

# Early Grade Prediction using Profile Data

by

Sumaiya Iqbal

16101189

Mahjabin Muntaha

16101246

Jerin Ishrat Natasha

19241035

Dewan Sakib

19341009

A thesis submitted to the Department of Computer Science and Engineering  
in partial fulfillment of the requirements for the degree of  
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering  
BRAC University  
April 2020

© 2020. BRAC University  
All rights reserved.

# Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at BRAC University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

**Student's Full Name & Signature:**

*Sumaiya*

---

Sumaiya Iqbal  
16101189

*Jin*

---

Jerin Ishrat Natasha  
19241035

*Mahjabin*

---

Mahjabin Muntaha  
16101246

*SA*

---

Dewan Sakib  
19341009

# Approval

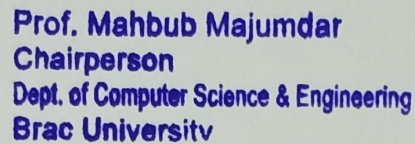
The thesis titled “Early Grade Prediction using Profile Data” submitted by

1. Sumaiya Iqbal (16101189)
2. Mahjabin Muntaha (16101246)
3. Jerin Ishrat Natasha (19241035)
4. Dewan Sakib (19341009)

Of Spring, 2020 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on April 07, 2020.

## Examining Committee:

Supervisor:  
(Member)



**Prof. Mahbub Majumdar**  
**Chairperson**  
**Dept. of Computer Science & Engineering**  
**Brac University**

Dr. Mahbubul Alam Majumdar  
Professor and Chairperson  
Department of Computer Science and Engineering  
BRAC University

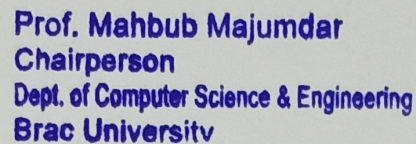
Program Coordinator:  
(Member)



---

Md. Golam Rabiul Alam  
Associate Professor  
Department of Computer Science and Engineering  
BRAC University

Head of Department:  
(Chair)



**Prof. Mahbub Majumdar**  
**Chairperson**  
**Dept. of Computer Science & Engineering**  
**Brac University**

Dr. Mahbubul Alam Majumdar  
Professor and Chairperson  
Department of Computer Science and Engineering  
BRAC University

## **Ethics Statement**

We, hereby declare that this thesis is based on the results we obtained from our work. Due acknowledgement has been made in the text to all other material used. This thesis, neither in whole nor in part, has been previously submitted by anyone to any other university or institute for the award of any degree.

## 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.

**Keywords:** Machine Learning Algorithms; Linear Regression; Decision Tree Regression; Gaussian Naïve Bayes; Decision Tree Classifier; Feature Importance; Chi-Square

## **Dedication**

Firstly, this thesis is dedicated to our parents for their love, effort, endless support and encouragement. Last but not the least, we are also thankful to our supervisor Dr. Mahbulul Alam Majumdar. Without his constant support and guidance it would have been impossible for us to conduct this research.

## Acknowledgement

Firstly, we would like to thank our supervisor Dr. Mahbubul Alam Majumdar for his support, feedback, motivation, 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.

# Table of Contents

<b>Declaration</b>	<b>i</b>
<b>Approval</b>	<b>ii</b>
<b>Ethics Statement</b>	<b>iii</b>
<b>Abstract</b>	<b>iv</b>
<b>Dedication</b>	<b>v</b>
<b>Acknowledgment</b>	<b>vi</b>
<b>Table of Contents</b>	<b>vii</b>
<b>List of Figures</b>	<b>x</b>
<b>List of Tables</b>	<b>xi</b>
<b>Nomenclature</b>	<b>xii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Thoughts behind the Prediction Model . . . . .	1
1.2 Problem Statement . . . . .	1
1.3 Research techniques . . . . .	2
<b>2 Literature Review</b>	<b>4</b>
2.1 Data Mining . . . . .	4
2.2 An Ensemble-Based Semi-Supervised Learning Approach . . . . .	5
2.3 Hybrid Classification Algorithms . . . . .	5
2.4 Data Driven System . . . . .	6
2.5 Machine Learning Techniques . . . . .	6
2.6 Microsoft Azure in EDM . . . . .	7
2.7 Multi-Relational Factorization Model . . . . .	7
2.8 Classification and Regression . . . . .	7
2.9 Supervised Learning Algorithms . . . . .	8
2.10 Neural Network and Classifiers . . . . .	8
2.11 At-Risk Student Prediction Model . . . . .	9
2.12 AutoML for Prediction Tasks . . . . .	9



<b>3</b>	<b>Background Study</b>	<b>10</b>
3.1	Academic Major . . . . .	10
3.2	Type of School . . . . .	10
3.2.1	Private School . . . . .	10
3.2.2	Public School . . . . .	10
3.3	Results of Previous Board Exams . . . . .	10
3.3.1	Secondary School Certificate . . . . .	10
3.3.2	Higher School Certificate . . . . .	12
3.3.3	O' Level . . . . .	13
3.3.4	A' Level . . . . .	13
3.4	Medium of Study . . . . .	14
3.4.1	Bangla Medium . . . . .	14
3.4.2	English Version . . . . .	14
3.4.3	English Medium . . . . .	14
3.5	Family Size . . . . .	15
3.6	Family Income . . . . .	15
3.7	Chronic Medical Condition . . . . .	15
3.8	CGPA . . . . .	15
3.9	Others . . . . .	16
3.10	Data Analysis with Various Algorithms . . . . .	16
3.10.1	Regression Algorithm . . . . .	17
3.10.2	Linear Regression . . . . .	17
3.10.3	Regression with Decision Tree . . . . .	17
3.10.4	Classification Algorithm . . . . .	17
3.10.5	Gaussian Naïve Bayes . . . . .	17
3.10.6	Decision Tree Classifier . . . . .	17
3.10.7	Chi-Square . . . . .	18
3.10.8	Pearson Correlation Coefficient . . . . .	18
3.11	Confusion Matrix . . . . .	18
<b>4</b>	<b>Questionnaire Description</b>	<b>19</b>
<b>5</b>	<b>Proposed Methodology</b>	<b>26</b>
5.1	Workflow . . . . .	26
5.2	Attribute and Dataset Description . . . . .	27
5.3	Dataset Pre-Processing . . . . .	29
5.4	Model Implementation . . . . .	32
5.5	Linear Regression . . . . .	32
5.6	Decision Tree Regression . . . . .	34
5.7	Gaussian Naïve Bayes . . . . .	35
5.8	Decision Tree Classification . . . . .	37
5.9	Chi-Square . . . . .	38
5.10	Pearson Correlation Coefficient . . . . .	39
<b>6</b>	<b>Result Analysis</b>	<b>40</b>
6.1	Accuracy . . . . .	40
6.2	Feature Analysis . . . . .	41
6.3	Comparison with Previous Works . . . . .	41

<b>7 Concluding Remarks</b>	<b>43</b>
<b>Bibliography</b>	<b>47</b>

# List of Figures

4.1	.....	19
4.2	.....	19
4.3	.....	20
4.4	.....	20
4.5	.....	20
4.6	.....	21
4.7	.....	21
4.8	.....	22
4.9	.....	22
4.10	.....	23
4.11	.....	23
4.12	.....	23
4.13	.....	24
4.14	.....	24
4.15	.....	25
4.16	.....	25
4.17	.....	25
5.1	Workflow of the proposed method for predicting students' grade . . .	26
5.2	An overview of all students CGPA after 9th semester from the dataset	29
5.3	Student count across categories . . . . .	31
5.4	Scatter plot for Linear Regression . . . . .	33
5.5	Bar Chart of Linear Regression . . . . .	33
5.6	Scatter plot for Decision Tree Regression . . . . .	34
5.7	Bar Chart of Decision Tree Regression . . . . .	35
5.8	Scatter plot for Gaussian Naïve Bayes . . . . .	36
5.9	Confusion Matrix for Gaussian Naïve Bayes . . . . .	36
5.10	Scatter Plot for Decision Tree Classifier . . . . .	37
5.11	Confusion Matrix of Decision Tree Classifier . . . . .	37
5.12	Bar Chart of Feature Importance using Chi-Square . . . . .	38
5.13	Feature Importance calculation using Pearson Correlation Matrix . .	39
6.1	Bar Chart Showing Accuracy of Four Different Algorithms . . . . .	40
6.2	Comparison of Result between Students from Different Mediums . . .	41

# List of Tables

3.1	Subjects of different groups in SSC . . . . .	11
3.2	Grading system of SSC . . . . .	11
3.3	Subjects of different groups in HSC . . . . .	12
3.4	Grading system of HSC . . . . .	13
3.5	Subjects of O' Level (IGCSE) . . . . .	13
3.6	Subjects of A' Level . . . . .	14
3.7	Grading Policy of BRAC University . . . . .	16
5.1	Student attributes and their domains . . . . .	28
5.2	Table of converting attributes domains into discrete variables . . . . .	30
5.3	Students' category, based on their CGPA . . . . .	31
5.4	Important features with Rank values . . . . .	38
6.1	Accuracy of different classification and Regression Algorithm . . . . .	41
6.2	Comparison with Previous Works on Student Grade Prediction . . . . .	42

# Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

*ANN* Artificial Neural Network

*BI* Business Intelligence

*CF* Cascade-forward Network

*DT* Decision Tree

*EDM* Educational Data Mining

*ELM* Extreme Learning Machine

*FFvariation* Feed-forward Network Variation

*FF* Feed-forward Network

*ID3* Iterative Dichotomiser 3

*KNN* K- nearest neighbor

*LMT* Logistic Model tree

*LR* Logistic Regression

*MAE* Mean Absolute Error

*MF* Matrix Factorization

*MRMF* Multi-Relational Matrix Factorization

*NB* Naïve Bayes

*RMSE* Root Mean Squared Error

*ROC* Receiver Operating Characteristic

*SSL* Semi-supervised learning

*SVM* Support Vector Machine

*WMMRF* Weighted Multi-Relational Matrix Factorization

# Chapter 1

## Introduction

### 1.1 Thoughts behind the Prediction Model

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 [21]. As a result, dropout rates tend to go higher as well. This is an undesirable situation for both the student and the university. So, it is important to identify the students at risk before it is too late. Additionally, students with unsatisfactory grades attempt for retakes of the courses. Therefore, the students fail to graduate on time. This is again a matter of concern for both the families of the students as it costs extra expenses for education and also delays the graduation period. Such events 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 [32].

### 1.2 Problem Statement

We know good grades are extremely important for students. For students who have just entered the university, the studying strategy plays an important role. As university is a place of experimental learning, the traditional method of studying such as memorizing whole textbooks or notes does not work. Most of the time students fail to identify the correct techniques for studying and they are unable to achieve satisfactory results. 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 [4], students' attendance [5], gender [3] 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. It is the responsibility of the teachers and the university to identify these problems on time because failing to do so will result in a large number of academically unsuccessful students.

