Comparison of Data Mining Algorithms for Predicting the Cancer Disease Using Python

Mehtab Mehdi[1], Kanika Pahwa[1] and Bharti Sharma[2]
[1]*SRM University, Sonipath*
[2]*DIT University Dehradun*

***Abstract*—Fundamentally, machine learning is the part of data science which is nothing but AI. We use machine learning algorithms for predicting the future results after analyzing the past data. This technique of data processing is called data analytics. Machine Learning algorithms are divided in three sections: Supervised, Unsupervised and Reinforcement. These algorithms are further subdivided in other sections. In this paper we are comparing these algorithms by which in future we could easily update the accuracy level of the ML algorithms. For doing this we used the healthcare data which has been uploaded on the kaggle. We implemented the machine learning algorithm using python programming language and calculated the accuracy level of each algorithm.**

Keywords: Ensemble Methods, Support Vector Machine, Random Forest, Decision Trees, Hadoop

I. Introduction

Data mining is the procedure of analysing unseen models of data according to different perspective for classification into valuable information, which is collected and bring together in frequent areas, such as data warehouses, for proficient analysis, data mining algorithms, facilitating business decision making and other information requirements to eventually cut cost and increase profits.

Data mining is also recognized as data finding and knowledge discovery. The data mining processes worked on ELT scheme means Extraction, Loading and Transformation. In these days, ML algorithms are using nearly all the places where a big data is store and develop. For example, banks typically use 'data mining' to find out their prospective customers who could be interested in credit cards, personal loans or insurances as well. Since banks have the transaction details and detailed profiles of their customers, they analyze all this data and try to find out patterns which help them predict that certain customers could be interested in personal loans or we can predict any particular disease using the past analysis on the bioinformatics datasets etc. Data mining engage well-organized data collection and warehousing as well as computer processing. Data mining is useful ML technique for creating the segments in data and calculating the probability of future outcome. These are called the machine learning algorithms. These ML algorithms are divided in three parts; those are supervised algorithms, unsupervised algorithms and reinforcement algorithms. These three algorithms further subdivided. I prepared a chart for these ML algorithms which is as in fig 1.

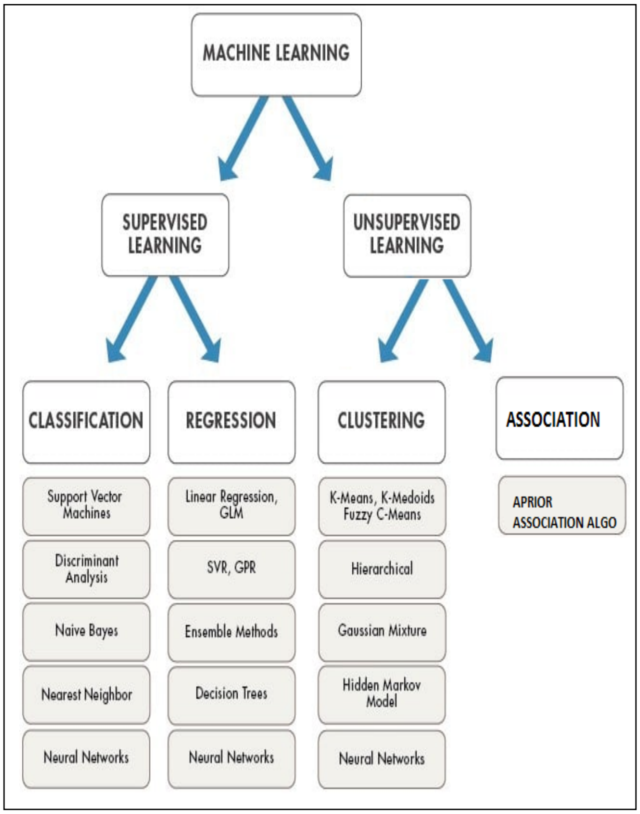

Fig. 1: Machine Learning Algorithms

In the present paper we are going to implement supervised and unsupervised algorithm and we will find the accuracy rate of each algorithm also we will create a table in which we will compare the accuracy rate and there score [3]

