European Journal of Molecular & Clinical Medicine
ISSN 2515-8260 Volume 7, Issue 2, 2020

COMPARATIVE STUDY OF MACHINE LEARNING ALGORITHMS FOR EVALUATING STUDENT ACADEMIC PERFORMANCE

Kanchan Sanyal
Asst.Teacher, Computer Application, Bhadrapur M.N.K High School, Birbhum, West Bengal, India

Dulal Kumbhakar
SACT-I, Department of BCA, Vivekananda Mahavidyalaya, Haripal, Hooghly, West Bengal, India.

Sunil Karforma
Professor, Department of Computer Science, The University of Burdwan,Golapbag, Bardhaman ,West Bengal, India

Abstract - Machine learning (ML) is transforming education and fundamentally changing teaching, learning and research. The ML technique helps the institution to utilize the resources in better ways and produces results in the best possible effective manner. The learning combines various processes like data preparation, classification, association, building models, training, clustering, prediction etc. to improve performance of students. It helps to the students to select any particular course based on their choices and previous performances. The main focus of this study is to analyze the various classification techniques over the educational data. The comparative study was conducted to predict the student performances based on some social variables (extracurricular activities, family education support, and desire for the higher education), previous exam grades and along with other attributes. In this paper, the Naive Bayes(NB) , Bayes Network(BN), Radial Bias Function (RBF), Multi-Layer Perceptron (MLP), Back Propagation Network(BPN), Random Forest(RF), J48, Radial Basis Function Network (RBFN) classification techniques were chosen for the experiment. After testing all the data over the mentioned classification we found that correctly classified instances percentage is 100% for Random Forest and it is highest compared to other classification algorithms.

Keywords: Machine learning, UCI repository dataset, classification Algorithms, J48, RBF, NB, BN, RBFN, RF, MLP.

1. INTRODUCTION

The term ML was first invented in 1959 by Arthur Samuel who is an American IBMer and explorer in the field of computer gaming and artificial intelligence [1]. ML is a subset of artificial intelligence (AI) that is used to enhance students' learning experience and can improve their progression and performance with learning. ML also helps teachers to create customized student's curriculum that provides the specific needs of the learners in efficiently [2].

The tasks of ML can be classified into the following three categories [3]:

- **Supervised learning:** Supervised learning is a powerful tool to classify and process data using machine language based on the classified labeled data. The data set is used in order to predict the classification of other unlabeled data through the use of machine learning algorithms. Here the mapping function from the input to the output is Y=f(X). Where x are the input variables and an output variable is Y. It can be predicted the output variables (Y) from new input data (x) to approximate the mapping function.
- **Unsupervised learning:** Unsupervised learning is a type of machine learning algorithm where only input data (X) and no corresponding have output variables. The most common unsupervised learning

method is cluster analysis, which is used for investigation data analysis to find hidden data grouping in order to learn more about the data.

- **Semi-Supervised Learning:** Semi-supervised learning is an approach to machine learning that combines a small amount of labeled data(Y) with a large amount of input data during training. So, a mixture of supervised and unsupervised techniques can be used.

In this part, we will briefly explain about the related works. Hilal Almarabeh has analyzed and evaluated the performance of the university students by applying different data mining classification techniques by using WEKA tool. Here NaiveBayes, Bayesian Network, ID3, J48 and Neural Network Different performance measures are used to compare the results and the experiment results shows that Bayesian Network classifier has the highest accuracy among the other classifiers [4]. S.V. Parmar and L. K. Sharma have predicted the student's academic performance based on social variable, pervious exam grades and other attributes in respect of the performance of the students. For this, they have used J48, Naïve Bayes, Bayes Net, Back propagation network and Radial Basis Function Network classification techniques for the experiment and this experiment result revealed that correct classify instance percent is 100% of Radial Basis Function Network regarding other classifiers [5]. S. Urkude and K. Gupta [6] have predicted the student's examination performance using statistical analytical testing method F1. In this context, F1 score is determined using support vector machines (SVM), Decision Tree and Naïve Bayes (NB) classification algorithms and the analysis is showed that the F1 score of SVM gives the better prediction accuracy rate compared to other two classification algorithms. Syeda Farha Shazmeen et al. [7] have worked with different data mining applications and various classification algorithms which are applied on different dataset to find out the accuracy rate of the algorithms. Here, data preprocessing techniques, feature selection and prediction of new class labels are applied for improving the performance. Jai Ruby and K. David have analyzed the prediction of the student's academic performance by applying classification algorithms such as ID3, REPTree, Simplecart, J48, NB Tree, BFTree, Decision Table, MLP and Bayesnet. In this case, the result shows the prediction accuracy rate above 68% for the dataset of the students [8]. Furthermore, G.Ayyappan has considered five ensemble classification techniques namely; Attribute Selected Classifier, Bagging, Classification through Regression, Weighted Instances Handler Wrapper and Multi Class Classifier for improving student's academic performance and preventing drop out also. Here, The data mining tool WEKA 3.8.1 is used and result showed that the Multi Class Classifier Classification technique has the highest accuracy rate (91.60%) compared to other techniques [9].

