THE PHILIPPINE-AMERICAN Academy of Science & Engineering

Prediction of Diabetes on Women using Decision Tree Algorithm

Jefferson Gante Sanidad[1], Argel Bandala[2], Bernard Gante Sanidad[3], Sheryl Sanidad Marfil[4]
[1]MS in Computer Science, Maharishi University of Management, Fairfield Iowa 52557 USA
[2]Assistant Professor, De La Salle University, Manila Philippines
[3]Professor, University of Santo Thomas, Manila, Philippines
[4]Teacher, Manuel L. Quezon Elementary School, Quezon City Philippines

ABSTRACT

Diabetes is a leading cause of death in most developed countries(2). It contributes to nearly 4 million deaths per year worldwide. This study used 8 measures prescribed by the World Health Organizations (WHO) as input variables to predict the onset of Diabetes mellitus. Data Mining techniques applied to this study are Decision Tree. The researchers used Weka to model the Prediction of Diabetes on Women using Decision Tree Algorithm.

KEYWORDS: body mass index, diabetes,, diastolic blood pressure, decision tree, plasma glucose, prediction

1 INTRODUCTION

Diabetes mellitus (or diabetes) is a chronic, lifelong condition that affects your body's ability to use the energy found in food. Normally, your body breaks down the sugars and carbohydrates you eat into a special sugar called glucose.

Glucose fuels the cells in your body. But the cells need insulin, a hormone, in your bloodstream in order to take in the glucose and use it for energy.

With diabetes mellitus, either your body doesn't make enough insulin, it can't use the insulin it does produce, or a combination of both.

Since the cells can't take in the glucose, it builds up in your blood. High levels of blood glucose can damage the tiny blood vessels in your kidneys, heart, eyes, or nervous system. That's why diabetes especially if left untreated can eventually cause heart disease, stroke, kidney disease, blindness, and nerve damage to nerves in the feet [1].

Diabetes is a leading cause of death in most developed countries [2]. It contributes to nearly 4 million deaths per year worldwide. There are three major types of diabetes:

Type 1 Diabetes

Type 1 diabetes *(previously known as insulin-dependent, juvenile or childhood-onset)* is characterized by deficient insulin production and requires daily administration of insulin. The cause of type 1 diabetes is not known and it is not preventable with current knowledge.

Symptoms include excessive excretion of urine (polyuria), thirst (polydipsia), constant hunger, weight loss, vision changes and fatigue. These symptoms may occur suddenly. Medical risks are associated with type 1 diabetes includes damage to the tiny blood vessels in your eyes *(called diabetic retinopathy)*, nerves *(diabetic neuropathy)*, and kidneys *(diabetic nephropathy)*. Even more serious is the increased risk of heart disease and stroke.

Type 2 Diabetes

Type 2 *diabetes (formerly called non-insulin-dependent or adult-onset)* results from the body's ineffective use of insulin. Type 2 diabetes comprises 90% of people with diabetes around the world [2],

but with the epidemic of obese and overweight kids, more teenagers are now developing type 2 diabetes and is largely the result of excess body weight and physical inactivity.

Symptoms may be similar to those of Type 1 diabetes, but are often less marked. As a result, the disease may be diagnosed several years after onset, once complications have already arisen.

Type2 diabetes is responsible for 90-95% of diabetic cases[2].Diabetes is a major risk factor for cardiovascular disease worldwide[3], approximately 50% of people with diabetes die of cardiovascular event. About 50- 80% of people with diabetes are unaware of their condition. [2]

Gestational Diabetes

Pregnancy, to some degree, leads to insulin resistance. A pregnant women is often diagnosed with Gestational Diabetes in middle or late pregnancy. Because high blood sugar levels in a mother are circulated through the placenta to the baby, gestational diabetes must be controlled to protect the baby's growth and development [1].

According to the National Institutes of Health, the reported rate of gestational diabetes is between 2% to 10% of pregnancies. Gestational diabetes usually resolves itself after pregnancy. Having gestational diabetes does, however, put mothers at risk for developing type 2 diabetes later in life. Up to 10% of women with gestational diabetes develop type 2 diabetes. It can occur anywhere from a few weeks after delivery to months or years later.

