IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

EMPIRICAL ESTIMATION OF CLASSIFICATION MODELS FOR PREDICTION OF DIABETIC RELATED DISEASES USING BIGDATA ON THE CLOUD AND HADOOP.

Dr.Vineetha K R, Associate Professor, Nehru College of engineering and research centre.

Nandhana P M, Department of MCA, Nehru College of engineering and research centre.

ABSTRACT - The rapid adoption of Information Technology(IT) in healthcare systems, the health data grows exponentially and it is available in different forms in different ways. Knowledge discovery and decision-making from such rapidly growing huge data is a challenge regarding both data organization and timely processing, which is a promising trend known as Big Data computing. Big Data computing is a new paradigm which combines large-scale computing with machine learning techniques is used to build predictive analytics for intrinsic information extraction. Cloud computing emerges as a service oriented computing model for processing large volumes of rapidly growing data at a faster scale which is a demand for BigData computing. Hence Big Data frameworks Hadoop and Spark are used to carry out Big Data tasks along with machine learning techniques. This thesis focuses on predictive analytics with machine learning to analyze Big Data making decisions about future complications of diabetic patients. The thesis discusses a framework for Big Data computing with Hadoop MapReduce in both standalone and in Cloud(AWS),as well Apache Spark in standalone and also demonstrating their effectiveness by their performances.

Keyword - Machine Learning, K-means, SVM, Diabetic, Hadoop, Big Data , Cloud.

I. INTRODUCTION

Here introduce the concept of Big Data, Big Data in healthcare industry, Diabetic Mellitus, Big data analytics, Predictive analytics, Data Mining and knowledge discovery, Cloud computing, Hadoop Map Reduce, and Machine learning techniques.

Big Data is a large data collection, yet it is growing exponentially over time. It is data that is so large and complex that none of the common data management tools can store or process it properly. Big Data, a collection of items from Social Networking, Mobile Computer, Statistics, and Clouds, is popularly known as SMAC. It is difficult to store and process such fast-growing data over a period of time using traditional tools. Major data analysis challenges include data capture, data retention, data analysis, search, sharing, transfer, visualization, testing, review, confidentiality, and data source. Current use of the term big data often refers to predictable analytics applications, user behavior statistics, or other advanced data analysis methods that extract value from big data, and rarely go to a specific data set size. IDC report says that, compare to other industry like manufacturing big data is expected to grow very fast in healthcare system. Present investigation is interested in the integration of high performance computing, Cloud and Big Data computing technologies for getting state of the art services from the medical and health-care industry. Predictive analysis involves lot of techniques for data mining and statistics using current and past data to predict future events. By utilizing in health care system, significant predictions can be made accurately by applying Big Data analytics in health care system. Diabetes Mellitus (DM) has become a global menace. It is a clinical disorder

due to the deficiency or ineffective nature of insulin. Diabetic Mellitus is a major health problem in countries like India and Asia Pacific region. Diabetes is due to either the pancreas not producing enough insulin or the cells of the body not responding properly to the insulin produced.

Big data analysis is the use of advanced analytical techniques against very large, diverse data sets that include structured, semi and unstructured data, from different sources, and in different sizes from terabytes to zettabytes.

Predictive analytics make predictions about a patient's future health status are based on their current health parameters and data combined with statistical modeling, data mining techniques and machine learning. Cloud computing is an on-demand access , via the Internet, access to computer resources — applications, servers (portable servers and virtual servers), data storage, development tools, communication capabilities, etc. — hosted on a remote data center managed by cloud services provider (or CSP) . The distributed Database for Cloud MapReduce is widely used to process large amounts of data while hiding complexity of parallel execution across hundreds of servers in the Cloud environment. Amazon Web Service (AWS) provides a service called EMR, Elastic Map Reduce as a remote service for storing and processing powerful data using the Map Reduce program on Amazon Cloud. Amazon EMR provides a Hadoop framework for using large amounts of data on Amazon EC2 that makes it faster and less expensive. MapReduce is a framework we can use to write applications to process large amounts of data, in parallel, in large hardware collections in a reliable way. Hadoop provides the required performance speed and storage space to extract useful information from large amounts of data. It splits the data into thousands of segments in a large number of machines to execute in parallel manner. The Hadoop architecture works with two major components namely HDFS and MapReduce to process high speed and store large amounts of data respectively. Present investigation is interested in discovering significant information from the larger data set using data mining. This process involves preparation and selection of data to be mined, cleaning the data, incorporating prior knowledge and drawing solutions from the observed results. Knowledge Discovery Databases(KDD) is involved in discovering useful information from huge data. In this specific algorithm, extracting information through data mining is an important step. Thus the overall aim of this process is to get useful information from raw data. The discovery of hidden useful knowledge from massive databases is possible by applying data mining. KDD is the process of getting important information from Big Dataset.

