# Analysis and Survey on Machine Learning Algorithms in Different Data Mining Process

M.MeenaKrithika, E.Ramadevi

**Abstract -** Recent world is fully filled with enormous data. The huge sum data put away from various data sources, which is expanded dependent on KDD from different data sources. Data sre produced from fundamental alongside valuable realities from data sources, different methods and technologies are involved to consolidate the huge measure of data sets. The principle yearn of data mining is to mine important data from tremendous measure of data and to recovery data. In data mining classification and clustering principle strategies to sort out and meet up obvious data in a major arrangement of data into fundamental assortment sets of gathering names. Machine learning and data mining are research regions of computer science whose brisk improvement is because of the advances in data examination research, growth in the database business and the subsequent market requirements for techniques that are fit for extricating important knowledge from enormous data stores. In this paper broke down three different machine learning algorithms utilizing data mining process. The exhibition of the Machine Learning algorithm is assessed utilizing different measurements, for example, Classification Ratio, Detection Ratio and Malicious Ratio.

**Keywords-**Machine Learning, Classification, Detection, Malicious, Decision, Naïve Bayes.

## 1. INTRODUCTION

Data mining is frequently used to apply to the two separate procedures of Knowledge discovery and prediction. Knowledge discovery gives express data that has a meaningful frame and can be comprehended by the client (ex: Association Rules Mining). Forecasting, or predictive modeling gives predictions of future occasions and might be straightforward and lucid in certain methodologies (ex: Rulebased frameworks) and murky in others, for example, neural networks. Data Mining depends on the utilization of true data. The essential point of the data mining is to find designs in the data that lead to better understanding of the data producing process and to valuable predictions. Data mining algorithms incorporate algorithms for Association rules, Sequential Patterns, Classification (Decision Trees, SVM, Bayesian Classification, Neural Networks), Prediction ( regression examination, Regression trees, Model Trees), Clustering, Collaborative sifting ( Web mining) and so on. Machine Learning, computational learning hypothesis utilized with regards to Data Mining, to indicate the use of conventional model-fitting or classification algorithms for predictive data mining. Machine learning is the investigation of computer algorithms that improve naturally through "understanding". It separates data from data naturally by computational and measurable strategies. Machine Learning is a strategy that can find already obscure regularities and patterns from assorted datasets.


- *Mrs. M. MeenaKrithikais currently workingAssistant professor in Department of computer science, NGM college, Pollachi-642001.*
- *Dr. E.Ramadeviis currently working Assistant professor in Department of computer applications, NGM college, Pollachi-642001.*


A significant focal point of machine learning research is to naturally figure out how to perceive complex patters and settle on keen decisions dependent on the data. Thus Machine learning is firmly identified with fields, for example, insights, likelihood hypothesis, data mining, design acknowledgment, man-made reasoning, versatile control and hypothetical computer science. The applications for machine learning incorporate machine discernment, computer vision, regular language handling, syntactic example acknowledgment, web indexes, restorative finding, bioinformatics, cerebrum machine interfaces and cheminformatics, distinguishing Visa extortion, securities exchange investigation, grouping DNA successions, discourse and handwriting acknowledgment, object acknowledgment in computer vision, game playing, programming building and robot headway.

