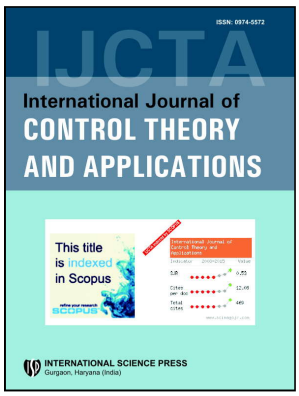


International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 10 • Number 14 • 2017

Prediction model of Diabetes Drug Using Hive and R

Nida Afreen Rizvi[1], Anjana Pandey[2], Ratish Agrawal[2] and Mahesh Pawar[2]

[1] Department of School of Information Technology RGPV, Bhopal, MP, India, Email: Mtech.nida271990@gmail.com
[2] Department of Information Technology UIT RGPV, Bhopal, MP, India, Email: anjanapandey@rgtu.net, ratishy@rgtu.net, mkpawar24y@gmail.com

Abstract: Today healthcare industries are moving towards analysis and processing huge health records. Due to the large un-structured behaviour of Big Data form health industry, it's required to analyse it as a structure manner while converting it into structure form.. Diabetic Mellitus (DM) is a major health disease in many developing countries such as India. In this paper, we are analysing the diabetic dataset by analysis algorithm in Hadoop/Map Reduce environment to develop a prediction model by which we can predict the diabetic complications according to the diabetic dataset. By prediction model we can predict the drug which is useful for the treatment, for analysis purpose we are uses hive on top of the hadoop/map-reduce and for generating graphs we use R .

Keywords: Hadoop , Diabetes mellitus[11], Hive, R.

1. INTRODUCTION

We are working with medical database, these medical data is huge i.e. big data so first we should understand what big data is, and what is importance of big data in health care. Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, data acuration, search, sharing, storage, transfer, visualization, querying, updating and information privacy. The term Bigdata is simply used for predictive analytics and many other advance techniques to extract data from different huge multiple datasets. Due to different advanced technique we can easily or confidently extract data which may lead in decision making and better decision may lead to good result or greater efficienct.

Next topic to study is hadoop, we are working with hadoop our data is loaded in hadoop hdfs and hive query runs through hadoop map reduce. Hadoop [1] is a free open source java based framework which is used to store large amount of data in hdfs and also used for process these large dataset. Hadoop works on master slave architecture in which many commodity hardware machine are connected to each other to form a cluster in which one machine is master and remaining machine are the slaves machine in which the data is stored. Hadoop uses HDFS for storing purpose.

For processing the large data which is stored in HDFS hadoop uses mapreduce algorithm in which it will launch the different task on different nodes by launching different mapper and to combine all these mapper

output it uses reducer to combine all the map output to combine in a single output. Mapreduce is capable to handle huge amount of data [2] by launching multiple mapper in parallel fashion by which the performance is high. Our data is run on top of hadoop through hive we should have a glance on hive.Apache Hive is a data warehouse infrastructure which runs on the top of Hadoop for providing data summarization, query, and analysis [4]. Hive is developed by Facebook; Apache hive is very useful in many analysis purpose and it uses by many companies for analysis purpose because it is very easy and required only one line of query and the hive shell in convert the simple query into complex mapreduce code so the mapreduce algorithm can run the code and give a result. Hive is used with many hadoop ecosystem like hbase, zookeeper ,sqoop. Apache hive can support many databases , hive internally uses derby database and for external database its support oracle, mysql, hbase, etc. After working with hive queries we have generated graphica; represtation of data though R.Executing R in the context of a MapReduce job provides various kind of graph , plot, pie chart through which we can easily analyse huge datasets. Problems that fit nicely into this model include "pleasingly parallel" scenarios. Here's a simple use case: Scoring a dataset against a model built in R. R provides great user interface by which we can easily load any type of datasets in R and R supports many packages by which we can easily analyse the dataset by ploting different graph, plots. R can easily integrate with hadoop and R also uses HDFS data and operate mapreduce algorithm to generate their result in a different analysis mode.This simulates the "apply" family of operators in R. Other tasks such as quantiles, crosstabs, summaries, data transformations and stochastic calculations (like Monte Carlo simulations) fit well within this paradigm.

2. RELATED WORK

KiyanaZolfaghar,AnkurTeredasai, enjutiBasuRai [7] real world medical data is distributed in heterogeneous form and to fetch information from that data is a difficult task so for that purpose techniques are required to retrieve information from that data. Hadoop provides an open source framework to access Big Data. In this paper Big Data solution for predicting risk for readmission congestive heart failure is presented in which for this purpose multicare health system's (MHS)[5] records of CHF is used to predicting risk of CHF. In this process national in patient datasets (NIS)[6] are used to check different parameters and MAHOUT [12]framework is used to categorize that data.

