Prediction on Diabetes Using Data mining Approach
Pardha Repalli, Oklahoma State University

Abstract
The main purpose of this paper is to predict how likely the people with different age groups are being affected by diabetes based on their life style activities and to find out factors responsible for the individual to be diabetic. Hence it is interesting to implement statistical techniques in medical field to understand which age group of people are being affecting by diabetes.

Introduction
Basic understanding on growth and factors affecting diabetes from external sources is required before building predictive models. Based on 2011 National Diabetes Fact Sheet 8.3% (25 million) of U.S population has diabetes. Diabetes is seventh leading cause of death according to U.S. death certificates and also, it is the major cause for heart stroke, kidney failure, nontraumatic lower-limb amputations and blindness [1]. People aged greater than 65 years are majorly affected by diabetes [2]. Our idea is to predict the diabetic cases and to find the factors responsible for diabetes using data mining approach.

Some of the interesting facts observed from the statistics given by the Centers for Disease Control are

- 26.9% of the population affected by diabetes are people whose age is greater than 65.
- 11.8% of all men aged 20 years or older are affected by diabetes.
- 10.8% of all women aged 20 years or older are affected by diabetes.

Data Understanding
The dataset used for analysis and modeling has 50784 records with 37 variables. In order to understand the behavior of all variables, a brief analysis was conducted to know the distribution and central tendency. A new variable age_new is computed as nominal variable, dividing in to three group’s young age, middle age and old age. Our target variable diabetes_diag_binary is a binary variable.

Some of the interesting facts observed after analyzing data are

- The number of diabetic people in dataset is around 5% and the remaining people are non-diabetic.
- The interesting fact observed in the analysis is that majority of the people affected by diabetes are people whose age is above 45 years which is 4.2% out of 5.1%.
• 34% of the population whose age is below 20 years is not affected by diabetes.
• 33.9% of the population whose age is above 20 and below 45 years is not affected by diabetes.
• 26.8% of the population whose age is above 45 years is not diabetic.

The variables considered for lifestyle activities are Last Dental checkups, Last checkup, Last Cholest_check, Last PSA test, Last PAP test, Last Breast Exam, Last Mammogram and Wears Seat Belts.

The roles of the variables are decided after analyzing the each and every variable present in the dataset. These can be clearly seen in Appendix A.

In general predictive algorithms work well when the data is normally distributed or symmetric but in real world mostly we get undistributed data. In order to improve the efficiency of predictive algorithms transformation of the variables which are not normally distributed is required. Interval variables are found to be right skewed and max normal transformation is applied for normal distribution. After the transformation with suitable formula, the transformed variable is found to be distributed normally with appropriate skewness and kurtosis.
The data is partitioned into two halves i.e., 50% of the data is training data and other 50% is validation data. The training data and validation data are almost equally divided as shown below. Various models are built on training data which will be discussed and those models are validated on validation data. On validation it is easy to find out how well the model performs on the new dataset. Hence validation is important while building models.

**Fig2: Data Partition**

Variable selection node is used to find out the most important input variables to predict the target variable. The important input variables to predict binary target variable are `High_Blood_Pressure`, `Cholest_Last_chck`, `Heart_disease`, `Los_all_teeth`, `Last_flushot`, `Years_Educ` and other group variables as shown in the figure.

**Fig3: Variable Selection**
Methodology

Cross Industry Standard Process for Data Mining methodology is used for building models. In this methodology most of the time is allotted on business understanding and data understanding and the remaining time is allotted on building models, assessment and deployment. Some of the models built are Decision trees and Regression models to predict the binary target variable. After the data manipulation and transformation of certain variables decision tree and regression models are built. Multiple models built varying properties are backward regression, forward regression, stepwise regression models, decision tree with entropy, decision trees are also build varying decision node properties based on misclassification rate as assessment measure. After running various models, model comparison node is used to select the best model for predicting binary target variable. The best model is selected based on the valid misclassification rate.

Results

Various decision models are built changing decision node properties varying leaf size, maximum Depth and maximum branch. Decision tree is built based on the misclassification rate as assessment measure. The best misclassification rate for the selected model is 0.50 for the validation data.

