AN EFFECTIVE DATA MANAGEMENT FRAMEWORK FOR HEALTHCARE: BIG DATA PERSPECTIVE

A Dissertation

Submitted

In Partial Fulfillment of the Requirements

for the Degree of

Master of Technology

In

Advanced Computing and Data Science

Submitted by

Sakshi Raj Singh

Roll No. 2001019002

Under the Supervision of

Dr. Mohammad Zunnun Khan

(Assistant Professor)

Department of Computer Science & Engineering

Faculty of Engineering

INTEGRAL UNIVERSITY, LUCKNOW, INDIA

July, 2022

INTEGRAL UNIVERSITY
इंटीग्रल विश्वविद्यालय

Accredited by NAAC. Approved by the University Grants Commission under Sections 2(f) and 12B of the UGC Act, 1956, MCI, PCI, IAP, BCI, INC, CoA, NCTE, DEB & UPSMF. Member of AIU. Recognized as a Scientific & Industrial Research Organization (SIRO) by the Dept. of Scientific and Industrial Research, Ministry of Science & Technology, Government of India.

CERTIFICATE

This is to certify that **Sakshi Raj Singh** (Enroll. No. 2020101956) has carried out the research work presented in the dissertation titled **"An Effective Data Management Framework for Healthcare: Big Data Perspective"** submitted for partial fulfillment for the award of the **Master of Technology Advanced Computing and Data Science** from **Integral University, Lucknow** under my supervision.

It is also certified that:

(i) This dissertation embodies the original work of the candidate and has not been earlier submitted elsewhere for the award of any degree/diploma/certificate.

(ii) The candidate has worked under my supervision for the prescribed period.

(iii) The dissertation fulfills the requirements of the norms and standards prescribed by the University Grants Commission and Integral University, Lucknow, India.

(iv) No published work (figure, data, table etc.) has been reproduced in the dissertation without express permission of the copyright owner(s).

Therefore, I deem this work fit and recommend for submission for the award of the aforesaid degree.

Dr. Mohammad Zunnun Khan
Dissertation Guide
(Assistant Professor)
Department of CSE,
Integral University, Lucknow
Date:
Place: Lucknow

DECLARATION

I hereby declare that the dissertation titled "**An Effective Data Management for Healthcare: Big Data Perspective**" is an authentic record of the research work carried out by me under the supervision of **Dr. Mohammad Zunnun Khan**, Department of Computer Science & Engineering, for the period from September2021 to july2022 at Integral University, Lucknow. No part of this dissertation has been presented elsewhere for any other degree or diploma earlier.

I declare that I have faithfully acknowledged and referred to the works of other researchers wherever their published works have been cited in the dissertation. I further certify that I have not willfully taken other's work, para, text, data, results, tables, figures etc. reported in the journals, books, magazines, reports, dissertations, theses, etc., or available at web-sites without their permission, and have not included those in this M. TECH dissertation citing as my own work.

Sakshi Raj Singh

(Enroll. No. **2020101956**

RECOMMENDATION

On the basis of the declaration submitted by “**Sakshi Raj Singh**”, a student of M.Tech CSE (Advanced Computing and Data Science), successful completion of Pre presentation on 24/6/2022 and the certificate issued by the supervisor, **Dr. Mohammad Zunnun Khan**, Assistant Professor, Computer Science and Engineering Department, Integral University, the work entitle An Effective Data Management For Healthcare :Big Data Perspective”, submitted to department of CSE, in partial fulfillment of the requirement for award of the degree of Master of Technology Advanced Computing and Data Science, is recommended for examination.

____________________________	____________________________
Program Coordinator Signature	HOD Signature
Dr. Faiyaz Ahamad	**Mrs. Kavita Agrawal**
Dept. of Computer Science &Engineering	Dept. of Computer Science & Engineering
Date:__________________	Date:____________________

COPYRIGHT TRANSFER CERTIFICATE

Title of the Dissertation: **An Effective Data Management for Healthcare: Big Data Perspective**

Candidate Name: **Sakshi Raj Singh**

The undersigned hereby assigns to Integral University all rights under copyright that may exist in and for the above dissertation, authored by the undersigned and submitted to the University for the Award of the Master of Technology Advanced Computing and Data Science degree.

The Candidate may reproduce or authorize others to reproduce material extracted verbatim from the dissertation or derivative of the dissertation for personal and/or publication purpose(s) provided that the source and the University’s copyright notices are indicated.

SAKSHI RAJ SINGH

ACKNOWLEDGEMENT

I am highly grateful to the Head of Department of Computer Science and Engineering for giving me proper guidance and advice and facility for the successful completion of my dissertation.

It gives me a great pleasure to express my deep sense of gratitude and indebtedness to my guide **Dr. Mohammad Zunnun Khan, Assistant Professor, Department of Computer Science and Engineering,** for his valuable support and encouraging mentality throughout the project. I am highly obliged to him for providing me this opportunity to carry out the ideas and work during my project period and helping me to gain the successful completion of my Project.

I am also highly obliged to the Head of department, **Mrs. Kavita Agarwal (Department of Computer Science and Engineering)** and PG Program Coordinator **Dr. Faiyaz Ahamad, Assistant Professor, Department of Computer Science and Engineering,** for providing me all the facilities in all activities and for his support and valuable encouragement throughout my project.

My special thanks are going to all of the faculties for encouraging me constantly to work hard in this project. I pay my respect and love to my parents and all other family members and friends for their help and encouragement throughout this course of project work.

SAKSHI RAJ SINGH

Table of Content

List of Tables

List of Figures

LIST OF ABBREVIATIONS AND SYMBOLS

PCA-	Principal Component Analysis
KNN-	K nearest neighbors
SVM-	Support Vector Machine
MSE-	Mean Squared Error
RMSE-	Root Mean Squared Error
MAE-	Mean Absolute Error
R2-	Root Squared
LN-	Linear Regression
NN-	Neural Network
RF-	Random Forests

ABSTRACT

Machine learning applied to electronic health records (EHRs) may yields actionable insights, from improving upon patient risk score system, to forecasting the beginning if illness, to optimizing hospital operations. Statistical models that harness the diversity and depth of EHRs derived data are still very uncommon and provide an attractive area for additional study. In this chapter, we give overview of how machine learning has been implemented in clinical contexts and describes the benefits if offers over conventional analytics approaches .We highlight the methodological and practical difficulties of employing machine learning in research and practices .Although there are numerous occasions in which machine learning can execute healthcare duties as well or better than humans, implementation concerns will preclude large-scale automation of healthcare professionals occupations for a significant duration. Ethical difficulties in the use of machine learning to healthcare are also highlighted.

Chapter 1

Introduction

1.1 Health Care System

The healthcare industry is one of the world's most significant and vast undertakings. Throughout the years, healthcare administration throughout the world has shifted from an infection-focused approach to a patient-focused one to a volume-based model, among other things (Agarwal, R &Dhār, V, 2014). Teaching the importance of healthcare and lowering the expense of healthcare is the guiding principle underlying the creation of a patient-centered healthcare delivery paradigm and a patient-focused thinking. The volume and interest in vast amounts of information in healthcare organizations are growing little by little, almost to the point of being non-existent. In order to provide excellent patient-centered care, it is critical to monitor and analyses the massive number of data sets generated (Agarwal, R, Khandelwal, 2015).

Historically, conventional techniques have become outmoded and are not suitable to break down massive amounts of information as the number and variety of information sources have increased at an alarmingly rapid pace over the last two decades The healthcare department generates an enormous amount of data, which necessitates the development of creative and cutting-edge tools and processes that are up to the challenge of handling the data and even better able to do so.

A community-based system serves as the foundation for the social insurance framework that is used by healthcare departments. This is as a result of the fact that it is comprised of a significant number of partners, such as physicians who specialize in a variety of fields, medical caretakers, research center technologists, and other individuals who collaborate in order to accomplish the common goals of lowering the cost of medication and the number of errors made while simultaneously increasing the level of high-quality healthcare experiences (Al Hamid, HA, 2017). The information that is produced by each of these partners comes from a variety of sources, including clinical notes and physical examinations, patient meetings and perceptions, tests performed at research facilities and imaging reports, medications, treatments, overviews, bills, and legal protection (Ardagna et. al 2016).

Daily, the pace at which information is created from diverse sources from many healthcare departments has increased dramatically, as seen by the increasing amount of data being generated. As a result, conventional dataset handling programmers are finding it more difficult to store, interpret, and break down this highly interconnected information. But in addition, there are significant processing advances in the form of innovative and efficient techniques and systems that are being developed in order to store, process, breakdown, and extract values from the massive and diverse medical information that is being created on a continual basis (Aruna, 2012).

As a result, the medical services framework is rapidly becoming into a large information sector. Over the course of the last several decades, medical services information has grown tremendously both in an organized and unstructured manner, driven primarily by the demands of an ever-expanding information-starved public and, more recently, the operational characteristics of e-health phases. The dangers of this multi-dimensional growth have prompted experts to create multiple new watchwords to describe Healthcare Big Data (HBD). This is a perilous development on many levels (Barberton, 2014).

However, it is not only the sheer quantity of information available, but also its variety, with particular emphasis on the types of sources that provide information and the types of objectives that seek it, that are very diverse and varied in the healthcare industry (Bani et. al 2016). The medical services workforce (doctors, clinical staff, and parental figures), benefit-giving organizations (including safety net providers), healing facilities with resources, clinicians and government controllers, drug stores and pharmaceutical manufacturers (including research and development groups), and therapeutic device manufacturers are among those included.

1.2 Big data

"Big data" is a word that refers to large amounts of data that are impossible to handle using normal software or internet-based platforms. Large volumes of data are symbolised by the term "big data." It performs far better than the amount of storage, processing, and analytical power that has historically been accessible. Despite the fact that there are a great deal of various definitions for big data, the notion developed by Douglas Laney is the one that is used the most often and is well recognized.

According to Laney's studies, the quantity of big data is expanding in three unique dimensions: in terms of volume, speed, and variety (known as the 3 Vs) (known as the 3 Vs). Big data refers to the large volumes of information that it holds, therefore the phrase "big data." The idea of big data comprises not only its bulk but also its velocity and diversity in addition to its most conspicuous component, its volume. When discussing data collection, the word "velocity" refers to the pace at which information is gathered and made accessible for further analysis. "Variety," on the other hand, refers to the many sorts of structured and unorganized information that each firm or system is capable of gathering, such as transaction-level information, video, audio, text, or log files. These three characteristics, known together as the "three V's," have become the accepted description of massive volumes of data.

The word "veracity" is still considered to be the most important component of the fourth V, despite the fact that other people have contributed a number of other Vs to this description.

Recent years have seen an explosion in interest throughout the globe in the concept of "big data," notably in the United States. Nearly every area of research, whether it be in the business world or the academic world, is responsible for the production and analysis of significant amounts of data (Katha et.al 2015).

The most challenging challenge in terms of administration is presented by this enormous accumulation of data, which may be both structured and disorganized at the same time. We need technically advanced apps and software in order to take use of high-end processing capability in a quick and cost-effective way. This is necessary in light of the fact that traditional software is unable to effectively handle massive amounts of data (Arora, 2013). In order to make sense of such an enormous amount of information, it would be necessary to use strategies based on artificial intelligence (AI), in addition to novel data fusion procedures. In point of fact, obtaining fully automated decision-making via the use of machine learning (ML) strategies like neural networks and other forms of artificial intelligence would be a big step forward in the field.

On the other hand, if appropriate software and hardware support are not available, massive amounts of data may be difficult to comprehend. This "endless sea" of data has to be navigated more effectively and new internet apps for quick data analysis need to be developed in order to get actionable insights into the issue. It is possible to leverage the information and insights generated from big data to make important social infrastructure components and services, such as health and safety, more aware, interactive and efficient. This can be accomplished by utilizing the appropriate storage and analytical tools in conjunction with big data (Al Hamid, 2017). In addition, the capacity to visualize vast quantities of data in a manner that is accessible to users will be an essential component in the development of civilization.

1.2.1 Big Data Characteristics

For the Big Data Commission at the Tech America Foundation, big data is defined as "huge quantities of high-velocity data that need the application of sophisticated methods and technology to enable for the acquisition of information and to store it and to distribute it and to manage and analyses it.". The V family is responsible for defining the properties of large data. Volume, variety, and velocity are the three most well-known V's in marketing. Veracity, visibility, variability, and value are some of the other V's that have lately been introduced (Olshannikova, 2015).

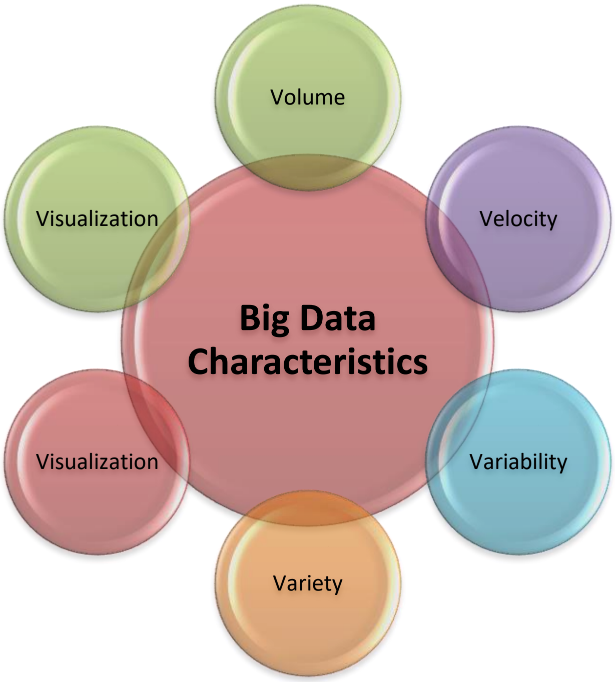


Fig 1.1: V’s Characteristics of Big data

- **Volume:** the significant amount of information that is gathered or produced.
- **Variety:** the many kinds of data collected from a great number of different sources. Structured, semi-structured, and unstructured data are the three primary categories that may be used to big data in the healthcare industry.

- **Velocity:** the rapidity with which the large amounts of data are being gathered, produced, and analyzed.
- **Veracity:** regardless of whether or not the data may be believed (the quality or reliability of data).
- **Variability:** a term that describes many forms of data whose meaning and dimensions are in a state of perpetual flux.
- **Visualization:** a method for making the vast and complicated facts more understandable by displaying them in graphs and charts.
- **Value:** Describes the importance of the information that has been produced or gathered.

1.3 Health care and Big Data

In the past, the healthcare business has often created vast volumes of data in order to satisfy regulatory and compliance obligations. This information has been put to use to improve the quality of treatment provided to patients (Bid good, 1997). Even if the vast majority of data is still stored in hard copy form, the current trend is toward the quick digitization of these massive volumes of data in the not too distant future. This is despite the fact that the vast majority of data already exists. Clinical decision support, sickness monitoring, and population health management are some examples of the kind of medical and healthcare jobs that might potentially benefit from the massive volumes of data that are referred to as "Big Data." Their main motivations are the need to comply with legal mandates and the prospect of enhancing patient care while lowering overall healthcare expenditures. Data generated by the U.S. healthcare system in 2011 was estimated to have exceeded 150 Exabyte's. If the present growth rate continues, big data for U.S. healthcare will approach the zettabyte (1021 gigabytes) size in the near future and the yottabyte scale within a few years (1024 gigabytes).

It's believed that Kaiser Permanente's electronic health records (EHRs), which include images and comments, contain between 26.5 and 44 petabytes of potentially valuable data, thanks to the health network in California having more than 9 million members (Wang, 2015). When it comes to electronic health data, "big data" refers to enormous and complicated data sets whose handling is beyond the capabilities of conventional software and/or hardware; they are also beyond the capabilities of traditional or frequently used data management tools and processes. In healthcare, the sheer volume of data and the speed at which it must be processed and analyzed to be meaningful are frightening. Patients' health and well-being are considered "big data" in the healthcare industry because of the collecting of all relevant information (Ambigavathi 2018). Information in electronic patient records (EPR) and clinical decision support systems (medical imaging, laboratory, pharmacy, and administrative data); sensor data generated by machines (such as vital sign monitoring); social media posts (such as Twitter feeds, blog posts, Facebook statuses, and web pages); and a lack of patience There is potential for the big data scientist amid the large quantity and variety of data available to him or her (Wang, 2018). It is possible for big data analytics to enhance treatment by identifying relationships and patterns and trends within data. This has the ability to save expenses while also increasing patient safety and reducing expenditures.

For obvious reasons, big data analytics in healthcare is a research topic within the discipline of data analytics in healthcare since it makes use of the explosion of data to extract insights that allow people to make better-informed choices. Healthcare practitioners and other stakeholders in the healthcare delivery system will be able to design diagnoses and treatments that are more comprehensive and insightful when significant volumes of data are synthesized and analyzed and the previously described relationships, patterns, and trends are uncovered (B. Durga Sri Et. al 2017). This, as one would expect, will result in higher quality care at lower costs, as well as improved overall outcomes.

Analyzing patient characteristics as well as the cost and result of care to discover the best clinically and cost-effective therapies is one of many instances where big data analytics might enhance outcomes. Analytic techniques (such as segmentation and predictive modelling) are used to identify patients who would benefit from a specific treatment (such as personalized medication), and these techniques are then applied to patient profiles in order to identify patients who would benefit from that treatment. We may establish new income streams by combining and synthesizing patient health records and claims data sets, which we then market to other parties. As an example, pharmaceutical firms may use our patient clinical records and claims data to find candidates for clinical trials (Luna, 2014). Payers are showing an increased interest in the design and distribution of mobile apps that provide assistance to patients in the management of their treatment, the identification of physicians, and the enhancement of their health. Through the use of analytics, payers are able to monitor patients' compliance with their prescribed medications and treatment plans, therefore identifying trends that lead to improvements in both individual and population health.

1.3.1 Advantages to healthcare

Companies in the healthcare industry, ranging from solo practices and multi-provider groups to large hospital networks and accountable care organizations, stand to benefit considerably from the digitalization, combining, and efficient use of big data (Dr V.P Gladys, 2018). It is possible to diagnose diseases at an earlier stage, when they are more amenable to treatment and more likely to be cured. Individual and population health can also be managed more effectively, and fraud in the health-care system can be detected more quickly and efficiently, among other things. Big data analytics may provide answers to a plethora of issues and problems.

There are several reasons that contribute to a patient's length of stay, including elective surgery, medical issues, and individuals at risk for medical complications or hospital-acquired infections and illness/disease progression (EMC Consulting). Big data analytics in healthcare may save the U.S. more than $300 billion annually, with an 8% reduction in national healthcare cost accounting for two-thirds of that amount (Mukesh Borana, 2015).

Clinical operations and R & D waste total $165 billion and $108 billion, respectively, according to the report. According to McKinsey, the following three domains might benefit from the use of big data: (see figure 1.1).

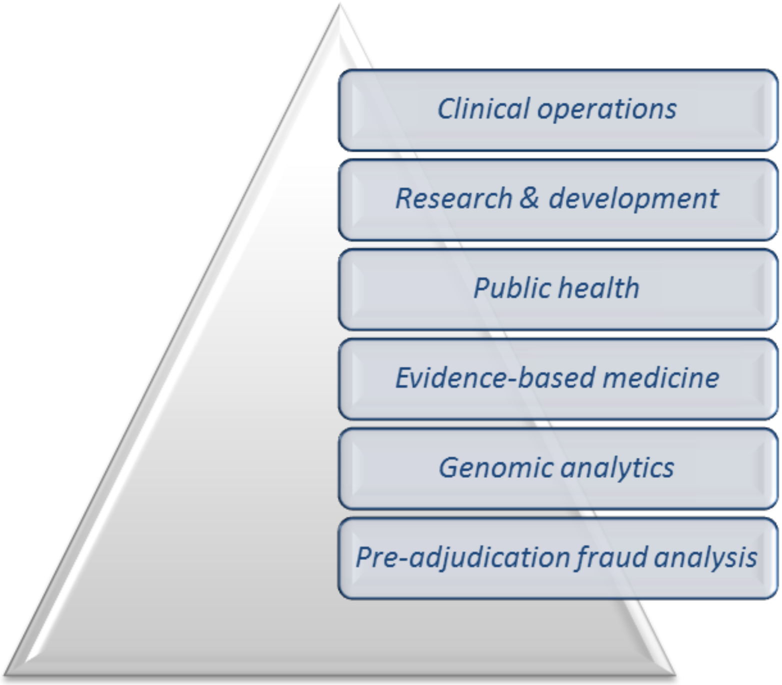


Fig 1.2: Big data Quantification Area

1.3 Healthcare as a big-data repository

The term "healthcare" refers to a multifaceted system that was developed with the main goal of preventing, diagnosing, and treating health-related disorders or impairments in human beings. This system is made up of a number of different services. Health professionals, such as physicians or nurses, health facilities, such as clinics and hospitals for the delivery of medications and other diagnostic or treatment technologies, and a financial institution that supports the first two components of the system make up the primary components of a healthcare system (MS.Minu, 2018). They come from a variety of health-related disciplines such as dentistry, medical science, midwifery and nursing as well as psychology and physiotherapy, amongst other specialties.

Depending on the severity of the ailment, the appropriate level of medical treatment will be determined. This might be any number of levels. Services include primary care consultation, an emergency department requiring specially trained staff (secondary care), advanced medical research and treatment (tertiary care), and specialized diagnostic or surgical procedures. They also offer services as an initial point of contact (quaternary care).

A patient's medical history, medical and clinical data, and any other private or secret medical information are all within the purview of medical experts at every level of the health care system. It doesn't matter whether you work in a hospital or clinic. Patients' medical records were often kept in either handwritten notes or typed reports, depending on the situation, in the past. E-health records have largely taken the role of paper-based records in today's world.

Even the results of a medical examination were recorded on paper and stored in a filing system that was difficult to navigate. In point of fact, this practice goes back thousands of years; the first case records that are known to exist were written in an Egyptian papyrus book that was dated to 1600 BC. Clinical case notes, in the words of Stanley Raiser, "capture the event of sickness as a story in which the patient, his or her family, and the doctor all play a part." (DeepthiYarmala).

The growth of computer systems and their capabilities have led to the digitalization of all clinical assessments and medical records in healthcare systems. These discoveries have shown the systems' potential. Electronic health records improve patient care and practitioner efficiency. The Institute of Medicine created the phrase "electronic health records" in 2003. The term refers to records that are kept electronically. The phrase "electronic health records" comes from the phrase "electronic health records," which comes from the phrase "electronic health records," etc. (MS. MI nu, 2018).

"Computerized medical records for patients that contain information about a person's physical or mental health and condition in the past, present, or future that is stored in electronic system(s) used to capture, transmit, receive, store, retrieve, link, and manipulate multimedia data for the primary purpose of providing healthcare and health-related services," is how Murphy and Heken and Waters define electronic health records (EHR) (MukeshBorna2015).

1.4 Issues of healthcare data

- In spite of the many advantages offered by healthcare data, the primary challenges that are now being faced by the healthcare services are becoming worse by the day. The following describe each of these:
- An ageing population • A substantial number of cases of medical mistakes
- Difficulty in gaining access to information regarding healthcare • Inefficiency in managing large amounts of data • Increased demand for high-quality and risk-free healthcare services despite the fact that resources (such as the number of physicians, hospitals, and laboratories, etc.) have remained the same.

1.5 Hadoop Distributed File Storage System

The Hadoop Distributed File System (HDFS) is a fundamental component of the Hadoop architecture. It is a distributed file system that stores a massive dataset as numerous files. This distributes the data over several workstations or clusters, making it very resilient to failure (Harshwardhan S, 2014). HDFS is an efficient answer to the problem of successfully storing large amounts of data since it was developed utilizing technology that is inexpensive [31]. The master-slave architecture is the foundation of HDFS. Name Nodes and Data Nodes are responsible for managing data storage in this architecture [1, as seen in Fig. 1.3].

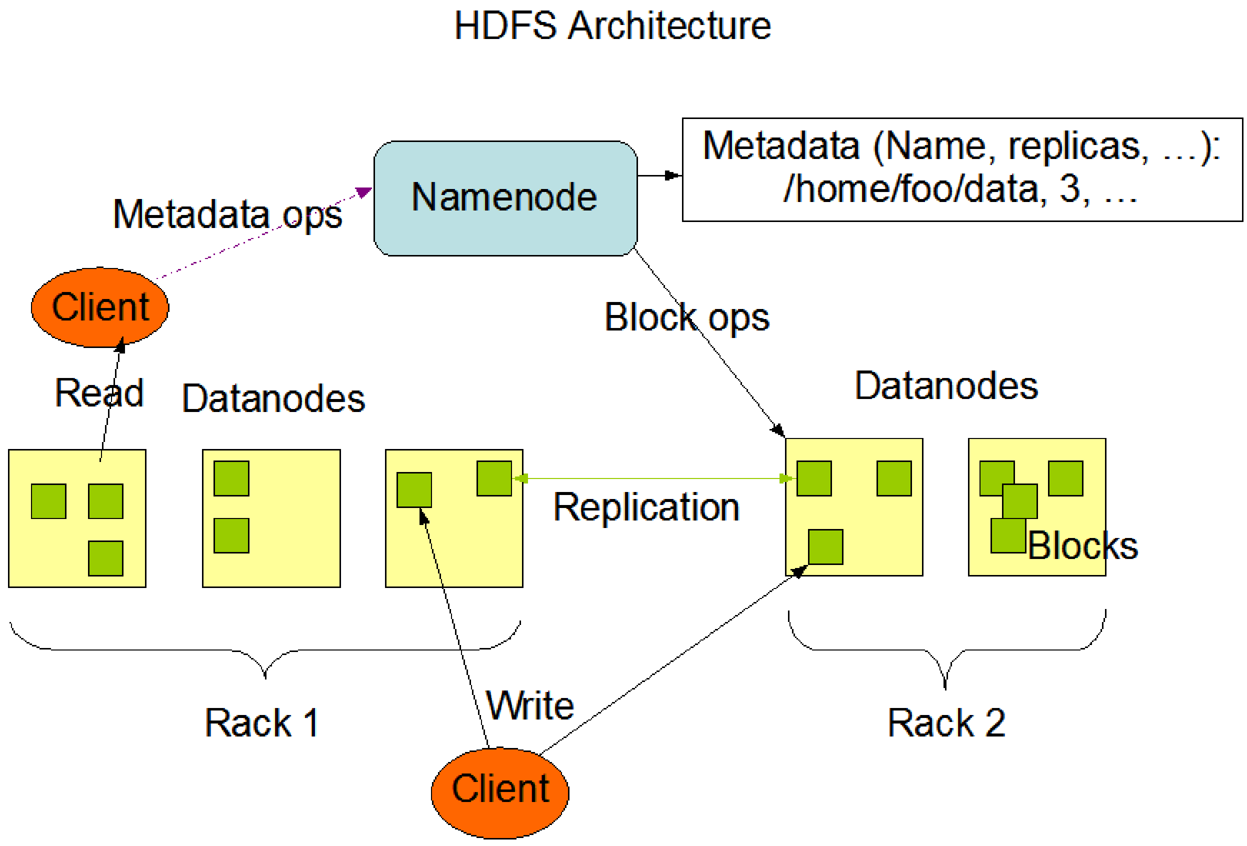


Fig 1.3: HDFS (https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html)

Neither the Name Node nor the Data Node are pieces of software that are intended to operate on commodity computers. Most of the time, the GNU/Linux operating system is used on these workstations (OS). HDFS is written in the Java programming language, and it may be operated on any system that is capable of running the Name Node or Data Node software. Because HDFS is written in Java, which is a highly portable programming language, it may be run on a broad variety of computers (Ivanilton Polato, 2014). A typical deployment consists of a dedicated workstation that is only devoted to running the Name Node software. On each of the other computers in the cluster, there is a separate instance of the Data Node programmer running. Although the design does not exclude operating many Data Nodes on the same system, in practice this is seldom the case in a real-world deployment.