We live in a technological era. Technology has made everything inconceivably easy for us. Nowadays, there is a lot of opportunity for a student to make life easier. Most of the time, because of the family situation, students have failed to use such resources or have little interest in them. Since they do not to utilize the opportunities, they fail to stand up to their expectations.

If a student gets proper instruction and practices modern learning methods, they can get a good CGPA. A study [7] indicated that if properly directed by parents and teachers, a student will have a successful academic career. Yet, only a few students get proper facilities. Others struggle to grasp the techniques. This paper is therefore focused on evaluating the factors influencing the academic performance of the students.

### 1.3 Research techniques

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 [19]. Additionally, machine learning techniques provide adaptability over statistics which is also beneficial [19]. 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 [29]. In this paper, a set of attributes such as GPA of Higher Secondary School Certificate (HSC) or A' Levels, Secondary School Certificate (SSC) or O' Levels grades, medium of study, time spent on studying each day, CGPA after 1st and 9th semester, time spent for commuting to the university, their chronic medical conditions (if any), class attendance 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 like CGPA after 1st and 9th semester, GPA of HSC/A' Level and SSC/O' Level and class attendance percentage in 1st semester. 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 data set is taken as input. The data set contains different types of information. The dataset is represented in tabular format, where each row represents a student and each column represents information about a student such as grades, gender, family income or medical information. 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.

The predicted grades were categorized into 3 types- Bad, Medium and Good. The corresponding GPA is considered to be less than 3.0 for Bad, less than 3.5 for Medium and 3.5 to 4.0 for Good. The prediction algorithm works on two target labels- "Predict" which is for the actual CGPA and "Predict\_discrete" which is for the CGPA categories.

Four algorithms have been used in total where two of them are Regression algorithms and two of them are Classification algorithms. The Regression algorithms applied on target label "Predict" includes Linear Regressions and Decision Tree Regression. The Classification algorithms applied on target label "Predict\_discrete" include Gaussian Naïve Bayes and Decision Tree Classifier.

The result of our model gave varying accuracy for the different algorithms used. The accuracy of Linear Regression was very poor which varied between 20%-50%, for Gaussian Naïve Bayes it varied between 50%-60%, Decision Tree Classifier – 95%-100%, Decision Tree Regression – 90%-100% which indicates that Decision Tree Classifier algorithm provided the most accurate results. 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. From the ranking results we obtained, it was seen that the attribute Medium of study affected the grades the most and GPA of SSC/O' Levels had the least impact on grades.

Many might think that a student with high CGPA towards the last few semesters must have had high CGPA initially after the first semester in their undergraduate life and vice versa. But our research has shown that CGPA after first semester had very little impact on the final CGPA of the students. This implies that students should not be discouraged if they perform poor at the start of their university life. With proper guidance from the teachers and with their own effort, they can improve their CGPA with time.

The later sections of our paper is organized in the following way. In section 2, we talked about related work in this field. Section 3 elaborates on the various machine learning techniques that have been used to predict the CGPA of the students. Section 4 describes the questionnaire that we used to collect students' profile data. In section 5, we discussed about the proposed methodology of our research paper. The analysis of our results have been shown in Section 6. At last in section 7 we concluded our work.



# Chapter 2

## Literature Review

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.

### 2.1 Data Mining

The final grade prediction based on a lot of information is challenging. The dataset contains an assortment of important information but there are moreover a few vital covered up information that most of the time cannot be extricated. Educational data processing theory, points to extricate valuable hidden information from large datasets [10], [12]. Data mining approach is additionally important to convert data into usable data, identify students' performance patterns, mode and after that on [12].

Tekin [10] 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 that are required to be effectively passed by the students for graduation. 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. His study revealed that all three algorithms performed well in foreseeing students' GPAs at graduation. Among all three methods the SVM technique yielded more precise predictions rate of 97.98%. The ELM method yielded the second most exact expectation rate (94.92%). Finally, NN showed the slightest exact prediction rate (93.76%).

## 2.2 An Ensemble-Based Semi-Supervised Learning Approach

Semi-supervised learning (SSL) is newer when compared with supervised and unsupervised learning [26]. It acts as a midway between supervised and unsupervised learning. Semi-supervised learning (SSL) offers a strong structure for unlabeled data when labels are constrained to get [25]. Ensemble process, on the other hand, is a learning technique that incorporates several models to diminish their classification errors yet on improve the precision of weak classifiers. Ensembles predictions are typically precise than single mode predictions [30].

Livieris et al. [24] proposed a technique to foresee students' performance within final exam. The technique comprised of three stages. Firstly, a dataset containing the performance of 799 understudies within the courses of "Mathematics" were collected. They chose the course "Mathematics" because it is taken into account one among the toughest courses within the field of science. The chosen attributes observed the students' execution on the primary and second academic semester. The attributes than categorized into 20 –point evaluating scale, where "0" was the least grade and "20" was that the most elevated grade. Within the following stage, the authors proposed an ensemble-based SSL algorithm. They assessed the performance of the developed algorithm which utilized an ensemble as base learner (signified as Vote). Within the final stage, authors compared their created ensemble-bases SSL algorithm with the foremost prevalent SSL algorithms by conducting a number of experiments. The SSL algorithms they chose were self-training, co-training and tri-training as base learners (signified as Best). The chosen supervised classifiers were Naïve Bayes (NB), Logistic Model Tree (LMT) and so on. The classification exactness of all learning algorithms was assessed by a regular strategy called stratified tenfold cross-validation. The proportion was labeled as separately 10%, 20% and 30%. After lots of experimental phases their experimental results showed that their proposed strategy was demonstrated to be viable for early understudy advance prediction as compared to a few existing semi-supervised learning technique. They analyzed the trend which appeared the SSL algorithms that utilized "Vote" as base classifier showed similar classification outcome unconventional of the proportion of the labeled data and dataset.

## 2.3 Hybrid Classification Algorithms

Kumar et al. [31] had talked about the utilization of big data in different areas of educational data mining such as enhancing students' performance, minimizing dropouts and numerous more. They too addressed methods employed in Educational Data Mining (EDM) and therefore the most recent technologies utilized in big data. They used hybrid classification algorithm in their research to predict student results. The dataset included 300 students' information. They selected the factors that indirectly influence a students' outcome like father's wage, the educational status of mothers, the usage of social systems, students' hours of study, educational cost and mother working status. Data were pre-processed in the first for better prediction result. They applied the dataset containing 17 attributes in "Weka tool". Weka

tool is created by utilizing java code and weka has been widely used in education to predict because of its effectiveness in exploring, evaluating and anticipating students' performance [8]. They used "CfsSubsetEval" for better classification and to pick proper for classification phase. CfsSubsetEval estimates the importance by taking into account each attributes individual predictive performance together with the degree of redundancy between them [20].

The chosen traits were applied to hybrid classification algorithm after pre-processing stage. A voting method was then employed in weka machine learning to choose two best classification algorithms. They were ID3 and J48. J48 algorithm produces decision trees which can be used for classification problems while ID3 is a simple algorithm for decision tree [6]. The tested result showed that the hybrid classification algorithm classified 188 occurrences accurately and classified 112 occurrences incorrectly. The classification accuracy was 62.667%. The authors concluded that the given data set was exactly predicted by the hybrid algorithm.

## 2.4 Data Driven System

In a study Rovira et al. [19] suggested a data driven framework to extricate valuable hidden data within the student scholastic information and hence make predictions of academic grades and dropout intention of students. The research consisted basically of two activities. The primary task was to forecast of students' dropout. Comparison of five separate classifier algorithms (Naïve Bayes, Logistic Regression, Support Vector Machines, Random Forest, and Adaptive Boosting) was done. After that, comparison among three distinctive approaches like Collaborative Filtering, Support Vector Regression and Linear Regression was done to anticipate the grades of students. The efficiency of the classifiers was measured according to the standard measures of accuracy, recall precision and F1. Mean Absolute Error (MAE) score. Grades of students were considered as feature vector for all the methods.

The dataset included only the students' grades. The authors split the data into train and test set. The values within the dataset were scaled in 0 to 10 ranges. After pre-processing the data and applying the classifier methods the authors revealed that Naïve Bayes and Logistic Regression performed well in predicting students' grades and dropout. They also had developed a visualization tool to better explain the obtained result.

## 2.5 Machine Learning Techniques

A number of considerations are taken into account while determining a grade for students. The use of social media as entertainment or research is two important factors that influence a student's study seriously. Altabrawee et al. [29] incorporated these two additional properties other than personal details of the students in their dataset. The students were labeled as weak (final 1st semester grade less than 60 out of 100) and good (final grade equal to or greater than 60). Machine learning techniques used in the study were Artificial Neural Network (ANN), Naïve Bayes, Logistic Regression and Decision Tree. The models were executed by RapidMiner

Studio Machine Learning Technique. The performance measure of the ROC index and the classification accuracy were used to compare the models. Authors defined Decision Tree model showed that not all the traits had an effect on classifying the student's status into Good or Weak and established that ANN model had the highest ROC performance and accuracy.

## 2.6 Microsoft Azure in EDM

Microsoft Azure is a lesser utilized machine learning technique to know the scholarly standard of the students. Anand et al. [13] built models using Microsoft Azure Machine Learning Studio as web services. The significance of the web services is that it can be accessed through custom applications and Business Intelligence (BI) software such as excel. 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 data provided needed to split for transformation of the data. First 70% data was used for training and therefore the remaining 30% data was used for model testing. The train model included ML algorithm and the input of 70% of split data. The score model included train model and 30% of split results. 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.

## 2.7 Multi-Relational Factorization Model

The factorization procedures are useful in case of sparse data [16]. Thai-Nghee et al. [45] suggested the technique of Matrix Factorization (MF) to predict the performance of students at courses in Algebra and Bridge to Algebra. In the research authors showed relationship between several variables, such as task provided to students, their ability to solve those tasks and many more. Multi-Relational Matrix Factorization (MRMF) accustomed exploit between students, tasks and their Meta data. They moreover proposed a Weighted Multi-Relational Matrix Factorization (WMRMF) to take into consideration the most connection which contains the target variable. The dataset contained information reflected the contact of log files between students and computer aided tutoring systems. They split the dataset into train set and test. 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 and WMRMF can perform nicely compared to the other methods, taking into account the multiple relationships between entities. Also, multi-relation approach can anticipate students' performance much better than Single-Relational MF.

## 2.8 Classification and Regression

When using finite sets, grade prediction using Linear Regression, Decision Trees, and Naïve Bayes classifier are effective. In a paper Pojon [18] made a distinction

between machine learning approaches and feature engineering techniques about how much they improve the accuracy of prediction. The data were split into train set (75%) and test set (35%). The dataset had baseline exactness, based on which the model was built to compare whether the model can make successful prediction or not. The first model was developed using the framework of Linear Regression. The created model was then applied to test dataset. The second and third models were constructed using Decision Tree method and Naïve Bayes. The procedure was as same as the previous demonstrate. The datasets were modified to improve the prediction efficiency. The modification method used was feature selection. And the result showed that, the Naïve Bayes classification produced the best results, followed by Decision Tree and Linear Regression.

