Early Grade Prediction Using Profile Data

Sumaiya Iqbal, Mahjabin Muntaha, Jerin Ishrat Natasha, and Dewan Sakib

***Abstract*—Universities are reputable institutions for higher education and therefore it is crucial that the students have satisfactory grades. Quite often it is seen that during the first few semesters many students dropout from the universities or have to struggle in order to complete the courses. One way to address the issue is early grade prediction using Machine Learning techniques, for the courses taken by the students so that the students in need can be provided special assistance by the instructors. Machine Learning Algorithms such as Linear Regression, Decision Tree Regression, Gaussian Naïve Bayes, Decision Tree Classifier have been applied on the data set to predict students' results and to compare their accuracy. The evaluated profile data have been collected from the students of 10th semester or above of the Computer Science department, BRAC University, Dhaka, Bangladesh. The Decision Tree Classifier technique has been found to perform the best in predicting the grade, closely followed by Decision Tree Regression and Linear Regression has performed the worst.**

***Index Terms*—Machine learning algorithms, linear regression, decision tree regression, Gaussian Naïve Bayes, Decision tree classifier, feature importance, Chi-Square.**

I. Introduction

A student may face many challenges while pursuing higher education. This includes poor academic performance where their incapability to cope up may influence them to withdraw the course [1]. As a result, dropout rates tend to go higher as well. So, it is important to identify the students at risk before it is too late. Such event call for the need of a dedicated support system for such students which can motivate them to strive for better results. This procedure will enable the instructors to predict students' performances and address the underlying learning difficulties which will be beneficial for both the students and the institution [2]. As university is a place of experimental learning, the traditional method of studying does not work. Most of the time students fail to identify the correct techniques for studying. Following the right strategy will certainly help them to score good marks. But this learning strategy is not the only factor that affects their marks in exams. Previously, a lot of study had been done to identify the factors that have the most impact on students' learning. The results of every research are quite different from each other, since different data had been used. Some of the data used include scholarly, family, medical, financial information of students and so on. Some works show that communication [3], students' attendance [4] and other factors have a significant connection to a student's academic success. It is also worth noticing that the factors influencing the students' education are not the same for everyone. A study [5] indicated that if properly directed by parents and teachers, a student will have a successful academic career. This paper is therefore focused on evaluating the factors influencing the academic performance of the students.

Machine learning models work on the provided data and they are not based on assumptions about the problem. This makes machine learning models more effective for predictive performance [6]. Machine learning techniques can help in forecasting the performance of the students so that appropriate measures can be taken soon enough for the students at risk. Our key emphasis is on comparing machine learning approaches and feature engineering strategies in terms of how much they improve the efficiency of predictions. We have conducted our research and collected the profile data of students with the help of a questionnaire which was answered by students of 10th semester or above from CSE/CS department of BRAC University, Dhaka, Bangladesh. One of the principal steps when implementing the machine learning techniques is the selection of the machine learning algorithms [7]. In this paper, a set of attributes such as medium of study, CGPA after 1st and 9th semester, their chronic medical conditions (if any) and a few other factors have been considered. In the next phase, the data collected were turned into discrete variables. Then normalization was done to corresponding data of the features. It observes which characteristics are more related to the improvement or failure of academic performance of the students. This thesis focuses on supervised learning, where the training dataset is taken as input. At last, the Machine Learning algorithm creates a model, which outputs which students are at risk by classifying them into 3 distinct categories. It also tries to predict the exact CGPA with the regression and classification algorithms. Four algorithms like Linear Regressions, Decision Tree Regression, Gaussian Naïve Bayes and Decision Tree Classifier have been used in total. The result of our model gave varying accuracy for the different algorithms used. Finally, Chi-Square based feature selection method and Pearson Correlation Coefficient was used to rank the top ten features for the Classifier and Regression algorithms respectively.

The later sections of our paper are organized in the following way. In Section II, we talked about related work in this field. Section III elaborates on the various machine learning techniques that have been used to predict the CGPA of the students. Section IV, describes the proposed methodology of our research paper. The analysis of our results has been shown in Section V. At last, in Section VI concludes the paper.

Manuscript received July 15, 2021; revised March 30, 2022.

The authors were with the Department of Computer Science and Engineering, BRAC University, 66 Mohakhali, Dhaka, Bangladesh (e- mail: sumaiyaiqbal1998@gmail.com, mahjabinmuntaha96@gmail.com, zerin.ishrat444@gmail.com, dewansakib@gmail.com).

doi: 10.18178/ijmlc.2022.12.5.1100

II. Related Work

Predicting the grade of students is considered the most critical activity in the field of Educational Data Mining (EDM). Estimating grades of student in a course is a key to determining students at risks as early as possible. So, it is important to build a prediction model that correctly predicts whether a student is going to pass a course or fail. There have been numerous studies in the field of EDM to predict the grade of student for identifying at risk students.

Tekin [8] followed a data mining strategy that includes data planning, creation and evaluation of the prediction model. He used a sample dataset of 127 unique undergraduate student records. The dataset comprised of the scores of 49 courses. Neural Network (NN), Support Vector Machine (SVM), and Extreme Learning Machine (ELM) classification algorithms were used to predict the students' GPA. The root-mean squared (RMS), the coefficient of multiple determinations and the coefficient of variation (COV) were used in evaluating the methods. Among all three methods the SVM technique yielded more precise predictions rate of 97.98%.

Microsoft Azure is a lesser utilized machine learning technique to know the scholarly standard of the students. Anand *et al*. [9] built models using Microsoft Azure Machine Learning Studio as web services. A dataset was prepared in their analysis and compiled into CSV format to provide training to machine and testing it. The data set included attendance for the students, ailment, past academic scores, students' study hours and so on. The accuracy of the developed framework was 67%. Later students' information will be stored in the online web system and the data stored will help the teachers, enhance the quality of education for students.

