International Journal of Engineering & Technology, 7 (4) (2018) 5986-5995

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i4.23542
Research paper

A model for improved performance prediction using ensemble-based hybrid classification approach on a multivariate student dataset

Anoopkumar M [1] *, Zubair Rahman A. M. J. Md [2]

[1] Research Scholar, Research and Development Centre, Bharathiar University, Coimbatore – 46, India
[2] Principal, Al-Ameen Engineering College, Erode, Tamilnadu, India
**Corresponding author E-mail: anoopkumar.m@gmail.com*

Abstract

Classification techniques have sensed substantial attention in Information Engineering and Technology for the performance prediction and optimisation since few decades. The discovered accuracy of the Classification Model helps the institutional practices and student's performances. In this paper, a novel Ensemble-based Hybrid Classification Approach (EHCA) has been proposed to be managed to produce improved performance prediction. The mining process with new attributes based on student behaviours has also been incorporated, since it creates a great impact on their academic performances. Moreover, the performance of the students is analysed with a set of classifiers in Educational Data Mining (EDM) namely, Naive Bayesian, Support-Vector-Machine (SVM) and J48. Futuristic-bound Ensemble approach is employed for enhancing the classifier performances. Here, the futuristic methods of ensembles of Bagging, Classification Boosting and Stacking are used for optimising the results with more precision. Further, the process of Ensemble-based Hybrid Classification is analysed and tested with the dataset collected from Kerala Technological Univer-sity-SNG College of Engineering (KTU_SNG). The results obtained are compared with the results obtained for utilized single classi-fiers and the EHCA on the basis of performance efficiency and classification accuracy. The work evidence the efficiency of the pro-posed approach and proves its reliability in Profound Performance Prediction and Optimisation.

Keywords: *Classification; Ensemble-Based Hybrid Classification; EHCA; Performance Prediction; Educational Data Mining EDM..*

1. Introduction

In present decade, the data mining techniques are widely used in different field ranges from marketing, finance, healthcare, security, government and education. Since it has been an ingrained field for finding meaningful pattern and relationships that make the user to derive knowledge and get larger value from the data. Data mining process can be effectively used in Educational systems for pattern discovery of students, student categorization and modelling, automation of prediction of their academic performances, which can be termed as the Educational Data Mining (EDM). EDM has been the greatly researched area in present scenario, though the results and observation have been utilized to provide ways to enhance the efficiency of the educational sectors and the accuracy of performance prediction of the traditional methodologies are not effective to the core to guarantee the earlier identification of student's criteria and intervention. Hence, the proposed model involves in the enhancement of student preservation and progression with higher knowledge and potential skills, which has a serious impact on society and economic growth.

As is auspicious, that the EDM is an emerging domain that concerned with proposing methodologies for surveying about the unique data categories that arrive from various educational backgrounds and utilizing those for betterment of students. With that concern, the proposed model Ensemble based Hybrid Classification Approach (EHCA). Based on the results obtained from the base classifiers, the further classification has been made with the ensemble-based technique. In current trends of data mining and machine learning, Ensemble modelling has been the most influential growth. The model includes the combination of multiple analytical classification models and then, fusing the results into single classification with more accuracy than the best of its elements [29]. The flow given Figure 1 shows the overall generic work process in the proposed model. Moreover, an ensemble of classifiers unifies the predictive results of multiple models based on two objectives.

1) The first objective is to enhance the accuracy rate of overall prediction results, when compared to the results obtained by using single classifier.
2) The second objective is to attain a considerable generalizability based on different dedicated classifiers involved in the process.

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

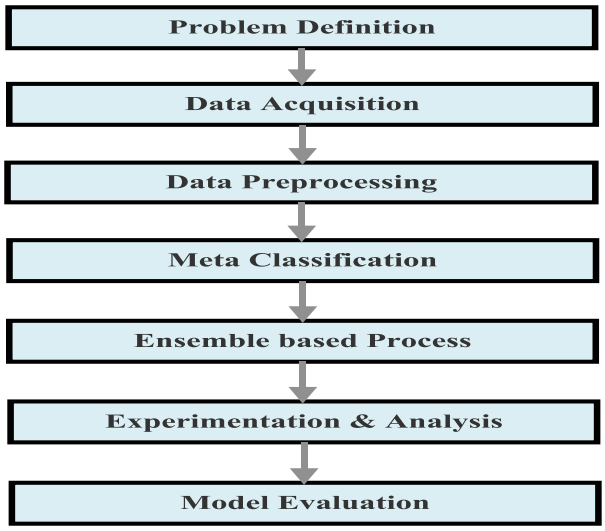


Fig. 1: Functions Involved in the Proposed Work.

Accordingly, an ensemble model can determine solutions in places where the single classification models may have some complications. Furthermore, the major essential principle is that an ensemble can choose a set of hypotheses out of some larger hypotheses space and fetch their prediction results together [23]. So, the philosophy of the ensemble classification model is that another base classifier that balances the errors made by one classifier. The general ensemble mechanism is portrayed in the Figure 2.

Here, for the effective construction of ensemble classifier, three major techniques are used: bagging, boosting and stacking. Bagging is a learning algorithm of ensemble model based on the bootstrap aggregation pattern. The second, boosting process concentrates on the instances of dataset that employs training each new instance from the previously identified errors for producing the predictive models. The final stacking is considered to be the stacked generalization, which is generated from the combination of several models in various ways by incorporating the meta-learner conceit. Moreover, WEKA is the tool used here for implementation, which is open source software that provides excellent framework for data mining and machine learning experimentations. With these notes, the proposed EHCA has been taken as exact tool for accurate student performance prediction with minimal error or misclassifications and also involved in identification of students at risk.

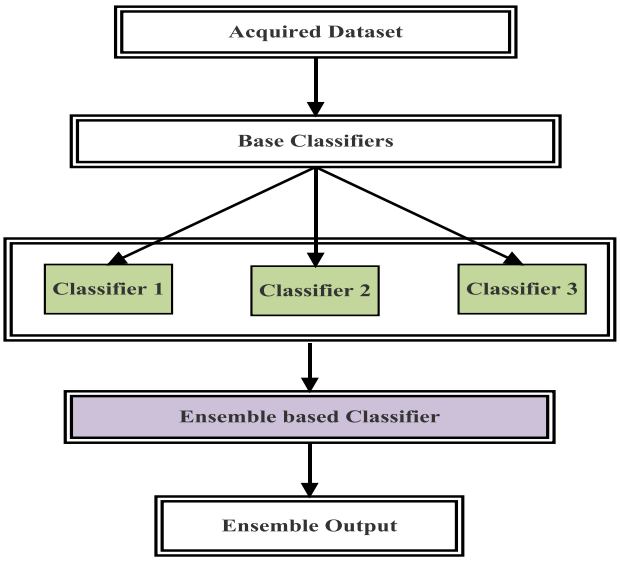


Fig. 2: Generic Ensemble Mechanism.

The remainder of this paper is organized into five sections: the second section gives a brief note about the problem statement. The third section deliberates about the related works using various classification methodologies and ensemble of classifiers in Educational Data Mining (EDM) in student performance analysis. The fourth section narrates the work process and framework of the proposed Ensemble based Hybrid Classification Approach (EHCA). Section five provides the experimental results and the comparative results of the proposed work and finally, section six concludes the paper with some points to future research scope.

