

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/369212801>

# Secondary School Result Prediction using Machine Learning Regressors and Primary-middle School Data

**Article** in Indonesian Journal of Electrical Engineering and Informatics (IJEI) · February 2023

DOI: 10.52549/ijeiei.v11i1.4302

CITATIONS

0

READS

12

4 authors, including:



**Jia Uddin**

Woosong University

154 PUBLICATIONS 866 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



A new machine learning approach to select adaptive IMFs of EMD [View project](#)



Suspicious Movement Detection in crowded areas [View project](#)

# Machine Learning based Stream Selection of Secondary School Students in Bangladesh

Shabbir Ahmad<sup>1</sup>, Md. Golam Rabiul Alam<sup>1</sup>, Jia Uddin<sup>2,\*</sup>, Md Roman Bhuiyan<sup>3</sup>, Tasnim Sakib Apon<sup>1</sup>

<sup>1</sup>Department of Computer Science and Engineering, Brac University, Bangladesh

<sup>2</sup>AI and Big Data Department, Woosong University, South Korea

<sup>3</sup>Department of Computing, Multimedia University, Malaysia

## Article Info

### Article history:

Received Nov 9, 2022

Revised Dec 13, 2022

Accepted Jan 28, 2023

### Keyword:

Regression analysis

Local interpretable model

agnostic explanations

Stream recommendation system

Bangladeshi secondary school

## ABSTRACT

In the Bangladeshi education system, there are three stages up to the secondary school certificate (SSC)- the primary (Primary Education Completion Certificate, or PEC), middle school (Junior School Certificate, or JSC), and SSC. A separate stream has to be chosen after the eighth grade, which could be any of the following streams: Science, Business Studies, and Humanities. The selection of a stream is very important for their future higher studies and career planning. Usually, students take the decision of selecting a stream based on PSC and JSC results only. To address this challenge, we have collected a dataset from different Bangladeshi schools, which consists of PSC and JSC students' records. There are 26 data for each student record including subject-wise student results, parent's academic qualification, parent's profession, parent's monthly income, sibling information, district, etc. In the experimental analysis, a series of machine learning regression algorithms have been utilized. Moreover, we have employed various performance metrics in order to validate our model's performance. The experimental results demonstrate that among the regressors, extreme gradient boosting algorithm's performance were superior in both science and humanities streams. In the business stream however, Support Vector Machine's performance is considerably better. It is expected that the analysis will help prospective students and stakeholders in their future decisions. Moreover, we have utilized Local Interpretable Model Agnostic Explanations that helps to increase the interpretability of the model.

Copyright © 2023 Institute of Advanced Engineering and Science.

All rights reserved.

## Corresponding Author:

Jia Uddin, Ph.D.

AI and Big Data Department

Endicott College

Woosong University, Daejeon, South Korea

jia.uddin@wsu.ac.kr

## 1. INTRODUCTION

In educational settings, the capacity to forecast a student's performance is critical. Students' academic success is influenced by a wide range of characteristics, including personal, social, psychological, etc. Machine learning appears to be a promising approach to achieving this goal. The machine learning approach makes patterns and links in massive amounts of data possible. There is a lot of information packed into a single piece of data. The data processing method is determined by the type of information the data generates. In the education industry, there is a great deal of data that may be used to provide useful information. This information and communication technology (ICT) aids the educational sector in collecting and compiling low-cost data. The amount of data saved in educational databases has grown considerably in recent years. Higher education institutions are motivating them to get to know their devoted students better. The best method is to properly maintain and process student databases [1].

In Bangladesh, the 9th-grade secondary school system is organized into three main categories. They are a group of science, business, and humanities students. The subjects of physics, chemistry, biology, and higher mathematics are the main priorities of the science stream. Business studies emphasize Accounting, Finance, and Banking, Business Entrepreneurship, Arts and crafts, Agriculture studies, whereas the Humanities stream studies Sociology, Geography, History, civics, Economics, Arts and crafts, and Agriculture studies, among other subjects [2]. However, a few subjects- Bangla, English, General Math, ICT, and Religious Studies—were generally applicable to all groups [3]. Choosing a group among Science, Business Studies, and Humanities in the 9th grade of secondary school is the most essential and crucial decision a student has to make. Stream selection is an important factor that affects a student's educational and personal life since there's a high possibility of dropping out if they can't cope up with their chosen stream's pressure [4]. There are many reasons why students drop out of secondary school in Bangladesh, including their perceptions of education, their prior employment history, their sociodemographic status (SDS), the size of their family, the number of siblings they have, a lack of food, the distance to their school, and bullying from other students or teachers [5]. Other factors that contribute to secondary school dropouts in Bangladesh include poor physical health; biased social norms; inadequate educational standards; economic hardship, geographic isolation, parental education, and family factors, unchecked population growth; unequal access to educational opportunities; early marriage and pregnancy of school-age girls; migration as a factor in school dropouts; relationship-related effects; and insecurity [6]. In our study, we generated a dataset keeping in mind the aforementioned reasons for dropouts. Students must be informed why choosing the incorrect separate stream would put them in danger in the future.

Students' difficulty in making decisions in secondary school is often the result of a lack of understanding in this area. In this decision-making process, it is seen that parents or teachers make decisions according to what they understand. A correct decision can be made by using a machine learning algorithm in stream selection to ensure the student's future. To solve the problem, the key contributions of our paper are as follows:

- As far as we have studied, there is no study on machine learning regression based stream selection on secondary school education in Bangladesh. Therefore, we have proposed a machine learning based stream selection of secondary school students in Bangladesh.
- We have applied a series of regression algorithms to predict individual students GPA for each stream and proposed the most appropriate stream for ninth-grade students. Among the regression methods, the extreme gradient boosting regressor shows higher accuracy than the other state-of-art algorithms.
- A dataset has been collected from the students of eleventh to twelfth grade or higher on which the proposed model has been built to infer the perfect stream for ninth-grade students. We made the dataset [29] public for reproducible research.
- Furthermore, we have utilized Local Interpretable Model Agnostic Explanations (LIME) as an explainable AI (XAI) that introduces interpretability to our proposed model.

## 2. RELATED WORKS

Nowadays, students' performance predictions are made on a broader scale using machine learning algorithms. Acharya et al. [7] described a machine learning issue in students' choice of universities. They contrasted various regression algorithms [30], including support vector, random forest, and linear regression. For a small dataset, linear regression performed better with a low MSE and a high R2 score. The results showed whether the chosen university was an ambitious or safe choice. In the paper, the author wanted to create more diverse profiles of students to enhance the size of their small dataset.

El Aissaoui et al. [8] put forth a multiple Linear Regression approach for creating a model that predicted student performance. The one produced utilizing the 'MARS' method is the most effective. In order to determine the elements that affect Moroccan university applicants' success on admission tests, the author would have preferred to have used a dataset that captures the characteristics of Moroccan university applicants.

Zulfiker et al. [9] discussed the students who were accepted each year into various universities in Bangladesh. They can improve their grades by taking the necessary action and forecasting their results before the final exam. Seven different machine learning techniques (Support Vector Machine [31], K-Nearest Neighbor (KNN), Logistic Regression, Decision Tree, AdaBoost, Multilayer Perceptron, and Extra Tree Classifier) were used in this study to forecast the students' final grades. This study achieved 81.72% accuracy, and the weighted voting classifier showed the best performance for classifying data. This research was conducted utilizing data from a single private university in Bangladesh. The author wants to expand their dataset by gathering information from more public and private universities. Besides this, for preprocessing, the author utilized discretization methods and oversampling techniques like Synthetic Minority Over-sampling Technique (SMOTE).

