Prediction of Student's Performance Using Support Vector Machine Classifier

Farhatul Janan
Department of Industrial and Production Engineering
Bangladesh University of Textiles
Tejgaon, Dhaka- 1208, Bangladesh
janan_ipe@butex.edu.bd

Sourav Kumar Ghosh
Department of Industrial and Production Engineering
Bangladesh University of Textiles
Tejgaon, Dhaka- 1208, Bangladesh
sourav@butex.edu.bd

Abstract

Analyzing students' performances based on both subjective and quantitative components is fundamental because sometimes these performances and so many other factors have led students to quit their studies or be dropped out of the institutions. This dropout rate is much higher in undergrad students than other educational stages. The first-year result of a student is very vital since within the larger part of cases this drives them to be either inspired or demotivated. So, the second-year result of an eminent university in Bangladesh is examined in this paper. This paper is basically based on finding the factors for students' distinctive sorts of performances and after that predicting students' results based on those six noteworthy variables. For this reason, support vector machines (SVM) has been utilized for classifying students' diverse levels of outcomes and anticipating students' performances. The input dataset for both training and testing were taken by merging the values obtained from two surveys done on students and experts utilizing an adaptive neuro-fuzzy interference system (ANFIS). The application of the proposed model can moreover be enhanced in predicting course-wise performances of the students and its precision can too be improved by adding new factors, increasing survey participants

Keywords

SVM Classifier, Fuzzy Logic, Performance Prediction, Classification Criteria.

1. Introduction

Dropping out of any stage of education is also a common phenomenon. Many factors play an important role in increasing dropout rates. One of the main reasons for dropping out of school is poor grades. This impacts their performance as many students are unable to adapt to the university's learning environment after arriving at university. There are other causes such as student involvement in politics and extracurricular activities. For these various known and unknown reasons, students' performances often tend to be poor, which affect outcomes. Therefore, it is necessary to analyze undergraduate grades to find the true root cause of students' different levels of performances. This study focuses primarily on revealing various factors that influence the performance at the undergraduate level.

Therefore, our primary inspiration behind this work was to assist students understand the characteristics that contribute to their poor results so that they can move improve their outcomes. When the major components are identified and checked, it'll give the students, course instructors, and others to enhance the environment. We have performed our study at the Bangladesh University of Textiles. This research is based on the results of second-year students (44th batch, 2018) of this university. For distinguishing the factors, we have surveyed students and a few experts (here experts mean distinctive course instructors related with students). Upon completion of the survey, factors and their scores were identified that showed the reasons for different levels of student performance. The scores of students and experts on a particular factor were merged using ANFIS analysis and then this modified data were used as input to this model. 80 percent of this data was used for training and the remaining 20 percent for testing, and finally, the

validation of the model (identification of factors and their ratings) was achieved by calculating the accuracy of the model. In this model, modified data were used instead of raw data and that's why this model performed better than traditional methods.

1.1 Objectives

The main objectives of this research is-

- To modify raw data of two surveys using fuzzy logic
- To use that modified data for the prediction of students' distinct level of outcomes using SVM classifier

2. Literature Review

Over the past few years, researchers have published many papers on identifying the reasons behind a student's performance. A few of these works incorporate only analyzing the traits which have direct or indirect impacts on students' results and some also incorporate predicting students' outcomes based on the considered traits utilizing diverse learning algorithms. Most of these learning calculations are artificial intelligence (AI) methods such as artificial neural systems, support vector machines, Bayesian classifiers, etc.

Socio-economic variables and entrance examination outcomes were utilized as the essential variables to anticipate the student's cumulative grade point average (CGPA) by applying ANN with the Levenberg–Marquardt algorithm [1]. The decision tree (DT) method worked more effectively than other machine learning (ML) algorithms on a case study of few undergraduate students in Kolkata [2]. Frequent pattern tree algorithm, ensemble semi-supervised learning (SSL) algorithm, recurrent neural network (RNN) and DT techniques were applied to predict student's results [3]–[7]. The classification algorithm occurring at two levels was applied for analyzing the anticipated graduation time where the passed or failed students were separated in the first level and three distinctive periods of graduations were classified within the following level [8]. SVM performed better than Bayesian Knowledge Tracing (BKT) in the prediction of students' problem-solving outcomes by appearing roughly 29 percent advancement, compared to the standard BKT method [9]. Methods were grouped into three groups such as fuzzy logic, data mining techniques, and hybrid [10]. A hybrid model was shaped utilizing Bayes network (BN) and Naive Bayes (NB) whereas generative models while SVM, C4.5, and Classification and Regression Tree (CART) were utilized as discriminative models [11]. Fuzzy ANFIS was utilized to change over numerous choice maker's evaluations into a final rating based on 78 fuzzy logic [12].

