IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-6, Issue-10, Oct 2018 E-ISSN: 2347-2693

Adopting Machine Learning Models for Data Analytics-A Technical Note

John Martin R[1*, 2], Swapna S.L[2], Sujatha S[3]

[1]School of Computer Science, Bharathiar University, Coimbatore, India
[2]Faculty of Computer Science & Information Technology, Jazan University, KSA
[3]Dept. of Computer Science, Bharathi Women's College (Autonomous), Chennai, India

Corresponding Author: jmartin.in@gmail.com

Available online at: www.ijcseonline.org

Accepted: 15/Oct/2018, Published: 31/Oct/2018

***Abstract*—** Data science is the most promising area in computer science today. Data science uses various methods and techniques to deal with large volume of data accumulated day by day. Predictive analytics is the prime concept in data science by processing these large volumes of data to make important predictions. This is being achieved through machine learning family of algorithms. This paper makes a note on the core concept of machine learning and the strategies to adopt suitable machine learning algorithms for the problems in data science. It also reviews different areas of machine learning applications in data science.

***Keywords*—** Data Science, Machine Learning, Supervised Learning, Reinforcement Learning

I. Introduction

Machine learning is now a data science technique that allows computers to process existing data through feature engineering and to forecast outcomes and future trends. Machine learning facilitates the computing systems to learn without being explicitly programmed [1].

When the predictions from machine learning occur, the apps and devices become smarter. While shopping online, machine learning persuades consumers by recommending other products we may like to purchase based on our search preferences. On swiping credit cards, machine learning helps to detect fraud by comparing the transaction to a database of transactions. Machine Learning is also emerged as a technique for cloud predictive analytics that enable us to quickly create and deploy predictive models as analytics solutions [2].

Machine learning is used in three wide categories of applications namely data exploration [3], descriptive analytics [4] and predictive analytics [5]. Data exploration is used to derive information from a large and unorganized data set to find characteristics for data analysis. Data mining is a data exploration phenomenon. Descriptive analytics is the process of analyzing a data set in order to examine what happened. Machine learning models for automated diagnosis of deceases using time-series EEG signals is a descriptive analytics [6]. Business and social networks analysis are also descriptive. Predictive analytics uses algorithms that analyze historical and/or live data to identify patterns or trends in order to forecast future events.

Following sections of this paper provide technical concepts on various machine learning models, algorithms and applications. Section-2 presents the overview of a machine learning model. Section-3 details the basic characteristics of machine learning algorithms with reference to existing literatures. Problem scenarios that are influencing the selection of machine learning algorithms and their classifications are presented in section-4. Section-5 describes the influencing factors which are used for choosing suitable machine learning algorithm. The last section of this technical paper featured for presenting the core application areas for data analytics.

II. Machine Learning Models

A machine learning model is a conceptualization of the problem one who tries to solve and the outcome one who predicts [7][8]. Models are being trained and analyzed from existing data. On training a model, known data sets is being used and make adjustments to the model based on the characteristics of the data to obtain the most accurate answer. The machine learning model is comprised of functional modules such as classifiers using an algorithm module that processes training data efficiently. Once the model is trained, test data is applied to evaluate the model. Known data is applied to evaluate the model to check whether the model predicts accurately. This phenomenon is illustrated in figure.1.

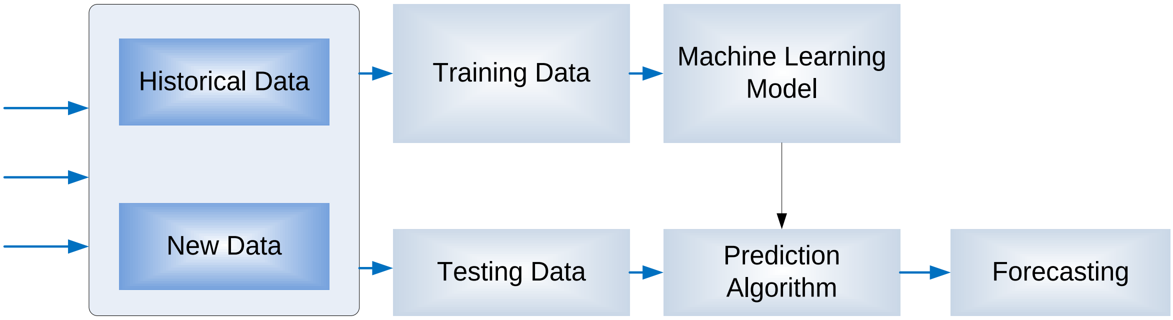

Figure-1: Training-Testing model

III. MACHINE LEARNING ALGORITHMS

A machine learning algorithm is self-contained semantic rules used to solve problems through data processing, mathematics, or automated reasoning. Machine learning algorithms are generally grouped into two classes: supervised and unsupervised [9][10].

