ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, Vol 9(47), DOI: 10.17485/ijst/2016/v9i47/106496, December 2016 ISSN (Online) : 0974-5645

Analysis of Diabetic Dataset and Developing Prediction Model by using Hive and R

Aanurag Kumar Srivastava[1*], Chandan Kumar and Neha Mangla[2]

[1]Atria Institute of Technology, Bangalore – 24, Karnataka, India; anuraags435@gmail.com, chandan3834@gmail.com
[2]Department of ISE, Atria Institute of Technology, Bangalore - 24, Karnataka, India; apj.neha@gmail.com

Abstract

Objectives: Diabetic is one of the most venerable disease spreading in the world, it will caused due to hereditary and also due to lack of diet. But if we analyze this disease then we can find some fact from the symptoms. Using these facts we can make a predicting model to predict the diabetic disease. By using this model the prediction of the diabetic will be easier and lots of benefits can be provided to the humanity. By sharing the information we extract from our model to the government will help the government for making the welfare program for the citizens. **Method and Analysis:** In this paper we have taken the sample of Pima Indian diabetic dataset which is having the 768 samples. So first of all that dataset will be given as input to hive so as to convert it into a formatted dataset. Then we will apply few queries on the formatted dataset in order to extract the useful information. Then we use the R tool in order to perform the statically analysis for generating the graph and also for calculating gini index and developing the prediction model, and efficiency of the model is also found. **Findings:** In our paper we have performed few queries on the diabetic dataset using hive such as finding the distinct values from the table and by finding it we can analyze the different attributes of the table and also time taken for analysis can also be calculated by default which is one of the positive points of using the hive. Then we will be using the r tool for statically analysis, as we all know picture speaks more than the word so by using the graph generated by r tool we can analyze the dataset easily and fast as compared to going through each rows of the dataset. We calculate gini index for attributes in order to find the inequality among the values using r tool. We also make the prediction model using KNN algorithm and we also find the accuracy of our model. These all things done by the use of r tool, which makes it simpler and also make the method easy to understand by the user to make prediction model and to calculate the efficiency of the model. By using the prediction model we can find the number of sample predictions made correctly. **Improvements:** We can improve the paper by doing the operations performed on large dataset such as millions of dataset in order to make paper more efficient. Our project efficiency is about 79% which can further be improved.

Keywords: Big-Data, Gini Index, Hadoop, Hive, K Nearest Neighbor, R

1. Introduction

In this paper we are analyzing the diabetic dataset using hadoop hive. Apache Hadoop is an open-source software framework written in Java for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. Apache Hive is data warehouse infrastructure built on top of Apache Hadoop for providing data summarization, ad-hoc query, and analysis of large datasets. Few data mining models developed as follows:

- Tried to analyze the diabetic dataset using hive and r and generated graphs which will be helpful for analyzing the dataset[1].
- Used regression based data mining technique for predictive analysis of diabetic treatment[2].
- Has made research to classify Diabetes Clinical data and predict the likelihood of a patient being affected

* Author for correspondence

with Diabetes. He applied different classification algorithm but found c4.5 is the best classification algorithm[3].

- This system will predict future state and generate useful information for effective decision-making[4].

As we are analyzing a big-data hive is a very important tool for analyzing big-data. As we all know the diabetic is spreading worldwide and if it is not handled perfectly then the disease will become worst. Due to lack of knowledge the people are not aware of this disease.

So in this paper we are trying to analyze the diabetic dataset to extract some facts and also by using these facts we try to analyze the dataset. We will also use a gini index to find the inequality among the data and also we will use the k nearest neighbor[5] to make a prediction model for predicting the diabetic disease. We are using Pima Indian diabetic dataset in order for analyzing the dataset. These are the table people living in phoenix Arizona in USA. This diabetic dataset will be consisting of the sample taken from a survey among the women in the tribe.

2. Material and Methods

Dataset analysis consists of several steps. These steps will be discussed in details in further topics.