## 2.9 Supervised Learning Algorithms

Supervised machine technique is the most common type of machine learning technique, where case input-output sets are vital for the learning function [21]. There are several supervised learning techniques available. In a study, Junejo and Eman [14] used supervised techniques and attempted to predict the final grade of more than 2000 students based on their class tests, lab scores, class participation, homework etc. prior to the final test. They primarily used supervised machine learning approaches such as ID3, K-Nearest Neighbor, Naïve Bayes, and Rule Induction method to classify students' grade. An important point noted from the analysis was that classifiers founded results were close to the actual labels. Predicted performance accuracy varying between 73 to 96 %. The conclusion authors made from their research was that both the Decision Tree based classifier performed significantly better than the other classifiers with Rule Induction taking the lead followed by ID3 classifier.

## 2.10 Neural Network and Classifiers

Al-Sudani et al. [28] used NN models (Feed-forward Network (FF), Cascade-forward network (CF), Feed-forward network variation (FF variation)) 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. Their test showed that the best result accomplished with FF network. 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 [11] carried a similar study on a large dataset (810 data with 43 attributes) with the aim expanding the accuracy of the neural network. Al-Sudani et al. [28] has proven that using an extended combination of students' domain information better accuracy can be found.

## 2.11 At-Risk Student Prediction Model

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. To improve the generalizability of the models and the predictive accuracy, they used a feature selection method to minimize the number of variables used in each demonstrate. 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.

## 2.12 AutoML for Prediction Tasks

In a research, Tsiakmaki et al. [32] recommended the use of Automated Machine Learning (AutoML) procedure for academic context. Their study primarily focused on predicting students' academic behaviors such as their execution in courses, dropout and pass/fail intention. Secondly, they limited the configuration space by having only tree and rule based classification algorithm. A custom plugin for Moodle was designed to collect data. The data included the performance of the students in three mandatory courses provided by Moodle. In their work the used algorithms were Naïve Bayes, PART, IBK, Sequential Minimal Optimization (SMO), Random Forest (RF) and Bagging. Where RF and SMO have been identified as best learning algorithms. All of their tests were performed using the AutoML framework. After much experimentation, their research showed that applying AutoML to educational data increases model efficiency over conventional approaches.

# Chapter 3

## Background Study

### 3.1 Academic Major

Academic Major is the scholarly specialty from which a student receives a bachelors or post-graduate degree. To receive an undergraduate or a postgraduate degree [35], a student must successfully pass all the requisite courses in a major discipline. Students usually select Academic Major from their field of interest. There are many major academics in different universities/colleges around the world. We decided to run the survey on CS/CSE major students of BRAC University for our research.

### 3.2 Type of School

#### 3.2.1 Private School

Private school is not funded by the State. Private schools are heavily dependent upon the students' monthly tuition fees and donations. The monthly tuition rates in private schools are higher than in the public schools, so they tend to provide better care and education generally. Private schools tend to be smaller in scale than the public schools. The ideologies also vary in several respects between the public school and private school.

#### 3.2.2 Public School

The biggest benefit of public school over private school is that they are sponsored by the State. The number of public schools are also more substantial. Primary and Secondary education in Bangladesh is entirely funded by the government in the Public Schools.

### 3.3 Results of Previous Board Exams

#### 3.3.1 Secondary School Certificate

The short form of Secondary School Certificate is SSC. SSC examination is conducted under the Bangladesh Central Board of Education. The exam is conducted in all of Bangladesh's districts and candidates across Bangladesh sit exactly at the

same time as the Board of Education decides. Every year an average of 2.1M students from 28000 institutions sit for SSC under Eight General Education boards in 3500 centers throughout the country. Students who completed all the classes (Class 1 to Class 10) successfully are eligible to sit for the SSC exam. There are 3 groups to choose from within SSC: i) Science ii) Commerce and iii) Arts. Students select any of these 3 groups at the beginning of Class 9 and after two years sit in SSC for their respective group exam. The following Table 3.1 [34] shows the subjects of different groups.

General Subjects	Science	Commerce	Arts
Bangla	Physics	Finance & Banking	Geography
English	Chemistry	Accounting	Civic & Citizenship
Math	Biology	Business Entrepreneurship	Economics
	Higher Math	General Science	General Science
	Information & Technology	Information & Technology	Information & Technology
	Islam & Moral Education	Islam & Moral Education	Islam & Moral Education
	Bangladesh & World	Home Science	Home Science
	Agriculture Studies	Agriculture Studies	Agriculture Studies
	Home Science	Music	Music

Table 3.1: Subjects of different groups in SSC

All students from any of the 3 groups must study the General Subjects. The remaining subjects depend on the group of study. The following Table 3.2 [33] shows the grading system of individual subjects in SSC.

Class Interval (Numbers Obtained)	Letter Grade	Grade Point
80-100	A+	5
70-79	A	4
60-69	A-	3.5
50-59	B	3
40-49	C	2
33-39	D	1

Table 3.2: Grading system of SSC

GPA (Grade Point Average) is calculated by adding all the subjects' grade point and dividing it by the total number of subjects. A student needs to pass all the subjects to earn a SSC Certificate.



### 3.3.2 Higher School Certificate

Higher School Certificate is the full form of HSC. HSC examination is conducted under the Bangladesh Central Board of Education. The exam is conducted in all districts of Bangladesh and candidates across Bangladesh sit exactly at the same time as the Education Board decides. Every year a total of 1.1M students from 9000 institutions sit for HSC under 10 Education boards around the country. Students who passed Class 11 & 12 successfully would be eligible to sit on the HSC test. HSC has also 3 groups to choose from: i) Science ii) Commerce and iii) Arts. At the beginning of college, students choose one of 3 groups and, after two years of study, sit for the HSC from their respective group. The subjects of different groups are shown in Table 3.3 [34].

General Subjects	Science	Commerce	Arts
Bangla	Physics	Accounting	Logic
English	Chemistry	Business Organization and Management	Economics
Information & Technology	Biology	Economics	Sociology
	Higher Math	Finance, Banking and Insurance	Geography
	Psychology (Optional)		Islamic History & Culture
	Geography (Optional)		Civic & Good Governance
	Statistics (Optional)		Social Work
	Engineering Draw and Work Practical (Optional)		History

Table 3.3: Subjects of different groups in HSC

All students from any of the 3 groups must study the General Subjects. The rest of the subjects are based on the group of study. GPA (Grade Point Average) of HSC is calculated by adding all the grade point of the subjects and dividing them with the total number of subjects. Everyone has to pass all of the subjects to earn HSC Certificate. The grading system of HSC is shown in Table 3.4 [33].

Class Interval (Numbers Obtained)	Letter Grade	Grade Point
80-100	A+	5
70-79	A	4
60-69	A-	3.5
50-59	B	3
40-49	C	2
33-39	D	1

Table 3.4: Grading system of HSC

### 3.3.3 O' Level

Ordinary Level is the full form of O' Level. It is also defined as IGCSE (International General Certificate of Secondary Education). It is a curriculum developed in the UK. British Council provides two separate O' Level/IGCSE syllabuses. They are: i) Edexcel International ii) Cambridge International Education (CIE) [42]. Most students usually choose 7-9 subjects in O' Level but they can choose a maximum of 14 subjects. Bangla Language, English Language and Mathematics are the compulsory subjects in all schools. Following Table 3.5 [38] shows some of the subjects of O' Level.

Bangla Language	English Language
Bangladesh Studies	Mathematics D
Pure Mathematics	Physics
Chemistry	Biology
Human Biology	Additional Mathematics
Accounting	Mathematics B
Computer Science	Economics
Art and Design	Business Studies

Table 3.5: Subjects of O' Level (IGCSE)

O' Level exams are held in two sessions – May/June and November for CIE and May/June and January for Edexcel. Students can give exams in any of the two sessions or both.

### 3.3.4 A' Level

In order to be admitted to University, one must pass A' Level. It is recognized worldwide. The subject lists of A' Level are given in Table 3.6 [38].

English Language	Physics
Chemistry	Mathematics
Biology	Human Biology
Accounting	Economics
Business Studies	IT/Computer Science
Art and Design	

Table 3.6: Subjects of A' Level

Grading System of O' Level/A' Level in Edexcel is not fixed. It varies depending on the difficulty level of the question papers in each session and marks are scaled where a student has to score a certain number of marks for each grade like A\*, A, B, C, D, E. These grades were then converted to the CGPA system according to the entry requirements for undergraduate admission in BRAC University. A minimum CGPA of 2.5 in O' Levels in five subjects and A' Levels in two subjects separately is required according to the scale (A=5, B=4, C=3 & D=2). Subjects with E grade does not qualify for the admission.

## 3.4 Medium of Study

### 3.4.1 Bangla Medium

Bangla Medium is Bangladesh's education system where textbooks are written in Bengali Language and the Education Board of Bangladesh suggested curriculum is followed. In Bangladesh, most student study in Bangla Medium. The number of Bangla Medium institutions is much higher than English Medium. In Bangla Medium, it is called Primary School from Class 1 to Class 5 and Class 6 to Class 10 is called High School. Students' must pass PSC (Primary Education School Certificate) at the end of primary school to enter High School. The grading of all classes is equivalent to all the board exams.

### 3.4.2 English Version

English Version shares a curriculum similar to Bangla Medium. The difference is that the textbooks are written in English Language. The textbooks contain almost the same materials as the text books of Bangla Medium but the words are written in English. In the English Version, it is called Primary School from Class 1 to Class 5 and Class 6 to Class 10 is called High School. Each student need to pass PSC (Primary Education School Certificate) to get into the High School. Grading system is also identical to all of the previously listed board examinations.

### 3.4.3 English Medium

Bangladesh has a number of renowned English Medium schools, particularly in Dhaka. They follow the UK education curriculum. Pearson Edexcel and Cambridge Assessment International Education are the most widely used syllabuses. Playgroup, KG-1 and KG-2 are typically a preparatory level in English medium schools. The core program begins at Grade I through Grade XII. In between, students must attend

O' Level to pass grade X and A' Level to complete school. A' Level is divided into two parts- Advanced Subsidiary (AS) and Advanced Level (A2). AS level and the A2 level altogether does the completion for the qualification of A' Level.

### **3.5 Family Size**

Family size can be defined by the number of number of individuals within the family. As a family, a group of people living together under the same roof and related to each other by blood, marriage, common-law union, adoption or relationship of support may be named. Understanding the family size is important in our analysis, as it impacts the education of a student in many ways. According to a study, Bangladesh's average household size in 2016 is 4.060 people [1].

### **3.6 Family Income**

Family earnings are considered overall allowance for family members aged 15 or older living in the same house. Wages, social security, child care, capital benefit from insurance, and dividends can be compensations [47]. From the viewpoint of Bangladesh, a family's average income is 602 USD per year [2]. Understanding family income is important for our research, since our country is not yet declared as a middle-income economy, but BRAC University's tuition fees are also the highest among all Universities of Bangladesh. While BRAC University has several funding programs, this cannot be managed to achieve by everyone. Family income thus certainly plays a significant role. We have ranged the family income from less than 30,000 BDT to more than 200,000 BDT in five groups.

