Diabetes Data Prediction Using Spark and Analysis in Hue Over Big Data

Bhargavi Chatragadda
Department of IT
VRSEC
Vijaywada, India
chatragaddabhargavi24@gmail.com

Supriya Kattula
Department of IT
VRSEC
Vijaywada, India
supriyakattula32@gmail.com

Geetha Guthikonda
Department of IT
VRSEC
Vijaywada, India
geetaguttikonda@gmail.com

***Abstract*— This paper deals in diabetes prediction by applying relevant data mining technique. The goal of data mining is knowledge extraction from the information that is stored in the dataset and also analyzing patterns. In the civilization there are number of health problems facing by the people and are not aware of the symptoms about why the sickness occurs. One of the health problems is the Diabetic Mellitus. The majority of the population is in front of with Diabetes. Now a day's people at younger age are also suffering with this problem. In this paper, we have used predictive analysis in HUE to foresee the diseases compose persistent and particulars related with it and the sort of behavior to be given. The dataset is collected from the Pima Indian database. This framework along with SVM Classification gives an effective method to count the number of persons who are suffering with diabetes.**

***Keywords*— Hadoop, Big Data, Predictive analysis, Spark**

I.INTRODUCTION

Due to unhealthy lifestyle in overeating and lack of physical activity lead to diabetes. That is due to high glucose in the body with uncontrolled food habits it is becoming more common problem in children. The human services information incorporates Electronic Health Reports (EHR) of patient's information, medical reports, specialist's solution, symptomatic information, therapeutic pictures, drug store data, medical coverage related information, information from social Medias and restorative diaries [2]. All these data all in all structures Big Data in social insurance. By utilizing the investigation of enormous information will deliver the anticipated outcomes for accepting the prototype to enhance the medicinal services and existence hope, appropriate treatment at beginning times requiring little to no effort. There are four parameters related with enormous information and is portrayed by four traits: volume, velocity, variety and veracity [3].

The human services industry is moving from announcing realities to revelation of bits of knowledge, toward getting to be data driven social insurance associations. Enormous statistics holds incredible potential to change the entire social insurance esteem chain from tranquilizes examination to patients minding quality [1]. Diabetes Mellitus or simply called Diabetes is the ailment in which the human body does not generate proper amount of insulin [4]. It is a standout amongst the most diligent sicknesses. It is caused for the most part due the inadequacy of insulin in the body. This insulin is created by the pancreas. There are mainly three types of diabetes mellitus. Type1 is one of the diabetes where the pancreas neglects to create enough insulin in the body which pulverizes the invulnerable framework. It is called as insulin-subordinate diabetes. It can be seen generally in youngsters and adolescents. Type2 is one of the diabetes where the pancreas delivers just some insulin in the body which isn't adequate for the body. It is called as insulin-protection diabetes. It is generally seen in grown-ups. The treatment for diabetes should be possible by utilizing indicated sedates and can't be forestalled for the most part. Type3 is the Gestational Diabetes which occurs mostly in the pregnant women. Gestational diabetes is generally developed in the sixth month i.e between 24 and 28 weeks and disappears after the baby is born. So women with this type of diabetes are more likely to have the type2 diabetes later in their life [4]. Because of the developing amorphous life of diabetic information shape wellbeing industry or every single other source, it is important to structure and accentuation the dimension into ostensible incentive with conceivable arrangement. Patient need to measure the glucose level in the blood and concentrate on their health at least 2 times a month for safety measures. The diabetes is the major problem the person suffers most. For the treatment of diabetes for a patient it must include insulin measurement and on the advice of a specialist, self checking of glucose level in the blood must be compulsory. For this the related data of a patient should be grouped in one record. Conveying a Health Information Exchange (HIE) can extricate medical data from a few unique archives and coordinate that information inside a solitary patient wellbeing record that all care suppliers can get to safely. Prescient examination is a strategy that joins an assortment of methods from information mining, insights, and diversion hypothesis that utilize the present and past information with measurable or other diagnostic models and techniques, to decide or foresee certain future occasions [5].