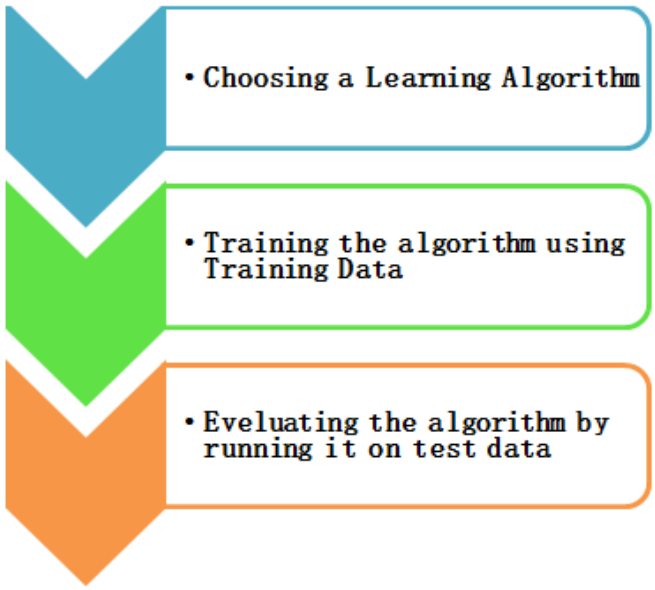



**Figure 1: Machine Learning Algorithm Process**

## 2. LITERATURE SURVEY

**[1] Wangyan Feng, Wenfeng Yan, Shuning Wu and Ningwei Liu** (2017) organize machine learning framework which use cloud document sharing access data to consequently identify client irregular conduct for potential insider danger. In the main stage, venture the entrance log data onto client/document and client/client relationship charts and apply three solo algorithms to produce exception markers. In the subsequent stage, outfit the anomaly pointers and presented the DWT with the Haar wavelet capacity to model clients' fleeting conduct for insider danger detection. A wavelet change maps a one-dimensional capacity into a two-dimensional capacity (for example "time" and "scale"). Officially, in a wavelet change examination application, the initial step is to receive a wavelet capacity, or mother wavelet. At that point, worldly examination is performed with a contracted, high-recurrence variant of the mother wavelet, while recurrence investigation is performed with an enlarged, low-recurrence adaptation of a similar wavelet. Since the first capacity can be spoken to as far as a wavelet development, data tasks can be performed utilizing only the relating wavelet coefficients. Wavelets are characterized by the wavelet work (for example the mother wavelet) and scaling capacity (additionally called the dad wavelet). The (is a capacity used to separate a given capacity into various scale segments, with some specific properties, for example, wave-like oscillation, integral worth is zero, and the integral estimation of its square arrangement is solidarity.

**Merits:**

- DWT with the Haar wavelet is able to effectively capture the temporal patterns of the user behaviors.

**Demerits:**

- More data sources, including but not limited to, active directory logs, windows logon logs, physical security logs, and printer logs, to better detect insider threats holistically

**[2] He Ming Yao , Yu Wei Qin, Li Jun Jiang** (2016) proposed the parasitic capacitance extraction comprehensively required by VLSI modeling is fathomed through the proposed new machine learning based strategy for minutes (ML-MoM). Mother into a machine learning process utilizing the artificial neural network (ANN). In light of this new elucidation, the traditional MoM explaining is transformed into a machine learning preparing process. Therefore, famous machine learning algorithms, for example, the multiple linear regression (MLR) are advantageously utilized to prepare the model and get the expected arrangements. The recently proposed algorithm is applied to parasitic capacitance extractions. The calculation utilizes the well known distributed computing stage – Amazon Web Service (AWS). Numerical tests exhibit fascinating execution favorable circumstances over the customary MoM. The beginning stage of new algorithm is to reinterpret MoM into a preparation model. It is a solitary layer perceptron (neural network structure). The network component Smn is utilized as the info preparing data. The charge thickness obscure is utilized as the weighting coefficient. The potential Vm is the yield of the neural network. Henceforth, unraveling MoM turns into another procedure: feed Smn to prepare the model. In the wake of preparing, can get the precise estimation of the weighting coefficient. Clearly all the more preparing data will deliver better exactness in the last arrangement. Subsequently, when the new algorithm is utilized, more field focuses than source focuses are required, which implies M>N. Thus, the MoM tackling process turns into a machine learning preparing process.

**Merits:**

- Machine learning algorithms such as MLR, software resources such as Python, and hardware resources such as Amazon Web Service, parasitic extractions can be executed effectively through modern machine learning technologies.
- A new gateway for modern computational electromagnetics (CEM) toward electromagnetic compatibility (EMC). Many new opportunities and possibilities could start from this point

**Demerits:**

- More data sources, including but not limited to, active directory logs, windows logon logs, physical security logs, and printer logs, to better detect insider threats holistically.

