Volume 5, Issue 4, 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Cardiovascular Disease Dataset Exploration Using Hive and R

Shetty Mamatha Gopal
Dept. of C.S&E
NMAMIT, Nitte
Affiliated to VTU, Belgaum, India

Assoc. Prof. Radhakrishna
Dept. of C.S&E
NMAMIT, Nitte
Karkala, Karnataka, India

Sadhana
Dept. of C.S&E
SCEM, Adyar
Affiliated to VTU, Belgaum, India

Abstract—huge amount of data gets collected in the hospitals which will be dumped into the medical data repository. A proper and careful study of this data can be carried out to extract some meaningful and interesting information which may help in carrying out various researches. Data mining is a field that deals with mining some meaningful facts from a large amount of data. In this paper an attempt is made to learn the heart disease data set and pull out some meaningful information out of it which can be used to build a prediction model.

Keywords— Cardiovascular Disease, Hive, HiveQL, R.

I. INTRODUCTION

Cardiovascular disease is generally referred to conditions that involve narrowed or blocked blood vessels that can lead to heart attack, chest pain (angina), or stroke. Other heart conditions, such as those that affect your heart's muscle, valves or rhythm, also are considered forms of heart disease. The world health organization terms that cardiovascular diseases are the number one reason for death across the globe, it is estimated that around 17.5 million deaths out of 56 million deaths around the world is due to cardiovascular diseases.

Hence this is a key issue and a cognizance regarding this disease is essential. A massive amount of data gets collected in the hospitals, most of them just gets stockpiled in some form of files which are never touched back; examination of these data may help in deriving some meaningful or important information.

A minor touch of data mining will help in producing stimulating facts which remains hidden otherwise, hence taking this into thought the cardiovascular disease, a thorough analysis of heart disease data set is performed using data mining technique.

II. RELATED WORK

The World Health Organization described that heart disease is the first principal cause of death in high and low income countries [6]. According to a statistics in Europe heart attacks, strokes account for 41% of all deaths [7].As per a report, in one fifth of Asian countries, most lives are lost to non-communicable diseases such as cardiovascular, cancers, and diabetes [8]. The report from World Health Organization states that, the total death due to chronic disease in India is 53% of all deaths [9]. Numerous people have developed prediction model to predict heart disease. The various work carried out are as follows.

Palaniapan et al [1] have developed a prototype Intelligent Heart Disease Prediction System, using techniques like Naïve Bayes, Decision Trees and Neural Network. Various attributes like age, blood pressure, blood sugar. This system is able to predict the possibility of patient getting heart disease.

Ranganatha S. et al [2] has worked on a system where algorithms had run on the information collected from the patients who were hospitalized for heart disease. The result of analysis was provided in the form of user understandable words and graph. Algorithm used in this research was ID3.ID3 outputs the result in the form of decision tree.

Dr. K Usha Rani [3] analyzed heart disease dataset using Neural Network approach. Parallel approach was also used to increase the efficiency of the classification process. The experimental results proved that neural networks technique provides output layers to speed up the learning process.

V. Manikantan et al [4] has used K-means clustering algorithm to cluster the heart disease database. In this work the heart disease database was preprocessed, the missing values and identical values were removed and an efficient approach for fragmenting and extracting substantial forms from the heart attack data warehouses was developed.

Miss. Manjusha B. Wadhonkar [5] has used artificial neural network approach to analyze the Heart disease dataset, to increase the efficiency of the classification process parallel processing approach is also applied. A binary heart disease dataset classifier was developed which could be used to assist doctors to group the data set of heart disease.

All the above researchers have been successful in analyzing the dataset related to heart disease using various approaches like Neural Network, various algorithms like K-means clustering. An attempt to analyze the dataset using Hive is done in this paper. The output of analysis through hive is presented in the form of graphs using R.