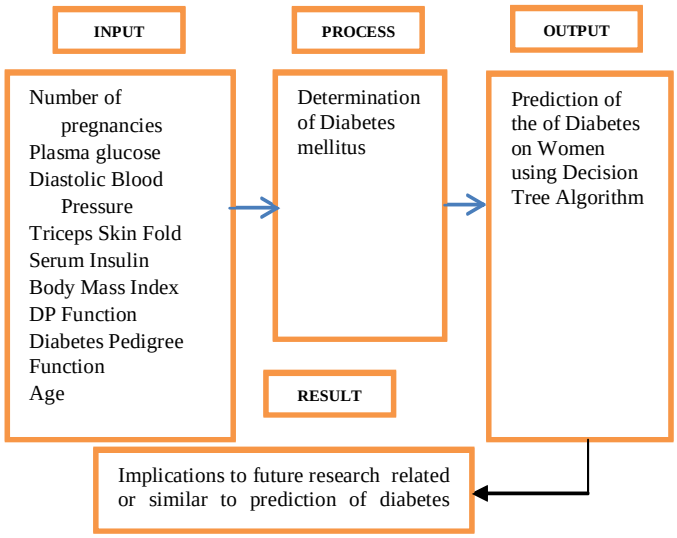
With gestational diabetes, risks to the unborn baby are even greater than risks to the mother. Risks to the baby include abnormal weight gain before birth, breathing problems at birth, and higher obesity and diabetes risk later in life. Risks to the mother include needing a cesarean section due to an overly large baby, as well as damage to heart, kidney, nerves, and eye. Symptoms of gestational diabetes are similar to Type 2 diabetes. Gestational diabetes is most often diagnosed through prenatal screening, rather than reported symptoms [3].

II. THEORETICAL FRAMEWORK OF THE STUDY

The first block represents the independent variable or the input[8]. It serves as stimulus variable used to determine the relationship to an observed phenomenon. Here, all the data are gathered in preparation for the intervening variable. The independent variable in the study is the analysis and evaluation of the existing condition of the respondents. The second block represents the intervening variable where all the data are analyzed, evaluated and processed. The third block represents the dependent variable to determine if the respondent has a diabetes or not, which is the study itself **"Prediction of Diabetes on Women using Decision Tree Algorithm".** Overall outcome of the study is the implications a future researchers can get out of the study in order to predict as to whether a woman has a diabetes or not . The study aimed to determine the **"Prediction of Diabetes in Women using Decision Tree Algorithm".**

Specifically, it sought to answer the following questions:

1. How many pregnancy does the respondents had?
2. What is the Plasma Glucose test result of the respondents?
3. What is the Diastolic Blood Pressure test result of the respondents?
4. What is the Triceps Skin Fold Thickness of the respondents?
5. What is the serum insulin test result of the respondents?
6. What is Body Mass Index (BMI) of the respondents?
7. What is the Diabetes pedigree result of the respondents?
8. What is the age of the respondents?



III. STUDY POPULATION

The population for this study was the Pima Indian population near Phoenix, Arizona USA. That population has been under continuous study since 1965 by the National Institute of Diabetes and Digestive and Kidney Diseases because of its high incidence rate of diabetes.[4,5,6] Each community resident over 5 years of age was asked to undergo a standardized examination every two years, which included an oral glucose tolerance test. Diabetes was diagnosed according to World Health Organization Criteria[10]; that is, if the 2 hour post-load plasma glucose was at least 200 mg/dl (11.1 mmol/l) at any survey examination or if the Indian Health Service Hospital serving the community found a glucose concentration of at least 200 mg/dl during the course of routine medical care [7]. In addition to being a familiar database to the investigators, this data set provided a well validated data resource in which to explore prediction of the date of onset of diabetes.

SELECTION OF THE VARIABLES

Eight variables were chosen to form the basis for forecasting the onset of diabetes within five years in Pima Indian women. Those variables were chosen because they have been found to be significant risk factors for diabetes among Pima's or other populations. The variables are:

1. Pregnancies: Number of pregnancies
2. PG Concentration: Plasma glucose at 2 hours in an oral glucose tolerance test
3. Diastolic BP: Diastolic Blood Pressure (mm Hg)
4. Tri Fold Thick: Triceps Skin Fold Thickness (mm)
5. Serum Ins: 2-Hour Serum Insulin (mu U/ml)
6. Body Mass Index: (weight in kg/ (height in m) ^{2})
7. DP Function: Diabetes Pedigree Function
8. Age: Age (years)

