diabetes, educational level and job status. |
| Bruno et al, 2014 | Italian Local Health Center | Decision tree classifier | HDL cholesterol, Gamma-glutamyl transferase (GT), venous blood, Cardiovascular, general visit, Glycated hemoglobin, Fundus oculi, gender, age |

# PREDICTION MODELS

A special emphasis has been placed on the assessment of individuals' risk of diabetes or pre-diabetes. As previously described, statistical analyses was performed in SAS version 9.4 Statistics for chi-square test and logistic regression. For decision tree analysis and comparison of the prediction models, SAS Enterprise Miner Version 13.1 was used.

## Logistic Regression Analysis

This aim at compared the performance of the different prediction methods for type 2 diabetes. Logistic regression can be applied to predict the probability of an event in subjects at risk of pre-diabetes. One tool was developed using the logistic regression model to identify individuals who have pre-diabetes as follows:

$$\text{Logit (p)} = \ln\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right) = \beta_0 + \beta_1 x_1 + \cdots \cdots + \beta_i x_i \tag{1}$$

Where p is the probability in individual has pre-diabetes, X ($x_1, \cdots x_i$) are the dichotomous explanatory variables, $\beta_0$ is constant, and $\beta_1$ to $\beta_i$ are the vector of regression coefficients corresponding to X. The term $\frac{\mathrm{p}}{1-\mathrm{p}}$ is known as the odds of the people who have pre-diabetes. The probability can be rewritten as

$$\mathrm{P} = [1 + \mathrm{e}\hat{}(-(\beta_0 + \beta_1 x_1 + \cdots \cdots + \beta_i x_i \ \ ))]^{-1} \tag{2}$$

In the logistic regression analysis, the odds of risk is usually used to predict the probability of the risk. As mentioned in part 3, there are many useful regression models used to predict diabetes or pre-diabetes. Logistic regression models and Cox Proportional hazards models could be simple or multivariable models. The function of Cox Proportional hazards model is as follows:

$$\lambda\ (t|X) = \lambda_0(t) \exp\ (\beta_1 x_1 + \cdots\cdots + \beta_i x_i) = \lambda_0(t) \exp\ (\ X\ \ \beta') \qquad (3)$$

In this case, $\lambda_0(t)$ denotes how the risk of event per unit time change from initial levels of concomitant variables. Based on prior understanding, the NHANES's data collected and reported in 2-year cycles. About 12,000 people pre 2-year cycle were invited to participate from a sampling frame that include all 50 states and the District of Columbia were selected. Unfortunately, for every 2-year survey cycle, the NHANES program selects new participants. This is the reason why the relationship between time and factors could not be found.

Multivariate regression analysis requires two or more dependent or target variables. But, in this study, the dependent variable was a binary categorical variable with two levels: 0 and 1, where 0 means normal and 1 means pre-diabetes. As mentioned, the independent variables have seventeen factors that were not time dependent. In order to understanding the complicated interactions between factors of pre-diabetes, 'enter' and 'forward' methods were used for testing predictors and developing the binary logistic regression.

**Decision Tree Analysis**

After logistic regression analysis, another important predictive model for pre-diabetes is decision trees. A decision tree is a powerful and popular data mining method, because it is systematic and is a graphical tool so one can see how to classify and predict variables. Decision trees are a typical technology in data mining that uses a logic model to output the classification method. By adopting a top-down approach and using the essential parts of the decision tree (decision node, branches, and leaves) the tree can be constructed.

According to the division of the different standards, the decision tree could be divided into different types. Based on the number of the branch in the interior nodes, two-level splits and multi-level splits are separated into two kinds of decision tree. In data mining, decision trees have two main types: classification tree and regression tree. The classification analysis predicts the result is the class to which the data belongs. Regression tree analysis predicts the output as a real number. In this study regression tree was used.

There are several popular decision trees algorithms used, such as ID3 (Qulinlan, 1986), Breiman et al.'s CART (1984), C4.5 (Qulinlan, 1993), and C5 (Qulinlan, 1993). ID3 stands for Iterative Dichotomiser 3 that was designed by Ross Quinlan for decision tree algorithms. ID3 is not optimal because it uses expected entropy reduction, not actual reduction. Therefore, after seven years, C4.5 and C5 algorithm were introduced by Qulinlan. C5.0 is significantly faster than C4.5. C5.0 algorithm require the target must be a categorical variable (i.e., nominal or ordinal). The CART is acronym for classification and regression trees. In general, the difference of CART and C4.5 are that tests in CART are always binary, but C4.5 allows two or more outcomes (Breiman, 1984). Unlike C4.5, the target of CART model is not just numeric variables but also character variables. CHAID (Chi-squared Automatic Interaction Detection) was published by Gordon V. Kass in 1980 (Magidson, 1994), which is especially applicable for categorical variables such as target. It can produce more than two branches in decision tree models. In SAS Enterprise Miner, their references list CART and CHAID were used in this system.

# RESULTS

## Results of Chi-square Test

In this study, chi-square tests were used to select risk factors from seventeen variables at p<0.05 significance level. It is well known chi-square was used to test the dependent variables. In order to fit the chi-square test, all continuous variables should be recoded to categorical variables. In other words, age and BMI were transformed into categorical variables, as shown below.

$$\text{Age level=}\begin{cases} 1 \text{ if } 20-39 \text{ years old} \\ 2 \text{ if } 40-59 \text{ years old} \\ 3 \text{ if } 60-79 \text{ years old} \\ 4 \text{ if } 80 \text{ years and older} \end{cases} \qquad \text{BMI level=}\begin{cases} 1=\text{Underweight } < 18.5\, kg/m^2 \\ 2=\text{Normal } 18.51-25 kg/m^2 \\ 3=\text{Overweight } 25.01-30 kg/m^2 \\ 4=\text{Obese} > 30 kg/m^2 \end{cases}$$

As the table 5 shows, in 2007-2008, a total of nine dependent variables showed statistically significant differences between normal and pre-diabetes individuals at a significance level of p<0.05. They are: age level (p<0.0001); body mass index (p<0.0001); citizenship status (p=0.0108); marital status (p=0.0015); risk for diabetes (p<0.0001); smoked (p=0.026); high cholesterol level (p<0.0001); hypertension status (p<0.0001); and diet (p=0.0031). The 2009-2010 data also have nine dependent variables shown in table 6, age level (p<0.0001); body mass index (p<0.0001); US armed forces (p=0.0178); total number of people in family (p=0.0042); risk for diabetes (p<0.0001); smoked (p=0.0004); high cholesterol level (p<0.0001); hypertension status (p<0.0001); and diet (p=0.0166) showed statistically significant differences. The 2011-2012 data shown in table 7: age level (p<0.0001); body mass index (p<0.0001); gender (p=0.0036); US armed forces (p=0.0006); citizenship status (p=0.0308); marital status (p=0.0006); total number of people in family (p=0.0246); annual family income (p=0.0369); risk for diabetes (p<0.0001); high cholesterol level (p<0.0001); hypertension status (p<0.0001); and diet (p=0.0072) that twenty variables showed statistically significant differences. The results of three sets of data were compared with the five variables age levels, BMI levels, risk for diabetes, high cholesterol levels, and hypertension status are significantly different in each year because the p-values are all less than 0.0001. According to the results of table 5-7, the figure 2 was made as following:

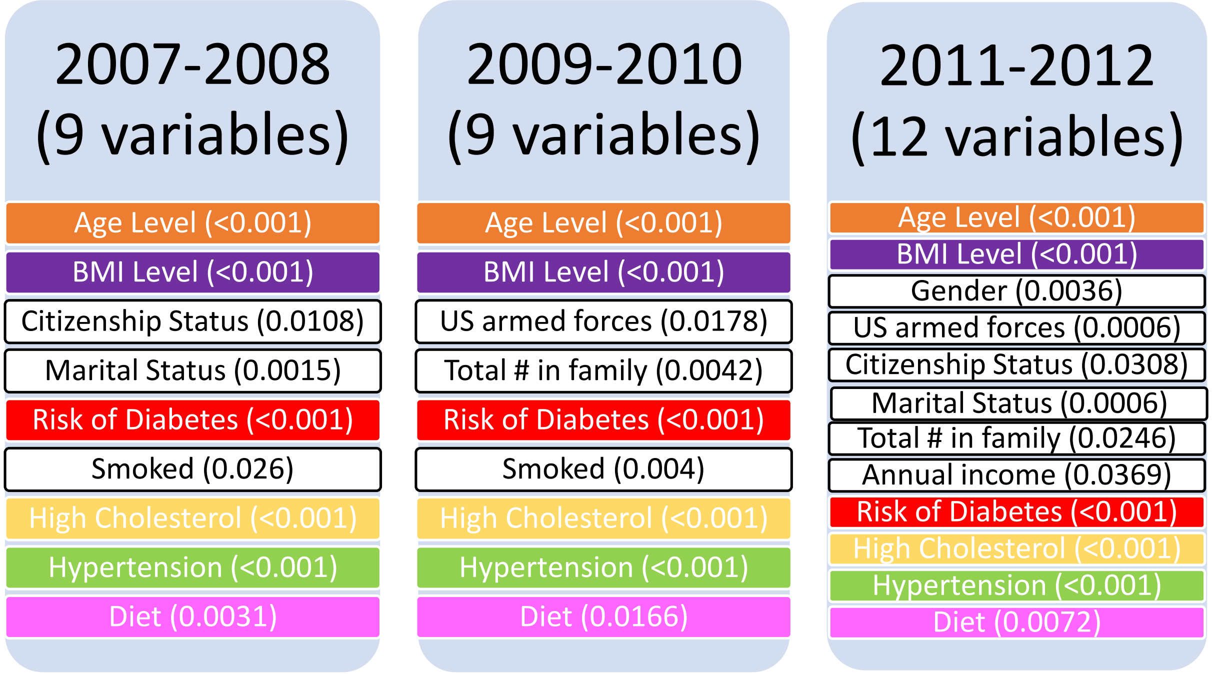


Figure 2. Significant variables were selected by Chi-square test

Combining three sets of data, thirteen variables were selected as risk factors in logistic regression and decision tree model. They are age, BMI, gender, served active duty in US armed forces, citizenship status, marital status, total number of people in the family, annual family income, ever told have health risk for diabetes, smoked, high cholesterol level, hypertension status, and diet. In other words, four variables (race, education level, born of U.S., and physical activity) were not associated with diagnosis of pre-diabetes at the p<0.05 significance level.

Table 5. Chi-square test analysis for 2007-2008

| **Factors** | **Factors Level** | **Pre-diabetes** | **Normal** | **Total** | **P-values** |
|---|---|---|---|---|---|
| | | 241 (8.07%) | 2744 (91.93%) | N=2985 | |
| **Age level** | 20-39 years old | 29 (12.03%) | 718 (26.17%) | 747 (25.023) | **<0.0001** |
| | 40-59 years old | 88 (36.51%) | 938 (34.18%) | 1026 (34.37%) | |
| | 60-79 years old | 105 (43.57%) | 870 (31.71%) | 975 (32.66%) | |
| | 80 years and older | 19 (7.89%) | 218 (7.94%) | 237 (7.94%) | |
| **Body mass index** | Underweight (<18.5 kg/$m^2$) | 2 (0.83%) | 46 (1.68%) | 48 (1.61%) | **<0.0001** |
| | Normal (18.51-25 kg/$m^2$) | 37 (15.35%) | 765 (33.81%) | 802 (26.87%) | |
| | Overweight (25.01-30kg/$m^2$) | 77 (31.95%) | 989 (36.04%) | 1066 (35.71%) | |
| | Obese (>30 kg/$m^2$) | 125 (51.87%) | 944 (34.4%) | 1069 (35.81%) | |
| Gender | Male | 111 (46.06%) | 1261 (45.95%) | 1372 (45.96%) | 0.9754 |
| | Female | 130 (53.94%) | 1483 (54.05%) | 1613 (54.04%) | |
| Race | Mexican American | 34 (14.11%) | 371 (13.52%) | 405 (13.57%) | 0.9290 |
| | Other Hispanic | 21 (8.71%) | 283 (10.31%) | 304 (10.18%) | |
| | Non-Hispanic White | 128 (53.11%) | 1446 (52.7%) | 1574 (52.73%) | |
| | Non-Hispanic Black | 47 (19.5%) | 537 (24.61%) | 584 (19.56%) | |
| | Other Race | 11 (4.56%) | 107 (3.9%) | 118 (3.95%) | |
| US armed forces | Yes | 44 (18.26%) | 403 (14.69%) | 447 (14.97%) | 0.1364 |
| | No | 197 (81.74%) | 2341 (85.31%) | 2538 (85.03%) | |
| Born of U.S. | Born in 50 US states or DC | 196 (81.33%) | 2159 (78.68%) | 2355 (78.89%) | 0.3343 |
| | Others | 45 (18.67%) | 585 (21.32%) | 630 (21.11%) | |
| **Citizenship status** | Yes | 232 (89.13%) | 2514 (91.62%) | 2746 (91.99%) | **0.0108** |
| | No | 9 (10.87%) | 230 (8.38%) | 239 (8.01%) | |
| Education level | Less than 9th grade | 24 (9.96%) | 277 (10.09%) | 301 (10.08%) | 0.5383 |
| | 9-11th grade | 42 (17.43%) | 390 (14.21%) | 432 (14.47%) | |
| | High school graduate/GED or equivalent | 59 (24.48%) | 621 (22.63%) | 680 (22.78%) | |
| | Some college or AA degree | 63 (26.14%) | 760 (27.7%) | 823 (27.57%) | |
| | College graduate or above | 53 (21.99%) | 696 (25.36%) | 749 (25.09%) | |
| **Marital status** | Married | 151 (62.66%) | 1600 (58.31%) | 1751 (58.66%) | **0.0015** |
| | Widowed | 24 (9.96%) | 250 (9.11%) | 274 (9.18%) | |
| | Divorced | 38 (15.77%) | 312 (11.37%) | 350 11.73%) | |
| | Separated | 9 (3.73%) | 82 (2.99%) | 91 (3.05%) | |
| | Never married | 16 (6.64%) | 345 (12.57%) | 361 (12.09%) | |
| | Living with partner | 3 (1.24%) | 155 (5.65%) | 158 (5.29%) | |

Table 5. Chi-square test analysis for 2007-2008 (continued)

| **Factors** | **Factors Level** | **Pre-diabetes**<br>241 (8.07%) | **Normal**<br>2744 (91.93%) | **Total**<br>N=2985 | **P-values** |
|---|---|---|---|---|---|
| Total number of people in family | 1 | 55 (22.82%) | 567 (20.66%) | 622 (20.84%) | 0.0762 |
| | 2 | 97 (40.25%) | 894 (32.58%) | 991 (33.2%) | |
| | 3 | 34 (14.11%) | 464 (16.91%) | 498 (16.68%) | |
| | 4 | 24 (9.96%) | 425 (15.49%) | 449 (15.04%) | |
| | 5 | 13 (5.39%) | 205 (7.47%) | 218 (7.3%) | |
| | 6 | 9 (3.73%) | 99 (3.61%) | 108 (3.62%) | |
| | 7 or more people in the Family | 9 (3.73%) | 90 (3.28%) | 99 (3.32%) | |
| Annual family income | $ 0 to $ 4,999 | 3 (1.24%) | 54 (1.97%) | 57 (1.91%) | 0.6322 |
| | $ 5,000 to $ 9,999 | 8 (3.32%) | 114 (4.15%) | 122 (4.09%) | |
| | $10,000 to $14,999 | 13 (5.39%) | 185 (6.74%) | 198 (6.63%) | |
| | $15,000 to $19,999 | 19 (7.88%) | 200 (7.29%) | 219 (7.34%) | |
| | $20,000 to $24,999 | 22 (9.13%) | 226 (8.24%) | 248 (8.31%) | |
| | $25,000 to $34,999 | 38 (15.77%) | 323 (11.77%) | 361 (12.09%) | |
| | $35,000 to $44,999 | 26 (10.79%) | 266 (9.69%) | 292 (9.78%) | |
| | $45,000 to $54,999 | 16 (6.64%) | 213 (7.76%) | 229 (7.67%) | |
| | $55,000 to $64,999 | 14 (5.81%) | 166 (6.05%) | 180 (6.03%) | |
| | $65,000 to $74,999 | 11 (4.56%) | 147 (5.36%) | 158 (5.29%) | |
| | $20,000 and Over | 9 (3.73%) | 101 (3.68%) | 110 (3.69%) | |
| | Under $20,000 | 6 (2.49%) | 28 (1.02%) | 34 (1.14%) | |
| | $75,000 to $99,999 | 22 (9.13%) | 281 (10.24%) | 303 (10.15%) | |
| | $100,000 and Over | 34 (14.11%) | 440 (16.03%) | 474 (15.88%) | |
| **Ever told have health risk for diabetes** | Yes | 99 (41.08%) | 316 (11.52%) | 415 (13.9%) | **<0.0001** |
| | No | 142 (58.92%) | 2428 (88.48%) | 2570 (86.1%) | |
| **Smoked** at least 100 cigarettes in life | Yes | 129 (53.53%) | 1264 (46.06%) | 1393 (46.67%) | **0.0260** |
| | No | 112 (46.47%) | 1480 (53.94%) | 1592 (53.33%) | |
| Physical activity | Less than 3 days Moderate activities a week | 174 (72.2%) | 2063 (75.18%) | 2237 (74.94%) | 0.3056 |
| | More than 3 days Moderate activities a week | 67 (27.8%) | 681 (24.82%) | 748 (25.06%) | |
| **High cholesterol level** | Yes | 145 (60.17%) | 1094 (39.87%) | 1239 (41.51%) | **<0.0001** |
| | No | 96 (39.83%) | 1650 (60.13%) | 1746 (58.49%) | |
| **Hypertension status** | High Blood Pressure | 142 (58.92%) | 954 (34.77%) | 1096 (36.72%) | **<0.0001** |
| | Borderline Hypertension | 16 (6.64%) | 72 (2.62%) | 88 (2.95%) | |
| | No | 83 (34.44%) | 1718 (62.61%) | 1801 (60.34%) | |
| **Diet** | Excellent | 14 (5.81%) | 263 (9.58%) | 277 (9.28%) | **0.0031** |
| | Very Good | 54 (22.41%) | 726 (26.46%) | 780 (26.13%) | |
| | Good | 99 (41.08%) | 1106 (40.31%) | 1205 (40.37%) | |
| | Fair | 52 (21.58%) | 529 (19.28%) | 581 (19.46%) | |
| | Poor | 22 (9.13%) | 120 (4.37%) | 142 (4.76%) | |

Table 6. Chi-square test analysis for 2009-2010

| **Factors** | **Factors Level** | **Pre-diabetes** | **Normal** | **Total** | **P-values** |
|---|---|---|---|---|---|
| | | 314 (9.35%) | 3043 (90.65%) | N=3357 | |
| **Age** | 20-39 years old | 45 (14.33%) | 824 (27.08%) | 869 (25.89%) | **<0.0001** |
| | 40-59 years old | 113 (35.99%) | 1138 (37.4%) | 1251 (37.27%) | |
| | 60-79 years old | 125 (39.81%) | 859 (28.23%) | 984 (29.31%) | |
| | 80 years and older | 31 (9.87%) | 222 (7.3%) | 253 (7.54%) | |
| **Body mass index** | Underweight (<18.5 kg/$m^2$) | 3 (0.96%) | 41 (1.35%) | 44 (1.31%) | **<0.0001** |
| | Normal (18.51-25 kg/$m^2$) | 45 (14.33%) | 808 (26.55%) | 853 (25.41%) | |
| | Overweight (25.01-30kg/$m^2$) | 92 (29.3%) | 1074 (35.29%) | 1166 (34.73%) | |
| | Obese (>30 kg/$m^2$) | 174 (55.41%) | 1120 (36.81%) | 1294 (38.55%) | |
| Gender | Male | 144 (45.986) | 1391 (45.71%) | 1535 (45.73%) | 0.9599 |
| | Female | 170 (54.14%) | 1652 (54.29%) | 1822 (54.27%) | |
| Race | Mexican American | 39 (12.42%) | 412 (13.54%) | 451 (13.43%) | 0.8450 |
| | Other Hispanic | 24 (7.64%) | 269 (8.84%) | 293 (8.73%) | |
| | Non-Hispanic White | 176 (56.05%) | 1684 (55.34%) | 1860 (55.41%) | |
| | Non-Hispanic Black | 56 (17.83%) | 525 (17.25%) | 581 (17.31%) | |
| | Other Race | 19 (6.05%) | 153 (5.03%) | 172 (5.12%) | |
| **US armed forces** | Yes | 59 (18.79%) | 422 (13.87%) | 481 (14.33%) | **0.0178** |
| | No | 255 (81.21%) | 2621 (86.13%) | 2876 (85.67%) | |
| Born of U.S. | Born in 50 US states or DC | 249 (79.3%) | 2330 (76.57%) | 2579 (76.82%) | 0.2750 |
| | Others | 65 (20.7%) | 713 (23.43%) | 778 (23.18%) | |
| Citizenship status | Yes | 287 (91.4%) | 2714 (89.19%) | 3001 (89.4%) | 0.2253 |
| | No | 27 (8.6%) | 329 (10.81%) | 356 (10.6%) | |
| Education level | Less than 9th grade | 32 (10.19%) | 260 (8.54%) | 292 (8.7%) | 0.3116 |
| | 9-11th grade | 45 (14.33%) | 408 (13.41%) | 453 (13.49%) | |
| | High school graduate/GED or equivalent | 70 (22.29%) | 652 (21.43%) | 722 (21.51%) | |
| | Some college or AA degree | 98 (31.21%) | 887 (29.15%) | 985 (29.34%) | |
| | College graduate or above | 69 (21.97%) | 836 (27.47%) | 905 (26.96%) | |
| Marital status | Married | 183 (58.28%) | 1734 (56.98%) | 1917 (57.1%) | 0.1621 |
| | Widowed | 30 (9.55%) | 268 (8.81%) | 298 (8.88%) | |
| | Divorced | 47 (14.97%) | 357 (11.73%) | 404 (12.03%) | |
| | Separated | 9 (2.87%) | 90 (2.96%) | 99 (2.95%) | |
| | Never married | 27 (8.6%) | 408 (13.41%) | 435 (12.96%) | |
| | Living with partner | 18 (5.73%) | 186 (6.11%) | 204 (6.08%) | |

Table 6. Chi-square test analysis for 2009-2010 (continued)

| **Factors** | **Factors Level** | **Pre-diabetes** | **Normal** | **Total** | **P-values** |
|---|---|---|---|---|---|
| | | 314 (9.35%) | 3043 (90.65%) | N=3357 | |
| **Total number of people in family** | 1 | 72 (22.93%) | 657 (21.59%) | 729 (21.72%) | **0.0042** |
| | 2 | 125 (39.81%) | 906 (29.77%) | 1031 (30.71%) | |
| | 3 | 41 (13.06%) | 443 (14.56%) | 484 (14.42%) | |
| | 4 | 31 (9.87%) | 458 (15.05%) | 489 (14.57%) | |
| | 5 | 24 (7.64%) | 318 (10.45%) | 342 (10.19%) | |
| | 6 | 8 (2.55%) | 118 (3.88%) | 126 (3.75%) | |
| | 7 or more people in the Family | 13 (4.14%) | 143 (4.7%) | 156 (4.65%) | |
| Annual family income | $ 0 to $ 4,999 | 7 (2.23%) | 79 (2.6%) | 86 (2.56%) | 0.3780 |
| | $ 5,000 to $ 9,999 | 20 (6.37%) | 123 (4.04%) | 143 (4.26%) | |
| | $10,000 to $14,999 | 25 (7.96%) | 225 (7.39%) | 250 (7.45%) | |
| | $15,000 to $19,999 | 16 (5.1%) | 182 (5.98%) | 198 (5.9%) | |
| | $20,000 to $24,999 | 22 (7.01%) | 216 (7.1%) | 238 (7.09%) | |
| | $25,000 to $34,999 | 44 (14.01%) | 333 (10.94%) | 377 (11.23%) | |
| | $35,000 to $44,999 | 25 (7.96%) | 290 (9.53%) | 315 (9.38%) | |
| | $45,000 to $54,999 | 26 (8.28%) | 257 (8.45%) | 283 (8.43%) | |
| | $55,000 to $64,999 | 18 (5.73%) | 207 (6.8%) | 225 (6.7%) | |
| | $65,000 to $74,999 | 14 (4.46%) | 145 (4.77%) | 159 (4.74%) | |
| | $20,000 and Over | 18 (5.73%) | 104 (3.42%) | 122 (3.63%) | |
| | Under $20,000 | 2 (0.64%) | 29 (0.95%) | 31 (0.92%) | |
| | $75,000 to $99,999 | 27 (8.6%) | 323 (10.61%) | 350 (10.43%) | |
| | $100,000 and Over | 50 (15.92%) | 530 (17.42%) | 580 (17.28%) | |
| **Ever told have health risk for diabetes** | Yes | 117 (37.26%) | 316 (10.38%) | 433 (12.9%) | **<0.0001** |
| | No | 197 (62.74%) | 2727 (89.62%) | 2924 (87.1%) | |
| **Smoked** at least 100 cigarettes in life | Yes | 169 (53.82%) | 1321 (43.41%) | 1490 (44.38%) | **0.0004** |
| | No | 145 (46.18%) | 1722 (56.59%) | 1867 (55.62%) | |
| Physical activity | Less than 3 days Moderate activities a week | 259 (82.48%) | 2508 (82.42%) | 2767 (82.42%) | 0.9769 |
| | More than 3 days Moderate activities a week | 55 (17.52%) | 535 (17.58%) | 790 (17.58%) | |
| **High cholesterol level** | Yes | 180 (57.32%) | 1141 (37.5%) | 1321 (39.35%) | **<0.0001** |
| | No | 134 (42.68%) | 1902 (62.5%) | 2036 (60.65%) | |
| **Hypertension status** | High Blood Pressure | 195 (62.1%) | 1108 (36.41%) | 1303 (38.81%) | **<0.0001** |
| | Borderline Hypertension | 26 (8.28%) | 128 (4.21%) | 154 (4.59%) | |
| | No | 93 (29.62%) | 1807 (59.38%) | 1900 (56.6%) | |

Table 6. Chi-square test analysis for 2009-2010 (continued)

| **Factors** | **Factors Level** | **Pre-diabetes** | **Normal** | **Total** | **P-values** |
|---|---|---|---|---|---|
| | | 314 (9.35%) | 3043 (90.65%) | N=3357 | |
| **Diet** | Excellent | 23 (7.32%) | 291 (9.56%) | 314 (9.35%) | **0.0166** |
| | Very Good | 83 (26.43%) | 686 (22.54%) | 769 (22.91%) | |
| | Good | 111 (35.35%) | 1316 (43.25%) | 1427 (42.51%) | |
| | Fair | 80 (25.48%) | 628 (20.64%) | 708 (21.09%) | |
| | Poor | 17 (5.41%) | 122 (4.01%) | 139 (4.14%) | |

Table 7. Chi-square test analysis for 2011-2012

| **Factors** | **Factors Level** | **Pre-diabetes** | **Normal** | **Total** | **P-values** |
|---|---|---|---|---|---|
| | | 322 (7.47%) | 3990 (92.53%) | N=4312 | |
| **Age** | 20-39 years old | 50 (15.53%) | 1676 (42.01%) | 1726 (40.02%) | **<0.0001** |
| | 40-59 years old | 125 (38.82%) | 1306 (32.73%) | 1431 (33.19%) | |
| | 60-79 years old | 124 (38.51%) | 816 (20.45%) | 940 (21.80%) | |
| | 80 years and older | 23 (7.14%) | 192 (4.81%) | 215 (4.99%) | |
| **Body mass index** | Underweight (<18.5 kg/$m^2$) | 0 | 93 (2.33%) | 93 (2.16%) | **<0.0001** |
| | Normal (18.51-25 kg/$m^2$) | 58 (18.01%) | 1349 (33.81%) | 1407 (32.63%) | |
| | Overweight (25.01-30kg/$m^2$) | 94 (29.19%) | 1300 (32.58%) | 1394 (32.33%) | |
| | Obese (>30 kg/$m^2$) | 170 (52.8%) | 1248 (31.28%) | 1418 (32.88%) | |
| **Gender** | Male | 132 (40.99%) | 1972 (49.42%) | 2104 (48.79%) | **0.0036** |
| | Female | 190 (59.01%) | 2018 (50.58%) | 2208 (51.21%) | |
| Race | Mexican American | 31 (9.63%) | 380 (9.52%) | 441 (9.53%) | 0.0193 |
| | Other Hispanic | 27 (8.39%) | 404 (10.13%) | 431 (10%) | |
| | Non-Hispanic White | 113 (35.09%) | 1542 (38.65%) | 1655 (38.38%) | |
| | Non-Hispanic Black | 106 (32.92%) | 982 (24.61%) | 1088 (25.23%) | |
| | Other Race | 45 (13.98%) | 682 (17.09%) | 727 (16.86%) | |
| **US armed forces** | Yes | 46 (14.29%) | 342 (8.57%) | 388 (9%) | **0.0006** |
| | No | 276 (85.71%) | 3648 (91.43%) | 3924 (91%) | |
| Born of U.S. | Born in 50 US states or DC | 235 (72.98%) | 2766 (69.32%) | 3001 (69.6%) | 0.1699 |
| | Others | 87 (27.02%) | 1224 (30.68%) | 1311 (30.4%) | |
| **Citizenship status** | Yes | 287 (89.13%) | 3378 (84.66%) | 3665 (85%) | **0.0308** |
| | No | 35 (10.87%) | 612 (15.34%) | 647 (15%) | |
| Education level | Less than 9th grade | 29 (9.01%) | 308 (7.72%) | 337 (7.82%) | 0.1948 |
| | 9-11th grade | 32 (9.94%) | 533 (13.36%) | 565 (13.1%) | |
| | High school graduate/GED or equivalent | 66 (20.5%) | 830 (20.8%) | 896 (20.78%) | |
| | Some college or AA degree | 115 (35.71%) | 1235 (30.95%) | 1350 (31.31%) | |
| | College graduate or above | 80 (24.84%) | 1084 (27.17%) | 1164 (26.99%) | |

Table 7. Chi-square test analysis for 2011-2012 (continued)

| **Factors** | **Factors Level** | **Pre-diabetes** | **Normal** | **Total** | **P-values** |
|---|---|---|---|---|---|
| | | 322 (7.47%) | 3990 (92.53%) | N=4312 | |
| **Marital status** | Married | 179 (55.59%) | 1928 (48.32%) | 2107 (48.86%) | **0.0006** |
| | Widowed | 22 (6.83%) | 252 (6.32%) | 274 (6.35%) | |
| | Divorced | 42 (13.04%) | 382 (9.57%) | 424 (9.83%) | |
| | Separated | 14 (4.35%) | 138 (3.46%) | 152 (3.53%) | |
| | Never married | 47 (14.06%) | 936 (23.46%) | 983 (22.8%) | |
| | Living with partner | 18 (5.59%) | 354 (8.87%) | 372 (8.63%) | |
| **Total number of people in family** | 1 | 69 (21.43%) | 958 (24.01%) | 1027 (23.82%) | **0.0246** |
| | 2 | 103 (31.99%) | 957 (23.98%) | 1060 (24.58%) | |
| | 3 | 58 (18.01%) | 661 (15.57%) | 719 (16.67%) | |
| | 4 | 44 (13.66%) | 686 (17.19%) | 730 (16.93%) | |
| | 5 | 28 (8.7%) | 380 (9.52%) | 408 (9.46%) | |
| | 6 | 13 (4.04%) | 182 (4.56%) | 195 (4.52%) | |
| | 7 or more people in the Family | 7 (2.17%) | 166 (4.16%) | 173 (4.01%) | |
| **Annual family income** | $ 0 to $ 4,999 | 7 (2.17%) | 175 (4.39%) | 182 (4.22%) | **0.0369** |
| | $ 5,000 to $ 9,999 | 16 (4.97%) | 227 (5.69%) | 243 (5.64%) | |
| | $10,000 to $14,999 | 30 (9.32%) | 347 (8.7%) | 377 (8.74%) | |
| | $15,000 to $19,999 | 26 (8.07%) | 296 (7.42%) | 322 (7.47%) | |
| | $20,000 to $24,999 | 24 (7.45%) | 311 (7.79%) | 335 (7.77%) | |
| | $25,000 to $34,999 | 33 (10.25%) | 455 (11.4%) | 488 (11.32%) | |
| | $35,000 to $44,999 | 39 (12.11%) | 378 (9.47%) | 417 (9.67%) | |
| | $45,000 to $54,999 | 12 (3.73%) | 287 (7.19%) | 299 (6.93%) | |
| | $55,000 to $64,999 | 25 (7.76%) | 187 (4.69%) | 212 (4.92%) | |
| | $65,000 to $74,999 | 18 (5.59%) | 199 (4.99%) | 217 (5.039%) | |
| | $20,000 and Over | 11 (3.42%) | 103 (2.58%) | 114 (2.64%) | |
| | Under $20,000 | 1 (0.31%) | 43 (1.08%) | 44 (1.02%) | |
| | $75,000 to $99,999 | 34 (10.56%) | 321 (8.05%) | 355 (8.23%) | |
| | $100,000 and Over | 46 (14.29%) | 661 (16.57%) | 707 (16.4%) | |
| **Ever told have health risk for diabetes** | Yes | 133 (41.3%) | 435 (10.9%) | 568 (13.17%) | **<0.0001** |
| | No | 189 (58.7%) | 3555 (89.1%) | 3744 (86.83%) | |
| Smoked at least 100 cigarettes in life | Yes | 138 (42.86%) | 1657 (41.53%) | 1795 (41.63%) | 0.6418 |
| | No | 184 (57.14%) | 2333 (58.47%) | 2517 (58.37%) | |

Table 7. Chi-square test analysis for 2011-2012 (continued)

| **Factors** | **Factors Level** | **Pre-diabetes**<br>322 (7.47%) | **Normal**<br>3990 (92.53%) | **Total**<br>N=4312 | **P-values** |
|---|---|---|---|---|---|
| Physical activity | Less than 3 days Moderate activities a week | 253 (78.57%) | 3269 (81.93%) | 3522 (81.68%) | 0.1340 |
| | More than 3 days Moderate activities a week | 69 (21.43%) | 721 (18.07%) | 790 (18.32%) | |
| **High cholesterol level** | Yes | 192 (59.63%) | 1085 (29.17%) | 1277 (29.62%) | **<0.0001** |
| | No | 130 (40.37%) | 2905 (72.81%) | 3035 (70.38%) | |
| **Hypertension status** | High Blood Pressure | 188 (58.39%) | 1149 (28.8%) | 1337 (31.01%) | **<0.0001** |
| | Borderline Hypertension | 18 (5.59%) | 151 (3.78%) | 169 (3.92%) | |
| | No | 116 (36.02%) | 2690 (67.42%) | 2806 (65.07%) | |
| **Diet** | Excellent | 26 (8.07%) | 410 (10.28%) | 436 (10.11%) | **0.0072** |
| | Very Good | 65 (20.19%) | 858 (21.5%) | 923 (21.41%) | |
| | Good | 134 (41.61%) | 1731 (43.38%) | 1865 (43.25%) | |
| | Fair | 68 (21.12%) | 809 (20.28%) | 877 (20.34%) | |
| | Poor | 29 (9.01%) | 182 (4.56%) | 211 (4.89%) | |

## Results of Logistic Regression

In binary logistic method, using ‘forward LR’ seven variables in the logistic regression equation were illustrated in table 8. Except age and body mass index are continuous variables, the others 11 factors are degree variables, Dummy variable should be used in logistic regression model. For hypertension status, high blood pressure=**1**; borderline hypertension=**2**; No=**3**. According to the results, the three logistic regression equations show similar results for five common variables: risk for diabetes, hypertension, high cholesterol level, age, and body mass index. These variables played an important role in the incidence of pre-diabetes in this research, as shown in table 8.

