

# ANALYSIS ON APPLICATIONS OF DATA MINING WITH CRYPTOGRAPHY FOR MEDICAL HEALTH CARE DATA

<sup>1</sup>Bommakanti Maneesh Kumar, <sup>2</sup>Sai Charan Ankith Gopiseti, <sup>3</sup>Mandava Lakshmi Ananya, <sup>4</sup>Bommakanti Vishwas

<sup>1</sup>*Bommakanti Maneesh Kumar, Advisory Associate solution Advisor Deloitte*

<sup>2</sup>*Sai charan Ankith Gopiseti, Analyst Deloitte consulting USI*

<sup>3</sup>*Mandava Lakshmi Ananya, Software engineer Applaud*

<sup>4</sup>*Bommakanti Vishwas, Senior software engineer, Wipro Limited*

\*\*\*

**ABSTRACT:** Day by day, the health care database is growing. With this database, the main issue is how to make these datasets helpful to society as a whole. Data mining is one of the alternatives. In data mining applications, the goal of data mining is to turn data into information. To find and disseminate relevant health information, data mining is used in healthcare systems. Data mining applications in the healthcare sector are described in detail in this research, which aims to reduce the complexity of analyzing health data transactions. This type of data mining method presents a number of challenges, including how to ensure that personal information about individuals is not disclosed to a third party at the moment of sharing it. Associative rule mining, k-anonymity, Data perturbation, condensation method, and cryptography are some of the approaches used to address this issue. Personal data privacy is the focus of the present work. Various hybrid approaches have been established in this study and implementation to provide privacy for various sorts of sensitive attributes in a database. As part of the initial phase of this project, the focus was on identifying distinct types of sensitive traits through a survey of people of different ages. To deploy hybrid cryptography techniques, vertically and horizontally partitioned databases were used in the second phase of this project. With respect to different database file sizes based on encryption time and memory usage before and after encryption, this third phase focuses on visualizing the results collected for various ways implemented in second phase.

**Key words:** Healthcare Systems, Data mining, Cryptography, Database.

## 1. INTRODUCTION

Data mining is designed to gather useful information from large databases or data stores. For commercial and scientific industries, data mining applications are used[1]. The scientific aspect of application data mining is at the core of this study. The nature of the data sets often differs greatly from traditional data mining applications in scientific data mining. Data mining applications, types of data used and data extracted in the healthcare industry are investigated during this work. In prevention and diagnosis of diseases, data mining algorithms in the health sector play a key role. Several

applications in the fields of data mining, including medical devices, pharmaceutical and hospital management, are found. The useful and secret information is to be found in the database. The mining of data is commonly known as the discovery of data. Discovery of knowledge is an interactive process involving an understanding of application, selection and generation of data sets, pretreatment and transformation of data. For a variety of applications, data mining was used, including marketing, customer connections, engineering and health analysis, predictions by specialists, mobile and mobile computing and Web mining.