To our best information, prediction of students' result using fuzzy ANFIS and SVM is yet to be done. In this paper, we have defined the problem as a multi-class classification issue. We have applied fuzzy logic which had merged distinctive sets of input data under certain rules to a single input database. After that, (SVM) has been utilized for the prediction of students' results.

The rest of the paper is organized as follows. In section 3, the methodologies of fuzzy logic and SVM are discussed. Data collection are demonstrated in Section 4. Results and analysis are shown in Section 5. Finally, in section 6, conclusions and its future directions are given.

3. Methods

This work centers on the investigation of diverse sorts of variables like psychological, individual, university facilities, learning environment, etc. that influence students' outcomes and prediction of student's performances based on these variables utilizing machine learning algorithms. Particularly, this work can be classified into four major steps as follows:

1. Identification of variables through two surveys, one for students and another for experts
2. Merging the raw data obtained from two surveys using fuzzy ANFIS to get the modified input data
3. Classification and prediction of student's outcomes based on SVM classifier
4. Evaluation of the precision of the model by analyzing the anticipated comes about with the actual results

In the first phase, a 21-question survey was conducted through Google's Answer Form, which received elemental ratings from students. Here, course teachers and student counselors were considered experts. In the next step, reduction of factors was performed through factor analysis and another expert study was conducted based on the selected factors. Next, the two survey data were merged using ANFIS model. Students were divided into 7 classes based on the GPA of the previous semester. Table 1 shows how students were divided into categories.

Table 1: GPA range of different classes

CGPA range	2.26-2.50	2.51-2.75	2.76-3.00	3.01-3.25	3.26-3.50	3.51-3.75	3.75-4.00

Class no.	1	2	3	4	5	6	7
Mean	2.46	2.66	2.89	3.13	3.38	3.62	3.86
Students no.	11	65	125	141	129	77	35

Then, 80 percent of these data were used as training data for both SVM classifier. The other 20 percent were used for testing. After that, results were analyzed with relative advantages and disadvantages. This whole process is shown in a flowchart in Figure 1.

