Diabetes Detection using Big Data Analytics with Logistic Regression Model

[1]Suramya S, [2]Srinivasachar G

[1]Student, [2]Assistant Professor
[1]Department of Computer Science and Engineering,
[1]Atria Institute of Technology, Bangalore, India.

Abstract: Diabetes Mellitus is one of the diseases that cannot be transferred from one person to another and has great impact on human life today. Due to the lifestyle and work schedule changes in the recent times, India houses lots of diabetic people in it. There are many other disorders connected to diabetes mellitus hence treating it in an early stage is very much necessary. The health care systems generate a large amount of data in less time, and these data can be both structured and unstructured. The main task here is to store, manage and analyze this data. Hence, big data analytics is used for enhanced insight, predictions about developing other disorders in the near future and also improves health care system by reducing the execution time and the optimal cost. The goal of this paper is to detect if a person would develop diabetes in future so that it can be obstructed or at least push it further for a few years. To do this we use Logistic Regression Statistical Model which minimizes the classification error. As there is a saying, "Prevention is better than Cure", so providing an alert regarding their vulnerability towards a specific medication is always helpful.

IndexTerms **-Diabetes, Big Data, Regression, Healthcare System, Statistical Model.**

I. INTRODUCTION

Diabetes Mellitus is an ancestral disorder which occurs due to frequent fluctuations of the glucose level in the human body where the production of hormone called insulin by pancreas is less. Insulin is very much essential in moving the glucose from the blood to cells and utilize it to produce energy. People getting diagnosed with this disease is increasing day by day worldwide. There are three types in diabetes, they are Type I, Type II and Gestational Diabetes. Type I is a persistent condition mostly identified in children below 16 years of age wherein the pancreas doesn't produce insulin at all. Hence, it is called insulin dependent i.e., insulin should be taken externally using injections. Type II is also a persistent condition identified in people above 16 years of age wherein the pancreas produce insulin but the cells will not be capable enough to use it and convert it into energy. Hence, it is called as non-insulin dependent and tablets will be recommended by the doctor. Gestational Diabetes is identified during pregnancy. It can be cured in few months with proper medication. Few symptoms of diabetes are frequent thirst and urination, fatigue, weight loss, skin related problems, etc. diabetic people will be very prone to develop other disorders like diabetic retinopathy, heart and kidney related problems and depression.

Big Data consists of data that demand cost-effective and innovative ways of data processing for enhanced insights and decision-making. In healthcare system, big data can be of immense use since there is lots of data entered. This data entered can be both structured and unstructured. In this paper, we mainly focus on storing, maintaining and analyzing of diabetes data. There are many algorithms used for diabetic analysis. We are using the logistic regression statistical model to detect if people would develop diabetes in the near future, and it can be obstructed or pushed further for a few years by following a diet plan and exercising daily.

II. EXISTING SYSTEM

At present, the technology developing rapidly has created space for innovative ideas and made human work easier. In the healthcare field, big data is of immense use for storing, maintaining, analyzing and predicting the data. The Diabetic data analysis with Predictive Method values the investigation of algorithm in Hadoop/ Map reduce program to analyze the stage of diabetes. With the help of Hadoop File System (HDFS), large data is stored and map reduce algorithm is implemented through which a decision tree is generated in top-down recursive separate and overcome approach. The decision tree uses a ratio to create the tree. The predictive method includes different phases like predictive analysis, collection of data, processing of data and analyzing the report. Later a chart is generated wherein if the fasting value is between 70 to 100 it is normal stage, between 101 to 125 the diabetes can be cured and if the value is 126 or above the diabetes cannot be cured, and the respective diabetic patient should be under strict medication. The flow diagram is as shown in Fig-1.

The diabetic patient data can also be analyzed using predictive method wherein a classification model is proposed with increased accuracy using the classifiers like decision trees, multilayer perception. Here using the two threshold two divisors and multilayer perception algorithms, type of diabetes can be analyzed i.e., if the patients are type I, type II or gestational diabetic depending on the symptoms, age, and if any ancestors were diabetic. The two threshold two divisors algorithm calculates the data length and block size and sets a range of chunks of data. This is sent as input to the multilayer perceptron algorithm where weights are multiplied, summed and sigmoid function is applied. The output is sent to each node in each layer until the network is produced. The main focus here is to increase the accuracy of the prediction. This system will provide the percentage of correct prediction.