Gopal A. Tathe, Pratik S. Patil, Sangram C. Parle [8] in medicare system traditional systems are not centralized or digitized, thus there is direct interaction between doctors, pharmaceutical stores, and manufacturing company. Thus most of the time doctors suggest medicine which is not economical. To resolve this problem a centralized decision making system is presented by the author. In this system the interaction between doctors, pharmaceutical stores, and manufacturing companies are performed through a centralized and digitized server, in this server four nodes are there doctors node, pharmaceutical store node, manufacturing company node and decision server node. In these node decision server node work as a master node, and monitor all the action that perform by the other node such as which medicine suggested for which disease. To implement this system Hadoop platform is used which provides fast analytical access to all the action performed by the system.

3. PROBLEM METHODOLOGY

In the existing system, we are loaded the raw data into the HDFS and then loaded it into a hive server and then we are start analyzing it with different hive queries.

In existing system it uses only distinct function in the query to analyse the dataset based on different attributes. In the exsisting system we are on analyse the diabetic data but not predict or take decision on the analysis result. The main problem in the existing system is we cannot predict or recommend any decision based on analysis result that by in proposed system we are creating a prediction model which is able to predict or to take a decision.

Our data contains two dataset, first is diabetic dataset[9] which we can analyse and create a second dataset from the analysis of diabetic dataset. The second dataset is prediction dataset [10]through which we can predict the drug.

3.1. Proposed system

1. We first download the pima Indian diabetic dataset from National Institute of Diabetes and Digestive and Kidney Diseases.
2. These dataset contains 8 attributes and total of 768 record in the dataset.
3. Then we loaded the dataset into hdfs and then into hive server through which we can run multiple distinct queries to analyse the dataset.
4. We can take the analysis result from existing system and according to the result we can make a prediction model with 5 attributes.
5. After the query execution we are loaded the prediction table into R to generate a graph according to the different packages based on attributes.

3.2. Data flow Diagram of Proposed System

In a proposed system configuration we are integrating Hive server on top of the hadoop 1.1.2 version by which we can analyse the whole datasets with the help of simple queries which run on hive server. Both the dataset are

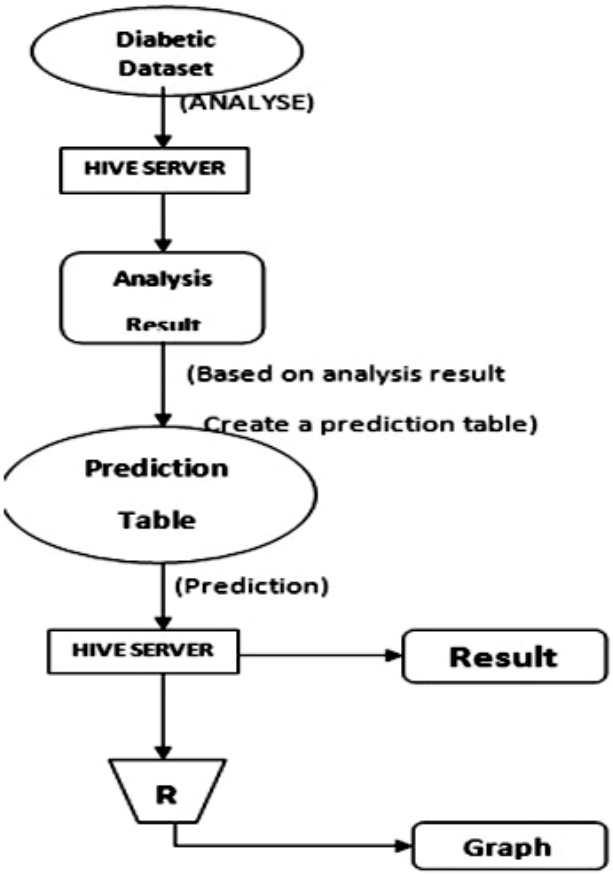


Figure 1: Data flow diagram of proposed system

applied to the Hive server , then we execute different queries on hive server to analyse or predict the output , the output of the hive is our actual output which we can take seperatly , But analysis is easier when we seen it in graphical options , for that reason we are integration R with hadoop and hive which generate graphs for different queries by which we can easily understand or analyse the datasets which helps us to predict the drug.