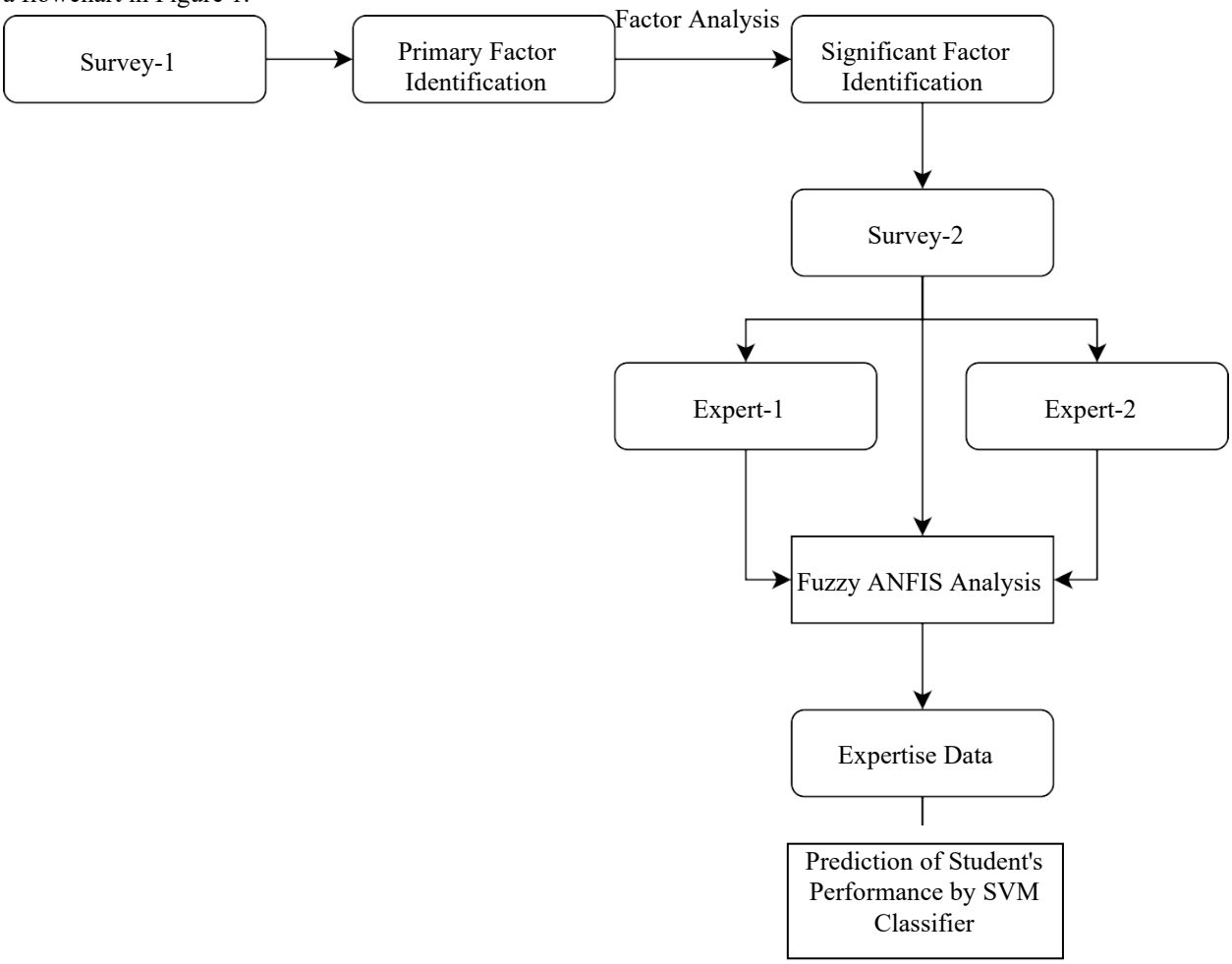

Figure 1: Process flowchart of the proposed model

3.1 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 [13].

Architecture: Its architecture contains four parts-

- Rule base: It contains IF-THEN rules provided by the experts which help to govern the decision-making system.
- Fuzzification: It is used for converting inputs which are called crisp numbers into fuzzy sets.
- 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.
- Defuzzification: It is used to convert the fuzzy sets obtained by the inference engine into a crisp value which is the ultimate output.

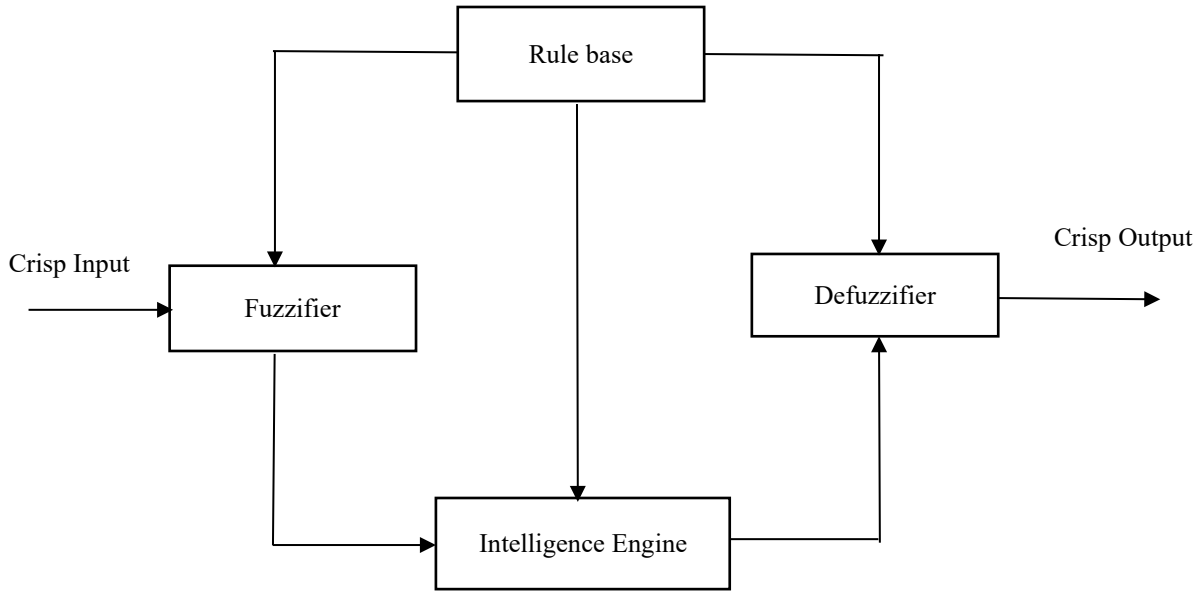

Figure 2: Fuzzy Logic structure

Membership Function: It 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 three types of fuzzifier:

- Singleton fuzzifier
- Gaussian fuzzifier
- 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.

