ISSN: 2366-1313

# THE POWER OF BIG DATA IN PREDICTIVE ANALYTICS

[#1]RAVULA HARITHA, *Research Scholar,*

[#2]Dr. G.THIPPANNA, *Supervisor,*

[#3]Dr. NALLA SRINIVAS, Co-*Supervisor,*

*Department of Computer Science and Engineering,*

**NIILM UNIVERSITY, KAITHAL, HARYANA, INDIA.**

**ABSTRACT**: Academics and businesses alike are intrigued by how Predictive Analytics might be used to improve company intelligence and create more accurate forecasts. Because the amount of data being processed grows on a daily basis, "big data" research becomes more challenging to manage using traditional analytical procedures. A lot of jumbled information comes from various sources. As a result, organizing information is critical for getting the most out of it. To do this, cutting-edge frameworks and procedures must be developed, allowing for intelligent and automated examination of massive databases in search of useful information. This review paper summarizes the research undertaken in the aforementioned subject. This paper explains how to use the Hadoop/Map-Reduce utility to perform predictive analytics on huge datasets. By using this process, one can make judgments based on the needs of the organization within a certain domain of data.

***Keywords****: predictive analytics, data mining, apache hadoop, R, apache sparks.*

## 1. INTRODUCTION

The concept of "big data" refers to the massive amount of information generated across numerous industries, with the internet and telecommunications serving as the key sources. Reliance Jio, Facebook, Google, Twitter, Instagram, and YouTube, among others, generate potentially zetabytes of data each day. Despite the fact that big data is a relatively new notion, enterprises faced major storage issues due to the massive amounts of data generated. The primary constraints were the costs associated with data collecting and storage. This is not a big concern at the moment. Every day, the telecommunications industry, which includes Reliance Jio, processes 16,000 gigabytes of data. In contrast, Facebook users trade and view 2.77 million videos and 31.25 million communications per minute on social media. Each second, Google processes 40,000 search results. YouTube receives 300 hours of new video content per minute. Scientists, physicians, and government analysts all require massive amounts of data to conduct their investigations. A significant increase in data volume has been noted on file-sharing websites and video-sharing platforms, such as YouTube. Big data is distinguished by its large volume, diverse content, and rapid evolution. The three Vs indicate three important elements of big data. The current exponential expansion of data volume classified as "high volume" accurately describes the current deluge. Organizations have considerable hurdles in storing and interpreting such data. The majority of big data is unstructured and comes in a variety of formats, including text, binary, audio, and video. Big data is being generated at an unprecedented rate and in real time, far exceeding the capabilities of traditional software solutions designed to handle the fast flow of data output.

Because of the expanding volume of data, academics and practitioners must develop novel data processing methodologies and models in order to obtain access to Big Data's vast resources. Predictive analytics is an excellent way to extract insights from massive information. It makes data collection easier and allows for predictions of future actions and developments. When combined with statistical analysis, predictive analytics and data mining provide an effective set of approaches for knowledge exploration. The goal of this research is to look into how big data and predictive analytics may be integrated. Predictive analytics is a specialty within the larger field of

ISSN: 2366-1313

data science, which encompasses data extraction. Analytics is a scientific subject that is frequently associated with the concept of business intelligence, from which the term "analytics" derives. It relates to how companies rationalize their judgments. Predictive analytics uses mathematical and statistical tools to identify patterns in data and extract decision-supporting information from it. Predictive analytics has numerous applications in a variety of fields, including academics and industry. Predictive analytics is the use of mathematics, statistics, and probability theory to improve machine learning, data modeling, and algorithm creation in computer science. The discipline is broad and provides several long-term opportunities.

Big Data is distinguished by its exponential rise in information generation and massive volume, which can exceed zettabytes. The complexities of big data originate from a variety of factors, including the procedures involved in information gathering, storage, retrieval, analysis, sharing, and visualization. The three V's, or characteristics of big data, are listed below. Figure 1 shows the fourth V, which symbolizes the data's ambiguity and unpredictability.

