Bound Model of Clustering and Classification (BMCC) for Proficient Performance Prediction of Didactical Outcomes of Students

Anoopkumar M[1*]
Research and Development Centre,
Bharathiar University,
Coimbatore – 46, Tamilnadu, India

A. M. J. Md. Zubair Rahman[2]
Principal, Head of the Institution,
Al-Ameen Engineering College,
Erode, Tamilnadu, India

Abstract*—In this era of High-Performance High computing systems, Large-scale Data Mining methodologies in the field of education have become a convenience to discover and extract knowledge from Databased of their respective educational archives. Typically, all educational institutions around the world maintain student data repositories. Attributes of students such as the name of the student, gender of student, age group (date of birth), religion, eligibility details, academic assessment details, etc. are kept in it. With this knowledge, in this paper, didactical data mining (DDM) is used to leverage the performance prediction of student and to analyse it proactively. As it is known, Classification and Clustering are the liveliest techniques in mining the required data. Hence, Bound Model of Clustering and Classification (BMCC) have been proposed in this research for most proficient educational data mining. Classification is one of the distinguished options in Data Mining to assign an object under some pre-defined classes according to their attributes, and hence it comes under a supervised learning problem. On the other side, clustering is considered as a non-supervised learning problem that involves in grouping up of objects with respect to some similarities. Moreover, this paper uses the dataset collected from Kerala Technological University-SNG College of Engineering (KTU_SNG) for performing the BMCC. An efficient J48 decision tree algorithm is used for classification and the k-means algorithm is incorporated for clustering here and is optimised with Bootstrap Aggregation (Bagging). The implementation has been done and analysed with a data mining tool called WEKA (Waikato Environment for Knowledge Analysis), and the results are compared with some most used classifications such as Bayes Classifier (NB), Neural Network (Multilayer Perceptron MLP) and J48. It is provable from the results that the model, proposed in this provides high Precision Rate (PR), accuracy and robustness with less computational time, though the sample data set includes some missing values.

Keywords*—Classification; clustering; precision rate; accuracy; j48 decision tree; bagging; educational data mining

I. INTRODUCTION

In general, data mining is the process of examining the large databases for extracting the new or required information. It can also be stated that it is the procedure of effective classification of folders with respect to the specified data patterns that are acquired from the dataset. There are enormous algorithms have been developed for retrieving the valuable information and knowledge discovery patterns, which is functional for decision support. In another word, data mining can also be known as KDD, which is Knowledge Discovery in Database, handles with the non-trivial retrieval of inherent, completely novel and valuable information from the databases [4]. The Fig. 1 contains the typical steps involved in the data mining process of retrieving valuable information.

In the present decade, there are rising research scopes in utilizing data mining techniques for education, henceforth for Didactical Data Mining (DDM) a similar one to Educational Data Mining (EDM). This newly developing interdisciplinary field EDM involves knowledge extraction from the data obtained from the educational environments [11]. The main intention is to have a good understanding about the students learning standards and identifying the procedure in which they study to enhance the educational results, to process with the obtained outcomes based on the educational phenomena. Moreover, the educational information system can store a wide range of prospective data that would be collected from the historical and multiple sources exist in the databases of distinctive educational institutions. The collected data are from different sources, formats and also at variant granularity levels and that may contain the personal or academic details of the students. In another option, the huge data can also be collected from the e-learning systems that are already provided with a huge amount of data from various institutions. For handling those data effectively, there is a need for effective implementation of an EDM technique. The following Fig 2 depicts the basic attributes that have to be combined with the Didactical Data Mining (DDM) or EDM system in addition to the typical data mining functions.

The main goal of the educational institutes is to afford a good quality of education and to enhance the power of managerial strategies. In order to accomplish the highest level of educational domain, analysing the major attributes of student's performance and discovering knowledge is a significant part. This can provide useful and beneficial recommendations to the academicians and the management to improve their decision-making structure, the performance of the students and teaching abilities of the tutors or faculties [7].

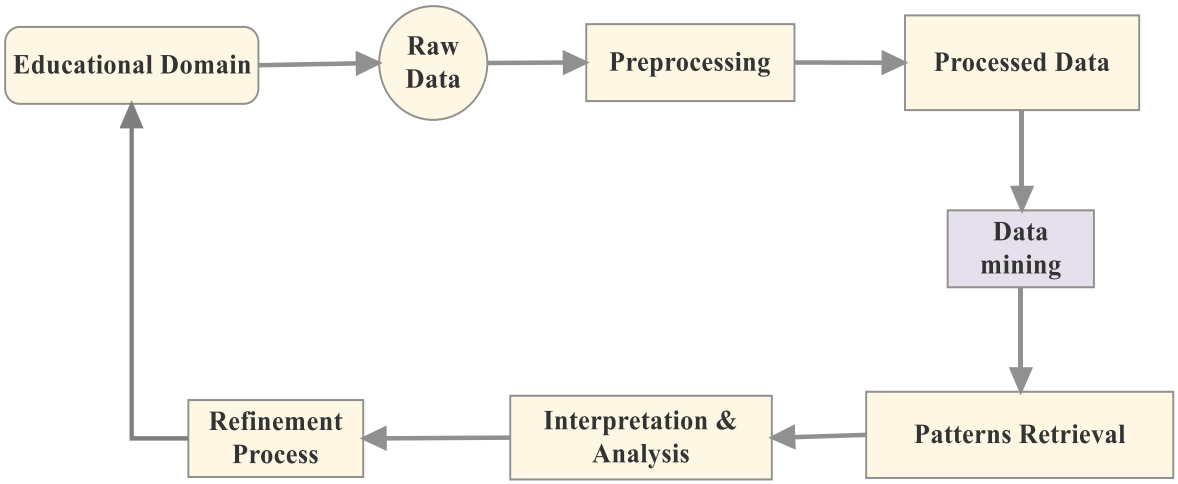

Fig. 1. Functions Involved in Typical Data Mining.