There exists another scheme which mainly focuses on five goals i.e., cost-effectiveness, comfortability, personalization, sustainability and smartness, and also suggests the treatment pattern. With the implementation of algorithms like support vector mechanisms and artificial neural networks, decision tree is generated through which suggestions on diet plan and exercises are given. This system consists of three main layers. The first layer senses the data entered and sends the result to the second layer where the person with diabetes will be diagnosed. Later this result is sent to the third layer which helps to share the data with the patient and doctor. The suggested treatment method and diet plan will be cross-checked by the doctor, and sent to the patient.

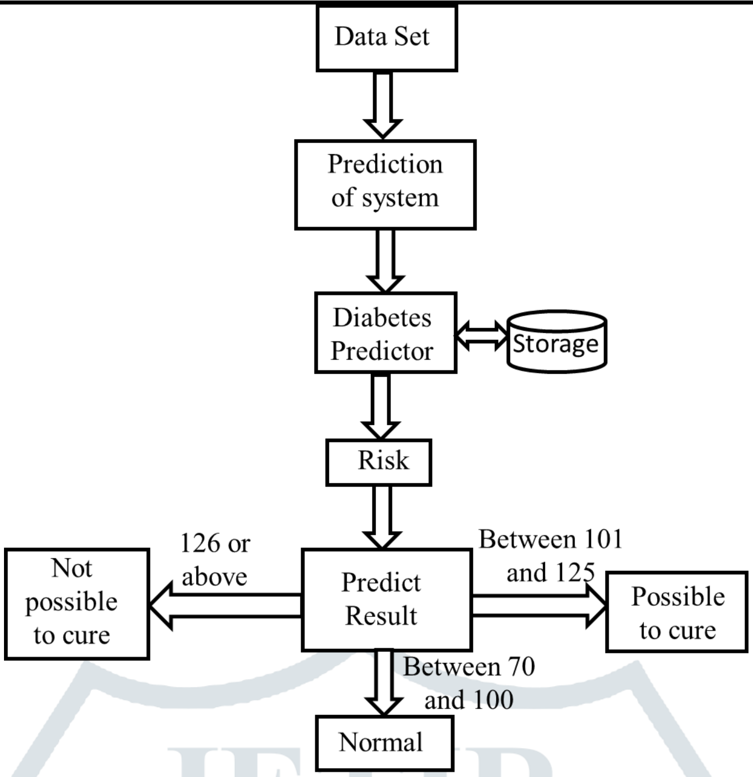

Fig-1: Flow diagram for diabetic prediction using predictive method

III. Proposed System

In the existing system, the main focus is on the diabetes prediction and analysis for people who already have diabetes and doesn't tell how it can be detected many years before. The risk of developing other disorders in diabetic patients is also not known. Hence, the goal of this paper is to detect if the person would develop diabetes in future so that it can be obstructed or at least push it further for a few years by following the right diet plan and exercise daily. For this we are using Logistic Regression Statistical Model.

Logistic Regression Model is a classification model in which the response variable is unambiguous. This algorithm is used for problems which are supervised. The main purpose of this model is to minimize the classification error. The working of this model is very simple and is one of the oldest statistical models. The logit function is obtained by taking probability of dependent variable with the natural logarithm. The logit function is used as a dependent variable and the net effect is observed. The general logistic function with independent variable on x-axis and logit dependent variable on y-axis is as shown in Fig-2,