II. Literature Reviewed

In the last decades, we found a lot of work has done on big data. Several of method has been developed to resolve this issue. Like as Map reduce method having the ability to split the data using the map and reduce function. By using the map function we can give the id 1 to every character or tokens. Jefferey Dean and Sanjay Ghemawat present a study on the MapReduce and told that how Google handle the peta bytes of data using MapReduce functionality.[1] Jianguo and their team wrote about the Random forest algorithm and there uses in the big data, they also included how ensemble modeling helps the spark to solve the big

Copyright © IEEE–2019 ISBN: 978-1-7281-3245-7

data problem. They have been proposed the parallel random forest algorithm with spark to solve the large scale data mining issue [2]

Our work focuses on the assessment of various algorithms of Machine learning. For that we found some good data mining research which helped us to do our work. Nigel and his team did very similar work. They compared the five ML algorithm for solving the IP traffic flow classification. The difference between his work and our work is that he has given only over view and didn't compare the accuracy, while we compared the accuracy of each algorithm and implemented in python.[3]. But overall that work is supportive to our work. Here, we are comparing the database of our own kaggle depositary where we have been stored our data. One more is done by Murat Soysal, They did work on the Network traffic classification and compared the performance of the ML algorithm [4]

III. Breif of Machine Learning Algorithms

Machine Learning is emerging and very hot topics in the software industry. Image processing, Audio/Video Processing, NLP or Text processing; Machine Learning is a very good solution of all the problems related to these technologies. Actually this is a part of data science, and work on data analytics using big data techniques. We first train to the machine and then ask the machine for prediction, this is the basic theme inside the machine learning algorithms. ML algorithm is divided in three flavours, Supervised, Unsupervised and Reinforcement.[5] We gave a brief overview of these algorithms and implemented these algorithms on our health data under the following sections.

A. Supervised Learning

$$Y = \text{Sign } F(X) = \begin{cases} -1 & \text{if } F(X) \\ +1 & \text{if } F(X) \end{cases}$$

Supervised learning is a way of machine learning by which we create the models on the labeled data set. In this learning technique we first give the training and then we find the accuracy of our algorithm. So we separate the data in two parts first is training data set and testing data set.[6]

1). Types of Supervised Algorithms

Supervised Algorithms are basically divided into two types

i. Classification Algorithms
ii. Regression Algorithms

i. *Classification Algorithm*

It categorizes the small problem into one set. We can categorize the kind of substance or objects by disposing the different feature to classify the exacting class. For example, anyone can simply categorize mobile phone into various types (Apple iphone4, black color, touch screen etc.) by joining different features (memory, mobile colour and shape, processer). If we have given a fresh mobile, everybody will tell that this mobile belong to that class because he or she knows the features of that model. We apply the same rule on the technology, like as if we classify the customers than we can classify them by age and their gender. We use the categorical variable(s) to classify the data. Following are the classification algorithms which we implemented on the healthcare data set. This data set is available on the kaggle.[7]

https://www.kaggle.com/mehtabind123/health-care-dataset.

➢ *Support Vector Machine*

The Support Vector Machine (SVM) is a separation classifier properly to which we can define as an unraveling hyper plane. So, on given labeled data i.e. supervised learning, the algorithm results a finest hyper plane who classify the distinct data among them. In 2D geometry, a hyper plane is a line bisect the geometry, in which each part have the similar type of data item.. The main technique use in the SVM classification is to split two plane classes of a training data set with a plane that make the most of the margin between them. SVM algorithm majorly use in the object detection or face detection. The benefit of SVM is that we can mould the SVM according to our convenience. Like as [8] write that dissolving of Oxygen using the least square SVM (LSSVM). So they use the SVM using the least square method which is different algorithm. Here D is representing the training data.