Hasan et al. [10] applied Naive Bayes, Sequential Minimal Optimization (SMO), and the Random Forest algorithm to help 9th-grade students choose the correct group (Science, Business, and Humanities) for their higher studies. The authors used random features in this paper from those higher-class students who had already gone through this process. This research achieved 84.9% accuracy for the Random Forest algorithm. The author used data from Bangladesh, and in the future, the author will integrate learning methodologies with database management systems and e-learning platforms on various international datasets to identify far better traits and factors as a result, which will improve the accuracy of the system's predictions.

Shahadat et al. [11] used Bayes-based, function-based, lazy-based, rule-based, and tree-based classifiers to remove irrelevant features to predict Higher Secondary Certificate examination results. This study found LMT performed best and only ten features needed to be emphasized to get a good result in HSC. The author claimed preparing or filtering data can enhance the proposed system's performance.

Ahammad et al. [12] predicted students' performance using the proposed model, which worked over students' Secondary School Certificate examination results. The authors conducted a comparative study among Naive Bayes, K-nearest Neighbors, Support Vector Machine, XG-boost, and Multi-layer Perceptron. In this study, MLP achieved 86.25% accuracy, and others had above 80% accuracy. In the future, the author intends to employ numerous neural network structures [32], including CNN and RNN, with a sizable dataset.

Hasib et al. [13] used a dataset from Portuguese school reports and surveys. The authors offered a predictive model using Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), XGBoost, and Naive Bayes for students' success in secondary education. Before applying classification models, imbalanced datasets were balanced using K-Means SMOTE (Synthetic Minority Oversampling Technique). This study found the highest accuracy of 96.89% for the Support Vector Machine. The author wanted to extend research on student performance in tertiary education using deep learning approaches.

Cortez et al. [14] addressed the prediction of Portuguese secondary school students' grades, which worked under a dataset that included Portuguese secondary students' two core subjects, past school grades, and demographic, social, and other school-related data. Decision Trees, Random Forests, Neural Networks, and Support Vector Machines with three different data mining goals (i.e., binary/5 – level, for example, classification and regression) were used for prediction purposes. In this study, the author found neural networks and support vector machines outperformed the decision tree and random forest. In this paper, they did not consider the following factors, which affect a student's performance: reasons for a student's choosing a particular school; a parent's employment; or alcohol use.

Karagiannopoulos et al. [15] used five wrapper feature selection methods- Forward Selection, Backward Selection, Best First Forward Selection, Best First Backward Selection, and Genetic Search Selection—over four regression algorithms- Regression Trees, Regression Rules, Instance-Based Learning Algorithms, and Support Vector Machines—to improve the performance of regression models. Although the forward selection wrapper approaches are less expensive in terms of computational effort and employ fewer characteristics for induction, they are less effective at improving the performance of a specific regression model. The issue of feature interaction can be solved by creating new features from the basic feature set. The author planned to introduce a hybrid feature selection method in a subsequent paper that combines the benefits of filter and wrapper selection methods.

In their study, Ramaswami et al. [16], developed a prediction model from the CHAID prediction model to find out highly influenced variables to help low achievers of higher secondary students studying in the Indian educational system. A total of 1000 datasets for the year 2006 were gathered from five different schools in three different districts of Tamilnadu. When compared to other models, the accuracy of the CHAID prediction model was judged to be good. Due to the small student sample sizes and the small geographic coverage of the schools in the several districts of the state of Tamilnadu, it was not possible to generalize the results in this paper.

Sharma et al. [17] worked on a movie review dataset for sentiment analysis. The authors used five feature selection methods- DF, IG, GR, CHI, and Relief -F and seven machine learning techniques- Naive Bayes, Support Vector Machine, Maximum Entropy, Decision Tree, K-Nearest Neighbor, Winnow, Adaboost for sentiment analysis. The Naive Bayes classifier produced superior results when employed with fewer features than the gain ratio and SVM when selecting emotive features. The performance of these machine learning techniques for sentiment classification across domains is planned as a focus of further research.

Ma et al. [18] predicted whether a student would be able to get a certificate using the open edX platform. First, they specifically divided the dataset's student attributes into three categories. Then, they used various feature selection techniques- Relief Algorithm, Information Gain, Gain Ratio, and Correlation Coefficient—to extract key, significant character traits from the remaining characters.

Doshi et al. [19] implemented the following classification techniques- NBTree, Multilayer Perceptron, Naive Bayes, and Instance-based-K- nearest neighbor to help students who can get success in the engineering

stream for their higher studies in the future. To find relevant features, authors used feature selection algorithms- Chi-square, InfoGain, and GainRatio). Then they applied a fast correlation-based filter on the given features. They conclude that FCBF provides the most significant output for feature relevance. The authors plan to use other feature selection methods that can be used on the dataset in the future.

### 3. PROPOSED METHODOLOGY

Initially, we conducted our data collection process. We have collected data in three streams that are science, business and humanities. After our data collection we have performed a series of data processing techniques that are null value removal/handling, data type handling, cleaning and finally data normalization. Our data is splitted into two segments. Training data is used to train the machine learning models whereas the testing data is utilized to evaluate the trained model. While evaluating our models we have considered a series of evaluation metrics. Moreover, after conducting a comprehensive comparison between the models we export the best fitted model. Finally, we employ LIME [28] that explains a model's prediction. Lastly, in order to select a stream for individual students, we predict their GPA in each of the streams and finally make a decision accordingly. Figure 1 represents the overall proposed model of this study.