2. Problem statement

The problem is defined specifically designing an Ensemble model for combining prediction efficiencies of classification algorithms such as Naive Bayes, SVM and J48 classifier for achieving better results in EDM. The developed model provides better rate of accuracy when compared with the results obtained using single classifier. The Ensemble based Hybrid Classification Approach (EHCA) has been analysed with various factors using the KTU_SNG student dataset containing 232 samples of with 45 attributes each (comprises both student's personal information and academic performance).

3. Related works

However, the importance of student's performance prediction has been on the utilization of their cognitive capability, log activities in Learning Management System (LMS) along with the student democratic attributes. Moreover, in [17], [21], [15] and [16], the authors have used demographic information as well as the student core for predicting the student performance, even many studies use machine learning methodologies such as Support Vector Machine (SVM), Artificial Neural Networks (ANN), etc.

In [20], [24], [2] and [8], the final grade of students is predicted using the log data derived from the web based system such as Learning Management System. The methodologies have used the attributes such as number of online sessions, login frequency, number of original and follow-up posts read or generated, content pages read. However, the most frequently used predictor variables derived from the LMS are completely based on the number of posts viewed, the amount spent online, study materials access and frequency of login.

In a different way, the works of [13], [26] and [11] utilized survey questionnaire techniques for collecting the student's personal and intrinsic data that are not clearly accessible over the database for the prediction of student's performance. Furthermore, the evaluation of the factors such as learning styles, personality traits, strategies of learning and motivational factors have also been analyzed. In a similar study processed in [25], three predictive models have been developed on the basis of survey-based retention methodology, framework of open data sources and internal database of institutions. Those methodologies have been compared and the performances are evaluates using the analytical models.

In general, there are number of studies in this research process have been accomplished that uses various methodologies and techniques for the evaluation of student's performance. These comprise Artificial Neural Networks (ANN) [21], decision tree algorithms [27], Naive Bayes [28] have used "key" democratic variables of students and their academic grade for the performance prediction of students in open university based on six different algorithms of linear regression, neural networks, model trees and support vector machine.

A hybrid algorithm has been implemented in [3] by using the concept of clustering and decision tree algorithm for the classification of the data samples. Specifically, the authors have used K-means clustering and Decision tree based classification for the proposal. In [30], the cluster and classification technique has been combined to improve the classification accuracy rate. Further in [14], Global Model for Classification (GMC) has been developed for enhancing the accuracy rate of classification in supervised learning. The design also includes the ensemble technique called bagging for better results.

In [22], it is stated that the J48 based classification has been the best decision tree induction algorithms for enhancing the predictive performance and also solving the disadvantages by pruning trees

method. The authors have introduced a novel decision tree algorithm based on the classifier called J48 and the reduced error pruning. Moreover, the pruning methodology reduces the computational complexity and over fitting in final classification, thereby improves the classification accuracy.

Further, the results of decision tree classifier can be enhanced by the ensemble method, since it this better than the results obtained from single method classification. However, the ensemble results are completely based on the base classifier selection [19]. Adaboost and bagging are the two ensemble concepts included in the paper for obtaining effective results of classification analysing diabetes patients.

In a different way, improved classification has been achieved for SONAR dataset using ensemble methods such as boosting, bagging and blending with J48 as base classifier in WEKA tool [1]. Multiclass classification has been performed using boosting and oversampling in [7]. The binarisation technique has been employed on the basis of One-Versus-All (OVA). Boosting is incorporated for solving the instances that are hard to learn and oversampling is for handling the issue of class imbalance problem. Furthermore, a new decision tree algorithm called NEC4.5 has been developed in [31]. This algorithm involved in training a neural network ensemble initially and the trained ensemble is induced to produce a new training set based in the desired class label replacement of the original training samples. Some additional training samples are also generated and added to the new training set of samples. The review work given in [4] has provided valuable information about the various classification models that are incorporated in the field of Educational Data Mining, for analyzing the student's academic performance and improve their results, in such a way improving the overall reputation of the educational institution.

In a different way, tuned J48 classification model has been described in [5]. Moreover, in that paper work, the classification results of models like Naive Bayes, Bayes Net, Multilayer Perceptron, SVM, REP Tree and Random Forest are evaluated and compared. The model produced results with 90.8% of accuracy rate. Further, Bound Model for Clustering and Classification (BMCC) is proposed and described in [6]. The work has been designed with the combined efficiencies of J48 decision tree based classification and k-means clustering. The process has been evaluated in WEKA tool and the results are compared with some conventional classification techniques. The model came up with 94.83% of accuracy for the base model and a slightly improved result of 97% with optimization.

4. Proposed methodology

The research scope concentrates on tracking and extracting student data for the case of enhancing the teaching methodologies and learning capabilities of students, which have been the significant goals in EDM. Hence, the capability to evaluate the academic performance of students is very crucial in the scenarios of educational domains. Perhaps, evaluation of student's academic performance is a challenging phase, since it depends on several factors such as personal, psychological, social-oriented, economic-centric and other dynamic environmental conditions. With those concerns, by analysing the all significant attributes, an efficient framework has been developed here to present the most informative knowledge representation.

Based on the valuable survey works and background research, the Ensemble based Hybrid Classification Approach (EHCA) has been developed. For the purpose of evaluation and experimentation, the recent real time data called KTU_SNG student dataset is collected from the Kerala Technological University-SNG College of Engineering (KTU_SNG). Moreover, some conversion process is carried out for the acquired dataset into a format, which is feasible for processing in WEKA tool. Moreover, it is to be given that the data set contains 232 samples with 45 attributes (i.e. total number of records are about 10440).

4.1. Ensemble based hybrid classification approach (EHCA)

The overall work process involved in Ensemble based Hybrid Classification Approach (EHCA) has been portrayed in the Figure 3. As an initial process, the required student information based on the designed attributes is acquired from the dataset. Then, it is given for data pre-processing that includes data cleaning and feature selection. The acquired data must undergo the aforementioned two processes before providing it for the process of data extraction. The overall intention is to predict the success and categorize Students_Nature under the Classes such as Outstanding, Excellent, Good, Average, low and very low. The following Table 1 illustrates some sample attributes along with their depictions and domain values obtained from the source dataset.

4.1.1. Data pre-processing

Data cleaning it the process involved in removing irrelevant attributes. There were attributes which seems to have less contributing in its nature or possesses a redundancy of existence. For an instance of student data, the attributes like community or financial status, etc. are not necessary to evaluate the intellectual performance of the students, since those domain values does not have any impact on that. But, those attributes are unavoidable on the dataset to be given as input to the proposed model. Accordingly, the missing values may also present in the dataset, which has to be removed for reducing the computational complexity over the mining process. Following that, the selection of significant parameters is taken for consideration. Feature selection is a process to select the most relevant attributes from the instances obtained from the complete dataset as a process of reducing dimensionality that helps for providing better classification results.

Moreover, two statistical methods are incorporated for determining the significance of each individual variable. Those are, Chi-square attribution evaluation for analyzing the qualitative student data based on the association between the instances and Information Gain attribute evaluation is to evaluate and treat the missing values into detached variables for effectiveness.