The presence of a single Name Node in a cluster makes the system's overall architecture much simpler to understand. Essentially, the Name Node is a middleman and a repository for all HDFS metadata. In order to prevent user data from passing via the Name Node, the system has been developed in a certain manner (IqbaldeepKaur, 2016).

1.6 The Importance of Research

Healthcare is often compared to other complex, adaptive systems, such as transportation networks.

When making judgments, it is critical to depend on data that is unstructured, organized, or semi-structured as a foundation. Health care companies are using Big Data technology to deal with all of the information on their patients in order to get a more comprehensive perspective of care coordination, health management, and result. Big Data contributes to the development of a sustainable healthcare system and the expansion of access to healthcare (Rahul Beakta, 2015).

Every day, trillions of quintillions of bytes of data are created all around the globe. Due to the exponential nature of the progress pace, we must review such vast amounts of data in an efficient and effective manner. Because of the diversity and amount (structured and unstructured) of data accessible online, it is difficult to properly extract vital information from all of the information available (Deepthi Yarmala, 2016). Furthermore, the amount and quality of data serve as the cornerstone of a sound analytical framework in every respect. The use of Big Data analytics is important in this context.

Indexing strategy is used to plan an access method to a searched item since there is a need for strategies that can search and access data items quickly. Indexing strategy is employed to plan an access technique to a searched item. It also controls how data is prepared in a storage system so that data retrieval may be conducted on the information. The most pressing demand for this study approach is to make it easier to retrieve trustworthy data contents from Big Data storage systems.

1.7 Motivations

Because of technology advancements, a large amount of data has been created in a wide range of formats. The cost of capturing, storing, processing, and analyzing data has increased. Academics and industry are both involved in making this procedure more affordable and efficient. The healthcare industry is one of the primary beneficiaries of data management.

The data generated by these entities would need the storage of information in an inefficient way. Data harmonization is essential for storing data in order to improve the quality of analytic outputs. Data derived from a variety of heterogeneous data sources, each with a distinct format and size. In order to offer a clear vision to the end-user, data harmonization works on the notion of storing, integrating, and loading multiple data of different types and volumes at the same time on a non-volatile platform in order to provide a clear vision to the user.

1.8 Research objectives

The proposed research retrieves big health data with respect to

quantification concept from different databases. The research includes quantification to:

- To identify the factors and areas where effective data management in healthcare is required.
- To identify the suited approach in data management for healthcare system.
- To develop an empirical framework for analytics on big data over the medical data.
- To Validate and Verify the proposed framework and model.

1.9 Thesis Organizations

The overall organization of the research work is given as follows:

In chapter 1, detailed introduction about the Big Data and their role are discussed. This chapter provides a detailed overview about the Big Data and the role of Big Data in health care environment. This section also provides the discussion about the various databases that supports Big Data handling.

In chapter 2, description about the various research methodologies that aim to perform Big Data analytics are dealt in this chapter. Also provides a detailed overview about the various research methods in terms of their working procedure.

In chapter 3, Discusses the problem of the study, methodology adopted for the study, sources of data, sample size, statistical techniques used for the analysis and brief profile of the various stages.

In Chapter 4, this chapter covers the use of big data analytics to the healthcare industry. Several distinct machine learning techniques are utilized in the research work presented here. A dataset including medical records of patients is gathered for the goal of conducting an experiment, and then a variety of machine learning techniques are performed on the information.

In chapter 5, Analyze the influencing factors for taking up big data, factors influencing the health care data, as well as problems faced by the HDFS.

In Chapter 6, Discusses the conclusions and the suggestions for the Future study in medical field and disease prediction at early stage.

Chapter 2

Preliminary Study

2.1 Overview

With the advent of technology and the increasing quantity of data that is coming into and leaving businesses on a daily basis, there is a need for data analysis that is both faster and more efficient. To make fast and effective judgments, it is no longer necessary to have a vast quantity of data on hand. The only way to effect change is for us to improve our ability to comprehend and use the data collected in many libraries by multiple institutions, as well as the data provided by people. In other words, if data is not properly analyzed, it becomes nothing more than a resource that is not used effectively.

When we talk about "big data," we are not only talking to the amount of data; rather, we are also referring to the capacity of that data. This is a very crucial distinction. It is difficult to comprehend and compile the results utilizing the data gathering techniques that are now in use due to the fact that the data sets are so extensive and diverse. The use of analytics to large amounts of data might be useful in this situation. Big Data Analytics involves collecting data from a wide range of sources, combining the data in a way that makes it possible for researchers to analyses it, and then delivering data items that are beneficial to the organization's target market or consumers.

The generation of medical data is now taking place from a large number of sources, some of which include people, mobile phones, body area monitors, hospitals, research institutions, healthcare professionals, and organizations. These are just few examples of the sources that are being used. To gather medical records and help in the administration of hospital outcomes that would otherwise be too large and complicated for conventional technologies, it is the utilization of the massive quantities of information provided by digital technology. This information is then used to collect the records. The use of electronic health records is one illustration of this use (EHRs). Big data is a phrase that refers to massive volumes of information that are created in the healthcare business as a consequence of the use of digital technologies. This information may then be used to provide better treatment for patients. This technology helps with the administration of hospital outcomes and gathers medical records, both of which would otherwise be too large and complex for the technologies that were previously available. The use of big data analytics in the area of medicine has led to a multitude of exciting innovations, many of which have the potential to save the lives of their respective patients. Electronic Health Records, data that was created by a computer or a sensor, health information exchanges, patient registries, portals, genetic databases, and public records are just a few examples of the various types of data that are utilized in healthcare application software. Other examples include public records, genetic databases, and data that was created by a computer or a sensor. They are vital data points in the healthcare system, and in order for their medical conditions to be treated, the healthcare system has to have access to data analytics of a high quality.

In the most recent issue of Biomed International Journal, an article authored by Ashwin Belle (2015) and a number of other researchers with the title "Big Data Analytics in Healthcare" was published. Ashwin Belle, along with a few other persons, contributed to the writing of this paper. Big data is a collection of data elements that have a certain size and speed, in addition to a variety of types and levels of complexity that require the investigation, adoption, and even invention of new hardware and software mechanisms in order to effectively process, interpret, and represent the data. According to the definition provided in this article, "big data" refers to a collection of data elements that have a certain size and speed. Their primary areas of academic interest were topics like medical image processing, physiological signal processing, and genetic data processing, among other things. This presentation will cover a wide range of topics, some of which include applications of big data in genomics, as well as a variety of problems and current techniques in the construction of monitoring systems that ingest both high fidelity waveform data and discrete data from non-continuous sources. Other topics that will be covered in this presentation include big data applications in genomics, as well as big data applications in genomics. In addition, this presentation will cover a broad spectrum of concerns and themes, making it one of the most comprehensive presentations of its kind. It was covered in the article Using Big Data Analytics how Hadoop data processing is one of the greatest alternatives to go with at the present trends, and how when it is used, it will give an additional advantage in terms of data analysis. This was because Hadoop data processing is one of the greatest alternatives to go with at the present trends. This point is argued by RevanthSonnati (2015) in the paper titled Improving Healthcare. Using Big Data Analytics went about how Hadoop data processing is one of the best options to go with at this point in time given the trends that are occurring. The purpose of this study paper was to provide a workable computer strategy that makes use of big data and analysis in order to enhance healthcare by promoting healthcare science, affordability, and accessibility. This was accomplished via the provision of a feasible computer approach. The major objective of the company is to be of service to society by using contemporary computational methodologies to evaluate and provide patient-centered medical care. It displays the data flow very clearly, beginning with the raw data in all of its many formats and continuing on via the Hadoop ecosystem and the analytical engines until it reaches the system's final goal. When doing an analysis of healthcare data, taking into account a patient's position on the map is of equal importance, as stated in the study.

Big Data Analysis is discussed in detail by Manpreet Singh (2017) and colleagues in their paper, BIG DATA ANALYTICS SOLUTION TO HEALTHCARE, in which they discuss the value of applying Big Data Analysis to patient care datasets for better insight in care coordination, health management, and patient engagement. It highlights their research in healthcare, as well as how big data may give solutions and have a broad variety of applications in biological issues. It comes to the conclusion that Big Data will have a stronger influence on the healthcare system than previously expected and will be a boon. When it comes to health research, predictive analysis appears to have a higher effect as scientific publications have the capacity to go viral, aid, and foresee an enormous quantity of emergency clinical circumstances.

2.2 Challenges of health analytics using Hadoop software

2.2.1 Cleansing

All the physicians and patients are interested in the cleanliness of the clinic and surgical suite, but they are not attentive in maintaining the data clean.

2.2.2 Data Cleansing: It is known as cleaning of data and it makes sure that the databases should not fluctuate, precise and appropriate information that should not violate.

2.2.3 Security: The first priority for all healthcare systems, especially in hacking of data, pulsation of prominent violations, blasphemer sequences. Data can be secured by HIPPA,

Which includes a list of specialized protection for organizations that stores PHI (Protected Health Information) which includes protocols, controlling the access, corroboration, monitoring and conyence safe.

Safeguards translates into common security approaches using present day Anti – virus software network, setting up of firewalls, encoding confidential information and several factors Attestation.

Health care systems should frequently remind their employees about essential data security protocols and should persistently review, who can access the valuable data resources in case of spiteful stakeholders.

2.2.4 Storage

As there is a rapid increase in the aggregate of healthcare information, the IT department faces challenges regarding critical price, surveillance and execution, but the front line clinicians don't have an idea of where their data is being kept. The prices and influences of on ground data centers are no longer bearable by some suppliers. While many institution are on ease with on ground data storing as it assures surveillance check, entry and uptime. While onsite server network is overpriced to calibrate, hard to support for and likely to cause information silos across different departments. Virtual storage has become an admired choice as it is affordable and trustworthy.

2.2.5 Governance

Particularly in the clinical sector, health care information has a lengthy duration of validity. Providers may be required to use anonymous datasets for research projects in situations where patient information may be kept for a minimum of six years. This raises concerns about current governance and case management, which is a major worry. Data may be utilized in a variety of applications, such as quality assessment and making comparisons. It is vital for researchers to understand who has used the data in the past, for what purpose, and for how long.

The most important component of a data governance strategy is the creation of comprehensive, accurate, and up-to-date metadata. It enables professionals to reproduce previous requirements, which allows them to conduct scientific study and establish standards, while also preventing data loss from occurring.

It is my responsibility to handle the improvement and mentoring of crucial metadata. Data governance will ensure that components have consistent definitions, configurations, and documentation throughout their lifecycle, from their inception to their decommissioning, and that they continue to be helpful to the purpose.

2.2.6 Querying

Persistent databases and robust governance mechanisms are required for any health care system that want to query information and get the results they anticipate. The ability to conduct data searches is critical for the provision of information and analytics. Institutions may query their information and get the results they are looking for with ease because to robust metadata and effective protocol management.

The ability to query information is a necessary pre-requisite for reporting and data analytics careers. Before they can conduct a successful and relevant analysis of their large data sets, healthcare organizations must first overcome the difficulties associated with data collection and storage. The first obstacle they must overcome is the existence of information silos as well as compatibility difficulties that prevent the query tools from obtaining access to the institution's information warehouse. If a dataset is saved in many forms or is preserved in a multi-walled off system, it is hard to get a complete profile of an institution's position and stature, or to obtain a complete health report on a single patient, from that dataset. Many organizations employ structured Query Language (SQL) to manage large information sets and similar databases, but this is only beneficial when the information given by the organization can be relied upon by the client. It is thus necessary to deliver absolute, unambiguous information in a well-organized manner.

2.2.7 Visualization

A clinician can consume data from the cleanliness and engaging data visualization, so they can easily use it.

Colour coding: most popular technique that provides an instant response (such as red-stop, yellow-caution, Green-go). Examples – Bar graphs, pie- charts, histograms, bar diagrams. Health care systems should review the good data presentations such as charts with proper sections and it should represent figures and their labeling information to reduce confusion. **Drawbacks**: overlapping text, low- quality graphs, this irritates users to avoid data.

2.2.8 Updating

Health care systems are not fixed, they require frequent updates to present day, such as address/marital status-might change rarely in their lifetime. It undergoes a great challenge for health care systems, to keep updating all the time, providers should have a clear idea of data about their updates and automation procedures without the damage of the quality information. While adopting an update, clinicians should not confuse to access patient decision making.

2.2.9 Reporting

Following the conclusion of the questioning technique, the contributor should prepare a document that is concise, convenient, and intelligible to the audience members who have gathered to witness it. The principles of data collection and the veracity of the data have a significant impact on the overall negative tone of the recorded reports. In the last phase of the process, a lack of sufficient data may result in a mistrust report being generated. The supplier should be able to understand and evaluate the differences between analysis and report generation. A report is always necessary in order to conduct an analysis. Few reports were found to be concentrating on the most recent popular topics, then summarizing or gaining the confidence of the reviewer to take a certain action. Institutions must be clear about the purpose of their reports and ensure that database managers are able to create the information they need to function properly.

This study includes information from Russom, P (2011) on Big Data. According to statistics from 2013, Facebook, a social networking site, has around 1.11 billion active accounts, with 751 million of those users accessing the site through a mobile device. A second example of Big Data is Flickr, which offers limitless picture uploads, the capacity to display HD video, infinite storage space and an endless video upload option, among other things. Flicker has a total of 87 million registered users, with more than 3.5 million uploading photographs every single day, according to the company.

Mukherjee (2012), a CD that was shared Apache Hadoop is used for large data analytics. "Big data analytics is the process of analyzing enormous amounts of data in order to extract relevant information and uncover hidden pathways. It is a reference to the Map Reduce Framework, which was created by Google. The Map Reduce Model is implemented using the Hadoop open source platform, which is free and open source.

It is the experimental work on big data challenges that is presented in Hadoop and Map Reduce (Osaka, 2013). It is shown in this study effort how to get the best possible solution by using a Hadoop cluster and the Hadoop Distributed File System (HDFS). The programming framework for parallel processing, Map Reduce, is being shown in order to demonstrate its use.

D. Garlasu et al (2013), when it comes to "a Big Data implementation based on Grid Computing," Grid Computing has the benefit of both storage capacity and processing power. It is used in conjunction with the Hadoop technology for the purpose of implementation. Grid computing is based on the notion of distributed computing and is used in many applications.

These have the advantage of having a large amount of storage capacity as well as processing power.

S.Vikram (2016), a review on Big Data and Methodology, presented several obstacles and concerns in the field of big data analysis and processing. They also provide an explanation of the basic research that has been conducted in relation to these technical challenges. Big-data analysis is the transformation of data in financial, operational, and commercial challenges utilizing discrete data sets, and it is becoming more popular. By using cloud-based virtualization technology, which is utilized to rapidly mine data sets, big-data approaches are able to add additional analysis to the data sets.

2.3 Scope delimitation and risks

When it comes to examining big data analytics, one can identify what has been specified and what are the criteria for picking the analytics and tools for big data. The focus of this study will be finding the weaknesses in reviewing big data analytics. The evaluation may identify which issues have been resolved and which issues still need to be addressed. Furthermore, it assists in telling researchers about what has been provided, which may open the door for them to do further analyses, given that big data is a popular issue these days and that many individuals are drawn to this notion.

Security and privacy problems, data capture difficulties, and obstacles in data processing and visualization are among the most major hurdles that must be overcome before big data analytics can have a beneficial impact on a wide range of businesses. Storage of huge volumes of data originating from a variety of sources is another critical issue that must be addressed but is not presently being handled with the tools now available. In order to handle some of the issues that particular industries, such as retail, banking, and healthcare, are experiencing, it is necessary to investigate and explore new analytical methods. This is where you come in.

Data visualization, predictive analytics, descriptive analytics, and diagnostic analytics are some of the tools that may be used to address the issues associated with acquiring and processing large amounts of data. Statistical models and artificial intelligence modelling are used by both organizations and individuals. Furthermore, machine learning algorithms may merge statistical and artificial intelligence methodologies in order to analyses vast volumes of data at rapid speed and with great accuracy. One solution to the storage problem is to make use of Hadoop (the Apache platform), which has the capability of processing very huge volumes of data. This is accomplished by dividing the data into smaller chunks and then allocating various portions of the information to other servers (nodes).

End-to-end encryption is a must for organization’s to keep an eye on the data's sources and to prevent unauthorized access to the data as it moves from one place to the next. Because many cloud service providers do not encrypt the data, companies have an obligation to check their cloud service providers. Data volume and encryption and decryption slow down the flow of information, which is why this is happening. Cloud service providers must be trusted by businesses. Big data privacy solutions protect personal data privacy when obtaining data from individuals who are unaware or difficult to get information from, such as personal interests, habits, and physiological traits. Acquiring data from individuals who are either unaware of or unwilling to provide it is a viable option.

Additionally, preserving personal private data that may be compromised during storage, transfer, and use, even if the data was obtained with the user's consent, is essential.

It is possible that there will be risks associated with this thesis's research, which will include the process of selecting a suitable subject, which will be determined by the important point of uncovering a personal practical or professional need or a personal urge to confront the research question.

2.4 Big data analytics

The expansion of big data continues apace, and a growing number of businesses are becoming interested in managing and analyzing data [42]. Organizations attempting to reap the benefits of big data are turning to big data analytics to help them make quicker and better choices. This is due to the fact that it is difficult to conduct analysis on massive datasets using the conventional strategies and infrastructure for data management (Constantia et al., 2015). As a direct consequence of this, there is a growing need for innovative methods and tools that are specifically tailored to the analysis of big data. The proliferation of large amounts of data is having an effect on all aspects of the data itself, as well as its collection, processing, and, eventually, the conclusions that may be taken from it [43]. The provision of big data tools and technologies can assist in the management of the otherwise exponential growth of network-produced data, as well as increase the ability of organizations to scale and capture the required data in order to reduce database performance issues. This is accomplished by reducing the burden on database administrators. The provision of tools and technology for large data (Elgendy, N. 2014). In addition to that, the meanings of several terms related to big data analytics are discussed.

When you open practically any major scientific or business publication published in the modern era, whether it is online or in print, you will almost always find a reference to data science, analytics, or big data, or any combination of the titles or words associated with these fields (Agarwal 2014). Researchers are examining big data definitions (Akter et al., 2016), while others examine the tools, methodologies, and processes that are necessary for data analysis (Russom, 2011).

2.5 Hadoop Technology

Hadoop is an open source architecture that is used to store the structured, semi-structured, unstructured, and quasi-structured data that is commonly referred to as Big Data. Hadoop is a distributed database management system. It generates meaningful results via the use of data analytics [47]. The ETL procedure is the industry standard for dealing with large amounts of data (Extract, Transform and Load). Extraction is the process of gathering data from numerous sources, transforming it to meet analytical requirements, and loading it into the appropriate systems in order to get meaningful value from it.

In addition to providing critical feedback/input to government entities, it also assists nonprofit groups. The information gathered is divided into two categories: operational data and analytical data. Different sorts of information are divided into two categories: Social Data is created from various social networking sites such as Facebook, Google Ads, and other similar sites. Transactional Data is generated from all everyday transactions. The data generated by industrial equipment, where sensors are installed in machines, data stored in a black box in the aviation industry, weblogs that track user behaviors, medical devices, smart meters, road cameras, satellites, games and many other Internet of Things applications are all examples of sensor or machine data (IoT).

Government institutions are increasingly being digitalized and Aadhaar-enabled in today's world. Applications that are Aadhaar-enabled would give better services and facilities to legitimate beneficiaries, as well as making it easier for individuals to engage in the digital economy, according to the government. Companies are increasingly turning to Hadoop technology in order to embrace digitalization in their organizations and reap the advantages that come with it. A distributed file system (also known as Hadoop) enables numerous concurrent processes to execute on various servers while dividing and transmitting data and files across different nodes. Hadoop was built in Java and is comprised of a Distributed File System (DFS).

In the event of a node failure, it becomes more efficient to process and restore the stored data without causing any delays. When processing or storing information in HDFS, there is a possibility of fraudulence occurring. Due to various Big Data issues concerning management, storage, processing, and security, it is necessary to deal with them individually.

Apache Hadoop is an excellent framework for processing, storing and analyzing large volumes of unstructured data – i.e., Big Data. It serves two purposes **(N. V. Chawla, 2013):**

1) Storing of the large volume of Data

2) Mapping and Reduction of Data stored in Hadoop.

The entire framework is explained diagrammatically considering its five layers and tools used in each layer. The five layers are:

- Data Source Connectors
- Data Store
- Processing
- Analytics
- Visualization

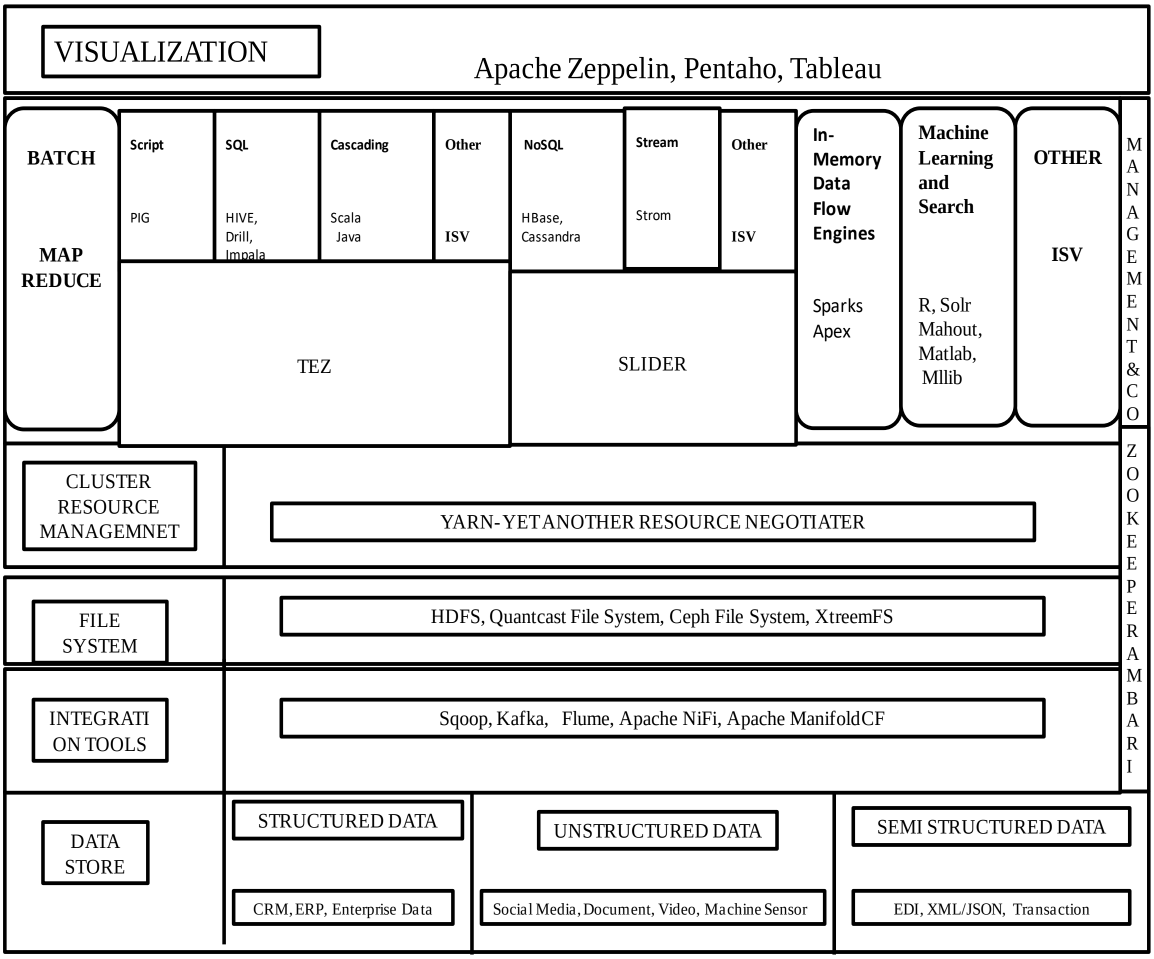

VISUALIZATION
Apache Zeppelin, Pentaho, Tableau
BATCH
MAP REDUCE
Script
PIG
SQL
HIVE, Drill, Impala
Cascading
Scala Java
Other
ISV
NoSQL
HBase, Cassandra
Stream
Strom
Other
ISV
In-Memory Data Flow Engines
Sparks Apex
Machine Learning and Search
R, Solr Mahout, Matlab, Mllib
OTHER
ISV
TEZ
SLIDER
MANAGEMENT & CO
ZOOKEEPER AMBARI
CLUSTER RESOURCE MANAGEMNET
YARN- YET ANOTHER RESOURCE NEGOTIATER
FILE SYSTEM
HDFS, Quantcast File System, Ceph File System, XtreemFS
INTEGRATI ON TOOLS
Sqoop, Kafka, Flume, Apache NiFi, Apache ManifoldCF
DATA STORE
STRUCTURED DATA
CRM, ERP, Enterprise Data
UNSTRUCTURED DATA
Social Media, Document, Video, Machine Sensor
SEMI STRUCTURED DATA
EDI, XML/JSON, Transaction

Fig.2.1. Hadoop Framework (McKinsey, 2011)

2.5.1 Need for Hadoop

Because of the proliferation of technology and the digitalization of its applications, there has been a meteoric rise in the amount of both structured and unstructured data that has been created, and this trend is only expected to continue. It has increased the demand for high storage capacity, management of information, accessing it, analyzing it, and the need to do all of this with security so that it can be analyzed and extracted without any loss of information. All of these requirements have led to an increase in the demand for high storage capacity **(Aditya B. Patel, 2012).** Organizations are moving the data on Hadoop architecture because of the following special features it has:

- The Capability of Storing and Processing a Diverse Collection of Complicated Datasets Across Multiple Distributed Systems
- The ability to perform parallel and multiple node computations quickly and reliably across all CPU cores.
- Fault Tolerance and High Availability, Capability to Handle Real-Time Node Failures and Ability to Redirect Traffic to Other Nodes to Handle Issues at the Application Layer.
- Storing and retrieving enormous amounts of data simultaneously without doing any data preprocessing.
- Capable of expanding in size, from a single machine all the way up to thousands of servers, thanks to its scalable nature.
- Servers may be dynamically added to or deleted from the clusters without causing any disruption to the functioning of the system.
- Economically advantageous due to the fact that Hadoop is an open source technology.
- Designed to work on all Java-based systems simultaneously.

2.5.2 Advantages of Hadoop

- The user is given the ability to easily develop and test distributed systems while using the Hadoop framework. It is efficient, and it automatically distributes the data and work among computers, which, in turn, makes use of the inherent parallelism of the CPU cores. In addition, it has a low memory footprint. **(Karthik Sridhar, 2015).**

- Hadoop does not depend on hardware to offer fault-tolerance and high availability (FTHA), but rather, the Hadoop library itself has been built to identify and manage errors at the application layer. This is because FTHA is a combination of fault tolerance and high availability.

- Pig script is used to perform read, filter, transform, join and write data without the need to know the complex programming skills. Thus it is helpful in data analytics to find the trend in large datasets. Hive acts as the relational database management, making it an essential part of the data access layer of Hadoop.

- Data Analytics is the process of finding the pattern in the voluminous structured and unstructured data and Hadoop has become one of the popular choices for the market analysts and researchers for data analysis.

