Scientific Journal of Impact Factor (SJIF): 5.71
e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 5, Issue 04, April -2018

Classification and Prediction of Diabetics using Weka and Hive Tool

Dr.C.S.Kanimozhi Selvi[1], Dr.S.V.Kogilavani[2], Dr.S.Malliga[3], D.Jayaprakash[4]

Department of Computer Science and Engineering
Kongu Engineering College, Perundurai, Erode-638060, Tamilnad, India

Abstract — *Data mining is one such field which tries to extract some interesting facts from huge data set. Since many years ago, the scientific community is concerned about how to increase the accuracy of different classification methods, and major achievements have been made so far. Besides this issue, the increasing amount of data that is being generated every day by various data logging methods raises more challenges. The objective of this project is to train historical diabetes data and classifies them. Classifiers like Naïve Bayes, Decision Tree, Decision Stump, K star and Random forest algorithms are trained very efficiently in a supervised learning setting. Though the chosen dataset is small in size, to learn the big data tools and to find the effectiveness of the tools on a small data set Hadoop and Hive is used in this paper.*

Keywords-*Classification, Diabetics, Big data, Hadoop, Hive*

I. INTRODUCTION

Medical professionals need a reliable prediction methodology to diagnose Diabetes. Classification is a data mining technique used to predict group membership for unseen data instances using the learned model[7][8]. Hence, the data mining algorithm called classification is used for diabetic dataset. The classification procedures on huge amounts of data usually referred as big data, on a distributed infrastructure is done using Hadoop Map Reduce.

Big data is a term for data sets that are so large or complex that traditional data processing applications are insufficient. Challenges include analysis, capture, search, sharing and storage. Accuracy in big data may lead to more confident decision making, and better decisions can result in greater operational efficiency, cost reduction and reduced risk.

Apache Hadoop is an open-source software framework for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. HDFS holds very large amount of data and provides easier access. To store such huge data, the files are stored across multiple machines. It is suitable for the distributed storage and processing. Hadoop provides a Command interface to interact with HDFS[5]. The basic architecture of Hadoop and HDFS is explained in Figure 1.

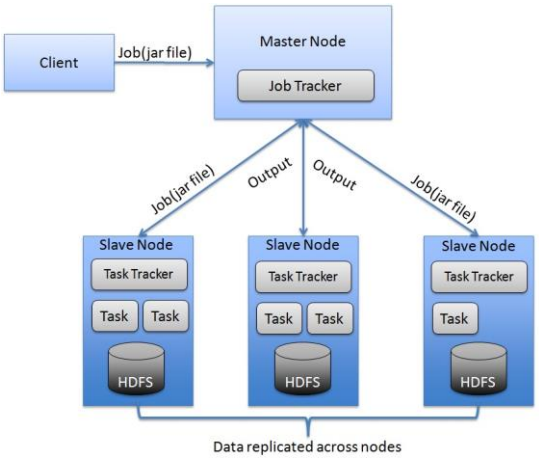

Figure 1. Basic architecture of Hadoop and HDFS

The proposed system uses Hadoop framework to load and process the Diabetes data. According to survey conducted, it is found out that million people have diabetes; by 2035 this may rise to 592million.The number of people with diabetes is increasing in every country, 80% of people with diabetes live in low- and middle-income countries. Every six seconds a person dies from diabetes. More than 21 million live births were affected by diabetes during pregnancy in 2013. There are three types of diabetics mellitus Type 1 DM results from the body's failure to produce insulin, and presently requires the person to inject insulin. Type 2 DM results from insulin resistance, a condition in cells fail to use insulin properly, Type 3, gestational diabetes occurs when pregnant women without a previous diagnosis of diabetes develop a high blood glucose level. Type2 diabetics are the most common type worldwide. Diabetes classification using data mining algorithm will help the medical practitioners in their diagnosing process. Modern

medicine generates a great deal of information which is deserted into the medical database. A proper analysis of such information may reveal some interesting facts, which may otherwise be hidden or go dissipate. Data mining is one such field which tries to extract some interesting facts from huge data set.

