International Journal of Advanced Science and Technology
Vol. 29, No. 1, (2020), pp. 970 - 978

# Predictive Method for Diabetic Medical Records Data Analysis Using Machine Learning and Hadoop

**[1]T. Ajay kumar, [2]Dr. B. Sankara Babu**
[1]PG Scholor, [2]Professor
[1,2]Department of Computer Science and Engineering
[1,2]Gokaraju Rangaraju Institute of Engineering and Technology
[1]ajaykumarthallapalli55@gmail.com, [2]sankarababu.b@griet.ac.in

***Abstract***

Presently days from social insurance businesses huge volume of information is creating. It is important to gather, store and procedure this information to find information from it and use it to take huge choices. Diabetic Mellitus (DM) is from the Non Communicable Diseases (NCD), and loads of individuals are experiencing it. Presently days, for creating nations, for example, India, DM has become a major medical problem. The DM is one of the basic diseases which has long haul difficulties related with it and furthermore pursues with different medical issues. With the assistance of innovation, it is important to fabricate a framework that store and break down the diabetic information and predict potential dangers likewise. Predictive investigation is a strategy that incorporates different information mining systems, ML algorithms and measurements those utilization present and past informational collections to pick up understanding and predict future dangers. In this work "machine learning calculation in Hadoop MapReduce environment are executed for Pima Indian diabetes informational index to discover missing qualities in it and to find designs from it. This work will have the option to predict kinds of diabetes are far reaching, related future dangers and as per the hazard level of patient the sort of treatment can be given".

***Keywords:*** Healthcare industry, Hadoop, MapReduce, ML, Predictive Analysis

## Introduction

Predictive examination which is help to human services associations to assess information on the past conduct and predict probability of future conduct to empower better choices and results of their patient[1]. Predictive models can settle on human choices increasingly viable and profoundly computerize a whole basic leadership process. It progressively, predictive examination utilizes information from the IOT to improve wellbeing and execution of patient results. Medicinal services industry faces numerous provokes that make us to realize the significance to build up the information investigation of the diabetes mellitus.

BigData is developing as an answer for the issues related with enormous measure of information. The huge measure of information produced would now be able to be utilized so as to give an internal perspective on what is truly occurring and recognize the developing patterns. Large Data can likewise be utilized in the field of medicinal services so as to make the framework increasingly powerful. Huge Data alludes to the enormous measure of information which might be organized or unstructured and can't be handled utilizing a social database model. Unstructured



ISSN: 2005-4238 IJAST

information alludes to the information that can't be put away in a specific line and section design. Huge Data additionally goes past the preparing limit of the ordinary database frameworks [1].

Medicinal services division information is developing past the managing limit of the human services associations and is relied upon to increment in the coming years. The greater part of the Healthcare information is frequently unstructured, and lives in imaging frameworks, medicinal remedy notes, protection claims information, Electronic Patient Record and so on. Joining organized and unstructured information for cutting edge investigation is basic to improve human services results. Due to information that are secluded in unique or contradictory arrangements or because of the need preparing capacity to load and inquiry huge datasets in an auspicious manner the Healthcare associations are not in a situation to use the advantages of the enormous arrangement of Healthcare information. With the assistance of cutting edge processing and versatility at a moderately minimal effort. Huge information arrangements regularly accompany set of creative information the board arrangements and systematic devices, when adequately executed can change the human services results [2].

The Healthcare information is developing quickly from an interior just as sources like cell phones mostly and other lot of ways we have in the society. The other therapeutic information structures like imaging, sensor perusing is additionally fuelling to the need of Big Data answers for deal with this enormous information accessible in the social insurance associations. The medicinal services industry needs to chip away at prediction, aversion and personalisation to improve their results. Diabetes otherwise called Diabetes Millet's is a malady that outcome in an excess of sugar in the blood, or high blood glucose. Diabetes of numerous kinds can prompt difficulties in numerous pieces of the body and increment the danger of passing on rashly. Commonness is expanding around the world, especially in low and center salary nations. Access to quality human services for these individuals makes diabetes a hazardous ailment [3].