5. EXPERIMENTAL RESULTS

As it was mentioned earlier Hive and R are used for the purpose of prediction. In Hive the data set should be first loaded to it, hence the diabetic data set is first loaded to it. The raw data which is loaded to hive is just a comma separated file. fig 2 shows the snapshot of the raw data which is loaded to Hive.

Once the file is loaded to Hive we can perform some prediction on the data set. We can predict the drug according to the different parameters which takes in the datasets. We will take huge number of records belonging to different class such as age , diabetes type, serum insuline and number of times pregnant.

Figure 2: Prediction dataset loaded in HDFS

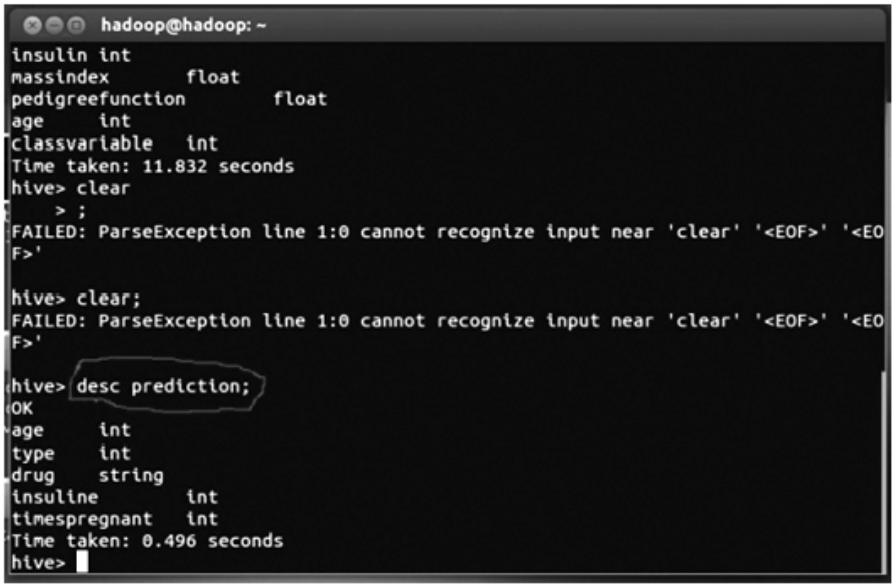


Figure 3: description of prediction table

Fig-3 shows the description of prediction table which shows all the attributes and their datatypes which comes in the prediction table. After loading the datasets we are start analysing the attributes through which we can predict drug.

After loading the datasets in to hive server we are execute different query to predict drug. Fig 4 shows the query to predict drug on the basis of age and type attributes. The query contains simple select statement by which we can select the drug from prediction table.

We can predict the drug information based on these four attributes and combination of these attributes. We perform some prediction of drug which will written in the tabular form This information is tabulated in table 1.

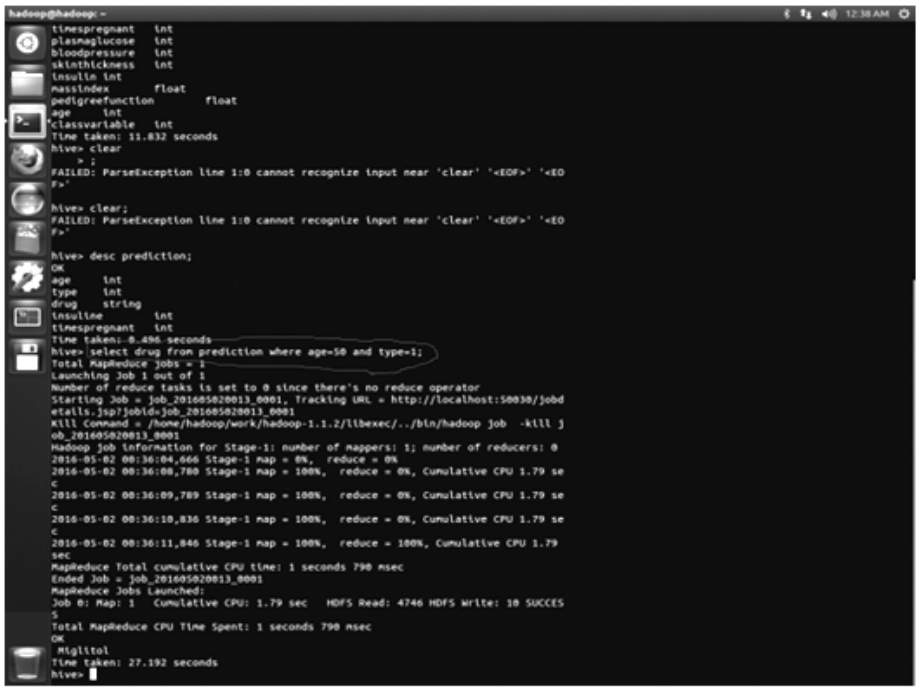
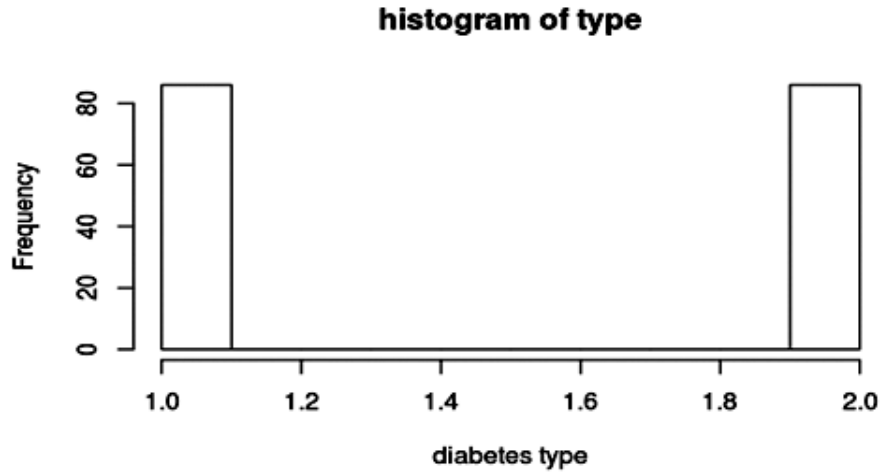