Machine learning technique involving automated learning from data set. It initially learns the knowledge and applies this knowledge to distribute for predictions. Machine learning models are classified into supervised, semi supervised and unsupervised learning algorithms based on the depth of knowledge. A machine learning algorithms could look at many more factors in patient's charts than doctors, and by adding additional features, there was once a widespread increase in the ability of the model to distinguish people who have complications of diseases, from people who don't.This study focuses on the importance of Big Data tools and machine learning algorithms for clustering and classification to develop a predictive model for predicting diabetic related diseases on Hadoop platform in standalone and in Cloud environment.

II. LITERATURE SURVEY

This Literature survey offers the review of literature on Diabetes prediction in healthcare systems using machine learning algorithms by using different technologies like BigData, predictive analytics, hadoop cluster, cloud computing, hadoopMapReduce.

N. Yuvaraj and K. R. SriPreethaa discussed the concept of "Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster". They describe the fact that health care systems are simply designed to meet the growing needs of the world. People around the globe suffer from various diseases that are extremely deadly. Diabetes is a common disease that causes blindness, kidney failure, heart attack, etc. Different machine learning methods are proposed in health care monitoring systems to predict various diseases and symptoms are available worldwide making the healthcare model work automatically and improve disease prediction accuracy. A distributed computer framework based on the Hadoop collection supports the processing and storage of large databases in the cloud space. In this case the use of diabetes-learning machine learning algorithms in collections based on hadoop. The results show that machine learning algorithms can produce more accurate diabetes predicting health systems. The Pima Indians Diabetes Database from the National Institute of Diabetes and Digestive Diseases is used to test algorithm performance.

B. Suvarnamukhi, M. Seshashayee discussed the "Big Data Processing System for Diabetes Prediction using Machine Learning Technique". In this study they explained that diabetes is not a deadly disease but one of the most threatening diseases in the world. Despite the existence of a limited number of diabetes predictors, large data-based predictions are rare. The use of BigData in the proposed work is extensive because, medical records from a variety of sources are extremely large and extracted and the necessary features intended for them are processed. The goal of this work is achieved by various stages such as data collection, pre-processing, selection of attributes and prediction. Diabetes predictions are made by the Extreme Learning Machine (ELM) classifiers. The performance of the proposed method is evaluated by varying the categories and methods available depending on the accuracy of the disease prediction, accuracy, recall and time management. From the test results, the effectiveness of the work is proven.

Ms Ashwini Abhale, Shruti Gulhane, Sandhya Budhewar, Swanali Jathar and Harshada Sonwane conducted research on "Predictive analysis of Diabetic Patient Data Using Machine Learning and Big Data". After the study they concluded, the healthcare industry

produces a huge amount of information about patients. Using BigData models data can collect, store and process data for information and use it to make important decisions. Diabetic Mellitus (DM) is caused by Non-Communicable Disease (NCD). Now, in developing countries like India, Diabetic Mellitus has become a major health problem and many people suffer from it. Diabetes mellitus causes chronic problems and has a variety of health problems. It is necessary to develop a system that maintains and analyzes diabetes data and predicts potential risks with the help of technology. Predictive analysis is a method that uses current and past data sets to predict future risks by combining various data mining techniques, machine learning algorithms and statistics. In this paper, they use a predictive analytical algorithm in Hadoop / Map Reduce the environment to predict the types of diabetes, related problems and the type of treatment to be provided.

