International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 01 | Jan 2020 www.irjet.net p-ISSN: 2395-0072

Predictive Analysis and Healthcare of Diabetes

Dr. Preeti Patil
HOD, Department of IT
DYPCOE, Akurdi

Amit Sharma, Pooja Singhal
Smriti Raina, Aniket Khaire
DYPCOE, Akurdi

Abstract: Diabetes mellitus is one of the major noncommunicable diseases which has a great impact on human life today. It is considered as one of the deadliest diseases today. If diabetes is left untreated, many health complications may occur.

A huge amount of data is gathered from the healthcare industry which is in a highly unstructured manner. It is necessary to structure the data into simple values. By applying computational analytics on the massive amount of data generated in healthcare system, will be used to create medical intelligence systems which will help medical prediction. This will create healthcare system which will be patient-centered and will reduce medical cost.

Machine learning (ML) is a method in which the system learns automatically from experience and improves the performance of the system to make more accurate predictions. To classify the patients into diabetic and non-diabetic we have developed and analyzed various predictive models. For this purpose, we used supervised machine learning algorithms namely linear kernel support vector machine (SVM-linear), *k*-nearest neighbor (*k*-NN), Decision Tree, etc.

Keywords: Big data, Hadoop, MapReduce, SVM, k-NN, Decision Tree.

I. INTRODUCTION

Today many people are getting affected by diabetes. With the help of this project people will become aware about diabetes and its risks and will be able to prevent it. This project will help the people to track their blood glucose levels and help them by providing proper healthcare to prevent or reduce their blood glucose level.

There are three main types of diabetes i.e. Type-1, Type-2 and Gestational diabetes. In Type-1 diabetes insulin is not produced in the body and the patient is required to take regular dose of insulin via injections. This is referred to as Insulin-Dependent Diabetes Mellitus (IDDM). About 10% of people have this type of diabetes. In Type-2 diabetes body develops insulin resistance and doesn't respond to insulin appropriately. Gestational diabetes develops during pregnancy, it causes high blood sugar that can affect your pregnancy and your baby's health.

Gathering huge amounts of data for medical use has been costly and time-consuming. With today's ML technologies, it becomes easier to collect such data and the data can also be converted into relevant critical insights. These insights can then be used to provide better healthcare. In this project we use predictive analysis models to provide a diabetes data interpretation system which is capable of detecting and monitoring the diabetic level of patients. Predictive analysis is a method, that incorporates a variety of techniques from data mining, statistics, and game theory that uses the current and past data with statistical or other analytical models and methods, to determine or predict certain future events.

II. LITERATURE SURVEY

In order to handle the huge amount of medical data gathered there was a need to develop an effective architecture. The architecture used in this project makes use of Hadoop for storing and handling of the big data collected. The MapReduce framework of Hadoop is used for performing computations on the huge amount of data for easy analysis.

We used the predictive analysis algorithms in Hadoop/Map Reduce environment to predict the diabetes types, complications associated with it and the type of treatment to be provided. The project consists of various machine learning algorithms called Decision Tree, SVM and Naive Bayes. The project focuses on an automated diabetes data interpretation system which combines different computing approaches in order to identify and highlight key clinical findings in the patient's diabetes data.

III. EXISTING WORK

A literature review shows many systems for diabetes prediction carried out by various methods. Many people have developed various prediction models using data mining to predict diabetes. Combination of various classification-regression models handles the missing and outlier data in the diabetic data set. They replaced these missing values with appropriate values which can be used for an effective analysis.

By using regression-based data mining techniques on the diabetes data, they discovered patterns using SVM algorithm that identify the best mode of treatment for diabetes across different ages. They concluded that the treatment of young aged patients can be delayed whereas the treatment of old aged patients must be done immediately.

The different big data technologies and research over healthcare are combined with efficiency. cost savings, etc. help in developing effective healthcare systems. The Hadoop usage in health care became very important in processing and managing large amounts of data.

IV. SYSTEM ARCHITECTURE

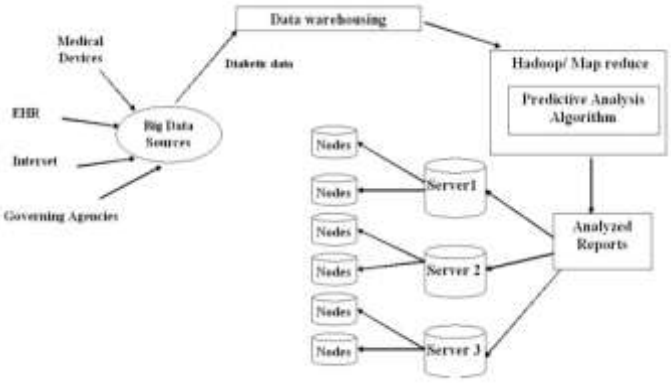

Fig.1 Proposed system architecture

The system architecture for the predictive system includes different phases like data collection, data warehousing, Map Reduce and analyzing reports. The system uses the predictive algorithms during Hadoop/MapReduce to predict the diabetes results.