Table 8. The importance of the 7 input variables in three logistic regression models

| Order | **2007-2008** | **2009-2010** | **2011-2012** |
|---|---|---|---|
| 1 | Risk for diabetes | Risk for diabetes | Risk for diabetes |
| 2 | Hypertension | Hypertension | Age |
| 3 | Age | High Cholesterol Level | High Cholesterol Level |
| 4 | Body Mass Index | Body Mass Index | Body Mass Index |
| 5 | High Cholesterol Level | Age | Hypertension |
| 6 | Marital status | Diet | Marital status |
| 7 | Diet | Smoked | Gender |

In order to see the importance of these variables are presented in table 8, where each variable is placed in order of its relative importance. To some extent, the three models have some consistency on the first five factors. If enter methods were used to test all variables, the p-value of the variable “diet” was equal to 0.0915 (close to 0.05) in the data from 2011-2012. The p-value of the variable “smoked” was equal to 0.0612 in the data from 2009-2010, but when variable “smoked” was entered to the logistic regression equation the whole p-value was less than 0.05. For the same reason, the p-value of variable “gender” was equal to 0.0658 in the data from 2011-2012, therefore “gender” was entered to the logistic regression equation.

Based on the analysis, “enter” and “forward” methods were applied to gain the receiver operating characteristic (ROC) curve respectively. The value of the area under the ROC curve (AUC), which measures the accuracy of the predictive model, ranges from 0 (0%) to 1 (100%), in which a higher value indicates a higher accuracy. The “forward” method was used to identify the accuracy of the logistic regression modeling for each 2-year cycle of data that included the 7 variables as shown in table 8. More specifically, the first area under the ROC curve for the data from 2007-2008 was 0.7796, the second area under the ROC curve for the data from 2009-2010 was 0.7696. The third area under the ROC curve for the data from 2011-2012 was 0.8125. Figure 3-5 shows the three comparisons of area under ROC curve.

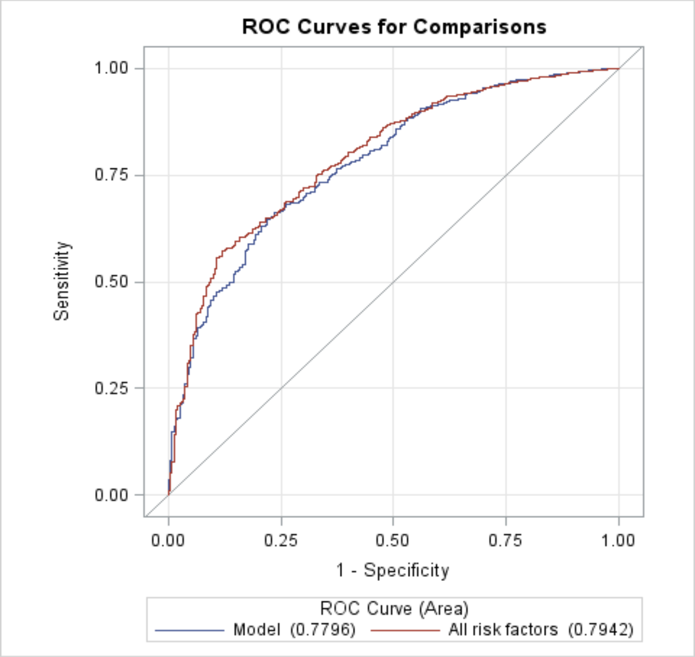


Figure 3. Comparison of AUC for logistic regression model in 2007-2008

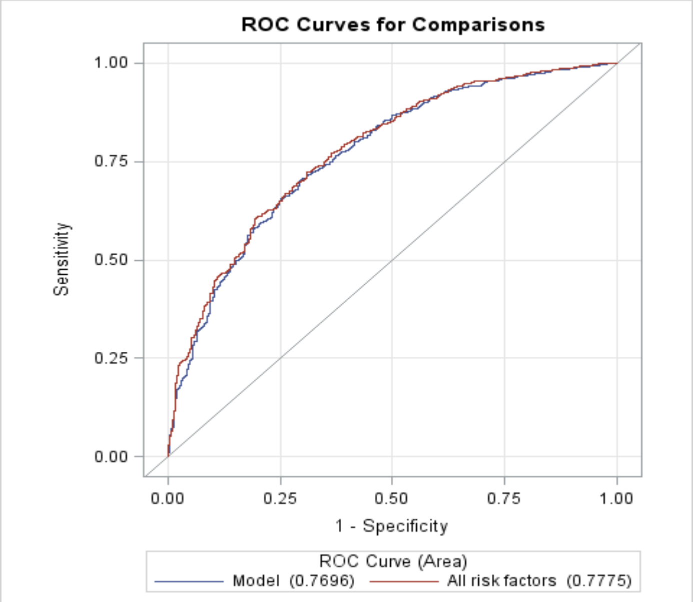


Figure 4. Comparison of AUC for logistic regression model in 2009-2010

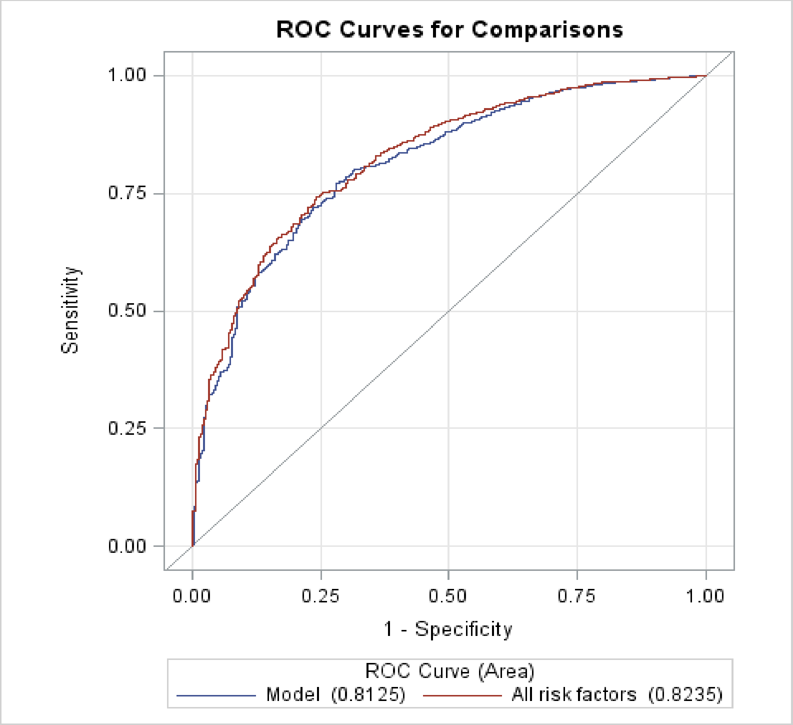


Figure 5. Comparison of AUC for logistic regression model in 2011-2012

Comparing the ROC curves of the model with only seven variables to the model with all risk factors, the value of AUC for all risk factors is slightly higher. The ROC statistics are compared in figure 6. Since some of factors are not significant, the logistic regression equation can be simplified by using the model with only seven variables. Because all ROC index greater than 76%, the logistic regression modeling using the "forward" method performs well at predict pre-diabetes.

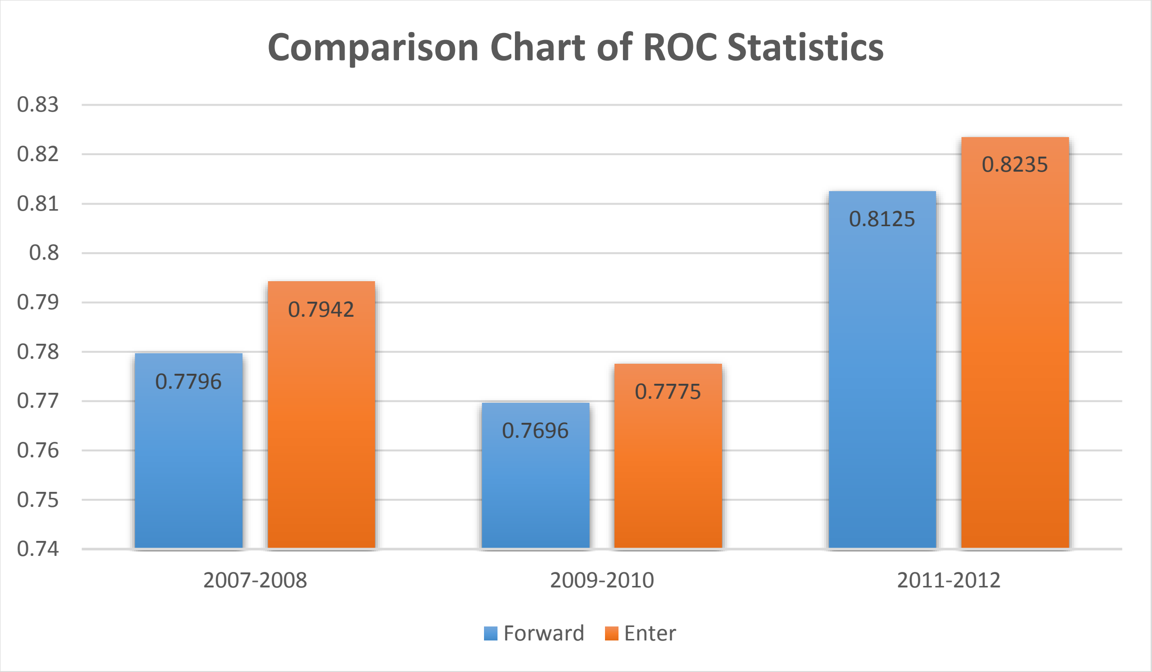


Figure 6. Comparison chart of logistic regression ROC statistics

According to the results of the forward logistic regression, table 9 illustrates the estimated coefficients $\boldsymbol{\beta}$ to predict the risk factors for pre-diabetes. The forward selection approach was applied to develop the logistic regression model (Equation 1), meanwhile the probability of pre-diabetes could be obtained from equation 2. All model variables meet the significance level of p<0.05. The estimator coefficients $\boldsymbol{\beta}$ for each 2-year cycle (e.g. $\boldsymbol{\beta}_{\boldsymbol{A}}$)from table 9 are used with equation 2, to estimate the probability that a person has pre-diabetes was gained. In order to clearly see the equations, table 9 was made, five common risk factors (age, risk for diabetes, high cholesterol level, hypertension, and BMI) were highlighted for all three datasets.

Table 9. Variables in the logistic regression equation

| | Parameter | 2007-2008<br>Estimate $\boldsymbol{\beta_A}$ | 2009-2010<br>Estimate $\boldsymbol{\beta_B}$ | 2011-2012<br>Estimate $\boldsymbol{\beta_C}$ |
|---|---|---|---|---|
| 0 | Intercept | 6.4636 | 5.3374 | 6.4510 |
| 1 | Gender (Male) | | | 0.2806 |
| | Gender (Female) | | | 0 |
| 2 | Age (20 years of age or older) | -0.0230 | -0.0186 | -0.0341 |
| 3 | Marital status (Married) | -1.4735 | | -0.1234 |
| | Marital status (Widowed) | -1.0596 | | 0.9272 |
| | Marital status (Divorced) | -1.4648 | | -0.0795 |
| | Marital status (Separated) | -1.6062 | | -0.1687 |
| | Marital status (Never married) | -0.9302 | | -0.1002 |
| | Marital status (Living with partner) | 0 | | 0 |
| 4 | Risk for diabetes (Yes) | -1.6940 | -1.5715 | -1.7608 |
| | Risk for diabetes (No) | 0 | 0 | 0 |
| 5 | Smoked at least 100 cigarettes in life (Yes) | | -0.2574 | |
| | Smoked at least 100 cigarettes in life (No) | | 0 | |
| 6 | High Cholesterol Level (Yes) | -0.5159 | -0.5033 | -0.8268 |
| | High Cholesterol Level (No) | 0 | 0 | 0 |
| 7 | Hypertension (Yes) | -0.6210 | -0.7303 | -0.5716 |
| | Hypertension (Borderline) | -1.2706 | -1.1078 | -0.6233 |
| | Hypertension (No) | 0 | 0 | 0 |
| 8 | Diet (Excellent) | 1.1530 | 0.4293 | |
| | Diet (Very good) | 0.8138 | -0.0418 | |
| | Diet (Good) | 0.8278 | 0.5633 | |
| | Diet (Fair) | 0.6736 | 0.1416 | |
| | Diet (Poor) | 0 | 0 | |
| 9 | Body Mass Index | -0.0378 | -0.0388 | -0.0405 |

### Results of Decision Tree

Using SAS Enterprise Miner version 13.1 for windows, two data mining models (logistic regression and decision tree) were performed and compared. According to three 2-year cycles of data shown in table 5-7, only less than one-tenth of total observations have pre-diabetes in raw data. These precise number are: 8.07%, 9.35%, and 7.47%. When the raw data were brought into the decision tree model, only a single leaf was generated, because the decision tree simply assumes everyone is normal with less than 10% error rate. In order to more accurately segment pre-diabetes based on the raw data, the proportion of the number of pre-diabetes in total raw data were increased compared to normal. The numbers of pre-diabetes in 2007-2008 and 2011-2012 have been hiked at 9 times and 2009-2010 has been hiked at 8 times. With this approach, less than 50% of observations were assumed to have pre-diabetes. All exact numbers shown in figure 7 following:

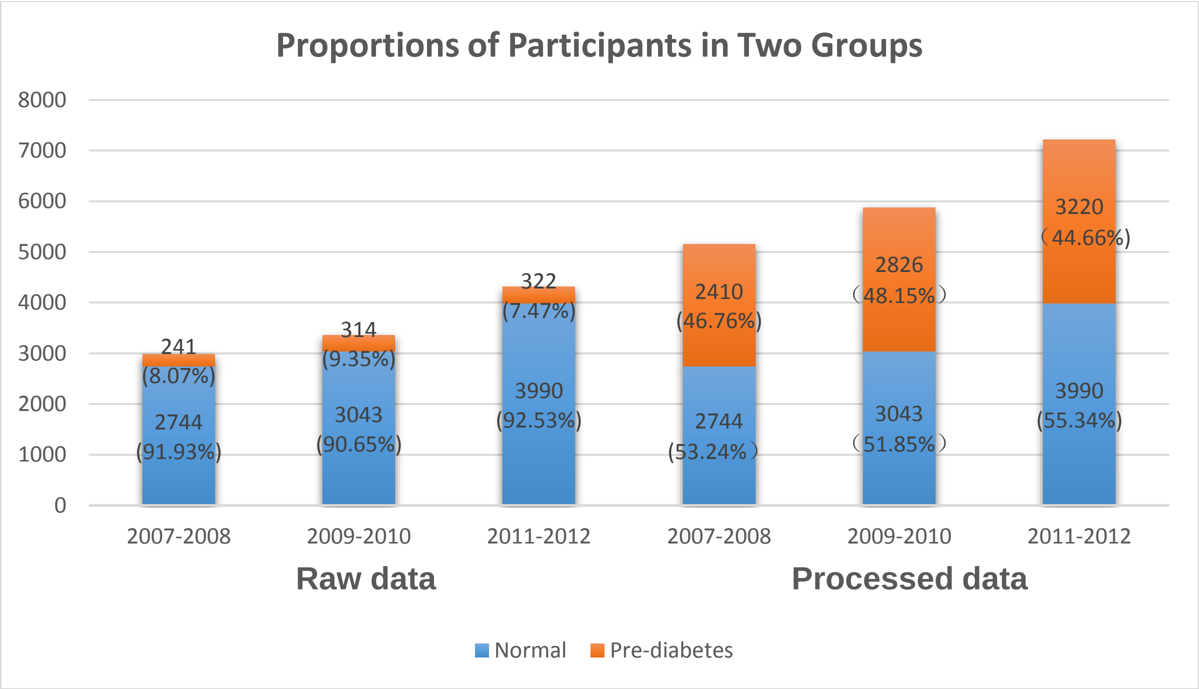


Figure 7. Proportions of participants in pre-diabetes and normal group

It displayed a nice graphical programming interface. As the figure 8 shows, each processed data node was connected with a decision tree partition node, then each processed data group was randomly divided into two parts, with 70% training data and 30% validation data. In each decision tree node, the maximum number of generations of nodes (maximum depth) was set as 10. In other words, the original generation node was the root node 0. In the next level, the generation was counted as 1. For decision tree node, in interactive sample part, sample size is

10000 and sample seed is 12345 as default value. In node part of decision tree, we set leaf size is 8, number of rules is 5, and number of surrogate rule is 4.

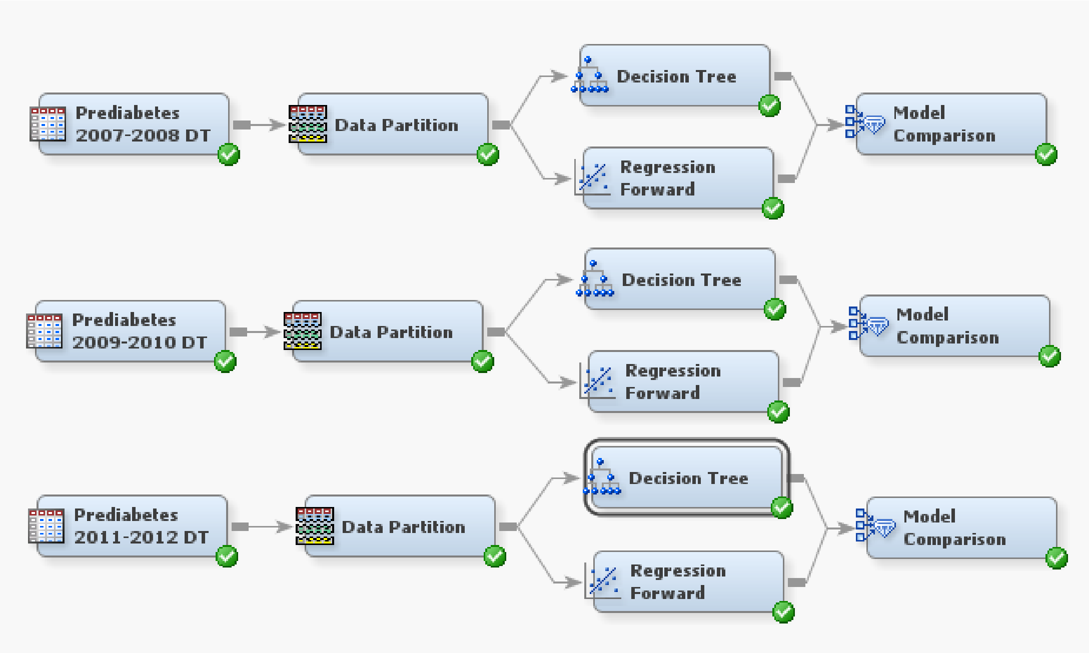


Figure 8. Comparison process flow diagram

Data from 2007-2008

From the output of the decision tree in data from 2007-2008, tree leaf report contains if-then logic to illustrate decision rules. Out of a total 81 nodes, eleven nodes (Nodes id: 25, 26, 28, 37, 51, 53, 63, 64, 67, 75, 79) predicted some groups (green box shown in figure 9) have lower risk with pre-diabetes. The percentage of training observations in the node with Pre-diabetes=0 was 1.00 (tested negative). Five nodes (red circle shown in figure 9) were took that predicted some groups have higher risk (>80%) with pre-diabetes, such as node 48, which predicted 8 observations with Pre-diabetes=1 was 1.00, as shown in the following code:

```
*------------------------------------------------------------*
If Total number Family is one of: 3 people in family
AND Smoked is one of: 2 means  Answer No for ‘Smoked at least 100 cigarettes in life’
```

AND Risk for diabetes is one of: 2 means Answer No for ‘Ever told have health risk for diabetes’

AND Diet is one of: 2 (Good=**2** for answer how healthy is the diet)

AND Annual Family Income is one of: 3, 5, 2, 7 ($ 5,000 to $ 9,999=**2**; $10,000 to $14,999=**3,** $20,000 to $24,999=**5**; $35,000 to $44,999=**7**)

AND Age < 41.5

Then

Tree Node Identifier = **48**

Number of Observations = 8

Predicted: Pre-diabetes=1 = 1.00

Predicted: Pre-diabetes=0 = 0.00

*------------------------------------------------------------*

If Risk for diabetes is one of: 1 means Answer Yes for ‘Ever told have health risk for diabetes’

AND Marital Status is one of: 5, 6 (Never married=**5**; Living with partner=**6**)

AND Hypertension is one of: 1, 2 (High Blood Pressure=**1**; Borderline Hypertension=**2**)

AND Annual Family Income is one of: 14, 5, 7 ($20,000 to $24,999=**5**; $35,000 to $44,999=**7,** $75,000 to $99,999=14)

AND Age >= 29.5

Then

Tree Node Identifier = **34**

Number of Observations = 25

Predicted: Pre-diabetes=1 = 0.92

Predicted: Pre-diabetes=0 = 0.08

*------------------------------------------------------------*

If Risk for diabetes is one of: 1 means Answer Yes for ‘Ever told have health risk for diabetes’

AND Hypertension is one of: 1, 2 (High Blood Pressure=**1**; Borderline Hypertension=**2**)

AND Body Mass Index >= 39.2

AND Age < 29.5

Then

Tree Node Identifier = **29**

Number of Observations = 8

Predicted: Pre-diabetes=1 = 0.88

Predicted: Pre-diabetes=0 = 0.13

*---------------------------------------------------------*

If Risk for diabetes is one of: 1 means Answer Yes for ‘Ever told have health risk for diabetes’

AND Marital Status is one of: 1, 3, 4, 2 (Married=**1**; Widowed=**2**; Divorced=**3**; Separated=**4**)

AND Hypertension is one of: 1, 2 (High Blood Pressure=**1**; Borderline Hypertension=**2**)

AND Age >= 29.5

Then

Tree Node Identifier = 31

Number of Observations = 524

Predicted: Pre-diabetes=1 = 0.87

Predicted: Pre-diabetes=0 = 0.13

*---------------------------------------------------------*

If Risk for diabetes is one of: 2 means Answer No for ‘Ever told have health risk for diabetes’

AND Hypertension is one of: 1, 2 (High Blood Pressure=**1**; Borderline Hypertension=**2**)

AND Diet is one of: 5 (Poor=**5**)

AND Age >= 41.5

Then

Tree Node Identifier = **22**

Number of Observations = 109

Predicted: Pre-diabetes=1 = 0.82

Predicted: Pre-diabetes=0 = 0.18

Data from 2009-2010

From the output of the decision tree in data from 2009-2010, tree leaf report contains if-then logic to illustrate decision rules. Out of a total 107 nodes, ten nodes (Nodes id: 32, 35, 67, 71, 73, 75, 77, 85, 87, 92) predicted some groups (green box shown in figure 10) have lower risk with pre-diabetes. The percentage of training observations in the node with Pre-diabetes=0 was 1.00 (tested negative). Six nodes (red circle shown in figure 10) were took that predicted some groups have higher risk (>80%) with pre-diabetes, such as node 15, which predicted 416 observations with Pre-diabetes=1 was 1.00, as shown in the following code:

*-------------------------------------------------------------*

If Risk for diabetes is one of: 1 means Answer Yes for ‘Ever told have health risk for diabetes’

AND Hypertension is one of: 2, 1 (High Blood Pressure=**1**; Borderline Hypertension=**2**)

AND Annual Family Income is one of: 15, 6, 12, 5, 3, 2 ($ 5,000 to $ 9,999=**2**; $10,000 to $14,999=**3**; $20,000 to $24,999=**5**; $25,000 to $34,999=**6**; $20,000 and Over=**12**; $100,000 and Over=**15**)

Then

Tree Node Identifier = **15**

Number of Observations = 416

Predicted: Pre-diabetes=1 = 0.90

Predicted: Pre-diabetes=0 = 0.10

*-------------------------------------------------------------*

If Risk for diabetes is one of: 1 means Answer Yes for ‘Ever told have health risk for diabetes’

AND Hypertension is one of: 3 means Answer No Hypertension Status

AND High Cholesterol is one of: 2 means Answer No High Cholesterol Status

AND Annual Family Income is one of: 12, 4, 8, 2, 1 ($ 0 to $ 4,999=**1**; $ 5,000 to $ 9,999=**2**; $15,000 to $19,999=**4**; $45,000 to $54,999=**8**; $20,000 and Over=**12**)

AND Age >= 36.5

Then

Tree Node Identifier = **45**

Number of Observations = 65

Predicted: Pre-diabetes=1 = 0.97

Predicted: Pre-diabetes=0 = 0.03

*-------------------------------------------------------------*

If Total number in Family is one of: 2, 1, 7 (1, 2, or 7 people in family)

AND Risk for diabetes is one of: 1 means Answer Yes for ‘Ever told have health risk for diabetes’

AND Hypertension is one of: 2, 1 (High Blood Pressure=**1**; Borderline Hypertension=**2**)

AND Body Mass Index >= 31.545

AND Annual Family Income is one of: 14, 4, 7, 9, 10, 8, 1 ($ 0 to $ 4,999=**1**; $15,000 to $19,999=**4**; $35,000 to $44,999=**7**; $45,000 to $54,999=**8**; $55,000 to $64,999=**9**; $65,000 to $74,999=**10**; $75,000 to $99,999=**14**)

Then

Tree Node Identifier = **49**

Number of Observations = 136

Predicted: Pre-diabetes=1 = 0.88

Predicted: Pre-diabetes=0 = 0.12

*------------------------------------------------------------*

If Smoked is one of: 1 means 'Smoked at least 100 cigarettes in life'

AND Risk for diabetes is one of: 2 means Answer No for 'Ever told have health risk for diabetes'

AND Hypertension is one of: 3 means Answer No Hypertension Status

AND Body Mass Index >= 37.25

AND Annual Family Income is one of: 15, 6, 14, 4, 7 ($15,000 to $19,999=**4**; $25,000 to $34,999=**6**; $35,000 to $44,999=**7**; $75,000 to $99,999=**14**; $100,000 and Over=**15**)

AND Age >= 49.5

Then

Tree Node Identifier = **51**

Number of Observations = 34

Predicted: Pre-diabetes=1 = 0.94

Predicted: Pre-diabetes=0 = 0.06

*------------------------------------------------------------*

If Risk for diabetes is one of: 1 means Answer Yes for 'Ever told have health risk for diabetes'

AND Marital Status is one of: 1 (Married=**1**)

AND Hypertension is one of: 3 means Answer No Hypertension Status

AND High Cholesterol is one of: 2 means Answer No High Cholesterol Status

AND Body Mass Index >= 32.68

AND Annual Family Income is one of: 15, 6, 7 ($25,000 to $34,999=**6**; $35,000 to $44,999=**7**; $100,000 and Over=**15**)

Then

Tree Node Identifier = **80**

Number of Observations = 32

Predicted: Pre-diabetes=1 = 0.81

Predicted: Pre-diabetes=0 = 0.19

*------------------------------------------------------------*

If Total number in Family is one of: 3, 4, 5 or 1 people in family

AND Smoked is one of: 2 means Answer No for ‘Smoked at least 100 cigarettes in life’

AND Risk for diabetes is one of: 2 means Answer No for ‘Ever told have health risk for diabetes’

AND Hypertension is one of: 3 means Answer No Hypertension Status

AND High Cholesterol is one of: 1 means Answer Yes for ‘High Cholesterol Status’

AND Body Mass Index < 37.25

AND Annual Family Income is one of: 12, 14, 9, 10, 8, 2 ($ 5,000 to $ 9,999=**2**; $45,000 to $54,999=**8**; $55,000 to $64,999=**9**; $65,000 to $74,999=**10**; $20,000 and Over=**12**; $75,000 to $99,999=**14**)

AND Age >= 49.5

Then

Tree Node Identifier = **106**

Number of Observations = 66

Predicted: Pre-diabetes=1 = 0.85

Predicted: Pre-diabetes=0 = 0.15

Data from 2011-2012

From the output of the decision tree in data from 2011-2012, tree leaf report contains if-then logic to illustrate decision rules. Out of a total 99 nodes, ten nodes (Nodes id: 21, 47, 49, 62, 65, 67, 78, 87, 93, 94) predicted some groups (green box shown in figure 11) have lower risk with pre-diabetes. The percentage of training observations in the node with Pre-diabetes=0 is close to 1.00 (tested negative). Five nodes (red circle shown in figure 11) were took that predicted some groups have higher risk (>80%) with pre-diabetes, such as, node 15 predicted 512 observations with Pre-diabetes=1 is close to 1.00, as shown in the following code:

```
*------------------------------------------------------------*
If Risk for diabetes is one of: 1 means Answer Yes for ‘Ever told have health risk for diabetes’
AND Hypertension is one of: 1, 2 (High Blood Pressure=1; Borderline Hypertension=2)
AND High Cholesterol is one of: 1 means Answer Yes for ‘High Cholesterol Status’
Then
 Tree Node Identifier = 15
 Number of Observations = 512
 Predicted: Pre-diabetes=1 = 0.91
 Predicted: Pre-diabetes=0 = 0.09
*------------------------------------------------------------*
If Risk for diabetes is one of: 1 means Answer Yes for ‘Ever told have health risk for diabetes’
AND High Cholesterol is one of: 2 means Answer No for ‘High Cholesterol Status’
AND Annual family income is one of: 7, 15, 12, 2, 14, 10, 3, 9 ($ 5,000 to $ 9,999=2; $10,000 to
$14,999=3; $35,000 to $44,999=7; $55,000 to $64,999=9; $65,000 to $74,999=10; $20,000 and
Over=12; $75,000 to $99,999=14; $100,000 and Over=15)
AND Age >= 43.5 or MISSING
Then
 Tree Node Identifier = 26
 Number of Observations = 236
 Predicted: Pre-diabetes=1 = 0.81
 Predicted: Pre-diabetes=0 = 0.19
*------------------------------------------------------------*
If Risk for diabetes is one of: 1 means Answer Yes for ‘Ever told have health risk for diabetes’
AND Hypertension is one of: 3 means Answer No Hypertension Status
AND High Cholesterol is one of: 1 means Answer Yes for ‘High Cholesterol Status’
AND Age >= 48.5
Then
 Tree Node Identifier = 29
```

Number of Observations = 130

Predicted: Pre-diabetes=1 = 0.87

Predicted: Pre-diabetes=0 = 0.13

*------------------------------------------------------------*

If Risk for diabetes is one of: 2 means Answer No for ‘Ever told have health risk for diabetes’

AND Hypertension is one of: 3 means Answer No Hypertension Status

AND High Cholesterol is one of: 2 means Answer No for ‘High Cholesterol Status’

AND Body Mass Index >= 49.25

AND Annual family income is one of: 6, 7, 5, 15, 12, 2, 8, 1 ($ 0 to $ 4,999=**1**; $ 5,000 to $ 9,999=**2**; $20,000 to $24,999=**5**; $25,000 to $34,999=**6**; $35,000 to $44,999=**7**; $45,000 to $54,999=**8**; $20,000 and Over=**12**; $100,000 and Over=**15**)

AND Age >= 45.5 or MISSING

Then

Tree Node Identifier = **95**

Number of Observations = 8

Predicted: Pre-diabetes=1 = 1.00

Predicted: Pre-diabetes=0 = 0.00

*------------------------------------------------------------*

If US Armed Forces is one of: 1 means Answer Yes for ‘Served active duty in US armed forces’

AND Risk for diabetes is one of: 2 means Answer No for ‘Ever told have health risk for diabetes’

AND Body Mass Index < 24.4

AND Annual family income is one of: 6, 7, 5, 15, 2, 14, 4, 3, 9 ($ 5,000 to $ 9,999=**2**; $10,000 to $14,999=**3**; $15,000 to $19,999=**4**; $20,000 to $24,999=**5**; $25,000 to $34,999=**6**; $35,000 to $44,999=**7**; $55,000 to $64,999=**9**; $75,000 to $99,999=**14**; $100,000 and Over=**15**)

AND Age < 61.5 AND Age >= 45.5

Then

Tree Node Identifier = **99**

Number of Observations = 25

Predicted: Pre-diabetes=1 = 0.80

Predicted: Pre-diabetes=0 = 0.20

*-----------------------------------------------------------*

Based on the above of results, nine factors were mentioned in the 2007-2008 decision tree to indicate which people have a high risk of pre-diabetes, their risk of diabetes, age (cut-point: 41.5 and 29.5), BMI (cut-point: 39.2), diet, smoked, total number in family, annual family income, marital status, and hypertension. For 2009-2010, risk of diabetes, hypertension, high cholesterol, age (cut-point: 36.5 and 49.5), BMI (cut-point: 37.25, 31.545, and 32.68), annual family income, marital status, smoked, and total number family play an important role in predicting. The majority of seven predictors are significant to indicate some groups of people who at a high risk of pre-diabetes: risk of diabetes, hypertension, high cholesterol, age (cut-point: 43.5, 45.5<age<61.5, and 48.5), BMI (cut-point: 24.4 and 49.25), annual family income, US Armed Forces. In summary, following the conditions of these nodes pre-diabetes status could be determined. In order to see the details of each node from figures 9-11, appendix A was made. Furthermore, appendix B shown the tree leaf report of SAS Enterprise Miner for three sets of data.