The factorization procedures are useful in case of sparse data [10]. Thai-Nghe *et al*. [11] suggested the technique of Matrix Factorization (MF) to predict the performance of students. The dataset contained information reflected the contact of log files between students and computer aided tutoring systems. Root Mean Squared Error (RMSE) had been used for assessment. First of all, they conducted their research by using "Students-Perform-Task" and secondly, using "Student-Applies-Skill" as the key reference. RMSE of the proposed system showed improvements in each case. Authors study revealed that the MRMF can perform nicely compared to the other methods, taking into account the multiple relationships between entities.

When using finite sets, grade prediction using Linear Regression, Decision Trees, and Naïve Bayes classifier are effective. In a paper Pojon [12] made a distinction between machine learning approaches and feature engineering techniques about how much they improve the accuracy of prediction. The dataset had baseline exactness, based on which the model was built to compare whether the model can make successful prediction or not. Three models they developed using the framework of Linear Regression, Decision Tree and Naïve Bayes respectively. And the result showed that, the Naïve Bayes classification produced the best results, followed by Decision Tree and Linear Regression. and is very common. Indeed, in spite of the fact that the dataset is exceptionally small, Gaussian NB performs very well. Its basic assumptions are that each predictor leads to the outcome independently and in equal terms.

Al-Sudani *et al*. [13] used NN models on a sample of 470 students and the dataset comprised of 9 pre-processing attributes. Alongside the students' entry qualification, they utilized social and statistic factors. They contrasted the best-performed NN model (FF) with classifiers such as Decision Tree (DT), Support Vector Machine (SVM) and K- nearest neighbor (KNN) to determine the overall performance of NN model. The accuracy of the classifiers differed greatly with NN. The NN delivered 83.7% statistically accurate best results. The neural network model proposed by the authors had been tasted in a small dataset. On the other hand, a study [14] carried a similar study on a large dataset (810 data with 43 attributes) with the aim expanding the accuracy of the neural network.

Marbouti *et al*. [15] attempted to identify at-risk students early utilizing predictive modeling methods. The dataset included data over 1600 students. Factors used in the prediction methods such as marks of tests, mid-term and homework learning objectives scores. They mainly used 7 predictive modeling approaches such as Naïve Bayes, Support Vector Machine, K-Nearest Neighbors, Logistic Regression, Multilayer Perception, Decision Tree and Ensemble method. The most accurate ones for identifying at-risk students were Ensemble model with 85% accuracy after analyzing seven predictive modeling methods.

III. Data Analysis with Various Algorithms

A. Regression Algorithm

A series of applied mathematics processes are performed to estimate the relation between a subordinate variable and one or more free variable that are known as Regression Analysis. To analyze our dataset, we used two regression algorithms- Linear Regression and Decision Tree Regression.

1) Linear Regression: For modeling the relationship between a dependent variable and one or more independent factors, Linear Regression is used. It is called simple Linear Regression, because there is only one explanatory variable.

2) Regression with Decision Tree: Decision Tree Regression algorithm is used to build the regression in the arrangement of a tree structure. Regression with DT splits the dataset into as many small subsets as possible and the corresponding decision tree incrementally grows [16].

B. Classification Algorithm

The computer program learns from the dataset and uses the learning to classify new observation, it is called machine learning classification method. We used Gaussian Naïve Bayes and Decision Tree Classification for our research work.

3) Gaussian Naïve Bayes: Gaussian Naïve Bayes (Gaussian NB) is an excellent classifier that works well with all sorts of datasets and is very common. Indeed, in spite of the fact that the dataset is exceptionally small, Gaussian NB performs very well.

4) Decision Tree Classifier: Decision Tree Classification algorithm is used to construct classification into the

arrangement of a tree structure. Eventually, a decision tree is created which has decision nodes and leaf nodes.

C. Chi-Square

The algorithm Chi-Square is used to check the relations between categorical variables. The Chi-Square tests null hypothesis is that there is no relationship on the populaces' categorical factors, they are autonomous. The Chi-Square calculation is most widely used when using a crosstabulation to determine Independence test [17].

D. Pearson Correlation Coefficient

Coefficients of correlation are used to quantify how strong association between two variables is [18]. Pearson's Correlation Coefficient measure is a number between-1 and 1 which represents a propensity to have a linear association for two random phenomena [19].

IV. Proposed Methodology

A. Workflow

We consider the most important aspect of a study in our short period of research work is the workflow. Fig. 1. provides a guideline for our work cycle. We followed this method to function in a designed way, or to make the best of the result.

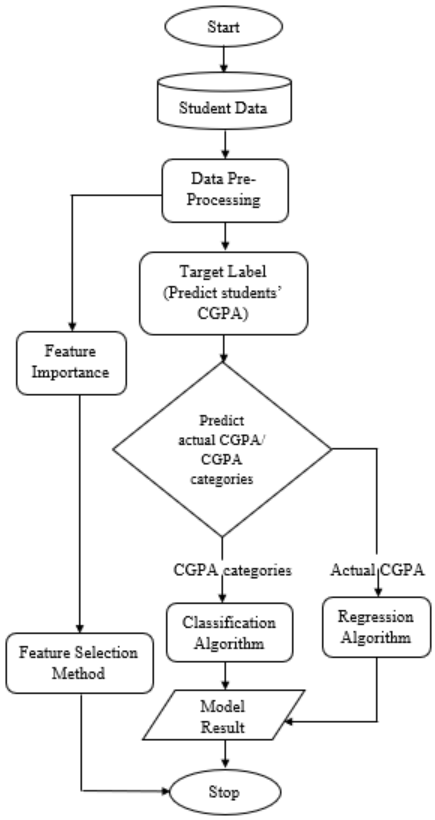

Fig. 1. Workflow of the proposed method for predicting students' grade.

B. Attribute and Dataset Description

The data collection used in this work is obtained from BRAC University's Department of Computer Science and Engineering. The dataset initially included 180 records of the students. The dataset has 17 attributes to it. The attributes can be divided into four categories that are gender, family related information, educational and personal information, and grades of students.

