Analysis of the Socio-economic Factors of Poor Academic Results and Predicting Probable Solutions of Major Factors

Shamima Akter Rozy

Department of Humanities and Social Science, Bangladesh University of Textiles

Sourav kumar Ghosh

Department of Industrial and Production Engineering, Bangladesh University of Textiles

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 recent pasts. At present, students’ dropout rate in university is gradually increasing and many bright students sometimes just cannot cope with the universities. This research is mainly based on finding the reasons for students’ different types of results and then predicting students’ performance based on those significant factors. Varied social and economic theories have been used for the identification of several factors for the surveys and stratified random sampling technique has been used for the collection of data. Significant factors were later identified using the analysis of variance (ANOVA) test. Then, two popular supervised machine learning algorithms have been used for classifying students’ different levels of results and predicting students’ performances, these are support vector machines (SVM) and random forests (RF) 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 two surveys done on students and experts using adaptive neuro-fuzzy inference system (ANFIS). The result exhibits that RF can perform the classification of multiple classes based on many distinguishing features with more confidence than SVM. Afterwards, important factors responsible for students’ poor performances were examined to find the probable solutions in contrast to those. This proposed model can also be applied to predict course-wise students’ performances.

Keywords: SVM, RF, Performance Prediction, Socio-economic factors, Machine Learning

1. 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. However the dropout rate increases as moving to upper stages especially in the undergraduate level. There are many socio-economic factors which may play a vital role in increasing the dropout rates. In the majority of the cases, when students’ results start degrading that is the main trigger point for increasing the probabilities of those students to be dropped out. This research is basically focused on finding these factors that affect students’ results at the undergraduate level.

University is a very important place for students where they shape their futures. It creates opportunities for students to fulfill their dreams and to achieve a purpose in life. Sometimes, the reverse case also occurs. After coming to university, many students cannot adapt themselves to the university’s environment for study where some other causes like some students become more involved in extra-curricular activities or students’ politics, some start to avoid studies because they dislike their departments, some don’t like the career they have ahead etc. Some students also have some financial problems, psychological problems, family issues etc. due to which their performances are hampered. Due to these various known and unknown reasons, students’ performances in the university in many cases tend to be low which in turn affects their results.

So, our main motivation behind this work was to help students, institutions, teachers, family members understand the attributes which are responsible for students’ poor results so that they can take necessary actions to improve their performance. If the major factors are identified and monitored, it will give the students, course teachers, and the administrations to ameliorate the study environment. On the other hand, if students can anticipate the reasons for degrading their results thus, they can work on those to improve their performances. Additionally, the institution (in this case university) itself can take some proper steps to improve students’ performances according to this study. We have performed our experiment at four universities i.e., Bangladesh University of textiles (BUTEX), Bangladesh University of Engineering and Technology (BUET), Dhaka University (DU), Jagannath University (JNU) in Bangladesh. For identifying the factors, we have done surveying on students and on some experts (here experts mean different course teachers associated with students). After completing the survey, factors and their ratings were identified which showed the reasons for students’ different levels of performance. We have merged the

Electronic copy available at: https://ssrn.com/abstract=4523122
ratings of students and experts on a particular factor on a particular student using fuzzy ANFIS analysis for using these modified data as input to the model. 80 percent of these data were used for training and the remaining 20 percent was used for testing (predicting students’ results based on those factors) and finally by computing the accuracy of the model, the validity of the formulation of the model (factors’ identification and their ratings) was accomplished. Recommendations for nullifying the impacts of those crucial factors were provided. In this model raw data is modified using expert opinions and fuzzy logic. For this reason, the model performed better than other traditional models which only considered raw data.

In this paper, we have formulated the problem as a multi-class classification problem. Different techniques are available to solve this problem. After analyzing the research works on classification problems and prediction of students’ performances, it was found that Random Forest (RF) classifier was used in many works and its performance was better than other methods. Using RF, important factors were identified, and probable solutions were suggested based on those factors. So, the main objective of this study was to identify the socio-economic factors behind students’ poor results and to find possible solutions to those factors.

The rest of the paper is organized as follows. Section 2 describes the related recent literary works. In section 3, the concepts of socio-economic theoretical frameworks, fuzzy logic, RF, and their methodologies are discussed. Section 4 covers the results and analysis including data description. Finally, in section 5, conclusions with discussions on the paper and its future directions are provided.

2. Literature Review

Over the years, many researchers have already done many works regarding analyzing the socio-economic factors responsible for students’ performance. The challenges faced by students in pre-entry, the initial entry and into university through various theoretical framework such as Spady’s sociological theory, Tinto’s integration theory, Bean’s psychological theory were discussed [1]. In these theoretical framework, various factors that can lead to poor academic performance, most notably individual student and university environment, interaction, attitudes, skills, academic and social system, students’ family background, academic potential, peer support, family responsibilities, finances, outside encouragements, lack of integration, adjustment, difficulty, isolation, learning and external obligation were pointed out. Different factors were identified for
the students of Urambo district’s community secondary responsible for their academic performance and various rules of teaching process and commitment of student learning process was also examined to improve the academic performance [2]. Some factors were also mentioned to mitigate academic performance such as providing conductive working environment to teachers, adequate supply of teaching and learning materials, motivation to teachers, a good education policy, use of proper teaching and learning methods, community participation in schools and good childcare. A study of 12 school principals, 36 teachers and 120 students of igembe’s government secondary schools was conducted and it was found that students’ academic performance were affected by child education, poverty, financial, material assistance and irregular income [3]. It was also emphasized that the importance of providing textbooks and various facilities to the needy students in order to improve their academic performance. Four hypotheses were used to highlight the factors affecting student academic result in some private colleges in Rawalpindi and Islamabad and it was noted that a student’s academic performance largely depends on the communication, learning facilities, proper guidance and family stress [4]. Biswas et.al. [5] study on the issues that affect the academic results of the graduate students at the Islamic University of Kushtia in Bangladesh. After performing a study on Islamic University of Kushtia in Bangladesh some important variables were identified which affect their academic results likely results of S.S.C and H.S.C, parental academic qualification, family income, class attendance, study time without class period, student’s internet uses for non-academic purposes, political status, and mobile phone using for non-academic purposes [5].