II. LITERATURE REVIEW

This chapter reviews few works related to classification of diabetes data.

A classification algorithms for big data analysis, using a map reduce approach is discussed in [4]. In their work, a tool within the scope of InterIMAGE Cloud Platform (ICP), which is an open-source, distributed framework for automatic image interpretation, is presented. They proposed the tool named ICP: Data Mining Package, is able to perform supervised classification procedures on huge amounts of data, usually it referred as big data, on a distributed infrastructure using Hadoop Map Reduce. The results of an experimental analysis using a SVM classifier on data sets of different sizes for different cluster configurations demonstrates the potential of the tool, as well as aspects that affect its performance.

An Application of Data Mining Methods and Techniques for Diabetes Diagnosis is proposed in [6]. Data mining is applied to find useful patterns to help in the important tasks of medical diagnosis and treatment[9][10]. This project aims for mining the relationship in Diabetes data for efficient classification.

In [2] Predictive Methodology for Diabetic Data Analysis in Big Data is stated using a Hadoop/Map Reduce algorithm. They used this algorithm to predict the diabetes mellitus disease complications associated with it and the type of treatment to be provided. Based on the analysis, this system provides an efficient way to cure and cares the patients with better outcomes like affordability and availability.

An imputation method using a hybrid combination of CART and Genetic Algorithm is proposed in [1]. The classical neural network model is used for prediction, on the preprocessed dataset.

III. PROPOSED SYSTEM

Due to the growing unstructured nature of Big Data form health industry, it is necessary to structure and emphasis its size into nominal value with possible solution. In this project, the dataset is discretized using Weka tool and all numerical values are converted into nominal values. The data set is also divided into training set and testing set: training set is to build the model and testing set is to determine the accuracy of the model.

The datasets are loaded into hive. A model is generated using Naive Bayes algorithm. Then the model is applied to test set and the accuracy is predicted. In our project, an attempt is made to classify the diabetic data set loaded in Hive and the accuracy of the model is predicted and compared against the benchmark algorithms.

3.1 Architectural Design

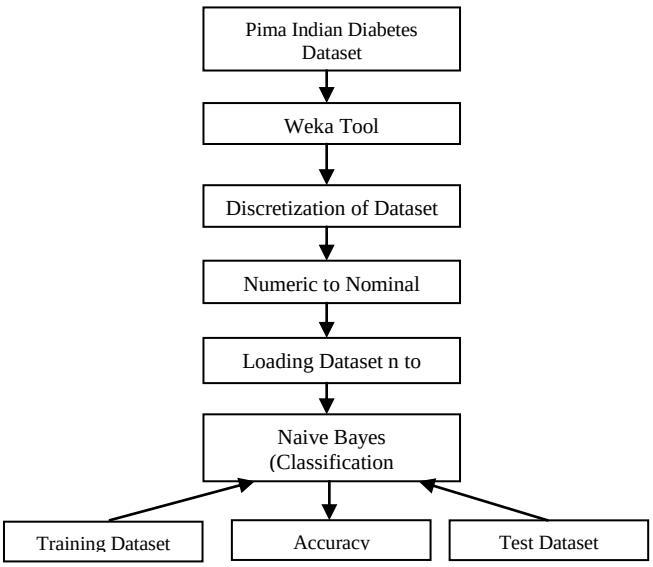

Figure 2. Flow Chart of Proposed System

Pima Indian Diabetes dataset [3] is used in this Project. The samples consist of 768 records with 8 attribute values and one of the two possible outcomes, namely whether the patient is tested positive for diabetes (indicated by output one) or not (indicated by zero). This data set was discritized using Weka tool. Cross validation is performed on the Dataset. Then the dataset is loaded into Hive. Model generated using Naive Bayes algorithm using Java code the model is applied to the test data. Accuracy is calculated all the Datasets.

Preprocessing

Data preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure data preprocessing transforms the data into a format that will be more easily effectively processed for the purpose of the user.