3.2 Prediction using Multiclass Support Vector Machine

Support vector machines are supervised machine learning methods with affiliated learning techniques for which it can be utilized for classifications and regressions. In other words, it can too be said that it is a discriminative classifier with a defined separating hyperplane. In two-dimensional space, it'll draw a hyperplane isolating two classes where either side of the hyperplane demonstrates a class. The principal aim is to draw a hyperplane in N-dimensional spaces (here N implies the number of variables) so that it can classify the data points. Hyperplanes ought to be at maximum distance from the data points so that future points can be classified with more certainty. SVM can moreover perform nonlinear classification utilizing 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 [14]. 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 [15].

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

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

For actualizing SVMs for more than two classes, two strategies are utilized. They are one against all (OAA) and one against one (OAO) [22]. Within the OAA strategy, to unravel an issue of n classes, n binary problems are solved rather than fathoming a single issue. Each classifier is basically utilized to classify one single class that's why values on that class will grant positive response and points on other classes will give negative values on that classifier. Within the case of OAO, for n course issues $\frac{n(n-1)}{2}$ SVM classifiers are built and each of them is prepared to partitioned one class from another. In the event that an unknown point is to be classified, each SVM votes for a class and the class with most extreme votes is considered as the ultimate result.

3.3 Framework for SVM Model

Feature Selection is extremely important for classification issues. In this work, variables which were recognized for students' diverse level of outcomes were the recognizing features for this multiclass classification issues. That's why these variables were utilized as the distinguishing features for isolating and predicting distinctive classes. The step by step breakdown of the method is given below-

- Identify the important variables for data points' separation by classes
- Divide the data points into the classes accordingly
- Training of SVM model with the assistance of training datasets and its associated class labels
- Test the data into the trained model
- Compare the predictions of classes gotten from the SVM with the genuine results.

4. Data Collection

4.1 Identification of Significant Factors

In this paper, a survey consisting of 21 questions was first conducted on second-year students (44th batch) of the Bangladesh University of Textiles to find their evaluations on each factor which could be connected to their diverse levels of outcomes. These 21 variables which were utilized in survey-1 are given in Table 2. After the completion of Survey-1, significant factors were determined based on student assessment on each factor. Analysis of variance (ANOVA) was utilized to check the correlation between different variables.

Table 2: List of all the factors for survey 1

Creating good notes	Group study	Adaptation to university
University facilities	Previous year questions	University environment
Exam Strategy	Class tests' marks	Discontent about the university
Time management	Fear of examinations	Political involvement
Class attendance marks	Hard questions	Family issues
Course difficulty level	Teachers' friendliness	Residential problems
Excessive inclination towards extra earning	Teachers 'hardness of checking exams' scripts	Personal problems (any kind of addictions or illegal activities)

For the factors' diminishment and distinguishing proof of critical variables, ANOVA has been utilized here. In Table 3, ANOVA tests' result is appeared where two variables, class attendance and teachers' friendliness were compared, and ANOVA with a 95 percent confidence interval was applied. As $F < F - crit$ or $p - value > 0.05$, the null hypothesis is accepted. That means the two factors are alike. Another case is given in Table 4, where components, exam strategy, and creating good notes, were compared in the same way. But the result here appears the opposite. As F>F-crit or p-value<0.05 which means null hypothesis is rejected and there's a difference between the variables.