At present, attention has increased in the prediction model where students’ results can be anticipated based on the studied attributes using different learning algorithms to find the validation of the model. Most of these learning algorithms are artificial intelligence (AI) techniques such as heuristics, machine learning, artificial neural networks, random forest, Bayesian classifiers, etc. There are much research on these instances. 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 [6]. Classifications algorithms such as the decision tree algorithm (J48), NB, Bayesian classifiers, k Nearest Neighbor (KNN) algorithm, and two rule learner’s algorithms (OneR and JRip) were used to estimate the student’s performance. The overall accuracy of the NB classifier was the greatest among other classifiers [7]. The support vector machine (SVM) classifier outperformed the KNN algorithm in

Electronic copy available at: https://ssrn.com/abstract=4523122
four different cases of a dataset while predicting student’s performance in the final examination [8]. SVM also excelled at Bayesian Knowledge Tracing (BKT) in predicting students’ problem-solving performance by showing approximately 29 percent improvement, compared to standard BKT method [9]. SVM casts as a derivative of the statistical learning theory which is different from NNs (Neural Networks) and performs efficiently for classification, regression, distribution estimation. SVM has high generalization ability with polynomial input, and it can easily classify the nonlinearly mapped data compared to other classical learning algorithms [10]. 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 [11]. SVM and RF were used to predict the first-year student’s performance [12][13][14]. First-year bachelor student’s data on a particular course over 8 years was analyzed with the help of DT, NB, and rule-based classifiers. It was posited (found) that a rule-based classifier gave the best prediction results compare to the other two classifiers [15]. 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 [16]. 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 [17][18][19][20] [21]. 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 [22]. The two-level classification algorithm was implemented to calculate the expected graduation time where the passed or failed students were differentiated in the first level and three different periods of graduations were classified in the next level [23]. A literature review of various techniques applied to predict student performance was studied. Techniques were grouped into three clusters such as fuzzy logic, data mining techniques, and hybrid [24]. 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 [25]. Sequential Minimal Optimization (SMO) and NB were combined into a hybrid model, SM Naive Bayes (SMNB) model was developed to predict the student performance. This hybrid model was surpassed by other individual models in accuracy [26]. 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 [27]. RF classifier outperformed C4.5 and NB algorithms in the prediction of the student’s success in the undergraduate program [28]. The linear random forest (LRF) has 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 [29]. Fuzzy ANFIS was used to convert multiple decision maker’s ratings into a final rating based on 78 fuzzy logics [30].

There is no existing research to our knowledge where the quantification of social and economic factors has taken into account for the prediction of students’ performances. This work was based on addressing this existing research gap. In this work, we have used social and economic theories to identify the factors responsible for students’ performances and introduced ANFIS model for merging the factors. Then we have used two learning algorithms, SVM and RF, for the prediction of students’ performances and compare their accuracies as the comparisons of these two methods were also not performed in this kind of educational research. So, our research can give new directions to the existing ones.

3. Methodology:

This research focuses on the identification and analysis of different types of socio-economic factors including psychological, personal, teaching impact, university facilities, learning environment etc. which affect students’ results and prediction of student’s performance based on these factors using machine learning algorithms. Basically, this project was performed in two stages. In the first stage, different factors responsible for poor academic performances were identified based on different social and economic theory/ sociological and economical frameworks [31-34]. Students’ performances were anticipated in the later stage based on those studied factors to identify the accuracy of the model. The whole process combining these two stages can be divided into the following steps.

1. Identify the factors from socio-economic point of view and prepare the survey questionnaire for performing the first survey on students.
2. Find the significant factors using ANOVA analysis from the survey results obtained from students.
3. Perform a second survey with only the significant attributes of experts.
4. Merging of factor rating of students’ and expert’s opinion by using fuzzy ANFIS.
5. Classification and prediction of student’s performance based on SVM and RF classifier.
6. Comparison of results and Probable solutions to the problem

In the 1st step, different factors affecting students’ performances were identified and analyzed using different social and economic theory i.e., Material Deprivation Theory, Cultural Deprivation Theory, Bean’s Model of Dropout Syndrome, Elger’s Theory of Performance etc. Then a survey questionnaire was prepared with those factors and stratified random sampling technique has been used for conducting the survey on students. After obtaining students’ responses from the survey, ANOVA analysis was done in the 2nd step to identify the fundamental factors. Another survey was conducted on experts with those selected factors. After that, fuzzy ANFIS model was developed to merge multi response ratings into a single rating point by consisting fuzzy logic. SVM and RF were applied to classify the student performance. Finally, the results were analyzed with relative advantages and disadvantages and probable suggestions were provided against those significant factors.

