Predictive Analysis for Classifying Type 2 Diabetes Patients Using SAS Enterprise Miner 14.1 and SAS Enterprise Guide 7.1

By:
Saurabh Sanjayrao Kokad
Oklahoma State University

ABSTRACT

In the past few years, the healthcare industry has produced massive amounts of data. Instead of making hard copies of the records, digitization of the records and computerized approach helped to produces these large data. The healthcare industry is moving from only reporting facts out of this data to discovery of the insights and predicting prospective trends by studying patterns in the past. This paper is summary of finding the pattern between diabetic patient demographics, classifying them into Primary, Secondary and Tertiary groups based on existing medical standard testing results and predicting the primary, secondary and tertiary classification of the patient with different demographical and standard health variables using theoretical classification as a target. A dataset containing 900,000 records with various 12 variables was obtained from Center for Health System Innovation (CHSI), Oklahoma State University(OSU). The predictive model used the input variables such as age, gender, race, marital status, A1C levels, glucose level, urban or rural status and the target variable category which classifies the 3 different diabetes levels by standard medical testing. A decision tree model was built using SAS Enterprise Miner and data analysis was done in SAS Enterprise Guide.

BACKGROUND

According to the World Health Organization (WHO), the chances of a 30-70-year-old person from the USA dying from one of the four non communicable disease (NCDs) – Stroke, Cancer, Respiratory disease or Diabetes – is 26%. The Global Status Report claimed that NCDs would be the cause of death for 52 million lives by 2030, globally. Around 8 million people died of NCD diseases in South-East Asia Region in 2011.

Diabetic Mellitus (DM) is one of the NCDs that is major health hazard in developing countries such as India. There are 2 major types of this disease. DM Type 1, which is genetic it results from the body's failure to produce insulin and requires the person to inject insulin. DM Type 2, which is not genetic, it results from insulin resistance by the human body, a condition in which cells fails to use insulin properly, sometimes combined with an absolute insulin deficiency. In 2011, around 41 million people aged 20-79 years had diabetes in the USA and this number is expected to increase to 87 million by 2030.

INTRODUCTION

The objective of this study is to research the disease Type 2 Diabetes Mellitus and to discover which demographics and health variables are contributing more towards Type 2 Diabetes for the population in the region of Oklahoma. Type 2 Diabetes Mellitus (DM) is a not genetic but a chronic disease in which there are high levels of sugar (glucose) in the blood. It is the most common form of diabetes. There are many risk factors and reasons related to this disease, the most dominant are, obesity (major risk factor), high cholesterol and blood pressure, sedentary lifestyle, unhealthy eating behaviours, and increased age. The lifelong difficulties of encountering type 2 diabetes may include (not limited to): high blood pressure (which causes your risk for heart attack and heart stroke), oral health complications, eye problems, loss of hearing ability, foot difficulties (nerve damage), skin infections, mental health issues and apparently early death.

The test called A1C, is a test that provides information about a person's average level of blood glucose (blood sugar) over the past 3 months. The A1C level below 5.7perecent is considered as normal, between 5.7 and 6.4 signals prediabetes. Type 2 Diabetes is officially diagnosed when ha two repeat A1C lab values are above 6.5 percent in at least recent 2 A1C tests.

This study examined the quantitative data provided by Centre for Health System Innovations (CHSI), type 2 diabetes is a very severe issue affecting the residents of Oklahoma. By understanding the risk factors and social determinants related to those with type 2 diabetes in CHSI Data, researchers and other health professionals can hope to this study improve the overall health of those with the disease, and slow or prevent others from progressing through the disease stages.

DATA DICTIONARY

Variable	Data Type	Description
Patient_ID	Nominal	This variable has the unique patient IDs
Encounter_ID	Nominal	This variable explains encounters of the diabetes.
Age_in_years	Interval	This variable explains the age of the patient.
Race	Nominal	This variable has patient's race.
Gender	Nominal	This variable has the gender of patient
Marital_Status	Nominal	This variable is categorical and explains patient's marital status
Urban_Rural_Status	Nominal	This categorical variable explains location status of patient
A1C_Numeric_Result	Interval	This continuous variable explains the A1C test results
Glucose_Numeric_result	Interval	This continuous variable explains the glucose test results
Category	Nominal (Target)	The target variable is categorical with different categories of diabetes Primary: No diabetes diagnosis (pre diabetic level) Secondary: New diagnosis of diabetes with no complications Tertiary: Diabetes with complications

DATA PREPERATION

Centre for Health System Innovations (CHSI), Oklahoma state University's department, transforming healthcare through creativity, innovation and entrepreneurship. This department focus specifically on rural health innovation, patient care innovation and data analytics.