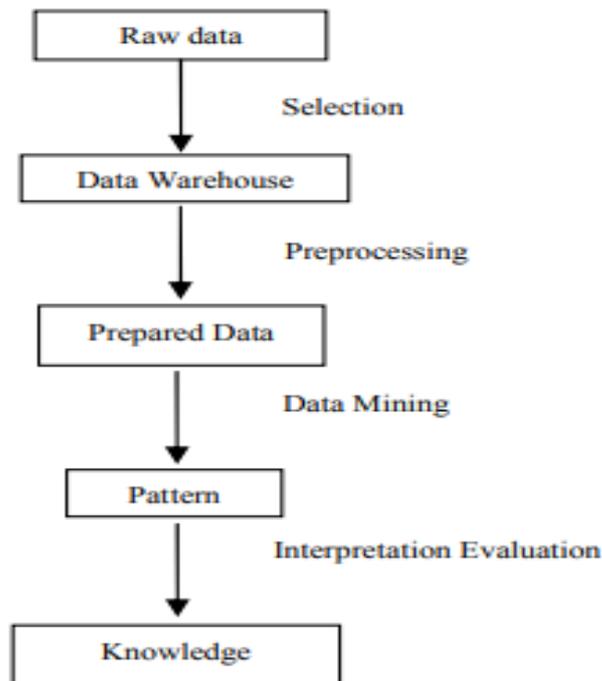


Fig. 1: Data mining role in the process of knowledge discovery

Medical data means databases, such as patient records, which store healthcare information. Many of these medical data are stored electronically during IT development. These databases have large volumes of data. Different sources may be provided for medical data, for example, X-rays, CT scans, MRIs, ultrasound, etc. Thus, there is an exponential increase in data volumes and databases needed for digital data to be stored[1]. In addition, raw medical data are typically enormous and unlike in nature, and can be gathered from various sources, such as physician observations and evaluations, patient images and interviews[2]. Different types of medical data. The images, data sets, signals, and so on can be used. In this scenario, due to research and development in the field of information collection tools, an enormous amount of information or data is visible electronically. There is a clear increase in database size for the storage of such large amounts of data or information[3].

### 1.1 Cryptography

It is now an all-pervasive mode for the transmission of multimedia data. The sensitive issue of ensuring data security is now very important with the advent of electronic commerce, in particular in an ever more open network environment of the modern era. In general, historical cryptography encryption technologies are commonly used to protect data security. The process of cryptographic applications can be defined as encryption. Cryptography allows the sender to store and transmit

useful information over the Internet (which is an insecure network). A single text is called data from Figure 2 below that can be read and understood without specific measurement. This simple and direct text is now encrypted to prevent intruders from using data. The result of this encryption process is a cypher text which is not understood. The recovery process from the cypher text is called the decryption process. the actual (i.e. plain text). The process of encryption and decryption is performed with either a public or private key or with both keys.

### 2. LITERATURE REVIEW

A review of literature is a text written with key contemporary knowledge points including substantial theoretical and methodological contributions to a particular subject.

The main topics of HianChyeKoh and Gerald Tan are data mining and their applications with main fields such as treatment efficiency, health care management, fraud and misuse detection and client relationship management[1].

JayanthiRanjan shows how data mining can find useful patterns in order to find observable patterns and remove them from these large data. This article shows how data mining in the pharmaceutical industry can improve the quality of decision making. Adverse drug reactions[2] have been addressed in the pharmaceutical industry.

The hybrid prediction system, M. Durairaj, K. Meena, consists of Rough Set Theory (RST) and the Medical Dispensation Medical Network (ANN). It explained the process for developing a new technique and software for the analysis of medical data to support competent solutions. Provide the hybrid RST and ANN-based analysis tool and indicative predictions. The on spermatological experiments are carried out with a view to predicting animal semen excellence. The proposed hybrid prediction system is used to prepare a medical database and train ANN to prevent production. To measure prediction accuracy, a comparison between observed and forecast cleavage rate[21] is used.

The progressive encryption system used to protect the data was Divya et al.[4]. The new Hadoop biosensor, Sunspot WLAN Architecture, ECC Digital Signature Algorithms, MySquare and Hadoop HDFS cloud storage are to be published in Hong Song Chen [7] research article.

ShwetaKharyad had various methods of data mining for breast cancer diagnosis and prognosis The decision tree is found to be the best predictor with 93,62% accuracy on benchmark data sets as well as SEER data sets[5].

Elias has spoken of Lemuye AIDS is an HIV disease that weakens the immune system of the body until simple infections that most healthy people resist cannot be combated. A priori algorithm is used to discover association rules. As an algorithm data mining tool, WEKA 3.6 is used. For the HIV prediction, the J48 classification is 81.8 percent accurate[6].

### 3. DATA MINING APPLICATIONS IN HEALTHCARE SECTOR

In data mining, valid, new and potentially useful data patterns are identified by us as the nontrivial process. Companies have the problem of overloading the information with the widespread use of databases and their explosive size growth. It becomes a problem for each company to use these large volumes of data effectively.

