International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 1 | ISSN : 2456-3307

The Study on Predictive Analysis Algorithm : Survey

Anandajayam P*[1], Dr. N. Sivakumar[2]
*[1]Research Scholar, Department of Computer Science and Engineering/ Pondicherry Engineering College, Tamil Nadu, India
[2]Assistant Professor, Department of Computer Science and Engineering/ Pondicherry Engineering College, Tamil Nadu, India

ABSTRACT

Nowadays the increase of data variety considered very controversy problem. So inventive methods are mandatory for analytics especially in big data where the data are very complex, structured, unstructured and semi structured. It is owing to a good deal of research which is carried out in Predictive, Prescriptive, Diagnostic, and Descriptive. Because of the increase in the huge volume of data this paper helps the researcher in analysing the prediction. Machine learning is one of the materialize ways to fabricate the analytic model for machines to learn from data and able to do analysis on prediction. The cue "big data analytics" can be simplified by the subsequent four manners: data, problem, methodology, and technology. In this paper, we discuss the study of predictive analytics. Predictive analytics is a prerequisite approach that handles the necessary quantum of potentially fragile data to predict the future possibilities, trends, and measures. Predictive analytics are composed of various mathematical and meticulous methods used to produce a new technique to predict future possibilities. This paper, scrutinizes about various predictive analytics algorithms with for and against in big data. The predictive algorithms have been explained in upcoming parts.

Keywords : Big Data Analytics, Predictive Analysis, Machine Learning Algorithm

I. INTRODUCTION

Big data is a group of abundant and complex data that describes based on structured, semi structured and unstructured data. Three dimensions of big data are volume, velocity, and variety. The challenges include capturing data, creating data, managing data, analysing and processing of data, sharing and transfer of data. Nowadays there are many kinds of data as being generated for each and every particular surrounding where they use various forms of data for some peculiar reasons.

Big data is categorized into two i) Storage and ii) Processing. But in this paper we choose Processing because the main reason for such popularity is the ease of use, scalability and failover properties and much research has been in processing because its process numerous data at the same time without any data collapse and processing is done by categorizing data based on its type such as structured, semi-structured and unstructured. In big data, the data are being stored and been processed using some tools such as Apache Hadoop, Apache Spark and Apache Hive ,which is a distributed computing framework modelled after Google MapReduce to process a large amount of data in parallel. Big data analytics are a process of collecting, organizing, analysing the management of the data for producing new information for the end user. Big data analytics has four operations: (i) acquisition (ii) assembly (iii) analyse and (iv) action [7] as it is largely involved in collecting data from different sources, manipulate it in a way that it becomes available to be consumed by analysts and in that prediction of data is done

because all data that are being stored are being predicted using various algorithms. It is a technique, tools, and technologies that use the date to find models that can anticipate outcome with a significant probability of accuracy. It gives the basis of predictive analysis based on the availability of high-quality data and effective sharing and it is based on i) Data collection ii) Deployment and iii) Statistics [1].In prediction, the model used is machine learning [2] because it is a subfield of computer science that deals with tasks such as pattern recognition, text analytics, and mathematical optimization. It is divided into three group of tasks they are i) Supervised Learning and ii) Unsupervised Learning, and iii)semi-supervised ,moving to supervised learning refers to a type of problem where there is an input data defined as a matrix and identities that have high probability densities with respect to individual classes[2]. It is better than unsupervised learning because the dissimilarity measure and fact such a method is not taking into task being solved and the distance metric is not adaptive.

The organization of this document is as follows: In Section 2: The overview of Big data and Predictive Analytics, Section 3: Literature survey, Section 4: Describes various predictive algorithms, Section 5: Conclusion of this paper.

II. OVERVIEW

Predictive modelling contain with statistical techniques from predictive analytics, data mining and machine learning which are used to analyse the present data and traditional historical data, facts to find the prognosis about the future the predictive analytics are used to identify risks and opportunities. A predictive model is the combination of data and mathematical process which helps to predict unobserved or unknown events. We overload with different type of method to convey predictive modelling process through clustering, decision tree, linear and logistic regression and SVM. The core based technique is regression analysis, which predicts the related values of single and multiple, correlated values based on manifest or contradict a particular assumption.