Supervised learning algorithms make predictions based on guided examples and past experiences. For example, hazardous guesses in the predictions of future stock prices with the past data. A supervised learning algorithm looks for patterns in the value labels. It can use any information that might be relevant—the day of the week, the season, the company's financial data, the type of industry, the presence of disruptive geopolitical events—and each algorithm looks for different types of patterns. Once the algorithm has found the best pattern it can, it uses that pattern to make predictions for unlabeled testing data—tomorrow's prices.

The most popular and widely used machine learning is supervised. All the modules in Microsoft’s Azure Machine Learning are supervised in nature. The specific types of supervised learning algorithms in Machine Learning include classification [11], regression [12], and anomaly detection [13].

On using the data to predict a category, supervised learning is also called classification. For instance, when assigning an image as a picture of either a 'cat' or a 'dog'. If there are only two choices, the classification will be called two-class or binomial classification. If there are more data classes on which the data either belongs to, it is known as multi-class classification. Whilst a value is being predicted with temporal data, supervised learning is called regression.

In situations to identify data points that are simply in unusual pattern, as any highly unusual credit card spending patterns in fraud detection. The idea that anomaly detection takes is to simply learn how normal object or activity looks like and to spot anything that is significantly different.

The clustering which is unsupervised in nature, data points have no labels to guide the classification [14]. The goal of these unsupervised learning algorithms is to organize the data in such a way by describing its structure. This is called grouping the data into clusters or alternate ways of organizing complex data so as to make them simpler and more organized.

Yet another type of machine learning called reinforcement learning [15], the algorithm gets to choose an action in response to each data point. Sometimes later the learning algorithm also gets a reward signal, showing how best the was the decision. By doing this, the algorithm changes and adopts its strategy to obtain the highest reward. Reinforcement learning is widely applied in the applications of robotics and also well suited for Internet of Things applications.

IV. SELECTION OF ALGORITHMS

Data science adopts various methods and algorithms rooted in few questions which are seeking for answer. Each case is addressed by using well established machine learning algorithms as illustrated in figure.2.

If the problem leads to a question “Is this A or B ?” then classification algorithms will an ideal choice. This kind of algorithms is called binomial or two-class classification as it is useful when the question has just two possible answers. When this question is rephrased to include more than two possible options like “Is this A or B or C or D, etc.?”, then it is called multiclass classification and can be used when you have several possible answers. Multiclass classification is modeled to choose the most likely one.

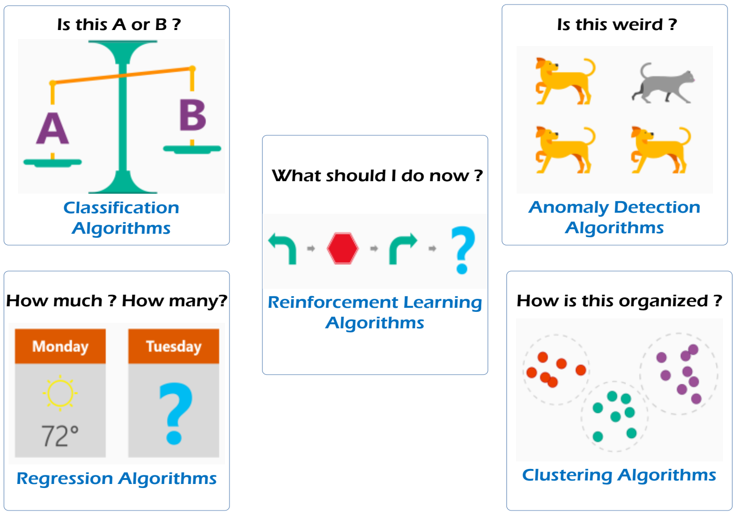

Figure.2: Machine Learning Types

On the situations while looking for the answers to the question "Is this weird?" uses anomaly detection algorithms. For instance, the bank which issues credit card analyzes purchase patterns of the customer, so as to alert the customer during the unusual purchase pattern which leads to possible fraud. Anomaly detection gives alarm for the unexpected or unusual events or behaviours. It also locates where to look for issues.

