Machine Learning based Stream Selection of Secondary School Students in Bangladesh

by

Shabbir Ahmad
19266003

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

Department of Computer Science and Engineering
BRAC University
May 2023

Declaration

It is hereby declared that

1. wrote the thesis on my own initiative while earning my degree at Brac University.
2. The thesis does not include any content that has already been published or authored by a third party, unless it is properly cited using complete and correct referencing.
3. The thesis does not include any content that has already been approved or submitted for consideration for another degree or diploma at a university or other institution.
4. I have thanked my primary sources of assistance.

Student's Full Name & Signature:

____

Shabbir Ahmad
19266003

Approval

The thesis titled "Machine Learning based Stream Selection of Secondary School Students in Bangladesh"

Submitted by:

Shabbir Ahmad (19266003)

Of Summer, 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of M.Sc. in Computer Science and Engineering on May 20, 2023.

Examining Committee:

Supervisor:
(Member)

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

Examiner:
(External)

Shamim H Ripon, PhD
Professor
Department of Computer Science and Engineering
East West University

Examiner:
(Internal)

Md. Ashraful Alam, PhD
Assistant Professor
Department of Computer Science and Engineering
BRAC University

Program Coordinator:
(Member)

Amitabha Chakrabarty, PhD
Associate Professor
Department of Computer Science and Engineering
School of Data and Sciences
Brac University

Chairperson:
(Chair)

Sadia Hamid Kazi, PhD
Associate Professor
Department of Computer Science and Engineering
BRAC University

Abstract

A strong civilization is built on a strong foundation, and education plays a vital role in acquiring the necessary information and skills for success in life. This thesis focuses on the education system in Bangladesh, which is divided into three levels: primary (PEC), middle school (JSC), and secondary school certificate (SSC). The selection of a stream after the eighth grade is crucial for students' higher studies and career planning, with three options available: Science, Business Studies, and Humanities.To address the challenge of stream selection based solely on PSC and JSC results, we have collected a dataset from various Bangladeshi schools, comprising student records that include subject-wise results, parent's academic qualification, parent's profession, parent's monthly income, sibling information, district, etc. In this study, we employ a series of machine learning regression algorithms to analyze the data.Furthermore, we utilize performance metrics and R2 scores to evaluate and validate the models' performance. Among the regressors, the gradient boosting algorithm demonstrates superior performance for the Science stream, achieving an R2 score of 0.34540. For the Business Studies stream, the Support Vector Machine exhibits significantly better performance with an R2 score of 0.534092. Finally, the Humanities stream shows excellent results with an R2 score of 0.80337 using extreme gradient boosting.To enhance the interpretability of our models, we leverage the Local Interpretable Model Agnostic Explanations (LIME) technique. The analysis and findings of this research are expected to assist prospective students and stakeholders in making informed decisions regarding stream selection, ensuring alignment with their future goals and aspirations.

Keywords: Regression analysis; Local interpretable model agnostic explanations; Stream recommendation system; Bangladeshi secondary school.

Dedication

This work is dedicated to my parents, loving wife, adorable son, siblings, and Dr. Golam Rabiul Alam, who guided and patiently supported me during this time.

Acknowledgement

First and foremost, i would want to respectfully thank and honor the Almighty for allowing me to live and complete tasks, such as the study work that is being given here.

It gives me great pleasure to express my sincere gratitude and respect to my supervisor, Dr. Golam Rabiul Alam, Professor, Department of Computer Science and Engineering, BRAC University, for his helpful suggestions, academic direction, ongoing inspiration, and gracious cooperation throughout the entire progress of this research work. Without his amazing guidance, I could not have completed my assignment.

Moreover, I want to thank Dr. Jia Uddin, Muntasir Ahmmed, Sayed Ahmed, Adil Mahmud Choudhury, Kazi Nurul Islam, Musleh Uddin, Dr. Md. Roman Bhuiyan, Tasnim Sakib and Abdullah Umar Nasib.

Finally, I'd like to express my gratitude to my family and friends for their support and encouragement.

Table of Contents

List of Figures

List of Tables

Nomenclature

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

AdaBoost Adaptive Boosting

CNN Convolutional Neural Network

EDA Exploratory Data Analysis

GLM Generalized Linear Model

GPA Grade Point Average

ICT Information and Communication Technology

JSC Junior School Certificate

KNN K-Nearest Neighbour

MAPE Mean Absolute Percentage Error

mRMR Minimum Redundancy and Maximum Relevance

MSE Mean Squared Error

PEC Primary Education Completion Certificate

R2Score R-Squared Score

RNN Recurrent Neural Network

SMOTE Synthetic Minority Over-sampling Technique

SSC Secondary School Certificate

SVM Support vector Machine

XAI Explainable Artificial Intelligence

XGBoost Extreme Gradient Boosting

LIME Local Interpretable Model Agnostic Explanations

Chapter 1

Introduction

1.1 Motivation

The process of selecting a student's secondary school stream is crucial in deciding their future educational and professional paths. The PEC and JSC results, which are used in the current process of stream selection, may not provide a comprehensive assessment of students' abilities and potential. To address this issue, a stream selection strategy that is data-driven and machine learning-based is becoming more and more important. This study makes use of machine learning techniques to examine vast volumes of data, identify important patterns and relationships, and develop predictive models that account for a variety of factors influencing students' preferences for and aptitudes in different fields. Not simply individual students will benefit from the study's findings; so will educational institutions and other stakeholders. Additionally, stream selection using machine learning approaches will improve educational practices in Bangladesh. In order to address the issues with Bangladesh's current stream selection processes, our research uses machine learning techniques to provide secondary school students with individualized recommendations.

1.2 Research Problem

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 [34]. However, a few subjects- Bangla, English, General Math, ICT, and Religious Studies—were generally applicable to all groups [25]. 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 [12]. 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 [29]. 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 [20]. 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.

1.3 Aims and Objectives

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:

1. 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.