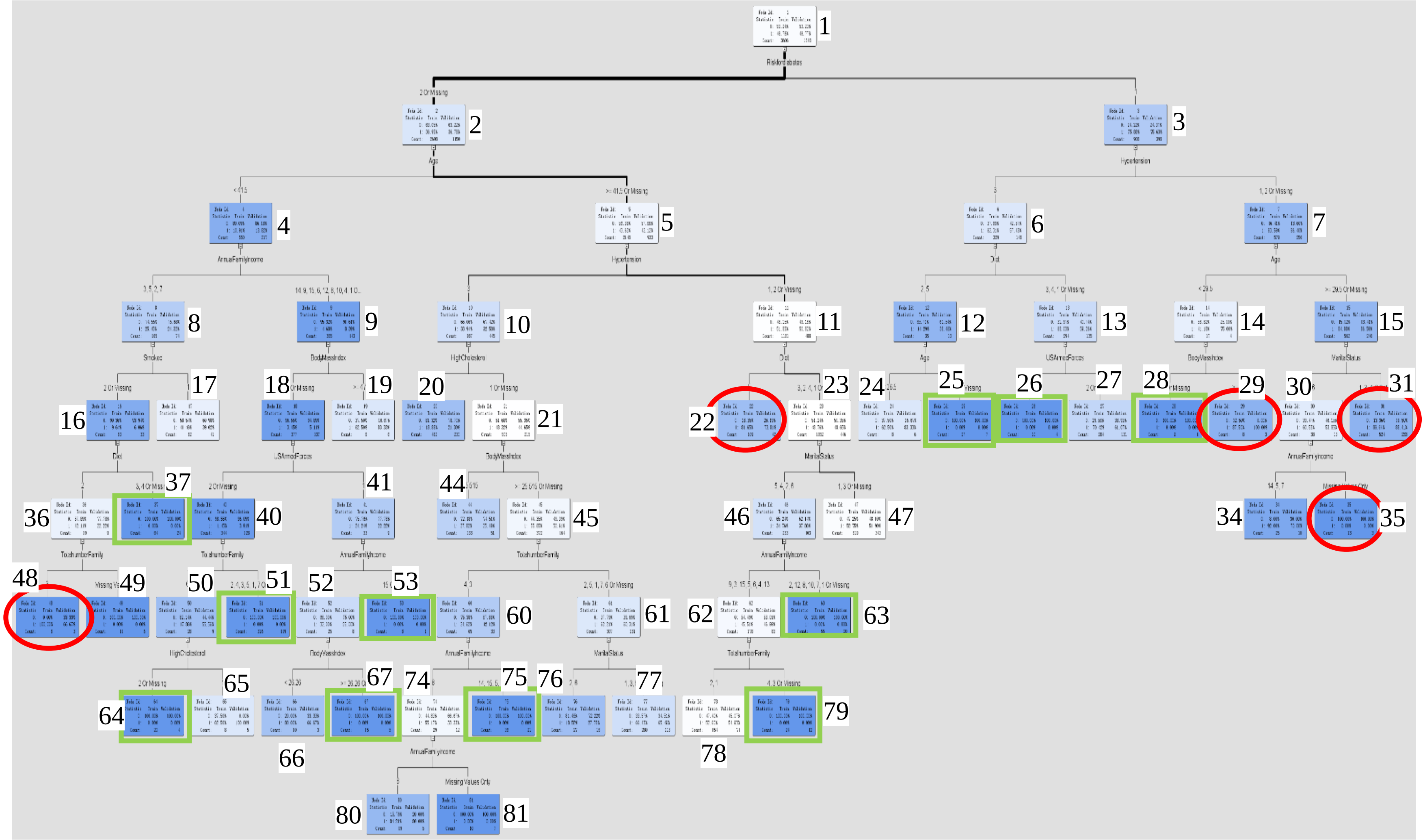


Figure 9. Decision tree for detecting pre-diabetes in 2007-2008's data

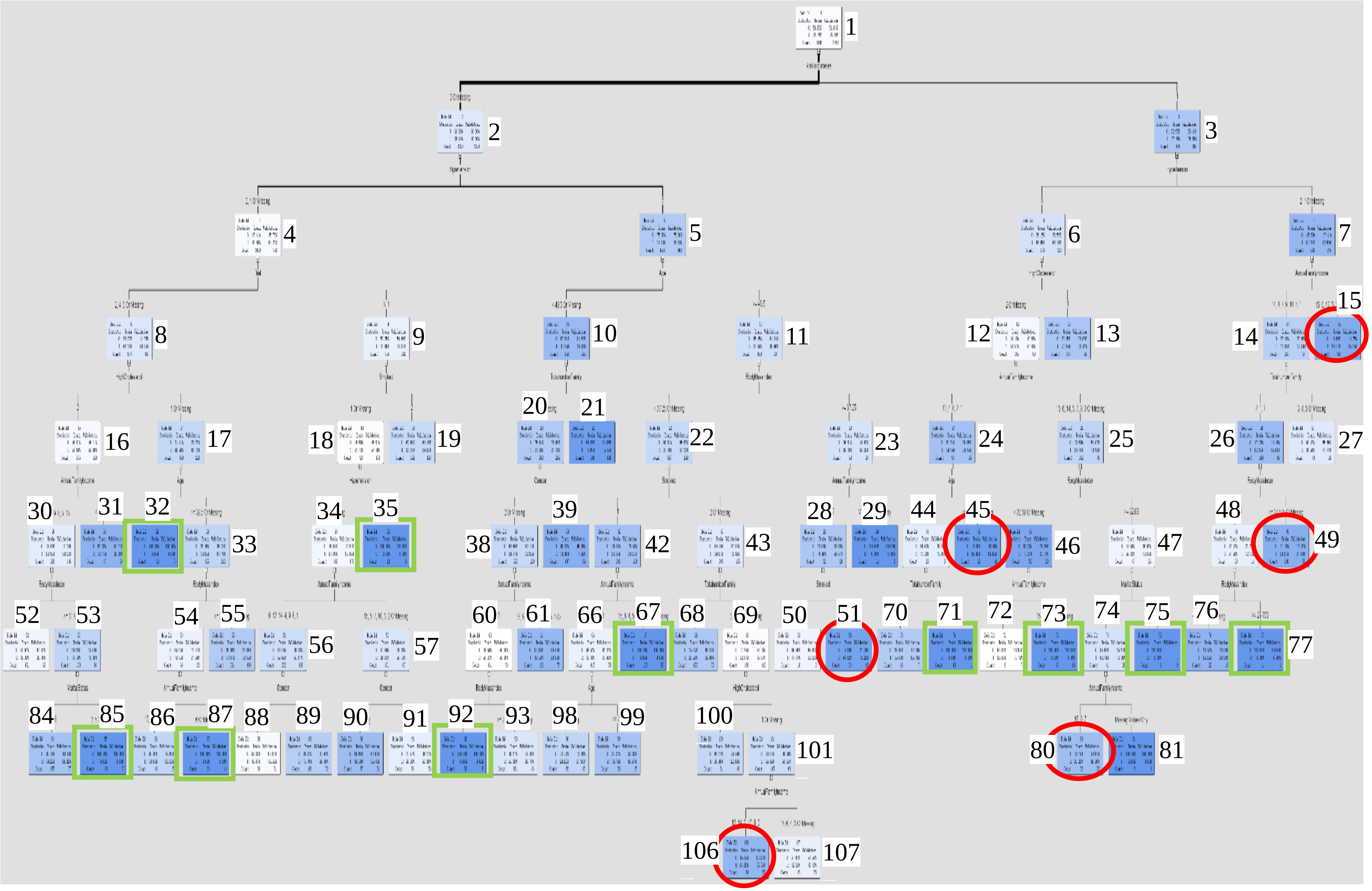


Figure 10. Decision tree for detecting pre-diabetes in 2009-2010's data

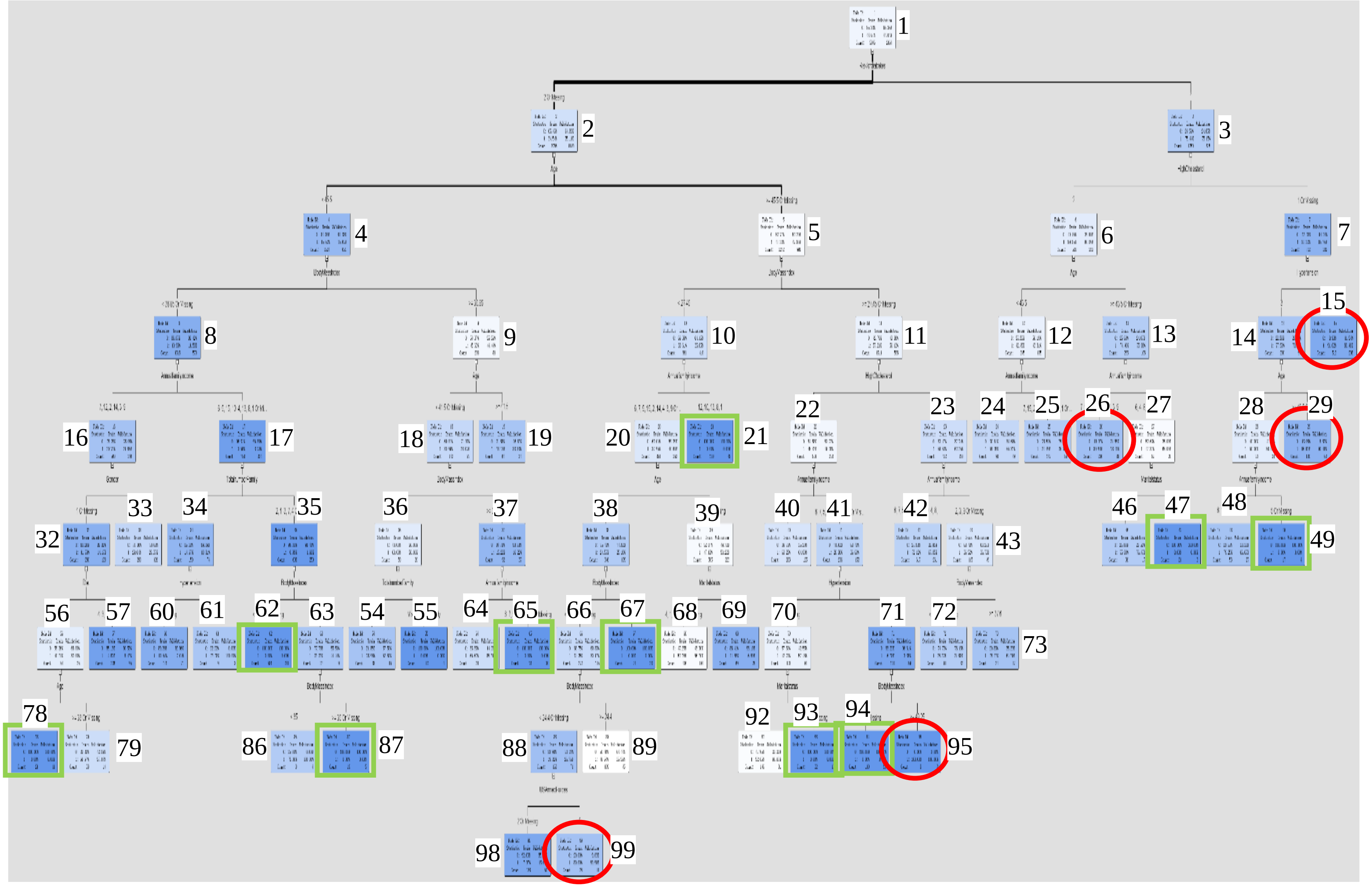


Figure 11. Decision tree for detecting pre-diabetes in 2011-2012’s data

## DISCUSSION THE RESULTS OF COMPARISON

Along with the cases of type 2 diabetes rapidly increasing, two data mining approaches (logistic regression and decision tree) are widely used to screen for pre-diabetes on order that more people should be predicted in early stage. As results section shown, risk of diabetes, hypertension, high cholesterol, age, and BMI play an important role in predicting for pre-diabetes. Except the obvious factor "risk for diabetes" was not considered as variables, the other four risk factors (hypertension, high cholesterol, age, and BMI) are in agreement with the most of the results of some previous studies.

As section 3 mentioned, for logistic regression models' literature reviews, 10 out of 51 papers (Schmidt et al, 2005, Borrell et al, 2007; Burke et al, 1999; Stern et al, 2002; Bang et al, 2009; Kanaya et al, 2005; Manson et al, 2000; Kahn et al, 2009; Wilson et al, 2007; Heikes et al, 2008) focused on U.S. population. Since Manson et al. (2000) only tested the relationship between cigarette smoking and the incidence of diabetes mellitus. Considering the other nine studies, a total of eight studies indicated age was considered as risk factor. But Willon et al. (2007) predicted the incident of diabetes mellitus in middle-aged individuals. It is the reason why variable "age" was not considered as risk factor in these two studies. In the other hand, for decision tree models' literature reviews, 3 out of 26 papers (Herman et al, 1995; Barriga et al, 1996; and Heikes et al, 2008) focused on U.S. population. All of these three studies indicated age was considered as risk factor in their models.

For variable "BMI", as previously mentioned, waist circumference is another indicator for BMI factor. So, BMI and waist circumference would be considered as one variable. Among the above nine studies, a total of six studies indicated BMI or waist circumference was considered as risk factor in their logistic regression models. But the other three studies, Burke et al. (1999), age, ethnic group, and neighborhood were significant (p<0.01) predictors of diabetes in their models. But the rising BMI was found to contribute significantly to the secular trend in diabetes incidence. Kanaya et al. (2005) indicated BMI was a poor marker for total adiposity in older adults. BMI was not independently associated with abnormal glucose intolerance for older adults age 71.3±9.8 years. Borrell et al. (2007) did not include BMI as a factor in their logistic regression modeling. In addition, for decision tree models' literature reviews, 3 out of 26 studies (Herman et al, 1995; Barriga et al, 1996; and Heikes et al, 2008) focused on U.S. population. All of these three papers indicated BMI (obesity, waist circumference, or weight) was considered a risk factor in their models.

For variable hypertension (blood pressure or systolic blood pressure), among the above nine studies, a total of seven studies indicated hypertension was considered as risk factor in their logistic regression models. But the other two studies, Burke et al. (1999) did not include blood pressure as a factor to predict the development of diabetes. Kanaya et al. (2005) predicted the development of diabetes in older adults. Comparing with the normal glucose tolerance, the abnormal glucose tolerance had a higher systolic blood pressure. But the diastolic blood pressure was no difference in two glucose subgroups. In addition, for decision tree models' literature reviews, only one study (Heikes et al, 2008) indicated hypertension was considered as risk factor in their model. Herman et al. (1995) compared two classification trees, the first classification tree by using simple questionnaire for community screening and the second classification tree also included hypertension. Their results shown these two classification trees have a same AUC value. Therefore, they just used the simple questionnaire not included hypertension to identify individuals at increased risk for undiagnosed diabetes. Barriga et al. (1996) mentioned hypertension was not selected as a risk indicator by using CART in their study.

Among the above nine studies, a total of four studies indicated high cholesterol was considered as a risk factor in their logistic regression models. But the other five studies, Burke et al. (1999), Kahn et al. (2009), and Bang et al. (2009) did not include high cholesterol as a factor to predict the development of diabetes. Kanaya et al. (2005) predicted the development of diabetes in older adults. Comparing with the normal glucose tolerance, HDL was no different in the two glucose subgroups. Heikes et al. (2008) build two detecting tools (logistic regression and CART) where variables of cholesterol were eliminated in models because of the large number of missing fields and low predictive value. In addition, for decision tree models' literature reviews, Herman et al. (1995) and Barriga et al. (1996) did not include high cholesterol as a factor to predict the development of diabetes. Heikes et al. (2008) eliminated the variables of cholesterol because of the large number of missing fields and low predictive value.

Based on the same processed data, decision tree (blue line) analysis has a higher ROC index than logistic regression (red line) in each of three sets of training and validation datasets. As figure 12 -14 shown, for training data of 2007-2008, the ROC indexes of the decision tree and the logistic regression were 83.7%, and 79.4%. For validation data of 2007-2008, the ROC indexes were 81.6% and 79.7%. For training data of 2009-2010, the ROC indexes of the decision tree and the logistic regression were 87%, and 78.4%. For validation data of 2009-2010, the ROC indexes were 83% and 76.4%. For training data of 2011-2012, the ROC indexes of the decision tree and the logistic regression were 88.5%, and 82.9%. For validation data of 2011-2012, the ROC indexes were 86.6% and 82.2%. According to the uniformity the results of training and validation data, they are all greater than 76%. According to the uniformity the results of training and validation data, in the following, whole dataset (training data 100%) were considered to compare and analyze the performance of two data mining methods.

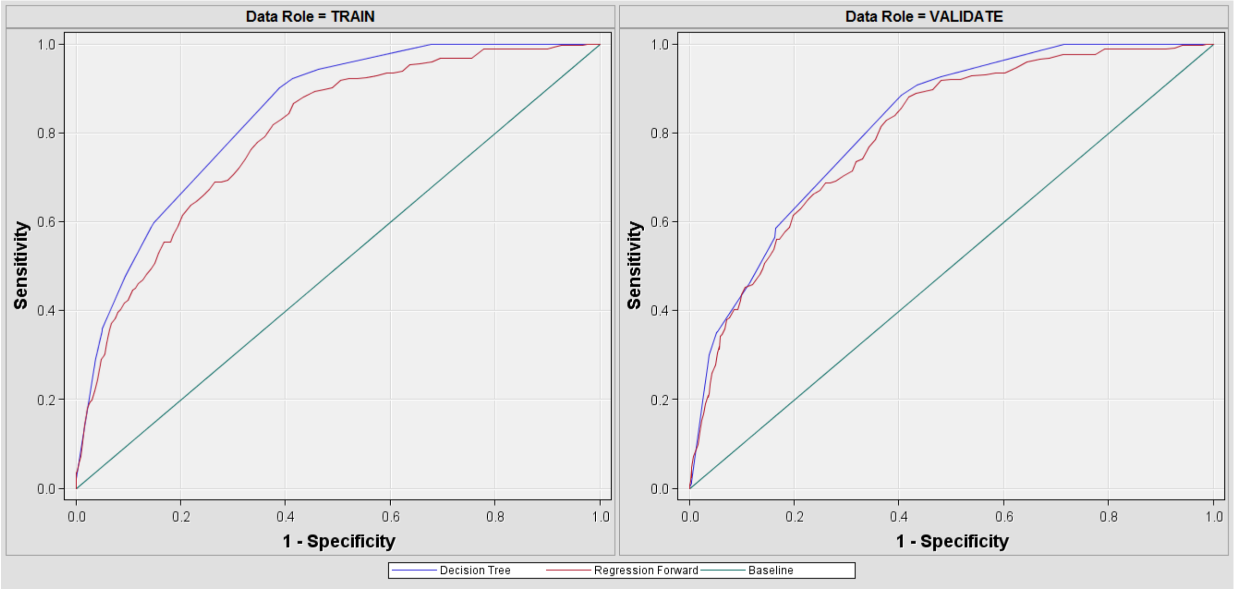


Figure 12. ROC curve of two models for 2007-2008 database

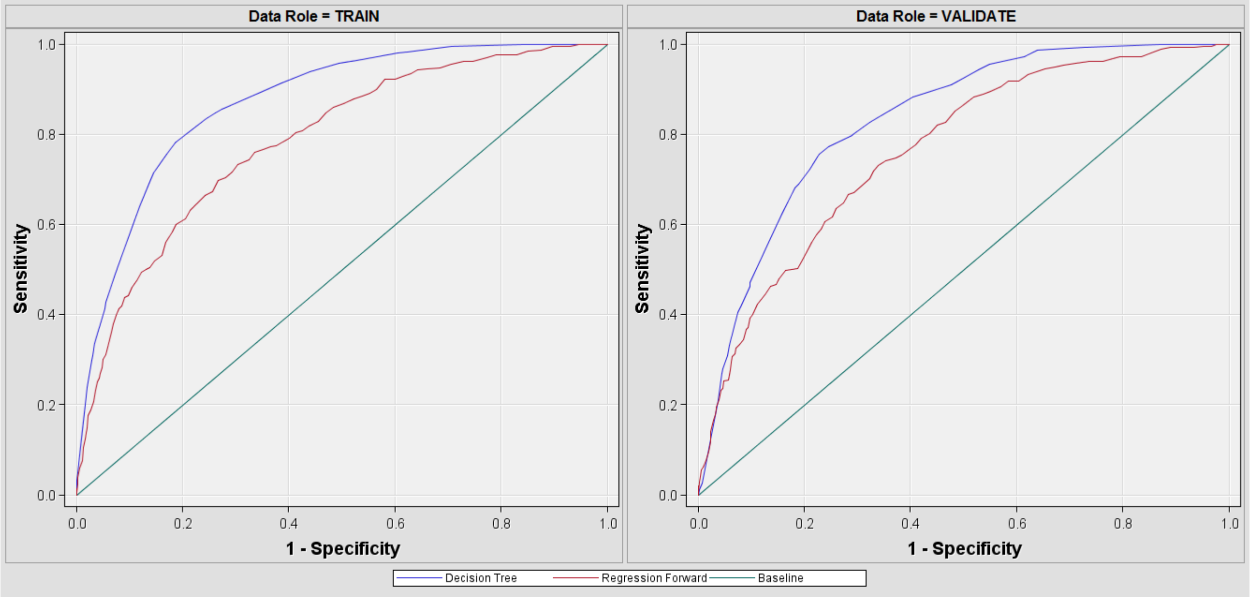


Figure 13. ROC curve of two models for 2009-2010 database

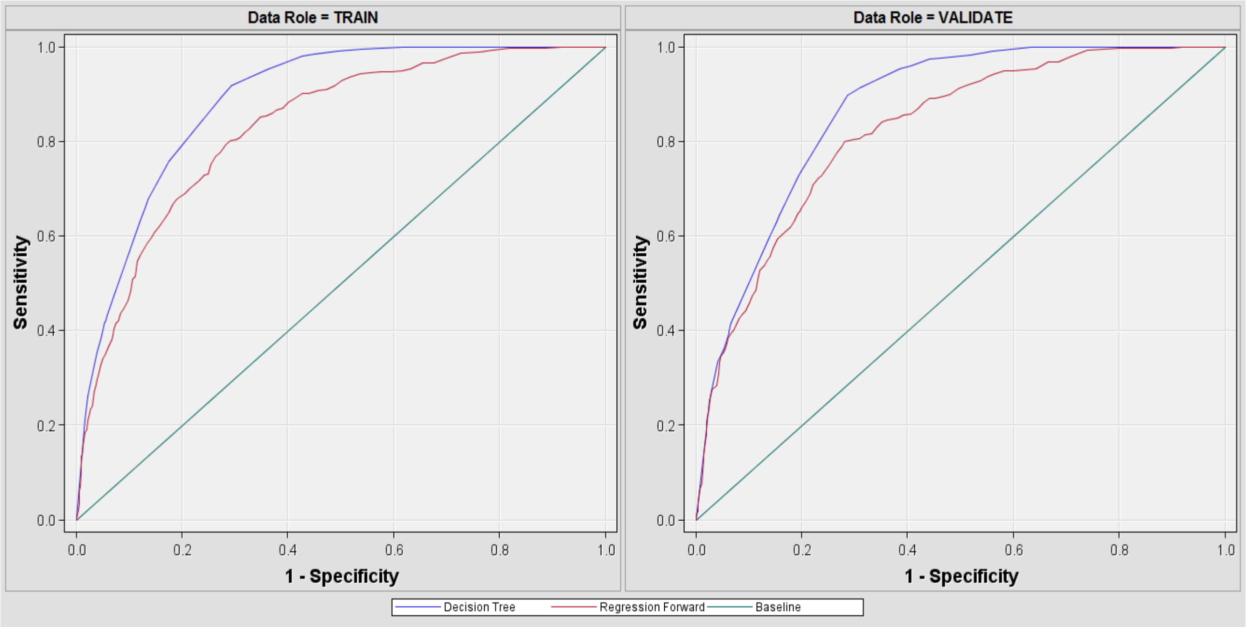


Figure 14. ROC curve of two models for 2011-2012 database

In addition to ROC values, in this study accuracy, sensitivity, specificity, also were used to evaluate the two models' performance. True positive (TP), true negative (TN), false positive (FP), and false negative (FN) were used to calculate accuracy, sensitivity, and specificity. The accuracy measures the percentage of correctly classified, which evaluate the predictive accuracy of the model. The formula takes the form: Accuracy = (TP+TN) / (TP+TN+FP+FN). Sensitivity also called the true positive rate, measures the percentage of positives correctly classified. Sensitivity is given by TP / (TP+FN). Specificity also called the true negative rate, measures the percentage of negatives correctly classified. The equation is Specificity = TN / (FP+TN). The details of four values (accuracy, sensitivity, specificity, and ROC) from training, validation, and whole dataset of each 2-year cycle were listed in table 10.

Table 10. Classification results and are under of ROC curve indices for two models

| | Logistic Regression | | | | | | Decision Tree | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2007-2008 | | 2009-2010 | | 2011-2012 | | 2007-2008 | | 2009-2010 | | 2011-2012 | |
| | Positive | Negative | Positive | Negative | Positive | Negative | Positive | Negative | Positive | Negative | Positive | Negative |
| Training dataset (70%) | | | | | | | | | | | | |
| True | 1111 | 1455 | 1379 | 1561 | 1579 | 2194 | 1519 | 1175 | 1546 | 1729 | 2013 | 2028 |
| False | 465 | 575 | 568 | 598 | 598 | 675 | 745 | 167 | 400 | 431 | 764 | 241 |
| Accuracy | 71.2% | | 71.6% | | 74.8% | | **74.7%** | | **79.8%** | | **80.1%** | |
| Sensitivity | 65.9% | | 69.8% | | 70.1% | | **90%** | | **78.2%** | | **89.3%** | |
| Specificity | **75.8%** | | 73.3% | | **78.6%** | | 61.2% | | **81.2%** | | 72.6% | |
| ROC | 79.4% | | 78.4% | | 82.9% | | **83.7%** | | **87%** | | **88.5%** | |
| Validation dataset (30%) | | | | | | | | | | | | |
| True | 479 | 629 | 565 | 656 | 651 | 948 | 641 | 491 | 641 | 706 | 867 | 854 |
| False | 195 | 245 | 258 | 284 | 250 | 315 | 333 | 83 | 208 | 208 | 344 | 99 |
| Accuracy | 71.6% | | 69.3% | | 73.9% | | **73.1%** | | **76.4%** | | **79.5%** | |
| Sensitivity | 66.1% | | 66.5% | | 67.4% | | **88.5%** | | **75.5%** | | **89.8%** | |
| Specificity | **76.3%** | | 71.8% | | **79.1%** | | 59.6% | | **77.2%** | | 71.3% | |
| ROC | 79.7% | | 76.4% | | 82.2% | | **81.6%** | | **83%** | | **86.6%** | |
| Whole dataset (training data 100%) | | | | | | | | | | | | |
| True | 1560 | 2091 | 1944 | 2198 | 2240 | 3136 | 2010 | 2215 | 2583 | 2235 | 3180 | 2983 |
| False | 653 | 850 | 845 | 882 | 854 | 980 | 529 | 400 | 808 | 243 | 1007 | 40 |
| Accuracy | 70.8% | | 70.6% | | 74.6% | | **82%** | | **82.1%** | | **85.5%** | |
| Sensitivity | 64.7% | | 68.8% | | 69.6% | | **83.4%** | | **91.4%** | | **98.8%** | |
| Specificity | 76.2% | | 72.2% | | **78.6%** | | **80.7%** | | **73.4%** | | 74.8% | |
| ROC | 79.8% | | 77.9% | | 82.8% | | **89.4%** | | **88.8%** | | **92%** | |

Based on the comparison between the models (figure 15) by using whole dataset, decision tree has a higher ROC index than logistic regression modeling. All ROC indexes (yellow box) for two data mining models were greater than 77% indicating both methods present a good prediction for pre-diabetes. The predictive accuracy of logistic regression modeling are greater than 70% completely on the whole three datasets, and the predictive accuracy of decision tree modeling are great then 83% completely on the whole three datasets. From the sensitivity analysis, the sensitivity of decision tree modeling are greater than 82%, and the sensitivity of logistic regression modeling are greater than 64%. For specificity analysis, the values of two data mining techniques are greater than 72%. In short, decision tree have the higher ROC indexes, accuracy and sensitivity values than logistic regression modeling. Except 2011-2012, specificity values of the decision tree are greater than logistic regression. Taken as a whole, the results of comparison indicated decision tree modeling is a better indicator to predict pre-diabetes.

