IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Issue-3, March 2019 E-ISSN: 2347-2693

Diabetes Prediction using Data Mining

Suhasini Vijaykumar[1*], Manjiri Moghe[2]

[1,2]MCA Student, University of Mumbai, India

Corresponding author: suhasini.kottur12@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i3.749753 | Available online at: www.ijcseonline.org

Accepted: 21/Mar/2019, Published: 31/Mar/2019

Abstract **-** Data Mining is a way to extract information from large amount of data. It brings out one conclusion by applying its efficient techniques. In today's world, it has helped many of the domains and growing its root by enhancing in its own way. In various data repositories, large medical datasets are available which are used in real world applications. Information is been generated by using various Data Mining techniques. Classification technique separates the information so as to generate useful content from it. It also helps in medical field to detect diseases such as diabetes which has affected various people from different countries. Insulin is main concept while taking into consideration the term 'Diabetes'. Insulin acts as glucose for energy. It is a Gateway to body cells and controls glucose level in our body. Diabetes is a disease in which level of glucose in blood increases. To make it easy and recover from most early stages, prediction is necessary. It is been done with the help of data mining. This study is significant of predicting diabetes and helping medical industry to grow.

Keywords – Health, Decision Tree, Diabetes, Prediction

I. INTRODUCTION

In recent times, the number of people suffering from diabetes is increasing day by day. It is a disease in which body does not produce insulin or use it properly. This increase the risks of developing, kidney disease, blindness, nerve damage, blood vessel damage and contribute to heart disease [1]. Diabetes is continuing to advance throughout the world and is around for thousands of years. Millions of people in the world have diabetes, and almost half of those people don't know that they have it. It is also most important to know which type of diabetes a particular person has so as to take medications accordingly.

Type 1 diabetes known as juvenile diabetes which occurs by when immunity system of our body damage cells releasing insulin [4], it removes insulin production from body. Here, patients produce almost no insulin or very little of it. The disease usually develops in children or young adults. So, patient takes insulin injections everyday in order to control level of glucose.

Type 2 diabetes known as insulin independent diabetes. Overall, 90% of all people suffering diabetes come under this type. It begins with insulin resistance, a condition in which muscle, liver, and fat cells do not use insulin well. So, your body needs more insulin to help glucose enter cells. Type 2 may remain undetected for many years and may be diagnosed when some complications appears or a urine glucose test or blood test is been done. It is not always, associated with obesity or overweight, which itself can cause insulin resistance and lead to high blood glucose levels. Here, diabetes can be controlled with regular exercise and healthy diet plans

Gestational diabetes is a type of diabetes consisting of high blood glucose level during pregnancy. It may lead to complications for mother and baby. Women having diabetes and their children are at an increased risk of developing type 2 diabetes later in life even though it may disappear after pregnancy[8]. Approximately half of women with a history of Gestational diabetes later may develop type 2 diabetes within five to ten years after delivery. Nowadays, large amount of information is collected from hospitals. Knowledge discovery is done through data mining. Decision tree helps in predicting early diagnosis of this disease. Detection of diabetes in pregnant women at an early stage helps in further complications. This paper mainly focuses on diabetes recorded in pregnant women. Here, Decision tree algorithm is been used to predict whether patient has diabetes or not.

II. LITERATURE REVIEW

Data mining is one type of research to find out relationship among the huge amount of information which is been stored in data warehouses. It is one type of investigation. Different data mining techniques has Numerous work that has been done related to lifestyle disease diagnosis. The algorithms, methods, dataset used by various authors and concluded results along with future work is carried out in finding out efficient methods of medical diagnosis. It is a technique of previously valid, useful, unknown patterns in huge databases. It is usually used by business intelligence organizations and is increasingly used in the sciences to extract information. It is been extracted from the enormous data sets generated by modern experimental and observational methods. Data ware house has large structure of data. It has more relevant information stored. It reaches to us by applying various types of techniques. Some of them include classification, clustering, regression. Clustering tells the similarities and differences in data. Regression identifies the relationship among datasets. Classification brings out useful information and helps to classify data. Decision Tree is mostly popular because of its methodology. It solves the problem by using tree representation[7]. The deeper the tree, more complex the decision rules and the fitter is the model. It breaks down data set into smaller subsets. The final result is a tree with decision nodes and leaf nodes. Leaf node represents a classification or decision. The root node is the best predicator of the tree which is at top. Decision Tree models can be converted to IF-THEN rules. It is been easily understood because of the reasoning process.

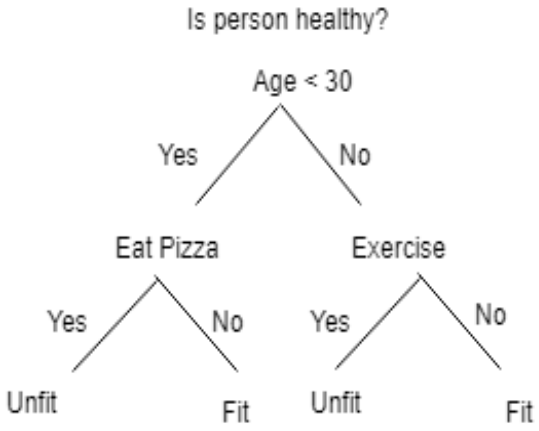

Fig 2.1 Example of Decision Tree

III. METHODOLOGY

The research framework for this project consists of two main phases which are Description and Pre-processing and then the implementation of it.