By utilizing BigData, “it is conceivable to predict the hazard required for the patient utilizing his/her past therapeutic history. Human services suppliers are digitizing their databases which clear route for the development of Big Data investigation. Utilizing calculations like Naive-Bayes and k-implies the prediction of hazard included should be possible. The prediction would empower the medicinal services suppliers to rapidly evaluate the patient's circumstance and furthermore give an understanding into patient's future if the flow circumstance wins as diabetes is an ailment which influences the patient. The hazard included can be evaluated by specialists and can base their treatment and furthermore the patient can be exhorted for way of life changes” [4].

## Types of Diabetes

***Type 1 Diabetes*** is called insulin-subordinate diabetes mellitus (IDDM) or adolescent beginning diabetes. Type1 happens for younger people who are below 30 years.

***Type 2 Diabetes*** is called non-NIDDM or grown-up beginning diabetes. Risk factors for Type 2 diabetes incorporate more established age, fatness, and family people.



ISSN: 2005-4238 IJAST

***Gestational Diabetes*** "is the third primary structure and happens when pregnant ladies without a past history of diabetes build up a high blood glucose level".

***Congenital Diabetes*** happens in human because of hereditary imperfections of insulin discharge, cystic fibrosis-related diabetes, and high portions of glucocorticoids prompts steroid diabetes.

## Related Work

The survey on earlier work gives numerous outcomes on examination of social insurance information which was done by various strategies, procedures. Numerous scientists have created and actualized different investigation and prediction models utilizing various information mining, information the executives and Hadoop strategies or mix of these systems.

Eswari et al. (2015) Through this examination the created framework can be predict diabetic sort's common and difficulties related with it. Based on such investigation, the framework can fix the patient by giving proper treatment as right on time as could be expected under the circumstances. The framework depends on Hadoop thus it is reasonable to any medicinal services association [1].

V. H. Bhat et al. (2009) proposed a methodology of combination of relapse, order, hereditary and "neural system which manages the missing qualities just as exception esteems in the diabetic informational index and supplanted the missing qualities by the relating property space. For prediction they utilized old style neural system model and applied it on the preprocessed informational collection" [2].

Sabibullah M. et al. (2013) "built up the prediction model dependent on delicate processing to locate the aggregated dangers of diabetic patients. They have utilized hereditary calculation for the experimentation on constant medicinal informational index. From the aftereffects of the tests the hazard level of patient and likewise the danger of heart stroke can be predicted. The created framework will assist the specialist with diagnosing the patient effectively" [4].

K. Rajesh and V. Sangeetha (2012) utilized characterization strategy to discover valuable data from diabetic informational collection. They utilized C4.5 calculation to discover patterns from the informational collection likewise for proficient order. They utilized Pima-Indian-Diabetes Data set for experimentation. While playing out the grouping creators didn't consider missing qualities in informational index [5].

Vaishnav and Dr. Patel (2015) checked on the various techniques for taking care of missing information. They explored various techniques, for example, K-Means, KNN, order and so on utilized for the missing qualities ascription and furthermore contrasted and their points of interest and drawbacks [7].

Heaps of research work utilizing various systems like information mining, Weka, Hadoop and its biological systems and so forth effectively is accomplished for investigating the human services information and growing great examination models [15]. For the investigation of diabetic information numerous creators favoured decision tree for the grouping, rules age, pattern acknowledgment and so forth [12].


ISSN: 2005-4238 IJAST

## Implementation Methodology

The predictive investigation framework design includes different stages like information assortment, missing qualities attribution, pattern revelation, and pattern coordinating and result examination. Figure1 portrays generally speaking engineering of proposed work.