2. 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, gradient boosting and suport vector regressor shows higher accuracy than the other state-of-art algorithms.

3. 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 [35] public for reproducible research.

4. Furthermore, we have utilized Local Interpretable Model Agnostic Explanations (LIME) as an explainable AI (XAI) that introduces interpretability to our proposed model.

1.4 Organization of the Report

This report is formatted as follows: Chapter 2 provided a description of the realted works for this thesis. Chapter 3 described about the dataset. Methodologies are discussed and briefly examined in Chapter 4. The Experimental Results and discussion are covered in Chapter 5. Chapter 6 presents the main conclusion of the thesis.

Chapter 2

Related Works

The field of machine learning (ML) is one that is expanding quickly and has the potential to transform the educational system. By training ML systems on big datasets of student data, it can be utilized to predict student performance. Large volumes of data can be processed rapidly and effectively by ML algorithms, and they can also spot patterns in data that are difficult for people to see. This can assist teachers in identifying students who are at danger of failing and in tailoring their instruction to each student. ML is a potential tool that can assist educators in early failure risk identification, individualized training for each student, and the identification of successful interventions to raise student outcomes. It can be used to identify kids who are likely to succeed in college, drop out of school, or fail a class. By providing students with the assistance they require and bridging the accomplishment gap, it has the ability to completely transform the educational system.

A number of researchers have used machine learning algorithms to predict student success in educational institutions. Acharya et al. [19] described a machine learning issue in students' choice of universities. They contrasted various regression algorithms[33], 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.[21] 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. [26] 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 [17], 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. [23] 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. [16] 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. [27] 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 [18], including CNN and RNN, with a sizable dataset.

Hasib et al. [32] 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. [5] 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. [3] 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. [6], 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. [7] 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. [15] 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. [9] 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.

The related works revealed several studies on predicting student outcomes and stream selection using machine learning techniques. However, most of these studies focused on factors other than primary and high school examination results (PEC and JSC) for stream selection. Our research paper aims to address this gap by specifically examining the impact of PEC and JSC results on stream selection in Bangladesh.

Acharya et al. [19] and El Aissaoui et al. [21] explored machine learning algorithms for choice of an better university and university students performance prediction, respectively. While they achieved good results, they did not consider primary and high school results. Zulfiker et al. [26] discussed about final grade of a single university in Bangladesh, but their focus was on improving grades rather than predicting streams based on primary and middle school exams.

Hasan et al. [23] focused on helping 9th-grade students choose their stream, but their study did not emphasize the importance of PEC and JSC results. Similarly, Shahadat et al. [16] predicted Higher Secondary Certificate examination results but did not utilize primary and middle school exam data.

Ahammad et al. [27] and Hasib et al. [32] predicted students' performance but did not specifically consider stream selection based on PEC and JSC results. Cortez et al. [5] predicted grades but did not explore the impact of primary and and high school results or other relevant factors.

Karagiannopoulos et al.[3] and Sharma et al. [7] focused on feature selection and model performance improvement but did not address stream selection based on primary and high school results.

Ramaswami et al.[6] and Ma et al.[15] explored prediction models but in different educational contexts, and they did not incorporate primary and high school results for stream selection. Doshi et al. [9] focused on engineering stream selection but did not consider primary and middle school exam results.

In summary, while several studies have investigated student performance prediction and stream selection using machine learning, none of them have extensively utilized PEC and JSC results or focused on stream selection in the context of Bangladesh. Moreover, the interpretability of the models using techniques like LIME has not been explored for stream selection. Therefore, our research aims to fill these gaps by examining the impact of PEC and JSC results on stream selection and leveraging LIME for model interpretability. Our features have been examined and approved by educationists and heads of institutions, making our research unique and valuable in the field of stream selection in Bangladesh.

Chapter 3

Dataset

Machine learning has totally altered how we look at data, empowering us to gain insightful knowledge and make precise predictions. In order to identify useful patterns and create models that improve various decision-making processes, this thesis investigates the use of machine learning algorithms on a range of dataset.

3.1 Data Collection

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 3.1 shows the list of attributes used to obtain student data.

3.2 Data Cleansing

We have fixed different names for the same organization after data preprocessing. 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

Table 3.1: Description of the attribute

Name of the attribute	Description
Gender	Student's Gender (0-Female, 1-Male)
Father's Highest Academic Qualification	Primary Examination Certificate-0, Junior School Certificate-1, Secondary School, Certificate-2, Higher Secondary Certificate-3, Bachelors-4, Masters-5, PhD-6
Mother's Highest Academic Qualification	Primary Examination 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
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
Father's average monthly income	Numeric
Mother's average monthly income	Numeric
How many siblings do you have	Numeric
District Currently you are living	Under Dhaka Division-0, Under Chottogram Division-1, Under Rajshahi Division-2, Under Khulna Division-3, Under Sylhet Division-4, Under Barishal Division-5, Under Rangpur Division-6 Under Mymenshingh 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 3.2: Stream and Division wise data collection

Stream Name/ Divisions	DHK	CTG	RAJ	KHU	SYL	BAR	RNP	MYNG	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

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), EDA has shown in Figure 3.1 to 3.6. We scaled our whole dataset. We have also done feature selection by Minimum Redundancy and Maximum Relevance (mRMR). Table 3.2 shows the data collection scenario for different streams and divisions of Bangladesh.

