Analysis of Factors Affecting Students' Academic Performance and Score Prediction Using Machine Learning

by

Md Shadat Hossain
ID: CSE2201025133

Kamrul Hasan Hridoy
ID: CSE2201025144

Zanantul Ferdus
ID: CSE2201025112

Raisa Joti
ID: CSE2201025101

Rayhan Ullah
ID: CSE2201025147

Supervised by
Oishika Khair Esha

Submitted in partial fulfilment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SONARGAON UNIVERSITY (SU)

January 2026

APPROVAL

The thesis titled **"Analysis of Factors Affecting Students' Academic Performance and Score Prediction Using Machine Learning"** submitted by **Md Shadat Hossain** (CSE2201025133), **Kamrul Hasan Hridoy** (CSE2201025144), **Zannatul Ferdus** (CSE2201025112), **Raisa Joti**(CSE2201025101) and **Rayhan Ullah** (CSE2201025147) to the Department of Computer Science and Engineering, Sonargaon University (SU), has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering and approved as to its style and contents.

Board of Examiners

Oishika Khair Esha — **Supervisor**
Lecturer
Department of Computer Science and Engineering
Sonargaon University (SU)

(Examiner Name and Signature) — **Examiner 1**
Department of Computer Science and Engineering
Sonargaon University (SU)

(Examiner Name and Signature) — **Examiner 2**
Department of Computer Science and Engineering
Sonargaon University (SU)

(Examiner Name and Signature) — **Examiner 3**
Department of Computer Science and Engineering
Sonargaon University (SU)

DECLARATION

We, hereby, declare that the work presented in this report is the outcome of the investigation performed by us under the supervision of **Oishika Khair Esha,** Lecturer**,** Department of Computer Science and Engineering, Sonargaon University, Dhaka, Bangladesh. We reaffirm that no part of this thesis has been or is being submitted elsewhere for the award of any degree.

Countersigned

(Oishika Khair Esha)
Supervisor

Signature

Md Shadat Hossain
ID:CSE220125133

Kamrul Hasan Hridoy
ID:CSE220125144

Zannatul Ferdus
ID:CSE220125112

Raisa Joti
ID:CSE220125101

Rayhan Ullah
ID:CSE220125147

ABSTRACT

In recent years, the increasing availability of educational data has created new opportunities for applying Machine Learning (ML) techniques to analyze and predict students' academic performance. Understanding the factors that influence student achievement is crucial for educators, institutions, and policymakers to improve learning outcomes and design effective academic interventions. This research focuses on analyzing how students' exam performance is affected by various demographic and educational factors using machine learning approaches.

The dataset used in this study was collected from a publicly available educational dataset consisting of 1,000 student records with demographic attributes such as gender, race or ethnicity, parental level of education, lunch type, and test preparation course completion, along with scores in mathematics, reading, and writing. Comprehensive data preprocessing techniques, including categorical encoding, feature scaling, and train–test splitting, were applied to prepare the dataset for modeling.

Exploratory Data Analysis (EDA) was conducted to identify patterns, correlations, and trends among the variables. Several machine learning models were then implemented and evaluated to predict students' academic performance, including traditional regression-based models and ensemble learning techniques. Model performance was assessed using standard evaluation metrics such as accuracy, mean squared error, and coefficient of determination (R^2).

The experimental results demonstrate that machine learning models can effectively capture the relationships between demographic factors and academic outcomes. The proposed approach achieved high predictive performance, highlighting the significant influence of test preparation courses, parental education level, and lunch type on students' exam scores. The findings suggest that ML-based predictive systems can serve as valuable decision-support tools in educational settings.

This study emphasizes the potential of machine learning to enhance educational analytics by providing data-driven insights into student performance, supporting early intervention strategies, and contributing to the development of personalized and inclusive learning environments.

ACKNOWLEDGMENT

At the very beginning, we would like to express our deepest gratitude to the Almighty Allah for granting us the strength, patience, and wisdom to successfully complete this research work within the scheduled time.

We feel truly privileged to have received the guidance and supervision of **Oishika Khair Esha** ,Lecturer, Department of Computer Science & Engineering, Sonargaon University. Their continuous support, valuable advice, and constructive feedback played a vital role in the successful completion of this project. Their encouragement and direction served as essential guidance throughout the research process.

We would also like to express our sincere appreciation to all the respected teachers of the Department of Computer Science & Engineering, Sonargaon University, for their continuous motivation, academic support, and for enriching our educational journey with knowledge and inspiration.