III. PROPOSED SYSTEM

A. Data Set

The data set used here for the study is from UCI Machine Learning Repository [9]. The database is the collection of data from four following locations: 1. Cleveland (Cleveland Data) 2. Hungary (Hungarian data) 3.Long Beach, CA (long-beach-va data) 4.Zurich, Switzerland (Switzerland data).This dataset is donated by, David W. Aha. It is a collection of medical analytical reports with values for 76 attributes, but all published experiments refer to using a subset of 14 of them. In the proposed system a detailed study of this dataset is done using hive and R. The various attributes and their description are shown in the table below [9].

Table2: Gini co-efficient of attributes

Name	**Type**	**Description**
Age	Continuous	Age in years
Sex	Discrete	1=Male 0=Female
Trestbps	Continuous	Resting blood pressure(in mm Hg)
Cp	Discrete	Chest pain type 1= typical angina 2=atypical angina 3=non-anginal pain 4=asymptomatic
Chol	Continuous	Serum Cholesterol in mg/dl
Fbs	Discrete	Fasting blood sugar> 120 mg/dl 1=true 0=false
Restecg	Discrete	Resting Electrocardiography results 0= Normal 1= ST-T wave abnormality 2= showing probable or define left ventricular hypertrophy by Estes' criteria
Thalach	Continuous	Maximum heart rate achieved
Exang	Discrete	Exercise induced angina 1=yes 0=no
Old peak ST	Continuous	Depression induced by exercise relative to rest
Slope	Discrete	The slope of the peak exercise segment 1=up sloping 2= flat 3= down sloping
Ca	Discrete	Number of major vessels colored by fluoroscopy that ranges between 0 and 3
Thal	Discrete	3=Normal 6=fixed defect 7=reversible defect

B. Architecture of proposed system

Fig 1 shows the architecture of the proposed system. The cardiovascular data set is fed as input to the system, which consist of of hive and R. This fresh data is just a collection of comma separated values put into a file, initially the data looks like, a junk of values. But a proper examination of this data set will disclose some stimulating facts. The raw data is given as input to hive, and then the input is analyzed and segregated based on different attribute. The yield obtained from hive is well structured data; this output is then given as input to R. R is one of languages used for doing statistical analysis. Pictures can sometimes communicate more than words and hence for the analyzed attribute-wise data using hive graphs using R.

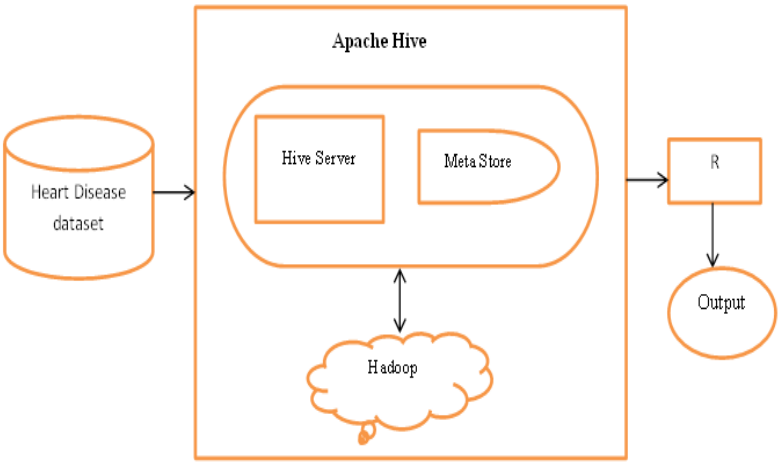

Fig. 1 Architecture of the proposed system

C. Hive and R

In the proposed system hive and R is used for the analysis of the data set. Hive is a Data Warehousing Solution built upon Hadoop. It provides a query language called HiveQL to manage the data. These queries can be used to access the data stored as files on Hadoop Distributed File system (HDFS) with whole Meta data repository. It helps in converting HiveQL statements into a set of map reduce jobs that are later executed on a Hadoop Cluster. Hive uses metadata in the relational database to support features like partitioning. In Hadoop contents of the HDFS can be viewed by writing Map Reduce code. In hive data is stored using standard tables with rows and columns. The table can be partitioned as per various attributes or dimensions. For example the patient records can be partitioned as per attributes like age, chest pain type. Later on queries can be written for the partitioned tables. Hive is an interesting project because it allows revealing the best portions of Hadoop, namely Map Reduce and data storage, to users who do not have any idea about map reduce. [11]