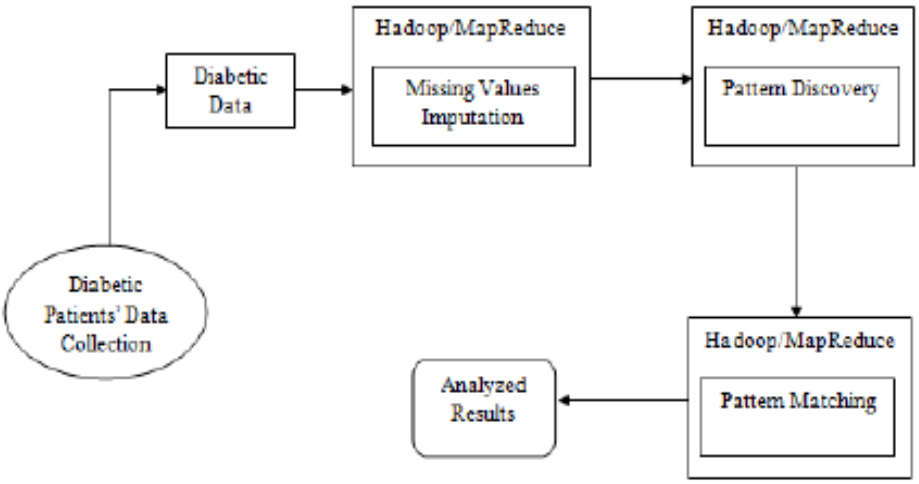


Fig. 1 Predictive Analysis Architecture

## Data Collection

The crude diabetic large information or informational index is given as contribution to the framework. The unstructured voluminous information can be acquired from different EHR/PHR, Clinical frameworks and outside sources (government sources, labs, drug stores, insurance agencies and so forth.), in different arrangements (level records, .csv, tables, ASCII/content, and so on.) and living at different areas [8].

## MapReduce Process

The cutting edge human services supplier is furnished with an Electronic Healthcare Records and a robotization device has enabledenormous measures of information age. This gathered information can be utilized to actualize large information examination. Aside from fundamental information being gathered present day frameworks additionally gather complex information from clinical preliminaries, explore and demonstrative tests.

Map Reduce is a programming model for parallel preparing of huge volume of information. This information can be anything besides it is explicitly intended to process the rundown of information. The primary topic idea of Map Reduce is to change rundown of information to rundown of yield information. In the event that the info information isn't coherent, at that point it is hard to see huge information input sets. All things considered we need a model that can form input information into comprehensible, justifiable yield list. As of late a few nontrivial Map Reduce calculations have risen, from registering the width of a diagram to actualizing the Expectation–Maximization (EM) calculation to bunch monstrous informational collections. Every one of these calculations gives a few bits of knowledge into what should be possible in a Map Reduce structure. In any case, there is an absence of thorough calculation examinations of


ISSN: 2005-4238 IJAST

the issues in question. In this work the creators have exhibiting a conventional model of calculation for Map Reduce and contrast it with the well known PRAM model [8] [9].

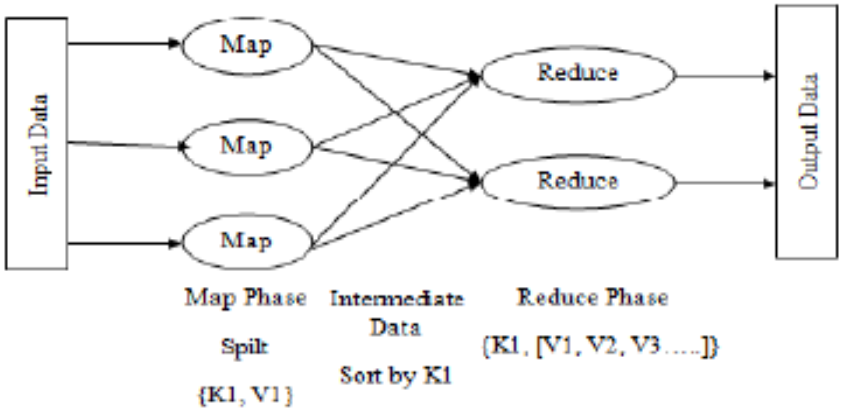


Fig. 2 MapReduce Architecture

The Map Reduce structure maps the <key, value> matches and creates a lot of yield sets of combining every one of the sets related halfway key, that is, Map Reduce works solely on <key, value> pair. The two activities in Map Reduce for example map and reduce originate from utilitarian programming dialects which pass works as contentions to different capacities. Map Reduce structure parts the information into fragments. These portions are then passed to various machines/groups for calculation. Map content, which is composed by client, runs on each machine to process segment of set of moderate <key, value> pair. The sets with same key are assembled by the Map Reduce library and passed reduce work for total. Reduce content, which is likewise composed by client, takes assortment of transitional key I alongside their set values for that keys and consolidations them as per the client indicated content into a littler arrangement of values. Subsequently, the reduce work totals the values of the assortment of middle of the road <key, value> sets having a similar key. The diagrammatic portrayal of work stream of Map Reduce is as appeared in Figure 2[10] [11].