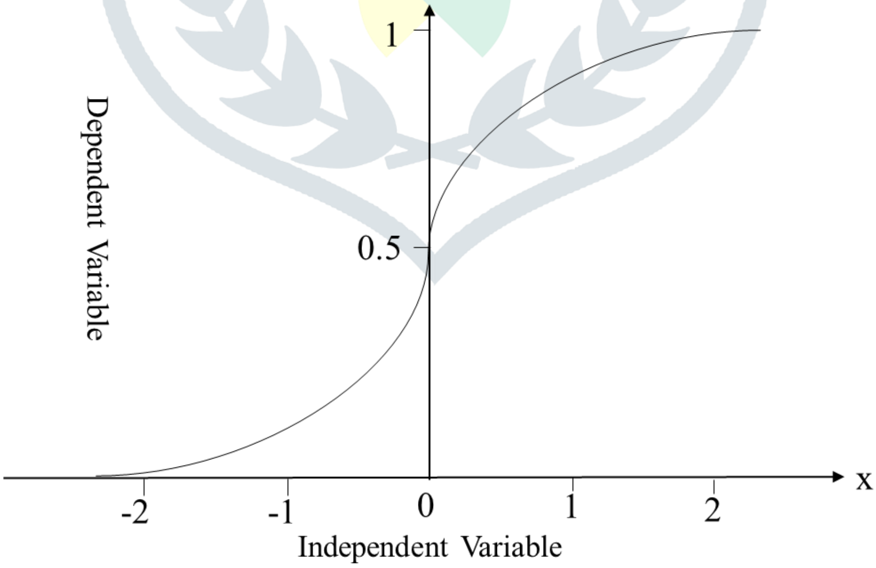

Fig-2: Graph depicting general logistic function

The error function is as follows:

$$Min \frac{1}{n}\sum_{i=0}^{n} \log(1 + e^{-y_i f(x_i)})$$

Here the coefficients are estimated and then logistic regression equation is obtained.

Where $f(x_i) = \alpha_1 x_{i1} + \alpha_2 x_{i2} + \cdots + \alpha_p x_{ip} = \sum_{k=1}^{p} \alpha_k x_{ik}$

In this problem, the x values will be assigned to age, ancestral disorder present, diabetes symptoms and α values are regression coefficients. The output of this method will be in the form of yes or no, 1 or 0, true or false.

The advantages of the regression model are, it is very easy to understand and provide simple algebraic equations. The strength is measured in terms of correlation coefficients. This model can match and beat the predictive power of the other techniques.

With the help of above explained model, we can detect if the person would develop diabetes in future or not. The main goal is to reduce the error and increase the accuracy of the prediction made. The variables entered to this system will be age, ancestor has diabetes or not, symptoms of diabetes like frequent urination and thirst, fatigue, weight loss, skin related problems, etc and these variables are independent. The Ockham's Razor principle can be utilized to fit the data well wherein there will be no need to multiply the variables entered. Here the unnecessary assumptions can be ignored. The data entered will be produced well by this principle and ready to do the required predictions. These data which are fit will be sent to the predictor where the logistic regression principle is used wherein if the value of data set is equal to or more than 0.5, then there are chances that the person would get diabetes in future and it can be obstructed or pushed further for few more years by following a diet plan and exercising daily. Else if the value of data set is less than 0.5, then there are chances that the person being detected with diabetes will be less. The flow diagram is as shown in Fig-3,