A family of machine learning algorithms that can predict the answer to the questions such as How much? or How many? is called regression. Regression algorithms widely applied for numerical predictions, such as weather forecasting, business projections etc.

The most promising area of machine learning is clustering. When you look for the response to "How is this organized?", the first choice is clustering algorithm. To know the structure of an existing data set, clustering methods could be used. It classifies data into natural clusters for easier interpretation. Usually clustering may not provide right answer. Common examples of clustering include: Which customers like the same types of products? Which is the common problem caused by the particular system? It is better to understand and predict the behaviours and events to realize How it is organized?

When we ask the query like "What should I do now?" could be addressed with a family of machine learning algorithms called reinforcement learning. Reinforcement learning algorithms learn from outcomes, and decide the further steps. In general, reinforcement learning is suitable for automated systems that have to make lots of small decisions without human guidance. These algorithms gather data as they go, learning from trial and error.

Data science can answer the five questions explored here by a separate family of machine learning algorithms. On the basis of problem scenario and other computational factors influencing the data, each class of algorithms are further classified and are summarized in figure 3.

V. Factors for Choosing an Algorithm

a) Accuracy

Unlike other computational algorithms, obtaining the most accurate answer possible isn't always necessary in machine learning. Depending on the application, sometimes an approximation is sufficient. In those cases more approximate methods can be adopted to reduce the processing time dramatically. Adoption of more approximate methods has the advantage to avoid over fitting.

b) Training time

The overall time span necessary to train a machine learning model varies between algorithms. Training time is closely related to accuracy; means one characteristically accompanies the other. Some algorithms are more sensitive to the number of data points which may influence training time.

c) Linearity

Many machine learning algorithms make use of linearity. In principle, linear classification algorithms assume that classes can be linearly separable by a straight line. For example, regression and support vector machines. Linear regression algorithms assume that data trends follow a straight line. In these algorithms, assumptions provide better results for some cases, but on others accuracy may come down.

There are problems with Non-linear class boundary which are relying on a linear classification algorithm would result in low accuracy. On the other way, data with a nonlinear trend by using a linear regression method would generate much larger errors than necessary. Linear algorithms are very popular and tend to be algorithmically simple and fast to train.