The main points behind this paper are as follows:

- Pima diabetic dataset is taken as a sample input.
- We perform some query on them using the hive works on hadoop platform.
- We use R tool for statical analysis of dataset and generation of prediction model.
- We use gini index for classification algorithm.
- For prediction we use k nearest neighbor algorithm.

2.1 Hadoop

It is an open-source framework that allows to store and process big data in a distributed environment across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. In our paper as we are using hive for analysis of diabetic dataset, the hadoop works as platform to support hive.

2.2 Hive

Itis a data-warehouse infrastructure built on top of Hadoop for summarization, query, data analysis, it was developed by Facebook, and Apache Hive is now used and developed by other companies such as Netflix and the Financial Industry Regulatory Authority. In our paper it is used in the analysis of diabetic dataset and performing few queries.

2.3 R Tool

It is a programming language used for statically computation and graphical support. In our paper we are using this for generation of graphs so as to analyze graph easily and also it helps in finding gini index and also in the development of the prediction model using knn algorithm.

2.4 Gini Index

It is a method used to measure the inequality among the sample. In our paper it is used for analysis of the inequality among the sample taken for a particular attribute.

Gini index ranges from 0 to 1 when it is 0 then it is perfect equality and when it is 1 it is perfect inequality. The area between Lorentz curve Figure 1 and line of equality represents the inequality, if it's more, more will be the inequality and if it's less nearer to the line of equality then it will be having less equality.

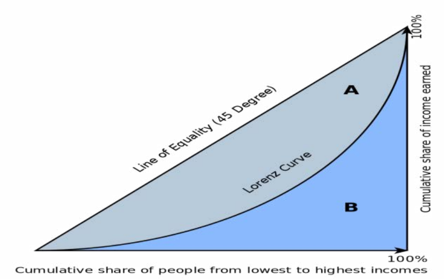

Figure 1. Lorenz curve.

2.5 KNN Algorithm

It is an algorithm which stores all available cases and classifies new cases based on similarity measures. A case is classified by a majority vote of its neighbors with the case being assigned to the class most common amongst its K Nearest Neighbors measured by a distance function. In our paper it is used to develop the prediction model.

Distance functions

1. Euclidians

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

2. Manhattan

$$\sum_{i=1}^{k} |x_i - y_i|$$

2.6 Histogram

Here in our paper it used to provide the analysis of diabetic dataset as picture speaks more than words so analysis of dataset will be easier using graph as compared to that of going through the table.

2.6.1 Confusion Matrix

A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. Here our model is KNN algorithm.

- True Positive (TP) - Measures the proportion of positives that are correctly identified.
- True Negative (TN) - Measures the proportion of negatives that are correctly identified.
- False Positive (FP) - Result that indicates a given condition has been fulfilled, when it actually has not been fulfilled.
- False Negative (FN) - It is where a test result indicates that a condition failed, while it actually was successful.

2.6.2 Evaluation Matrix

In the following subsection, all metrics can be derived from the four basic cardinalities of the so-called confusion matrix, namely the True Positives (TP), the False Positives (FP), the True Negatives (TN), and the False Negatives (FN).

- **Sensitivity:** It is also called the true positive rate, measures the proportion of positives that are correctly identified as such (e.g., the percentage of sick people who are correctly identified as having the condition.

$$\text{Sensitivity} = \text{TPR} = \frac{TP}{TP + FN}$$

- **Specificity:** It is also called the true negative rate, measures the proportion of negatives that are correctly identified as such (e.g., the percentage of healthy people who are correctly identified as not having the condition.

$$\text{Specificity} = \text{TNR} = \frac{TN}{TN + FN}$$

- **Precision:** A measure of degree to which the same result would be produced over different segmentation sections.

$$\text{Precision} = \text{PPV} = \frac{TP}{TP + FP}$$

2.6.3 Statistical Analysis

We used Pima Indian diabetic dataset which contains 768 samples. We used r tool for statistical analysis of the diabetic dataset.

