https://doi.org/10.48048/tis.2022.5070

Application of the Classification Algorithms on the Prediction of Student's Academic Performance

Sourav Kumar Ghosh[1], Farhatul Janan[1,*] and Ishtiyaque Ahmad[2]

[1]Department of Industrial and Production Engineering, Bangladesh University of Textiles, Dhaka 1208, Bangladesh
[2]University of California, Santa Barbara, California 93106, United States

([*]Corresponding author's e-mail: janan_ipe@butex.edu.bd)

Received: 4 December 2020, Revised: 18 August 2021, Accepted: 25 August 2021

Abstract

Measuring student performance based on both qualitative and quantitative factors is essential because many undergraduate students could not be able to complete their degree in the recent past. The first-year result of a student is very important because in the majority of cases this drives the students to be either motivated or demotivated. So, the first-year student performance of a renowned university in Bangladesh is investigated in this paper. This research is mainly based on finding the factors for students' different types of results and then predicting students' performance based on those 11 significant factors. For this purpose, 2 popular supervised machine learning algorithms have been used for classifying students' different levels of results and predicting students' performances, those are support vector machines (SVM) classifier and random forest classifier (RFC) which are tremendously used in classification and regression analysis. The input dataset for both training and testing were taken by merging the values obtained from 2 surveys done on students and experts using an adaptive neuro-fuzzy interference system (ANFIS). RF has outperformed SVM in predicting students' performances. According to factor analysis, students' effort (Factor-11) is the significant factor. This proposed model can also be applied to predict course-wise students' performances and its precision can also be greatly improved by adding new factors.

Keywords: SVM Classifier, Random forest classifier, Fuzzy logic, Performance prediction, Classification criteria

Introduction

Education is one of the basic needs of human beings. It has many levels such as early childhood education (level 0), primary education (level 1), lower secondary education (level 2), upper secondary education (level 3), postsecondary non-tertiary education (level 4), short-cycle tertiary education (level 5), bachelor's or equivalent level (level 6), master's or equivalent level (level 7) and so on. Dropout from any educational stage is also a common phenomenon. There are many factors that play a vital role in increasing dropout rates. Degradation of performance or result is one of the major reasons for student dropouts. This research is mainly focused on finding different factors that affect freshmen students' results at the undergraduate level.

After coming to university, many students cannot adapt themselves to the university's environment for study and thus it affects their performances. There are also some other causes like students' involvement in politics, extra-curricular activities, etc. Due to these various known and unknown reasons, students' performances in the university in many cases tend to be poor which in turn affects their results. So, the performance of undergraduate students should be analyzed to find the real root cause of the problem related to a student's performance.

So, our main motivation behind this work was to help students understand the attributes which are responsible for their poor results so that they can improve their performance. If the major factors are identified and monitored, it will give the students, course teachers, and the administrators to ameliorate the study environment. On the other hand, if students can anticipate the reasons for degrading their results, they can work on those to improve their performances. That's why we had chosen 1st-year students for our research. We have performed our experiment at the Bangladesh University of Textiles.

The results of the recent first year (45th batch, 2018) students at this university were taken for this research. For identifying the factors, we have surveyed students and some experts (here experts mean different course teachers associated with students). After completing the survey, factors and their ratings were identified that showed the reasons for students' different levels of performance. We have merged the ratings of students and experts on a particular factor using ANFIS for modifying these datasets as single input dataset to the model. 80 % of these data were used for training and the remaining 20 % was used for testing and finally, by computing the accuracy of the model, the validity of the formulation of the model (factors' identification and their ratings) was accomplished. In this model, raw data are modified using expert opinions and fuzzy logic. For this reason, the model performed better than other traditional models which only considered raw data.

Over the recent years, researchers have published many works on analyzing the reasons behind a student's performance. Some of these research works include only analyzing the attributes which have direct or indirect effects on students' performance and some also include predicting other students' results based on the studied attributes using different learning algorithms. Most of these learning algorithms are artificial intelligence (AI) techniques such as artificial neural networks, random forest, Bayesian classifiers, etc.

Naive Bayesian (NB) data mining technique was applied to predict the student performance based on 19 attributes such as gender, food habit, the medium of teaching, family status, family income, students' grade, and so on [1]. SVM also excelled at Bayesian Knowledge Tracing (BKT) in predicting students' problem-solving performance by showing approximately 29 percent improvement, compared to the standard BKT method [2]. Factor reduction was implemented by correlation-based feature selection (CBFS), chi-square-based feature evaluation (CBFS), and information gain attribute evaluation (IGATE). The decision tree (DT) algorithm worked more efficiently than other machine learning (ML) algorithms on a case study of some undergraduate students in Kolkata [3]. Student performances' prediction based on particle swarm optimization (S3PSO) was proposed to predict student performance which performed better than other ML methods like SVM, KNN, and so on [4]. Frequent pattern tree algorithm, ensemble semi-supervised learning (SSL) algorithm, recurrent neural network (RNN) and DT techniques were employed to forecast student's academic performance [5-9]. Socio-economic factors and entrance examination results were used as the primary factors to predict the student's cumulative grade point average (CGPA) by applying ANN with the Levenberg–Marquardt algorithm [10]. The 2-level classification algorithm was implemented to calculate the expected graduation time where the passed or failed students were differentiated in the first level and 3 different periods of graduations were classified in the next level [11]. A literature review of various techniques applied to predict student performance was studied. Techniques were grouped into 3 clusters such as fuzzy logic, data mining techniques, and hybrid [12]. A hybrid model was formed utilizing Bayes network (BN) and Naive Bayes (NB) as generative models while SVM, C4.5, and Classification and Regression Tree (CART) were employed as discriminative models [13]. Student performance was forecasted by random forest algorithm considering e-learning environment where lab total, assignment submission, mid-term was assessed as the principal attributes [14]. Students' academic success and their selection of majors were also predicted constructing 2 types of classifiers using random forests [15]. On a study of profile characterization of Chilean students, RFC were able to successfully identify the factors which could explain students' performances and the factor named "parents' educational expectations" was identified to be the key factor for students' outstanding performance [16]. The linear random forest (LRF) have advantages over least squared linear regression, neural networks, epsilon support vector regression, KNN regression, regression tree, regression random forest, gradient descent boosted trees, and linear decision tree algorithms in the context of learning ability, algorithm robustness, and feasibility of the hypothesis space [17]. Fuzzy ANFIS was used to convert multiple decision maker's ratings into a final rating based on 78 fuzzy logic [18]. Fuzzy ANFIS was used in the conversion of multiple experts' and students' ratings into single dataset [20,21].

RFC is actively using in many literary works for prediction related problems [19] while its performance is promising in the forecasting of students' results [20]. On the other hand, the use of SVM which is well known for classification [21] and regression [22] is also observed in the anticipation of students' academic performances. To our best knowledge, prediction of students' progress using fuzzy ANFIS and machine learning algorithms, RFC and SVM is yet to be done. Using RFC and SVM along with fuzzy logic could demonstrate which method performs better in this scenario and give a new direction toward education data mining.

In this paper, we have formulated the problem as a multi-class classification problem. Different machine learning-based techniques are available to solve this type of problem. After analyzing the research works on classification problems and prediction of students' performances, it was found that

both the Support Vector Machine (SVM) and Random Forest (RF) classifiers were separately used in many works and their performances were better than other methods [23,24]. So, we aimed to implement these 2 methods to compare their accuracy. And we have also introduced fuzzy logic which had merged different sets of input databases under certain rules to a single input database.