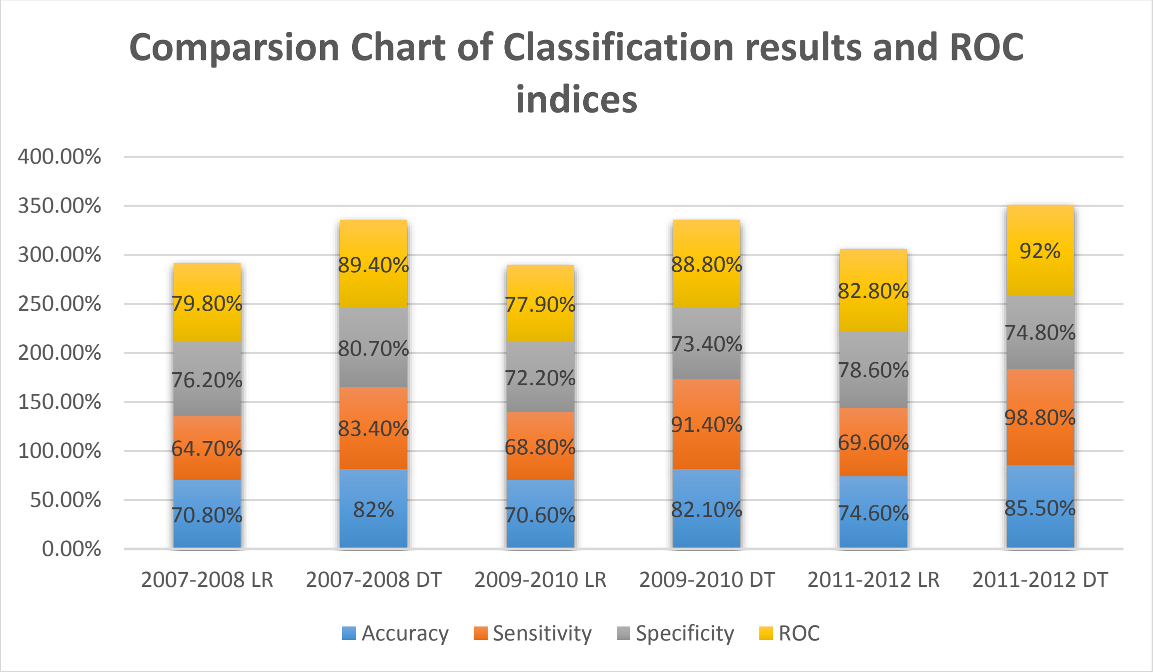


Figure 15. Comparison chart of classification results and ROC indices from whole dataset

# CONCLUSION

In conclusion, logistic regression and decision tree models were applied to assess the risk of pre-diabetes. Based on the data of National Health and Nutrition Examination Survey (NHANES), a chi-square test was used to select which variables are significant for predicting pre-diabetes. In this step, 13 risk factors (age, BMI, gender, served active duty in US armed forces, citizenship status, marital status, total number of people in the family, annual family income, ever told have health risk for diabetes, smoked, high cholesterol level, hypertension status, and diet) were chosen from 17 total variables of raw datasets. The two data mining approaches analyze the three sets of databases (2007-2008, 2009-2010, and 2011-2012) that include these 13 factors. As it was previously mentioned in this study, section of results of decision tree (page 48), that less than 10% of people suffer from pre-diabetes and in order to better fit the decision tree model, the proportion of patients with pre-diabetes were expanded from a total of 10% to 9 or 10 times that values. Having considered the changes of proportion, it appears that the effect of significant predictors were also expanded by about ten times. In summary, the final results indicated that decision tree modeling performed better on ROC indexes, accuracy and sensitivity.

Using the processed data, logistic regression and decision tree models were applied in SAS Enterprise miner 13.1. As research studies previously mentioned (Meng et al, 2013; Heikes et al, 2008), the results show that the decision tree is a better indicator to predict pre-diabetes. But considering that the data have been changed for the purposes of this study, comparison of logistic regression by using raw data and processed data showed that they have a similar ROC value as shown in figure 16.

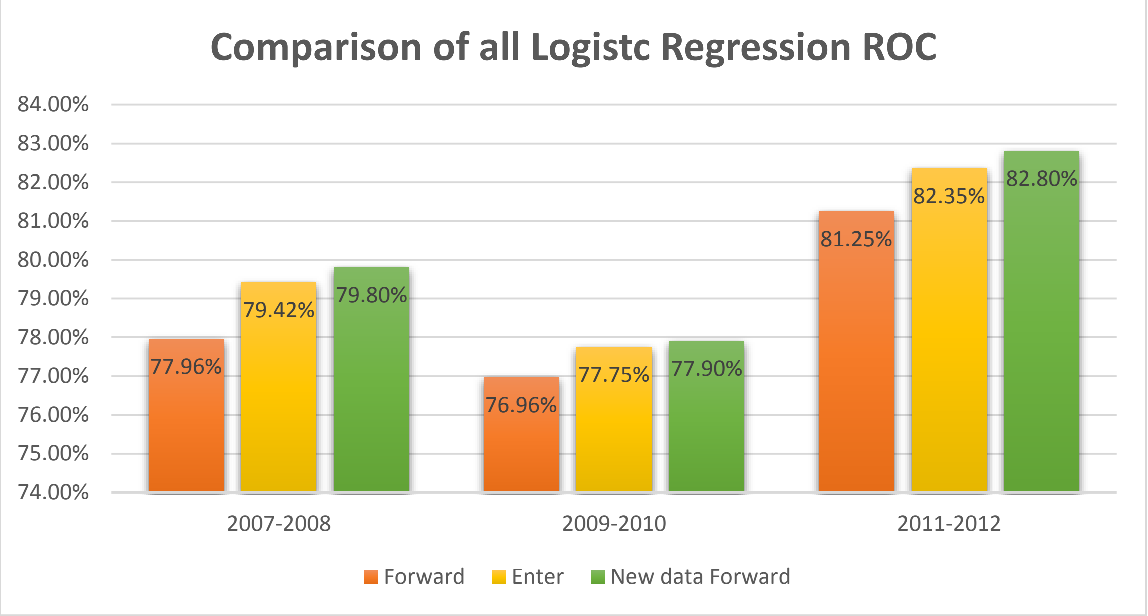


Figure 16. Comparison chart of all logistic regression ROC indices

On the other hand, considering the limit of the diabetes dataset, factors such as sleeping and drinking variables may be in included in these models. In addition, future studies should including factors such as fast glucose tolerance test, casual glucose tolerance test, diastolic blood pressure, serum insulin, triceps skin thickness, as part of minimal modeling to predict diabetes. As previously mentioned in the equation 3 (page 33), if the dataset followed a group population during several years, Cox proportional hazards model could be used. In additional, population modeling and Markova modeling could also be used.

According to the results of these two models (logistic regression and decision tree), for U.S. populations, five common risk factors (age, BMI, risk for diabetes, hypertension status, and high cholesterol levels) have significant association with pre-diabetes. The obvious factor “risk for diabetes” could be excluded, then age, BMI, hypertension status, and high cholesterol levels played a key role in predicting pre-diabetes. As previously mentioned in table 1 “The type 2 diabetes screening criteria”, a total of six guidelines (ADA, WHO, HIS, VA/DoD, ICSI and IDF) indicated the variable “BMI ≥25 kg/$m^2$” was the risk factor for screening diabetes. A total of five guidelines indicated age as the predictor for screening diabetes, from which, ADA, VA/DoD and IDF recommend individuals age ≥45 years should be screened, WHO suggested individuals age ≥35 years should be screened, and CMS individuals age ≥65 years should be screened. There are 8 guidelines indicating variable “hypertension (>140/90)” was the risk factors for screening diabetes. And VA/DoD indicated the variable “HDL cholesterol <40 mg/dl” was the risk factors for screening diabetes, and the other seven guidelines indicating the variable “HDL cholesterol <35 mg/dl” was the predictors for screening diabetes.

As expected, age, gender, race are factors that patients can’t change, but individuals can sharply lower their chances of developing the diabetes through modest weight loss and lowering their BMI with more physical activities. Everyone could follow these variables in models to determine their status of pre-diabetes. In particular, if more people could consider early prevention or the prompt management of pre-diabetes then it will help to effectively reduce the total costs and lost work and wages for people with diagnosed diabetes. In the meantime, it could help to reduce the incidence of diabetes nationwide.

# REFERENCES

1. Aekplakorn W, Bunnag P, Woodward M, Sritara P, Cheepudomwit S, Yamwong S, Yipintsoi T, Rajatanavin R, "A risk score for predicting incident diabetes in the Thai population," *Diabetes Care*, 2006, 29, pp. 1872-1877.

2. Al Khalaf MM, Eid MM, Najjar HA, Alhajry KM, Doi SA, Thalib L, "Screening for diabetes in Kuwait and evaluation of risk scores," *East Mediterr Health J*, 2010, 16, pp. 725-731.

3. Al-Lawati JA, Tuomilehto J, "Diabetes risk score in Oman: a tool to identify prevalent type 2 diabetes among Arabs of the Middle East," *Diabetes Res Clin Pract,* 2007, 77, pp. 438-444.

4. Al Jarullah, Asma A, "Decision tree discovery for the diagnosis of type II diabetes," *Innovations in Information Technology (IIT), 2011, International Conference on*. IEEE, 2011.

5. Amatul, Zehra, et al., "A Comparative Study on the Pre-Processing and Mining of Pima Indian Diabetes Dataset," *ICSEC 2014 FSKKP*, 2013, pp. 1-10.

6. American Diabetes Association, "Economic costs of diabetes in the US in 2012," *Diabetes care*, 2013, *36*(4), pp. 1033-1046.

7. American Diabetes Association, "Standards of medical care in diabetes—2014," *Diabetes care* 37, Supplement 1, 2014, S14-S80.

8. Baan CA, Ruige JB, Stolk RP, Witteman JCM, Dekker JM, Heine RJ, Feskens EJM, "Performance of a predictive model to identify undiagnosed diabetes in a health care setting," *Diabetes Care*, 1999, 22, pp. 213-219.

9. Balkau B, Lange C, Fezeu L, Tichet J, de Lauzon-Guillain B, Czernichow S, Fumeron F, Froguel P, Vaxillaire M, Cauchi S, Ducimeti ère P, Eschw ège E, "Predicting diabetes: clinical,

biological, and genetic approaches: data from the Epidemiological Study on the Insulin Resistance Syndrome (DESIR)," *Diabetes Care*, 2008, 31, pp. 2056-2061.

10. Bang H, Edwards AM, Bomback AS, Ballantyne CM, Brillon D, Callahan MA, Teutsch SM, Mushlin AI, Kern LM, "Development and validation of a patient self-assessment score for diabetes risk," *Ann Intern Med*, 2009, 151, pp. 775-783.
11. Barriga, Katherine J., et al., "Population screening for glucose intolerant subjects using decision tree analyses," *Diabetes research and clinical practice* 34, 1996, S17-S29.
12. Bindraban NR, van Valkengoed IGM, Mairuhu G, Holleman F, Hoekstra JBL, Michels BPJ, Koopmans RP, Stronks K, "Prevalence of diabetes mellitus and the performance of a risk score among Hindustani Surinamese, African Surinamese and ethnic Dutch: a cross-sectional population-based study," *BMC Public Health*, 2008, 8, p. 271.
13. Bonora, E., Kiechl, S., Willeit, J., Oberhollenzer, F., Egger, G., Meigs, J. B., ... & Muggeo, M., "Population-based incidence rates and risk factors for type 2 diabetes in White individuals: The Bruneck Study," *Diabetes*, 2004, *53*(7), pp. 1782-1789.
14. Borrell LN, Kunzel C, Lamster I, Lalla E, "Diabetes in the dental office: using NHANES III to estimate the probability of undiagnosed disease," *J Periodontal Res*, 2007, 42, pp. 559-565.
15. Breault, Joseph L., Colin R. Goodall, and Peter J. Fos., "Data mining a diabetic data warehouse." *Artificial Intelligence in Medicine*, 2002, 26.1, pp. 37-54.
16. Breiman L., Friedman J. H., Olshen R. A., and Stone C. J., "Classification and Regression Trees," Wadsworth International Group, Belmont, California, 1984.
17. Bruno, Giulia, et al., "A Clustering-Based Approach to Analyse Examinations for Diabetic Patients." *Healthcare Informatics (ICHI), 2014, IEEE International Conference on*. IEEE, 2014.

18. Burke, J. P., Williams, K., Gaskill, S. P., Hazuda, H. P., Haffner, S. M., & Stern, M. P., "Rapid rise in the incidence of type 2 diabetes from 1987 to 1996: results from the San Antonio Heart Study," *Archives of Internal Medicine*, 1999, *159*(13), pp. 1450-1456.

19. Cabrera de Le ón A, Coello SD, del Cristo Rodr íguezP érez M, Medina MB, Almeida Gonz ález D, Diaz BB, de Fuentes MM, Aguirre-Jaime A, "A simple clinical score for type 2 diabetes mellitus screening in the Canary Islands," *Diabetes Res Clin Pract*, 2008, 80, pp. 128-133.

20. Carlsson S, Midthjell K, Grill V, "Smoking is associated with an increased risk of type 2 diabetes but a decreased risk of autoimmune diabetes in adults: an 11-year follow-up of incidence of diabetes in the Nord- Trondelag study," *Diabetologia*, 2004, 47(11), pp. 1953-1956.

21. Centers for Disease Control and Prevention (CDC), "National diabetes statistics report," Atlanta (GA), 2014.

22. Centers for Medicare and Medicaid Services: Diabetes-related services.

23. Chaturvedi V, Reddy KS, Prabhakaran D, Jeemon P, Ramakrishnan L, Shah P, Shah B, "Development of a clinical risk score in predicting undiagnosed diabetes in urban Asian Indian adults: a population-based study," *CVD Prev Control*, 2008, 3, pp. 141-151.

24. Chen L, Magliano DJ, Balkau B, Colagiuri S, Zimmet PZ, Tonkin AM, Mitchell P, Phillips PJ, Shaw JE, "AUSDRISK: an Australian Type 2 Diabetes Risk Assessment Tool based on demographic, lifestyle and simple anthropometric measures," *Med J Aust*, 2010, 192, pp. 197-202.

25. Chien K, Cai T, Hsu H, Su T, Chang W, Chen M, Lee Y, Hu FB, "A prediction model for type 2 diabetes risk among Chinese people," *Diabetologia*, 2009, 52, pp. 443-450.

26. Chung, Sukyung, et al., "Reconsidering the Age Thresholds for Type II Diabetes Screening in the US," *American journal of preventive medicine,* 2014, 47.4, pp. 375-381.

27. Daveedu raju Adidela, Lavanya Devi.G, Jaya Suma.G, Appa Rao Allam, "Application of Fuzzy ID3 to Predict Diabetes," International Journal of Advanced Computer and Mathematical Sciences, 2012, Vol 3, Issue 4, pp.541-545, ISSN 2230-9624.

28. Heikes, K. E., Eddy, D. M., Arondekar, B., & Schlessinger, L., "Diabetes Risk Calculator: A simple tool for detecting undiagnosed diabetes and pre-diabetes," *Diabetes Care*, 2008, *31*(5), pp. 1040-1045.

29. Herman, William H., et al., "A new and simple questionnaire to identify people at increased risk for undiagnosed diabetes," *Diabetes Care,* 1995, 18.3, pp. 382-387.

30. Hippisley-Cox J, Coupland C, Robson J, Sheikh A, Brindle P, "Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore," BMJ, 2009, 338, p. b880.

31. Hische, Manuela, et al., "Decision trees as a simple-to-use and reliable tool to identify individuals with impaired glucose metabolism or type 2 diabetes mellitus," *European Journal of Endocrinology*, 2010, 163.4, pp. 565-571.

32. Hu, F. B., Manson, J. E., Stampfer, M. J., Colditz, G., Liu, S., Solomon, C. G., & Willett, W. C., "Diet, lifestyle, and the risk of type 2 diabetes mellitus in women," *New England Journal of Medicine,* 2001, *345*(11), pp. 790-797.

33. Huang, Yue, et al., "Feature selection and classification model construction on type 2 diabetic patients' data," *Artificial intelligence in medicine,* 2007, 41.3, pp. 251-262.

34. IDF, Clinical Guidelines Task Force, "Global Guideline for Type 2 Diabetes: recommendations for standard, comprehensive, and minimal care," *Diabetic medicine: a journal of the British Diabetic Association,* 2006, 23.6 p. 579.

35. Indian Health Service, "Standard of care and clinical practice recommendations: type 2 diabetes," 2011.

36. Gao WG, Qiao Q, Pitkäniemi J, Wild S, Magliano D, Shaw J, Söderberg S, Zimmet P, Chitson P, Knowlessur S, Alberti G, Tuomilehto J, "Risk prediction models for the development of diabetes in Mauritian Indians," *Diabet Med,* 2009, 16, pp. 996-1002.

37. Gao WG, Dong YH, Pang ZC, Nan HR, Wang SJ, Ren J, Zhang L, Tuomilehto J, Qiao Q, "A simple Chinese risk score for undiagnosed diabetes," *Diabet Med*, 2010, 27, pp. 274-281.

38. Garber, Alan J., et al., "AACE comprehensive diabetes management algorithm 2013," *Endocrine Practice,* 2013, 19.2, pp. 327-336.

39. Glümer C, Carstensen B, Sabdbaek A, Lauritzen T, Jørgensen T, BorchJohnsen K, "A Danish diabetes risk score for targeted screening," *Diabetes Care*, 2004, 27, pp. 727-733.

40. Gray LJ, Taub NA, Khunti K, Gardiner E, Hiles S, Webb DR, Srinivasan BT, Davies MJ, "The Leicester Risk Assessment score for detecting undiagnosed type 2 diabetes and impaired glucose regulation for use in a multiethnic UK setting," *Diabet Med,* 2010, 27, pp. 887-895.

41. Griffin SJ, Little PS, Hales CN, Kinmonth AL, Wareham NJ, "Diabetes risk score: towards earlier detection of type 2 diabetes in general practice," *Diabetes Metab Res Rev,* 2000, 16, pp. 164-171.

42. Gupta AK, Dahlof B, Dobson J, Sever PS, Wedel H, Poulter NR, "Anglo-Scandinavian Cardiac Outcomes Trial Investigators: Determinants of new onset diabetes among 19,257 hypertensive patients randomized in the Anglo-Scandinavian Cardiac Outcomes Trial-Blood

Pressure Lowering Arm and the relative influence of antihypertensive medication," *Diabetes Care*, 2008, 31, pp. 982-988.

43. Kahn HS, Cheng YJ, Thompson TJ, Imperatore G, Gregg EW, "Two risks coring systems for predicting incident diabetes mellitus in U.S. adults age 45 to 64 years," *Ann Intern Med*, 2009, 150, pp. 741-751.

44. Kanaya AM, Wassel Fyr CL, de Rekeneire N, Schwartz AV, Goodpaster BH, Newman AB, Harris T, Barrett-Connor E, "Predicting the development of diabetes in older adults: the derivation and validation of a prediction rule," *Diabetes Care*, 2005, 28, pp. 404-408.

45. Karegowda, et al., "Rule based Classification for Diabetic Patients using Cascaded K-Means and Decision Tree C 4. 5," *International Journal of Computer Applications,* 2012, 45.12.

46. Karthikeyani, V., et al., "Comparative of Data Mining Classification Algorithm (CDMCA) in Diabetes Disease Prediction," *International Journal of Computer Applications*, 2012, 60.12, pp. 26-31.

47. Katz MH, "Multivariable analysis: a primer for readers of medical research," *Ann Intern Med*. 2003; 138 (8), pp. 644–650.

48. Kavitha, K., and R. M. Sarojamma, "Monitoring of diabetes with data mining via CART Method," *International Journal of Emerging Technology and Advanced Engineering*, 2012, 2.11, pp. 157-162.

49. Kawakami N, Takatsuka N, Shimizu H, et al., "Effects of smoking on the incidence of non-insulindependent diabetes mellitus: replication and extension in a Japanese cohort of male employees," *Am J Epidemiol*. 1997, 145 (2), pp. 103-109.

50. Keesukphan P, Chanprasertyothin S, Ongphiphadhanakul B, Puavilai G., "The development and validation of a diabetes risk score for high-risk Thai adults," *J Med Assoc Thai,* 2007, 90, pp. 149-154.

51. Ko G, So W, Tong P, Ma R, Kong A, Ozakit R, Chow C, Cockram C, Chan J., "A simple risk score to identify Southern Chinese at high risk for diabetes," *Diabet Med*, 2010, 27 pp. 644-649.

52. Kolberg JA, Jørgensen T, Gerwien RW, Hamren S, McKenna MP, Moler E, Rowe MW, Urdea MS, Xu XM, Hansen T, Pedersen O, Borch-Johnsen K., "Development of a type 2 diabetes risk model from a panel of serum biomarkers from the Inter99 cohort," *Diabetes Care,* 2009, 32, pp. 1207-1212.

53. Lavanya, D., and K. Usha Rani, "Performance evaluation of decision tree classifiers on medical datasets." *IJCA) International Journal of Computer Applications*, 2011, 26.4.

54. Lindström, J., & Tuomilehto, J., "The Diabetes Risk Score: A practical tool to predict type 2 diabetes risk," *Diabetes care*, 2003, *26*(3), pp. 725-731.

55. Liou, Fen-May, Ying-Chan Tang, and Jean-Yi Chen, "Detecting hospital fraud and claim abuse through diabetic outpatient services," *Health care management science*, 2008, 11.4, pp. 353-358.

56. Luo, Senlin, et al., "A Risk Assessment Model for Type 2 Diabetes in Chinese," 2014, e104046.

57. Magidson, Jay, "The CHAID approach to segmentation modeling: Chi-squared automatic interaction detection," *Advanced methods of marketing research*, 1994, pp. 118-159.

58. Management of Diabetes Mellitus Update Working Group. "VA/DoD clinical practice guideline for the management of diabetes mellitus. Version 4.0."*Update August*, 2010.

59. Manson, JoAnn E., et al., "A prospective study of cigarette smoking and the incidence of diabetes mellitus among US male physicians," *The American journal of medicine*, 2000, 109.7, pp. 538-542.

60. Meng, X. H., Huang, Y. X., Rao, D. P., Zhang, Q., & Liu, Q., "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors," *The Kaohsiung journal of medical sciences*, 2013, *29*(2), pp. 93-99.

61. Miyaki, Koichi, et al., "Novel statistical classification model of type 2 diabetes mellitus patients for tailormade prevention using data mining algorithm," *Journal of epidemiology,* 2002, 12.3, pp. 243-248.

62. Mochan, Eugene, and Mark H. Ebell, "Risk-assessment tools for detecting undiagnosed diabetes," *American family physician,* 2009, 80.2, p. 175.

63. Mohan V, Deepa R, Deepa M, Somannavar S, Datta M, "A simplified Indian Diabetes Risk Score for screening for undiagnosed diabetic subjects," *J Assoc Physicians India*, 2005, 53, pp. 759-763.

64. Tamez-Pérez, Héctor Eloy, et al., "AACE comprehensive diabetes management algorithm 2013 endocrine practice," *Endocrine Practice,* 2013, 19.4, pp.736-737.

65. Tan, Evelyn, Jennifer Polello, and Lisa J. Woodard, "An Evaluation of the Current Type 2 Diabetes Guidelines: Where They Converge and Diverge," *Clinical Diabetes,* 2014, 32.3, pp. 133-139.

66. Tayek, John A, "Is weight loss a cure for type 2 diabetes?" *Diabetes care*, 2002, 25.2, pp. 397-398.

67. Perry IJ, Wannamethee SG, Walker MK, et al., "Prospective study of risk factors for development of noninsulin dependent diabetes in middle aged British men," *BMJ*. 1995; 310 (6979), pp. 560-564.

68. Pires de Sousa AG, Pereira AC, Marquezine GF, Marques do NascimentoNeto R, Freitas SN, Nicolato RLdC, Machado-Coelho GL, Rodrigues SL, Mill JG, Krieger JE, "Derivation and external validation of a simple prediction model for the diagnosis of type 2 diabetes mellitus in the Brazilian urban population," Eur J Epidemiol, 2009, 24, pp. 101-109.

69. Quinlan, J. Ross, "Induction of decision trees." *Machine learning* 1.1, 1986, pp. 81-106.

70. Quinlan, J. Ross, "Simplifying decision trees." *International journal of man-machine studies*1987, 27.3, pp. 221-234.

71. Quinlan J., "C4.5: programs for machine learning, " San Mateo, CA: Morgan Kaufmann, 1993.

72. Radha, P., and B. Srinivasan, "Predicting Diabetes by cosequencing the various Data Mining Classification Techniques," *International Journal of Innovative Science, Engineering & Technology*, Vol. 1 Issue 6, August 2014.

73. Ramachandran A, Snehalatha C, Vijay C, Wareham NJ, Colagiuri S, "Derivation and validation of diabetes risk score for urban Asian Indians," *Diabetes Res Clin Pract*, 2005, 70 pp. 63-70.

74. Ramezankhani, Azra, et al., "Applying decision tree for identification of a low risk population for type 2 diabetes. Tehran Lipid and Glucose Study," *Diabetes research and clinical practice*, 2014, 105.3 pp. 391-398.

75. Riethof, M., P. L. Flavin, and B. Lindvall, "Diagnosis and management of type 2 diabetes mellitus in adults." *Institute for Clinical Systems Improvement (ICSI)*, 2012.

76. Ruige JB, de Neeling JND, Kostense PJ, Bouter LM, Heine RJ, "Performance of an NIDDM screening questionnaire based on symptoms and risk factors," *Diabetes Care,* 1997, 20, pp. 491-496.

77. Sankaranarayanan, Sriram, and T. Pramananda Perumal, "A Predictive Approach for Diabetes Mellitus Disease through Data Mining Technologies," *Computing and Communication Technologies (WCCCT), World Congress on*. IEEE, 2014.

78. Schmidt, M. I., Duncan, B. B., Bang, H., Pankow, J. S., Ballantyne, C. M., Golden, S. H., ... & Chambless, L. E., "Identifying Individuals at High Risk for Diabetes The Atherosclerosis Risk in Communities study," *Diabetes care,* 2005, *28*(8), pp. 2013-2018.

79. Schulze MB, Hoffmann K, Boeing H, Linseisen J, Rohrmann S, Möhlig M, Pfeiffer AF, Spranger J, Thamer C, Häring HU, Fritsche A, Joost HG, "An accurate risk score based on anthropometric, dietary, and lifestyle factors to predict the development of type 2 diabetes," *Diabetes Care,* 2007, 30, pp. 510-515.

80. Stern MP, Williams K, Haffner SM, "Identification of persons at high risk for type 2 diabetes mellitus: Do we need the oral glucose tolerance test?" *Ann Intern Med,* 2002, 136, pp. 575-581.

81. Sugimori H, Miyakawa M, Yoshida K, et al., " Health risk assessment for diabetes mellitus based on longitudinal analysis of MHTS database," *JMed Syst*, 1998, 22(1), pp. 27-32.

82. Sun F, Tao Q, Zhan S., "An accurate risk score for estimation 5-year risk of type 2 diabetes based on a health screening population in Taiwan," *Diabetes Res Clin Pract*, 2009, 85, pp. 228-234.

83. Tabaei, B. P., & Herman, W. H., "A multivariate logistic regression equation to screen for diabetes development and validation," *Diabetes Care*, 2002, *25*(11), 1999-2003.

84. Toussi, Massoud, et al., "Using data mining techniques to explore physicians' therapeutic decisions when clinical guidelines do not provide recommendations: methods and example for type 2 diabetes," *BMC medical informatics and decision making*, 9.1, 2009, p. 28.

85. Tuomilehto J, Lindström J, Hellmich M, Lehmacher W, Westermeier T, Evers T, Brückner A, Peltonen M, Qiao Q, Chiasson JL, "Development and validation of a risk-score model for subjects with impaired glucose tolerance for the assessment of the risk of type 2 diabetes mellitus: the STOP-NIDDM risk-score," *Diabetes Res Clin Pract*, 2010, 87, pp. 267-274.

86. Van Belle G., "Biostatistics: A Methodology for the Health Sciences. Hoboken, NJ: Wiley-Interscience," 2004.

87. Visalatchi, G., S. J. Gnanasoundhari, and M. Balamurugan., "A Survey on Data Mining Methods and Techniques for Diabetes Mellitus," *chart*, 2014, 2025, pp. 299-1.

88. Vohra, Rajan, and Anshul Arora, "Prediction of Diabetes Using J48 And ID3," July, 2014.

89. Waki K, Noda M, Sasaki S, et al., "Alcohol consumption and other risk factors for self-reported diabetes among middle-aged Japanese: a population based prospective study in the JPHC study cohort I," [published correction appears in *Diabet Med*. 2005; 22(6): 818]. *Diabet Med*. 2005, 22(3), pp. 323-331.

90. Wilson PWF, Meigs JB, Sullivan L, Fox CS, Nathan DM, D'Agostino RB Sr, "Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study," *Arch Intern Med*,2007, 167, pp. 1068-1074.

91. Woolthuis, Erwin P. Klein, et al., "Yield of opportunistic targeted screening for type 2 diabetes in primary care: the diabscreen study," *The Annals of Family Medicine*, 2009, 7.5, pp. 422-430.

