ISSN- 2394-5125 VOL 7, ISSUE 17, 2020

A Survey on Diabetic Analysis on Big data and Machine Learning

Dr Akash Saxena
akash27saxena@gmail.com
Mahendra Singh Panwar
(panwarpanwar80@gmail.com)
Compocom Institute Of Information Technology and Management, Jaipur

Abstract— The medical industry contains huge amount of data. It is a large in size and very difficult to predict a disease in traditional methods. Diabetics Mellitus is non communicable diseases. It targets more in middle income countries.We are using Big data analyticsto predict thediabetes data accurately.Big data analytics creates an awareness about the diabetes among the patients. It helps a patient to resolve and care the diseases with Electronic Health Records. Based on dataset,big data predict a upcoming risk in diabetes and provide a treatment accordingly. The machine learning algorithms, namely Support Vector Machine, Naïve Bayes, Random Forest, ANN, Decision tree, Simple CART algorithms are analyzed on diabetesdisease dataset.

Keywords—Big data, Hadoop, MapReduce, HDFS, Machine Learning

I. INTRODUCTION

In this paper we are provide a awareness on diabetes risk among patients and machine learning algorithm gives a accurate results on prediction while comparing with another. Data metrics helps to predict a type of diabetes diseases.

II. BIG DATA

Big data provide an access to manage and store a large amount of data. It performsa less processing compare to the other traditional platforms. Using big data we can access structured, semi structured and unstructured data. It provides a result in low cost. Big data provides better results
without using super computer and high cost [1]. Based on current and past records, big data analyses a future risk and cure diabetes. Big data takes a better decision and strategic move [5]. Big data V's arevolume, velocity, variety, variability, veracity, visualization and value, all the 6 V's of Big Data which has its great importance for further use.

ADVANTAGES OF BIGDATA IN HEALTHCARE

- Predicts a diseases accurately
- Easy monitoring the ElectronicHealth Record.
- Helps to take decision correctly bydoctor.
- Hospital visit can be reduced.
- Doctor can keep track of patientthrough SMART technologies.
- Keep a people healthy.

III. DIABETES

Diabetic Mellitus is a long term metabolic diseases. It affects low and middle income countries [13]. It increases in developing countries like India. In 2030 India will become a "Diabetic Capital". DM contains three types, they are

- DM type 1: Body failure which willnot produce insulin.
- DM type 2: Insulin will not help usto reduce diabetes
- DM type3: Female will be affectedby diabetic during pregnancy level

Due to the diabetes the people will get heart attack, kidney diseases and blindness. Diabetes will occur because of low physical labor. Diabetics will occur lack of physical activity improper food and genetic disorder, lifestyle changes [1]. People affected with diabetic from 8.2%-18.6% in urban area andfrom 2.4%-9.2% in rural area with time frame of 16 years (1992-2008). In 2007 statistic report India Ranks First in the list with 49 million peoples are affected withdiabetes [6]. After 50 years of age Diabetics II is considered as a Rich man disease [7]. Almost half of all death due to high blood

ISSN- 2394-5125 VOL 7, ISSUE 17, 2020

glucose occurs before the age of 70 years. WHO projects diabetes will be the Seventh leading cause of death in 2030 [9]. The symptoms of diabetes like thirsty, feelingtired, itching skin infection and headache [11].

MINIMIZE THE CAUSES OF HIGH BLOOD GLUCOSE

- Focus on physical fitness
- Exercising on regular basis
- Lower intake of salt
- Regular take medicine.
- No alcohol and no smoke

Increase a healthy diet to reduce thelevel of high blood pressure.

