Classification Techniques : A study on the topic of Supervised Machine Learning Algorithms

Monesa Thoguluva Janardhanan
Computer Science Department
University of North Carolina at Charlotte
Charlotte, United States of America
mthogulu@uncc.edu

***Abstract*—The main aim of this paper is to present the ideas grasped from the study of the Classification Techniques of Machine Learning. Classification techniques are a supervised Machine Learning approach which helps in classifying the sets of categories to which the data belongs to. This have wide applications in many industries like Healthcare, Text Mining, Image analysis etc. This paper presents the most commonly used Machine Learning Classification Techniques such as Logistic Regression, Decision Trees, Random Forest and Naïve Bayes Classification. 18 research papers were randomly selected for this study which mostly focuses on the topics of Sentiment analysis, Image Classification and Healthcare Industry. Some also deals with the comparison of these techniques for the same research in order to find out the most accurate algorithm. The classification techniques overview and accuracy are mentioned at the Literature Review section of this paper.**

Keywords—Classification, Logistic Regression, Decision Trees, Random Forest, Naïve Bayes Classification.

I. Introduction

Machine Learning Techniques are numerously used in every field as this empowers the software to learn, explore and analyses without any human intervention [19]. "Google says Machine Learning is the future". Machine Learning has its applications in enormous fields now including Healthcare Industry, Recommendation Systems, Management, Data Mining, Image Processing, Prediction Systems etc. [19]. There are three different types of Machine Learning Techniques — Supervised, Semi-supervised, Unsupervised. Supervised Algorithm, as the name stands for the machine needs to be supervised first and then test it. The data sets are split into two, one is the training data set and the other is the testing data set [20]. Most commonly 70% of data would be used for training and the rest 30% would be used for testing. The training data set trains the machine and once a model is created, we use the testing data set to test the data. From the references used we can infer that the most widely used supervised algorithms are Support vector machine, Logistic Regression, Linear Regression, Decision Trees, Random Forest, Naïve Bayes Classification, K-Nearest Neighbor, Neural Networks [2]. Semi-Supervised learning, as the name stands for the machine is trained for a very small data sets and tested for huge data sets. Most commonly only 30% of data will be used for training and the remaining 70% will be used for testing data set. Continuity assumption, Cluster assumption, Manifold assumption, Generative models, Low-density separation, Graph-based methods, Heuristic approaches are some of the assumptions which are used by Semi-supervised learning [21]. Unsupervised Algorithm, as the name stands for the machine is not trained for anything. The system has to draw inferences from the data and propose a result. The most common unsupervised learning method is cluster analysis and reinforcement learning [21]. This paper deals with some of the classification techniques used in Supervised Machine Learning Algorithms.

II. Classification Techniques

A. Logistic Regression

Logistic Regression is an effective and inductive supervised learning algorithm that can be applied to many numbers of fields. It is a discriminative classifier which is very simple and involves less computation. It predicts the binary outcome on a given set of independent variables [1]. Logistic regression classification is a linear regression after normalization of logistic equation. This normalization method suppresses too large and too small results (usually noise) so as to ensure that the mainstream results are not ignored. At the same time the model is easy to explain, easy to extract rules and also has good robustness to noise and inference. Logistics regression is a form of regression which allows prediction of outcome variables by combination of continuous and discrete predictors [2]. The range of logistic function is always between 0 and 1. So a conversion function is always needed to make it lie between 0 and 1. A typical logistic function is in S shape as shown below in the Fig 1. [1].

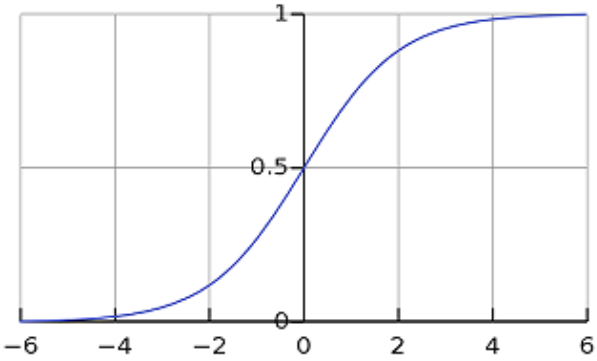

Figure 1: Logistic Function Curve [1]

B. Decision Tree

Decision Tree is a supervised machine learning algorithm for classification and regression problems and is very useful in predicting future instances [8]. It offers many advantages over other classification techniques because of its ease of use and simplicity. Decision trees are powerful and popular tool for classification and prediction. The attractiveness of DT is due to the fact that, it represents rules [8]. DT is a classifier in the form of tree structure where each node is either a leaf node or a decision node. DT induction algorithm is an approximate discrete function method which can yield lots of

useful expressions [10]. It is one of the most important methods for classification.