2.6 The Need of Hadoop for Healthcare Applications

There are generally a lot of crises among the general public when it comes to the problem of Big Data these days, and this is no exception. It thus makes perfect sense that this Big Data is also present in the Healthcare applications when seen in the correct context. Anusha B (2015) was tasked with investigating the applications of big data in the critically important field of healthcare. For example, if we speak about how big data is being utilized in healthcare applications, we may think of services such as healing illnesses, increasing profits, avoiding epidemics, enhancing the quality of people's lives, preventing deaths, and lowering total expenses. Almost all of the laboratories and hospitals make use of big data in order to minimize the total expenses of providing services. We will also do some research on how Big Data is used in the Healthcare Industry: Based on research conducted by the business, the US Healthcare industry created about 150 billion gigabytes of data in 2011, which is comparable to 150 Exabyte’s of data. The data generated by the US Healthcare industry in 2011 was evaluated (Groot, 2010). Furthermore, according to the findings of the research, data was created via a variety of techniques, including patient care, record keeping, and other criteria. Since the beginning of this year, there has been a significant rise in data, which has climbed to around $1.2 trillion in the Healthcare applications (S. V. Nuti, 2014). The majority of big data in healthcare systems is derived from very large healthcare datasets. Furthermore, big datasets are very challenging to maintain when utilizing certain standard methods, as previously stated. It is difficult for them to keep track of such a big number of data sets. Utilizing very outdated data management systems makes it very difficult to make use of the information. The notion of Big data in the healthcare systems is very essential since the data not only increases in bulk, but it also comes in a variety of formats such as unstructured, semi-structured, and organized, and it grows at an extremely rapid pace or speed.

At the moment, data analytics on this sort of data is relatively challenging, and data administration is not very straightforward. As a result, we may conclude that data analytics has emerged as a significant concern in today's healthcare systems. It also leads to the notion of Big Data analysis technology, which is necessary since there are several issues associated with it (N. V. Chawla, 2013).

The data in the healthcare industry is increasingly being stored in digital form, as opposed to the printed versions that were formerly used. Consequently, at a time when data is being more digitalized, it is critical to understand how big data might be used in the Healthcare Industry. Population management, clinical decision-making, and disease monitoring are the three most important applications of Big Data in the healthcare business (A. Gandomi, 2015). The healthcare business has just recently begun to explore the potential of big data. Over 80% of the data in the healthcare industry is in unstructured formats. In this particular instance, the Healthcare sector has a significant task in investigating and managing the situation. The healthcare industries are looking forward to how they will handle this unstructured data in such a manner that it will be beneficial for them for a variety of objectives in the healthcare business in the near future. The information included in the (M. Viceconti, 2015).

The Healthcare Business is growing at an alarming pace, and it will be a major issue for the industry in the future years to adapt its practices in order to keep up with the growing amount of data. Although they cannot benefit solely from it, they do need approaches that will assist them in monitoring big data and making the most of it so that they may more readily exploit big data in the healthcare sector.

The rising digitalization of healthcare information is being analyzed using innovative methodologies in order to enhance the quality of treatment, the outcomes of health care, and the costs of healthcare. Organizations must examine internal and external patient data in order to more properly assess risk and results. Meanwhile, many customers are attempting to boost data openness in order to generate fresh insight knowledge. In their research, Praveen Kumar et al (2014) claimed that Hadoop, which is based on Map Reduction, is a strong tool for managing a large quantity of information. Fault tolerant methods may be used in conjunction with this echo system. Large-scale clinical data analytics, as described by Emad A Mohammed and colleagues (2014), would stress the modelling of whole interacting processes in clinical contexts, with the clinical data sets themselves being evolutions of very large-scale datasets. In 2014, Arantxa Duque Barrachina et al claimed that huge datasets may be utilized to identify large datasets by using Hadoop approaches, which they described as follows: K. Divya and colleagues (2014) utilized a progressive encryption approach to safeguard the information. Using a novel Hadoop-based biosensor Sunspot wireless network architecture, ECC digital signature algorithm, MySQL database, and Hadoop HDFS cloud storage, Sabia and SheetalKalra (2014) developed a novel Hadoop-based biosensor Sunspot wireless network architecture, which security administrators can use to protect and manage key data. The SWOT (Strengths, Weaknesses, Opportunities, and Threats) study conducted by Lidong

Wang and colleagues (2015) revealed that Radio Frequency Identification Technology (RFID) is a promising technology (RFID).

2.7 Critical Observations

The growing number of patients, along with the advent of novel symptoms and illnesses, has made health monitoring and evaluation a difficult duty for medical professionals and hospitals in today's environment. Many health-based sensing applications face significant hurdles due to the processing of large and diverse data gathered by biomedical sensors, as well as the need for patient categorization and illness detection. As a result, the combination of remote sensing devices with big data technology has been shown to be an efficient and low-cost option for healthcare applications. We can get to the conclusion that Hadoop technology is the most effective method of simplifying difficulties in the healthcare system by using Hadoop technology. It can be applied quickly and easily in healthcare systems since it eliminates all of the constraints of conventional healthcare systems. Consequently, we can provide individualized therapies to all patients at very little cost, and we can evaluate the data in a single data store to come up with a permanent therapy for the specific ailment in question. Most importantly, Hadoop technology has provided a solution to the big data challenge that has plagued healthcare organizations. It will be employed in the future in all healthcare systems to make the task easier for everyone involved.

CHAPTER 3

PROPOSED FRAMEWORK FOR HEALTH CARE DATA PRIVACY USING HDFS SYSTEM

3.1 Behavior Statement

In recent research efforts, numerous healthcare frameworks have been advocated for the purpose of managing enormous volumes of diverse data derived from a variety of data sources in order to generate important patterns and trends. This section examines the aforementioned cloud data frameworks and focuses on their contributions to the field of healthcare. An applied architecture framework for the healthcare system that makes use of cloud data analytics has been suggested by industry professionals.

The Data Source layer, the Transformation layer, the cloud Data platform layer, and the Analytical layer are some of the layers that make up the framework. The data source layer focuses its attention primarily on the internal and external data sources that are associated with healthcare and may be situated in a variety of different locations and presented in a variety of different formats. These data sources may be accessed through a variety of different channels. It is the responsibility of the transformation layer to perform tasks such as the extraction, transformation, and loading of data into the cloud data platform utilizing a variety of data staging approaches, including those involving middleware and data warehousing processes.

The Map-Reduce programming style is used by the layer of the cloud platform that is comprised of numerous Hadoop ecosystem tools. These tools are used to conduct certain operations on the Hadoop Distributed File System (HDFS). Operations like as querying, reporting, online analytical processing, and data mining methods are carried out by the analytical layer.

In addition, the authors have provided an overview of a variety of platforms and technologies that may be used to analyzed cloud-based healthcare data. The emphasis of the suggested architectural framework is only on the theoretical elements, despite the fact that it is a pioneering one in the context of cloud data for the field of healthcare. The suggested framework has not been used to develop any experimental or evaluation plans at this time. Can provide an unrivalled level of protection for the data repository used by contemporary computer systems and operated through cloud computing.

Hadoop is the primary platform for organizing data stored in the Cloud, and it answers the challenge of how to make the data meaningful for analytical purposes.

The provision of secure medical care ought to both raise the general standard of medical treatment and lower its associated costs. In the course of our study, we came up with a plan for an infrastructure that guarantees the safety of patient medical records.

Despite the fact that diverse frameworks are intended to meet particular healthcare objectives, the frameworks themselves are well suited for adopting standard architectural guidelines in order to carry out activities such as data collection, data preprocessing, data analysis, interpretation, and visualization.

Experts such as data scientists must be very careful when picking the right tools for each step of the framework's design and deployment because of the healthcare framework's domain-specific nature. These reasons are related to its usage within a health care context: Here, we'll provide an overview of the different technologies that are used to do tasks such as integrating data, injecting intelligence, searching and indexing, stream data processing, and data visualization. Included in these activities are: They all have a role in the effective performance of their obligations.

Recent advancements in digital technology have an influence on the field of healthcare, and the transition to the keeping of electronic patient records may represent a paradigm change. The amount of information pertaining to healthcare is quite high, and this might lead to an expansion of its nature in terms of protection, complexity, multiplicity, and appropriateness, which ultimately results in processing.

The data must be handled and preserved in accordance with certain statutory requirements, as well as the potential to enhance treatment, safeguard lives, and reduce costs associated with cloud data storage. Continue to provide a wide range of use cases, including clinical decision making, insurance for the health care industry, disease surveillance monitoring, health and population management, difficult or emerging event control or monitoring, and patient optimization for diseases affecting multiple organ systems.

The use of healthcare information maintenance digital technologies in the healthcare sector not only offers a wide variety of benefits and prospects, but it also makes it possible to address a greater number of obstacles and problems. In point of fact, the unease regarding the security of private and confidential data, as well as the protection of one's privacy, is being alleviated each year as a result of the proliferation of new technologies in the healthcare industry. Cloud computing, wireless sensor networks, and information interchange of healthcare data are among these technologies.

In addition, healthcare organizations discovered that different techniques, such as reactive, bottom-up, and centric technical ways to uncover privacy and Strict security requirements would be insufficient to protect the healthcare industry and its patients' personal information. All healthcare organizations should take a realistic and constructive attitude, take a high protection strategy, and use computations to guard against breaches of patient information and other types of security incidents. Taking privacy and security into consideration, this should be done.

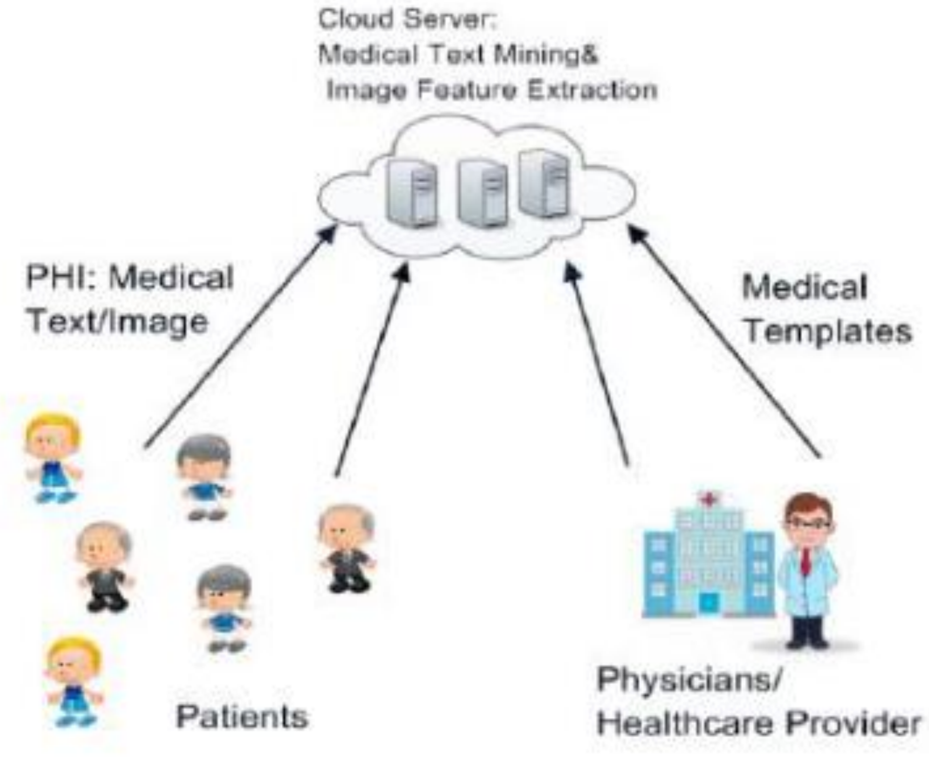


Fig 3.1: Cloud storage assistance for e-Healthcare system [1]

A novel cloud computing method that can safely store private information pertaining to the healthcare industry is beginning to emerge in the medical environment. Moving toward electronic health records and storing them in a cloud storage environment is a good place to start if you're looking to tackle a wide variety of healthcare sectors. A healthcare cloud service provider's ability to simplify digital medical information storage exchanges across multiple clinics and hospitals may be limited, even though it creates data to represent the healthcare record medical Centre.

Cloud computing, which is based on private and encrypted cloud data service providers, is becoming the preferred method for the storing of data by healthcare organizations. The selection of cloud computing service providers is based on the elimination of repetitive tasks associated with administration of infrastructure and the reduction of costs associated with maintenance and development. The fact that healthcare information is stored in the cloud environment makes it possible to provide treatment that is both methodical and effective. This is accomplished by retrieving previous information regarding the medical histories of a variety of patients via accessing a variety of databases using authentication provided by the medical cloud environment in order to provide appropriate information regarding the health problem of each individual patient.

The cloud computing infrastructure for the health care industry that ensures the confidentiality and safety of patient information. Machine learning is used in healthcare to preserve and secure data during processing and transport. The machine learning classification technique is superior for categorizing healthcare information, and it's employed in cloud computing to ensure information protection and preservation. The patient's medical history is the most sensitive and important piece of information to keep private throughout the diagnostic evaluation and decision-making processes.

3.1.1 Cloud Integrating Open Data

Our contention, which is supported by the research that has been done previously as well as the experiences that have been gained in the field, is that the movement of data integration applications into clouds, particularly public clouds, can cause existing information system architectures to become destabilized. There will always be expenses associated with the redesign of a new system, and depending on the circumstances, it may not even be achievable at all. The possible benefits of an all-encompassing data virtualization, which will be outlined in the next chapter, need to be weighed against the associated risks and costs. In order to appraise and evaluate the various solutions and offers for cloud data integration, we are going to carry out a design study in which we will attempt to discover answers to the following primary challenges:

- How can an application for the universal secure virtualization of health cloud data be designed and constructed such that it may merge data from inside organizations with data from external sources?

- What are the requirements in terms of both the technical and organizational aspects?

- How can we reintegrate and modify the data integration infrastructures that are already in place?

- Are private cloud integration solutions capable of acting as a transitional solution that can subsequently be used in conjunction with secure cloud data integration solutions?

3.2 Proposed Framework

Figure 3.2 depicts the framework for securing cloud data. This framework is based on the data privacy model that will describe the specifics of each component and apply the necessary security technologies for implementation between components in the privacy data at cloud environment. Figure 3.2. The framework for securing cloud data. The following describes the access control procedure that is used to provide flexible service on each component:

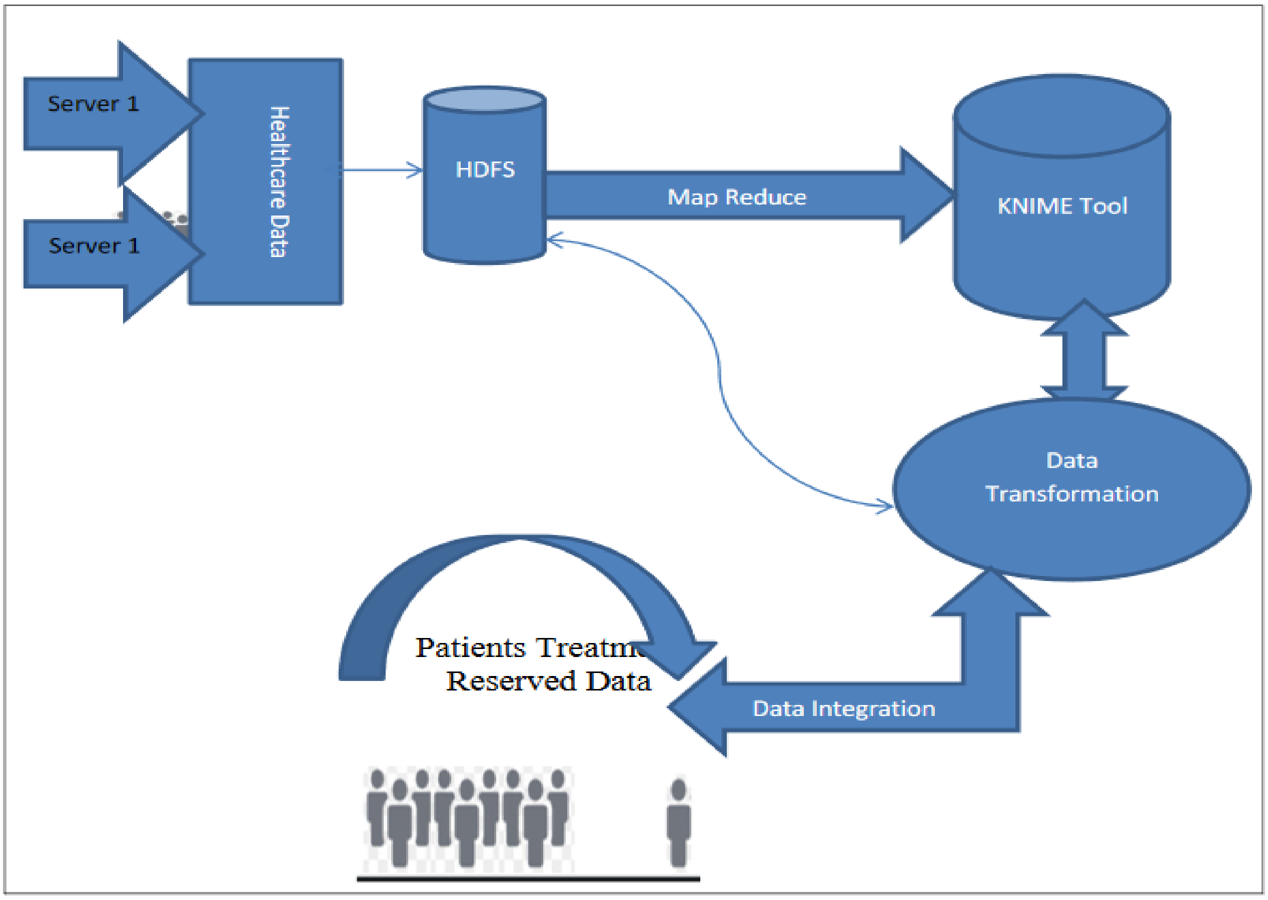


Fig 3.2: Proposed Framework

3.2.1 Health Care Data:

Historically, the healthcare industry has had a poor track record of making effective use of technology, particularly in the area of enhancing the quality of care provided to patients. Healthcare has entered the sixteenth year after the millennium, although many systems still depend on paper for notification and decision-making, medical records are one of the most common examples of systems that run manually. This is due to the fact that the number of people working in healthcare has not kept pace with technological advancements. The healthcare sector is quite different from other industries, and the primary distinctions between it and other industries may be broken down into three distinct categories.

To begin, this industry is subject to extensive government supervision, which includes restrictions designed to keep patients safe. The costs associated with high-risk mistakes that might occur in the healthcare business are higher than those associated with errors in other industries, and lastly, this industry is made up of a large number of individual units, such as hospital administration personnel, laboratories, and patients. Any false criteria might cause serious harm or even death. Due to the delicate nature of healthcare and patient data, any misleading criteria will trigger this.

As a consequence, the delicate nature of the data handling process may be preserved even after the use of cutting-edge technology.

The cost dynamic is the primary or primary strength of using cloud computing in the healthcare industry. Because these plans adhere to the host initially, there are no start-up costs associated with them, which results in a significant reduction in the running costs that are incurred (Tamil Ilakkiya, 2015). When compared to the information technology sector, the healthcare industry is in a better position to weather the heavy costs since it has less resources at its disposal. This circumstance is seen as a positive for the healthcare industry.

In addition, access to the patient's information may be gained anytime, anywhere, and in a manner that is extremely slick and simple. This infers, in a roundabout way, an increase in the alliance between patients and physicians, as well as an intensification of the quality of services provided to patients. Any criterion that is deceptive will create serious damage and might even lead to life or death in certain cases because of the exceptional privacy of healthcare and the security of patients' data. This makes the data itself sensitive.

As a consequence, the adoption of new technology may result in the management of sensitive data being carried out with less haste. The healthcare industry as a whole is undergoing widespread reform, which in turn leads the healthcare information technology (HIT) to undergo modernization. Cloud computing is, without a doubt, at the core of this change and serves as a conduit for the route that reform is taking.

3.2.2 Connectivity with HDFS:

Hadoop clusters are seeing an increase in the number of applications being executed on them daily. This is because businesses have discovered a model that is straightforward and effective, as well as one that functions well in a decentralized setting (Saeid, 2015). The model is constructed such that it may function well across thousands of computers and enormous data sets while employing technology that is readily available. HDFS and Map Reduce together make up a paradigm that is scalable, fault-tolerant, and conceals all the complexity involved in cloud data security.

Because of Hadoop's growing popularity, it is vital to have a solid grasp of the underlying technological intricacies. Because of this reason, we were compelled to investigate Hadoop and all of its constituent parts in more detail. To successfully extract the necessary information from a massive volume of unstructured data, one must first tackle the challenging process of evaluating, inspecting, and processing the data.

Throughout this chapter, we will describe Hadoop and its components in great depth. These include map Reduce and Hadoop DFS (HDFS). HDFS is a distributed file system with three nodes: Name Node, Data Node, and Secondary Name Node. The Hadoop Distributed File System (HDFS) allows dependable, scalable, fault-tolerant data storage on commodity hardware.

It does this by spreading storage and processing over large clusters by combining storage resources that may expand based on requests and queries while staying affordable and within budget. It collaborates closely with map Reduce to do this. HDFS is not architecture-specific;

it can read data in any format, including text, photos, and videos, for example, and will automatically optimize it for high bandwidth streaming (Maslin, 2015).

HDFS's ability to tolerate errors is the primary benefit it offers. HDFS is able to offer a scale-out storage solution for Hadoop in addition to ensuring rapid data movement between nodes and allowing Hadoop to continue providing service even in the event that individual nodes fail to function. This reduces the likelihood of a catastrophic failure occurring.

The HDFS serves as the foundation for the creation of the file management system for cloud computing. The fundamental concept behind HDFS is developed from the architecture and principles of HDFS. According to the qualities of HDFS, the system is made up of three components: the cluster of primary service control center, the storage service cluster, and the client. And the data nodes that make up the storage service cluster are geared up with the qualities of being able to handle a high volume as well as a widespread dispersion of data. In the file system, it is the data storage center that has a big capacity for storing information.

The processing of real-time data access activities is primarily within the purview of the data nodes. In addition, the data will not be subjected to any extensive processing by the storage services. Massive quantities of storage space are made available in order to guarantee the accuracy and accessibility of the data, but this is only done in response to commands issued by the name node and requests for data access activities made by users (G. Federico, 2005).

Figure 3.2 presents the Framework in its entirety. The overall structure is composed of three distinct components. The interface with users, which is the most essential factor in determining the quality of the system's user experience, is mostly the responsibility of the client side. The client side of the system is responsible for directly delivering service requests from all types of users to the master service. The data in the file is fragmented thanks to the response information provided by the control center. Based on the redundancy strategy of the

HDFS data block, the data is transferred to the associated storage nodes in accordance with the parameters that are applied to the block size. The metadata mapping information that is given by the control servers is used to combine the block data that is scattered throughout the many data nodes that make up the download system into a single full file.

3.3 MAP Reduce:

Map Reduce allows for a large volume of data to be handled quickly and conveniently while maintaining a high degree of parallelism. In addition, these applications are capable of running on clusters comprised of commodity hardware, making it an appropriate candidate for scaling. Java serves as the foundation for Map Reduce. The dataflow for the typical Map Reduce operation looks like what is depicted in Figure 3.3.

In the Map task, individual items are partitioned into tuples, which are sometimes referred to as key-value pairs (R. Zhang, 2011).

These tuples are considered intermediate and will be used as an input for the next step of the reduce job. After that, the Reduce job will merge it into a less number of tuples. The Reduce job cannot be begun until after the Map task has been finished.

For building applications that reliably and fault-tolerantly handle large amounts of data in parallel on commodity hardware resources, Hadoop Map Reduce is a framework. These applications may be written with the help of the Map Reduce framework (J. Rama Prabha, 2019). In the beginning, the data is split up into individual pieces by a Map Reduce task. These chunks are then processed in parallel by Map jobs.

The outputs of the maps are first passed through the framework's sorting process, and then they are utilized as input for the reduction operations. The input data and the output data from the job are typically stored in the same area inside the file system. This is the case in the majority of circumstances. The framework is accountable for the scheduling, monitoring, and re-execution of tasks in the event that they are not completed properly the first time.

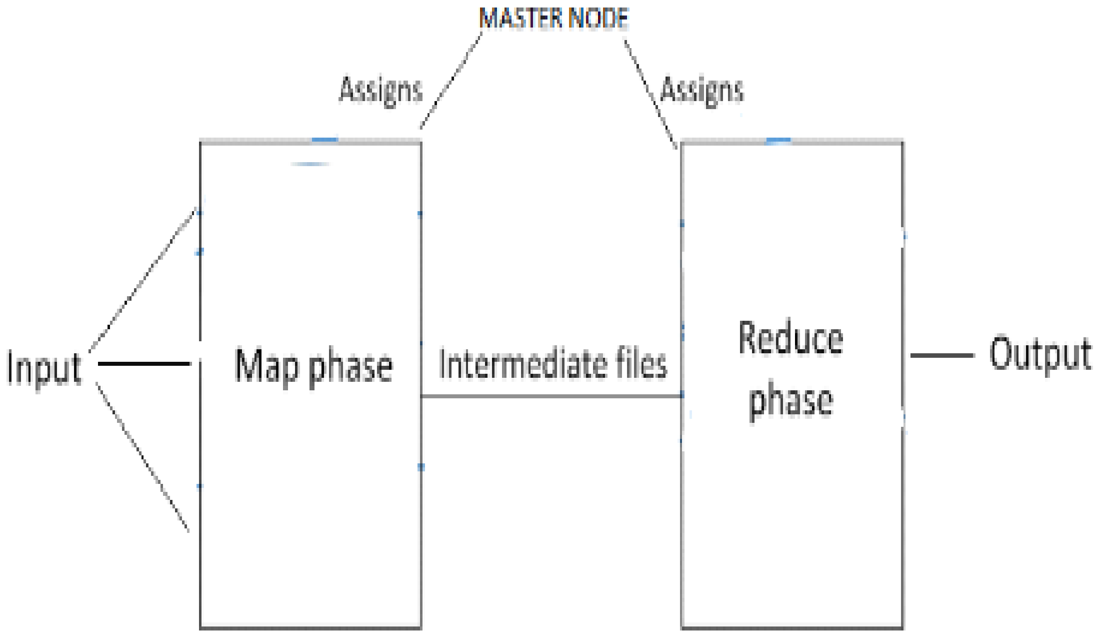


Fig 3.3: The general Map Reduce dataflow

3.4 Map Reduce Core Functions

a) Input reader

The input is broken up into smaller sections or blocks. After that, a Map function is given responsibility for these blocks.

b) Map function

Each individual element is partitioned into tuples, which are also referred to as key-value pairs.

c) Shuffle and Sort

A function of partitioning using the provided key and the specified number of reducers, it locates the appropriate reducer.

The map's intermediate outputs are arranged in descending order according to the results of this comparison function.

d) Reduce function

Combines tuples at intermediate levels to produce a more manageable group of tuples, which is then sent to output.

e) Output writer

Gives file output.

In Map Reduce, the processing of a reducer cannot begin until all of the mapping jobs have been finished. The fact that reducers have to wait for longer than necessary is the most significant disadvantage of using this method. To put it another way, this is not an effective or efficient use of the resources that are available.

3.5 Data Transformation:

In the modelling process, specialized domain specialists construct the data model. They gather the data, clean it up, and format it since some of the mining tools will only take data in a certain format. In addition to this, they provide brand new derived qualities, such as an ever-aging value. During the phase known as "data preparation," the data are modified several times and in no particular sequence. In this phase, common responsibilities include preparing the data for the modelling tool by choosing tables, records, and characteristics (Then, 2012). The interpretation of the data has not been altered in any way. Transformations that take place in the cloud need the dynamic redistribution of resources across cloud infrastructure. This transfer of data from one data processor to another without the express agreement of the data subject constitutes a risk to the data subject's right to privacy from a legal point of view. In order to keep people's faith in cloud computing, its infrastructure providers need to provide various degrees of assurance and responsibility to the people whose data is being stored there. In situations involving C Cloud Transformation, there are often numerous suppliers involved, and it is necessary that responsibility be passed down the chain.