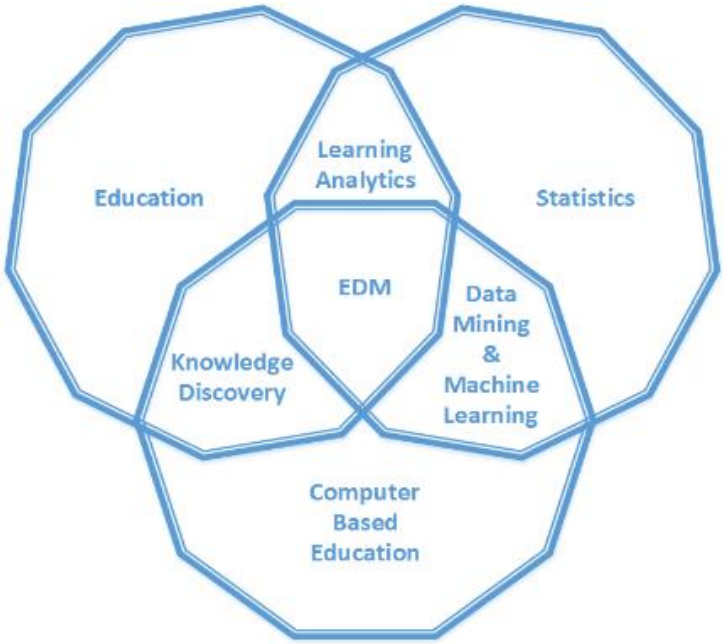

Fig. 2. Basic Factors of DDM/ EDM.

In general, EDM uses many data mining algorithms like decision tree algorithm, neural networks, rule induction, naive Bayesian and so on. On using these methodologies, the powerful knowledge can be retrieved using classification, association rules and clustering. In this paper, the main objective is about the combination of Classification and Clustering, the well-known model for enhancing the precision of the results. The framed work Bound Model of Clustering and Classification (BMCC), concentrates on the classification of yearly-growing data sets using J48 decision tree classifier since it is comparably faster and more exact than the other classification techniques. And also, a decision tree can be transformed into simple and understandable classification rules. Furthermore, clustering is the course of grouping up of similar objects and that comes under the unsupervised pattern classification and K-means algorithm is incorporated for clustering. In this paper, an efficient machine learning tool called WEKA (Waikato Environment for Knowledge Analysis) has been used for the implementation of classification and clustering algorithms. The dataset acquired here is the KTU_SNG student dataset that contains 232 samples with 60 attributes for each. The proposed (BMCC) model is implemented with this data set in WEKA environment for producing précised results to analyse and monitor the student's performance and take organisational decision for the betterment of students.

The leftovers of this paper are systematized as follows: Section 2 describes the problem statement in short. Section 3 considers the related works based on Educational Data Mining. Section 4 presents the operations involved in the proposed (BMCC) model. The results and discussions are given in section 5 and finally, section 6 concludes the work with some key points for future enhancement.

II. Problem Statement

The problem is defined specifically designing a model for predicting the performance of KTU Students. The work [21] carried out in the field of EDM have identified the use of a Tuned J48. There were studied which used clustering [6] along with classification to enhance its effectiveness [23]. Here in this, it is enquiring the scope of these combinations in predicting student performance in KTU by combining the Tuned J48 classification algorithm with the K-Means. The bound model has been analysed with various factors referred in section IV and using the KTU_SNG student dataset containing 232 samples of with 60 attributes each (comprises both personal information and academic performance). This is a research enquiry about how best the model can serve the expectation of stakeholders since it is new University established in the year 2015 and the status of Datasets are also in an infant stage.

III. Related Works

Because of the potentials of data mining to educational domains, it has become the most efficient research area. There are so many works have been done based on this. In [1] a valuable survey work on data mining has been provided. The authors have discussed the EDM as a hopeful research [16] field and some particular advancement that are not provided by other fields. In another work [12], a case study is provided, in which educational data mining has been used to examine the learning behaviour of students. The main intention of the study was to describe the significance of the EDM in enhancing student performance in higher secondary education. The authors have used the data from the warehouse and the gathered data comprises both the personal and academic records of the students. Data and records gathered up from e-learning management systems (e-LMS) have also been incorporated. The described the study work on the source of applying some classification techniques using decision tree algorithms and association rules for obtaining different kinds of knowledge. All outliers had been detected for analysis and finally, the student's records are presented with the knowledge discovered to enhance the overall performance of the students. A valuable and efficient review work has been done by the authors in [16]. The paper has discussed various concepts and techniques involved in EDM. Moreover, the comprehensive analysis has been made for the methodologies used for faculty and the student performance evaluation, which helps the institution.

In [2], it is stated that the classification techniques are included to help in enhancing the standard of the educational system by analysing the student data to handle the main factors that may have great change on the student performances in specializations. The classification rules are derived on the basis of decision tree making, and then the derived rules are examined and evaluated. It permits the students to establish the final grade under specializations. Another work in [5], also discussed the classification in data mining for student's performance evaluation. The contribution of the work was to extract the information from the database that defines the status of the students at the end of all semesters. The collected student information includes data about class test performances, assignments, seminars and attendances. The analysis helps in the identification of students who require special care and need some counselling from the mentors. Likewise, the concept provided in [3] involved in the prediction of student's enrolment based on the academic data using classification rules. The derived rules are evaluated by various methods and permit the university management to organize the required resources for the newly enrolled students. The process aids in providing the earlier information about the category of the students going to be admitted and the areas to be concentrated in their specializations to maintain them in a better way. WEKA tool has been used for the implementation and examination of classification techniques such as Rule-based classification, Naive Bayes and Decision Tree [17]. The paper involved in obtaining efficient results, using classification in required data extraction from large databases. The process is limited to handle the missing values and incomplete data in the collection.

In [6], the authors explained the J48 decision tree classification algorithm clearly. Moreover, the decision tree has been used for classifying the diabetes person's data. They have also discussed the basic steps involved in decision tree algorithm and computations. In a variant form, the performance of the instructor has been evaluated in [20]. Initially, the analysis has been made with four classifying techniques such as Support Vector Machine, Decision Tree algorithms, Artificial Neural Networks and discriminant analysis for developing an efficient classifier. The results have been stated that the decision tree algorithm provided more précised classification. A Quality model has been developed in [18] to process better classification and student performance examination. The author has done a great work by implementing various data mining technique to develop a predictive model for categorizing students based on their study rate and personal information. CART (Classification and Regression Tree), ID3 algorithm, CHAID (Chi-Squared Automatic Interaction Detection), and C4.5 all these four decision tree methods have been used for classification and the detailed comparative analysis has been made with that.

In order to categorize the failure patterns of students, the authors of [8] used the mining process based on association rules. The major contribution of the work was to detect the background connection between the failed courses and the equivalent course suggestions for improving the performance of low-grade students. The association rules derived from the algorithm provide some hidden relationships about the failed courses which could make a base knowledge for academic designers in taking academic choices. This could pave a way for modification and re-construction of curriculum for the improvement of student's study rate and decrease the failure rates. K-means clustering algorithm has been utilized by the authors in [10], for determining the learning activities of the students that may include the class quizzes, internal examination results and assignments. There clustered data will be provided by their respective tutors prior to the commencement of final examination. This may help the tutors to decrease the fail ratio of the students by taking corrective actions at the right time to help the student's progression.