The studied approaches for predicting models include semi parametric regression, time series modelling exponential smoothing Bayesian statistics time-varying spline decomposition techniques transfer functions gray dynamic models and judgmental predicting. These predicts are usually characterized by their time horizon: (i) short-term predict for ensuring system stability,(ii) medium-term predict for maintenance scheduling, and (iii) long-term predict for network planning[3]. While a point predict provides an estimate of an expected value of the future demand, probabilistic previse contain additional valuable information. Having access to prediction intervals, such as a lower and upper boundary of the values of datasets distribution or a prediction density, would inform the decision maker of the uncertainty inherent in the prognosticating. Quantification of this predicts uncertainty is essential for managing the risk associated with decision making. Probabilistic methods which are able to capture the various factors that govern the electricity demand and predict its peak. These models are strongly dependent on the availability of historical data and need a large dataset to produce accurate results. By using such a regression model, Lei and Hu[4] showed that a single variable linear model is able to predict the energy consumption in hot and cold weather conditions. Finding concealed experiences and example with help of information strategies and apply watched example to questions in the past, present and future.

A. Predictive Analytics

Prediction is the process of defining real-time regression analysis and machine learning analysis to predict future measures using some techniques, tools, and technologies in a defined manner so that our future work will be processed on that prediction. Predictive analytics belongs to the family of big data that accord with extracting information from data

and using it to predict trends and behaviour patterns with some facts and evidence for future [5]. Prediction is used to make inference about the present and to realize our past and to check about our future. It is used in wide area of research in different opinion that provides new mechanisms and many approaches that can be applied in any environment. The types of predictive model are i) Classification, ii) Regression, iii) Clustering, iv) Collaborative Filtering, v)Dimensionally Reduction, vi) Optimization Primitive A predictive model is a function that takes input variables applies a formula to predict the outcome.

$$Y = f(x) \quad (1)$$

It is described as follows:

Y- Output as predicted.

f - Prediction Function.

x- Input for prediction.

III. LITERATURE SURVEY

In recent years it has been found that there is the increase in rate possibilities for big data in retail. "The Role of Big Data and Predictive Analytics in Retailing ", Researchers are working towards this field to examines the opportunities in and possibilities arising from big data in retailing identify. It is based on computational approaches that analyze and Methodology of Bayesian analysis techniques data borrowing, updating, augmentation and hierarchical modeling to a smart application of statistical tools and domain knowledge combined with theoretical insights. Historical theory-agnostic predictive analytics tools are likely to have the larger impact and lesser bias. The main part of the role of big data and predictive analytics in a retail context is set to rise in importance, aided by newer sources of data and large-scale correlation techniques that of theory, domain knowledge, and smart application of extant statistical tools are likely to continue undiminished. The remakes are with the Improve data quality, less data compression and transformation [35].

The design, application and evaluation of a predictive scheduling framework aiming at fast and distributed stream data processing." Performance Modeling and Predictive Scheduling for Distributed Stream Data Processing", Researchers are working in the distributed stream data processing system in which features the perspective analysis and method as topology-aware modeling for performance prediction and predictive scheduling. It is presently an effective algorithm to assign threads to machines under the guidance of prediction results. Merits and demerits of performance predictive analytics enhance the topology-aware the prediction method offers an average accuracy and the average processing time is reduced when compared to the Storm's default scheduler[36].

Data mining methods for exploring the main process is to collect, extract and store the valuable information is extracted. The two main objectives of predictive analytics are Regression and Classification. The probabilities of events Predictive analytics is the roof of advanced analytics, which is to predict the future events. They analyze and Methodology of Data mining technique, analytical and statistical techniques. This approach outperforms more conventional data mining methods in terms used in predictive analytics for modeling and forecasting is based on presented inspection focus towards the predictive analytics, regression techniques and forecasting in knowledge discovery domain. Remakes efficient in choosing marketing methods and Helpful in social media analytics [37].

The presented predictive analytics mainly focuses on opportunities, applications, trends & challenges in Knowledge discovery domain. Proposed an efficient imputation method using a Classification & Regression is the two main objectives of predictive analytics. The model is built by data mining tools and techniques is used to determine the probable future outcome of an event or the likelihood of a situation occurring. Predictive analytics are composed of

various statistics, analytical techniques and Data-mining techniques used to analyze characteristic improving models. Remarks are the major problem associated with scaling algorithms is that Challenges of Predictive Analytics in Knowledge discovery domain [38].