II.RELATED WORK

There are number of factors which the diabetes effects the body. So, Data mining is used by numerous individuals to create different expectation models utilizing information anticipate diabetes. New predictions were made based on patterns that were analyzed [6]. The traditional neural system is utilized for forecast, on the prescribed dataset. In prescient investigation of diabetic treatment utilizing relapse support information mining procedures to diabetes information, they find designs utilizing SVM calculation that distinguish the most excellent method of conduct for treatment crosswise over various age [7]. At an early stages drug treatment for patients

978-1-5386-2440-1/18/$31.00 ©2018 IEEE

must be inferred. Proper diagnosing can lower cost and in the old age gathering ought to be endorsed medicate treatment instantly. Anticipation and alignment of diverse sort of diabetes utilizing C4.5 order calculation was completed using Pima Indians Diabetes Database [4]. Point by point investigation of the Pima diabetic informational index was completed proficiently utilizing of Hive and R. In this investigation we can determine a few fascinating certainties, which can be utilized to build up the forecast models [8]. The prescient investigation works in three regions, for example, Operations administration, Medical administration and biomedicine, and System outline and arranging.

Selection of huge information in medicinal services fundamentally expands security and patients essential concerns and prescribes the patient data to be put away in information focus' with fluctuating levels of security. The study of New England Journal of Medicine tells that one out of five patients experience the ill effects of preventable readmissions. Therefore, 1% of the populace represents 20% of all US medicinal services uses nearly and 25% for more than 80% of all consumptions [9]. Different huge information innovation stack and research over human services joined with proficiency cost reserve funds, etc., are

Clarified in better social insurance. The Hadoop utilization in human services turned out to be more critical to practice the data and to embrace the expansive scale information administration exercises. The examination on the consolidated process and capacity can elevate the expenditure adequacy to be picked up utilizing HUE [10].

BIG DATA

The process used for the traditional mining of information is called the Big data. The data that is not processed by the relational databases can be structured by using Big data, that utilizes parallel concepts. The raw data has no value so it should be processed so as to be valuable. Now a days, in any industry big data has the way to analyze and control the information [11]. It reduces the cost of treatment by predicting Outbreaks, avoiding unnecessary syndromes so as to improve the quality of life. Enormous Data is ordinary information that is immense in estimate with heaps of data in various configuration and heaps of clamor that can't be mined utilizing the customary framework[12]. Sam Madden expressed that the information are too huge, too quick, too hard and excessively perplexing, making it impossible to examine with the current framework which is known as Big Data. The way toward putting away, dissecting, overseeing and envisioning the information is exceptionally troublesome. As per Marko Grobelnik Big Data is fundamentally the same as Small-information, Big Information requires a totally new devices and strategies to investigate and take care of numerous true issues in a superior and a proficient way.

HADOOP

The essential tool that is resourcefully useful for backup and analyzing huge amount of information is Hadoop. Because of its uniqueness capability novel methods can be used for health care information to reduce the economy for improvement in analysis. The clinical information in Healthcare is purely unstructured. So it is an opportunity to find insights and analyze to handle large amount of data.

HUE:

HUE stands for Hadoop User Experience. It is a open source web interface that supports Apache Hadoop and its eco systems and a workbench for browsing, querying and visualizing the data.

HIVE:

The complicated analysis for the Map Reduce is done in Hive QL. The basic SQL operations are used in HIVE. Hive does not support OTP and row wise updates. It is used to process structured data in Hadoop. It is a platform to develop SQL type scripts to do Map Reduce operations.

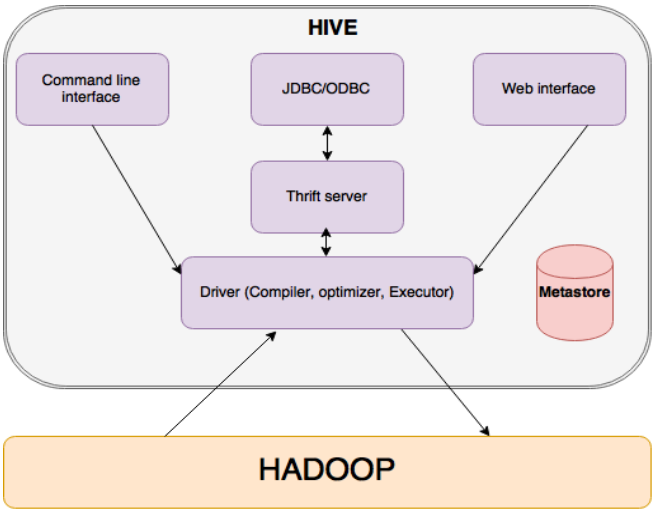


*Fig*1: Hive Architecture

SPARK:

The general open source engine for large data processing is Spark. In Memory, it runs hundred times faster than Hadoop Map reduce. This is the main feature of Spark. It can run on Hadoop's Yarn technology. Parallel apps can be developed using 80 high-level operators. Apache Spark is an extremely quick group figuring innovation, intended for quick calculation. It depends on Hadoop Map Reduce and it stretches out the Map Reduce model to productively utilize it for more kinds of calculations, which incorporates intuitive inquiries and stream preparing. The fastest cluster computing framework is Apache Spark, which is designed for fastest estimation. Is has the foundation of Hadoop Map Reduce and also extends the Map reduce representation for further computations. This increases the speed of the processor for an application. There are number of features which support Spark. Speed, supports multiple languages, data streaming, Machine learning etc. The basic data structure of Spark is Resilient Distributed Dataset. For parallel handling and creating large dataset, Map Reduce is implemented. Defect tolerance and work distribution can be easy with parallel

computation using high-level operators. The large amount of workload can be covered by Spark. To improve the quality of health care Apache Spark is the heartbeat of many applications. Based on the Past medical information, Apache Spark is used to analyze the patient's data for the health issues people are facing.

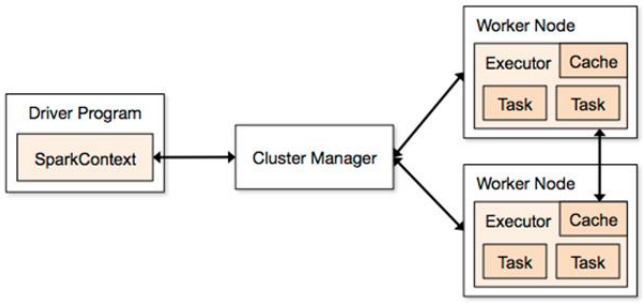


Fig 2: Spark Architecture

III. METHODOLOGY

The following is the flow diagram. Here we take the input from the PID (Pima Indian Dataset). The collected data is sent into the data warehouse for extracting, Loading and processing the data.

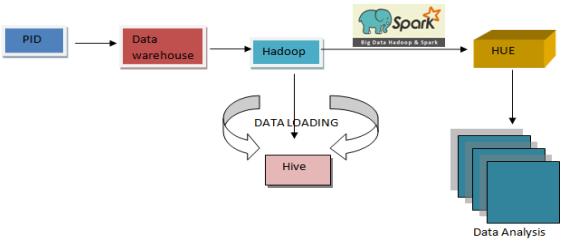
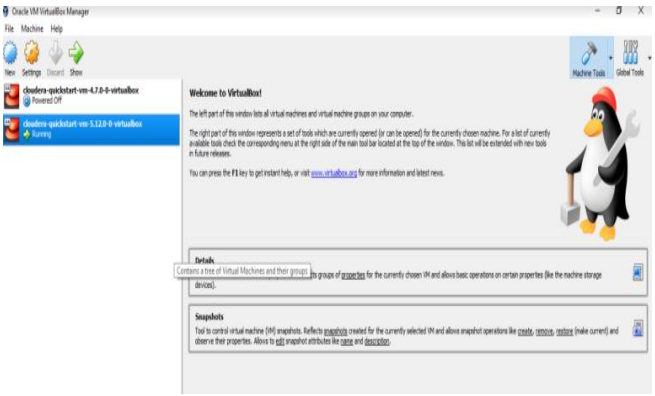