• In terms of volume, existing storage and analytical methods are insufficient for managing large datasets. The data is around petabytes in size.

• Diversity: Information generated using a variety of organized and semi-structured media, such as emails and blogs.

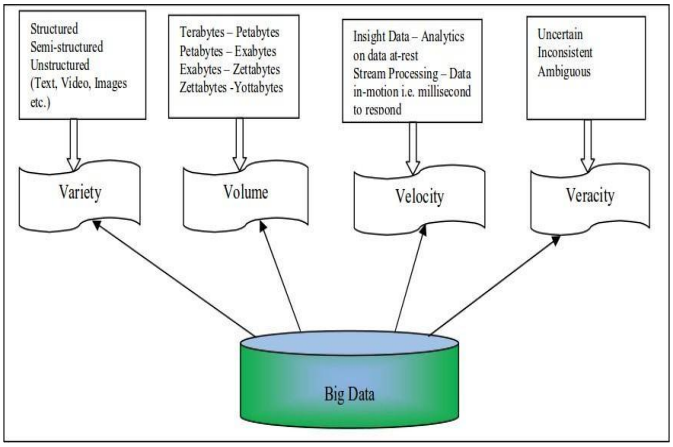


Figure 1 depicts the distinctive features of Big Data.

Examples include images, sensory data, social media posts, and usage statistics.

• Velocity: Unlike slow buildup, data is generated quickly using real-time queries to offer timely and relevant information.

• Veracity: The data contains components of ambiguity, inconsistency, and uncertainty.

## 2. ANALYSIS ALGORITHMS

The process of analyzing enormous data sets to uncover key patterns and principles is known as "data analysis." The major goal is to develop a fruitful way for merging a computer's computational skills with human visual perception in order to find patterns. Data analysis aims to effectively manage and modify large databases. Data analysis and knowledge discovery are closely related and often used interchangeably. As shown in Figure 2, data analysis entails the use of several processes to extract information from data. Data analysis is the process of collecting information from data using two different learning approaches: supervised and unsupervised learning.

**Supervised Learning:**

Supervised learning is the process of drawing conclusions from labeled training data in the context of data mining. Supervised learning divides the variables under examination into two categories: explanatory (independent) factors and one or more dependent variables. The basic goal of the study, like regression analysis, is to establish a correlation between the dependent and independent variables. To use directed data analysis approaches, you must first know the values of the dependent variables for a large section of the dataset.

**Unsupervised Learning:**

Unsupervised learning is the process of identifying hidden structures or patterns in unlabeled data. Unsupervised learning handles all variables similarly, making no distinction between dependent and independent variables.

Despite the term "undirected data mining," a defined goal must be met. This goal could range from general data reduction to more specific tasks like clustering. Discriminant analysis and cluster analysis are distinct from unsupervised learning and supervised learning. A large number of

ISSN: 2366-1313

available values and a precise characterization of the target variable are prerequisites for supervised learning. Unsupervised learning is commonly used when the goal variable is unknown or there are only a few instances of the objective variable recorded. Predictive analysis seeks to forecast probabilities and projected trends.

Throughout the estimating phase, outcomes are evaluated using a specified value. The estimation method is used to calculate the values of unknown continuous variables such as height, income, and credit card balance using the currently available input data. An association rule is one that establishes associations between a collection of elements, such as "occur together" or "one implies the other".

Data visualization is a popular way for analyzing data and providing descriptive insights. Creating meaningful visualizations can be difficult, but because people are skilled at deciphering visual signals, a correctly chosen image can express concepts more accurately than a thousand association rules. Traditional data visualization methods include Crystal Reports, SQL Server Reporting Services, and Excel-based reports. This equipment' processing capabilities were limited for massive datasets. Tableau, Domo, and Power BI are just a few of the many tools available for visualizing large amounts of data. Because of their user-friendly interfaces and high-quality graphics, these technologies allow users to experience data in novel ways. These gadgets process massive amounts of data.