V. DATA COLLECTION

The data is collected from various sources such as Medical devices, EHR (Electronic Health Records), Governing agencies, etc. The data from these different sources will help create a large dataset for effective analysis.

VI. DATA WAREHOUSING

In this phase massive unstructured data warehoused into single unit in which, data from various sources is cleansed, accumulated and made ready for further processing. Integration of various EHRs can help in identifying.

VII. HADOOP

Hadoop is an open source tool from Apache. Hadoop can store and process extremely large amounts of data across a large cluster of computers. Hadoop allocates data to numerous clusters, each of which solves different parts of the larger problem and then combines them for the final result.

Nowadays organization are producing huge of amount of data at rapid rate. A survey conducted on data generation says that Facebook produces 600TB of data per day. To handle such data a system is required which can store large amounts of structured and unstructured data. The conventional filesystem is incapable of storing and processing such huge data. Hadoop has been developed to handle such huge amounts of data and provide high computation capabilities. Hadoop uses two main components to do its job: Hadoop Distributed File System and Map/Reduce.

VIII. HADOOP DISTRIBUTED FILE SYSTEM

The Hadoop distributed file system (HDFS) is a distributed file system for the Hadoop framework. HDFS is responsible for the storage of big data in Hadoop. In HDFS the data is divided and stored across various nodes of the cluster.

HDFS has five services as follows:

Name Node: HDFS consists of one Name Node. It is also known as the Master Node. The name node contains the metadata of the data stored in HDFS. It is responsible for mapping the data to the data nodes and contains information such as in which data node is the required data stored.

Data Node: The data node is the place where the data is actually stored. A Data Node stores data in it as the blocks. It is also known as the slave node. The data nodes are responsible for providing read and write requests from the clients. The data nodes also perform block creation, deletion, and replication when instructed by the name node.

Secondary Name Node: This can be considered as the backup node in Hadoop. The secondary name node is used to create a checkpoint for the file system's status. In case the primary name node fails the secondary name node can be used to restore the systems to the previous stable state.

Job Tracker: Job Tracker is used for processing the data. Job Tracker receives the requests for Map Reduce execution from the client. Job tracker interacts with the Name node to know about the location of the data which is to be processed.

Task Tracker: Task tracker is the Slave Node for the Job Tracker. The Job Tracker assigns tasks to the Task Tracker. The Task Tracker then performs the task and reports back to the Job Tracker.

IX. MAP REDUCE

The Map Reduce is a programming model provided by Hadoop that allows complex computations on huge amount of data. Every Map Reduce program has at least one Mapper and one Reducer function.

The Map() function is present on the master node and divides the input data into smaller subsets of data. These smaller subsets are provided as inputs to the worker nodes that process the smaller tasks and return the results back to the master node.

The Reduce( ) function collects the results from the worker nodes combines them to produce a final result.

X. METHODOLOGY

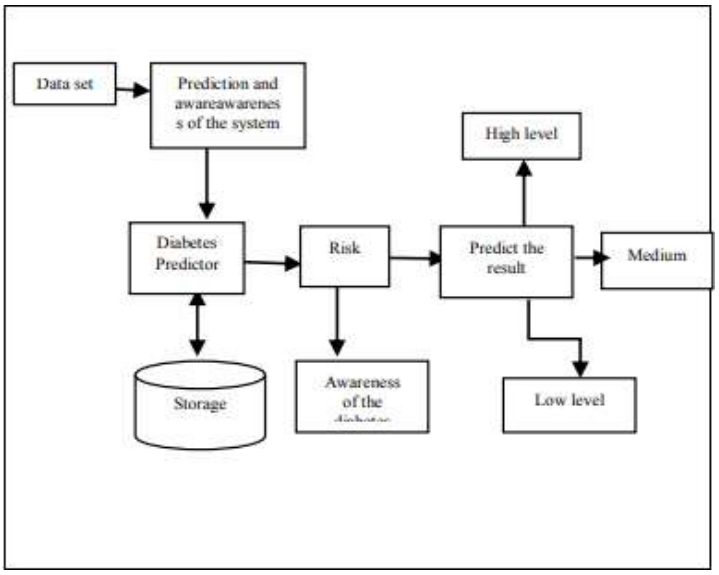

Fig.2 Proposed procedure model

The proposed procedure is summarized in Fig.2. It shows all the steps that will be carried out by the predictive system.