However, Machine Learning has a significant role in education. The Knowledge and the techniques which are used in machine learning will have a great impact on making a successful and effective decision to improve the student's academic performance evaluation. In this study, Seven Classification algorithms based on the UCI data set [10] which contains 395 instances along with 32 attributes have been used. We have made several comparison based on the accuracy, Error measurements, Performance metrics and reliability to find out the best possible classifier algorithm. Our experiment results show that the Random Forest (RF) has the best performance with 100% accuracy rate compared to other classifiers.

The paper is organized as follows. Section 2 describes the methodology of this comparative study paper. Section 3 describes error measurement and classification accuracy. Section 4 describes the experiment and results. Section 5 concludes the paper.

2. METHODOLOGY

A. DATA SET

The data set approach for student performance in secondary education of two Portuguese schools is used. The attributes of the data set include student grades, demographic, social and school related characteristics and it was gathered by using school reports and questionnaires. Two datasets are supplied regarding the student performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). In [10], the two datasets were customized under classification and regression tasks. There is the important that the target attribute G3 has a strong correlation with attributes G2 and G1. This happens because G3 is the final

year grade which is obtained at the 3rd period, while G1 and G2 correspond to the 1st and 2nd period grades. It is more difficult to determine the predict G3 without G2 and G1, but such prediction is much more important regarding student performance.

B. WEKA

This is a tried and tested open source machine learning software that has a built in graphical user interface for the implementation of machine learning algorithms. WEKA [11] is basically used for teaching, learning, research and industrial applications though its built-in machine learning tools. In this study, we have used WEKA 3.8.4 for accessing the different types of classification algorithms.

C. CLASSIFICATION ALGORITHMS:

WEKA 3.8.4 consists of different classification algorithms. These are listed below:

a. **Naive Bayes**

This is a popular classification algorithm based on conditional probability. This algorithm is best suited for real time prediction and text classification. In this context, Bayesian reasoning is used to decision making that deals with probability inference in order to gather the knowledge of prior events by predicting events through rule base [12].

b. Bayes Network

Bayes Network is a valuable algorithm which is used to identify the unfairness in the data set and applies accurate techniques to mitigate the unfairness in the data set.

c. Multilayer Perceptron

Multilayer Perceptron is one of the commonly used artificial neural network algorithms. It generates set of outputs from set of inputs with several layers of input nodes .The input and output nodes are interconnected via a direct graph [13].

d. Back Propagation Network

It is a common technique of teaching artificial neural networks how to implement a given assignment using supervised learning. The idea behind Back propagation is to reduce the errors by tuning the weights of the data until the artificial neural network learns the training data. It is also used to make the model reliable by increasing the generalization concept [14].

e. Radial Basis Function Network

Radial Basis Function Network is a particular type of artificial neural network that uses radial basis functions (RBF) as activation functions. The output of the network associates a weight value with each of the RBF neurons, and multiplies the neuron's activation by this weight and then adds it to the total response. Radial basis function networks have many applications including function approximation, time series prediction, classification, and system control [15].

f. J48

J48 is one of the most important classification algorithms which offer stability among the precision, speed, and clarification of the outputs. J48 algorithm is also used to generate the decision tree for tracking weak students. Therefore, J48 is used for both of classification and prediction operations [16].

g. Random Forest

Random forest or random decision forest is an ensemble technique that uses a collection of multitude decision trees during the training time. Each decision tree is exposed to a subset of the data and generates output as mode of class (classification) or mean prediction (regression) [17].

3. ERROR MEASUREMENT AND CLASSIFICATION ACCURACY

The different error measures are used to determine the classification technique. These are explained as follows [4, 18]:

I. Mean Absolute Error (MAE):

This error measures the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight. The MAE is denoted by the following formula:

MAE $= \sum_{i=1}^{n} \frac{|x-y|}{n}$, Where x is the predicted value and y is the actual value.