Where

$$\mathbf{D} = \{(\mathbf{X_1}, \mathbf{Y_1}), \mathbf{i} = \mathbf{1, 2, \ldots, n}, \mathbf{X_1 R^M}, \; \mathbf{y_1} \in \{-1, +1\} \}$$

Which has data example Xi and label Yi, So our objective function would be

$$F(X_i) = (X_i . W_i) + b = \sum_{i=1}^{m} X_i \; W_i + b \quad \ldots\ldots.(2)$$

We use the signum function to assign the model on (X,Y) sample:

$$Y = \text{Sign } F(X) = \begin{cases} -1 & \text{if } F(X) \leq 0 \\ +1 & \text{if } F(X) > 0 \end{cases}$$

When we implemented the above algorithm using the python code we got the following result: (Fig 2) In this data the diagnosis is the categorized variable by which we can predict the cancer type depending on the particular features. SVM can classify these two types of data and we can show it by the scatter matrix as shown in figure 2

Accuracy of SVM classifier on training set: 0.93

Accuracy of SVM classifier on test set: 0.95

Copyright © IEEE–2019 ISBN: 978-1-7281-3245-7

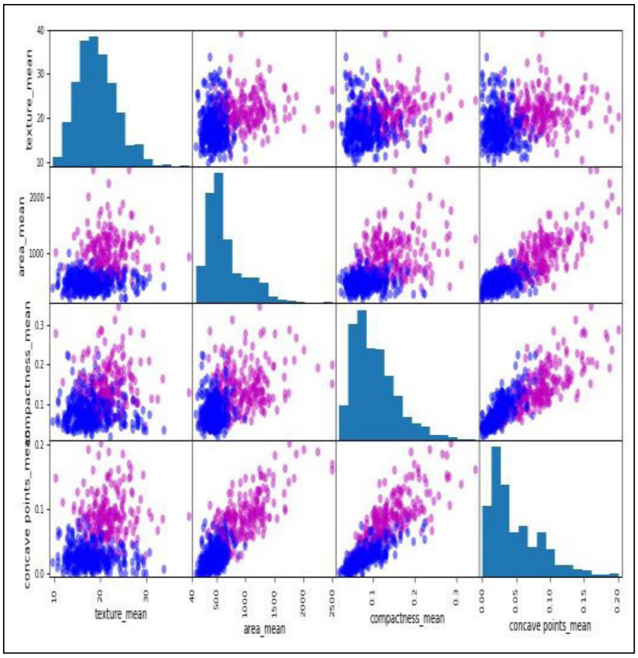

Fig. 2: Support Vector Machine

➢ *Naive Bayes Classifier*

Naive Bayes Classifier is based on probabilistic approach to distribute the data in a particular class. This is the part of the random experiment. NBC depend on the class label which lies inside the finite set. Bayes theorem is the behind of the NBC algorithm. This ML algorithm is famous in the sentimental analysis or text analysis.[9] NBC approach is

$V_{x,y} = \text{arg max}_{vj\ \varepsilon\ v}\ P\ (Vj) \sum P(aj|vj)$ ...(1)

We usually calculate P(ai|vj) by the formula (2):

$P(ai|vj) = (sc + sp) / (r + s)$ …(2)

where:

s = that training data where v = vj

sc = no. of samples where v = vj and a = ai

p = P(ai|vj)

r = corresponding sample size

NBC classifier used the SVM classifier. This theory uses the probability theory of random experiments. For example if we want to separate the positive and negative statement in a text so we use the NBC for that.

When we apply NBC on our data we found only 44.45665% accuracy. So NBC is not suitable to classify this data.