3.1 Flowchart:

At first, socio-economic factors responsible for students’ performance were identified utilizing theoretical frameworks of sociology. Then, surveying was done on students with those factors. After obtaining students’ feedback, significant factors were identified using ANOVA analysis. A second survey was conducted on experts utilizing these significant factors. After the completion of these surveys, ratings of each attribute marked by students were merged with the factor ratings given by the experts with the help of Fuzzy ANFIS analysis. Then, 80 percent of these merged data were used as training data for both SVM and RF classifier. Another 20 percent were used for testing. After that the accuracy of both the methods were checked with the actual results of the students.
3.2 Theoretical and Conceptual Framework

3.2.1 Material deprivation theory

Material deprivation refers to the inability to add the basic resources and services of life. Gibson and Asthana [31] discuss the impact of material deprivation on educational achievement and point
out the correlations between them. They also talked about the way poverty has a negative impact on educational performance. They have tried to highlight various issues in their theory such as higher levels of sickness in poorer homes, less able to afford hidden costs, tuition fees and loans, less access to pre-school facilities, more likely to have part time jobs [32].

3.2.2 Cultural deprivation theory

Cultural deprivation theory refers to the inferior norms, skills, values, knowledge of the lower social classes that hinders many from pursuing their good academic performance. It is often the case that working class parents and students are not too worried about their later educational life. Some of the reason behind the poor academic performance of the students such as lack of interest, less able to help their children with homework, restricted speech code, immediate gratification rather than deferred the underclass a higher percentage of single parent families [32]

3.2.3 Bean’s Model of dropout syndrome

The three factors mentioned in Bean’s model of dropout syndrome are academic factors, social psychological factors, and environmental factors. Among the academic factors he has discussed are pre-matriculation, academic performance, and academic issues. Student’s goals, utility alienation, faculty contact in psychological factors and how the issue of finance and opportunity to transfer in environmental factors affect dropouts [33].

3.2.4 Elger’s theory of performance

The theory of performance tries to identify how performance can be improved while explaining performance. This theory explains performance area, components of performance as well as primary domains that affect performance in a variety way [34]. Emphasis has been placed on quality increase, capacity increase, knowledge increase, skills increase to increase performance.

A gist of the above four theories is illustrated in Table 1 where it is seen that in each theories different types of factors having direct or indirect effect on students’ results were mentioned. These factors were divided into four categories: economic, psychological, cultural, institutional environment which are shown in Table 2.
Table 1: Different social theories of poor academic results

<table>
<thead>
<tr>
<th>Name of the Theory</th>
<th>Highlighted Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Material Deprivation Theory</td>
<td>➢ Higher level of poorer homes&lt;br&gt;➢ Less able to afford hidden costs&lt;br&gt;➢ Tuition fees and loans&lt;br&gt;➢ Less access to facilities&lt;br&gt;➢ More likely to have part time jobs</td>
</tr>
<tr>
<td>Cultural Deprivation Theory</td>
<td>➢ Lack of interest&lt;br&gt;➢ Restricted speech code&lt;br&gt;➢ Immediate gratification rather than deferred&lt;br&gt;➢ Single parent families</td>
</tr>
<tr>
<td>Bean’s Model of Dropout Syndromes</td>
<td>➢ Academic factors&lt;br&gt;➢ Social psychological factors&lt;br&gt;➢ Environmental factors</td>
</tr>
<tr>
<td>Elger’s Theory of Performance</td>
<td>➢ Various level of performance&lt;br&gt;➢ Identify performance area&lt;br&gt;➢ Different primary domain&lt;br&gt;➢ Increase level of skill&lt;br&gt;➢ Increase level of knowledge</td>
</tr>
</tbody>
</table>

Table 2: Socio-economic Factors behind poor academic results

<table>
<thead>
<tr>
<th>Category</th>
<th>Highlighted Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Economic Factors</td>
<td>➢ Managing of academic costs</td>
</tr>
<tr>
<td>Psychological Factors</td>
<td>➢ Lack of parental support&lt;br&gt;➢ Lack of teachers’ support&lt;br&gt;➢ Lack of classmates’ support&lt;br&gt;➢ Conflict interest about career&lt;br&gt;➢ Fear of examination&lt;br&gt;➢ Negative effects on pre-marital relationship</td>
</tr>
<tr>
<td>Cultural Factors</td>
<td>➢ Imbalanced family pattern&lt;br&gt;➢ Difference opinion about ideology management&lt;br&gt;➢ Increased tendency to break social rules and regulation&lt;br&gt;➢ Abuse of social media&lt;br&gt;➢ Overuses of electronic devices</td>
</tr>
</tbody>
</table>
Institutional Environmental Factors

- Lack of campus facilities
- Lack of campus recreational activities
- Lack of quality education services
- Relationship pattern with university staff
- Political instability
- Tendency to be absent

3.3 ANOVA Analysis:

A common approach to figure out the similarity and dissimilarity among variables/samples is to perform the analysis of variance (ANOVA). It can compare these samples and depict how different these samples are. This technique of comparing variables based on their means is called ANOVA. ANOVA is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples.