A short overview on the attributes used to build the model for predicting the grade of the students. Our research was conducted on the students who are currently in their 10th semester or above and doing their major in the CSE and CS area. First, it is normal to find that there is a disparity in the study habits of boys and girls in higher studies and that gender plays a role in deciding the students' academic achievement [20]. Also, medium of study and type of school in higher secondary level play a crucial role in students' academic success. Students who do not have enough grasp over the English language or those with a Bengali Medium background often struggle to understand class lectures conducted in English at University. Family size is often related to academic success, because parents may pay more attention to the academic performance of their children if there is a smaller number of family members. Higher secondary results also impact the student's success, as students with weak outcomes can have trouble understanding classes and end up with bad grades.

A student who stays away from university, which can adversely affect their academic performance as a traffic jam, is a major issue in our country. The family income of students has a significant effect on the output of the students because most students choose to attend the universities in Dhaka and the tuition and other expenses in Dhaka city are high. The student's number of hours studied also influences the standard of his academic results, as does his class attendance [20]. A significant factor is the outcome of 1st semester graduates. On the other hand, a student should have a secure CGPA after the 9th semester which will help us determine whether their outcome has degraded. Table I lists the attributes and their scope used to construct the dataset. Fig. 2. provides an overview of the CGPA of all students after the 9th semester and will give us an overview of the academic performance of the current students. We can see that the majority of students have a CGPA of at least 3, which falls between the medium and good spectrum according to our model.

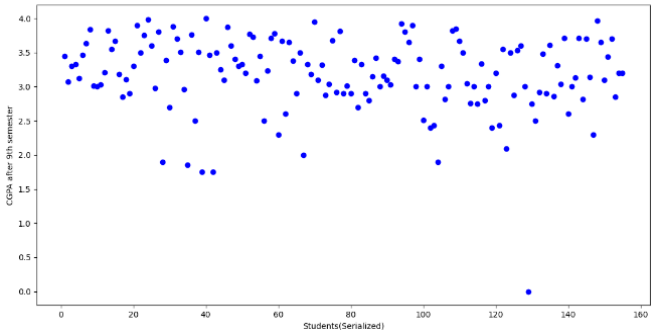

Fig. 2. An overview of all students CGPA after 9th semester from the dataset.

C. Dataset Pre-Processing

As we collected the data through survey via google form, there were some data which we could not use due to missing value or wrong information. To be precise, we had a total of 180 entries from which we could use 155 entries. We gave a unique id (Serial no.) to every data which were usable for further research. Then we started the pre-processing of our dataset. The first step of dataset pre-processing is to turn

data into discrete variables. We had to convert the domains of the attributes into discrete variables to make those work. Table II below shows which attributes domains converted into which discrete variables.

TABLE I: STUDENT ATTRIBUTES AND THEIR DOMAINS

Attribute Number	Attribute Name	Description	Domain
1	Semester	Students' semester (10th semester or above)	Yes, No
2	Major	Students' field of study	CSE, CS
3	Gender	Students' gender	Male, Female
4	Medium	Student's medium of study in higher secondary level	Bangla Medium, English Version, English Medium
5	SSC_GPA	Student's Grade point average of SSC examination	5.0, >=4.5, >=4.0, >=3.5, >=3.0
6	HSC_GPA	Student's Grade point average of HSC examination	5.0, >=4.5, >=4.0, >=3.5, >=3.0
7	Grade (O' Level)	Student's O' Level Grade in 5 subjects	A/A*= 5.0, B= 4.0, C= 3.0, D= 2.0
8	Grade (A' Level)	Student's A' Level Grade in 2 subjects	A/A*= 5.0, B= 4.0, C= 3.0, D= 2.0
9	F_size	Student's family size	Numerical value
10	C_time	Number of minutes student needs to come to university	Less than 15 minutes, 15 minutes<C_time<=30 minutes, 30 minutes<C_time<=60 minutes, 60 minutes<C_time<=120 minutes, More than 120 minutes
11	F_income	Students' family income	Less than 30000 Taka, 30000 Taka<F_income<=60000Taka, 60000 Taka<F_income<=100000Taka, 100000Taka<F_income<=200000Taka, More than 200000 Taka
12	M_condition	Whether student has any chronic medical condition	Yes, No
13	S_time	Numbers of hour student studies each day	Less than 0.5-hour, 0.5-hour<S_time<=1.0-hour, 1.0-hour<S_time<=2.0-hour, More than 2.0-hour
14	Attendance	Student's attendance percentage in 1st semester	Less than 70%, 75%, 80%, 85%, 90%
15	School	Students' type of school at higher secondary level	Public, Private
16	Result_1	Student's CGPA after 1st semester	Numerical Value [0.0<=Result_1<=4.0]
17	Result_2	Student's CGPA after 9th semester	Numerical Value [0.0<=Result_2<=4.0]

TABLE II: TABLE OF CONVERTING ATTRIBUTES DOMAINS INTO DISCRETE VARIABLES

Attribute Name	Domain	Discrete Variables
Gender	Male Female	0 1
Medium	Bangla Medium English Version English Medium	0 1 2
C_time	Less than 15 minutes, 15 minutes<C_time<=30 minutes 30 minutes<C_time<=60 minutes 60 minutes<C_time<=120 minutes More than 120 minutes	0 1 2 3 4
F_income	Less than 30000 Taka 30000 Taka<F_income<=60000 Taka 60000 Taka<F_income<=100000 Taka 100000Taka<F_income<=200000Taka More than 200000 Taka	0 1 2 3 4
M_condition	No Yes	0 1
S_time	Less than 0.5-hour 0.5-hour<S_time<=1.0-hour 1.0-hour<S_time<=2.0-hour More than 2.0-hour	0 1 2 3
School	Public Private	0 1

The next step is to normalize the data. In Machine Learning, normalization is used for preparing data. Normalization adjusts the numeric column value to a rising scale within a dataset. We normalized some of the attributes of our dataset, the attributes are shown below:

- SSC_GPA,
- Grade (O' Level),
- HSC_GPA,
- Grade (A' Level),
- Attendance,
- Result_1 and
- Result_2.