## Missing Value Imputation

Missing information make different issues when investigating and handling information in the database. “Missing information is an issue related with information mining research. Missing information happens because of no trait related for any case, or the values are not important or the values are not gathered appropriately at the hour of information has been oppressed. The missing values in a database can influence the precision and execution of the classifier which brings about trouble to extricate the significant data, loss of productivity. It might be exceptionally hard to get the information quality mining results from the fragmented informational collections, subsequently this missing holes need to the treated” [7][13].

In the Pima-Indian-diabetes informational index there is different trait “which have invalid values. It is unimaginable that any patient who has 0 circulatory strain or 0 plasma glucose level in his body. Subsequently it is important to ascribe missing values for good arrangement results or it will cause wrong characterization results”. The missing values can be credited by utilizing


ISSN: 2005-4238 IJAST

various systems, for example, arrangement bunching. We utilized order based method to supplant missing values with property mean.

## Decision Tree

For prediction purposes and classifications DT is one of the popular and powerful tools. Decision rules are nothing but rules that are interpreted by humans to make well informed decisions. It returns actionable knowledge that can be used by humans. There are certain key requirements of DT. First, it needs an expressible attribute values that are clearly specified. For instance, values like code, mild, hot are specified for an attribute related to weather. Second, there needs to clearly defined target classes may be multi-class or Boolean. The learning model of DT needs sufficient training data.

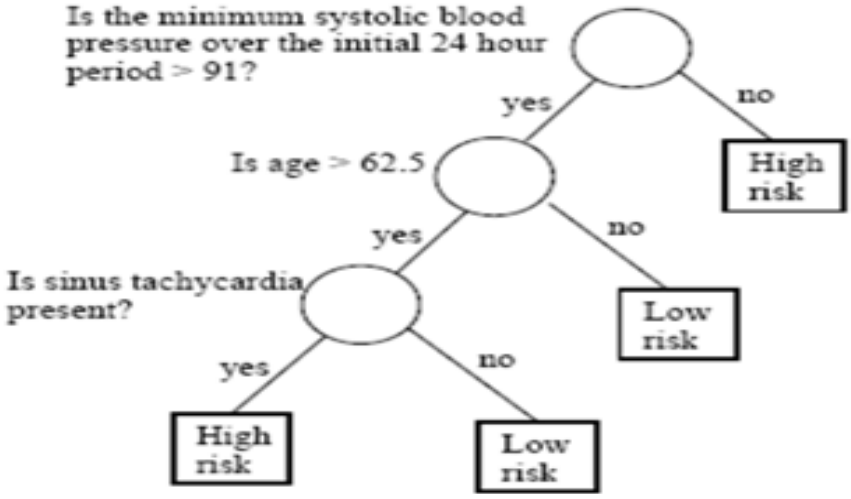


Figure 3: Shows decision tree for a healthcare dataset

There are three rules in the DT. The first rule is related to blood pressure of human. The second rule is related to age of the person while the third rule is related to the presence of sinus tachycardia. The target classes include low risk and high risk. Every condition has two possibilities like yes and no. This algorithm works effortlessly for both categorical and continuous data. The given population is divided into multiple sets. It computes entropy of every attribute. The attributes with minimum entropy and maximum information gain are used to split data for generating decisions. The entropy and gain are computed as in Eq. (1) and Eq. (2).

$$Entropy(S) = \sum_{i=0}^{n} -\text{pi} \log 2 \text{ pi} \quad (1)$$

$$Gain(\text{S}, \text{A}) = Entropy(S) - \sum_{v \in Values\ (A)} \frac{|\text{Sv}|}{|\text{S}|} Entropy(Sv) \quad (2)$$

At last here the inward hubs contain the attributes while the branches speak to the consequence of each test on every hub. DT is broadly utilized for grouping purposes since it needn't bother with much information in the field or setting the parameters for it to work.