Minyechil Alehegn and Rahul Joshi make a research on "Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach" and conclude that, machine learning (MLT) techniques used to predict medical databases and machine learning (ML) have the ability to answer questions. A large medical database is accessible to various data storage units that used to predict future risk. Diabetes Disease (DD) is currently one of the leading causes of death in the world. Different data mining techniques are used by different researchers at different times to collect and predict symptoms in medical data. A total of 768 cases, data set in PIDD (Pima Indian Diabetes Data Set). In this system the best-known forecasting algorithms operate at KNN, Naïve Bayes, Informal Forest, and J48. By using these algorithms it creates a ensemble hybrid model by combining individual techniques / methods into one to increase efficiency and accuracy.

Ayman Mir and Sudhir N Dhage are make a 374esearch in "Diabetes Disease Prediction Using Machine Learning on Big Data of Healthcare" and explained that, The healthcare sector is a highly diverse field of research with rapid technological advancement and ever-growing data. In order to deal with large amounts of health care data we need Big Data Analytics which is a emerging trend in the healthcare domain. Millions of patients seek treatment worldwide through various procedures. Analyzing the trends in treating patients for a specific disease will help to make informed and effective decisions to improve the overall quality of health care. Machine learning techniques are promising method that can help diagnose disease early and can help doctors make diagnostic decisions. This paper aims to create a classification model using the WEKA diabetes forecasting tool using Naive Bayes, Vector Support Machine, Random Forest and Simple CART algorithm. The study hopes to recommend the best algorithm based on the effective outcome of predicting diabetes. The test results for each algorithm used in the database have been tested. It is noted that the Vector Support Machine has been very effective in predicting disease with high accuracy.

III. DATASET

a. Traditional diabetic Dataset
b. Real time diabetic dataset

Traditional Dataset (Pima Indians Diabetic Dataset)

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

Real time dataset (Big Data)

The present investigation used simplified diabetes risk rating for the identification of newly diagnosed diabetic patients in various districts. The collected dataset was large data set in millions. Simulation is the method to do it which generates a sample that resembles smaller dataset.

IV. METHODOLOGY

1. Data Acquisition

The dataset acquisition comes under data collection. It is the process of collecting and measuring information on targeted variables in an established system, allowing one to answer relevant questions and evaluate results. Diabetic Big Data or data set is provided as input into the system. The structured or unstructured input data is obtained from Clinical programs and external sources such as government, laboratories, pharmacies, insurance companies, etc., in various formats. The appropriate attributes are used to analyze data and predict accurate results.

2. Data Pre-processing

Data pre-processing is deals with data preparation by data cleaning in order to improve the quality of the data for efficiency and ease of the mining process. Data preparation is the process of gathering, combining, structuring and organizing data so it can be used in business intelligence (BI), analytics and data visualization applications. The process of manage the incomplete data and irrelevant data and also manage the noisy data. It also handle the missing values kown as data cleaning.

3. Data Modelling

In data Modeling, data is transformed for the analysis. Once the data used for this study has been fully analyzed, the K-means (unregulated) and SVM (supervised) machine learning algorithms have been used and MapReduce to predict diabetes-related disease.
Hadoop is an open source framework used to store and

process large data sets and manages data duplication and node failure.

Hadoop consists of four main modules:

- Hadoop Distributed File System (HDFS) – HDFS provides much better data than standard file systems, in addition to high error tolerance and native support for large data sets.
- However Other Service Speaker (YARN) – Controls and monitors cluster nodes and resource usage. Organizes task and job.
- Map Reduce – A framework that helps programs perform parallel calculations. The map function takes input data and converts it into databases that can be computed in pairs of key value. Map output is used to streamline tasks to integrate output and give the desired result.
- Hadoop Common – Provides standard Java libraries that can be used in all modules.

Apache Spark is a distributed processing machine used for large records. It utilizes in-memory caching, and optimized query execution for instant analytic queries in against to information of any length. It supports code reuse throughout multiple workloads. It aggregates data and divide it throughout a server cluster, where it could then be computed and either moved to a one-of-a-kind statistics store or run thru an analytic model. The person does not need to define where precise files are sent or what computational resources are used to store or retrieve documents.

4. Data Evaluation

After data modeling, the evaluation was done using RStudio which was carried out through following techniques :

i. Classification techniques (C5.0)
ii. Regression Analysis (Linear regression)

C4.5 is the algorithm used to produce the decision tree used for division, so C4.5 is called the statistical classifier. This algorithm uses the advantage of information gain as a divition criteria. Can receive data with categories or numerical values. To manage continuous values, it produces a threshold and then divides the attributes by the values above the limit and the values equal to or below the limit. This C4.5 algorithm can easily handle missing values. Missing attribute values are not used in gain calculation in C4.5.