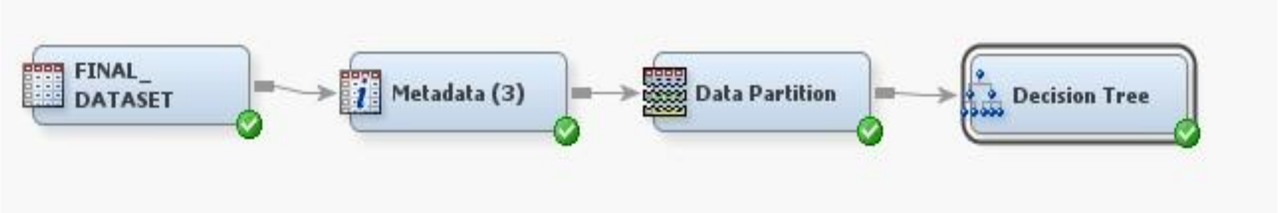
For this study, Healthcare data on Type-2 diabetes from CHSI have been used. Data was divided into three segments as Type-2 diabetes data, glucose test data, A1C level data. Type 2 diabetes data had patient details with their diagnosis, encounter and visits. A1C level data had A1C level Percentage details. Glucose test data had the glucose test details of the patients.

The Glucose data and A1C data was joined with patient ID to get the consolidated data with all the information on Glucose and A1C level of patients. And this data was joined with type 2 diabetes data to import patient demographics details in the consolidated data.

A1C Test	- a1c_diagnosis_comp
	- a1c_encounter_comp
	- a1c_lab_comp
	- a1c_medication_comp
Glucose Fasting	- glucose_diagnosis_comp
	- glucose_encounter_comp
	- glucose_lab_comp
	-glucose_medication_comp
Type 2 Diabetes	- diabetes_diagnosis_comp
	-diabetes_encounter_comp
	- diabetes _lab_comp
	- diabetes_medication_com

METHODOLOGY

This study uses the modelling technique SEMMA (Sample, Explore, Modify, Model and Assess). First, the data was sampled and analyzed using SAS Enterprise Guide. Then data was portioned into two stratified samples (Training 30%, Validation 70%) within SAS Enterprise Miner.

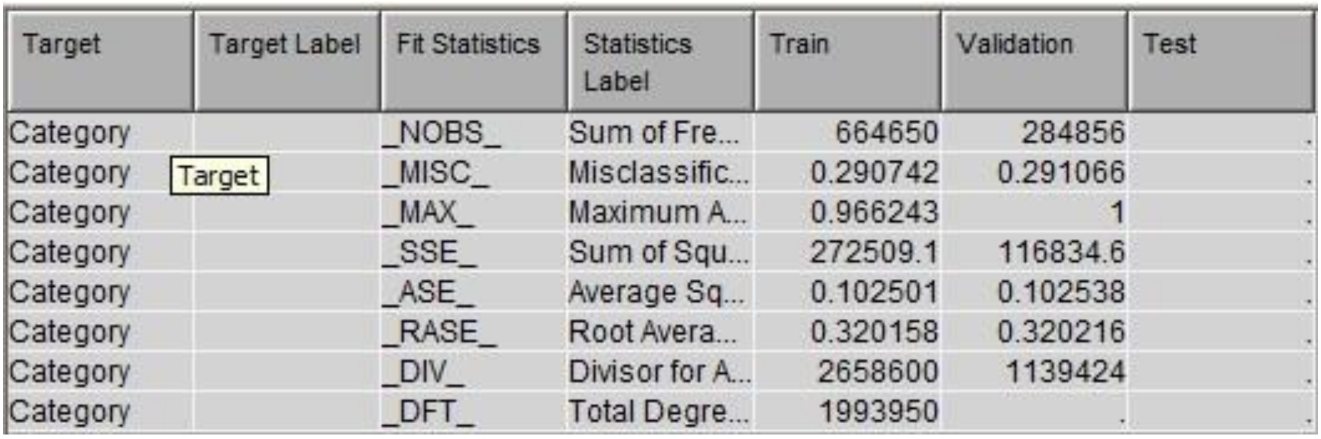


This approach uses training data to build the model and validation data is used to test the accuracy of the model. This approach gives the honest assessment of the model.

