ISSN No:-2456-2165

Data Analytics Applied on Perennial Disease using Mapreduce System

S. Kalaiarasi[1]
Chirag A[2], Mohammed Kaareem Khan R[3], Ulagappan R[4], Vignesh M[5]
[1]Assistant Professor (O.G), Department of Computer Science & Engineering,
SRM Institute of Science & Technology, Ramapuram
Chennai, India[1].
[2345]UG Student, Department of Computer Science & Engineering,
SRM Institute of Science & Technology, Ramapuram
Chennai, India[2345].

Abstract:- Presently most of the people are affected by aperennial disease which is "Diabetes". A great deal of research is currently being taken place. A model is proposed to foreshow existing system by implementing clustering and classifications techniques which are implemented to determine the type of diabetes. We diagnose diabetes based on the records of patient's data by which we analyse the seriousness of the diabetes. In this method for clustering the entire dataset into three clusters, where cluster-0 is used for gestational diabetes, cluster-1 is used for type-1 diabetes, cluster-2 is used for type-2 diabetes by using Naïve Bayesian algorithm. The Classification model gives the clustered dataset which further classifies patient's level of diabetes as mild, moderate and severe. To diagnose diabetes, performance analysis of various algorithms is done. The result is presented by showing the performance of various classification algorithms.

Keywords:- *Classification, Sorting, Diagnosis of Diabetes, Naïve Bayes, Random tree.*

I. INTRODUCTION

Diabetes is caused when there's a lack of insulin in a person's blood. It is usually referred to as diabetes, but it is meant as Diabetes mellitus. People who are diagnosed with Diabetes Mellitus are called ad "Diabetics".

A person with high blood sugar may have symptoms like frequent urination, increased level of thirst, and increased hunger. If diabetes if left untreated it may cause many difficulties. complications of diabetic are ketoacidosis, nonketotic hyperosmolar coma, or death. Over a long period of time various other complications arise that include heart disease, stroke, serious kidney failure, foot ulcers, and retinal damage. Pre-Diabetes occurs when there is an increase in the sugar level in the blood.

Diabetes occurs mainly due to the pancreas is not producing sufficient amount of insulin or if the cells of the body is not responding properly to the produced insulin. There are three types of Diabetes Mellitus:

If a person's pancreas fails to produce sufficient insulin, then it results in Type 1 Diabetes Mellitus. It is referred to as "insulin-dependent diabetes mellitus" (IDDM) or "juvenile diabetes". Younger people who are below 20 years of age are usually affected. People suffering from type 1 diabetes suffer throughout their life and rely on insulin injections to maintain the imbalance in their insulin levels. The affected people must follow a strict diet and exercises as instructed by their dietician.

If a person has an insulin resistance, a condition where the cells fail to respond to the insulin in an appropriate way. The level of insulin may also reduce as the disease advances. This is termed as "non-insulin-dependent diabetes mellitus" (NIDDM) or also known as "adult-onset diabetes". The maximum cause of this type of diabetes is due to excessive body weight and lack of exercise.

The third type of diabetes mellitus is the Gestational Diabetes where it occurs to pregnant women without previous history of diabetes, where they develop very high blood sugar levels in them. A recent study shows that around 18% of pregnant women have diabetes. There is also the risk of developing the Gestational diabetes when pregnancy occurs during older age period.

The prime reason for Type 2 Diabetes is Obesity. By doing regular exercises and by maintaining a proper diet, the type 2 diabetes can be avoided. If there is no effect in the reduction of glucose level by maintenance of body, then medicines can be prescribed. A survey conducted by the National Diabetes Statistics Report states that around 29.23 million people or 9.41% of the U.S population have the type 2 diabetes.

As per the survey taken on 2015, it was stated that an estimated 416 million people were suffering from diabetes, where 90% of the cases were under the type 2 Diabetes mellitus. Early death risk is almost doubled if the person is suffering due to diabetes. Almost 2.0 to 5.0 million deaths occurred due to diabetes from the year 2013 to 2016. Globally diabetes costed an estimated amount of US$613 billion, and US$245 billion alone in the United States in 2014.