Important input variables selected by the decision tree are high_Blood_Pressure, Cholest_last_Check, Adult_Bmi, Last_Flu_shot, heart_attack_diagonise and other variables. The average square error for the selected model is 0.043 which is low error. Based on the English rules obtained from the results of decision tree model, the input variables selected are shown in Appendix B. These are factors responsible for an individual to be diabetic. The most important variable that has major effect on diabetes is high blood pressure. Also cholest_last_Check, Adult_Bmi, Last_Flu_shot and heart_attack_diagonise have significant impact on effecting diabetes. The people with whose high_blood_pressure_diagonise value is \(2, -1\) then the probability of not affected by diabetes is 98% and in the other case the probability to be affected by diabetes is 2\%.
Conclusion

Overall, the selected model for prediction of binary target variable is decision tree with low misclassification rate. It is evident that people with age above 45 years are mostly affected by diabetes. Hence those suggested visiting clinic for checkup, dental checkup and cholesterol checkup frequently to control the diabetes. Young and middle age people are also suggested visiting clinic to check whether they have diabetes. The factors responsible for the individual to be diabetic are high blood pressure, last_cholesterol_check, Adult_BMI, Last_flu-shot and heart_attack. Hence it can be said that as the age of the person increases the probability of getting affected by diabetes also increases.
References

2. American Diabetes Association

Contact Information

Pardha Repalli, Oklahoma State University, Email Id: pardha.s.repalli@okstate.edu

Pardha Repalli is a Master’s student in Management Information System at Oklahoma State University. He is a SAS Certified Base Programmer for SAS 9 and SAS Certified Predictive Modeler for Enterprise Miner 6.1
Appendix

A.