Description and Pre-processing-

This study is taking into consideration diabetes affecting the pregnant women from Pima Indian Diabetes Dataset [3]. The main aim of the study is to predict whether the patient is diabetic or not. The data mining classification is been applied to Diabetes Dataset.

1.1 Dataset Description

No. of Attributes	**No. of Instances**
9	768

The dataset describes attributes which takes record of women as input who has diabetes. Certain factors affecting the diabetes are mentioned here.

1.2 Attributes Description

Attributes	**Values**
a. Number of times pregnant	Preg
b. Plasma Glucose Concentration	Plas
c. Diastolic Blood Pressure	Pres
d. Skin Fold Thickness	Skin

e. Insulin	Insu
f. BMI	Mass
g. Diabetes Pedigree Function	Pedi
h. Age(years)	Age
i. Class Variable(0 or 1)	Class

Record of female patients is taken into consideration in this dataset. All values are numeric in nature. Pregnancy tells the count of pregnancy in women.

The 'tight diabetes control' is been maintained by people with diabetes. Glucose quantity is actually very small in blood and body fluids.

The diastolic Blood pressure, is in the arteries when the heart rests between beats. Blood pressure which is lower than 80 is normal.

The tester pulls the fold of skin away from the underlying muscle and pinches the skin at the location site so only the skin and fat tissue are being held. Skinfold thickness is been measured by Special skinfold callipers.

Insulin is released from the pancreatic beta cells when carbohydrates are ingested, and, to a lesser degree, protein. It is the primary hormone that responds to what we eat. Insulin is also secreted when the stomach stretches, regardless of food type. Insulin signals for the storage of sugars

BMI(Body Mass Index) is been calculated by measuring height and weight of a person.

$$BMI = \frac{Weight}{Height^2}$$

Diabetes Pedigree Function (pedi) provides some data on diabetes mellitus history in relatives and the genetic relationship of those relatives to the patient.

Implementation—

Dataset is in the CSV format. Such format is generally used for analysis and calculations, as the data is available in the sequence. ID3 algorithm is been used which is used as its attribute selection measure. The attribute with the highest information gain is chosen as splitting attribute for any node. It uses Entropy and Information Gain to construct Decision Tree. Firstly the pre-processing is done where the data is in the CSV(Comma Separated Value) format. Then it goes to the Decision Tree Algorithm. The calculations are been done on the attributes by applying certain methods and using ID3 algorithm of the data mining. The result tells us the prediction of dataset and classify women into diabetic and non-diabetic as described in Fig 3.1

It can act as a powerful tool in the healthcare industry to determine the percentage of such disease and patients can take care of themselves prior so that non can go out of control.

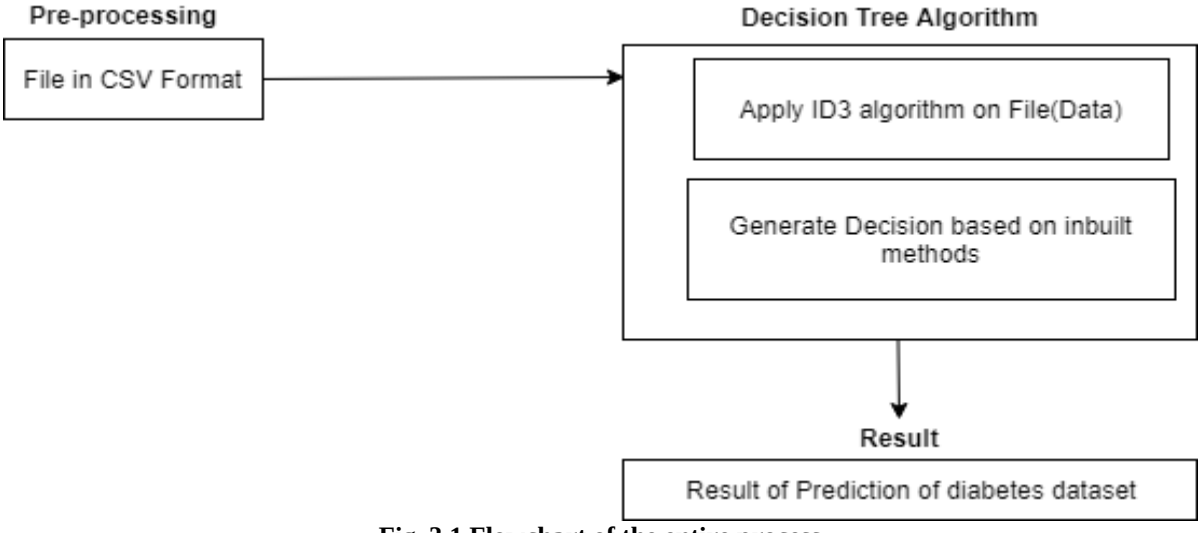

Fig. 3.1 Flowchart of the entire process

Entropy

H(S) is the measure of uncertainty in the dataset S.

$$H(S) = \sum[-\mathrm{p(x)}\log_2\mathrm{p(x)}]$$

Where,
S = current dataset where entropy is calculated.
X= set of classes in S
P(x) = proportion of number of elements in class X to the number of elements in set S.