➢ *K Nearest Neighbors Algorithms (KNN)*

KNN (K- Nearest Neighbours) is one of many (supervised learning) algorithms used in data mining and machine learning, it's a classification algorithm where the learning is base on "how similar" is a data (a vector) from other. The k-nearest neighbour's algorithm tries a simple loom to execute categorization. When examined with our data, it looks during the training data and finds the k training samples that are closest to the new example. In [7] top 10 ML algorithm researchers give the first rank to the KNN algorithm. This algorithm can use with the help of PCA (Principle Component Analysis) which is a reinforcement algorithm. [10] then it assign the most common label from all those k-labels to the test samples. When we applied this algorithm on our healthcare dataset so we got the accuracy is 0.966666666667.

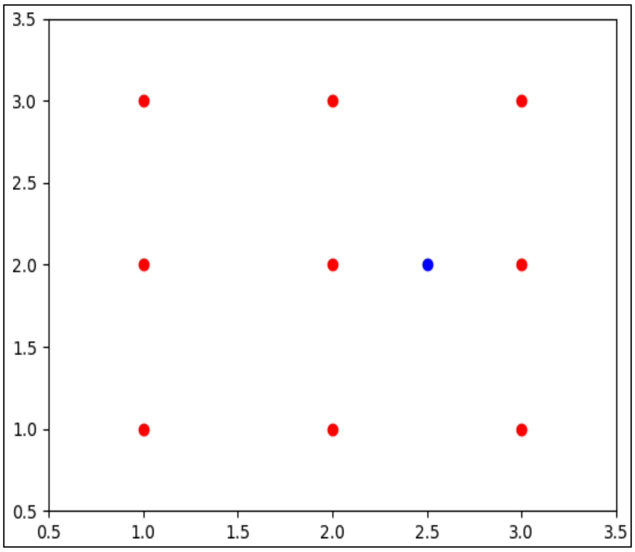

Fig. 3: KNN Classifier

ii. *Regression*

Regression is a statistical calculation used in business, healthcare, and other disciplines that attempts to determine the power of the relationship between one dependent variable (generally denoted by Y) and a series of other changing variables (independent variables x1,x2,x3....).

➢ *Linear Regression*

Linear regression is a basic technique of Machine Learning supervised algorithm. Regression means to forecast the dependent variable using independent variable. By using this technique we can easily identify which of the feature has the highest weight and which one is not taking part in the predictive analysis. The simple regression equation is the linear equation i.e. Y=A+BX where A is constant, B is regression coefficient and X is value of independent variable. So by the above equation we can calculate the dependent variable Y using different values of X.

In our research work, we have one column named type of disease which have two values either M or B, so we take this column as a y variable, rest others are independent variable, On the basis of linear regression we found which feature is most important for calculating the y. Than we found the accuracy using test data by KFold method, The calculated accuracy of Linear Regression is: 0.9717081730669989.

Copyright © IEEE–2019 ISBN: 978-1-7281-3245-7

➢ *Logistic Regression*

As the name suggests itself, Logistic!! where we do work on some logic i.e. logistic regression based on the logit function that generates the value either 0 or 1. Logistic Regression works on the discrete data and uses the sigmoid function to calculate the score. When we did the prediction using the logistic regression we got the following results. Figure 4

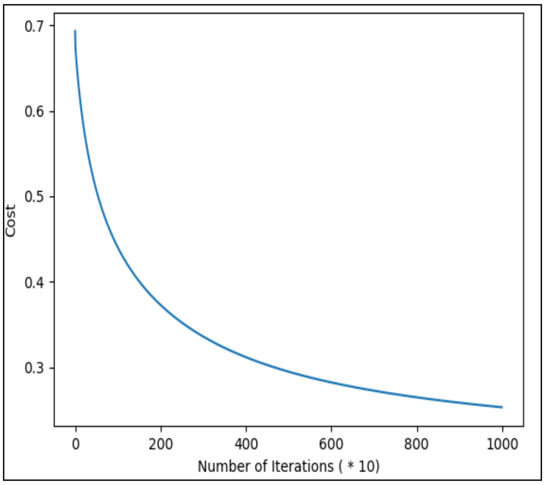

Fig. 4: Cost Vs Iterations

This algorithm predicted the output with following accuracy rate:

Train accuracy: 91.95979899497488 %

Test accuracy: 91.81286549707602 %

B. *Unsupervised Learning*

Unsupervised learning can't be use in regression because it is unidentified what the output values could be, therefore making it not possible to train the machine how we usually do. Now relate this framework to machine learning. Usual datasets in ML have labels (like as: the answer key), and track the logic of "X leads to Y." For example: we might wish for to figure out if people with more Twitter followers normally make higher salaries. We think that our input (Twitter followers) might guide to our output (salary), and we try to estimate what that association is.