3.6 Data Integration

Integration of information systems is a crucial responsibility that falls on the shoulders of software architects and developers. Both academic institutions and private businesses have been active in this area of research for at least the last two decades. This has led to the development of a vast number of customized software architectural patterns and solutions for the integration of information systems. These patterns and products address specific areas of integration (Tamil Ilakkiya, 2015).

This section will concentrate on data integration as one of these components and will investigate the potential and problems presented by the application of the cloud paradigm to this subject. We will argue that integrating data via the cloud is probably not a silver bullet for all of an organization's integration problems, and what needs to be investigated is whether kind of technological architecture is the greatest match for an organization given its demands (Rolim, 2010). For application areas such as business intelligence (BI), customer relationship management (CRM), and master data management (MDM), advanced data integration technologies and solutions that solve the inherent integration difficulties are a necessity.

Read-only applications used to be the primary emphasis of these kinds of solutions, which means that the integrated data view did not alter the data in any way. Read-and-write situations, on the other hand, which include the modification of integrated data, seem to be becoming increasingly significant.

3.6 Summary

Cloud computing is vulnerable to a broad range of security risks because it uses a wide range of technologies, including networks, databases, operating systems, virtualization, resource scheduling, transaction management, load balancing, concurrency control, and memory management. There are thus similar worries about cloud computing in terms of security. Using big data analytics and mobile cloud computing, we provide a framework for secure Health Information Systems (HISs) in this chapter. A high degree of integration, interoperability, and sharing of electronic health records is provided by the framework, which is accessible to healthcare providers, patients, and practitioners. Cloud computing enables consumers to have quick Internet connection, share electronic health records (EHRs), and provide access to them. The use of big data analytics helps to evaluate patient data so that the appropriate intervention may be provided to the appropriate patient at the appropriate time. The suggested framework implements a number of security restrictions and access controls, which together ensure the data's integrity, confidentiality, and privacy. The framework that has been suggested has as its ultimate goal the introduction of a new generation of HISs that are able to provide consumers healthcare services that are both of high quality and cheap cost. This will be achieved by using a variety of technologies, including mobile computing, cloud computing, and big data analytics, in conjunction with one another. In the not too distant future, our intention is to build and execute HIS based on the framework that was suggested.

Chapter 4

Model Development

4.1 Personalized Healthcare Model

The use of medicine is essential to both the upkeep and extension of life. Because not all of the body's systems are clinically the same, a person's medication has to be tailored to their particular body system in order to be effective. While Remdisivir and Tocilizumab have been shown to be effective for one subset of patients, it has been found that the same subset of patients with almost similar clinical characteristics is unable to prevent another subset from developing from a mild or moderate state to a severe stage. This phenomenon has been attributed to the fact that one set of medicines works for one category of patients. Because it takes a more "customized" approach, personalized medicine presents an opportunity to find a solution to this problem. Alternative names for it include precision medicine, tailored medicine, and personalized medicine.

Machine Learning (ML) and Artificial Intelligence (AI) are often grouped together (AI). Nevertheless, machine learning (ML) is a subfield of AI that discovers variable patterns of data in order to anticipate or categories patterns that are hidden or invisible. This information, in turn, may be used for exploratory data analysis, data mining, and data modelling. The machine learning algorithms point to the prospect of locating drugs with specific targets based on clinical, genetic, laboratory, nutritional, and lifestyle-related data.

Throughout the course of this essay, the terms machine learning (ML) and artificial intelligence (AI) as well as personalized medicine will be used interchangeably. These terms will also be referred to as precision medicine, individualized medicine, and customized medicine. It is expected that those who read this paper have a fundamental understanding of medical jargon, Python, and data science.

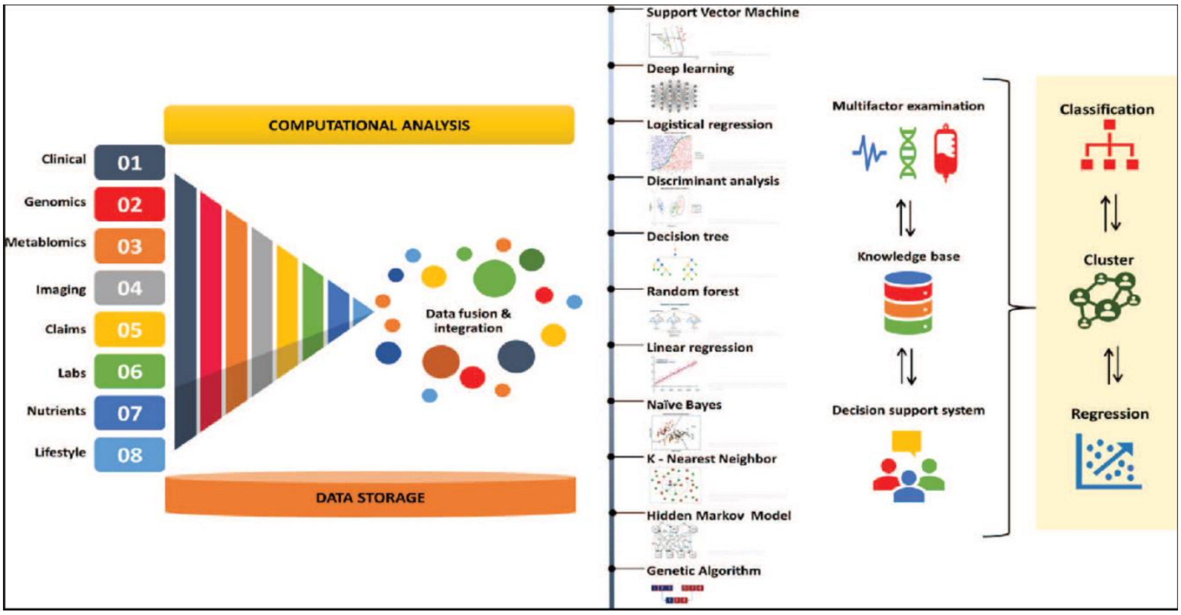


Fig 4.1: Steps of Personalized Healthcare Model

Steps

1. Classification – The two methods for supervised learning classification that are used most often are known as Logistic Regression and Naive Bayes.

2. Regression – Linear Regression is the most common supervised learning regression algorithm.

3. Clustering – K-means Algorithm, Mean Shift Algorithm, and Hierarchical Clustering are the common algorithms. These are all unsupervised, i.e. target variable is not available.

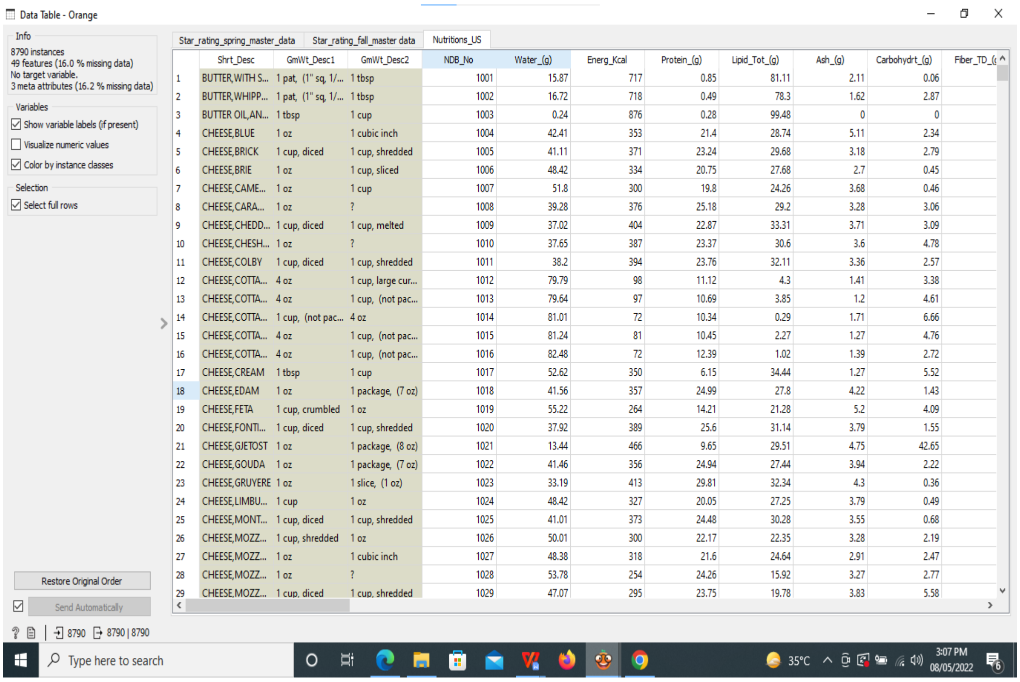
4. Classification and Regression combined – The Support Vector Machine (SVM), Decision Tree, Random Forest, and K-Nearest Neighbors are all examples of supervised machine learning algorithms that may be used in the context of solving predictive classification and regression issues respectively.

Reinforcement learning is another highly significant sort of machine learning that can be used both when there is a category target variable present and when there is no target variable present. This type of learning may be used in any circumstance. It has several potential applications in the fields of automobiles and in marketing that is optimized. It is a kind of algorithm known as semi-supervised learning.

4.2 Dataset

This dataset was produced by individuals who participated in a distributed survey using Amazon Mechanical Turk between December 3rd and December 5th, 2016. Thirty users who were eligible for a Fit bit provided their approval for the transmission of their personal tracker data. This data included minute-level output for tracking of their physical activity, heart rate, and sleep. Individual reports may be analyzed using the export session ID (column A) or the timestamp as a starting point (column B). The difference in output may be attributed to individuals' unique monitoring habits and preferences, as well as the usage of various Fit bit tracker models.

https://www.kaggle.com/datasets/arashnic/fitbit



Data Table - Orange

Info
8790 instances
49 features (16.0 % missing data)
No target variable.
3 meta attributes (16.2 % missing data)

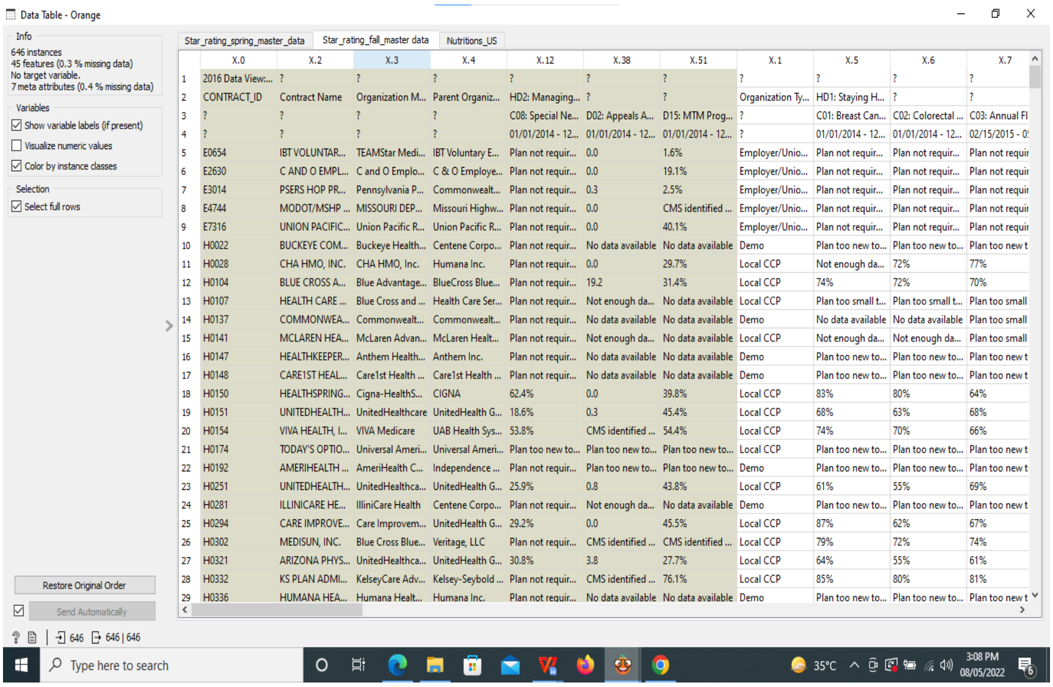
Variables
☑ Show variable labels (if present)
□ Visualize numeric values
☑ Color by instance classes

Selection
☑ Select full rows

	Shrt_Desc	GmWt_Desc1	GmWt_Desc2	NDB_No	Water_(g)	Energ_Kcal	Protein_(g)	Lipid_Tot_(g)	Ash_(g)	Carbohydrt_(g)	Fiber_TD_(
1	BUTTER,WITH S...	1 pat, (1" sq, 1/...	1 tbsp	1001	15.87	717	0.85	81.11	2.11	0.06	
2	BUTTER,WHIPP...	1 pat, (1" sq, 1/...	1 tbsp	1002	16.72	718	0.49	78.3	1.62	2.87	
3	BUTTER OIL,AN...	1 tbsp	1 cup	1003	0.24	876	0.28	99.48	0	0	
4	CHEESE,BLUE	1 oz	1 cubic inch	1004	42.41	353	21.4	28.74	5.11	2.34	
5	CHEESE,BRICK	1 cup, diced	1 cup, shredded	1005	41.11	371	23.24	29.68	3.18	2.79	
6	CHEESE,BRIE	1 oz	1 cup, sliced	1006	48.42	334	20.75	27.68	2.7	0.45	
7	CHEESE,CAME...	1 oz	1 cup	1007	51.8	300	19.8	24.26	3.68	0.46	
8	CHEESE,CARA...	1 oz	?	1008	39.28	376	25.18	29.2	3.28	3.06	
9	CHEESE,CHEDD...	1 cup, diced	1 cup, melted	1009	37.02	404	22.87	33.31	3.71	3.09	
10	CHEESE,CHESH...	1 oz	?	1010	37.65	387	23.37	30.6	3.6	4.78	
11	CHEESE,COLBY	1 cup, diced	1 cup, shredded	1011	38.2	394	23.76	32.11	3.36	2.57	
12	CHEESE,COTTA...	4 oz	1 cup, large cur...	1012	79.79	98	11.12	4.3	1.41	3.38	
13	CHEESE,COTTA...	4 oz	1 cup, (not pac...	1013	79.64	97	10.69	3.85	1.2	4.61	
14	CHEESE,COTTA...	1 cup, (not pac...	4 oz	1014	81.01	72	10.34	0.29	1.71	6.66	
15	CHEESE,COTTA...	4 oz	1 cup, (not pac...	1015	81.24	81	10.45	2.27	1.27	4.76	
16	CHEESE,COTTA...	4 oz	1 cup, (not pac...	1016	82.48	72	12.39	1.02	1.39	2.72	
17	CHEESE,CREAM	1 tbsp	1 cup	1017	52.62	350	6.15	34.44	1.27	5.52	
18	CHEESE,EDAM	1 oz	1 package, (7 oz)	1018	41.56	357	24.99	27.8	4.22	1.43	
19	CHEESE,FETA	1 cup, crumbled	1 oz	1019	55.22	264	14.21	21.28	5.2	4.09	
20	CHEESE,FONTI...	1 cup, diced	1 cup, shredded	1020	37.92	389	25.6	31.14	3.79	1.55	
21	CHEESE,GJETOST	1 oz	1 package, (8 oz)	1021	13.44	466	9.65	29.51	4.75	42.65	
22	CHEESE,GOUDA	1 oz	1 package, (7 oz)	1022	41.46	356	24.94	27.44	3.94	2.22	
23	CHEESE,GRUYERE	1 oz	1 slice, (1 oz)	1023	33.19	413	29.81	32.34	4.3	0.36	
24	CHEESE,LIMBU...	1 cup	1 oz	1024	48.42	327	20.05	27.25	3.79	0.49	
25	CHEESE,MONT...	1 cup, diced	1 cup, shredded	1025	41.01	373	24.48	30.28	3.55	0.68	
26	CHEESE,MOZZ...	1 cup, shredded	1 oz	1026	50.01	300	22.17	22.35	3.28	2.19	
27	CHEESE,MOZZ...	1 oz	1 cubic inch	1027	48.38	318	21.6	24.64	2.91	2.47	
28	CHEESE,MOZZ...	1 oz	?	1028	53.78	254	24.26	15.92	3.27	2.77	
29	CHEESE,MOZZ...	1 cup, diced	1 cup, shredded	1029	47.07	295	23.75	19.78	3.83	5.58	

8790 | 8790 | 8790

Fig 4.2: Dataset table (a)



Data Table - Orange

Info
646 instances
45 features (0.3 % missing data)
No target variable.
7 meta attributes (0.4 % missing data)

Variables
☑ Show variable labels (if present)
□ Visualize numeric values
☑ Color by instance classes

Selection
☑ Select full rows

	X.0	X.2	X.3	X.4	X.12	X.38	X.51	X.1	X.5	X.6	X.7
1	2016 Data View:...	?	?	?	?	?	?	?	?	?	?
2	CONTRACT_ID	Contract Name	Organization M...	Parent Organiz...	HD2: Managing...	?	?	Organization Ty...	HD1: Staying H...	?	?
3	?	?	?	?	C08: Special Ne...	D02: Appeals A...	D15: MTM Prog...	?	C01: Breast Can...	C02: Colorectal ...	C03: Annual Fl
4	?	?	?	?	01/01/2014 - 12...	01/01/2014 - 12...	01/01/2014 - 12...	?	01/01/2014 - 12...	01/01/2014 - 12...	02/15/2015 - 0
5	E0654	IBT VOLUNTAR...	TEAMStar Medi...	IBT Voluntary E...	Plan not requir...	0.0	1.6%	Employer/Unio...	Plan not requir...	Plan not requir...	Plan not requir
6	E2630	C AND O EMPL...	C and O Emplo...	C & O Employe...	Plan not requir...	0.0	19.1%	Employer/Unio...	Plan not requir...	Plan not requir...	Plan not requir
7	E3014	PSERS HOP PR...	Pennsylvania P...	Commonwealt...	Plan not requir...	0.3	2.5%	Employer/Unio...	Plan not requir...	Plan not requir...	Plan not requir
8	E4744	MODOT/MSHP ...	MISSOURI DEP...	Missouri Highw...	Plan not requir...	0.0	CMS identified ...	Employer/Unio...	Plan not requir...	Plan not requir...	Plan not requir
9	E7316	UNION PACIFIC...	Union Pacific R...	Union Pacific R...	Plan not requir...	0.0	40.1%	Employer/Unio...	Plan not requir...	Plan not requir...	Plan not requir
10	H0022	BUCKEYE COM...	Buckeye Health...	Centene Corpo...	Plan not requir...	No data available	No data available	Demo	Plan too new to...	Plan too new to...	Plan too new t
11	H0028	CHA HMO, INC.	CHA HMO, Inc.	Humana Inc.	Plan not requir...	0.0	29.7%	Local CCP	Not enough da...	72%	77%
12	H0104	BLUE CROSS A...	Blue Advantage...	BlueCross Blue...	Plan not requir...	19.2	31.4%	Local CCP	74%	72%	70%
13	H0107	HEALTH CARE ...	Blue Cross and ...	Health Care Ser...	Plan not requir...	Not enough da...	No data available	Local CCP	Plan too small t...	Plan too small t...	Plan too small
14	H0137	COMMONWEA...	Commonwealt...	Commonwealt...	Plan not requir...	No data available	No data available	Demo	No data available	No data available	Plan too small
15	H0141	MCLAREN HEA...	McLaren Advan...	McLaren Healt...	Plan not requir...	Not enough da...	No data available	Local CCP	Not enough da...	Not enough da...	Plan too small
16	H0147	HEALTHKEEPER...	Anthem Health...	Anthem Inc.	Plan not requir...	No data available	No data available	Demo	Plan too new to...	Plan too new to...	Plan too new t
17	H0148	CARE1ST HEAL...	Care1st Health ...	Care1st Health ...	Plan not requir...	No data available	No data available	Demo	Plan too new to...	Plan too new to...	Plan too new t
18	H0150	HEALTHSPRING...	Cigna-HealthS...	CIGNA	62.4%	0.0	39.8%	Local CCP	83%	80%	64%
19	H0151	UNITEDHEALTH...	UnitedHealthcare	UnitedHealth G...	18.6%	0.3	45.4%	Local CCP	68%	63%	68%
20	H0154	VIVA HEALTH, I...	VIVA Medicare	UAB Health Sys...	53.8%	CMS identified ...	54.4%	Local CCP	74%	70%	66%
21	H0174	TODAY'S OPTIO...	Universal Ameri...	Universal Ameri...	Plan too new to...	Plan too new to...	Plan too new to...	Local CCP	Plan too new to...	Plan too new to...	Plan too new t
22	H0192	AMERIHEALTH ...	AmeriHealth C...	Independence ...	Plan not requir...	Plan too new to...	Plan too new to...	Demo	Plan too new to...	Plan too new to...	Plan too new t
23	H0251	UNITEDHEALTH...	UnitedHealthca...	UnitedHealth G...	25.9%	0.8	43.8%	Local CCP	61%	55%	69%
24	H0281	ILLINICARE HE...	IlliniCare Health	Centene Corpo...	Plan not requir...	Not enough da...	No data available	Demo	Plan too new to...	Plan too new to...	Plan too new t
25	H0294	CARE IMPROVE...	Care Improvem...	UnitedHealth G...	29.2%	0.0	45.5%	Local CCP	87%	62%	67%
26	H0302	MEDISUN, INC.	Blue Cross Blue...	Veritage, LLC	Plan not requir...	CMS identified ...	CMS identified ...	Local CCP	79%	72%	74%
27	H0321	ARIZONA PHYS...	UnitedHealthca...	UnitedHealth G...	30.8%	3.8	27.7%	Local CCP	64%	55%	61%
28	H0332	KS PLAN ADMI...	KelseyCare Adv...	Kelsey-Seybold ...	Plan not requir...	CMS identified ...	76.1%	Local CCP	85%	80%	81%
29	H0336	HUMANA HEA...	Humana Healt...	Humana Inc.	Plan not requir...	No data available	No data available	Demo	Plan too new to...	Plan too new to...	Plan too new t

646 | 646 | 646

Fig 4.2: Dataset table (b)

The training data set and the test data set are each sent in the form of two individual files apiece. The first one, training/test variants, gives information on the genetic mutations, and the second one, training/test text, gives the clinical evidence (text) that our human specialists utilized to classify the genetic mutations. Both of these are referred to as the training set. Both of these collections of data are used in the process of training and validating the system. Using the ID field, any option is accessible to us.

Therefore, the genetic mutation (row) with ID=15 in the file training variants was categorized using the clinical evidence (text) from the row with ID=15 in the file training text. This was done in order to ensure accuracy.

Last but not least, to ramp up the excitement! In order to avoid manual labelling, some of the data for the test was created by machine. We are going to provide us all of the findings that your classification algorithm produced, and we are going to disregard the samples that were created by the machine.

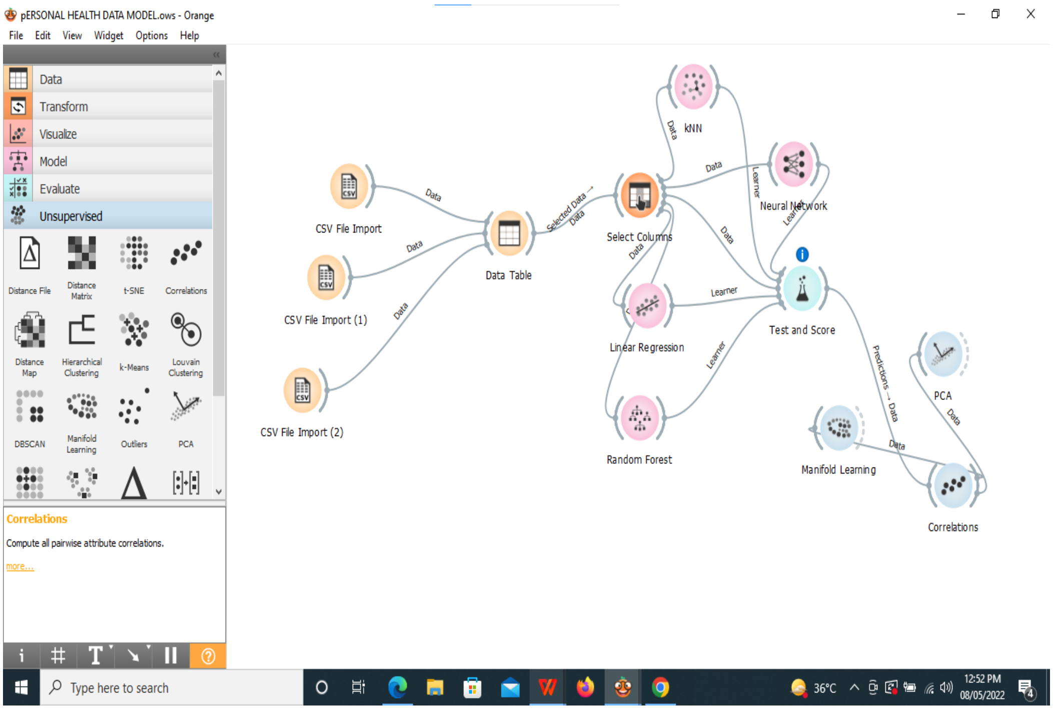


Fig 4.3: Model developed on Orange

The figure 4.4 is a implantation proposed model using Open source UI of Python, i.e. Orange, here we have collected all CSV files in one data table for development of basic model and further in second step select the column from the data and in third step applied

KNN

Neural Network

And LINEAR REGRATION

Random Forest algorithms

On the basis of these algorithm of ML generated the test and Score table as a result and shown in chapter 5.

4.3 Public Healthcare Model

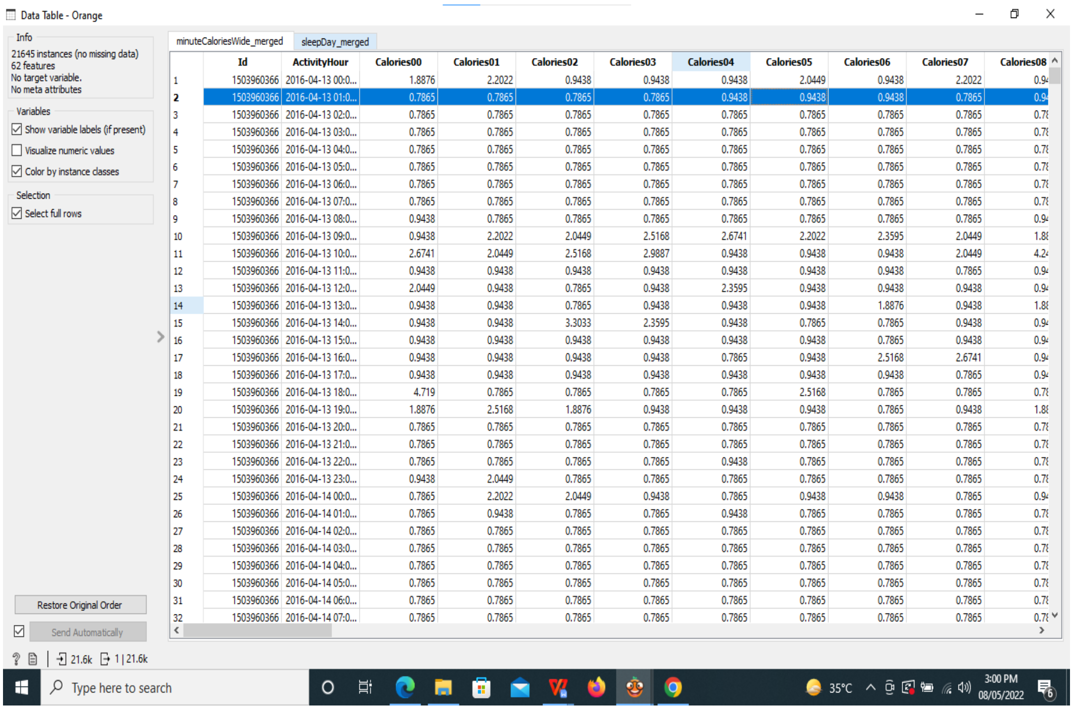
For many years, a critical dilemma in the field of public health care data analytics has existed. In order to determine how to provide the highest possible level of care, organizations that provide public health services need to be able to effectively store, examine, and interpret data. In the field of health care, there is a diverse selection of tools for data analytics, including clinical and operational applications that may assist companies in capturing health data for the purpose of increasing medical treatment. Data pertaining to medical treatment is compiled from a wide range of computer systems and mobile devices, including online patient portals, electronic medical records, and health monitoring equipment. As a direct consequence of this, data may be found in a variety of forms, such as clinical notes and medical photographs, and it can even be unstructured at times. Data governance includes master data management, which compiles all of an organization's master data into a single, trustworthy source of information that can be utilized to enhance patient care and ensure their safety. The term "data analytics" refers to the process of doing some kind of analysis on the data by making use of both quantitative and qualitative methods in order to look for trends and patterns within the data. The sophisticated expertise required "to gather, manage, analyses, interpret, and convert data into accurate, consistent, and timely information" should be had by analysts who work with health data.