**[3] Jyun-Shun Liang, Kerwin Wang** (2017) proposed a novel instrument to remove the vibration features of pivoting machinery. The instrument comprises of a sound range analyzer, a sign preparing circuit and a solitary board computer. The machine learning utilized in incorporates Random Forest, K-Nearest Neighbor Classifier (KNN), and Support Vector Machine (SVM). Right off the bat, Random Forest is one of the solid classifiers in machine learning. It has numerous focal points, for example, low-summed up blunder, valuable for data mining, and so forth. Besides, KNN is a regular of sluggish learning algorithm. The "languid" implies that it doesn't take in discriminative capacity from database yet retains database. Be that as it may, it would cause KNN "overfitting" in high dimensional space, the features should be shortened through dimensionality decrease. Thirdly, SVM is an algorithm that changes unique space to higher-dimensional space. It utilizes super vectors to discover hyperplane from the preparation data. In addition, SVM has been generally utilized in shortcoming classification. In utilized Random Forest to remove the significant features of the vibration range. Developed the KNN model to realtime order diverse mechanical vibration. The preparation data used to build the model was taken from the engine while it was lingering or penetrating wood or plastic examples. The Random Forest model doesn't experience the ill effects of overfitting. Then again, the KNN model fits the preparation data well with two significant features. The SVM model gives a high exactness on ordering the preparation data. The decision capacity of the two models is delineated. The decision limit of the RBF part SVM model isn't tight with the goal that the classifier won't cause high speculation mistake on inconspicuous data.

**Merits:**

- Robust with regard to the search space; for instance, classes don't have to be linearly

separable. Classifier can be updated online at very little cost as new instances with known classes are presented.

**Demerits:**

- KNN algorithm is that it is a lazy learner, i.e. it does not learn anything from the training data and simply uses the training data itself for classification.

**[4] HAN LIU1, MIHAELA COCEA, WEILI DING** (2017) proposed handling of picture features in classification undertakings. Specifically, audit two primary kinds of feature selection systems, in particular filter and wrapper. C4.5 is an algorithm of decision tree learning. The idea of this algorithm includes the assessment of features on an iterative premise. Specifically, a decision tree has a root hub and various inward hubs. Every one of these hubs is marked a feature and otherwise called a non-leaf hub, which shows that at every cycle of decision tree learning, a feature should be chosen towards naming a non-leaf hub of a tree. The C4.5 algorithm utilizes entropy or data gain towards heuristic selection of features, which shows that the chose feature is judged heuristically as the best one in the feature set. Following the fruition of a learning task, as a decision tree works in a white box way, the scholarly tree can show which features are added into it as non-leaf hubs, for example those features, which are not added into the tree, are made a decision as immaterial as per the heuristic qualities. Based on the above proclamation, C4.5 can accomplish self-assessment of features during the time spent learning a decision tree and after the fruition of the learning. Self-assessment of features during the time spent learning are expected to limit the computational multifaceted nature of feature selection. Likewise, as showed, C4.5 performs comparatively or far better than the other well known learning algorithms in picture classification. In this area, position the assessment of features with regards to decision tree learning.

**Merits:**

- A significant advantage of a decision tree is that it forces the consideration of all possible outcomes of a decision and traces each path to a conclusion. It creates a comprehensive analysis of the consequences along each branch and identifies decision nodes that need further analysis.

**Demerits:**

- They are unstable, meaning that a small change in the data can lead to a large change in the structure of the optimal decision tree. They are often relatively inaccurate. Many other predictors perform better with similar data.