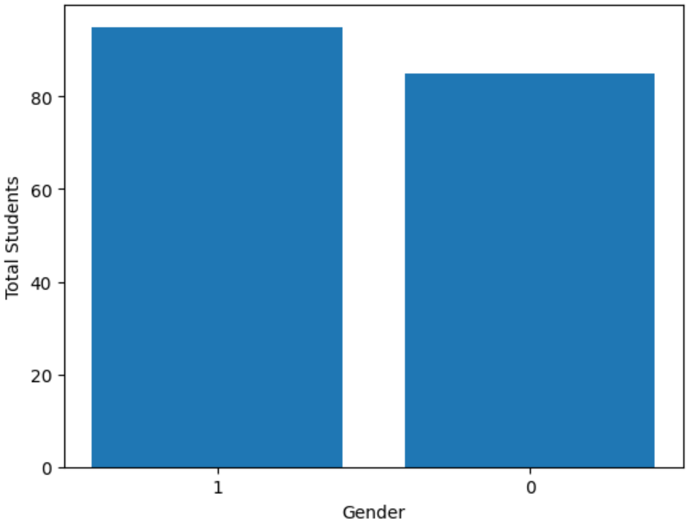


Figure 3.1: Exploratory Data Analysis on Gender distribution.

In Figure 3.1, a comparison is presented depicting the number of data records for males and females. It is evident from the graph that the number of female records slightly surpasses that of male records. This observation highlights a marginally higher representation of females within the dataset.

The analysis conducted in Figure 3.2 reveals an interesting finding: the average monthly income of fathers does not seem to have any influence on the overall re-

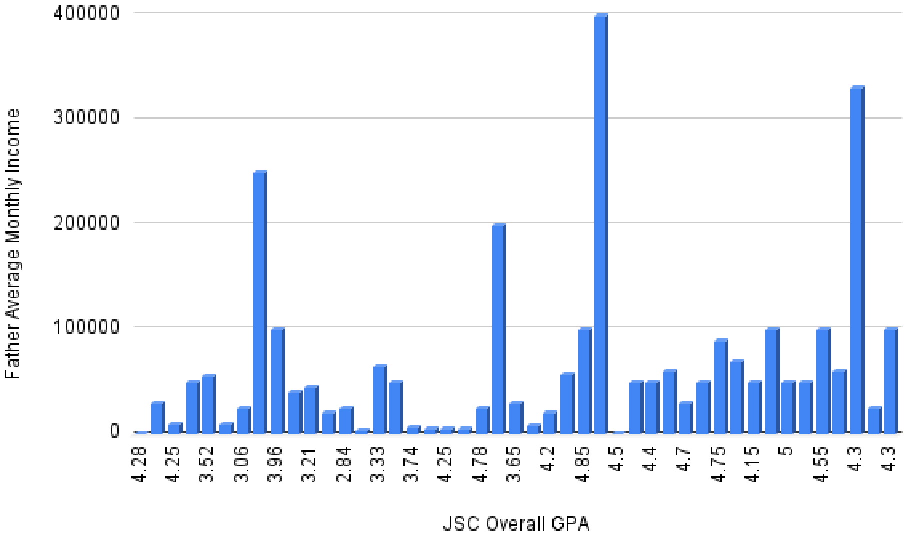


Figure 3.2: Exploratory Data Analysis on students GPA with respect to Fathers Average Monthly Income.

sults of the JSC (Junior School Certificate) examination. This contradicts the common notion that a father's income affects academic performance. Surprisingly, our dataset does not exhibit any discernible impact of father's income on JSC results.

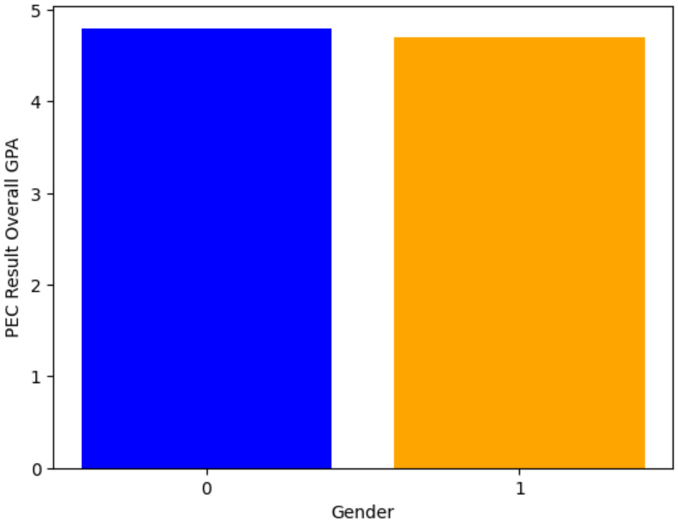


Figure 3.3: Exploratory Data Analysis on Gender distribution with respect to PEC Result Overall GPA.

The findings presented in Figure 3.3 reveal that there is a similarity in the PEC (Primary Education Completion) overall GPA between male and female students. However, it is worth noting that female students exhibit slightly better performance compared to their male counterparts. Although the overall GPAs are similar, the data suggests a marginal advantage for female students in terms of academic achievement.

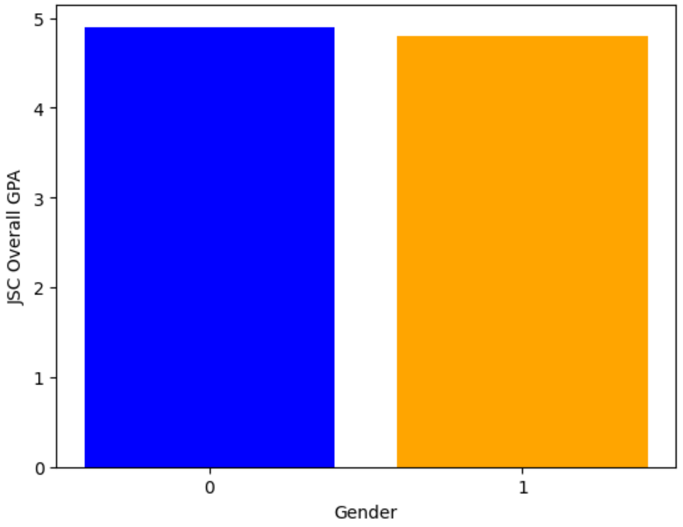


Figure 3.4: Exploratory Data Analysis on Gender distribution with respect to JSC Result Overall GPA.