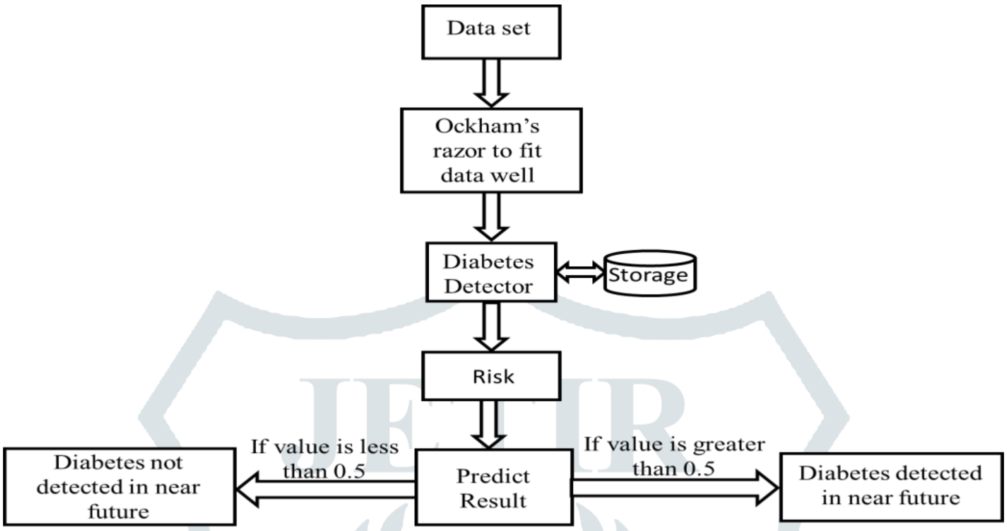
Data set
Ockham's razor to fit data well
Diabetes Detector
Storage
Risk
If value is less than 0.5
Predict Result
If value is greater than 0.5
Diabetes not detected in near future
Diabetes detected in near future

Fig-3: Flow diagram for diabetic detection with logistic regression statistical model

The roc curve abbreviated as receiver operating characteristic curve can be obtained with the help of the values found. The below graph shows the roc curve which is above straight line in blue color. The straight line refers to guessing the detection of diabetes randomly. The x-axis is false rate and y-axis is true rate.

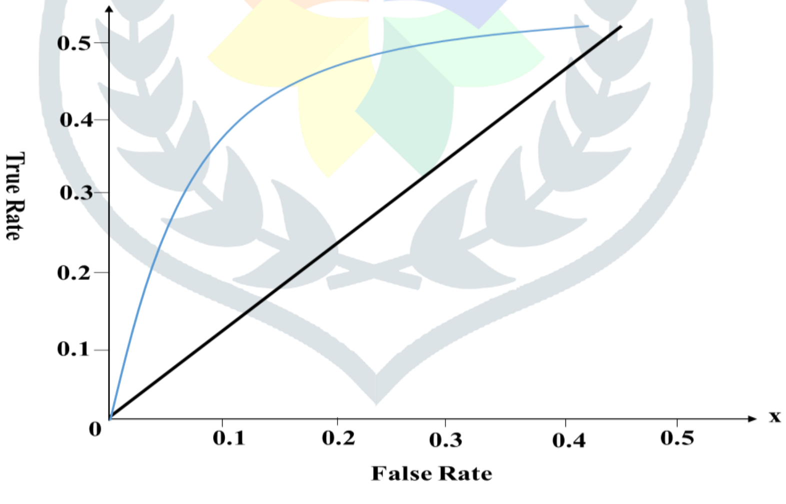

Fig-4: Graph depicting roc curve

IV. Conclusion

This paper explains about the diabetic detection using big data analytics algorithm. Diabetes may not be a non-curable disease, but is very dangerous if not treated at the right time. Hence, detecting the trace of diabetes presence beforehand is an advancement in healthcare using the present technology. Big data analytics with logistic regression statistical model for detecting if a person would develop diabetes in future is helpful for achieving good health. This model is very accurate compared to guessing the detection of diabetes randomly. It is fast compared to other algorithms like the map reduce and multiple perceptions methods. This work can be enhanced further like if the diabetic person is at the risk of developing other related disorders.

References

[1] Mrs. Ashwini Abhale, Shruti Gulhane, Sandhya Budhewar, Swanali Jathar, Harshada Sonwane, "Predictive Analysis of Diabetic patient data using Machine Learning & Big data".

[2] Thanga Prasad. S, Sangavi. S, Deepa. A, Sairabanu. F, Ragasudha. R, "Diabetic Analysis in big data using Predictive method".

[3] Min Chen, Jun Yang, Jiehan Zhou, Yixue Hao, Jing Zhang, Chan-Hyun Youn, "5G Smart Diabetes- Towards Personalized Diabetes Diagnosis with Healthcare Big Data Clouds", IEEE April 2018.

[4] Scott A. Czepiel, "Maximum Likelihood Estimation of Logistic Regression Models, Theory & Implementation".