For prediction analysis, Decision tree approach is used to predict the category of the patient whether patient is in non-diabetic phase, pre-diabetic phase or diabetic phase. For this analysis various nominal and interval variables were crunched.

MODEL ASSESMENT

Decision Tree model was built as explained above. This model was built to predict which demographics and health factors leads to which one of the diabetes category. The selection criteria were misclassification rate which can be seen in below table is low.

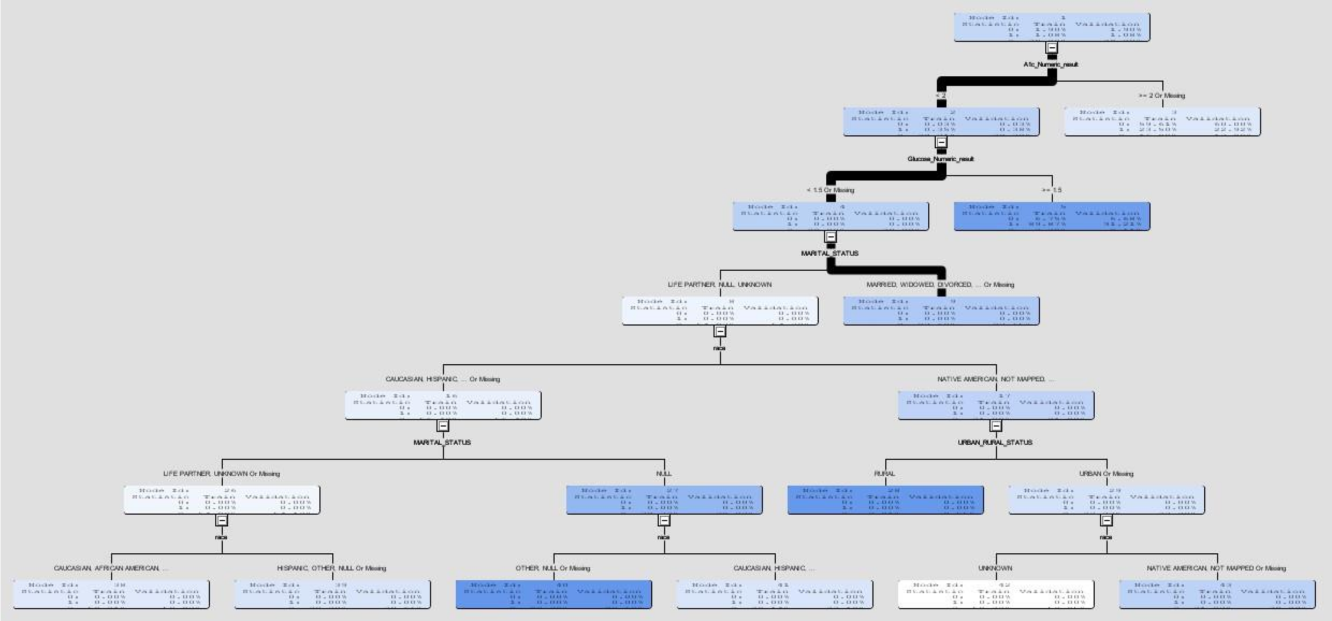


Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Category		_NOBS_	Sum of Fre...	664650	284856	.
Category	Target	_MISC_	Misclassific...	0.290742	0.291066	.
Category		_MAX_	Maximum A...	0.966243	1	.
Category		_SSE_	Sum of Squ...	272509.1	116834.6	.
Category		_ASE_	Average Sq...	0.102501	0.102538	.
Category		_RASE_	Root Avera...	0.320158	0.320216	.
Category		_DIV_	Divisor for A...	2658600	1139424	.
Category		_DFT_	Total Degre...	1993950	.	.

This conveys that, this model predicts the category 71% accurately.