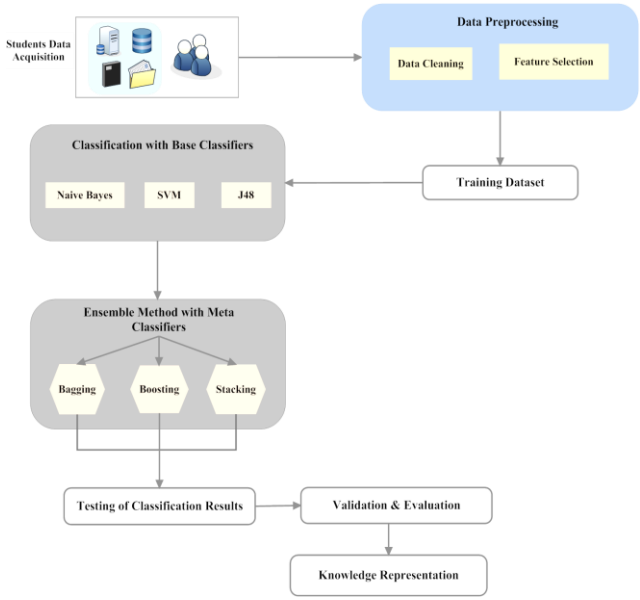


Fig. 3: Framework of Ensemble based Hybrid Classification Approach.

Table 1: Sample Attributes from the Obtained KTU_SNG Dataset for Processing EHCA

Attributes	Description	Possible Values
Gender	Student's sex	{M, F}
Age Group	Based on DOB	{between 17 to 25}
Blood Group	Group of Blood for the case of Emergencies	Varying for Samples
Family Income	Financial Status of the family for Scholarship purposes	{<10,000, (10,000-30,000), (30,000-50,000), >50,000}

Economically Backward	Financial Stability of family	{True, False}
Extracurricular	National/International level sports player, athlete, dancer or musician	{Yes, No}
Category	Communal Category for Scholarship purposes	{OC/FC, OBC, SC/ST}
Admission Type	Mode of admission for Scholarship	{Merit, Management}
High School Final Grade	Educational Excellence in School level	{A, A+, B, B+, C, C+}
Qualification Grade	Educational Excellence in Plus Two level	{low, average, good, best, excellent}
Program	Degree information	B.E/B. Tech
Branch	Department of the student	{CS, CE, ME, ECE, EEE, NASB}
Internal Marks	Class test performances	Based on internal class performances and Attendance
Lab Performance	Lab Activities	Based on Lab Attendance and Record books
Attendance	Total Attendance Percentage	Varies from 70 to 100 (change to 75 to 100)
Total Credits	Marks obtained	Given in credit rating
Students_Nature	Nature class of Students	{Outstanding, Excellent, Good, Average, Low, Verylow}
Result	Calculated from marks	{Pass, Fail}

4.1.2. Classification using base classifiers

Support Vector Machine (SVM) classifier, Naive Bayes classification model and the J48 decision tree are the three base classifiers used in this proposed model. Moreover, they are briefly described below:

Support Vector Machine (SVM)

As is well-known, SVM is a learning methodology for handling the issues over object detection and pattern recognition, and also for evaluation and mapping-up of linear and non-linear functions. Moreover, a set of hyperplanes are constructed in the high-dimensional space for better classification outcomes. The optimum hyperplane which maximize margin of support vectors are selected and used.

However, in the application of EDM, specifically, the predictive analytics is still considered to be limited. As mentioned in the Fig. 4, the two margins denote the distance between the training data termed as support vectors and the solid line called the hyperplanes. From that, the SVM classification algorithm involves in finding the best optimal hyper plane that classifies or separates the data exactly.

Naive Bayes Classification

Naive Bayes classification algorithm is a simple probabilistic classifier that computes a set of probabilities based on the frequency count and the combinations of values provided in the dataset. The Naive Bayes classifier is dependent on the Bayes' Theorem with the liberal assumptions between predictors.

J48 Decision Tree based Classification

J48 is an open source java implementation of C4.5 in the Weka tool. For classifying the new instance from the dataset, the classification algorithm initially required to create a decision tree oriented to the attribute or domain values of the accessible training data. Hence, when it obtains a set of instances or training dataset, it recognizes the attributes and their values that provide discrimination of several instances. Moreover, in this algorithm, the classification is processed continually till it attains the pure leaf; hence the results obtained must be as accurate as possible.

4.1.3. Ensemble classification using meta classifiers

Ensemble model based classification has become the most dominant development in Data mining at present. The methodology involves in uniting multiple classification models and then, producing the results into one generally more precise than the best of its base classifiers. An Ensemble based classifier combines the predictions from multiple methodologies based on two goals:

1) Boosting the overall classification accuracy compared to the single base classifier.
2) Achieving a better generalizability based on various specialized classifiers included.

Accordingly, an ensemble can determine solutions where a single classifier may have some difficulties. The main objective of the proposed EHCA is that of selecting a set of hypotheses out of all avail and unites their determinations into one. The common philosophy behind the EHCA is that one base classifier balances the error made by the other.

Here, three major approaches of Ensemble methods are used for the classifier construction, bagging, boosting and stacking. The approaches are explained below:

Bagging

Bagging is considered to be an ensemble learning algorithm that is working on the basis of bootstrap aggregation. It generally takes the base models within the ensemble and assigns equal weights to all. It is stated as an easy algorithm for implementation and also affords better performance results. The majority voting technique that pools all the classifier results together and the class with greatest vote for each instance is considered as the final result. The overall working demonstration of bagging technique is illustrated in the Figure 4.

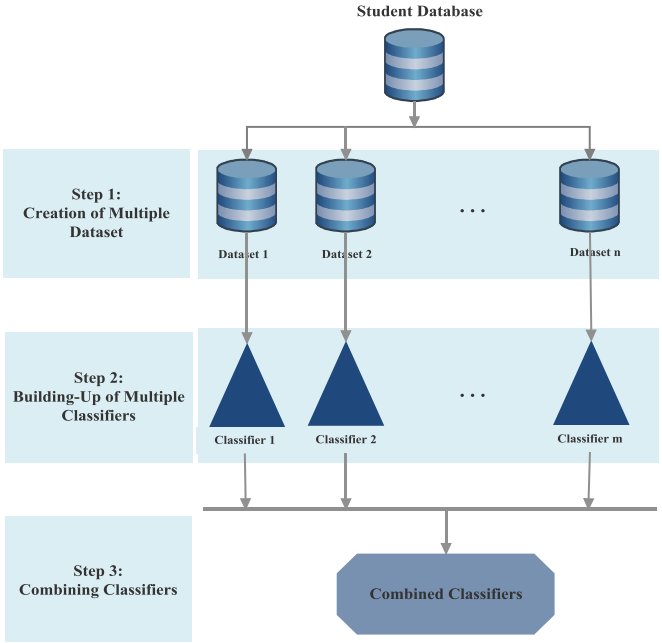


Fig. 4: General Work Process in Bagging.

Boosting

Boosting is a slightly different technique from bagging. It generally involves in enhancing the performance of any base classifier and also minimizes the error or misclassification of the weak one. Further, the boosting method concentrates on the instances for dataset that includes training each new instance models from the erroneous classification of the previous model for generating the predictive patterns. Like bagging, boosting can also be employed to similar algorithm and uses voting majority strategy for decision making.

Stacking

Stacking is an ensemble learning algorithm can also be stated as stacked generalization that combines multiple approaches with the inclusion of meta-learner concept. It works on the basis of constructing different learners that are utilized to develop an intermediate prediction that has become the input for the meta-classifier for final result. It also helps in reducing the generalization error rate and also enhances the performance accuracy. The pictorial representation of the process of stacking is given in the Figure 5.