**[5] MohdNaimMohd Ibrahim, MohdZalimanMohdYusoff** (2015) proposed procedure to arrange tweets notion utilizing Naïve Bayes methods dependent on coaches' discernment into three classes; positive, negative or nonpartisan. The tweets which comprise catchphrase 'Maybank' and 'Malaysia' is being get from twitter stream utilizing Python Twitter library named Twython. Every one of the tweets are put away in database utilizing MongoDB. 50 of most significant tweets for each 'Maybank' and 'Malaysia' catchphrase is being chosen for this investigation. These two watchword is picked for showing capacity of Naïve Bayes for assessment classification in political and business condition. The vast majority of the tweets are blended language among Malay and English sentences/word. This is nature of current internet based life pattern among netizens to stirred up dialects. The tweets are being separated to two sections, one sections which comprise 25 tweets for preparing reason and another part which comprise 25 tweets for classification utilizing Naïve Bayes. The tweets that are chosen for each part is randomly picked. The 25 tweets are appeared to the coach for them to prepare. The preparation is finished via mentor to choose notion whether Positive, Negative or Neutral that the coach think/feel for each tweet. The selection interface is appeared in screen shot beneath. At that point the mentor need to click 'Train' catch to spare the preparation. These preparation data will be utilized in following stage for classification of the tweets opinion. The staying 25 tweets will be characterized their slant utilizing Naïve Bayes algorithm. The preparation data is being utilized as features to the Naïve Bayes to figure the likelihood of outstanding tweets for classification reason. The programmed classification is being finished by using Python NLTK Library

**Merits:**

- Super simple, you're just doing a bunch of counts. If the NB conditional independence assumption actually holds, a Naive Bayes classifier will converge quicker than discriminative models like logistic regression, so you need less training data.

**Demerits:**

- Data scarcity. Chances of loss of accuracy. Zero Frequency i.e. if the category of any categorical variable is not seen in training data set then model assigns a zero probability to that category and then a prediction cannot be made.

## 3. ALGORITHMS FOR SIMPLE ANALYSIS

Three different machine learning algorithms has taken for simple analysis. the three algorithms are then compared with some quality metrics such as Classification Ratio, Detection Ratio and Malicious Ratio. The experimental results from weka tools are displayed for analyzing different results.

### 3.1 K-Nearest Neighbors Algorithm

In design acknowledgment, the k-Nearest Neighbors algorithm (or k-NN for short) is a nonparametric technique utilized for classification and regression.[1] In the two cases, the info comprises of the k nearest preparing models in the component space. The yield relies upon whether k-NN is utilized for classification or regression: In k-NN classification, the yield is a class participation. An article is grouped by a dominant part vote of its neighbors, with the item being appointed to the class generally basic among its k nearest neighbors (k is a positive whole number, ordinarily little). In the event that k = 1, at that point the article is basically relegated to the class of that solitary nearest neighbor. In k-NN regression, the yield is the

property estimation for the article. This worth is the normal of the estimations of its k nearest neighbors. k-NN is a kind of occasion based learning, or languid learning, where the capacity is just approximated locally and all calculation is conceded until classification. The k-NN algorithm is among the most straightforward of all machine learning algorithms.

### 3.2 Naive Bayes:

It is a likelihood based classification procedure. It considers all highlights autonomous of one another. It computes likelihood of each element freely for a specific class mark. Naïve Bayes is utilized in this paper for malware prediction utilizing web traffic data. These are the means behind the Naïve Bayes algorithm:

1. Preparing data set is taken as information.
2. Highlights are extricated from that preparation data. In this paper web traffic data comprises of 43 highlights.
3. At that point from the preparation data for each component Naïve bayes figures likelihood that in the event that element has specific worth, at that point the dataset class be will malicious or not.
4. In the event that each component has constrained potential qualities, at that point above probabilities can be determined. Be that as it may, if the huge number of qualities is there for each element, scope of qualities can likewise be taken.
5. At that point for each line of test data set after the preparation stage. Based all things considered probabilities determined from preparing data decision is taken.

**Advantages:**

a) Since this methodology depends on the likelihood it very well may be applied to a wide assortment of spaces and results can be utilized from multiple points of view.
b) Doesn"trequire enormous measure of data before preparing to start.
c) These algorithms are computationally quick to decide.