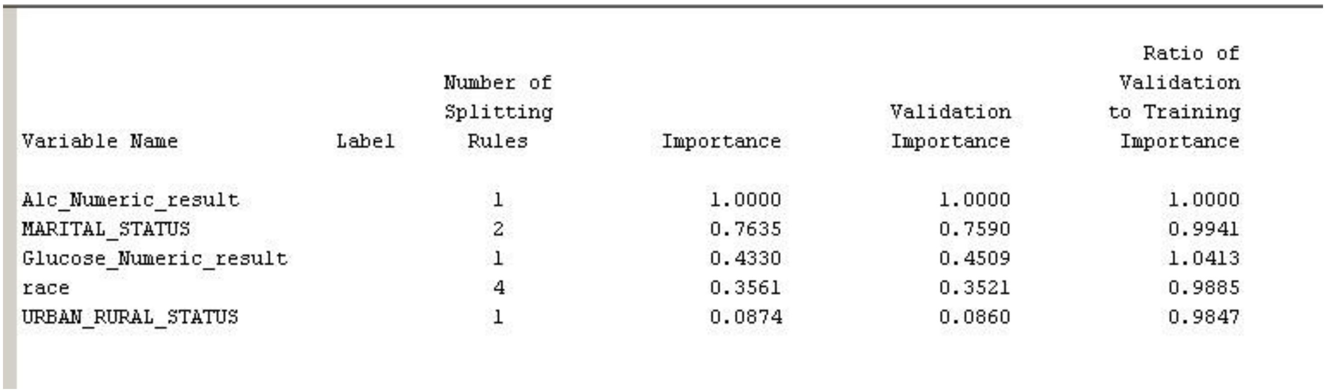
RESULTS

The following decision tree model results shows that numeric_results of AIC is the primary variable that helps to predict the category.


A1c_Numeric_result
< 2
>= 2 Or Missing
Glucose_Numeric_result
< 1.5 Or Missing
>= 1.5
MARITAL_STATUS
LIFE PARTNER, NULL, UNKNOWN
MARRIED, WIDOWED, DIVORCED, ... Or Missing
race
CAUCASIAN, HISPANIC, ... Or Missing
NATIVE AMERICAN, NOT MAPPED, ...
MARITAL_STATUS
URBAN_RURAL_STATUS
LIFE PARTNER, UNKNOWN Or Missing
NULL
RURAL
URBAN Or Missing
race
race
race
CAUCASIAN, AFRICAN AMERICAN, ...
HISPANIC, OTHER, NULL Or Missing
OTHER, NULL Or Missing
CAUCASIAN, HISPANIC, ...
UNKNOWN
NATIVE AMERICAN, NOT MAPPED Or Missing

And then decision tree shows the following branches of the model and which variables helps to predict the category of the diabetes patient. If the patient is married, widowed or divorced (the levels of marital status out of 14 levels) and the patient is native American and the glucose numeric result is less than 1.5 then that patient has the highest probability of diagnosing diabetes.

The following table shows the variable importance for this model. The results show that to predict the category of the diabetic patient, A1C_Numeric_result is most important variable followed by marital status, Glucose_numeric_result and race of the patient.



```
                                                                                              Ratio of
                                   Number of                                                Validation
                                   Splitting                             Validation        to Training
Variable Name             Label      Rules            Importance         Importance         Importance

Alc_Numeric_result                     1                  1.0000             1.0000             1.0000
MARITAL_STATUS                         2                  0.7635             0.7590             0.9941
Glucose_Numeric_result                 1                  0.4330             0.4509             1.0413
race                                   4                  0.3561             0.3521             0.9885
URBAN_RURAL_STATUS                     1                  0.0874             0.0860             0.9847
```

The ratio of the validating to training importance for A1C_Numeric_result is 1.00 and that for marital status is 0.9941 which conveys that all the variables have almost equal importance in the training data and the validation data as well.

CONCLUSION

The wide-ranging analysis on structured and unstructured data has provided some deep insights to study the patterns in the patient’s demographics and health related variables. Age, A1C level, Marital Status are the important predictors to predict the classification of the certain patient. Deeper insights with age, race and regional information was understood using SAS enterprise guide. A1C level is found to be significantly important to predict the classification of the patient.

REFERENCES

1) Center for Health System Innovation, Okalahoma State University
2) Predictive Modelling with SAS Enterprise Miner by Kattamuri Sarma
3) World Health Organization (WHO)

CONTACT INFORMATION

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

Saurabh Sanjayrao Kokad
Oklahoma State University
Stillwater, OK, 74074
Saurabh.kokad@okstate.edu