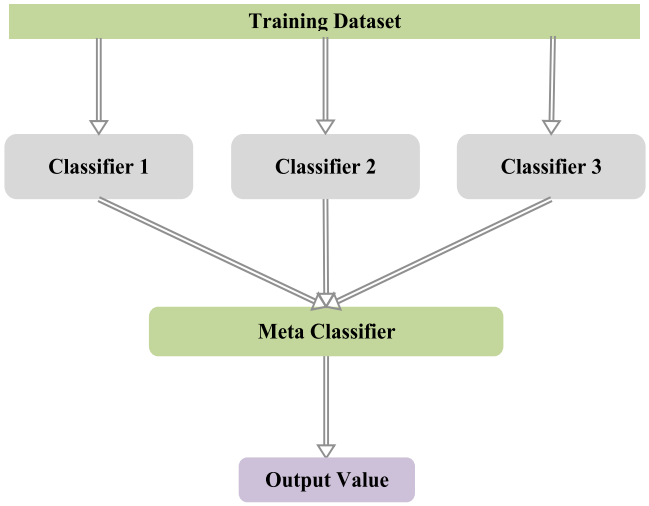


Fig. 5: Ensemble Model Using Stacking Approach.

The base stage models are trained depended on a full training set, then the meta model is trained with the outcomes of the base models as feature sets. The algorithm given in Table 2 runs over stacking.

Table 2: Stacking-Algorithm Summarized

Algorithm	Stacking
1:	Input: training data $D = \{x_i, y_i\}_{i=1}^{m}$
2:	Ouput: ensemble classifier H
3:	*Step 1: learn base-level classifiers*
4:	for $t = 1$ to T do
5:	learn h_t based on D
6:	end for
7:	*Step 2: construct new data set of predictions*
8:	for $i = 1$ to m do
9:	$D_h = \{x_i', y_i\}$, where $x_i' = \{h_1(x_i), ..., h_T(x_i)\}$
10:	end for
11:	*Step 3: learn a meta-classifier*
12:	learn H based on D_h
13:	return H

4.2. Decisive metrics for performance evaluation

The parameters such as specificity and sensitivity are used to measure the performance of the proposed methodology [10]. The Table 3 exemplifies a sample Confusion Matrix of analysis. The main factor that is used here is producing ROC (Receiver Operating Characteristics) graph that are significantly used for the determination of cut-off value for the specific process of accurate classification. Generally, the graph is plotted between the obtained values of True Positive Rate (TPR) and False Positive Rate (FPR).

Table 3: Sample Confusion-Matrix

Observed		Predicted Positive	Negative
	Positive	TP (# of TPs)	FN (# of FNs)
	Negative	FP (# of FPs)	TN (# of TNs)

The resulted outcome would be positive or negative. Here, the results of the ensemble classification are determined by the values of True Positive (TN), False Positive (FP), True Negative (TN) and False Negative (FN) values. Further, the performance analysis depends on the following evaluations.

The rate of sensitivity is described as the possibility of the classification result to be positive when there is the classification is appropriate and it is computed as follows.

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

Specificity rate is another decisive factor that is defined as the test result is negative, in a specific class.

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

The retrieval of positive predictions is called as precision. In particular, it is defined as the ratio of the predicted true positives out of all actually positive results. The formula is given as follows:

$$\text{Precision Rate (PR)} = \frac{\text{True Positive (TP)}}{\text{True Positive+False Positive}} \quad (3)$$

Accuracy is defined as the ratio of total number of accurately classified samples in to the total sum of acquired instances. Scientifically, it can be defined as,

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} * 100 \quad (4)$$

F-measure is a significant parameter for evaluating the proficiency of the proposed EHCA. It combines the TPR and the Precision Rates (PR) into an instant measure of performance. The equation is stated as follows,

$$\text{F_Measure} = \frac{2 * \text{TPR} * \text{PR}}{\text{TPR+PR}} \quad (5)$$

The precision and recall and its harmonic average are termed as the F1 score, where the best value of F1 score is at 1 (the perfect precision and recall) and the worst case is at 0.

Besides False Positive Rate (FPR) is also calculated for the false interpretation rate as follows.

$$\text{False Positive Rate} = 1 - \text{specificity} = 1 - \text{TNR} \quad (6)$$

T-Test: The final evaluation of this experimental work for classifier evaluation is done using Paired T-Test for classifiers. Weka workbench experiment options are used to test its effectiveness. Weak learning model and strong learning model can easily be identified from the outputs. Percent Correct is set as the comparison filed all through the test. This t-tester assumes the samples are independent.

5. Results

The experimental analysis has been carried out with the acquired KTU_SNG student data set which contains 232 samples with 45 performing attributes (i.e. total number of records are about 10440). The dataset is an initially pre-processed, contributory feature selected and feature extracted using Info-Gain-Ranker method to eliminate non-performing rank 0 attributes. However, majority of performance parameters are though now seeming ineffective cannot be eliminated as it has correlations and variations with other courses. Moreover, the data file is saved in the file format called Comma Separated Value (CSV) in MS-Excel and then, converted to the file format called Attribute Relation File Format (ARFF) inside the WEKA environment that is feasible for that environment.

Moreover, the proposed EHCA has been examined based on the decisive factors described in the section 4.2. The results are systematically verified and tested for a Paired T-Test in Weka favouring against various outcomes such as Percent Correct, F-Measure, Weighted average F-Measure…etc. A few are discussed in this following section. And, the results are compared with the individual base classifiers used in the process such as Naive Bayes, SVM and J48, and the previous work of the author called Bound Model for Clustering and Classification (BMCC).

However, in this work, the ensemble classifier technique is enforced for enhancing the predictive and the classification results with the combination of heterogeneous classifiers. The combination also includes the weighted accuracy rate and the classifier diversity. Hence, the results outperform the results obtained from the traditional classifiers. The experimental results show that the proposed Ensemble based classification model provides prospective results with highest rate of accuracy and precision, and the adaptability among the individual utilization of classification techniques used in base and ensemble classifiers. The Precision Rate and accuracy are computed on the basis of True Positive (TN), True Negative (TN), False Positive (FP) and False Negative (FN) as per the equations (3)

and (4), given in the section 4.2. Specificity, Sensitivity and the ROC analysis are also carried out and the result is interpreted. The following discussion on this help to explore more.

6. Discussions

The proposed work of Ensemble-based Hybrid Classification Approach (EHCA) have been experimented and implemented in WEKA. The Table 4, Table 5 and Table 6 depicts the extracted best results obtained for the proposed techniques of Ensemble method of classification such as bagging, boosting and stacking respectively for the class Result status.

Table 4: Results Obtained for Ensemble-Bagging

Total Number of Instances			= 232	Values in Percentage		
Correctly Classified Instances			225	96.98 %		
Incorrectly Classified Instances			7	3.02 %		
TP Rate	FP Rate	Precision Rate	Recall Rate	F-Measure	ROC Area	Class
0.984	0.083	0.978	0.984	0.981	0.979	Pass
0.917	0.016	0.936	0.917	0.926	0.982	Fail
Weighted Avg.						
0.97	0.069	0.97	0.97	0.97	0.98	

Table 5: Results Obtained for Ensemble-Boosting

Total Number of Instances			= 232	Values In Percentage		
Correctly Classified Instances			227	97.85 %		
Incorrectly Classified Instances			5	2.15 %		
TP Rate	FP Rate	Precision Rate	Recall Rate	F-Measure	ROC Area	Class
0.995	0.083	0.979	0.995	0.987	0.984	Pass
0.917	0.005	0.978	0.917	0.946	0.984	Fail
Weighted Avg.						
0.978	0.067	0.978	0.978	0.978	0.984	