Data Discretization using Weka Tool

Many Machine learning algorithms perform discretization of continuous data before performing a feature selection operation. WEKA uses discretization by default. When Weka starts Weka GUI chooser window is opened. When clicking on the Explorer button, the Weka Knowledge Explorer window is viewed. In that window the needed dataset is selected. Then the classify tab is selected to do discretization.

Numeric to Nominal Conversion

Naive Bayes algorithm can only be applied on the nominal attributes. So convert the dataset form the numeric to nominal after discritiztion.In the discretization, it covert all attributes into the ranges. Then, the ranges are replaced with the strings. Now the dataset is converted to nominal.

Loading the Dataset into HIVE

The preprocessed dataset is divided into training sets and testing sets. To perform 3 fold cross validation, 3 training sets and 3 testing sets are generated. Then the datasets are loaded into hive.

Model Generation using Classification Algorithms

Classification algorithms like Naïve Bayes, Decision tables, Random Forest, Decision Stump, and KStar are executed on PIMA Diabetics dataset. The classifier is tested on two different data sets from the Pima Indian Diabetes Dataset. 3-fold cross validation is performed to calculate the accuracy of the developed classifiers. The different Accuracies obtained are compared to get the best accuracy of the classifiers. It was observed from the comparisons that the naive Bayes classifier's results are very comparable to other algorithms. The popularity of naive Bayes classifier has increased and is being adopted by many because of its simplicity, computational efficiency, and its good performances for real-world problems.

IV. SYSTEM IMPLEMENTATION

4.1 Hadoop Installation

To store the dataset in Hadoop Distributed File System (HDFS), Hadoop should be installed. Hadoop-1.2.1is installed using the following steps:

Step 1: Install SSH then need to update the repository before open ssh is detected. Give yes if asked for dependencies
sudo apt-get update
sudo apt-get install openssh-server

Step2: Need to generate keys in the Terminal
ssh-keygen
Enter file name: empty
Enter passphrase: empty
ssh localhost
GIVE YES
Password: ******
It gets connected
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized keys
ssh localhost
CONNECTED WITHOUT PASSWORD!!!!!!

Step 3: installation of java can be done through the following commends.
For offline install: place tar.gz file in downloads folder
sudo mkdir -p /usr/lib/jvm/
sudo tar xvf ~/Downloads/jdk-7u67-linux-x64.tar.gz -C /usr/lib/jvm
cd /usr/lib/jvm
sudo ln -s jdk1.7.0_67 java-1.7.0-sun-amd64
sudo update-alternatives --config java

Step 4: Now, the java path needs to be set
sudo nano $HOME/.bashrc
Add the below lines
export JAVA_HOME="/usr/lib/jvm/jdk1.7.0_67"
export PATH="$PATH: $JAVA_HOME/bin"
Save and Exit

Step 5: Check the Java Path
exec bash
$PATH

The new paths have added should also appear here

Step 6: Directory is changed to home
cd ~

Step 7: installing hadoop, place the Hadoop file in the home directory
(CAN'T DO IT IN DOWNLOADS AS DONT HAVE PERMISSION TO CREATE FOLDERS/FILES IN THAT DIRECTORY)
cd ~
tar -zxvf hadoop-1.2.1-bin.tar.gz
sudo cp -r hadoop-1.2.1 /usr/local/Hadoop
Password: *******

Step 8: Check the hadoop directory existance
cd /usr/local/hadoop

Step 9: change back to home directory
cd ~

Step10: Need to set a hadoop path
sudo nano $HOME/.bashrc
Go to last line and paste
export HADOOP_PREFIX=/usr/local/hadoop
export PATH=$PATH:$HADOOP_PREFIX/bin
Save and Exit

Step11: To test the path
exec bash
$PATH
The new paths have added should also appear here

Step12: configuring hadoop environment
cd /usr/local/hadoop/conf
sudo nano hadoop-env.sh
Add
export JAVA_HOME=/usr/lib/jvm/java-7-openjdk-i386

