Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 11, November-2019

Smart Health Care Prediction for Cardiovascular Disease-A Systematic Review

Sathya Balaji
Assistant Professor, CSE
KPR Institute of Engineering and Technology,
Coimbatore, India

***Abstract*—Big Data alludes to huge amounts of data or pieces of information made by the digitization of everything that gets merged and broken down by explicit advancements. Applied to human services, it will utilize explicit wellbeing information of a populace (or of a specific individual) and conceivably help to prevent epidemics, fix ailment, and chop down costs. Big Data application has a ton of positive and life sparing outcomes. Big data characterizing features incorporate the capacity to deal with massive volume of information and variety at the high speed. Predictive analysis assists doctors with settling on information driven choices within seconds and improve treatment for patients. It should be possible for the patients with complex chronicles, and experiencing various conditions. Heart disease is a pervasive illness which causes demise over the world. A ton of research is going on predictive analytics using machine learning techniques to uncover better choice making. Big Data analysis encourages extraordinary chances to anticipate future wellbeing status from wellbeing parameters and give best outcomes. Naive Bayes method can be utilized for the heart disease prediction in patients and make forecasts on the test information to predict the classification. It gives probabilistic classification based on Bayes theorem between the highlights. The result reveals that it gives better accuracy to predict the diseases. Hadoop spark can be used as big data computing tool on healthcare data. The results in prediction ensured that the system improved significantly in terms of accuracy, CPU utilization, and processing time. It can be done to predict different patients' future health conditions. It takes the training dataset to estimate the health parameters necessary for classification. The results show the early disease detection to figure out future health of patients.**

Index Terms— Big Data, Machine Learning, Naïve Bayes' Theorem, Prediction, Hadoop-Spark

I. INTRODUCTION

Healthcare big data refers to gathering, breaking down and utilizing consumer, patient, physical, and clinical information that is excessively immense or complex to be comprehended by conventional means of data processing. Rather, big data is regularly handled by machine learning algorithms and data scientists [1]. The ascent of healthcare big data comes in light of the digitization of medicinal services data and the ascent of significant worth based consideration, which has urged the business to utilize information investigation to settle on vital business choices. Faced with the difficulties of healthcare data – such as volume, velocity, variety, and veracity – health systems need to embrace innovation equipped for gathering, storing, and dissecting this data to deliver significant bits of knowledge.

Big data has become progressively powerful in healthcare because of three significant moves in the social insurance industry: the tremendous measure of information accessible, developing medicinal services costs, and an emphasis on commercialization. Big data empowers wellbeing frameworks to transform these difficulties into chances to give customized tolerant adventures and quality care.

Heart attack is one of the most well-known diseases faced by the patients these days. The fundamental goal of dealing with the cardio vascular disease is to survey the huge number of datasets, analyze and absorb the data that is utilized for predict, manage and treat the chronic diseases such as heart attacks. Data mining, Visualization and Hadoop are the tools or techniques used for handling large volume of data's in datasets [2][3].

The primary purpose for heart attack is a blockage which causes blood stream to one of the coronary arteries, vital channels through which blood travels to the muscles of the heart, to become diminished or blocked. It makes the heart muscles to get quickly denied of red platelets which take the necessary oxygen necessary for sustaining life and awareness in the human body. Within a few minutes, heart muscles got arrested, which prompts to the individuals death. A hard substance called Plague which is comprised of numerous cells and cholesterol (fat) is formed and it causes the blockage of coronary arteries [6].