### A. Clustering

Clustering is an unsupervised learning technique that does not rely on pre-defined classes. Clustering is the process of categorizing data objects into discrete clusters, with data within each cluster being comparable but data from other clusters being dissimilar. A cluster is made up of data objects that have features that distinguish them from the other objects in the cluster while also sharing some qualities. Clustering includes calculating the distance between each pair of items and determining the degree of dissimilarity between entities. The metrics include the Minkowski, Manhattan, and Euclidean distances.

This study focuses on the well-known partition-based clustering algorithm K-means. K-means is an iterative, numerical, unsupervised, and non-deterministic clustering algorithm. Within the k-means algorithm, each cluster is represented by the mean value of the elements that comprise it. To maximize inter-cluster similarity while decreasing intra-cluster similarity, divide a set of n items into k clusters. Calculating the mean value of the items in a cluster determines comparative similarity.

The algorithm consists of two discrete phases.

The second step assigns each element in the dataset to the centroid nearest to it. The initial phase involves a random selection of k centroids, where k is a predefined value.

Euclidean distance is used to calculate the distance between each data point and the cluster's centroid. This study looks at three different algorithms, including the K-means approach, that improve the system's speed and efficiency by overcoming its limits and eventually generating the optimal number of clusters.

### B. Classification and Prediction

Classification is the process of evaluating a novel entity's attributes in order to classify it to a predetermined category. The classification problem has well-defined classes, and the training set includes reclassified examples. Create a model that can classify unclassified data. This study focuses on several categorization methods, including k Nearest Neighbors, Support Vector Machine, linear regression, and random forest.

The k-Nearest Neighbor (kNN) approach is a powerful and simple non-parametric model used for classification and regression. The kNN approach has been identified as one of the top ten most influential algorithms in data mining.This study is mostly focused on classification assignments. Because of its progressive learning process, kNN requires the retention of all instances of training data. Following that, it evaluates a certain distance or similarity measure

ISSN: 2366-1313

for each unobserved situation and training example, selecting the k cases that are most similar to them. It is important to do this technique iteratively for each input sample in contrast to the entire training dataset. To categorize unlabeled features, nearest neighbor classifiers use the class of the labeled features that are the most similar to them.

Support Vector Machines (SVMs) can be thought of as boundary-separating surfaces that separate unique data points, with each data point representing an example plotted in three dimensions based on its feature values. The primary goal of a Support Vector Machine (SVM) is to generate a hyperplane, which is a linear boundary that splits the data on both sides in a fairly consistent manner. SVM learning is a hybrid approach that combines linear regression modeling and kNN instance-based learning. The synergy is extremely powerful, allowing SVMs to express complex connections effectively.

Support Vector Machines (SVMs) are highly versatile and can be used for a variety of learning tasks, such as numerical prediction and classification. The issue of algorithms is not covered in this review study.

The goal of regression analysis is to find the relationship between one or more numerical independent variables (predictors) and a single numerical dependent variable (the variable to predict). Assume that there is a linear relationship between the independent and dependent variables. Lines can be represented in slope-intercept form, as seen in the equation y = a + bx, where y is the dependent variable and x is the independent variable. In this equation, the variable b denotes the magnitude of the line's ascent with each unit increment in the x-coordinate. "a" represents the value of "y" in the absence of an attribute "x" value. The term "intercept" refers to the point where a line crosses the vertical axis. Regression equations allow data to be represented and analyzed by using a consistent slope-intercept structure.

This study focuses on the linear regression model, which employs straight lines. Linear regression has a derivative called logistic regression.

The Random Forest technique is commonly used in classification tasks involving high-dimensional data. The system has the ability to assess the value of possible predictors using its integrated variable importance metrics. This approach works well with a wide range of regression problems, including those involving nominal, metric, and survival response variables. Random forests combine robustness and adaptability to create a unified machine learning technique. Random forests can process large datasets by using a small, stochastic subset of the total feature collection. In circumstances when the "curse of dimensionality" could render conventional models ineffectual, this is especially useful. Table I summarizes the advantages and disadvantages of the previously mentioned algorithm.