92. World Health Organization. *Guidelines for the prevention, management and care of diabetes mellitus*, 2006.

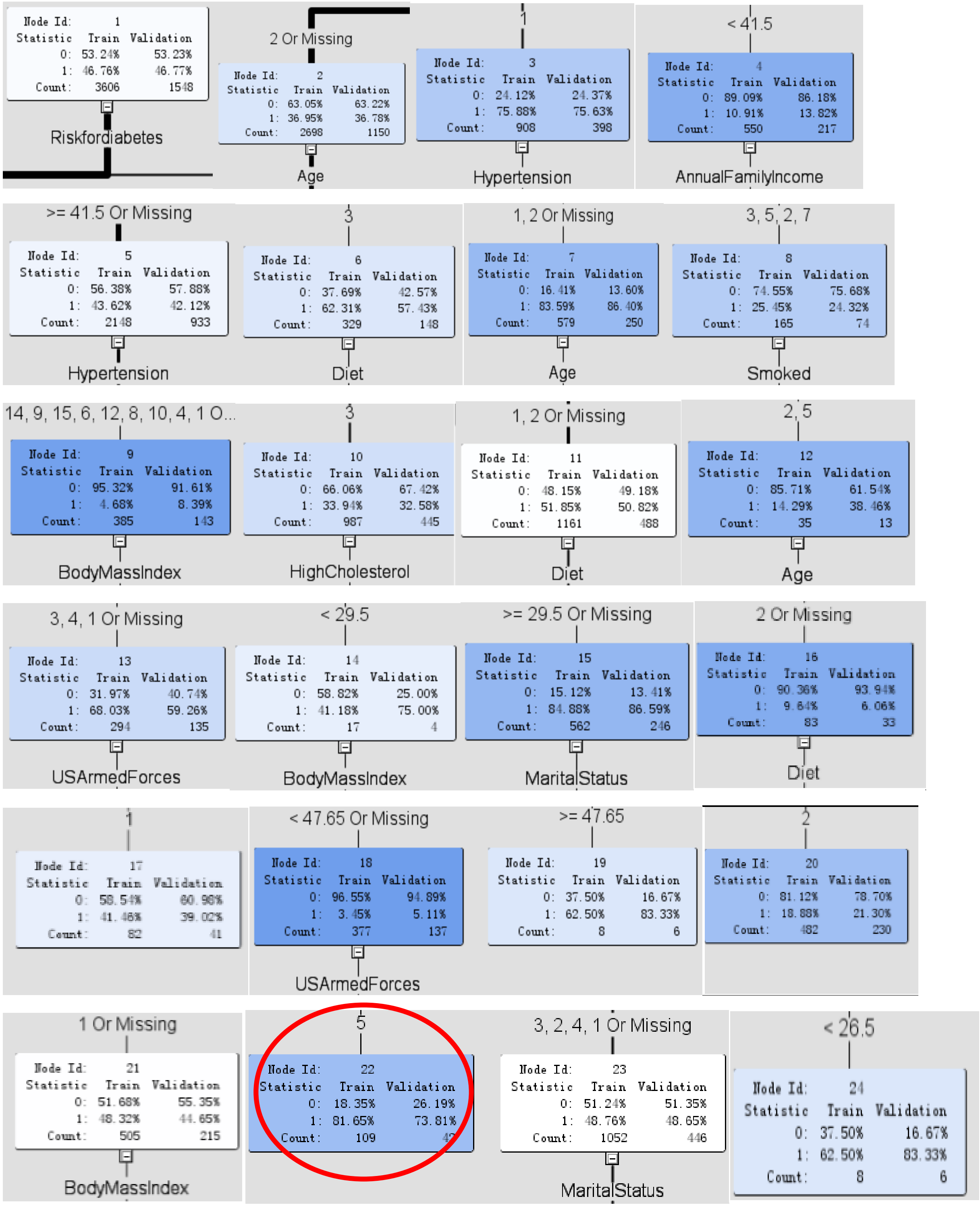
93. Xie J, Hu D, Yu D, Chen CS, He J, Gu D., "A quick self-assessment tool to identify individuals at high risk of type 2 diabetes in the Chinese general population," *J Epidemiol Community Health*, 2010, 64, pp. 236-242.

94. Xin, Zhong, et al. "A simple tool detected diabetes and pre-diabetes in rural Chinese." *Journal of clinical epidemiology,* 2010, 63.9, pp. 1030-1035.

95. Yang, W., Lu, J., Weng, J., Jia, W., Ji, L., Xiao, J., ... & He, J., "Prevalence of diabetes among men and women in China," *New England Journal of Medicine,* 2010, *362*(12), pp. 1090-1101.

# APPENDIX A

## Decision Tree Nodes of 2007-2008



Node Id: 1

| Statistic | Train | Validation |
|---|---|---|
| 0: | 53.24% | 53.23% |
| 1: | 46.76% | 46.77% |
| Count: | 3606 | 1548 |

Riskfordiabetes

2 Or Missing

Node Id: 2

| Statistic | Train | Validation |
|---|---|---|
| 0: | 63.05% | 63.22% |
| 1: | 36.95% | 36.78% |
| Count: | 2698 | 1150 |

Age

1

Node Id: 3

| Statistic | Train | Validation |
|---|---|---|
| 0: | 24.12% | 24.37% |
| 1: | 75.88% | 75.63% |
| Count: | 908 | 398 |

Hypertension

< 41.5

Node Id: 4

| Statistic | Train | Validation |
|---|---|---|
| 0: | 89.09% | 86.18% |
| 1: | 10.91% | 13.82% |
| Count: | 550 | 217 |

AnnualFamilyIncome

>= 41.5 Or Missing

Node Id: 5

| Statistic | Train | Validation |
|---|---|---|
| 0: | 56.38% | 57.88% |
| 1: | 43.62% | 42.12% |
| Count: | 2148 | 933 |

Hypertension

3

Node Id: 6

| Statistic | Train | Validation |
|---|---|---|
| 0: | 37.69% | 42.57% |
| 1: | 62.31% | 57.43% |
| Count: | 329 | 148 |

Diet

1, 2 Or Missing

Node Id: 7

| Statistic | Train | Validation |
|---|---|---|
| 0: | 16.41% | 13.60% |
| 1: | 83.59% | 86.40% |
| Count: | 579 | 250 |

Age

3, 5, 2, 7

Node Id: 8

| Statistic | Train | Validation |
|---|---|---|
| 0: | 74.55% | 75.68% |
| 1: | 25.45% | 24.32% |
| Count: | 165 | 74 |

Smoked

14, 9, 15, 6, 12, 8, 10, 4, 1 O...

Node Id: 9

| Statistic | Train | Validation |
|---|---|---|
| 0: | 95.32% | 91.61% |
| 1: | 4.68% | 8.39% |
| Count: | 385 | 143 |

BodyMassIndex

3

Node Id: 10

| Statistic | Train | Validation |
|---|---|---|
| 0: | 66.06% | 67.42% |
| 1: | 33.94% | 32.58% |
| Count: | 987 | 445 |

HighCholesterol

1, 2 Or Missing

Node Id: 11

| Statistic | Train | Validation |
|---|---|---|
| 0: | 48.15% | 49.18% |
| 1: | 51.85% | 50.82% |
| Count: | 1161 | 488 |

Diet

2, 5

Node Id: 12

| Statistic | Train | Validation |
|---|---|---|
| 0: | 85.71% | 61.54% |
| 1: | 14.29% | 38.46% |
| Count: | 35 | 13 |

Age

3, 4, 1 Or Missing

Node Id: 13

| Statistic | Train | Validation |
|---|---|---|
| 0: | 31.97% | 40.74% |
| 1: | 68.03% | 59.26% |
| Count: | 294 | 135 |

USArmedForces

< 29.5

Node Id: 14

| Statistic | Train | Validation |
|---|---|---|
| 0: | 58.82% | 25.00% |
| 1: | 41.18% | 75.00% |
| Count: | 17 | 4 |

BodyMassIndex

>= 29.5 Or Missing

Node Id: 15

| Statistic | Train | Validation |
|---|---|---|
| 0: | 15.12% | 13.41% |
| 1: | 84.88% | 86.59% |
| Count: | 562 | 246 |

MaritalStatus

2 Or Missing

Node Id: 16

| Statistic | Train | Validation |
|---|---|---|
| 0: | 90.36% | 93.94% |
| 1: | 9.64% | 6.06% |
| Count: | 83 | 33 |

Diet

1

Node Id: 17

| Statistic | Train | Validation |
|---|---|---|
| 0: | 58.54% | 60.98% |
| 1: | 41.46% | 39.02% |
| Count: | 82 | 41 |

< 47.65 Or Missing

Node Id: 18

| Statistic | Train | Validation |
|---|---|---|
| 0: | 96.55% | 94.89% |
| 1: | 3.45% | 5.11% |
| Count: | 377 | 137 |

USArmedForces

>= 47.65

Node Id: 19

| Statistic | Train | Validation |
|---|---|---|
| 0: | 37.50% | 16.67% |
| 1: | 62.50% | 83.33% |
| Count: | 8 | 6 |

2

Node Id: 20

| Statistic | Train | Validation |
|---|---|---|
| 0: | 81.12% | 78.70% |
| 1: | 18.88% | 21.30% |
| Count: | 482 | 230 |

1 Or Missing

Node Id: 21

| Statistic | Train | Validation |
|---|---|---|
| 0: | 51.68% | 55.35% |
| 1: | 48.32% | 44.65% |
| Count: | 505 | 215 |

BodyMassIndex

5

Node Id: 22

| Statistic | Train | Validation |
|---|---|---|
| 0: | 18.35% | 26.19% |
| 1: | 81.65% | 73.81% |
| Count: | 109 | 42 |

3, 2, 4, 1 Or Missing

Node Id: 23

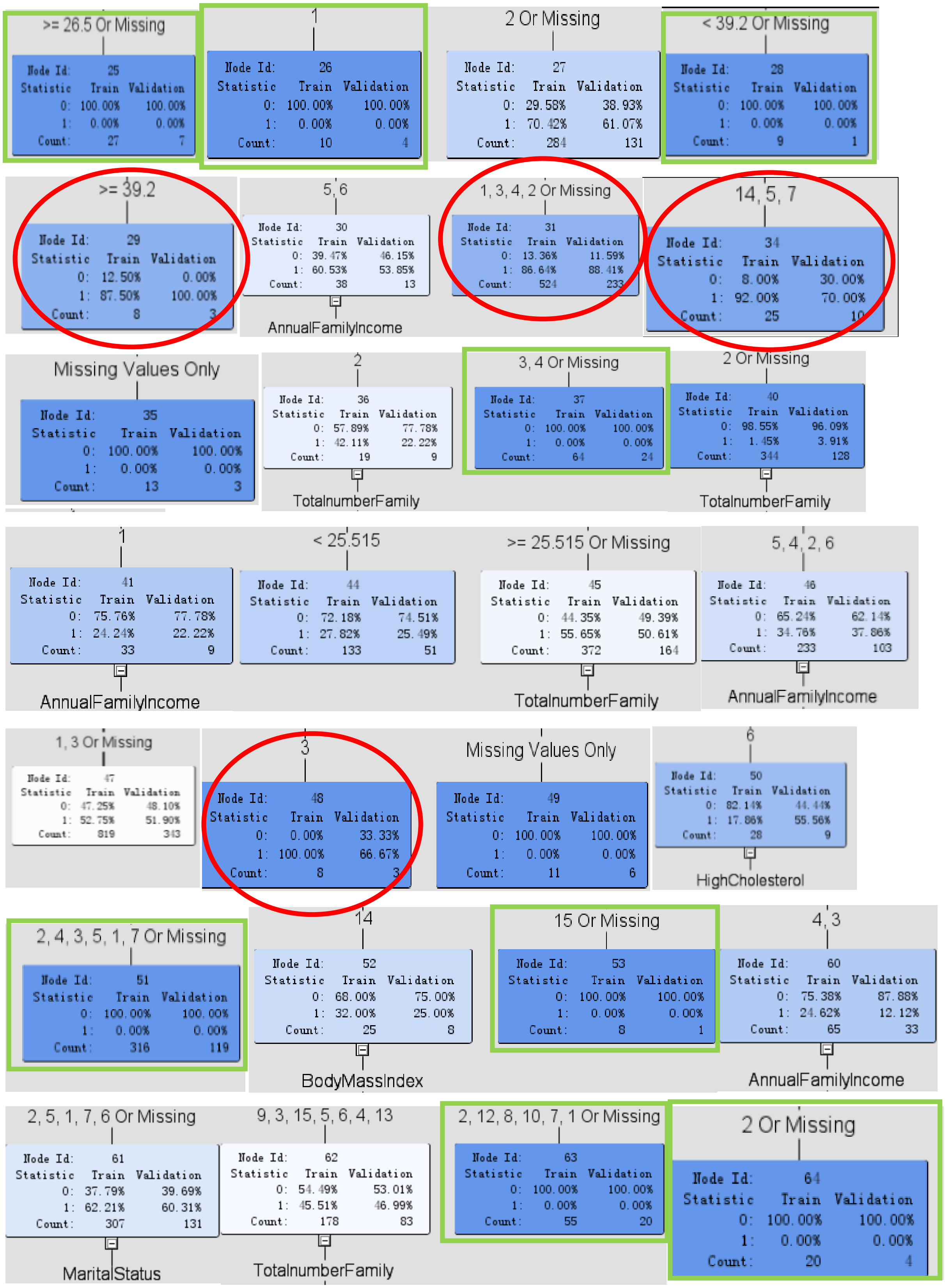
| Statistic | Train | Validation |
|---|---|---|
| 0: | 51.24% | 51.35% |
| 1: | 48.76% | 48.65% |
| Count: | 1052 | 446 |

MaritalStatus

< 26.5

Node Id: 24

| Statistic | Train | Validation |
|---|---|---|
| 0: | 37.50% | 16.67% |
| 1: | 62.50% | 83.33% |
| Count: | 8 | 6 |



```
>= 26.5 Or Missing
Node Id:      25
Statistic   Train  Validation
       0:  100.00%   100.00%
       1:    0.00%     0.00%
   Count:      27         7
```

```
1
Node Id:      26
Statistic   Train  Validation
       0:  100.00%   100.00%
       1:    0.00%     0.00%
   Count:      10         4
```

```
2 Or Missing
Node Id:      27
Statistic   Train  Validation
       0:   29.58%    38.93%
       1:   70.42%    61.07%
   Count:     284       131
```

```
< 39.2 Or Missing
Node Id:      28
Statistic   Train  Validation
       0:  100.00%   100.00%
       1:    0.00%     0.00%
   Count:       9         1
```

```
>= 39.2
Node Id:      29
Statistic   Train  Validation
       0:   12.50%     0.00%
       1:   87.50%   100.00%
   Count:       8         3
```

```
5, 6
Node Id:      30
Statistic   Train  Validation
       0:   39.47%    46.15%
       1:   60.53%    53.85%
   Count:      38        13
AnnualFamilyIncome
```

```
1, 3, 4, 2 Or Missing
Node Id:      31
Statistic   Train  Validation
       0:   13.36%    11.59%
       1:   86.64%    88.41%
   Count:     524       233
```

```
14, 5, 7
Node Id:      34
Statistic   Train  Validation
       0:    8.00%    30.00%
       1:   92.00%    70.00%
   Count:      25        10
```

```
Missing Values Only
Node Id:      35
Statistic   Train  Validation
       0:  100.00%   100.00%
       1:    0.00%     0.00%
   Count:      13         3
```

```
2
Node Id:      36
Statistic   Train  Validation
       0:   57.89%    77.78%
       1:   42.11%    22.22%
   Count:      19         9
TotalnumberFamily
```

```
3, 4 Or Missing
Node Id:      37
Statistic   Train  Validation
       0:  100.00%   100.00%
       1:    0.00%     0.00%
   Count:      64        24
```

```
2 Or Missing
Node Id:      40
Statistic   Train  Validation
       0:   98.55%    96.09%
       1:    1.45%     3.91%
   Count:     344       128
TotalnumberFamily
```

```
1
Node Id:      41
Statistic   Train  Validation
       0:   75.76%    77.78%
       1:   24.24%    22.22%
   Count:      33         9
AnnualFamilyIncome
```

```
< 25.515
Node Id:      44
Statistic   Train  Validation
       0:   72.18%    74.51%
       1:   27.82%    25.49%
   Count:     133        51
```

```
>= 25.515 Or Missing
Node Id:      45
Statistic   Train  Validation
       0:   44.35%    49.39%
       1:   55.65%    50.61%
   Count:     372       164
TotalnumberFamily
```

```
5, 4, 2, 6
Node Id:      46
Statistic   Train  Validation
       0:   65.24%    62.14%
       1:   34.76%    37.86%
   Count:     233       103
AnnualFamilyIncome
```

```
1, 3 Or Missing
Node Id:      47
Statistic   Train  Validation
       0:   47.25%    48.10%
       1:   52.75%    51.90%
   Count:     819       343
```

```
3
Node Id:      48
Statistic   Train  Validation
       0:    0.00%    33.33%
       1:  100.00%    66.67%
   Count:       8         3
```

```
Missing Values Only
Node Id:      49
Statistic   Train  Validation
       0:  100.00%   100.00%
       1:    0.00%     0.00%
   Count:      11         6
```

```
6
Node Id:      50
Statistic   Train  Validation
       0:   82.14%    44.44%
       1:   17.86%    55.56%
   Count:      28         9
HighCholesterol
```

```
2, 4, 3, 5, 1, 7 Or Missing
Node Id:      51
Statistic   Train  Validation
       0:  100.00%   100.00%
       1:    0.00%     0.00%
   Count:     316       119
```

```
14
Node Id:      52
Statistic   Train  Validation
       0:   68.00%    75.00%
       1:   32.00%    25.00%
   Count:      25         8
BodyMassIndex
```

```
15 Or Missing
Node Id:      53
Statistic   Train  Validation
       0:  100.00%   100.00%
       1:    0.00%     0.00%
   Count:       8         1
```

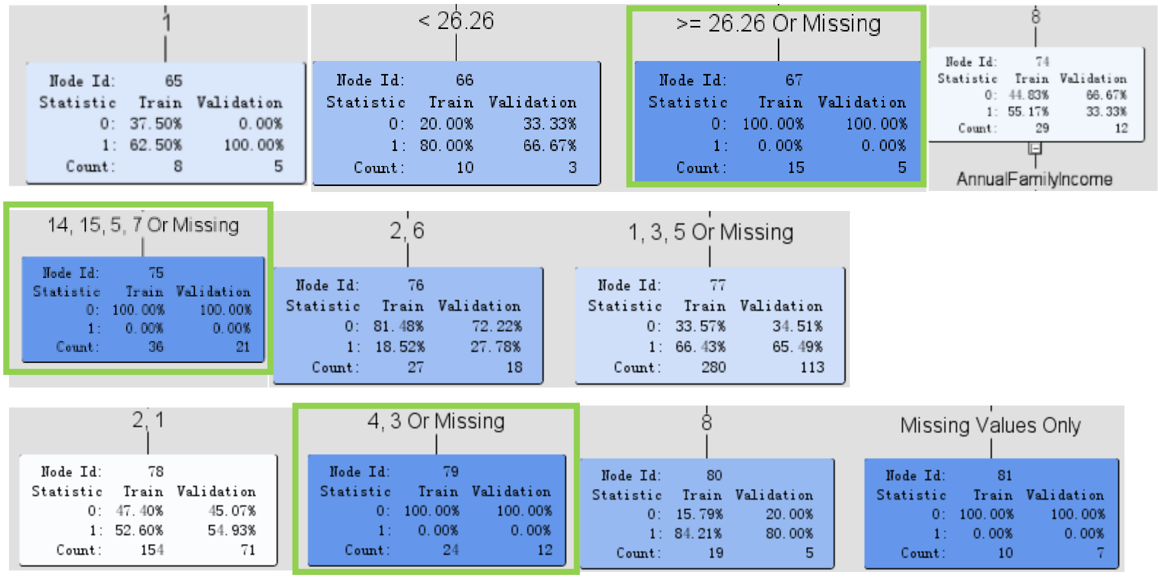
```
4, 3
Node Id:      60
Statistic   Train  Validation
       0:   75.38%    87.88%
       1:   24.62%    12.12%
   Count:      65        33
AnnualFamilyIncome
```

```
2, 5, 1, 7, 6 Or Missing
Node Id:      61
Statistic   Train  Validation
       0:   37.79%    39.69%
       1:   62.21%    60.31%
   Count:     307       131
MaritalStatus
```

```
9, 3, 15, 5, 6, 4, 13
Node Id:      62
Statistic   Train  Validation
       0:   54.49%    53.01%
       1:   45.51%    46.99%
   Count:     178        83
TotalnumberFamily
```

```
2, 12, 8, 10, 7, 1 Or Missing
Node Id:      63
Statistic   Train  Validation
       0:  100.00%   100.00%
       1:    0.00%     0.00%
   Count:      55        20
```

```
2 Or Missing
Node Id:      64
Statistic   Train  Validation
       0:  100.00%   100.00%
       1:    0.00%     0.00%
   Count:      20         4
```



1

```
Node Id:       65
Statistic   Train   Validation
       0:  37.50%        0.00%
       1:  62.50%      100.00%
   Count:       8            5
```

< 26.26

```
Node Id:       66
Statistic   Train   Validation
       0:  20.00%       33.33%
       1:  80.00%       66.67%
   Count:      10            3
```

>= 26.26 Or Missing

```
Node Id:       67
Statistic   Train   Validation
       0: 100.00%      100.00%
       1:   0.00%        0.00%
   Count:      15            5
```

8

```
Node Id:       74
Statistic   Train   Validation
       0:  44.83%       66.67%
       1:  55.17%       33.33%
   Count:      29           12
```

AnnualFamilyIncome

14, 15, 5, 7 Or Missing

```
Node Id:       75
Statistic   Train   Validation
       0: 100.00%      100.00%
       1:   0.00%        0.00%
   Count:      36           21
```

2, 6

```
Node Id:       76
Statistic   Train   Validation
       0:  81.48%       72.22%
       1:  18.52%       27.78%
   Count:      27           18
```

1, 3, 5 Or Missing

```
Node Id:       77
Statistic   Train   Validation
       0:  33.57%       34.51%
       1:  66.43%       65.49%
   Count:     280          113
```

2, 1

```
Node Id:       78
Statistic   Train   Validation
       0:  47.40%       45.07%
       1:  52.60%       54.93%
   Count:     154           71
```

4, 3 Or Missing

```
Node Id:       79
Statistic   Train   Validation
       0: 100.00%      100.00%
       1:   0.00%        0.00%
   Count:      24           12
```

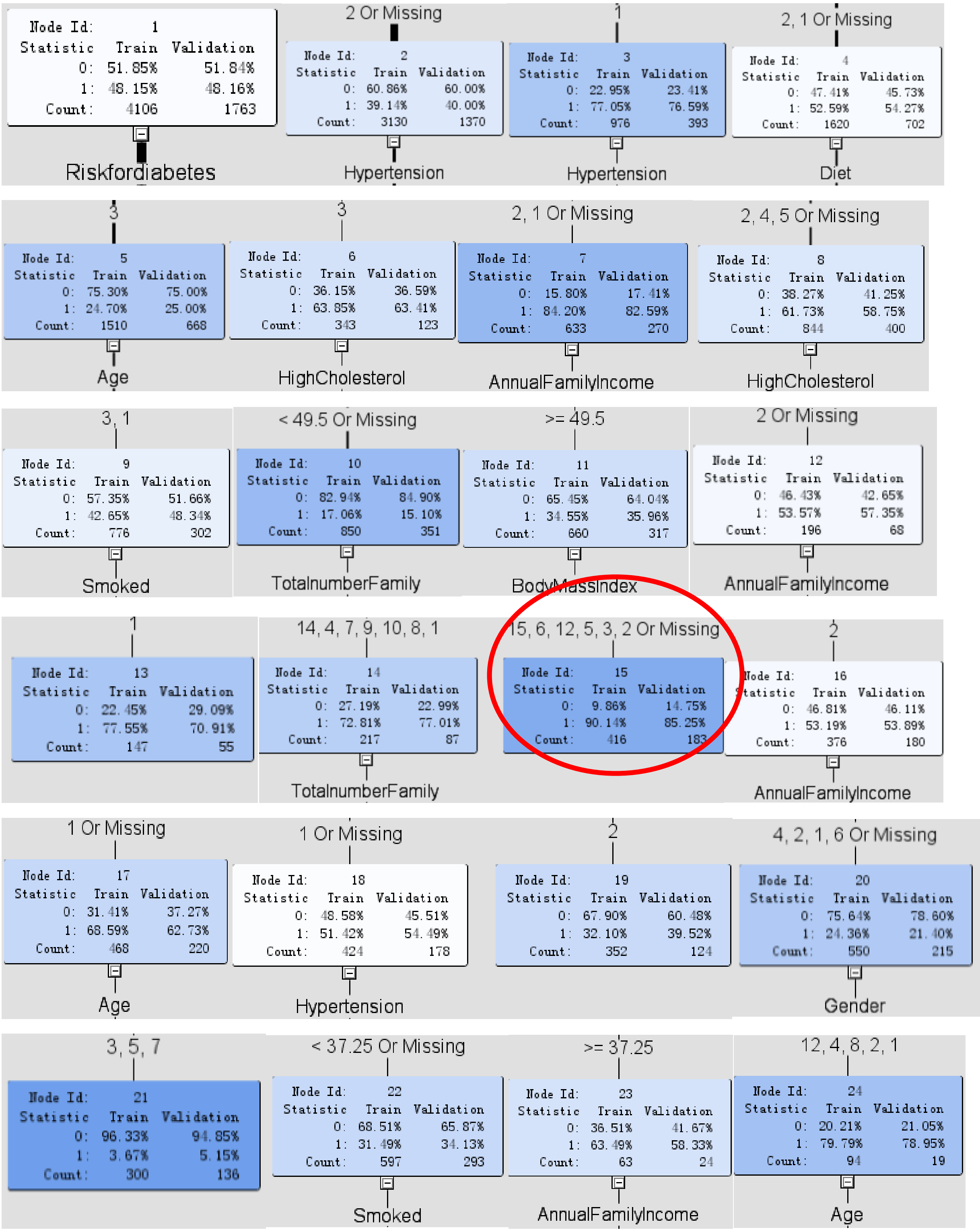
8

```
Node Id:       80
Statistic   Train   Validation
       0:  15.79%       20.00%
       1:  84.21%       80.00%
   Count:      19            5
```

Missing Values Only

```
Node Id:       81
Statistic   Train   Validation
       0: 100.00%      100.00%
       1:   0.00%        0.00%
   Count:      10            7
```

## Decision Tree Nodes of 2009-2010



```
Node Id:        1
Statistic   Train  Validation
       0:  51.85%      51.84%
       1:  48.15%      48.16%
   Count:    4106        1763
```

Riskfordiabetes

2 Or Missing

```
Node Id:        2
Statistic   Train  Validation
       0:  60.86%      60.00%
       1:  39.14%      40.00%
   Count:    3130        1370
```

Hypertension

1

```
Node Id:        3
Statistic   Train  Validation
       0:  22.95%      23.41%
       1:  77.05%      76.59%
   Count:     976         393
```

Hypertension

2, 1 Or Missing

```
Node Id:        4
Statistic   Train  Validation
       0:  47.41%      45.73%
       1:  52.59%      54.27%
   Count:    1620         702
```

Diet

3

```
Node Id:        5
Statistic   Train  Validation
       0:  75.30%      75.00%
       1:  24.70%      25.00%
   Count:    1510         668
```

Age

3

```
Node Id:        6
Statistic   Train  Validation
       0:  36.15%      36.59%
       1:  63.85%      63.41%
   Count:     343         123
```

HighCholesterol

2, 1 Or Missing

```
Node Id:        7
Statistic   Train  Validation
       0:  15.80%      17.41%
       1:  84.20%      82.59%
   Count:     633         270
```

AnnualFamilyIncome

2, 4, 5 Or Missing

```
Node Id:        8
Statistic   Train  Validation
       0:  38.27%      41.25%
       1:  61.73%      58.75%
   Count:     844         400
```

HighCholesterol

3, 1

```
Node Id:        9
Statistic   Train  Validation
       0:  57.35%      51.66%
       1:  42.65%      48.34%
   Count:     776         302
```

Smoked

< 49.5 Or Missing

```
Node Id:       10
Statistic   Train  Validation
       0:  82.94%      84.90%
       1:  17.06%      15.10%
   Count:     850         351
```

TotalnumberFamily

>= 49.5

```
Node Id:       11
Statistic   Train  Validation
       0:  65.45%      64.04%
       1:  34.55%      35.96%
   Count:     660         317
```

BodyMassIndex

2 Or Missing

```
Node Id:       12
Statistic   Train  Validation
       0:  46.43%      42.65%
       1:  53.57%      57.35%
   Count:     196          68
```

AnnualFamilyIncome

1

```
Node Id:       13
Statistic   Train  Validation
       0:  22.45%      29.09%
       1:  77.55%      70.91%
   Count:     147          55
```

14, 4, 7, 9, 10, 8, 1

```
Node Id:       14
Statistic   Train  Validation
       0:  27.19%      22.99%
       1:  72.81%      77.01%
   Count:     217          87
```

TotalnumberFamily

15, 6, 12, 5, 3, 2 Or Missing

```
Node Id:       15
Statistic   Train  Validation
       0:   9.86%      14.75%
       1:  90.14%      85.25%
   Count:     416         183
```

2

```
Node Id:       16
Statistic   Train  Validation
       0:  46.81%      46.11%
       1:  53.19%      53.89%
   Count:     376         180
```

AnnualFamilyIncome

1 Or Missing

```
Node Id:       17
Statistic   Train  Validation
       0:  31.41%      37.27%
       1:  68.59%      62.73%
   Count:     468         220
```

Age

1 Or Missing

```
Node Id:       18
Statistic   Train  Validation
       0:  48.58%      45.51%
       1:  51.42%      54.49%
   Count:     424         178
```

Hypertension

2

```
Node Id:       19
Statistic   Train  Validation
       0:  67.90%      60.48%
       1:  32.10%      39.52%
   Count:     352         124
```

4, 2, 1, 6 Or Missing

```
Node Id:       20
Statistic   Train  Validation
       0:  75.64%      78.60%
       1:  24.36%      21.40%
   Count:     550         215
```

Gender

3, 5, 7

```
Node Id:       21
Statistic   Train  Validation
       0:  96.33%      94.85%
       1:   3.67%       5.15%
   Count:     300         136
```

< 37.25 Or Missing

```
Node Id:       22
Statistic   Train  Validation
       0:  68.51%      65.87%
       1:  31.49%      34.13%
   Count:     597         293
```

Smoked

>= 37.25

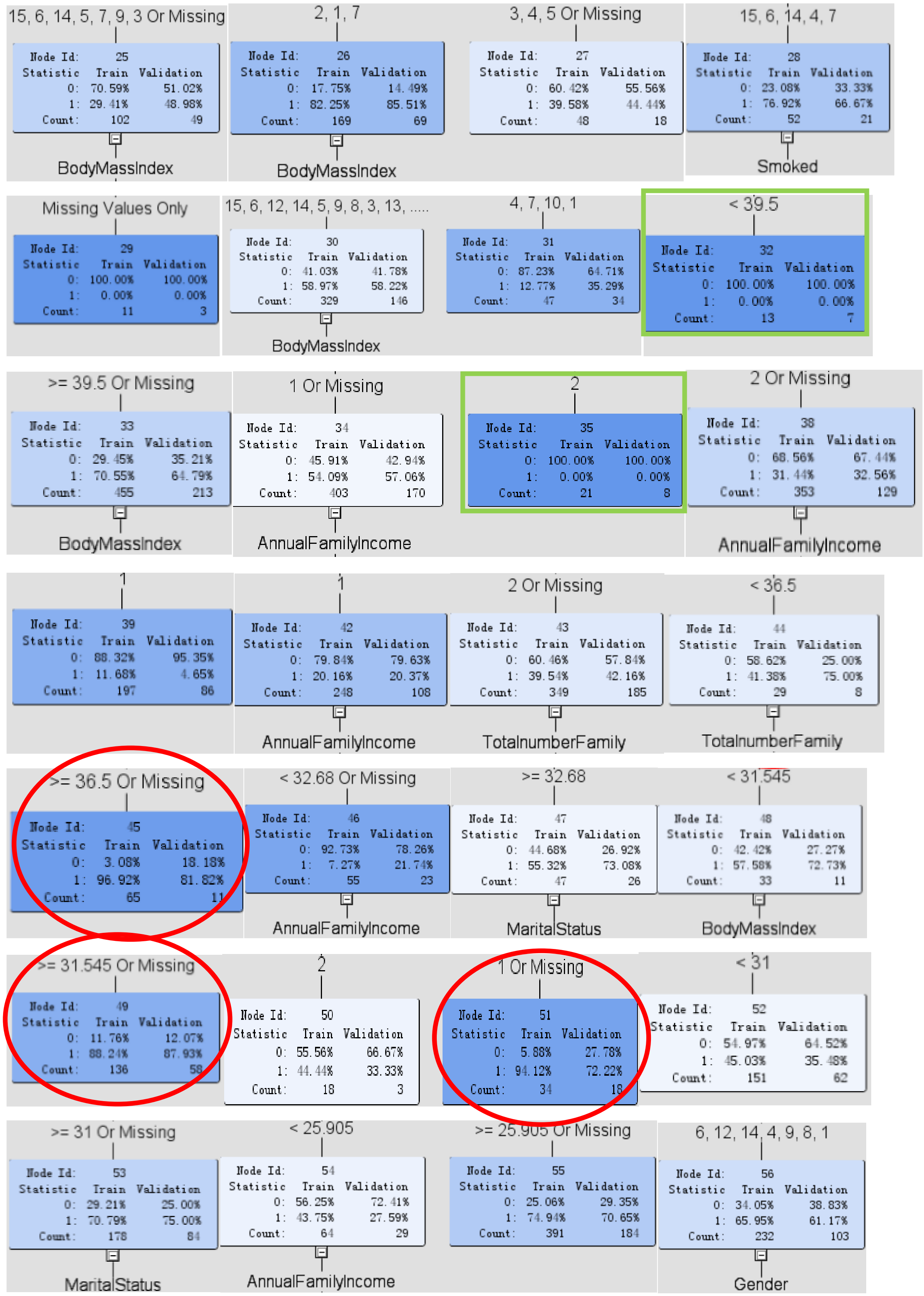
```
Node Id:       23
Statistic   Train  Validation
       0:  36.51%      41.67%
       1:  63.49%      58.33%
   Count:      63          24
```