II. Root Mean Squared Error (RMSE):

RMSE is also known as root mean squared deviation that measures the average magnitude of the error. It's the square root of the average of squared differences between prediction and actual observation. RMSE is denoted by the following formula:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - y\right)^2}$$, Where X_i is the predicted value and y is the modeled value.

III. Relative Absolute Error (RAE):

RAE is the total absolute error and normalizes it by dividing by the magnitude of the exact value. The absolute error measures the difference between the actual value and individual measured value. The RAE denoted by:

$$\text{RAE} = \frac{\sum_{i=1}^{n}\left|X_i - a_i\right|}{\sum_{i=1}^{n}\left|\bar{a} - a_i\right|}$$, Where X_i is the predicted value, a_i is the actual value and $\bar{a}$ is the arithmetic mean.

IV. Root Relative Squared error (RRSE):

RRSE is relative to what it would have been if a simple predictor had been used and it takes the total squared error and normalizes it by dividing by the total squared error of the simple predictor. RRSE is denoted by the following formula:

$$\text{RRSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - a_i\right)^2}{\sum_{i=1}^{n}\left(a_i - \bar{a}\right)^2}}$$, Where x_i is the predicted value, a_i is the actual value, $\bar{a}$ is the Arithmetic mean.

Classification Terminology for Accuracy Prediction

Accuracy is termed as the number of the correct classification instances over total number given instances. The accuracy is determined through the confusion matrix. Table 1 shows the confusion matrix that shows the ways in which classification model is confused when it makes predictions [19, 20].

Confusion Matrix		**Actual (Target)**	
		Predicted x	Predicted y
Model	True x	TP	FN
	True y	FP	TN

Table 1: Confusion Matrix

- True Positive (TP): Observation is positive, and is predicted to be positive.
- False Negative (FN): Observation is positive, but is predicted negative.
- True Negative (TN): Observation is negative, and is predicted to be negative.
- False Positive (FP): Observation is negative, but is predicted positive.

From Table 1 TP Rate, FP Rate, Precision, Recall, F-measures and Accuracy can be calculated.

True positive Rate (TP Rate): TP Rate or Sensitivity measures the rate of positives of the correctly classified instances.

$$\text{TP Rate or Sensitivity} = \frac{TP}{TP + FN} \quad (1)$$

False Positive Rate (FP Rate): FP Rate measures the rate of negativity of correctly classified instances.

$$\text{FP Rate} = \frac{FP}{FP + TN} \quad (2)$$

Precision: The Proportion of correctly classified instances which are predicted positive.

$$\text{Precision, P} = \frac{TP}{TP + FP} \quad (3)$$

Recall: The Proportion of Correctly Classified instances which are actually positive.

$$\text{Recall, R} = \frac{TP}{TP + FN} \quad (4)$$

F-Measures: The F-measure is the harmonic mean of precision and recall. It is a variant of accuracy that not affected by negatives.

$$\text{F-Measures} = \frac{2 * P * R}{P + R} \quad (5)$$

Where P is the Precision and R is the Recall.

Accuracy: The closeness or the proximity of actual value with respect to measured value is defined as the Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (6)$$

Kappa Statistics: The Kappa statistic (or value) is a metric that compares an Observed Accuracy with an Expected Accuracy (random chance). The kappa statistic is used not only to evaluate a single classifier, but also to evaluate classifiers among themselves.

$$\text{Kappa Statistics} = \frac{OA - EA}{1 - EA} \quad (7)$$

Where OA = observed accuracy, EA = expected accuracy.