Materials and methods

This research focuses on the identification and analysis of different types of factors including psychological, personal, teaching impact, university facilities, learning environment, etc. that affect students' results and prediction of student's performance based on these factors using machine learning algorithms. Specifically, this work can be classified into 4 major steps as follows:

1) Identification of factors through a student survey and an expert survey
2) Modification of factor rating by applying expert's opinion with the help of fuzzy ANFIS
3) Classification and prediction of student's performance based on SVM and RF classifier
4) Evaluation of the accuracy of the 2 models by analyzing the predicted results with the actual results

In the 1st step, a survey consisting of 31 questions was done via google response form to get the ratings of copious factors from students. Here the course teachers and student advisors were considered experts. In the following step, factors were reduced by factor analysis, and based on selected factors, another survey of experts was conducted. Then the ANFIS model was developed to convert multi-response ratings into a single rating point by consisting of fuzzy logic. Students were separated into 7 classes based on their GPA in the previous term. **Table 1** shows how students were separated into classes. After that SVM and RFC were applied using train and test data to predict the student performance. Finally, the results were analyzed with relative advantages and disadvantages.

Table 1 GPA range of different classes.

CGPA range	Class no.	Mean	Students no.
2.26 -2.50	1	2.46	11
2.51 -2.75	2	2.66	65
2.76 -3.00	3	2.89	125
3.01 -3.25	4	3.13	141
3.26 -3.50	5	3.38	129
3.51 -3.75	6	3.62	77
3.75 -4.00	7	3.86	35

Process flowchart

At first, surveying was done on students to identify the primary factors which affect their performance. After performing factor analysis based on students' responses to the survey and experts' opinions, important factors were identified. Then, the second survey was done on experts who gave ratings of the effect of each factor on each student. After the completion of these surveys, ratings of each attribute were merged with the factor ratings given by the experts with the help of Fuzzy ANFIS analysis. Then, 80 % of these merged data were used as training data for both SVM and RF classifiers. The other 20 % was used for testing. After that, the accuracy of both methods was checked with the actual results (grades) of the students. The full process flowchart of the work is shown in **Figure 1**.

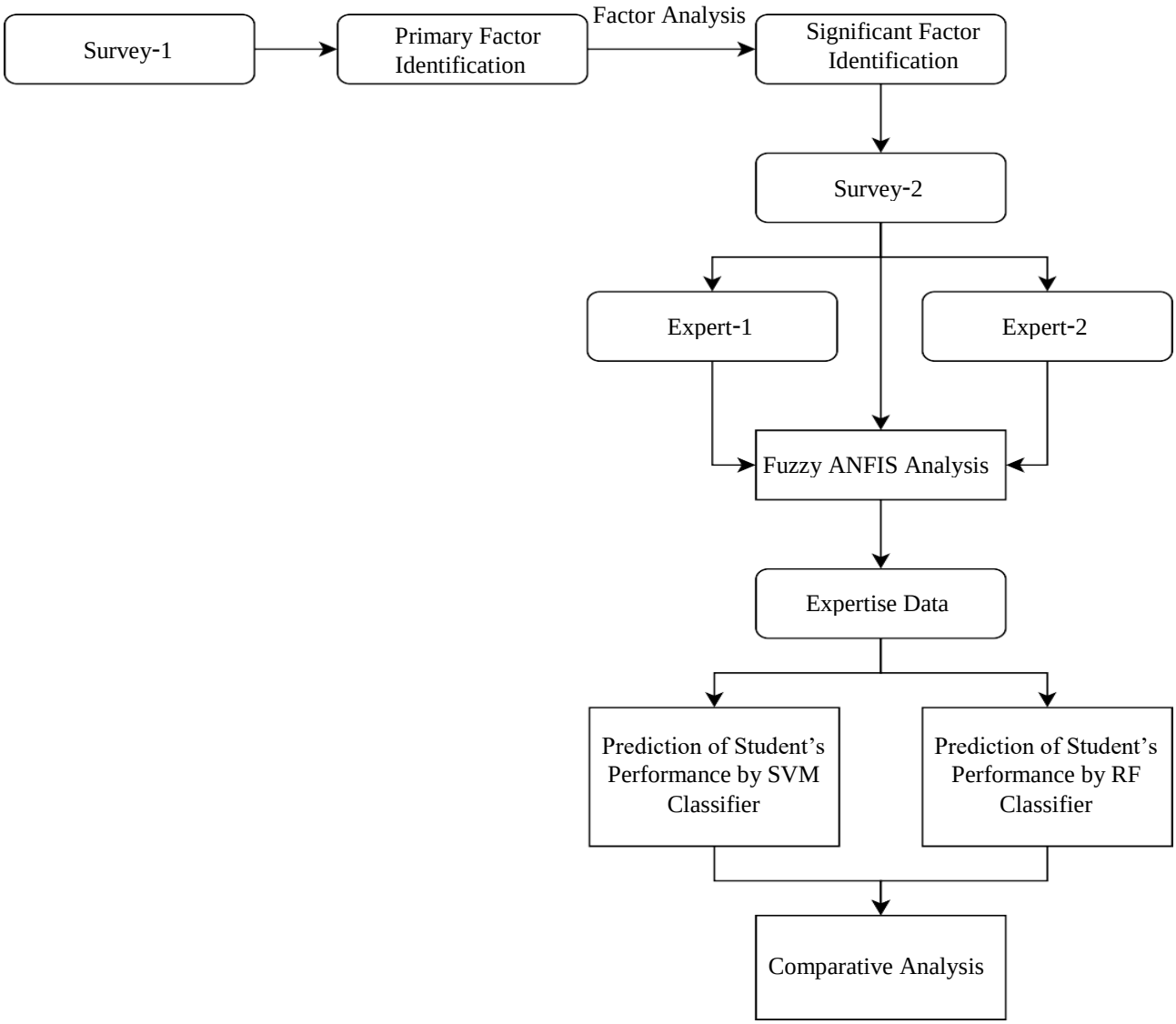


Figure 1 Process flowchart of the proposed model.

Identification of Significant Factors

In this paper, a survey of 31 questions was first performed on first-year students (45th batch) of the Bangladesh University of Textiles to identify their ratings on each factor which could be linked to their different levels of performance. These 31 factors which were used in survey-1 are shown in **Table 2**. After performing survey-1, significant factors were identified from the ratings of the students on each factor. There are various hypothesis testing methods available to capture the relationships between variables [25]. Here, Analysis of variance (ANOVA) was used for this purpose to check the correlation between different factors as ANOVA has better control over type 1 error [26].

Table 2 List of all the factors for survey 1.

Creating good notes	Group study	Adaptation to university	Self-confidence
University facilities	Previous year questions	University environment	Effort
Assignment submission	Class tests' marks	Discontent about the university	Motivation
Proper knowledge	Preparation time	Discontent about the department	Procrastination
Time management	Fear of examinations	Political involvement	Teaching methods
Class attendance marks	Hard questions	Family issues	Overconfidence
Course difficulty level	Teachers' friendliness	Residential problems	Exam strategy
Excessive inclination towards extra earning	Teachers 'hardness of checking exams' scripts	Personal problems (any kind of addictions or illegal activities)	

For the reduction of factors and identification of significant factors, ANOVA has been used here. In **Table 3**, ANOVA tests' result is shown where 2 factors, proper knowledge, and time management were compared, and ANOVA with a 95 % confidence interval was applied. As $F < F\text{-}crit$ or $p\text{-}value > 0.05$, the null hypothesis is accepted, which means the 2 factors are alike. Another example is given in **Table 4**, where factors, exam strategy, and effort, were compared with the same confidence interval. But the result here shows the opposite. As $F > F\text{-}crit$ or $p\text{-}value < 0.05$ that means a null hypothesis is rejected and there is a difference between the factors.