Table 6: Results Obtained for Ensemble-Stacking

Total Number of Instances			= 232	Values In Percentage		
Correctly Classified Instances			227	97.85%		
Incorrectly Classified Instances			5	2.15 %		
TP Rate	FP Rate	Precision Rate	Recall Rate	F-Meaure	ROC Area	Class
0.995	0.083	0.979	0.995	0.987	0.977	Pass
0.917	0.005	0.978	0.917	0.946	0.977	Fail
Weighted Avg.						
0.978	0.067	0.978	0.978	0.978	0.977	

Each of the tables handle different ensemble methods with an equally competent result. ROC of ensemble-bagging, ensemble-boosting and ensemble-stacking are analysed to produce an average of 0.9803, a most considerable result of ROC. It is even generating as pointed in (1) and (6) an average True Positive Rate (TPR) of 0.974 and False Positive Rate (FPR) of 0.068. From the observation, for the method of bagging, the accuracy rate as in (4) is obtained as 96.98%. And, for boosting, it is enhanced as 97.85%. Finally, for stacking, the results are as similar as boosting and calculated as 97.85%. As an average, the overall accuracy rate of the proposed model EHCA is evaluated as 97.5%. It shows that the Ensemble based Hybrid Classification Approach provides better précised results on classifying instances than others and helpful for efficient student performance evaluation.

For the purpose of effective communication, it is to be considered that, the false positive rate or the value of sensitivity would be nearly 0. As per the results provided in the following figures the sensitivity rate obtained for the proposed model very close to 0, when compared to the remaining methods. Another supposition is also to be taken for Precision Rate, which is to be closer with the value 1.

The Fig. 6 illustrates the comparison for the classification models, at Weka implementation, such as Naive Bayes, SVM, J48, BMCC, Ensemble-Bagging, Ensemble-Boosting and Ensemble-Stacking, which are incorporated in the proposed work. From the comparative analysis, the Ensemble model with boosting provides higher rate of classification than others.

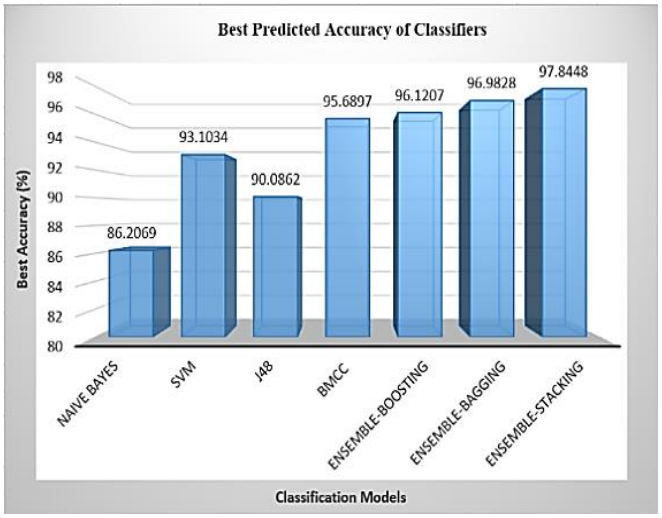


Fig. 6: Comparison of Accuracy Rate between Classification Models.

However, the experimental environment has shown considerable results, the later test conditions of the same might generate a rise or a fall in the result. Hence, the experimental result is tested for reasons of its performance using Paired T-Test as given in the Figure 7 and the result ascertains the notion and have shown a lifted result of each models.

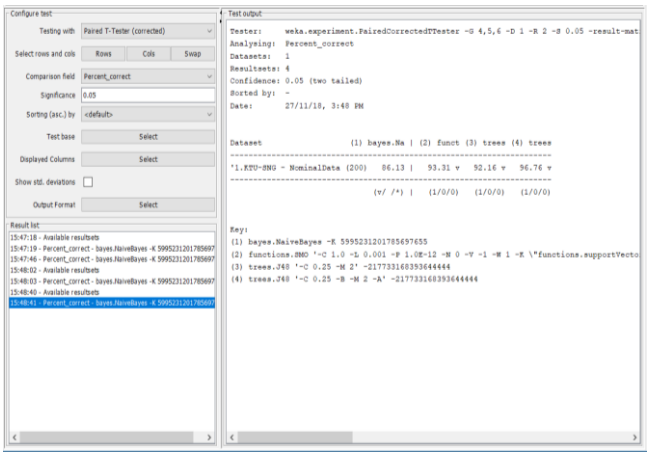


Fig. 7: Comparison of Percent Correct (Paired T-Test) Rate Between Base Classifications.

Further, the results and its operating environment conditions are analysed and a sample is illustrated in Figure 8.

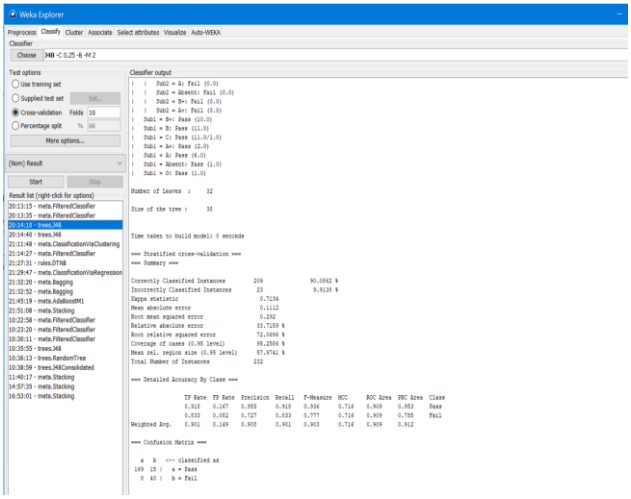


Fig. 8: Classification Results for J48 Classification In WEKA.

The base accuracy and the ensemble accuracy of each methods of the listed models are studied and a propagated accuracy is found. The SVM among the other have shown no major difference at the initial observations. The average incremental change it has produced is around 1.3 to 5.6 percentage except SVM techniques as is given in Table 7.

Table 7: Base and Ensemble Accuracy of Classifiers

Naive Bayes	Accuracy (%)
Base Accuracy	86.2069
Ensemble-Bagging,	87.069
Ensemble-Boosting	91.8103
Average	88.36207

SVM (SMO)	Accuracy (%)
Base Accuracy	93.1034
Ensemble-Bagging	92.6724
Ensemble-Boosting	93.1034
Average	92.95973

J48	Accuracy (%)
Base Accuracy	90.0862
Ensemble-Bagging	91.8103
Ensemble-Boosting	94.8276
Average	92.24137

BMCC	Accuracy (%)
Base Accuracy	95.6897
Ensemble-Bagging	96.9828
Ensemble-Boosting	97.8448
Average	96.8391

Profound futuristic methods of voting, bootstrap aggregation and stacking of ensemble learning have been utilised to create a setup environment for Hybrid Implementation proposed in EHCA and many of the result are as given as in table 8. The generated result is categorised into low medium and high accuracies for the convenience of analysis.

Table 8: Ensemble Based Hybrid Classification Accuracy Sample

EHCA Stages	Low Accuracy	Medium Accuracy	High Accuracy	Best Avg. Accuracy
EHCA-Bagging,	91.8	92.7	96.9	93.80
EHCA-Boosting	93.1	94.2	97.8	95.03
EHCA-Stacking	96.6	97	97.8	97.13

For providing the evidence of the efficiency of the proposed methodology, the comparative evaluation has been carried out. From the results, it is to be stated that, the ensemble method based on stacking has taken more time for execution, since it needed to perform more number of iterations. As it is mentioned, it provided 97.13 % of accurate results. Among all, boosting and stacking method of ensemble classification provide better results, as it is taken as rounded 97.8% maximum. The Figure 9, Figure 10, Figure 11 and Figure 12 portray the screen shots of the experimental implementation of base classifications and the proposed model and the execution in the Weka Knowledge Workflow tool. It helped to have a sovereign implementation of the design of the proposed work and has standardised to increase the easiness and the analysis of the results.