Big Data guarantees large scale examinations of results, designs, temporal trends, and correlations at a population level. This sort of investigation helps in distinguishing epidemic outbreaks, foreseeing high-risk patients, listing frequency of re-admissions, and assigning triage (treatment order) to approaching patients. EHRs are immersed with quantitative (lab result and medical test values), qualitative (text-based documents, pictures etc.) and value-based (visits record, prescription record) information. Given the abundance of rich datasets that are stored, there is degree for a lot of interoperability within health and social frameworks. The integration of traditional patient related longitudinal information (prescription history, complaint list, family health history) with social determinants of health (wages, education, caste/religion, residence state) offers an perceptive and flawless solution for identifying and reducing public-health ills. With such a system in place, the patient can not only have access to personalized

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 11, November-2019

care but also be an active stakeholder in his health and well-being[8]. Moreover, in the future, incorporation of systems biology – the demonstration of complex biological frameworks – with EHR data can provide a clear way for customized medication, treatment and care that is exclusive to every patient. Therefore, supervision of patients, hospital supply chains, medical manpower, asset assignment, and minimization of unnecessary expenses are some significant value additions that big data brings to the health sector and delivery of care services to the people. Given the lack of resources, lack of skilled medical practitioners and increasing economic efficiency, the case for the consumption of big data to search countless data sets for gathering crucial trends and experiences is more squeezing than any time in recent memory.

Smart healthcare supports and add to the basis of big data, and contains multiple products such as home health care, wearable health care, and bio-transplant health care. For patients needing monitoring at home, home health care systems use a sensor that is installed in the home that manages the individuals' health along with their smart phones. In case of wearable health care, sensors are worn on the human body, providing personalized service through the measurement, transmission and analysis of biological signal of the user's body in real time.

II. ANALYSIS OF BIG DATA USING HADOOP

Apache Hadoop is an open source software structure for storage and huge scale processing of data-sets on clusters of hardware. It provides a software framework for distributed storage and processing of big data using the MapReduce programming model. It was originally intended for computer clusters built from commodity hardware and also found use on clusters of higher-end hardware. All the modules in Hadoop are structured with a basic statement that hardware malfunction are normal events and ought to be consequently dealt with by the framework [3].

Hadoop runs a small number of applications on distributed systems with huge number of nodes involving petabytes of data. It has a distributed file system, called Hadoop Distributed File System or HDFS, which empowers quick information move among the nodes.

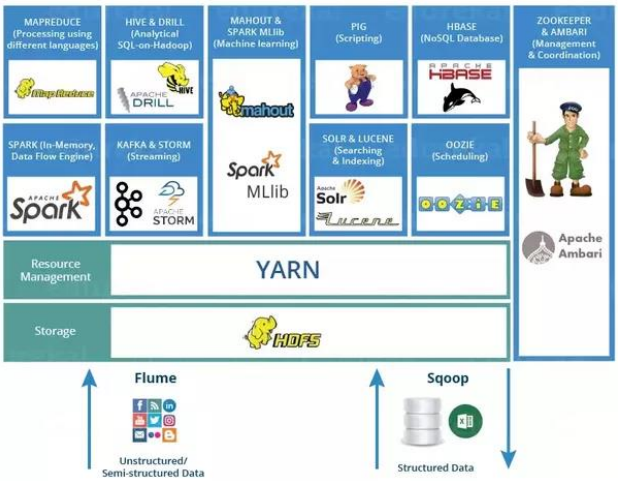

Fig 1: Apache Hadoop framework

Hadoop Distributed File System (Hadoop HDFS):

HDFS provides a storage layer for Hadoop. It is appropriate for distributed storage and handling, i.e. while the information is being stored it first get disseminated & then it proceeds. HDFS Provides a direction line interface to associate with Hadoop. It provides streaming access to file system datawhich includes file permission and authentication[4].

HBase:

It facilitate to store data in HDFS. It is a NoSQL database or non-relational database. HBase mainly used when you need random, real-time, read/write access to your big data. It provides support to the high volume of data and high throughput. In HBase, a table can have a large number of sections.

Sqoop:

A sqoop is a tool intended to move data between Hadoop and NoSQL. It is managed to introduce data from relational databases such as Oracle and MySQL to HDFS and export data from HDFS to relational database.

Flume:

Flume gathers the event data and transfers it to HDFS. It is preferably appropriate for event data from multiple systems. After the data is moved into HDFS, it is processed and the processed data is moved into SPARK.

SPARK:

An open source cluster computing structure which provides 100 times quicker performance as compared to MapReduce. Spark run in the Hadoop cluster & process data in HDFS.