Like any other kind of hypothesis in statistics, ANOVA works with two hypotheses, i.e., a Null hypothesis and an Alternate hypothesis. The Null hypothesis is valid when all the sample means are equal, that means all the samples are alike. Thus, they can be considered as a part of a larger set of the population. On the other hand, the alternate hypothesis is valid when at least one of the sample means is different from the rest of the sample means. In mathematical form, they can be represented as:

\[ H_0 : \mu_1 = \mu_2 = \cdots = \mu_L \quad \text{Null hypothesis} \]
\[ H_1 : \mu_1 \neq \mu_m \quad \text{Alternate hypothesis} \]

Where \( \mu_1 \) and \( \mu_m \) belong to any two sample means out of all the samples considered for the test. In other words, the null hypothesis states that there is no significant difference among the samples. Whereas the alternate hypothesis states that at least one of the sample means is different from another.

**Types of Tests**

There are two main types of tests in ANOVA: one-way and two-way. Two-way tests can be with or without replication.
• **One-way ANOVA between groups:** It is used when two groups are tested to see if there’s a difference between them.

• **Two-way ANOVA without replication:** It is used when there is one group and double testing is performed on that same group before and after one or more incidents. For example, if testing is done on one set of individuals before and after they take a medication to see if it works or not.

• **Two-way ANOVA with replication:** It is utilized when there are two groups, and the members of those groups are doing more than one thing. For example, two groups of patients from different hospitals trying two different therapies.

### 3.4 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 modern computer is based. A computer can only take precise inputs and give output as TRUE or FALSE which resembles with humans’ YES or NO. Inventor of fuzzy logic, Dr. Lotfi Zadeh first observed that unlike computers, human’s decision modeling are a range of possibilities between YES and NO, such as [35]-

<table>
<thead>
<tr>
<th>Certainly Yes</th>
<th>Cannot Say</th>
<th>Possibly Yes</th>
<th>Possibly No</th>
<th>Certainly No</th>
</tr>
</thead>
</table>

The fuzzy logic works on these possible levels of inputs to give a specific output.

Architecture: Its architecture shown in Figure 2 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 to convert inputs i.e., crisp numbers into fuzzy sets. Crisp inputs are basically the exact inputs measured by sensors and passed into the control system for processing, such as temperature, pressure, rpm’s, etc.
• Inference Engine: It determines the matching degree of the current fuzzy input with respect to each rule and stimulates human reasoning based on 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 inference engine into a crisp value which is the ultimate output.

![Fuzzy Logic structure diagram](image)

**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 fuzzifiers:

- Singleton fuzzifier
- Gaussian fuzzifier
- Trapezoidal or triangular fuzzifier

In varied problems defined by fuzzy logic, experts are needed to form the imposed fuzzy rules. Sometimes, a learning model like artificial neural network (ANN) method becomes necessary as experts don’t always share their knowledge behind imposing rules. That’s why ANFIS structure was developed by incorporating ANN into fuzzy method. This model overcomes one of the primary downsides of fuzzy systems, which is the difficulty in constructing fuzzy rules, while maximizing the benefit of fuzzy systems, which is the interpretability of relationships between a specified output and a collection of inputs. In the case of neural networks, the opposite is true. This ANFIS method solves the problem of intelligence techniques not being able to describe the fundamental behavior of learning networks.

In this paper, ANFIS has been used as we were able to identify the logic for imposing rules with the active participation of the experts. The ANFIS model has been established where a set of inputs
and outputs were given then rules were developed using the learning techniques of this model. At first, membership functions of each input and output were specified by selecting the type of fuzzifiers and adjusting its range. Then the rules were developed.

3.5 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 two-dimensional space, it draws a hyperplane separating two 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 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 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 there by a binary classifier.

3.5.1 Binary SVM:
For separating two classes, the optimal separating hyperplane (OSH) maximizes the distances between the two nearest data points in order to classify two classes. Assume that dataset for training is represented by a set, \( j = \{ (x_i, y_i) \}_{i=1}^l \), here \( (x_i, y_i) \in \mathbb{R}^{n+1} \), \( l \) 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 [30].

\[
\min \frac{1}{2} \| w \|^2 \\
\text{s.t.} \ y_i (w^T \varphi(x_i) + b) \geq 1 \quad i = 1, 2, \ldots, l
\]

3.5.2 Multiclass SVM:
For implementing SVMs for more than two classes, two methods are used. One is one against all (OAA) and another one is one against one (OAO). In 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 one single class, that’s why points in that class will give positive response and points belonging to other classes will give negative values on that classifier. If points belong to class \( A_1 \), only the SVM classifier which is trained to separate this class will give positive response. In case
of OAO, for n class problems \(\frac{n(n-1)}{2}\) SVM classifiers are constructed and each of them is trained to separate one class from another. If an unknown point is to be classified, each SVMs vote for different class and maximum vote of one class is the main result.

3.5.3 Framework for SVM Model:
Selection of feature is very important for classification problem. In this work, factors which had been identified for students’ different levels of performance by the survey were the distinguishing features for this multiclass classification problem. That’s why these factors were used as features for separating and predicting different classes. 80 percent of the merged ratings were used as training data and students were divided into seven classes based on CGPA. After training the model, the remaining 20 percent data were used for testing. Then, the prediction of classes of tested data were checked with the accurate result of the students. The step-by-step breakdown of the model is given below-

- Identification of features for data points’ separation by classes
- Divide the data points into the classes accordingly
- Training of SVM model with the help of training data and its class labels
- Testing of data into the trained model
- Comparison of classes’ predictions obtained from the SVM model with the actual results.