The target of this study is to predict students' grades. After calculating students' grades, each student will be categorized into self-labeled category as "Bad", "Medium" and "Good" based on their calculated CGPA. Table III shows the category of student based on their CGPA.

TABLE III: STUDENTS' CATEGORY, BASED ON THEIR CGPA

Category	Condition (CGPA based)
Bad	CGPA< 3
Medium	CGPA< 3.5
Good	3.5<=CGPA<=4

In Fig. 3. we showed visually the total number of students in the categories listed. According to our analysis, the minority of the students fall into the Bad category, while the majority fall into the Medium category.

The categorized target label has already shown in the

Table III. So, there are two target labels we need to predict. They are:

1) *Predict- the Actual CGPA and*
2) *Predict_discrete- CGPA Categories.*

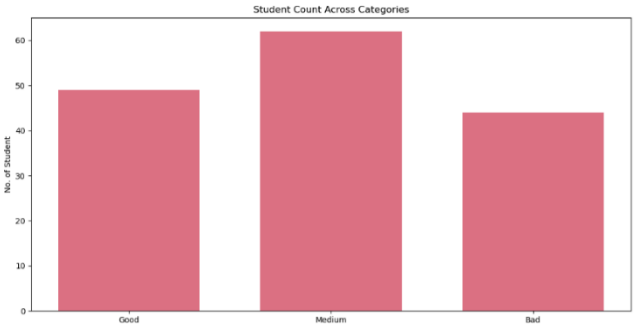

Fig. 3. Student count across categories, based on their CGPA.

D. Model Implementation

Machine learning is the method of learning from instances or more broadly speaking a set of rules to construct a classifier that can be used to generalize from new instances. Creating a classifier is a two-step operation. The classifier model is built in the first step using a given set of data. This move is called training. The second step called testing decides the correctness of the classification rules defined in the previous step [20]. If the classifier's accuracy is above an appropriate limit [21] then the classifier model built in the first stage can be used to classify new data records.

Regression and Classification methods can be categorized into supervised learning techniques. In this study: i) Linear Regression and Decision Tree regression analysis algorithms on the target label "Predict" and ii) Gaussian Naïve Bayes and Decision Tree classification analysis algorithms on the target label "Predict_discrete" are used to achieve better performance and more precise structural results that could be used to evaluate target labels based on pre-determined condition. In addition, two feature selection approaches (Chi-Square and Pearson Correlation Coefficient) have been used on the features to classify essential features that have the most effect on research and outcome of a student.

E. Linear Regression

Linear regression is one of the easiest models to apply on dataset in machine learning. Linear regression is a type of straightforward regression analysis where the number of free variables is one and a linear relationship exists between the independent(x) variable and the dependent(y) [22]. A line can be plotted based on the given data points, which models the best points. The line can be modelled according to the linear Equation (1) [22] below.

$$y = a_0 + a_1 * x \quad (1)$$

The linear regression algorithm has the concept of determining the best values for a_0 and a_1. Scatter plot for Linear Regression for our dataset is represented in Fig. 4.

In Fig. 4. the diagram depicts a weaker link between X and Y. Due to the weaker relationship between X and Y, the points on the graph are more scattered around the trend line. The amount of scattering also shows the model's accuracy for Linear Regression. The more the scattering, the less accurate the model is.

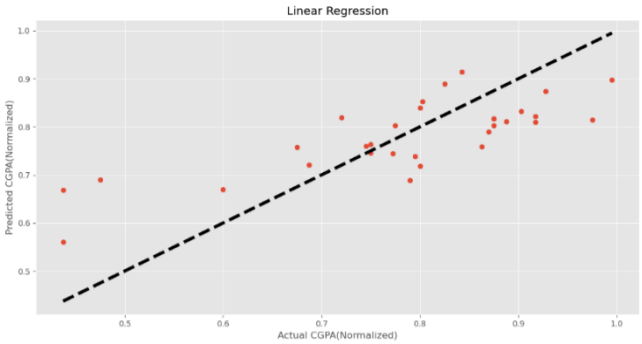

Fig. 4. Scatter plot for Linear Regression on student grade prediction using profile data.

Fig. 5 shows the Bar Chart for Linear Regression algorithm where horizontal axis shows the Student (Serial no.) and vertical axis shows student's CGPA after 9th semester. Here, the Actual CGPA is depicted with a blue bar and the Predicted CGPA is depicted with a green bar.

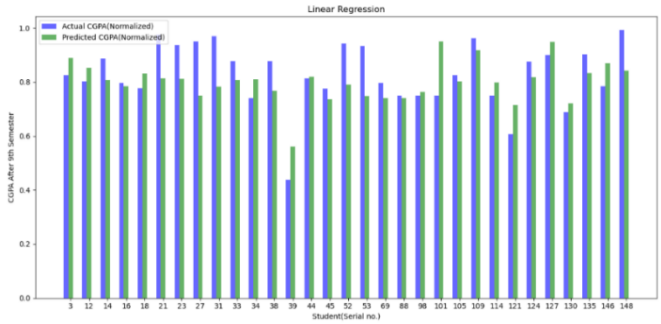

Fig. 5. Bar Chart of Linear Regression for student grade prediction.

F. Decision Tree Regression

Decision Tree is a supervised machine learning approach. In Decision Tree Regression the target variable will take continuous values usually real numbers. Entropy and Information Gain are used to construct a Decision Tree regressor [23], which can be calculated using the Equations (2) and (3).

$$\text{Entropy}(T, X) = \sum_{c \in X} P(c)E(c) \quad (2)$$

$$\text{InformationGain}(T, X) = Entropy(T) - Entropy(T, X) \quad (3)$$

In the above equations, Entropy (T, X) = conditional entropy of T given variable X.

