Analysis of Diabetic Data Set Using Hive and R

Sadhana¹, Savitha Shetty²
CSE, NMAM Institute of technology, Nitte, Karnataka, India.

Abstract— Modern medicine generates a great deal of information which is deserts in to 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. In this paper an attempt is made to analyse the diabetic data set and derive some interesting facts from it which can be used to develop the prediction model.

Keywords-Diabetes mellitus, hyperglycemia, Hive, R.

I. INTRODUCTION
Diabetes mellitus has been defined as a clinical syndrome characterized by hyperglycemia, due to deficiency or diminished effectiveness of insulin [5]. Diabetes mellitus has become a global menace. The world health organization has estimated the number of diabetics in the world by 2025 may reach up to 60 million and India's contribution to it would be 30 million. Hence this is a major issue and an awareness regarding this disease is essential.

A huge amount of data gets accumulated in the hospitals, most of them just get stored in some form of files which are never touched back, if these data are analysed properly they help in deriving some interesting facts. A small touch of data mining will help in generating interesting facts which remained unrevealed otherwise, hence taking into consideration the diabetes mellitus a detailed analysis of diabetic data set is performed using data mining technique.

II. RELATED WORK
According to survey conducted it is found out that 382 million people have diabetes; by 2035 this may rise to 592 million. The number of people with type 2 diabetes is increasing in every country, 80% of people with diabetes live in low- and middle-income countries. The greatest numbers of people with diabetes are between 40 and 59 years of age.175 million people with diabetes are undiagnosed. Diabetes caused 5.1 million deaths in 2013; every six seconds a person dies from diabetes. More than 21 million live births were affected by diabetes during pregnancy in 2013. [6] Many people have developed various prediction models using data mining to predict diabetes. A few of the models developed using data mining are as follows:


The support vector machine algorithm was used for experimental analysis. The data sets Datasets of Non Communicable Diseases (NCD) was analysed for finding out the effectiveness of different treatment types for different age groups. The arrived at a conclusion that drug treatment for patients in the young age group can be delayed whereas; patients in the old age group should be prescribed drug treatment immediately.

Velide Phani Kumar et al [2] analyzed diabetes data using various data mining techniques which involved, Naive Bayes, J48(C4.5) JRip ,Neural networks, Decision trees, KNN,Fuzzy logic and Genetic Algorithms based on accuracy and time. They found that that out of various data mining techniques which were employed to analyze the diabetes data. J48 (C4.5) took least time.

Rupa Bagdi et al [3] developed a decision support system which combined the strengths of both OLAP and data mining. This system would predict the future state and generate useful information for effective decision-making. They also compared the result of the ID3 and C4.5 decision tree algorithms. The system could discover hidden patterns in the data and it also enhanced real-time indicators and discovered bottlenecks and it improved information visualization.

K. Rajesh et al [4] carried out a research to classify Diabetes Clinical data and predict the likelihood of a patient being affected with Diabetes. The training dataset used for data mining classification was the Pima Indians Diabetes Database they applied Different classification techniques and found out that c4.5 classification algorithm was the best algorithm to classify the data set.

All the above researchers have been successful in analysing the diabetic data set and developing good prediction models. But most of them used tools like weka and oracle data miner and few of them used tools like Tanagra. In this paper an attempt is made to make analysis of diabetic data set using Hadoop and R.

III. PROPOSED SYSTEM
A. Data set
The data set used for the purpose of this study is Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Diseases. This database database, donated by Vincent Sigillito, is a collection of medical diagnostic reports of 768 examples from a population living near Phoenix, Arizona, USA.
The samples consist of examples 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 is analysed used hive and R.

### B. Architecture of proposed system

Fig 1 shows the architecture of the proposed system. The diabetic data set is given as input to the system, which comprises of hive and R. The raw data is just a file consisting of comma separated values, for the first time when we look into it, it just looks like a junk of data. But a proper analysis of this data set will reveal some interesting facts. The raw input is given as input to hive, the data set is analysed and partitioned based on different attribute the output which is obtained from hive is well formatted data, and then this output is given as input to R. It is one of the best languages which is used for statistical computing as well as for generating graphs. As we all know that pictures speak more than words, after analysing the data using hive the graphs are generated for each data set using R.

### B. Hive and R

In this research we have used hive and R for the purpose of analysing the data set. Hive is a Data Warehousing Solution built on top of Hadoop. It Provides SQL-like query language named HiveQL. It is used for querying tables which are stored as flat files on Hadoop Distributed File system (HDFS) with complete Meta data repository. It helps in translating HiveQL statements into a set of map reduce jobs which are then executed on a Hadoop Cluster. To support features like schemas and data partitioning, Hive keeps the metadata in a relational database. In Hadoop to know about the contents of the HDFS one needs to write Map Reduce programs. Whereas hive supports a data model using standard tables with rows and columns. Importantly it supports partitioning the table on a particular dimension. For example we could store Patient details and partition on Age factor. This allows to later creating queries on an organized data model.

Hive is an interesting project because it allows exposing the best parts of hadoop, namely Map Reduce and data storage, to end users who may have no idea about map reduce or no interest in writing Map Reduce programs. [8]

### IV. EXPERIMENTAL RESULTS

As it was mentioned earlier Hive and R are used for the purpose of analysis. 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.

**Fig. 1 Architecture of the proposed system**

**Fig. 2 snapshot of raw data loaded to hive**

Once the file is loaded to Hive we can perform some analysis on the data set. We found out that there out of 768 records in the data set out of that 500 records were belonging to class 0 and 268 were belonging to class 1 and the time taken in hive to perform this analysis was just 19.086 seconds. Then the data set was partitioned into based on the attributes. There are eight attributes in the data set and for each data set the number of distinct values was found out and the time taken in hive to perform this analysis was recorded. This information is tabulated in table 1.
We have also calculated the gini index for each attribute after the attribute is analysed using Hive. The Gini coefficient measures the inequality among values of a frequency distribution. A Gini coefficient of zero expresses perfect equality where all values are the same. Formula to compute gini index is as follows:

\[ \text{Gini} = 1 - \sum_{i=0}^{c-1} \left( p(i/t) \right)^2 \] [9]

In the formula c indicates the number of class and p(i) indicates the number of values that belong to each class for given attribute value and t indicates the total no of records for the given attribute value. Table 2 shows the minimum gini index values of each attribute.

**Table 2**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Gini Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of times pregnant</td>
<td>0.300877</td>
</tr>
<tr>
<td>Plasma glucose concentration</td>
<td>0.132653</td>
</tr>
<tr>
<td>Serum Insulin</td>
<td>0.165289</td>
</tr>
<tr>
<td>Diastolic BP</td>
<td>0.152778</td>
</tr>
<tr>
<td>Diabetes pedigree</td>
<td>0.375000</td>
</tr>
<tr>
<td>Body Mass Index</td>
<td>0.197531</td>
</tr>
<tr>
<td>Age</td>
<td>0.146133</td>
</tr>
<tr>
<td>Triceps skin fold thickness</td>
<td>0.132653</td>
</tr>
</tbody>
</table>

Fig 3 and Fig 4 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 class distribution for each attribute.

The graphical representation can be used to interpret some interesting facts for example in the first attribute i.e. number of times pregnant the minimum value is 0 and the maximum value is 17 the graph shows that as the value for this attribute increases all the records have positive value for diabetes which indicates that if the count increases then the possibility of being affected by diabetes is more. Similarly the other attributes can be analysed using the graphs generated.

**V. 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 only the analysis is carried out but the information which was revealed can be further used to develop efficient prediction models.

REFERENCES