The Diabetes Pedigree Function (DPF) was developed by Smith, et al. [17] to provide a synthesis of the diabetes mellitus history in relatives and the generic relationship of those relatives to the

subject. The DPF uses information from parents, grandparents, siblings, aunts and uncles, and first degree cousins. It provides a measure of the expected genetic influence of affected and unaffected relatives on the subject's eventual diabetes risk. *See [17] for details.*

CASE SELECTION

Diabetes was defined as a plasma glucose concentration greater the 200 mg/dl two hours following the ingestion of 75 gm. of a carbohydrate solution. Cases were drawn from the pool of examinations which met the following criteria:

i. The subject was female.
ii. The subject was ≥21 year of age at the time of the index examination. An index examination refers to the study that was chosen for use in this model. It does not necessarily correspond to the chronologically first examination for this subject.
iii. Only one examination was selected per subject. That examination was one that revealed a non-diabetic GTT and met one of the following two criteria:
    a. Diabetes was diagnosed within five years of the examination, OR
    b. A GTT performed five or more years later failed to reveal diabetes mellitus.
iv. If diabetes occurred within one year of an examination, that examination was excluded from the study to remove from the forecasting model those cases that were potentially easier to forecast.

In 75% of the excluded examinations, DM was diagnosed within six months. Using these criteria, 768 examinations were selected. From those, 576 were selected randomly to be used in the training or learning set and the remaining 192 cases became the forecasting set.

Our hypothesis was that Decision could learn to forecast whether a given individual would develop diabetes mellitus within five years given the value of the eight input variables.

IV. DECISION TREE

Decision trees [10] [11] are a simple, but powerful form of multiple variable analysis. They provide unique capabilities to supplement, complement, and substitute for :

- traditional statistical forms of analysis (such as multiple linear regression)
- a variety of data mining tools and techniques (such as neural networks)
- recently developed multidimensional forms of reporting and analysis found in the field of business intelligence

Decision trees are produced by algorithms that identify various ways of splitting a data set into branch-like segments. These segments form an inverted decision tree that originates with a root node at the top of the tree. The object of analysis is reflected in this root node as a simple, one-dimensional display in the decision tree interface. The name of the field of data that is the object of analysis is usually displayed, along with the spread or distribution of the values that are contained in that field. A decision tree can reflect both a continuous and categorical object of analysis. The display of this node reflects all the data set records, fields, and field values that are found in the object of analysis. The discovery of the decision rule to form the branches or segments underneath the root node is based on a method that extracts the relationship between the object of analysis (that serves as the target field in the data) and one or more fields that serve as input fields to create the branches or segments. The values in the input field are used to estimate the likely value in the target field. The target field is also called an outcome, response, or dependent field or variable.

Once the relationship is extracted, then one or more decision rules can be derived that describe the relationships between inputs and targets. Rules can be selected and used to display the decision tree, which provides a means to visually examine and describe the tree-like network of relationships that characterize the input and target values. Decision rules can predict the values of new or unseen observations that contain values for the inputs, but might not contain values for the targets.

Each rule assigns a record or observation from the data set to a node in a branch or segment based on the value of one of the fields or columns in the data set. Fields or columns that are used to create the rule are called inputs. Splitting rules are applied one after another, resulting in a hierarchy of branches within branches that produces the characteristic inverted decision tree form. The nested hierarchy of branches is called a

decision tree, and each segment or branch is called a node. A node with all its descendent segments forms an additional segment or a branch of that node. The bottom nodes of the decision tree are called leaves (or terminal nodes). For each leaf, the decision rule provides a unique path for data to enter the class that is defined as the leaf. All nodes, including the bottom leaf nodes, have mutually exclusive assignment rules; as a result, records or observations from the parent data set can be found in one node only. Once the decision rules have been determined, it is possible to use the rules to predict new node values based on new or unseen data. In predictive modeling, the decision rule yields the predicted value.

IV. RESULTS

As in [17], the entire sample of 768 was first randomly divided into a training and test sample, with 576 cases used for training, and 192 for test. The training sample had 378 subjects without diabetes, and 198 subjects with diabetes. For the test set, 122 subjects did not develop diabetes, while 70 did.

The researchers undertake the predictive accuracy of the decision tree.