1) Decision Tree Algorithm [6]:

a) Place the best attribute of the data set at the root of the tree. The Attributes are selected based on Information Gain.

b) Split the training set into subsets. Each subset should contain data with the same value for an attribute.

c) Repeat the steps on each subset until leaf nodes are created in all the branches of the tree.

2) Decision Tree Pseudocode [6]:

The below image displays about the decision tree. Pseudocode

Input: an attribute-valued dataset D
1: Tree = {}
2: **if** D is "pure" OR other stopping criteria met **then**
3: terminate
4: **end if**
5: **for all** attribute a ϵ D **do**
6: Computer information-theoretic criteria if we split on a
7: **end for**
8: a_{best} = Best attribute according to above computed criteria
9: Tree = Create a decision node that tests a_{best} in the root
10. D_v = Induced sub-datasets from D based on a_{best}
11: **for all** D_v **do**
12: $Tree_v$ = C4.5(D_v)
13: Attach $Tree_v$ to the corresponding branch of Tree
14: **end for**
15: **return** Tree

Figure 2: Decision Tree Pseudocode [6]

C. Random Forest

The algorithm of Random Forest was first proposed by Ho in 1995. The so-called random forest algorithm is actually a combined classifier that contains multiple decision trees which can solve problems such as classification and regression. Ho believes that as long as the decision tree distinguishes the oblique hyper plane better accuracy can be obtained without over training [12]. It is a method which contains multiple basic algorithms. Each basic algorithm can be expressed and implemented in a decision tree. Each decision tree has number of leaf nodes and the depth of the tree, the random forest can combine the calculation results produced by multiple decision trees to optimize the results [12]. The greater the number of trees in the forest gives more precise result [13]. There are two steps in the Random Forest Algorithm. The first one is Random Forest Formation and the other one is to make forecast from the random forest classifier constructed in the first stage.

1) Random Forest Formation Algorithm [13]:

a) From total "m" features randomly select "K" features where k << m.

b) Among the "K" features, determine the node "d" using the best rift point.

c) Split the node into daughter nodes using the best rift.

d) Repeat the 1 to 3 steps until "l" number of nodes has been reached.

e) Build forest by repeating steps 1 to 4 for "n" number times to create "n" number of trees.

2) Random Forest Prediction Algorithm [13]:

a) Takes the test features and uses the rules of each randomly produced decision tree to anticipate the outcome and stores the outcome (target).

b) Appraise the votes for each anticipated target.

c) Consider the highest chosen anticipated target as the final forecast from the random forest algorithm.

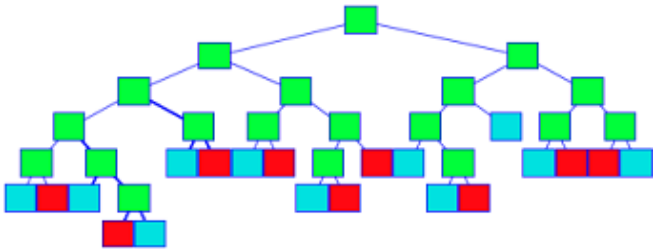

Figure 3: Classification Tree Topology [12]

D. Naïve Bayes Classifier

Naïve Bayes Classifier is a simple technique which works well for textual data. It considers feature which are conditionally independent of each other. Because of the independence supposition, the parameters for each attribute can be learned independently and this enormously streamlines learning particularly when the data is large [4].

Bayes Theorem is as follows:

$$P(X \mid Y) = P(Y \mid X) * P(X) / P(Y)$$

Here, P (X | Y) denotes the Posterior; P (Y | X) the likelihood; P (X) the prior and P (Y) is the Evidence. The condition posed here when the probability is considered is that the probability of P (Y) should never be zero [16]. The naïve proposed could be used in various applications like Game Prediction to predict the future of playing with the weather data that we have, and also in News Categorization naïve could apply the classifier for classification of news contents based on news code, another application is in Spam Filtering [5] naïve could classify the emails with spam and non-spam and its probability [16]. The Naïve Bayes has been developed by C. and G. Salton [1] and also S. Roberson and K. Spark [2] in the 1970 respectively [18].