Finally, we would like to convey our heartfelt gratitude and deepest appreciation to our parents for their unconditional love, constant encouragement, and moral support, which have been a continuous source of strength throughout our academic life.

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ML	Machine Learning
CSV	Comma-Separated Values
EDA	Exploratory Data Analysis
LR	Linear Regression
LASSO	Least Absolute Shrinkage and Selection Operator
RR	Ridge Regression
KNN	K-Nearest Neighbors
DT	Decision Tree
RF	Random Forest
XGB	Extreme Gradient Boosting
CB	CatBoost
AB	AdaBoost
MAE	Mean Absolute Error
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
R^2	Coefficient of Determination
TTS	Train-Test Split

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 INTRODUCTION

1.1 Background of the Study

Education is widely recognized as a fundamental pillar for social and economic development because it enhances human capital, improves employability, and supports long-term economic growth. Students' academic performance plays a crucial role in shaping their future opportunities, career paths, and personal development, as academic results often determine access to higher education and professional careers. In recent years, educational institutions have increasingly adopted data-driven approaches to evaluate student performance due to the limitations of traditional evaluation methods. The rapid advancement of technology has led to the generation of large volumes of educational data, creating opportunities for advanced analytical techniques.

Machine Learning (ML), a subset of Artificial Intelligence (AI), enables systems to learn patterns from data and make accurate predictions without explicit programming. ML has gained significant attention in the education sector because it can analyze complex relationships among multiple factors affecting student performance. The availability of structured and publicly accessible educational datasets on platforms such as **Kaggle** enables reliable, transparent, and reproducible research. By applying ML techniques to such datasets, educators and researchers can identify academic trends, predict performance, and detect at-risk students, thereby supporting targeted interventions and improved learning outcomes.

1.2 Problem Statement

Students' academic performance is influenced by a combination of demographic, social, and educational factors. Traditional evaluation methods often fail to capture the complex interactions among these variables. Factors such as **gender, race or ethnicity, parental level of education, type of lunch**, and **participation in test preparation courses** can have a significant impact on students' exam scores. However, understanding the relative importance of these factors and their combined effect remains a challenge.

The problem addressed in this research is to develop an effective machine learning-based system that can analyze these factors and accurately predict students' academic performance. Such a predictive system can assist educational institutions in making informed decisions and improving academic planning.

1.3 Motivation

The motivation behind this study arises from the growing need to improve educational outcomes through data-driven decision-making. Early identification of students who may struggle academically allows institutions to provide timely support and personalized learning strategies. Machine learning models offer a powerful means to predict performance trends and uncover hidden patterns within educational datasets.

Moreover, understanding the influence of socio-economic and demographic factors can help reduce educational disparities and promote fairness in academic evaluation. This research aims to contribute to the development of intelligent educational systems that support both students and educators.

1.4 Objectives of the Study

The main objectives of this research are as follows:

1. To analyze the impact of demographic and educational factors on students' academic performance.
2. To apply and compare multiple machine learning algorithms for predicting students' exam scores.
3. To evaluate model performance using standard regression metrics such as MAE, RMSE, and R^2 score.
4. To identify the most effective model for student performance prediction based on experimental results.

1.5 Scope of the Study

This study focuses on predicting students' academic performance using a structured dataset containing demographic attributes and exam scores. The scope of the research includes:

- Exploratory Data Analysis (EDA) to understand data distribution and relationships.
- Data preprocessing, including feature encoding and scaling.
- Implementation of multiple regression-based machine learning models.
- Performance evaluation and comparative analysis of models.

The study is limited to the available features in the dataset and does not consider external factors such as psychological conditions, teaching quality, or institutional infrastructure.

1.6 Significance of the Study

This research holds significance in several aspects. It demonstrates the practical application of machine learning techniques in educational analytics and provides insights into the factors affecting academic performance. The findings can assist educators and policymakers in designing data-driven strategies to enhance student learning outcomes. Additionally, the proposed approach can be extended to other educational datasets and contexts.

1.7 Organization of the Thesis

The remainder of this thesis is organized as follows:

- **Chapter 1: Introduction** – Introduces the research background, objectives, problem statement, and

  significance of the study.

- **Chapter 2: Data Description and Exploratory Data Analysis (EDA) – Presents the dataset, feature**

  descriptions, and preliminary analysis of the data.

  .