C5.0 is an extension of the C4.5 algorithm which is also an extension of ID3. It is a partition algorithm commonly used in the Big Data set. It is equivalent to C4.5 in speed, memory and efficiency. The C5.0 model not only works by separating a sample based on a field that provides maximum information gain but also separates samples by a large imfomation gain field.

Regression Analysis, It is a statistical method that allows you to examine the relationship between two or more variables and is the art of measuring straight lines in data patterns. Line regression analysis is used to predict the value of a variable based on the value of other variables. The variable want to predict is known as dependent variable. The variables use to predict the value of some variables are known as independent variables.

5. Data Validation

Data Validation was done with the following metrics :

i. Performance
ii. Pearson's Linear Correlation Coefficient

Performance

• Evaluate the performance of classification models, using accuracy (ACC), sensitivity (SN, also called memory) and specificity (SP).
• These measures can be calculated by the number of true positive(TP), false positive (FP), true negative (TN), and false negative (FN) in each category.

- Classification accuracy (ACU) is the most commonly used method of performance measurement. Accuracy is calculated by taking the ratio of the truly divided samples (true negative, actual positive) and the total number of samples.
  - Accuracy = really divided samples / number of samples
- Test methods, which are used to measure performance, are sensitivity and specificity. Sensitivity is calculated by dividing the true positive (TP) samples into the total positive sample (TP) and the false negative (FN) samples.
  - Sensitivity = TP / (TP + FN)
- Specification is calculated by separating negative (TN) samples from the total number of true negative and false (FP) samples.
  - Specificity = TN / (TN + FP)

Pearson's Linear Correlation Coefficient

The Pearson coefficient is the method used to present the linear correlation. It is mainly related to the linear relationship between random variables r, and the value of r is between – 1 and 1. The value of r from -1 to 0 indicates a perfect negative linear relationship between variables, 0 indicates no linear correlation between variables, and a scale of 0 to 1 indicates positive linear relationships between variables.

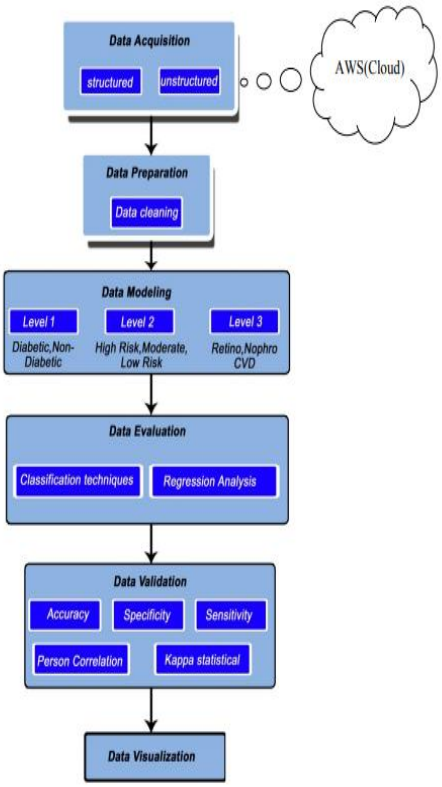


Figure 1 : Workflow of Methodology

Techniques used in modeling

1. **K-means**

K-Means Clustering is an algorithm used to solve clustering problems in machine learning or data science. Allows us to combine data into different groups and an easy way to find groups of categories on a non-labeled database without the need for any training.

Step-1: To determine the number of clusters, select the number K.

Step-2: Select Select K points or random centroids. (It can be something from the input database).

Step-3: Assign a data point to the nearest centroid, which will form pre-defined K clusters.

Step-4: Calculate the variance and place a new centroid for each collection.

Step-5: Repeat step 3 for redistributing each data point to the nearest new centroid for each collection.

Workflow of Methodology

Step-6: In the event of any redeployment, go to step 4 and go to FINISH.

Step-7: The model is ready.