Table 3 ANOVA table of 2 correlated factors.

Factors		**Sum of Squares, SS**	***df***	***MS***	***F***	***p-value***	***F-crit***	**Hypothesis**	**Result**
Proper Knowledge	Between groups	24.49485	1	24.49	0.0511	0.821	3.849	Accepted	Factors Correlated
Time Management	Within groups	557811.5	1164	479.21					
	Total	557836	1165						

Table 4 ANOVA table of 2 uncorrelated factors.

Factors		**Sum of Squares, SS**	***df***	***MS***	***F***	***p-value***	***F-crit***	**Hypothesis**	**Result**
Exam Strategy	Between groups	8215	1	8215.28	17.20	3.6E-05	3.8494	Rejected	Factors uncorrelated
Effort	Within groups	555889	1164	477.56					
	Total	564104	1165						

In this way, using the ANOVA tests, correlated and uncorrelated factors were identified. Then correlated factors were eliminated and uncorrelated factors were taken as significant factors (factors responsible for students' different levels of performances) which were used for the whole classification problem. Eleven significant factors were found which are shown in **Table 5**.

Table 5 List of 11 significant factors.

Factors' no.	**Factors' name**	**Factors' no**.	**Factors' name**
1	Creating good notes	7	Political Involvement
2	Exam strategy	8	Personal problems
3	Class tests marks	9	Fear of examinations
4	Teaching methods	10	Procrastinations
5	Hard questions	11	Effort
6	Adaptation to university		

Fuzzy ANFIS analysis

Fuzzy logic imitates in a way that resembles human reasoning. It is an approach where computing is based on different "degrees of truth" rather than Boolean (1, 0) logic on which a modern computer is based. A computer can only take precise inputs and give output as TRUE or FALSE which resembles a human's YES or NO [27]. Its architecture contains 4 parts as shown in **Figure 2**.

1) Rule base: It contains IF-THEN rules provided by the experts which help to govern the decision-making system.

2) Fuzzification: It is used for converting inputs which are called crisp numbers into fuzzy sets.

3) Inference engine: It determines the matching degree of the current fuzzy input with respect to each rule and stimulates human reasoning on the basis of these rules then it decides which rules are to be fired according to the input field. Then, the fired rules are combined to form the control actions.

4) Defuzzification: It is used to convert the fuzzy sets obtained by the inference engine into a crisp value which is the ultimate output.

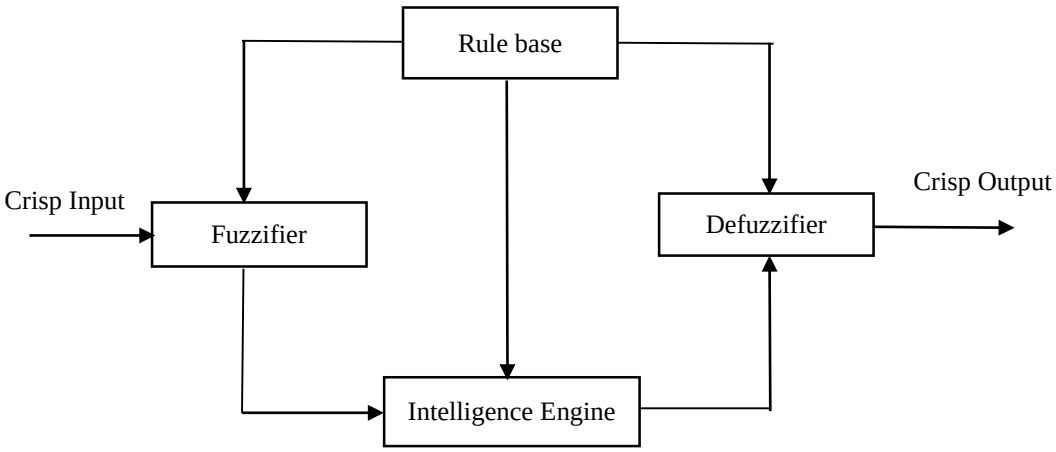


Figure 2 Fuzzy logic structure.

Membership function is a graph that defines how each point in the input and output space is mapped to membership value between 0 and 1. There are largely 3 types of fuzzifier:

1) Singleton fuzzifier

2) Gaussian fuzzifier

3) Trapezoidal or triangular fuzzifier

In this paper, the fuzzy ANFIS model has been established where a set of inputs and outputs are given then rules are added. At first, triangular membership functions are used for each input and output data. Then rules are added. Train data was transformed through the fuzzy ANFIS model.

Merging of factors' ratings by fuzzy analysis

After finding 11 significant factors using ANOVA analysis, another survey was conducted on 2 experts (one expert was course teacher and another expert was course coordinator) with those factors. Now, there were 3 ratings for each factor and fuzzy was used to merge these ratings. To merge these ratings in fuzzy analysis, different rules were imposed on inputs (students' and experts' ratings) to get the possible output (merged value). **Figure 3** shows the demonstrations of these rules on inputs and how they affect the outputs. In **Figure 4**, the relationship between input parameters and the output parameter of fuzzy ANFIS is shown in 3D graphs. After that, combining all these outputs, finally, the merged ratings were identified.

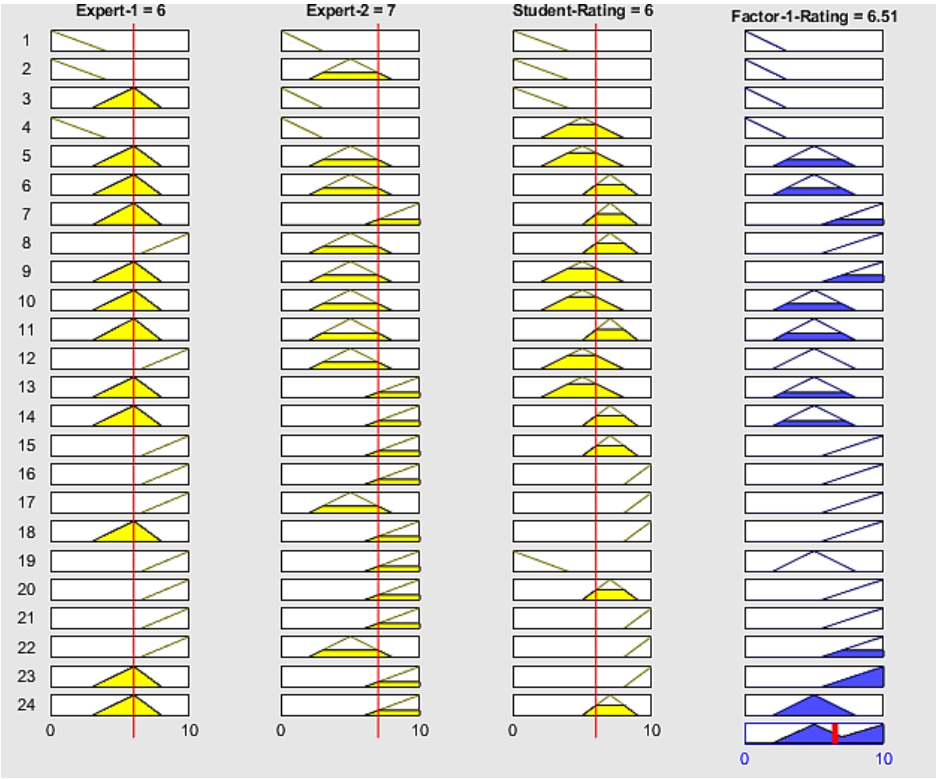


Figure 3 Demonstration of rules on fuzzy ANFIS.