- **Chapter 3: Machine Learning Models** – Explains the algorithms used and model implementation details.
- **Chapter 4: Results and Discussion** – Presents experimental results, model comparisons, and analysis.

- **Chapter 5: Model Evaluation and Analysis** – Provides in-depth evaluation of model performance,

  ensemble learning impact, and best model selection.

- **Chapter 6: Limitations and Ethical Considerations** – Discusses the study's limitations, ethical

  issues, and practical considerations for responsible use of predictive models.

- **Chapter 7: Conclusion and Future Work** – Summarizes the research findings, highlights key

  contributions, and suggests directions for future research in student performance prediction.

CHAPTER 2 DATASET DESCRIPTION AND EXPLORATORY DATA ANALYSIS

2.1 Dataset Description

The dataset used in this research was obtained from a publicly available source on Kaggle titled **"Students Performance in Exams"**. The dataset contains academic and demographic information of students and is widely used for educational data analysis and performance prediction tasks.

The dataset consists of **1000 student records** with **8 attributes**, representing both categorical and numerical features. It provides exam scores along with background information, which makes it suitable for analyzing the factors that influence students' academic performance.

Dataset Summary

- **Total number of instances:** 1000
- **Total number of features:** 8
- **Categorical features:** 5
- **Numerical features:** 3
- **Missing values:** None

The dataset is clean and well-structured, which allows effective application of machine learning algorithms without extensive data cleaning.

2.2 Feature Description

The dataset contains demographic, socio-economic, and academic attributes. A brief description of each feature is presented in Table 2.1.

Table 2.1: Description of Dataset Features

Feature Name	Description
gender	Gender of the student (male/female)
race/ethnicity	Student's ethnic group (Group A to Group E)
parental level of education	Highest education level of the student's parents
lunch	Type of lunch received by the student (standard/reduced)
test preparation course	Whether the student completed a test preparation course
math score	Score obtained in Mathematics exam
reading score	Score obtained in Reading exam
writing score	Score obtained in Writing exam

In this study, the exam scores (math, reading, and writing) are considered as the target variables for performance prediction, while the remaining features are used as input variables.

2.3 Data Preprocessing

Data preprocessing is a crucial step in machine learning to ensure that the data is suitable for modeling. The following preprocessing steps were applied in this study:

2.3.1 Handling Missing Values

The dataset was examined for missing or null values. No missing values were found, so no imputation techniques were required.

2.3.2 Encoding Categorical Variables

Machine learning models require numerical inputs. Therefore, categorical features such as gender, race/ethnicity, parental level of education, lunch type, and test preparation course were transformed into numerical representations using one hot encoding techniques.

2.3.3 Feature Scaling

Feature scaling was applied to ensure that numerical variables contribute equally to model training. Standardization techniques were used where necessary to normalize the range of numerical values.

2.3.4 Train–Test Split

To evaluate model performance, the dataset was divided into training and testing sets. **80% of the data** was used for training the models, while the remaining **20%** was used for testing.

2.4 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to understand the distribution of data, identify patterns, and analyze relationships between features and target variables.

2.4.1 Distribution of Exam Scores

Histograms were used to analyze the distribution of math, reading, and writing scores. The results show that the scores are approximately normally distributed, with most students scoring between 60 and 80.

2.4.2 Gender-wise Performance Analysis

A comparison of exam scores based on gender revealed that female students tend to perform better in reading and writing, while male students slightly outperform in mathematics.

2.4.3 Impact of Test Preparation Course

Students who completed the test preparation course achieved significantly higher scores across all subjects compared to those who did not complete the course. This indicates that preparatory programs have a positive impact on academic performance.

2.4.4 Correlation Analysis

A correlation matrix was generated to examine the relationships among numerical features. A strong positive correlation was observed between reading and writing scores, indicating that students who perform well in reading are likely to perform well in writing.

2.5 Insights from Exploratory Data Analysis

Based on the EDA results, the following insights were obtained:

- Test preparation courses have a strong positive influence on student performance.
- Reading and writing scores are highly correlated.
- Parental level of education and lunch type show moderate influence on exam scores.
- Demographic factors contribute to performance variation among students.

These insights highlight the importance of socio-economic and educational factors in predicting academic performance.