3.6 Prediction using Random Forest Classifier:
Random Forest (RF) is another supervised machine learning algorithm which is also used for both classification and prediction; however, it is mainly applied for classification applications. Forest means trees and the more the trees the more robust the forest is. In random forest classification method, this model creates different decision trees based on data samples and when new data points are inserted for its class prediction, each decision tree gives one prediction and finally best solution is selected by voting. For an input vector \((x)\), each decision tree will give a vote. Then, \(C^B_{rf} = \text{majorityvote}\{C_b(x)\}\biggr|_1^B\) where \(C_b(x)\) is the prediction of class on \(b^{th}\) random forest tree and \(C^B_{rf}\) is the final prediction selecting the majority vote. The main concept behind this model is simple but a powerful one. It is an ensemble method because a large number of uncorrelated models (trees) working as a community definitely outperforms the working of a single model (tree). The reason for this wonderful effect is that the models protect each other from their errors. Choice of attribute
selection and pruning methods are necessary for the design of decision trees. There are many attribute selection methods but the most frequently used attribute selection measures in decision tree induction are Information Gain Ration Criterion [33] and Gini Index [34]. RFC uses the Gini Index method for its attributes’ selection which measures the impurity of an attribute with respect to its classes. For a given training set \( P \), selecting a sample case randomly and to predict its class as \( C_i \), the Gini index can be written as-

$$\sum \sum_{j \neq i} \left( \frac{f(C_i, T)}{|T|} \right) \left( \frac{f(C_j, T)}{|T|} \right)$$

Here, \( f(C_i, T)/|T| \) is the probability that a selected case belongs to class \( C_i \). For generating a prediction model, the RFC needs the definition and insertion of two parameters, the number of classification trees desired and number of predicting variables which 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 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.
3.6.1 Framework for RFC Model: The framework for RFC is represented by a flowchart in Figure 3.

Figure 3: Flowchart of RFC model
Input data with various features and an output attribute with different levels is split into two datasets: training dataset and testing dataset. Then bootstrap aggregating and attribute bagging are developed to form a randomly selected decision trees by minimizing the misclassification rate. Finally, the testing dataset is examined to predict the class.

4. Results and Discussions

4.1 Data Collection

Students’ academy results and performance are influenced by a number of important factors. Through this research, many factors have been identified, all of which have been discussed together in socio-economic factors. These characterizing variables of students’ performance are shown in Table 2. In this project, the initial survey questionnaires were prepared based on these factors. The survey was conducted on the students of four universities, Bangladesh University of Textiles (BUTEX), Bangladesh University of Engineering and Technology (BUET), Dhaka University (DU), Jagannath University (JNU) to identify the ratings on each factor which could be linked to students’ different levels of performance. Survey results of students’ response on each variable are discussed below.

4.1.1. Economic factors

Economic factors are one of the main factors behind good or bad academic results. Being financially supportive makes it possible to focus on studying smoothly. But in many cases, it is seen that due to the inability of the students’ family, they have to go through financial crisis which makes it difficult for them to manage their academic costs.

Managing Academic Costs

In the survey, students were asked how well they or their family can manage the necessary academic costs. From the survey responses shown in Figure 4, it is noticed that majority of DU students has answered that they can manage their academic costs excellently while most BUTEX and BUET respondents have answered that their condition in managing costs is average.
Peace of mind is necessary for human development by improving their learning and instructional process. Without this, various states like mental stress, depression and over expectation etc. arise and bring bad results for the students. Research has shown that students do not only have poor academic results due to their own psychological problems, but also their whole life is affected by this. These factors are:

**Lack of parental support**

According to the research data, maximum parents provide adequate support to their children to manage their education well but sadly, some parents can’t. Figure 5 shows the statistics of four universities’ students in this context which shows that DU respondents are getting their parents support warmly. In the case of parental support, the conditions of the respondents of JNU are ahead in moderate possibilities. Above all, with the help of the parents, the inaccessible path of a student's education becomes much easier, so the sincerity and cooperation of the parents needs to be further enhanced.
Lack of teachers’ support

A supportive relationship between teachers and students should exist in order to create a better learning atmosphere for the students. Figure 6 shows the opinion of the students on how much support they are getting from teachers. The distance between the teacher and the students is largely responsible for the poor academic results of the students.

![Teachers’ Support](https://ssrn.com/abstract=4523122)

Figure 6: Teachers’ Support

Lack of classmates’ support

Support of classmates as well as teachers is necessary for students as students spend most of the daytime with their classmates. Above all, with the help of classmates, it is possible for students to overcome many difficult problems in their lives. Research survey in Figure 7 has shown that classmates’ supports conditions are very high in BUTEX students, JNU and DU respondents think that they are also getting enough support from their classmates while classmate support for BUET students is in moderate condition.

![Classmates’ Support](https://ssrn.com/abstract=4523122)

Figure 7: Classmates’ Support
Conflict of interest about career:

Conflicting interest in one's career is one of the reasons why one cannot perform well in one's own department and subject. Because in maximum cases, they can’t choose their subjects or fields or distinctive career paths, rather these are imposed on them. Thus, conflict of interest in career arises. Figure 8 shows that the highest respondents of BUTEX strongly agree that they have conflict of interest about their career, while respondents of DU think they have very few problems with career conflict.