The data presented in Figure 3.4 demonstrates a comparable JSC (Junior School Certificate) overall GPA between male and female students. Notably, female students have a slight edge over their male counterparts in terms of academic performance. Although the overall GPAs are in close proximity, the findings suggest a slight advantage for female students in terms of their JSC results.

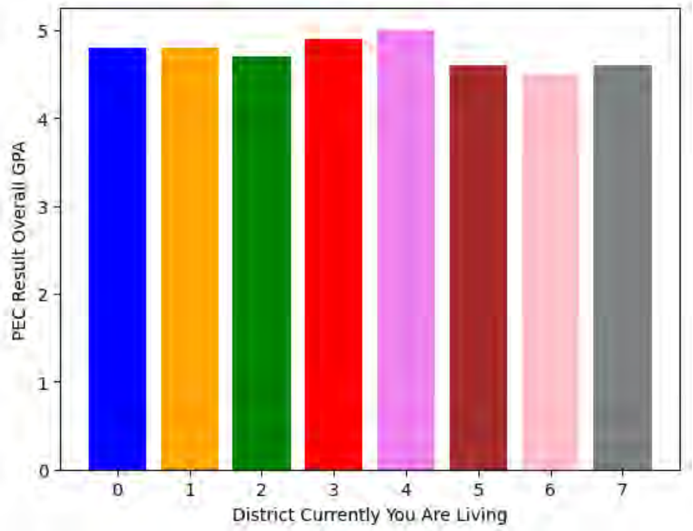


Figure 3.5: Exploratory Data Analysis on PEC overall results based on the district that students live in.

The results depicted in Figure 3.5 and Figure 3.6 exhibit remarkably similar patterns regarding the PEC (Primary Education Completion) and JSC (Junior School Certificate) overall GPAs based on the district where students currently reside. The data indicates a high degree of consistency between the two figures. The performance of students, as reflected in their overall GPAs, appears to be influenced by the district they live in. This correlation holds true for both PEC and JSC results, suggesting a significant impact of the district on students' academic achievements at both stages.

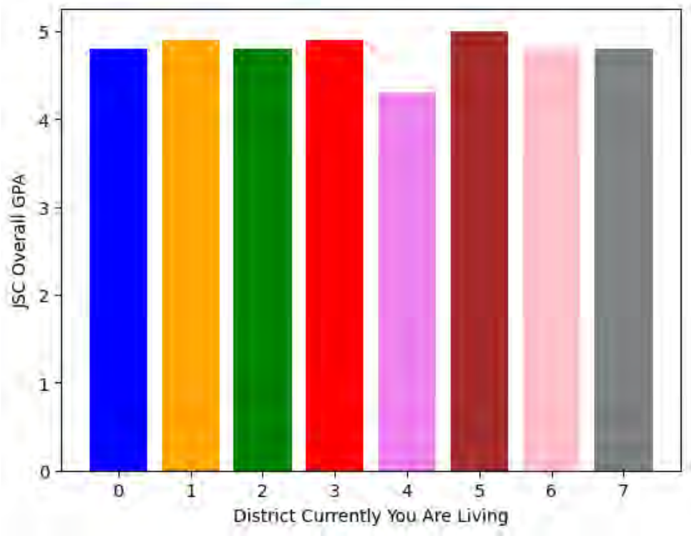


Figure 3.6: Exploratory Data Analysis on JSC overall results based on the district that students live in.

In summary, the figures (3.1 to 3.6) illustrate that the number of female data records is slightly higher, father's average monthly income has no impact on JSC overall results, female students perform slightly better than male students in PEC and JSC overall GPA, and the district that the students currently live in has no significant impact on the PEC overall GPA and JSC overall GPA.

Chapter 4

Methodology

This study intends to alleviate the difficulty of stream selection for Bangladeshi ninth-grade students. Machine learning algorithms, such as linear regression, support vector machine regression, random forest regression, adaptive boosting, gradient boosting, and extreme gradient boosting, are applied in the methodology. The study makes use of information from more advanced students who have already gone through the stream selection procedure. The goal of the study is to pinpoint crucial elements and characteristics that might greatly increase the predictability of stream selection. The results of this study will help improve decision-making and ensure that students successfully transition into the educational programs they want to pursue.

We first collected data from three streams: science, business, and humanities. We then performed data processing techniques, including null value removal/handling, data type handling, cleaning, and data normalization. Data scaling was necessary for our dataset because the data lacked diversity, and most students had similar GPAs. Scaling was also necessary to prevent machine learning algorithms from being biased and improve the convergence of the implemented models in our study. We used the mRMR method for feature selection. At the end of our study, we compared the results with all features we collected and the features derived from mRMR. We divided our dataset into two segments: training data was used to train the machine learning models while testing data was used to evaluate the trained model. Our study used several machine learning algorithms, including linear regression, support vector machine regression, random forest regression, adaptive boosting, gradient boosting, and extreme gradient boosting. We evaluated our models using metrics such as mean squared error, mean absolute percentage error, explained variance, mean Poisson deviance, mean gamma deviance, and R2 score. After comprehensively comparing the models, we exported the best-fitted model. Finally, we employed LIME [14], which explains a model's prediction. We predicted the GPA of individual students in each of the streams and made a decision accordingly to select a stream. Figure 4.1 represents the overall proposed model of this study.