2.6 Summary

This chapter presented a detailed description of the dataset, feature explanations, data preprocessing steps, and exploratory data analysis. The findings from EDA provide a strong foundation for applying machine learning algorithms in the next chapter. In the following chapter, various machine learning models and methodologies used for predicting students' academic performance are discussed in detail.

	gender	race_ethnicity	parental_level_of_education	lunch	test_preparation_course
0	female	group B	bachelor's degree	standard	none
1	female	group C	some college	standard	completed
2	female	group B	master's degree	standard	none
3	male	group A	associate's degree	free/reduced	none
4	male	group C	some college	standard	none

lunch	test_preparation_course	math_score	reading_score	writing_score
standard	none	72	72	74
standard	completed	69	90	88
standard	none	90	95	93
free/reduced	none	47	57	44
standard	none	76	78	75

Fig 2.2.1: Samples of dataset in CSV format

Interpretation: The figures above present a selection of sample records from the dataset, illustrating the structure and format of each feature, including demographic and academic attributes. This representation provides a clear understanding of how the data is organized, which is essential for subsequent exploratory data analysis, preprocessing, and the application of machine learning models. By examining these samples, researchers can verify data quality, identify potential issues such as missing values or inconsistent entries, and plan appropriate strategies for feature encoding and scaling, ensuring reliable and accurate model development.

CHAPTER 3
MACHINE LEARNING MODELS AND METHODOLOGY

3.1 Introduction

In this chapter, the methodology for predicting students' academic performance using machine learning algorithms is described. The study implements several regression-based machine learning models to analyze the relationship between demographic and educational features and students' exam scores. This chapter presents the workflow of the predictive modeling process, details of each model used, and the evaluation metrics applied to assess model performance.

3.2 Methodology Overview

The methodology of this research consists of the following steps:

1. **Data Collection:** Using the publicly available dataset from Kaggle, containing 1000 student records with demographic and academic features.
2. **Data Preprocessing:** Handling missing values, encoding categorical variables, feature scaling, and splitting the dataset into training and testing sets (80:20 ratio).

3. **Model Selection:** Implementing multiple regression-based machine learning algorithms to predict exam scores.
4. **Model Training:** Training each model using the training dataset.
5. **Evaluation:** Assessing model performance using standard metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2).
6. **Comparison and Analysis:** Comparing model performance to identify the best-performing algorithm.

Figure 3.1 below illustrates the end-to-end workflow of the predictive modeling process.

3.3 Machine Learning Models Used

The following machine learning models were applied in this study:

3.3.1 Linear Regression (LR)

Linear Regression is a simple regression model that establishes a linear relationship between independent variables and the dependent variable. It serves as a baseline model for predicting students' exam scores.

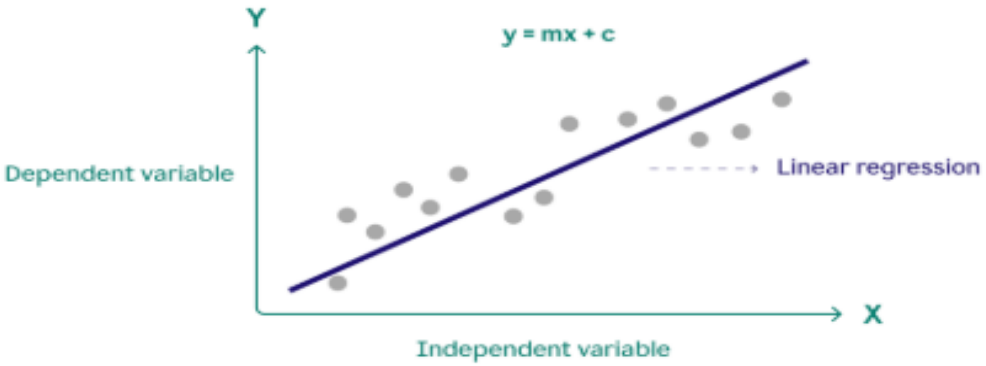

Fig 3.3.1: Linear Regression Working Principle

3.3.2 Lasso Regression

Lasso Regression (Least Absolute Shrinkage and Selection Operator) introduces L1 regularization, which penalizes the absolute size of coefficients. This helps in feature selection and prevents overfitting.