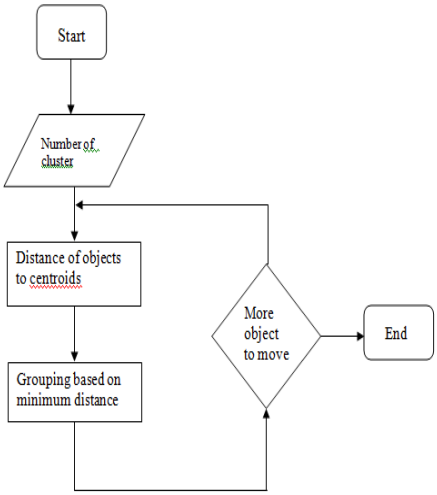


Figure 2 : Flow chart of K-means

2. MRK-SVM

The proposed algorithm is the hybrid model for the Big Dataset using the combination of clustering(K-means) and classification(SVM) techniques processed on Hadoop platform.

Steps to implement MRK-SVM algorithm:

Step 1: Read input data from HDFS.
Step 2: Input data is acquired by Mapper.
Step 3: Introduce the central location of the clusters and specify the nearest cluster for each data point.
Step 4: Set the location of each set in the description of all data points, otherwise repeat the steps above in two steps until they merge.
Step 5: Now the combined results data is used to create variables by creating a training database and creating our data model.
Step 6: Prediction is made on this new model and the test database where the database is trained.
Step 7: Finish the process with the predictive output file.

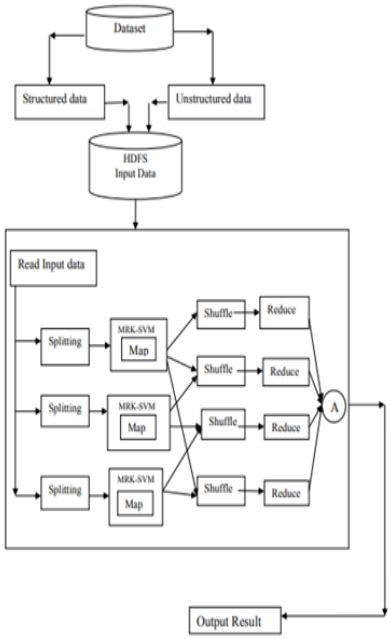


Figure 3 : Flow diagram of MRK-SVM

V. RESULT ANALISYS

1. Analysis of Diabetic small data in Hadoop MapReduce

The Pima Indian Diabetic data set with 768-instances was used as a small data set, loaded into the Hadoop framework in the file system input when MapReduce divides inputs into a few pieces. A 768 standard data set, the Pima Indian Diabetic data set, was loaded into the Hadoop framework in place of input files in the distributed file system where MapReduce divides the input into a few pieces. Map operations process input data and generate intermediate data.

The output of the map phase was sorted to produce an intermediate data. The reduce function processed the data in each partition and combine the intermediate values. The Hadoop framework collect the output of the reduce function to the output files on the disk. The dataset was classified as diabetic with 268 and non-diabetic with 500 in 16.842seconds.

2. Analysis of Diabetic Big Data in Hadoop MapReduce

The health care industry produce a extensively large amount of data so we can manage it using Big dataset models. Hence the Pima Indian diabetic dataset was replicated to 10, 100, 1000 times so that the data has been increased to 7680, 76800, 7,68,000 bytes. The map and reduce functions was loaded in the location of the input and output files in the Hadoop framework. MapReduce functions splits the submitted data in the input files into several pieces. The map tasks processed the input data and produced intermediate key value. The output of the map phase was sorted out and created an intermediate key. The processing of data in the each partition and merging of intermediate values are done by reduce functions. The output of the reduce function collected to the output files on the disk by the Hadoop framework. The replicated datasets 7680, 76,800, 7,68,000 bytes were classified as diabetic and non-diabetic within 80, 123 and 158 seconds respectively.

3. Analysis of Diabetic small data in RStudio

Consider a small Pima Indian Diabetic data set which contain 768 instances that loaded d into input csv file. Then R script was used to process the loaded data and remove the fuzzy data using data cleaning preprocess. The cleaned data was used for classifying the data into diabetic and non-diabetic. After processing and evaluation of the data, , the dataset was classified using the collected dataset. The dataset was classified as diabetic with 268 and non-diabetic with 500 in 15 seconds.