AnnualFamilyIncome

12, 4, 8, 2, 1

```
Node Id:       24
Statistic   Train  Validation
       0:  20.21%      21.05%
       1:  79.79%      78.95%
   Count:      94          19
```

Age



15, 6, 14, 5, 7, 9, 3 Or Missing

```
Node Id:      25
Statistic   Train   Validation
       0:  70.59%     51.02%
       1:  29.41%     48.98%
   Count:     102         49
```

BodyMassIndex

2, 1, 7

```
Node Id:      26
Statistic   Train   Validation
       0:  17.75%     14.49%
       1:  82.25%     85.51%
   Count:     169         69
```

BodyMassIndex

3, 4, 5 Or Missing

```
Node Id:      27
Statistic   Train   Validation
       0:  60.42%     55.56%
       1:  39.58%     44.44%
   Count:      48         18
```

15, 6, 14, 4, 7

```
Node Id:      28
Statistic   Train   Validation
       0:  23.08%     33.33%
       1:  76.92%     66.67%
   Count:      52         21
```

Smoked

Missing Values Only

```
Node Id:      29
Statistic   Train   Validation
       0: 100.00%    100.00%
       1:   0.00%      0.00%
   Count:      11          3
```

15, 6, 12, 14, 5, 9, 8, 3, 13, .....

```
Node Id:      30
Statistic   Train   Validation
       0:  41.03%     41.78%
       1:  58.97%     58.22%
   Count:     329        146
```

BodyMassIndex

4, 7, 10, 1

```
Node Id:      31
Statistic   Train   Validation
       0:  87.23%     64.71%
       1:  12.77%     35.29%
   Count:      47         34
```

< 39.5

```
Node Id:      32
Statistic   Train   Validation
       0: 100.00%    100.00%
       1:   0.00%      0.00%
   Count:      13          7
```

>= 39.5 Or Missing

```
Node Id:      33
Statistic   Train   Validation
       0:  29.45%     35.21%
       1:  70.55%     64.79%
   Count:     455        213
```

BodyMassIndex

1 Or Missing

```
Node Id:      34
Statistic   Train   Validation
       0:  45.91%     42.94%
       1:  54.09%     57.06%
   Count:     403        170
```

AnnualFamilyIncome

2

```
Node Id:      35
Statistic   Train   Validation
       0: 100.00%    100.00%
       1:   0.00%      0.00%
   Count:      21          8
```

2 Or Missing

```
Node Id:      38
Statistic   Train   Validation
       0:  68.56%     67.44%
       1:  31.44%     32.56%
   Count:     353        129
```

AnnualFamilyIncome

1

```
Node Id:      39
Statistic   Train   Validation
       0:  88.32%     95.35%
       1:  11.68%      4.65%
   Count:     197         86
```

1

```
Node Id:      42
Statistic   Train   Validation
       0:  79.84%     79.63%
       1:  20.16%     20.37%
   Count:     248        108
```

AnnualFamilyIncome

2 Or Missing

```
Node Id:      43
Statistic   Train   Validation
       0:  60.46%     57.84%
       1:  39.54%     42.16%
   Count:     349        185
```

TotalnumberFamily

< 36.5

```
Node Id:      44
Statistic   Train   Validation
       0:  58.62%     25.00%
       1:  41.38%     75.00%
   Count:      29          8
```

TotalnumberFamily

>= 36.5 Or Missing

```
Node Id:      45
Statistic   Train   Validation
       0:   3.08%     18.18%
       1:  96.92%     81.82%
   Count:      65         11
```

< 32.68 Or Missing

```
Node Id:      46
Statistic   Train   Validation
       0:  92.73%     78.26%
       1:   7.27%     21.74%
   Count:      55         23
```

AnnualFamilyIncome

>= 32.68

```
Node Id:      47
Statistic   Train   Validation
       0:  44.68%     26.92%
       1:  55.32%     73.08%
   Count:      47         26
```

MaritalStatus

< 31.545

```
Node Id:      48
Statistic   Train   Validation
       0:  42.42%     27.27%
       1:  57.58%     72.73%
   Count:      33         11
```

BodyMassIndex

>= 31.545 Or Missing

```
Node Id:      49
Statistic   Train   Validation
       0:  11.76%     12.07%
       1:  88.24%     87.93%
   Count:     136         58
```

2

```
Node Id:      50
Statistic   Train   Validation
       0:  55.56%     66.67%
       1:  44.44%     33.33%
   Count:      18          3
```

1 Or Missing

```
Node Id:      51
Statistic   Train   Validation
       0:   5.88%     27.78%
       1:  94.12%     72.22%
   Count:      34         18
```

< 31

```
Node Id:      52
Statistic   Train   Validation
       0:  54.97%     64.52%
       1:  45.03%     35.48%
   Count:     151         62
```

>= 31 Or Missing

```
Node Id:      53
Statistic   Train   Validation
       0:  29.21%     25.00%
       1:  70.79%     75.00%
   Count:     178         84
```

MaritalStatus

< 25.905

```
Node Id:      54
Statistic   Train   Validation
       0:  56.25%     72.41%
       1:  43.75%     27.59%
   Count:      64         29
```

AnnualFamilyIncome

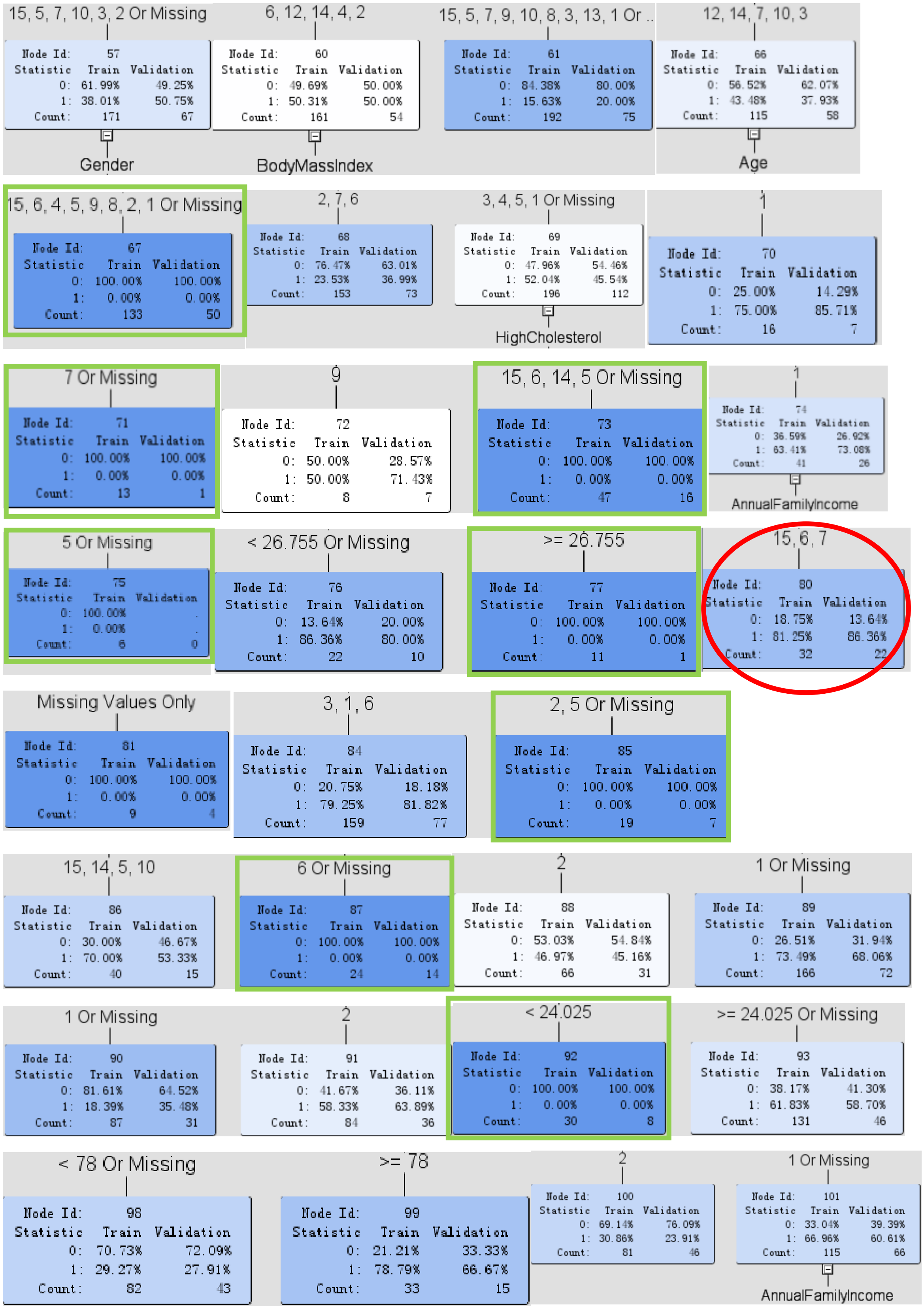
>= 25.905 Or Missing

```
Node Id:      55
Statistic   Train   Validation
       0:  25.06%     29.35%
       1:  74.94%     70.65%
   Count:     391        184
```

6, 12, 14, 4, 9, 8, 1

```
Node Id:      56
Statistic   Train   Validation
       0:  34.05%     38.83%
       1:  65.95%     61.17%
   Count:     232        103
```

Gender



```
15, 5, 7, 10, 3, 2 Or Missing
Node Id:      57
Statistic   Train   Validation
   0:      61.99%     49.25%
   1:      38.01%     50.75%
Count:       171         67
Gender
```

```
6, 12, 14, 4, 2
Node Id:      60
Statistic   Train   Validation
   0:      49.69%     50.00%
   1:      50.31%     50.00%
Count:       161         54
BodyMassIndex
```

```
15, 5, 7, 9, 10, 8, 3, 13, 1 Or ..
Node Id:      61
Statistic   Train   Validation
   0:      84.38%     80.00%
   1:      15.63%     20.00%
Count:       192         75
```

```
12, 14, 7, 10, 3
Node Id:      66
Statistic   Train   Validation
   0:      56.52%     62.07%
   1:      43.48%     37.93%
Count:       115         58
Age
```

```
15, 6, 4, 5, 9, 8, 2, 1 Or Missing
Node Id:      67
Statistic   Train   Validation
   0:     100.00%    100.00%
   1:       0.00%      0.00%
Count:       133         50
```

```
2, 7, 6
Node Id:      68
Statistic   Train   Validation
   0:      76.47%     63.01%
   1:      23.53%     36.99%
Count:       153         73
```

```
3, 4, 5, 1 Or Missing
Node Id:      69
Statistic   Train   Validation
   0:      47.96%     54.46%
   1:      52.04%     45.54%
Count:       196        112
HighCholesterol
```

```
1
Node Id:      70
Statistic   Train   Validation
   0:      25.00%     14.29%
   1:      75.00%     85.71%
Count:        16          7
```

```
7 Or Missing
Node Id:      71
Statistic   Train   Validation
   0:     100.00%    100.00%
   1:       0.00%      0.00%
Count:        13          1
```

```
9
Node Id:      72
Statistic   Train   Validation
   0:      50.00%     28.57%
   1:      50.00%     71.43%
Count:         8          7
```

```
15, 6, 14, 5 Or Missing
Node Id:      73
Statistic   Train   Validation
   0:     100.00%    100.00%
   1:       0.00%      0.00%
Count:        47         16
```

```
1
Node Id:      74
Statistic   Train   Validation
   0:      36.59%     26.92%
   1:      63.41%     73.08%
Count:        41         26
AnnualFamilyIncome
```

```
5 Or Missing
Node Id:      75
Statistic   Train   Validation
   0:     100.00%          .
   1:       0.00%          .
Count:         6          0
```

```
< 26.755 Or Missing
Node Id:      76
Statistic   Train   Validation
   0:      13.64%     20.00%
   1:      86.36%     80.00%
Count:        22         10
```

```
>= 26.755
Node Id:      77
Statistic   Train   Validation
   0:     100.00%    100.00%
   1:       0.00%      0.00%
Count:        11          1
```

```
15, 6, 7
Node Id:      80
Statistic   Train   Validation
   0:      18.75%     13.64%
   1:      81.25%     86.36%
Count:        32         22
```

```
Missing Values Only
Node Id:      81
Statistic   Train   Validation
   0:     100.00%    100.00%
   1:       0.00%      0.00%
Count:         9          4
```

```
3, 1, 6
Node Id:      84
Statistic   Train   Validation
   0:      20.75%     18.18%
   1:      79.25%     81.82%
Count:       159         77
```

```
2, 5 Or Missing
Node Id:      85
Statistic   Train   Validation
   0:     100.00%    100.00%
   1:       0.00%      0.00%
Count:        19          7
```

```
15, 14, 5, 10
Node Id:      86
Statistic   Train   Validation
   0:      30.00%     46.67%
   1:      70.00%     53.33%
Count:        40         15
```

```
6 Or Missing
Node Id:      87
Statistic   Train   Validation
   0:     100.00%    100.00%
   1:       0.00%      0.00%
Count:        24         14
```

```
2
Node Id:      88
Statistic   Train   Validation
   0:      53.03%     54.84%
   1:      46.97%     45.16%
Count:        66         31
```

```
1 Or Missing
Node Id:      89
Statistic   Train   Validation
   0:      26.51%     31.94%
   1:      73.49%     68.06%
Count:       166         72
```

```
1 Or Missing
Node Id:      90
Statistic   Train   Validation
   0:      81.61%     64.52%
   1:      18.39%     35.48%
Count:        87         31
```

```
2
Node Id:      91
Statistic   Train   Validation
   0:      41.67%     36.11%
   1:      58.33%     63.89%
Count:        84         36
```

```
< 24.025
Node Id:      92
Statistic   Train   Validation
   0:     100.00%    100.00%
   1:       0.00%      0.00%
Count:        30          8
```

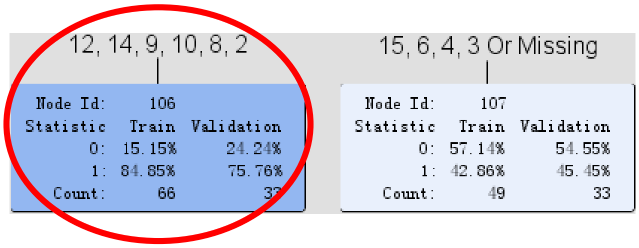
```
>= 24.025 Or Missing
Node Id:      93
Statistic   Train   Validation
   0:      38.17%     41.30%
   1:      61.83%     58.70%
Count:       131         46
```

```
< 78 Or Missing
Node Id:      98
Statistic   Train   Validation
   0:      70.73%     72.09%
   1:      29.27%     27.91%
Count:        82         43
```

```
>= 78
Node Id:      99
Statistic   Train   Validation
   0:      21.21%     33.33%
   1:      78.79%     66.67%
Count:        33         15
```

```
2
Node Id:     100
Statistic   Train   Validation
   0:      69.14%     76.09%
   1:      30.86%     23.91%
Count:        81         46
```

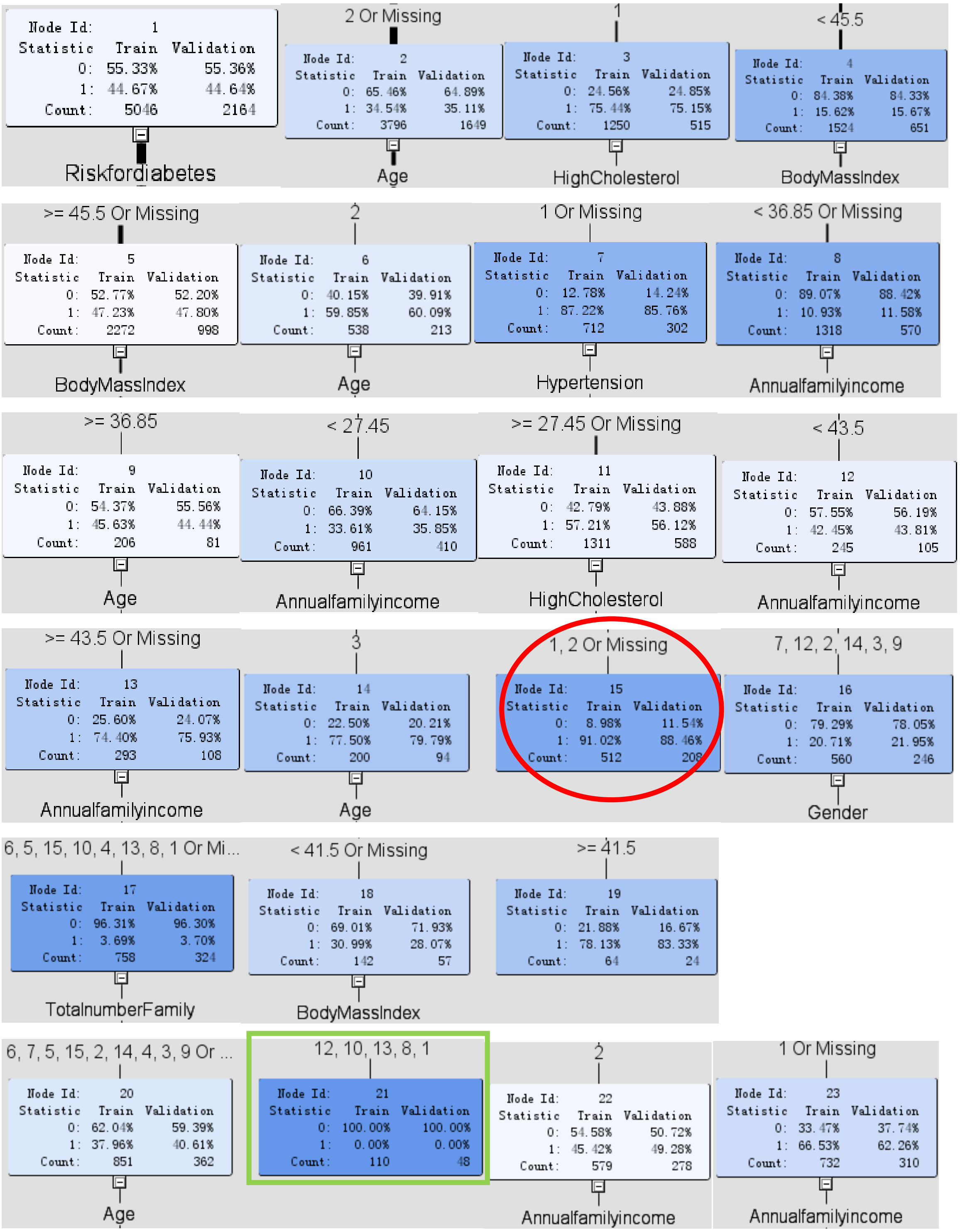
```
1 Or Missing
Node Id:     101
Statistic   Train   Validation
   0:      33.04%     39.39%
   1:      66.96%     60.61%
Count:       115         66
AnnualFamilyIncome
```



| Node Id: | 106 | |
|---|---|---|
| Statistic | Train | Validation |
| 0: | 15.15% | 24.24% |
| 1: | 84.85% | 75.76% |
| Count: | 66 | 33 |

| Node Id: | 107 | |
|---|---|---|
| Statistic | Train | Validation |
| 0: | 57.14% | 54.55% |
| 1: | 42.86% | 45.45% |
| Count: | 49 | 33 |

## Decision Tree Nodes of 2011-2012



```
Node Id:        1
Statistic   Train   Validation
       0:  55.33%      55.36%
       1:  44.67%      44.64%
   Count:   5046        2164
```
Riskfordiabetes

2 Or Missing
```
Node Id:        2
Statistic   Train   Validation
       0:  65.46%      64.89%
       1:  34.54%      35.11%
   Count:   3796        1649
```
Age

1
```
Node Id:        3
Statistic   Train   Validation
       0:  24.56%      24.85%
       1:  75.44%      75.15%
   Count:   1250         515
```
HighCholesterol

< 45.5
```
Node Id:        4
Statistic   Train   Validation
       0:  84.38%      84.33%
       1:  15.62%      15.67%
   Count:   1524         651
```
BodyMassIndex

>= 45.5 Or Missing
```
Node Id:        5
Statistic   Train   Validation
       0:  52.77%      52.20%
       1:  47.23%      47.80%
   Count:   2272         998
```
BodyMassIndex

2
```
Node Id:        6
Statistic   Train   Validation
       0:  40.15%      39.91%
       1:  59.85%      60.09%
   Count:    538         213
```
Age

1 Or Missing
```
Node Id:        7
Statistic   Train   Validation
       0:  12.78%      14.24%
       1:  87.22%      85.76%
   Count:    712         302
```
Hypertension

< 36.85 Or Missing
```
Node Id:        8
Statistic   Train   Validation
       0:  89.07%      88.42%
       1:  10.93%      11.58%
   Count:   1318         570
```
Annualfamilyincome

>= 36.85
```
Node Id:        9
Statistic   Train   Validation
       0:  54.37%      55.56%
       1:  45.63%      44.44%
   Count:    206          81
```
Age

< 27.45
```
Node Id:        10
Statistic   Train   Validation
       0:  66.39%      64.15%
       1:  33.61%      35.85%
   Count:    961         410
```
Annualfamilyincome

>= 27.45 Or Missing
```
Node Id:        11
Statistic   Train   Validation
       0:  42.79%      43.88%
       1:  57.21%      56.12%
   Count:   1311         588
```
HighCholesterol

< 43.5
```
Node Id:        12
Statistic   Train   Validation
       0:  57.55%      56.19%
       1:  42.45%      43.81%
   Count:    245         105
```
Annualfamilyincome

>= 43.5 Or Missing
```
Node Id:        13
Statistic   Train   Validation
       0:  25.60%      24.07%
       1:  74.40%      75.93%
   Count:    293         108
```
Annualfamilyincome

3
```
Node Id:        14
Statistic   Train   Validation
       0:  22.50%      20.21%
       1:  77.50%      79.79%
   Count:    200          94
```
Age

1, 2 Or Missing
```
Node Id:        15
Statistic   Train   Validation
       0:   8.98%      11.54%
       1:  91.02%      88.46%
   Count:    512         208
```

7, 12, 2, 14, 3, 9
```
Node Id:        16
Statistic   Train   Validation
       0:  79.29%      78.05%
       1:  20.71%      21.95%
   Count:    560         246
```
Gender

6, 5, 15, 10, 4, 13, 8, 1 Or Mi...
```
Node Id:        17
Statistic   Train   Validation
       0:  96.31%      96.30%
       1:   3.69%       3.70%
   Count:    758         324
```
TotalnumberFamily

< 41.5 Or Missing
```
Node Id:        18
Statistic   Train   Validation
       0:  69.01%      71.93%
       1:  30.99%      28.07%
   Count:    142          57
```
BodyMassIndex

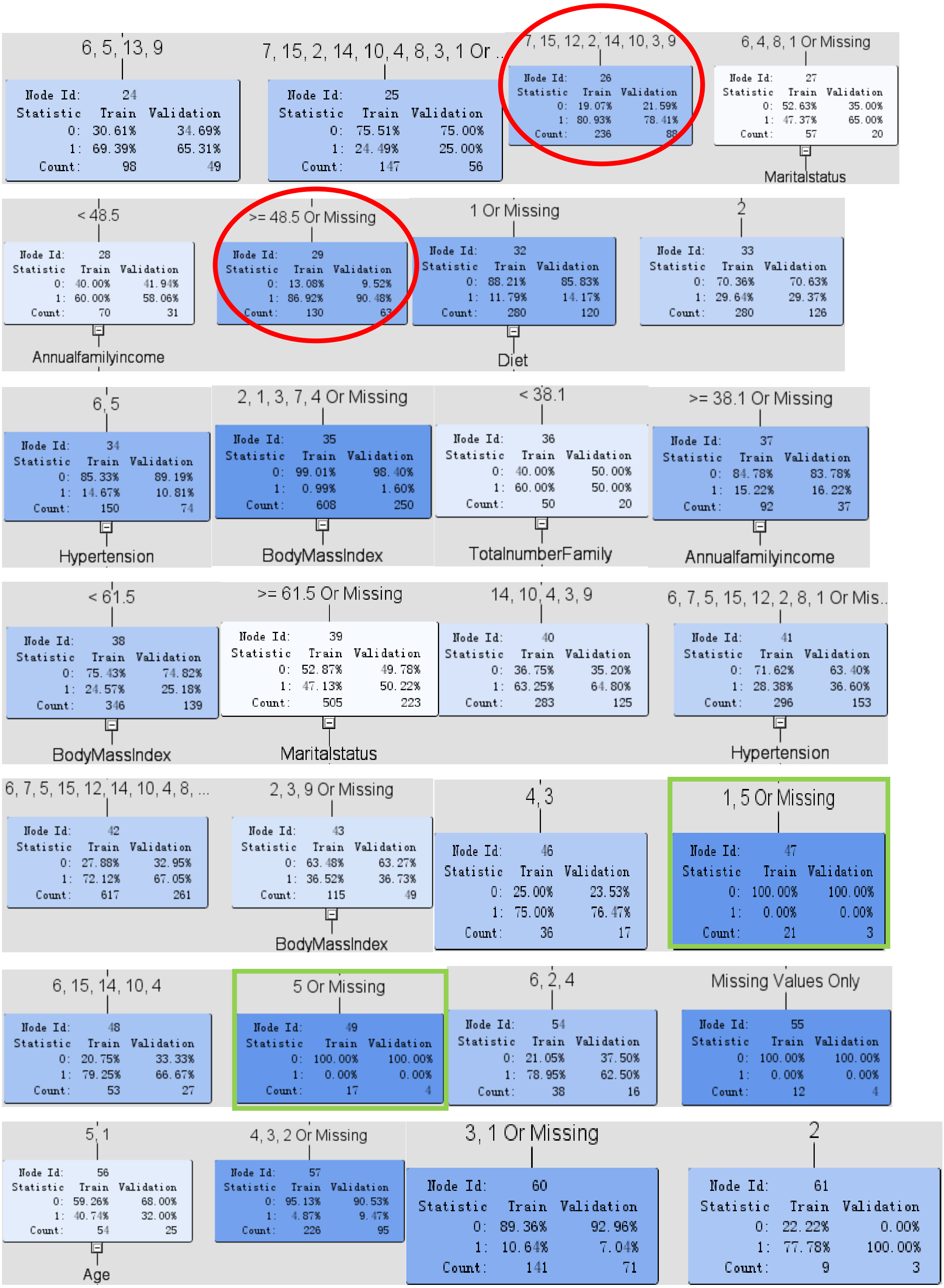
>= 41.5
```
Node Id:        19
Statistic   Train   Validation
       0:  21.88%      16.67%
       1:  78.13%      83.33%
   Count:     64          24
```

6, 7, 5, 15, 2, 14, 4, 3, 9 Or ...
```
Node Id:        20
Statistic   Train   Validation
       0:  62.04%      59.39%
       1:  37.96%      40.61%
   Count:    851         362
```
Age

12, 10, 13, 8, 1
```
Node Id:        21
Statistic   Train   Validation
       0: 100.00%     100.00%
       1:   0.00%       0.00%
   Count:    110          48
```

2
```
Node Id:        22
Statistic   Train   Validation
       0:  54.58%      50.72%
       1:  45.42%      49.28%
   Count:    579         278
```
Annualfamilyincome

1 Or Missing
```
Node Id:        23
Statistic   Train   Validation
       0:  33.47%      37.74%
       1:  66.53%      62.26%
   Count:    732         310
```
Annualfamilyincome



```
6, 5, 13, 9
Node Id:      24
Statistic   Train   Validation
       0:  30.61%       34.69%
       1:  69.39%       65.31%
   Count:      98           49
```

```
7, 15, 2, 14, 10, 4, 8, 3, 1 Or ...
Node Id:      25
Statistic   Train   Validation
       0:  75.51%       75.00%
       1:  24.49%       25.00%
   Count:     147           56
```

```
7, 15, 12, 2, 14, 10, 3, 9
Node Id:      26
Statistic   Train   Validation
       0:  19.07%       21.59%
       1:  80.93%       78.41%
   Count:     236           88
```

```
6, 4, 8, 1 Or Missing
Node Id:      27
Statistic   Train   Validation
       0:  52.63%       35.00%
       1:  47.37%       65.00%
   Count:      57           20
Maritalstatus
```

```
< 48.5
Node Id:      28
Statistic   Train   Validation
       0:  40.00%       41.94%
       1:  60.00%       58.06%
   Count:      70           31
Annualfamilyincome
```

```
>= 48.5 Or Missing
Node Id:      29
Statistic   Train   Validation
       0:  13.08%        9.52%
       1:  86.92%       90.48%
   Count:     130           63
```

```
1 Or Missing
Node Id:      32
Statistic   Train   Validation
       0:  88.21%       85.83%
       1:  11.79%       14.17%
   Count:     280          120
Diet
```

```
2
Node Id:      33
Statistic   Train   Validation
       0:  70.36%       70.63%
       1:  29.64%       29.37%
   Count:     280          126
```

```
6, 5
Node Id:      34
Statistic   Train   Validation
       0:  85.33%       89.19%
       1:  14.67%       10.81%
   Count:     150           74
Hypertension
```

```
2, 1, 3, 7, 4 Or Missing
Node Id:      35
Statistic   Train   Validation
       0:  99.01%       98.40%
       1:   0.99%        1.60%
   Count:     608          250
BodyMassIndex
```

```
< 38.1
Node Id:      36
Statistic   Train   Validation
       0:  40.00%       50.00%
       1:  60.00%       50.00%
   Count:      50           20
TotalnumberFamily
```

```
>= 38.1 Or Missing
Node Id:      37
Statistic   Train   Validation
       0:  84.78%       83.78%
       1:  15.22%       16.22%
   Count:      92           37
Annualfamilyincome
```

```
< 61.5
Node Id:      38
Statistic   Train   Validation
       0:  75.43%       74.82%
       1:  24.57%       25.18%
   Count:     346          139
BodyMassIndex
```

```
>= 61.5 Or Missing
Node Id:      39
Statistic   Train   Validation
       0:  52.87%       49.78%
       1:  47.13%       50.22%
   Count:     505          223
Maritalstatus
```

```
14, 10, 4, 3, 9
Node Id:      40
Statistic   Train   Validation
       0:  36.75%       35.20%
       1:  63.25%       64.80%
   Count:     283          125
```

```
6, 7, 5, 15, 12, 2, 8, 1 Or Mis...
Node Id:      41
Statistic   Train   Validation
       0:  71.62%       63.40%
       1:  28.38%       36.60%
   Count:     296          153
Hypertension
```

```
6, 7, 5, 15, 12, 14, 10, 4, 8, ...
Node Id:      42
Statistic   Train   Validation
       0:  27.88%       32.95%
       1:  72.12%       67.05%
   Count:     617          261
```

```
2, 3, 9 Or Missing
Node Id:      43
Statistic   Train   Validation
       0:  63.48%       63.27%
       1:  36.52%       36.73%
   Count:     115           49
BodyMassIndex
```

```
4, 3
Node Id:      46
Statistic   Train   Validation
       0:  25.00%       23.53%
       1:  75.00%       76.47%
   Count:      36           17
```

```
1, 5 Or Missing
Node Id:      47
Statistic   Train   Validation
       0: 100.00%      100.00%
       1:   0.00%        0.00%
   Count:      21            3
```

```
6, 15, 14, 10, 4
Node Id:      48
Statistic   Train   Validation
       0:  20.75%       33.33%
       1:  79.25%       66.67%
   Count:      53           27
```

```
5 Or Missing
Node Id:      49
Statistic   Train   Validation
       0: 100.00%      100.00%
       1:   0.00%        0.00%
   Count:      17            4
```

```
6, 2, 4
Node Id:      54
Statistic   Train   Validation
       0:  21.05%       37.50%
       1:  78.95%       62.50%
   Count:      38           16
```