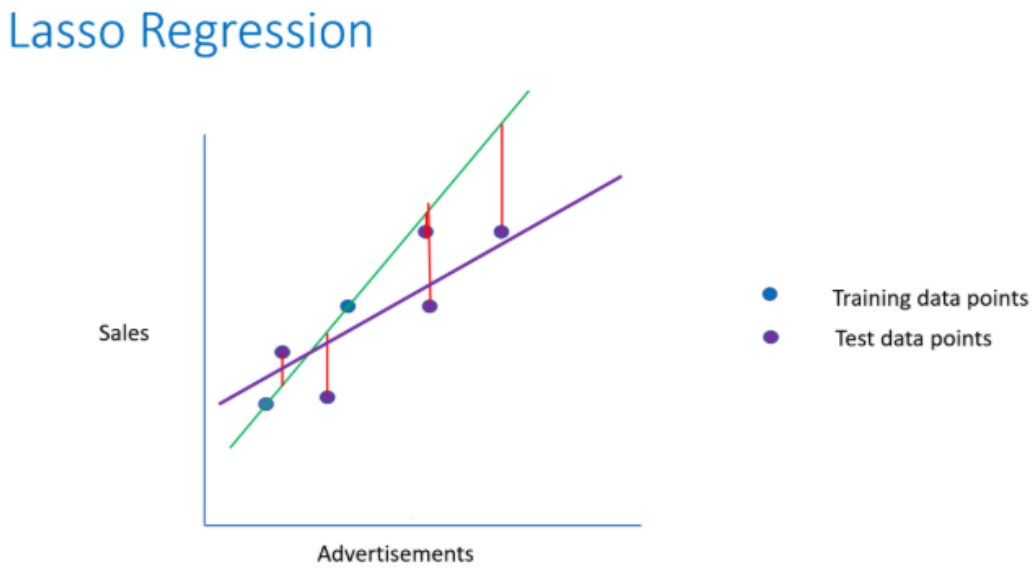

Fig 3.3.2:Lasso Regression Working Principle

3.3.3 Ridge Regression

Ridge Regression applies L2 regularization, penalizing the square of the coefficients to reduce model complexity and overfitting, improving generalization.

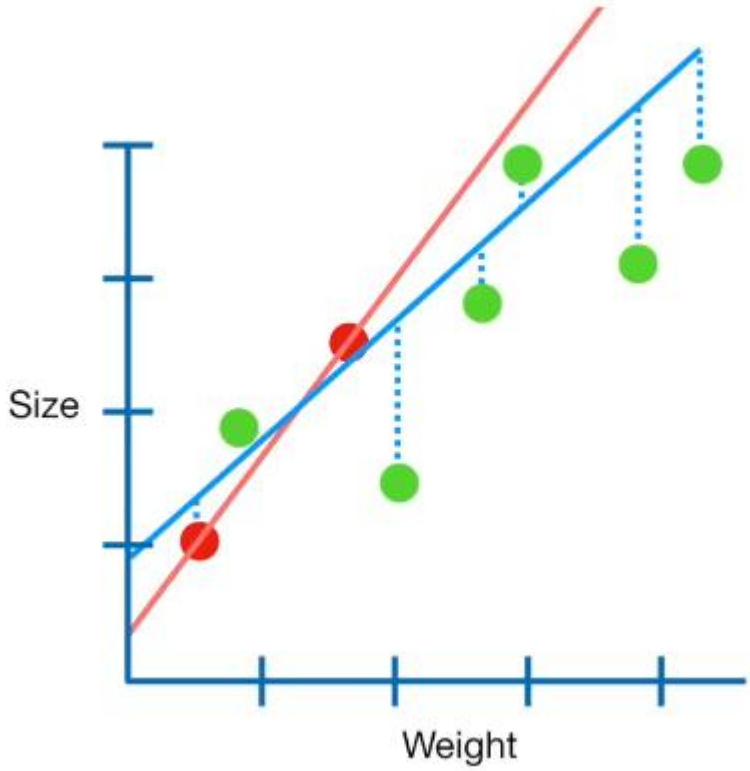

Fig 3.3.3:Ridge Regression Working Principle

3.3.4 K-Nearest Neighbors Regressor (KNN)

KNN Regressor predicts the value of a target variable based on the average of the k-nearest neighbors in the feature space. It is non-parametric and effective for datasets with local patterns.