ISSN: 2005-4238 IJAST

## Predictive Pattern Matching

At whatever point the warehoused dataset was sent to Hadoop framework, promptly the map reduce task is performed. In mapping stage, "the Master Node parts huge data into littler tasks for various Worker Nodes. Figure 2. Conveys the precise activity of predictive pattern coordinating framework. The Master node is one comprises of Name Node (NN) and Job Tracker (JT), which consistently utilizes the map and reduce task. The Worker Node or Slave Node gets the request from the Master Node, process the pattern coordinating task for diabetes data with the assistance of Data Node – Same Machine (DN) and Task Tracker (TT). The predictive coordinating is the way toward contrasting the investigated edge value and the acquired value. In the event that the pattern coordinating procedure was finished by all Worker Nodes dependent on the prerequisite, it was put away in middle plates. This procedure is known as nearby compose. On the off chance that the reduce task was started by Master Node, all other distributed Worker Nodes will peruse the prepared data from middle of the road plates. In light of the inquiry got from Client through Master Node, the reduce task will be acted in Worker Node". The outcomes acquired from the reduce stage will be disseminated in different servers.

## Results Analysis

***Input Data set:*** So as to find "patterns from the data set, we have given pima diabetes data set to C4.5 calculation dependent on decision tree. Before offering contribution to C4.5 calculation, we applied calculation 1 on the data set for the missing value attribution. After that as per the standard scope of each trait all the numeric values in data set are changed over into string with the goal that it ought to be anything but difficult to translate the yield of C4.5 algorithm".

***Selection of Attributes:*** There are four attributes chose to reduce intricacy of results. Qualities are chosen based on following criteria.
1) Attributes that are useful to analyze Diabetic patient and risk related with the patient.
2) The traits having less number of missing values in it to show signs of improvement exactness.

The means engaged with data handling are Information Extraction, Feature Selection and Predictive Modelling Information Extraction: Here patients medicinal services records are gathered from different sources in clinics and afterward sorted out as a Structured-EHR. Highlight Selection: From the gathered patient's record, the most significant highlights required for diabetic prediction and displaying is separated.

***Predictive Modelling:*** We have built a predictive model to predict whether the patient have diabetes or not.

From the examination of the considerable number of patterns plainly, "when plasma is high the majority of the occasions understanding is of diabetic class and when plasma is low the greater part of the occasions persistent is of non diabetic class. At the point when plasma is medium both diabetic and non from the above diagram we can compose patterns that will be useful to characterize the patients into diabetic or non diabetic class and furthermore ready to predict hazard level of every patient".