Fig 3: Flow Diagram

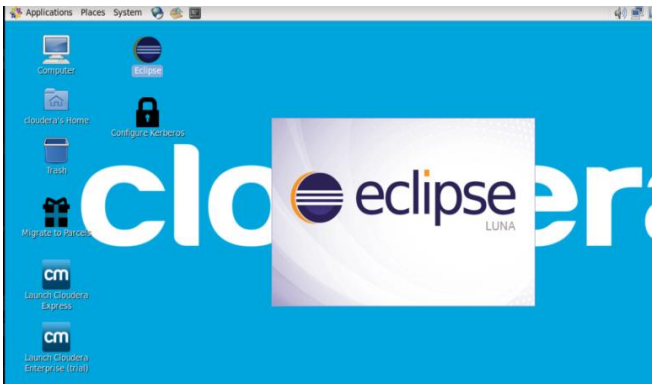
Then the code is written in Java in Hadoop in which the predictive analysis algorithm is used to analyze the clinical dataset. We write queries in hive and create the directories in Hue.

IV. IMPLEMENTATION

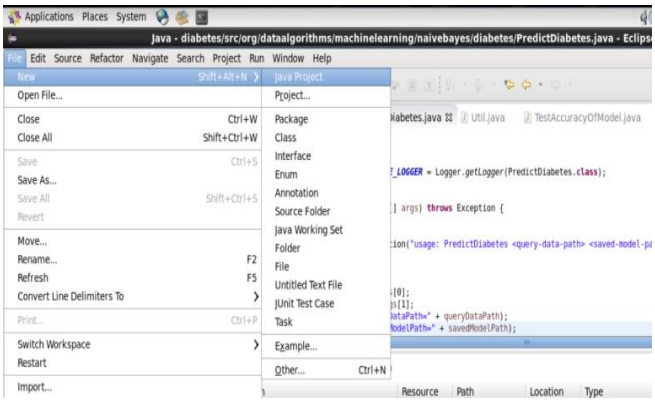
Step1: Install Cloudera-quickstart-vm 5.12.0.0 on your computer.



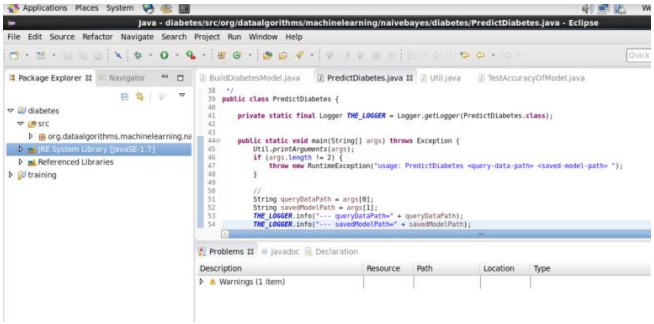
Step 2: After installation, open the cloudera window and click on eclipse icon.



Step 3: Create a new project and write the code.

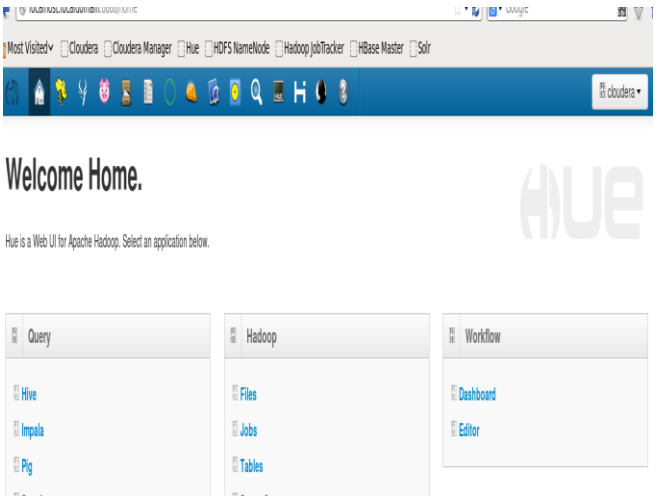


Next create a new package, new class, import the jar files and add external jars.