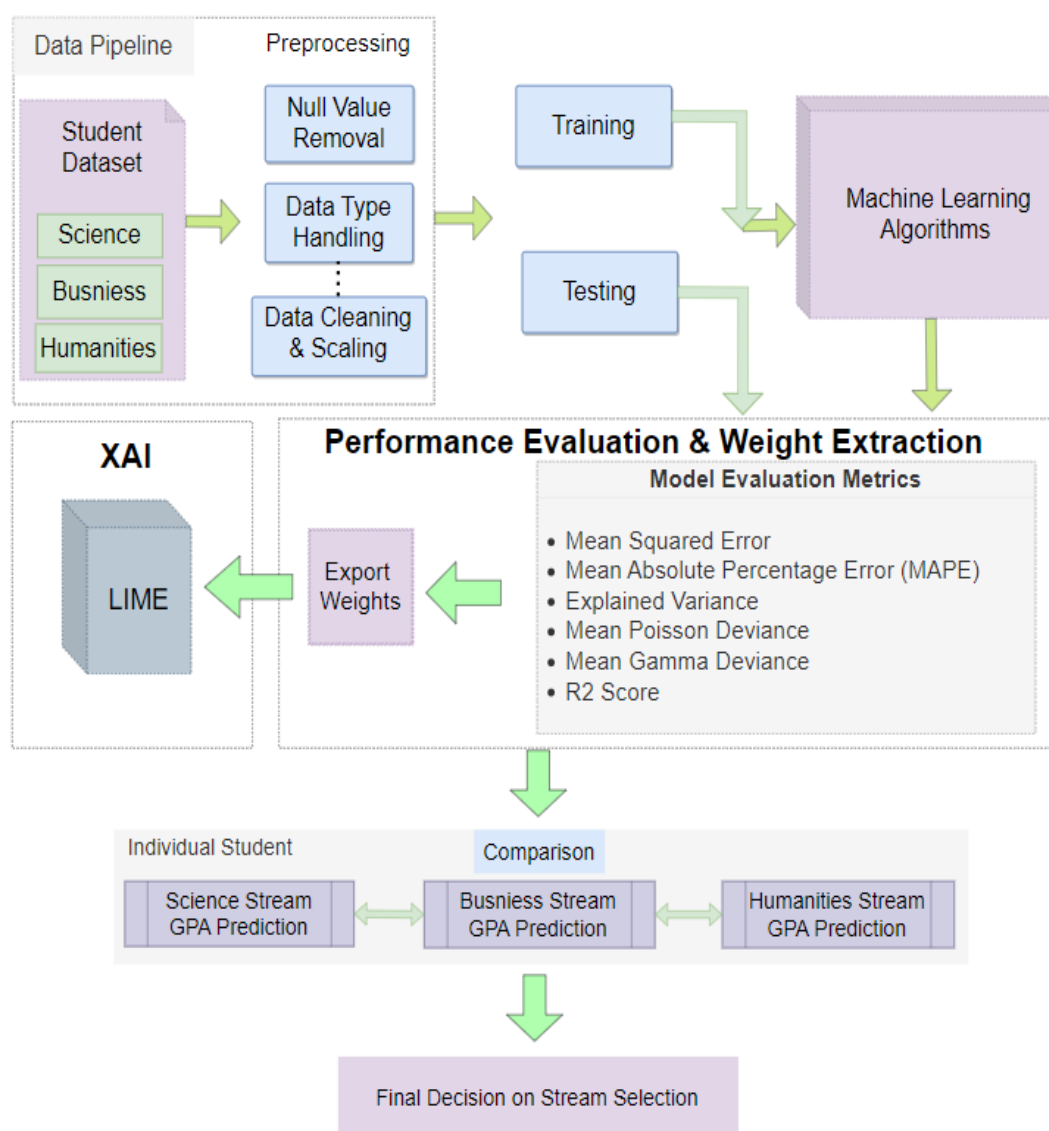


Figure 1. Proposed framework for predicting the best sperate stream for secondary school students in Bangladesh.

Table 1. Description of the attribute

Name of the attribute	Description
Gender	Student's Gender (0-Female, 1-Male)
Father Highest Academic Qualification	Primary Examination Completion Certificate-0; Junior School Certificate-1; Secondary School Certificate-2; Higher Secondary Certificate-3; Bachelors-4; Masters-5; Phd-6
Mother Highest Academic Qualification	Primary Examination Completion Certificate-0; Junior School Certificate-1; Secondary School Certificate-2; Higher Secondary Certificate-3; Bachelors-4; Masters-5; Phd-6
Father's Profession	Government Service, Teacher, Driver, Contractor, Accountant, Doctor, Mechanic, Lawyer, Tailor, Salesman, Banker, Artist, retired (govt. service), Retired (private service)-0; Business, Farmer, Fisherman, Politician, Cook-1; Unemployed, Labor, Others-2
Mother's Profession	Government Service, Teacher, Driver, Contractor, Accountant, Doctor, Mechanic, Lawyer, Tailor, Salesman, Banker, Artist, retired (govt. service), Retired (private service)-0; Business, Farmer, Fisherman, Politician, Cook-1; Housewife, Unemployed, Labor, Others-2
Father's average monthly income	Numeric
Mother's average monthly income	Numeric
How many siblings do you have	Numeric
District Currently you are living	District Under Dhaka Division-0, District Under Chottogram Division-1, District Under Rajshahi Division-2, District Under Khulna Division-3, District Under Sylhet Division-4, District Under Barishal Division-5, District Under Rangpur Division-6, District Under Mymensingh Division-7
PEC Result Overall GPA	Numeric
PEC Bangla	Numeric
PEC English	Numeric
PEC Mathematics	Numeric
PEC Religion	Numeric
PEC BGS	Numeric
PEC Science	Numeric
JSC Overall GPA	Numeric
JSC Bangla	Numeric
JSC English	Numeric
JSC Mathematics	Numeric
JSC BGS	Numeric
JSC ICT	Numeric
JSC Religion	Numeric
JSC Science	Numeric
Group SSC	Science-0; Business Studies-1; Humanities-2
Overall GPA SSC	Numeric

Table 2. Stream and Division wise data collection

Stream Name/ Divisions	Dhaka	Chottogram	Rajshahi	Khulna	Sylhet	Barisal	Rangpur	Mymensingh	Total
Science	84	18	31	4	1	4	24	8	174
Business Studies	56	13	13	2	22	1	2	1	110
Humanities	2	2	1	0	96	2	0	0	103
Total	142	33	45	6	119	7	26	9	387

### 3.1. Dataset

In our research, we have collected data from those students who are now in eleventh to twelfth grade or have already passed secondary and higher secondary levels. Otherwise, it is impossible to understand which separate stream is the perfect choice for them—a survey done by Google Form with 27 questions and a face-to-face interview. Later, we discussed with the principals and academic counselors from Cambrian School and College, Dhaka, Winsome School and College, and a few parents from both schools the features we used to develop a dataset. After the discussion, we used 26 features to create the dataset. We have developed the datasets mostly from Bangladesh International School and College Jeddah, Kingdom of Saudi Arabia, Cambrian School and College, Dhaka, Ghatla High School, Begumganj, Noakhali, and Jalalabad School and College, Sylhet. From Ghatla High School, Begumganj, Noakhali, and Jalalabad School and College, Sylhet. We received the data in an excel sheet format. For the science stream, we have been able to collect 174 students' data, of which 90 records are for male students and 84 for female students. In the Business Studies stream, we have 78 records for male students and 32 records for female students, for a total of 110 students' data. For the Humanities stream, we have collected 67 male students' data and 36 female students' data from a total of 103 students' data. Table 1 shows the list of attributes used to obtain student data.