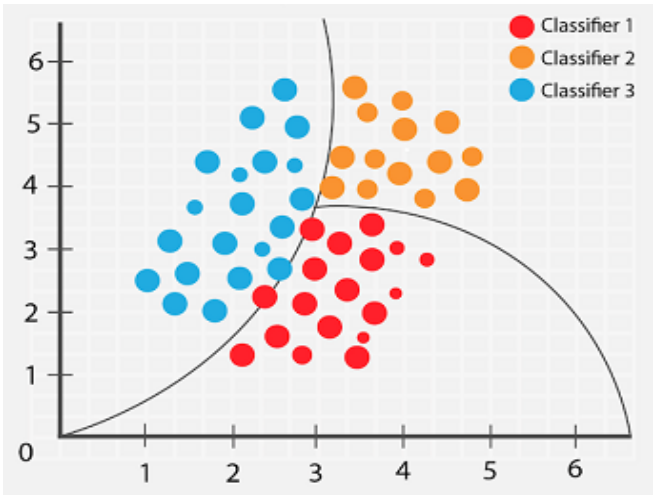

Figure 4: Naïve Bayes Classifier [22]

III. Literature Review

A. Prediction of Liver Disease using Classification Algorithms [1]

Reference [1] focuses on prediction of liver disease using different classification algorithms namely Logistic Regression, Support Vector Machine, K-Nearest Neighbor. In this experiment the data set was divided into training and testing set. The ratio of training and testing set was 70% and 30% respectively. Performance accuracy is measured by two methods. One is by confusion matrix and the other is sensitivity. From the confusion matrix the accuracy of K-nearest neighbor model, Logistic Regression Model, Support Vector Machine Model was 73.97%, 73.97%, 71.97% respectively. And the corresponding sensitivity was 0.317, 0.195, 0.195. From this experiment Logistic regression and K-Nearest Neighbor has the highest accuracy but Logistic Regression has the highest sensitivity. Therefore, the paper states that Logistic Regression is appropriate for predicting Liver disease.

B. Liver Patient Classification using Logistic Regression [2]

Reference [2] also focuses on Liver Patient prediction using Logistic Regression. It states that Logistic regression have proved its significance on this data set by achieving better classification accuracy than NBC (Naïve Bayes Classifier), C4.5 (Decision Tree), SVM (Support Vector Machine), ANN (Artificial Neural Network), and KNN (K Nearest Neighbors). The labeled data set is published on UCI machine learning repository as "Indian Liver Patient Records". The proposed Logistic regression-based approach achieved an accuracy of 74% with less than 1 minutes of execution time for building the classification model. Table 1 in explains about the accuracy of different classification algorithm which are used for this study.

	Classification algorithm	Accuracy
1	Logistic regression	74%
2	Ann	71.59%
3	C 4.5	68.69%
4	Knn	62.89%
5	Svm	58.26%
6	Nbc	56.52%

Table 1: Comparisons of different machine learning algorithms [2]

C. Logistic regression modeling for context-based classification [3]

Reference [3] is based on Logistic Regression approach to concept/document for Information Retrieval. The research has used 150 topics from the TIPSTER collection. The research states that the Logistic regression coupled with cross validation function is an effective machine learning algorithm. The research focuses on the topics such as Environment, Finance, International Economics, Law and Government, Military, Political, Science and Technology etc. The machine was trained to classify the topics when a document is given. The regression model is built by adding one variable at each iteration. The heuristic which was used is as shown in the below Figure 5.

1. For variable_name = 1; variable_name = n; ++variable_name;
   {calculate the crossvalidation criterion for variable_name added to the current best model}
2. Select the additional best predictor (variable_name with the lowest prediction error)
3. Set the best model to prior best model + next additional variable
4. Go to Step 1

Figure 5: Heuristic for building Regression Model [3]

D. Research on Logistic Regression Algorithm of Breast Cancer Diagnose Data by Machine Learning [4]

Reference [4] is detection of Breast Cancer by Logistic Regression. The results showed that when two features maximum texture and maximum perimeter are selected the classification accuracy is 96.5%. The cancer data set used in this research is taken from the Wisconsin breast cancer data set in the University of California at Irvine's machine learning data collection. The data set has a total of 569 records. Radius, Perimeter, Area, Texture, Compactness, Smoothness, Concavity, Concave points, symmetry, fractal dimension were taken to classify the data sets.