Additional categories may be divided for healthcare data mining applications:

a. Diagnosis and prediction of diseases – Data mining in the healthcare industry, diagnostics and disease forecasts has an important purpose. Using healthcare data mining has helped medical practitioners improve their health services[15]. The selection of a wrong treatment for a patient that could damage the patient's health cannot expend time or money.

b. Ranking of various hospitals – In order to classify all the data from different hospitals, data mining techniques

are used to study. Due to its capability for treating patients with severe diseases, different organisations, i.e. higher ranking hospitals are more suitably designed to treat high-risk patients because it is their highest priority, classifies the various hospitals.

c. Better treatment techniques – By comparing all treatment methods, both the doctor and the patient can select the best treatment option. You can choose the best treatment techniques both as efficiency and cost. The side-effects of various treatments can also be detected through data mining and thus reduce the risk for patients[6].

d. Effective treatments– In order to analyse the effectiveness of treatments, data collection is applied by comparing factors such as causes, symptoms and side effects. For example, the results of treatments of various patients who have been treated with different drugs can be compared. This allows us to determine which treatment is effective for the health and the costs of the patient.

e. Better quality services provided to patients– With technological advances, we have already stored large numbers of data in digital form. Data mining can help in the extraction of many interesting unknown patterns when applied to this massive medical data. By using these patterns, the quality of services and treatment for patients can be improved. Data mining also helps to understand the needs of patients and to improve their treatment[6].

f. Medical Device Industry: One important point of the healthcare system is medical equipment. This one is mainly used for the best communication work. The development of mobile healthcare apps providing comfort and safety to keep patients' vital signal continuously monitored is provided by mobile communications and wireless, low-cost sensors[11]. Lightweight single pass data stream mining algorithms can carry out in-board mobile devices real-time analyses, taking account of available resources such as the battery charge and the storage available.

g. Hospital Management Organizations including modern hospitals can generate huge amounts of data and collect them. Use of data mining in hospital-based information system data that visualise the temporal behaviour of global hospital activities[12].

Hospital management in three layers:

- Services for hospital management
- Services for medical staff
- Services for patients

#### 4. PROPOSED METHODOLOGY

We propose Advanced Encryption Standard to overcome the problems in Cryptography Base 64 (AES). Computer power is growing today, and hackers need stronger algorithms in order to deal with attacks. AES fulfils this requirement. It is frequently used with three common 128, 192, 256 bit keys. It is a standard. AES is built into software and hardware, and even on small devices such as smartphones, it works fast and efficiently. We use a 128bit block and 128, 192,256 bit key, respectively, with a larger block size and larger keys. In the longer term, AES will ensure greater safety, because in practical use it is considered unbreakable.