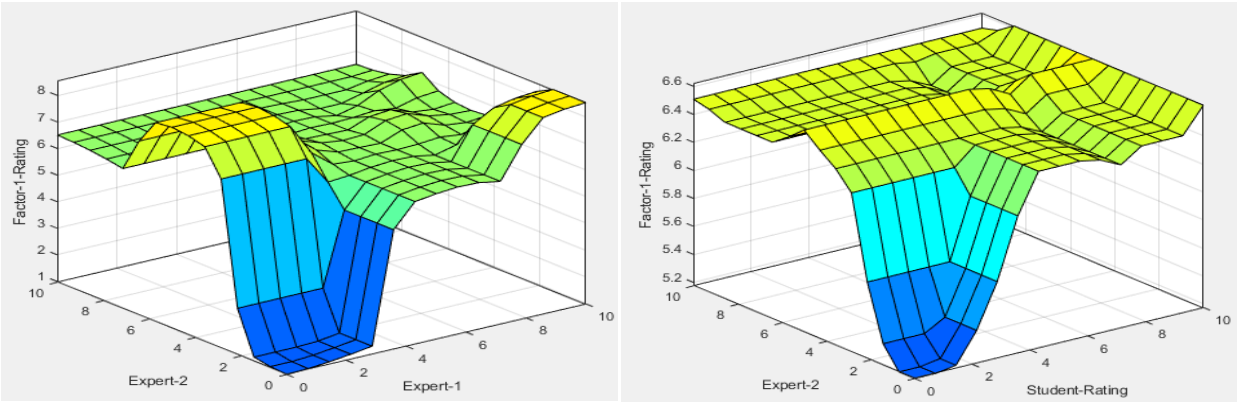


Figure 4 Factor rating with respect to expert-1 vs expert-2 and student rating vs expert-2.

Prediction using multiclass support vector machine

Support vector machines are supervised machine learning techniques with associative learning algorithms for which it can be used for data classifications and regressions. In other words, it can also be said that it is a discriminative classifier with a defined separating hyperplane. In 2-dimensional space, it will draw a hyperplane separating 2 classes where either side of the hyperplane indicates a class. The main objective is to draw a hyperplane in N-dimensional spaces (here N means the number of features) so that it can distinctly classify the data points. Hyperplanes should be at maximum distance from the data points so that future points can be classified with more confidence. SVM can also perform nonlinear classification using the Kernel trick. The main idea behind this kernel trick is to map these data to higher dimensional feature spaces so that they can be separated by a binary classifier [28]. Assume that dataset for training is represented by a set, $j = \{(x_i, y_i)\}_{i=1}^{l}$, here $(x_i, y_i) \in R^{n+1}$, l is the number of samples, n is the number of features and a class label $y_i = \{-1,1\}$. The separating hyperplane which is defined by the parameters w and b can be obtained by solving the following convex optimization problem [29].

$$\text{Min } \frac{1}{2}\|w\|^2$$

$$s.t.\, y_i(w^T\varphi(x_i) + b) \geq 1 \qquad i = 1, 2, \ldots\ldots, l$$

For implementing SVMs for more than 2 classes, 2 methods are used. They are one against all (OAA) and one against one (OAO) [30]. In the OAA method, to solve a problem of n classes, n binary problems are solved instead of solving a single problem. Each classifier is mainly used to classify 1 single class that's why points on that class will give positive responses and points belonging to other classes will give negative values on that classifier. In the case of OAO, for n class problems $\frac{n(n-1)}{2}$ SVM classifiers are constructed and each of them is trained to separate 1 class from another. If an unknown point is to be classified, each SVM votes for a class and the class with maximum votes is considered as the final result.

Framework for SVM model

The selection of features is very important for classification problems. In this work, factors which were identified for students' different level of performance by the survey were the distinguishing features for this multiclass classification problems. That is why these factors were used as features for separating and predicting different classes. The step by step breakdown of the model is given below;

1) Identification of features for data points' separation by classes
2) Division of the data points into the classes accordingly
3) Training of SVM model with the help of training data and its class labels
4) Testing of data into the trained model
5) Comparison of classes' predictions obtained from the SVM model with the actual results.

Prediction using random forest classifier

For generating a prediction model, the RFC needs the definition and insertion of 2 parameters, the number of classification trees desired, and several predicting variables that are used in each node to grow the trees. For each node, the best split is done by searching selected features. Thus, RFC consists of *N* decision trees where *N* is a user-defined value about the number of trees to be grown. When new data points are to be classified, these are passed down to all those trees and then it chooses its class by maximum votes out of *N* votes [31].

Framework for RFC model

The framework for the RFC model is represented by a flowchart in **Figure 5**. Input data with various features and an output attribute with different levels are split into 2 datasets: Training dataset and testing dataset. Then bootstrap aggregating and attribute bagging are developed to form a randomly selected decision tree by minimizing the misclassification rate. Finally, the testing dataset is examined to predict the class.