E. Using Random Forest Algorithm for Breast Cancer Diagnosis [12]

Reference[12] uses Random Forest algorithm for detecting breast cancer. It states that random forest algorithm can combine the characteristics of multiple eigenvalues and the combined results of multiple decision trees can be used to improve prediction accuracy. First K-training subsets are formed which forms K-decision trees. For random forest CART algorithm is used. It uses GINI coefficient method for node splitting. The data set used in this paper is taken from the Wisconsin breast cancer data set in the University of California at Irvine's machine learning data collection. The data set has a total of 569 records. This data set is same as that of Reference[3] and the accuracy is 95%.

F. Sentiment classification on big data using Naïve Bayes and logistic regression [5]

Reference [5] focuses on sentiment analysis of twitter reviews and classify them whether it is a positive or negative comment. Naïve Bayes Classification algorithm and Logistic Regression algorithm have been used, and they are compared for the accuracy, precision and computation time. The research has been implemented on the top of Hadoop along with Mahout. This research states that the logistic regression is better compared to Naïve Bayes in overall. The data set which has been used is the real time twitter review and it contains equal number of positive and negative reviews which makes the machine learning algorithm easy to classify the reviews. The below Table II gives the detailed accuracy of both the algorithms.

Parameters	Naïve Bayes	Logistic Regression
Dataset Size	6 MB	6 MB
Accuracy%	66.667	76.767
Precision%	69.23	73.575
Computation time (Mili-sec)	15732	73.575

Table 2 : Comparison result of two algorithms [5]

G. Using Decision Tree Classification Algorithm to Predict Learner Typologies for Project-Based Learning [6]

Reference [6] uses Decision Tree classification algorithm for predicting Learner's Typologies. The algorithm used is J48 Decision Tree. This paper also explains about 8 other research based on Decision Tree which is used to classify the performance of the students or predict the grade of the student etc. It states that most of the decision tree algorithms focused on the prediction of academic performance of students but there is significantly no research paper to predict the learner's typologies based on decision tree. The accuracy of this model is 96.19%. The research predicted the learner's groupings after getting the class labels through KDD Process.

H. Hierarchical decision tree classification of SAR data with feature extraction method based on spatial variations [7]

Reference [7] proposed a binary decision tree classification model for extracting the features on SAR (Synthetic Aperture Radar) data. The below Figure 6 shows the binary decision tree logic.

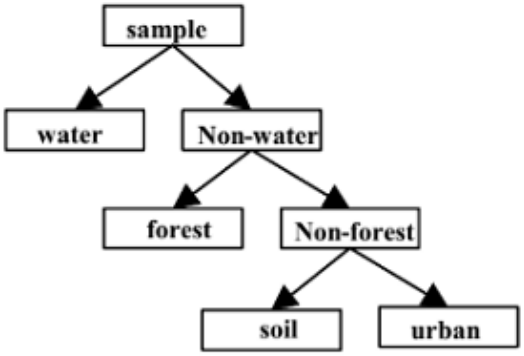

Figure 6: Binary Decision Tree Logic [7]

The aim of the research is to create a decision-making model which can be applied to every pixel of the given image and classify. It states that since only two classes are assigned to each node (binary tree) the computation time and the accuracy of classification process are improved. The final output of the given image is as shown in Figure 7.