1). *Types of Unsupervised Algorithm*

Unsupervised Algorithms are basically divided into two types

i. Clustering Algorithm

ii. Association Algorithms

i. *Clustering Algorithm*

Clustering is a very common technique for prediction of arrangement of different type of data using bunch. These bunch of data calls clusters. This is the task of recognizing the subgroups by that we can filter those data which presents in the unlike clusters. It means we create the uniform clusters of each data points. These cluster usually based on the Euclidean distance or Manhattan distance. All the data in a single cluster is similar whose similarity measures from the features of data. Features are very important component to create the similar data clusters. The application of clustering is in market segmentation; where we try to find consumers that are like to each other.

➢ *K Means Algorithm*

K means algorithm based on iteration by this we create K non overlap unique sub group of data (clusters). The data point created by the K-Means Algorithm always unique and inter related to each other. It creates the cluster on wide distance. The data points must follow some rule like the data points come to each other by using the sum of square distance or smallest distance from the centeroid of the cluster. We found little variation inside clusters. similar data should belongs to the one cluster.

The technique of kmeans algorithm is based on Expectation Maximization or simply we can EM. In the E(Expectation)-step it assign the data to its neighbor while in the M(Maximazation) calculate the centeroid of all cluster.

The objective function is:

$$J = \sum_{i=1}^{m}\sum_{k=1}^{K} w_{ik}\left\|x^{i} - \mu_{k}\right\|^{2} \tag{1}$$

In this function w_{ik}=1 for xi data point if it is inside the k cluster else, w_{ik}=0. We assume μk is the centeroid As treating the μk fixed we minimize J with respect to w_{ik}. After that we fix the w_{ik} and minimize J with respect to μk. In tne E step we differentiate J with respect to μk and update the clusters. After that we do same process with μk and again calculate the centeroid. Then we differentiate J w.r.t. μk and recompute the centroid when we are getting difference with respect to w_{ik} it is Expectation and when we are finding the difference with respect to μk, it is the Maximization step.

So, Expectation is:

$$\frac{\partial J}{\partial w_{ik}} = \sum_{i=1}^{m}\sum_{k=1}^{K}\left\|x^{i} - \mu_{k}\right\|^{2}$$

$$\Rightarrow w_{ik} = \begin{cases} 1 & \text{if } k = argmin_{j}\left\|x^{i} - \mu_{j}\right\|^{2} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

We assign the data point xi to its neighboring subgroup which is following the rule that is sum of squared distance from cluster's centeroid.[11]

Now Maximization will be:

$$\frac{\partial J}{\partial \mu_{k}} = 2\sum_{i=1}^{m} w_{ik}\left(x^{i} - \mu_{k}\right) = 0$$

$$\Rightarrow \mu_{k} = \frac{\sum_{i=1}^{m} w_{ik}x^{i}}{\sum_{i=1}^{m} w_{ik}} \tag{3}$$

When we execute the k means algorithm on our datasets we got the following result. We got the accuracy 65. 334.

ii. *Association Algorithm*

- *Aprior Association Algorithm*

Association between data is another way to use the unsupervised machine learning algorithm. This is more similar to Market Basket Analysis. Market Basket analysis read the behavior of customer, According to this analysis similar item can purchase more time faster if they come together. Aprior Association algorithm is also based on market basket analysis, According to the aprior algorithm similar items associate and they can show the association between frequent and un frequent items [13]

On this data we didn't implement the Aprior Algorithm as this data is not suitable for this algorithm. Also, This we can conclude in our futre research work to implement Aprior Association Algorithm on the health care data.