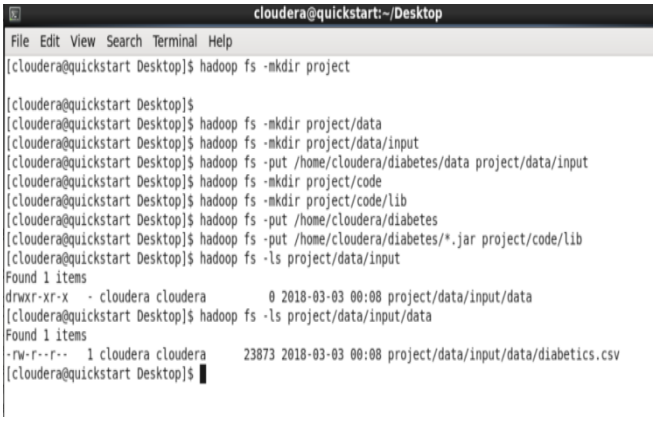


Step 4: Then open Hue and give the login details like username as cloudera and password cloudera.

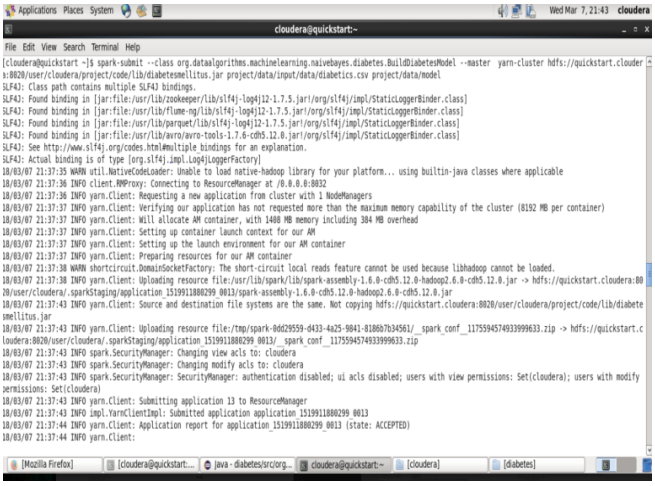
Step 5: The hue homepage appears.



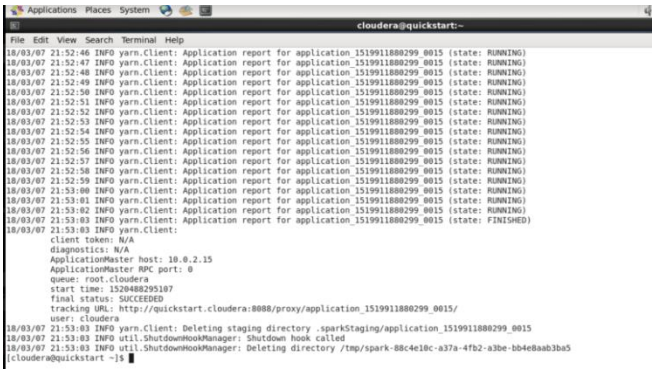
Step6: In hue home page click file browser and click new button and select Directory**.** We can also create the Directory in the cloudera terminal and then describe the dataset.

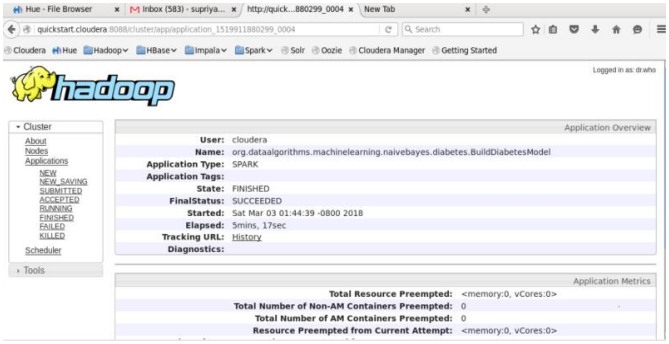


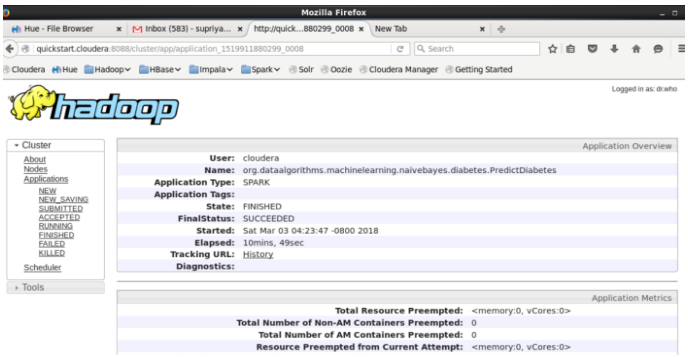
Step 7: Then run the query for the Build Model program with its path.