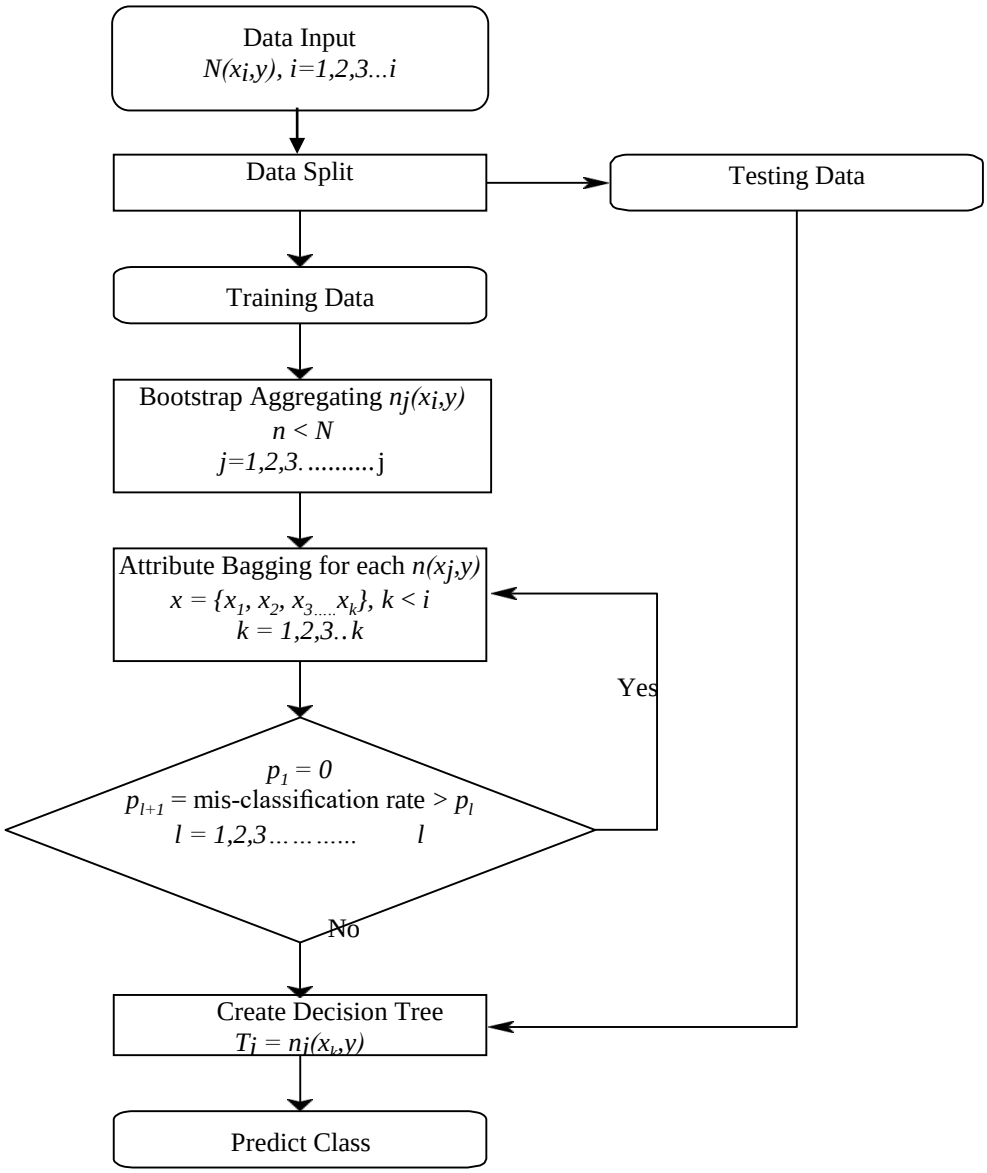


Figure 5 Flowchart of the RFC model.

Results and discussion

Prediction of classes using SVM classifier

After merging all the datasets to a single dataset using fuzzy analysis, 80 % of these data were used as a training dataset for 2 machine learning algorithms (SVM and RF classifier). In SVM, the significant factors were used as distinguishing features for class separation.

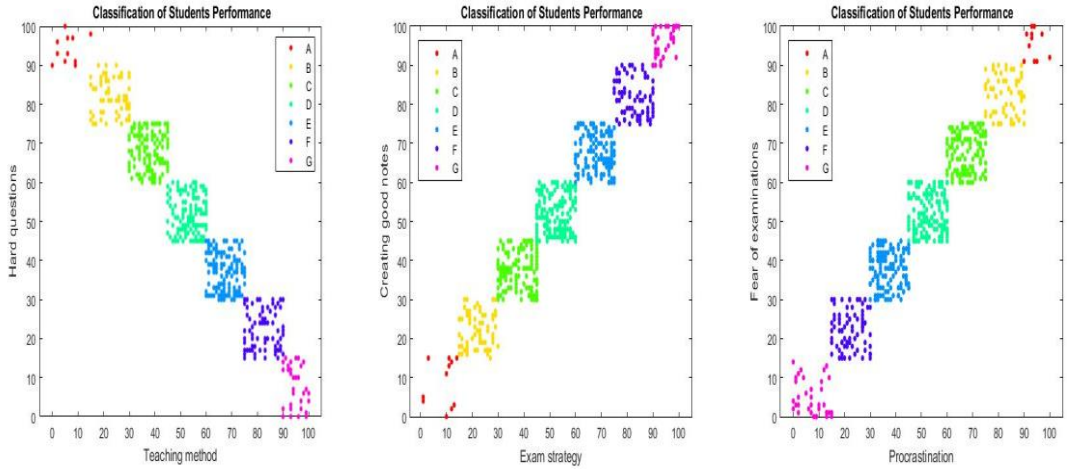


Figure 6 Classification of students based on factors (hard questions and teaching method, creating good notes and exam strategy, fear of examinations and procrastinations).

Three cluster diagrams are shown in **Figure 6** using training data considering any 2 factors. The classes are very well separated from each other. For example, using only 2 factors (hard questions and teaching method) we can see that each class has occupied a specific region in the graph. Thus, the class of future data falling in any of these regions can be predicted with high accuracy. Considering all these significant factors, SVM can classify each class more distinctively. The SVM model was also tuned for performance optimization. Here, tuning was performed by varying the values of 3 parameters, namely, - epsilon, gamma, and cost functions. Epsilon is a measure of misclassification errors. Usually, SVM optimizes the model stepwise against a certain measure, and each time it converges then stops to see whether the model is good enough or not. In this way, the strictness of the model's optimization is done by epsilon. In SVM, gamma means the width or slope of the kernel function. With the low value of gamma, the decision region becomes broad which lowers the accuracy of SVM while the higher value of gamma improves the accuracy of SVM. Cost function implies the amount by which misclassification of training examples should be penalized in a particular model. For the higher value of cost, separating the hyperplane's margin will be smaller thus it can classify the training examples more correctly. And when its value is less, a reverse case occurs where the model misclassified more points. **Figure 7** shows the effect of different values of epsilon and cost functions on the performance of SVM. As the region gets darker, the better performance of SVM can be achieved and it clearly shows that performance level does not depend on epsilon but becomes excellent after certain values of cost. On the other hand, it is seen that the performance of SVM does not depend on cost but becomes extremely good when gamma value crosses a certain point.

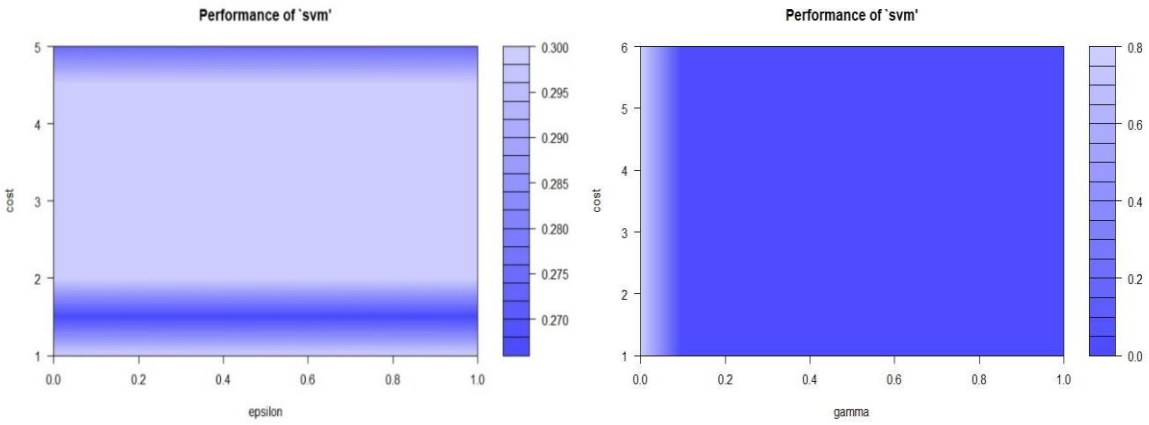


Figure 7 Tuning of the model by changing cost, gamma, and epsilon functions' values.

The values of cost, gamma, and epsilon were varied from 1 to 6, 0 to 1, and 0 to 1, respectively to generate 600 different models. Among those models, the best model is represented in **Table 7** based on performance. The error rate varies with the model parameters as shown in **Figure 8**.

Table 7 Best predicting models after tuning of models.