Dataset

Context

The Cleveland, Hungary, Switzerland, and Long Beach V databases make up this data collection, which was compiled in 1988 and dates back to that year. It has 76 properties, including the attribute that was predicted, however all of the published tests only relate to employing a subset of 14 of those features. The existence of cardiac disease in the patient is referred to as the "target" field in this context. It is an integer where 0 indicates there is no illness and 1 indicates there is disease.

https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset



Info
21645 instances (no missing data)
62 features
No target variable.
No meta attributes

Variables
☑ Show variable labels (if present)
□ Visualize numeric values
☑ Color by instance classes

Selection
☑ Select full rows

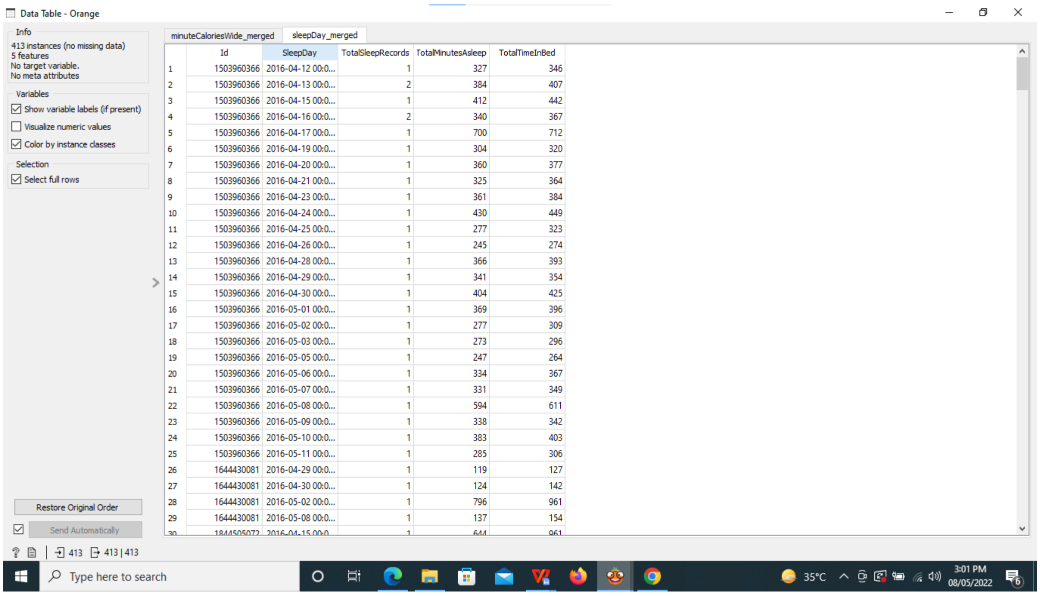
☑ Send Automatically

21.6k | 1 | 21.6k

minuteCaloriesWide_merged | sleepDay_merged

	Id	ActivityHour	Calories00	Calories01	Calories02	Calories03	Calories04	Calories05	Calories06	Calories07	Calories08
1	1503960366	2016-04-13 00:0...	1.8876	2.2022	0.9438	0.9438	0.9438	2.0449	0.9438	2.2022	0.94
2	1503960366	2016-04-13 01:0...	0.7865	0.7865	0.7865	0.7865	0.9438	0.9438	0.9438	0.7865	0.94
3	1503960366	2016-04-13 02:0...	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.78
4	1503960366	2016-04-13 03:0...	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.78
5	1503960366	2016-04-13 04:0...	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.78
6	1503960366	2016-04-13 05:0...	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.78
7	1503960366	2016-04-13 06:0...	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.78
8	1503960366	2016-04-13 07:0...	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.78
9	1503960366	2016-04-13 08:0...	0.9438	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.94
10	1503960366	2016-04-13 09:0...	0.9438	2.2022	2.0449	2.5168	2.6741	2.2022	2.3595	2.0449	1.88
11	1503960366	2016-04-13 10:0...	2.6741	2.0449	2.5168	2.9887	0.9438	0.9438	0.9438	2.0449	4.24
12	1503960366	2016-04-13 11:0...	0.9438	0.9438	0.9438	0.9438	0.9438	0.9438	0.9438	0.7865	0.94
13	1503960366	2016-04-13 12:0...	2.0449	0.9438	0.7865	0.9438	2.3595	0.9438	0.9438	0.9438	0.94
14	1503960366	2016-04-13 13:0...	0.9438	0.9438	0.7865	0.9438	0.9438	0.9438	1.8876	0.9438	1.88
15	1503960366	2016-04-13 14:0...	0.9438	0.9438	3.3033	2.3595	0.9438	0.7865	0.7865	0.9438	0.94
16	1503960366	2016-04-13 15:0...	0.9438	0.9438	0.9438	0.9438	0.9438	0.9438	0.7865	0.9438	0.94
17	1503960366	2016-04-13 16:0...	0.9438	0.9438	0.9438	0.9438	0.7865	0.9438	2.5168	2.6741	0.94
18	1503960366	2016-04-13 17:0...	0.9438	0.9438	0.9438	0.9438	0.9438	0.9438	0.9438	0.7865	0.94
19	1503960366	2016-04-13 18:0...	4.719	0.7865	0.7865	0.7865	0.7865	2.5168	0.7865	0.7865	0.78
20	1503960366	2016-04-13 19:0...	1.8876	2.5168	1.8876	0.9438	0.9438	0.9438	0.7865	0.9438	1.88
21	1503960366	2016-04-13 20:0...	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.78
22	1503960366	2016-04-13 21:0...	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.78
23	1503960366	2016-04-13 22:0...	0.7865	0.7865	0.7865	0.7865	0.9438	0.7865	0.7865	0.7865	0.78
24	1503960366	2016-04-13 23:0...	0.9438	2.0449	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.78
25	1503960366	2016-04-14 00:0...	0.7865	2.2022	2.0449	0.9438	0.7865	0.9438	0.9438	0.7865	0.94
26	1503960366	2016-04-14 01:0...	0.7865	0.9438	0.7865	0.7865	0.9438	0.7865	0.7865	0.7865	0.78
27	1503960366	2016-04-14 02:0...	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.78
28	1503960366	2016-04-14 03:0...	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.78
29	1503960366	2016-04-14 04:0...	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.78
30	1503960366	2016-04-14 05:0...	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.78
31	1503960366	2016-04-14 06:0...	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.78
32	1503960366	2016-04-14 07:0...	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.7865	0.78

Fig 4.4: Data Table (a)



Data Table - Orange

Info
413 instances (no missing data)
5 features
No target variable.
No meta attributes

Variables
☑ Show variable labels (if present)
□ Visualize numeric values
☑ Color by instance classes

Selection
☑ Select full rows

minuteCaloriesWide_merged | sleepDay_merged

	Id	SleepDay	TotalSleepRecords	TotalMinutesAsleep	TotalTimeInBed
1	1503960366	2016-04-12 00:0...	1	327	346
2	1503960366	2016-04-13 00:0...	2	384	407
3	1503960366	2016-04-15 00:0...	1	412	442
4	1503960366	2016-04-16 00:0...	2	340	367
5	1503960366	2016-04-17 00:0...	1	700	712
6	1503960366	2016-04-19 00:0...	1	304	320
7	1503960366	2016-04-20 00:0...	1	360	377
8	1503960366	2016-04-21 00:0...	1	325	364
9	1503960366	2016-04-23 00:0...	1	361	384
10	1503960366	2016-04-24 00:0...	1	430	449
11	1503960366	2016-04-25 00:0...	1	277	323
12	1503960366	2016-04-26 00:0...	1	245	274
13	1503960366	2016-04-28 00:0...	1	366	393
14	1503960366	2016-04-29 00:0...	1	341	354
15	1503960366	2016-04-30 00:0...	1	404	425
16	1503960366	2016-05-01 00:0...	1	369	396
17	1503960366	2016-05-02 00:0...	1	277	309
18	1503960366	2016-05-03 00:0...	1	273	296
19	1503960366	2016-05-05 00:0...	1	247	264
20	1503960366	2016-05-06 00:0...	1	334	367
21	1503960366	2016-05-07 00:0...	1	331	349
22	1503960366	2016-05-08 00:0...	1	594	611
23	1503960366	2016-05-09 00:0...	1	338	342
24	1503960366	2016-05-10 00:0...	1	383	403
25	1503960366	2016-05-11 00:0...	1	285	306
26	1644430081	2016-04-29 00:0...	1	119	127
27	1644430081	2016-04-30 00:0...	1	124	142
28	1644430081	2016-05-02 00:0...	1	796	961
29	1644430081	2016-05-08 00:0...	1	137	154
30	1844505072	2016-04-15 00:0...	1	644	961

Restore Original Order
☑ Send Automatically
413 413 | 413

Fig 4.4: Data Table (b)

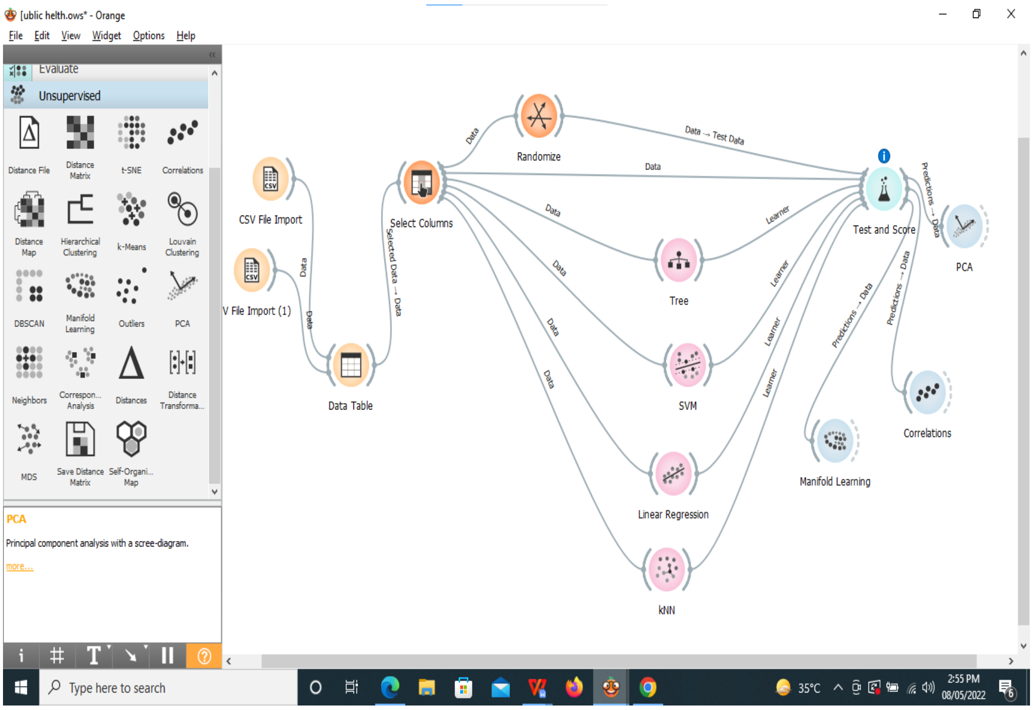


Fig 4.5: Public Model developed on Orange

The figure 4.7 is an implantation proposed model using Open source UI of Python, i.e. Orange, here we have collected all CSV files in one data table for development of basic model and further in second step select the column from the data and in third step applied KNN Neural Network and LINEAR REGRATION, Random Forest algorithms On the basis of these algorithm of ML generated the test and Score table as a result and shown in chapter 5.

4.4 Summary

The evaluated algorithms each have distinct challenges to overcome throughout their training. The complexity of training DT-based algorithms, also known as RF, is much lower. This holds true for both straightforward regression trees and more complex ensembles of trees (RF). When there are extremely little data, RF performs much better than NN and SVM, both of which become more complicated in this situation. SVMs are constructed using a variety of kernel types; hence, the parameter combinations that need to be optimized vary from kernel type to kernel type. Nevertheless, it is essential that a strong emphasis be placed on the fact that broad generalizations cannot be made about the superiority of any strategy for resolving all sorts of situations. This is due to the fact that the performance of the approaches may differ when used to various datasets; hence, broad generalizations on the superiority of any method cannot be drawn.

The RF algorithm is very sensitive to even minute shifts in the training dataset; as a result, it may be unstable at times and has a tendency to over fit the model. KNN is simple to design and comprehend, but it has the primary problem of becoming noticeably slower as the quantity of the data being used rises, and it is difficult to determine the appropriate value of K for the KNN classifier. Herein lies KNN's primary flaw. In classification, the NN technique is less often utilized because of its high computing cost. Overtraining and noise have little influence on either SVM or RF, demonstrating their ability to handle data that isn't uniform. In image classification research and applications, support vector machines (SVM) and random forests (RF) are becoming more prominent among the nonparametric approaches. As input datasets get bigger, classification times for NN and KNN are projected to climb more than those for SVM and RF. Due to the time-consuming parameter tuning procedure, the many types of neural network architectures available, and the high number of algorithms used for training NN, most researchers prefer SVM or RF as simpler methods that consistently achieve results with high accuracies and are often times faster.

Chapter 5

Result and Comparison

5.1 Personalized healthcare model

When training the data, the KNN Algorithm was used with 5 neighbors, and the Euclidean metric was employed on the Uniform Weight Neural Network that was used, which had 100 hidden layers, and the maximum number of iterations was 200.

The liner regression is applied using a lasso regularization strength of L1 (alpha = 0.0001) and an elastic net mixing ratio of 0.50:0.50

The Random Forest algorithm is used, with 10 different trees and 5 different qualities being taken into consideration at each split.

5.1.1 Result of Personalized healthcare model

Model	**MSE**	**RMSE**	**MAE**	**R2**
Random Forest	46111189.287	6790.522	3880.087	0.461
Neural Network	86140449.441	9281.188	7024.140	-0.007
Linear Regression	86139613.302	9281.143	7024.113	-0.007
KNN	64860573.371	8053.606	4753.905	0.242

Table 5.1.

5.1.2: Compare: Model by coefficient of variation of the RMSE

	Random Forest	**Neural Network**	**Linear Regression**	**K Nearest Neighbor**
Random Forests		**0.000**	**0.000**	**0.000**
Neural Network	**1.000**		**0.832**	**1.000**
Linear Regression	**1.000**	**0.168**		**1.000**
K Nearest Neighbor	**1.000**	**0.000**	**0.000**	

Table 5.2

5.2 Test and Score

The mean squared error (MSE) is a measurement that takes the average of the squares of the mistakes or variances (the difference between the estimator and what is estimated).

The root square of the arithmetic mean of the squares of a group of integers is what is known as the root mean square error (RMSE) (a measure of imperfection of the fit of the estimator to the data)

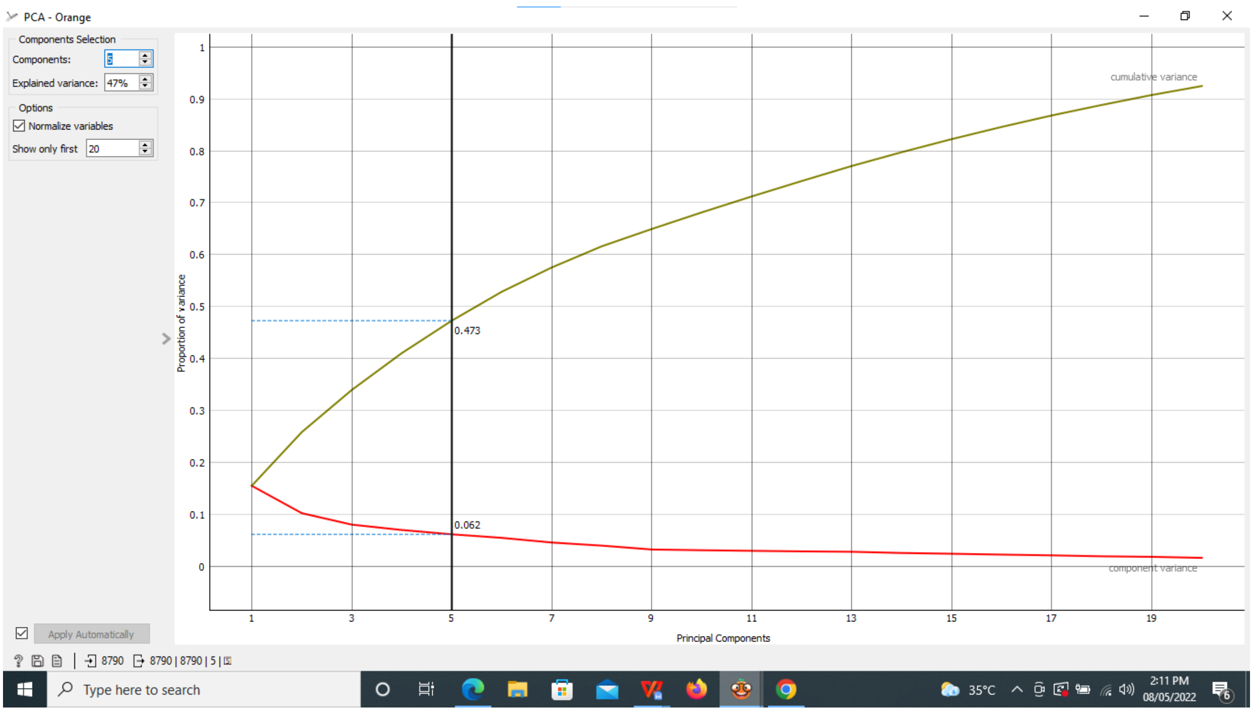
The MAE is a metric that determines how accurate forecasts and predictions are in relation to actual results.

The percentage of the total variation in the dependent variable that can be predicted based on the independent variable is what is meant to be represented by the R2 statistic.

The CVRMSE is the RMSE that has been normalized by the actual values' mean value.

Train time is the total amount of time, measured in seconds, spent instructing models.

Test time refers to the total amount of time, measured in seconds, spent validating models.



PCA - Orange

Components Selection

Components: 5

Explained variance: 47%

Options

☑ Normalize variables

Show only first 20

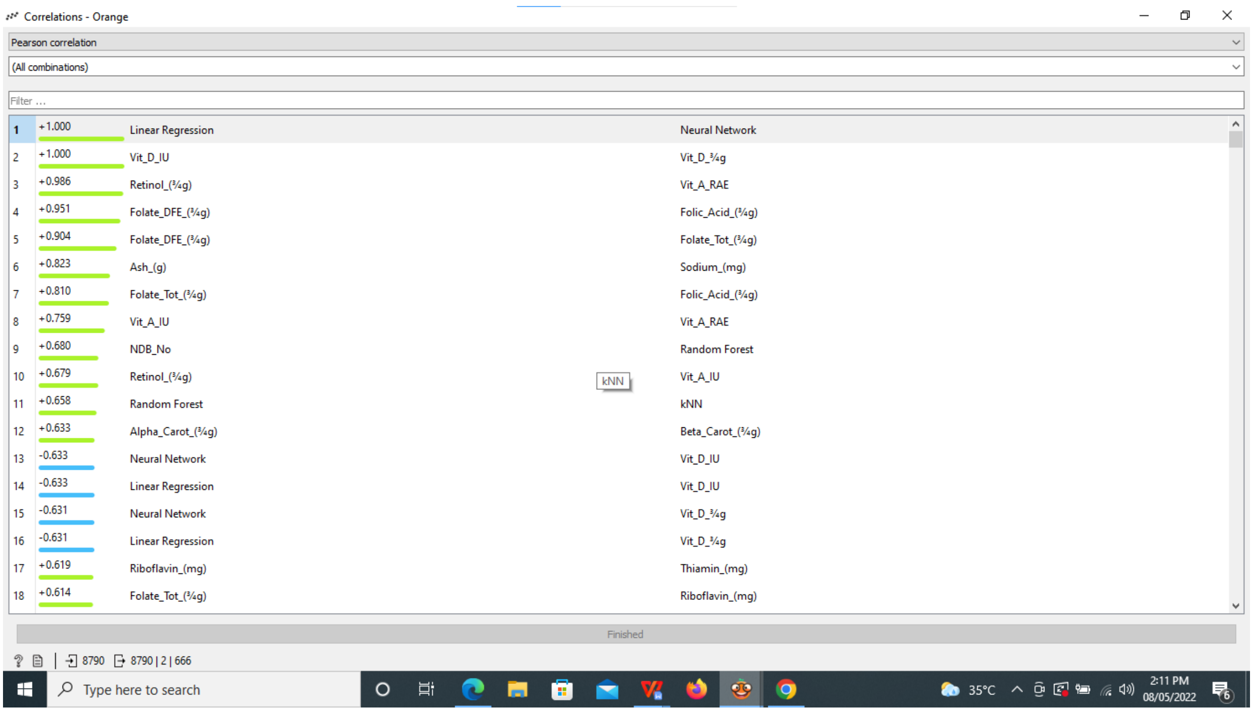
☑ Apply Automatically

8790 | 8790 | 8790 | 5

Fig 5.3: PCA (a)

5.3 PCA

PCA transforms the data into a dataset called principle components, which consists of variables that are not connected with one another. PCA widget displays a graph (scree diagram) showing a degree of explained variance by best principal components and allows the user to interactively set the number of components to be included in the output dataset. In addition, PCA widget allows the user to select which variables should be included in the output dataset.



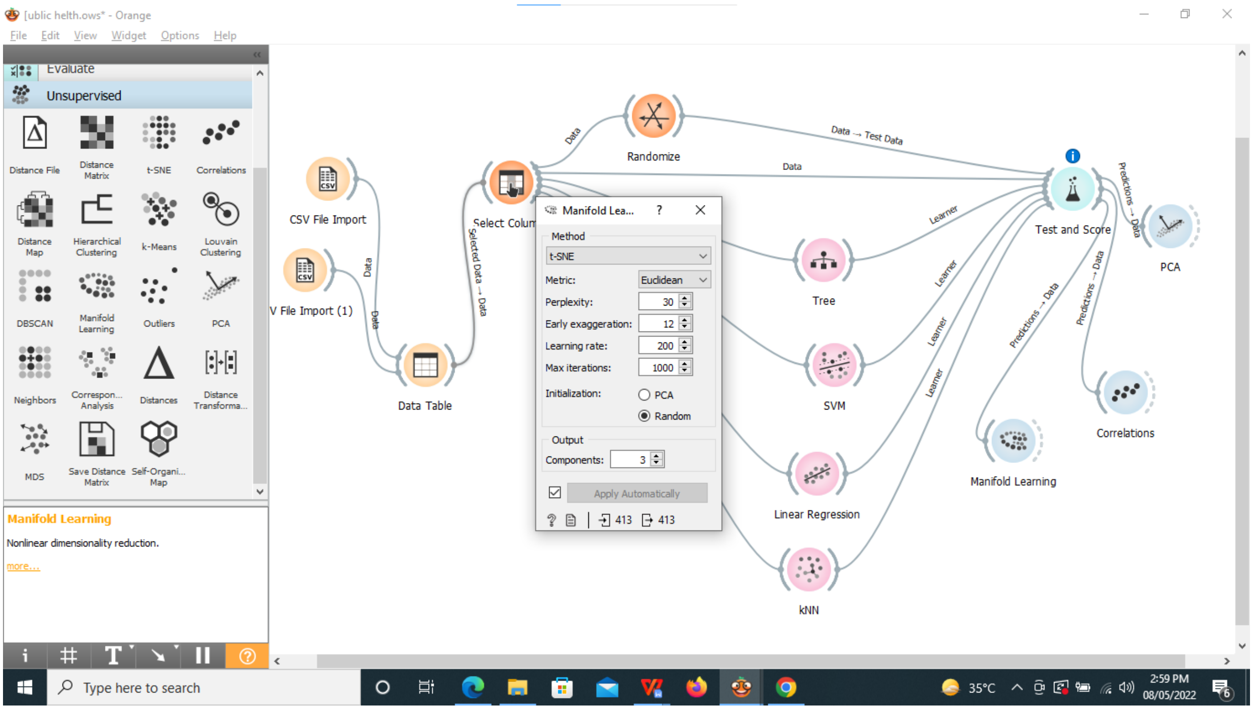
1	+1.000	Linear Regression	Neural Network
2	+1.000	Vit_D_IU	Vit_D_¾g
3	+0.986	Retinol_(¾g)	Vit_A_RAE
4	+0.951	Folate_DFE_(¾g)	Folic_Acid_(¾g)
5	+0.904	Folate_DFE_(¾g)	Folate_Tot_(¾g)
6	+0.823	Ash_(g)	Sodium_(mg)
7	+0.810	Folate_Tot_(¾g)	Folic_Acid_(¾g)
8	+0.759	Vit_A_IU	Vit_A_RAE
9	+0.680	NDB_No	Random Forest
10	+0.679	Retinol_(¾g)	Vit_A_IU
11	+0.658	Random Forest	kNN
12	+0.633	Alpha_Carot_(¾g)	Beta_Carot_(¾g)
13	-0.633	Neural Network	Vit_D_IU
14	-0.633	Linear Regression	Vit_D_IU
15	-0.631	Neural Network	Vit_D_¾g
16	-0.631	Linear Regression	Vit_D_¾g
17	+0.619	Riboflavin_(mg)	Thiamin_(mg)
18	+0.614	Folate_Tot_(¾g)	Riboflavin_(mg)

Fig 5.4: Correlation (a)

Calculates Pearson or Spearman correlation scores for each pair of characteristics in a dataset that is being analyzed by the Correlations tool. These approaches can only identify relationships with a monotonic pattern.

Proceed to the pair that has the strongest inverse correlation, which is DIS-NOX. Now link the Scatter Plot to the Correlations function, and specify two outputs: Data to Data and Features to Features [60]. Take note of the manner in which the feature pair is included straight into the scatter plot. It would seem that the two characteristics do not have a positive correlation with one another.

In this model we have observed following thing



Manifold Lea... ?

Method

t-SNE

Metric: Euclidean

Perplexity: 30

Early exaggeration: 12

Learning rate: 200

Max iterations: 1000

Initialization: ○ PCA ● Random

Output

Components: 3

☑ Apply Automatically

413 413

Fig 5.5: Public Data model Result

KNN algorithm is applied with 5 neighbors and Euclidean metric is utilized on Uniform Weight

Neural Network is applied with 100 hidden layer and maximum iteration were 200 while training the data.

Liner regression is applied with Lasso regularization (L1) strength (alpha=0.0001) and elastic net mixing 0.50:0.50

Random Forest is applied with 10 no of trees and 5 number of attributes considered at each split.