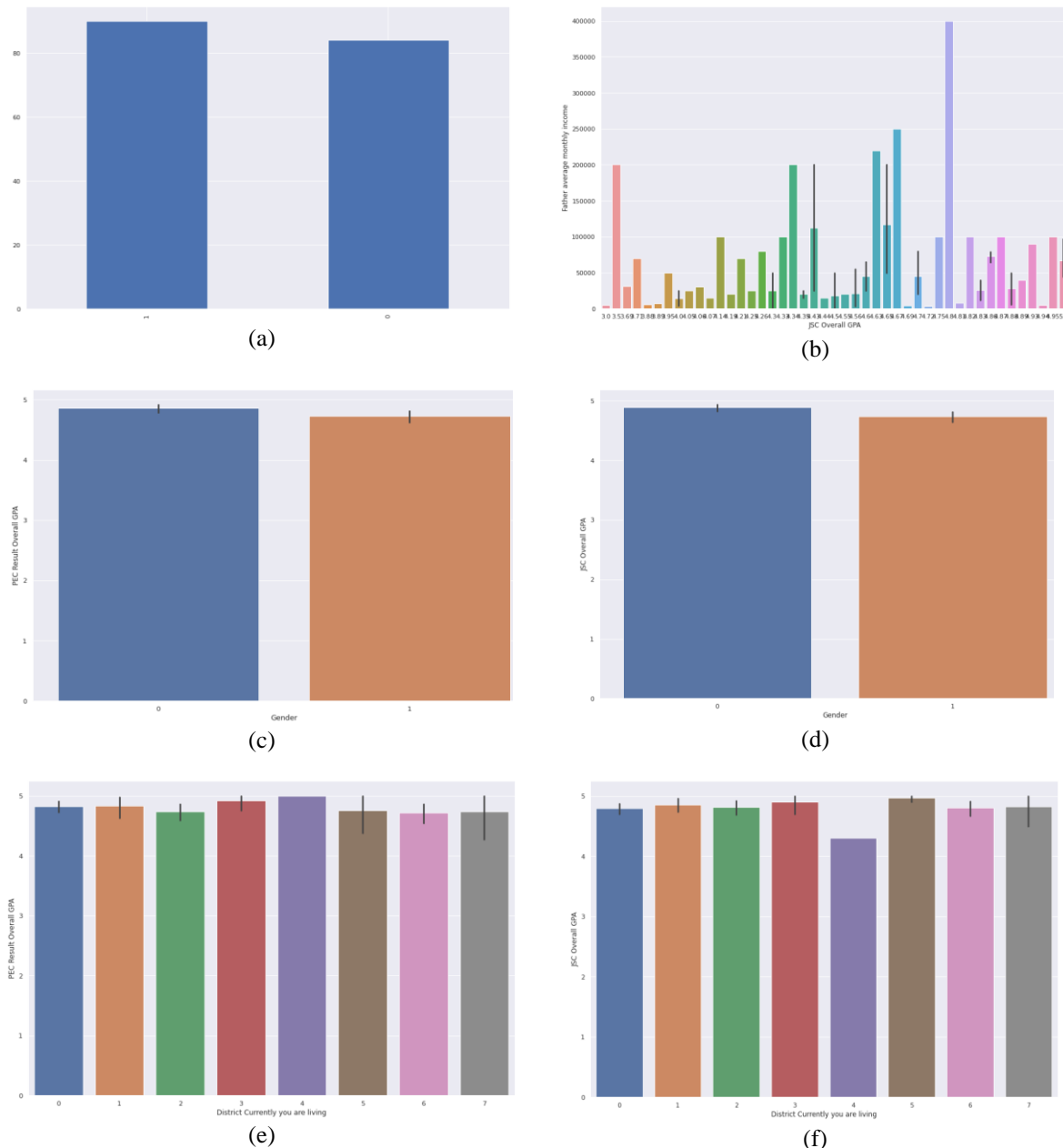


Figure 2. Exploratory Data Analysis on collected dataset. Here, (a) denotes Gender distribution. (b) represent a students GPA with respect to Fathers Average Monthly Income. (c) depicts the Gender distribution with respect to PEC Result Overall GPA. (d) reveals the Gender distribution with respect to JSC Result Overall GPA. (e) represents PEC overall results based on the district that students live in. (f) reflects the similar plot with JSC overall result.

### 3.2. Data Cleansing

We have fixed different names for the same organization after data pre-processing. It has been seen that the student has studied at Bangladesh International School and College Jeddah. Still, while writing the institution's name, the student has written Bangladesh International School, Bangladesh International School and College, or Bangladesh School Jeddah. As Bangladesh International School and College Jeddah are located in Saudi Arabia, this school is affiliated with the Gulshan police station in Dhaka. We got 253 data from google forms, and the rest were collected from Ghatla High School, Begumganj, Noakhali, and Jalalabad School & College, Sylhet in our prepared excel sheet format. As many as 23 records have been excluded from the datasets due to having too many null values. At last, we were able to collect 387 students' data.

We have done the dataset cleaning process, and, after doing Exploratory Data Analysis (EDA), we scaled our whole dataset. Table 2 shows the data collection scenario for different streams and divisions of Bangladesh.

### 3.3. Regression Techniques

#### 3.3.1. Linear Regression:

A case model with only one independent variable is called simple linear regression. Simple linear regression identifies a variable's dependence.  $y = \beta_0 + \beta_1 x + \varepsilon$ . Simple regression [20] can tell the difference between how the dependent variables affect each other and how the independent variables affect each other.

#### 3.3.2. Support Vector Machine Regression:

The goal of the support vector regression algorithm (SVR) is to find the predictor variables' most flat mathematical functions whose difference from the target is less than  $R+$  for all the training data. This function forms the core of a tube that is  $R+$  away from both margins. In contrast to the hard margin, the soft margin hyperplane SVR lets you go outside of  $R+$  by adding slack variables. Using the Gaussian radial basis function (RBF) kernel makes the model less linear and more able to change. Since the feature space has an infinite number of dimensions, the primal form cannot be used to solve the optimization problem in this case. However, the dual form obtained by using Lagrange multipliers can be used.

The RBF kernel function's support vectors' radius of impact, as well as the hyperparameters  $C$ , which penalize points outside the -tube, were all included in the technique for optimizing the hyperparameters. The libsvm implementation is used by the scikit-learn module. For the more extensive training dataset, a subsampling without replacement approach called pasting was used because the fit time complexity is more than quadratic with the number of samples [21].

#### 3.3.3. Random Forest Regression:

A machine learning ensemble approach using randomized decision trees is called Random Forest. Given that the result is derived from the multiple decision tree scores generated by bagging or Bootstrapping, subsampling, or random forest, is an ensemble method. In the case of regression, the unexpected forest result is the average scores from the randomized decision trees. A decision tree is an algorithm for machine learning that divides the space of predictor variables into groups of target variables that are similar to each other. In regression, the split flow stops when a further sub-partition is thought to not change the mean square error of the target variables in a significant way. The rules for making decisions at the tree's leaves, which are the tree's last nodes, back up the predictions about the target variable. In a randomized decision tree, the best way to split at each node is chosen by a random variable. The number of trees in the forest, the least number of samples needed to be at a leaf, and the minimum number of samples needed to divide an internal node were all considered in the hyperparameter optimization process [21].

#### 3.3.4. Adaptive Boosting:

AdaBoost, short for Adaptive Boosting, is a powerful ensemble learning algorithm that is used to improve the performance of weak learners [23]. It is a meta-algorithm that can be applied to any type of learning algorithm, such as decision trees, neural networks, or support vector machines. The basic idea behind AdaBoost is to iteratively train a series of weak learners, such as decision stumps, and give more weight to the samples that were misclassified by the previous weak learners. This process continues until a desired level of accuracy is achieved or a maximum number of weak learners is reached. The final output is a weighted sum of the predictions made by the weak learners. AdaBoost is particularly useful when the data contains a large number of samples with a small number of features, and the data is noisy or unbalanced. Because AdaBoost gives more weight to the misclassified samples, it can help to focus on the difficult examples and improve the performance of the final model. The algorithm has two main components: The weak learner: This is the base learning algorithm that is used to create the ensemble. It should be a simple algorithm that can be trained quickly and has low variance. Common choices include decision stumps, which are single-level decision trees, and perceptrons. The weight update: This is the mechanism by which the algorithm assigns higher weights to the misclassified samples. After each weak learner is trained, the samples are re-weighted so that the samples that were misclassified have a higher weight. AdaBoost is also computationally efficient and easy to implement, as it only requires a small number of parameters to be set.