3. Result and Discussion

- First we will create the table by using the database where we are going to store our dataset. The table will be consisting of the attribute name followed by the data type as in Figure 2.
- Then we will load the dataset into the table Figure 3.
- Our dataset look like Figure 4.
- We perform few queries on the dataset as finding distinct values Figure 5, finding number of positive and negative sample from the dataset.
- Now we will go to our R tool where we will first take our dataset as an input as shown in command line 1 from Figure 6.
- Then we will find the summary of dataset using command shown in line 2 from Figure 6 from below, the summary will show what is max, min, median values for a particular attribute.
- But in order to perform the k nearest neighbor we will need to normalize the values of attribute. So that the values lies between 0 to 1 using commands between line 4-7 shown in Figure 6.
- Then we will generate the histogram using command from line-8, for the attribute. As picture speaks more than words it will be easy to go through the dataset using the graph. Graphs we can see in Figure 7.
- We also generate the Lorentz curve in order to find the equality among the different attributes using command from line 11 from Figure 6 and graph in Figure 7
- We will also find the gini index using command line 9-10 from Figure 6.
- Then we will make our prediction model using KNN algorithm[6] using the normalized dataset by using command line 12-19 from Figure 6.

- Cross table we can see below which is used for prediction model Figure 8.
- The test data consisted of 268 observations. Out of which 169 cases have been accurately predicted (TN->true Negatives) in nature which constitutes 63.1%. Also, 43 out of 268 observations were accurately predicted (TP-> True Positives) in nature which constitutes 16%. Thus a total of 43 out of 268 predictions where TP i.e., True Positive in nature.
- There were 43 cases of False Negatives (FN).There were 13 cases of False Positives (FP) meaning 13 cases were actually non diabetic in nature but got predicted as diabetic.
- The total accuracy of the model is 79.10 % ((TN+TP)/268).

We can find accuracy by using 20th command from Figure 6.

```
training@localhost:~
File  Edit  View  Terminal  Tabs  Help
[training@localhost ~]$ hive
Hive history file=/tmp/training/hive_job_log_training_201603190223_322439004.txt
hive> show databases;
OK
anurag_db
anurags
default
hello
nabi
Time taken: 5.679 seconds
hive> use hello;
OK
Time taken: 0.045 seconds
hive> show tables;
OK
diabetic2
sample1
Time taken: 0.534 seconds
hive> create table diab(seq int,npg int,glu int,bp int,skin float,bmi float,ped
float,age int,type int)
    > row format delimited
    > fields terminated by ',';
OK
Time taken: 0.727 seconds
```

Figure 2. Commands for showing databases, tables and creating tables.

```
_1.6
    > fields terminated by ',';
OK
Time taken: 0.727 seconds
hive> desc diab;
OK
seq     int
npg     int
glu     int
bp      int
skin    float
bmi     float
ped     float
age     int
type    int
Time taken: 0.587 seconds
hive> load data local inpath '/home/training/Desktop/dataset.txt' into table dia
b;
Copying data from file:/home/training/Desktop/dataset.txt
Copying file: file:/home/training/Desktop/dataset.txt
Loading data to table hello.diab
OK
Time taken: 0.852 seconds
hive>
```

Figure 3. Commands for loading dataset into table.