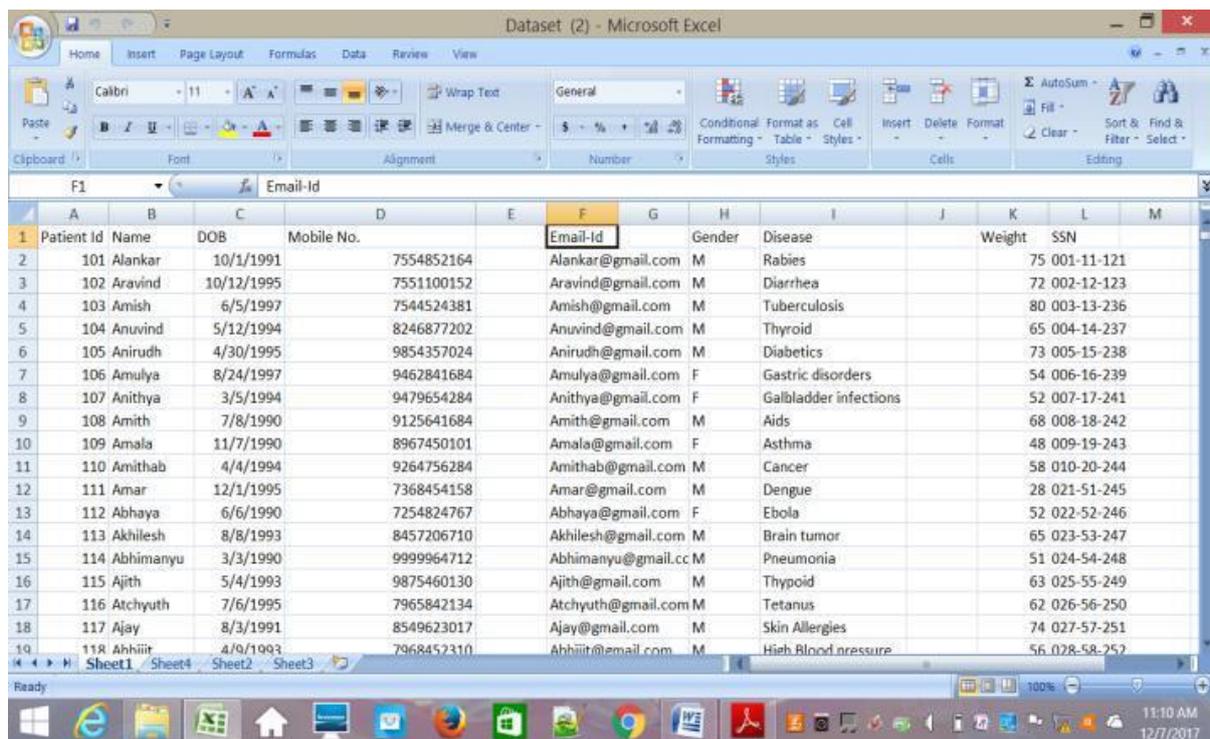
#### A. Research Methodology

Big data analysis in healthcare is used to de-identify the following platforms and tools. The procedure is as follows:

1. Data collection
2. Hadoop Cluster and Map Reduce
3. Experiments

#### B. Data Collection

Big data is regarded as a peta byte in this paper. Big Data can be used in many fields, like web and social networking, machinery, massive exchanges and biometric data sensor information. We have collected patient data, such as patient names, email identifications, mobile no, ssn, sex, weight, etc. Sample data sets shown in Figure 2.



Patient Id	Name	DOB	Mobile No.	Email-Id	Gender	Disease	Weight	SSN
101	Alankar	10/1/1991	7554852164	Alankar@gmail.com	M	Rabies	75	001-11-121
102	Aravind	10/12/1995	7551100152	Aravind@gmail.com	M	Diarrhea	72	002-12-123
103	Amish	6/5/1997	7544524381	Amish@gmail.com	M	Tuberculosis	80	003-13-236
104	Anuvind	5/12/1994	8246877202	Anuvind@gmail.com	M	Thyroid	65	004-14-237
105	Anirudh	4/30/1995	9854357024	Anirudh@gmail.com	M	Diabetics	73	005-15-238
106	Amulya	8/24/1997	9462841684	Amulya@gmail.com	F	Gastric disorders	54	006-16-239
107	Anithya	3/5/1994	9479654284	Anithya@gmail.com	F	Galbladder infections	52	007-17-241
108	Amith	7/8/1990	9125641684	Amith@gmail.com	M	Aids	68	008-18-242
109	Amala	11/7/1990	8967450101	Amala@gmail.com	F	Asthma	48	009-19-243
110	Amithab	4/4/1994	9264756284	Amithab@gmail.com	M	Cancer	58	010-20-244
111	Amar	12/1/1995	7368454158	Amar@gmail.com	M	Dengue	28	021-51-245
112	Abhaya	6/6/1990	7254824767	Abhaya@gmail.com	F	Ebola	52	022-52-246
113	Akhilesh	8/8/1993	8457206710	Akhilesh@gmail.com	M	Brain tumor	65	023-53-247
114	Abhimanyu	3/3/1990	9999964712	Abhimanyu@gmail.com	M	Pneumonia	51	024-54-248
115	Ajith	5/4/1993	9875460130	Ajith@gmail.com	M	Thypoid	63	025-55-249
116	Atchyuth	7/6/1995	7965842134	Atchyuth@gmail.com	M	Tetanus	62	026-56-250
117	Ajay	8/3/1991	8549623017	Ajay@gmail.com	M	Skin Allergies	74	027-57-251
118	Abhiit	4/9/1993	7968452310	Abhiit@gmail.com	M	High Blood nressure	56	028-58-252

Fig. 2. Sample Dataset

#### C. Hadoop Cluster and Map Reduce

Hadoop is a software framework which enables large data sets to be processed via large computer clusters. Hadoop Distributed File System is a distributed Java file system which collects all sorts of data without previous organisation. A parallel software model for large amounts of data is the cartographical decrease. From the HDFS to the Map reduction the Hadoop Cluster is

connected. This allows us to implement the Hadoop cluster programme.