#### 3.3.5. Gradient Boosting:

Gradient Boosting combines the predictions of multiple weak learners to create a stronger model [24]. It is a boosting algorithm that uses gradient descent to minimize the loss function. This method is used to improve the performance of a model by iteratively adding new models that are trained to correct the errors



made by the previous models. The basic idea behind Gradient Boosting is to train a sequence of weak models, such as decision trees, and add them together in a weighted manner. The algorithm starts with an initial model, typically a simple model such as a decision tree with one leaf. Then it iteratively trains new models and adds them to the ensemble, with each new model focusing on the samples that were misclassified by the previous models. The final output is a weighted sum of the predictions made by all the models in the ensemble. Gradient Boosting has several advantages: It is a powerful technique that can be used to improve the performance of a wide range of models, including decision trees, linear models, and neural networks. It is robust to overfitting, because it introduces randomness by training the models on different subsets of the data. It can handle a variety of data types, such as categorical and numerical features, and it can also handle missing data. It has two main components: The weak learner: This is the base learning algorithm that is used to create the ensemble. It should be a simple algorithm that can be trained quickly and has low variance. Common choices include decision trees, which are decision stumps with more than one level. The loss function: This is the mechanism by which the algorithm measures the error of the current ensemble. It is used to guide the training of new models by identifying the samples that are misclassified by the current ensemble. Gradient Boosting is computationally expensive and may require a lot of memory to store the multiple models of the ensemble, but it is widely used in many practical application and it is known for its good performance in many competitions and real-world problems

### 3.3.6. Extreme Gradient Boosting:

XGBoost is a gradient boosting algorithm that uses decision trees as its base model [25]. It is an implementation of gradient boosting framework. The main difference between XGBoost and other gradient boosting libraries is that XGBoost uses a more regularized model formalization to control over-fitting, which gives it better performance. One of the key features of XGBoost is its ability to handle missing values and irrelevant features. It can automatically learn the best missing value and feature interactions, and it can also handle large datasets with a large number of features. XGBoost also includes a number of other features that make it a powerful tool for machine learning and data science. Tree pruning: XGBoost uses a cost complexity parameter, known as "gamma," to control tree pruning. This allows the algorithm to automatically find the optimal trade-off between model complexity and performance. Regularization: XGBoost includes both L1 and L2 regularization, which helps to prevent overfitting. Column subsampling: XGBoost can randomly subsample the columns of the input data, which can help to reduce overfitting and improve performance. Early stopping: XGBoost includes an early stopping feature, which allows the algorithm to stop iterating once the performance on a validation set starts to deteriorate. Cross-validation: XGBoost can automatically perform cross-validation, which makes it easy to tune the model's hyperparameters. Built-in evaluation metrics: XGBoost includes a number of built-in evaluation metrics, such as error, log loss, and area under the ROC curve, which makes it easy to evaluate model performance. Speed: XGBoost is highly optimized and can be run on distributed systems. It is significantly faster than other gradient boosting libraries. Overall, XGBoost is a powerful and versatile tool that can be used for a wide range of machine learning tasks. It is widely used in industry and academia and can be integrated into various platforms and tools.

## 4. EXPERIMENTAL RESULT AND ANALYSIS

In the experimental evaluation, we employed a series of evaluation methods such as mean squared error, MAPE, mean absolute percentage error, explained variance, mean poisson deviance, mean gamma deviance, and R2 score. This section is divided into two subsections. In the first section, we discuss the utilized performance metrics and in the second section we conduct an analysis of our findings.

### 4.1. Performance Metrics

#### 4.1.1. Mean Squared Error:

Mean Squared Error (MSE) is a commonly used loss function for regression problems [26]. It measures the average squared difference between the predicted values and the true values. The MSE is widely used in practice because it is easy to compute and interpret. A lower MSE indicates a better fit between the predicted and true values, and it can be used to compare different models and select the best one. However, it can be sensitive to outliers, meaning if there are some extreme values in the dataset, it can affect the final MSE value, therefore in some cases other loss functions like Mean Absolute Error (MAE) are preferred. Equation 1 represents the mean squared error expression [26].

$$MSE = 1/n * \sum (y_i - \hat{y}_i)^2 \quad (1)$$

#### 4.1.2. Explained Variance:

Explained Variance is a statistical measure that quantifies the proportion of the total variance in the dependent variable that is explained by the independent variables in a regression model [27]. It is typically represented as a value between 0 and 1, where a value of 1 indicates that the model perfectly explains the variance in the target variable, and a value of 0 indicates that the model does not explain any of the variance. The explained variance is commonly computed using the R-squared statistic, which is defined as [27]:

$$R^2 = 1 - (\text{Sum of Squared Residuals} / \text{Total Sum of Squares}) \quad (2)$$

Where Sum of Squared Residuals is the sum of the squared differences between the predicted values and the true values, and Total Sum of Squares is the sum of the squared differences between the true values and the mean of the true values.

#### 4.1.3. Mean Absolute Percentage Error (MAPE):

Mean Absolute Percentage Error as a performance metric [22]. Errors are defined as discrepancies between the actual or observed value and the projected value. In statistics, it is referred to as a measure of a prediction technique's predictive accuracy. The MAPE decreases as the outlook gets better. The MAPE value can be calculated using the formula[22]:

$$\text{MAPE} = \frac{1}{n} \sum_{t=1}^n \left( \frac{A_t - F_t}{A_t} \right) \quad (3)$$

In MAPE, initially, it finds the absolute difference between Actual Value (A) and Estimated/Forecast Value (F). After applying the mean function, MAPE can be expressed as a percentage.

#### 4.1.4. Mean Poisson Deviance:

Mean Poisson Deviance is a measure of goodness of fit for Poisson regression models. Poisson regression is a type of generalized linear model (GLM) that is used to model count data, such as the number of occurrences of an event. The Poisson distribution is often used to model count data because it has the property of equating the mean and variance of the distribution, which is often the case with count data. Mean Poisson Deviance is a measure of the discrepancy between the predicted values and the observed values. It is calculated as:

$$\text{Deviance} = 2 * \sum (y_i * \log(y_i/y_i^{\wedge}) - (y_i - y_i^{\wedge})) \quad (4)$$

where  $y_i$  is the observed count,  $y_i^{\wedge}$  is the predicted count, and the summation is taken over all observations. Mean Poisson Deviance is similar to the residual deviance in other GLM models. It measures the difference between the observed and predicted values using the log-likelihood ratio. A smaller deviance indicates a better fit of the model to the data.

#### 4.1.5. Mean Gamma Deviance:

Mean Gamma Deviance is a measure of goodness of fit for Gamma regression models. Gamma regression is a type of generalized linear model (GLM) that is used to model continuous data that is positively skewed and has a positive mean, such as response time, income, or cost. The Gamma distribution is often used to model such types of data. Mean Gamma Deviance is a measure of the discrepancy between the predicted values and the observed values. It is calculated as:

$$\text{Deviance} = 2 * \sum (y_i * \log(y_i/y_i^{\wedge}) - (y_i/y_i^{\wedge}) + \log(y_i^{\wedge})) \quad (5)$$

where  $y_i$  is the observed value,  $y_i^{\wedge}$  is the predicted value, and the summation is taken over all observations.

