# Comprehensive Review On Supervised Machine Learning Algorithms

Hemant Kumar Gianey
Thapar University,
Patiala, Punjab

Rishabh Choudhary
Global Technical Campus,
Jaipur, Rajasthan.

*ABSTRACT:*

**Machine learning is an area of computer science in which the computer predicts the next task to perform by analyzing the data provided to it. The data accessed by the computer can be in the form of digitized training sets or via interaction with the environment. The algorithms of machine learning are constructed in such a way as to learn and make predictions from the data unlike the static programming algorithms that need explicit human instruction. There have been different supervised and unsupervised techniques proposed in order to solve problems, such as, Rule-based techniques, Logic-based techniques, Instance-based techniques, stochastic techniques. The primary objective of our paper is to provide a general comparison among various state-of-the-art supervised machine learning algorithms.**

***Keywords: Machine learning, rule-based techniques, stochastic techniques, logic-based techniques, instance-based techniques.***

## I. INTRODUCTION

Machine learning empowers system with the ability to learn automatically and get better with experience without being explicitly programmed. The algorithms of machine learning are useful in areas where deploying explicitly written algorithms with high speed performance are unfeasible. A simple task such as sorting of numbers is easy and can be performed by giving some numbers as input and getting an ordered list as an output. Here we know what to give as input and what procedure has to be followed to obtain the desired output. But certain tasks are not easy to comprehend such as filtering of emails to differentiate between legitimate emails and spam mails. Here we know the input to be provided and the output is in the form of true or false, but the instructions that need to be given to the program to perform these actions are not clear. Such unique situations where there is no specific algorithm to achieve success, we take the help of data and instruct the machine to analyze the data and make an intelligent sense of this data [1].

**Few Applications of Machine Learning:** [2]

- Classification of texts or documents. e.g: Filtering spam messages [3].
- Speech recognition [4].
- Computer vision tasks such as image recognition and face detection [5].
- Self-driving vehicles [6].
- Web page ranking like for search purposes [7].
- Collaborative filtering [8].
- Medical diagnosis [9].
- Computation biology application [10].
- Recommendation systems, search engines, information extraction systems [11].

## II. Learning Strategies

Machine learning employs the following strategies

### 1. Supervised learning:

### A. Regression:

#### i. Linear Regression:

In simplest terms we can say that in linear regression we add the inputs multiplied by some constants to obtain the output. It creates a correlation between Y, a dependent variable, and X, which can be multiple independent variables, using a straight line (regression line).

The general equation can be written as -

$Y = a + bX$, Where,

Y – Dependent Variable,
X – Explanatory Variable

#### ii. Support Vector Machine Regression:

If we are given a particular training set, say $\{(x1, y1),\ldots,(xi, yi)\} \subset X \times R$, here **X** -> space of input patterns. Our goal in SV regression is to search for a fitting function f(x), having deviation less than ε from the target ($y_i$) acquired for the relating training data set. The function should be reasonably flat. Or it can be said that any error less than ε is fine [12]. The linear function (**f**) -

$$f(x) = (w, x) + b \text{ with, } w \in X , b \in R$$

Here $(\cdot , \cdot)$ represents the dot product of **X**, flatness in this case is described by **w.** In order to make sure we need to keep the norm to a minimum.

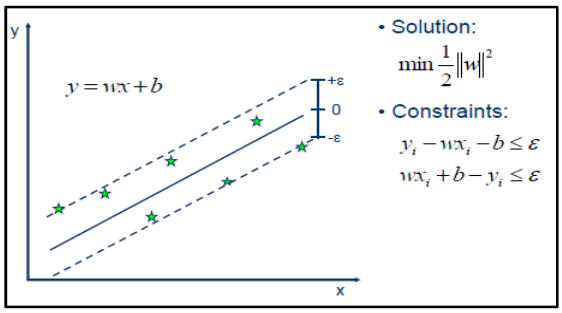


**Figure 1: Support vector machine (SVM)**

**Source: SVM algorithm [13]**

**iii. Decision Tree Regression:**

This regression mechanism works by breaking down a dataset in smaller sub datasets and subsequently related decision tree is developed in an incremental way. Finally a tree having decision nodes and leaf nodes is obtained. The tree has a root node which is the topmost decision node corresponding to the best predictor [14].