### **3.7 Chronic Medical Condition**

Chronic Medical Conditions are chronic illness or illness last for at least three months or longer. Chronic medical problems require medical checkups and medicines on a daily basis. Chronic medical disease includes heart disease, cancer, asthma, diabetes and so on. In our survey, we asked the students whether they have any chronic medical problem because it affects ones' grade.

### **3.8 CGPA**

CGPA in its full form is the Cumulative Grade Point Average. The academic results CGPA range from 0.00 to 4.00. GPA and CGPA do has some variations. GPA is the academic performance of one particular semester, whereas CGPA is the average of the academic performance measured so far. For an undergraduate or a post-graduate student, CGPA is most relevant. BRAC University offers its students a specific grading system. The grading policy of BRAC University is given in Table 3.7 [43].

Class Interval (Numbers Obtained)	Letter Grade	Grade Point	Remarks
90 - 100	A	4.0	Excellent
85 - <90	A-	3.7	
80 - <85	B+	3.3	
75 - <80	B	3.0	Good
70 - <75	B-	2.7	
65 - <70	C+	2.3	
60 - <65	C	2.0	Fair
57 - <60	C-	1.7	
55 - <57	D+	1.3	
52 - <53	D	1.0	Poor
50 - <52	D-	0.7	
<50	F	0.0	Failure

Table 3.7: Grading Policy of BRAC University

Collecting the students' CGPA is one of most significant aspects of our work.

### 3.9 Others

Attendance represents the attention and participation of a student in learning. The association of attendance with the academic performance of students is both positive and negative. We categorized the attendance percentage according to BRAC University's attendance policy and asked the students first semester's attendance percentage.

We know that, traffic jam is an enormous problem in our region. Because of traffic jam, Dhaka City's average traffic speed has fallen from 21 kmph to 7 kmph in the last 10 years and could drop to 4 kmph by 2035 which is very slow [23]. Typically, it is slower than the speed of walking. That is why we collected information about the commute time from home to university of a student.

We also gathered information about the study time of a student other than the university hours. We categorized it into 4 parts which starts with less than half an hour, incrementing half an hour on every category, which ends at more than two hours daily.

### 3.10 Data Analysis with Various Algorithms

The implementation of suitable machine learning algorithms to construct exact predictive models is therefore of vital importance for researchers [32]. After evaluating and considering the complexity of machine learning methods we chose appropriate algorithms for our research work or it would be hard to understand and describe the results obtained.

### **3.10.1 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. The regression analysis is wide used for prediction and estimating. Additionally, regression algorithms are accustomed infer casual relationships between the independent and subordinate variables. To analyze our dataset, we used two regression algorithms- Linear Regression and Decision Tree Regression.

### **3.10.2 Linear Regression**

For modeling the relationship between a dependent variable (scalar response) and one or more independent factors (explanatory factors), Linear Regression is used. It is called simple Linear Regression, because there is only one explanatory variable. On the other hand, if the explanatory variable is more than one than it is called multiple Linear Regression. In Linear Regression, the relationships are modeled using linear predictor models whose unknown model parameters have been estimated from data.

### **3.10.3 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. Eventually, a Decision Tree is developed which has decision nodes and leaf nodes. A decision node comprises more than one branch, which represents values for the attributes evaluated. Leaf nodes are used to represent decision about numeric goal [36].

### **3.10.4 Classification Algorithm**

The computer program learns from the dataset and uses the learning to classify new observation, it is called machine learning classification method. For machine learning there are seven classification algorithms; we use for our research work are Gaussian Naïve Bayes and Decision Tree Classification.

### **3.10.5 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. Compared to other more advanced machine learning algorithms, Gaussian Naïve Bayes is an extremely fast learner. Gaussian NB can solve the spam filtering and other real world issues. Its basic assumptions are that each predictor leads to the outcome independently and in equal terms.

### **3.10.6 Decision Tree Classifier**

Decision Tree Classification algorithm is used to construct classification into the arrangement of a tree structure. Unlike regression with DT, Decision Tree Classifi-

cation algorithm often split the dataset into as many small subsets as possible and builds a linked decision tree incrementally. Eventually, a decision tree is created which has decision nodes and leaf nodes. There is more than one branch in a decision node which represents values for the attributes evaluated. The leaf nodes are used to reflect classification or decision. There are multiple nodes and each node asks a different question, depending on the next node generated information and the very last node being the data expected. The sequence of the nodes is determined by the attribute of the information gain of each condition.

### **3.10.7 Chi-Square**

The algorithm Chi-Square is frequently 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 (otherwise called a bivariate table) to determine Independence test. Crosstabulation demonstrates at the same time the dimensions of two absolute factors, with the convergence of the groups of the factors occurring in the table cells [46].

### **3.10.8 Pearson Correlation Coefficient**

Coefficients of correlation are used in statistics to quantify how strong association between two variables is [40]. Pearson's Correlation Coefficient, by far the most common correlation measure, is a number between -1 and 1 which represents a propensity to have a linear association for two random phenomena [27]. Presumption made by Pearson Correlation is at least one attribute which must follow a normal distribution [40].

A positive correlation by Pearson shows that two random variables takes a significant value simultaneously. The negative Pearson Correlation is opposite to the positive Pearson Correlation. The association with a negative Pearson means that when one random variable takes a large value, the other takes a low value. [27]

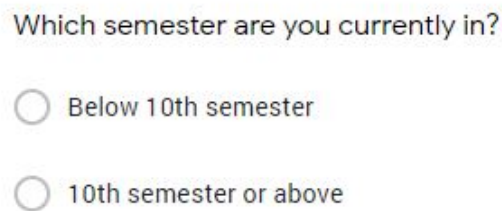
## **3.11 Confusion Matrix**

A Confusion Matrix defines a model's output or accuracy in a table form. Which makes the confusion matrix very simple to grasp, a good understanding of the outcome is also provided by the graphical representation of confusion matrix. Performance estimation with confusion matrix is helpful if there are in more than one output class. The accuracy of any machine learning algorithm can be correctly calculated using the confusion matrix.

# Chapter 4

## Questionnaire Description

We conducted a survey to collect data for our dataset. We gathered responses from BRAC University students where they answered a number of questions. Such responses helped us to figure out important factors that influence the grades of the students.



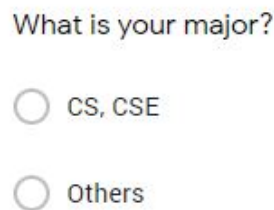
Which semester are you currently in?

Below 10th semester

10th semester or above

Figure 4.1

Figure 4.1 shows a question which asked participants about their semester. This is the primary question as we are considering the students who are in or above 10 semester for better predictive outcome.



What is your major?

CS, CSE

Others

Figure 4.2

Figure 4.2 shows a question where students were questioned about their major as the study is for the students doing their major in the CS/CSE field.



What is your gender?

- Male
- Female
- Other

Figure 4.3

Figure 4.3 shows the basic details we need for our study. It is one of the important questions of our study and has been included to identify whether grades vary with the gender.

What was your medium of study?

- Bangla Medium
- English Version
- English Medium

Figure 4.4

What was your type of School at Higher Secondary Level?

- Private
- Public

Figure 4.5

Figure 4.4 and Figure 4.5 asks the participants about their educational background. This is an important topic as students from diverse fields of educational medium may face various kinds of challenges at the university and thus have an impact on their grades.

What was your SSC GPA?

- 5.0
- $\geq 4.5$
- $\geq 4.0$
- $\geq 3.5$
- $\geq 3.0$

Figure 4.6

What was your HSC GPA?

- 5.0
- $\geq 4.5$
- $\geq 4.0$
- $\geq 3.5$
- $\geq 3.0$

Figure 4.7

Figure 4.6 and Figure 4.7 shows the question about students secondary and higher secondary school results. These questions were for students from the Bengali Medium background.

What was your average O'Level grade in five subjects? [A\*/ A= 5.0, B= 4.0, C=3.0, D=2.0]

- 5.0
- $\geq 4.5$
- $\geq 4.0$
- $\geq 3.5$
- $\geq 3.0$
- $\geq 2.5$

Figure 4.8

What was your average A'Level grade in two subjects?[A\*/ A= 5.0, B= 4.0, C=3.0, D=2.0]

- 5.0
- $\geq 4.5$
- $\geq 4.0$
- $\geq 3.5$
- $\geq 3.0$
- $\geq 2.5$

Figure 4.9

The above two questions in Figure 4.8 and Figure 4.9 were for students from the English medium background. A student with good GPA in SSC/HSC/O' Level/A' Level can be expected to perform well in the university courses or there might be no relationship between these two stages of academics whatsoever. The last 4 Figures (Figure 4.6, 4.7, 4.8 & 4.9) were added to study this relationship.

What is the monthly income of your family?

- More than 200000
- More than 100000
- More than 60000
- More than 30000
- Less than 30000

Figure 4.10

Figure 4.10 asked about the monthly family income of a student. Many parents cannot afford the expenses of education from Private Universities. As a result, the students contribute to these expenses by spending hours in part time jobs or tuition. This might possibly affect their grades as a consequence of studying fewer hours for the courses.

How many members do you have in your family?

Short answer text

---

Figure 4.11

Figure 4.11 asked about the number of members in the family of a student. In Bangladesh there are two types of family- Nuclear family and joint family. Joint families have more members than the nuclear ones. Also, some students have more siblings than others in spite of having a nuclear family. With increase in family members, the studying environment in the household changes and thus can affect grades.

Do you have any chronic medical condition that hampers you day to day life?

- YES
- NO

Figure 4.12

A student with chronic medical condition may not be able to give their best effort while completing a course. So, Figure 4.12 of the questionnaire considered the medical condition factor.

What was your attendance percentage in first semester?

- 90%
- 85%
- 80%
- 75%
- 70%
- Less than 70%

Figure 4.13

Figure 4.13 asked the attendance percentage in the first semester. Attending classes regularly is important for understanding the topics of the courses and so it is one of the most important factors in achieving good grades.

How many hours do you study each day? (in hours)

- More than 2.0
- More than 1.0
- More than 0.5
- Less than 0.5

Figure 4.14

Academic performance depends significantly on the effort and time spent studying for the topics. Students who spend more hours studying each day tend to do better in exams. In Figure 4.14, students were asked about the number of hours they study each day.

How much time do you need to come to University? (In Minutes)

- More than 120
- More than 60
- More than 30
- More than 15
- Less than 15

Figure 4.15

Figure 4.15 asked about the time taken for the commute to university. Some students live far from the campus and may have to leave early for classes.

What was your CGPA after 1st semester?

Short answer text

---

Figure 4.16

What is your CGPA after 9th semester?