Spark has the following major components:

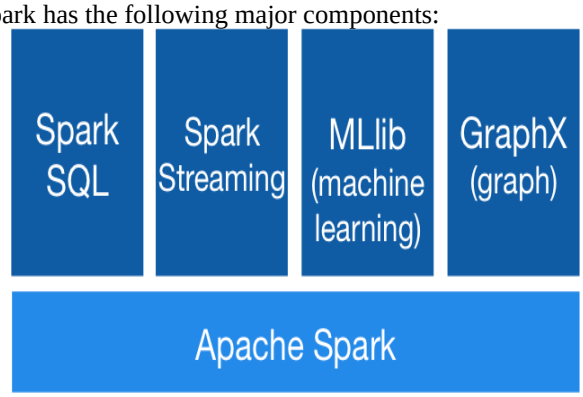

Fig 2: SPARK Framework

Hadoop MapReduce:

Another structure that processes the information based on JAVA and the MapReduce programming model. Many tools such as Hive, Pig build on Map Reduce Model. It is wide & mature fault tolerance structure and the most usually utilized framework. After the data processing, an analysis is done by the open-source data flow system called Pig[3][4].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 11, November-2019

Pig:
An open-source dataflow structure which is primarily used for Analytics and translate pig script to Map-Reduce code and save producer from writing Map-Reduce code.

Impala:
It is a high-performance SQL engine which runs on a Hadoop cluster. It is best for interactive analysis and has a very low latency which can be calculated in milliseconds. Impala supports a dialect of a sequel. So, data in HDFS modeled as a database table.

Hive:
It is an abstraction wrap on top of the Hadoop. It's very similar to the Impala and ideal for data processing and ETL (extract, transform and load) operations. Impala is chosen for ad-hoc queries, and hive executes queries using Map-Reduce. Hive is appropriate for structured information.

Hue:
Hue is an short form for Hadoop user experience. It is an open-source web interface for analyzing data with Hadoop. We can execute the following operations using Hue.
1. Upload and browse data
2. Query a table in Hive and Impala
3. Run Spark and Pig jobs
4. Workflow search data.

Hue makes Hadoop accessible to use. It also offers an editor for the hive, impala, MySQL, Oracle, Postgre SQL, Spark SQL and Solar SQL.

Four stages of big data processing:

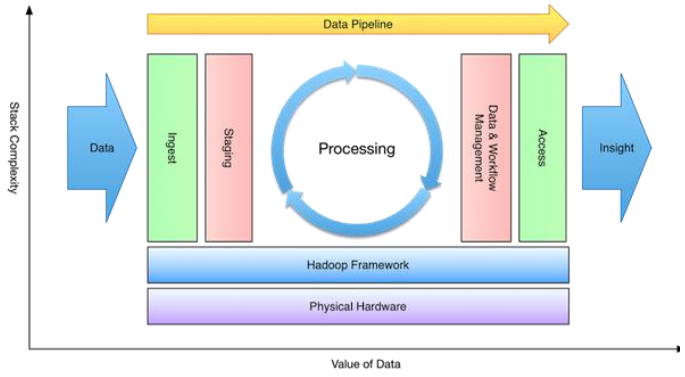

Fig 3: Stages of Big data Processing

The first stage is Ingested, where data is ingested or transferred to Hadoop from a variety of resources such as relational databases system or local files. Sqoop moves data from our RDMS (Relational Database) to HDFS[5][3].
The second stage is processing. In this stage, the data is stored and processed. Spark and MapReduce perform data processing.
The third stage is analyzing and the data is interpreted by the processing framework such as Pig, Hive & Impala. Pig convert the data using Map and Reduce and it is more apt to structured data.
The fourth stage is accessed which is performed by a tool such as Hue and Cloudera search. In this stage, the analyzed data can be accessed by users. Hue is web-interface for exploring data.