ISSN No:-2456-2165

As per the reports provided by the International Diabetes Federation, there were about 367 million people in 2011 who were suffering with type 2 diabetes, and they also predicted that the figures may be increased drastically to 553 million by 2030. Around 80% of the people affected by diabetes usually belonged to the middle and low-income. countries. The side effects to the people who have high blood sugar level can have heart disease, strokes, kidney failure and diabetic retinopathy. By 2025 the figures will be increased, for the people who have been affected by type 2 diabetes. In India the figures of diabetes have reduced by 2.8% in rural area when it's compared to the urban area. The Indian Diabetic Risk Score (IDRS) has estimated that a person who has a normal blood pressure but also has high Indian diabetic risk score is declared to be hypertensive or diabetic. Of all the diabetics 90% belong to the category of type 2 and the remaining 10% belong to the category type 1 and the gestational diabetes.

Clustering and classification are some of the data mining techniques that can be used to determine the health status of diabetic patients. Cluster analysis or the clustering technique is a process where the objects are grouped in the form of certain sets, where the objects are of the same type in the group (clusters) when compared to other groups (clusters). This clustering is a common technique for statistical data analysis, where it is used in various fields which incorporates machine learning, image processing, pattern recognition, data retrieval, bioinformatics, data compression and also in computer graphics.

Classification is a technique which is based on the supervised machine learning approach. The classification technique assigns target classes to different objects or groups. The classification is of a two-step process. The first process is the Model construction, where the training dataset in a database is evaluated. The second process is the Model usage, where the above constructed model is being used for the classification purpose. The accuracy of the classification is calculated based on the percentage of the test samples or test dataset that is classified.

The type of diabetes and its harshness level for every patient is determined by applying some of the data mining techniques such as sorting and classification.

Sorting and classification is a few of the data mining techniques that can be used to study the health of a diabetic patient. When a set of objects are grouped such that there are similarities in the closest group than the other groups then it is known as clustering. It is also a technique that is very common for analysis of statistical data. It's implemented in various fields like includes machine learning, pattern recognition and etc.

Classification is a technique where a target class has different objects assigned to it. There are two processes model construction, which is used to evaluate the training dataset and model usage, where the model that is constructed is used for classification. the accuracy of the classification is based on the classified test samples.

The diagnosis of the diseases and the severity level for every patient is done using data mining techniques such as clustering and classification.

This paper contains various sections. The 2nd sections contains literature study. While section 3 contains the methodology and the experimental results are given in section 4.the 5th section concludes the work.

II. RELATED WORKS

There are about 350 million people suffering from diabetes and the studies show that it will be one of the leading causes of death worldwide by 2030 according the world health organization. During the next 10 years the death due to diabetes is expected to rise by about 50 %. The number of people suffering from diabetes is increasing in every country and almost 4 out of 5 people live in the low and middle-income countries and half of those people don't even know that they suffer from diabetes. Obesity, overweight and the lack of physical activity and be the main cases of the epidemic.

An approach known as Cole was presented by Gao et al, which helps to identify diabetes at an early age. Cole runs multiple miner agents as well as combination agent. It improves the knowledge of the data which is presented using different methods.

Rajesh used various classification algorithms like ID3, C4.5, LDA, Naïve Bayes to diagnose diabetes using a given dataset. C4.5 was concluded as the best algorithm with 91% accuracy and less error rate of 0.938.

Adidela used fuzzy ID3 method to present the type of diabetes. the data set is first clustered and classification algorithms are applied and then the diseases is predicted.

Patil classified type II diabetes using the apriori algorithm. four association rules for class value "yes "and class value "no" and ten association rules were presented.

Ananthapadmanaban developed SVM and Naïve Bayes classification algorithms to detect diabetic retinopathy. The latter algorithm had a higher accuracy rate.

G. Parthiban et al applied Naïve Bayes method to diagnose heart related problems that occur in diabetic patients

P. Padmaja proposed a model that used various clustering techniques to characterize diabetes data and analyzed that to get different evaluations.

Proper exercise and physical activities and a strict diet can help keep diabetes from increasing. medicines are prescribed for severe conditions.

III. METHODOLOGY

To predict the type of diabetes simple K-Means method is used. Similarly, classification algorithms such as Naïve Bayes, Random Tree are used for predicting the risk factors of diabetes patient in the proposed model.