Fear of examination

Fear of examination is another reason behind poor test results, especially in semester final exams. According to the research study shown in Figure 9 BUET and JNU respondents have said that their fear of examination is very high. On the other hand, BUTEX and DU respondent’s fear of examination are in moderate position.
Negative effects of pre-marital relationship

In today’s changing society, it is seen that many people are getting involved in pre-marital relationships. Figure 10 shows the statistics of respondents and what they think about the negative impacts of pre-marital relationships in their personal life. According to the study, respondents of the DU and JNU are agree that there are more negative effects on pre-marital relationship in their personal life on the other hand, respondents of the BUTEX and BUET are facing less negative impact on their personal life.

![Negative effects on pre-marital relationship](image)

Figure 10: Negative effects on pre-marital relationship

4.1.3 Cultural factors:

Cultural factors were identified based on many contexts like family, social media, and the use of additional electronic devices etc. Some issues like cultural adaptation and cultural lag affect students’ results indirectly and directly.

**Imbalanced family pattern:**

Imbalance in family patterns is increasing day by day in the present society and it also has a big impact on the socialization process of children. In this criterion, many respondents agree that this factor has significant effects in their life. This study data showed that about 14%, 9% and 7% of DU, JNU, BUTEX respondents strongly agreed with the negative impacts of this factor.

**Difference of opinion about ideology management**

When a child grows up in a family, in many cases they seem to be different from their parents’ ideology. In many cases, these ideological differences also exist compared with other family members. From the survey data portrayed in Figure 11, it is seen that that almost every university respondent had the highest neutral position in this criterion of different opinions about ideology management, but this variable was observed higher in JNU respondents. Students facing these
ideological issues in the family often become frustrated and stressed, which in turn has a negative impact on their academic results. In many cases these problems lead to suicidal tendencies among students.

Figure 11: Difference of opinion about ideology management

**Disobeying social rules and regulations**

A major obstacle to proper education is not being able to strictly maintain social norms or getting involved in various deviant or criminal activities. From Figure 12, it is observed that these types of deviant behaviors were less in BUET and more in DU. Followed by DU, JNU also had a high rate in disobeying social rules. If these types of behaviors are developed in a student, his/her studies will be greatly hampered. Thus, the performance of the student will be affected.

Figure 12: Disobeying social rules and regulations

**Abuse of social media**

According to the survey data, almost every university’s respondent had agreed about their involvement in social media and pointed out that a lot of their time was spent on maintaining social media. In many cases they spent four to five hours effortlessly on social media which later had a
negative impact on their study. According to the survey data, their usage of social media has increased many times over due to conducting online educational activities in the Corona Pandemic.

**Overuses of electronic devices**

The use of electronic devices such as mobile phones, computers, laptops, and tabs is increasing day by day among the students. Sometimes it is for learning purposes and in the majority of cases they overuse these electronic devices for entertainment causes. According to the survey data shown in Figure 13, procrastination of BUTEX respondents were very high due to overuses of electronic devices. Also, many of the respondents of BUET, DU and JNU respectively have agreed with the overuse of electronic devices. Also, the respondents of each university agree that they have a high impact on them. The interest in reading books, newspapers, journals is declining day by day due to the use of over electronic devices.

![Procrastination due to over uses of electronic devices](https://ssrn.com/abstract=4523122)

**4.1.4 Institutional Environment Factors**

Factors like lack of campus facilities, campus recreational activities and quality educational services, political instability etc. fall under this category which indirectly and in some cases directly affect students’ performances.

**Lack of Campus Facilities**

One of the main reasons behind the poor results of students is not providing proper campus facilities. Figure 14 reveals that the majority of the respondents of BUET and DU think that their campus facilities are in excellent position. The condition of BUTEX in this criterion seems average from the obtained data.
Lack of campus recreational activities

Various types of recreational activities need to be established to increase students' interest in learning. These activities provide opportunities for students to nurture themselves, improve their self-esteem and reduce anxiety and depression. Thus, their motivation towards learning also increases. From Figure 15, it is seen that the highest number of respondents of DU has agreed about their excellent condition in recreational facilities. Followed by DU, BUET and JNU comes next. The position of BUTEX is lower compared to the other three universities in this context.

Lack of quality education services

From Figure 16, it can be observed that, majority of BUET respondents think that they are in an excellent position in terms of quality education. The number of respondents of DU and JNU who think their excellency in terms of quality education are also high. While the majority of BUTEX respondents think that their quality of education services is average.
Relationship pattern with university staffs

In order for academic life to run smoothly, good relations should be maintained with the staff of the university. From Figure 17, it is noted that the majority of respondents of all the universities think there are in good relationship with their university staffs. This number is higher than BUET and lowest in DU.

Political instability

Political instability often serves as a barrier to sound teaching. According to the survey data shown in Figure 18, 38 respondents of BUTEX out of 100 said political instability exists highly in the university which affects their studies and exams. The majority of students at other universities have also claimed the same thing. Only a few students have mentioned the opposite.
Along with political instability, the tendency of students to be absent from class is widely observed. In many cases, they are absent from class due to various political issues or involvement. According to survey data illustrated in Figure 19, majority of students from DU, JNU, BUTEX has opinionated that their tendency to be absent from the class is very low while this tendency is higher in case of BUET students compared with other three universities.