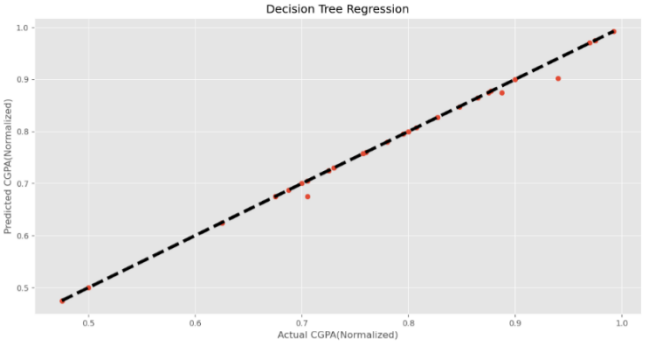

Fig. 6. Scatter plot for decision tree regression.

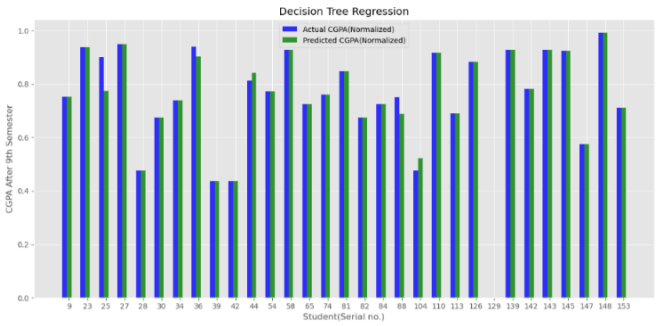

Fig. 7. Bar chart of decision tree regression.

In Fig. 6. we showed a Scatter plot for Decision Tree Regression. Where the x-axis was labelled with the Actual CGPA and y-axis shows Predicted CGPA. As most of the plotted data points lie above the regression line, there is a positive association between X and Y.

Fig. 7. shows the Bar Chart for Decision Tree Regression algorithm. Here, we compared Actual CGPA to Predicted CGPA. In this case Actual and Predicted CGPAs are nearly comparable.

G. Gaussian Naïve Bayes

A Gaussian Naïve Bayes classifier is a type of predictive machine learning algorithm. This algorithm is based on Bayes theorem. Gaussian Naïve Bayes algorithm is really fast as compared to other machine learning classifiers. This is why the application of this algorithm to large datasets is simpler. Important assumptions of this algorithm are the independence of the features and their equal contribution to performance. Equation (4) [24] indicates the probability of an occurrence considering the probability of another already existing occurrence.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \tag{4}$$

where, A is the event we will find the probability of and B is the event that has already taken place [24]. Scatter plot for Gaussian Naïve Bayes classifier for our dataset is represented in Fig. 8.

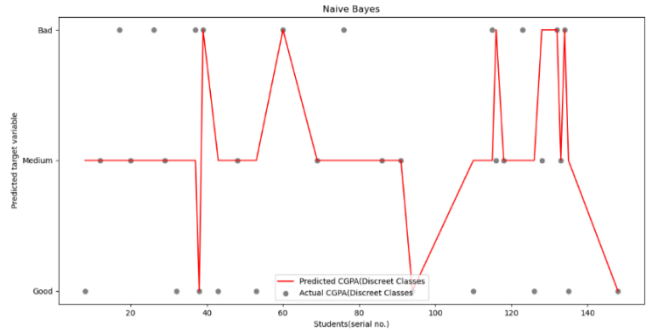

Fig. 8. Scatter plot for Gaussian Naïve Bayes.

In Fig. 8. the x-axis was labelled with the Students (serial no) and y-axis shows Predicted Target Variable. From the diagram dots that are not connected to the line are the misclassified instances from Gaussian Naïve Bayes algorithm. The Fig. 9. shows the Confusion Matrix for Gaussian Naïve Bayes Classifier algorithm. In the confusion matrix, we can see the amount of data for Bad is relatively high. The result is not best, but acceptable.

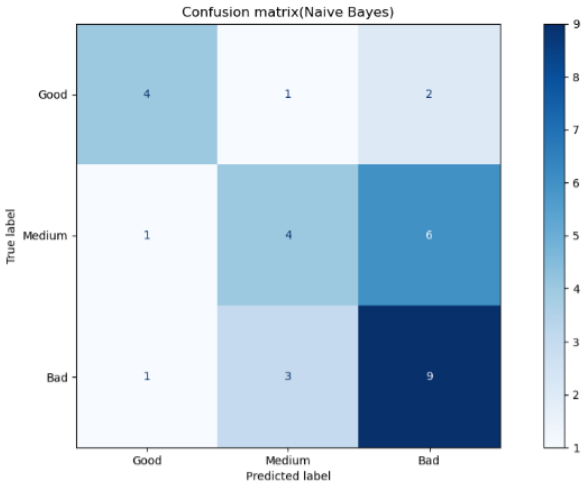

Fig. 9. Confusion Matrix for Gaussian Naïve Bayes.

H. Decision Tree Classification

Decision tree Classifier is a supervised learning method. Information gain is the most popular attribute selection approach for Decision Tree. Information gain can be calculated using the following Equation (5) [25].

$$Info(D) = -\sum_{i=1}^{m} Pi \log_2 Pi \tag{5}$$

where, *Pi* is the probability of how uncertain we are about the data. Fig. 10. shows the scatter plot. The classifier algorithm properly categorized all of the data in our dataset. Fig. 11. shows the Confusion Matrix of Decision Tree Classifier. The only class that performs better is Medium with a score of 13. In the Confusion Matrix, the higher value of the diagonal indicates that most of the data were predicted correctly.