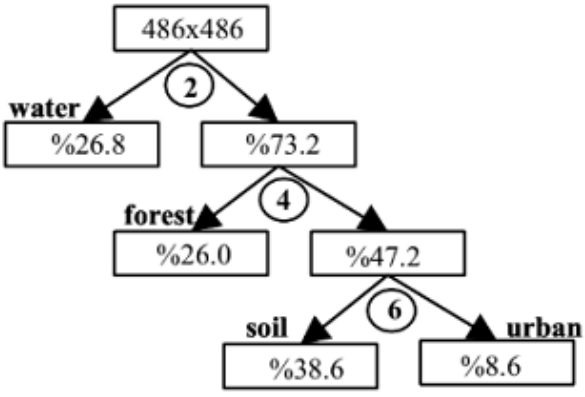

Figure 7: Classification Result [7]

I. Feature Selection for Classification Using Decision Tree [8]

Reference [8] in their study used Decision Tree algorithm for Pattern Recognition. It proposes a framework which determines the best eigenvectors of two main human positions which says whether they are standing or not standing. The preprocessing extracts the silhouette of the person using binary image extraction process. This study uses three rules to replace the p-dimensional feature space with a much smaller m-dimensional feature space. A collection of 200 images of various human postures are taken as the data set with no restriction on the facing. This research states that the KG-rule and Scree test can be used as guide in the optimal feature selection.

J. A Kind of Fuzzy Decision Tree Based on the Image Emotion Classification [9]

Reference [9] is based on Image Emotion Classification using Decision Tree Algorithms. It classifies the images using two of decision tree algorithms the Min-Ambiguity and the FID3 Algorithm. The accuracy of the image classified are displayed as shown in the Table 3.

Algorithm	No of images	No of correct	Precision
The Min-Ambiguity	250	229	0.916
The FID3-Algorithm	250	224	0.896

Table 3: The result of classification based on Image emoticon [9]

K. A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms [10]

Reference [10] proposed the use of decision tree C4.5. Algorithm, bagging with decision tree C4.5 algorithm and Bagging with Naïve Bayes algorithm to identify the heart disease of the patient and compares the effectiveness among them. This paper states that many decision tree algorithms is proposed for this research and one of the famous one is ID3. The choice of split attribute is based on information entropy. The algorithm that is used in this research is C4.5 which is an improved extension of ID3 in means of computing efficiency, deals with continuous variables, handles attributes missing values and avoid over fitting and perform other function. The accuracy of all the three approaches is given in the Table 4.

	DT C 4.5	Bagging + DT C 4.5	Bagging + NBS
Precision	78.93%	79.53%	82.50%
Recall	72.02%	73.72%	80.29%
F-Measure	75.32%	76.52%	81.38%
TPR	72.02%	73.72%	80.29%
FPR	15.52%	15.32%	13.75%

Table 4: Performance Result of three models [10]

L. Predicting Opioid Use Disorder (OUD) Using A Random Forest [11]

Reference [11] demonstrates how machine learning can be used to predict adult risk for Opiod Use Disorder. After the model is trained by numerous factors like gender, age, race, income, employment, education, first use of alcohol before 18 years, first use of marijuana before 18 years etc. The machine identified that the early initiation of marijuana emerges as a dominant factor for developing OUD in adult life. The random forest classifier can predict adults likes to develop OUD accurately (avg. sensitivity = 0.81, avg specificity = 0.76, avg AUC = 0.86).

M. A Machine Learning Approach for Heart Rate Estimation from PPG Signal using Random Forest Regression Algorithm [13]

Reference [13] proposes a new method of Random Forest Regression Algorithm to estimate the Heart Rate (HR) from wearable devices. During physical exercise the measurement is seriously affected due to motion artifacts. It proposes a multi-model machine learning approach (MMMLA). It first separates the noisy and non-noisy data using K-Means Clustering and then Random Forest Regression algorithm is used to estimate the HR.

N. Random Forest Algorithm for the Prediction of Diabetes [14]

Reference [14] paper's objective is to develop a system which can perform an early prediction of diabetes for a patient using Random Forest Regression algorithm technique. The data sets that are used are taken from UCI learning repository. The first step of this study was to select R features from the total features "m" where R << m. Then the next step was to find the best split point from the R features. Then the nodes are split into further children nodes until "I" number of noes have been reached. Then they have repeated the same process to build n number of trees and finally predicted the result. This study states that the accuracy was about 83% which is better compared to other algorithms like Naïve Bayes (80.37%), REP Trees (78.5%), Logistic Regression (77%).

O. Predicting soil heavy metal based on Random Forest model [15]

Reference [15] compares three machine learning models like Support Vector Machine, Random Forest and Extreme Learning Machine for prediction of soil heavy metal. Thirty samples of Cd, As, Pb concentration were taken and were divided into training and testing sample in the ratio of 2:1. Twenty training samples were trained separately by the above-mentioned algorithms and the remaining 10 were tested. This study states that for the prediction of heavy metals Random Forest is best compared to other algorithms.