Plasma Glucose (PG) concentration is the best criteria or the starting node of the decision tree with a 73.83% predictive accuracy as shown in the Table 2. As shown in Figure 6, if the PG test is less than or equal to 127, it is further tested on the Body Mass Index. With 8 variable-predictor to diabetes, it is observed that some of the variable is repeatedly tested and the tree further grows. Among the test conducted, The highest predictive accuracy obtained from 5 variables and 3 variables with a predictive accuracy of 75%.

With 5 variables with a 74.87% predictive accuracy as shown in table 3, It starts taking prediction based on PG concentration as shown in Figure 7, that is if the PG test result is greater than 127, it will go further to test the BMI, if the BMI is less than or equal to 29.9 then that subject-respondent is healthy, otherwise the responded is predicted to have a diabetes. As show in Figure 7if the BMI <= 26.4, the respondent is healthy, otherwise additional predictor-variable needs to be evaluated.

Prediction can be reduced to 3 variable-indicator with predictive accuracy of 74.87% as shown in table 4, In Figure 8, it starts taking prediction based on PG concentration if the test result is <= 127, then the subject-respondent is healthy, otherwise if the test result is >=127, it will further analyze the BMI, that is if the BMI <=29.9 the subject-respondent is healthy, otherwise predicted to have a diabetes.

ACKNOWLEDGMENT

The authors wish to express their sincere and profound gratitude, a genuine feeling of admiration to the persons who extended their untiring efforts to work together and made valuable contribution to the success of this study

Heartfelt thanks to Dr. Anil K. Maheshwari, their professor in Data Mining course, for his invaluable guidance and help in the researchers pursuit of this study. His willingness to share substantial ideas for the improvement of this study will never be forgotten.

To the researcher's family who were the researchers source of strength and inspirations which brought about the success of this piece of work and for always being supportive in all their endeavors.

To our Almighty God, for being the source of wisdom, unconditional love and blessings.

REFERENCES

[1] Diabetes Health Center. (n.d.). from http://www.webmd.com. Retrieved 2015

[2] International Diabetes Federation: www.idf.org,

[3].World Health Organization: from www.who.int Retrieved January 5, 2015

[4] Bennett, PMH., T.A. Burch, and M. Miller. 1971. Diabetes mellitus in American (Pima) Indians. Lancet

[5] Knowler, W.C.et al. Diabetes incidence and prevalence in Pima. 1978.

[6] Knowler, W.C., DJ. Pettitt, PJ. Savage, and P.H. Bennett 1981. Diabetes incidence in Pima Indians: contnbutions of obesity and parental diabetes. Am J Epidemiol 113:144-156.

[7] Anil Maheshwari. Business Intelligence and Data Mining. Business Experts Press. 2014

[8] Jefferson Gante Sanidad, B. G. S, et al. Learning Style and Students' Perception of Satisfaction on Web-based Learning Environment at Arabian Gulf University-Kingdom of Bahrain.

[9] del Rosario, J, Sanidad, J. et al. Modelling and Characterization of a Maze-Solving Mobile Robot Using Wall Follower Algorithm. Applied Mechanics and Materials, 446, 1245-1249.

[10] World Health Organization, Report of a Study Group:Diabetes Mellitus. World Health Organization Technical Report Series. Geneva, 727, 1985.

[11] Barry de Ville, Decision Trees for Business Intelligence and Data Mining, SAS Press, 2013

[12] Minsky, M., and S. Papert. Perceptrons.Cambridge, MA: MIT Press. 1969

[13] Muralli S. Shanker. Using Neural Networks to Predict the Onset of Diabetes Mellitus. Kent State University,

[14] Lippmann, R.P. An Introduction to computing with neural nets. IEEE ASSP Magazine, 4, 2-22

Confusion Matrix	Sick	Healthy
Sick	160	108
Healthy	93	407

Table 2. Decision Tree with 8 variables.

Confusion Matrix	Sick	Healthy
Sick	149	119
Healthy	74	426

Table 3. Decision Tree with 5 variables.

Confusion Matrix	Sick	Healthy
Sick	148	120
Healthy	72	428

Table 4. Decision Tree with 3 variables.

Number of variables	Predictive Accuracy
8	73.83%
5	75%
3	75%

Table 5. Predictive Accuracy with different input variables.