In [9], ID3 and C4.5 classification algorithm are used for the performance prediction of students. They have also discussed the missing data issues in classification. The causes for non-availability of data are given as follows:

- Malfunction of Equipment's
- Deletion due to a discrepancy with other stored information
- No-data enrolment because of some misconception
- No-data registration

Moreover, in [13], a valuable case study has been provided by analysing the classification based data mining techniques for the application-oriented analysis of education sectors. The academic data sets are examined with the classification algorithms such as Naive Bayes (NB), J48, Multilayer Perceptron (MLP), SMO (Sequential Minimal Optimization) and REPTree (Reduced Error Pruning Decision Tree) by using the WEKA tool. A research work has been discussed about the various constraints influenced in student performance evaluation and grade computation for alerting the students for final examinations [19]. Moreover, the developed methodology is mainly used for grade calculation that paves a way for developing the coaching techniques to the students.

In [15], the performances of undergraduate and PG students of two different institutions are utilized in decision tree classification and Bayesian Network algorithms. This aids to identify the students for scholarships and also for special care. The comparative results concluded that the decision tree algorithm provides results with more precision than the Bayesian Network algorithm. In another work [14], a case study has been performed with 300 student samples that are obtained from the Punjab University. The study results in finding a dignified correlation in some factors such as parent's educational factors, the income of the family and the performance of individual students.

As classification as a prominent of techniques identified in [16], in the [21] the work is taken J48 as the prerogative of research over educational data and to have an extended analysis on *Naïve Bayes, Bayes Net, Multilayer Perceptron, SVM, REPTree and Random Forest* classifications with a *Tuned J48.* And the proposed model has shown overall betterments among the other ones over the datasets used. The study generates a realisation that a varied and better implementation of J48 can make differences in the performance prediction in a realistic environment.

IV. Proposed Model

Through the valuable review of previous works and discussions on EDM for analysing student's performance, a number of attributes influenced in the process are being identified and defined effectively. Those attributes are characterized as input variables. For this concern, the recent real-time data called KTU_SNG student dataset is collected from the Engineering College. Some manual techniques are used for figure out the data and the data is transformed into a format that is feasible for processing in WEKA tool. Following that, the selection of parameters and features are taken into consideration. Moreover, it is to be stated that the data set contains 13920 records (i.e. 232 samples with 60 attributes). Table I shows the sample of the data.

The Fig 3 portrays the generic data mining methodology that comprises data acquisition, data pre-processing, data mining and pattern extraction processes. Initially, data pre-pre-processing has been done with the acquired data from the database before providing the data set for the mining process (i.e. removing irrelevant attributes). For an instance, some factors like caste or category, etc. are not required for analysing the intellectual performance of the students, since they come under the personal data. Following that, the missing values are removed from the data set in order to reduce the complexities on mining. The overall motive is to categorize Nature_of_Student under the categories such as Outstanding, Excellent, Good, Average, low and very low. The following Table I presents some sample attributes and their descriptions along with the domain values obtained from the source database.

A. Bound Model of Clustering and Classification (BMCC Model)

Classification is described as the process of identifying a set of models that determined and differentiate classes and concepts of data for using the mining model to detect the unknown classes. On the other end, clustering is examined with the similarities between the features present in the objects. The following Fig 4 depicts the overall framework for the combined function of clustering and classification. The Bond Model of Clustering and Classification, (BMCC) model works effectively even if the dataset has some missing values. Furthermore, the block diagram presents the steps involved in the evaluation and comparative analysis to identify the Nature_of_Student and Performance for enhancing the overall academic performance of the institution.

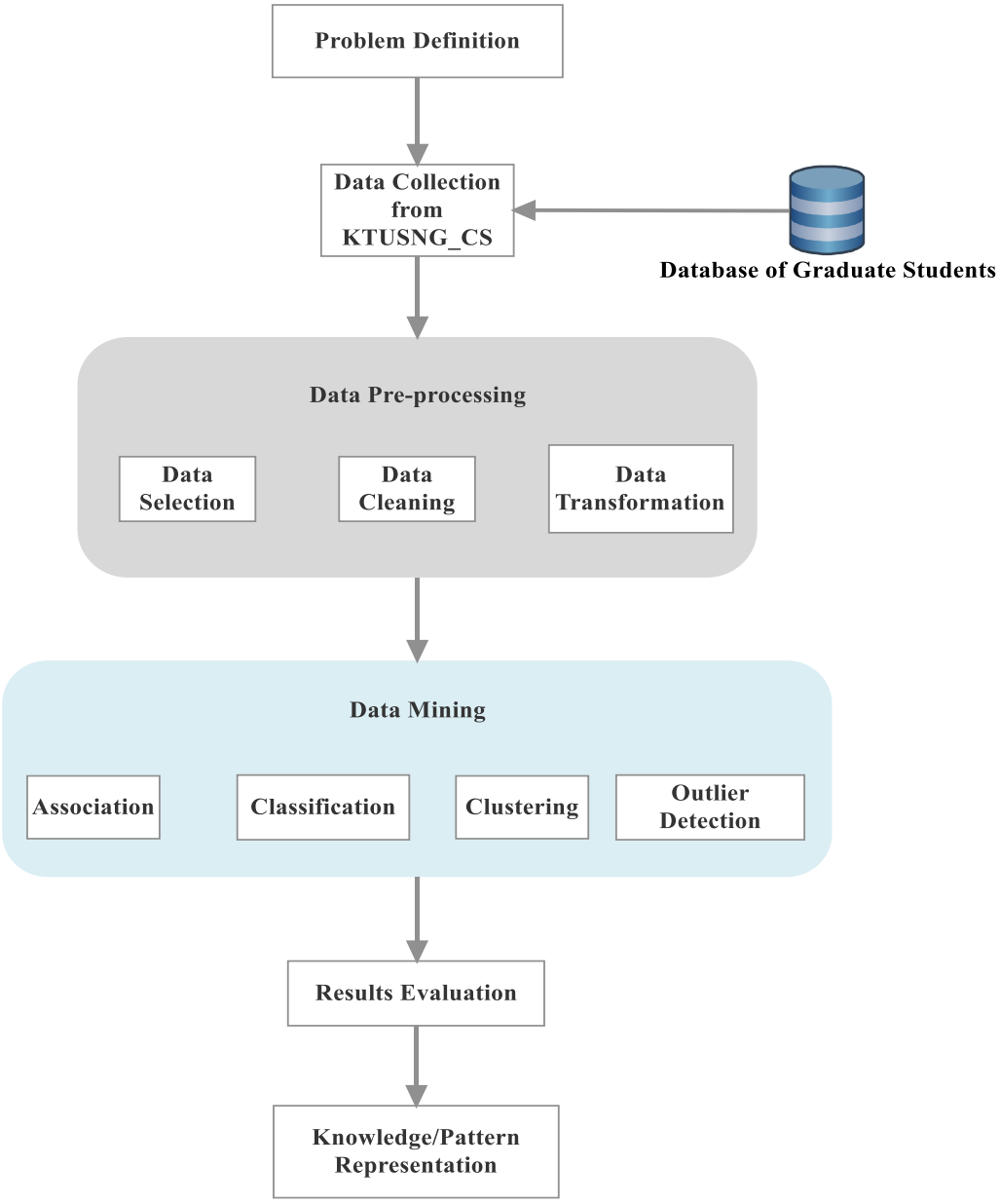

Fig. 3. The Flow of Generic Data Mining Methodology.