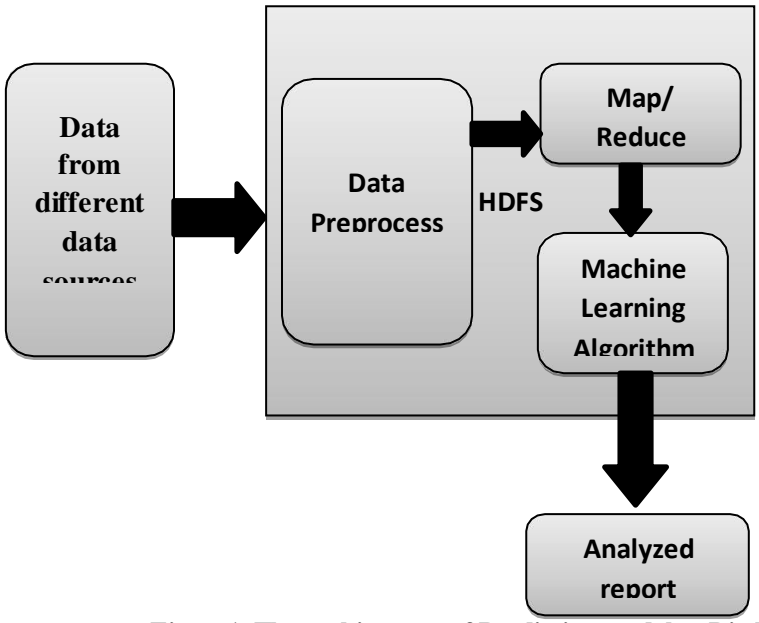

Figure1. The architecture of Prediction modelon Bigdata.

IV. TECHNIQUES

Hadoop:

Hadoop is a open source software [2]. It written in java [7]. Hadoop Map/Reduce executes parallel in system. The map reduce consist two function, map and reduce function. The map function split the datainto rows along with key ad value pairs. Thereduce function is fed a output. The input and output are stored in file system.

Hadoop Distributed File System (HDFS):

Hadoop distributed file system transfera data in distributed system [1]. Store a data in HDFS using STDIN, STDOUT. It contains master/slave architecture. Themaster node give an instruction to data node and slave node will read/write operation and send back to master node. HDFS methods like distributing, storing, processing, file permission and authentication [11].

MAP( ) - STDIN

- Collect a data after preprocessing
- Remove punctuation, hash tag andrepeated data's using stopward().
- After that split() is used to convert a data into word and send to the reducer.

REDUCER() – STDOUT

- Count the records
- Provide the result in statistical report.

ISSN- 2394-5125 VOL 7, ISSUE 17, 2020

PREPROCESSING:

- Convert unstructured data to structureddata
- Finding missing value in text.
- Less missing value will provideaccuracy in better.

SUPPORT VECTOR MACHINE:

Support vector machine is used for classification and regression task. SVMfollows a supervised learning. SVM finds a hyperplane optimally between the various classes of data[4].

NEURAL NETWORK

Neural network takes a input data for processing the information and make the decision.Neural network contain three layers. They are input layer, hidden layerand output layer [14]. Neural network consistset of connected input and output data inwhich each connection has a weight associated with It [9].

NAÏVE BAYES:

Naïve Bayes is a classification method that makes use of Bayesian theorem [9]. It built the probability independently of each attribute that can affect the data. Bayesian method is more popular in medical research because of better prediction [15].

RANDOM FOREST:

Random forest is a classification and regression model. Random forest generates a simple decision trees and decide which labelto return. Random forest gives better results.RF performs prediction and regression [14].

SIMPLE CART ALGORITHM:

Simple Cart method contains a classification and regression analysis. CART stands for Classification and Regression Tree algorithm. It constructs a decision treesusing ancient data [12].

DECISION TREE:

Decision tree is a classification,regression model. Decision tree contains tree structure where each internal node and non-leaf node has a attribute [8]. Each branch represents the outcome of test. The final result holds by leaf node. The decision tree follows a supervised learning. In supervised learning, the input and outputs are known [3].

ASSOCIATED NEURAL NETWORK

Classification is applied in data using associated neural network. ANN includes feedforward neural networks and K-closest neighbor strategy, which is used to sort the trained data [10].

V. DATA METRICSSTATISTICAL ASSESSMENT

Statistical assessment used to predictthe type of diabetes. It contains metrics are accuracy, sensitivity, f-measure, recall [8]. TP,TN,FP and FN denotes true positive as diabetics, true negative as non-diabetics, false-positive non diabetic as diabetic, false- negative diabetic as non-diabetic[10]. Precision is measured as a ratio of true prediction and recall is measured as a ratio of true prediction. Accuracy is calculated correctly among the classes. Kappa is a ratioof prediction related to class. Sensitivity assets positive prediction and Specificity assets negative prediction [13].