5.2.1 Public healthcare Model result

Test and Score for Public healthcare Model Result

Model	MSE	RMSE	MAE	R2
KNN	547337916800216064	2335514461.763	1888835517730	-0.292
Tree	6219026910608169984.000	2439797688.388	71928187338879	-0.649
SVM	5301557836533572608.000	2302511202.60	1843191422206	-0.252
Linear Regression	4118599453201383424	2029433283.52	1616870091602	0.027

Table 5.3

The MSE is calculated by taking the average of the squares of any mistakes or variances (the difference between the estimator and what is estimated).

RMSE, or root mean square error, is calculated by taking the square root of the arithmetic mean of the squares of a group of values (a measure of imperfection of the fit of the estimator to the data)

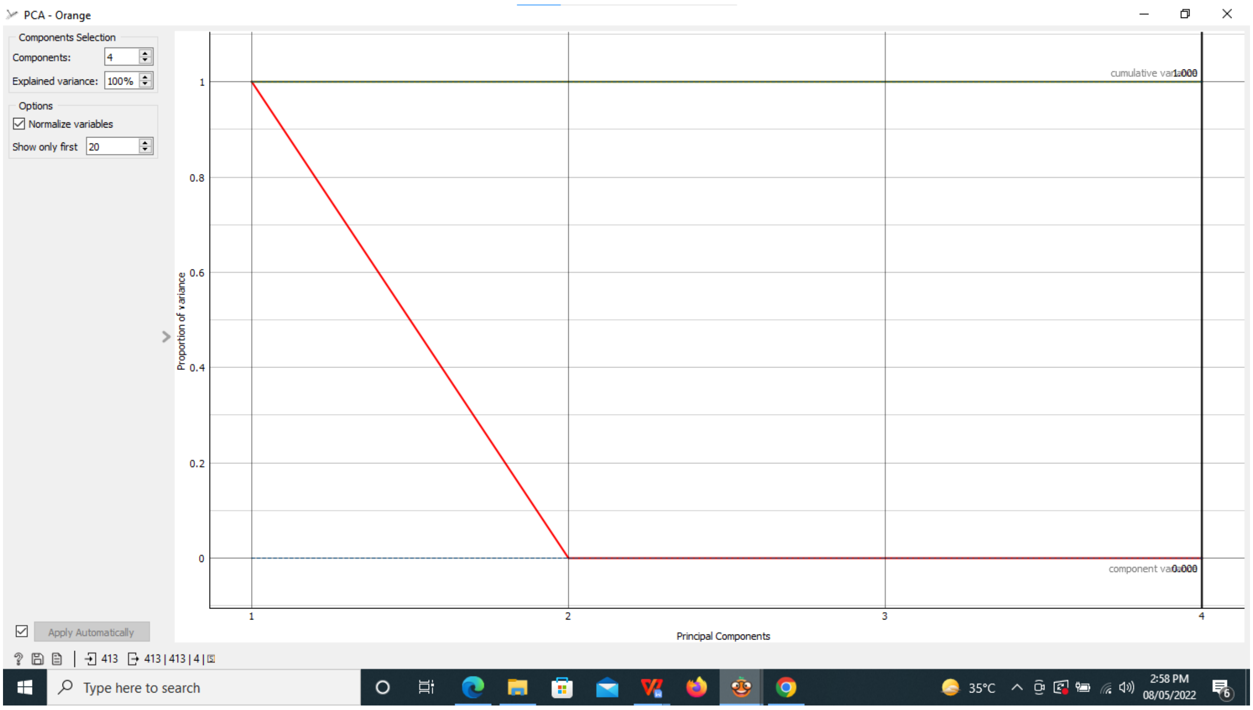
The MAE is a measurement that determines how well forecasts or projections are able to anticipate future events [62].

The coefficient of determination, or R2, may be understood as the fraction of the variation in the dependent variable that can be predicted from the independent variable.

The CVRMSE is the RMSE with the actual values' mean value used to normalize it.

Train time is the total amount of time in seconds that is utilized for training models.

The total amount of time, measured in seconds, spent testing various models.



PCA - Orange

Components Selection

Components: 4

Explained variance: 100%

Options

☑ Normalize variables

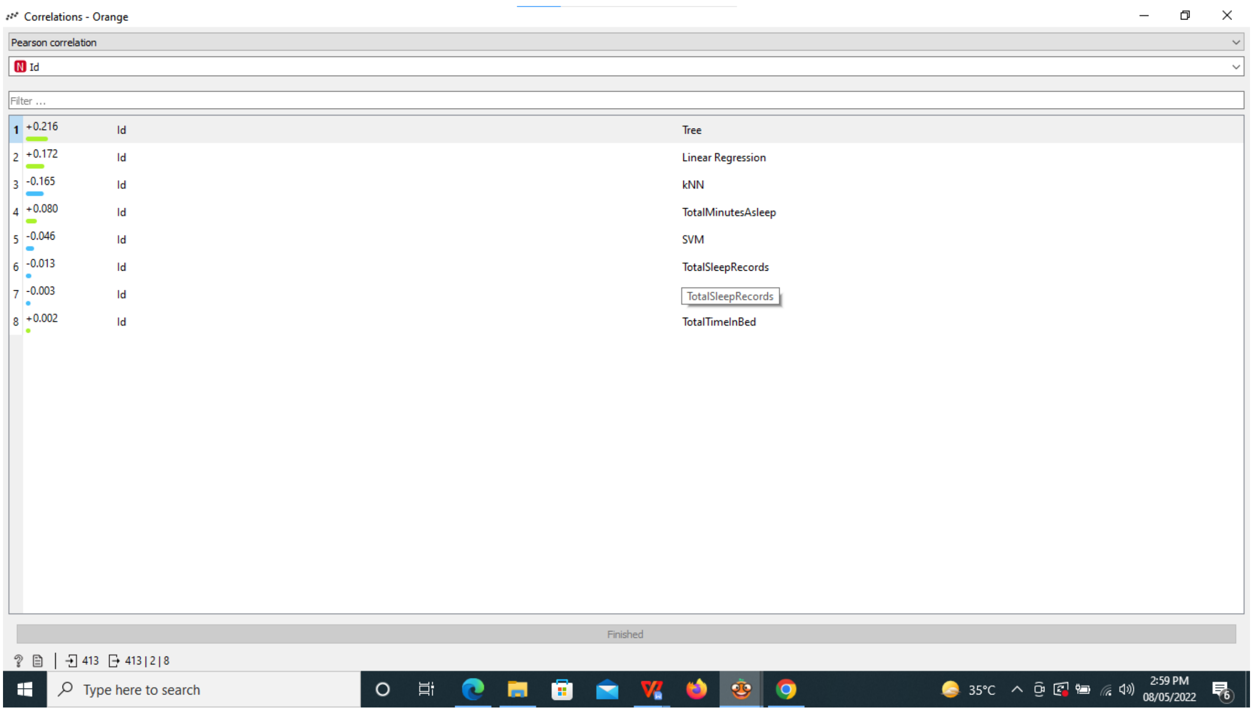
Show only first 20

☑ Apply Automatically

413 | 413 | 413 | 4

Fig 5.4: PCA (b)

PCA transforms the data into a dataset called principle components, which consists of variables that are not connected with one another. PCA widget displays a graph (scree diagram) showing a degree of explained variance by best principal components and allows the user to interactively set the number of components to be included in the output dataset. In addition, PCA widget allows the user to select which variables should be included in the output dataset.


Correlations - Orange

Pearson correlation

N Id

Filter ...

1	+0.216	Id	Tree
2	+0.172	Id	Linear Regression
3	-0.165	Id	kNN
4	+0.080	Id	TotalMinutesAsleep
5	-0.046	Id	SVM
6	-0.013	Id	TotalSleepRecords
7	-0.003	Id	TotalSleepRecords
8	+0.002	Id	TotalTimeInBed

Finished

413 413 | 2 | 8

Correlation (b)

Correlations computes Pearson or Spearman correlation scores for all pairs of features in a dataset. These methods can only detect monotonic relationship.

Go to the most negatively correlated pair, DIS-NOX. Now connect Scatter Plot to **Correlations** and set two outputs, Data to Data and Features to Features. Observe how the feature pair is immediately set in the scatter plot. Looks like the two features are indeed negatively correlated.

5.3 Observation and findings of Both Public and Private healthcare model:

We have observed following things

This model helps to improve –

5.4 Accessibility and Responsiveness

Six articles documented that a significant proportion of outpatient services in low- and middle-income countries appeared to be provided by the private sector [15]–[18]. However, the percentage of total visits varied substantially across countries and income levels [15].

5.5 Quality of Health Care

Nine retrospective chart reviews and survey-based studies found that diagnostic accuracy and adherence to medical management standards were worse among private than public sector care providers [37]–[45].

Patient Outcomes

Public sector provision was associated with higher rates of treatment success for tuberculosis and HIV [61]–[64] as well as vaccination [65]-[67]

5.6 Accountability, Transparency and Regulation

Data on this theme tended to be unavailable from the private sector. No papers were found to describe any systematic collection of outcome data from entirely private sector sources. One recent independent review of Ghana's private sector referred to the private sector as a “black box,” with a dearth of information on delivery practices and outcomes [22]

Comparison of table For Personalized and Public Model

	Public Healthcare Model				Private Healthcare Model			
Model	MSE	RMSE	MAE	R2	MSE	RMSE	MAE	R2
KNN	54733279168.00	2339514461.00	18888835517.00	(0.29)	86140449.44	9281.19	7024.14	(0.01)
Tree	62190269106.00	2493797688.00	1928187338.00	(0.47)	86133613.30	9281.14	7024.11	(0.01)
SVM	5301557836533.00	2302511202.00	1843191422.00	(0.25)	64860573.37	8053.61	4753.91	0.42
Linear rig ration	411859945320.00	2029433293.00	1616870091.00	0.03	461111189.00	6790.52	3880.09	0.46

Table 5.4

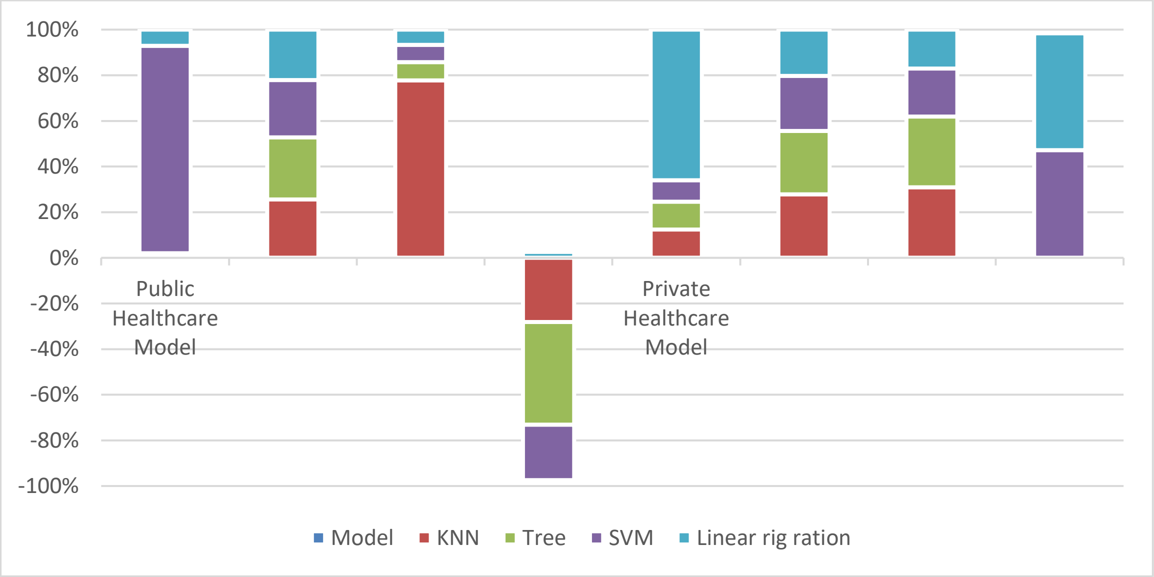


Fig 5.5: Comparison Result

Chapter 6

Conclusion and Future work

6.1 Conclusion

Big data analytics make use of the dichotomy that exists between structured and unstructured data sources. Conquering the challenge of transitioning to a data environment that is integrated is common knowledge. It is interesting to note that the fundamental concept of big data primarily depends on the concept that the more information there is, the more insights one may obtain from this information, and the better one can forecast what will happen in the future. The big data healthcare industry is forecast to expand at a pace that is comparable to exponential growth by a variety of reputable consulting organizations and health care corporations. On the other hand, in a very short period of time, we have seen a wide range of analytics that are now being used and which have shown substantial influence on the decisions made in the healthcare business as well as its overall performance. Because of the exponential rise of medical data coming from a variety of fields, computational specialists have been obliged to devise novel ways in order to evaluate and understand the vast amounts of data that are being collected within a certain length of time. There has been some progress made in integrating computational systems for signal processing from the perspective of research as well as that of practicing medical practitioners. Consequently, the next major objective may be to construct an accurate model of the human body by fusing together physiological data with various "-omics" methodologies. This one-of-a-kind concept has the potential to expand our understanding of medical conditions and maybe contribute to the creation of innovative diagnostic tools. The steadily increasing amount of genetic data that is now accessible, including the inherent hidden flaws that are caused by experiments and analysis procedures, requires more attention. However, there are chances to implement systemic changes within the healthcare research at each and every stage of this vast process.

Data scientists are faced with a difficulty when it comes to the proper integration and execution of the large number of medical data that has been acquired from a variety of platforms. It has been argued that a revolution in healthcare is necessary in order to bring together bioinformatics, health informatics, and analytics in order to encourage more effective and individualized therapies. In addition, new methods and technologies need be created in order to grasp the type of the data (structured, semi-structured, or unstructured), the complexity of the data (dimensions and characteristics), and the volume of the data in order to extract information that is relevant. Big data's most valuable feature is the fact that it can be used in almost any way imaginable. Big data has only been around for a few years, but in that time it has already ushered in a number of significant innovations throughout the health care industry. These innovations range from medical data management to drug discovery programmers for complex human diseases such as cancer and neurodegenerative disorders. It's easy to observe how the EHR system, which collects, administers, and makes use of patient data, has evolved significantly since the late 2000s in the healthcare business to highlight this argument. Many people question whether or not big data will replace skilled workers, subject matter experts, and intellectuals. We, on the other hand, believe that big data will complement and strengthen the current pipeline of healthcare innovations and developments. One can plainly observe how the health care industry is shifting from one with a bigger volume base to one with a greater emphasis on individualized treatment or a more focused area. Therefore, it is very necessary for professionals and technicians to have an understanding of this developing scenario. It is possible to make a forecast that big data analytics will get closer to developing a predictive system in the following year. This would imply making projections about an individual's health in the future based on the data that is now available or already collected (such as EHR-based and Omits-based). In a similar vein, one might hypothesize that the development of population health information may result from the accumulation of structured information collected from a particular location. Big data, when viewed in its entirety, will make it simpler to provide medical care by facilitating the prediction of epidemics (in relation to the health of populations), delivering early warnings of disease conditions, and assisting in the discovery of novel biomarkers and intelligent therapeutic intervention strategies for an improved quality of life. This will all lead to a reduction in the burden placed on medical professionals.

References

1. **https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html**
2. Health data". McGraw-Hill Concise Dictionary of Modern Medicine. McGraw-Hill. 2002.
3. Agarwal, R &Dhār, V, 2014,‗Big Data, data science and analytics: The opportunity and challenge for IS research, vol. 25, no. 3, pp. 443-448.
4. Agarwal, R, Khandelwal, A & Stoica, I, 2015,‗Succinct: Enabling Queries on Compressed Data', In NSDI, vol. 15, pp. 337-350.
5. Al Hamid, HA, Rahman, SMM, Hossain, MS, Almogren, A &Alamri, A, 2017,‗A security model for preserving the privacy of medical Big Data in a healthcare cloud using a fog computing facility with pairing-based cryptography', IEEE Access, vol. 5, pp. 22313-22328.
6. Ardagna, CA, Damiani, E, Frati, F &Rebeccani, D, 2016, ‗A configuration-independent score-based benchmark for distributed databases', IEEE Transactions on Services Computing, vol. 9, no. 1, pp. 123-137.
7. Arora, R &Aggarwal, RR, 2013, ‗Modeling and querying data in mongo dB', International Journal of Scientific and Engineering Research, vol. 4, no. 7, pp. 141-144.
8. Aruna, S, Nandakishore, LV &Rajagopalan, SP, 2012, ‗Cloud based decision support system for diagnosis of breast cancer using digital mammograms', International Journal of Computer Applications (IJCA), vol. 1, pp. 1-3.
9. Barberton, E, Gribaudo, M &Iacono, M, 2014, ‗Performance evaluation of NoSQL big-data applications using multi-formalism models', Future Generation Computer Systems, vol. 37, pp. 345-353.

10. Olshannikova, E.; Ometov, A.; Koucheryavy, Y.; Olsson, T. Visualizing Big Data with augmented and virtual reality: Challenges and research agenda. J. Big Data 2015, 2, 22.

11. Bani-Salameh, H.; Jeffery, C.; Hammad, M. Developers' social networks-tools analysis based on the 3Cs model. Int. J. Netw.

    Virtual Organ. 2013, 13, 159–175

12. Luna, D.R.; Mayan, J.C.; Garcia, M.J.; Almerares, A.A.; Househ, M. Challenges and potential solutions for big data implementations in developing countries. Yearb. Med. Inform. 2014, 23, 36–41.

13. Wang, Y.; Kung, L.; Byrd, T.A. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. Technol. Forecast. Soc. Chang. 2018, 126, 3–13.

14. Wang, L.; Wang, G.; Alexander, C.A. Big data and visualization: Methods, challenges and technology progress. Digit. Technol. 2015, 1, 33–38.

15. VistA. Available online: https://worldvista.org/AboutVistA (accessed on 5 February 2021).

16. Bidgood, W.D., Jr.; Horii, S.C.; Prior, F.W.; Van Syckle, D.E. Understanding and using DICOM, the data interchange standard for biomedical imaging. J. Am. Med. Inform. Assoc. JAMIA 1997, 4, 199–212

17. Dr.V.P.GladisPushparathi, S.DivyaBarathi, S.Kavitha, S.ShaktiNivetha, ―Clincial Decision Support System using Hadoop‖, International Journal of Innovative Research in Science, Engineering and Technology, Volume 7, Special Issue 2, March 2018.

18. MukeshBorana, Manish Giri, Sarangkamble, KiranDeshpande , ShubhangiEdake ―Healthcare Data Analysis using Hadoop‖, International Research Journal of Engineering and Technology , Volume 02, Issue 07,October 2015.

19. MS.Minu, IshanMeena, Pratyush, R. Aravind, VijaydityaSarker, ―Healthcare Analysis Using Hadoop Framework‖, International Journal for Science and Advance Research In Technology, Volume 4, Issue 10, October 2018.
20. DeepthiYaramala ―Healthcare Data Analytics using Hadoop‖ , Thesis, San Diego University.
21. B. Durga Sri, K.Nirosha, M. Padmaja, ―Healthcare Analysis Using Hadoop‖, International Journal Of Current Engineering And Scientific Research (IJCESR), Volume-4, Issue-6, 2017
22. Rahul Beakta, ―Big Data and Hadoop: A Review Paper‖, Research Gate, Volume 2, Special Issue 2, 2015.
23. Iqbaldeep Kaur, NavneetKaur, AmandeepUmmat, JaspreetKaur, NavjotKaur, ―Research Paper on Big Data and Hadoop‖, Internal Journal of Computer Science and Technology, Volume 7, Issue 4, October-Dec 2016.
24. Ivanilton Polato, ReginaldoRé, Alfredo Goldman , Fabio Kon A comprehensive view of Hadoop research A systematic literature review‖, Journal of Network and Computer Applications, Elsevier, 2014.
25. Harshawardhan S. Bhosale ,Prof.Devendra P. Gadekar,‖A Review Paper on Big Data and Hadoop‖, International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014
26. Ambigavathi, Big Data Analytics in Healthcare, 2018 Tenth Int. Conf. Adv. Comput., pp. 269276, 2018.
27. Belle, R. Thiagarajan, S. M. R. Soroushmehr, F. Navidi, D. A. Beard, and K. Najarian, Big Data Analytics in Healthcare, Biomed Res. Int., vol. 2015, pp. 116, Nov. 2015, doe: 10.1155/2015/370194.
28. R. Sonnati, Improving Healthcare Using Big Data Analytics, Improv. Health. Using Big Data Anal., vol. 6, no. 3, pp. 142146, 2015.

29. M. Singh, N. Delhi, V. Bhatia, R. Bhatia, and D. Specialist, Big data analytics, pp. 239241, 2017.
30. Kavitha. G, IV B.Tech IT, Dr. D.Prabha, Clinical Data Analyticsin Big Data Using Hadoop, International Journal of Scientific &Engineering Research, Volume 6, Issue 5, May-2015, 24-27.
31. Alison Bolen, SAS Insights Editor, How do you know if you're ready for Hadoop?, available online at:sas.com/en_in/insights/articles/big-data/ready-for-hadoop.htm
32. Constantiou, I.D. and Kallinikos, J., 2015. New games, new rules: big data and the changing context of strategy. Journal of Information Technology, Volume 30, pp. 44-57.
33. Elgendy, N. and Elragal, A., 2014. Big data analytics: a literature review paper. sol., Springer, champ, pp. 214-227
34. Agarwal, R. and Dhār, V., 2014. Big data, data science, and analytics: The opportunity and challenge for IS research. IS research Journal.
35. Akter, S., Wamba, S.F., Gunasekaran, A., Dubey, R. and Childe, S.J., 2016. How to improve firm performance using big data analytics capability and business strategy alignment? International Journal of Production Economics, pp. 113-131.
36. Russom, P., 2011. Big data analytics, s.l.: TDWI best practices report, fourth quarter.
37. S.VikramPhaneendra&E.Madhusudhan Reddy "Big Data- solutions for RDBMS problems- A survey" In 12th IEEE/IFIP
38. Network Operations & Management Symposium (NOMS 2010) (Osaka, Japan, Apr 19{23 2013).
39. Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, W., (18-22 Dec.,2012) , "Shared disk big data analytics with Apache Hadoop".
40. Aditya B. Patel, Manashvi Birla, Ushma Nair,(6-8 Dec. 2012),"Addressing Big Data Problem Using Hadoop and Map Reduce"
41. Garlasu, D.; Sandulescu, V; Halcu, I.; Neculoiu, G.; (17-19 Jan. 2013),"A Big Data implementation based on Grid Computing", Grid Computing.
42. Efficient Capabilities of Processing of Big Data using Hadoop Map Reduce, Praveen Kumar1, Dr Vijay Singh Rathore, International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 6, June 2014.
43. Vinod Ganeshan,Head-BFSI Vertical, Hitachi Data Systems, " How Big Data is going to Impact Businesses in the Near Future" PC Quest, July 2015, Pg.No.40-41

44. "Big Data: The next frontier for innovation, competition, and productivity",McKinsey Global Institute, May 2011, p. 11: http://www.mckinsey.com/Insights/MGI/Research/Technology_andInnovation/Big_data_The_next_frontier_for_innovation.
45. Karthik Sridhar, Founder & CEO data Culture," Making Sense of Big data for SME‟s" PC Quest, July 2015, pg.no. 51
46. Anusha B. Dhakite and Prof. Sameer Y. Thakur, " A Survey on Hadoop Technology and Its Role in Information Technology" 2015, Volume 5, Issue 4, www.ijarcsse.com
47. Roshani K. Chaudhari and Prof. D. M. Dakhane, "Contribution of Hadoop to Big Data Problems" April 2015, volume 5, Issue 4, www.ijarcsse.com
48. Groot, S &Kitsuregawa, M 2010, 'Jumbo: beyond Map Reduce for workload balancing', in: VLDB, PhD workshop.
49. S. V. Nuti, B. Wayda, I. Ranasinghe, S. Wang, R. P. Dreyer, S. I. Chen, et al., The Use of Google Trends in Health Care Research: A Systematic Review, PLoS One. 9 (2014) e109583. Doe: 10.1371/journal. pone.0109583.
50. N. V. Chawla, D. A. Davis, Bringing big data to personalized healthcare: a patient-centered framework, J. General Internal Med. 28 (2013) 660-665. Doi: 10.1007/s11606-013-2455-8.
51. A. Gandomi, M. Hider, beyond the hype: big data concepts, methods, and analytics, Int. J. Inf. Manage. 35 (2015) 137–144.
52. M. Viceconti, P. Hunter, R. Hose, Big Data, Big Knowledge: Big Data for Personalized Healthcare, IEEE J. Biomed. Heal. Informatics. 19 (2015) 1209–1215.
53. Praveen Kumar, et al., "Efficient Capabilities of Processing of Big Data using Hadoop Map Reduce", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, No. 6, 2014.
54. Emad A. Mohammed, et al. "Applications of the Map Reduce programming Frame work to clinical Big Data analysis: current landscape and future trends", Big Data Mining, 2014.
55. Arantxa Duque Barrachina and AislingO Driscoll, "A Big Data methodology for categorizing technical support requests using Hadoop and Mahout", Journal of Big Data, 2014.

56. K. Divya, N. Sadhasivam, “Secure Data Sharing in Cloud Environment Using Multi Authority Attribute Based Encryption”, International Journal of Innovative Research in Computer and Communication Engineering,Vol. 2, No. 1, 2014.

57. Sabia and Sheetal Kalra, “Applications of Big Data: Current Status and Future Scope”, International Journal of Computer Applications, Vol. 3, No. 5, pp. 2319-2526, 2014.

58. Hongsong Chen and Zhongchuan Fu, “Hadoop -Based Healthcare Information System Design and Wireless Security Communication Implementation”, Hindawi Publishing Corporation Mobile Information Systems, 2015

59. J. Rama Prabha’ “Security in Cloud Health Care”, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-4, November 2019.

60. https://www.researchgate.net/publication/309755898_A_Study_on_MapReduce_Challenges_and_Trends/fulltext/5a22fb6ba6fdcc8e8666e2dd/A-Study-on-MapReduce-Challenges-and-Trends.pdf?origin=publication_detail

61. R. Zhang and L. Liu, “Security Models and Requirements for Healthcare Application Clouds”, IEEE

62. 3rd International Conference on Cloud Computing, 2010.

63. G. Federico, R. Meili, and R. Scoville, “Extrapolating Evidence of Health Information Technology Savings and Costs”, Santa Monica, Calif.: RAND Corporation, MG-410-HLTH, 2005.

64. American Standards for Testing and Materials (ASTM), “Standard Guide for Properties of a Universal Healthcare Identifier (UHID)”, West Conshohocken, Pa.: ASTM, E1714-00, October 10, 2000.

65. Maslin, M., &Ailar, R., (2015). Cloud Computing Adoption in the Healthcare Sector: A SWOT Analysis, Asian Social Science, 11(10), pp. 12-17.

66. Saeid, A., Zohreh, S., Ali, T., Steven, R., Abdullah, G., Samee, U. K., (2015). Cloud Adoption in Malaysia: Trends, Opportunities, and Challenges, IEEE Cloud Computing, IEEE Computer Society.

67. Tamil Ilakkiya N. S. (2015). Role of Cloud in Improving Patient Care, International Journal of Advanced Research in Computer Science and Software Engineering, 5(3), pp. 171-175.

68. Rolim, C. O., Koch, F. L., Westphall, C. B., Werner, J., Fracalossi, A., & Salvador, G. S. (2010). A cloud computing solution for patient's data collection in healthcare institutions. In e-Health, Telemedicine, and Social Medicine, ETELEMED'10, Second International Conference on IEEE, pp. 95-99.