IV. EXPERIMENTAL RESULTS

As it is specified before, Hive and R are used for the purpose of analysis in this research. The data set must be first loaded into hive from a file. It is simply a collection of comma separated values. Fig 2 shows the snapshot of the fresh data which is loaded to Hive.

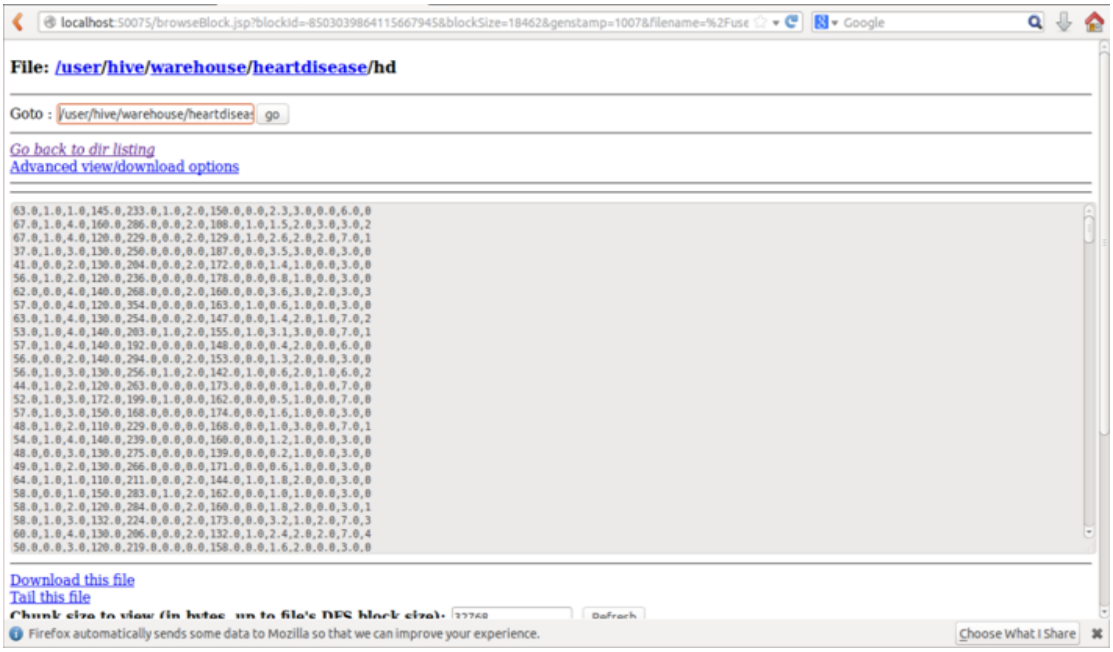

Fig. 2 Snapshot of raw data loaded into hive

Once the file is loaded to Hive we can start studying the data set. Then the data set is divided into different tables based on the attributes. There are 14 attributes in the data set and for each attribute the number of distinct values was found .This information is shown in fig 3.