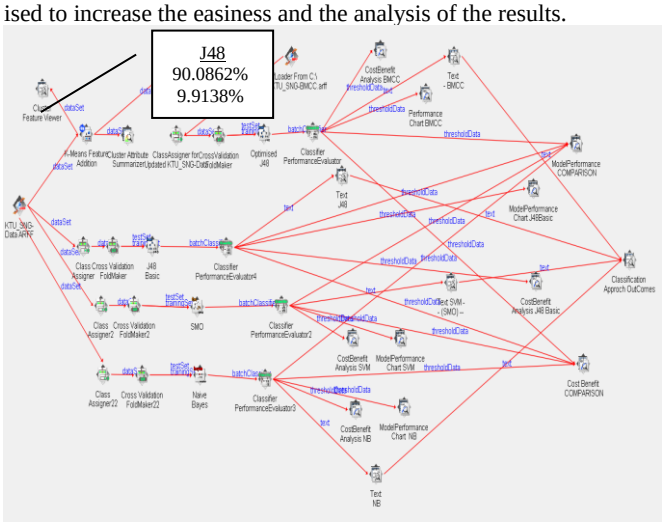


Fig. 9: Environment Setup for Base Classification and BMCC in WEKA.

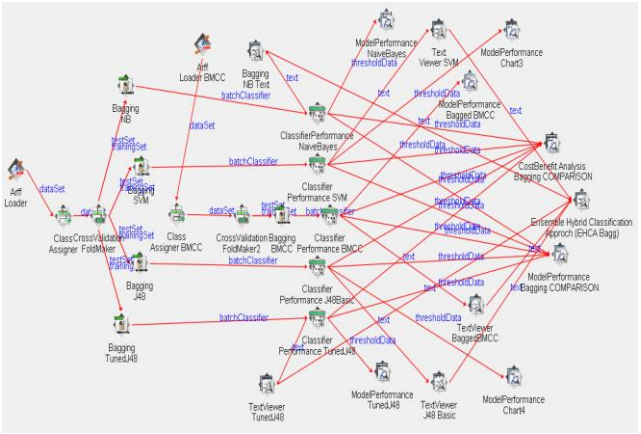


Fig. 10: Environment Setup for EHCA Ensemble-Bagging Classification in WEKA.

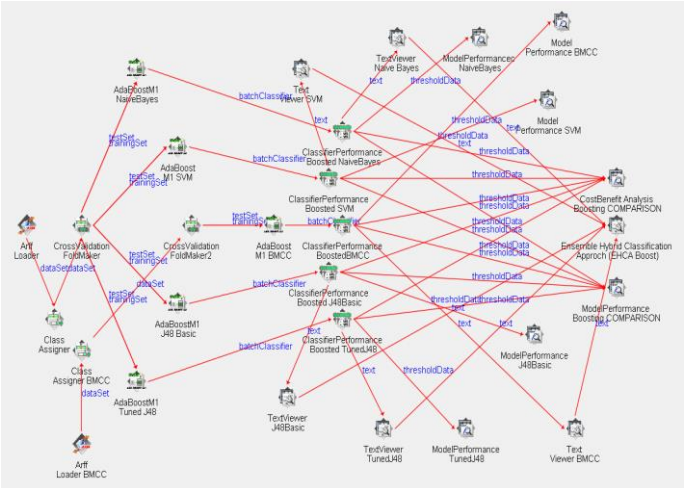


Fig. 11: Environment Setup for EHCA Ensemble-Boosting Classification in WEKA.

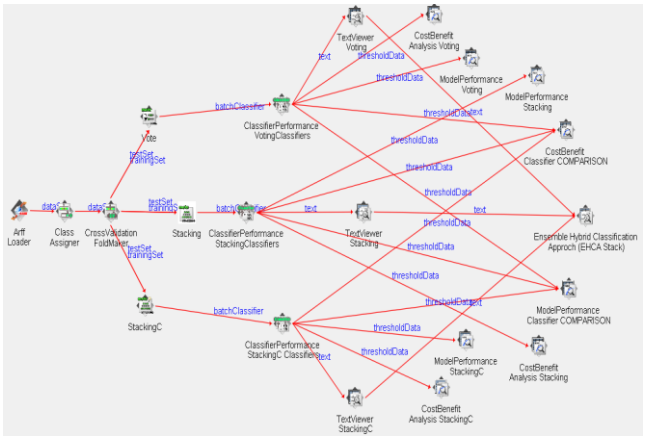


Fig. 12: Environment Setup for EHCA Ensemble-Stacking Classification in WEKA

Knowledge Workflow implementation has been used for the proposed work and the result of variations of models used in this and related models are depicted in Table 9

Table 9: Final Accuracy at Base Experiment

Method	J48	BMCC	EHCA
Accuracy (%)	90.08%	95.68%	96.98%

The result of all models are observed and analysed for the basic measures given in Section 4.2 for the evaluation of classifier performance. Sensitivity as in (1), Specificity as in (2) and Accuracy as given in (4) gain considerable interest in the same. The initial experimental result of BMCC and EHCA are being analysed and a graphical presentation is made in Figure 13.

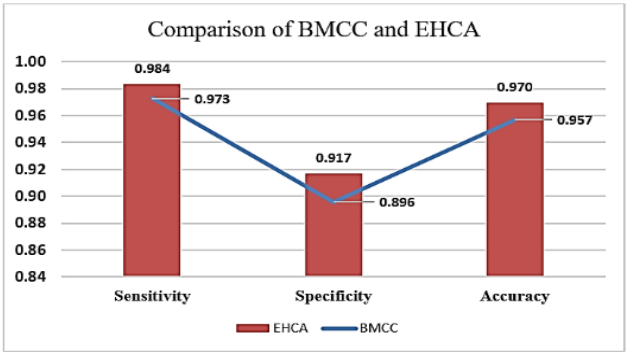


Fig. 13: Base Comparison Percentages of BMCC and EHCA.

The model could able to manage to minimise the Cost-Benefit at 69.38 gain threshold with 97.8448% accuracy as given in Figure 14.

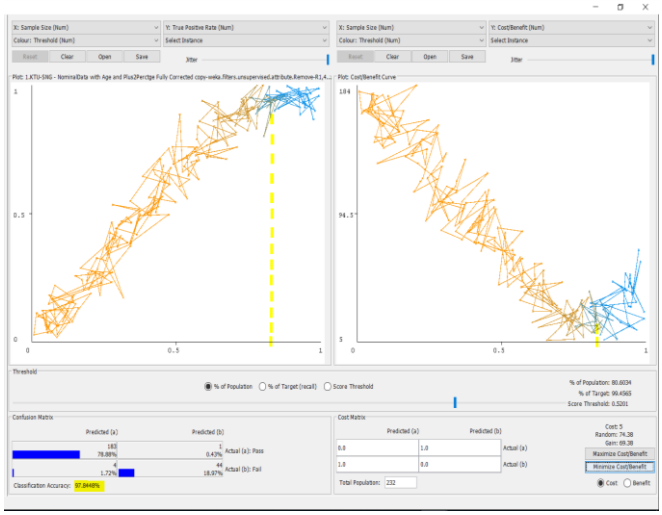


Fig. 14: Cost-Benefit of EHCA with Optimum Gain.