Short answer text

---

Figure 4.17

The questions in Figure 4.16 and Figure 4.17 have been asked to monitor the change in CGPA of a student along the graduation period. Right after joining university, a student might need some time to adapt to the new system. In our research, in many cases it was seen that CGPA after first and ninth semester varied drastically which implies that students succeed to increase their CGPA after a period of 1 or 2 years.

All the questions that were asked might have both direct and indirect impact on students' grades. After analysis of these factors we will be able to propose a better prediction model to identify the students at risk. Detailed description of the factors and their relationship with students' academic performance will be described in the next section.

# Chapter 5

## Proposed Methodology

### 5.1 Workflow

We consider the most important aspect of a study in our short period of research work is the workflow. Figure 5.1 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. The first phase in our workflow was data- selection. The data was then pre-processed. Next, the pre-processed data were used to apply ML algorithm and to extract important features from it. For ranking feature importance, feature selection methods has been used. Now, there were two target labels (Predict, Predict\_discrete) we wanted to predict. We applied regression algorithms on target label “Predict”. Whereas, classifier algorithms were implemented on target label “Predict\_discrete” which is a discretized version of the original target label. Lastly, a comparison was made with the results obtained to see whether the algorithms would correctly predict CGPA for students.

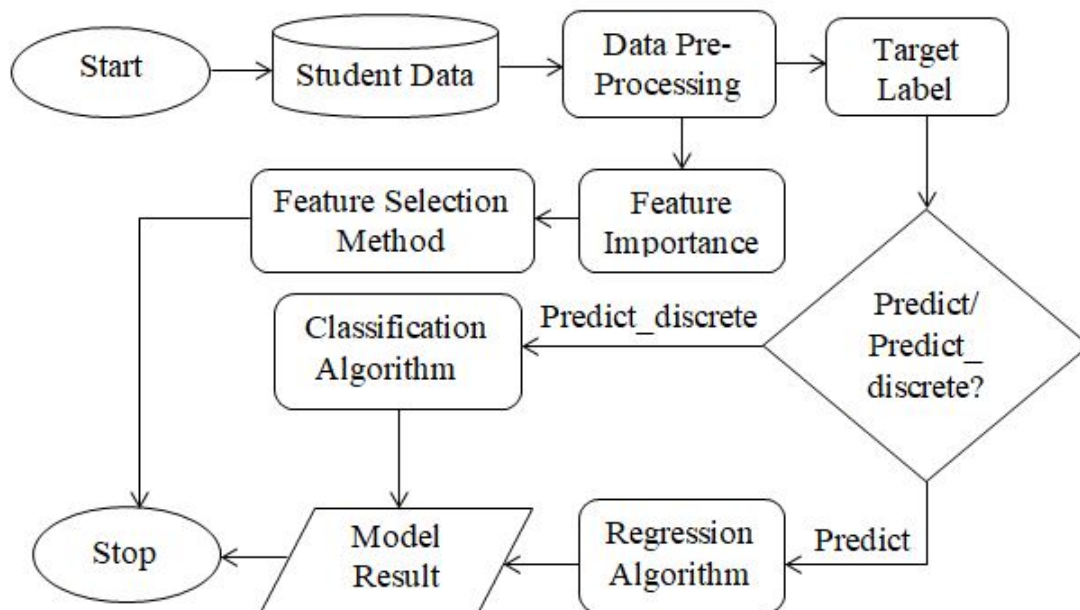


Figure 5.1: Workflow of the proposed method for predicting students' grade

## 5.2 Attribute and Dataset Description

The data collection used in this work is obtained from BRAC University's Department of Computer Science and Engineering. The data were acquired from the following source, a student survey. 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. The dataset was then split into train set and test set. The proportion of the train and test set was 80% and 20% respectively.

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 who are 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 [9]. Also, medium of study and type of school in higher secondary level play a crucial role in students' academic success over the years; students from the private and English medium background have generally been seen doing better in academics. Whereas, it may be difficult for public and Bengali medium students to understand classes because the language of instruction is mainly English in 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 [9]. A significant factor is the outcome of 1st semester graduates, representing their result in the next semester. 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 5.1 lists the attributes and their scope used to construct the dataset.



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, $\geq 4.5$ , $\geq 4.0$ , $\geq 3.5$ , $\geq 3.0$
6	HSC_GPA	Student's Grade point average of HSC examination	5.0, $\geq 4.5$ , $\geq 4.0$ , $\geq 3.5$ , $\geq 3.0$
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<=60000 Taka, 60000 Taka <F_income<=100000 Taka, 100000 Taka<F_income<=200000 Taka, 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 5.1: Student attributes and their domains

The following Figure 5.2 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.

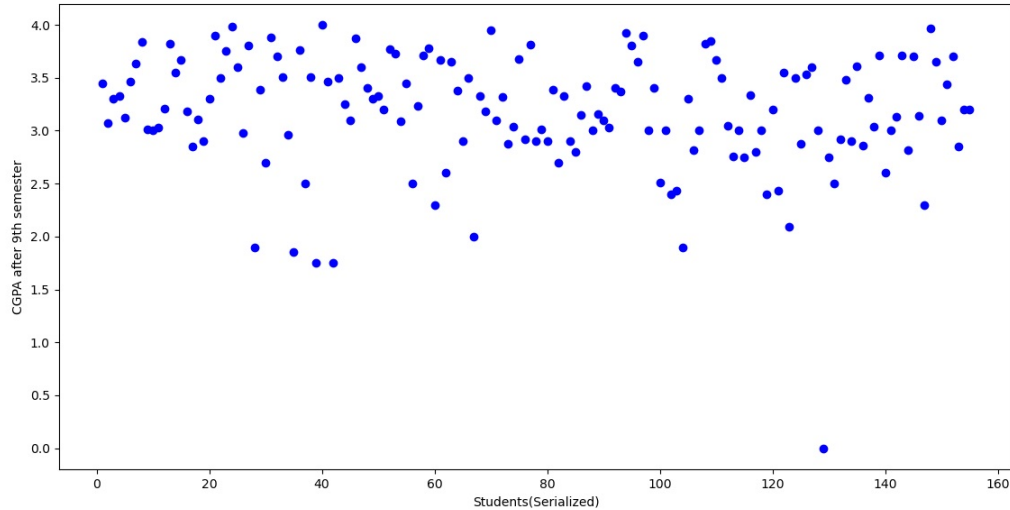


Figure 5.2: An overview of all students CGPA after 9th semester from the dataset

### 5.3 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. The Table 5.2 below shows which attributes domains converted into which discrete variables.

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

Table 5.2: Table of converting attributes domains into discrete variables

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 5.3 shows the category of student based on their CGPA.

Category	Condition (CGPA based)
Bad	$CGPA < 3$
Medium	$CGPA < 3.5$
Good	$3.5 \leq CGPA \leq 4$

Table 5.3: Students' category, based on their CGPA

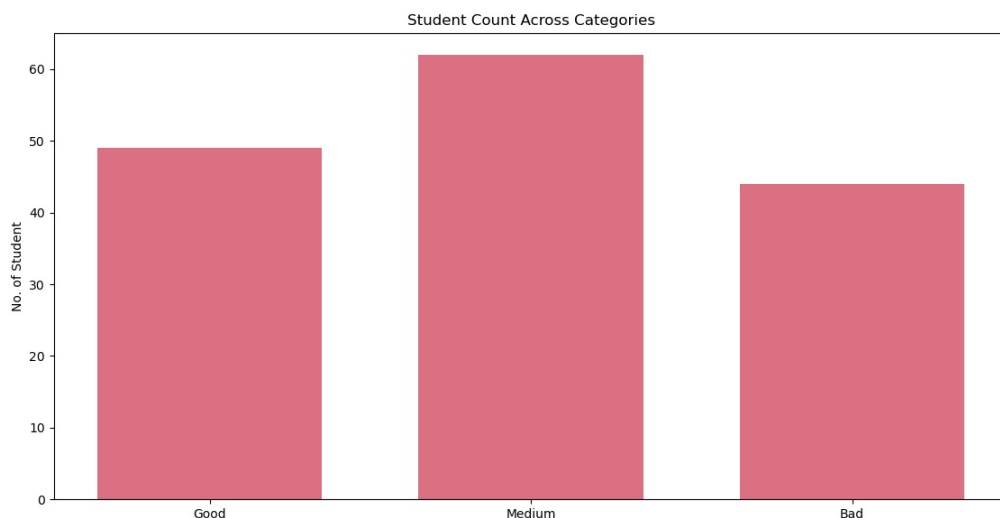


Figure 5.3: Student count across categories

In Figure 5.3 we showed visually the total number of students in the categories listed.

The categorized target label has already shown in the Table 5.3. So, there are two target labels we need to predict. They are:

1. Predict- the Actual CGPA and
2. Predict\_discrete- CGPA Categories.

## 5.4 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 [9]. If the classifier's accuracy is above an appropriate limit [6] 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.

## 5.5 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 (5.1) [22] below.

$$y = a_0 + a_1 * x \quad (5.1)$$

The linear regression algorithm has the concept of determining the best values for a  $a_0$  and  $a_1$ . Scatter plot for Linear Regression for our dataset is represented in Figure 5.4.

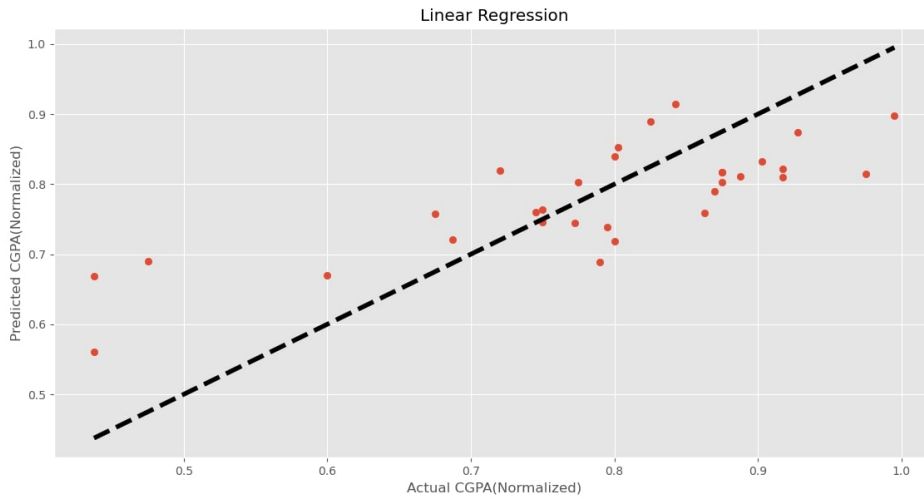


Figure 5.4: Scatter plot for Linear Regression

In Figure 5.4 we showed a Scatter plot for Linear Regression on student grade prediction using profile data. Where the x-axis was labelled with the normalized Actual CGPA and y-axis shows Predicted CGPA (Normalized) which is target label “Predict\_discrete” for our study.

Figure 5.5 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, we labelled Actual CGPA with blue color and Predicted CGPA with green color.