4. EXPERIMENT AND RESULTS

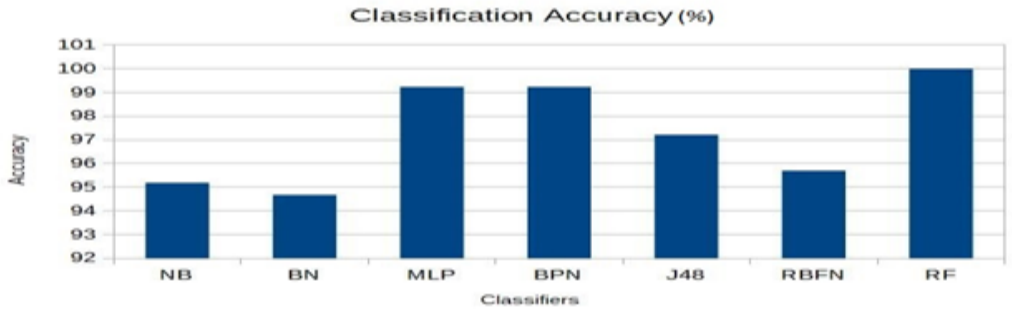
In this study, the experiment has been done on UCI Student Performance dataset consisting of 395 instances [10]. We have used 7 selected classification algorithms for this experiment. These are : Bayesian Net(BN), Naive Bayes(NB) ,J48, Neural Network (Multilayer Perceptron(MLP) , Back Propagation Net(BPN)) , Radial Basis Function Network(RBFN) , Random Forest(RF). The performances of the

mentioned classifiers using WEKA 3.8.4 are shown in the following Table 2. In the Table 2, Random Forest has the more correct classified instances than all other classification algorithms, which is usually the best accuracy classifier for our model.

Criteria	Classifiers						
	NB	**BN**	**MLP**	**BPN**	**J48**	**RBFN**	**RF**
Accuracy (%)	95.19	94.68	99.24	99.24	97.22	95.7	100
Correctly Classified Instances	376	374	392	392	384	378	395
Incorrectly Classified Instances	19	21	3	3	11	17	0

Table 2: Performances of the classifiers

The following graphical representation represents the Random Forest is the best classifier for student academic performance based on the UCI dataset. Comparing all other classification algorithms based on the accuracy parameter, Random Forest performs best among the all.



Graph 1: Accuracy performance

For the analysis of Errors, we have introduced RMSE and RRSE as parameters. Fig. 1 and Fig. 2 show the analysis of all the classifiers based on RMSE and RRSE. It is clearly observed that the Random Forest gives better performance for the given data set to minimize the error rate based on RMSE and RRSE for our model.

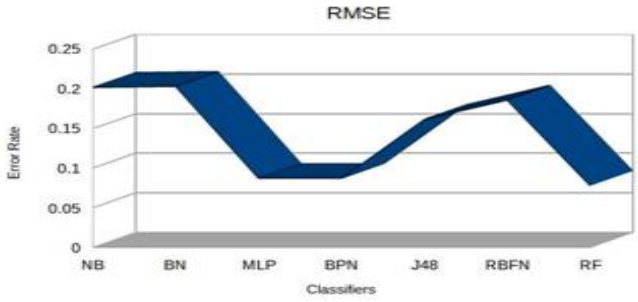


Fig. 1: RMSE metrics

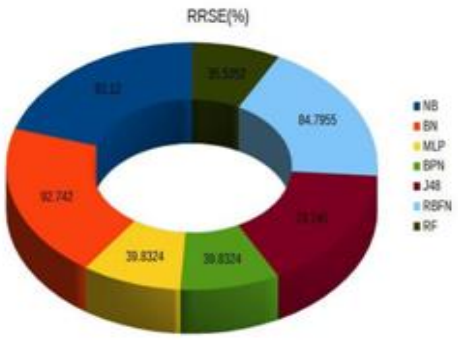


Fig. 2: RRSE Metrics

The following Table shows the Error measurements while testing is done for each classifier models. Random Forest (RF) minimizes the all type's error measurements except RAE in efficient manner. Here, two key error measurement factors RMSE and RRSE are considered in our model and the Neural Network (MLP, BPN) classifiers perform well while minimizing RAE. In rest of the measurement parameters like MAE, RMSE, RRSE, the Random Forest (RF) referred as the most promising in minimizing the Errors.

Criteria	NB	BN	MLP	BPN	J48	RBFN	RF
Mean absolute error(MAE)	0.0586	0.0801	0.0096	0.0096	0.0523	0.0694	0.0338
Root Mean Squared Error(RMSE)	0.202	0.2033	0.0873	0.0873	0.1617	0.1859	0.0779
Relative Absolute Error(%)(RAE)	59.69	81.5498	9.773	9.773	53.2619	70.675	34.4258
Root Relative Squared Error(%)(RRSE)	92.12	92.742	39.8324	39.8324	73.745	84.7955	35.5352

Table 3: Error Measurement

Metrics like True Positive (TP), False Positive (FP), Precision, Recall, F-measures are the key parameters to determine the accuracy of a classifier. The following figure shows the performance metrics regarding each classifier. Comparing and analyzing the results of each classifier, Random Forest outperforms compared to other classifiers.