The advantages and disadvantages of algorithms are outlined in Table 1.

| Algorithms | Strengths | Weaknesses |
|---|---|---|
| K-means | • K-Means gives tighter clusters than hierarchical clustering, especially if the clusters are globular.<br>• It is highly flexible and can be adapted to address nearly all of its shortcomings with simple adjustments.<br>• It is fairly efficient and performs well at dividing the data into useful clusters. | • It is difficult to predict K-Value.<br>• It is not guaranteed to find the optimal set of clusters because it choose an element of random chance.<br>• It does not work well with data of different size and different density. |
| KNN | • It effective if training data is large.<br>• Makes no assumptions about the underlying data distribution.<br>• Fast training phase.<br>• It is Robust to noisy data. | • Required to determine the value of number of nearest neighbors 'k'.<br>• Slow classification phase.<br>• Requires a large amount of memory.<br>• Computation cast is high. |
| SVM | • It can be used for numeric prediction or classification problems.<br>• It is not influenced by noisy data and not very prone to overfitting.<br>• maximizes margin, so the model is slightly more robust compare to linear regression.<br>• It supports kernels, so you can model even non-linear relations. | • It is difficult to finding the best model requires testing of various combinations of kernels and model parameters.<br>• It is slow in training and testing.<br>• SVM have high algorithmic complexity and huge memory requirements of the required quadratic programming in large-scale tasks. |
| Random Forest | • It is one of the most accurate learning algorithms available. It produces a highly accurate, classifier,for many data sets.<br>• It runs efficiently on large databases.<br>• It also offers an experimental method for detecting variable interactions. | • The model is not easily interpretable.<br>• It has been observed to overfit for some datasets with noisy classification/regression tasks. |

All the about analytics technique are available in almost most of the programming language like R, Python etc. R is an amazing data science programming tool to run statistical data analysis on models and translating the results of analysis

ISSN: 2366-1313

into colorful graphics. There is no doubtthat R is the most preferred programming tool for statisticians, data scientists, data analysts and data architects but it drops short when working with large datasets. One major disadvantage with R programming language is that all objects are loaded into the main memory of a single machine. Large datasets of size petabytes cannot be loaded into the primary memory; this is when Hadoop integrated with R language is an ideal solution. To adapt to the in-memory, single machine limitation of R programming language, data analyst have to limit their data analysis to a sample of data from the large data set. This limitation of R programming language comes as a major barrier when dealing with big data. As we know that, R is not very scalable, the core R engine can process only limited amount of data.

To the contrary, distributed computing frameworks like Hadoop are scalable for complex operations and tasks on large datasets (petabyte range), Apache Spark etc.

## 3. TOOLS AND TECHNIQUE

An range of analytics approaches are available for a variety of programming languages, including Python, R, and others. R is a powerful data science programming language that makes it easy to visualize research results and conduct statistical data analysis on models. R is definitely the preferred programming language for architects, data scientists, statisticians, and analysts. When dealing with large datasets, however, its speed is limited. One notable disadvantage of the R programming language is that each object is kept in the main memory of a single computer. When working with petabyte-sized datasets, putting them into main memory is not feasible. Hadoop integration with the R programming language offers a suitable answer in such cases. To avoid the R programming language's in-memory, single-machine constraint, data analysts must restrict their study to a subset of the large data set. Working with large datasets is particularly tough because to the R programming language's constraints. It is commonly accepted that R's scalability is limited since the underlying R engine can only process a finite quantity of data at once.

Distributed computing systems, such as Apache Spark and Hadoop, can handle petabyte-scale datasets and sophisticated computational operations.