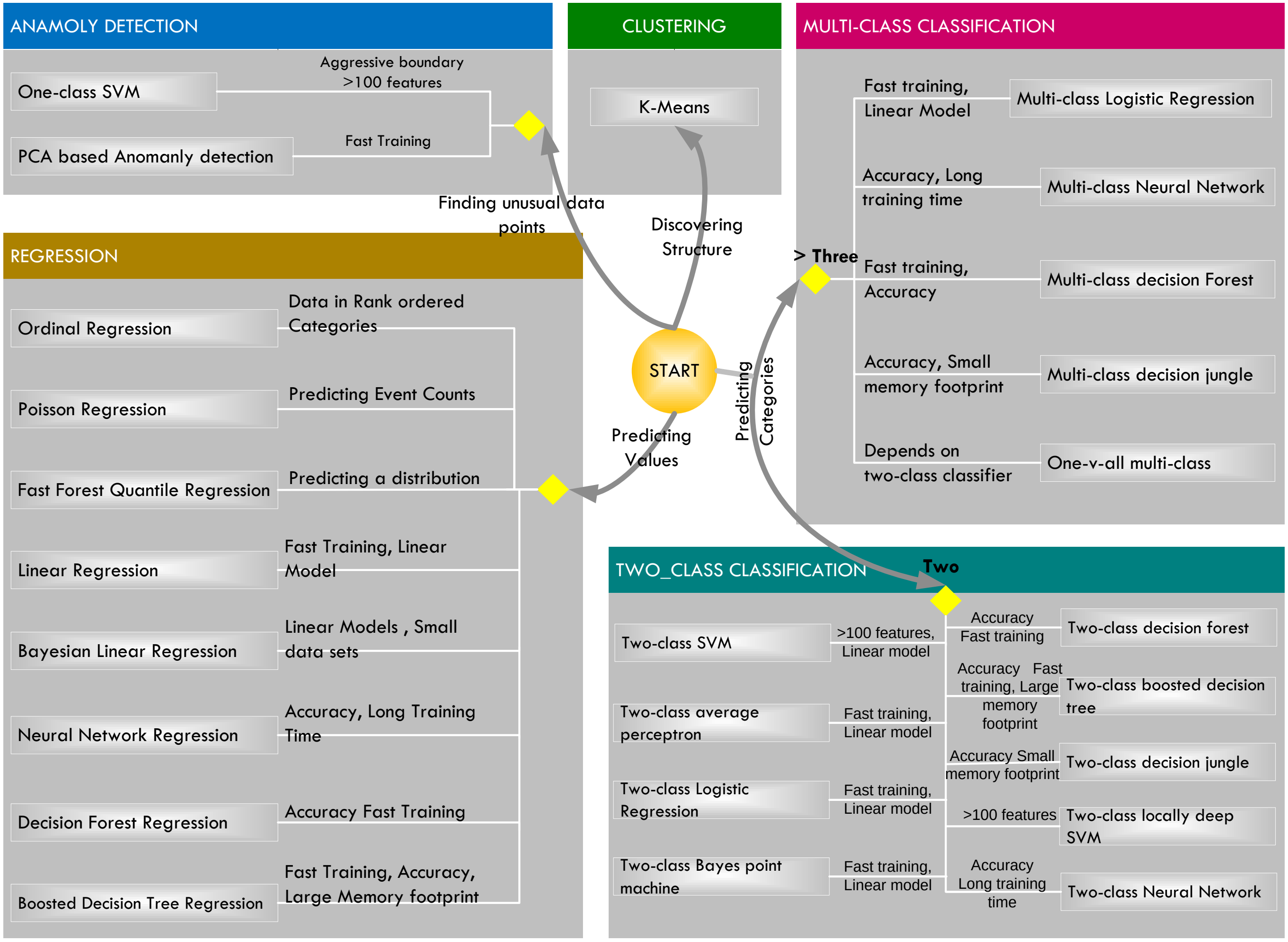
ANAMOLY DETECTION
One-class SVM
Aggressive boundary >100 features
PCA based Anomanly detection
Fast Training
CLUSTERING
K-Means
MULTI-CLASS CLASSIFICATION
Fast training, Linear Model
Multi-class Logistic Regression
Accuracy, Long training time
Multi-class Neural Network
> Three
Fast training, Accuracy
Multi-class decision Forest
Accuracy, Small memory footprint
Multi-class decision jungle
Depends on two-class classifier
One-v-all multi-class
Finding unusual data points
Discovering Structure
START
Predicting Values
Predicting Categories
REGRESSION
Ordinal Regression
Data in Rank ordered Categories
Poisson Regression
Predicting Event Counts
Fast Forest Quantile Regression
Predicting a distribution
Linear Regression
Fast Training, Linear Model
Bayesian Linear Regression
Linear Models , Small data sets
Neural Network Regression
Accuracy, Long Training Time
Decision Forest Regression
Accuracy Fast Training
Boosted Decision Tree Regression
Fast Training, Accuracy, Large Memory footprint
TWO_CLASS CLASSIFICATION
Two
Two-class SVM
>100 features, Linear model
Two-class average perceptron
Fast training, Linear model
Two-class Logistic Regression
Fast training, Linear model
Two-class Bayes point machine
Fast training, Linear model
Accuracy Fast training
Two-class decision forest
Accuracy Fast training, Large memory footprint
Two-class boosted decision tree
Accuracy Small memory footprint
Two-class decision jungle
>100 features
Two-class locally deep SVM
Accuracy Long training time
Two-class Neural Network

Figure. 3. Machine Learning Algorithms

d) Number of parameters

Parameters are the cognitive knob a data scientist gets to tune when setting up an algorithm. Parameters are numeric values that could affect the algorithm's behavior, such as error tolerance or number of iterations, or any other factors which determines how the algorithm should behave. The training time and accuracy of the algorithm may sometimes sensitive to getting the right settings of parameters. Normally, algorithms with large numbers as parameters need the most trials to find a good combination of parameter values. The positive aspects of having more parameters are that an algorithm will have greater flexibility and can often achieve better accuracy if the right combination of parameter settings.

e) Number of features

The cases with particular types of data like textual data in genetics, large number of features are extracted as compared to the number of data points. The number of features can be less in some learning algorithms, may lead to unfeasibly long training time. For instance, Support Vector Machines are particularly well suited to this case.

VI. Machine Learning Applications

Pattern Recognition

The most common uses of machine learning algorithms are in pattern/image recognition. There are many more applications in place where the machine learning is used

[16]. On using with digital images, the measurements depict the outputs of each pixel in the image. On handling with black and white images, each pixel intensity serves as one measurement. In the case of coloured image, each pixel is represented by providing three measurements to the intensities of three primary colour components ie. RGB.