Step 8: Next run the query for the Predict Model program.

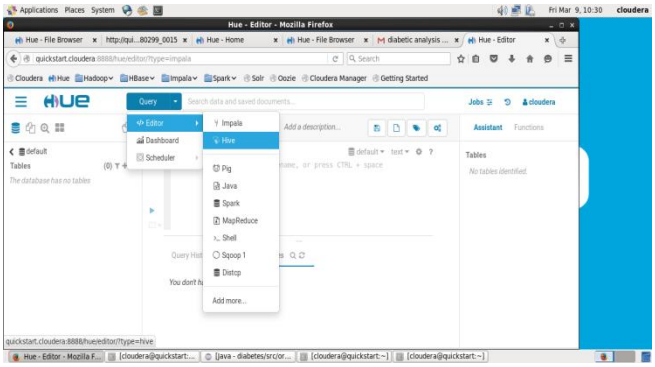




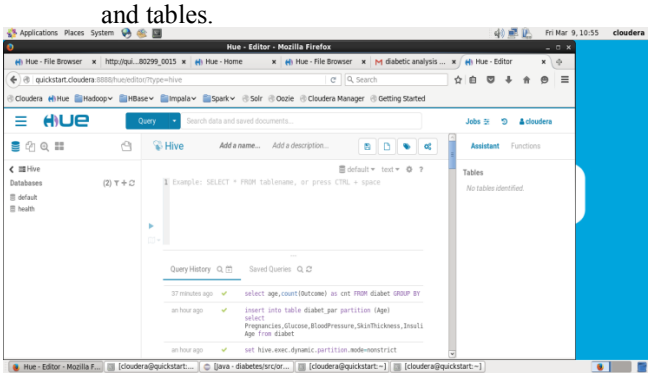


The results shows that the clusters are succeeded and now we show the analysis part in the Hue environment by writing queries in hive.

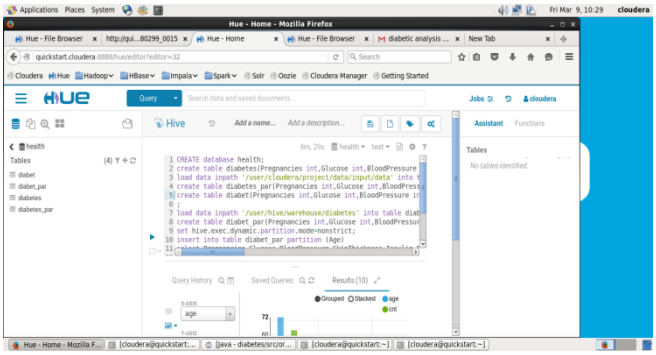
Step9: Open the Hue browser and click on Query->Editor->Hive



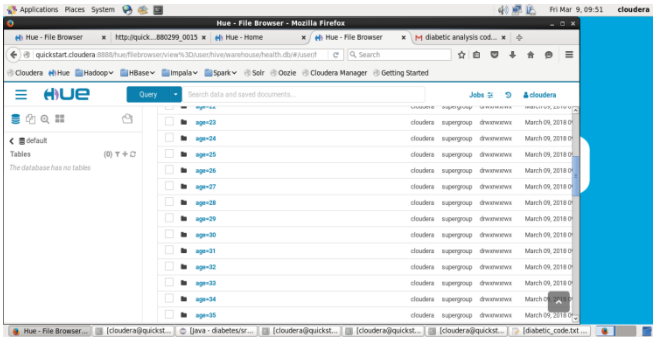
Step10: The query editor opens and then creates the database and tables.



Step 11: After creating tables load the dataset into the table and then partition the table.