Mainly the research of the work is move to highlight the big data issues and challenges faced by the Dynamic Energy Management employed in SG networks it brief description of the most commonly used data processing methods in the literature, and proposes a promising direction for future research in the field. It is based on computational approaches that analyze and techniques of Dynamic energy management (DEM) and high performance computing (HPC) used to data analysis modeling. Mainly three part described algorithm i) Design and development of algorithms, ii) Design of machine learning (ML) iii) Development of novel data-aware. Merits are Improve economic efficiency, reliability, and sustainability, high performance, insisting on cost efficiency and security and Real-time monitoring and forecasting system. Demerit are processing time takes more [39].

G. Kumaresan discussed about the main applications which depends thoroughly on big data, predictive analytics is the process of analysis to predict the concealed pattern and association among data. Its dynamic nature of reviewing the effects of design features, the objective is to provide the exhaustive view of different predictive analytics applications and approaches. Methods focused with dissimilar perspectives based on applications and data variety used Statistical and analytical techniques. The implications will be based on pattern predictions and different evolutionary techniques. Using temporal pattern prediction is improved for consumer and when they will make online purchase again. Remarks are Privacy of Data Analysis of User Data and Scaling of User Data [40].

“Tobias Schoenherr and Cheri Speier-Pero”, deliberate about how to train our next-generation data scientists and experiences in developing and implementing one of the first MS degree programs in predictive analytics. They discuss the results of a recent large-scale survey the minds of many supply chain management (SCM) professionals, complemented by our experiences in developing, implementing, and administering it should be provided effectively an assessment of the current state of the field processing. ."Data Science, Predictive Analytics, and Big Data in Supply Chain Management: Current State and Future Potential". It uses to analyze and Methodology of Supply Chain Management (SCM). The statistics indicated that these differences were not statistically significant merits are Recognized essential advantages and significant obstructions to SCM prescient examination and Offering knowledge into the future capability of information science, prescient examination, and enormous information in SCM. Research of the further infuses curricula with predictive analytics components [41].

Big Data proposed and specifically identified design to explain the different Technological advancement enables supervised machine learning algorithms and unsupervised machine learning algorithm like deep neural networks, support vector machines, decision trees, naive Bayes algorithm, support vector machine (SVM), K-means, Linear Regression, Logistic Regression and random forest algorithm to detect the intrusions in the synchrophasor data set it classified into Normal, Attack and disturbance operations. The objective of the proposed work and this is achieved by applying the feature selection and dimensionality reduction method are Deep Neural Networks(DNN), Support Vector Machines(SVM), Random Forest(RF), Decision Trees(DT) and its processing another methodology has to be Apache spark the processing time. They mainly discuss designing of hybrid intrusion detection system is to increase the chance of detecting and classify the attacks into the category. Remake are Reduce the prediction time taken by the proposed and highest accuracy for the raw dataset [43].

The technology mainly based on the objective of machine learning is to discover knowledge and make intelligent decision methods and technology progress of machine learning in Big Data are also presented. The machine learning system includes classification, regression, clustering, and density estimation and it has been many different approaches include decision tree learning, association rule learning, artificial neural networks, support vector machines (SVM), Clustering, Bayesian networks, and genetic algorithms more explanation and Challenges of machine learning applications in Big Data are discussed. The Methodology is the various parts discuss topic Machine learning algorithms, supervised clustering, unsupervised clustering, and Semi-supervised clustering. Remakes are obtained Good performance and Moderate size of data sets [44].

"Hina Gulati", discussed about analysis of the data set using data mining algorithms Predicting student's dropout reasons can be difficult task due to multiple factors that can affect the decision. Furthermore the absolute technique is long and time consuming it also gathered data is from different sources. Researching on many problems in education, analyzing student data and deriving useful knowledge is known as Educational Data Mining using methodology as data mining and Pre-processing. After preparing the data for mining, classification algorithms are applied and by analysis of the decision tree and induction rule methods executed merits has most effective way to analyze student performance and help in identifying reasons for drop-out and demerits takes long and time consuming[45].

IV. DIFFERENT PREDICTIVE ALGORITHM

A. Decision Trees

Decision Tree is a supervised learning algorithm method for inductive research over data. Decision trees are minimized powerful form of multiple variable analysis and Classification model. The Decision trees are handles both categorical data and numerical data and it is represented as a Graphical tree structure Decision tree algorithm like ID3, C4.5, J48, CART[9][11] The input and output variable are homogeneous set to split the samples. They provide different function based capabilities to complement, supplement and substitute for (i)Traditional statistical form of analysis such as multiple linear regression.(ii) A variety of data mining tools and techniques (such as neural networks). (iii) Recently developed multidimensional forms of reporting and analysis. There are two types of decision trees are binary variable decision tree and continuous variable decision tree. The binary variable has a binary target variable and the continuous variable has a continuous target variable. Handling processing large data set approaches such as parallel, distributed, scalable and Meta decision tree and Other issues Optimization of Decision Tree learning like incremental induction of decision tree and oblique decision trees [10].There is some terminology related to the decision to decision trees are the root node, splitting, decision node, terminal node, pruning and child node. These are the powerful prediction method to solve the problem.