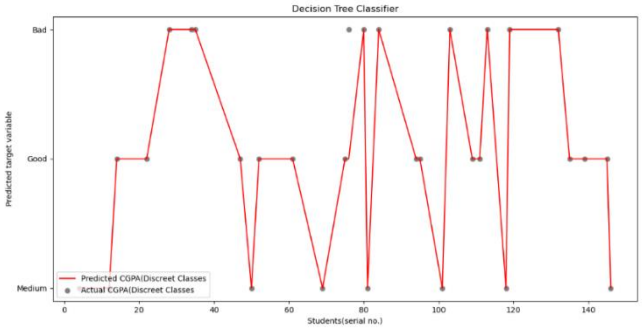

Fig. 10. Scatter plot for decision tree classifier.

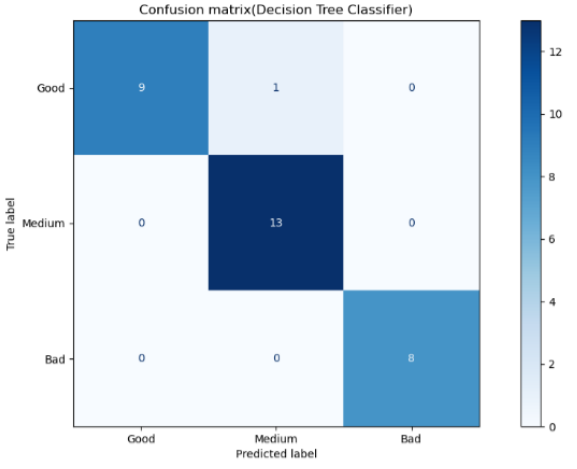

Fig. 11. Confusion matrix of decision tree classifier.

I. Chi-Square

The Chi-Square is widely used to determine Tests of Independence using a crosstabulation. The calculation of Chi-Square is very elementary. To identify the top rank features, we applied Chi-Square algorithm on our data. Table IV shows the ranking result.

TABLE IV: IMPORTANT FEATURES WITH RANK VALUES

Serial no.	Feature Name	Ranked Values
1	Medium of Study	14.298395
2	Gender	4.6742440
3	Chronic medical condition	4.446018
4	Students' study hour (in hours)	3.738780
5	Time needed to come to university	3.264150
6	Number of family member	3.090286
7	Monthly family income	2.6074411
8	CGPA after 1st semester	1.158824
9	Type of School at Higher Secondary Level	0.962738
10	SSC/O' Level result	0.415696

From the Table IV above, it is noted that the student's medium of study is the most important aspect that often affects their grades, while the result of SSC/O' Level has less effect on the determination of grades than others.

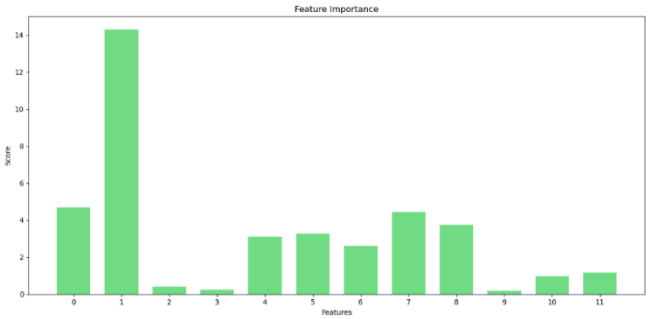

Fig. 12. Bar Chart of Feature Importance using Chi-Square.

Fig. 12 shows the Bar Chart for ranking of important features using Chi-Square. The points from the above Bar Chart indicates the following features:

- 0: Gender,
- 1: Medium of study,
- 2: SSC/O' Level result,
- 4: Number of family member,
- 5: Time needed to come to university,
- 6: Monthly family income,
- 7: Chronic medical condition,
- 8: Students study hour (in hours),
- 10: Type of school at Higher Secondary Level and
- 11: CGPA after 1st semester.

J. Pearson Correlation Coefficient

Pearson correlation highlight a one-dimensional relation between two variables [26]. To calculate the feature importance, we also applied Pearson correlation on our data. The result we find by Pearson correlation as seen in Fig. 13.

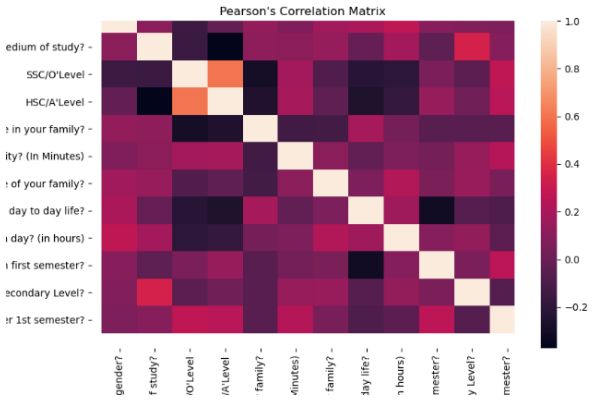

Fig. 13. Feature Importance using Chi-Square feature importance calculation using Pearson Correlation Matrix.

The Pearson coefficient's optimum value ranges from -1 to + 1. When the coefficient value goes above the optimum level, it would be called swift. And if the value goes to 0, then it is referred to as weak. In our analysis, none of the features surpass the optimum limit required, which will then be known as the significant features for the classifier algorithm. So, each of the features has equal importance at exactly the same degree.

V. RESULT ANALYSIS

A. Accuracy

Accuracy is one of the most useful metrics used by machine learning algorithms for testing models trained. Higher precision in our analysis indicates reliability of the models to determine accurately whether the student is at risk or not.

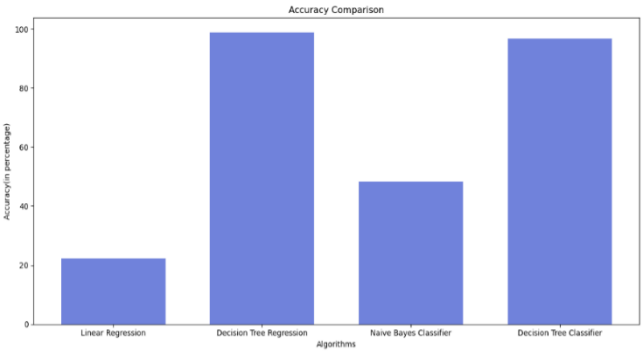

Fig. 14. Bar chart showing accuracy of four different algorithms.