C. *Ensemble Model*

Ensemble technique is based on the grouping technique. In this technique we take the group of different or similar classifier. We usually split the data into different groups and give that data to each classifier of each group. It divides the one dataset into numerous random data sets. Than we use voting technique to choose the uppermost results. After voting we get the resultant data. So many times we use the decision trees to classify all the groups of data. Due to using so many trees this algorithm looks like a forest. That's why we call this algorithm Random Forest. (Figure 5)

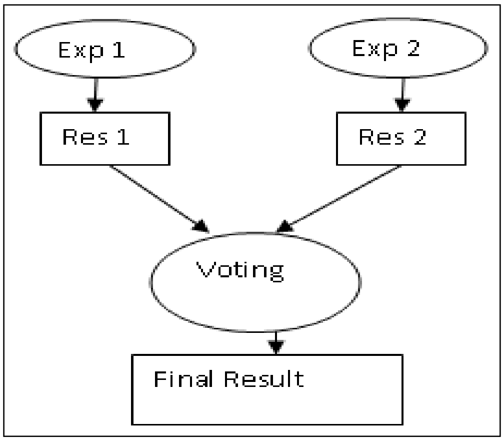

Fig. 5: Ensemble Approach

We use the random forest algorithm to classify and prediction in our health care data, we got following result. In the random forest algorithm we use numerous decision trees to classify the data. After that we use the voting technique for the final result.[12].

Random Forest Algorithm on Healthcare Dataset

Trees: 1

Scores: [90.2654867256637, 87.61061946902655, 90.2654867256637, 93.80530973451327, 87.61061946902655]

Mean Accuracy: 89.912%

Trees: 5

Scores: [95.57522123893806, 92.92035398230088, 97.34513274336283, 92.92035398230088, 89.38053097345133]

Mean Accuracy: 93.628%

Trees: 10

Scores: [95.57522123893806, 94.69026548672566, 94.69026548672566, 89.38053097345133, 96.46017699115043]

Mean Accuracy: 94.159%

D. *Hadoop Model*

For managing the big data hadoop gives a very useful technique to store and process the data. This is a master slave frame work. Hadoop works to clean and pruning the data set. It has two jobs one is to store the data which call hadoop distributed file system (hdfs) and another one is for processing for that we use MapReduce framework. In MR there are two things one is job tracker while another one is task tracker. Job tracker is master daemon and task tracker is slave daemon of hadoop. In the figure we had shown the MR process to calculate the number of words in a file. Map and reduce both programs are written in java.

IV. Experimental Results

- *Data Set*

We used the data set of https://www.kaggle.com/mehtabind123/health-care-dataset. In this data set there are 24 fields when we clean this data set we filter the id features and worked only rest of the 23 features. One categorical feature name diagnosis is there who has two different values B and M means Benign and Malignant. We mapped it by 0 and 1 by which our algorithm can predict easily for the future.

- *Data Implementation*

We cleaned the data by the python numpy and pandas package. All visualization we did by matplotlib library .we also removed the missing value if it was there in the dataset. This process refreshes the last data and gives the accurate result of any data mining algorithm. We select an algorithm and applied that algorithm on the cleaned data. We used the sklearn library of python programming to create the different model

- *Software for Experiment*

This research work we used the python programming for validating our research. This project can use to track the data of health care patients.

V. Results

For this we prepared a table where we wrote the algorithm name and that's accuracy:

Table 1: Comparison of Different Algorithm.