4. Analysis of Diabetic Big Data in RStudio

The health care industry produce a large amount of data about the patients. So it require a Big data set model to process these data. So the replicated Pima Indian diabetic datasets with size 7680, 76800, 768000 bytes in the form of csv file was loaded as input file into R. Then data cleaning preprocess was carried out so as to remove the fuzzy data. The cleaned data was used for diabetic and non-diabetic data classification process. After processing the data, evaluation was carried out. The replicated dataset with size 7680,76800 and 768000 bytes was classified as diabetic and non-diabetic with time as 68, 378 and 527 seconds respectively.

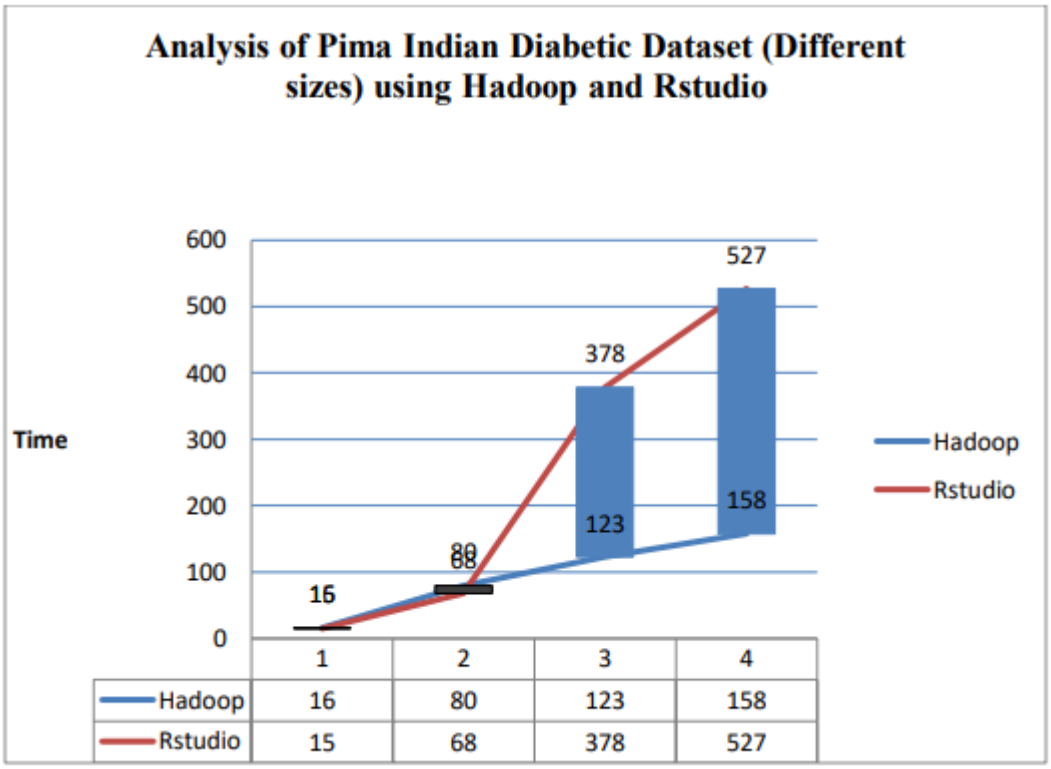
Table 1 : Analysis of small dataset using Hadoop and RStudio

Dataset Name	No. of Instances	Classification of Dataset		Tools	Time taken for Analysis (Seconds)
		Diabetic	Non-Diabetic		
Pima Indian Diabetic data set	768	268	500	Hadoop(MapReduce)	16.84
				RStudio	15

Table 2 : Analysis of BigData dataset using Hadoop and RStudio

Dataset Name	No. of instances	Tools (Time taken for analysis(Seconds)	
		Hadoop	RStudio
Pima Indian Diabetic data set	768	16	15
	7680	80	68
	76800	123	378
	768000	158	527

Result 1: Processing time of different size dataset using Hadoop and RStudio



VI. CONCLUSION

Hadoop MapReduce, Apache Spark and RStudio are Big Data tools, were tested with small and Big Data. when it is in real time data processing as well when the memory resources is sufficient Apache Spark was better among these for Big Data analysis. Apache Hadoop Map Reduce may help better When access amount of memory is required, considering huge performance gap.The speed of the processing time was achieved fastly by using MRK-SVM algorithm.RStudio is a programming language for statistical computing and was used to validate the models statistically for the predicted output. The Big Data was analyzed very fast by using Cloud due to its distributed and parallel processing in Hadoop environment available on Amazon Web Services. By using Cloud, the memory constraints in processing the Big data in standalone can be overcome the speed of processing was increased and the response time was decreased for processing of large datasets by using Cloud computing and MapReduce together.