The **ID3** (Iterative Dichotomiser 3) is the basic algorithm used to build the decision tree. The ID3 algorithm uses Standard Deviation Reduction (SDR) to construct the decision tree. Steps involved in SDR:

i. First, we calculate the standard deviation of the target.

ii. After this, we split the datasets on the different attributes. The resulting standard deviation for each branch is than subtracted from the standard deviation before the split. This is SDR.

$$\mathbf{SDR(T,X) = S(T) - S(T,X)}$$

iii. The attribute having the largest SDR as the decision node is to be selected.

iv. Dataset has to be divided based on the values of selected attributes. We further split a branch set if the standard deviation is greater than 0.

v. The process keeps on running in recursion until all the data is processed.

**iv. Random Forest Regression:**

It is an extra cover of randomness to bagging. Unlike a normal tree, random forest splits each node using the best among a subset of predictors randomly chosen at that node. It is an easy to operate technique because it has very few parameters – the number of trees in the forest and variables at each node in the random subset. [15].

The algorithm steps of RF Regression:

i. Using the actual data, take $n_t$ bootstrap sample.

ii. For every bootstrap sample, a regression tree has to be grown with some alterations: sample $m_t$ of the predictors randomly at each node and choose the best split amongst the variables.

iii. Estimate the latest data by cumulating the predictions of the $n_t$ trees (average for regression).

**v. LASSO Regression:**

LASSO is short for 'least absolute shrinkage and selection operator'. By shrinking some coefficients and setting others coefficients to zero, it keeps the quality features of subset selection and also ridge regression. LASSO regression has the ability to reduce the variability and improve the precision of linear regression models and also it penalizes the entire size of regression coefficients [16].

$$\sum_{i=1}^{N}\left\{y_i - \sum_{j=0}^{M} w_j x_{ij}\right\}^2 + \lambda \sum_{j=0}^{M} |w_j|$$

This regression uses absolute values in the penalty function, instead of squares which makes some of the parameter predictions to be precisely zero.

**B. Classification:**

**i. Logistic regression:**

It is a reliable procedure to solve binary classification problem. Logistic regression is used to predict the probability of an outcome having only two values. The core of logistic regression is the logistic function - a S-shaped curve taking any real-valued number and mapping that number in a value between 0 and 1, though never precisely at 0 and 1. [17].

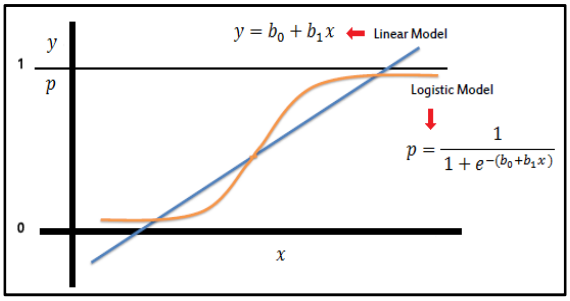


**Figure 2: The steepness of the curve.**

**Source: Logistic regression [18]**

**ii. K Nearest Neighbors:**

This method is known for its simplicity because of the factors such as the ease of interpreting and the low calculation time. It basically stores the cases that are available and categorizes new cases based on the homogeneity basis such as distance function. The object is categorized by a majority vote of its neighbors and the result is usually class integration. After this, object is allotted to a class which has the greatest similarity amongst the K nearest neighbors [19]. Some functions for distance are,

Euclidean - $\sqrt{\sum_{i=1}^{k}(x_i - y_i)}$

Manhattan - $\sum_{i=1}^{k}|x_{i-}y_i|$

Minkowski - $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

In the case of categorical variables Hamming distance must be used.