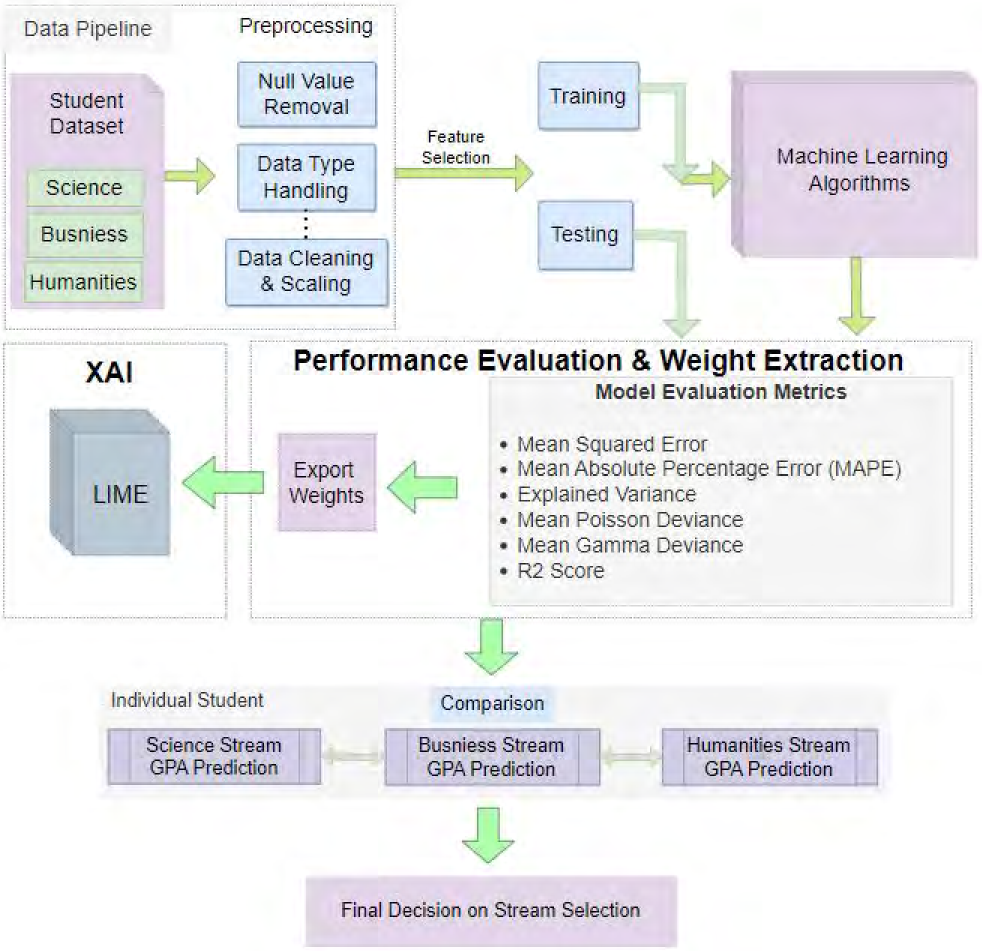


Figure 4.1: Proposed framework for predicting the best separate stream for secondary school students in Bangladesh.

4.1 Feature Selection

4.1.1 Minimum Redundancy and Maximum Relevance

Using the relationship between a feature and the target class, the Minimum Redundancy and Maximum Relevance (mRMR)[11] filter method chooses features that are most relevant to the target classes. A method of estimating that seeks to maximize the dependence between the joint distribution of the class label and the chosen features is known as mRMR. mRMR employs mutual knowledge within a process to satisfy optimization criteria. The formula for mRMR are as follows

$$mRMR(x) = R(x, y) - \frac{1}{k} \sum_{x' \in S} D(x, x') \tag{4.1}$$

4.2 Model Specification

4.2.1 Linear Regression

A case model with only one independent variable is called simple linear regression. Simple linear regression identifies a variable's dependence.

$$y = \beta_0 + \beta_1 x + \epsilon \tag{4.2}$$

Simple regression [24] can tell the difference between how the dependent variables affect each other and how the independent variables affect each other.

4.2.2 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 [28]. It is possible to formulate the SVM regression[22] problem as follows:

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i \tag{4.3}$$

subject to:

$$y_i - w^T x_i - b \leq \epsilon + \xi_i \tag{4.4}$$

$$w^T x_i + b - y_i \leq \epsilon + \xi_i \tag{4.5}$$

$$\xi_i \geq 0, i = 1, 2, ..., n \tag{4.6}$$

4.2.3 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 [28]. A random forest regression [22] model's prediction can be calculated as follows:

$$\hat{y} = \frac{1}{M} \sum_{i=1}^{M} f_i(x) \tag{4.7}$$

4.2.4 Adaptive Boosting

AdaBoost, short for Adaptive Boosting, is a powerful ensemble learning algorithm that is used to improve the performance of weak learners [4]. 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. An AdaBoost model's [2] prediction can be calculated as follows:

$$\hat{y} = \sum_{i=1}^{M} w_i f_i(x) \tag{4.8}$$

4.2.5 Gradient Boosting

Gradient Boosting combines the predictions of multiple weak learners to create a stronger model [8]. 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. A Gradient Boosting Regression model's prediction can be calculated as follows:

$$\hat{y} = \sum_{m=1}^{M} \gamma_m h_m(x) \tag{4.9}$$

4.2.6 Extreme Gradient Boosting

XGBoost is a gradient boosting algorithm that uses decision trees as its base model [10]. 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. An XGBoost regression model's prediction can be mathematically represented as follows:

$$\hat{y}i = \sum k = 1^K f_k(x_i) \tag{4.10}$$

Chapter 5

Experimental Results and Discussion

The details of how our methodology was used will be covered in this part, followed by a review of the outcomes it generated.

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.

5.1 Performance Metrics

5.1.1 Mean Squared Error

Mean Squared Error (MSE) is a commonly used loss function for regression problems [13]. 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 4.1 represents the mean squared error expression.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2 \tag{5.1}$$

5.1.2 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 [1]. 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:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5.2}$$

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.

5.1.3 Mean Absolute Percentage Error (MAPE)

Mean Absolute Percentage Error as a performance metric [31]. 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:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \tag{5.3}$$

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.

5.1.4 Mean Poisson Deviance

Mean Poisson Deviance [30] 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:

$$\text{Deviance} = 2\sum_{i=1}^{n}\left(y_i \log\left(\frac{y_i}{\hat{y}_i}\right) - (y_i - \hat{y}_i)\right) \tag{5.4}$$

where y_i is the observed count, y_i 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.