In face detection, the two major categories could be face vs. no face present. A separate category for each person with its characteristics is stored in a database of several individuals. In the case of character recognition, segmentation of a piece of writing into smaller images is done, each containing a single character. Different machine learning algorithms are being adopted for pattern recognition problems [17] based on the purpose to be served.

Speech /Signal Processing
Automatic speech recognition is a prime topic today in IT industry. It includes, discrete word recognition, continuous speech recognition and speech generation.
Handling of speech using machine learning approaches is promising as with signal processing applications [18]. In speech recognition, the spoken words are given as inputs to the applications. A set of numerical values that represent the speech signal could be used as features in this application. The speech signal is being segmented into portions that contain distinct words or phonemes. Each segment is being identified by the intensities or energy at various time-frequency bands.

Medical Diagnosis
Machine learning helps in solving diagnostic and prognostic problems in a wide variety of medical domains such as biomedical signal processing [19] [20], diagnostic decision support systems [21] etc. Analysis of clinical parameters and their combinations for prognosis is being carried out using machine learning methods for forecasting of disease progression, medical knowledge extraction, treatment planning and support, and patient management.
Machine learning is also used for data analysis, such as detection of normality in the data by dealing with anomalies using improper data, interpretation of continuous data used in the Intensive Care Unit, and for intelligent alarming resulting in efficient monitoring of patients. It also improves the accuracy of medical diagnosis by analyzing the data of patients.

Statistical Arbitrage
Statistical arbitrage is a successful automated trading strategy that is crucial for short term and can have large number of securities. Such kinds of problems are being implemented by applying trading algorithms for a set of securities on account of financial parameters like historical correlations and general economic variables [22]. This can be viewed as an estimation problem with the assumption that the prices will close to a historical average. Adopting suitable machine learning algorithms to obtain an index arbitrage strategy is now a useful technique in data science. In specific, linear regression and support vector regression (SVR) are fruitful methods in for this scenario.

Learning Associations
Learning association is a business strategy for the process of developing insights into various associations between products. For example, how seemingly dissimilar products may form an association to one another by analyzing the buying behaviours of customers. One most important application of machine learning is learning associations in business environment. Machine learning algorithms are employed to study the association between the products people buy [23].

Information Extraction
Another application of machine learning called Information or knowledge extraction. It is to dig out structured information from unorganized data from number of online sources such as web pages, communities, blogs, business reports, and even e-mails [24][25]. The relational database systems keep track of the extracted information in an organized manner. This is the key process in big data industry [26]. Extraction principle is used to aid the business industry much by providing the data in the way they want in real time.

Predictive Analytics
One of the well-known areas in data analytics today is prediction or forecasting [27]. As in banking, computing the probability of any of loan applicants faulting the loan repayment based on the past history. The probability of the fault is being computed by classifying the existing data in definite classes. It is guided by a set of classification rules prescribed by the analysts based on the purpose. By using the classified data the probability is estimated across all sectors for varied purposes.

VII. CONCLUSION

Subsequent to data engineering, building machine learning models and applying machine learning algorithms are the two essential steps for data analytics. The technical stuff compiled through this paper will be considered as a primary source of knowledge for machine learning researchers to choose appropriate algorithms for their problems. Even though it provides shadow knowledge on machine learning concepts, the essentials on building machine learning models for the problems related to data analytics are revealed.

CONFLICT OF INTERESTS

No potential conflict of interest was noticed by the authors.

REFERENCES

[1]. J.D. Kelleher, B.M. Namee, A. D’Arcy, “Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies” The MIT Press, 2015.

[2]. H. Chen, R.H.L. Chiang, V.C. Storey, “Business Intelligence and Analytics: From Big Data to Big Impact”, MIS Quarterly, Vol. 36, No. 4, pp. 1165-1188, 2012.

[3]. T.M. Mitchell, “Machine learning and data mining”, Commun. ACM, Vol. 42, No.11, pp.30-36, 1999.

[4]. J. Lin, A. Kolcz, “Large-scale machine learning at twitter”, In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data (SIGMOD '12). ACM, New York, NY, USA, pp.793-804, 2012.

[5]. S. Ryu, “Predictive Analytics: The Power to Predict Who Will Click Buy, Lie or Die”, Healthc Inform Res., vol.19, No.1, pp.63-65, 2013.

[6]. L.Wang, et al., “Automatic Epileptic Seizure Detection in EEG Signals Using Multi-Domain Feature Extraction and Nonlinear Analysis”, Entropy, vol.19, no.6, 222, 2017.