Parameters	Kernel	Cost	Gamma	Epsilon	Ntree	m_{try}
Best SVM model	Radial	1	0.1	0	-	-
Best RFC model	-	-	-	-	150	3

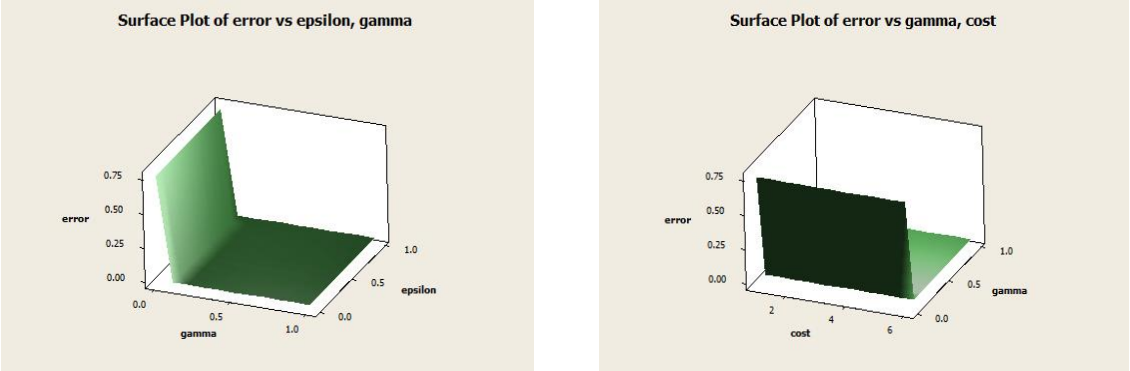


Figure 8 3D view of error vs model parameter.

Prediction of classes' using RF classifier

Random forest was created using a training dataset where 500 trees were built. In **Figure 9**, one of the random trees is presented where voting for each node is depicted. For instance, node-3 has maximum votes for class 7.

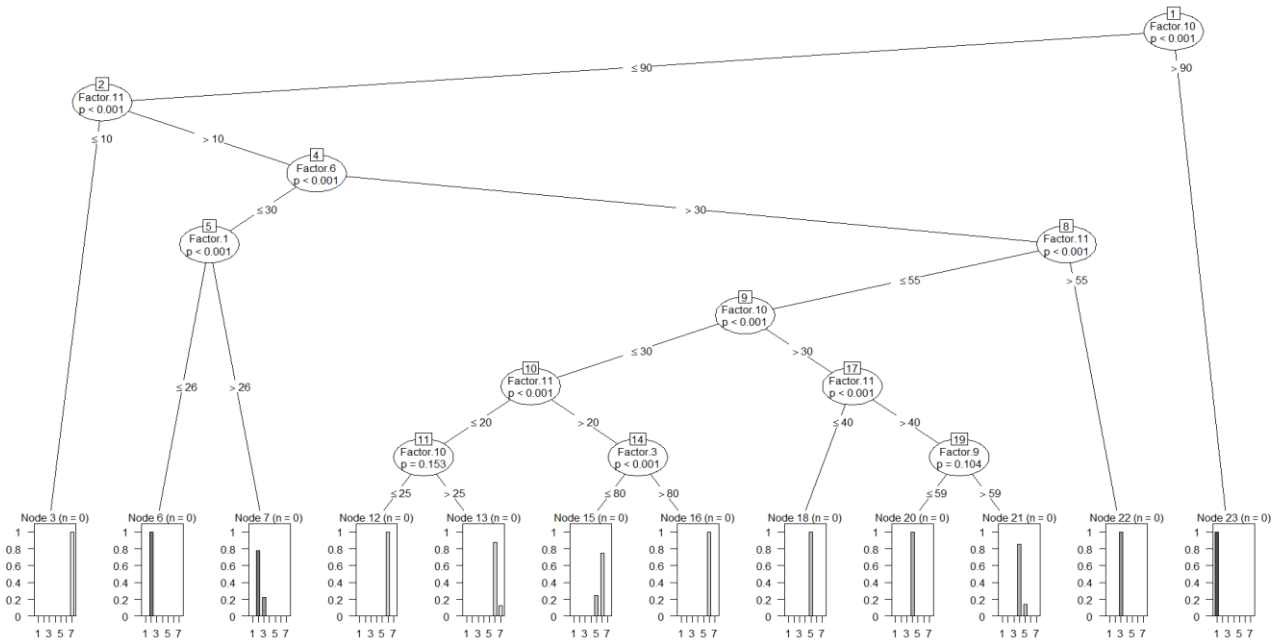


Figure 9 One of the random forest trees with probable outcomes.

After that, error rate was calculated which is shown in the **Figure 10**. We can see that above 150 trees, the error rate becomes constant. That is why 150 trees were selected to build the RF model. The error rate for 150 trees is also given in **Figure 10**.

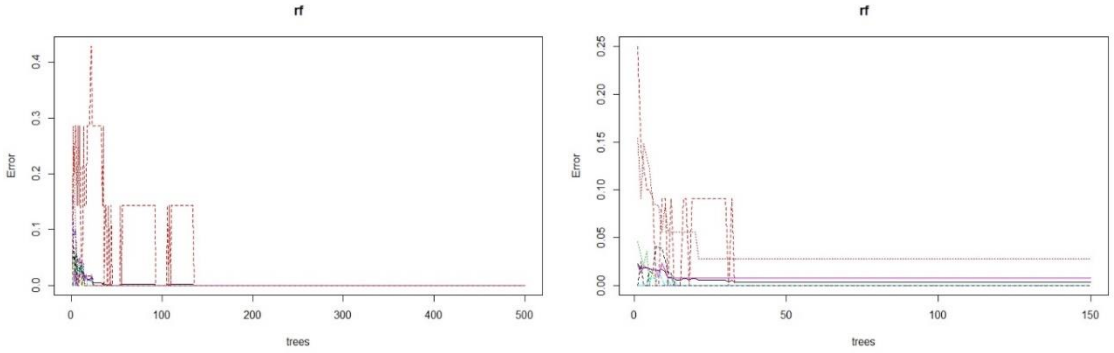


Figure 10 Error rate concerning no of trees.

Another important parameter of RF classifier is the number of attributes used in attribute bagging process which is m_{try}. Analyzing the OEB error to find out the suitable m_{try}. The relationship between OEB error and m_{try} is depicted in **Figure 11**. The lowest OEB error is found when the value of m_{try} is less than 3.

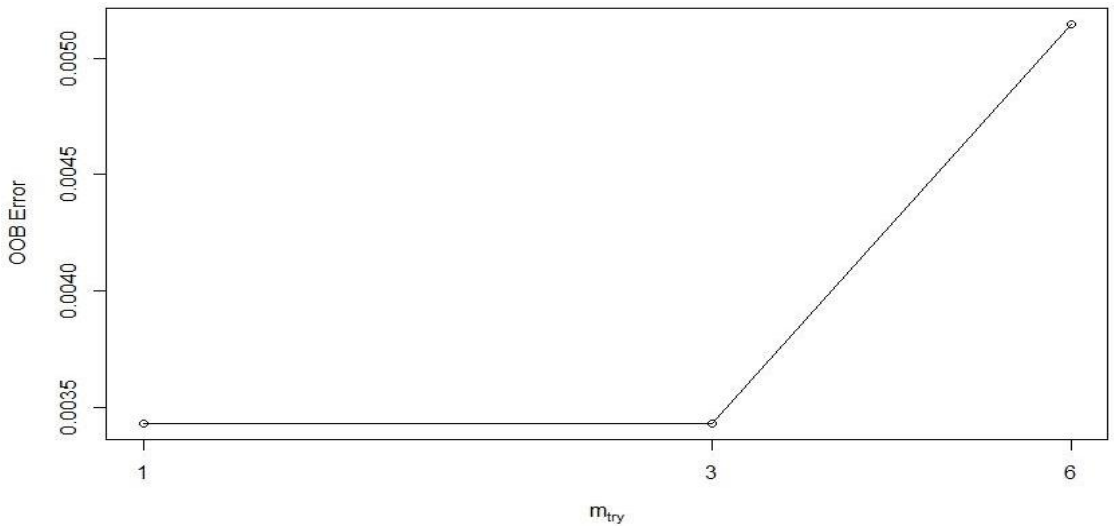


Figure 11 OEB error vs m_{try} graph.