Advantage:

- ✓ Decision trees are capable of handling errors and missing value in the datasets.
- ✓ Decision trees are considered to be a nonparametric method. This means that decision trees have no assumptions about the space distribution and the classifier structure.

Disadvantage:

- ✓ nsufficient in applying value to predicts continues value.
- ✓ The greedy characteristic of decision trees leads to another disadvantage that should be pointed out.

B. Logistic Regression

This algorithm is used to predict the outcome of events or facts that are not continuous in process.

This algorithm applies a logistic function to a linear combination of features to predict the outcome of a categorical dependent variable based on predictor variables. It helps to estimate the possibility of falling into a specific level of the categorical dependent variable based on the given predictor variables [12]. Based on the nature of response, logistic regression is stratified into 3 types:

B.1. Binary Logistic Regression

Binary logistic regression is for the specific case when the response variable has only two possible values: yes or no[19].The data analytics to predict the probability and characterizes of the statistics desired outcome, where OLS regression was applied to data with a binary dependent variable described quasi-formally Predicted logit of

$$y = [\beta 1\chi 1 + \beta 2\chi 2 + \ \dots + €] \qquad (2)$$

Where β is a logistic coefficient, X is a variable. It is prime fit for binary classification (as 0 or 1), Binomial or binary logistic regression refers to the instance in which the observed outcome can have only two possible type. To explore the relative influence of Sequence function and categorical independent variables on your dependent variable and to assess interaction effects between the independent variables [21].

B.2.Multi-nominal Logistic Regression

Multinomial logistic regression model is an extension of the binomial logistic regression model. Multinomial Logistic regression refers to cases where the outcome can have three or more possible types and it is most straightforward interpretation

$$\Pr(Y_i = K) = 1 - \sum_{K=1}^{K-1} \Pr(Y_i = K)\, e^{\beta_{k.x_i}} \ \rightarrow \Pr(Y_i = K) = \frac{1}{1+\sum_{K=1}^{K-1} e^{\beta_{k.x_i}}} \qquad (3)$$

• Logit Function for Y= 0 relative to logit function Y= 2

• Logit Function for Y= 1 relative to logit function Y= 2 where Y is a reference group, Multinomial logistic regression has also been used in the study of examining factors affecting family economic status among families in Kerman city in Iran which has indicated that some of the variables are stronger than others as a predictor to family economic status [22]. This performance of the model depends upon the fact of classification accuracy.

B.3. Ordinal Logistic Regression

The OLS method which is commonly used to predict dependent variable based on terms of one or more independent variables. it is extended the technique of a multiple logistic regression analysis to research situation where the outcome variable is categorical thereby modeling the survival of infancy. There are various ordinal logit models to compare dependent variable categories (i) Proportional Odds Model (POM), (ii) Non-Proportional Odds Model (NPOM) (iii) Partial Proportional Odds Model (PPOM)[23]. The logistic function is

$$h(x) = 1\ (1 + {}^{-}ex) \qquad (4)$$

Applications:

It is used in real-world such as:

- ✓ Credit Scoring
- ✓ Measuring the success rates of marketing campaigns
- ✓ Predicting the revenues of a certain product
- ✓ Is there going to be an earthquake on a particular day?
- ✓ Advantages:
- ✓ Manage perplex and tests interaction.
- ✓ Easier to inspect and less complex.
- ✓ Durable algorithm as the independent variables need not have equal variance or normal distribution.

Disadvantages:

- ✓ When the training data is scarce and high dimensional, in such situations a logistic model may overfit the training data.
- ✓ Logistic regression algorithms cannot predict continuous outcomes.

For instance, logistic regression cannot be applied when the goal is to determine how heavily it will rain because the scale of measuring rainfall is continuous. Data scientists can predict heavy or low rainfall but this would make some compromises with the precision of the dataset. • Logistic regression algorithms require more data to achieve stability and meaningful results. These algorithms require minimum of 50 data points per predictor to achieve stable outcomes.