Table 3: ANOVA table of two correlated factors

	Sum of Squares, SS	*df*	*MS*	*F*	*p-value*	*F-crit*	Hypothesis	Result
Between groups	0.085763	1	0.085763	0.000176	0.989429	3.84946	Rejected	Factors Correlated
Within groups	568431.2	1164	488.343					
Total	568431.3	1165						

Table 4: ANOVA table of two uncorrelated factors

	Sum of Squares, SS	*df*	*MS*	*F*	*p-value*	*F-crit*	Hypothesis	Result
Between groups	7641.702	1	7641.702	15.96127	6.87E-05	3.84946	Accepted	Factors uncorrelated
Within groups	557282.8	1164	478.7653					
Total	564924.5	1165						

In this way, utilizing the ANOVA tests, correlated and uncorrelated variables were recognized. At that point correlated components were eliminated and uncorrelated variables were considered as critical components (components dependable for students' distinctive levels of outcomes) which were utilized for the total classification problem. 6 critical variables were found which are appeared in Table 5.

Table 5: List of 6 significant factors

Factors' no.	Factors' name
1	Creating good notes
2	Exam strategy
3	Class tests marks
4	Personal problems
5	Fear of examinations
6	Political Involvement

4.2 Merging of factors' ratings by Fuzzy analysis

After finding 6 significant factors using ANOVA analysis, another survey was conducted on two experts (one expert was course teacher and another expert was course coordinator) with those factors. Now, there were three 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.

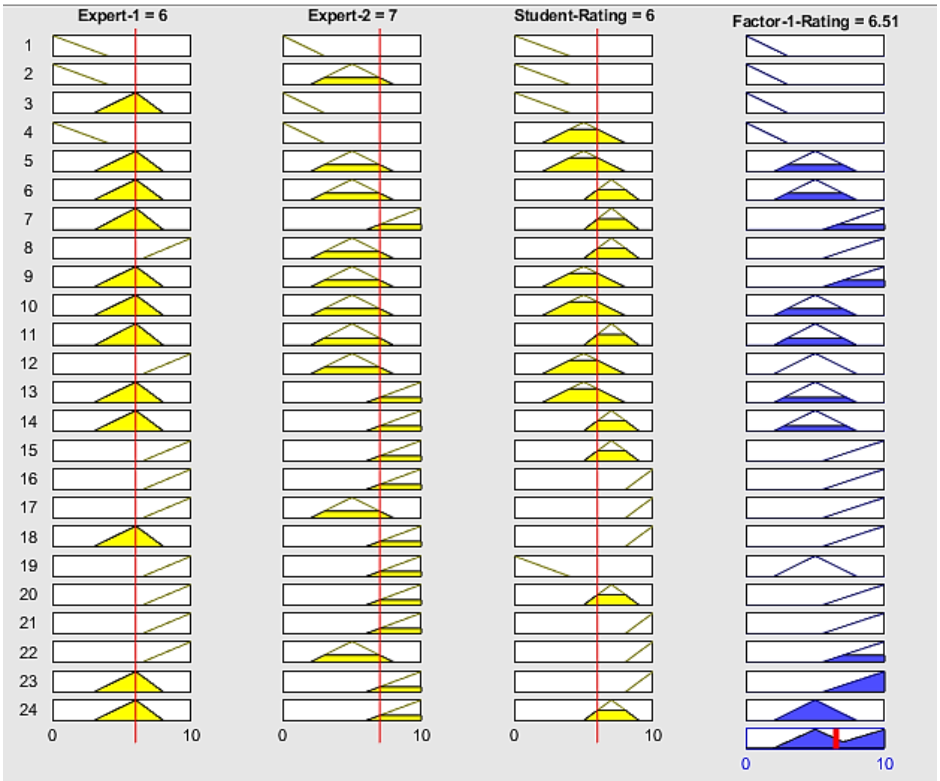

Figure 3: Demonstration of rules on fuzzy ANFIS

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.