#### 4.1.6. R2 Score:

The R-squared (R2) score is a statistical measure that represents the proportion of the variance in the dependent variable (also known as the target variable) that is explained by the independent variables (also known as the predictors or features) in a regression model. It is a value between 0 and 1, where a value of 1 indicates that the model perfectly explains the variance in the target variable, and a value of 0 indicates that the model does not explain any of the variance. The R-squared score is calculated as:

$$R^2 = 1 - (\text{Sum of Squared Residuals} / \text{Total Sum of Squares}) \quad (6)$$

Where Sum of Squared Residuals is the sum of the squared differences between the predicted values and the true values, and Total Sum of Squares is the sum of the squared differences between the true values and the mean of the true values. The R-squared score is a commonly used measure of goodness of fit for linear regression models, and it can also be applied to other types of models. It is a measure of how well the model fits the data, a high R-squared value means that the model fits the data well and a low R-squared value means that the model does not fit the data well.

Table 3. Performance Evaluation for Science stream.

Machine Learning Algorithm	Mean Squared Error	Mean Absolute Percentage Error (MAPE)	Explained Variance	Mean Poisson Deviance	Mean Gamma Deviance	R2 Score
Support Vector Machine Regression	0.254334	0.043381	0.338168	0.014052	0.003059	0.308250
Random Forest Regression	0.259202	0.041277	0.288627	0.014754	0.003247	0.281518
Linear Regression	0.297445	0.051826	0.138872	0.018923	0.004060	0.053868
Gradient Boosting Regressor	0.247409	0.043440	0.345616	0.013275	0.002885	0.345409
ADA Boost Regressor	0.294958	0.051792	0.069652	0.01913	0.004221	0.069623
<b>Extreme Gradient Boosting</b>	0.255625	0.040922	0.307820	0.014305	0.003139	0.301212

Table 4. Performance Evaluation for Business Studies stream.

Machine Learning Algorithm	Mean Squared Error	Mean Absolute Percentage Error (MAPE)	Explained Variance	Mean Poisson Deviance	Mean Gamma Deviance	R2 Score
Random Forest Regression	0.384224	0.068950	0.514893	0.036536	0.009134	0.4863788
Linear Regression	0.429913	0.078906	0.357799	0.044692	0.010942	0.356961
Gradient Boosting Regressor	0.511562	0.067908	0.541848	0.034929	0.008811	0.511562
Extreme Gradient Boosting	0.523223	0.068744	0.551153	0.034099	0.008584	0.523223
ADA Boost Regressor	0.416501	0.076927	0.457631	0.042029	0.010665	0.416501
<b>Support Vector Machine Regression</b>	0.365942	0.069041	0.558916	0.033906	0.008706	0.534092

#### 4.2. Performance Evaluation

We have done our experiment by using Jupyter Notebook and Python code for each stream separately, with separate datasets. In our research, 70% of the data is used for training and 30% is for testing. We have utilized six different machine learning models for this study. Our findings for each of the streams is presented below.

Table 5. Performance Evaluation for Humanities stream.

Machine Learning Algorithm	Mean Squared Error	Mean Absolute Percentage Error (MAPE)	Explained Variance	Mean Poisson Deviance	Mean Gamma Deviance	R2 Score
Support Vector Machine Regression	0.376961	0.082050	0.533314	0.039947	0.011560	0.533312
Random Forest Regression	0.256137	0.056065	0.803241	0.018689	0.005515	0.784535
Linear Regression	0.262854	0.071188	0.776502	0.021169	0.006587	0.773084
Gradient Boosting Regressor	0.270212	0.056971	0.786403	0.019917	0.005586	0.760203
ADA Boost Regressor	0.277654	0.060603	0.776809	0.024455	0.008073	0.746813
<b>Extreme Gradient Boosting</b>	0.244682	0.055422	0.817297	0.017635	0.005366	0.803374

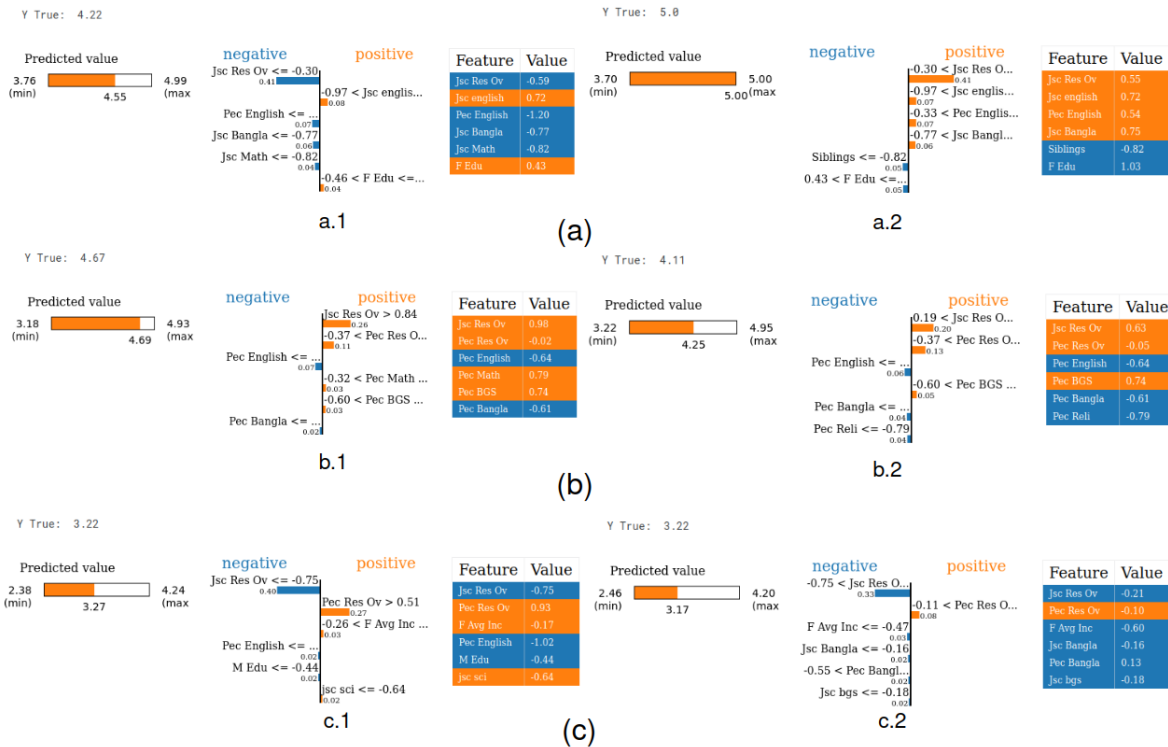


Figure 3. Local Interpretable Model Agnostic Explanations prediction. Here, science stream is represented by (a), business stream by (b) and finally humanities stream by (c).

#### 4.2.1. Science Stream

Table 3 represents our findings in the science stream. Here, Linear Regression has the highest Mean Squared Error and Mean Absolute Percentage Error. Random Forest Regression and Support Vector Machine Regression however performed moderately well. In this stream, Gradient Boosting Regressor managed to gain the lowest mean squared error. However, it has a slightly higher MAPE and R2 score. In terms of explained variance, mean possession deviance and mean gamma deviance it has a moderate score. Extreme Gradient Boosting however has a more decent score in all of the considered metrics. Considering all the metrics our findings state that, Extreme Gradient Boosting is the better performing model in the stream.