C. Linear Regression

This algorithm is used to predict the outcome for continuous process in the set of input variables (a) that are used to determine the output variable (b). A relationship exists between the input variables and the output variable [25]. The goal of ML is to quantify this relationship. In Linear Regression, the relationship between the input variables (a) and output variable (b) is expressed as an equation of the form

$$b = xa + y \quad (5)$$

a is the input b is the outcome x is the intercept y is the slope of the line

Applications:

It is used in real-world such as:

- ✓ Financial services
- ✓ Estimating sales and marketing
- ✓ Used in field of environmental science Advantages:
- ✓ It is one of the easiest machine learning algorithms to explain to others.
- ✓ It is easy of use as it requires less tuning.
- ✓ It is the mostly widely used technique that runs fast.

Disadvantages:

- ✓ Linear Regression Is Sensitive to Outliers
- ✓ Data Must Be Independent
- ✓ It cannot be used to predict non continuous outcome.

D. Support Vector Machine

Support Vector Machines is a supervised machine learning algorithm is one of best machine learning algorithms, which was proposed in 1990s and is on the concept of decision planes that define decision boundaries and also used for both classification problems and regression challenges. It is one of the most dominant classifications of the algorithm to solve the classification problem. A decision plane is one that separates the class after the input data between a set of objects having different class memberships. Then, we perform classification by finding the hyper plane that differentiates the two classes very well. It built the model to predict classes. There are two types of classes are linear SVM classifier and non-linear SVM classifier [13].

D.1. Linear SVM Classifier

In this linear model the data are a point to separate a gap. It is used to predict the straight hyper plane which is dividing into two classes. It points the hyper plane to maximize the distance from hyper plane to the nearest data point. So, it is called as maximum margin hyperplane.It finds the decision boundary that maximizes the margin of the position from the negative training datasets. It finds the decision boundary that maximizes the margin of the position from the negative training datasets [14].

D.2. Non-Linear SVM classifier

In the data world, the data is spread up to extend. To solve this problem separation of data into different classes on basis of the straight linear hyper plane. In this type, it applies the kernel trick to maximum-margin hyper planes. This function is used to build a high dimensional feature space [15].

D.3. Regression SVM

$y = f(x) + noise$, in this where f is the function is the input variable to find the functional form for f to predict the new cases in the SVM that which has not presented before.

Applications:

It is used in real-world such as:

- ✓ Used in Face detector

- ✓ Used to categorization of text and hypertext
- ✓ Used in field of Bioinformatics
- ✓ Used as a Remote Homology detection Advantage:
- ✓ It is Robust and meticulous model to solve the prediction problem. It uses the kernel trick to build in expert knowledge about the problem in the kernel. SVM is defined by a convex optimization problem for which there are efficient methods.

Disadvantage:

- ✓ It expresses the biggest limitation of the support vector approach lies in the kernel.
- ✓ The second limitation is speed and size, both in training and testing requires more memory.

In a way the SVM moves the problem of over fitting from optimizing the parameters to model selection. Sadly kernel models can be quite sensitive to over fitting the model selection criterion. It is sensitive to noise.

E. Random forest Algorithm

Random forest algorithm is the supervised machine learning algorithm. This algorithm creates forest with the number of trees and creates high accuracy. It consists of many decision tree and output by individual tree. It used divide and conquer method to improve performance. The random forest starts with the technique of decision tree. In the decision tree input is entered as traverses down the tree and gets the data into smaller sets [16]. This algorithm is used for both classification and regression to solve the problem. Each tree is constructed using the following algorithm [18]. Let the number of cases be N and the number of variables in the classifier be M. The number of m input variables is used to determine the decision at a node of the tree and m should be much less than M. Choose set of tree by choosing n times with replacement from all N available cases. Then Use the rest of the cases to estimate the error of the tree, by predicting their classes. For each node it chooses randomly m variables and Calculate the best split based on m variables in the data set. The relationship between the two trees in the forest increase the forest error rate [17].A tree with the lower error rate is called strong classifier. It increases the strength of individual trees and decreases the forest error rate and reduces both the relationship and strength between the optimum ranges of m.

Advantage:

- ✓ It can handle thousand of input variable without variable deletion
- ✓ It can produce the highly accurate classifier.
- ✓ It achieves maximum productivity on large datasets
- ✓ It avoid over fitting problem

Disadvantage:

- ✓ The classification of random forest is difficult for human to interrupt
- ✓ It observed to over fit some dataset with noisy classification task.