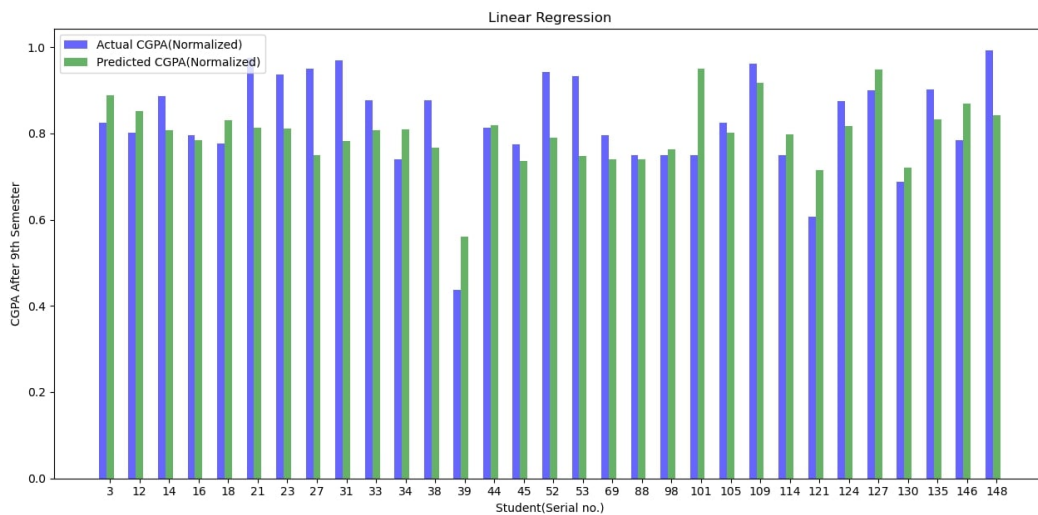


Figure 5.5: Bar Chart of Linear Regression

## 5.6 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 [17], which can be calculated using the Equations (5.2) and (5.3).

$$Entropy(T, X) = \sum_{c \in X} P(c)E(c) \quad (5.2)$$

$$InformationGain(T, X) = Entropy(T) - Entropy(T, X) \quad (5.3)$$

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

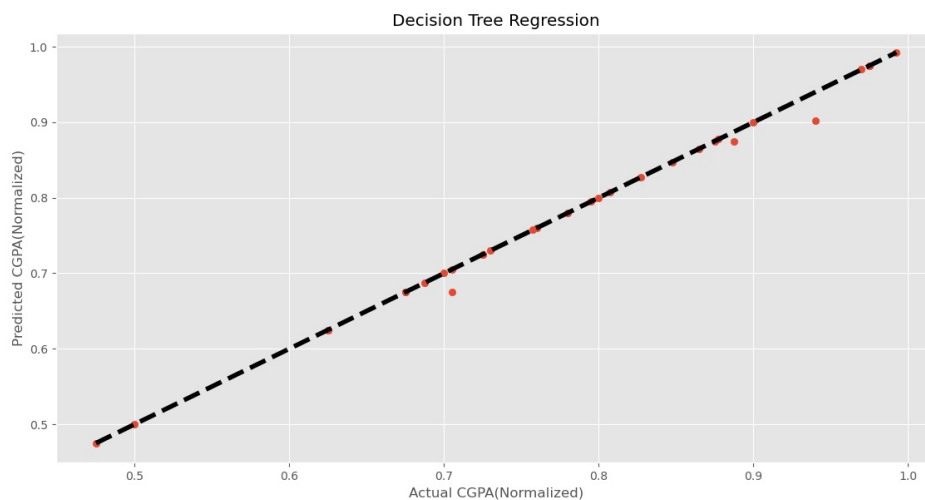


Figure 5.6: Scatter plot for Decision Tree Regression

In Figure 5.6 we showed a Scatter plot for Decision Tree Regression. Where the x-axis was labelled with the normalized Actual CGPA and y-axis shows Predicted CGPA (Normalized).

Figure 5.7 shows the Bar Chart for Decision Tree Regression algorithm where horizontal axis shows the Student (Serial no.) and vertical axis shows students' CGPA after 9th semester. Here, we labelled Actual CGPA with blue color and Predicted CGPA with green color.

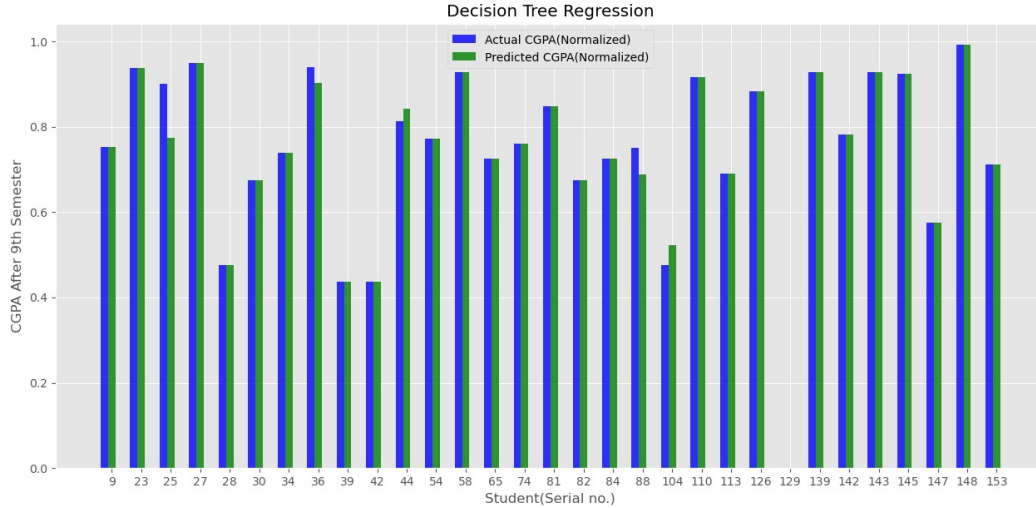


Figure 5.7: Bar Chart of Decision Tree Regression

## 5.7 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 (5.4) [41] indicates the probability of an occurrence considering the probability of another already existing occurrence.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (5.4)$$

Where, A is the event we will find the probability of and B is the event that has already taken place [41]. Scatter plot for Gaussian Naïve Bayes classifier for our dataset is represented in Figure 5.8.



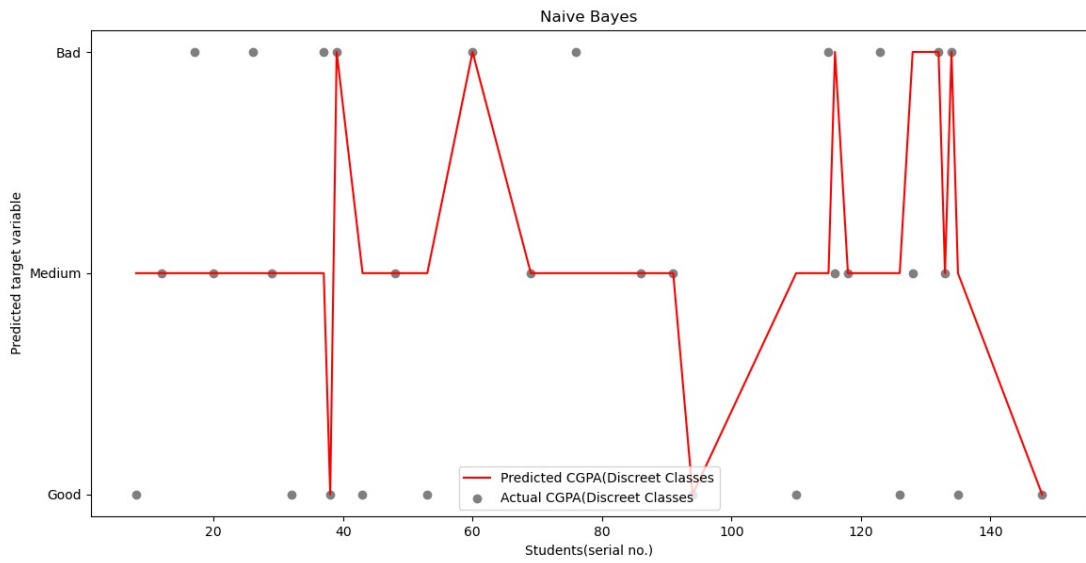


Figure 5.8: Scatter plot for Gaussian Naïve Bayes

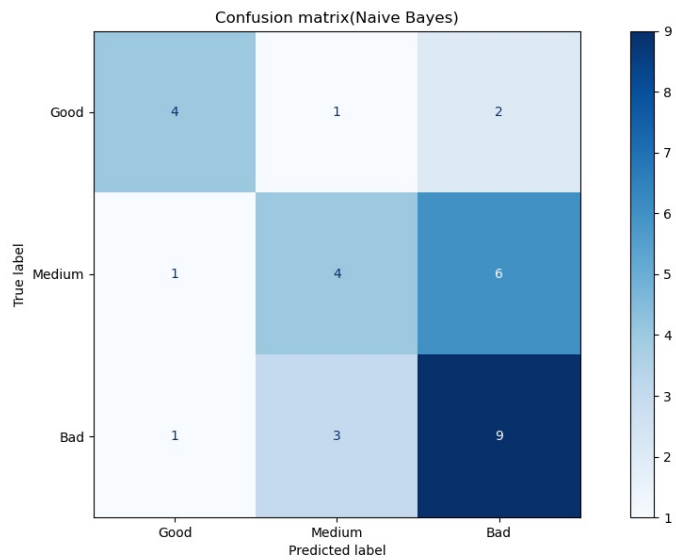


Figure 5.9: Confusion Matrix for Gaussian Naïve Bayes

In Figure 5.8 we display a Gaussian Naïve Bayes Scatter plot on prediction of the student grade using profile data. Where the x-axis has been labelled with Students (Serial no.) and y-axis, has been labelled with Predicted Target Variable. The Figure 5.9 shows the Confusion Matrix for Gaussian Naïve Bayes Classifier algorithm for the target label.

## 5.8 Decision Tree Classification

Decision tree Classifier is the method of predictive analysis used in supervised machine learning. Information gain is the most popular attribute selection approach for Decision Tree. Information gain can be calculated using the following Equation (5.5) [39].

$$Info(D) = - \sum_{i=1}^m P_i \log_2 P_i \quad (5.5)$$

Where, pi is the probability which gives us the idea of how uncertain we are about the data. The following Figure 5.10 shows the scatter plot and Figure 5.11 shows the confusion matrix of Decision Tree Classifier.