Apache Hadoop is a widely recognized framework for distributed computing. It facilitates the decentralized processing of large datasets over multiple computer clusters by utilizing simple programming approaches. The system is designed to grow organically from a single server to a large number of devices capable of processing and storing data on their own.

To achieve high availability, the library is specifically designed to detect and address issues at the application layer, rather than relying on hardware. This allows a collection of computers to provide a highly available service despite the failure of a single system.

Apache Spark is a framework intended for clustered computing. It is a versatile and speedy platform that can handle enormous amounts of data processing. Spark beats Hadoop by a factor of ten, speeding up the development of iterative machine learning systems. This article does not cover how to use Spark and Apache Hadoop for machine learning-based big data analysis. The primary focus is on how these frameworks can be used to analyze Big Data.

### A. Classification and Prediction

Apache Hadoop is an open source framework that allows for the processing and analysis of massive amounts of data. Hadoop permits the decentralized processing of massive amounts of data across vast networks of low-cost machines. Hadoop has piqued the interest of nearly all academics and business experts because to its multiple benefits and distinguishing features.

ISSN: 2366-1313

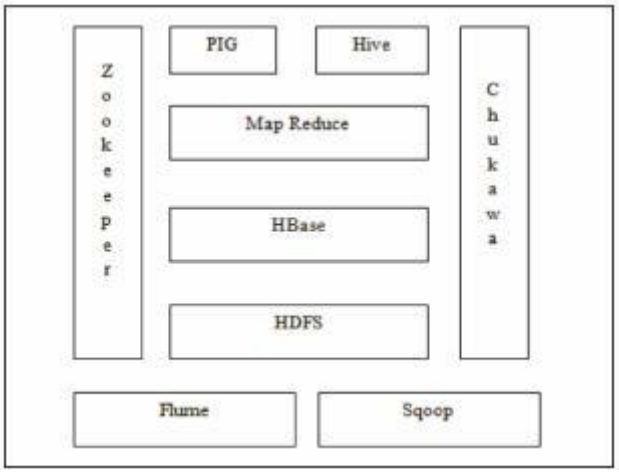


Figure 2 illustrates the structure of the Hadoop software library.

Hadoop's broad computational library includes a large number of modules, such as Map Reduce, HBase, HDFS, Hive, and Pig. The following section discusses the various Hadoop architecture components seen in Figure 2.

Sqoop and Apache Flumes are two examples of data integration tools that are used for data gathering. Flume and Sqoop's primary job is to collect data from several sources and index it in a central store.

HDFS, or Hadoop Distributed File System, operates on low-cost hardware and is based on the Google File System (GFS). HDFS consists of a single Name Node and multiple Data Nodes. The Name Node manages the file system's metadata. The storage of physical data is the responsibility of data nodes.

HBase and Google Big Table are columnar databases with similar functionalities. Hbase can manage Hadoop Map-Reduce's input and output.

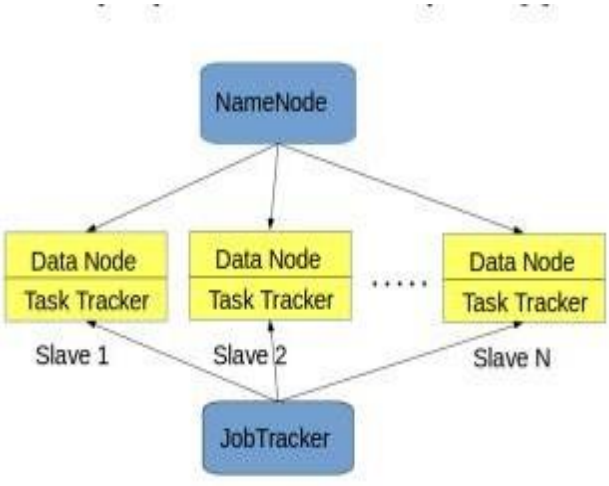


Figure 3 depicts a master/slave map reduction model.