#### D. Experiment

Hadoop is a framework for open source writing software that requires enormous data processing from the Apache Software Foundation, in Java. Parallel to this, large clusters work on clusters with thousands of computer nodes. It is also extremely reliable and tolerant to fault in

data processing. After the completion of Hadoop installation, Hadoop can be setup into a cloud-era operating system, and the daemons automatically start the HDFS process. Map Reduce is an algorithm or design

for fast processing of enormous amounts of data. It can be divided into Mapper and Reducer according to its name[14]. The hadoop application has demonstrated in Figure 3.

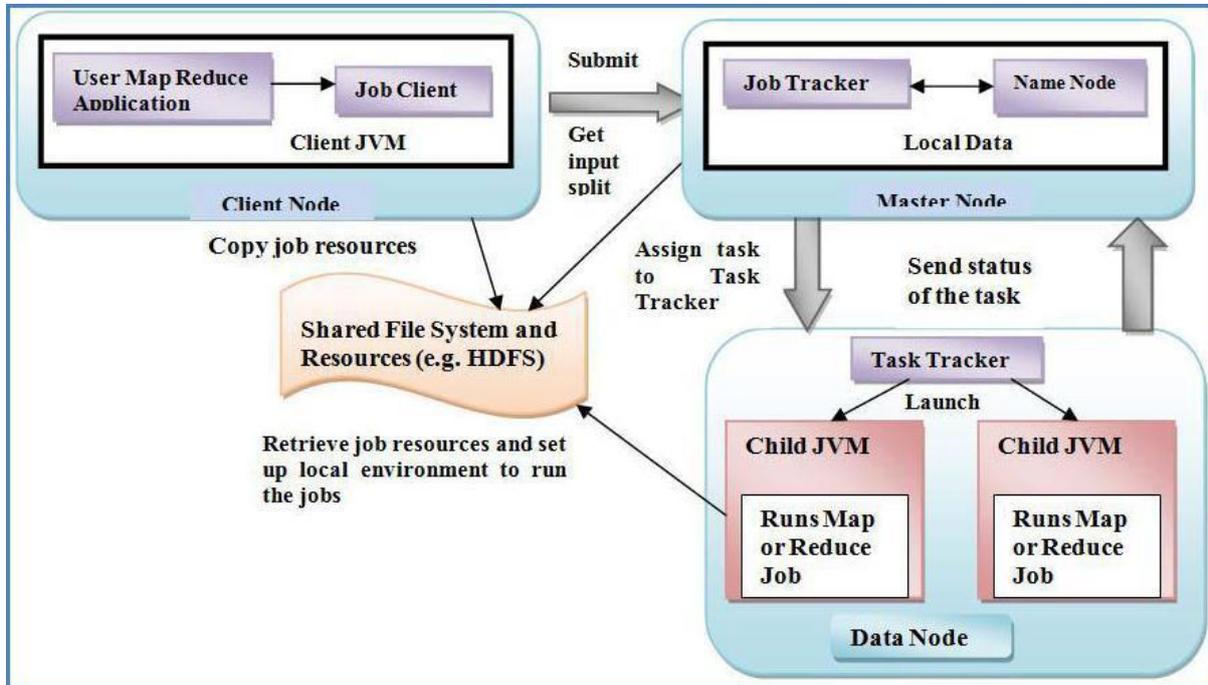


Fig. 3. Hadoop applications and infrastructure interactions.

### 5. RESULTS AND DISCUSSION

This chapter is a thorough comparative study by several researchers on data mining applications in the health

sector. The main way of predicting the positive results of healthcare data is to use data mining tools. To determine the level of accuracy of different health issues, different data mining tools are used.