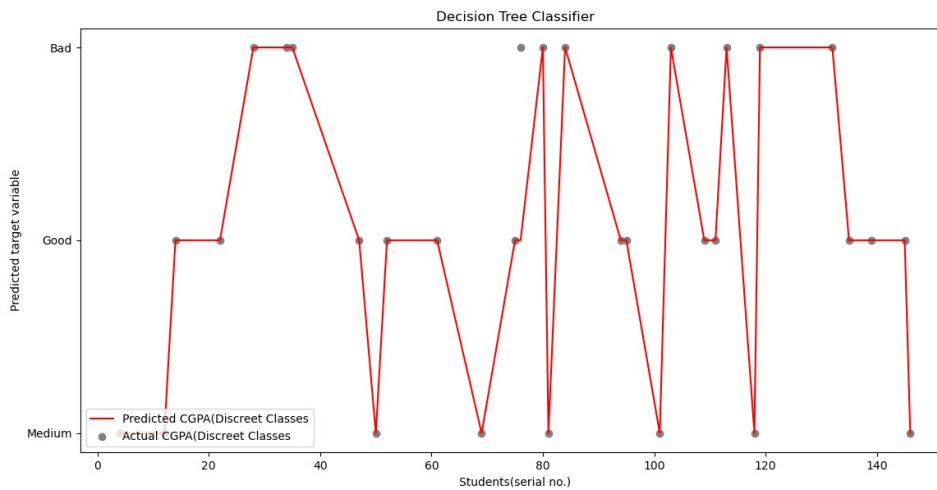


Figure 5.10: Scatter Plot for Decision Tree Classifier

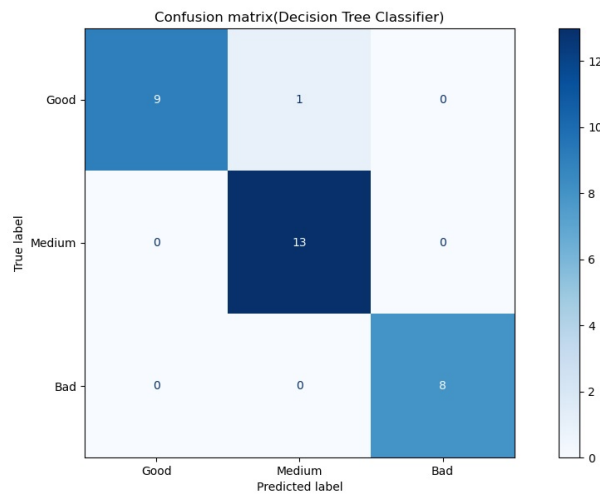


Figure 5.11: Confusion Matrix of Decision Tree Classifier

## 5.9 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 5.4 shows the ranking result.

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

Table 5.4: Important features with Rank values

From the Table 5.4 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.

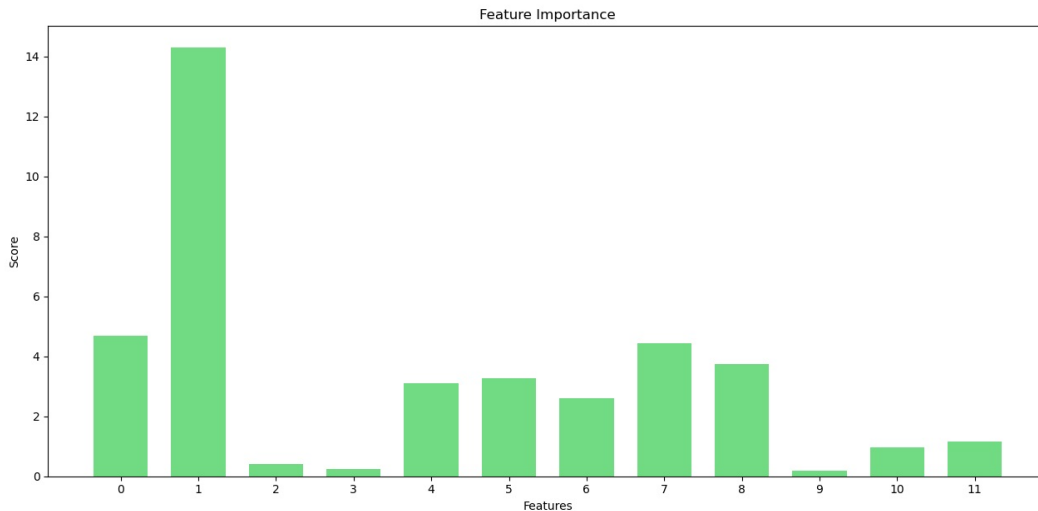


Figure 5.12: Bar Chart of Feature Importance using Chi-Square

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 Higer Secondary Level and
- 11: CGPA after 1st semester.

## 5.10 Pearson Correlation Coefficient

Pearson correlation highlight a one-dimensional relation between two variables [37]. Pearson correlation coefficient can measure the dependency between two random variables, thus it is one of the popular algorithm in machine learning. To calculate the feature importance, we also applied Pearson correlation on our data. The result we find by Pearson correlation as seen in Figure 5.13.

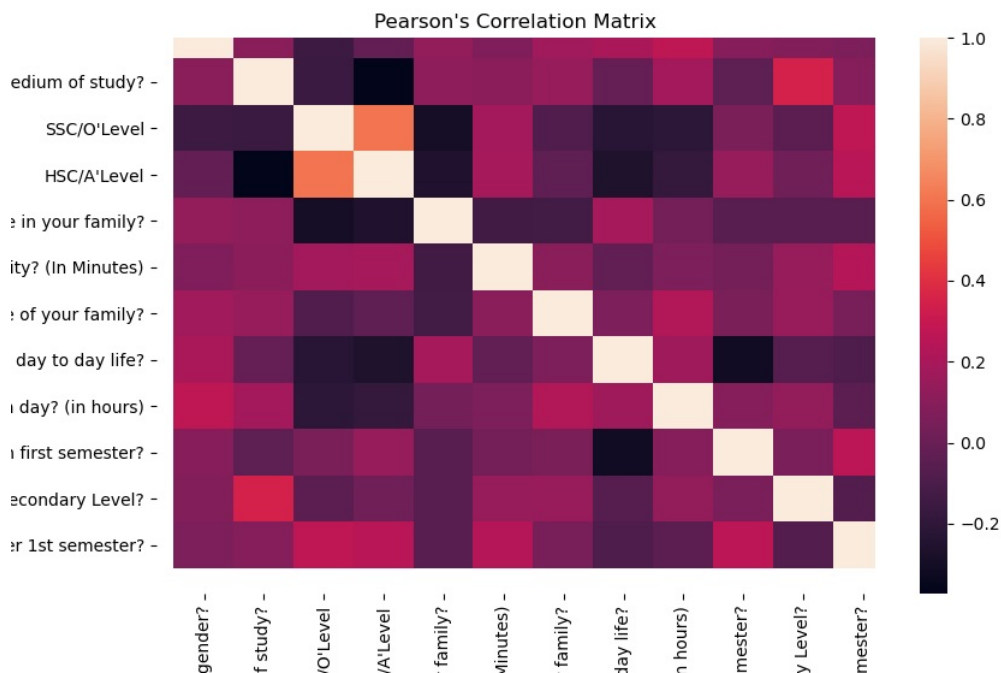


Figure 5.13: Feature Importance calculation using Pearson Correlation Matrix

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.

# Chapter 6

## Result Analysis

### 6.1 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.

Table 6.1 shows the of Accuracy for different classification and regression algorithm we used on different student data. The classification algorithms we used are Gaussian NB and DT Classification. Where, Decision Tree Classifier (DT Classifier) offers the highest result in classification with a precision varies within 95 to 100%. Contrarily, LR and DT 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%.

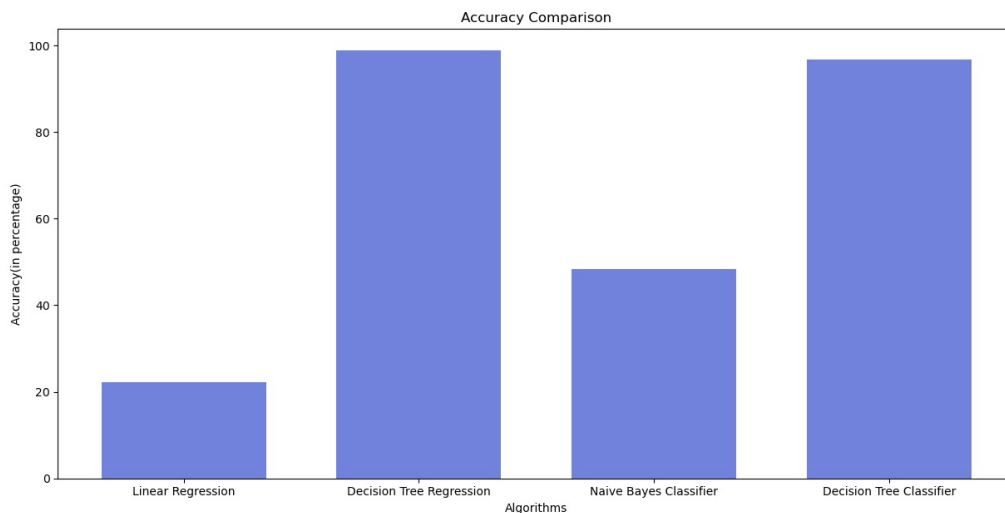


Figure 6.1: Bar Chart Showing Accuracy of Four Different Algorithms

Model	Accuracy
DT Regression	90% - 100%
LR	20% - 50%
Gaussian NB	50% - 60%
DT Classifier	95% - 100%

Table 6.1: Accuracy of different classification and Regression Algorithm

## 6.2 Feature Analysis

The most important feature in determining students' performance is Medium of study. Which is obtained by applying Chi-Square based feature selection method. Figure 6.2 shows the comparison result between students from different mediums gained by implementing Chi-Square algorithm on the dataset. Where we labelled Bangla Medium with blue color, English Version with Green color and English Medium with Red color.

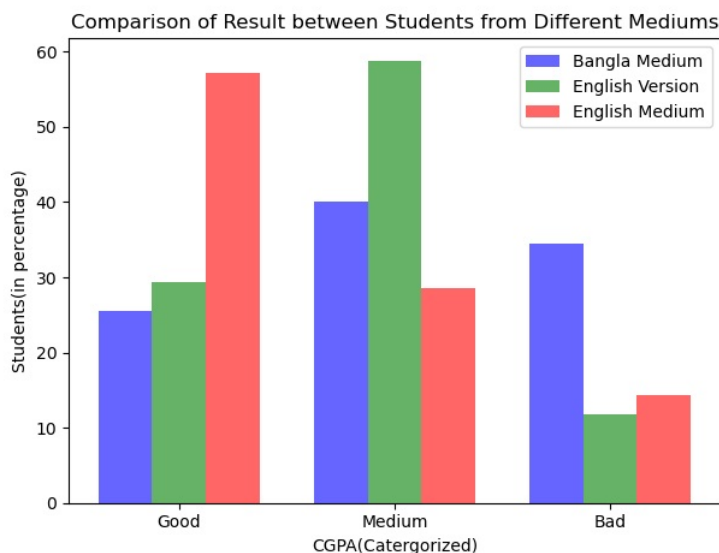


Figure 6.2: Comparison of Result between Students from Different Mediums

Figure 6.2 summed 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.