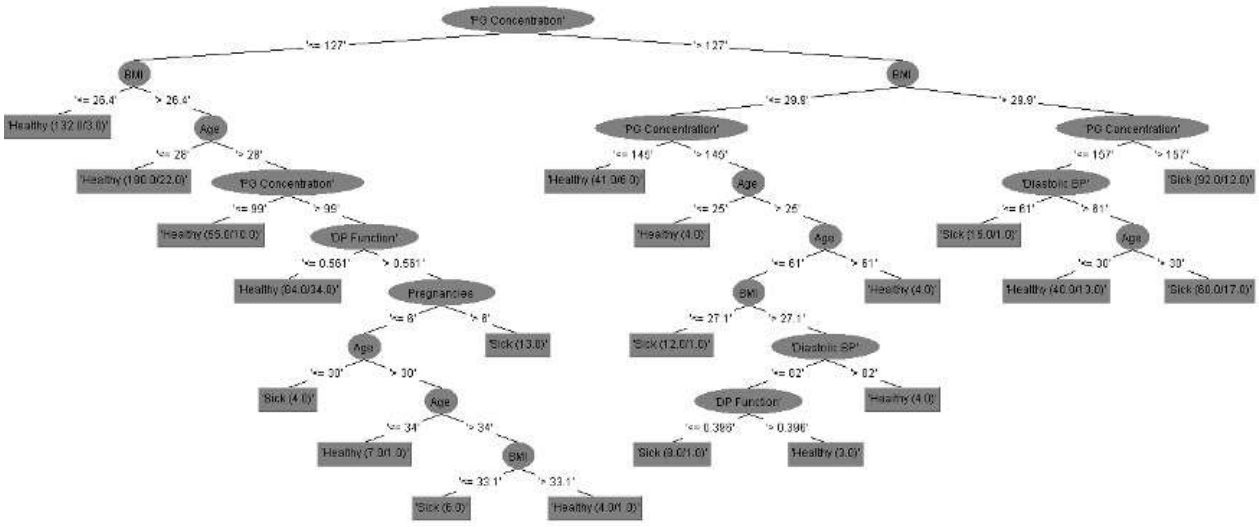


Figure 6. Decision Tree with 8 variables.

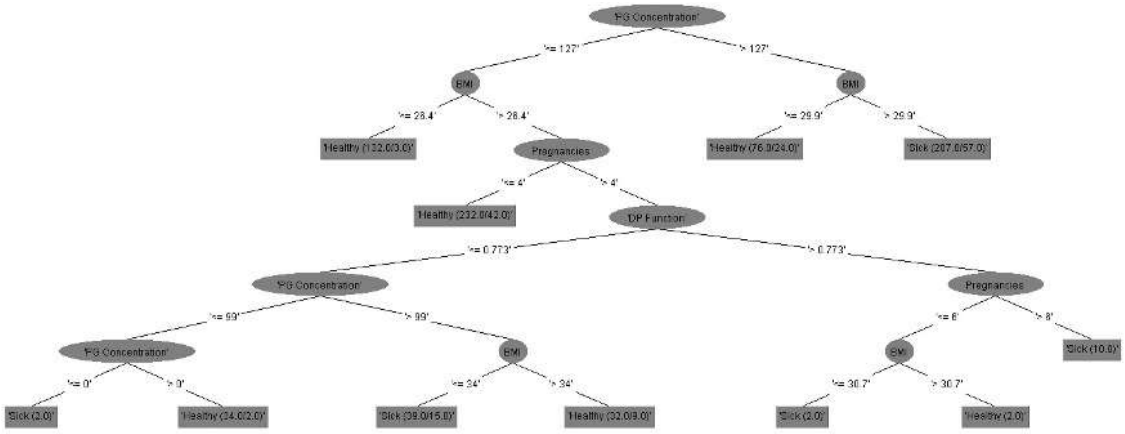


Figure 7. Decision Tree with 5 variables.

.

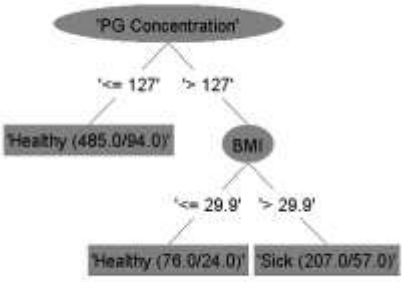


Figure 8. Decision Tree with 3 variables