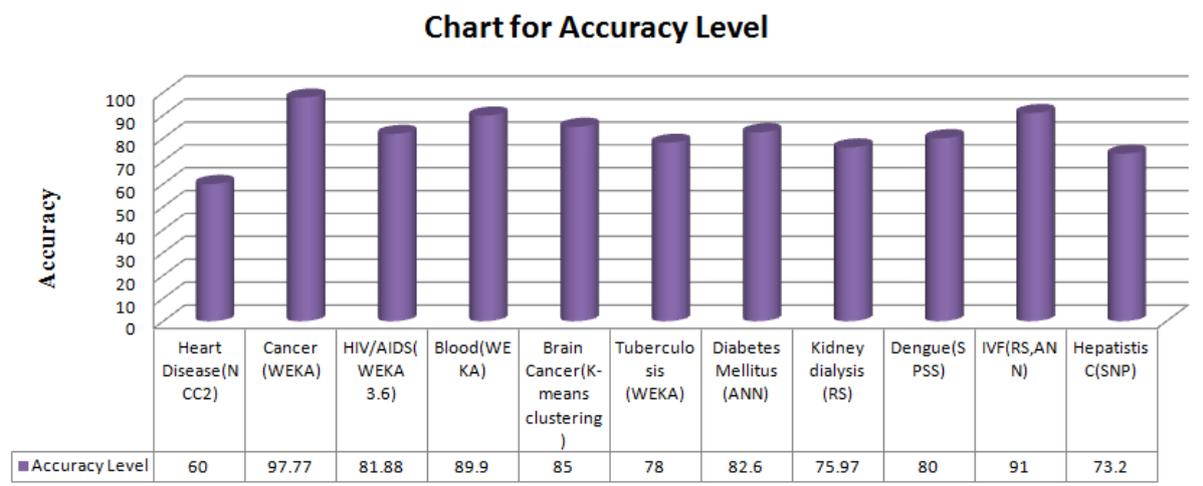


Fig. 4. Data mining tools for diagnosis chart for accuracy level

The data mining tools and the level of accuracy are shown in Figure 4 of this table with health problem

values. This graph compares prediction accuracy levels for various applications in data mining.

## Percentage of Accuracy in Estimating Success

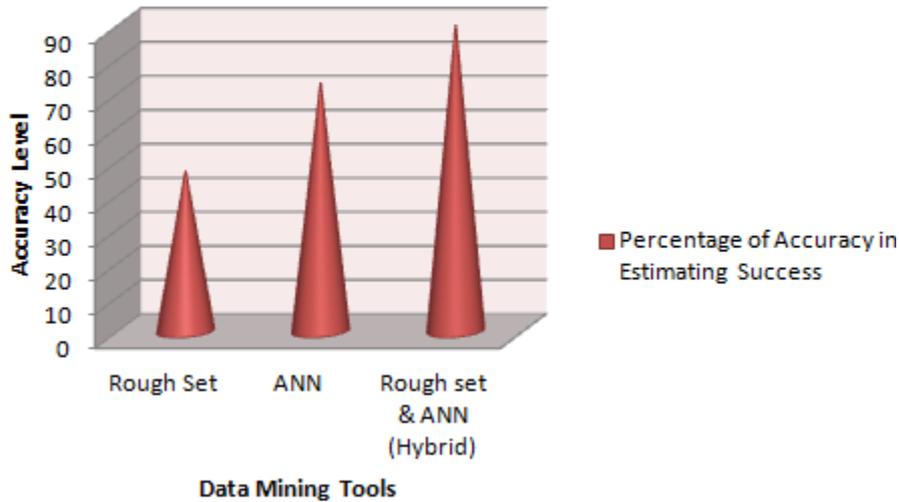


Fig. 5. Rough Set, ANN and Hybrid technology success rate.

The prediction accuracy of this hybrid ANN/RST approach is approximately 90%. This comparison results for the success rate prediction of IVF treatments in three separate data mining applications are presented in Fig. 5.