**iii. Naïve Bayesian:**

It is based on the probabilistic model of Bayes theorem, and easy to set up as complex iterative parameter estimation is almost none, making it viable to use for large sets of data [20]. Given the class variables, the value of a certain characteristic is assumed to be independent of the value of any other characteristic by the naïve Bayesian classifiers. We can calculate the Posterior probability P (A|B)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

P (A|B) - Posterior probability of the class when predictor is given (attribute),
P (A) - Prior probability of the class,
P (B|A) - Probability of the predictor when class is given,
P (B) - Prior probability of the predictor

iv. **Decision Tree (DT) Classification:**

In the decision tree classification, the ID3 algorithm uses Entropy and Information Gain to construct a decision tree instead of Standard Deviation Reduction method. Entropy is used to calculate the homogeneity of the sample. For entropy to be zero, the sample has to be totally homogeneous and this happens if the sample is divided in equal parts [21].

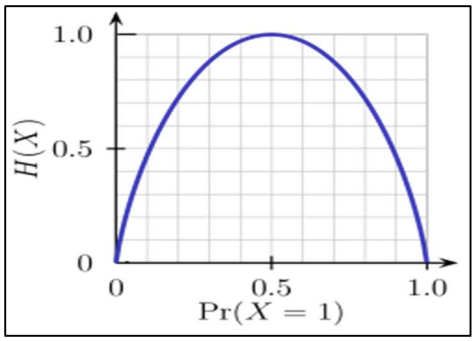


**Figure 3: Entropy of a decision tree**

**Source: Entropy of a DT algorithm [22]**

Algorithm for Information Gain:

i. Calculate Entropy of the target.
ii. We split the dataset based on different attributes and then calculate the entropy for each branch. After this we add it proportionally which gives us total entropy after the split. The total entropy after the split is subtracted from the total entropy before the split and the result is Information Gain.
iii. Attribute with largest Information Gain is the decision node.
iv. A branch with entropy 0 is a leaf node.
v. A branch with entropy greater than 0 needs further splitting.

**v. One-R:**

"One Rule" is a classification algorithm which first produces a single rule for each of the predictors in the data, than the 'one rule' is selected based on the rule having smallest total error. It is a fairly simple and accurate algorithm. To make a rule for the estimator, a table of frequency has to be made for each of the predictors against their targets. The rules produced by One-R are nearly accurate to the state-of-the-art classifying algorithms and can be easily interpreted by the humans [23].

Algorithm for One-R -

i. First, count the frequency of each value of target (class).
ii. Select the class with highest frequency.
iii. That class is assigned to the predictor by the rule.
iv. Next, calculation of total error of the rules for every estimator.
v. The estimator having small gross error has to be selected.

**vi. Zero-R:**

It is among the most simple classification method and relies on target while ignoring all the predictors. 'Zero Rule' classification

method in simple terms, guesses the class with the majority. While having no prediction ability, 'Zero Rule' can be used to determine a baseline performance as a standard for some different classification methods.

Algorithm for Zero-R is simple: Make a table of frequency and choose its most recurrent value [24].

**III. Methodologies carried out for Experiment**

German credit data set has been used. This data set is selected keeping in mind the diversity of the data and data type to be considered during the implication of the methodologies carried out during the test, including categorical data, integer data, nominal data, and in some the mixture of these data types.

Some classification algorithms have been used in order to manipulate the different outcomes of the tools used over different data sets and data types. These algorithms are first defined in detail. The output of the algos is used to verify and analyze the result.

WEKA tool is used for implementation. It is open source tool available freely over the internet. Weka provides a package for data mining system which includes all the features required for the data mining process to be carried. This includes a very good application of all classifier algorithms so far concluded. This it has a very wide application and functionality. It is a java based tool which is allows the service to the user to be worked upon by either graphical User Interface or simple command prompt.