### 3.3 Decision tree:

This kind of classifier models data with the assistance of a tree. Tree is having highlights as the inward hubs and edges show the estimations of highlights. And edges isolated hubs dependent on the qualities. All the leaf hubs of the decision tree speaks to a class which is relied upon to be acquired on the off chance that we have every one of the highlights having particular qualities which are in the way from the root to that class having middle of the road include hubs. Probably the most mainstream decision tree algorithms are ID3, C4.5, CART. ID3 is one of the most straightforward decision tree approaches it utilizes idea of data gain as the parting criteria. C4.5 is the development of ID3. It takes a shot at the guideline of addition proportion.

**Advantages:**

All decision tree approaches are simple to understand and easy to interpret.

## 4. EXPERIMENTAL RESULTS

Three different algorithms are used for machine learning comparison.

**Classification Ratio**

| K-Nearest Neighbor Algorithm | Decision Tree Algorithm | Naïve Bayes Algorithm |
|---|---|---|
| **0.02** | 0.09 | 0.04 |
| **0.05** | 0.14 | 0.08 |
| **0.09** | 0.19 | 0.13 |
| **0.14** | 0.25 | 0.19 |
| **0.19** | 0.3 | 0.22 |

**Table 1: Comparison table of Classification Ratio**

The comparison table of classification ratio explains the different values of K-Nearest Neighbor algorithm, Decision Tree Algorithm and Naïve Bayes Algorithm. While comparing these algorithms the decision tree algorithm is better than the other. The K-Nearest Neighbor Algorithm values are starts from 0.02 to 0.19, The decision tree algorithm values are starts from 0.09 to 0.3, naïve bayes algorithm values are starts from 0.04 to 0.22.

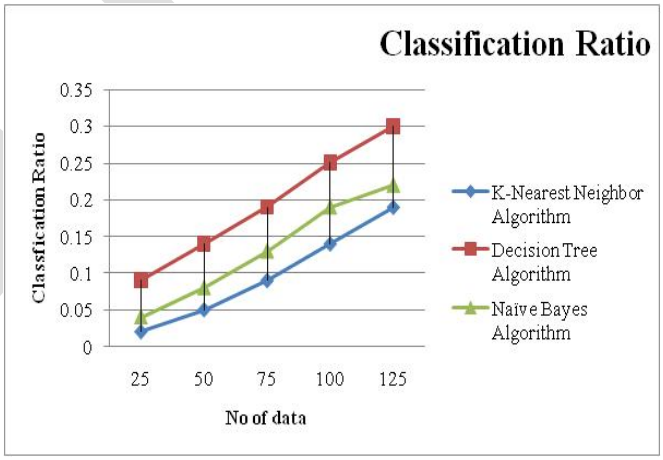



**Figure 2: Comparison chart of Classification Ratio**

The comparison chart of classification ratio shows the values of K-Nearest Neighbor algorithm, Decision Tree Algorithm and Naïve Bayes Algorithm. No of data in X axis and classification ratio in Y axis. The K-Nearest Neighbor Algorithm values are starts from 0.02 to 0.19, The decision tree algorithm values are starts from 0.09 to 0.3, naïve bayes algorithm values are starts from 0.04 to 0.22. the decision tree algorithm values are higher than the other algorithm.

**Detection Ratio**

| K-Nearest Neighbor | Decision Tree | Naïve Bayes |
|---|---|---|

| Algorithm | Algorithm | Algorithm |
|---|---|---|
| 2.2 | 0.09 | 1.3 |
| 2.7 | 0.14 | 1.8 |
| 3.5 | 0.19 | 2.1 |
| 4.4 | 0.25 | 2.4 |
| 5.1 | 0.3 | 3 |

**Table 2: Comparison table of Detection Ratio**

The comparison table of detection ratio explains the different values of K-Nearest Neighbor algorithm, Decision Tree Algorithm and Naïve Bayes Algorithm. While comparing these algorithms the K-Nearest Neighbor algorithm is better than the other. The K-Nearest Neighbor Algorithm values are starts from 2.2 to 5.1, The decision tree algorithm values are starts from 0.09 to 0.3, naïve bayes algorithm values are starts from 1.3 to 3.