Accuracy = (TP+TN)/ (TP+TN+FP+FN)

Sensitivity = TP/(TP+FN)Specificity = TN/(TN+FP)

Error rate = (FP+FN)/(TN+FN+FP+TP)

Precision = TP / (TP+FP)Recall = TP / (TP+FN)

Kappa = (accuracy-expaccuracy) / (1-expaccuracy)

ISSN- 2394-5125 VOL 7, ISSUE 17, 2020

VI. TABLE STRUCTURE

S.NO	AUTHOR	PUBLICATION YEAR	TECHNIQUES	TOOLS	DATASET	ACCURACY/ CONCLUSION
1.	J.Ramsingh, V.Bhuvaneswari	2016	Hadoop Distributed Files System(HDFS), Hadoop Map Reduce, YARN.	R-Hadoop, Python	1300 instance and 19 attributes	Overall people are less aware of diabetes.
2.	Dr Saravana Kumar, Eswari, Sampath, Lavanya	2015	Predictive Analysis System	Hadoop/ Map Reduce	Electronic Health Report, Clinical Report, data from social media, medical journals	Cure the patient with low expenses.
3.	Gauri D. Kalyankar, Shivananda R. Poojara, Nagaraj V. Dharwadkar	2017	Predictive Analysis System	Hadoop Map/Reduce	768 records and 8 attribute 1 class variable.	It concluded Hadoop MapReduce based Machine learning algorithm applied to find a missing value and discover pattern for future risk.
4.	P. Amudha, S. Sivakumari	2018	Map Reduced Based SVM Classification, MR-NB(Naïve Bayes), MR-KNN.	Hadoop Map Reduce, SVM Classification	50000 instances and 12 features	MR-SVM algorithm executes better than other algorithms
5.	Prashant Johri, Tanya Singh, Sanjoy Das, Shipra Anand	2017	Predictive Analysis System	Hadoop, Hive, R	Electronic Health Record, Clinical report, data from social media.	The ramp up of diabetic patient will also deserve money loss of the nation as whole.
6.	J. Ramsingh, V. Bhuveneswari	2018	DAEE- Data Analysis and Evaluation Engine Framework	Hadoop Map/Reduce	Dataset collected from tweets, WhatsApp and survey	Early diagnosis is done to reduce the impact of diabetes affecting money loss and economic development.
7.	J Ramsingh, Dr. V. Bhuvaneswari	2017	HDFS, Hadoop Map/Reduce, YARN.	Hadoop Map Reduce, Pig	3000 instances, 19 attributes	Average of 20-25 percent of people is aware about diabetes.
8.	C.B. Sivaparhipan, N. Karthikeyan, S. Karthik	2018	Statistical assessment system model	Hadoop Map Reduce	Electronic Health Record	The main intension of big data assessment to cure a diabetic with healthcare secure.

ISSN- 2394-5125 VOL 7, ISSUE 17, 2020

9.	N. Yuvaraj, K.R. Sripreetha.	2017	Neural Network, Support Vector Machine, Decision tree, Naïve Bayes, Random forest	Hadoop, Machine learning	75,664 patients, 13 attributes	The Random forest algorithm is a highest accuracy than decision tree and naïve bayes algorithm.
10.	AHMAD ALI AIZUBI	2018	Normalization, Ant bee colony approach, SVM-trained multilayer neural network, Associated neural network.	Hadoop, classification algorithms	Electronic Health Record	Final output is the feature selection and classification algorithm achieved high accuracy, sensitivity, and specific and minimum error rate.
11.	ThangaPrasd, Sangavi S, Deepa A, Sairabanu F, Ragasudha R.	2017	Hadoop HDFS, Big data, Map Reduce, Data warehousing, Decision Tree	Hadoop	Electronic Health Record	Big data analyses provide a good health result.
12.	Ayman Mir, Sudhir	2018	Naïve Bayes, SVM, Random Forest, Simple CART algorithm.	WEKA	768 instances and 9 attributes	SVM provide a better result compared with other Machine Learning algorithms.
13.	P. Suresh Kumar, S. Pranavi	2017	Random Forest, LDA, CART,K-NN, SVM	Hive, R	650 records	Random Forest algorithm provides data more correctly and accurately.
14.	Quan Zou, Kaiyang Qu, Yumei Leo, Dehui Yin, Ying Ju and Hua Tang	2018	Decision tree, Random Forest, Neural networks.	Java, MATLAB	14 and 18 attributes	Random Forest provides better results.
15.	J. Archenna and Dr. E.A.Mary Anita	2017	Healthcare Recommender System Framework, Bayesian Network.	Spark	1000 records	Prediction and recommendation system was studied.