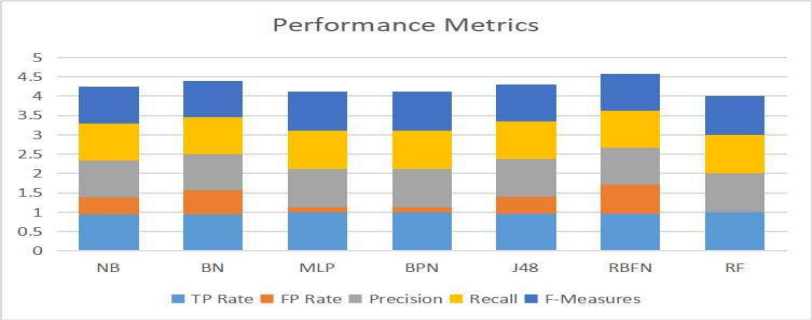


Fig. 3: Classifier Performance metrics

The weighted average of each metric for all the classifiers is presented in the following table.

Classifier	Performance Metrics				
	TP Rate	FP Rate	Precision	Recall	F-Measures
NB	0.952	0.429	0.953	0.952	0.952
BN	0.947	0.618	0.941	0.947	0.943
MLP	0.992	0.142	0.992	0.992	0.992
BPN	0.992	0.142	0.992	0.992	0.992
J48	0.972	0.427	0.97	0.972	0.969
RBFN	0.957	0.76	0.951	0.957	0.944
RF	1	0	1	1	1

Table 4: Weighted Average of Class label metric.

From Table 4, it is clear that RF gives highest TP Rate, Precision, Recall, F-Measures which indicates the highest level of Accuracy for RF. The FP Rate is 0(zero) indicating no incorrect classification found in RF.

Above all of these measures, we are interested to measure the inter classification or inter rate reliability between the categorical data prescribed in our working data set for each of the classification algorithm. For this purpose, we introduce another metric 'Kappa Statistics' which measures the inter rate reliability between the groups of data.

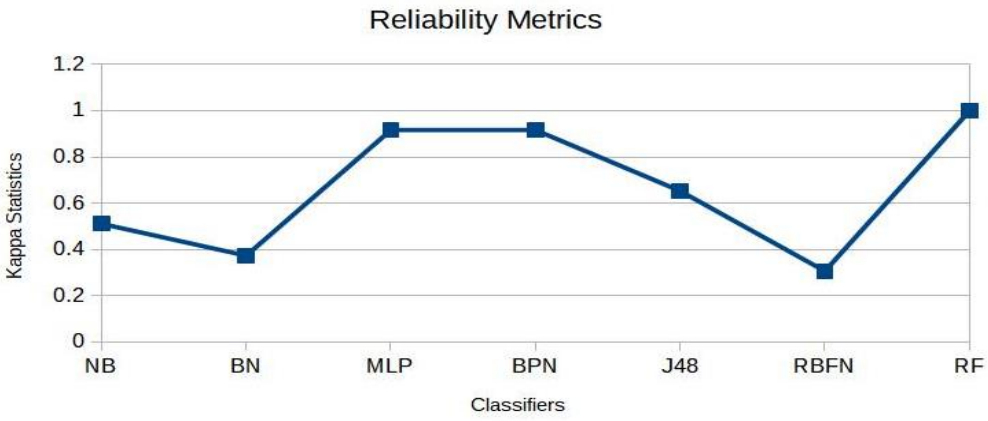


Fig. 4: Reliability Measurement

From the Fig. 4, it can be shown that RF is most promising for finding the inter reliability between the categorical or qualitative data in respect of our chosen data set.

5. CONCLUSION

It is very important to timely predict the student academic performance tendency in educational institutions. Several well-known unsupervised as well as supervised algorithms classification algorithms are used to improve the student academic performance regarding assignment evaluation in education sector. Classification plays an important role for identifying a particular pattern in the data and it can be applied in effective manner in educational data. The main aim of this comparative study paper is to analyze the various classification algorithms over the educational data to predict the student academic performances based on some social attributes such as extracurricular activities, family education support, Desire for the higher education and along with other attributes. In this study paper, we have considered the Naive Bayes(NB) , Bayes Network(BN), Radial Bias Function (RBF),Multi-Layer Perceptron (MLP),Back Propagation Network(BPN), Random Forest(RF), J48,Radial Basis Function Network (RBFN) classification techniques for this experiment. After several comparisons are made based on UCI data using above mentioned classification algorithms, we have observed that correctly classified instances percentage is 100% for Random Forest and it is the highest accuracy rate compared to other classification algorithms. In future, more data sets will be collected and compared their results with several classification techniques and some machine learning techniques like prediction, integration, regression, clustering and association.