Step13: Search the following commend
#export HADOOP_OPTS
export HADOOP_OPTS=-Djava.net.preferIPv4Stack=true
Save and exit
Configuration
USE NAMENODE NAME INSTEAD OF HNNAME
ESPECIALLY FOR MULTINODE DEPLOYMENT PROVIDE THE NAME OF NAMENODE AND NOT YOUR SYSTEM NAME

Step14: Add the below coding on the xml sites.
sudo nano core-site.xml

```
<configuration>
        <property>
                <name>fs.default.name</name>
                        <value>hdfs://decoy:10001</value>
        </property>
        <property>
                <name>hadoop.tmp.dir</name>
                <value>/usr/local/hadoop/tmp</value>
        </property>
 </configuration>
```

Save and Exit
sudo nano mapred-site.xml

```
 <configuration>
        <property>
                <name>mapred.job.tracker</name>
                <value>decoy:10002</value>
        </property>
</configuration>
```

Create Temp Directory:

sudo mkdir /usr/local/hadoop/tmp
pwd:hadoop
sudo chown anand /usr/local/hadoop/tmp
sudo chown anand /usr/local/hadoop
Formatting the DFS:
Don't do if creating a multinode deployment
hadoop namenode -format

Step15: check for success message, start all process
start-all.sh

Step16: Java process status: shows all running process
jps

Step17: To see the details from GUI web interface,go to firefox ,hadoop administration change the namenode name accordingly
http://welcome-HP-246-Notebook-PC:50070/dfshealth.jsp
Look for live nodes and browse the dfs file system
HADOOP JOB TRACKER
CHANGE THE NAMENODE NAME ACCORDINGLY
http://welcome-HP-246-Notebook-PC:50030/jobtracker.jsp
Step18: To stop all process
Stop-all.sh

4.2 HIVE Installation

Step1: Download stable version of hive that is compatible for current version of Hadoop.
It will be extracted to the same home directory. Now it will be move it to /usr/local/hive.
sudo mv hive-0.11.0 /usr/local/hive

Step2: To change directory into home, unpack the tar to extract hive
tar -zxvf hive-0.11.0.tar.gz

Step3: Now move the untar file into /usr/local/hive
s**udo mv hive-0.11.0 /usr/local/hive**

Step4: Configuration of Hive, Open a bash.rc and add a below line
sudo nano $HOME/.bashrc
Add these lines
export HIVE_PREFIX=/usr/local/hive
export PATH=$PATH:$HIVE_PREFIX/bin

Step 5: To check the path of the Hive
$PATH

Step6: To get into hive Prompt
hive
Now open a hive Prompt:
hive>

4.3 Preprocessing Using Weka

1. Starts Weka – you get the Weka GUI chooser window.
2. Click on the Explorer button and you get the Weka Knowledge Explorer window.
3. Click on the "Open File" button and open an ARFF file.
4. Click on Choose and select filters/unsupervised/attribute/Discretized. Then click on the area right of the Choose button.

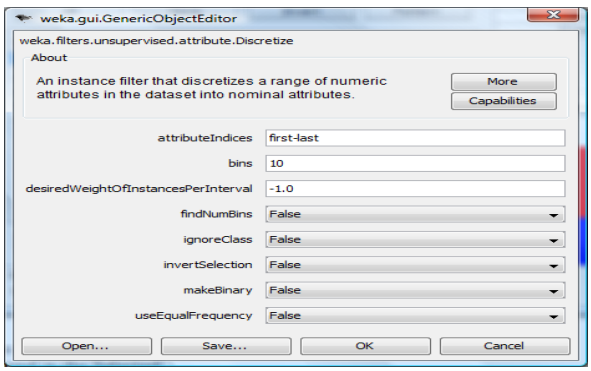

Figure 3. Numeric to Nominal conversion

4.4 Loading Dataset into HIVE

The data set can be divided into Training dataset and Test dataset. Then the dataset is loaded into hive using the following commends. First create the hive a table using the following commends.