## 6.3 Comparison with Previous Works

Previously there has been many research works in the field of Educational Data Mining. In Table 6.2 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.

Parameters	Kumar et al. [31]	Altabrawee et al. [29]	Tekin [10]	Al-Sudani et al. [28]	Junejo et. al. [14]	Present Study
Features	Personal, demographic information and students' usage of social media	Grades, Personal information and student satisfaction	Scores of vocational and cultural courses	Academic, institutional, demographic, psychological and financial information	Sessional grades	Gender, personal, educational and family related information
Sample size	300	161	127	481	2500	155
Machine Learning Algorithm used	ID3 and J48	ANN, NB, LR and DT	NN, SVM and ELM	NN, DT, KNN and SVM	ID3, KNN, NB and Rule Induction	DT Classifier, Gaussian NB, DT Regression and LR
Best Algorithm	ID3 and J48	ANN	SVM	NN	Rule Induction	DT Classifier and DT Regression
Accuracy of prediction	62.667%	77.04%	97.98%	83.7%	73-96%	95-100% for DT Classifier and 90-100% for DT Regression
Metrics	Accuracy	ROC index	RMS, COV and $R^2$	Accuracy	Accuracy, precision and recall	Accuracy

Table 6.2: Comparison with Previous Works on Student Grade Prediction

# Chapter 7

## Concluding Remarks

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 which included Linear Regression, Decision Tree Regression, Gaussian Naive Bayes and Decision Tree Classifier Algorithm. Among these Decision Tree Classifier consistently performed better than others where accuracy varied between 95-100%. 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.



# Bibliography

- [1] *Bangladesh: Hies: Average household size: Economic indicators: Ceic*, 1970. [Online]. Available: <https://www.ceicdata.com/en/bangladesh/household-income-and-expenditure-survey-number-of-household-by-size/hies-average-household-size>.
- [2] *Bangladesh household income per capita [2000 - 2020] [data charts]*, 1970. [Online]. Available: <https://www.ceicdata.com/en/indicator/bangladesh/annual-household-income-per-capita>.
- [3] A. Díaz, “Personal, family, and academic factors affecting low achievement in secondary school”, *Electronic Journal of Research in Educational Psychology and Psycho Pedagogy*, vol. 1, Apr. 2003. [Online]. Available: [https://www.researchgate.net/publication/255581992\\_Personal\\_family\\_and\\_academic\\_factors\\_affecting\\_low\\_achievement\\_in\\_secondary\\_school](https://www.researchgate.net/publication/255581992_Personal_family_and_academic_factors_affecting_low_achievement_in_secondary_school).
- [4] N. Harb and A. El-Shaarawi, “Factors affecting students’ performance”, *Journal of Business Education*, vol. 82, no. 5, pp. 282–290, 2007. [Online]. Available: <https://mpira.ub.uni-muenchen.de/id/eprint/13621>.
- [5] A. Raychaudhuri, M. Debnath, S. Sen, and B. G. Majumder, “Factors affecting students’ academic performance: A case study in agartala municipal council area”, 2010.
- [6] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. 2011, ISBN: 9780123814791.
- [7] I. Mushtaq and S. N. Khan, “Factors affecting students’ academic performance”, *Global Journal of Management And Business Research*, vol. 12, no. 9, 2012, ISSN: 2249-4588. [Online]. Available: <https://journalofbusiness.org/index.php/GJMBR/article/view/721>.
- [8] A. A. Aziz, N. H. Ismail, and F. Ahmad, “Mining students’ academic performance”, *Journal of Theoretical and Applied Information Technology*, vol. 53, no. 3, 2013, ISSN: 1992-8645. [Online]. Available: [https://www.researchgate.net/publication/258124336\\_Mining\\_students'\\_academic\\_performance](https://www.researchgate.net/publication/258124336_Mining_students'_academic_performance).
- [9] A. Acharya and D. Sinha, “Early prediction of students performance using machine learning techniques”, *International Journal of Computer Applications*, vol. 107, no. 1, 2014, ISSN: 0975 – 8887.
- [10] A. Tekin, “Early prediction of students’ grade point averages at graduation: A data mining approach”, *Eurasian Journal of Educational Research*, pp. 207–226, 54 2014.

- [11] 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, ISSN: 2035-4983.
- [12] A. Algarni, “Data mining in education”, *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 6, 2016. DOI: 10.14569/IJACSA.2016.070659.
- [13] 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, ISSN: 2349 5359.
- [14] K. Junejo and E. Eman, “Grade prediction using supervised machine learning techniques”, Mar. 2016.
- [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”, vol. 103, pp. 1–15, 2016, *Computers & Education*. DOI: <https://doi.org/10.1016/j.compedu.2016.09.005>.
- [16] 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.
- [17] R. Jain, *Decision tree. it begins here*. 2017. [Online]. Available: [https://medium.com/@rishabhjain\\_22692/decision-trees-it-begins-here-93ff54ef134](https://medium.com/@rishabhjain_22692/decision-trees-it-begins-here-93ff54ef134).
- [18] M. Pojon, “Using machine learning to predict student performance”, M.S. thesis, Faculty of Natural Sciences, Software Development, University of Tampere, 2017.
- [19] S. Rovira, E. Puertas, and L. Igual, “Data-driven system to predict academic grades and dropout”, *PLoS ONE*, vol. 12, no. 2, 2017. DOI: 10.1371/journal.pone.0171207.
- [20] M. Zaffar, M. A. Hashmani, and K. S. Savita, “Performance analysis of feature selection algorithm for educational data mining”, in *2017 IEEE Conference on Big Data and Analytics (ICBDA)*, 2017, pp. 7–12. DOI: <https://doi.org/10.1109/icbdaa.2017.8284099>.
- [21] S. Boonman, “Early prediction in students’ performance in a distance learning university”, 2018, [Online]. Available: <http://arno.uvt.nl/show.cgi?fid=147602>.
- [22] R. Gandhi, *Introduction to machine learning algorithms: Linear regression*, 2018. [Online]. Available: <https://towardsdatascience.com/introduction-to-machine-learning-algorithms-linear-regression-14c4e325882a>.
- [23] A. A. Haider, *Traffic jam: The ugly side of dhaka’s development*, 2018. [Online]. Available: <https://www.thedailystar.net/opinion/society/traffic-jam-the-ugly-side-dhakas-development-1575355>.
- [24] I. E. Livieris, K. Drakopoulou, T. A. Mikropoulos, V. Tampakas, and P. Pintelas, “An ensemble-based semi-supervised approach for predicting students’ performance”, *Springer International Publishing AG, part of Springer Nature, Research e-Learn. ICT Education*, 2018. [Online]. Available: <https://doi.org/10.1177/0735633117752614>.

- [25] A. Oliver, A. Odena, C. Raffel, E. D. Cubuk, and I. J. Goodfellow, “Realistic evaluation of deep semi-supervised learning algorithms”, 32nd Conference on Neural Information Processing Systems (NeurIPS), Montréal, Canada, 2018. [Online]. Available: <https://papers.nips.cc/paper/7585-realistic-evaluation-of-deep-semi-supervised-learning-algorithms.pdf>.
- [26] Y Reddy, V. Pulabaigari, and E. B., “Semi-supervised learning: A brief review”, *International Journal of Engineering Technology*, vol. 7, p. 81, Feb. 2018. DOI: 10.14419/ijet.v7i1.8.9977.
- [27] Y.-L. K. Samo, *The black swans in your market neutral portfolios (part i)*, 2018. [Online]. Available: <https://towardsdatascience.com/the-black-swans-in-your-market-neutral-portfolios-part-i-7521683a7317>.
- [28] S. Al-Sudani and R. Palaniappan, “Predicting students’ final degree classification using an extended profile”, *CrossMark, Education and Information Technologies*, 2019. DOI: <https://doi.org/10.1007/s10639-019-09873-8>.
- [29] 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. DOI: <https://doi.org/10.29196/jubpas.v27i1.2108>.
- [30] M. Imran, S. Latif, D. Mehmood, and M. S. Shah, “Student academic performance prediction using supervised learning techniques”, *International Journal of Emerging Technologies in Learning*, vol. 14, no. 14, pp. 92–104, 2019, ISSN: 1863-0383. DOI: <http://dx.doi.org/10.3991/ijet.v14i14.10310>.
- [31] 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, ISSN: 2277-3878. [Online]. Available: <https://www.ijrte.org/wp-content/uploads/papers/v8i4/D8241118419.pdf>.
- [32] M. Tsiakmaki, G. Kostopoulos, S. Kotsiantis, and O. Ragos, “Implementing automl in educational data mining for prediction tasks”, 2019, 10(1),90. DOI: <https://doi.org/10.3390/app10010090>.
- [33] *Grading system in bangladesh - gpa cgpa school college university*, 2020. [Online]. Available: <https://www.mediabangladesh.net/grading-system-bangladesh/>.
- [34] S. Jamal, *Nctb book download. nctb all book download*, 2020. [Online]. Available: <https://teachingbd24.com/nctb-book-download/>.
- [35] Wikipedia contributors, *Academic major — Wikipedia, the free encyclopedia*, [Online; accessed 3-April-2020], 2020. [Online]. Available: [https://en.wikipedia.org/w/index.php?title=Academic\\_major&oldid=945112914](https://en.wikipedia.org/w/index.php?title=Academic_major&oldid=945112914).
- [36] [Online]. Available: <https://saedsayad.com/decision.tree.reg.htm>.
- [37] *Correlation coefficient: Simple definition, formula, easy calculation steps*. [Online]. Available: <https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/>.
- [38] *Curriculum*. [Online]. Available: <https://hurdcointschool.com/curriculum/>.
- [39] *Decision tree classification in python*. [Online]. Available: <https://www.datacamp.com/community/tutorials/decision-tree-classification-python>.

- [40] *Home*. [Online]. Available: <http://sites.utexas.edu/sos/guided/inferential/numeric/bivariate/cor/>.
- [41] *Naive bayes classification using scikit-learn*. [Online]. Available: <https://www.datacamp.com/community/tutorials/naive-bayes-scikit-learn>.
- [42] *O & a levels*. [Online]. Available: <http://orangebangladesh.co.uk/o-levels/>.
- [43] *Policies and procedures: Brac university*. [Online]. Available: <https://www.bracu.ac.bd/academics/policies-and-procedures>.
- [44] *Simple linear regression and correlation*. [Online]. Available: [https://www.statsdirect.com/help/regression\\_and\\_correlation/simple\\_linear.htm](https://www.statsdirect.com/help/regression_and_correlation/simple_linear.htm).
- [45] N. Thai-Nghe, L. Drumond, and T. H. L. Schmidt-Thieme, “Multi-relational factorization models for predicting student performance”.
- [46] *Using chi-square statistic in research*. [Online]. Available: <https://www.statisticssolutions.com/using-chi-square-statistic-in-research/>.
- [47] *What is family income? definition and meaning*. [Online]. Available: <http://www.businessdictionary.com/definition/family-income.html>.