File Edit View Terminal Tabs Help

3	115	66	39	140	38.1	0.15	28	0
6	194	78	0	0	23.5	0.129	59	1
4	129	60	12	231	27.5	0.527	31	0
3	112	74	30	0	31.6	0.197	25	1
0	124	70	20	0	27.4	0.254	36	1
13	152	90	33	29	26.8	0.731	43	1
2	112	75	32	0	35.7	0.148	21	0
1	157	72	21	168	25.6	0.123	24	0
1	122	64	32	156	35.1	0.692	30	1
10	179	70	0	0	35.1	0.2	37	0
2	102	86	36	120	45.5	0.127	23	1
6	105	70	32	68	30.8	0.122	37	0
8	118	72	19	0	23.1	1.476	46	0
2	87	58	16	52	32.7	0.166	25	0
1	180	0	0	0	43.3	0.282	41	1
12	106	80	0	0	23.6	0.137	44	0
1	95	60	18	58	23.9	0.26	22	0
0	165	76	43	255	47.9	0.259	26	0
0	117	0	0	0	33.8	0.932	44	0
5	115	76	0	0	31.2	0.343	44	1
9	152	78	34	171	34.2	0.893	33	1
7	178	84	0	0	39.9	0.331	41	1
1	130	70	13	105	25.9	0.472	22	0
1	95	74	21	73	25.9	0.673	36	0
1	0	68	35	0	32.0	0.389	22	0
5	122	86	0	0	34.7	0.29	33	0
8	95	72	0	0	36.8	0.485	57	0
8	126	88	36	108	38.5	0.349	49	0
1	139	46	19	83	28.7	0.654	22	0
3	116	0	0	0	23.5	0.187	23	0

inter outside or press Ctrl+Alt.

Figure 4. Structured dataset.

```
Cloudera_training_VM_1.6
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapred.reduce.tasks=<number>
Starting Job = job_201603112118_0014, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=job_201603112118_0014
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -kill job_201603112118_0014
2016-03-12 00:22:04,233 Stage-1 map = 0%,  reduce = 0%
2016-03-12 00:22:09,395 Stage-1 map = 100%,  reduce = 0%
2016-03-12 00:22:24,851 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_201603112118_0014
OK
17
Time taken: 29.265 seconds
hive> SELECT COUNT(DISTINCT PGL) AS some_alias FROM diabetic1;
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapred.reduce.tasks=<number>
Starting Job = job_201603112118_0015, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=job_201603112118_0015
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -kill job_201603112118_0015
2016-03-12 00:23:14,955 Stage-1 map = 0%,  reduce = 0%
2016-03-12 00:23:20,236 Stage-1 map = 100%,  reduce = 0%
2016-03-12 00:23:31,501 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_201603112118_0015
OK
136
Time taken: 27.619 seconds
hive>
```

Figure 5. Queries performed on dataset using hive.

```
1  pima<-read.table('C:\\Users\\lenovo\\Desktop\\pima.txt',header=T)
2  summary(pima)
3  head(pima)
4  normalize<-function(x){return((x-min(x))/(max(x)-min(x)))}
5  normalize(c(1,2,3,4,5,6,7,8))
6  pima_n<-as.data.frame(lapply(pima[,c(1,2,3,4,5,6,7,8)],normalize))
7  summary(pima_n)
8  hist(pima$age,ylim=c(0,300),xlim=c(0,100),xlab="age",col="green")
9  library(ineq)
10 ineq(pima$age,type='Gini')
11 plot(Lc(pima$age))
12 pima_train<-pima_n[1:500, ]
13 pima_test<-pima_n[501:768, ]
14 pima_trainlab<-pima[1:500,9]
15 pima_testlab<-pima[501:768,9]
16 library(class)
17 library(gmodels)
18 pima_pred<-knn(train=pima_train,test=pima_test,cl=pima_trainlab,k=27)
19 CrossTable(x=pima_testlab,y=pima_pred,prop.chisq=FALSE)
20 mean(pima_pred==pima_testlab)
```

Figure 6. Queries performed on dataset using R.