TABLE I. SAMPLE ATTRIBUTES OF THE ACQUIRED DATASET WITH DESCRIPTIONS AND DOMAIN VALUES

ATTRIBUTES	DESCRIPTION	DOMAIN VALUES
Gender	Student's sex	{M,F}
DOB	Date of Birth	Varying for samples
Mother Tongue	For language fluency	{Malayalam, Tamil}
Economically Backward	The financial status of the family	{True, False}
Handicapped	Health-Based analysis	{True, False}
Admission Type	Type of Admission	{Regular, Lateral}
Branch	Department of the student	{CS, CE, ME, ECE, EEE, NASB}
Attendance	Regularity	Based on the attendance count
Total Credits	Marks obtained	Given in percentage
Result	Calculated from marks	{PASS, FAIL}

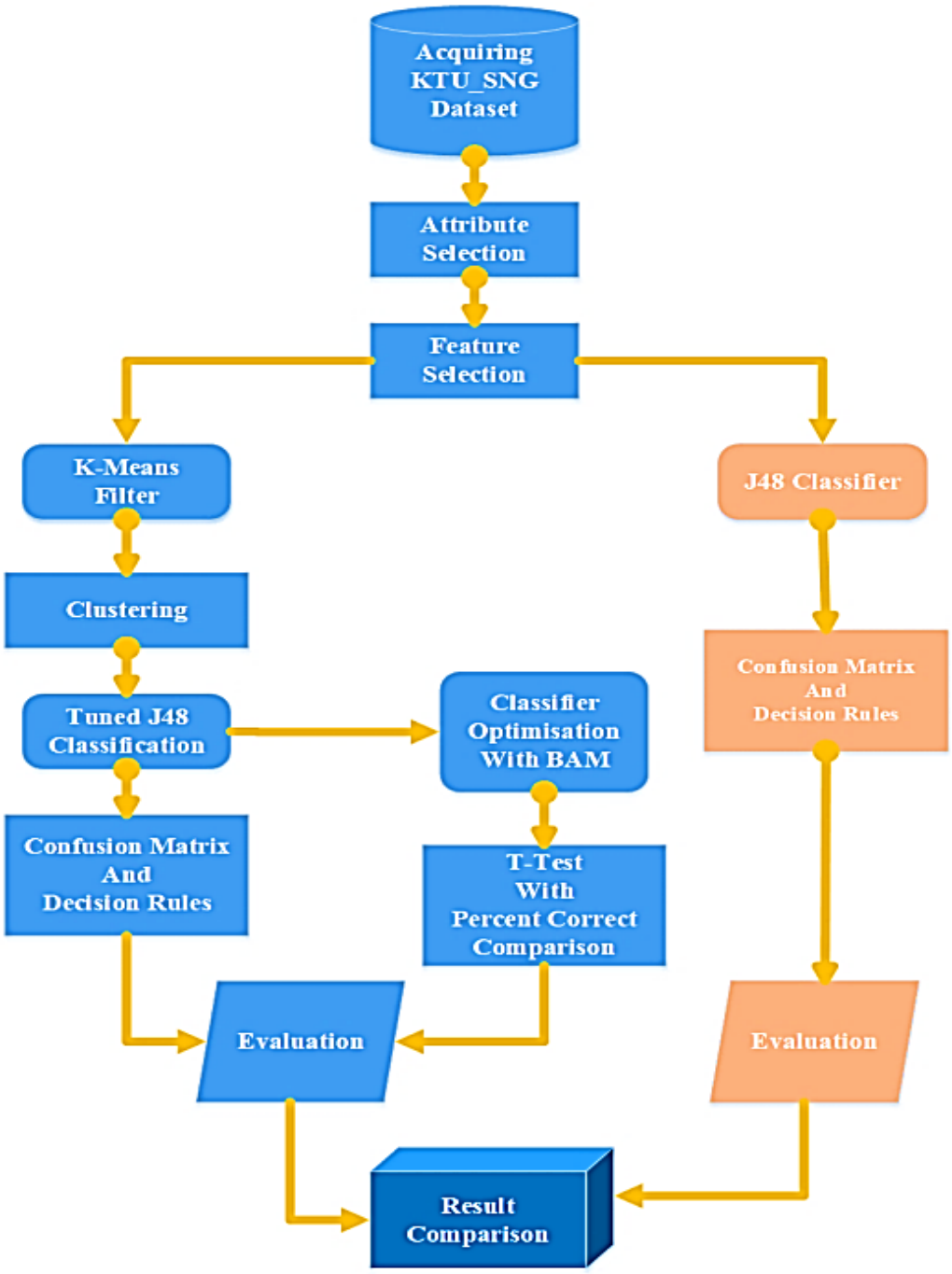

Fig. 4. The Framework of (BMCC) Model.

A tuned Classification technique [21] applied here is the J48 classifier using the WEKA tool. The tuning process may opt for a various combination of base J48, as demo choices. And, K-Means clustering is used here as a feature enhancement tool for natural clustering of classes, with a number of clusters equal to the number of classes [23]. incorporated here for the integration with classification. In this work, the finest classification rules are identified for classifying the students under Outstanding, Excellent, Good, Average, low and very low using the data mining tool.

The model is optimised using the Bootstrap Aggregation Model (BAM) for reducing the complexity in performance prediction. The result is compared with Base BMCC and experimental tests are also undertaken. Also, these results are referred to compare with j48 algorithms to check the efficiency of the proposed model against a base classifier. Other well-known base models in this domain, like Naïve Bayes and MLP, are also tested against the results

1) Developing classifiers (J48): Basically, J48 is the advancement of the ID3 algorithm as mentioned earlier. There

are some additional features in J48 such as accounting missing values, pruning, rule derivation, etc. In the WEKA tool, the J48 classifier is processed with open source Java implementation. In this algorithm, the classification is performed constantly till it obtains the pure leaf; hence the results obtained must be as appropriate as possible.