ISSN- 2394-5125 VOL 7, ISSUE 17, 2020

VII. LITERATURE REVIEW

J. Ramsingh, V. Bhuvaneswari et al. described their by research people have less awareness of Diabetic Mellitus. There is a compulsion to make the people to aware of diabetes. Younger generation is the most affected generation of the world. Hadoop Map/Reduce is used to monitor the Diabetesawareness.

Dr. Saravana Kumar, Eswari Sampath,Lavanya et al. briefed that Predictive Analysis Algorithm helps to cure a patient with low cost. Hadoop Map/Reduce is used to predict the diabetes disease. Based on the results they will be given the treatment to conquer the diseases.

Gauri D. Kalyankar, Shivananda R. Poojara,Nagaraj V. Dharwadkar et al. described that Predictive analysis is used to predefine the type of diabetes. Predictive analyses are machine learning and Hadoop Map/Reduce.

According to the severity level of patient thetreatment can be provided.

P. Amudha, S. Sivakumariconcluded et al. that Big data analytics with Hadoop access alarge amount of data. SVM algorithm performing well among the general algorithms. It gives precision, on time and error less output.

Prashant Johri, Tanya Singh, Sanjoy Das,Shipra Anand et al.stated that nation willaffect by financial loss by increasing diabetics patient on day by day. Using predictive analysis we were discover the future level of problem and to improve, it will get down the factors which causes it.

J. Ramsingh, V. Bhuvaneswari et al. described that diabetes among the peoples more in India compare to other countries. Due to lack of fitness and change of habits are the reasons for diabetes in India. Proactive diagnosis should be done to decrease the impact of diabetes affecting health and wealth of country.

J. Ramsingh, Dr.V. Bhuvaneswariet al. concluded that the people have no sufficient awareness of diabetes. Average age of 20-25percent of people is well aware about the diabetes. The older and young group are moderate aware of diabetes.

C. B. Sivaparthipan, N. Karthikeyan, S. Karthiket al. stated that the statistical assessment model assist to assess the outcome of diabetes among patients. The main intention is cure diabetes in healthcare firm. Accuracy and f-measure the result is precision compare to other methods.

N. Yuvaraj, K.R. SriPreethaa et al. mentioned that the machine learning algorithm supports to predict the results correctly. From the machine learningalgorithm the Random Forest algorithmprovides the more accuracy compare to the decision tree and naïve bayes algorithm.

Ahmad Ali AlZubiet al. defined that the SVM Neural Network provides high accuracy, sensitivity and specificity and less error rate.

Thanga Prasad. S, Sangavi S, Deepa. A, Sirabanu. F, Ragasudha. R et al. appreciatedthat the big data analysis enhances the healthcare system to maintain the cost and readmission patient. Big data analysis provides a good health outputs.

Ayma Mir, Sudhir N. Dhage et al. resolved that the machine learning algorithm is used to make a decision on diabetes accurately. SVM algorithms provide a more accuracy compare to other algorithms.

P. Suresh Kumar, S. Pranaviet al. discovered that the different machine learning algorithms are finest predicting algorithms. The Random Forest algorithm is predicting the data more accurate and correctly.

Quan Zou, Kaiyang Qu, Yumei Leo, Dehui Yin, Ying Ju and Hua Tanget al. stated that there is not much variance in Random forest,Decision Tree, Neural Network but RandomForest provides better output than other algorithms.

J.Archenaa and Dr. E.A. Mary Anita et al. briefed that the big data analysis is used to make effective healthcare data and reduce the treatment and unexpected cost.