Table V shows the of accuracy for different classification and regression algorithm we used on different student data. The classification algorithms we used are Gaussian Naïve Bayes and Decision Tree Classification. Here, Decision Tree Classifier (DT Classifier) offers the highest result in classification with a precision varies within 95 to 100%. Contrarily, Linear Regression and Decision Tree Regression also been used. The most reliable model developed using the regression algorithm is the Decision Tree Regression (DT Regression), which has an accuracy varies between 90% to 100%. For our dataset, however, the Linear Regression technique has the lowest accuracy because it only properly predicts a few data points.

TABLE V: ACCURACY OF DIFFERENT CLASSIFICATION AND REGRESSION ALGORITHM

Model	Accuracy
Decision Tree Regression	90% - 100%
Linear Regression	20% - 50%
Gaussian Naïve Bayes	50% - 60%
Decision Tree Classifier	95% - 100%

B. Feature Analysis

The most important feature in determining students' performance is Medium of study. This is obtained by applying Chi-Square based feature selection method. Fig. 15 shows the comparison result between students from different mediums gained by implementing Chi-Square algorithm on the dataset.

Fig. 15 sums this up- students who are of English medium background perform significantly well followed by students of the English version. In all the three mediums, the result of Bangla Medium students is relatively low.

C. Comparison with Previous Works

Previously there has been many research works in the field of Educational Data Mining. Each research uses a different machine learning approach, with varied results. In Table VI we showed the comparison among our proposed model and the few existing early grade prediction based research work where they worked with different dataset and attributes.

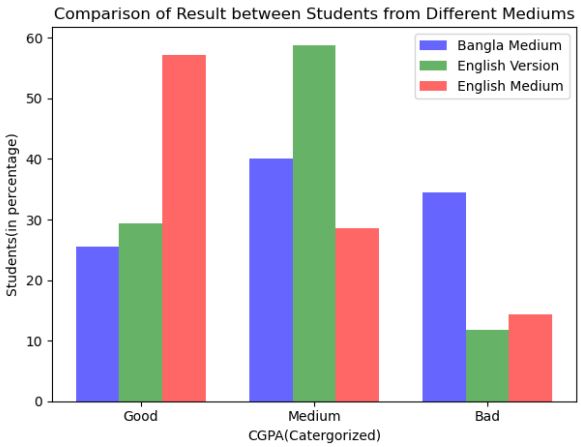

Fig. 15. Comparison of result between students from different mediums.

TABLE VI: COMPARISON WITH PREVIOUS WORKS ON STUDENT GRADE PREDICTION

Parameters	Kumar *et al.* [27]	Altabrawee *et al.* [7]	Junejo *et al.* [28]	Present Study
Features	Personal, demographic information and students' usage of social media	Grades, Personal information and student satisfaction	Sessional grades	Gender, personal, educational and family related information
Sample size	300	161	2500	155
Machine Learning Algorithm used	ID3 and J48	ANN, NB, LR and DT	ID3, KNN, NB and Rule Induction	DT Classifier, Gaussian NB, DT Regression and LR
Best Algorithm	ID3 and J48	ANN	Rule Induction	DT Classifier and DT Regression
Accuracy of prediction	62.667%	77.04%	73-96%	95-100% for DT Classifier and 90-100% for DT Regression
Metrics	Accuracy	ROC index	Accuracy, precision and recall	Accuracy

VI. CONCLUSION

We have conducted this research to predict an early estimation of the students' grades. Four Machine learning algorithms were applied on the pre-processed data. Among the used algorithms Decision Tree Classifier consistently performed better than others. The most inconsistent results were provided by Linear Regression with an accuracy of no more than 20-50%. The Chi-Square algorithm used on the data provided us with the top rank features, where a student's medium of study was the most important factor affecting the grades and SSC/O' Level GPA was the least. It should also be noted that the Pearson Correlation Coefficient was not effective in determining feature importance.

The findings of the research work are associated with a few limitations. First of all, the sample data that we worked on is small in size and does not cover large diversity which affects the results. Secondly, the data collected is disproportionate. For example, in our data set there were greater numbers of data of the Male gender than Female whereas the CSE/CS department of BRAC University has students of both gender in almost equal number. Also, there were fewer data for English Medium and English Version students than Bengali Medium. Finally, the data has been collected from the CS/CSE department only whereas in reality other departments' results may not match with the CSE students' dataset.

In future, we plan to work with our methodology on a bigger and more diversified dataset. Our work can be implemented in universities of other developing countries like Bangladesh. Our work will aid students to work on their shortcomings while doing the course and before it is too late. The educators can alert the students beforehand and provide them guidance accordingly. To conclude, the proposed method can improve academic results and thus bring massive changes to our education system.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Sumaiya, Jerin, Sakib collected the data; Sakib, Mahjabin analyzed the data; Mahjabin developed the model and carried out the experiments; All authors conducted the research; Sumaiya, Jerin, Sakib wrote the paper; All authors discussed the results and approved the final version; All authors contributed equally.

ACKNOWLEDGMENT

Firstly, we would like to thank our supervisor Dr.

Mahbubul Alam Majumdar for his support, feedback, guidance and contribution in conducting this research. We are grateful to him for his supervision in completing our research.

We are also grateful to Anal Acharya and Devadatta Sinha respectively from Department of Computer Science, St Xavier's College, Kolkata, India and Department of Computer Science and Engineering, University of Calcutta, Kolkata, India for their article "Early Prediction of Students Performance using Machine Learning Techniques" which was published 14 December, 2014 in International Journal of Computer Applications. We followed their article to conduct our research.

Lastly, we thank BRAC University to give us the opportunity to conduct this research and for giving us the chance to complete our Bachelor degree.