The partial dependency of individual factors on different classes was also investigated. For example, the partial dependency of factor-1 for 2 distinct classes is shown in **Figure 12**. It depicts that if the factor value is between 60 to 80 it gives a more accurate value for predicting class 5 while it predicts more accurately the class 3 when it is below 45.

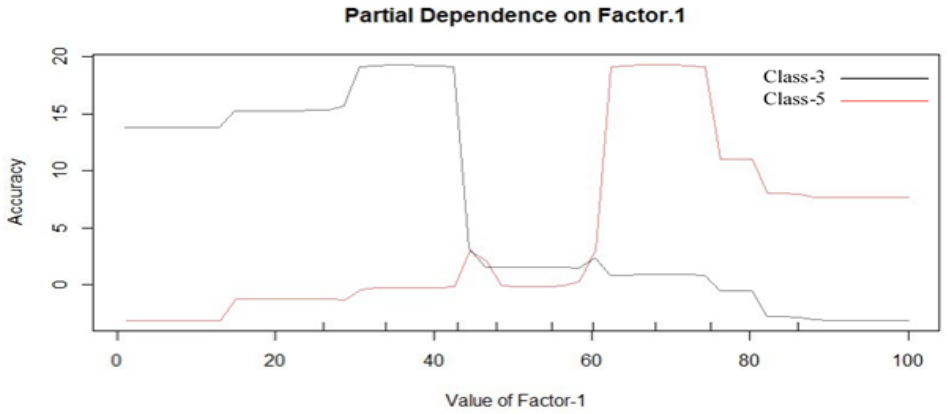


Figure 12 Partial dependency on factor 1.

Comparative analysis

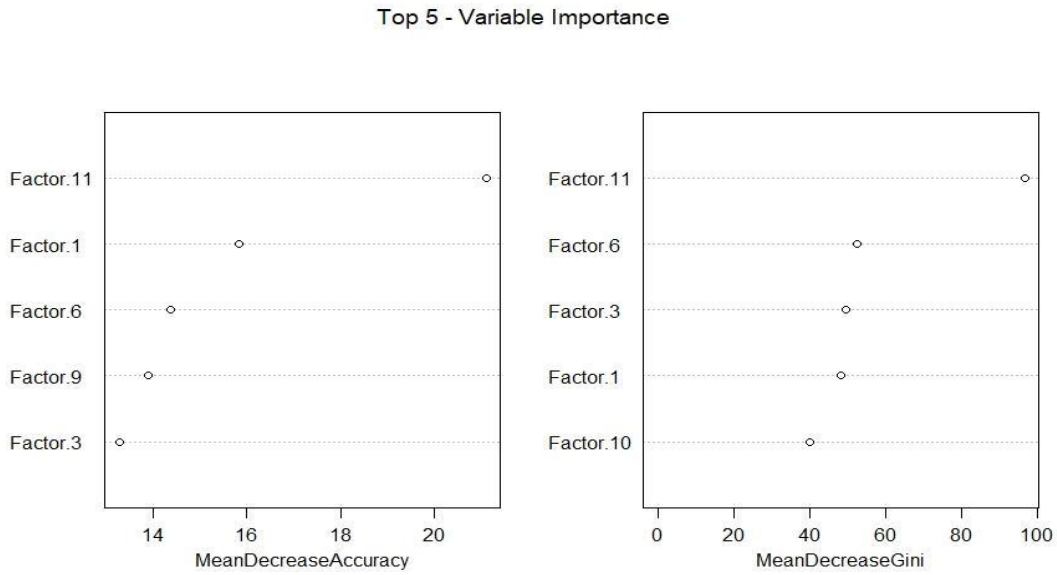
The accuracy of the SVM model is 81.25%. The confusion matrix for this model is presented in **Table 8**. On the other hand, the accuracy of the RF classifier is 96.88%. The confusion matrix for this model is shown in **Table 9**. Both models cannot predict class 6 accurately. The important factors were also analyzed using RFC which is depicted in **Figure 13**. We can see the top 5 important variables. **Figure 13(a)** shows how the model performance degrades without each variable.

Table 8 Confusion matrix for SVM model.

		Predicted Class by SVM						
		1	**2**	**3**	**4**	**5**	**6**	**7**
Actual Class	1	1	0	0	0	0	0	0
	2	0	6	0	0	0	0	0
	3	0	0	15	1	0	0	0
	4	0	0	1	1	1	0	0
	5	0	0	0	2	2	1	0
	6	0	0	0	0	0	0	0
	7	0	0	0	0	0	0	1

Table 9 Confusion matrix for RFC model.

		Predicted Class by RFC						
		1	2	3	4	5	6	7
Actual Class	1	1	0	0	0	0	0	0
	2	0	6	0	0	0	0	0
	3	0	0	16	0	0	1	0
	4	0	0	0	4	0	0	0
	5	0	0	0	0	3	0	0
	6	0	0	0	0	0	0	0
	7	0	0	0	0	0	0	1



(a) (b)

Figure 13 Top 5 factors that affect student performance most.

Factor-11 (effort of the student) is the most important factor that is responsible for better prediction of student performance. In **Figure 13(b)**, it measures how pure the nodes are at the end of the tree without each variable. For factor-11, 6, 3, 1, & 10 mean decreases in Gini is very high. Thus, these factors are responsible for the student's performance. Authority needs to focus on those factors so that the students' performance can enhance.

Conclusions

In this work, we propose a hybrid model for predicting first-year university students' performance in the final examinations. The fuzzy ANFIS is incorporated with RFC and SVM to develop that model. Our experimental results illustrated that our proposed model is proved to be effective and pragmatic for the accurate prediction of students' progress, as compared to some traditional machine learning algorithms. The prediction of RF and SVM model is 96.88 and 81.25 %, respectively. So, the RFC outperforms the SVM classifier in terms of accuracy. It is also found that the students' effort and creating good notes (factor-11 and factor-1) are the 2 key factors that have a great impact on students' success. So, early detection of these factors' ratings could yield valuable insights for a better educational environment and assistance to students' performance.

The main focus of this study was to predict students' academic performances based on varied attributes so that students can improve their conditions for achieving good results. Because in many cases, the main reason for dropout of students is when they lack in motivation due to their poor academic results. If these poor academic results of students can be improved or prevented, students won't be demotivated. That's why first year students were considered for this research work because if the attributes for obtaining better results can be installed in the very first year of students they will be aware of the influence varied factors have on their academic performance and can act accordingly to prevent the situation. For better understanding of prediction of students' overall results up to graduation level, second, third and fourth year should be considered for data collection because some other factors like result of previous year/ years, adaptation in the university etc. should be included in that case. But prediction of students' results up to graduation level was not the primary focus in this work that's why only first year students were considered. Though, data were collected from the engineering students but it can be applied in any branches of education in the university level. Because the model formulation and identification of relevant factors responsible for students' academic results were not dependent on a particular department or branch of the education.