Algorithm Name	Accuracy Rate
Linear Regression	67.33
Logistic Regression	91.11
Support Vector Machine	89.34
K Nearest Neighbors	91.67
Decision Tree	94.3
K Means	65.12
Aprior Association Algorithm	68.22
Random Forest Algorithm(10 trees)	97.55

Further we can do same work by using Genetic Algorithm and Neural Network for classifying the health care data

VI. Discussion

In this research we uses different supervised and supervised algorithm on the health care data and find the accuracy of each step. We presented the table which is defining the accuracy of each algorithm. We implemented these algorithms using python. at the last of the paper. We gave the training data 80% while 20% is the test data. Sample size is very important part for defining the quality of classifier.

VII. Conclusions And Future Work

We also used hadoop by which it processed by the MR and we use pig to processed and store the data in the HDFS.

From our resultant table 1 it is clear that the accuracy of ensemble method is comparatively better than other. Also if we use any other algorithm than we will not find the so accurate result as compare to random forest.

By this work one thing is confirmed that till now the result of ensemble model is better than the other. But if we use the Genetic Algorithm with ensemble approach so we can produce the better accuracy rate than this. In the future we will merge the ensemble approach with some of better classifiers like SGD or GA approach

References

[1] Jeffrey Dean, Sanjay Ghemawat "MapReduce: simplified data processing on large clusters" Volume 51 Issue 1, January 2008 Pages 107-113.Communications of the ACM - 50th anniversary issue: 1958 – 2008

[2] Jianguo Chen, Kenli Li , Zhuo Tang, Kashif Bilal, Shui Yu, Chuliang Weng, Keqin LiA "Parallel Random Forest Algorithm for Big Data in a Spark Cloud Computing Environment", IEEE Transactions on Parallel and Distributed Systems

[3] Nigel Williams, Sebastian Zander, Grenville Armitage "A Preliminary Performance Comparison of Five Machine Learning Algorithms for Practical IP Traffic Flow Classification", ACM SIGCOMM Computer Communication Review, Volume 36, Number 5, October 2006

[4] MuratSoysalaEce GuranSchmidtb "Machine learning algorithms for accurate flow-based network traffic classification: Evaluation and comparison" Performance Evaluation Volume 67, Issue 6, June 2010, Pages 451-467

[5] Fabrizio SebastianiConsiglio Nazionale delle Ricerche, Pisa, Italy "Machine learning in automated text categorization" published in ACM Computing Surveys (CSUR), Volume 34 Issue 1, March 2002 Pages 1-47

[6] [6] SB Kotsiantis, I Zaharakis "Supervised machine learning: A review of classification techniques", Emerging Artificial Intelligence Applications in Computer Engineering, IOS Press 2007

[7] Xindong Wu,Vipin Kumar, J. RossQuinlan,Joydeep Ghosh,Qiang Yang,Hiroshi Motoda, "Top 10 algorithms in data mining" Knowledge and Information Systems January 2008, Volume 14, Issue 1, pp 1–37

[8] Salim Heddama, Ozgur Kisi "Modeling daily dissolved oxygen concentration using least square support vector machine, multivariate adaptive regression splines and M5 model tree", Journal of Hydrology Volume 559, April 2018, Pages 499-509

[9] Shuo Xu "Bayesian Naïve Bayes classifiers to text classification", Volume: 44 issue: 1, : February 1, 2018, page(s): 48-59

[10] Serpen, Gursel, Aghaei, Ehsan "Host-based misuse intrusion detection using PCA feature extraction and kNN classification algorithms" Intelligent Data Analysis, vol. 22, no. 5, pp. 1101-1114, 2018

[11] YeonJoo Jeong Jihang Lee John Moonong Hoon Shin, Wei D. Lu* "K-means Data Clustering with Memristor Networks" ACS Publications, June 2018

[12] KamaldeepSingha,Sharath ChandraGuntukub, AbhishekThakur "Big Data Analytics framework for Peer-to-Peer Botnet detection using Random Forests" Information Sciences Volume 278, 10 September 2014, Pages 488-49

[13] Agrawal, R., Imielienski, T., and Swami, A. (1993). Mining Association Rules between Sets of Items in Large Databases. In: Proc. Conf. on Management of Data, 207–216. New York: ACM Press