<table>
<thead>
<tr>
<th>Name</th>
<th>Hidden</th>
<th>Role</th>
<th>New Role</th>
<th>Level</th>
<th>New Level</th>
<th>New Order</th>
<th>New Report</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADULT_BMI</td>
<td>N</td>
<td>Input</td>
<td>Input</td>
<td></td>
<td>Interval</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>ANGINA_DIASTON</td>
<td>Default</td>
<td>Input</td>
<td>Input</td>
<td>Interval</td>
<td>Nominal</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>ASTHMA_DIASTON</td>
<td>Default</td>
<td>Input</td>
<td>Input</td>
<td>Interval</td>
<td>Nominal</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>BLD_STOOL</td>
<td>N</td>
<td>Input</td>
<td>Interval</td>
<td>Nominal</td>
<td>Default</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>CENBISUS_REN</td>
<td>Default</td>
<td>Input</td>
<td>Input</td>
<td>Interval</td>
<td>Nominal</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>CHILD_BMI</td>
<td>N</td>
<td>Input</td>
<td>Input</td>
<td>Interval</td>
<td>Nominal</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>CHOLEST_LBN</td>
<td>Default</td>
<td>Input</td>
<td>Input</td>
<td>Interval</td>
<td>Ordinal</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>CURRENTLY_INF</td>
<td>Default</td>
<td>Input</td>
<td>Input</td>
<td>Interval</td>
<td>Nominal</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>DENTAL_CHEK</td>
<td>Default</td>
<td>Input</td>
<td>Input</td>
<td>Interval</td>
<td>Ordinal</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>DIABETES_DIASTON</td>
<td>Default</td>
<td>Input</td>
<td>Target</td>
<td>Interval</td>
<td>Binary</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>EMPHYSEMA</td>
<td>N</td>
<td>Input</td>
<td>Input</td>
<td>Interval</td>
<td>Nominal</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>HEART_ATTAIN</td>
<td>Default</td>
<td>Input</td>
<td>Input</td>
<td>Interval</td>
<td>Nominal</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>HEART_DISEASE</td>
<td>Default</td>
<td>Input</td>
<td>Input</td>
<td>Interval</td>
<td>Nominal</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>HIGHEST_DEN</td>
<td>Default</td>
<td>Input</td>
<td>Input</td>
<td>Interval</td>
<td>Ordinal</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>HIGH_BLOODPRESS</td>
<td>Default</td>
<td>Input</td>
<td>Input</td>
<td>Interval</td>
<td>Nominal</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>ID_OAP</td>
<td>N</td>
<td>Input</td>
<td>Input</td>
<td>Interval</td>
<td>Nominal</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>JOINT_PAIN</td>
<td>N</td>
<td>Input</td>
<td>Input</td>
<td>Interval</td>
<td>Nominal</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>LAST_BREATH</td>
<td>N</td>
<td>Input</td>
<td>Input</td>
<td>Interval</td>
<td>Ordinal</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>LAST_CHECK</td>
<td>Default</td>
<td>Input</td>
<td>Input</td>
<td>Interval</td>
<td>Ordinal</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>LAST_FLUSH</td>
<td>N</td>
<td>Input</td>
<td>Input</td>
<td>Interval</td>
<td>Ordinal</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>LAST_MAMMON</td>
<td>Default</td>
<td>Input</td>
<td>Input</td>
<td>Interval</td>
<td>Ordinal</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>LAST_PAP_MON</td>
<td>Default</td>
<td>Input</td>
<td>Input</td>
<td>Interval</td>
<td>Ordinal</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>LAST_PAP_SAN</td>
<td>N</td>
<td>Input</td>
<td>Input</td>
<td>Interval</td>
<td>Ordinal</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>LIGT_STRA</td>
<td>N</td>
<td>Input</td>
<td>Input</td>
<td>Interval</td>
<td>Ordinal</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>MARRIAGE_STRA</td>
<td>Default</td>
<td>Input</td>
<td>Input</td>
<td>Interval</td>
<td>Ordinal</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>NUMB_VISITSN</td>
<td>Default</td>
<td>Input</td>
<td>Input</td>
<td>Interval</td>
<td>Ordinal</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>OTHER_HIS</td>
<td>N</td>
<td>Input</td>
<td>Input</td>
<td>Interval</td>
<td>Ordinal</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>PERSON_LIN</td>
<td>Default</td>
<td>Input</td>
<td>Input</td>
<td>Interval</td>
<td>Ordinal</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>SERVED_HOT</td>
<td>N</td>
<td>Input</td>
<td>Input</td>
<td>Interval</td>
<td>Ordinal</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>SEX</td>
<td>N</td>
<td>Input</td>
<td>Input</td>
<td>Interval</td>
<td>Ordinal</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>SIGMODIGEON</td>
<td>Default</td>
<td>Input</td>
<td>Input</td>
<td>Interval</td>
<td>Ordinal</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>STROKE_DIASTON</td>
<td>Default</td>
<td>Input</td>
<td>Input</td>
<td>Interval</td>
<td>Ordinal</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>WEARS_EYE</td>
<td>N</td>
<td>Input</td>
<td>Input</td>
<td>Interval</td>
<td>Ordinal</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>WEAR_HEAR</td>
<td>N</td>
<td>Input</td>
<td>Input</td>
<td>Interval</td>
<td>Ordinal</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>WEAR_BEAT</td>
<td>N</td>
<td>Input</td>
<td>Input</td>
<td>Interval</td>
<td>Ordinal</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>YEAR_EDUC</td>
<td>N</td>
<td>Input</td>
<td>Input</td>
<td>Interval</td>
<td>Ordinal</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>ave new</td>
<td>N</td>
<td>Input</td>
<td>Input</td>
<td>Interval</td>
<td>Ordinal</td>
<td>Default</td>
<td>Default</td>
</tr>
</tbody>
</table>

B.