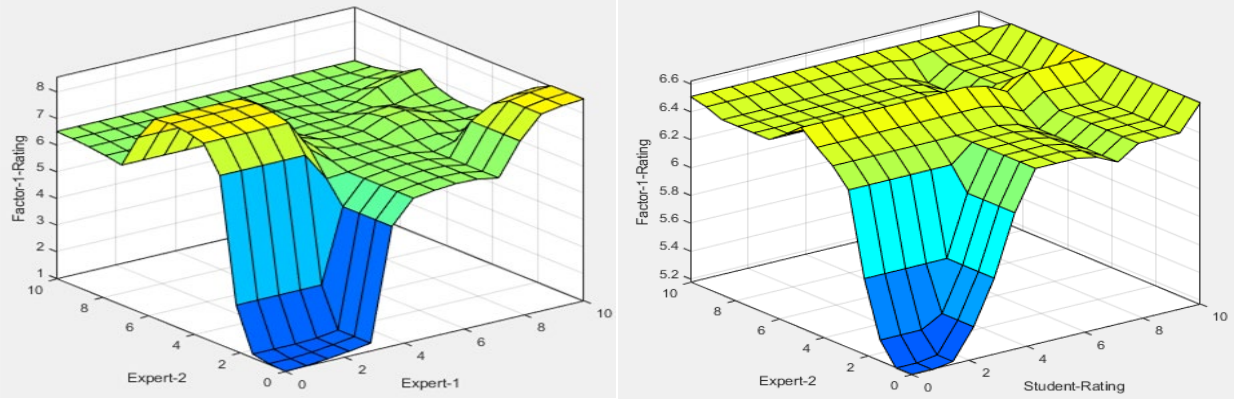

Figure 4: Factor rating with respect to expert-1 vs expert-2 and student rating vs expert-2

5. Results and Discussion

5.1 Prediction of classes using SVM classifier

After combining all the data sets into one data set using fuzzy analysis, 80 percent of that data was used as a training data set for SVM. In SVM, significant factors were used as distinguishing characteristics to separate classes.

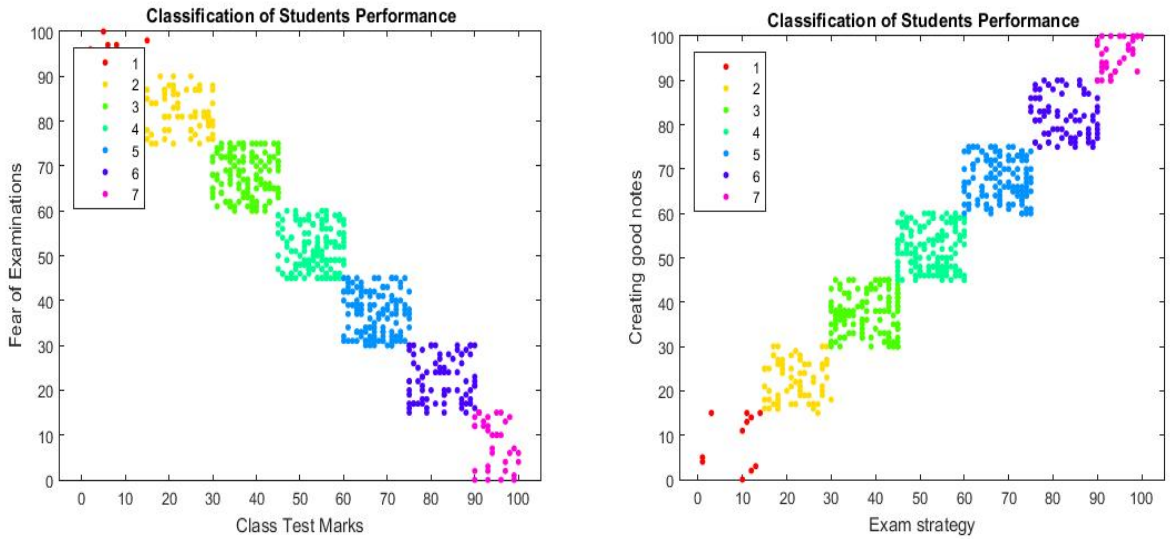

Figure 5: Classification of students based on factors (Fear of Examinations and Class Test Marks, Creating good notes and Exam Strategy,)

Two cluster diagrams are shown in Figure 5 using training data considering any two factors. The classes are very well separated from each other. For example, using only two factors (Fear of Examinations and Class Test Marks) 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 three 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 margin of hyperplane 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 6 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.