5.1.5 Mean Gamma Deviance

Mean Gamma Deviance [30] 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:

$$Deviance = 2\sum_{i=1}^{n} y_i \log\left(\frac{y_i}{\hat{y}_i}\right) - \left(\frac{y_i}{\hat{y}_i}\right) + \log(\hat{y}_i) \tag{5.5}$$

where y_i is the observed value, y_i is the predicted value, and the summation is taken over all observations.

5.1.6 R2 Score

The R-squared (R2) score [19] 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:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5.6}$$

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.

5.1.7 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 [13]. 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

Table 5.1: Performance Evaluation for Science Stream

ML Algorithm	Mean Squared Error	Mean Absolute % Error	Explained Variance	Mean Poisson Deviance	Mean Gamma Deviance	R2 Score
SVM Regression	0.25433	0.04338	0.33816	0.01405	0.00305	0.30825
RF Regression	0.25920	0.04127	0.28862	0.01475	0.00324	0.28151
Linear Regression	0.29744	0.05182	0.13887	0.01892	0.00406	0.05386
ADA Boost Regression	0.29495	0.05179	0.06965	0.01913	0.00422	0.06962
Extreme Gradient Boosting	0.25562	0.04092	0.30782	0.01430	0.00313	0.30121
Gradient Boosting Regression	0.24740	0.04344	0.34561	0.01327	0.00288	0.34540

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.

5.2 Performance Evaluation

We have done our experiment by using Jupyter Notebook and Python code for each stream separately, with separate datasets. In our research, 80% of the data is used for training and 20% is for testing. We have utilized six different machine learning models for this study. Our findings for each of the streams is presented below.

5.2.1 Test Cases

Figure 5.1 shows the full scenario and how our proposed model will work. After feeding data into the proposed model, it predicts each student's GPA for each stream. The highest GPA for any stream will be chosen as the student's proposed stream.

5.2.2 Science Stream

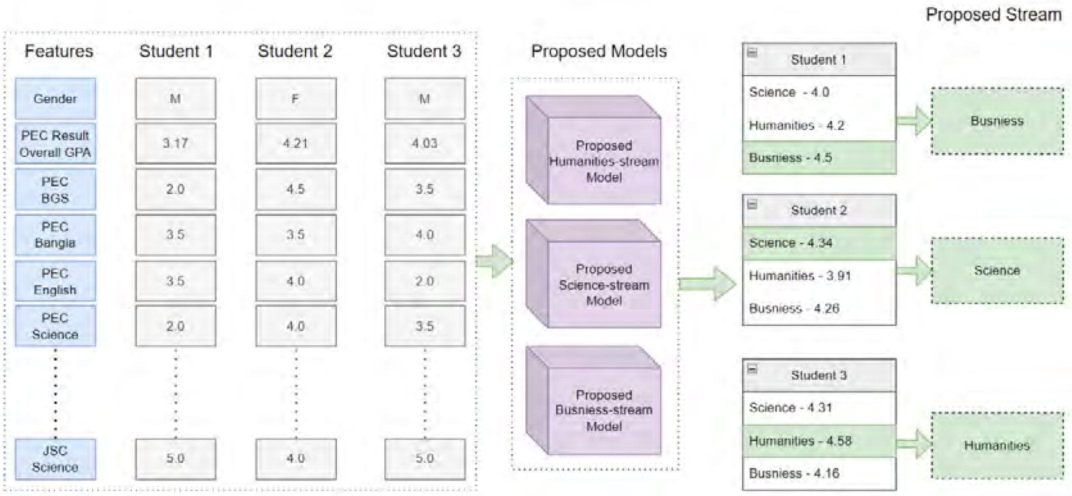
Table 5.1 represents our findings in the science stream. Here, Linear Regression has the highest Mean Squared Error and Mean Absolute Percentage Error. Random Forest Regression and Support Vector Machine Regression however performed moderately well. In this stream, Gradient Boosting Regressor managed to gain the lowest mean squared error. However, it has a slightly higher MAPE and R2 score.

Table 5.2: Performance Evaluation for Business Studies stream

ML Algorithm	Mean Squared Error	Mean Absolute % Error	Explained Variance	Mean Poisson Deviance	Mean Gamma Deviance	R2 Score
RF Regression	0.384224	0.068950	0.514893	0.036536	0.009134	0.486378
Linear Regression	0.429913	0.078906	0.357799	0.044692	0.010942	0.356961
Gradient Boosting Regression	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 Regression	0.416501	0.076927	0.457631	0.042029	0.010665	0.416501
SVM Regression	0.365942	0.069041	0.558916	0.033906	0.008706	0.534092

Table 5.3: Performance Evaluation for Humanities stream

ML Algorithm	Mean Squared Error	Mean Absolute % Error	Explained Variance	Mean Poisson Deviance	Mean Gamma Deviance	R2 Score
SVM Regression	0.37696	0.08205	0.53331	0.03994	0.01156	0.53331
RF Regression	0.25613	0.05606	0.80324	0.01868	0.00551	0.78453
Linear Regression	0.26285	0.07118	0.77650	0.02116	0.00658	0.773084
Gradient Boosting Regressor	0.27021	0.05697	0.78640	0.01991	0.00558	0.76020
ADA Boost Regression	0.27765	0.06060	0.77680	0.02445	0.00807	0.746813
Extreme Gradient Boosting	0.24468	0.05542	0.81729	0.01763	0.00536	0.80337



Features	Student 1	Student 2	Student 3
Gender	M	F	M
PEC Result Overall GPA	3.17	4.21	4.03
PEC BGS	2.0	4.5	3.5
PEC Bangla	3.5	3.5	4.0
PEC English	3.5	4.0	2.0
PEC Science	2.0	4.0	3.5
⋮	⋮	⋮	⋮
JSC Science	5.0	4.0	5.0

Figure 5.1: Test Cases for proposed Model

In terms of explained variance, mean poisson deviance and mean gamma deviance it has a moderate score. Extreme gradient boosting has a high score for mean squared, Mean Poisson Deviance, and Mean Gamma Deviance in comparison with the Gradient Boosting regressor. Considering all the metrics our findings state that, Gradient Boosting Regressor is the better performing model in the stream.