MapReduce is a powerful computational framework designed for large-scale data analysis. Figure 3 shows that each Map-Reduce cluster node has a master Job Tracker and a subordinate Task Tracker. The Master is largely responsible for assigning captive labor. It is also in charge of overseeing tasks and completing those that have proven unsuccessful. Task fulfillment is the subordinate's responsibility, as delegated by the master.

SQL is similar to high-level declarative languages like Pig Latin and Hive. Pig Latin simplifies data flow procedures, whereas Hive streamlines ad hoc data retrieval and summarizing.

Zookeeper and Chukwa are critical components for managing and controlling distributed applications built on the Hadoop architecture.

## 4. LITERATURE REVIEW

Ng, Ghoting, Kenny, and others have campaigned for the use of predictive analytics in the healthcare industry since the introduction of electronic health records, which permitted the quick accumulation of massive amounts of different data. They proposed developing a prognostic model for the examination of medical data as a line of action. The researchers developed a technology called parallel predictive modeling (PARAMO). It works in parallel by grouping the assignments into a graph using topological sorting and creating a graph to represent the interdependencies between them.

The implementation of Hadoop Map Reduce technology enabled parallel execution. The platform's performance was then examined using multiple EHR datasets, and the results showed a considerable improvement in the system's computing power. This study demonstrated the effectiveness of using predictive modeling approaches to streamline and speed labor processes in the healthcare sector.

S. G. Manikandan and S. Ravi created the Hadoop

ISSN: 2366-1313

architecture as a method of managing huge volumes of data, recognizing that traditional data management, warehousing, and analysis systems lack the tools required for big data analysis. A better understanding of the market and customers may enable firms to use Map Reduce on Hadoop and HDFS, allowing for more informed business decisions and gaining a competitive advantage. This article examines the steps needed in configuring a Map-Reduce task to handle massive amounts of data.

A. Jalanila and N. Subramanian evaluated the SAS® Text Miner, Python, and R programming tools for performance, usability, and visualization capabilities. SAS® Text Miner uses predictive modeling to identify trends in text data. R and Python, both open-source programming languages, are used to analyze and understand data. The author compared these three tools using the Random Forest (RF) and Support Vector Machines (SVM) models, which are both available in all three tools.To start with.

Table II summarizes the algorithms' merits and weaknesses.

| | Python | R | SAS® Text Miner |
|---|---|---|---|
| SVM | 0.62 | 0.633 | 0.53 |
| RF | 0.6 | 0.619 | 0.64 |
| User expertise | Experienced | Intermediate | Beginner |
| Ease of use | Fair | Fair | Good |
| | | | |
| Performance | Good | Fair | Good |
| Visualizations | Fair | Fair | Good |

Table 2 depicts a graphical comparison of the important criteria. The RF model performed similarly across all three instruments. In comparison to Python or R, the SAS® Text Miner implementation of the SVM model had lower accuracy. Analysts' levels of skill with SAS® Text Miner, R, and Python range from seasoned to intermediate to rookie.

X. Wu et al. provide the HACE theorem, a framework that defines the characteristics of the Big Data revolution and proposes a strategy for managing Big Data through data mining. According to the HACE theorem, big data has three fundamental characteristics: 1) its size and diversity of origins; 2) its autonomous operation under decentralized and distributed control; and 3) its complexity and dynamic nature in terms of the data it contains and the interconnections between it. In order to analyze Big Data, the author examined many challenges at the data, model, and system levels. The use of high-performance computing systems that are properly built to maximize the benefits of Big Data is crucial for Big Data mining.

Complex data circumstances, such as unknown or missing values, typically arise at the data level as a result of various data collecting contexts and the availability of multiple information sources. Furthermore, the author elaborated on research endeavors and large-scale data extraction projects. For example, the author noted the Big Data program of the US National Science Foundation (NSF), which began in 2012 and included the introduction of the BIGDATA proposal during President Obama's administration. A federal program can be credited with sparking various beneficial projects researching the foundations of Big Data management. The author also emphasized the importance of distributed frameworks such as Spark and Apache Hadoop for huge data mining.