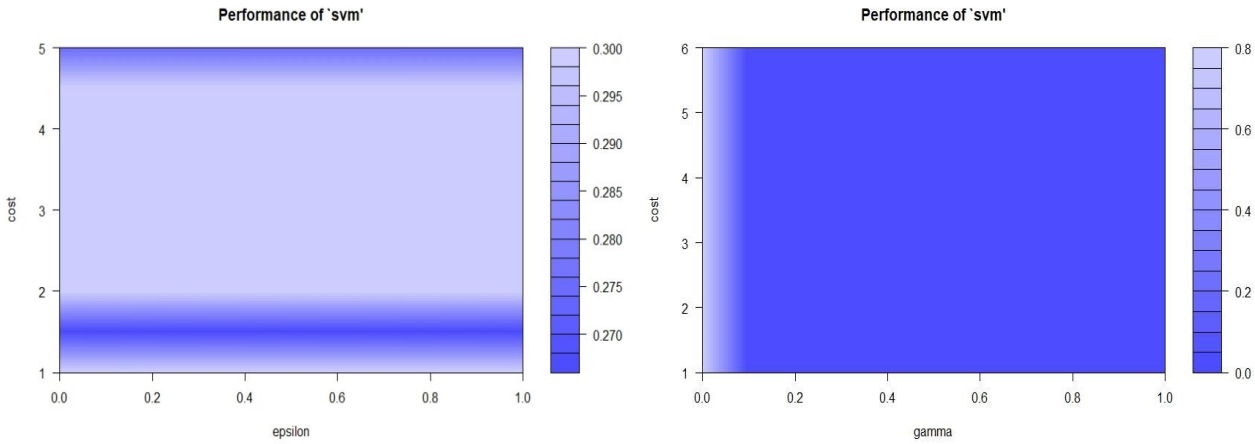

Figure 6: Tuning of the model by changing cost, gamma, and epsilon functions' values

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 6 based on performance. The error rate varies with the model parameters as shown in Figure 7.

Table 6: 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

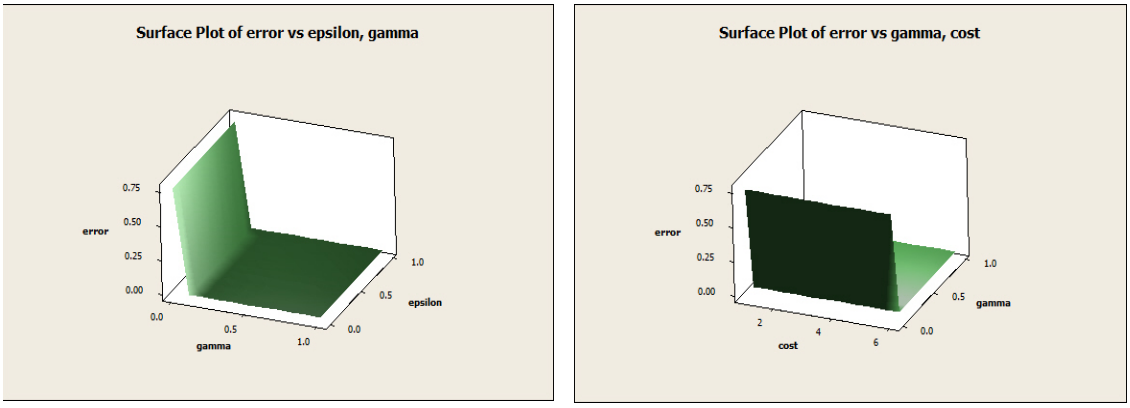

Figure 7: 3D view of error vs model parameter

5.2 Comparative Analysis

The accuracy of the SVM model is 81.25%. The confusion matrix for this model is presented in Table 7. The SVM model cannot predict class 6 accurately.

Table 7: 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

6. Conclusion

In this work, we propose a hybrid model to predict the performance of second-year university students in the final examinations. The fuzzy ANFIS is integrated with SVM to develop this model. Our experimental results show that our proposed model is more effective and practical for the accurate prediction of students' progress, as compared to

some traditional machine learning algorithms. The prediction of SVM model is 81.25%. Therefore, early detection of these factors' ratings can provide valuable insights to help improve the learning environment and students' performance.

Finally, we point out that the attributes of students applied to our work are not limited, but more new features can be introduced into our database to improve the quality of our models. Not only new features but also more experts can be added to learn more about element classification. Different artificial intelligence (AI) techniques can be used to increase prediction accuracy. The proposed model can also be used to analyze the performance of senior students. However, considering the situation of other students at the university, more data can be integrated, which will be more versatile. In addition, student progress can be assessed by subject.

References

[1] E. T. Lau, L. Sun, and Q. Yang, "Modelling, prediction and classification of student academic performance using artificial neural networks," *SN Applied Sciences*, vol. 1, no. 9, 2019, doi: 10.1007/s42452-019-0884-7.

[2] A. Acharya and D. Sinha, "Early Prediction of Students Performance using Machine Learning Techniques," *International Journal of Computer Applications*, vol. 107, no. 1, pp. 37–43, 2014, doi: 10.5120/18717-9939.