Information Gain-

When the dataset is split on an attribute the Information Gain is calculated based on entropy.

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

IV. RESULT AND DISCUSSION

Decision tree algorithm is used for generating the result. The sample dataset is been taken into consideration and with the help of python language as backend, final result is been calculated by applying function related to decision tree. All the attributes are taken into account and final result is been displayed by the bar chart. Here, one can easily discriminate between the patients. This technique is much more useful in the health industries for prediction. With this ratio we come to know about the severity of the disease and what care should be taken in near future. Accuracy is been measured by considering all the instances used in the diabetes dataset. The result describes that the risk is less as diabetes detected in patients are less and those not having it are more. It also tells it can be eradicated by taking proper care and going through certain useful healthy measures. The performance depends upon people suffered from diabetes or not. It predict that by using data mining as the root one can reduce the risk in near future.

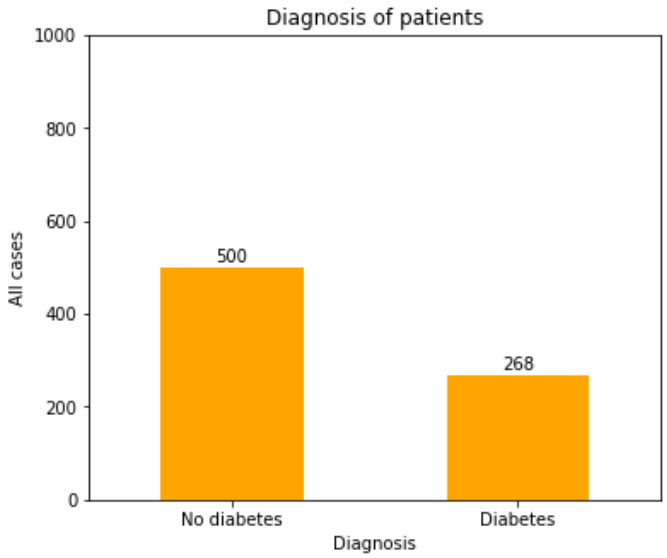

Fig 4.1 Bar graph showing diabetes detection

V. CONCLUSION

Data mining techniques brings out the Hidden patterns which are been extracted from various type of data. This study is different from others as it mainly focuses on dataset of pregnant women. Decision Tree algorithm produces accurate result and end up with the conclusion of prediction. Diagnosis of such disease should be done early so as to deny the further problems which may occur in near future. In generation of this research work, details of the disease and its types are also understood in a better way. Further, in future work may increase by including complex classification models of data mining and predicting everything at an early stage so that there will be certain amount of reduction in the number of diabetic patients all over the world.

REFERENCES

[1]. E.I.Mohamed, R.Linderm, G.Perriello,N.Daniele, S.J.Poppl, & A.DeLorenzo. "Predicting type 2 diabetes using an electronic nose based artificial neural network analysis," Diabetes nutrition & metabolism, 15(4),215–221.202.

[2]. Frank, A. & Asuncion, A. (2010). UCI Machine Learning Irvine, CA: University of California, School of Information and Computer Science.

[3] Khyati K.Gandi, Prof. Nilesh B. Prajapati. *'Study of Diabetes Prediction using Feature Selection and Classification'* International Journal of Engineering Research & Technology. Vol-3. Issue-2. Feb-2014

[4] T.monika Singh, Rajashekar shashtry *'Prediction of Diabetes Using Probability Approach'* International Journal of Engineering Research & Technology. Vol-4. Issue-2. Feb-2017

[5] Data Mining Concepts and Techniques by Jiawei Han and Micheline Kamber [Book]

[6] Predicting Diabetes: Clinical, Biological, and Genetic Approaches Data from the Epidemiological Study on the Insulin Resistance Syndrome (DESIR)

[7] Priya B. Patel, Parth P. Shah, Himanshu D. Patel. *'Analyze Data Mining Algorithms for Predicition of Diabetes'* International Journal of Engineering Development and Reseach. Vol-5. Issue-3. 2017.

[8] Marije Lamain – de Ruiter, Anneke Kwee, Christiana A. Naaktgeboren, Arie Franx, Karel G. M. Moons and Maria P. H. Koster *'Prediction models for the risk of gestational diabetes: a systematic review'* Diagnostic and Prognostic Research 2017 1:3