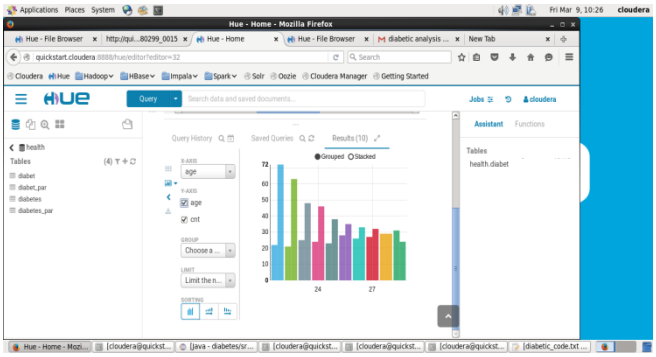


Step12: Give the query as “set.hive.dynamic.partition” for the multiple partitions.



V.RESULTS AND OBSERVATIONS

Based on the queries given we get the result in the form of graphs.



Based on the dataset and the graphs obtained, we analyse that the 72 patients have diabetes mellitus at the age 22. 63 patients have diabetes mellitus at the age 21. 48 patients have diabetes mellitus at the age 25. 29 patients have diabetes mellitus at the age 29.

VI.CONCLUSION

The objective of study investigates diabetic treatment in health care industry utilizing huge information examination. The outline of prescient investigation arrangement of diabetic treatment may give progressive information and analysis

capitulate the best outcomes in medicinal services. By utilizing Spark we have anticipated the diabetes composes predominant, which sexual orientation and race has greater plausibility for being influenced by Diabetes. Apache Spark can also be used to process genomic sequence by reducing the point in time. Spark can also be used with R in the future.

REFERENCES

[1] Dr. Saravana Kumar NM ,Eswari T,Sampath P, Lavanya S, "Predictive Methodology for Diabetic Data Analysis in Big Data" in 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15).

[2] Muni kumar N, Manjula R,"Role of Big Data Analytics in Rural Health Care – A Step Towards Svasth Bharath", International Journal of Computer Science and Information Technologies, vol 5(6), pp 7172-7178, 2014.

[3] Wullianallur Raghupathi, and Viju Raghupathi, "Big data analytics in healthcare: promise and potential", Health Information Science and Systems, vol. 2(3) pp. 2-10, 2014.

[4] K. Rajesh, V. Sangeetha, "Application of Data Mining Methods and Techniques for Diabetes Diagnosis" in International Journal of Engineering and Innovative Technology (IJEIT) Vol 2(3), 2012.

[5] Nishchol Mishra, Dr.Sanjay Silakari, "Predictive Analytics: A Survey, Trends, Applications, Oppurtunities & Challenges", International Journal of Computer Science and Information Technologies, vol. 3(3), 4434-4438 4434, 2012.

[6] V. H. Bhat, 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..

[7] Abdullah A. Aljumah, Mohammed Gulam Ahamad, Mohammad Khubeb Siddiqui, "Application of data mining: Diabetes health care in young and old patients", Journal of King Saud University – Computer and Information Sciences, vol. 25, pp. 127–136, 2012.

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

[9] Andrew Pearson, Qualex Asia, "Predictive Analytics for the Healthcare Industry", Andrew Pearson, Qualex Asia Limited, 2012.

[10] D. Peter Augustine, "Leveraging Big Data analytics and Hadoop in Developing India's Health Care Services", International Journal of Computer Applications, vol 89(16), pp 44-50, 2014.

[11] Monica Korlapati Geetha Guttikonda, Sneha Cherukuri, Chandra Naga Sravanthy, Mohammad Irfanullah "PREDICTION OF HEART DISEASE AND STRATEGIC DECISION MAKING FOR PHI OF MEDICAL DATASET" International Journal of Latest Trends in Engineering and Technology Volume 8 Issue 3 Pages 45-50

[12] Prasanna Komara Geetha Guttikonda, Madhavi Katamaneni "PROVIDING BETTER MEDICAL HEALTHCARE SERVICES USING DATAMINING TECHNIQUES" IJIET Volume 7 Issue 2 Pages 491-496