Steps in the Algorithm:

- If some instances fit into the same class of the leaf in tree representation is provided with the label under the same Classification.
- The potential data is evaluated for every attributes provided in the sample and then the Gain value is calculated.
- The current selection criterion provides the best attribute and that could be selected for effective branching.

a) Gain Calculation

The gain computation is dependent on the entropy measure of the data disorders. The Entropy $\vec{y}$ is calculated as,

$$Entropy\ (\vec{y}) = \sum_{b=1}^{n} \frac{|y_a|}{|\vec{y}|} \log \frac{|y_a|}{|\vec{y}|} \quad (1)$$

$$Entropy\ (b/\vec{y}) = \frac{|y_a|}{|\vec{y}|} \log \frac{|y_a|}{|\vec{y}|} \quad (2)$$

And the gain value is calculated from the entropy computations,

$$Gain(\vec{y}, b) = Entropy(\vec{y}) - Entropy\ (b/\vec{y}) \quad (3)$$

The gain value can be maximized by dividing the entropy values by using the split function $\vec{y}$ through the value of b.

b) Pruning

This is the significant process in classification. Some attributes are there in all data sets which may differ from other neighbourhood attribute and may not be well-defined. The classification operation has to be performed with these instances and the decision tree is to be formed. Here, pruning in incorporated to reduce the classification errors, which can be occurred due to the categorization in the training set. Specifically, pruning is performed for the simplification of the tree.

TABLE II. SAMPLES OF KTU_SNG DATASET UNDER CLASSIFICATION

Attendance Percentage	**Total Credit**	**Result**	**Nature_of_Student**
More than 90%	90 and above	Pass	**Outstanding**
More than 80%	Between 80 and 90	Pass	**Excellent**
More than 75%	Between 70 and 80	Pass	**Good**
More than 75%	Between 60 and 70	Pass	**Average**
More than 70%	Between 45 and 60	Pass	**Low**
More than 70%	Below 45	Fail	**Very low**

Here, J48 involves in developing decision trees from a training data set as in ID3 algorithm, based on the information entropy theory. The training data set is considered as TS={ts1, ts2, ..., tsn} of some earlier classified instances. Each sample tsi= a1, a2... is a vector, where a1, a2... denotes the attributes of the samples. The significance of the decision tree is to project the data with a good precision rate. The algorithm chooses an effective feature of the data that presents at each tree node, capable of splitting the sample data set into two subclasses or sets to be developed with one class or another. The samples under classification are provided in Table II, based on the classification of the Nature_of_Student under classes such as Outstanding, Excellent, Good, Average, Low and Very Low.

2) K-Means clustering: K-means clustering is a traditional clustering methodology that combined similar items in large data sets into groups. It involves selecting the initial centroids and determine the 'K' number of clusters by conveying the given instances to the near most centroids detected. Computation of Euclidean distance is the distance measure used in k-means clustering for centroids detection. Further, the computation is as follows,

$$F = \sum_{i=1}^{K} \sum_{j=1}^{n} \left\| x_j^{(i)} - CF_i \right\|^2 \quad (4)$$

Where 'F' is the objective function, 'K' is the number of clusters and 'n' is the samples. The centroid function is denoted as 'CF' and 'x' be the specific case of the instances.

B. Parameters and Performance Evaluation

The two most used parameters such as specificity and sensitivity are used to measure the performance of the proposed methodology [25]. The specificity can also be termed as True Positive Rate (TPR). Hence; the Specificity and sensitivity are computed as follows,

$$Specificity = \frac{True\ Negative\ (TN)}{False\ Positive\ (FP) + True\ Negative\ (TN)} = TPR = Recall \quad (5)$$

$$Senitivity = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Negative\ (FN)} \quad (6)$$

$$Error\ Rate = \frac{FP}{FP+TN} \quad (7)$$

1) Confusion matrix: The following Fig 5 depicts the confusion matrix for three class case. When a set of objects are evaluated, the results are effectively counted and the confusion matrix has been prepared. It can also be denoted as the contingency table. In the figure, the table has shown the correct decisions of classifier presents on the major diagonal, whereas the errors are at the rest.

The column denotes the Actual class and the rows denote the predictions. An edge is given as TP, whether it is positive or negative as same as the predictions respectively. FP denotes the results that are not predicted accurately. TN is representing the correctly predicted results, whereas, the FN values denote the falsely predicted with the preferred network.

		Actual		
		Positive	Negative	Non-Existent
Predicted	Positive	TP	FP	FN
	Negative	FP	TP	FN
	Non-Existent	FP	FP	TN

Fig. 5. Confusion Matrix.

2) Precision rate (PR): The retrieval of positive predictions is called precision. In particular, it is computed as the number of appropriate classification samples belong to 'A' divided by a number of samples categorized as under the class 'A'. Hence, it is defined as the ratio of the predicted true positives out of all actually positive results. The formula is given as follows:

$$Precision\ Rate\ (PR) = \frac{True\ Positive\ (TP)}{True\ Positive + False\ Positive} \quad (8)$$

3) Accuracy: Accuracy of computation is plainly defined as the relationship of a total number of correctly classified samples into the total sum of adopted samples. Mathematically, it can be defined as,

$$Accuracy = \frac{TP+TN}{FN\ ((TP+1)+(1+TN))} * 100 \quad (9)$$

4) F-Measure: F-measure is a significant parameter for evaluating the proficiency of the proposed model. It combines the TPR and the Precision Rates (PR) into an instant measure of performance. The equation is given as,

$$F - Measure = \frac{2*TPR*PR}{TPR+PR} \quad (10)$$

5) 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.

V. Experimental Results and Discussions

In this section of the experimental analysis, the results obtained for the proposed (BMCC) model corresponding to the parameters given in the (section 4.2) are compared with the traditionalistic classification techniques such as Naive Bayes, Multilayer Perceptron (MLP) and J48. Moreover, for the evaluation, KTU_SNG student dataset which contains 13920 records (includes 232 samples with 60 attributes) has been taken. For implementation, the training data set is classified through the WEKA tool with J48 decision-making classifier algorithm. The Bound Model of Clustering and Classification of data mining has initially employed with the K-Means algorithm for clustering as feature addition. It helps to remove the class attributes through the unsupervised learning process with the implemented training sample, following that Tuned J48 classifier has been implemented on the instances for classifying the student's performance.

The model use clustering as filter feature addition for the subsequent Tuned J48 classifier in this model. The K-Means cluster divides the sample into K clusters and the provision to specify the number of clusters help it used in classification. It is going to be one another strong feature during the classification.