97.8448%

Previous work of the author [6] produced a max. of 97.41% during the experimental stage and a 95.08% average accuracy at the testing. Experimental the result of this work has slightly improved to 96.83% and the proposed model has come up with 97.50% avg. max of EHCA in this research. The Table 10 investigating research abreast that the proposed model conceit of its performance with others.

Table 10: Final Accuracy Comparison at Weka Implementation

Method	J48	BMCC	EHCA
Accuracy (%)	92.24%	96.83%	97.50%

As mentioned in section 4.2, Recall or True Positive Rate (TPR) and Specificity or True Negative Rate (TNR) of the BMCC and EHCA are tabulated and compared to analyse how best are they with one another according to their recall rate and specificity value as given in Figure 15. The sensitivity 1.0 is rated best, whereas 0.0 is the worst case.

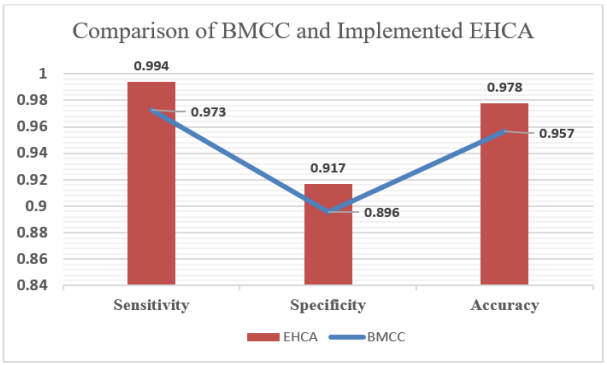


Fig. 15: Comparison Percentages of BMCC and Implemented EHCA.

ROC analysis as per Section 4.2 is being analysed using tools and the result of the same is given in Figure 16. Anything greater than 90% are often treated as good with proof of arguments. From the plots it is evident that the curve is tending towards '1' and hence possess an advisable merit.



Fig. 16: Sample ROC Area Curve of EHCA.

The values of ROC Area of the two competing models BMCC and EHCA have been analysed for its class accuracy. Though it found seemingly similar, EHCA came up with slightly better coverage. The same is given in the Figure 17.

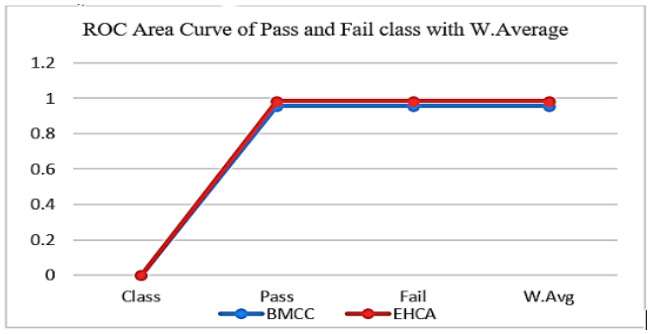


Fig. 17: ROC Area Curve of Pass and Fail.

Among the studies related to educational data mining by the author in paper [5], use of J48 and related methods substantiated in [18] [9] [12] have produced an accuracy of 86% and 92% where as a varied implementation has produced 99.87%. The author in the paper [6] have given a special mention about an unrealistic specific result produced by a combined model of J48 using optimisation with 99.6% accuracy. The work proposed in this study was produced highest 96.98 % (almost 97%) during the exploration and 97.5% avg. max at the implementation and a highest of 97.8%.

Further, the implication is stated that it is better to select the vital attributes vigilantly, which would be useful for the prediction of students at-risk from the obtained dataset and utilize sources to enhance their performance.

7. Conclusion and future work

Academic Excellence is being a significant concern among the academic institutions around the world. This paper has presented an efficient classification mode Ensemble based Hybrid Classification Approach (EHCA) based on Educational Data Mining with some new attributes oriented instances, includes students' academic, personal and behavioural feature. These kinds of features are correlated to the learner activities in EDM. Mostly used methods for the performance prediction and profiling are classification and statistical grouping. Here in this research, the classification methodologies like Naive Bayes, Support Vector Machine and J48 decision tree algorithms are used as base classifiers. In addition, Ensemble based methods are also incorporated for enhancing the accuracy of the classification results. For that, the common ensemble methods such as bagging, boosting and stacking methods are employed. The evaluated results show that proposed model produces high rate of accuracy that the rest. Hence, it can be stated that the ensemble based hybrid classification (EHCA) provides better results than the inclusion of single classification model.

It is expected that enhanced results may be followed in the future works, and work can be carried out in a way that use maximum data possible by combining some better reinforced data mining and sentiment measure of stakeholder interests with optimisation methods may provide a little more relevant values for datasets, which may help to enhance the performance and profiling accuracy.

Acknowledgement

This research was supported by SNG College of Engineering (SNGCE) Kerala affiliated to KTU, under the External Research Support Policy of institution by providing adequate data for this research.

References

[1] Aakash Tiwari, Aditya Prakash, "Improving classification of J48 algorithm using bagging, boosting and blending ensemble methods on SONAR dataset using WEKA", Int. J. of Engg. and Tech.l Res. (IJETR) ISSN: 2321-0869, Volume-2, Issue-9, September 2014, pp. 207-209.

[2] Agudo-Peregrina, Á.F., Iglesias-Pradas, S., Conde-González, M.Á. and Hernández-García, Á. (2014), "Can We Predict Success from Log Data in VLEs? Classification of Interactions for Learning Analytics and Their Relationship with Performance in VLE-Supported F2F and Online Learning", Comput.s in Hum. Behav., Vol. 31, pp. 542-550. https://doi.org/10.1016/j.chb.2013.05.031.

[3] Akanksha Ahlawat1, Bharti, Bharti Suri, "Improving Classification in Data Mining Using Hybrid Algorithm", IEEE Trans. on Know. and Data Engg, VOL. 17, pp.237-246.

[4] Anoopkumar M and A. M. J. Md. Zubair Rahman, "A Review on Data Mining Techniques and Factors Used in Educational Data Mining to Predict Student Amelioration" 2016 IEEE Int. Conf. on Data Min. and Adv. Comput.g (SAPIENCE), March-2016 pp. 122-133. https://doi.org/10.1109/SAPIENCE.2016.7684113.

[5] Anoopkumar M and A. M. J. Md. Zubair Rahman, "Model of Tuned J48 Classification and Analysis of Performance Prediction in Educational Data Mining", (IJAER) Int. J. of Applied Engg. Res. ISSN 0973-4562 Volume 13, Number 20 (2018) pp. 14717-14727.

[6] Anoopkumar M and A. M. J. Md. Zubair Rahman, "Bound Model of Clustering and Classification (BMCC) for Proficient Performance Prediction of Didactical Outcomes of Students" (IJACSA) Int. J. of Adv. Comput. Sci. and Appl.s, Vol. 9, No. 11, 2018, pp. 1-9. https://doi.org/10.14569/IJACSA.2018.091133.

[7] Ayon Sen, Md. Monirul Islam, Kazuyuki Murase, and Xin Yao, "Binarisation with Boosting and Oversampling for Multiclass Classification", IEEE Trans. on Cybernet.s, Vol. 46, No. 5, May 2016, pp. 1078-1091. https://doi.org/10.1109/TCYB.2015.2423295.