F. Spectral Clustering

The spectral clustering algorithm is a delightful method of the algorithm in which the data is let to analyze its spectrum by constructing an affinity matrix using the given data to get a clear clustering from the eigenvector values. Spectral clustering is the most widely used technique for data analytics and it implements easily, solved efficiently by standard linear algebra software [20]. It is more powerful and specialized clustering algorithm. Spectral Clustering has been large-scale used in many areas, As a typical unsupervised learning method, many clustering applications can be found in fields including in the statistics, machine learning, pattern recognition, data mining, and image processing[26]. It is status attractive and challenging to deal with clustering problems. (1) The data sets contain a large number data; (2) Clusters are of widely differing size and shapes; (3) The data dimensionality is very high [27].

Applications: It is used in real time application as follows:

- ✓ Segmenting the given image based on color, texture, pixel size.
- ✓ Separation of speech from noise. Advantages:
- ✓ High in performance yield
- ✓ Ability to group non vector data
- ✓ Good in time and space complexity
- ✓ Implementation is done easily

Disadvantage:

- ✓ Crunching cost for large datasets
- ✓ Fragile in choosing parameters

G. Principal component Analysis

It is a strategy that uses an impertinent variation to convert a set of statement of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components [25].

Applications:

- Used in finance
- Used in agriculture
- Used in plant breeding
- Used to change detection

Advantages:

- It analyzes the data with lusty tools.
- The data are found and squashed.
- It can reduce the amount of dimension without losing data.

Disadvantage:

- The result is difficult to expound because of all linear combination of variables

H. Streaming Clustering

The term data stream is the capacity to develop fast sequence of information. The concept of data streaming is more suitable than a datasets model to access large amount of data set stored in secondary memory where performance required linear. (i) It comprises of ongoing flow of huge data sets. (ii) It rapidly produces data that occurs in real time with speed counter requirements. (iii) Miscellaneous access to the data stream is impossible to process it and is able to access the data once.(iv) Storage of the data stream is restricted so only a summary of the data can be saved to find the crucial data is a challenging task and(v) it is multidimensional so algorithms are required to mine streaming data[29].The methods of data streaming clustering are Hierarchical methods, Partitioning methods, Grid-based methods, Density-based methods, Model-based methods which are described below [30].

H.1.Hierarchicalmethods:

Clustering techniques in hierarchical, which can be divided in two methods namely heap or cluster and divisive. It merges a set of n objects into general categories and divides n objects into smaller clusters in order. However in hierarchical agglomerative clustering (HAC) is more used frequent method with the option of manually determining the number of clusters [28]. Online divisive agglomerative clustering (ODAC) is a time series data stream clustering technique used to handle concept of both heap and divisive hierarchical methods.

H.2.Partitioning methods:

The partitioning techniques such as k-median and k-means are the data stream clustering. The k-median-based clustering algorithm is the Stream L Search algorithm which has been proposed for clustering high quality data streams. It is part of two sequences starting with the determination of sample size by the STREAM algorithm. Then ,when the size of the sample is larger than the outcome determined from a predefined equation, the L SEARCH algorithm is then applied.The k-means algorithm is used to create binary data stream clusters for Several experiments to modified algorithm is far better than the scalable k-means approach[31].

H.3.Grid-based methods:

Grid-based clustering algorithms such as Wave Cluster have very unique characteristics of processing time and it is not dependent on the number of data points, which makes them fast [32]. These algorithms use a multi-resolution grid structure and this structure separates an objects space into a predefined number of cells. Density-based methods: It as ability to detect arbitrary shaped

clusters and also have the ability to handle noise and they require time to scan raw data. According to such algorithms do not require prior knowledge of the number of clusters (k) unlike k-means algorithms that need to be given the number of clusters in advance [33]. Advantage:

- ✓ It is scalable
- ✓ It is sturdy
- ✓ Good in speed and storage capacity Disadvantage:
- ✓ Suffers in ability to handle difficult tasks

H.4 Singular Value Decomposition:

The singular value decomposition (SVD) is a resolution of actual or compound matrix. It is the concept of Eigen disintegration of a positive semi definite normal matrix to any matrix via an extension of the polar dissolution [34].

$$M = \sum V^* \qquad (6)$$

where U is an m unitary matrix over K P is a diagonal m n matrix with non-negative real numbers on the diagonal, V is an n unitary matrix over K, and V ∗ is the conjugate transpose of V.