Since similar studies are found in these areas, a famous model of clustering and classification have been opted with an optimisation to improve its performance. There were feature selection and ensemble models [24] of approaches. The optimisation is implemented using Bootstrap Aggregation Model. During the optimisation of BMCC, there are used choices of J48 as the demo options of the combination. How improvements have produced in BMCC are analysed and presented hereafter.

The experimental results illustrate that the proposed (BMCC) model provides potential results with utmost Precision Rate (PR), accuracy and the adaptability among the individual utilization of classification and clustering techniques in Table III, Table IV. The Fig. 6 illustrates the performance of BMCC and Improved BMCC. The Fig. 7 depicts the distribution of students according to their classes provided by the data set. The Precision Rate and accuracy are evaluated on the basis of True Positive, True Negative, False Positive and False Negative as per the (8) and (9), given in the previous sections B.2 and B.3.

The Table III depicts the comparative analysis based on the results of the parameters evaluated. Form the observation, it is apparent that the error rate for (BMCC) model is **0.095** which is comparably low than others. And, the accuracy rate is **94.83%** which is most desirable than other models. It shows that the (BMCC) model provides better-précised results than others and helpful for efficient student classification and improvement of student's performance.

For computation purpose, it is to be considered that, the false positive rate or the value of sensitivity would be **0**. As per the results provided in Table III and Table IV, the sensitivity rate obtained for the combined model is the lowest (closer to 0), when compared to the other techniques. Another assumption is also to be made for Precision Rate as 1. And from the results provided in Table III, it is noticeable that the proposed model comprises a higher rate of precision than the others.

The experiment is also conducted for identifying Nature_of_Student, the Class of students, have also done, however, the results were not competing with performance prediction of students. And the details are slightly suspended from this paper as it focuses on a binary classification problem. The result of the study regarding nature had shown a significant rise in FN rate in many experiments, and it is assumed to be due to the perceived quality of the internal assessments and its recording. Students attitude and their sentiments towards activities are not considered to grade students in KTU. It may be a work proposed to analyse the impact of sentiment analysis in the field of student classification problems, and thereby a solution may be devised permanently.

TABLE III. PARAMETERS BASED PERFORMANCE EVALUATION

Parameters	Naive Bayes	MLP	J48	BMCC
TP	161	176	170	177
FP	5	10	8	5
TN	43	38	40	43
FN	23	8	14	7
Error Rate	0.121	0.078	0.095	0.052
Accuracy	87.93%	92.24%	90.52%	94.83%
Precision	0.97	0.947	0.956	0.973
Recall	0.875	0.957	0.924	0.962
F1 Measure	0.921	0.952	0.94	0.968
Sensitivity	0.88	0.96	0.92	0.96
Specificity	0.90	0.79	0.83	0.90

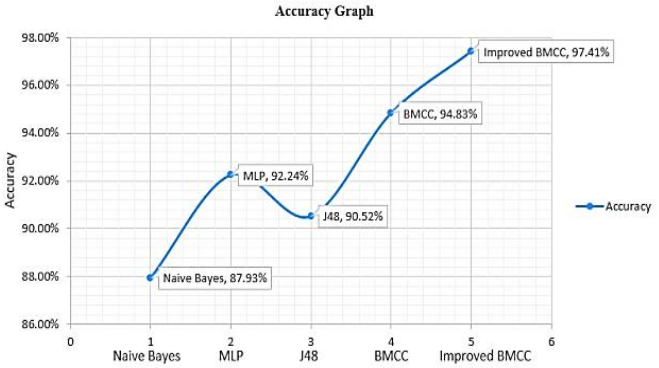

Fig. 6. Performance Accuracy of BMCC and Improved BMCC with Others.

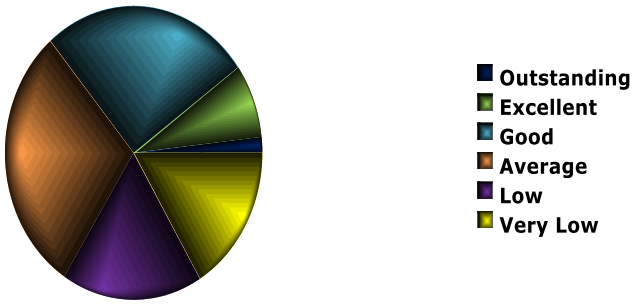

Fig. 7. Distribution of Students According to their Classes.

The following Table IV depicts the values obtained for the proposed (BMCC) models with overall classification results under "PASS", "FAIL". It is portrayed from the table that out of 232 instances, 220 are classified accurately under the suitable class and the rest are not accurately classified as given in the Base BMCC. Whereas the improved one has 226 and 6 as correctly classified and misclassified correspondingly to its credits. The optimum estimated results of the evaluation parameters are given in Table IV. Further, the Fig.9 shows the results obtained with Classification using (BMCC) Model and implemented with and without optimisation in the Weka Workbench environment.

TABLE IV. RESULTS FOR EVALUATION PARAMETERS FOR CLASSIFYING "PASS" OR "FAIL"

Total Number of Instances= 232			**Values In - Dataset Base BMCC**		**Values In - Dataset Improved BMCC**	
Correctly Classified			220	**94.80%**	226	**97.4%**
Incorrectly Classified			12	**5.20%**	6	**2.60%**
TP	**FP**	**PR**	**RC**	**F-M**		**CLASS**
0.962	0.962	0.973	0.962	0.967		PASS
0.896	0.038	0.86	0.896	0.878		FAIL
Weighted Average:						
0.948	**0.09**	**0.949**	**0.948**	**0.949**		
0.984	0.063	0.984	0.984	0.984		PASS
0.938	0.016	0.938	0.938	0.938		FAIL
Weighted Average:						
0.974	**0.053**	**0.974**	**0.974**	**0.974**		

The decision tree based classification sample and rules are given in Fig 8.

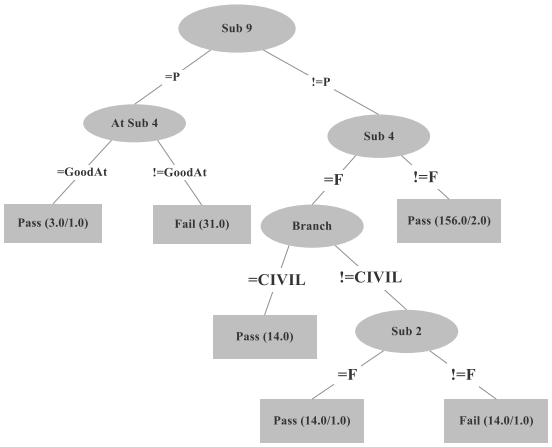

Fig. 8. Decision-Tree-based Classification Sample.