[8] Cerezo, R., Sánchez-Santillán, M., Paule-Ruiz, M.P. and Núñez, J.C. (2016), "Students' LMS Interaction Patterns and Their Relationship with Achievement: A Case Study in Higher Education", Comput.s & Edu., Vol. 96, pp. 42-54. https://doi.org/10.1016/j.compedu.2016.02.006.

[9] Cufoglu, M. Lohi and K. Madani, "A Comparative Study of Selected Classifiers with Classification Accuracy in User Profiling," 2009 WRI World Congress on Computer Science and Information Engineering, Los Angeles, CA, 2009, pp. 708-712. doi: 10.1109/CSIE.2009.954. https://doi.org/10.1109/CSIE.2009.954.

[10] D. L. Gupta, A. K. Malviya, Satyendra Singh, "Performance Analysis of Classification Tree Learning Algorithms", Int. J. of Comp. Appli. (0975 – 8887), Volume 55– No.6, October 2012. https://doi.org/10.5120/8762-2680.

[11] Fariba, T.B. (2013), "Academic Performance of Virtual Students Based on Their Personality Traits, Learning Styles and Psychological Wellbeing: A Prediction", Proc.-Soc.l and Behav.l Sci., Vol. 84, pp. 112-116. https://doi.org/10.1016/j.sbspro.2013.06.519.

[12] G. Kaur and A. Chhabra, "Improved J48 Classification Algorithm for the Prediction of Diabetes," Int. J. Comput. Appl., vol. 98, no. 22, pp. 13–17, 2014. https://doi.org/10.5120/17314-7433.

[13] Gray G., Mcguinness C., Owende P. (2016) "Non-Cognitive Factors of Learning as Early Indicators of Students At-Risk of Failing in Tertiary Education. In: Khine M.S., Areepattamannil S. (eds) Non-cognitive Skills and Factors in Educational Attainment". Contem.y Approach. to Res. in Learn. Innov.s. Sense Publishers, Rotterdam, pp. 199-237. https://doi.org/10.1007/978-94-6300-591-3_10.

[14] Hina Anwar, Usman Qamar, and AbdulWahab Muzaffar Qureshi, "Global Optimization Ensemble Model for Classification Methods", Hindawi Publishing Corp. Sci.c Worl. J. Vol 2014, Article ID 313164, pp. 1-9. https://doi.org/10.1155/2014/313164.

[15] Hoe, A.C.K., Ahmad, M.S., Hooi, T.C., Shanmugam, M., Gunasekaran, S.S., Cob, Z.C. and Ramasamy, A. (2014), "Analyzing Students' Records to Identify Patterns of Students' Performance", Res. and Inno. in Info. Sys. (ICRIIS), 2013 Int. Conf. on IEEE, Kuala Lumpur, pp. 544-547. https://doi.org/10.1109/ICRIIS.2013.6716767.

[16] Ikbal, S., Tamhane, A., Sengupta, B., Chetlur, M., Ghosh, S. and Appleton, J. (2015), "On early prediction of risks in academic performance for students", IBM J. of Res. and Develop.t, Vol. 59. No. 6, pp. 1-5. https://doi.org/10.1147/JRD.2015.2458631.

[17] Kotsiantis, S.B. and Pintelas, P.E. (2005), "Pred. Students' Marks in Hellenic Open University", Adv. Learn. Technol.s, ICALT 2005, 5th IEEE Int. Conf. on IEEE, Washington, DC, July 5-8, pp. 664-668.

[18] M. S. Halawa, M. E. Shehab and E. M. R. Hamed, "Predicting student personality based on a data-driven model from student behavior on LMS and social networks," 2015 Fifth International Conference on Digital Information Processing and Communications (ICDIPC), in IEEE, Sierre, 2015, pp. 294-299. https://doi.org/10.1109/ICDIPC.2015.7323044.

[19] Md. Rajib Hasan, Fadzilah Siraj, and Mohd Shamrie Sainin, "Improving ensemble decision tree performance using Adaboost and Bagging" AIP Conf. Proc.s 1691, 030008 (2015). https://doi.org/10.1063/1.4937027.

[20] Minaei-Bidgoli, B., Kashy, D.A., Kortemeyer, G. and Punch, W.F. (2003), "Predicting Student Performance: An Application of Data Mining Methods with an Educational Web-Based System", IEEE, Front.s in Edu. FIE 2003 33rd Annual, Westminster, CO. Vol. 1, pp. 1-13. https://doi.org/10.1109/FIE.2003.1263284.

[21] Oladokun, V.O., Adebanjo, A.T. and Charles-Owaba, O.E. (2008), "Predicting Students' Academic Performance Using Artificial Neural Network: A Case Study of an Engineering Course", The Pacific J. of Sci. and Technol., Vol. 9. No. 1, pp. 72-79.

[22] Prerna Kapoor1, Reena Rani, "Efficient Decision Tree Algorithm Using J48 and Reduced Error Pruning" Int. J. of Engg. Res. and Gen. Sci. Volume 3, Issue 3, May-June 2015, pp. 1613-1621.

[23] Rokach, L. "Ensemble-based classifiers," Artif. Intelli. Rev., Vol. 33, pp. 1-39, 2010. https://doi.org/10.1007/s10462-009-9124-7.

[24] Romero, C., López, M.I., Luna, J.M. and Ventura, S. (2013), "Predicting Students' Final Performance from Participation in On-Line Discussion Forums", Comput.s & Edu.n, Vol. 68, pp. 458-472. https://doi.org/10.1016/j.compedu.2013.06.009.

[25] Sarker, F., Tiropanis, T. and Davis, H.C. (2013), "Exploring Student Predictive Model That Relies on Institutional Databases and Open Data Instead of Traditional Questionnaires", Proc.s of the 22nd Int. Conf. on WWW. ACM, pp. 413-418. https://doi.org/10.1145/2487788.2487955.

[26] Sembiring, S., Zarlis, M., Hartama, D., Ramliana, S. and Wani, E. (2011), "Prediction of Student Academic Performance by an Application of Data Mining Techniques", Inte. Conf. on Manag.t and Artif.l Intelli. IPEDR, Vol. 6, pp. 110-114.

[27] Swamy, M.N. and Hanumanthappa, M. (2012), "Predicting Academic Success from Student Enrolment Data Using Decision Tree Technique", Int. J. of App. Info. Sys., Vol. 4. No. 3, pp. 1-6. https://doi.org/10.5120/ijais12-450654.

[28] Trstenjak, B. and Donko, D. (2014), "Determining the Impact of Demographic Features in Predicting Student Success in Croatia", 37th Int. Conv.n on Info. and Comm.n Tech., Electronics and Microelectronics (MIPRO), IEEE, pp. 1222-1227. https://doi.org/10.1109/MIPRO.2014.6859754.

[29] Wang, X. "Modeling Entrance into STEM Fields of Study among Students Beginning at Beginning at Community Colleges and Four-Year Institutions," Res. in High. Edu.n, 54 (6), 664-669, September 2013. https://doi.org/10.1007/s11162-013-9291-x.

[30] Yaswanth Kumar Alapati, "Combining Clustering with Classification: A Technique to Improve Classification Accuracy" Int. J. of Comput. Sci. Engg. (IJCSE), Vol. 5 No.06 Nov 2016, pp. 336- 338.

[31] Zhi-Hua Zhou and Yuan Jiang, "NeC4.5: Neural Ensemble Based C4.5", IEEE Trans. on Knowl. and Data Engg., VOL. 16, NO. 6, Jun 2004, pp. 770-773. https://doi.org/10.1109/TKDE.2004.11.