Zongheng Yang, Shivaram Venkataraman, and colleagues developed a R interface for Apache Spark that allows users to perform detailed data analysis with SparkR, Spark's distributed compute engine. Numerous additions to the well-known statistical programming language R make machine learning and data processing more efficient. The R runtime is limited to a single thread and can only handle data sets that can be stored in a single processor's RAM. These limits make it difficult to undertake interactive data analysis with R. The author introduces the SparkR R package, which

ISSN: 2366-1313

serves as an interface to Apache Spark. SparkR uses Spark's distributed compute engine to enable the analysis of enormous volumes of data directly from the R shell.

Ping Sun and colleagues developed the breakthrough data mining approach RFDM (RHadoop-based Fuzzy Data Mining), which accelerates the fuzzy data mining process and lowers costs. A Hadoop-based architecture was created to fulfill the demands of data extraction from massive datasets. RFDM supports a wide range of algorithms, including Apriori, KNN, SVM, Neural Network, Bayesian Network, K-means, Density-based Spatial Clustering of Applications with Noise, and many more.

## 5 . CONCLUSIONS AND FUTURE WORK

This article covered some predictive analytics approaches. Because of the exponential development in the volume of data handled on a daily basis, big data analysis necessitates a different technique than traditional analytics. However, the full potential of big data is yet to be realized. We investigated how individuals in various big data sectors analyze huge amounts of data.

We may make the most of large data by deploying distributed frameworks like Hadoop and Spark, which are smart and scalable tools. Spark and Apache Hadoop will let predictive analytics reach their full potential.

## REFERENCES


[1] R Bender and U Grouven. Ordinal logistic regression in medical research. 1997.

[2] Harshawardhan S Bhosale and Devendra P Gadekar. A Review Paper on Big Data and Hadoop. International Journal of Scientific and Research Publications, 4(1):2250–3153, 2014.

[3] Http://hadoop.apache.org/. Welcome to Apache Hadoop.

[4] ArunJalanila and Nirmal Subramanian. Comparing SAS® Text Miner,Python, R: Analysis on Random Forest and SVM Models for Text Mining. 2016 IEEE International Conference on Healthcare Informatics (ICHI), pages 316–316, 2016.

[5] JyotiNandimath, Ekata Banerjee, AnkurPatil, PratimaKakade, SaumitraVaidya, and DivyanshChaturvedi. Big data analysis using Apache Hadoop. 2013 IEEE 14th International Conference on Information Reuse & Integration (IRI), pages 700–703, 2013.

[6] Kenney Ng, AmolGhoting, Steven R. Steinhubl, Walter F. Stewart, Bradley Malin, and Jimeng Sun. PARAMO: A PARAllel predictive MOdeling platform for healthcare analytic research using electronic health records. Journal of Biomedical Informatics, 48:160–170, 2014.

[7] Anand V Saurkar, VaibhavBhujade, and PritiBhagat. A Review Paper on Various Data Mining Techniques. 4(4):98–101, 2014.

[8] Ping Sun, Lei Xu, and Hongfei Fan. RHadoop-based fuzzy data mining:Architecture, design and system implementation. 2016.

[9] ShivaramVenkataraman, Zongheng Yang, Davies Liu, Eric Liang, XiangruiMeng, ReynoldXin, Ali Ghodsi, Michael Franklin, Ion Stoica,andMateiZaharia. SparkR: Scaling R Programs with Spark. Sigmod, page 4, 2016.

[10] Xindong Wu, Xingquan Zhu, Gong Qing Wu, and Wei Ding. Data mining with big data. IEEE Transactions on Knowledge and Data Engineering, 26(1):97–107, 2014.

[11] JyotiYadav and Monika Sharma. A Review of K-mean Algorithm. International Journal of Engineering Trends and Technology, 4(7):2972–2976, 2013.

ISSN: 2366-1313