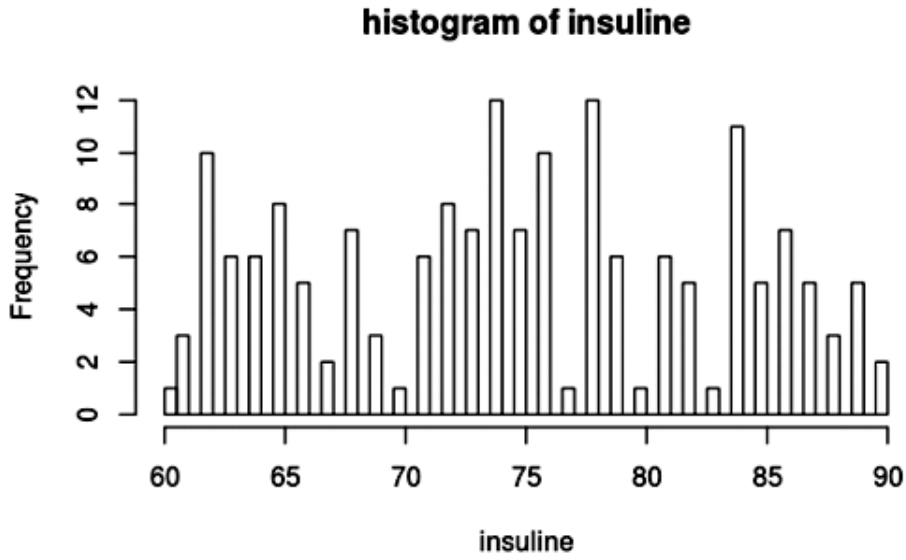
Figure 4: Query to predict drug

Table 1
Drug Prediction for different attributes

Attribute	*Time taken in hive (seconds)*
Age	59.454
Age and diabetes type	39.550
Age and Serum Insulin	32.886
Age and times pregnant	34.406
Age, type and serum insulin	36.302
Age,type and times pregnant	41.218
Type and serum insulin	43.841
Type and timespregnant	35.636
Type, serum insulin and timespregnant	36.225

Fig 5 shows the snapshot of the graphs generated using R. Both the graphs are plotted as follows, x axis displays the different values for the attribute and y axis shows the frequency distribution of a quantitative variable.