5.2.3 Business Studies Stream

Our insights in the business stream are presented in Table 5.2. Although Gradient Boosting was the best performing model in the science stream, in this stream despite having a moderate MAPE score, it has the second 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.

5.2.4 Humanities Stream

In the Humanities stream, Table 5.3 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.

5.3 LIME Use Case Scenario

While utilizing LIME we have used our best fitted model. Figure 5.2 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

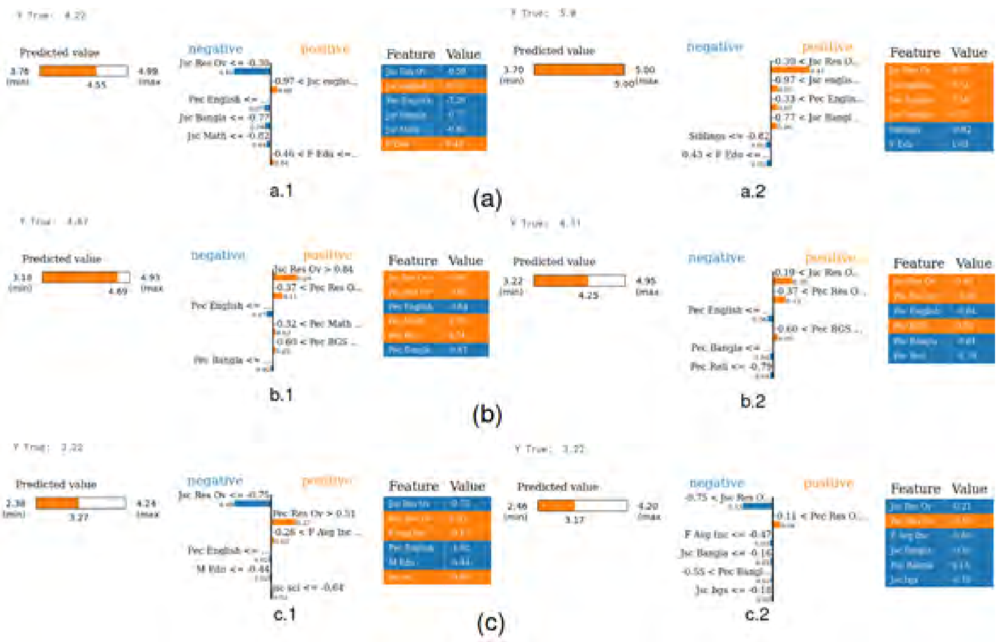


Figure 5.2: Local Interpretable Model Agnostic Explanations prediction. Here, science stream is represented by (a), business stream by (b) and finally humanities stream by (c).

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 4.67 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.

5.4 Factor Analysis

After utilizing mRMR, we have the top ten features: JSC Mathematics, Mother Profession, District Currently you are living, PEC Religion, JSC English, JSC Science, JSC Overall GPA, Father Highest Academic Qualification, JSC BGS, JSC Bangla. Table 5.4 provides insightful findings regarding the performance of different streams using mRMR-selected features. The results indicate that the science stream exhibits lower scores for R2, MSE, and Explained Variance, suggesting suboptimal performance in these metrics when utilizing gradient boosting. However, other metrics show minor changes in scores. Similarly, the Business Studies stream shows relatively lower scores across various metrics, except for MAPE and Mean Pois-

Table 5.4: Performance Evaluation after Utilizing mRMR

ML Algorithm	Mean Squared Error	Mean Absolute % Error	Explained Variance	Mean Poisson Deviance	Mean Gamma Deviance	R2 Score
Gradient Boosting Regression -Science	**0.23462**	0.507585	**0.24632**	0.01542	0.00334	**0.23787**
SVM Regression -Business Studies	**0.32688**	0.08450	**0.42429**	0.04401	**0.01110**	**0.38500**
Extreme Gradient Boosting-Humanities	**0.19279**	0.05603	**0.80646**	0.01787	**0.00528**	**0.79491**

son Deviance, where Support Vector Machine Regression performs comparatively better. On the other hand, the Humanities stream demonstrates minimal changes in MAPE and Mean Poisson Deviance, while other metrics indicate lower scores specifically in relation to Extreme Gradient Boosting. These observations shed light on the distinct performance patterns exhibited by different streams, providing valuable insights into their respective strengths and weaknesses when considering the mRMR-selected features.

Chapter 6

Conclusion

In conclusion, this research addressed the challenge of stream selection in the education system of Bangladesh by developing a machine learning model based on a comprehensive dataset from various schools. By considering 26 features and employing regression algorithms, we achieved superior performance in predicting stream outcomes. The gradient boosting algorithm performed well for the Science stream, while Support Vector Machine regression excelled in predicting results for the Business stream. Extreme gradient boosting showed excellent results for the Humanities stream. The use of the LIME technique enhanced the interpretability of our models. These findings contribute to informed decision-making for stream selection, aligning students' choices with their future goals and aspirations in the education system of Bangladesh.

6.1 Limitations

There are several restrictions on this thesis and more research has to be taken into account. The dataset utilized, which is modest in size and has a low variety, may not accurately reflect every student in Bangladesh, among the shortcomings. Additionally, the information predominantly focuses on records from elementary and middle schools, leaving out crucial elements like student hobbies and extracurricular activities that affect stream choice.