Application:

- ✓ Used in extracting pattern
- ✓ Used in compressing the image
- ✓ Used as a web search engine
- ✓ Advantage:
- ✓ It has produced a better compact picture ratio
- ✓ Disadvantage:
- ✓ It is not fast from the crunching view
- ✓ The problem strong due to high and large work of associate calculations.

V. CONCLUSION

In this survey, various concepts include big data analytics, predictive analytics techniques have been studied. We considered some of the predictive algorithms and anatomized the prime algorithm as support vector machine (SVM) algorithm as in our focus of perspective SVM is one of the premier algorithms that provides high accuracy of classification and easily solves any sort of datasets. It uses kernel trick to associate skilled data concerning the matter kernel. The kernel function plays the role of dot product in features space. It has the ability to manage vast feature spaces and versatile in choosing similar functions. It works extremely well free edge of the partition. It is compelling in high dimensional spaces. It is viable in situations where a number of measurements are more prominent than the number of tests. It utilizes a subset of preparing point in choice capacity. It is memory proficient. It is used in many real world problems like text categorization, image classification, bioinformatics, and handwritten character recognition.

VI. REFERENCES

[1]. V.Kavya and S.Arumugam. A Review On Predictive Analytics In Data Mining. International Journal of Chaos, Control, Modelling and Simulation Vol.5, No.1/2/3, September 2016

[2]. Lidong Wang and Cheryl Ann Alexander, Machine Learning in Big Data. International Journal of Mathematical, Engineering and Management Sciences Vol. 1, No. 2, 5261, ISSN: 2455-774, 2016.

[3]. R.J. Hyndman and S.Fan, Density forecasting for long-term peak electricity demand, IEEE Trans. Power Syst, PP.1142-1153, May 2010.

[4]. F. Lei and P. Hu, A baseline model for office building energy consumption in hot summer and cold winter region ,Proceedings of International Conference on Management and Service Science. PP.1-4, May 2009.

[5]. G. Kumaresan and P. Rajakumar Predictive Analytics Using Big Data: A Survey, International Journal of Management, Information Technology and Engineering.ISSN:2348-0513; ISSN (Online): 2454-471X; Vol. 3, Issue 9, PP.61-68, Sep 2015.

[6]. Jiang Zheng, Aldo Dagnino, An Initial Study of Predictive Machine Learning Analytics on Large Volumes of Historical Data for Power

System Applications, IEEE International Conference on Big Data PP. 978-1-4799, Vol .1, 2014.

[7]. M.D. Anto Praveena1 Dr. B. Bharathi2, A Survey Paper on Big Data Analytics,.

[8]. Sunpreet Kaur and Sonika Jindal, A Survey on Machine Learning Algorithms, International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2763 Issue 11, Volume 3 (November 2016).

[9]. S.Nagaparameshwara Chary, A Survey on Comparative Analysis of Decision Tree Algorithms in Data Mining International Conference on Innovative Applications in Engineering and Information Technology (ICIAEIT), Volume.3,Special Issue.1,March.2017.

[10]. Dipak V.Patil and R.S Bichkar, Issues in Optimization of Decision Tree Learning: A Survey International Journal of Applied Information System-ISSN:2249-0868 Foundation of Computer Science FCS ,Volume.3-5,July,2012.

[11]. Mr.Brija in R Patel and Mr.Kushik K Rana, A Survey on Decision Tree Algorithm For Classification International Journal of Engineering Development and Research, Volume 2, Issue 1, ISSN: 2321-9939, 2014.

[12]. Marina Komaroff and Noven Pharmaceuticals Multinomial Logistic Regression Models", PharmaSUG International Journal of Engineering Development and Research, Volume 2, Issue 1, ISSN: 2341-9239, 2017.

[13]. Ashis Pradhan Support Vector Machine -A Survey International Journal of Emerging Technology and Advanced Engineering ,Volume 2, ISSN 2250-2459,August 2012.

[14]. https://www.analyticsvidhya.com/blog/2017/09 /understaing-support-vector-machine-example-code

[15]. http://www.statsoft.com/Textbook/Support-Vector-Machines

[16]. Ashis Pradhan Luckyson Khaidem Predicting the direction of stock market prices using random forest, to appear in Applied Mathematical Finance, 2016.

[17]. http://citeseerx.ist.psu.edu/viewdoc/downloaddoi=10.1.1.442.2759$&$rep=rep1$&$type=pdf

[18]. https://www.analyticsvidhya.com/blog/2014/06 / introduction-random-fores-simplified

[19]. Susan Walsh Binary Logistic Regression What, When, and How JMP Discovery Conference 2016.