REFERENCES

[1] S. Boonman. (2018). Early prediction in students' performance in a distance learning university. [Online]. Available: http://arno.uvt.nl/show. cgi?fid=147602
[2] M. Tsiakmaki, G. Kostopoulos, S. Kotsiantis, and O. Ragos, "Implementing autoML in educational data mining for prediction tasks," *Applied Sciences*, vol. 10, no. 1, p. 90, 2019.
[3] N. Harb and A. El-Shaarawi, "Factors affecting students' performance," *Journal of Business Education*, vol. 82, no. 5, pp. 282–290, 2007.
[4] A. Raychaudhuri, M. Debnath, S. Sen, and B. G. Majumder, "Factors affecting students' academic performance: A case study in Agartala municipal council area," *Bangladesh e-Journal of Sociology,* 2010.
[5] I. Mushtaq and S. N. Khan, "Factors affecting students' academic performance," *Global Journal of Management and Business Research*, vol. 12, no. 9, 2012.
[6] S. Rovira, E. Puertas, and L. Igual, "Data-driven system to predict academic grades and dropout," *PLoS ONE*, vol. 12, no. 2, 2017.
[7] H. Altabrawee, O. A. J. Ali, and S. Q. Ajmi, "Predicting students' performance using machine learning techniques," *Journal of University of Babylon, Pure Applied Science*, vol. 27, no. 1, 2019.
[8] A. Tekin, "Early prediction of students' grade point averages at graduation: A data mining approach," *Eurasian Journal of Educational Research*, pp. 207–226, 2014.
[9] V. Anand, S. Kumar, and A. N. Madheswari, "Students results prediction using machine learning techniques," *International Journal of Advanced Science and Applications*, vol. 3, no. 2, pp. 325–329, 2016.
[10] Z. Iqbal, J. Qadir, A. N. Mian, and F. Kamiran, "Machine learning based student grade prediction: A case study," Department of Computer Science and Department of Electrical Engineering, Information Technology University, Lahore, Pakistan, 2017.
[11] N. Thai-Nghe, L. Drumond, and T. H. L. Schmidt-Thieme, "Multi-relational factorization models for predicting student performance," *Computer Science*, 2011,
[12] M. Pojon, "Using machine learning to predict student performance," M.S. thesis, Faculty of Natural Sciences, Software Development, University of Tampere, 2017.
[13] S. Al-Sudani and R. Palaniappan, "Predicting students' final degree classification using an extended profile," *CrossMark, Education and Information Technologies*, 2019.
[14] A. Siri, "Predicting students' dropout at university using artificial neural networks," *Italian Journal of Sociology of Education*, vol. 7, no. 2, pp. 225–247, 2015.
[15] F. Marbouti, H. A. Diefes-Dux, and K. Madhavan, "Models for early prediction of at-risk students in a course using standards-based grading," *Computers & Education*, vol. 103, pp. 1–15, 2016.
[16] Saedsayed. [Online]. Available: https://saedsayad.com/decision tree reg.htm
[17] Using chi-square statistic in research. [Online]. Available: https://www.statisticssolutions.com/using-chi-square-statistic-in-research/
[18] Pearson Correlation and Linear Regression. [Online]. Available: http://sites.utexas.edu/sos/guided/inferential/numeric/bivariate/cor/
[19] Y.-L. K. Samo. (Aug. 2018). [Online]. Available: https://towardsdatascience.com/the-black-swans-in-your-market-neutral-portfolios-part-i-7521683a7317
[20] A. Acharya and D. Sinha, "Early prediction of students performance using machine learning techniques," *International Journal of Computer Applications*, vol. 107, no. 1, 2014.
[21] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed., 2011.
[22] R. Gandhi, "Introduction to machine learning algorithms: Linear regression," *Towardsdatascience*, May 2018.
[23] R. Jain, "Decision tree. It begins here," *Medium*, June 2017.
[24] Naive bayes classification using scikit-learn. [Online]. Available: https://www.datacamp.com/community/tutorials/naive-bayes-scikit-learn
[25] Decision tree classification in python. [Online]. Available: https://www.datacamp.com/community/tutorials/decision-tree-classification-python
[26] "Correlation coefficient: Simple definition, formula, easy calculation steps." [Online]. Available: https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/
[27] A. D. Kumar, R. P. Selvam, and V. Palanisamy, "Prediction of student performance using hybrid classification," *International Journal of Recent Technology and Engineering*, vol. 8, no. 4, 2019.
[28] K. Junejo and E. Eman, "Grade prediction using supervised machine learning techniques," in Proc. *Conference: 4th Global Summit on Education*, 2016.

Sumaiya Iqbal received her bachelor of science degree in computer science and engineering from BRAC University, Dhaka, Bangladesh in 2020.

At present, she is making preparations for pursuing a master's degree soon. Her fields of interest include Software engineering and digital communications.

Besides that, she is now working as a technical executive at ekShop, a2i, Dhaka, Bangladesh. Her job focuses on web design and software development.

Mahjabin Muntaha received her bachelor of science degree in computer science and engineering from BRAC University, Dhaka, Bangladesh in 2020.

She is an associate creative engineer at Orbitax Bangladesh Limited, Dhaka, Bangladesh.

Currently, she is taking preparations for pursuing a master's degree. Her areas of interest are Machine Learning with applications in Artificial Intelligence.

Jerin Ishrat Natasha received her bachelor of science degree in computer science from BRAC University, Dhaka, Bangladesh in 2020.

She is currently working as a junior software engineer at AusAsia Group, an Australian based IT Company. Her work focuses on software development projects and quality assurance engineering.

She now plans on attaining more in-depth knowledge through a Graduate Master's degree specialized in advanced software engineering.

Dewan Sakib received her bachelor of science degree in computer science from BRAC University, Dhaka, Bangladesh in 2020.

He is working as a content analyst at Bigo Technology LTD.

He is currently planning to pursue his master's in near future. His areas of interest are machine learning, artificial intelligence and data science.