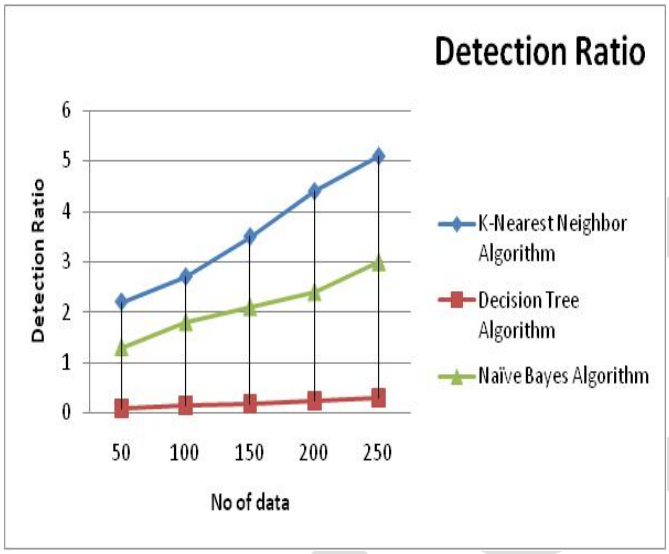



**Figure 3: Comparison chart of Detection Ratio**

The comparison chart of detection ratio shows the values of K-Nearest Neighbor algorithm, Decision Tree Algorithm and Naïve Bayes Algorithm. No of data in X axis and detection ratio in Y axis. . The K-Nearest Neighbor Algorithm values are starts from 2.2 to 5.1, The decision tree algorithm values are starts from 0.09 to 0.3, naïve bayes algorithm values are starts from 1.3 to 3.TheK-Nearest Neighbor algorithm values are higher than the other algorithm.

**Malicious Ratio**

| K-Nearest Neighbor Algorithm | Decision Tree Algorithm | Naïve Bayes Algorithm |
|---|---|---|
| 31.9 | 39 | 26.77 |
| 37.7 | 45 | 31.98 |
| 42.6 | 49 | 34.56 |
| 50.4 | 55 | 38.92 |
| 55.23 | 58 | 44.56 |

**Table 3: Comparison table of Malicious Ratio**

The comparison table of malicious ratio explains the different values of K-Nearest Neighbor algorithm, Decision Tree Algorithm and Naïve Bayes Algorithm. While comparing these algorithms the Naïve Bayesalgorithm is better than the other. The K-Nearest Neighbor Algorithm values are starts from 31.9 to 55.23, The decision tree algorithm values are starts from 39 to 58, naïve bayes algorithm values are starts from 26.77 to 44.56.

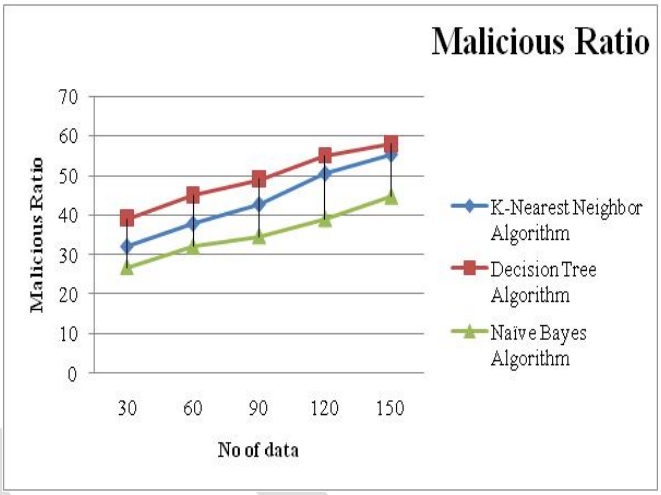



**Figure 4: Comparison chart of Malicious Ratio**

The comparison chart of malicious ratio shows the values of K-Nearest Neighbor algorithm, Decision Tree Algorithm and Naïve Bayes Algorithm. No of data in X axis and malicious ratio in Y axis. The K-Nearest Neighbor Algorithm values are starts from 31.9 to 55.23, The decision tree algorithm values are starts from 39 to 58, naïve bayes algorithm values are starts from 26.77 to 44.56.The Naïve Bayes algorithm values are higher than the other algorithm.