Node = 3

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

if HIGH BLOOD PRESSURE DIAGNOSIS IS ONE OF: 2, -1 or MISSING
then

Tree Node Identifier = 3
Number of Observations = 21285
Predicted: DIABETES_DIAG_BINARY=0 = 0.98
Predicted: DIABETES_DIAG_BINARY=1 = 0.02

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

Node = 5

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

if HOW LNG CHOLEST LST CHK >= 2
AND HIGH BLOOD PRESSURE DIAGNOSIS IS ONE OF: 1
then

Tree Node Identifier = 5
Number of Observations = 784
Predicted: DIABETES_DIAG_BINARY=0 = 0.92
Predicted: DIABETES_DIAG_BINARY=1 = 0.08

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

Node = 8
*-------------------------------------------------------------*
if Transformed: ADULT BMI < 3.54065 or MISSING
AND HOW LNG CHOLEST LST CHK <= 1 or MISSING
AND HIGH BLOOD PRESSURE DIAGNOSIS IS ONE OF: 1
then
Tree Node Identifier = 8
Number of Observations = 2480
Predicted: DIABETES_DIAG_BINARY=0 = 0.82
Predicted: DIABETES_DIAG_BINARY=1 = 0.18

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

Node = 36
*-------------------------------------------------------------*
if Transformed: ADULT BMI >= 3.54065
AND HOW LNG LAST FLU SHT <= 2 or MISSING
AND HOW LNG CHOLEST LST CHK <= 1 or MISSING
AND HIGH BLOOD PRESSURE DIAGNOSIS IS ONE OF: 1
AND HEART ATTACK DIAGNOSIS IS ONE OF: 1
then
Tree Node Identifier = 36
Number of Observations = 59
Predicted: DIABETES_DIAG_BINARY=0 = 0.31
Predicted: DIABETES_DIAG_BINARY=1 = 0.69

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

Node = 37
*-------------------------------------------------------------*
if Transformed: ADULT BMI >= 3.54065
AND HOW LNG LAST FLU SHT <= 2 or MISSING
AND HOW LNG CHOLEST LST CHK <= 1 or MISSING
AND HIGH BLOOD PRESSURE DIAGNOSIS IS ONE OF: 1
AND HEART ATTACK DIAGNOSIS IS ONE OF: 2 or MISSING
then
Tree Node Identifier = 37
Number of Observations = 429
Predicted: DIABETES_DIAG_BINARY=0 = 0.60
Predicted: DIABETES_DIAG_BINARY=1 = 0.40

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

Node = 39

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

if YEARS OF EDUCATION >= 12 or MISSING
AND Transformed: ADULT BMI >= 3.54065
AND HOW LNG LAST FLU SHT >= 3
AND HOW LNG CHOLEST LST CHK <= 1 or MISSING
AND HIGH BLOOD PRESSURE DIAGNOSIS IS ONE OF: 1
then
    Tree Node Identifier = 39
    Number of Observations = 234
    Predicted: DIABETES_DIAG_BINARY=0 = 0.82
    Predicted: DIABETES_DIAG_BINARY=1 = 0.18

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

Node = 74

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

if YEARS OF EDUCATION <= 11
AND Transformed: NUM OFFICE-BASED PROVIDER VISITS < 0.10029 or MISSING
AND Transformed: ADULT BMI >= 3.54065
AND HOW LNG LAST FLU SHT >= 3
AND HOW LNG CHOLEST LST CHK <= 1 or MISSING
AND HIGH BLOOD PRESSURE DIAGNOSIS IS ONE OF: 1
then
    Tree Node Identifier = 74
    Number of Observations = 113
    Predicted: DIABETES_DIAG_BINARY=0 = 0.63
    Predicted: DIABETES_DIAG_BINARY=1 = 0.37

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

Node = 75

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

if YEARS OF EDUCATION <= 11
AND Transformed: NUM OFFICE-BASED PROVIDER VISITS >= 0.10029
AND Transformed: ADULT BMI >= 3.54065
AND HOW LNG LAST FLU SHT >= 3
AND HOW LNG CHOLEST LST CHCK <= 1 or MISSING
AND HIGH BLOOD PRESSURE DIAGNOSIS IS ONE OF: 1
then
Tree Node Identifier  = 75
Number of Observations = 9
Predicted: DIABETES_DIAG_BINARY=0  = 0.22
Predicted: DIABETES_DIAG_BINARY=1  = 0.78

Acknowledgement:

The dataset used for this study is provided by SAS for Data Mining Shootout 2010 and I thank SAS for allowing me to use the dataset.