### A. Preliminary Dataset Preparation

The data collected from the HSCIC (Health & Social Care Information Center) can include patient dummy patient health data, patient identification, date of birth, Email ID, sex, diseases and SSN, and can be used as a part of this work. The data is kept in the CSV (Comma Separated Value) file.

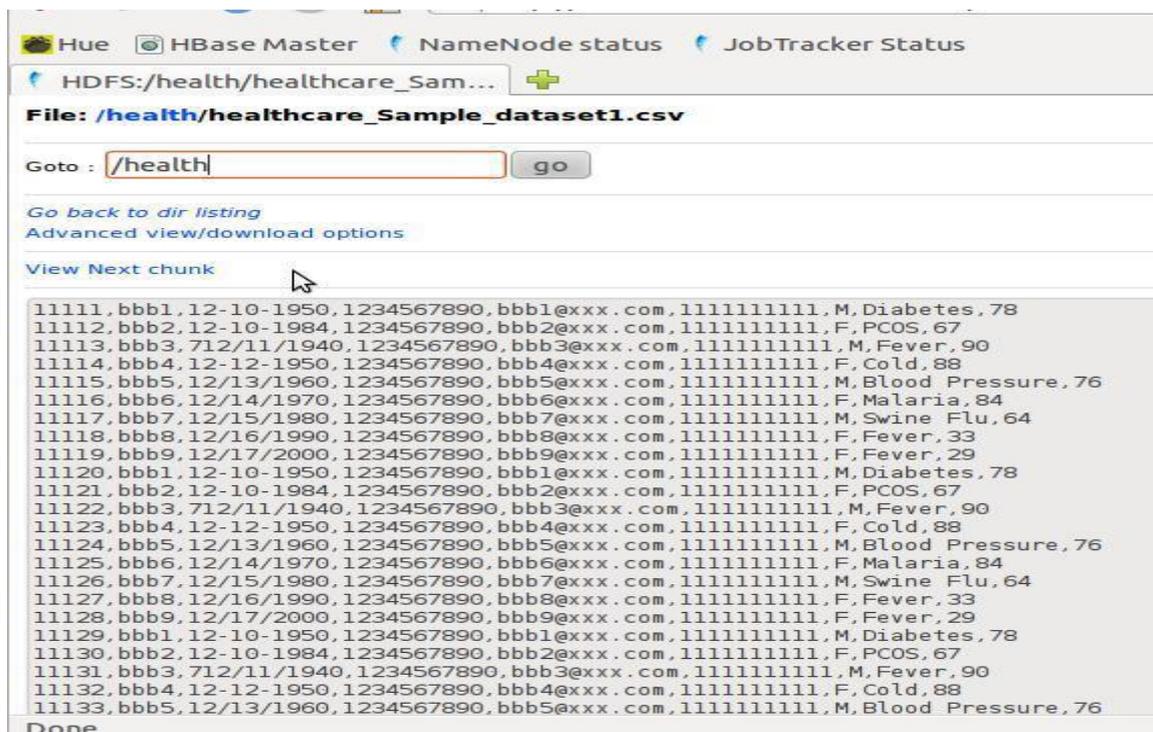


Fig. 6. Health care\_sample\_Dataset1 (plain text)

```

cloudera@cloudera-vm: ~
File Edit View Search Terminal Help
DeIdentifyData.class      health.pig
DeIdentifyData.java       pigsteps
DeIdentifyData$Map.class  steps.txt
deidentify script.pig
cloudera@cloudera-vm:~$ export CLASSPATH=${HADOOP_HOME}/hadoop-core-0.20.2-cdh3
u6.jar:${CLASSPATH}
cloudera@cloudera-vm:~$ export CLASSPATH=${HADOOP_HOME}/commons-codec-1.4.jar:${
{CLASSPATH}
cloudera@cloudera-vm:~$ javac DeIdentifyData.java
DeIdentifyData.java:157: warning: unmappable character for encoding UTF8
      * declared as a NullWritable when you don't need to use that po
sition; it effectively stores a constant empty value.
      ^
DeIdentifyData.java:157: warning: unmappable character for encoding UTF8
      * declared as a NullWritable when you don't need to use that po
sition; it effectively stores a constant empty value.
      ^
2 warnings
cloudera@cloudera-vm:~$ jar cvf DeIdentifyData1.jar DeIdentifyData*.class
added manifest
adding: DeIdentifyData.class(in = 2419) (out= 1266)(deflated 47%)
adding: DeIdentifyData$Map.class(in = 2798) (out= 1248)(deflated 55%)
cloudera@cloudera-vm:~$