[3] P. A. Patil and R. v. Mane, "Prediction of Students Performance Using Frequent Pattern Tree," in *Proceedings of the 6th International Conference on Computational Intelligence and Communication Networks*, 2014, pp. 1078–1082, doi: 10.1109/CICN.2014.226.

[4] I. E. Livieris, V. Tampakas, N. Kiriakidou, T. Mikropoulos, and P. Pintelas, "Forecasting Students' Performance Using an Ensemble SSL Algorithm," in *International Conference on Technology and Innovation in Learning, Teaching and Education*, 2018, pp. 566–581, doi: 10.1007/978-3-030-20954-4_43.

[5] F. Okubo, A. Shimada, T. Yamashita, and H. Ogata, "A Neural Network Approach for Students' Performance Prediction," in *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 2017, pp. 598–599, doi: 10.1145/3027385.3029479.

[6] A. B. Raut and A. A. Nichat, "Students Performance Prediction Using Decision Tree Technique," *International Journal of Computational Intelligence Research*, vol. 13, no. 7, pp. 1735–1741, 2017, [Online]. Available: http://www.ripublication.com.

[7] 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, ICCE 2017, New Zealand*, 2017, pp. 439–444.

[8] V. Tampakas, I. E. Livieris, E. Pintelas, N. Karacapilidis, and P. Pintelas, "Prediction of students' Graduation time using a two-level classification algorithm," in *International Conference on Technology and Innovation in Learning, Teaching and Education*, 2018, pp. 553–565, doi: 10.1007/978-3-030-20954-4_42.

[9] Y.-J. Lee, "Predicting Students' Problem Solving Performance using Support Vector Machine.," *Journal of Data Science*, vol. 14, no. 2, pp. 231–244, 2016, [Online]. Available: http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=116345930&site=ehost-live.

[10] M. Chauhan and V. Gupta, "Comparative Study of Techniques Used in Prediction of Student Performance," *World Scientific News*, vol. 113, pp. 185–193, 2018.

[11] A. Daud, M. D. Lytras, N. R. Aljohani, F. Abbas, R. A. Abbasi, and J. S. Alowibdi, "Predicting student performance using advanced learning analytics," in *Proceedings of the 26th international conference on world wide web companion, Perth, Australia*, 2017, pp. 415–421, doi: 10.1145/3041021.3054164.

[12] S. K. 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 (ICERIE) Sylhet, Bangladesh*, 2019, pp. 833–840.

[13] L. A. Zadeh, "Fuzzy Logic," *Computer*, vol. 21, no. 4, pp. 83–93, 1988.

[14] C. cortes and V. Vapnik, "Support-Vector Networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1109/64.163674.

[15] S. Salzberg, "Book review: C4. 5: by j. ross quinlan. inc., 1993. programs for machine learning morgan kaufmann publishers," *Machine Learning*, vol. 16, pp. 235–240, 1994, doi: 10.1016/S0019-9958(64)90259-1.

Biography / Biographies

Farhatul Janan is a Lecturer in the Department of Industrial and Production Engineering at Bangladesh University of Textiles, Dhaka, Bangladesh. She previously worked as a Lecturer in the Department of Industrial and Production Engineering at Military Institute of Science and Technology, Dhaka, Bangladesh. She earned B.Sc. in Industrial and Production Engineering from Bangladesh University of Engineering and Technology, Dhaka, Bangladesh. She is an

ongoing student of M.Sc. in Industrial and Production Engineering in Bangladesh University of Engineering and Technology. Her research interests include operations research, machine learning and regression analysis, data science, supply chain management and logistics systems optimization, engineering economics and decision theory.

Sourav Kumar Ghosh is a lecturer in Industrial and Production Engineering at Bangladesh University of Textiles (BUTEX). He earned B.Sc. in Industrial and Production Engineering from Bangladesh University of Engineering and Technology (BUET) in 2017. He is a former lecturer in textile engineering at Primeasia University. He is currently enrolled in a Master's program in Industrial and Production Engineering at Bangladesh University of Engineering and Technology (BUET). He has published six journal papers and eight conference papers. S. K. Ghosh has completed several research projects under UGC. His research interests include machine learning, supply chain optimization, operation research, parameter optimization of CNC machines, renewable energy, and lean manufacturing.