In conclusion, we point out that the students' attributes implemented in our work are not bounded rather more new attributes can be introduced in our database to improve the quality of our model. Maybe

there are varied hidden factors which were overlooked in our work or day by day new factors/ problems will be emerging which will have direct or indirect impact on students' results. Thus, identification and addition of new factors will increase the accuracy of prediction of students' performances. Not only new attributes but also more experts can be added to get more insight about factor ratings. This proposed model can be extended to analyze the senior students' performance also. However, more data could be integrated by considering other university students' conditions, which would be more versatile. In addition, students' progress can be evaluated subject-wise.

Acknowledgements

We would like to thank faculty members and students (45th batch, 2018) of Bangladesh University of Textiles, Bangladesh to participate in the survey without which we would not be able to conduct our research.

References

[1] T Devasia, TP Vinushree and V Hegde. Prediction of students performance using educational data mining. *In*: Proceedings of the 2016 International Conference on Data Mining and Advanced Computing, Ernakulam, India. 2016, p. 91-5.

[2] YJ Lee. Predicting students' problem solving performance using support vector machine. *J. Data Sci.* 2016; **14**, 231-44.

[3] A Acharya and D Sinha. Early prediction of students performance using machine learning techniques. *Int. J. Comput. Appl.* 2014; **107**, 37-43.

[4] SM Hasheminejad and M Sarvmili. S3PSO: Students' performance prediction based on particle swarm optimization. *J. AI Data Min.* 2019; **7**, 77-96.

[5] PA Patil and RV Mane. Prediction of students performance using frequent pattern tree. *In*: Proceedings of the 6th International Conference on Computational Intelligence and Communication Networks, Bhopal, India. 2014, p. 1078-82.

[6] IE Livieris, V Tampakas, N Kiriakidou, T Mikropoulos and P Pintelas. *Forecasting students' performance using an ensemble SSL algorithm*. *In*: MA Tsitouridou, JA Diniz and T Mikropoulos (Eds.). Technology and innovation in learning, teaching and education. Vol 993. Springer, Cham, Switzerland, 2018, p. 566-81.

[7] F Okubo, A Shimada, T Yamashita and H Ogata. A neural network approach for students' performance prediction. *In*: Proceedings of the 7th International Learning Analytics & Knowledge Conference, Vancouver BC, Canada. 2017, p. 598-9.

[8] AB Raut and AA Nichat. Students performance prediction using decision tree technique. *Int. J. Comput. Intell. Res.* 2017; **13**, 1735-41.

[9] F Okubo, T Yamashita, A Shimada and S Konomi. Students' performance prediction using data of multiple courses by recurrent neural network. *In*: Proceedings of the 25th International Conference on Computers in Education, Christchurch, New Zealand. 2017, p. 439-44.

[10] ET Lau, L Sun and Q Yang. Modelling, prediction and classification of student academic performance using artificial neural networks. *SN Appl. Sci.* 2019; **1**, 982.

[11] V Tampakas, IE Livieris, E Pintelas, N Karacapilidis and P Pintelas. Prediction of students' graduation time using a two-level classification algorithm. *In*: M Tsitouridou, JA Diniz and T Mikropoulos (Eds.). Technology and innovation in learning, teaching and education. Vol 993. Springer, Cham, Switzerland, p. 553-65.

[12] M Chauhan and V Gupta. Comparative study of techniques used in prediction of student performance. *World Sci. News.* 2018; **113**, 185-93.

[13] A Daud, MD Lytras, NR Aljohani, F Abbas, RA Abbasi and JS Alowibdi. Predicting student performance using advanced learning analytics. *In*: Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia. 2017, p. 415-21.

[14] Y Abubakar and NBH Ahmad. Prediction of students' performance in e-learning environment using random forest. *Int. J. Innovat. Comput.* 2017; **7**, 1-5.

[15] C Beaulac and JS Rosenthal. Predicting university students' academic success and major using random forests. *Res. High. Educ.* 2019; **60**, 1048-64.

[16] MB Sanzana, SS Garrido and CM Poblete. Profiles of Chilean students according to academic performance in mathematics: An exploratory study using classification trees and random forests. *Stud. Educ. Eval.* 2015; **44**, 50-9.

[17] Y Ao, H Li, L Zhu, S Ali and Z Yang. The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling. *J. Petrol. Sci. Eng.* 2018; **174**, 776-89.
[18] SK Ghosh, N Zoha and F Sarwar. A generic MCDM model for supplier selection for multiple decision makers using fuzzy TOPSIS. *In*: Proceedings of the 5th International Conference on Engineering Research, Innovation and Education, Sylhet, Bangladesh. 2019, p. 833-40.
[19] D Wu, C Jennings, J Terpenny, RX Gao and S Kumara. A comparative study on machine learning algorithms for smart manufacturing: Tool wear prediction using random forests. *J. Manuf. Sci. Eng.* 2017; **139**, 071018.
[20] SK Ghosh and F Janan. Prediction of student's performance using random forest classifier. *In*: Proceedings of the 11th Annual International Conference on Industrial Engineering and Operations Management, Singapore. 2021, p. 7088-100.
[21] F Janan and SK Ghosh. Prediction of student's performance using support vector machine classifier. *In*: Proceedings of the 11th Annual International Conference on Industrial Engineering and Operations Management, Singapore. 2021, p. 7078-88.
[22] EM Jordaan and GF Smits. Robust outlier detection using SVM regression. *In*: Proceedings of the IEEE International Joint Conference on Neural Networks, Budapest, Hungary. 2004.
[23] H Al-Shehri, A Al-Qarni, L Al-Saati, A Batoaq, H Badukhen, S Alrashed, J Alhiyafi and SO Olatunji. Student performance prediction using support vector machine and k-nearest neighbor. *In*: Proceedings of the 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering, Windsor ON, Canada. 2017.
[24] T Mahboob, S Irfan and A Karamat. A machine learning approach for student assessment in e-learning using quinlan's C4.5, naive bayes and random forest algorithms. *In*: Proceedings of the 19th International Multi-Topic Conference, Islamabad, Pakistan. 2016, p. 1 - 8.
[25] E Mooi, M Sarstedt and I Mooi-Reci. Hypothesis testing & ANOVA. *In*: E Mooi, M Sarstedt and I Mooi-Reci (Eds.). Market research. Springer, Singapore, 2018, p. 153-214.
[26] MT Sow. Using ANOVA to examine the relationship between safety & security and human development. *J. Int. Bus. Econ.* 2014; **2**, 101-6.
[27] LA Zadeh. Fuzzy logic. *Computer* 1998; **21**, 83-93.
[28] C Cortes and V Vapnik. Support-vector networks. *Mach. Learn.* 1995; **20**, 273-97.
[29] SL Salzberg. C4.5: Programs for machine learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Mach. Learn.* 1994; **16**, 235-40.
[30] S Cheong, S Oh and S Lee. Support vector machines with binary tree architecture for multi-class classification. *Neural Inform. Process. Lett. Rev.* 2004; **2**, 47-51.
[31] L Breiman. Random forests. *Mach. Learn.* 2001; **45**, 5-32.