VII. REFERENCES

1. Mrs.Ashwini Abhale ,Shruti Gulhane ,Sandhya Budhewar ,Swanali Jathar ,Harshada Sonwane , Predictive analysis of Diabetic Patient Data Using Machine Learning and Big Data,http://www.ijrat.org/downloads/Conference_Proceedings/NCRITSI-2K19/NCRITSI2K19-08.pdf

2. Minyechil Alehegn, Rahul Joshi, Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach, https://www.irjet.net/archives/V4/i10/IRJET-V4I1077.pdf

3. B. Suvarnamukhi, M. Seshashayee, Big Data Processing System for Diabetes Prediction using Machine Learning Technique , https://www.ijitee.org/wp-content/uploads/papers/v8i12/L35151081219.pdf

4. Ayman Mir, Sudhir N Dhage, Diabetes Disease Prediction Using Machine Learning on Big Data of Healthcare, https://ieeexplore.ieee.org/abstract/document/8697439

5. N. Yuvaraj, K. R. SriPreethaa, Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster,https://link.springer.com/article/10.1007/s10586-017-1532-x

6. http://docs.aws.amazon.com/AmazonS3/latest/dev/Welcome.html

7. http://hadoop.apache.org/

8. "Analysis of Diabetic Data Set using Hive and R", International Journal of Emerging Technology and Advanced Engineering, Sadhana and Savitha Shetty, 2014.

9. Vikram Phaneendra, E.Madhusudhan Reddy, "Big Data- solutions for RDBMS-Research Problems ". At the Network Operations & Management Symposium.

10. Albert Bifet, "Big Data Mine in Real Time" Informatica

11. Real-Time Healthcare Analytics on Apache Hadoop using Spark http://www.intel.com/content/dam/www/public/uen/documents/white-papers/big-data-real time health care-analyticswhitepaper .pdf.

12. Spark MLib, Apache Spark performance,https://spark.apache.org/mllib/.

13. Rajesh Jangade and Ritu Chauhan, Big data with integrated Cloud computing for healthcare anaytics computing for Sustainable Global Development (INDIA.Com),2016. http://ieeexplore.ieee.org/document/7725023/.

14. Ashish Ghosh, Aggregation pheromone density based data clustering, Information Sciences.

15. http://ftp.ics.uci.edu/pub/ml-repos/machine-learning databases/pima-indiansdiabetes, 2003. ISSN: 2278-0181

16. http://feliperego.github.io/blog/2015/10/23/Interpreting-Model-Output-In-R.

17. Samira Daneshyar and Majid Razmjoo, "Large-Scale Data Processing Using Mapreduce In Cloud Computing Environment", International Journal on WebService Computing (IJWSC), Vol.3, No.4, December 2012.

18. http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html

19. http://www.lifenscience.com/bioinformatics/sensitivity-specificity-accuracy

20. http://docs.aws.amazon.com/AmazonS3/latest/dev/Welcome.html

21. http://d0.awsstatic.com/whitepapers/Big_Data_Analytics_Options_on_AWS.pdf

22. K. Michael, and K. W. Miller, Big Data: New opportunities and New Challenges, IEEE Computer, 46 (6) (2013), pp. 22-24.

23. https://www.ijarcsse.com/docs/papers/Volume_4/5_May2014/V4I5-0391.pdf

24. Sagiroglu, S. Sinanc,D."Big Data: A Review".Collaboration Technologies and Systems (CTS), International Conference on, 42(47):20-24. 2013.

25. Garlasu, D.; Sandulescu, V. ; Halcu, I. ; Neculoiu, G. A Big Data implementation based on Grid Computing", Grid Computing.2013.