## CONCLUSION

This paper indicates different classification methods utilized in numerous fields, for example, Decision Tree, Nearest Neighbor, Naïve Bayes. Generally Decision trees and Naïve Bayes have distinctive operational profiles, where one is exceptionally precise the other isn't and the other way around. Then again, decision trees and rule classifiers have a comparative operational profile. Different algorithms will be joined for grouping the data set. This paper gives compressive diagram of different machine learning methods utilized in various fields of data mining. Different machine learning methods exhibited different results that may be useful for future innovative researches in machine learning algorithm.

## REFERENCES

[1] Wangyan Feng, Wenfeng Yan, Shuning Wu and Ningwei Liu, “Wavelet Transform and Unsupervised Machine Learning to Detect Insider Threat on Cloud FileSharing”, ©2017 IEEE.

[2] He Ming Yao , Yu Wei Qin, Li Jun Jiang, “Machine Learning Based MoM (ML-MoM) For Parasitic Capacitance Extractions”, ©2016 IEEE.

[3] Jyun-Shun Liang, Kerwin Wang, “Vibration Feature Extraction Using Audio Spectrum Analyzer based Machine Learning”, ©2017 IEEE.

[4] HAN LIU1, MIHAELA COCEA, WEILI DING, “DECISION TREE LEARNING BASED FEATURE EVALUATION AND SELECTION FOR IMAGE CLASSIFICATION”, ©2017 IEEE.

[5] MohdNaimMohd Ibrahim, MohdZalimanMohdYusoff, “Twitter Sentiment Classification Using Naïve Bayes Based on Trainer Perception”, ©2015 IEEE.

[6] Akoglu L., Tong H. and Koutra D. (2015). Graph based anomaly detection and description: a survey. Data Mining and Knowledge Discovery 29.3: 626-688, 2015.

[7] Basso M., Hobert K. A. and Mann J. (2016). Magic quadrant for enterprise file synchronization and sharing, 2016.

[8] D. Jung, Z. Zhang, M. Winslett, Vibration Analysis for IoT Enabled Predictive Maintenance. In 2017 IEEE 33rd ICDE, pages 1271-1282.

[9] A. Moosavian, G. Najafi, H. Nadimi, M. Arab, Estimation of Engine Friction Using Vibration Analysis and Artificial Neural Network. In 2017 ICMSC, pages 130-135.

[10] A. Boudiaf, A. Djebala, H. Bendjma, A. Balaska, A. Dahane, A Summary of Vibration Analysis Techniques for Fault Detection and Diagnosis in Bearing. In 2016 8th ICMIC, pages 37-42.

[11] J. Lou, J. Liao, Y. Wei, Y. Yang, G. Li, Experimental Identification and Vibration Control of A Piezoelectric Flexible Manipulator Using Optimal Multi-Poles Placement Control. In Appl. Sci. 2017,7,309.

[12] W. Xu, J. Zhang, Q. Zhang, X. Wei, Risk prediction of type II diabetes based on random forest model. In 2017 AEEICB, pages 382-386.

[13] ] S. Zhang, X. Li, M. Zong, X. Zhu, R. Wang, Efficient kNN Classification with Different Numbers of Nearest Neighbors. In 2017 IEEE Transactions on Neural Networks and Learning Systems, pages 1-12.

[14] J. Xiang, Y. Zhong, A Novel Personalized Diagnosis Methodology Using Numerical Simulation and an Intelligent Method to Detect Faults in a Shaft. In Appl. Sci. 2016, 6, 414.

[15] S. Raschka, Python Machine Learning. Packt Publishing, Inc., 2015, pages 171-17.

1.

IJSTR