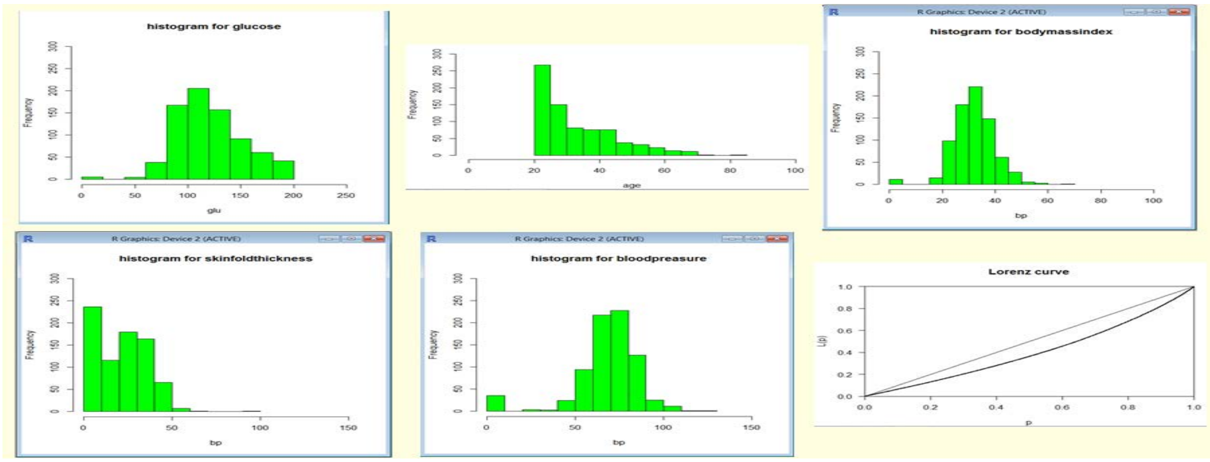

Figure 7. Graphs generated using R.

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|

Total Observations in Table:  268

             | pima_pred
pima_testlab |         0 |         1 | Row Total |
-------------|-----------|-----------|-----------|
           0 |       169 |        13 |       182 |
             |     0.929 |     0.071 |     0.679 |
             |     0.797 |     0.232 |           |
             |     0.631 |     0.049 |           |
-------------|-----------|-----------|-----------|
           1 |        43 |        43 |        86 |
             |     0.500 |     0.500 |     0.321 |
             |     0.203 |     0.768 |           |
             |     0.160 |     0.160 |           |
-------------|-----------|-----------|-----------|
Column Total |       212 |        56 |       268 |
             |     0.791 |     0.209 |           |
-------------|-----------|-----------|-----------|
```

Figure 8. Cross table for prediction model.

4. Conclusion

In this paper we tried to analyze the diabetic dataset and make a prediction model for it. But this particular paper can be referred as to analyze for some other disease like cancer with more symptoms and the important facts which we get can be shared with government for making any important program for humanity.

5. Acknowledgement

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of the people who made it possible whose constant guidance and encouragement crowned our efforts with success.

I express my gratitude to Dr. Neha Mangla, Department of Information Science and Engineering, and Atria Institute of Technology for her valuable support.

I would like to express my gratitude to Mr. Suhas A. Bhyratae, Assi. Prof., and Mrs. Shanthi Mahesh Assoc. Prof. Department of Information Science and Engineering for their support.

I am indeed very happy to greatly acknowledge the persons involved in lending their help to make this paper. At this juncture, I would like to thank my parents and all my friends for their total support and encouragement.

6. References

1. Sadhana SS, Shetty S. Analysis of diabetic dataset using hive and R. 2014.
2. Aljumah AA, Ahamad MG, Siddiqui MK. Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University – Computer and Information Sciences. 2013; 25:127–36.
3. Rajesh K, Sangeetha V. Application of data mining methods and techniques for diabetes diagnosis. IJEIT. 2012 Sept; 2(3).
4. Bagdi R. Patil P. Diagnosis of diabetes using OLAP and data mining integration. International Journal of Computer Science and Communication Networks. 2014; 2(3): 314-22.
5. Available from: http://www.analyticsvidhya.com/blog/2015/08/learning-concept-knn-algorithms programming/
6. Available from: http://www.josechristian.com/programming/inequality-and-lorenz-curve-r/
7. Available from: http://www.r-bloggers.com/gini-index-and-lorenz-curve-with-r/
8. Available from: http://www.saedsayad.com/k_nearest_neighbors.html