P. Classification and Optimization Scheme for Text Data using Machine Learning Naïve Bayes Classifier [16]

Reference [16] proposes Naïve Bayes Algorithm for text classification in Hadoop Map Reduce instead of K-Means clustering. The study states that as Hadoop uses K-Means clustering the time complexity of the process increases. Text Classification is most important in Information Retrieval System. It says that the Hadoop Map Reduce for text classification has the problem of slow processing as there are too many data points in which K-means can become very complex. The study has an accuracy of 72.13% using Gaussian Naïve Bayes algorithm.

Q. Twitter sentiment classification using Naive Bayes based on trainer perception [17]

Reference [17] focuses on classification of Text document based on Naïve Bayes using N-Gram features. Feature extraction was an important step in this study where the words were being split into 2-gram, 3-gram and 4-gram words. A total of 1150 documents were used in this scope of study and these documents were divided into 5 separate categories. The study states that the best performance is obtained for 3-grams and the performance of the proposed approach drops with increasing number of grams. This was mainly due to increase in number of keywords when n-gram increases. The performance of this study achieved a success rate of 92%.

R. Classification of Text Documents based on Naive Bayes using N-Gram Features [18]

Reference [18] paper's objective is to classify the twitter reviews using Naïve Bayes algorithm. This study predicts the total number of positive reviews and the negative reviews based on the Trainer Perception. The study also indicates that many supervised techniques such as Nearest Neighbor, Naïve Bayes, TF-IDF and Support Vector Machine (SVM) are proven to have good results. The research states that the accuracy of this is about 90% with a standard deviation of 14%. It also says that by training and verifying the classification by the same person they could achieve high degree of accuracy using Naïve Bayes Technique.

IV. Conclusion

This paper demonstrates about the classification algorithms such as Logistic Regression (LR), Decision Trees (DT), Random Forest (RF), Naïve Bayes classification (NBS) which are used in machine learning technique. This research presents the overview of 18 research papers which are picked at random for this study. Out of 18 papers, 5 were related to medical field; 3 were related to Image Classification; 5 were related to prediction at an early stage and 5 were related to Sentiment Analysis. Table 5 displays the fields and the reference were mentioned in the []. Near to the reference the algorithm which are proposed for the research are mentioned.

Medical field	Image Classification	Prediction Analysis	Sentimental Analysis
[1] LR	[7] DT	[10] DT+NBS	[16] NBS
[2] LR	[8] DT	[11] RT	[16] NBS
[4] LR	[9] DT	[14] RT	[16] NBS
[12] RT		[15] RT	[16] NBS+LR
[13] RT		[6] RT	[16] LR

Table 5: The best accuracy of algorithm and their field using the references.

From the above table, we can infer that Logistic Regression algorithm is widely used in the medical field like the predicting the liver disease or breast cancer etc. For Image classification, decision tree algorithms are widely used as mentioned in the above table. As the decision tree algorithms are very effective in predicting future instance and the Random forest is nothing but a combined classifier that contains multiple decision trees all the prediction research have used either Decision Trees or Random Forest. And also, from the above table we can infer that the Naïve Bayes algorithm is widely used for Text classification.

The data set of research [4] and [12] are similar and both are predicting the breast cancer. The former uses the Logistic Regression while the latter uses Random Forest and the results have proved that Logistic Regression is better. Similarly, LR has proved better in predicting Liver disease compared with SVM, ANN, KNN, NBC in Reference [2]. Reference [14] states that Prediction has better results with Random Forest than NBC, REP, LR. In Reference [16] study

says that NBS is better algorithm than K-Means which are used in MapReduce.

REFERENCES

[1] k. Thirunavukkarasu, A. S. Singh, M. Irfan and A. Chowdhury, "Prediction of Liver Disease using Classification Algorithms," 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 2018, pp. 1-3.

[2] S. H. Adil, M. Ebrahim, K. Raza, S. S. Azhar Ali and M. Ahmed Hashmani, "Liver Patient Classification using Logistic Regression," 2018 4th International Conference on Computer and Information Sciences (ICCOINS), Kuala Lumpur, 2018, pp. 1-5.