histogram of type
Frequency
0
20
40
60
80
1.0
1.2
1.4
1.6
1.8
2.0
diabetes type


histogram of insuline
Frequency
0
2
4
6
8
10
12
60
65
70
75
80
85
90
insuline


histogram of Age
Frequency
0
10
20
30
40
0
20
40
60
80
100
Age

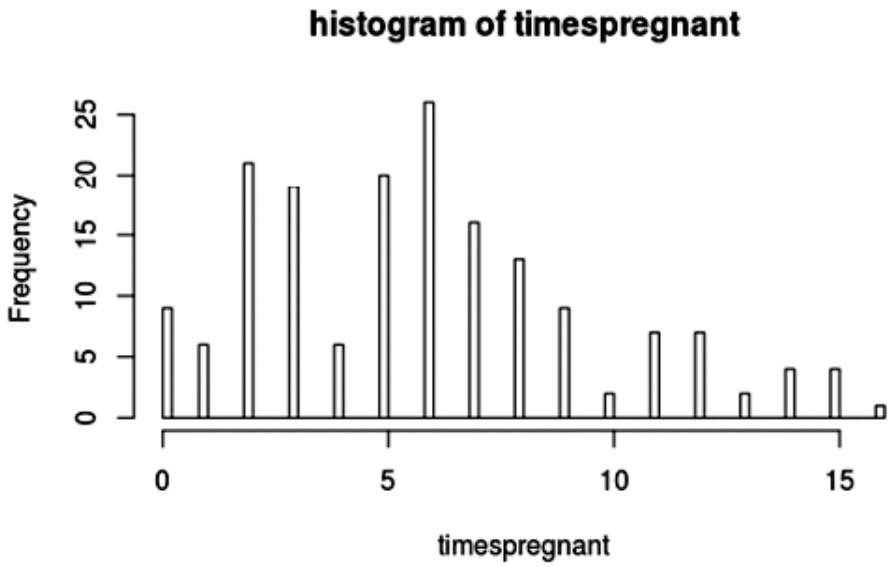


Figure 5: Histogram diagram of different attributes using R.

A histogram consists of parallel vertical bars that graphically shows the frequency distribution of a quantitative variable. The area of each bar is equal to the frequency of items found in each class. In the data set prediction, the histogram of the different attribute variable is a collection of parallel vertical bars showing the number of attributes classified according to their durations.

The graphical representation can be used to interpret some interesting facts for example in the first attribute i.e. diabetes type and its shows that there is two types of diabetes which is type 1 and type 2, The second attribute shows different insulin values.

6. CONCLUSIONS

A detailed analysis of the diabetic data set was carried out efficiently with the help of hive and R. The facts which were revealed during the process can be used for developing some prediction models.In this work we can analyse the diabetes data sets using hive and R and also develop a prediction dataset using which we can predict the diabetes drug using the different attributes of analysed datasets.

REFERENCES

[1] Divyakant Agrawal, UC Santa Barbara, Philip Bernstein, Microsoft Elisa Bertino, Purdue Univ. " Big Data White pdf", from Nov 2011 to Feb 2012

[2] Wullianallur Raghupathi and Viju Raghupathi, "Big data analytics in healthcare: promise and potential",*Health Information Science andSystems 2014.*

[3] Apache Hive https://en.wikipedia.org/wiki/Apache_Hive

[4] Nambiar, R.; Bhardwaj, R.; Sethi, A.; Vargheese, R., "A look at challenges and opportunities of Big Data analytics in healthcare," in Big Data, 2013 IEEE International Conference on , vol., no., pp.17-22, 6-9 Oct. 2013

[5] Wullianallur Raghupathi, Viju Raghupathi "Big Data Analytics in Healthcare: Promise and Potential" http://www.hissjournal.Com/content/2/1/3 Health Information Science and Systems 2014, 2:3

[6] Kiyana Zolfaghar, Naren Meadem, Ankur Teredesai, Senjuti Basu Roy "Big Data Solutions for Predicting Risk-of-Readmission for Congestive Heart Failure patients" 2013 IEEE International Conference on Big Data.

[7] Gopal A. Tathe, Pratik S. Patil, Sangram C. Parle "Healthcare Analytics using Hadoop" IJSRD - International Journal for Scientific Research & Development| Vol. 4, Issue 02, 2016 | ISSN (online): 2321-0613

[8] Sadhana, SavitaShety "Analysis of diabetic data set using Hive and R " IJETAE, July 2014.

[9] Donzé J. Aujesky D., Williams D., Schnipper J.L, MD. ¯Potentially avoidable 30-day hospital readmissions in medical patients: Derivation and validation of a prediction model. JAMA Internal Medicine, 173(8):632-638, Apr. 2013.

[10] Definition, Diagnosis and Classification of Diabetes Mellitus and its Complications, Report of a WHO Consultation Part 1: Diagnosis and Classification of Diabetes Mellitus World Health Organization Department of Non communicable Disease Surveillance, Geneva, 1999.

[11] UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html.

[12] Apache Hadoop. http://hadoop.apache.org.