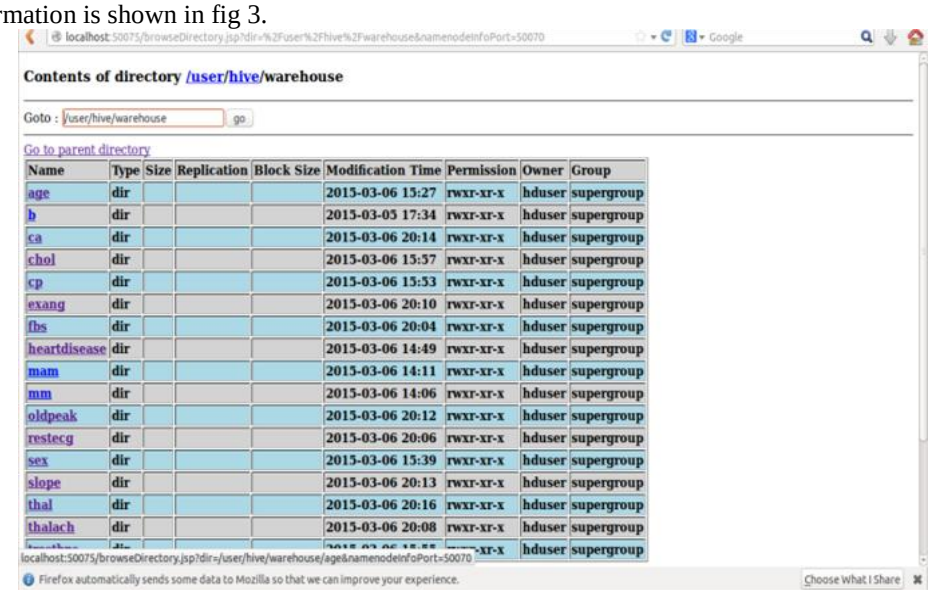

Fig. 3 Snapshot of partitioned data loaded as per attributes

Once the analysis is done the Gini Index is calculated for each attribute. The Gini coefficient is the measure of 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 given below:

$$Gini = 1 - \sum_{i=0}^{c-1} [p(i/t)]^2 \quad [12]$$

In the formula c denotes the number of class and p (i) denotes the number of values that belong to each class for the 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.

Table2: Gini co-efficient of attributes

Attribute	**Gini co-efficient**
Age	0.18000007
Cp	0.31119996
Restecg	0.55865973
Slope	0.41450113
Exang	0.49130142
Fbs	0.63818276
Sex	0.4295888
Ca	0.42762136
Thal	0.3753085
Chol	0.44444442
Thalach	0.31999993
Trestbps	0.44444442
Oldpeak	0.19753087

Fig 4 shows the snapshot of the graphs generated using R. Both the graphs are plotted as follows, x axis represents the various values for the attribute and y axis depicts the class distribution for each attribute. These graphs can be used to understand some stimulating facts for example in the first attribute i.e. age, the minimum value is 29 and the maximum value is 77. The graph clearly shows the various classes for each value of age i.e., for age value 29, the possible class value was 0. Similarly the other attributes can be analyzed using the graphs generated.