Various algorithms such as SVM, linear regression, decision tree, k-means clustering, etc. will be used in the predict the result step.

- **Pattern discovery:**

  For treatment of diabetes it is important to test the various patterns such as blood glucose levels, insulin dosage, body mass index (BMI), number of times pregnant, age, etc.

  This pattern discovery must include the following:
  - Association rule mining- Association between diabetic type and laboratory results.
  - Clustering- clustering of similar patterns of usage, etc.
  - Classification- Classification of health risk value by the level of patient health condition.

XI. ALGORITHMS

Some of the algorithms used in the system are as follows:

- **Decision Tree**

  A decision tree or a classification tree is a tree in which each internal (non-leaf) node is labeled with an input feature.

  It is a flowchart like structure that works on yes or no questions. It provides two paths of operations for every input.

- **SVM**

  SVM is a classification algorithm which is use to classify between two classes. It uses hyperplanes to classify between the two classes.

  Given a two-class training sample the aim of a support vector machine is to find the best separating hyperplane between the two classes

- **Naïve Bayes'**

  It is a family of classifiers which is based on the Bayes' algorithm. It defines that all the features are independent and unrelated to each other. The state of one feature in a class does not affect the state of another feature.

  Using Bayes theorem posterior probability P(C|X) can be calculated as:

  $P(C|X) = (P(X|C)\ P(C))/P(X)$

  where,
  P(C|X) = target class's posterior probability
  P(X|C) = predictor class's probability
  P(C) = class C's probability being true
  P(X) = predictor's prior probability.

- **Clustering**

  Clustering is a machine learning technique used for grouping the data points. Given a set of datapoints we can classify each data point into a specific group. This helps in distinguishing the datapoints from each other. The data points which are in the same group have similar properties whereas the datapoints which are in different groups have different properties.

  One of the most common and effective clustering algorithms is k-means clustering. In k-means clustering there are *k* number of clusters and all the datapoints are classified into these *k* clusters.

- **Linear regression**

  Linear regression is a method to model a relationship between one dependent variable and one or many independent variable(s). There are two types of linear regression: Simple linear regression and Multiple linear regression.

  In linear regression the data is modeled such that the value of an unknown (dependent) variable can be identified using the known (independent) variable.

X. CONCLUSION

Big data analytics provide better results in an effective and affordable manner. Hadoop provides easy mechanisms for handling and processing of big data. This system will analyze the health records of the patients and help the patient understand their current diabetic status. It will also provide healthcare to the patients to prevent or reduce diabetes.

This system can be used in both the rural as well as urban areas. By employing location aware healthcare service, anyone from rural area can get proper treatment at low cost.

XI. REFERENCES

[1] Thanga Prasad. S, Sangavi. S, Deepa. A, Sairabanu. F, Ragasudha. R, "Diabetic Data Analysis in Big Data With Predictive Method"

[2] Dr Saravana Kumar N M, Eswari T, Sampath P & Lavanya S, 2015. "Predictive Methodology for Diabetic Data Analysis in Big Data"

[3] Deepti Sisodia , Dilip Singh Sisodia, 2018. "Prediction of Diabetes using Classification Algorithms"

[4] Michael G. Kahn, Dijia Huang; Stephen A. Bussmann, Charlene A. Abrams, James C. Beard, "DIABETES DATA ANALYSIS AND INTERPRETATION METHOD"

[5] Riccardo Schiaffini, Claudia Brufani, Beatrice Russo, Danilo Fintini, Antonella Migliaccio, Lia Pecorelli, Carla Bizzarri, Vincenzina Lucidi, Marco Cappa, 2010. "The predictive role of continuous glucose monitoring system"

[6] Abdullah A. Aljumah, Mohammed Gulam Ahamad, Mohammad Khubeb Siddiqui, "Application of data mining: Diabetes health care in young and old patients"

[7] K. Rajesh, V. Sangeetha, "Application of Data Mining Methods and Techniques for Diabetes Diagnosis"

[8] Wullianallur Raghupathi, and Viju Raghupathi, "Big data analytics in healthcare: promise and potential"

[9] Muni kumar N, Manjula R,"Role of Big Data Analytics in Rural Health Care – A Step Towards Svasth Bharath"