6.2 Future Works

There are various ways to build on this study in future work. The model's ability to predict stream selection can be improved by implementing deep learning techniques like CNN and RNN. The findings will also be more generalizable if the dataset is expanded to include a bigger volume of data from other schools and areas. To have a fuller picture of how decisions are made, it is also crucial to take into account other significant factors including students' hobbies and professional goals. It is possible to do more research to improve the stream selection accuracy of machine learning models, which will help stakeholders and students make well-informed judgments.

Bibliography

[1] K. E. O'Grady, "Measures of explained variance: Cautions and limitations.," *Psychological Bulletin*, vol. 92, no. 3, p. 766, 1982.

[2] G. Ridgeway, D. Madigan, and T. S. Richardson, "Boosting methodology for regression problems," in *Seventh International Workshop on Artificial Intelligence and Statistics*, PMLR, 1999.

[3] M. Karagiannopoulos, D. Anyfantis, S. Kotsiantis, and P. Pintelas, "Feature selection for regression problems," *Educational Software Development Laboratory, Department of Mathematics, University of Patras, Greece*, 2004.

[4] D. P. Solomatine and D. L. Shrestha, "Adaboost. rt: A boosting algorithm for regression problems," in *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541)*, IEEE, vol. 2, 2004, pp. 1163–1168.

[5] P. Cortez and A. M. G. Silva, "Using data mining to predict secondary school student performance," 2008.

[6] M. Ramaswami and R. Bhaskaran, "A chaid based performance prediction model in educational data mining," *arXiv preprint arXiv:1002.1144*, 2010.

[7] 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*, 2012, pp. 1–7.

[8] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in neurorobotics*, vol. 7, p. 21, 2013.

[9] M. Doshi, "Correlation based feature selection (cfs) technique to predict student perfromance," *International Journal of Computer Networks & Communications*, vol. 6, no. 3, p. 197, 2014.

[10] T. Chen, T. He, M. Benesty, *et al.*, "Xgboost: Extreme gradient boosting," *R package version 0.4-2*, vol. 1, no. 4, pp. 1–4, 2015.

[11] N. Rachburee and W. Punlumjeak, "A comparison of feature selection approach between greedy, ig-ratio, chi-square, and mrmr in educational mining," in *2015 7th international conference on information technology and electrical engineering (ICITEE)*, IEEE, 2015, pp. 420–424.

[12] N.-B. Sara, R. Halland, C. Igel, and S. Alstrup, "High-school dropout prediction using machine learning: A danish large-scale study.," in *ESANN*, vol. 2015, 2015, 23rd.

[13] A. De Myttenaere, B. Golden, B. Le Grand, and F. Rossi, "Mean absolute percentage error for regression models," *Neurocomputing*, vol. 192, pp. 38–48, 2016.

[14] M. T. Ribeiro, S. Singh, and 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, pp. 1135–1144.

[15] 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 1st International Conference on E-Education, E-Business and E-Technology*, 2017, pp. 36–41.

[16] N. Shahadat, M. Rahman, S. Ahmed, and B. Rahman, “Predicting higher secondary results by data mining algorithms with vbr: A feature reduction method,” in *2017 4th International Conference on Advances in Electrical Engineering (ICAEE)*, IEEE, 2017, pp. 164–169.

[17] J. Uddin, F. N. Arko, N. Tabassum, T. R. Trisha, and 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)*, IEEE, 2017, pp. 1–4.

[18] R. A. Khan, J. Uddin, S. Corraya, and 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, 2018.

[19] 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)*, IEEE, 2019, pp. 1–5.

[20] M. N. I. Sarker, M. Wu, and M. A. Hossin, “Economic effect of school dropout in bangladesh,” *International journal of information and education technology*, vol. 9, no. 2, pp. 136–142, 2019.

[21] O. El Aissaoui, Y. El Alami El Madani, L. Oughdir, A. Dakkak, and Y. El Allioui, “A multiple linear regression-based approach to predict student performance,” in *Advanced Intelligent Systems for Sustainable Development (AI2SD’2019) Volume 1-Advanced Intelligent Systems for Education and Intelligent Learning System*, Springer, 2020, pp. 9–23.

[22] L. H. Alamri, R. S. Almuslim, M. S. Alotibi, D. K. Alkadi, I. Ullah Khan, and N. Aslam, “Predicting student academic performance using support vector machine and random forest,” in *2020 3rd International Conference on Education Technology Management*, 2020, pp. 100–107.

[23] 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 *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, IEEE, 2020, pp. 1–6.

[24] D. Maulud and A. M. Abdulazeez, “A review on linear regression comprehensive in machine learning,” *Journal of Applied Science and Technology Trends*, vol. 1, no. 4, pp. 140–147, 2020.

[25] M. R. Megha, *No science, arts or commerce in secondary education: A good idea?* Nov. 2020. [Online]. Available: https://www.thedailystar.net/shout/news/no-science-arts-or-commerce-secondary-education-good-idea-2000413.

[26] 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*, vol. 11, no. 3, 2020.

[27] 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)*, vol. 19, no. 1, 2021.

[28] 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*, vol. 26, no. 2, pp. 1527–1547, 2021.

[29] 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*, vol. 47, no. 2, pp. 173–190, 2021.

[30] W. Chen, D. Sharifrazi, G. Liang, S. S. Band, K. W. Chau, and A. Mosavi, "Accurate discharge coefficient prediction of streamlined weirs by coupling linear regression and deep convolutional gated recurrent unit," *Engineering Applications of Computational Fluid Mechanics*, vol. 16, no. 1, pp. 965–976, 2022.

[31] 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.

[32] 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)*, IEEE, 2022, pp. 0399–0405.

[33] M. S. I. Khan, N. Islam, J. Uddin, S. Islam, and 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*, vol. 34, no. 8, pp. 4773–4781, 2022.

[34] [Online]. Available: https://dhakaeducationboard.gov.bd/.

[35] *Group prediction by regressor-dataset.xlsx*. [Online]. Available: https://docs.google.com/spreadsheets/d/1Az5vyGnDrzhM_xZzYIGZ1LEmy5a14bR2/edit.