Create table testsample(ntp string, pcc string, si, string, dbp string, dp string, bmi string, age string, tsft string,class string)

Row Format Delimited

Fields Terminated By ','

Lines Terminated By '\n';

Then load the Data sets into the hive using the following commends

Load data local inpath '/home/welcome/data/testsample.txt' into table testsample;

```
6,148,72,35,0,33.6,0.627,50,1
1,85,66,29,0,26.6,0.351,31,0
8,183,64,0,0,23.3,0.672,32,1
1,89,66,23,94,28.1,0.167,21,0
0,137,40,35,168,43.1,2.288,33,1
3,78,50,32,88,31.0,0.248,26,1
10,115,0,0,0,35.3,0.134,29,0
2,197,70,45,543,30.5,0.158,53,1
4,111,72,47,207,37.1,1.390,56,1
1,164,82,43,67,32.8,0.341,50,0
10,139,80,0,0,27.1,1.441,57,0
1,189,60,23,846,30.1,0.398,59,1
5,166,72,19,175,25.8,0.587,51,1
7,100,0,0,0,30.0,0.484,32,1
0,118,84,47,230,45.8,0.551,31,1
7,107,74,0,0,29.6,0.254,31,1
```

Figure 4. Loading Data set into Hive

4.5 Performance

Performance is measured in terms of Accuracy, True Positive Rate, False Positive Rate, and Precision