### 4.2 Identification of Significant Factors

After performing survey-1, significant factors were identified from the ratings of the students on each factor. Analysis of variance (ANOVA) was used for this purpose to check the correlation between different factors. For the reduction of factors and identification of significant factors, ANOVA had been used here. In Table 3, ANOVA tests’ result is shown where two factors, disobeying social rules and regulations, and abuse of social media 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 example is given in Table 4, where factors, fear of examination and lack of teachers' support, were compared with the same confidence interval. But result here shows the opposite. As, \( F > F_{\text{crit}} \) or \( p - value < 0.05 \) that means a null hypothesis is rejected and there is a difference between the factors.

Table 3: ANOVA table of two correlated factors

<table>
<thead>
<tr>
<th>Factors</th>
<th>Sum of Squares, SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p-value</th>
<th>F-crit</th>
<th>Hypothesis</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disobeying social rules and regulations Between groups</td>
<td>24.49485</td>
<td>1</td>
<td>24.49</td>
<td>0.0511</td>
<td>0.821</td>
<td>3.849</td>
<td>Accepted</td>
<td>Factors Correlated</td>
</tr>
<tr>
<td>Abuse of social media Within groups</td>
<td>557811.5</td>
<td>1164</td>
<td>479.21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>557836</td>
<td>1165</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4: ANOVA table of two uncorrelated factors

<table>
<thead>
<tr>
<th>Factors</th>
<th>Sum of Squares, SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p-value</th>
<th>F-crit</th>
<th>Hypothesis</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fear of examination Between groups</td>
<td>8215</td>
<td>1</td>
<td>8215.28</td>
<td>17.20</td>
<td>3.6E-05</td>
<td>3.8494</td>
<td>Rejected</td>
<td>Factors uncorrelated</td>
</tr>
<tr>
<td>Lack of teachers’ support Within groups</td>
<td>555889</td>
<td>1164</td>
<td>477.56</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>564104</td>
<td>1165</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 performance) which were used for the whole classification problem. 11 significant factors were found which are shown in Table 5.
Table 5: List of 11 significant factors

<table>
<thead>
<tr>
<th>Factors’ no.</th>
<th>Factors’ name</th>
<th>Factors’ no.</th>
<th>Factors’ name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Lack of teachers’ support</td>
<td>7</td>
<td>Political Instability</td>
</tr>
<tr>
<td>2</td>
<td>Managing of academic costs</td>
<td>8</td>
<td>Lack of quality education services</td>
</tr>
<tr>
<td>3</td>
<td>Overuse of electronic devices</td>
<td>9</td>
<td>Lack of classmates’ support</td>
</tr>
<tr>
<td>4</td>
<td>Lack of campus facilities</td>
<td>10</td>
<td>Tendency to be absent from class</td>
</tr>
<tr>
<td>5</td>
<td>Imbalanced family pattern</td>
<td>11</td>
<td>Fear of Examination</td>
</tr>
<tr>
<td>6</td>
<td>Conflict of interest about career</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4.3 Merging of factors’ ratings by Fuzzy analysis

After finding 11 significant factors using ANOVA analysis, another survey was conducted on two experts 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 20 shows the demonstrations of these rules on inputs and how they affect the outputs.

Figure 20: Demonstration of rules on fuzzy ANFIS

How to factor rating is affected by any two input ratings are shown in Figure 21. After combining all these outputs, finally, the merged ratings were identified.

Electronic copy available at: https://ssrn.com/abstract=4523122
Figure 21: Factor rating with respect to expert-1 vs expert-2 and student rating vs expert-2

4.4 Prediction of classes’ using SVM classifier

After merging all the datasets to a single dataset using fuzzy analysis, 80 percent of these data were used as a training dataset for two machine learning algorithms (SVM and RF classifier). In SVM, significant factors were used as distinguishing features for classes’ separation. Students were separated into seven classes based on their GPA in the previous term. Table 6 shows how students were separated into classes.

<table>
<thead>
<tr>
<th>CGPA range</th>
<th>2.26-2.50</th>
<th>2.51-2.75</th>
<th>2.76-3.00</th>
<th>3.01-3.25</th>
<th>3.26-3.50</th>
<th>3.51-3.75</th>
<th>3.75-4.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class no.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>

The tuning of the SVM model was also done to increase the accuracy of the classes’ prediction. Figure 22 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 22: Tuning of the model by changing cost, gamma, and epsilon functions’ values

The value of cost, gamma, and epsilon was varied from 1 to 6, 0 to 1, and 0 to 1 respectively to consist of 600 different models. Among those models, the best model is represented in Table 7 based on performance.

Table 7: Best predicting models after tuning of models

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Kernel</th>
<th>Cost</th>
<th>Gamma</th>
<th>Epsilon</th>
<th>Ntree</th>
<th>m_{try}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best SVM model</td>
<td>Radial</td>
<td>1</td>
<td>0.1</td>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Best RFC model</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>150</td>
<td>3</td>
</tr>
</tbody>
</table>

4.5 Prediction of classes’ using RF classifier:

Random forest was created using a training dataset where 500 trees were built. In Figure 23, one of the random trees is presented.
After that error rate was calculated which is shown in the figure. We can see that above 150 trees, the error rate became constant. That is why 150 trees were selected to build the RF model. The error rate for 150 trees is also given in Figure 24.