III. NAIVE BAYES CLASSIFIER IN BIG DATA

Naive Bayes is a probabilistic method for creating classifiers. The characteristic statement of the Naive Bayes classifier is to think about the value of a particular characteristic which is independent of the value of any other characteristic, specified with the class variable[7].
In spite of the oversimplified statement mentioned previously, naive Bayes classifiers have good consequences in complex real-world circumstances. A benefit of naive Bayes is that it only need a small amount of training data to calculate approximately the parameters essential for classification and that the classifier can be trained incrementally[8].
Naive Bayes is a conditional probability model: given a crisis case to be classified, represented by a vector x = (x1, …, xn) in place of some n features (independent variables), it gives to this instance probabilities for each of K possible results or classes.

$$p(C_k|x_1, \ldots\ldots, x_n)$$

The crisis with the above formulation is that if the number of features n is large or if a feature can take on a huge number of values, then basing such a model on probability tables is infeasible. We therefore reformulate the model to make it simpler. By means of Bayes theorem, the conditional probability can be decomposed as

$$p(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)}$$

This means that beneath the above independence assumptions, the conditional distribution above the class variable C is

$$p(C_k|x_1, \ldots\ldots, x_n) = \frac{1}{Z}p(C_k)\prod_{i=1}^{n} p(x_i|C_k)$$

where the evidence Z = p(x) is a scaling factor dependent only on x1, …, xn, that is a stable if the values of the feature variables are recognized. One frequent law is to pick the theory that is most feasible; this is known as the maximum a posteriori or MAP decision rule. The corresponding classifier, a Bayes classifier, is the function that assigns a class label as follows

$$\hat{y} = argmax\, p(C_k)\prod_{i=1}^{n} p(x_i|C_k)$$

Implementing the algorithm in R is a easy process. The following case reveals how train a Naive Bayes classifier and use it for prediction in a spam filtering crisis.
By using Naives bayes, hidden knowledge can be uncovered and removed along with diseases (heart attack, cancer and diabetes) from a historical heart disease database. It can reply difficult queries for analyzing disease

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 11, November-2019

and thus help healthcare practitioners to make smart clinical choices which traditional decision support systems cannot. By giving efficient treatments, it also assist to reduce the treatment costs[7][8].

IV. CONCLUSION

In this review paper, an overview is provided on Big Data, Hadoop and Naïve Bayes Classifier. An overview to big data test is specified and a variety of chance and applications of big data has been discussed. This paper explains the Hadoop Framework and its components HDFS and Map reduce. The Hadoop Distributed File System (HDFS) is a distributed file system planned to run on hardware. Hadoop plays an significant role in Big Data. Naive Bayesian Classification technique eliminates the unseen knowledge from a precedent heart disease database. This is the well-organized model to predict patients with heart disease. This model could answer complex question, each with its own potency with respect to ease of model explanation, access to detailed information and accuracy.

V. REFERENCES

[1] S.Vikram Phaneendra & E.Madhusudhan Reddy "Big Data-solutions for RDBMS problems-A Survey" In 12th IEEE/IFIP Network Operations & Management Symposium (NOMS 2010) (Osaka, Japan, Apr 19{23 2013).

[2] Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N " Analysis of Bidgata using Apache Hadoop and Map Reduce" Volume 4, Issue 5, May 2014" 27

[3] Aditya B. Patel, Manashvi Birla, Ushma Nair, (6-8 Dec. 2012),"Addressing Big Data Problem Using Hadoop and Map Reduce".

[4] Kyong-Ha Lee Hyunsik Choi "Parallel Data Processing with Map Reduce: A Survey" SIGMOD Record, December 2011 (Vol. 40, No. 4)

[5] Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, "Shared disk big data analytics with Apache Hadoop", 2012, 18-22

[6] Mrs.G.Subbalakshmi, "Decision Support in Heart Disease Prediction System using Naive Bayes", Indian Journal of Computer Science and Engineering.

[7] Fayyad, U: "Data Mining and Knowledge Discovery in Databases: Implications for scientific databases", Proc. of the 9th Int. Conf. on Scientific and Statistical Database Management, Olympia, Washington, USA, 2-11, 1997.

[8] Giudici, P.: "Applied Data Mining: Statistical Methods for Business and Industry", New York: John Wiley, 2003.

[9] Han, J., Kamber, M.: "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006.