69. Then, T. S., Liu, C. H., Chen, T. L., Chen, C. S., Bau, J. G., & Lin, T. C. (2012). Secure Dynamic access control scheme of PHR in cloud computing. Journal of medical systems, 36(6), pp. 4005-4020.

APPENDIX

PLAGIARISM CHECK REPORT

plag dissertation

by National Printers

Submission date: 21-Jul-2022 03:25AM (UTC-0400)

Submission ID: 1857235058

File name: DISSERTATION_SAKSHI.docx (2.31M)

Word count: 19891

Character count: 108793

plag dissertation

ORIGINALITY REPORT

14%	12%	7%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.ijert.org Internet Source	1%
2	iul.ac.in Internet Source	1%
3	buildmedia.readthedocs.org Internet Source	1%
4	www.ijaprr.com Internet Source	1%
5	journalofbigdata.springeropen.com Internet Source	1%
6	www.pubmedcentral.nih.gov Internet Source	1%
7	Submitted to Delhi Technological University Student Paper	1%
8	www.diva-portal.org Internet Source	1%
9	docplayer.net Internet Source	<1%

AN EFFECTIVE DATA MANAGEMENT FRAMEWORK FOR

HEALTHCARE USING HDFS: A REVIEW

Sakshi Raj Singh,

P G Student Department of CSE Integral University,
Lucknow, Uttar Pradesh, India
sarajsing@student.iul.ac.in

Mohammad Zunnun Khan

Associate Professor Department of CSE Integral University,
Lucknow, Uttar Pradesh, India
zunnunkhan@gmail.com

Abstract

This paper provides a broad overview of the relationship between big data and healthcare. In the healthcare industry, big data architecture and techniques can be used to manage data growth. The first step toward a better understanding of how big data affects healthcare is to conduct an empirical study. A growing number of healthcare professionals are utilizing big data. Machine learning and big data are difficult to predict in the healthcare industry. When it comes to disease diagnosis, machine learning and big data analytic have ignored privacy and security.

***Keywords*: Big Data Analytic, Healthcare, Electronic Health Record, Electronic Medical Record**

1. The Main Text

The healthcare industry has decided to abandon a conservative approach to diagnosis and treatment after decades of doing so. Big Data-based solutions are becoming increasingly popular as a result of the rise of chronic diseases, globalization, technological advancements, and a push for evidence-based medicine.

Patient-centered care can be delivered using Big Data solutions that provide a complete picture of each patient. Patients will benefit from improved medical care as a result. In an EBM treatment decision-making process, no doctors are involved; instead, scientific evidence is used to produce a quantifiable result.

The concept of big data has been useful in a variety of industries, and its technologies have been put to good use in a variety of fields. [1] The use of big data in health care has a lot of potential. Because of the massive amount of data that must be sorted through in order to make sense of it, data management is critical in healthcare.

The use of mobile and wearable sensors has aided the proliferation of data sources in healthcare. Traditional data analysis methods are becoming increasingly ineffective as the volume and variety of medical data grows. Prescriptive analytics, as well as descriptive, diagnostic, and predictive analytics, are used in the healthcare industry.

Descriptive Analytics. Data can be analyzed using descriptive statistics. Reporting current events requires describing and critiquing the current state. There are a variety of methods that can be used at this level of analysis. Typical descriptive analytics tools include histograms and charts.

Diagnostic Analysis. It's time to take stock. One of the most important aims of this article is to provide an explanation for how and why certain events took place. Clustering and decision trees are frequently used in the diagnosis of patients, as are other diagnostic analysis techniques.

Predictive Analytics. Having the ability to predict what will happen in the future. Prediction, trend identification, and the calculation of uncertain outcome probabilities are just a few of the skills demonstrated. The likelihood of a patient having a complication can be predicted using data. Predictive models are frequently built using machine learning.

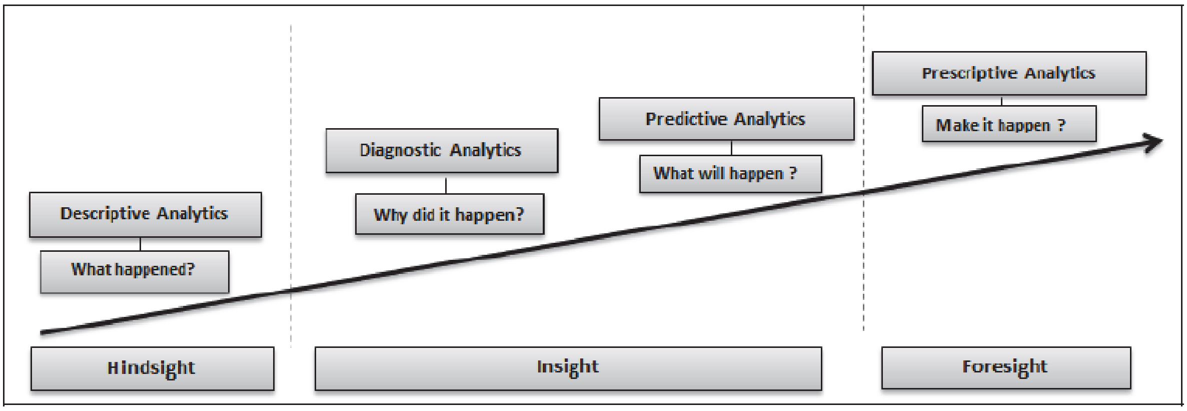


Fig 1Analytics for healthcare domain

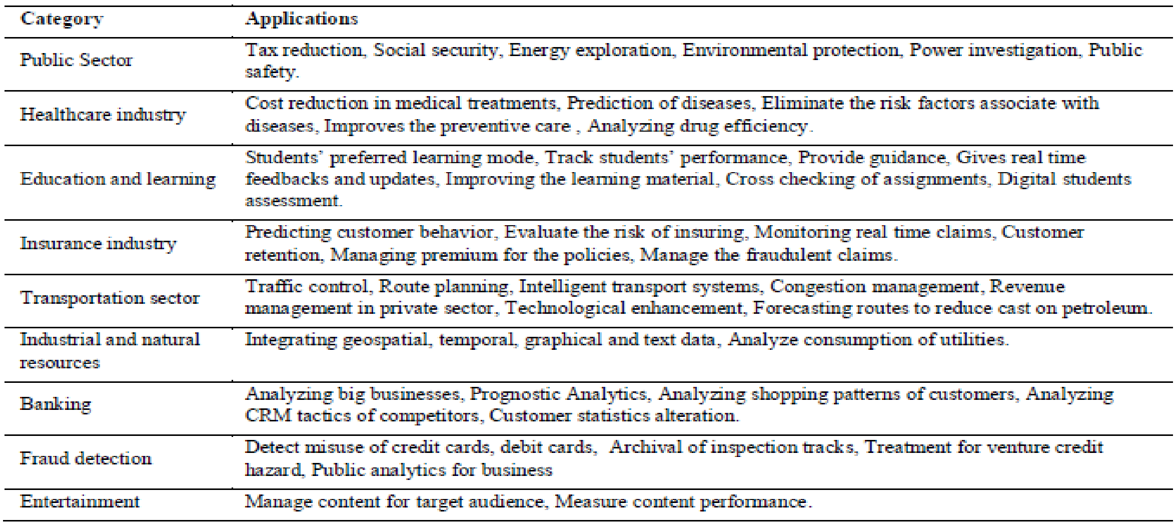
2. Data Management -Based Healthcare Systems

It's impossible to deny that researchers are in awe of the potential of big data to improve healthcare analytics. A wide range of diseases can now be diagnosed and treated using health informatics because of recent advances in big data health informatics [13]. Security and privacy concerns have been raised as a result of this investigation. Raghupathi et al. [14] discussed the challenges and architecture of big data health analytics. Big data healthcare systems are vulnerable, according to another study, if safety and privacy issues aren't taken care of right away. In addition, a large amount of data is generated by healthcare systems that can be used to improve patient care. Analyses healthcare data using data mining techniques such as survival analysis and patient similarity. It is possible to perform multidimensional analysis on large medical data sets using this framework.

Data mining is used to analyze a large amount of information in this framework. Smart health, a platform for healthcare analytics, was built with the help of ICT. Wiki-Health can be used to analyze sensor data. Data storage and querying are all included in this platform. All aspects of the data lifecycle must be handled by an application. A query and analysis layer is in charge of data storage and retrieval. Asserts that privacy and data protection issues have complicated the creation of such platforms. An intelligent healthcare management system.

2.1. *Data management Applications*

Big data has been a hot topic since 2010 and is still doing so. A growing number of businesses, organizations, and individuals are turning to big data for a variety of purposes. According to figure2, the most common uses of big data are shown



Category	Applications
Public Sector	Tax reduction, Social security, Energy exploration, Environmental protection, Power investigation, Public safety.
Healthcare industry	Cost reduction in medical treatments, Prediction of diseases, Eliminate the risk factors associate with diseases, Improves the preventive care , Analyzing drug efficiency.
Education and learning	Students' preferred learning mode, Track students' performance, Provide guidance, Gives real time feedbacks and updates, Improving the learning material, Cross checking of assignments, Digital students assessment.
Insurance industry	Predicting customer behavior, Evaluate the risk of insuring, Monitoring real time claims, Customer retention, Managing premium for the policies, Manage the fraudulent claims.
Transportation sector	Traffic control, Route planning, Intelligent transport systems, Congestion management, Revenue management in private sector, Technological enhancement, Forecasting routes to reduce cast on petroleum.
Industrial and natural resources	Integrating geospatial, temporal, graphical and text data, Analyze consumption of utilities.
Banking	Analyzing big businesses, Prognostic Analytics, Analyzing shopping patterns of customers, Analyzing CRM tactics of competitors, Customer statistics alteration.
Fraud detection	Detect misuse of credit cards, debit cards, Archival of inspection tracks, Treatment for venture credit hazard, Public analytics for business
Entertainment	Manage content for target audience, Measure content performance.

Fig 2 Applications of Data management in various categories

3. SUGGESTED ELECTRONIC HEALTH RECORD

This section of our paper will go over the requirements for our system and the various components in greater detail. We'll compare and contrast the most popular open-source projects in the final section.

3.1. *System description*

We still have a long way to go before we can implement our management system rules after researching and studying the various stakeholders involved in data collection and storage in a centralised server that medical organizations can access. The Ehr management system's specifications and requirements are shown in Figure 3.

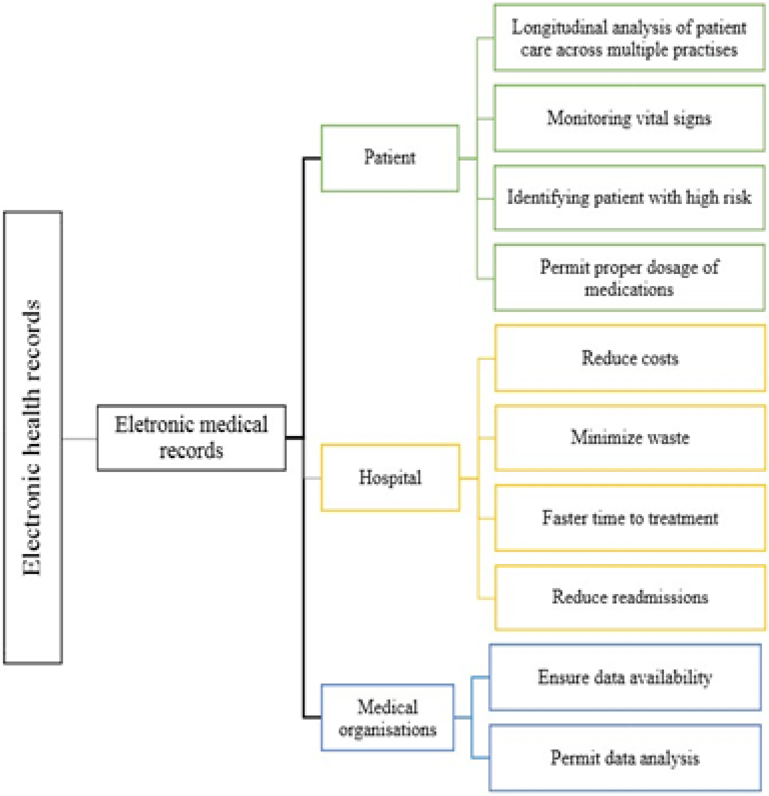

Electronic health records
Eletronic medical records
Patient
Longitudinal analysis of patient care across multiple practises
Monitoring vital signs
Identifying patient with high risk
Permit proper dosage of medications
Hospital
Reduce costs
Minimize waste
Faster time to treatment
Reduce readmissions
Medical organisations
Ensure data availability
Permit data analysis

Fig 3 EHR management system needs

This diagram depicts the features that our software will require. Furthermore, it is critical that medical institutions safeguard their own interests.

An ideal system would detect high-risk patients in real time and administer the appropriate medication dose based on the patient's previous medical history. As a result of this implementation, medical facilities will save money and resources, and they will be able to process data more quickly.

3.2.*Architecture overview*

An EHR system capable of interacting with multiple EMRs would be included in our design. The design of a system that will allow EMRs and the EHR repository to exchange health data is the focus of this phase of our project. A discussion of electronic medical records will follow an introduction to our architecture (EMRs). Because of the lack of coordination in the sector, health care in developing countries will be used as a case study. Others in the industry perform unnecessary medical tests and procedures.

4. LITERATURE REVIEW

Yu, H., & Wang, D. (2012) more and more patient data is being collected and stored by newer health care IT systems. EMR or PHR data can be used to provide medical advice to patients, while the results of data analysis can be used in scientific research. Because of this, a relational database management system is incapable of meeting the needs of large-scale health care data sets. Archived data stores can benefit from improved performance, scalability, and fault tolerance thanks to a Hadoop-based solution described in this paper. We've laid the groundwork for effective data management. [1]

Benhlima, L. (2018) big data-driven healthcare monitoring systems are examined in this paper. Big data processing methods and technologies have also been extensively discussed. Large amounts of data generated in real time by a wide range of medical sources can now be processed by a big data architecture in the healthcare industry. Based on big data analytics, this is the plan. After creating a generic big data architecture for healthcare, it created batch and stream processing to allow for simultaneous generation of accurate predictions and dashboards. [2]

Shakil, K. A. et al [2020] Medical technology has advanced to the point where an ever-increasing volume of electronic medical data poses a security risk. Medical records are stored in a variety of formats in the healthcare data management system. There is a massive amount of unstructured and inaccessible data as a result. Many hospitals in the United States have multiple locations. Patients' health data from various locations may need to be combined from time to time for research purposes. A cloud-based healthcare management system can thus facilitate the efficient management of health-care data. However, the most pressing concern with a cloud-based healthcare system is security. This category of criminal activity includes identity theft, tax fraud, bank fraud, insurance fraud, medical fraud, and the defamation of well-known patients. [3]

Rallapalli, S. [2016] the analysis of large amounts of data is a major challenge for healthcare organizations. In recent years, the variety of data generated by healthcare devices has increased dramatically. Data must be processed and effectively analyzed in order to make better decisions. With cloud computing, data storage, processing, and analysis can all be done on-demand. Traditional data processing systems can no longer keep up with the sheer volume of data. To improve performance and solve scalability issues, we need a better distributed system on the cloud. [4]

Khennou, F et al (2016) the amount of data generated on a daily basis is rapidly increasing across a wide range of industries. To properly store, process, and analyze these massive amounts of data, creative problem solving is required. Because of the variety of data we deal with, the time it takes to analyze it, and the speed at which it is collected, the concept of big data has a lot of value in the healthcare industry. In this paper, we hope to present a new health data architecture model. The framework supports an unstructured medical data storage and management framework based on the multi-agent paradigm. With the integration of the mobile agent model into the Hadoop ecosystem, we will be able to instantly connect multiple health repositories. [5]

Babar, M.et al (2019) the unbroken amplification of a flexible urban setup is hampered by Big Data processing. In a smart city, making decisions based on the vast amounts of data generated is difficult. The goal of Big Data analytics is to analyze large amounts of data. Conventional methods are no longer effective due to the massive amount of data.[6]

Bani-Salameh et al (2021) Software developers and data scientists use big data to uncover new insights and develop better solutions for improving healthcare and patient safety. Big data analytics (BDA) is gaining popularity due to its importance in healthcare decision-making. The implementation of big data analytics and management in Jordanian healthcare organizations is examined in this article. In Jordan, there are numerous challenges and limitations to managing and analyzing big data in the health sector, which are discussed here. This conceptual framework can be used to analyses health big data. According to the conceptual framework proposed in this paper, we could merge the current health information system (HIS) and HIE, which could help us gain insights from our massive datasets and reduce resource waste. By using the framework to process the collected data to develop knowledge and support decision-making, health care quality can be improved for both the community and individuals. [7]

Mathew, P. S et al (2015) the amount of data generated in the healthcare industry is growing at an exponential rate, and this trend is expected to continue in the near future. It's not uncommon for the vast majority of healthcare data (e.g., medical records and insurance claims) to be unstructured, stored in silos, and dispersed across multiple systems. It is critical to integrate and factor in these disparate data sets in order to improve healthcare outcomes. Healthcare organizations are unable to take advantage of the wealth of data they possess, either because it is stored in silos in incompatible formats or because of a lack of processing power that prevents them from quickly loading and querying large datasets. [8]

Kaur, P., Sharma et al (2018) this paper gives a basic introduction to big data and how it can be applied to healthcare. The use of big data architecture and techniques in the healthcare industry can effectively manage data growth. To better understand the role of big data in the healthcare sector, the first step is to conduct an empirical study. There has been a lot of work done with big data in the healthcare sector. It's difficult to imagine how machine learning and big data will affect the healthcare industry. [9]

De Silva et al [2015] the healthcare industry generates a large amount of medical, clinical, and omics data with varying levels of complexity and features. Clinical decision-support is gaining traction as medical institutions and regulatory bodies seek to better manage this data for more effective and efficient healthcare delivery and quality-assured outcomes. The amassing of data across all stages, from disease diagnosis to palliative care, provides additional evidence of the opportunities and challenges of effective data management, analysis, prediction, and optimization techniques as part of knowledge management in clinical environments. [10]

5. About Data Management

The term "Big Data" is frequently associated with the technology that enables its use. The size of the dataset and the complexity of operations required for its processing impose strict memory storage and computational performance requirements. According to Google Trends, the most popular search term associated with the term "Big Data" is "Hadoop." Hadoop is a free and open-source framework for processing large amounts of data using a variety of programming models across a distributed network of computers. HDFS is a file system that allows data dispersed across multiple machines to be accessed without having to deal with the complexity inherent in their dispersed nature. Map Reduce is a programming model for efficiently implementing distributed and parallel algorithms, and HDFS is its file system. While Map Reduce (Dean & Ghemawat) was created as an Apache open source project, both HDFS (Shvachko et al. 2010) and HDFS (Dean & Ghemawat 2008) were proposed by Google (Ghemawat et al. 2003). (Ghemawat et al. 2003). This demonstrates Google's importance in the development of current Big Data thinking. Hadoop has a number of modules and libraries that can be used in conjunction with HDFS and Map Reduce to meet a variety of coordination and analysis needs as well as workflow design requirements for Big Data applications.

6. OPPORTUNITIES FOR DATA MANAGEMENT SOLUTIONS IN HEALTH CARE

The data management solutions can be used in the health care to get innovative outcomes in the following areas:

- Clinical decision support - BDA technologies can be used to predict outcomes [6] or recommend alternative treatments to clinicians and patients at the point of care.
- Personalized care - Predictive data mining and analytic solutions can be used to diagnose disease symptoms in patients before they appear. Patients who are elderly or disabled can wear sensors on their clothing to monitor drug efficacy in real time.
- Public and population health - BDA solutions that mine web-based and social media data can predict flu outbreaks and population health trends.
- Fraud Detection – Predictive models such as decision trees, neural networks, and regression analysis are used in anti-fraud measures (to name a few). [17].
- Secondary usage of health data [8] - Medical records for non-medical purposes Clinical data analysis can now be used to identify patients with rare diseases, research treatment options, and evaluate clinical outcomes.
- Evidence based medicine [8]:

Evidence-based medicine doctors make diagnoses based on statistical studies [8]. Doctors can make the best decisions possible using a combination of intuition and the most recent scientific evidence.

7. CONCLUSION

The healthcare industry is ready to embrace new technologies in order to improve patient care and lower healthcare costs. As a result of digitization, healthcare providers are now producing massive amounts of digital data. The proper use of health-care data can result in better outcomes at lower costs. Researchers are given analytic tools to help them make the most of the massive amounts of healthcare data they have at their disposal. If the right tools are used, data analytics in healthcare has the potential to produce positive results. In the future, researchers may look into some of the issues raised here

Figures are to be sequentially numbered in Arabic numerals. The caption must be placed below the figure. Typeset in 8 pt Times Roman with baselineskip of 10 pt. Long captions are to be justified by the “page-width”'.Use double spacing between a caption and the text that follows immediately, e.g. Fig. 1.

Previously published material must be accompanied by written permission from the author and publisher.

References

[1] Yu, H., & Wang, D. (2012, August). Research and implementation of massive health care data management and analysis based on hadoop. In 2012 Fourth International Conference on Computational and Information Sciences (pp. 514-517). IEEE.

[2] Benhlima, L. (2018). Big data management for healthcare systems: architecture, requirements, and implementation. Advances in bioinformatics, 2018.

[3] Shakil, K. A., Zareen, F. J., Alam, M., & Jabin, S. (2020). BAMHealthCloud: A biometric authentication and data management system for healthcare data in cloud. Journal of King Saud University-Computer and Information Sciences, 32(1), 57-64.

[4] Rallapalli, S., Gondkar, R., & Ketavarapu, U. P. K. (2016). Impact of processing and analyzing healthcare big data on cloud computing environment by implementing hadoop cluster. Procedia Computer Science, 85, 16-22.

[5] Khennou, F., Khamlichi, Y. I., & Chaoui, N. E. H. (2016, October). Designing a health data management system based hadoop-agent. In 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt) (pp. 71-76). IEEE.

[6] Babar, M., Arif, F., Jan, M. A., Tan, Z., & Khan, F. (2019). Urban data management system: Towards Big Data analytics for Internet of Things based smart urban environment using customized Hadoop. Future Generation Computer Systems, 96, 398-409.

[7] Bani-Salameh, H., Al-Qawaqneh, M., & Taamneh, S. (2021). Investigating the Adoption of Big Data Management in Healthcare in Jordan. Data, 6(2), 16.

[8] Mathew, P. S., & Pillai, A. S. (2015, March). Big Data solutions in Healthcare: Problems and perspectives. In 2015 International conference on innovations in information, embedded and communication systems (ICIIECS) (pp. 1-6). IEEE.

[9] Kaur, P., Sharma, M., & Mittal, M. (2018). Big data and machine learning based secure healthcare framework. Procedia computer science, 132, 1049-1059.

[10] De Silva, D., Burstein, F., Jelinek, H. F., & Stranieri, A. (2015). Addressing the complexities of big data analytics in healthcare: The diabetes screening case. Australasian Journal of Information Systems, 19.

[11] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," Health Information Science and Systems, vol. 2, article 3, 2014.

[12] I. Olaronke and O. Oluwaseun, "Big data in healthcare: Prospects, challenges and resolutions," in Proceedings of the 2016 Future Technologies Conference, FTC 2016, pp. 1152–1157, usa, December 2016.

[13] A. Belle, R.Thiagarajan, S. M. R. Soroushmehr, F. Navidi, D. A. Beard, and K. Nigerian, "Big data analytics in healthcare," BioMed Research International, vol. 2015, Article ID 370194, 2015.

[14] J. Sun and C. K. Reddy, "Big data analytics for healthcare," in Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, p. 1525, Chicago, Ill, USA, August 2013.

[15] M. Bochicchio, A. Cuzzocrea, and L. Vaira, "A big data analytics framework for supporting multidimensional mining over big healthcare data," in Proceedings of the 15th IEEE International Conference on Machine Learning and Applications, ICMLA 2016, pp. 508–513, usa, December 2016.

[16] S. Sakr and A. Elgammal, "Towards a Comprehensive Data Analytics Framework for Smart Healthcare Services," Big Data Research, vol. 4, pp. 44–58, 2016.

[17] Y. Li, C. Wu, L. Guo, C.-H. Lee, and Y. Guo, "Wiki-health: A big data platform for health sensor data management," Cloud Computing Applications for Quality Health Care Delivery, pp. 59–77, 2014.

[18] N. Poh, S. Tirunagari, and D.Windridge, "Challenges in designing an online healthcare platform for personalized patient analytics," in Proceedings of the 2014 IEEE Symposium on Computational Intelligence in Big Data, CIBD 2014, usa, December 2014.

Implementation of Personalized and Public Healthcare Model: Big Data Perspective

Sakshi Raj Singh[1], Mohammad Zunnun Khan[2], Faiyaz Ahmad[3]
Department of CSE Integral University, Lucknow, India

E- mail : [1]srssingh2022@gmail.com [2]zunnun@iul.ac.in
[3]faiyaz@iul.ac.in

Abstract

Machine learning applied to electronic health records (EHRs) may yield actionable insights, from improving upon patient risk score systems, to forecasting the beginning of illness, to optimizing hospital operations. Statistical models that harness the diversity and depth of EHR-derived data are still very uncommon and provide an attractive area for additional study. In this chapter, we give an overview of how machine learning has been implemented in clinical contexts and describe the benefits it offers over conventional analytic approaches. We highlight the methodological and practical difficulties of employing machine learning in research and practice. Although there are numerous occasions in which ML can execute healthcare duties as well or better than humans, implementation concerns will preclude large-scale automation of healthcare professional occupations for a significant duration. Ethical difficulties in the use of ML to healthcare are also highlighted.

Index Terms Health Care Data, Machine Learning Approach, clinical decision support

1 INTRODUCTION

Volume-based, amongst other changes (Agarwal, R & Dhār, V, 2014). In order to create a patient-centered healthcare delivery paradigm and a patient-focused way of thinking, the overarching guiding concept that must be adhered to is to educate people about the significance of healthcare while also reducing the cost of healthcare. In healthcare companies, both the volume of information and the interest in large quantities of data are gradually increasing, but they are getting closer and closer to the point when they won't exist at all. In order to offer outstanding treatment that is focused on the patient, it is essential to keep track of and analyses the enormous amount of data sets that are created (Agarwal, R, Khandelwal, 2015). Throughout the course of history, conventional methods have become antiquated and are no longer appropriate for breaking One of the most important and extensive businesses in the world is the healthcare sector. In the course of time, healthcare administration all over the globe has undergone a number of shifts, including moving from an approach that centered on infections to one that centered on patients to one that was down massive amounts of information. This is because the number and variety of information sources have increased at an alarmingly rapid pace over the course of the last two decades. The healthcare department generates an enormous amount of data, which necessitates the development of creative and cutting-edge tools and processes that are up to the challenge of handling it and even more so. A community-based system serves as the foundation for the social insurance framework that is used by healthcare departments.