#### 4.2.2. Business Stream

Our insights in the business stream are presented in Table 4. Although Extreme Gradient Boosting was the best performing model in the science stream, in this stream despite having a moderate MAPE score, it has the highest mean squared error. Among the other machine learning models Random Forest and Linear Regression's performance were notable. Support Vector Machine's performance however was the most superior.

#### 4.2.3. Humanities Stream

In the scientific stream, Table 5 summarizes our conclusions. Although Support Vector Machine was the better performing model in the previous stream, in this stream its performance decayed. Except Support Vector Machine, the remaining algorithms all performed noticeably better. Extreme Gradient Boosting has the lowest mean squared error and mean absolute percentage error in this stream. Our findings conclude that Extreme Gradient Boosting is the better model in this region.

### 4.3. Local Interpretable Model Agnostic Explanations

It is a model-agnostic method, which means that it can be used to explain the predictions of any machine learning model, regardless of its architecture or underlying assumptions [26]. The main idea behind LIME is to approximate the behavior of a complex model in a local neighborhood around a specific instance. It does this by generating a simplified, interpretable model that is only valid in the vicinity of the instance in question. This allows the user to understand why the model made a specific prediction for a particular instance, even if the global behavior of the model is complex and difficult to interpret. The algorithm works by perturbing the input instance and generating a new dataset that is locally similar to the original instance. It

then fits a simple interpretable model, such as a linear model or a decision tree, to this new dataset. The coefficients of this model can be used to determine the relative importance of each feature in the original instance's prediction. LIME can also be used to generate human-readable explanations of a model's predictions by visualizing the decision boundary of the simple interpretable model. This can be a useful tool for building trust in a model and gaining insight into its behavior. Overall, LIME is a powerful technique for interpreting and understanding the predictions of any machine learning model, and it can be a valuable tool for building more transparent and trustworthy models. While utilizing LIME we have used our best fitted model.

Figure 3 represents the LIME predictions of our testing data. Here, (a) represents the science stream. From a.1 we can visualize that whereas Y true value is 4.22 our model has predicted 4.55. LIME has successfully explained which features play the most important roles for this prediction. Here, Jsc's overall result has the most negative value for the model's prediction causing the model to predict far off. However, in a.2 our model was precise in predicting its output. Here, the true value was 5.0 and the model's predicted value is also 5.0. Here, JSC overall result, JSC English and JSC Bangla feature has the most positive values. In science stream, utilizing LIME it is revealed that these three values become the most important features in predicting a students SSC result. (b) denotes the business stream. In b.1 whereas our model has predicted 4.69, the true value is 4.69 and this slight misinterpretation of the prediction is because of the PEC English feature. In b.2 PEC English, PEC Bangla, PEC Religion has a negative impact on our model's prediction causing the model to predict 4.25 where the true value is 4.11. Finally, (c) represents the humanities stream. In c.1 we can visualize that once again the JSC Res overall result had a negative impact on our model. On the other hand the PEC Res overall became the most important feature in this prediction. Lastly in c.2, once again JSC Res overall, PEC Res overall acted similarly.

## 5. CONCLUSION

The aim of this study is to develop a machine learning model that helps students and stakeholders find a proper stream for a student. A total of 26 features were considered in the prediction analysis. Among the 6 regression algorithms, Extreme Gradient Boosting more accurately predicts the results for Science and Humanities streams and Support Vector Machine regression for Business stream.

The dataset used in this paper has a limited diversity and is small in volume. In the future, the research can be extended by considering a high volume of data and other state-of-the-art regression algorithms.

## ACKNOWLEDGEMENTS

This research is funded by Woosong University Academic 2023.

## REFERENCES

- [1] P. Shruthi, and B. P. Chaitra. "Student performance prediction in education sector using data mining", 2016.
- [2] SSC Routine 2022 PDF All Education Board, July 2022. Available. <https://dhakaeducationboard.gov.bd/data/20220731181556699548.pdf>, Accessed July 31, 2022.
- [3] No science, arts or commerce in secondary education: A good idea?, November 24, 2020 (Accessed July 28, 2022). Available. <https://www.thedailystar.net/shout/news/no-science-arts-or-commerce-secondary-education-good-idea-2000413>
- [4] N. B., Sara, H. Rasmus, I. Christian, and A. Stephen, "High-school dropout prediction using machine learning: A Danish large-scale study." In ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence, pp. 319-24. 2015.
- [5] M.A. Rahman, "Factors Leading to Secondary School Dropout In Bangladesh: The Challenges to Meet the Sdg's Targets." Journal of the Asiatic Society of Bangladesh, Science 47, no. 2, pp. 173-190, 2021.
- [6] M.N.I. Sarker, M. Wu, and M.A. Hossin, "Economic effect of school dropout in Bangladesh", International journal of information and education technology, 9(2), pp.136-142, 2019.
- [7] M.S. Acharya, A. Armaan, and A.S. Antony, "A comparison of regression models for prediction of graduate admissions," In 2019 international conference on computational intelligence in data science (ICCIDS) (pp. 1-5). IEEE, 2019.
- [8] O. El Aissaoui, Y. El Alami El Madani, L. Oughdir, A. Dakkak, and Y. El Alloui, "A multiple linear regression-based approach to predict student performance," In International Conference on Advanced Intelligent Systems for Sustainable Development (pp. 9-23). Springer, Cham, 2019.
- [9] M.S. Zulfiker, N. Kabir, A.A. Biswas, P. Chakraborty, and M.M. Rahman, "Predicting students' performance of the private universities of Bangladesh using machine learning approaches," International Journal of Advanced Computer Science and Applications, 11(3), pp. 672-679, 2020.
- [10] R. Hasan, M.K.A. Ovy, I.Z. Nishi, M.A. Hakim, and R. Hafiz, "A Decision Support System of Selecting Groups (Science/Business Studies/Humanities) for Secondary School Students in Bangladesh," In 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE, 2020.