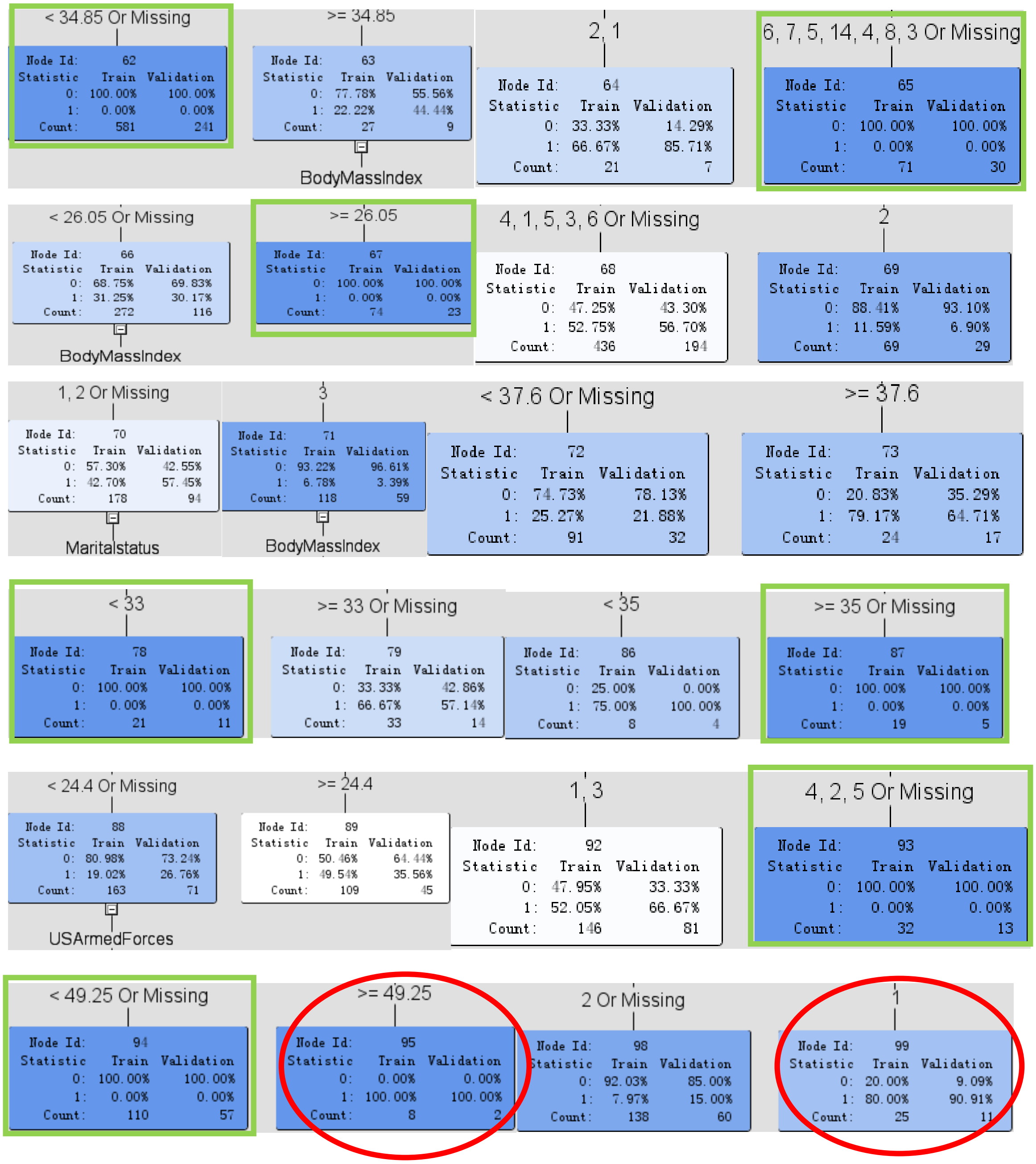
```
Missing Values Only
Node Id:      55
Statistic   Train   Validation
       0: 100.00%      100.00%
       1:   0.00%        0.00%
   Count:      12            4
```

```
5, 1
Node Id:      56
Statistic   Train   Validation
       0:  59.26%       68.00%
       1:  40.74%       32.00%
   Count:      54           25
Age
```

```
4, 3, 2 Or Missing
Node Id:      57
Statistic   Train   Validation
       0:  95.13%       90.53%
       1:   4.87%        9.47%
   Count:     226           95
```

```
3, 1 Or Missing
Node Id:      60
Statistic   Train   Validation
       0:  89.36%       92.96%
       1:  10.64%        7.04%
   Count:     141           71
```

```
2
Node Id:      61
Statistic   Train   Validation
       0:  22.22%        0.00%
       1:  77.78%      100.00%
   Count:       9            3
```


< 34.85 Or Missing
Node Id: 62
Statistic Train Validation
0: 100.00% 100.00%
1: 0.00% 0.00%
Count: 581 241
>= 34.85
Node Id: 63
Statistic Train Validation
0: 77.78% 55.56%
1: 22.22% 44.44%
Count: 27 9
BodyMassIndex
2, 1
Node Id: 64
Statistic Train Validation
0: 33.33% 14.29%
1: 66.67% 85.71%
Count: 21 7
6, 7, 5, 14, 4, 8, 3 Or Missing
Node Id: 65
Statistic Train Validation
0: 100.00% 100.00%
1: 0.00% 0.00%
Count: 71 30
< 26.05 Or Missing
Node Id: 66
Statistic Train Validation
0: 68.75% 69.83%
1: 31.25% 30.17%
Count: 272 116
BodyMassIndex
>= 26.05
Node Id: 67
Statistic Train Validation
0: 100.00% 100.00%
1: 0.00% 0.00%
Count: 74 23
4, 1, 5, 3, 6 Or Missing
Node Id: 68
Statistic Train Validation
0: 47.25% 43.30%
1: 52.75% 56.70%
Count: 436 194
2
Node Id: 69
Statistic Train Validation
0: 88.41% 93.10%
1: 11.59% 6.90%
Count: 69 29
1, 2 Or Missing
Node Id: 70
Statistic Train Validation
0: 57.30% 42.55%
1: 42.70% 57.45%
Count: 178 94
Maritalstatus
3
Node Id: 71
Statistic Train Validation
0: 93.22% 96.61%
1: 6.78% 3.39%
Count: 118 59
BodyMassIndex
< 37.6 Or Missing
Node Id: 72
Statistic Train Validation
0: 74.73% 78.13%
1: 25.27% 21.88%
Count: 91 32
>= 37.6
Node Id: 73
Statistic Train Validation
0: 20.83% 35.29%
1: 79.17% 64.71%
Count: 24 17
< 33
Node Id: 78
Statistic Train Validation
0: 100.00% 100.00%
1: 0.00% 0.00%
Count: 21 11
>= 33 Or Missing
Node Id: 79
Statistic Train Validation
0: 33.33% 42.86%
1: 66.67% 57.14%
Count: 33 14
< 35
Node Id: 86
Statistic Train Validation
0: 25.00% 0.00%
1: 75.00% 100.00%
Count: 8 4
>= 35 Or Missing
Node Id: 87
Statistic Train Validation
0: 100.00% 100.00%
1: 0.00% 0.00%
Count: 19 5
< 24.4 Or Missing
Node Id: 88
Statistic Train Validation
0: 80.98% 73.24%
1: 19.02% 26.76%
Count: 163 71
USArmedForces
>= 24.4
Node Id: 89
Statistic Train Validation
0: 50.46% 64.44%
1: 49.54% 35.56%
Count: 109 45
1, 3
Node Id: 92
Statistic Train Validation
0: 47.95% 33.33%
1: 52.05% 66.67%
Count: 146 81
4, 2, 5 Or Missing
Node Id: 93
Statistic Train Validation
0: 100.00% 100.00%
1: 0.00% 0.00%
Count: 32 13
< 49.25 Or Missing
Node Id: 94
Statistic Train Validation
0: 100.00% 100.00%
1: 0.00% 0.00%
Count: 110 57
>= 49.25
Node Id: 95
Statistic Train Validation
0: 0.00% 0.00%
1: 100.00% 100.00%
Count: 8 2
2 Or Missing
Node Id: 98
Statistic Train Validation
0: 92.03% 85.00%
1: 7.97% 15.00%
Count: 138 60
1
Node Id: 99
Statistic Train Validation
0: 20.00% 9.09%
1: 80.00% 90.91%
Count: 25 11

# APPENDIX B

## 2007-2008 Tree Leaf Report

```
*------------------------------------------------------------*
 Node = 17
*------------------------------------------------------------*
if Smoked IS ONE OF: 1
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND AnnualFamilyIncome IS ONE OF: 3, 5, 2, 7
AND Age < 41.5
then
 Tree Node Identifier   = 17
 Number of Observations = 82
 Predicted: Prediabetes=1 = 0.41
 Predicted: Prediabetes=0 = 0.59

*------------------------------------------------------------*
 Node = 19
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 2 or MISSING
AND BodyMassIndex >= 47.65
AND AnnualFamilyIncome IS ONE OF: 14, 9, 15, 6, 12, 8, 10, 4, 1 or MISSING
AND Age < 41.5
then
 Tree Node Identifier   = 19
 Number of Observations = 8
 Predicted: Prediabetes=1 = 0.63
 Predicted: Prediabetes=0 = 0.38

*------------------------------------------------------------*
 Node = 20
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 3
AND HighCholesterol IS ONE OF: 2
AND Age >= 41.5 or MISSING
then
 Tree Node Identifier   = 20
 Number of Observations = 482
 Predicted: Prediabetes=1 = 0.19
 Predicted: Prediabetes=0 = 0.81

*------------------------------------------------------------*
 Node = 22
```

```
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 1, 2 or MISSING
AND Diet IS ONE OF: 5
AND Age >= 41.5 or MISSING
then
 Tree Node Identifier     = 22
 Number of Observations = 109
 Predicted: Prediabetes=1 = 0.82
 Predicted: Prediabetes=0 = 0.18

*------------------------------------------------------------*
 Node = 24
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 1
AND Hypertension IS ONE OF: 3
AND Diet IS ONE OF: 2, 5
AND Age < 26.5
then
 Tree Node Identifier     = 24
 Number of Observations = 8
 Predicted: Prediabetes=1 = 0.63
 Predicted: Prediabetes=0 = 0.38

*------------------------------------------------------------*
 Node = 25
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 1
AND Hypertension IS ONE OF: 3
AND Diet IS ONE OF: 2, 5
AND Age >= 26.5 or MISSING
then
 Tree Node Identifier     = 25
 Number of Observations = 27
 Predicted: Prediabetes=1 = 0.00
 Predicted: Prediabetes=0 = 1.00

*------------------------------------------------------------*
 Node = 26
*------------------------------------------------------------*
if USArmedForces IS ONE OF: 1
AND Riskfordiabetes IS ONE OF: 1
AND Hypertension IS ONE OF: 3
AND Diet IS ONE OF: 3, 4, 1 or MISSING
then
```

```
 Tree Node Identifier     = 26
 Number of Observations = 10
 Predicted: Prediabetes=1 = 0.00
 Predicted: Prediabetes=0 = 1.00

*------------------------------------------------------------*
 Node = 27
*------------------------------------------------------------*
if USArmedForces IS ONE OF: 2 or MISSING
AND Riskfordiabetes IS ONE OF: 1
AND Hypertension IS ONE OF: 3
AND Diet IS ONE OF: 3, 4, 1 or MISSING
then
 Tree Node Identifier     = 27
 Number of Observations = 284
 Predicted: Prediabetes=1 = 0.70
 Predicted: Prediabetes=0 = 0.30

*------------------------------------------------------------*
 Node = 28
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 1
AND Hypertension IS ONE OF: 1, 2 or MISSING
AND BodyMassIndex < 39.2 or MISSING
AND Age < 29.5
then
 Tree Node Identifier     = 28
 Number of Observations = 9
 Predicted: Prediabetes=1 = 0.00
 Predicted: Prediabetes=0 = 1.00

*------------------------------------------------------------*
 Node = 29
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 1
AND Hypertension IS ONE OF: 1, 2 or MISSING
AND BodyMassIndex >= 39.2
AND Age < 29.5
then
 Tree Node Identifier     = 29
 Number of Observations = 8
 Predicted: Prediabetes=1 = 0.88
 Predicted: Prediabetes=0 = 0.13

*------------------------------------------------------------*
```

```
  Node = 31
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 1
AND MaritalStatus IS ONE OF: 1, 3, 4, 2 or MISSING
AND Hypertension IS ONE OF: 1, 2 or MISSING
AND Age >= 29.5 or MISSING
then
  Tree Node Identifier     = 31
  Number of Observations = 524
  Predicted: Prediabetes=1 = 0.87
  Predicted: Prediabetes=0 = 0.13

*------------------------------------------------------------*
  Node = 34
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 1
AND MaritalStatus IS ONE OF: 5, 6
AND Hypertension IS ONE OF: 1, 2 or MISSING
AND AnnualFamilyIncome IS ONE OF: 14, 5, 7
AND Age >= 29.5 or MISSING
then
  Tree Node Identifier     = 34
  Number of Observations = 25
  Predicted: Prediabetes=1 = 0.92
  Predicted: Prediabetes=0 = 0.08

*------------------------------------------------------------*
  Node = 35
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 1
AND MaritalStatus IS ONE OF: 5, 6
AND Hypertension IS ONE OF: 1, 2 or MISSING
AND AnnualFamilyIncome equals Missing
AND Age >= 29.5 or MISSING
then
  Tree Node Identifier     = 35
  Number of Observations = 13
  Predicted: Prediabetes=1 = 0.00
  Predicted: Prediabetes=0 = 1.00

*------------------------------------------------------------*
  Node = 37
*------------------------------------------------------------*
if Smoked IS ONE OF: 2 or MISSING
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND Diet IS ONE OF: 3, 4 or MISSING
AND AnnualFamilyIncome IS ONE OF: 3, 5, 2, 7
AND Age < 41.5
then
 Tree Node Identifier    = 37
 Number of Observations = 64
 Predicted: Prediabetes=1 = 0.00
 Predicted: Prediabetes=0 = 1.00

*------------------------------------------------------------*
 Node = 44
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 3
AND HighCholesterol IS ONE OF: 1 or MISSING
AND BodyMassIndex < 25.515
AND Age >= 41.5 or MISSING
then
 Tree Node Identifier    = 44
 Number of Observations = 133
 Predicted: Prediabetes=1 = 0.28
 Predicted: Prediabetes=0 = 0.72

*------------------------------------------------------------*
 Node = 47
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 2 or MISSING
AND MaritalStatus IS ONE OF: 1, 3 or MISSING
AND Hypertension IS ONE OF: 1, 2 or MISSING
AND Diet IS ONE OF: 3, 2, 4, 1 or MISSING
AND Age >= 41.5 or MISSING
then
 Tree Node Identifier    = 47
 Number of Observations = 819
 Predicted: Prediabetes=1 = 0.53
 Predicted: Prediabetes=0 = 0.47

*------------------------------------------------------------*
 Node = 48
*------------------------------------------------------------*
if TotalnumberFamily IS ONE OF: 3
AND Smoked IS ONE OF: 2 or MISSING
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND Diet IS ONE OF: 2
AND AnnualFamilyIncome IS ONE OF: 3, 5, 2, 7
AND Age < 41.5
then
 Tree Node Identifier     = 48
 Number of Observations = 8
 Predicted: Prediabetes=1 = 1.00
 Predicted: Prediabetes=0 = 0.00

*------------------------------------------------------------*
 Node = 49
*------------------------------------------------------------*
if TotalnumberFamily equals Missing
AND Smoked IS ONE OF: 2 or MISSING
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND Diet IS ONE OF: 2
AND AnnualFamilyIncome IS ONE OF: 3, 5, 2, 7
AND Age < 41.5
then
 Tree Node Identifier     = 49
 Number of Observations = 11
 Predicted: Prediabetes=1 = 0.00
 Predicted: Prediabetes=0 = 1.00

*------------------------------------------------------------*
 Node = 51
*------------------------------------------------------------*
if USArmedForces IS ONE OF: 2 or MISSING
AND TotalnumberFamily IS ONE OF: 2, 4, 3, 5, 1, 7 or MISSING
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND BodyMassIndex < 47.65 or MISSING
AND AnnualFamilyIncome IS ONE OF: 14, 9, 15, 6, 12, 8, 10, 4, 1 or MISSING
AND Age < 41.5
then
 Tree Node Identifier     = 51
 Number of Observations = 316
 Predicted: Prediabetes=1 = 0.00
 Predicted: Prediabetes=0 = 1.00

*------------------------------------------------------------*
 Node = 53
*------------------------------------------------------------*
if USArmedForces IS ONE OF: 1
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND BodyMassIndex < 47.65 or MISSING
AND AnnualFamilyIncome IS ONE OF: 15 or MISSING
AND Age < 41.5
```

```
then
  Tree Node Identifier     = 53
  Number of Observations = 8
  Predicted: Prediabetes=1 = 0.00
  Predicted: Prediabetes=0 = 1.00

*--------------------------------------------------------*
  Node = 63
*--------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 2 or MISSING
AND MaritalStatus IS ONE OF: 5, 4, 2, 6
AND Hypertension IS ONE OF: 1, 2 or MISSING
AND Diet IS ONE OF: 3, 2, 4, 1 or MISSING
AND AnnualFamilyIncome IS ONE OF: 2, 12, 8, 10, 7, 1 or MISSING
AND Age >= 41.5 or MISSING
then
  Tree Node Identifier     = 63
  Number of Observations = 55
  Predicted: Prediabetes=1 = 0.00
  Predicted: Prediabetes=0 = 1.00

*--------------------------------------------------------*
  Node = 64
*--------------------------------------------------------*
if USArmedForces IS ONE OF: 2 or MISSING
AND TotalnumberFamily IS ONE OF: 6
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND HighCholesterol IS ONE OF: 2 or MISSING
AND BodyMassIndex < 47.65 or MISSING
AND AnnualFamilyIncome IS ONE OF: 14, 9, 15, 6, 12, 8, 10, 4, 1 or MISSING
AND Age < 41.5
then
  Tree Node Identifier     = 64
  Number of Observations = 20
  Predicted: Prediabetes=1 = 0.00
  Predicted: Prediabetes=0 = 1.00

*--------------------------------------------------------*
  Node = 65
*--------------------------------------------------------*
if USArmedForces IS ONE OF: 2 or MISSING
AND TotalnumberFamily IS ONE OF: 6
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND HighCholesterol IS ONE OF: 1
AND BodyMassIndex < 47.65 or MISSING
AND AnnualFamilyIncome IS ONE OF: 14, 9, 15, 6, 12, 8, 10, 4, 1 or MISSING
AND Age < 41.5
then
 Tree Node Identifier     = 65
 Number of Observations = 8
 Predicted: Prediabetes=1 = 0.63
 Predicted: Prediabetes=0 = 0.38

*------------------------------------------------------------*
 Node = 66
*------------------------------------------------------------*
if USArmedForces IS ONE OF: 1
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND BodyMassIndex < 26.26
AND AnnualFamilyIncome IS ONE OF: 14
AND Age < 41.5
then
 Tree Node Identifier     = 66
 Number of Observations = 10
 Predicted: Prediabetes=1 = 0.80
 Predicted: Prediabetes=0 = 0.20

*------------------------------------------------------------*
 Node = 67
*------------------------------------------------------------*
if USArmedForces IS ONE OF: 1
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND BodyMassIndex < 47.65 AND BodyMassIndex >= 26.26 or MISSING
AND AnnualFamilyIncome IS ONE OF: 14
AND Age < 41.5
then
 Tree Node Identifier     = 67
 Number of Observations = 15
 Predicted: Prediabetes=1 = 0.00
 Predicted: Prediabetes=0 = 1.00

*------------------------------------------------------------*
 Node = 75
*------------------------------------------------------------*
if TotalnumberFamily IS ONE OF: 4, 3
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 3
AND HighCholesterol IS ONE OF: 1 or MISSING
AND BodyMassIndex >= 25.515 or MISSING
AND AnnualFamilyIncome IS ONE OF: 14, 15, 5, 7 or MISSING
AND Age >= 41.5 or MISSING
then
 Tree Node Identifier   = 75
 Number of Observations = 36
 Predicted: Prediabetes=1 = 0.00
 Predicted: Prediabetes=0 = 1.00

*------------------------------------------------------------*
 Node = 76
*------------------------------------------------------------*
if TotalnumberFamily IS ONE OF: 2, 5, 1, 7, 6 or MISSING
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND MaritalStatus IS ONE OF: 2, 6
AND Hypertension IS ONE OF: 3
AND HighCholesterol IS ONE OF: 1 or MISSING
AND BodyMassIndex >= 25.515 or MISSING
AND Age >= 41.5 or MISSING
then
 Tree Node Identifier   = 76
 Number of Observations = 27
 Predicted: Prediabetes=1 = 0.19
 Predicted: Prediabetes=0 = 0.81

*------------------------------------------------------------*
 Node = 77
*------------------------------------------------------------*
if TotalnumberFamily IS ONE OF: 2, 5, 1, 7, 6 or MISSING
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND MaritalStatus IS ONE OF: 1, 3, 5 or MISSING
AND Hypertension IS ONE OF: 3
AND HighCholesterol IS ONE OF: 1 or MISSING
AND BodyMassIndex >= 25.515 or MISSING
AND Age >= 41.5 or MISSING
then
 Tree Node Identifier   = 77
 Number of Observations = 280
 Predicted: Prediabetes=1 = 0.66
 Predicted: Prediabetes=0 = 0.34

*------------------------------------------------------------*
 Node = 78
*------------------------------------------------------------*
if TotalnumberFamily IS ONE OF: 2, 1
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND MaritalStatus IS ONE OF: 5, 4, 2, 6
AND Hypertension IS ONE OF: 1, 2 or MISSING
AND Diet IS ONE OF: 3, 2, 4, 1 or MISSING
AND AnnualFamilyIncome IS ONE OF: 9, 3, 15, 5, 6, 4, 13
AND Age >= 41.5 or MISSING
then
 Tree Node Identifier     = 78
 Number of Observations = 154
 Predicted: Prediabetes=1 = 0.53
 Predicted: Prediabetes=0 = 0.47

*---------------------------------------------------------*
 Node = 79
*---------------------------------------------------------*
if TotalnumberFamily IS ONE OF: 4, 3 or MISSING
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND MaritalStatus IS ONE OF: 5, 4, 2, 6
AND Hypertension IS ONE OF: 1, 2 or MISSING
AND Diet IS ONE OF: 3, 2, 4, 1 or MISSING
AND AnnualFamilyIncome IS ONE OF: 9, 3, 15, 5, 6, 4, 13
AND Age >= 41.5 or MISSING
then
 Tree Node Identifier     = 79
 Number of Observations = 24
 Predicted: Prediabetes=1 = 0.00
 Predicted: Prediabetes=0 = 1.00

*---------------------------------------------------------*
 Node = 80
*---------------------------------------------------------*
if TotalnumberFamily IS ONE OF: 4, 3
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 3
AND HighCholesterol IS ONE OF: 1 or MISSING
AND BodyMassIndex >= 25.515 or MISSING
AND AnnualFamilyIncome IS ONE OF: 8
AND Age >= 41.5 or MISSING
then
 Tree Node Identifier     = 80
 Number of Observations = 19
 Predicted: Prediabetes=1 = 0.84
 Predicted: Prediabetes=0 = 0.16

*---------------------------------------------------------*
 Node = 81
*---------------------------------------------------------*
```

```
if TotalnumberFamily IS ONE OF: 4, 3
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 3
AND HighCholesterol IS ONE OF: 1 or MISSING
AND BodyMassIndex >= 25.515 or MISSING
AND AnnualFamilyIncome equals Missing
AND Age >= 41.5 or MISSING
then
 Tree Node Identifier    = 81
 Number of Observations = 10
 Predicted: Prediabetes=1 = 0.00
 Predicted: Prediabetes=0 = 1.00
```