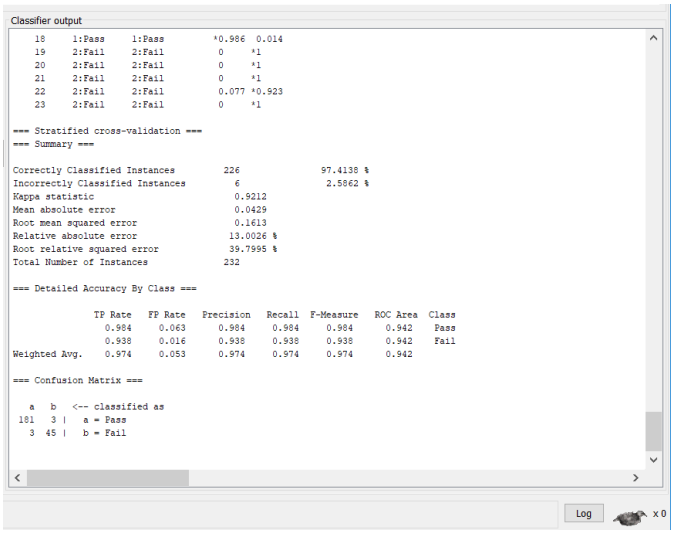

Fig. 9. A sample Result of (BMCC) Model Implementation in WEKA Tool.

TABLE V. AVERAGE ERROR RATE OF CLASSES COMPARED TO (BMCC) CLASSIFICATION MODEL

Method	Naive Bayes	MLP	J48	Base BMCC	Improved BMCC
Error Rate	0.108	0.174	0.148	0.09	0.053

TABLE VI. RESULTS OF RELATED EXPERIMENTS

Method	J48	Base BMCC	Improved BMCC
Accuracy	90.52%	94.83%	97.41%

From the outset of this work, it is clear that the utilisation of classification techniques is considered. Especially the accuracy and precision of each of them considered more. From Table V, it is clear that the proposed bound model of clustering and classification has produced the most competent outputs. A random improvisation is also tried with the BMCC and a marginal improvement has shown in it. However, a detailed observation may be done later with more option to improve the proposed one.

Whatever being implemented had shown an appealing progress in the experiments, even though few variations in results based on the filtering used at the pre-processing had also shown. It may be considered and experimented with more kinds of data from various environments collected. However, it is intuitively suggested to try-out for an optimum feature extraction and pre-processing for these kinds of researches. The results of variations of models used in this and related models are depicted in Table VI.

Based on the experimental analysis given in this paper, it is explicit that the Bound Model of Clustering and Classification provides better results in classifying instances with a better rate of accuracy.

During the experiment in this paper, subsamples of KTU_SNG Data are prepared using different filters such as Gain Ratio, Info Gain, Correlation Attribute filtering...etc. Also, the actual data along with the data prepared using Weka for testing and training are used during the bootstrap optimisation of BMCC. It shows an incremental uplifting of accuracy during bootstrapping (Bagging-Tuned J48). Similar to every bagging its learning curve started from a minimum of Base BMCC rate and progressed to achieve a score of 99.569% accuracy. The score initially felt a little exaggerating as the database size and the stability of it is taken into account. However, the Max-Rate (99.569%) may be reached at the year goes by and the features of database tend to become normal form. The comparison of the last few observations of optimised BMCC is given in Table VII.

TABLE VII. OPTIMISED PERFORMANCE OUTCOME OF BMCC

Model	Experiment 1	Experiment 2	Experiment 3	Experiment 4
BMCC	94.3966%	97.8448%	99.569%	99.569%

TABLE VIII. OPTIMISED PERFORMANCE OUTCOME OF BMCC TEST

Model	Exp. T-Tester1	Exp. T-Tester 2	Exp. T-Tester 3	Exp. T-Tester 4
BMCC	93.09%	93.44%	95.08%	98.71%

During the performance test, the same had produced comparatively lover figures with the experimental models are given in Table VIII.

The model could Minimise the Cost-Benefit at 75.72 gain threshold with 99.569% accuracy as given in Fig 10.

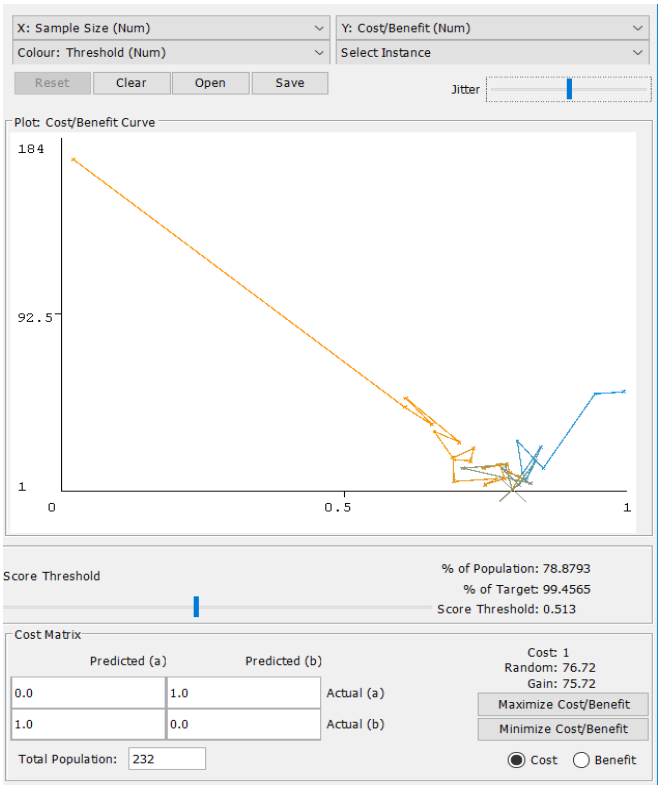

Fig. 10. Minimised Cost-Benefit of (BMCC) at Optimum Gain.

Interesting properties of Reinforcement Learning [22] in the form of a reutilising model, for information extraction, might ameliorate BMCC. Hence, non-rationally believed that it provides an overriding Educational Data Mining technique or DDM technique for the managing performance of the students.