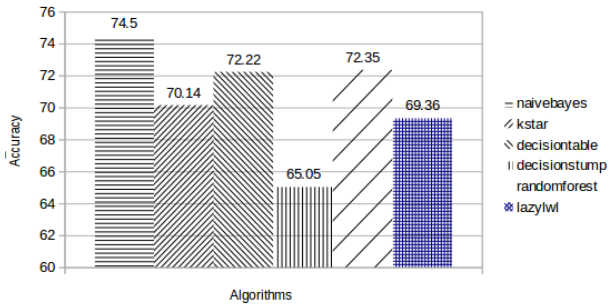

Figure 5. Comparison chart for Accuracy

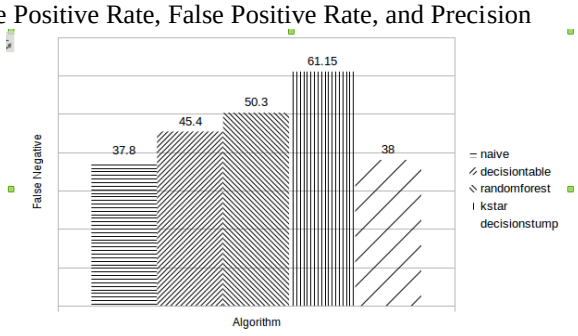

Figure 6. False Negative Rates on Diabetic Dataset

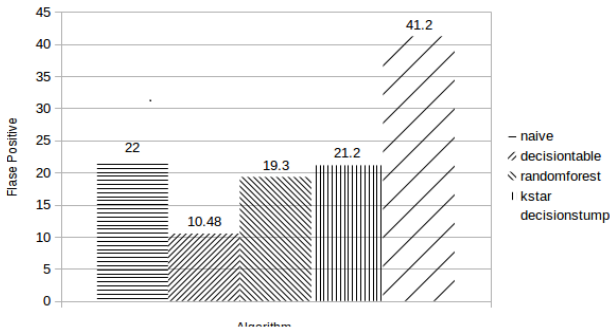

Figure 7. False Positive Rate on Diabetic Dataset

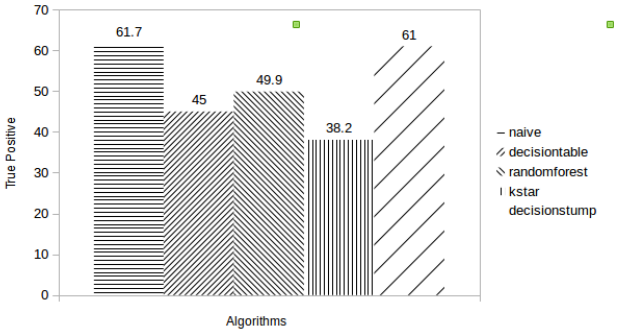

Figure 8. True Positive Rates on Diabetic Dataset

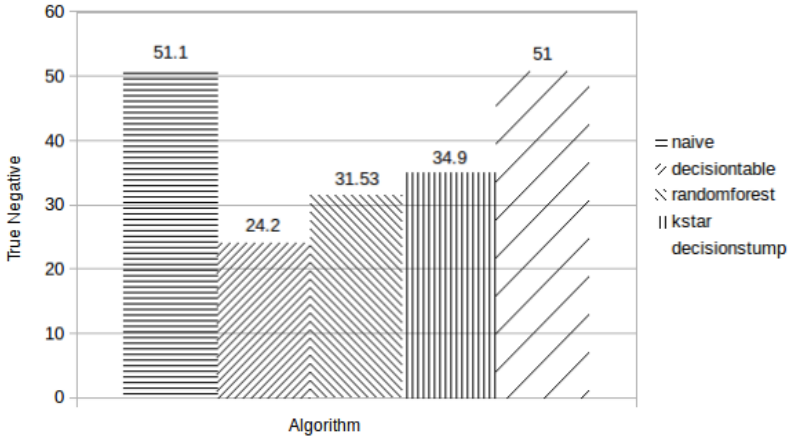

Figure 9. True Negative Rates on Diabetic Dataset

This shows that Naïve Bayes algorithm comparatively better than other algorithms like Decision tables, Random Forest, Decision Stump, and KStar.

V. CONCLUSION

A detailed analysis of diabetes data set was carried out efficiently with the help of Hadoop and Hive. The facts which were revealed during the process can be used to develop some prediction models. Only the analysis is carried out but the information which was revealed can be further used to develop prediction models. In this work, classification algorithms were applied on the small dataset. In future, it can be extended to classify large datasets. Also, the classification algorithms which deal with unstructured data can be implemented and evaluated.

REFERENCES

[1] V. H. Bhatt, P. G. Rao, and P. D. Shenoy, "An Efficient Prediction Model for Diabetic Database Using Soft Computing Techniques," Architecture,Springer-Verlag Berlin Heidelberg, pp. 328-335, 2009

[2] Eswari, T., P. Sampath, and S. Lavanya. "Predictive methodology for diabetic data analysis in big data." Procedia Computer Science 50 (2015): 203-208.

[3] http://www.ics.uci.edu/~mlearn/MLRepository.html.

[4] Ayma, Victor Andres, R. S. Ferreira, Patrick Nigri Happ, Dário Augusto Borges Oliveira, Gilson AOP Costa, Raul Queiroz Feitosa, A. Plaza, and Paolo Gamba. "On the architecture of a big data classification tool based on a map reduce approach for hyperspectral image analysis." In Geoscience and Remote Sensing Symposium (IGARSS), 2015 IEEE International, pp. 1508-1511. IEEE, 2015.

[5] Pakize.S. and Gandomi.A. 2014Comparative Study of Classification Algorithms Based on Map Reduce Model. International Journal of Innovative Research in Advanced Engineering, 1(7), pp.215-254

[6] K.Rajesh,v.Sangeetha,"Application of Data Mining Methods and Techniques for diabetes Diagnosis" in International Journal of Engineering and Innovative Technology(IJEIT) Volume 2,Issue 3,September2014.

[7] Phyu T. (2009), "Survey of classification techniques in data mining", Proceedings of the International Multiconference of Engineering and Computer Scientist (IMECS), vol.

[8] Sadhana, Savitha Shetty, "Analysis of Diabetic Data Set Using Hive and R", International Journal of the Emerging and Advanced Engineering, Vol 4(7), 2014

[9] Pratap, Anju, and C. S. Kanimozhiselvi. "Application of Naive Bayes dichotomizer supported with expected risk and discriminant functions in clinical decisions—Case study." In Advanced Computing (ICoAC), 2012 Fourth International Conference on, pp. 1-4. IEEE, 2012.

[10] Pratap, Anju, and C. S. Kanimozhiselvi. "Predictive assessment of autism using unsupervised machine learning models." International Journal of Advanced Intelligence Paradigms 6.2 (2014): 113-121.