[20]. http://www.kyb.mpg.de/fileadmin/userupload/files/publications/attachments/luxburg06 TR v2 4139%5b1%5d.pdf

[21]. Sofia Strombergsson Binary Logistic Regression and its application to data from a study of children's recognition of their own recorded voices, Term paper in Statistical Methods, Spring 2009.

[22]. Sidhakam Bhattacharyya Comparative Analysis using Multinomial Logistic Regression International Conference on Business and Information Management (ICBIM), IEEE, 2014.

[23]. Erkan ARI and Zeki YILDIZ Parallel Lines Assumption In Ordinal Logistic Regression And Analysis Approaches, International Interdisciplinary Journal of Scientific Research Volume 3, Dec 2014.

[24]. Swati Gupta A Regression Modeling Technique on Data Mining, International Journal of Computer Applications, ISSN: 0975-8887, Volume 116-No.9, April 2015, Dec 2014.

[25]. https://www.dezyre.com/article/top-10-machine-learning-algorithms/ 202.

[26]. S.V.Suryanarayana, A Survey: Spectral Clustering Applications and its Enhancements. International Journal of Computer Science and Information Technologies Vol. 6, No.185-189, ISSN: 0975-9646,2015.

[27]. Cuimei Guo and Sheng Zheng, A Survey on Spectral Clustering. Proceedings of 2010 Conference on Dependable Computing (CDC2010) ,November 20-22, 2010.

[28]. J. Han, M. Kamber, and J. Pei, Data mining: concepts and techniques: Morgan kaufmann. 2006.

[29]. S. Ding, F. Wu, J. Qian, H. Jia, and F. Jin, Research on data stream clustering algorithms. Artificial Intelligence Review, pp. 1-8, 2013

[30]. H. Yang, D. Yi, and C. Yu, Cluster Data Streams with Noisy Variables. Communications in Statistics-Simulation and Computation, 2014.

[31]. C.C. Aggarwal and P. S. Yu, A framework for clustering uncertain data streams. In Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on, 2008, pp. 150-159.

[32]. G. Sheikholeslami, S. Chatterjee, and A. Zhang, WaveCluster: a wavelet based clustering approach for spatial data in very large databases. The VLDB Journal, vol. 8, pp. 289-304, 2000.

[33]. W.-K. Loh and Y.-H. Park, A Survey on Density-Based Clustering Algorithms. in Ubiquitous Information Technologies and Applications, ed: Springer, pp. 775-780, 2014

[34]. Alexander J. Stimpson and Mary L. Cummings, Assessing Intervention Timing in Computer-Based Education Using Machine Learning Algorithms. ScienceDirect, 2017.

[35]. Eric T. Bradlow and Manish Gangwar, The Role of Big Data and Predictive Analytics in Retailing. ScienceDirect, 2017.

[36]. Teng Li and Jian Tang, Performance Modeling and Predictive Scheduling for Distributed Stream Data Processing. IEEE,2016.

[37]. Kavya.V, Arumugam.S, A Review On Predictive Analytics In DataMining. International Journal of Chaos, Control, Modelling and Simulation (IJCCMS), 2016.

[38]. Nishchol Mishra and Dr.Sanjay Silakari, Predictive Analytics: A Survey, Trends, Applications, Oppurtunities & Challenges. (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (3), 4434- 4438, ISSN: 0975-9646, 2012.

[39]. Panagiotis D.Diamantoulaki, Big Data Analytics for Dynamic Energy Management in Smart Grids. ScienceDirect, 2015.

[40]. G. Kumaresan,P. Rajakumar, Predictive Analytics Using Big Data: A Survey. IJMITE, 2015.

[41]. Tobias Schoenherr and Cheri Speier-Pero, Data Science, Predictive Analytics, and Big Data in Supply Chain Management. Journal of Business Logistics, 2015.

[42]. Xing He, A Big Data Architecture Design for Smart Grids Based on Random Matrix Theory. IEEE, 2015.

[43]. Vimalkumar K, A Big Data Framework for Intrusion Detection in Smart Grids Using Apache Spark. IEEE, 2017.

[44]. Lidong Wang, Cheryl Ann Alexander, Machine Learning in Big Data. International Journal of Mathematical, Engineering and Management Sciences, 2016.

[45]. [45] Hina Gulati, Predictive Analytics Using Data Mining Technique, *2nd International Conference on Computing for Sustainable Global Development,* New Delhi, 2015, pp. 713-716.,2015