IV. DISCUSSION

Proposed model of this paper contains Three steps.

- Step 1: The data is first processed
- Step 2: Using Simple K-Means algorithm sort the data as type-2 diabetes, Type-1 diabetes and Gestational diabetes.
- Step 3: Classification algorithms are used to describe the risk level of diabetes of a particular patient

V. DATASET USED

ATTRIBUTE	DISCRIPTION	TYPE
Gender	Considered as Male=1 Female=0	Numeric
Insulin Dependent	Considered as min=50and max=500	Numeric
HbA1c	Considered as min=13 and max=19	Numeric
Plasma	Considered as min=2 and max=11	Numeric
Systolic	Blood pressure considered as min=30 and Max=370	Numeric
Mass	BMI Considered as min=1 and max=200	Numeric
Diastolic	Blood pressure Considered as min=60 and max=350	Numeric
Bg	Blood group Considered as 0= ‘O’, 1= ‘A’, 2 = ‘B’, 3 = ‘AB’	Numeric
Age	Considered as min=1 and max=125	Numeric
Pregnancy	Considered as 1= yes 0= no	Numeric
Pedigree	Considered as 0= no family history and 1= family history	Numeric
Living area	Considered Living area as 0=Urban and 1= Rural	Numeric
Food habit	Considered as 0=Healthy and 1=Moderate and 2= Junk Food	Numeric
Job type	Considered as 0= stressed job and 1= unstressed job	Numeric

Table 1. Attributes And Data Used

VI. ACCURACY MEASURE

For this work Naive Bayes, Random Tree and algorithms are used. Inter Cross validation is used to obtain the process. Accuracy shows how the datasets are classified. Recall and precision is used to measure the accuracy.

Accuracy = (TP + TN)/ (TP + TN + FP + FN).
Precision = TP/ (TP + FP).
Recall = TP/ (TP + FN).
TN - Negative tuples.
TP - Positive tuples.
FN - Incorrectly classified negative tuples.
FP - Incorrectly classified positive tuples.

Classifier	Precision			Recall		
	Mild	Moderate	Severe	Mild	Moderate	Severe
Naïve Bayes	1	0.87	0.89	1	0.76	0.87
Random Tree	0.921	0.926	0.985	0.952	0.934	0.978

Table 2. Recall And Precision Values

VII. RESULTS

WEKA tool is used to execute system. There are three stages:

- *Stage 1*- the dataset is uploaded and pre-processedusing Map reduce system.
- *Stage 2*- Once the data is sorted into types of diabetesthen it is used to find the type of diabetes for all patient.
- *Stage 3*- Once the classification is performed, the riskof diabetes is predicted.

(A). PERFORMANCE OF MAP REDUCE SYSTEM

The Map Reduce System sorts the entire dataset into 3 clusters as

- gestational diabetes
- type-1 diabetes
- type-2 diabetes

The elapsed time for construction of the models is 0.18 seconds. After pre-processing 620 entries 359 were in type -2, 115 were in type-1 and 146 were in gestational diabetes.

(B). Classification Algorithm Performance

The sorted dataset of every patient's risk levels of diabetes is classified as mild, moderate and severe. Then classification algorithms are used.

Diabetes Type	Number of Patients	Risk Level	Number of Patients
Type-0 Diabetes (Gestational)	145	Mild	60
		Moderate	56
		Severe	27
Type-1 Diabetes	114	Mild	24
		Moderate	53
		Severe	35
Type-2 Diabetes	358	Mild	55
		Moderate	167
		Severe	135

Table 2. Risk Levels of Diabetes

Classifier	Error Rate	Accuracy Value
Naive Bayes	0.092	90.8
Random Tree	0.037	96.2