[7]. R.S. Michalski, J.G. Carbonell, T.M. Mitchell, “Machine Learning An Artificial Intelligence Approach”, 1983

[8]. A. Vellido, J.D. Martín-Guerrero, P.J.G. Lisboa, “Making machine learning, models, interpretable”, In Proc. European Symposium On Artificial Neural Networks, Computational Intelligence and Machine Learning, 2012.

[9]. S.B. Kotsiantis, “Supervised Machine Learning: A Review of Classification Techniques”. In Proceedings of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies, Ilias Maglogiannis, Kostas Karpouzis, Manolis Wallace, and John Soldatos (Eds.). IOS Press, Amsterdam, The Netherlands, pp.3-24, 2007.

[10]. J. Dougherty, R. Kohavi, M. Sahami, “Supervised and Unsupervised Discretization of Continuous Features”, In Proceedings of the Twelfth International Conference on Machine Learning, San Francisco (CA), pp.194-202, 1995.

[11]. T.G. Dietterich,” Ensemble Methods in Machine Learning”, In: Multiple Classifier Systems, MCS 2000, Lecture Notes in Computer Science, vol 1857, Springer, Berlin, Heidelberg. 2000.

[12]. A.J. Smola, & B. Schölkopf, “A Tutorial on Support Vector Regression”, Statistics and Computing, vol.14, No.199, 2004.

[13]. P.K. Chan, M.V. Mahoney, M.H. Arshad, “A machine learning approach to anomaly detection”, (CS-2003-06). Melbourne, FL. Florida Institute of Technology, 2003.

[14]. R. Xu, D. Wunsch, "Survey of clustering algorithms," in IEEE Transactions on Neural Networks, vol. 16, no. 3, pp. 645-678, May 2005.

[15]. M.L. Littman, A.W. Moore, L.P. Kaelbling, “Reinforcement Learning: A Survey”, Journal of Artificial Intelligence Research, Vol.4, pp.237-285,1996.

[16]. M. Nasser, Nasrabadi, "Pattern Recognition and Machine Learning," Journal of Electronic Imaging, vol.16, No.4, 049901, 2007.

[17]. K.C. Fu, “Sequential Methods in Pattern Recognition and Machine Learning”, Vol.52, Academic Press, 1968.

[18]. M.M. Richter, S. Paul, “Signal Processing and Machine Learning with Applications”, Springer International Publishing, 2018.

[19]. J.L. Cabra, D. Mendez, Luis C. Trujillo. “Wide Machine Learning Algorithms Evaluation Applied to ECG Authentication and Gender Recognition”, In Proceedings of the 2018 2nd International Conference on Biometric Engineering and Applications (ICBEA '18). ACM, New York, NY, USA, pp.58-64, 2018.

[20]. R. John Martin, S. Sujatha, S.L. Swapna, “Multiresolution Analysis in EEG Signal Feature Engineering for Epileptic Seizure Detection”, International Journal of Computer Applications, Vol.180, No.17. pp.14-20, February 2018.

[21]. R.A. Miller, A. Geissbuhler, “Diagnostic Decision Support Systems”, In: Berner E. (eds) Clinical Decision Support Systems. Health Informatics. Springer, NY, pp. 99-125, 2016.

[22]. E. Chong, C. Han, Frank C. Park, “Deep Learning Networks for Stock Market Analysis and Prediction: Methodology, Data Representations, and Case Studies”, Expert Systems with Applications, Vol;.83, pp.187-205, 2017.

[23]. X. Z. Zhang, "Building Personalized Recommendation System in E-Commerce using Association Rule-Based Mining and Classification," In Proc. International Conference on Machine Learning and Cybernetics, Hong Kong, pp. 4113-4118, 2007.

[24]. D. Freitag, “Machine Learning for Information Extraction in Informal Domains”. Machine Learning, Vol.39, pp.169–202, 2000.

[25]. Rakesh. S.Shirsath, Vaibhav A.Desale, Amol. D.Potgantwar, "Big Data Analytical Architecture for Real-Time Applications", International Journal of Scientific Research in Network Security and Communication, Vol.5, Issue.4, pp.1-8, 2017.

[26]. V.K. Gujare, P. Malviya, "Big Data Clustering Using Data Mining Technique", International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.2, pp.9-13, 2017

[27]. K.S. Shin, T.S. Lee, H.J. Kim, “An application of support vector machines in bankruptcy prediction model”, Expert Systems with Applications, Vol.28, No.1, pp.127-135,2005.