Another important parameter of RF classifier is the number of attributes used in attribute bagging process which is $m_{try}$. Analysing the OEB error to find out the suitable $m_{try}$. The relationship
between OEB error and $m_{try}$ is depicted in Figure 25. The lowest OEB error is found when the value of $m_{try}$ is less than 3.

![Figure 25: OEB error vs $m_{try}$ graph](image)

The partial dependency of individual factors on different classes was also investigated. For example, the partial dependency of factor-1 for two distinct classes is shown in Figure 26. 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 26: Partial dependency on factor-1](image)
4.6 Results

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 can not predict class 6 accurately.

Table 8: Confusion matrix for the SVM model

<table>
<thead>
<tr>
<th>Actual Class</th>
<th>Predicted Class by SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 0 0 0 0 0 0</td>
</tr>
<tr>
<td>2</td>
<td>0 6 0 0 0 0</td>
</tr>
<tr>
<td>3</td>
<td>0 15 1 0 0 0</td>
</tr>
<tr>
<td>4</td>
<td>0 1 1 1 0 0</td>
</tr>
<tr>
<td>5</td>
<td>0 2 2 1 0 0</td>
</tr>
<tr>
<td>6</td>
<td>0 0 0 0 0 0</td>
</tr>
<tr>
<td>7</td>
<td>0 0 0 0 0 0 1</td>
</tr>
</tbody>
</table>

Table 9: Confusion matrix for the RFC model

<table>
<thead>
<tr>
<th>Actual Class</th>
<th>Predicted Class by RFC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 0 0 0 0 0 0</td>
</tr>
<tr>
<td>2</td>
<td>0 6 0 0 0 0</td>
</tr>
<tr>
<td>3</td>
<td>0 16 0 0 0 1</td>
</tr>
<tr>
<td>4</td>
<td>0 0 4 0 0 0</td>
</tr>
<tr>
<td>5</td>
<td>0 0 0 3 0 0</td>
</tr>
<tr>
<td>6</td>
<td>0 0 0 0 0 0</td>
</tr>
<tr>
<td>7</td>
<td>0 0 0 0 0 0 1</td>
</tr>
</tbody>
</table>

The important factors were also analyzed using RFC which is depicted in Figure 27. We can see the top 5 important variables. The 1st graph tests how worse the model performs without each variable.

Figure 27: Top five factors that affect student performance most
Factor-11 (Fear of examination) is the most important factor that is responsible for better prediction of student performance. In the 2nd graph, it measures how pure the nodes are at the end of the tree without each variable. For factor-11, 6, 3, 1, & 10 mean decrease 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.

4.7 Recommendations

Eleven significant factors responsible for students’ academic performances were identified using ANOVA analysis. After that, five crucial factors (Factor-11, 6, 3, 1, 10) out of 11 were determined using RF. To tackle the impacts produced by these factors on students’ results, some probable solutions for students, family members, institutions and academic members are proposed here.

Suggestions for students: Students should

- Ask questions to the teachers if they do not understand the content of the lesson as well as keep notes of the topics taught in the class.
- Spontaneously participate in academic activities (class) as well as increase their vocational activities.
- Conduct various competitive education activities to increase their motivation level in educational activities and increase their participation.
- Try to make the right decision without getting frustrated with one’s career development.
- Overcome the fear of exams, by increasing their interest in regular studies so that they do not have to take too much pressure in one or two days before exams.
- Avoid excessive use of mobile phones, computers, laptops, and games.
- Restrain from overindulgence in social media.

Suggestions for family members: Family members should

- Try to properly meet the basic academic needs of the child.
- Maintain a good relationship with the child and listen to them so that they never feel alone.
- Mutually increase their discipline and respect for each other so that the norms, ideas, values, personality of the child are properly developed through family socialization.
- Not impose ambitions on children about their future and do not provide additional discipline for their failures.
• Increase family trust, fraternity, confidence, attachment, involvement, commitment so that family instability, parental antiracism does not affect on the children’s academic performance.

**Suggestions for institutions and academic members:** Institutions and academic members should

• Create the right learning environment for students and provide solutions by identifying the issues that hinder academic performance.
• Increase learning activities in classroom by bridging the gap between students and teachers.
• Implement various activities such as infrastructural development, curriculum activities, seminar, training, workshop in a timely manner.
• Abolish criminal activities, deviant behaviors, harassment, political abuse on campus in order for students to study properly.

6. Conclusion

In this work, we propose a hybrid model for predicting four 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 has 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 fear of examination (factor-11) is the most important factor that has a great impact on students’ success. Five most crucial factors were also detected including Factor- 11. So, early detection of these factors’ ratings could yield valuable insights for a better educational environment and assistance to students’ performance. Based on these factors, probable suggestions are also provided.

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. Not only new attributes but also more experts can be added to get more insight into factor ratings. Besides, deep learning algorithms like Convolutional Neural Network (CNN), Recurrent Neural Networks (RNNs) can be implemented for comparative analysis. 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.


Funding

This work was supported by the Bangladesh University of Textiles (BUTEX) Research Fund (FY: 2020-2021)

References