[3] J. R. Brzezinski and G. J. Knafl, "Logistic regression modeling for context-based classification," Proceedings. Tenth International Workshop on Database and Expert Systems Applications. DEXA 99, Florence, Italy, 1999, pp. 755-759.

[4] L. Liu, "Research on Logistic Regression Algorithm of Breast Cancer Diagnose Data by Machine Learning," 2018 International Conference on Robots & Intelligent System (ICRIS), Changsha, 2018, pp. 157-160.

[5] Prabhat and V. Khullar, "Sentiment classification on big data using Naïve Bayes and logistic regression," 2017 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, 2017, pp. 1-5.

[6] E. Gyimah and D. K. Dake, "Using Decision Tree Classification Algorithm to Predict Learner Typologies for Project-Based Learning," 2019 International Conference on Computing, Computational Modelling and Applications (ICCMA), Cape Coast, Ghana, 2019, pp. 130-1304.

[7] N. G. Kasapoglu, B. Yazgan and F. Akleman, "Hierarchical decision tree classification of SAR data with feature extraction method based on spatial variations," IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No.03CH37477), Toulouse, 2003, pp. 3453-3455 vol.6.

[8] N. M. Tahir, A. Hussain, S. A. Samad, K. A. Ishak and R. A. Halim, "Feature Selection for Classification Using Decision Tree," 2006 4th Student Conference on Research and Development, Selangor, 2006, pp. 99-102.

[9] Z. Juanjuan, L. Huijun, L. Yue and C. Junjie, "A Kind of Fuzzy Decision Tree Based on the Image Emotion Classification," 2012 International Conference on Computing, Measurement, Control and Sensor Network, Taiyuan, 2012, pp. 167-170.

[10] M. C. Tu, D. Shin and D. Shin, "A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms," 2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, Chengdu, 2009, pp. 183-187.

[11] Wadekar, "Predicting Opioid Use Disorder (OUD) Using A Random Forest," 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), Milwaukee, WI, USA, 2019, pp. 960-961.

[12] Dai, R. Chen, S. Zhu and W. Zhang, "Using Random Forest Algorithm for Breast Cancer Diagnosis," 2018 International Symposium on Computer, Consumer and Control (IS3C), Taichung, Taiwan, 2018, pp. 449-452.

[13] S. S. Bashar, M. S. Miah, A. H. M. Z. Karim, M. A. Al Mahmud and Z. Hasan, "A Machine Learning Approach for Heart Rate Estimation from PPG Signal using Random Forest Regression Algorithm," 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox'sBazar, Bangladesh, 2019, pp. 1-5.

[14] K. VijiyaKumar, B. Lavanya, I. Nirmala and S. S. Caroline, "Random Forest Algorithm for the Prediction of Diabetes," 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), Pondicherry, India, 2019, pp. 1-5.

[15] W. Ma, K. Tan and P. Du, "Predicting soil heavy metal based on Random Forest model," 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, 2016, pp. 4331-4334.

[16] Venkatesh and K. V. Ranjitha, "Classification and Optimization Scheme for Text Data using Machine Learning Naïve Bayes Classifier," 2018 IEEE World Symposium on Communication Engineering (WSCE), Singapore, Singapore, 2018, pp. 33-36.

[17] M. N. M. Ibrahim and M. Z. M. Yusoff, "Twitter sentiment classification using Naive Bayes based on trainer perception," 2015 IEEE Conference on e-Learning, e-Management and e-Services (IC3e), Melaka, 2015, pp. 187-189.

[18] M. BAYGIN, "Classification of Text Documents based on Naive Bayes using N-Gram Features," 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), Malatya, Turkey, 2018, pp. 1-5.

[19] SAS. (2020) – Machine Learning : What is it and Why it matters https://www.sas.com/en_us/insights/analytics/machine-learning.html

[20] Adian Wison (2019) : A brief intoduction to Supervised Learning - https://towardsdatascience.com/a-brief-introduction-to-supervised-learning-54a3e3932590

[21] Wikipidia : Semi Supervised Learning - https://en.wikipedia.org/wiki/Semi-supervised_learning

[22] Yang S (2019) : TowardsDataScience - https://towardsdatascience.com/ introduction-to-na%C3%AFve-bayes-classifier-fa59e3e24aaf