**Credit data-set:** This dataset is for classification of credit risk in Germany. The dataset contains 1,000 observations on 20 attributes (7 numerical and 13 categorical). The target variable consists of 2 classes: 1 for creditworthy and 0 for not creditworthy. The dataset is from Daimler-Benz, Forchung Ulm, Germany

1. Title: German Credit data
2. Number of Instances: 1000
3. Number of Attributes german: 20 (7 numerical, 13 categorical)

Number of Attributes german.numer: 24 (24 numerical)

**IV. Parameters used for evaluating tool performance:**

Only considering the value of accuracy obtained by any classifier as the measure of evaluation of performance of the classifying algorithm is not the correct way. The accuracy of the classifier is just the value for the instances classified to be belonging to their actual class. It does not introduce to the other specificities of the classifier like; the relation between the data attributes, the measure of correct distribution of data instances to each and every class possible, the number of the positive outcomes from the among all received positive outcomes, and several other. These parameters are also much needed during the evaluation of the performance of any classifier which further has helped with the comparative study of selected tools in this research work.

Following are some parameters described which are used in the evaluation process.

**i) Accuracy, confusion matrix and Recall:**

A confusion matrix is a sort of table which defines the number of data instances which are wrongly classified and which are truly classified. This is an nXn matrix where 'n' is the number of classes defined for the data set.

| | | Predicted | |
|---|---|---|---|
| | | 0 | 1 |
| Actual | 0 | TN | FP |
| | 1 | FN | TP |

**Figure 4: Confusion Matrix**

This confusion matrix serves as the basis for deriving values for almost every other parameter used. The columns represent the values which are predicted by the classifier and rows represent the actual values or class labels to which the data object actually belongs.

The values in the cells are as:-

**True negative:** which is the proportion of the negative cases which were classified correctly?
Calculated as: TN/TN+FP

**False Positive:** which is the proportion of negative cases which were classified incorrectly as positive?
Calculated as: FP/TN+FP

**False Negative:** which is the proportion of positive cases which were classified incorrectly as negative?
Calculated as: FN/FN+TP

**True Positive or Recall:** proportion of positive cases that were correctly classified.
Calculated as: TP/FN+TP

**Accuracy:** it denotes the number of correct predictions made by the classifier. And it is calculated as:

TN+TP/TN+FP+FN+TP

**Precision (Confidence):**
Precision is the rate of positives which were predicted positive and in actually too were positive. This value is given as:

TP/FP+TP

**Lift and ROC curves:**
These are the two visual measures of determining the performance of the classifier.
**a) Lift curve** :
**a) Lift curve:**
Lift curve provides us with the measure to determine the effectiveness of the classifier model so generated. It is the ratio of the result obtained with or without the classifier model. It is the curve on the graph of population threshold and the rate of positive responses received. The graph consists of a baseline in the middle and the performance of the classifier is evaluated on the basis of the lift curve formed on either the upper side of the baseline or on the lower side.
If it is on the upper side then is considered good and the area under the curve is measured which in this case would be more. And if the curve is under the baseline then it is not a good classifier.

At any point on graph it denotes the cumulative gain for the percentile of targeted population.

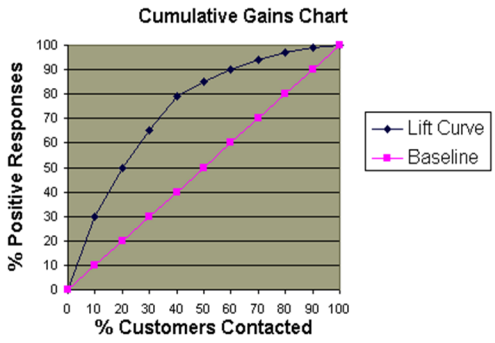


**Figure 5 : Lift curve**

**b) ROC (Receiver operating Curve):**
This is the curve which is just the visual representation of the true positive and false positive value which on graph helps to determine the significance of the Classifier model.
The x axis is the false positive values and y axis is the true positive values. At any point on graph the significance of the model for the data object can be seen.
If the classifier is just excellent then it is denoted by the point(0,1) on the graph where '0' means the FP rate is null and '1' means TP rate is for all correctly classified.