ISSN: 2005-4238 IJAST

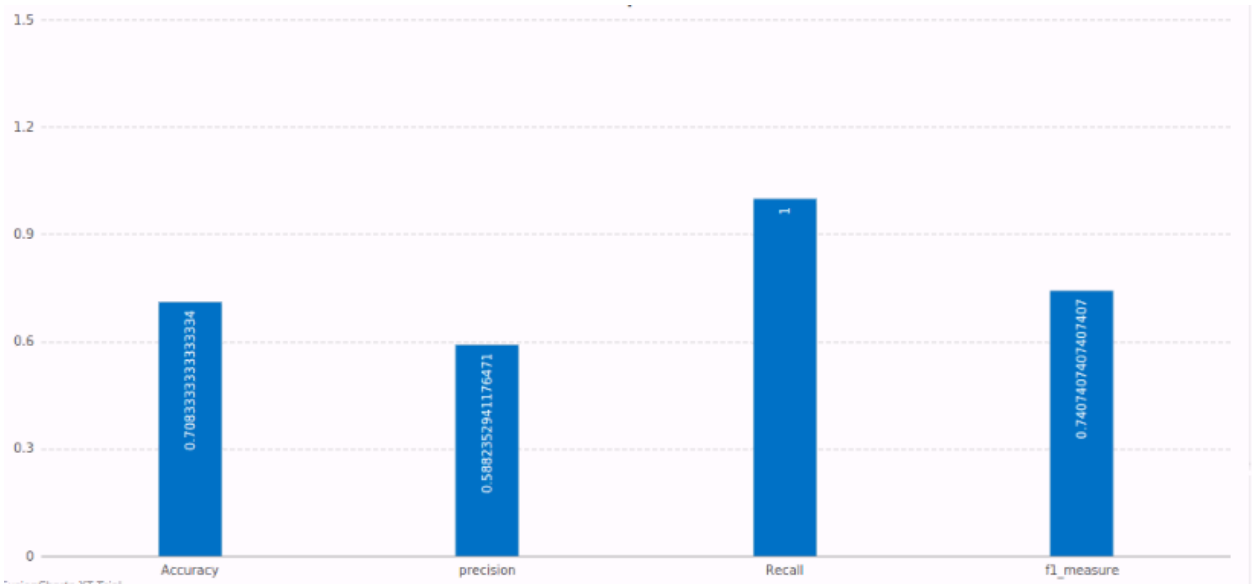


Fig: Diabetics prediction based on decision tree algorithm

## Conclusion and Future Work

In this study paper principally centered around predict Diabetes mellitus by utilizing enormous data examination. As indicated by this examination the majority of the paper covers the algorithm Hadoop MapReduce environment to predict the diabetes types common and difficulties of patient with it. In every one of the paper they utilized distinctive dataset for examination. For the estimation they utilized decision tree calculation. Thus, the outcome is very great. It is conceivable to future improve the diabetes mellitus to utilize any machine learning calculation/for proficiency of the prediction. We will utilize ANN in light of the fact that it shows better outcome when contrast and SVM.

## References

[1] N. W. T. Marty Kohn, "Transforming Healthcare through Big Data," IEEE, vol. 18, no. 6, pp. 21-45, 2015.
[2] P. H. Marco Viceconti, "Big Data, Big Knowledge: Big Data for Personalized Healthcare," IEEE Journal, vol. 16, no. 4, pp. 1209-1215, 2015.
[3] A. K. Chaitanya Kaul, "Comparative Study on Healthcare Prediction Systems using Big Data," IEEE Journal, vol. 5, no. 4, pp. 1-7, 2015.
[4] K. K. a. R. Rani, "Managing Data in Healthcare Information Systems," IEEE Journal, vol. 48, no. 3, pp. 52-59, 2015.
[5] O. Shenyin, "Big Data for Modern Industry: Challenges and Trends," IEEE Journal, vol. 103, no. 2, pp. 143-146, 2015.
[6] G. C. Tung Chee-chen, "A Method for Calculating the probability of Diabetes based on large data," IEEE Journal, vol. 27, no. 9, pp. 13-17, 2014.
[7] Rajnik L. Vaishnav , Dr. K. M. Patel, "Analysis of Various Techniques to Handling Missing Value in Data set, International Journal of Innovative and Emerging Research in Engineering"
Volume 2, Issue 2, 2015



ISSN: 2005-4238 IJAST

[8] Wei Dai, Wei Ji, "A MapReduce Implementation of C4.5 Decision Tree Algorithm," IJDTA Vol.7, No.1 (2014), pp.49-60
[9] "Machine Learning tutorials and examples https://www.toptal.com/machine-learning/machinelearningtheory-an-introductory-primer".
[10] Anish Talwar, Yogesh Kumar, "Machine Learning: An artificial intelligence methodology," IJECS ISSN:2319-7242 Volume 2 Issue 12, Dec.2013
[11] Brona Brejova, Tomas Vina, Ming Li, "Pattern Discovery: Methods and Software," Technical Report CS-2000-22, Dept. of Computer Science, University of Waterloo.
[12] Dr.Rajni Jain, "Rule Generation Using Decision Trees," IASRI.
[13] Md. Geaur Rahman, Md. Zahidul Islam, "A Decision Tree-based Missing Value Imputation Technique for Data Pre-processing," "Proceedings of the 9-th Australasian Data Mining Conference (AusDM'11)", Ballarat, Australia.
[14] Gauri D.Kalyankar, Shivananda R Poojara, N V Dharwadkar,"Weblog Analysis Using Hadoop" National Research Symposium on Computing - RSC 2016, ISBN: 978-81-931456-1-8, Dec. 19-20, 2016.
[15] Sadhana, Savitha Shetty, "Analysis of Diabetic Data Set Using Hive and R," IJETAE, vol 4(7), 2014.
[16] A.Ravishankar Rao, Atul Chhabra, Rajarshi Das, Vikash Ruhil, "A framework for analyzing publicly available healthcare data," IEEE 2015.



ISSN: 2005-4238 IJAST