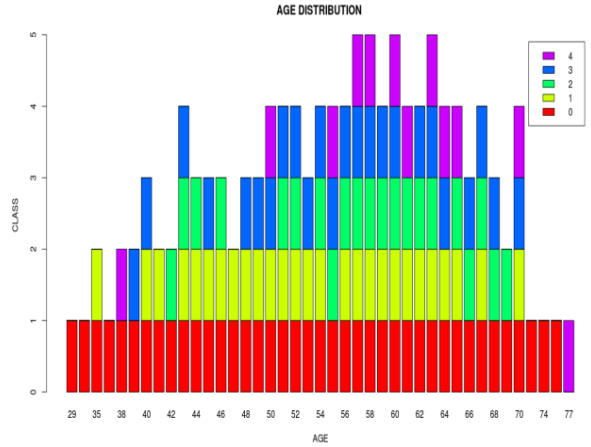

Fig. 4 Barplot for age attribute generated using R

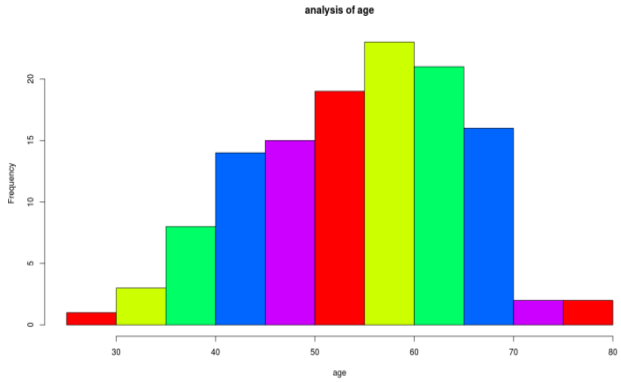

Fig. 5 Snapshot of graphs generated using R

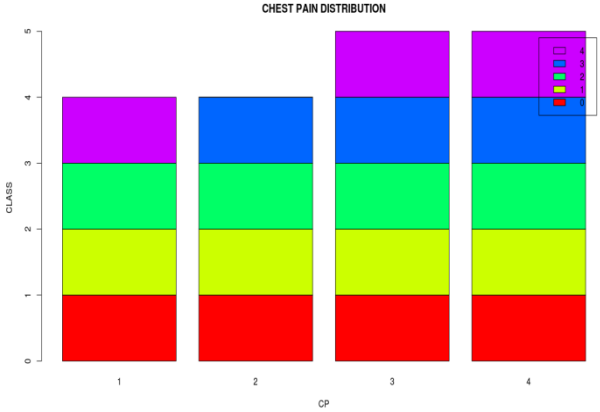

Fig. 6 Snapshot of graph generated for attribute CP using R

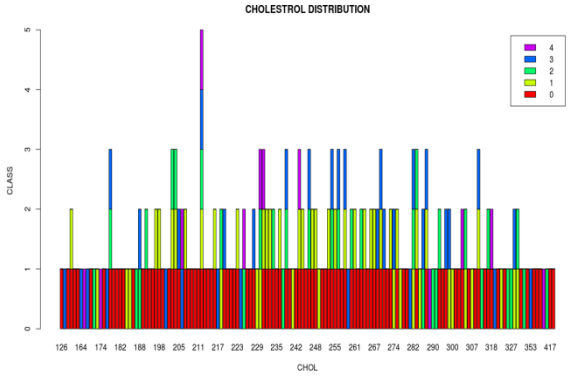

Fig. 7 Snapshot of graph generated for attribute Chol using R

V. CONCLUSION

A detailed study of the cardiovascular disease data set was carried out efficiently with the help of hive and R. The facts which were discovered during the process can be used for building some prediction models.

REFERENCES

[1] Palaniappan S., Awang R., "Intelligent heart disease prediction system using data mining techniques" in Computer Systems and Applications, 2008.AICCSA 2008. IEEE/ACS International Conference on March 31 2008-April 4 2008, 108-115.

[2] Ranganatha, S., Pooja Raj H J, Anusha C, Vinay S K, "Medical data mining and analysis for heart disease data set using classification techniques", in Research & Technology in the Coming Decades(CRT 2013), National Conference on Challenges

[3] Dr. K. Usha Rani, "Analysis of heart disease dataset using neural network approach", in International Journal of Data Mining & Knowledge Management Process(IJDKP) Vol.1, September 2011.

[4] V.Manikantan & S. Latha, "Predicting the Analysis of Heart Disease symptoms Using Medicinal Data Mining Methods", in International Journal on Advanced Computer Theory and engineering (IJACTE), ISSN: 2319-2526, Volume-2, Issue-2, 2013.

[5] Miss. Manjusha B. Wadhonkar, Prof. P. A. Tijare and Prof. S. N. Sawalkar ," Classification of Heart Disease dataset using Multilayer feed forward back propagation Algorithm" in International journal of Application or Innovation in Engineering & management(IJAIEM),Vol. 2, Issue 4, April 2013.

[6] http://www.who.int/mediacentre/factsheets/fs317/en/

[7] http://www.epha.org/a/2352

[8] http://www.unescap.org/stat/data/syb2009/9.Health-risks-causes-of-death.asp

[9] http://www.who.int/chp/chronic_disease_report/media/india.pdf

[10] https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[11] http://www.oceantara.com/overview-of-hive-for-hadoop/

[12] http://www-users.cs.umn.edu/