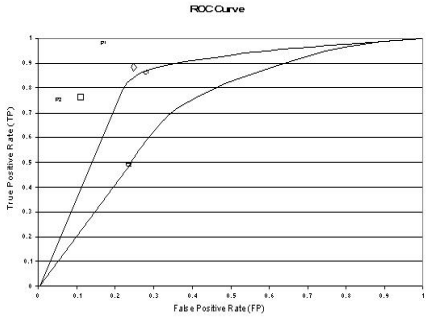


**Figure 6: ROC curve**

**V. Result Observation**
In this section the observed result of the experiment are presented and evaluated. The analyses of each parameter on data collected for each algorithm are presented in table1 and figures from 7 to 14.

**Lift curve Graphs:**
Following are the Lift curves obtained for Credit dataset

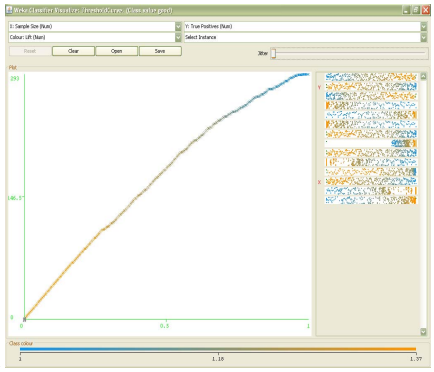

**Figure 7: Lift curves for k nearest neighbor for Credit dataset**

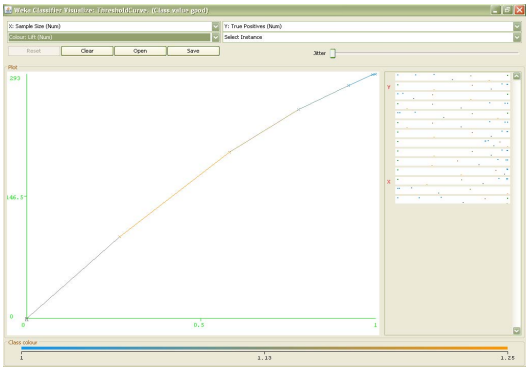

**Figure 8: Lift curves for Decision Tree for Credit dataset**

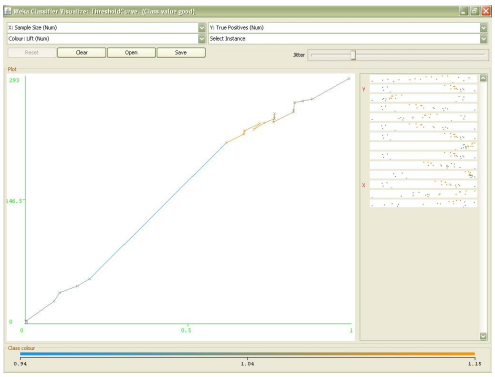

**Figure 9: Lift curves for NB over Credit dataset**

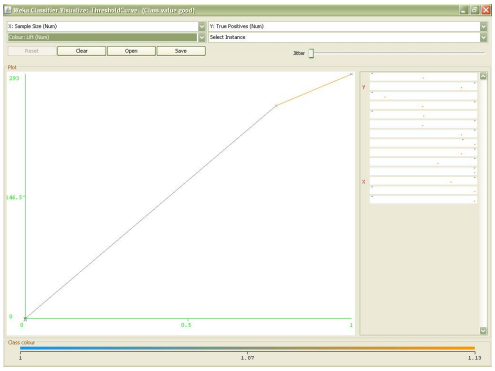

**Figure 10: Lift curves for Support Vector Machine over Credit dataset**

**ROC curve Graphs**

Following are the ROC curves obtained for dataset

**Figure 11: ROC curves for credit data set for NB**

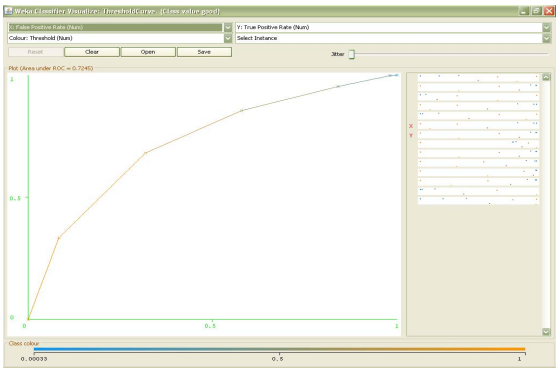

**Figure 12: ROC curves for Credit dataset for KNN**

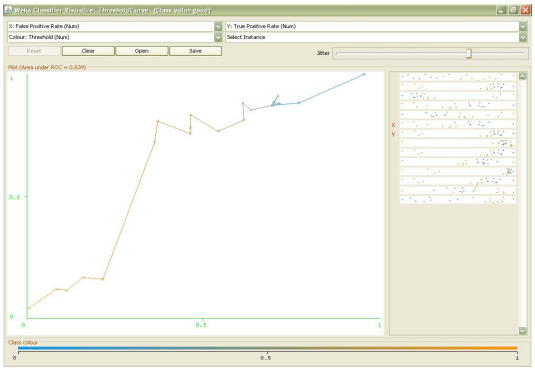

**Figure 13: ROC curves for Decision tree for credit data set.**

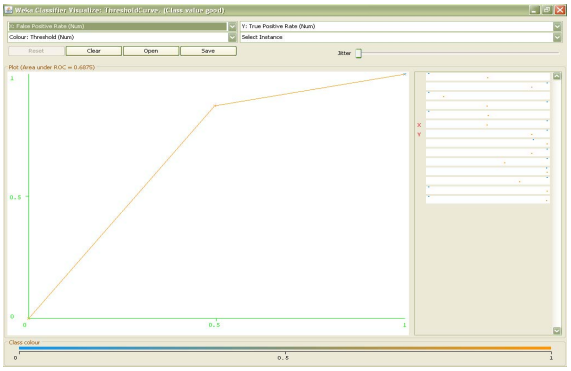

**Figure 14: ROC curves for credit data set for SVM**

## VI. Conclusion:

This paper discusses the commonly used supervised algorithms. The primary goal was to prepare a comprehensive review of the key ideas and present different techniques