**2009-2010 Tree Leaf Report**

```
*------------------------------------------------------------*
 Node = 13
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 1
AND Hypertension IS ONE OF: 3
AND HighCholesterol IS ONE OF: 1
then
 Tree Node Identifier    = 13
 Number of Observations = 147
 Predicted: Prediabetes=1 = 0.78
 Predicted: Prediabetes=0 = 0.22

*------------------------------------------------------------*
 Node = 15
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 1
AND Hypertension IS ONE OF: 2, 1 or MISSING
AND AnnualFamilyIncome IS ONE OF: 15, 6, 12, 5, 3, 2 or MISSING
then
 Tree Node Identifier    = 15
 Number of Observations = 416
 Predicted: Prediabetes=1 = 0.90
 Predicted: Prediabetes=0 = 0.10

*------------------------------------------------------------*
 Node = 19
*------------------------------------------------------------*
if Smoked IS ONE OF: 2
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 2, 1 or MISSING
AND Diet IS ONE OF: 3, 1
then
 Tree Node Identifier     = 19
 Number of Observations = 352
 Predicted: Prediabetes=1 = 0.32
 Predicted: Prediabetes=0 = 0.68

*------------------------------------------------------------*
 Node = 21
*------------------------------------------------------------*
if TotalnumberFamily IS ONE OF: 3, 5, 7
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 3
AND Age < 49.5 or MISSING
then
 Tree Node Identifier     = 21
 Number of Observations = 300
 Predicted: Prediabetes=1 = 0.04
 Predicted: Prediabetes=0 = 0.96

*------------------------------------------------------------*
 Node = 27
*------------------------------------------------------------*
if TotalnumberFamily IS ONE OF: 3, 4, 5 or MISSING
AND Riskfordiabetes IS ONE OF: 1
AND Hypertension IS ONE OF: 2, 1 or MISSING
AND AnnualFamilyIncome IS ONE OF: 14, 4, 7, 9, 10, 8, 1
then
 Tree Node Identifier     = 27
 Number of Observations = 48
 Predicted: Prediabetes=1 = 0.40
 Predicted: Prediabetes=0 = 0.60

*------------------------------------------------------------*
 Node = 29
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 3
AND BodyMassIndex >= 37.25
AND AnnualFamilyIncome equals Missing
AND Age >= 49.5
then
 Tree Node Identifier     = 29
 Number of Observations = 11
 Predicted: Prediabetes=1 = 0.00
 Predicted: Prediabetes=0 = 1.00

*------------------------------------------------------------*
 Node = 31
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 2, 1 or MISSING
AND HighCholesterol IS ONE OF: 2
AND Diet IS ONE OF: 2, 4, 5 or MISSING
AND AnnualFamilyIncome IS ONE OF: 4, 7, 10, 1
then
 Tree Node Identifier     = 31
 Number of Observations = 47
 Predicted: Prediabetes=1 = 0.13
 Predicted: Prediabetes=0 = 0.87

*------------------------------------------------------------*
 Node = 32
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 2, 1 or MISSING
AND HighCholesterol IS ONE OF: 1 or MISSING
AND Diet IS ONE OF: 2, 4, 5 or MISSING
AND Age < 39.5
then
 Tree Node Identifier     = 32
 Number of Observations = 13
 Predicted: Prediabetes=1 = 0.00
 Predicted: Prediabetes=0 = 1.00

*------------------------------------------------------------*
 Node = 35
*------------------------------------------------------------*
if Smoked IS ONE OF: 1 or MISSING
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 2
AND Diet IS ONE OF: 3, 1
then
 Tree Node Identifier     = 35
 Number of Observations = 21
 Predicted: Prediabetes=1 = 0.00
 Predicted: Prediabetes=0 = 1.00

*------------------------------------------------------------*
```

```
 Node = 39
*------------------------------------------------------------*
if TotalnumberFamily IS ONE OF: 4, 2, 1, 6 or MISSING
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 3
AND Gender IS ONE OF: 1
AND Age < 49.5 or MISSING
then
 Tree Node Identifier   = 39
 Number of Observations = 197
 Predicted: Prediabetes=1 = 0.12
 Predicted: Prediabetes=0 = 0.88

*------------------------------------------------------------*
 Node = 45
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 1
AND Hypertension IS ONE OF: 3
AND HighCholesterol IS ONE OF: 2 or MISSING
AND AnnualFamilyIncome IS ONE OF: 12, 4, 8, 2, 1
AND Age >= 36.5 or MISSING
then
 Tree Node Identifier   = 45
 Number of Observations = 65
 Predicted: Prediabetes=1 = 0.97
 Predicted: Prediabetes=0 = 0.03

*------------------------------------------------------------*
 Node = 49
*------------------------------------------------------------*
if TotalnumberFamily IS ONE OF: 2, 1, 7
AND Riskfordiabetes IS ONE OF: 1
AND Hypertension IS ONE OF: 2, 1 or MISSING
AND BodyMassIndex >= 31.545 or MISSING
AND AnnualFamilyIncome IS ONE OF: 14, 4, 7, 9, 10, 8, 1
then
 Tree Node Identifier   = 49
 Number of Observations = 136
 Predicted: Prediabetes=1 = 0.88
 Predicted: Prediabetes=0 = 0.12

*------------------------------------------------------------*
 Node = 50
*------------------------------------------------------------*
if Smoked IS ONE OF: 2
```

```
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 3
AND BodyMassIndex >= 37.25
AND AnnualFamilyIncome IS ONE OF: 15, 6, 14, 4, 7
AND Age >= 49.5
then
 Tree Node Identifier     = 50
 Number of Observations = 18
 Predicted: Prediabetes=1 = 0.44
 Predicted: Prediabetes=0 = 0.56

*------------------------------------------------------------*
 Node = 51
*------------------------------------------------------------*
if Smoked IS ONE OF: 1 or MISSING
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 3
AND BodyMassIndex >= 37.25
AND AnnualFamilyIncome IS ONE OF: 15, 6, 14, 4, 7
AND Age >= 49.5
then
 Tree Node Identifier     = 51
 Number of Observations = 34
 Predicted: Prediabetes=1 = 0.94
 Predicted: Prediabetes=0 = 0.06

*------------------------------------------------------------*
 Node = 52
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 2, 1 or MISSING
AND HighCholesterol IS ONE OF: 2
AND Diet IS ONE OF: 2, 4, 5 or MISSING
AND BodyMassIndex < 31
AND AnnualFamilyIncome IS ONE OF: 15, 6, 12, 14, 5, 9, 8, 3, 13, 2 or MISSING
then
 Tree Node Identifier     = 52
 Number of Observations = 151
 Predicted: Prediabetes=1 = 0.45
 Predicted: Prediabetes=0 = 0.55

*------------------------------------------------------------*
 Node = 55
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 2 or MISSING
```

```
AND Hypertension IS ONE OF: 2, 1 or MISSING
AND HighCholesterol IS ONE OF: 1 or MISSING
AND Diet IS ONE OF: 2, 4, 5 or MISSING
AND BodyMassIndex >= 25.905 or MISSING
AND Age >= 39.5 or MISSING
then
 Tree Node Identifier     = 55
 Number of Observations = 391
 Predicted: Prediabetes=1 = 0.75
 Predicted: Prediabetes=0 = 0.25

*------------------------------------------------------------*
 Node = 61
*------------------------------------------------------------*
if TotalnumberFamily IS ONE OF: 4, 2, 1, 6 or MISSING
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 3
AND Gender IS ONE OF: 2 or MISSING
AND AnnualFamilyIncome IS ONE OF: 15, 5, 7, 9, 10, 8, 3, 13, 1 or MISSING
AND Age < 49.5 or MISSING
then
 Tree Node Identifier     = 61
 Number of Observations = 192
 Predicted: Prediabetes=1 = 0.16
 Predicted: Prediabetes=0 = 0.84

*------------------------------------------------------------*
 Node = 67
*------------------------------------------------------------*
if Smoked IS ONE OF: 1
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 3
AND BodyMassIndex < 37.25 or MISSING
AND AnnualFamilyIncome IS ONE OF: 15, 6, 4, 5, 9, 8, 2, 1 or MISSING
AND Age >= 49.5
then
 Tree Node Identifier     = 67
 Number of Observations = 133
 Predicted: Prediabetes=1 = 0.00
 Predicted: Prediabetes=0 = 1.00

*------------------------------------------------------------*
 Node = 68
*------------------------------------------------------------*
if TotalnumberFamily IS ONE OF: 2, 7, 6
```

```
AND Smoked IS ONE OF: 2 or MISSING
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 3
AND BodyMassIndex < 37.25 or MISSING
AND Age >= 49.5
then
 Tree Node Identifier     = 68
 Number of Observations = 153
 Predicted: Prediabetes=1 = 0.24
 Predicted: Prediabetes=0 = 0.76

*------------------------------------------------------------*
 Node = 70
*------------------------------------------------------------*
if TotalnumberFamily IS ONE OF: 1
AND Riskfordiabetes IS ONE OF: 1
AND Hypertension IS ONE OF: 3
AND HighCholesterol IS ONE OF: 2 or MISSING
AND AnnualFamilyIncome IS ONE OF: 12, 4, 8, 2, 1
AND Age < 36.5
then
 Tree Node Identifier     = 70
 Number of Observations = 16
 Predicted: Prediabetes=1 = 0.75
 Predicted: Prediabetes=0 = 0.25

*------------------------------------------------------------*
 Node = 71
*------------------------------------------------------------*
if TotalnumberFamily IS ONE OF: 7 or MISSING
AND Riskfordiabetes IS ONE OF: 1
AND Hypertension IS ONE OF: 3
AND HighCholesterol IS ONE OF: 2 or MISSING
AND AnnualFamilyIncome IS ONE OF: 12, 4, 8, 2, 1
AND Age < 36.5
then
 Tree Node Identifier     = 71
 Number of Observations = 13
 Predicted: Prediabetes=1 = 0.00
 Predicted: Prediabetes=0 = 1.00

*------------------------------------------------------------*
 Node = 72
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 1
```

```
AND Hypertension IS ONE OF: 3
AND HighCholesterol IS ONE OF: 2 or MISSING
AND BodyMassIndex < 32.68 or MISSING
AND AnnualFamilyIncome IS ONE OF: 9
then
 Tree Node Identifier     = 72
 Number of Observations = 8
 Predicted: Prediabetes=1 = 0.50
 Predicted: Prediabetes=0 = 0.50

*---------------------------------------------------------*
 Node = 73
*---------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 1
AND Hypertension IS ONE OF: 3
AND HighCholesterol IS ONE OF: 2 or MISSING
AND BodyMassIndex < 32.68 or MISSING
AND AnnualFamilyIncome IS ONE OF: 15, 6, 14, 5 or MISSING
then
 Tree Node Identifier     = 73
 Number of Observations = 47
 Predicted: Prediabetes=1 = 0.00
 Predicted: Prediabetes=0 = 1.00

*---------------------------------------------------------*
 Node = 75
*---------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 1
AND MaritalStatus IS ONE OF: 5 or MISSING
AND Hypertension IS ONE OF: 3
AND HighCholesterol IS ONE OF: 2 or MISSING
AND BodyMassIndex >= 32.68
AND AnnualFamilyIncome IS ONE OF: 15, 6, 14, 5, 7, 9, 3 or MISSING
then
 Tree Node Identifier     = 75
 Number of Observations = 6
 Predicted: Prediabetes=1 = 0.00
 Predicted: Prediabetes=0 = 1.00

*---------------------------------------------------------*
 Node = 76
*---------------------------------------------------------*
if TotalnumberFamily IS ONE OF: 2, 1, 7
AND Riskfordiabetes IS ONE OF: 1
AND Hypertension IS ONE OF: 2, 1 or MISSING
AND BodyMassIndex < 26.755 or MISSING
AND AnnualFamilyIncome IS ONE OF: 14, 4, 7, 9, 10, 8, 1
then
 Tree Node Identifier    = 76
 Number of Observations = 22
 Predicted: Prediabetes=1 = 0.86
 Predicted: Prediabetes=0 = 0.14

*------------------------------------------------------------*
 Node = 77
*------------------------------------------------------------*
if TotalnumberFamily IS ONE OF: 2, 1, 7
AND Riskfordiabetes IS ONE OF: 1
AND Hypertension IS ONE OF: 2, 1 or MISSING
AND BodyMassIndex < 31.545 AND BodyMassIndex >= 26.755
AND AnnualFamilyIncome IS ONE OF: 14, 4, 7, 9, 10, 8, 1
then
 Tree Node Identifier    = 77
 Number of Observations = 11
 Predicted: Prediabetes=1 = 0.00
 Predicted: Prediabetes=0 = 1.00

*------------------------------------------------------------*
 Node = 80
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 1
AND MaritalStatus IS ONE OF: 1
AND Hypertension IS ONE OF: 3
AND HighCholesterol IS ONE OF: 2 or MISSING
AND BodyMassIndex >= 32.68
AND AnnualFamilyIncome IS ONE OF: 15, 6, 7
then
 Tree Node Identifier    = 80
 Number of Observations = 32
 Predicted: Prediabetes=1 = 0.81
 Predicted: Prediabetes=0 = 0.19

*------------------------------------------------------------*
 Node = 81
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 1
AND MaritalStatus IS ONE OF: 1
AND Hypertension IS ONE OF: 3
AND HighCholesterol IS ONE OF: 2 or MISSING
AND BodyMassIndex >= 32.68
AND AnnualFamilyIncome equals Missing
then
 Tree Node Identifier     = 81
 Number of Observations = 9
 Predicted: Prediabetes=1 = 0.00
 Predicted: Prediabetes=0 = 1.00

*------------------------------------------------------------*
 Node = 84
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 2 or MISSING
AND MaritalStatus IS ONE OF: 3, 1, 6
AND Hypertension IS ONE OF: 2, 1 or MISSING
AND HighCholesterol IS ONE OF: 2
AND Diet IS ONE OF: 2, 4, 5 or MISSING
AND BodyMassIndex >= 31 or MISSING
AND AnnualFamilyIncome IS ONE OF: 15, 6, 12, 14, 5, 9, 8, 3, 13, 2 or MISSING
then
 Tree Node Identifier     = 84
 Number of Observations = 159
 Predicted: Prediabetes=1 = 0.79
 Predicted: Prediabetes=0 = 0.21

*------------------------------------------------------------*
 Node = 85
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 2 or MISSING
AND MaritalStatus IS ONE OF: 2, 5 or MISSING
AND Hypertension IS ONE OF: 2, 1 or MISSING
AND HighCholesterol IS ONE OF: 2
AND Diet IS ONE OF: 2, 4, 5 or MISSING
AND BodyMassIndex >= 31 or MISSING
AND AnnualFamilyIncome IS ONE OF: 15, 6, 12, 14, 5, 9, 8, 3, 13, 2 or MISSING
then
 Tree Node Identifier     = 85
 Number of Observations = 19
 Predicted: Prediabetes=1 = 0.00
 Predicted: Prediabetes=0 = 1.00

*------------------------------------------------------------*
 Node = 86
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 2, 1 or MISSING
AND HighCholesterol IS ONE OF: 1 or MISSING
AND Diet IS ONE OF: 2, 4, 5 or MISSING
AND BodyMassIndex < 25.905
AND AnnualFamilyIncome IS ONE OF: 15, 14, 5, 10
AND Age >= 39.5 or MISSING
then
 Tree Node Identifier     = 86
 Number of Observations = 40
 Predicted: Prediabetes=1 = 0.70
 Predicted: Prediabetes=0 = 0.30

*------------------------------------------------------------*
 Node = 87
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 2, 1 or MISSING
AND HighCholesterol IS ONE OF: 1 or MISSING
AND Diet IS ONE OF: 2, 4, 5 or MISSING
AND BodyMassIndex < 25.905
AND AnnualFamilyIncome IS ONE OF: 6 or MISSING
AND Age >= 39.5 or MISSING
then
 Tree Node Identifier     = 87
 Number of Observations = 24
 Predicted: Prediabetes=1 = 0.00
 Predicted: Prediabetes=0 = 1.00

*------------------------------------------------------------*
 Node = 88
*------------------------------------------------------------*
if Smoked IS ONE OF: 1 or MISSING
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 1 or MISSING
AND Gender IS ONE OF: 2
AND Diet IS ONE OF: 3, 1
AND AnnualFamilyIncome IS ONE OF: 6, 12, 14, 4, 9, 8, 1
then
 Tree Node Identifier     = 88
 Number of Observations = 66
 Predicted: Prediabetes=1 = 0.47
 Predicted: Prediabetes=0 = 0.53

*------------------------------------------------------------*
 Node = 89
*------------------------------------------------------------*
if Smoked IS ONE OF: 1 or MISSING
```

```
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 1 or MISSING
AND Gender IS ONE OF: 1 or MISSING
AND Diet IS ONE OF: 3, 1
AND AnnualFamilyIncome IS ONE OF: 6, 12, 14, 4, 9, 8, 1
then
 Tree Node Identifier   = 89
 Number of Observations = 166
 Predicted: Prediabetes=1 = 0.73
 Predicted: Prediabetes=0 = 0.27

*------------------------------------------------------------*
 Node = 90
*------------------------------------------------------------*
if Smoked IS ONE OF: 1 or MISSING
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 1 or MISSING
AND Gender IS ONE OF: 1 or MISSING
AND Diet IS ONE OF: 3, 1
AND AnnualFamilyIncome IS ONE OF: 15, 5, 7, 10, 3, 2 or MISSING
then
 Tree Node Identifier   = 90
 Number of Observations = 87
 Predicted: Prediabetes=1 = 0.18
 Predicted: Prediabetes=0 = 0.82

*------------------------------------------------------------*
 Node = 91
*------------------------------------------------------------*
if Smoked IS ONE OF: 1 or MISSING
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 1 or MISSING
AND Gender IS ONE OF: 2
AND Diet IS ONE OF: 3, 1
AND AnnualFamilyIncome IS ONE OF: 15, 5, 7, 10, 3, 2 or MISSING
then
 Tree Node Identifier   = 91
 Number of Observations = 84
 Predicted: Prediabetes=1 = 0.58
 Predicted: Prediabetes=0 = 0.42

*------------------------------------------------------------*
 Node = 92
*------------------------------------------------------------*
if TotalnumberFamily IS ONE OF: 4, 2, 1, 6 or MISSING
```

```
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 3
AND Gender IS ONE OF: 2 or MISSING
AND BodyMassIndex < 24.025
AND AnnualFamilyIncome IS ONE OF: 6, 12, 14, 4, 2
AND Age < 49.5 or MISSING
then
 Tree Node Identifier   = 92
 Number of Observations = 30
 Predicted: Prediabetes=1 = 0.00
 Predicted: Prediabetes=0 = 1.00

*------------------------------------------------------------*
 Node = 93
*------------------------------------------------------------*
if TotalnumberFamily IS ONE OF: 4, 2, 1, 6 or MISSING
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 3
AND Gender IS ONE OF: 2 or MISSING
AND BodyMassIndex >= 24.025 or MISSING
AND AnnualFamilyIncome IS ONE OF: 6, 12, 14, 4, 2
AND Age < 49.5 or MISSING
then
 Tree Node Identifier   = 93
 Number of Observations = 131
 Predicted: Prediabetes=1 = 0.62
 Predicted: Prediabetes=0 = 0.38

*------------------------------------------------------------*
 Node = 98
*------------------------------------------------------------*
if Smoked IS ONE OF: 1
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 3
AND BodyMassIndex < 37.25 or MISSING
AND AnnualFamilyIncome IS ONE OF: 12, 14, 7, 10, 3
AND Age < 78 AND Age >= 49.5 or MISSING
then
 Tree Node Identifier   = 98
 Number of Observations = 82
 Predicted: Prediabetes=1 = 0.29
 Predicted: Prediabetes=0 = 0.71

*------------------------------------------------------------*
 Node = 99
```

```
*------------------------------------------------------------*
if Smoked IS ONE OF: 1
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 3
AND BodyMassIndex < 37.25 or MISSING
AND AnnualFamilyIncome IS ONE OF: 12, 14, 7, 10, 3
AND Age >= 78
then
 Tree Node Identifier   = 99
 Number of Observations = 33
 Predicted: Prediabetes=1 = 0.79
 Predicted: Prediabetes=0 = 0.21

*------------------------------------------------------------*
 Node = 100
*------------------------------------------------------------*
if TotalnumberFamily IS ONE OF: 3, 4, 5, 1 or MISSING
AND Smoked IS ONE OF: 2 or MISSING
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 3
AND HighCholesterol IS ONE OF: 2
AND BodyMassIndex < 37.25 or MISSING
AND Age >= 49.5
then
 Tree Node Identifier   = 100
 Number of Observations = 81
 Predicted: Prediabetes=1 = 0.31
 Predicted: Prediabetes=0 = 0.69

*------------------------------------------------------------*
 Node = 106
*------------------------------------------------------------*
if TotalnumberFamily IS ONE OF: 3, 4, 5, 1 or MISSING
AND Smoked IS ONE OF: 2 or MISSING
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 3
AND HighCholesterol IS ONE OF: 1 or MISSING
AND BodyMassIndex < 37.25 or MISSING
AND AnnualFamilyIncome IS ONE OF: 12, 14, 9, 10, 8, 2
AND Age >= 49.5
then
 Tree Node Identifier   = 106
 Number of Observations = 66
 Predicted: Prediabetes=1 = 0.85
 Predicted: Prediabetes=0 = 0.15
```

```
*------------------------------------------------------------*
 Node = 107
*------------------------------------------------------------*
if TotalnumberFamily IS ONE OF: 3, 4, 5, 1 or MISSING
AND Smoked IS ONE OF: 2 or MISSING
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 3
AND HighCholesterol IS ONE OF: 1 or MISSING
AND BodyMassIndex < 37.25 or MISSING
AND AnnualFamilyIncome IS ONE OF: 15, 6, 4, 3 or MISSING
AND Age >= 49.5
then
 Tree Node Identifier     = 107
 Number of Observations = 49
 Predicted: Prediabetes=1 = 0.43
 Predicted: Prediabetes=0 = 0.57
```

**2011-2012 Tree Leaf Report**

```
*------------------------------------------------------------*
 Node = 15
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 1
AND Hypertension IS ONE OF: 1, 2 or MISSING
AND HighCholesterol IS ONE OF: 1 or MISSING
then
 Tree Node Identifier     = 15
 Number of Observations = 512
 Predicted: Prediabetes=1 = 0.91
 Predicted: Prediabetes=0 = 0.09

*------------------------------------------------------------*
 Node = 19
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 2 or MISSING
AND BodyMassIndex >= 36.85
AND Age < 45.5 AND Age >= 41.5
then
 Tree Node Identifier     = 19
 Number of Observations = 64
 Predicted: Prediabetes=1 = 0.78
 Predicted: Prediabetes=0 = 0.22

*------------------------------------------------------------*
```

```
 Node = 21
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 2 or MISSING
AND BodyMassIndex < 27.45
AND Annualfamilyincome IS ONE OF: 12, 10, 13, 8, 1
AND Age >= 45.5 or MISSING
then
 Tree Node Identifier   = 21
 Number of Observations = 110
 Predicted: Prediabetes=1 = 0.00
 Predicted: Prediabetes=0 = 1.00

*------------------------------------------------------------*
 Node = 24
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 1
AND HighCholesterol IS ONE OF: 2
AND Annualfamilyincome IS ONE OF: 6, 5, 13, 9
AND Age < 43.5
then
 Tree Node Identifier   = 24
 Number of Observations = 98
 Predicted: Prediabetes=1 = 0.69
 Predicted: Prediabetes=0 = 0.31

*------------------------------------------------------------*
 Node = 25
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 1
AND HighCholesterol IS ONE OF: 2
AND Annualfamilyincome IS ONE OF: 7, 15, 2, 14, 10, 4, 8, 3, 1 or MISSING
AND Age < 43.5
then
 Tree Node Identifier   = 25
 Number of Observations = 147
 Predicted: Prediabetes=1 = 0.24
 Predicted: Prediabetes=0 = 0.76

*------------------------------------------------------------*
 Node = 26
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 1
AND HighCholesterol IS ONE OF: 2
AND Annualfamilyincome IS ONE OF: 7, 15, 12, 2, 14, 10, 3, 9
AND Age >= 43.5 or MISSING
```

```
then
  Tree Node Identifier     = 26
  Number of Observations = 236
  Predicted: Prediabetes=1 = 0.81
  Predicted: Prediabetes=0 = 0.19

*------------------------------------------------------------*
  Node = 29
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 1
AND Hypertension IS ONE OF: 3
AND HighCholesterol IS ONE OF: 1 or MISSING
AND Age >= 48.5 or MISSING
then
  Tree Node Identifier     = 29
  Number of Observations = 130
  Predicted: Prediabetes=1 = 0.87
  Predicted: Prediabetes=0 = 0.13

*------------------------------------------------------------*
  Node = 33
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 2 or MISSING
AND Gender IS ONE OF: 2
AND BodyMassIndex < 36.85 or MISSING
AND Annualfamilyincome IS ONE OF: 7, 12, 2, 14, 3, 9
AND Age < 45.5
then
  Tree Node Identifier     = 33
  Number of Observations = 280
  Predicted: Prediabetes=1 = 0.30
  Predicted: Prediabetes=0 = 0.70

*------------------------------------------------------------*
  Node = 40
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 2 or MISSING
AND HighCholesterol IS ONE OF: 2
AND BodyMassIndex >= 27.45 or MISSING
AND Annualfamilyincome IS ONE OF: 14, 10, 4, 3, 9
AND Age >= 45.5 or MISSING
then
  Tree Node Identifier     = 40
  Number of Observations = 283
  Predicted: Prediabetes=1 = 0.63
```

```
 Predicted: Prediabetes=0 = 0.37

*------------------------------------------------------------*
 Node = 42
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 2 or MISSING
AND HighCholesterol IS ONE OF: 1 or MISSING
AND BodyMassIndex >= 27.45 or MISSING
AND Annualfamilyincome IS ONE OF: 6, 7, 5, 15, 12, 14, 10, 4, 8, 1
AND Age >= 45.5 or MISSING
then
 Tree Node Identifier     = 42
 Number of Observations = 617
 Predicted: Prediabetes=1 = 0.72
 Predicted: Prediabetes=0 = 0.28

*------------------------------------------------------------*
 Node = 46
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 1
AND Maritalstatus IS ONE OF: 4, 3
AND HighCholesterol IS ONE OF: 2
AND Annualfamilyincome IS ONE OF: 6, 4, 8, 1 or MISSING
AND Age >= 43.5 or MISSING
then
 Tree Node Identifier     = 46
 Number of Observations = 36
 Predicted: Prediabetes=1 = 0.75
 Predicted: Prediabetes=0 = 0.25

*------------------------------------------------------------*
 Node = 47
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 1
AND Maritalstatus IS ONE OF: 1, 5 or MISSING
AND HighCholesterol IS ONE OF: 2
AND Annualfamilyincome IS ONE OF: 6, 4, 8, 1 or MISSING
AND Age >= 43.5 or MISSING
then
 Tree Node Identifier     = 47
 Number of Observations = 21
 Predicted: Prediabetes=1 = 0.00
 Predicted: Prediabetes=0 = 1.00

*------------------------------------------------------------*
```

```
 Node = 48
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 1
AND Hypertension IS ONE OF: 3
AND HighCholesterol IS ONE OF: 1 or MISSING
AND Annualfamilyincome IS ONE OF: 6, 15, 14, 10, 4
AND Age < 48.5
then
 Tree Node Identifier     = 48
 Number of Observations = 53
 Predicted: Prediabetes=1 = 0.79
 Predicted: Prediabetes=0 = 0.21

*------------------------------------------------------------*
 Node = 49
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 1
AND Hypertension IS ONE OF: 3
AND HighCholesterol IS ONE OF: 1 or MISSING
AND Annualfamilyincome IS ONE OF: 5 or MISSING
AND Age < 48.5
then
 Tree Node Identifier     = 49
 Number of Observations = 17
 Predicted: Prediabetes=1 = 0.00
 Predicted: Prediabetes=0 = 1.00

*------------------------------------------------------------*
 Node = 54
*------------------------------------------------------------*
if TotalnumberFamily IS ONE OF: 6, 2, 4
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND BodyMassIndex < 38.1 AND BodyMassIndex >= 36.85
AND Age < 41.5 or MISSING
then
 Tree Node Identifier     = 54
 Number of Observations = 38
 Predicted: Prediabetes=1 = 0.79
 Predicted: Prediabetes=0 = 0.21

*------------------------------------------------------------*
 Node = 55
*------------------------------------------------------------*
if TotalnumberFamily equals Missing
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND BodyMassIndex < 38.1 AND BodyMassIndex >= 36.85
AND Age < 41.5 or MISSING
then
 Tree Node Identifier    = 55
 Number of Observations = 12
 Predicted: Prediabetes=1 = 0.00
 Predicted: Prediabetes=0 = 1.00

*------------------------------------------------------------*
 Node = 57
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 2 or MISSING
AND Gender IS ONE OF: 1 or MISSING
AND Diet IS ONE OF: 4, 3, 2 or MISSING
AND BodyMassIndex < 36.85 or MISSING
AND Annualfamilyincome IS ONE OF: 7, 12, 2, 14, 3, 9
AND Age < 45.5
then
 Tree Node Identifier    = 57
 Number of Observations = 226
 Predicted: Prediabetes=1 = 0.05
 Predicted: Prediabetes=0 = 0.95

*------------------------------------------------------------*
 Node = 60
*------------------------------------------------------------*
if TotalnumberFamily IS ONE OF: 6, 5
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 3, 1 or MISSING
AND BodyMassIndex < 36.85 or MISSING
AND Annualfamilyincome IS ONE OF: 6, 5, 15, 10, 4, 13, 8, 1 or MISSING
AND Age < 45.5
then
 Tree Node Identifier    = 60
 Number of Observations = 141
 Predicted: Prediabetes=1 = 0.11
 Predicted: Prediabetes=0 = 0.89

*------------------------------------------------------------*
 Node = 61
*------------------------------------------------------------*
if TotalnumberFamily IS ONE OF: 6, 5
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 2
AND BodyMassIndex < 36.85 or MISSING
AND Annualfamilyincome IS ONE OF: 6, 5, 15, 10, 4, 13, 8, 1 or MISSING
AND Age < 45.5
then
 Tree Node Identifier     = 61
 Number of Observations = 9
 Predicted: Prediabetes=1 = 0.78
 Predicted: Prediabetes=0 = 0.22

*------------------------------------------------------------*
 Node = 62
*------------------------------------------------------------*
if TotalnumberFamily IS ONE OF: 2, 1, 3, 7, 4 or MISSING
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND BodyMassIndex < 34.85 or MISSING
AND Annualfamilyincome IS ONE OF: 6, 5, 15, 10, 4, 13, 8, 1 or MISSING
AND Age < 45.5
then
 Tree Node Identifier     = 62
 Number of Observations = 581
 Predicted: Prediabetes=1 = 0.00
 Predicted: Prediabetes=0 = 1.00

*------------------------------------------------------------*
 Node = 64
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 2 or MISSING
AND BodyMassIndex >= 38.1 or MISSING
AND Annualfamilyincome IS ONE OF: 2, 1
AND Age < 41.5 or MISSING
then
 Tree Node Identifier     = 64
 Number of Observations = 21
 Predicted: Prediabetes=1 = 0.67
 Predicted: Prediabetes=0 = 0.33

*------------------------------------------------------------*
 Node = 65
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 2 or MISSING
AND BodyMassIndex >= 38.1 or MISSING
AND Annualfamilyincome IS ONE OF: 6, 7, 5, 14, 4, 8, 3 or MISSING
AND Age < 41.5 or MISSING
then
 Tree Node Identifier     = 65
 Number of Observations = 71
```

```
  Predicted: Prediabetes=1 = 0.00
  Predicted: Prediabetes=0 = 1.00

*------------------------------------------------------------*
  Node = 67
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 2 or MISSING
AND BodyMassIndex < 27.45 AND BodyMassIndex >= 26.05
AND Annualfamilyincome IS ONE OF: 6, 7, 5, 15, 2, 14, 4, 3, 9 or MISSING
AND Age < 61.5 AND Age >= 45.5
then
  Tree Node Identifier    = 67
  Number of Observations = 74
  Predicted: Prediabetes=1 = 0.00
  Predicted: Prediabetes=0 = 1.00

*------------------------------------------------------------*
  Node = 68
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 2 or MISSING
AND Maritalstatus IS ONE OF: 4, 1, 5, 3, 6 or MISSING
AND BodyMassIndex < 27.45
AND Annualfamilyincome IS ONE OF: 6, 7, 5, 15, 2, 14, 4, 3, 9 or MISSING
AND Age >= 61.5 or MISSING
then
  Tree Node Identifier    = 68
  Number of Observations = 436
  Predicted: Prediabetes=1 = 0.53
  Predicted: Prediabetes=0 = 0.47

*------------------------------------------------------------*
  Node = 69
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 2 or MISSING
AND Maritalstatus IS ONE OF: 2
AND BodyMassIndex < 27.45
AND Annualfamilyincome IS ONE OF: 6, 7, 5, 15, 2, 14, 4, 3, 9 or MISSING
AND Age >= 61.5 or MISSING
then
  Tree Node Identifier    = 69
  Number of Observations = 69
  Predicted: Prediabetes=1 = 0.12
  Predicted: Prediabetes=0 = 0.88

*------------------------------------------------------------*
```

```
 Node = 72
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 2 or MISSING
AND HighCholesterol IS ONE OF: 1 or MISSING
AND BodyMassIndex < 37.6 AND BodyMassIndex >= 27.45 or MISSING
AND Annualfamilyincome IS ONE OF: 2, 3, 9 or MISSING
AND Age >= 45.5 or MISSING
then
 Tree Node Identifier   = 72
 Number of Observations = 91
 Predicted: Prediabetes=1 = 0.25
 Predicted: Prediabetes=0 = 0.75

*------------------------------------------------------------*
 Node = 73
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 2 or MISSING
AND HighCholesterol IS ONE OF: 1 or MISSING
AND BodyMassIndex >= 37.6
AND Annualfamilyincome IS ONE OF: 2, 3, 9 or MISSING
AND Age >= 45.5 or MISSING
then
 Tree Node Identifier   = 73
 Number of Observations = 24
 Predicted: Prediabetes=1 = 0.79
 Predicted: Prediabetes=0 = 0.21

*------------------------------------------------------------*
 Node = 78
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 2 or MISSING
AND Gender IS ONE OF: 1 or MISSING
AND Diet IS ONE OF: 5, 1
AND BodyMassIndex < 36.85 or MISSING
AND Annualfamilyincome IS ONE OF: 7, 12, 2, 14, 3, 9
AND Age < 33
then
 Tree Node Identifier   = 78
 Number of Observations = 21
 Predicted: Prediabetes=1 = 0.00
 Predicted: Prediabetes=0 = 1.00

*------------------------------------------------------------*
 Node = 79
*------------------------------------------------------------*
```

```
if Riskfordiabetes IS ONE OF: 2 or MISSING
AND Gender IS ONE OF: 1 or MISSING
AND Diet IS ONE OF: 5, 1
AND BodyMassIndex < 36.85 or MISSING
AND Annualfamilyincome IS ONE OF: 7, 12, 2, 14, 3, 9
AND Age < 45.5 AND Age >= 33 or MISSING
then
 Tree Node Identifier   = 79
 Number of Observations = 33
 Predicted: Prediabetes=1 = 0.67
 Predicted: Prediabetes=0 = 0.33

*------------------------------------------------------------*
 Node = 86
*------------------------------------------------------------*
if TotalnumberFamily IS ONE OF: 2, 1, 3, 7, 4 or MISSING
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND BodyMassIndex < 35 AND BodyMassIndex >= 34.85
AND Annualfamilyincome IS ONE OF: 6, 5, 15, 10, 4, 13, 8, 1 or MISSING
AND Age < 45.5
then
 Tree Node Identifier   = 86
 Number of Observations = 8
 Predicted: Prediabetes=1 = 0.75
 Predicted: Prediabetes=0 = 0.25

*------------------------------------------------------------*
 Node = 87
*------------------------------------------------------------*
if TotalnumberFamily IS ONE OF: 2, 1, 3, 7, 4 or MISSING
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND BodyMassIndex < 36.85 AND BodyMassIndex >= 35 or MISSING
AND Annualfamilyincome IS ONE OF: 6, 5, 15, 10, 4, 13, 8, 1 or MISSING
AND Age < 45.5
then
 Tree Node Identifier   = 87
 Number of Observations = 19
 Predicted: Prediabetes=1 = 0.00
 Predicted: Prediabetes=0 = 1.00

*------------------------------------------------------------*
 Node = 89
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 2 or MISSING
AND BodyMassIndex < 26.05 AND BodyMassIndex >= 24.4
AND Annualfamilyincome IS ONE OF: 6, 7, 5, 15, 2, 14, 4, 3, 9 or MISSING
AND Age < 61.5 AND Age >= 45.5
then
 Tree Node Identifier     = 89
 Number of Observations = 109
 Predicted: Prediabetes=1 = 0.50
 Predicted: Prediabetes=0 = 0.50

*------------------------------------------------------------*
 Node = 92
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 2 or MISSING
AND Maritalstatus IS ONE OF: 1, 3
AND Hypertension IS ONE OF: 1, 2 or MISSING
AND HighCholesterol IS ONE OF: 2
AND BodyMassIndex >= 27.45 or MISSING
AND Annualfamilyincome IS ONE OF: 6, 7, 5, 15, 12, 2, 8, 1 or MISSING
AND Age >= 45.5 or MISSING
then
 Tree Node Identifier     = 92
 Number of Observations = 146
 Predicted: Prediabetes=1 = 0.52
 Predicted: Prediabetes=0 = 0.48

*------------------------------------------------------------*
 Node = 93
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 2 or MISSING
AND Maritalstatus IS ONE OF: 4, 2, 5 or MISSING
AND Hypertension IS ONE OF: 1, 2 or MISSING
AND HighCholesterol IS ONE OF: 2
AND BodyMassIndex >= 27.45 or MISSING
AND Annualfamilyincome IS ONE OF: 6, 7, 5, 15, 12, 2, 8, 1 or MISSING
AND Age >= 45.5 or MISSING
then
 Tree Node Identifier     = 93
 Number of Observations = 32
 Predicted: Prediabetes=1 = 0.00
 Predicted: Prediabetes=0 = 1.00

*------------------------------------------------------------*
 Node = 94
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 3
AND HighCholesterol IS ONE OF: 2
AND BodyMassIndex < 49.25 AND BodyMassIndex >= 27.45 or MISSING
AND Annualfamilyincome IS ONE OF: 6, 7, 5, 15, 12, 2, 8, 1 or MISSING
AND Age >= 45.5 or MISSING
then
 Tree Node Identifier   = 94
 Number of Observations = 110
 Predicted: Prediabetes=1 = 0.00
 Predicted: Prediabetes=0 = 1.00

*------------------------------------------------------------*
 Node = 95
*------------------------------------------------------------*
if Riskfordiabetes IS ONE OF: 2 or MISSING
AND Hypertension IS ONE OF: 3
AND HighCholesterol IS ONE OF: 2
AND BodyMassIndex >= 49.25
AND Annualfamilyincome IS ONE OF: 6, 7, 5, 15, 12, 2, 8, 1 or MISSING
AND Age >= 45.5 or MISSING
then
 Tree Node Identifier   = 95
 Number of Observations = 8
 Predicted: Prediabetes=1 = 1.00
 Predicted: Prediabetes=0 = 0.00

*------------------------------------------------------------*
 Node = 98
*------------------------------------------------------------*
if USArmedForces IS ONE OF: 2 or MISSING
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND BodyMassIndex < 24.4 or MISSING
AND Annualfamilyincome IS ONE OF: 6, 7, 5, 15, 2, 14, 4, 3, 9 or MISSING
AND Age < 61.5 AND Age >= 45.5
then
 Tree Node Identifier   = 98
 Number of Observations = 138
 Predicted: Prediabetes=1 = 0.08
 Predicted: Prediabetes=0 = 0.92

*------------------------------------------------------------*
 Node = 99
*------------------------------------------------------------*
if USArmedForces IS ONE OF: 1
AND Riskfordiabetes IS ONE OF: 2 or MISSING
AND BodyMassIndex < 24.4 or MISSING
AND Annualfamilyincome IS ONE OF: 6, 7, 5, 15, 2, 14, 4, 3, 9 or MISSING
AND Age < 61.5 AND Age >= 45.5
then
 Tree Node Identifier     = 99
 Number of Observations = 25
 Predicted: Prediabetes=1 = 0.80
 Predicted: Prediabetes=0 = 0.20
```