VI. CONCLUSION AND FUTURE ENHANCEMENT

The proposed (BMCC) model is an effective technique for the Proficient Performance Prediction (PPP) of educational datasets and mining classification that helps in successfully identifying the huge data sets of educational domains. Decision tree J48, proposed BMCC and its improved model came up with **90.52%, 94.83%** and **97.41%** accuracy respectively in predicting the performance of students in the KTU_SNG-Dataset. The evaluation results are evidence for the Bound Model of Clustering and Classification technique provides results with more precision than the traditional classification methodologies such as Naïve Bayes, MLP and

base J48, for categorizing the given instances with their specific classes and attributes. A substantial increase of **7.0%** is noted here. As is well known, clustering operations belong to the unsupervised learning process; hence, the classes are developed on the basis of formed clusters to which the samples belong to. And then, the unknown datasets are classified on the basis of the decision rules, which are acquired by employing a classification technique on these clusters. The classification results obtained from this combined methodology brings utmost precision rate and accuracy with minimal error rate and cost benefit. The **99.569%** accuracy at the experiment and the **98.71%** accuracy at test obtained by the proposed model serve the best result compared to all other. The result shows that the Optimised BMCC outperformed all other methods in this research experiment. And hence it may be a good choice for the future also to enhance the model with optimisation and booting methodologies to enhance classification accuracy in Educational Data Mining. And, the potential results are valuable for the academicians for analysing the student's performances and making decisions or arrangements according to that to enhance their quality along with the institution.

As it is discussed performance improvisation methods and its applications to BMCC and a dependable, refined and standardised information extraction technique for it are going to be a good choice of future work. Also, in future, the combined work can be extended and integration can also be done with the association or ensemble techniques and the results can be analysed with some other efficient models. Furthermore, the methodology can be applied and examined with some other interesting areas like the sentiment of students towards their activities, sports, medicine and so on, which have to deal with huge datasets.

REFERENCES

[1] Romero, C. and Ventura, S. "Educational Data Mining: A Survey from 1995 to 2005", Exp.t Sys. with App. (33), (2007), pp. 135-146.

[2] Al-Radaideh, Q., Al-Shawakfa, E. and Al-Najjar, M. "Mining Student Data Using Decision Trees", The 2006 Int. Arab Conf. on Info. Technol. (ACIT'2006) – Conference Proceedings, (2006).

[3] Shannaq, B., Rafael, Y. and Alexandro, V. "Student Relationship in Higher Education Using Data Mining Techniques", Glob. J. of Comput. Sci. and Technol., (2010), vol. 10, no. 11, pp. 54-59.

[4] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann, 2000.

[5] Baradwaj, B. and Pal, S. "Mining Educational Data to Analyse Student s' Performance", Int. J. of Adv. Comput. Sci. and Appl., (2011), vol. 2, no. 6, pp. 63-69.

[6] Gaganjot Kaur and Amit Chhabra, "Improved J48 Classification Algorithm for the Prediction of Diabetes", Int. J. of Comput. Appl. (0975 – 8887), July 2014, Volume 98 – No.22, pp. 13-17.

[7] Kumar, V. and Chadha, A. "An Empirical Study of the Applications of Data Mining Techniques in Higher Education", Int. J. of Adv. Comput. Sci. and Appl., (2011), vol. 2, no. 3, pp. 80-84

[8] Chandra, E. and Nandhini, K. "Knowledge Mining from Student Data", Europ. J. of Sci. Res., (2010), vol. 47, no. 1, pp. 156-163.

[9] Kalpesh Adhatrao, Aditya Gaykar, Amiraj Dhawan, Rohit Jha and Vipul Honrao, "predicting students' performance using id3 and c4.5 classification algorithms", Int. J. of Data Mining & Knowl. Manag. Proc. (IJDKP) Vol.3, No.5, September 2013, pp. 39-52.

[10] Ayesha, S., Mustafa, T., Sattar, A. and Khan, I. "Data Mining Model for Higher Education System", Europ. J. of Scienti. Res., vol. 43, no. 1, (2010), pp. 24-29.

[11] Romero, C., Ventura, S. and Garcia, E. "Data mining in course management systems: Moodle case study and tutorial", Comput. & Educ., vol. 51, no. 1, (2008), pp. 368-384.

[12] El-Halees, A. "Mining Students Data to Analyse Learning Behavior: A Case Study", The 2008 Int. Arab Conf. of Info. Technol. (ACIT2008) – Conf. Proc.s, University of Sfax, Tunisia, Dec 15- 18. (2008).

[13] Parneet Kaura, Manpreet Singh, Gurpreet Singh Josanc, "Classification and prediction based data mining algorithms to predict slow learners in the education sector", 3rd Int. Conf. on Rec. Tren. in Comput. 2015(ICRTC-2015), Science Direct, Procedia Comput. Sci. 57, (2015), pp. 500 – 508.

[14] Syed Tahir Hijazi, and S. M. M. Raza Naqvi, "Factors affecting students' performance: A Case of Private Colleges", Bangladesh e-J. of Sociol., Volume3.Number1, January 2006.

[15] Nguyen Thai Nghe, Paul Janecek, and Peter Haddawy, "A Comparative Analysis of Techniques for Predicting Academic Performance", In Proc.s of the 37th ASEE/IEEE Front. in Educ. Conf., 2007 Pp. 7- 12.

[16] 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," IEEE Int. Conf. on Data Min.g and Adv. Comput., Ernakulam, 2016, pp. 122-133.

[17] Fadhilah Ahmad*, Nur Hafieza Ismail and Azwa Abdul Aziz, "The Prediction of Students' Academic Performance Using Classification Data Mining Techniques" Appli. Math. Sci.s, Vol. 9, 2015, no. 129, 6415 – 6426.

[18] Amjad Abu Saa, "Educational Data Mining & Students' Performance Prediction" (IJACSA) Int. J. of Adv. Comput. Sci. and Appl., 2016Vol. 7, No. 5.

[19] V. Ramesh, P. Parkavi, K.Ramar "Predicting Student Performance: A Statistical and Data Mining Approach," Int. J. of Comput. Appl. (0975 – 8887), February 2013, Volume 63– No.8.

[20] MUSTAFA AGAOGLU, "Predicting Instructor Performance Using Data Mining Techniques in Higher Education", IEEE Access, May 13, 2016, Digital Object Identifier 10.1109/ACCESS.2016.2568756.

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

[22] Narasimhan, K., Yala, A., & Barzilay, R. (2016). "Improving Information Extraction by Acquiring External Evidence with Reinforcement Learning". Proceed. of the 2016 Conf. on Empir. Meth. in Nat. Lang. Proc. pages 2355–2365, Austin, Texas, November 1-5, 2016 https://doi.org/10.18653/v1/D16-1261

[23] Kyriakopoulou, A., Kalamboukis, T.: Using clustering to enhance text classification. In SIGIR 2007, 30th Annual International ACM SIGIR Conf. on Res. and Devel.t in Info. Ret., 2007. pp805–806.

[24] Pushpalata Pujari, Jyoti Bala Gupta, "Improving Classification Accuracy by Using Feature Selection and Ensemble Model", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-2, May 2012.

[25] 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.