for every supervised learning method. The paper makes it clear that every algorithm differs according to area of application and no algorithm is more powerful than the other in different scenarios. The choice of an algorithm should be made depending on the type of problem given to us and the data available. The accuracy can be increased by using two or more algorithm together in suitable conditions.

**Table 1: Parameters used for supervised algorithms**

| Parameters used/Algorithm used | Logistic Regression | Support vector Machine | Random forest | Naive Bayesian | One-R | Zero-R |
|---|---|---|---|---|---|---|
| Accuracy | 75.2 | 75.1 | 76.4 | 75.4 | 66.1 | 70 |
| Precision | 0.741 | 0.738 | 0.751 | 0.743 | 0.608 | 0.490 |
| Recall | 0.752 | 0.751 | 0.764 | 0.754 | 0.661 | 0.700 |
| F-measure | 0.744 | 0.741 | 0.744 | 0.746 | 0.620 | 0.576 |
| TP-Rate | 0.752 | 0.751 | 0.764 | 0.754 | 0.661 | 0.700 |
| FP-Rate | 0.398 | 0.410 | 0.440 | 0.393 | 0.614 | 0.700 |
| MCC | 0.379 | 0.371 | 0.386 | 0.385 | 0.061 | 0.000 |
| ROC area | 0.785 | 0.671 | 0.791 | 0.787 | 0.524 | 0.500 |
| PRC area | 0.798 | 0.681 | 0.810 | 0.797 | 0.591 | 0.580 |

## VII. References:


[1] Domingos, P. "A few useful things to know about machine learning", Communications of the ACM, 55(10),2012 pp.1.

[2] Mohri, M., Rostamizadeh, A. and Talwalker, A. "Foundations of machine learning", Cambridge, MA: MIT Press,2012.