VIII. RANDOM TREE

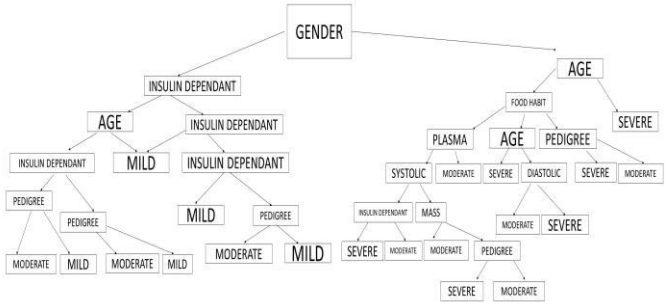

Fig 1:- Random Tree

An inference can be drawn from this tree that depicts certain attributes like plasma, age, pedigree and food habit are prominent in diagnosing diabetes. The tree shows information of diabetic patients where their insulin dependent level is lower than 188.6mmol, the diabetic level of a patient would be mild if their insulin dependent value is greater than 188.6mmol.The tree describes when the age of the patient is less than 35 with pedigree =1, then the patient may have moderate level of type-2 diabetes where patient has low insulin level and if the patient's age is greater than 35 with pedigree=0 and HbA1c value greater than 5%, its termed as severe level of diabetes.

IX. ACKNOWLEDGEMENT

All my special thanks to our guide and mentor Mrs. S. Kalaiarasi for her kind support and suggestions. It would not have been possible without her help. Also, would like to extend our sincere thanks to all the faculty members of the department of Computer Science and Engineering, SRM Ramapuram for their help.

X. CONCLUSION

Diabetes is considered a perennial disease among the humans. The only way to prevent it is to create awareness among the humans about diabetes the major cause of type 1 and type 2 diabetes which intern lead to heart problems, kidney diseases and eye related problems. Prevention or control of gestational diabetes is important. Women who have them are more are likely to develop type-2 than the women who don't. The children of such a mother may have the risk of obesity and type-2 diabetes. All of these difficulties could be easily cured if we constantly maintain our blood sugar levels. Datamining techniques help and draw a path to predict diabetes. These techniques can be tried out in various other

medical fields which would improve the diagnosis of other diseases also. It helps doctors and hospitals by improving the healthcare intelligence and hence the readmission of patients and help the system as a whole. The proposed model helps to enhance the medical intelligence of healthcare systems and thus helps predict diseases more accurately and thus help to find ways to improve a patient's health overall.

REFERENCES

[1]. Type1diabetes.Availablefrom:http://www.diabetes.org/diabetesbasics/type
[2]. National Diabetes Statistics Report. 2014. Available from:http://www.cdc.gov/diabetes/pubs/statsreport14/national-diabetes-reportweb.pdf
[3]. Type-2 diabetes in India: Challenges and possible. solutions.Availablefrom:http://www.apiindia.org/medicine_update_2013/chap40.pdf
[4]. JaliMV, HiremathMB. Diabetes. Indian Journal of Science and Technology. 2010 Oct; 3(10).
[5]. Ferreira D, Oliveira A, Freitas A. Applying data mining techniques to improve diagnoses in neonatal jaundice. BMC Med InformatDecis Making. 2012; 12:143. DOI: 10.1186/1472-6947-12-143.
[6]. National Center for Chronic Disease Prevention and Health Promotion. Gestational Diabetes. Centers for Disease Control and Prevention. U.S. Department of Health and Human Services; 2011. Available from: http://www.cdc.gov/
[7]. AfrandP,Yazdani NM, Moetamedzadeh H, NaderiF,Panahi MS. Design and implementation of an expert clinical system for diabetes diagnosis. Global Journal of Science, Engineering and Technology; 2012.
[8]. AljarullahAA. Decision tree discovery for the diagnosis of type II diabetes. International Conference on Innovative in Information Technology; 2011. p. 303–7.
[9]. LanordM, Stanley J, Elantamilan D, Kumaravel TS. Prevalence of Prehypertension and its Correlation with Indian Diabetic Risk Score in Rural Population. Indian Journal of Science and Technology. 2014 Oct; 7(10):1498–503.
[10]. Han Kamber M. Data mining concepts and techniques.. 2nd ed. Amsterdam, Netherlands: Elsevier Publisher; 2006. p. 383–5.
[11]. Gao, Denzinger J, James RC. CoLe: A cooperative data mining approach and its application to early diabetes detection. Proceedings of the 5th International Conference on Data Mining (ICDM'05); 2005.
[12]. Rajesh K, Sangeetha V. Application of data mining methods and techniques for diabetes diagnosis. International Journal of Engineering and Innovative Technology (IJEIT). 2012; 2(3):224–9