```

Fig. 7. Result of added manifest

**B. Preliminary Data Analysis**

The dataset can be found in CSV format (Comma Separated Value). The results of this project are to encrypt simple text into encrypted data using Advanced Encryption Standard AES (AEC) algorithms. Here, we set the class path for Hadoop jar files by using the single node system. The sample data set in HDFS is shown in Figure 6. The execution process is shown in Figure 7.

**CONCLUSION**

This paper aimed at comparing the various application of data mining for useful information extraction in the health sector. Data Mining is a challenge but reduces human effort and enhances the diagnostic accuracy drastically by forecasting diseases. Efficient application cryptography tools could reduce human resources and expertise costs and time constraints. Excellent, irrelevant and massive data are also used for exploring the knowledge from medical data. In this scenario it's very useful and interesting tools for data mining in the exploration of medical information. From this study, it is observed that a combination of more than one method of data mining than one technique used for the diagnosis or prediction of medical disease can produce more promising results.

**REFERENCES**

- [1]. HianChyeKoh and Gerald Tan, –Data Mining Applications in Healthcare||, journal of Healthcare Information Management – Vol 19, No 2.
- [2]. JayanthiRanjan, –Applications of data mining techniques in pharmaceutical industry||, Journal of Theoretical and Applied Technology, (2007).
- [3]. RubanD.Canlas Jr., MSIT., MBA , – Data mining in Healthcare: Current applications and issues||.
- [4]. K. Srinivas , B. Kavitha Rani and Dr. A. Govrdhan, –Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks|| International Journal on Computer Science and Engineering (2010).
- [5]. ShwetaKharya, –Using Data Mining Techniques ForDiagnosis And Prognosis Of Cancer Disease||, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012.
- [6]. EliasLemuye, –Hiv Status Predictive Modeling Using Data Mining Technology||.

[7]. Arvind Sharma and P.C. Gupta –Predicting the Number of Blood Donors through their Age and Blood Group by using Data Mining Tool|| International Journal of Communication and Computer Technologies Volume 01 – No.6, Issue: 02 September 2012.

[8]. Arun K Punjari, –Data Mining Techniques||, Universities (India) Press Private Limited, 2006.

[9]. Margaret H.Dunham, –Data Mining Introductory and Advanced Topics||, Pearson Education (Singapore) Pte.Ltd.,India. 2005.

[10]. PrasannaDesikan, Kuo-Wei Hsu, JaideepSrivastava, –Data Mining For Healthcare Management||, 2011SIAM International Conference on Data Mining, April, 2011.

[11] D. Peter Augustine, “Leveraging Big Data Analytics and Hadoop in Developing India’s Health Care Services” International Journal of Computer Applications, Vol. 89, No. 16, 2014.

[12] Muni Kumar N, Manjula R., et al., “Role of Big Data Analytics in Rural Health Care – A Step Towards SvasthBharath “, International Journal of Computer Science and Information Technologies, Vol. 5, No. 6, pp. 7172-7178, 2014.

[13] HarshawardhanS.Bhosale, Prof. DevendraP.Gadekhar”, A Review Paper on Big Data and Hadoop”, International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014.

[14] DasariMadhav, B.V. Ramana, “De-identified Personal Health Care System Using Hadoop” International Journal of Electrical and Computer Engineering(IJECE) Vol. 5, No. 6, December 2015, pp. 1492-1499 ISSN: 2088-8708.

[15] Lidong Wang, Cheryl Ann Alexander, “Medical Applications and Healthcare Based on Cloud Computing” International Journal of Cloud Computing and Services Science (IJ\_CLOSER), vol. 2, No. 4, pp. 217- 225,2014.