[3] Nguyen, T. and Shirai, K. "Text Classification of Technical Papers Based on Text Segmentation", Natural Language Processing and Information Systems, 2013,pp.278-284.

[4] Deng, L. and Li, X. "Machine Learning Paradigms for Speech Recognition: An Overview", IEEE Transactions on Audio, Speech, and Language Processing, 21(5), 2013, pp.1060-1089.

[5] Siswanto, A., Nugroho, A. and Galinium, M. "Implementation of face recognition algorithm for biometrics based time attendance system", 2014 International Conference on ICT For Smart Society (ICISS).

[6] Chen, Z. and Huang, X. "End-to-end learning for lane keeping of self-driving cars", 2017 IEEE Intelligent Vehicles Symposium (IV).

[7] Yong, S., Hagenbuchner, M. and Tsoi, A. "Ranking Web Pages Using Machine Learning Approaches", 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.

[8] Wei, Z., Qu, L., Jia, D., Zhou, W. and Kang, M. "Research on the collaborative filtering recommendation algorithm in ubiquitous computing", 2010 8th World Congress on Intelligent Control and Automation.

[9] Kononenko, I. "Machine learning for medical diagnosis: history, state of the art and perspective", Artificial Intelligence in Medicine, 23(1), 2011, pp.89-109.

[10] Jordan, M. "Statistical Machine Learning and Computational Biology",IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2007).

[11] Thangavel, S., Bkaratki, P. and Sankar, A. "Student placement analyzer: A recommendation system using machine learning", 4th International Conference on Advanced Computing and Communication Systems (ICACCS-2017).

[12] Byun, H. and Lee, S., "Applications of Support Vector Machines for Pattern Recognition: A Survey". Pattern Recognition with Support Vector Machines, 2002, pp.214-215.

[13] Support vector machine regression algorithm [Online], http://chem-eng.utoronto.ca/~datamining/dmc/support_vector_machine_reg.htm, last access 22.08.2017.

[14] Kotsiantis, S. "Decision trees: a recent overview. Artificial Intelligence Review", 39(4), 2011, pp.262-267.

[15] Andy Liaw and Matthew Wiener "Classification and Regression by randomForest", R News, ISSN 1609-363, vol. 2/3, December 2002, pp. 18-22.

[16] Tibshirani, R."Regression shrinkage and selection via the lasso: a retrospective", Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(3), 2011, pp.273-282.

[17] Brownlee, J. "Logistic Regression for Machine Learning - Machine Learning Mastery", [online] Machine Learning Mastery. Available at: http://machinelearningmastery.com/logistic-regression-for-machine-learning/ [Accessed 12 Aug. 2017].

[18] The steepness of the curve of logistic regression [Online], http://chem-eng.utoronto.ca/~datamining/dmc/logistic_regression.htm, last access 22.08.2017.

[19] Bicego, M. and Loog, M, "Weighted K-Nearest Neighbor revisited", 23rd International Conference on Pattern Recognition (ICPR), 2016, pp. 1642-1647.

[20] Ting, K. and Zheng, Z. "Improving the Performance of Boosting for Naive Bayesian Classification. Methodologies for Knowledge Discovery and Data Mining", 1999, pp.296-298.

[21] Peng Ye, "The decision tree classification and its application research in personnel management", Proceedings of 2011 International Conference on Electronics and Optoelectronics, 2011, pp. 1-4.

[22] Entropy of a decision tree classification algorithm [Online], http://chem-eng.utoronto.ca/~datamining/dmc/decision_tree.htm, last access 22.08.2017.

[23] Muda, Z., Yassin, W., Sulaiman, M. and Udzir, N. "Intrusion detection based on k-means clustering and OneR classification", 2011 7th International Conference on Information Assurance and Security (IAS).

[24] Kerdegari, H., Samsudin, K., Ramli, A. and Mokaram, S. "Evaluation of fall detection classification approaches", 2012 4th International Conference on Intelligent and Advanced Systems (ICIAS2012).