VIII. DISCUSSION

Big data plays a vital role in healthcare system. It supports to predict diseases accurately. The social media is used to create a awareness of diabetes. The dataset are collected from Electronic Health Report, Clinical report, doctor prescription, diagnostic report, medical images, pharmacyinformation, insurance data and data from social media helps to predict a type of diabetes and future risk. The instances and features like Name, DOB, Occupation, Foodhabits, Food causes diabetes, Food Control diabetes, Symptoms, Plasma, glucose density, serum insulin, blood pressure, history of diabetes, BMI, age, Number of pregnancy are collected from WhatsApp, tweets through social media. These attributes are used to predict the diabetes. The machine learning algorithms are used topredict the disease correctly. From thestudied the SVM and Random forest algorithms provides a better results compare to other algorithms. The lack of physical exercise, improper diet and life style changes are increases the diabetes. The increase of diabetic patient will affect the financial of country. The middle age groupis less aware of diabetes, the senior people and young generation are moderate aware ofdiabetes. The working people have aware ondiabetes than farmer and homemakers. Based on location people have less aware of diabetes. Exercise and healthy diet is more effective to manage diabetes.

ISSN- 2394-5125 VOL 7, ISSUE 17, 2020

IX. CONCLUSION

An Awareness rate is very less inpeoples mind about the diabetes. Big data analysis helps to cure and care the patient with economy cost. We can avoid the impact of diabetes in future by doing proactive diagnosis to build the country in economy mode with less risk. In this paper the prediction model and awareness on diabetes was studied.

References

[1] J. Ramsingh, V. Bhuvaneswari, "Data Analytic on Diabetic awareness with Hadoop Streaming using Map Reduce in Python", IEEE International Conference on Advances in Computer Applications (ICACA), 2016.

[2] Dr. Saravana kumar N M, Eswari T, Sampath P &Lavanya S, "Predictive Methodology for Diabetic Data Analysis in Big Data", 2nd International Symposium on Big Data and Cloud Computing (ISBCC), 2015.

[3] Gauri D. Kalyankar, Shivananda R. Poojara, Nagaraj V. Dharwadkar, "Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop", International conferences on I-SMAC (IoT in Social, Mobile, Analytics andCloud),2017.

[4]P. Amudha, S. Sivakumari, "Big data Analytics Using Support Vector Machine", International Conference on Soft-computing and Network Security (ICSNS), 2018

[5] Prashant Johri, Tanya Singh, Sanjoy Das, Shipra Anand, "Vitality of Big Data Analytics in Healthcare Department", IEEE International Conference on InfocomTechnologies and Unnned System (ICTUS), 2017.

[6]J. Ramsingh, V. Bhuvaneswari, "A Big Data framework to analyze risk factors of diabetes outbreak in Indian population usinga MapReduce algorithm", Proceedings of the Second International Conference onIntelligent Computing and Control Systems (ICICCS), 2018.

[7] J. Ramsingh, Dr.V. Bhuvaneswari, "Social Networking Data Analytics On Diabetes Using Pig –A BigData Tool", International journal of Emering Trends & Technology in Computer Science (IJETTCS), Volume 6, Issue 3, May-June 2017.

[8] C. B. Sivaparthipan, N. Karthikeyan, S. Karthik, "Designing statistical assessment healthcare information system for diabetics analysis using big data", Springer, 2018.

[9] N. Yuvaraj, K.R. SriPreethaa, "Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster", Springer, 2017.

[10] Ahmad Ali AlZubi, "Big data analytic diabetics using map reduce andclassification techniques", Springer, 2018.

[11]Thanga Prasad. S, Sangavi S, Deepa. A, Sirabanu. F, Ragasudha. R, "Diabetic Data Analysis In Big Data With Predictive Method", 2017.

[12]Ayma Mir, Sudhir N. Dhage, "Diabetes Disease Prediction using Machine Learning on Big Data of Healthcare", Fourth International Conference on ComputingCommunication Control and Automation (ICCUBEA), 2018

[13] P. Suresh Kumar, S. Pranavi, "PerformanceAnalysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics", International Conference onInfocom Technologies and Unnned System (ICTUS), 2017.

[14] Quan Zou, Kaiyang Qu, Yumei Leo, Dehui Yin, Ying Ju and Hua Tang, "Predicting Diabetes Mellitus With Machine Learning Techniques", Frontiers in Genetics, 2018.

[15]J.Archenaa and Dr. E.A. Mary Anita, "Health Recommender System using Big data analytics", Journal of Management Science and Business Intelligence, 2017.