- [11] N. Shahadat, M. Rahman, S. Ahmed, and B. Rahman, "Predicting higher secondary results by data mining algorithms with VBR: A feature reduction method," In 4th International Conference on Advances in Electrical Engineering (ICAEE) (pp. 164-169). IEEE, 2017.
- [12] K. Ahammad, P. Chakraborty, E. Akter, U.H. Fomey and S. Rahman, "A comparative study of different machine learning techniques to predict the result of an individual student using previous performances," International Journal of Computer Science and Information Security (IJCSIS), 19(1), 2021.
- [13] K.M. Hasib, F. Rahman, R. Hasnat, and M.G.R. Alam, "A Machine Learning and Explainable AI Approach for Predicting Secondary School Student Performance," In 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC) (pp. 0399-0405). IEEE 2022.
- [14] P. Cortez, and A.M.G., Silva, "Using data mining to predict secondary school student performance," 2008.
- [15] P. Cortez, and M. Karagiannopoulos, D. Anyfantis, S.B. Kotsiantis, and P.E. Pintelas, "Feature selection for regression problems," Educational Software Development Laboratory, Department of Mathematics, University of Patras, Greece, 2004.
- [16] Ramaswami, M. and Bhaskaran, R., "A CHAID based performance prediction model in educational data mining," arXiv preprint arXiv:1002.114, 2010.
- [17] A. Sharma, and S. Dey, "A comparative study of feature selection and machine learning techniques for sentiment analysis," In Proceedings of the 2012 ACM research in applied computation symposium (pp. 1-7), 2012.
- [18] C. Ma, B. Yao, F. Ge, Y. Pan, and Y. Guo, "Improving prediction of student performance based on multiple feature selection approaches," In Proceedings of the 2017 International Conference on E-Education, E-Business and E-Technology (pp. 36-41), 2017.
- [19] M. Doshi, "Correlation based feature selection (CFS) technique to predict student Performance," International Journal of Computer Networks & Communications, 6(3), p.197, 2014.
- [20] D. aulud, and A.M. bdulazeez, "Review on linear regression comprehensive in machine learning," Journal of Applied Science and Technology Trends, 1(4), pp.140-147, 2020.
- [21] R. Costa-Mendes, T. Oliveira, M. Castelli, and F. Cruz-Jesus, "A machine learning approximation of the 2015 Portuguese high school student grades: A hybrid approach," Education and Information Technologies, 26(2), pp.1527-1547, 2021.
- [22] A.A. Elrahman, T.H.A. Soliman, A.I. Taloba, and M.F. Farghally, "A Predictive Model for Student Performance in Classrooms Using Student Interactions With an eTextbook," arXiv preprint arXiv:2203.03713, 2022.
- [23] D.P. Solomatine, D.L. Shrestha, "AdaBoost RT a boosting algorithm for regression problems," In2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541) 2004 Jul 25 (Vol. 2, pp. 1163-1168). IEEE.
- [24] A. Natekin, A. Knoll, "Gradient boosting machines, a tutorial. Frontiers in neurorobotics," 2013 Dec 4; vol. 7, no. 21.
- [25] T Chen, T He, M Benesty, V Khotilovich, Y Tang, H Cho, K Chen, "Xgboost: extreme gradient boosting. R package version 0.4-2. 2015 Aug 1; vol. 1, no. 4, pp. 1-4.
- [26] A De Myttenaere, B Golden, B Le Grand, F Rossi, "Mean absolute percentage error for regression models," Neurocomputing. 2016 Jun 5;vol. 192, pp. 38-48.
- [27] KE O'Grady, "Measures of explained variance: Cautions and limitations," Psychological Bulletin. 1982 Nov; vol. 92, no. 3, pp. 766.
- [28] MT Ribeiro, S Singh, C Guestrin, "Why should i trust you?," Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining 2016 Aug 13, pp. 1135-1144.
- [29] Group prediction by Regressor-Dataset.xlsx. (n.d.). Google Docs. [https://docs.google.com/spreadsheets/d/1Az5vyGnDrzhM\\_xZzYIGZ1LEmy5a14bR2/edit](https://docs.google.com/spreadsheets/d/1Az5vyGnDrzhM_xZzYIGZ1LEmy5a14bR2/edit)
- [30] M. S. I. Khan, N. Islam, J. Uddin, S. Islam, M. K. Nasir, "Water quality prediction and classification based on principal component regression and gradient boosting classifier approach," Journal of King Saud University-Computer and Information Sciences, 2022, vol. 34,no. 8, pp. 4773-4781.
- [31] J. Uddin, F. N. Arko, N. Tabassum, T. R. Trisha, F. Ahmed, "Bangla sign language interpretation using bag of features and Support Vector Machine," In 2017 3rd International Conference on Electrical Information and Communication Technology (EICT), 2017, pp. 1-4.
- [32] R. A. Khan, J. Uddin, S. Corraya, J. Kim, "Machine vision based indoor fire detection using static and dynamic features," International Journal of Control and Automation, vol. 11, no. 6, pp. 87-98.

## BIOGRAPHIES OF AUTHORS






**Shabbir Ahmad** is a graduate student at Bangladesh's Brac University studying computer science and engineering. He graduated from the International Islamic University in Chittagong, Bangladesh, with a B.Sc. in Computer Science and Engineering. At Cambrian College in Bangladesh, he teaches information and communication technology.





**Md. Golam Rabiul Alam** is a Professor of Computer Science and Engineering Department of BRAC University. He received Ph.D. in Computer Engineering from Kyung Hee University, South Korea in 2017. He served as a Post-doctoral researcher in Computer Science and Engineering Department, Kyung Hee University, Korea from March 2017 to February 2018. Dr. Rabiul Alam received B.S. and M.S. degrees in Computer Science and Engineering, and Information Technology from Khulna University and University of Dhaka respectively. Dr. Rabiul Alam has published around 70 research articles in reputed journals, and conference proceedings. He has also 3 registered patents on ambient assisted living, mobile cloud computing and mobile fog computing, respectively. He is the reviewer of IEEE Communications Magazine, IEEE Transactions on Network and Service Management, IEEE Access, Elsevier Journal on Vehicular Communications, Elsevier Journal of Systems Architecture, IEEE/IFIP IM (International Symposium on Integrated Network Management), and IEEE/IFIP NOMS (Network Operations and Management Symposium), IEEE ICC (International Conference on Communications), IEICE/KICS APNOMS (Asia-Pacific Network Operations and Management Symposium), KIISE ICOIN (International Conference on Information Networking), ECCE (International Conference on Electrical, Computer and Communication Engineering), ACM ICUIMC (International Conference on Ubiquitous Information Management and Communication-IMCOM), and TPC member of ICCA (International Conference on Computer and Applications, UAE). Dr. Rabiul Alam's research interests include Healthcare-IoT networks, Mental-health informatics, Ambient intelligence systems, Mobile-cloud computing, Edge computing, Affective computing, and UAV Image processing. He is a member of the IEEE Computer Society, Consumer Electronics Society, and Industrial Electronics Society and received several best paper awards at national and international research conferences.



**Dr. Jia Uddin**    is as an Assistant Professor, Department of [AI and Big Data](#), Endicott College, Woosong University, Daejeon, South Korea. He received Ph.D. in Computer Engineering from University of Ulsan, South Korea, M.Sc. in Electrical Engineering (Specialization: Telecommunications), Blekinge Institute of Technology, Sweden, and B.Sc. in Computer and Communication Engineering, International Islamic University Chittagong, Bangladesh. He was a visiting faculty at School of computing, Staffordshire University, United Kingdom. He is an Associate Professor (now on leave), Department of Computer Science and Engineering, Brac University, Dhaka, Bangladesh. His research interests are Industrial Fault Diagnosis, Machine Learning/Deep Learning based prediction and detection using multimedia signals.



**Dr. Md Roman Bhuiyan** has completed his Doctor of Philosophy (Ph.D.) by research in information technology from Multimedia University, Cyberjaya, Malaysia in 2022. He received the B.Sc. (Eng.) in computer science and engineering from Uttara University, Bangladesh, in 2015, and the Masters of Software Engineering from the faculty of computer science and engineering, FTMS College Malaysia in collaboration with Leeds Beckett University, UK, in 2017. His current research interests include Artificial Intelligence, Algorithm Design, Computer Vision, Deep Learning, Machine Learning, Image, and Video Analysis.