REFERENCES

1. Wikipedia, “Machine learning”. Retrieved from: //https://en.wikipedia.org/wiki/Machine_learning#cite_note-Samuel-11.
2. Ibtehal Talal Nafea “Machine Learning in Educational Technology”. //http://dx.doi.org/10.5772/intechopen.72906.
3. Jason Brownlee, “Supervised and Unsupervised Machine Learning Algorithms”, blog, 2016. Retrieved from: //https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms//.
4. Hilal Almarabeh, “Analysis of Students' Performance by Using Different Data Mining Classifiers”, I.J. Modern Education and Computer Science, 2017, 8, 9-15.
5. S.V. Parmar and L. K. Sharma, “comparative study of supervised learning for student performance evaluation”, International Journal of Computer Engineering & Technology (IJCET),Volume 9, Issue 2, 2018, pp. 32–38.

6. Shubhangi Urkude and Kshitij Gupta, "Student Intervention System using Machine Learning Techniques", International Journal of Engineering and Advanced Technology (IJEAT), Volume-8, Issue-6S3, 2019, pp. 2061-2065.
7. Syeda Farha Shazmeen et al., "Performance Evaluation of Different Data Mining Classification Algorithm and Predictive Analysis", IOSR Journal of Computer Engineering (IOSR-JCE), Volume 10, Issue 6, 2013, pp. 01-06.
8. Jai Ruby and K. David, "Predicting the Performance of Students in Higher Education Using Data Mining Classification Algorithms - A Case Study", International Journal for Research in Applied Science & Engineering Technology (IJRASET), Vol. 2 Issue XI, 2014, pp. 173-180.
9. G.Ayyappan, "Ensemble Classifications for Student Academics Performance Data Set", Indian Journal of Computer Science and Engineering (IJCSE), Vol. 10 No. 1, 2019, pp. 31-34.
10. P. Cortez and A. Silva, "Using Data Mining to Predict Secondary School Student Performance", In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12.
11. Weka, University of Waikato, New Zealand. Retrieved from: //http://www.cs.waikato.ac.nz/ml/weka/
12. M. Mayilvaganan and D. Kalpanadevi, "Comparison of Classification Techniques for predicting the performance of Students Academic Environment", 2014 International Conference on Communication and Network Technologies (ICCNT), pp. 113-118.
13. Ijaz Khan et al., "Tracking Student Performance in Introductory Programming by Means of Machine Learning" 2019 IEEE, pp. 1-6.
14. V.Ramesh, P.Parkavi and P.Yasodha, "Performance Analysis of Data Mining Techniques for Placement Chance Prediction", International Journal of Scientific & Engineering Research, Volume 2, Issue 8, August-2011, pp. 1-8.
15. Wikipedia, "Radial basis function network". Retrieved from: //https://en.wikipedia.org/wiki/Radial_basis_function_network#:~:text=In%20the%20field%20of%20mathematical,basis%20functions%20as%20activation%20functions.&text=Radial%20basis%20function%20networks%20have,%2C%20classification%2C%20and%20system%20control.
16. Alaa Khalaf Hamoud et al., "Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis", International Journal of Interactive Multimedia and Artificial Intelligence, Vol. 5, No. 2, 2018, pp. 26-31.
17. Vladimir L. Uskav et al., Machine Learning – based Predictive Analysis of Student Academic Performance in STEM Education, IEEE, 2019, pp. 1370-1376.
18. J. Sujatha and S.P. Rajagopalan, "Performance Evaluation of Machine Learning Algorithms in the Classification of Parkinson Disease Using Voice Attributes", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 21 (2017) pp. 10669-10675.
19. Jesse Davis and Mark Goadrich, "The Relationship between Precision-Recall and ROC Curves", University of Wisconsin-Madison, 1210 West Dayton Street, Madison, WI, 53706 USA.
20. Julius Sim and Chris C Wright, "The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements", Physical Therapy, Volume 85, Issue 3, 1 March 2005, Pages 257–268, https://doi.org/10.1093/ptj/85.3.257.