This is as a result of the fact that it is comprised of a significant number of partners, such as physicians who specialize in a variety of fields, medical caretakers, research Centre technologists, and other individuals who collaborate in order to accomplish the common goals of lowering the cost of medication and the number of errors made while simultaneously increasing the level of high-quality healthcare experiences (Al Hamid, HA, 2017). The information that is produced by each of these partners comes from a variety of sources, including clinical notes and physical examinations, patient meetings and perceptions, tests performed at research facilities and imaging reports, medications, treatments, overviews, bills, and legal protection. As can be observed from the ever-increasing quantity of data that is being produced, the rate at which information is created on a daily basis from a wide variety of sources by many different healthcare departments has greatly grown. As a consequence of this, typical dataset handling programmers are having a harder time storing, interpreting, and dissecting this kind of information because of its complex web of interconnections. However, in addition to this, there have been significant processing advancements made in the form of innovative techniques and systems that are being developed in order to store, process, break down, and extract values from the massive amount of diverse medical information that is continually being produced. These techniques and systems can be found in the ongoing development of the healthcare industry (Aruna, 2012). As a direct consequence of this, the framework for medical services is quickly developing into a significant information industry. The demands of an ever-expanding public that is information-starved and, more recently, the operational characteristics of e-health phases have been the primary drivers behind the tremendous growth of organized and unstructured medical services information over the course of the past several decades. This growth has occurred in both an organized and unstructured manner. The risks associated with this expansion on several fronts have caused industry professionals to coin a number of new buzzwords to define "Big Data" in the healthcare sector (HBD). This is a development that is risky on many different fronts (Barbierato, 2014). However, the sheer quantity of information that is available is not the only thing that is very diverse and varied in the healthcare industry. The information itself is also very diverse and varied, with a particular emphasis on the different types of sources that provide information and the different types of objectives that seek it. Included in this group are members of the workforce of the medical services industry (doctors, clinical staff, and parental figures), benefit-giving organizations (including safety net providers), healing facilities with resources, government controllers, drug stores and pharmaceutical manufacturers (including research and development groups), and manufacturers of therapeutic devices. The statistical method known as machine learning involves fitting models to data and allowing models to 'learn' by being trained with this data. Machine learning is one of the most frequent types of artificial intelligence (AI). According to a study conducted by Deloitte in 2018 among 1,100 managers in the United States whose firms were already pursuing AI, 63 percent of enterprises questioned were using machine learning in their operations. There are a lot of different implementations of this general AI concept, which is at the heart of a lot of different AI techniques. Precision medicine is the most common application of traditional machine learning in the field of healthcare. Precision medicine involves determining which treatment protocols are most likely to be successful on a patient based on a variety of patient attributes and the context in which the treatment is being administered. The vast majority of applications involving machine learning and precision medicine need a training dataset for which the end variable is known; this kind of learning is referred to as supervised learning. The neural network is an example of a more complicated kind of machine learning. A technique that has been accessible since the 1960s has been well established in healthcare research for many decades. It has also been utilized for categorizing applications, such as assessing whether or not a patient would get a certain illness. It looks at issues in terms of the inputs, the outputs, and the weights of the variables or 'features' that correlate the inputs with the outcomes. It has been compared to the method in which neurons interpret signals, however this comparison to the operation of the brain is not very strong. Deep learning, also known as neural network models, are among the most complicated types of machine learning [5]. These models include several layers of characteristics or variables that are used to make predictions. Because of the increased computing power of today's graphics processing units and cloud architectures, such models may include hundreds or even tens of thousands of features that have been concealed from view. The detection of possibly malignant tumors in radiological pictures is one of the most prevalent applications of deep learning in the healthcare industry (Luna, 2014). Radionics, also known as the identification of clinically significant signals in imaging data that are beyond what can be observed by the human eye, is one field that is progressively benefiting from the use of deep learning. In oncology-related image analysis, radionics and deep learning are most often seen in conjunction with one another.

II Big Data and Health Care

Big data has been efficiently exploited in the business sector for the detection of behavioral patterns of customers in order to build novel commercial services and solutions. This has been made possible thanks to the core value of big data. The use of big data serves predictive analytical approaches and machine learning platforms (Sarkar et al., 2017) in the field of healthcare, which allows for the supply of sustainable solutions like the execution of treatment plans and tailored medical care (Sarkar et al., 2017). (Bid good, 1997) did a comparison of the big data that is created from the business sector with the big data that is generated from the healthcare sector under various qualities and their values. Instead of volume, velocity, and variety, they rethought the characteristics of big data in healthcare as silos, security, and variety. These are the three qualities that constitute big data in healthcare. The term "Silo" refers to the legacy database that stores public healthcare information and is kept in locations such as hospitals owned by various parties. The security element alludes to the additional diligence that is required for the upkeep of healthcare data. The diversity characteristic denotes the presence of healthcare data in a range of formats, including structured, unstructured, and semi-structured data respectively. The field of healthcare has, from the point of view of the many stakeholders involved, been subject to a variety of pragmatic alterations at various phases since the birth of big data analytics and the technology connected with it (Sukumar, 2015). In addition to the analysis of legacy sources, which includes patient medical history, diagnostic and clinical trials data, drug effectiveness index, and other relevant information, the impact of big data on the healthcare industry has led to the discovery of new data sources, such as social media platforms, telematics, wearable devices, and so on. When various data sources and analytics are combined, it offers a useful source of information for health care researchers, which may help them work toward the development of innovative healthcare solutions. The healthcare industry has always generated enormous amounts of data, which have been put to use to make improvements to patient care. This data is created as a result of record keeping, compliance and regulatory duties, and patient care (Bidgood, 1997). Even while the great majority of data is still stored in hard copy form at this time, there is a growing movement toward the rapid digitization of these enormous amounts of data in the not too distant future. These enormous amounts of information, collectively referred to as "big data," have the potential to facilitate a wide variety of functions within the fields of medicine and healthcare. Some examples of these functions include clinical decision support, disease surveillance, and population health management. They are being pushed forward as a result of obligatory requirements as well as the possibility of achieving cost savings while simultaneously enhancing the standard of healthcare service. Reports indicate that the data generated by the healthcare system in the United States alone topped 150 Exabyte's in the year 2011. If the present growth rate remains the same, the amount of big data used in healthcare in the United States will soon exceed the petabyte scale (1021 gigabytes), and not long after that, it will reach the yottabyte scale (1024 gigabytes). It is estimated that the health network Kaiser Permanente, which is based in California and has more than 9 million members, has between 26.5 and 44 petabytes of potentially rich data from electronic health records (EHRs) stored on its servers. Examples of this type of data include photos and annotations (Wang, 2015). Big data refers to electronic health data sets that are so large and complex that it is difficult (if not impossible) to manage them with traditional software and/or hardware. Not only is it difficult (if not impossible) to manage them with traditional software and/or hardware, but it is also difficult (if not impossible) to manage them with traditional or commonly used tools and methods for data management (Youssef, 2015). The sheer volume of big data in the healthcare industry is intimidating, but that's not the only reason why; the sheer diversity of data types and the speed with which it has to be processed and evaluated in order for it to be helpful are also major factors. According to McKinsey, the use of big data may be able to aid in the reduction of waste and inefficiency in the following three areas (see figure 1.1).

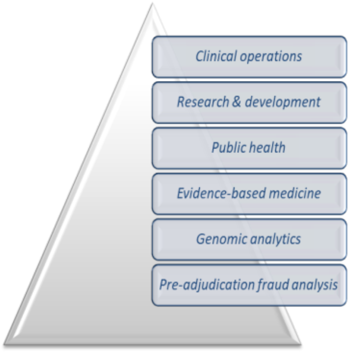


Fig 1 Big data Quantification Area

III Proposed Model

Public health data analytics has been a major issue for years. In order to provide high-quality treatment, public health care institutions must be able to organize, analyses, and interpret data [16]. Healthcare companies may benefit from a broad variety of data analysis technologies, including clinical and operational applications that can aid in the collection of health data. Online patient portals, electronic medical records, and health-tracking gadgets are just a few of the tools used to gather health-care data. Clinical notes and medical photographs are two examples of data that may be in various forms or even unstructured [15]. This includes master data management, which links master data to a single and dependable source of data for use in improving patient care and ensuring patient safety. In the proposed model, the data is analyzed using quantitative and qualitative methodologies in order to identify trends and patterns. Health data analysts should be able to "acquire, manage, analyses, interpret, and convert data into accurate, consistent, and timely information" if they are to be effective.

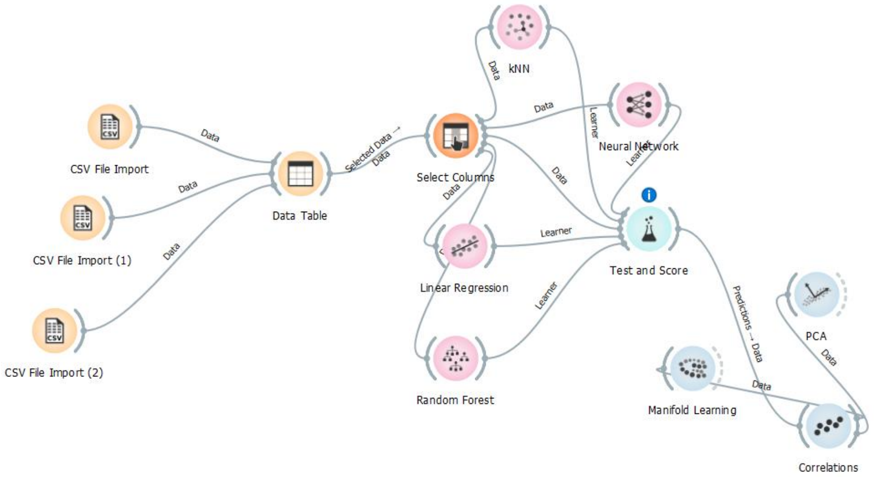
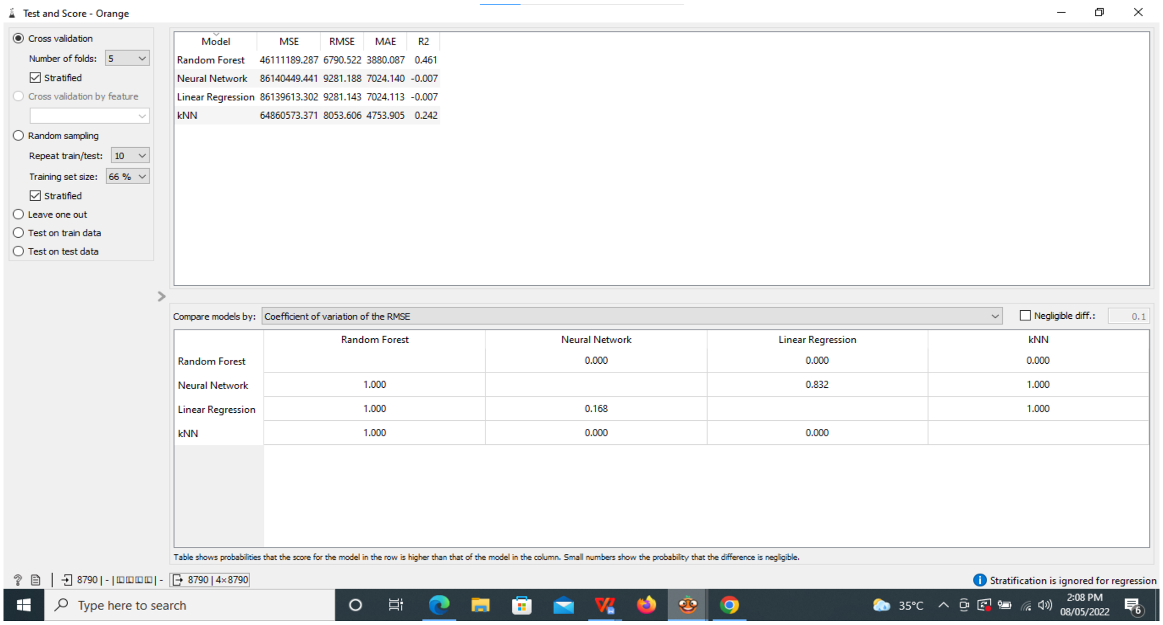


Figure 2: Flow chart of Proposed Model

Knn Algorithm is applied with 5 neighbours and Eculidean metric is utilized on Uniform Weight. Neural Network is applied with 100 hidden layerand maximum iteration was 200 while training the data. Liner regression is applied with Lasso regularization (L1) strength (alpha=0.0001) and elastics net mixing 0.50:0.50. Random Forest is applied with 10 no of trees and 5 number of attributes considered at each split.



Test and Score - Orange

● Cross validation

Number of folds: 5

☑ Stratified

○ Cross validation by feature

○ Random sampling

Repeat train/test: 10

Training set size: 66 %

☑ Stratified

○ Leave one out

○ Test on train data

○ Test on test data

Model	MSE	RMSE	MAE	R2
Random Forest	46111189.287	6790.522	3880.087	0.461
Neural Network	86140449.441	9281.188	7024.140	-0.007
Linear Regression	86139613.302	9281.143	7024.113	-0.007
kNN	64860573.371	8053.606	4753.905	0.242

Compare models by: Coefficient of variation of the RMSE

□ Negligible diff.: 0.1

	Random Forest	Neural Network	Linear Regression	kNN
Random Forest		0.000	0.000	0.000
Neural Network	1.000		0.832	1.000
Linear Regression	1.000	0.168		1.000
kNN	1.000	0.000	0.000	

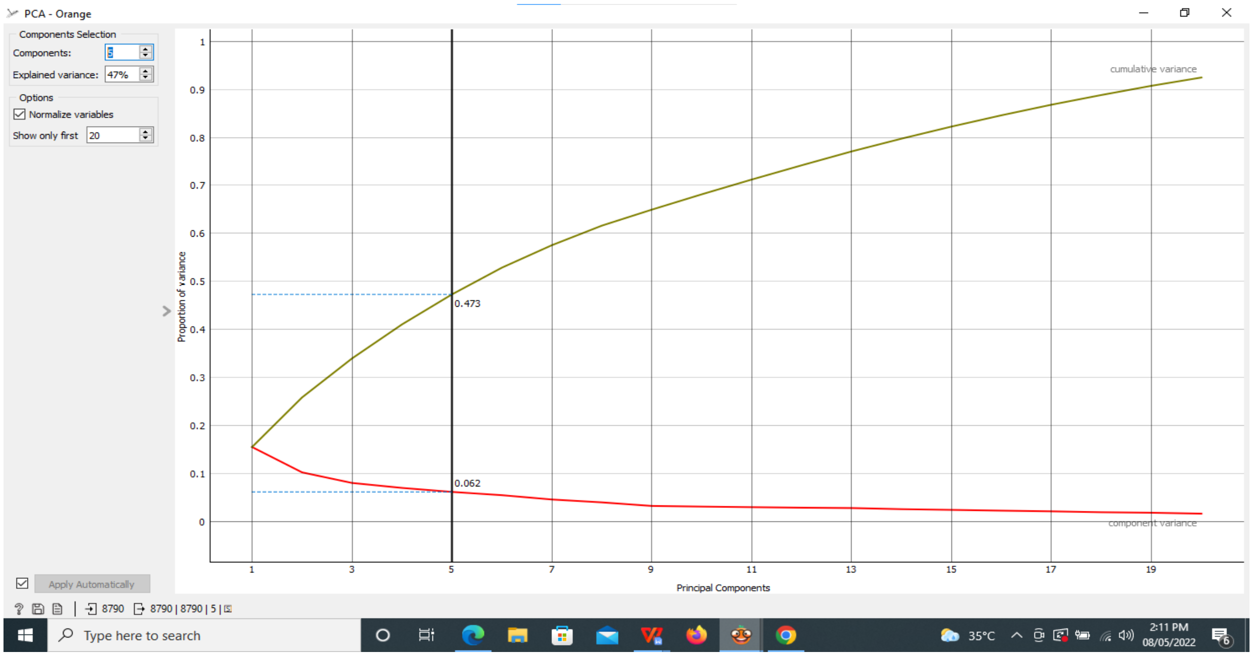
Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

8790 | - | - 8790 | 4×8790

Stratification is ignored for regression

Figure 3: Model description

In figure 3, MSE measures the average of the squares of the errors or deviations (the difference between the estimator and what is estimated). RMSE is the square root of the arithmetic mean of the squares of a set of numbers (a measure of imperfection of the fit of the estimator to the data) MAE is used to measure how close forecasts or predictions are to eventual outcomes. R2 is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable. CVRMSE is RMSE normalized by the mean value of actual values. Train time as cumulative time in seconds used for training models and Test time as cumulative time in seconds used for testing models.



PCA - Orange

Components Selection

Components: 5

Explained variance: 47%

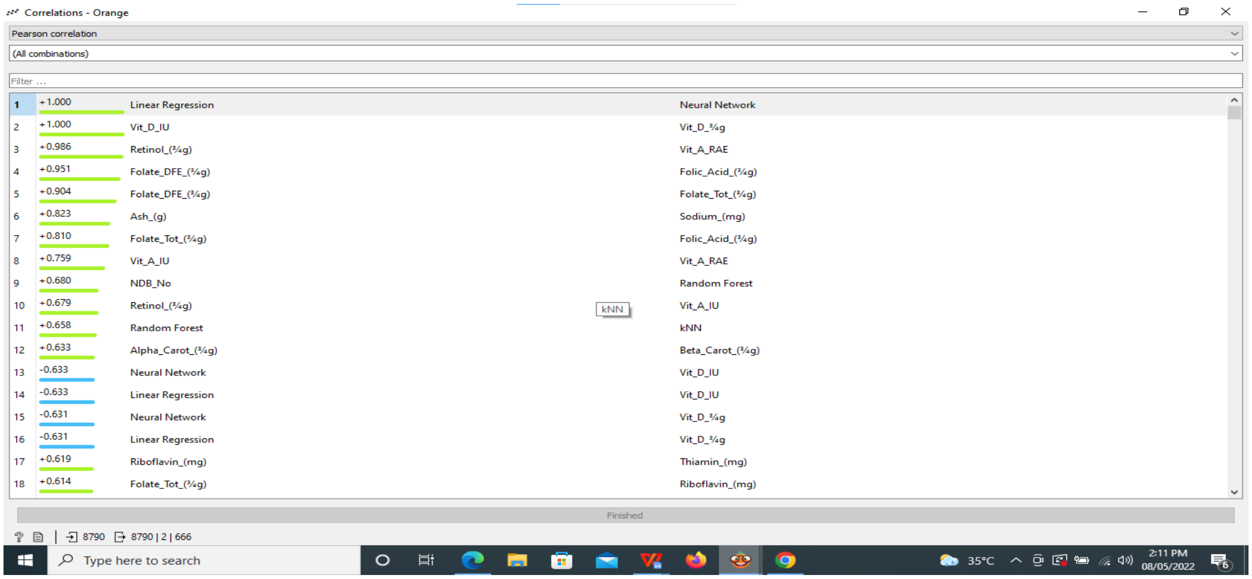
Options

☑ Normalize variables

Show only first 20

Figure 4: PCA evaluation

In figure 4, transforms the data into a dataset with uncorrelated variables, also called principal components. PCA widget displays a graph (scree diagram) showing a degree of explained variance by best principal components and allows to interactively set the number of components to be included in the output dataset.



Correlations - Orange

Pearson correlation

(All combinations)

1	+1.000	Linear Regression	Neural Network
2	+1.000	Vit_D_IU	Vit_D_¾g
3	+0.986	Retinol_(¾g)	Vit_A_RAE
4	+0.951	Folate_DFE_(¾g)	Folic_Acid_(¾g)
5	+0.904	Folate_DFE_(¾g)	Folate_Tot_(¾g)
6	+0.823	Ash_(g)	Sodium_(mg)
7	+0.810	Folate_Tot_(¾g)	Folic_Acid_(¾g)
8	+0.759	Vit_A_IU	Vit_A_RAE
9	+0.680	NDB_No	Random Forest
10	+0.679	Retinol_(¾g)	Vit_A_IU
11	+0.658	Random Forest	kNN
12	+0.633	Alpha_Carot_(¾g)	Beta_Carot_(¾g)
13	-0.633	Neural Network	Vit_D_IU
14	-0.633	Linear Regression	Vit_D_IU
15	-0.631	Neural Network	Vit_D_¾g
16	-0.631	Linear Regression	Vit_D_¾g
17	+0.619	Riboflavin_(mg)	Thiamin_(mg)
18	+0.614	Folate_Tot_(¾g)	Riboflavin_(mg)

Figure 5: Correlation Outcomes

In figure 5, Correlations compute Pearson or Spearman correlation scores for all pairs of features in a dataset. These methods can only detect monotonic relationship. Go to the most negatively correlated pair, DIS-NOX. Now connect

Scatter Plot to Correlations and set two outputs, Data to Data and Features to Features. Observe how the feature pair is immediately set in the scatter plot. Looks like the two features are indeed negatively correlated.

Discussion

Big data analytics takes use of the discrepancy between organized and unorganized data. The challenge of moving to an integrated data environment is well-known. Intriguingly, the fundamental premise of big data is based on the premise that, the more data there is, the more insights can be gleaned from it and the more accurate forecasts can be made. Several reputable consulting organizations and health care providers correctly predict that the big data healthcare business will develop at an exponential pace in the near future. However, in a short period of time, we have seen a wide range of analytics being used that have had a substantial influence on the healthcare industry's decision-making and performance. Computing professionals have been obliged to devise novel techniques to deal with the ever-increasing volume of medical data from diverse sources in a limited period of time due to the exponential expansion of this data. Scientific research as well as clinical practice has seen an increase in the use of computer systems for signal processing It's possible to build an accurate representation of a person's physical structure by integrating physiological data with "-omits" approaches. We can learn more about illness conditions and maybe build new diagnostic tools thanks to this revolutionary notion. We need to keep an eye on the ever-increasing quantity and quality genetic data, as well as the inherent flaws that arise from experiments and analysis. However, there are chances to implement systemic changes in healthcare research at each stage of this lengthy process. Table 1 and graph 6 show the results of our comparison of the two study modules (public and private module).

T(0.25)**able 1: Comparative study**

	Public Healthcare Model				Private Healthcare Model			
Model	MSE	RMSE	MAE	R2	MSE	RMSE	MAE	R2
KNN	54733279168.00	2339514461.00	18888835517.00	(0.29)	86140449.44	9281.19	7024.14	(0.01)
Tree	62190269106.00	2493797688.00	1928187338.00	(0.47)	86133613.30	9281.14	7024.11	(0.01)
SVM	5301557836533.00	2302511202.00	1843191422.00		64860573.37	8053.61	4753.91	0.42
Linear rig ration	411859945320.00	2029433293.00	1616870091.00	0.03	461111189.00	6790.52	3880.09	0.46

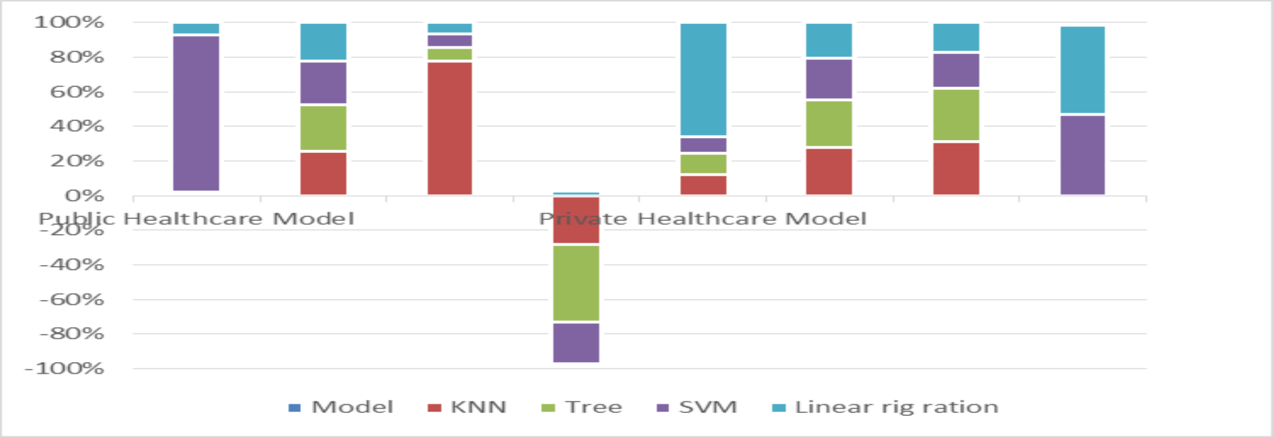


Figure 6: Critical observation

References

1. User Testing Healthcare Chabot apps are on the rise but the overall customer experience (cx) falls short according to a User Testing report. San Francisco: User Testing, 2019.
2. Davenport TH, Kirby J. Only humans need apply: Winners and losers in the age of smart machines. New York: Harper Business, 2016.
3. Char DS, Shah NH, and Magnus D. Implementing machine learning in health care – addressing ethical challenges. N Engle J Med 2018;378:981–3.
4. Fakoor R, Ladhak F, Nazi A, Huber M. Using deep learning to enhance cancer diagnosis and classification. A conference presentation The 30th International Conference on Machine Learning, 2013.
5. Health data, McGraw-Hill Concise Dictionary of Modern Medicine. McGraw-Hill. 2002.
6. Agarwal, R &Dhar, V, 2014, ‗Big Data, data science and analytics: The opportunity and challenge for IS research, vol. 25, no. 3, pp. 443-448.
7. Agarwal, R, Khandelwal, A &Stoica, I, 2015, ‗Succinct: Enabling Queries on Compressed Data', In NSDI, vol. 15, pp. 337-350.
8. Al Hamid, HA, Rahman, SMM, Hossain, MS, Algren, A &Alamri, A, 2017, ‗A security model for preserving the privacy of medical Big Data in a healthcare cloud using a fog computing facility with pairing-based cryptography', IEEE Access, vol. 5, pp. 22313-22328.
9. Ardagna, CA, Damiani, E, Frati, F &Rebeccani, D, 2016, ‗A configuration-independent score-based benchmark for distributed databases', IEEE Transactions on Services Computing, vol. 9, no. 1, pp. 123-137.
10. Arora, R & Aggarwal, RR, 2013, ‗Modeling and querying data in mongodb', International Journal of Scientific and Engineering Research, vol. 4, no. 7, pp. 141-144.

11. Aruna, S, Nandakishore, LV &Rajagopalan, SP, 2012, ‗Cloud based decision support system for diagnosis of breast cancer using digital mammograms', International Journal of Computer Applications (IJCA), vol. 1, pp. 1-3.
12. Barbierato, E, Gribaudo, M &Iacono, M, 2014, ‗Performance evaluation of NoSQL big-data applications using multi-formalism models', Future Generation Computer Systems, vol. 37, pp. 345-353.
13. Olshannikova, E.; Ometov, A.; Koucheryavy, Y.; Olsson, T. Visualizing Big Data with augmented and virtual reality: Challenges and research agenda. J. Big Data 2015, 2, 22.
14. Bani-Salameh, H.; Jeffery, C.; Hammad, M. Developers' social networks-tools analysis based on the 3Cs model. Int. J. Netw.Virtual Organ. 2013, 13, 159–175
15. Luna, D.R.; Mayan, J.C.; Garcìa, M.J.; Almerares, A.A.; Househ, M. Challenges and potential solutions for big data implementa-tions in developing countries. Yearb. Med. Inform. 2014, 23, 36–41.
16. Wang, Y.; Kung, L.; Byrd, T.A. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. Technol. Forecast. Soc. Chang. 2018, 126, 3–13.
17. Wang, L.; Wang, G.; Alexander, C.A. Big data and visualization: Methods, challenges and technology progress. Digit. Technol. 2015, 1, 33–38.
18. VistA. Available online: https://worldvista.org/AboutVistA (accessed on 5 February 2021).
19. Bidgood, W.D., Jr.; Horii, S.C.; Prior, F.W.; Van Syckle, D.E. Understanding and using DICOM, the data interchange standard for biomedical imaging. J. Am. Med. Inform. Assoc. JAMIA 1997, 4, 199–212
20. Sarkar, B.K., 2017. Big data for secure healthcare system : a conceptual design. Complex Intell. Syst. 3, 133–151
21. Sukumar, S.R., Natarajan, R., Ferrell, R.K., 2015. Quality of big data in health care. Int. J. Health Care Qual. Assur. 28, 621–634.
22. Youssef, A.E., 2014. A Framework for secure healthcare systems based on big data analytics in mobile cloud computing environments. Int. J. Ambient Syst. Appl. 2, 1–11.