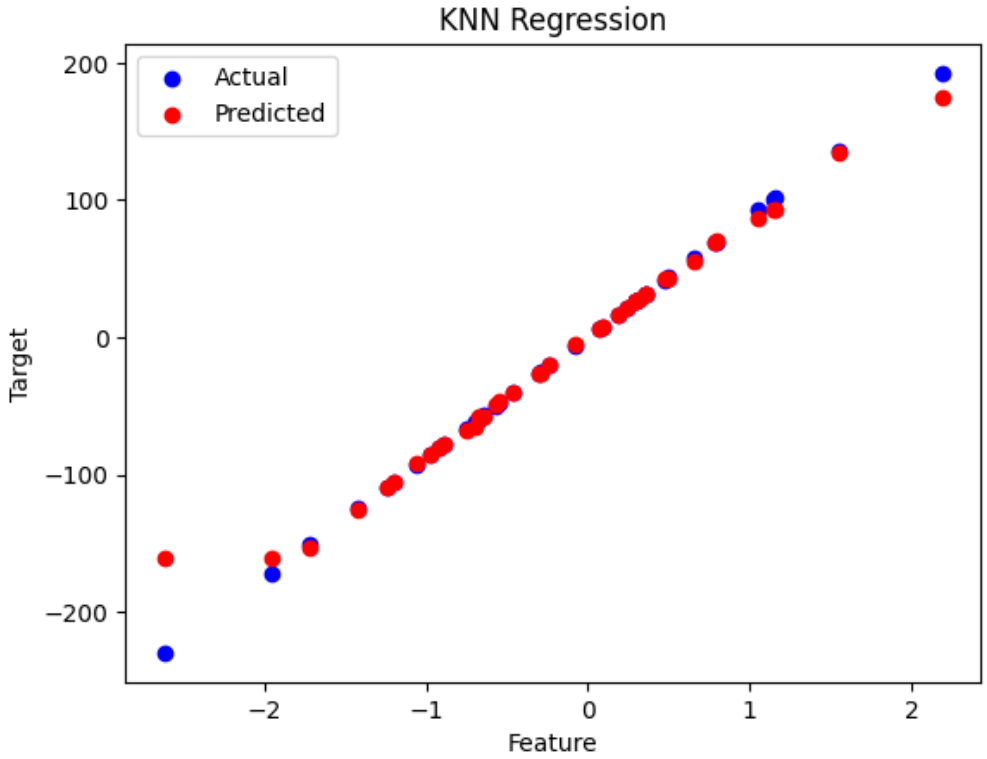

Fig 3.3.4:KNN Regressor Working Principle

3.3.5 Decision Tree Regressor (DT)

Decision Tree Regressor splits the data into subsets based on feature values, forming a tree structure. It captures non-linear relationships between features and exam scores.

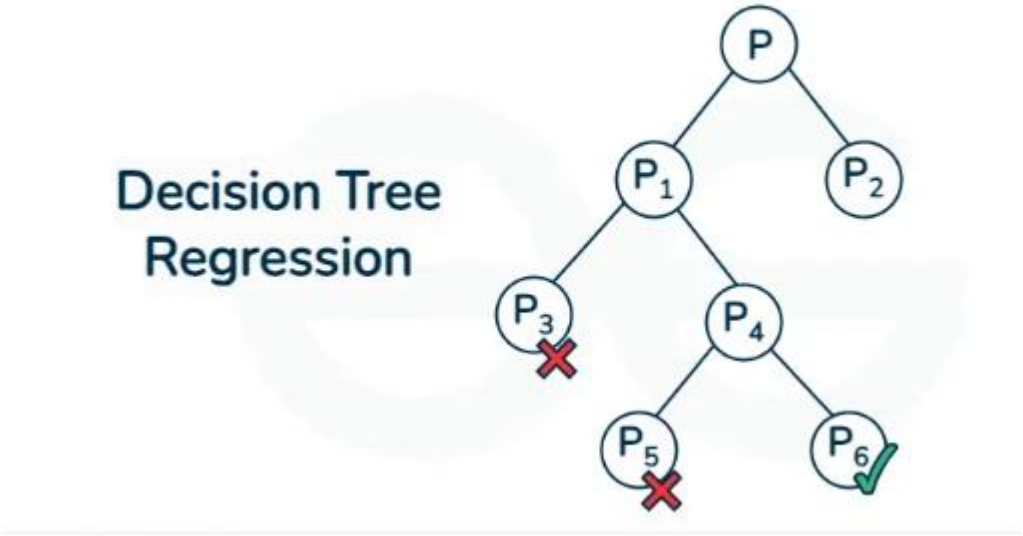

Fig 3.3.5:Decision Tree Regressor Working Principle

3.3.6 Random Forest Regressor (RF)

Random Forest Regressor is an ensemble learning method that constructs multiple decision trees and combines their predictions. It reduces overfitting and improves accuracy.

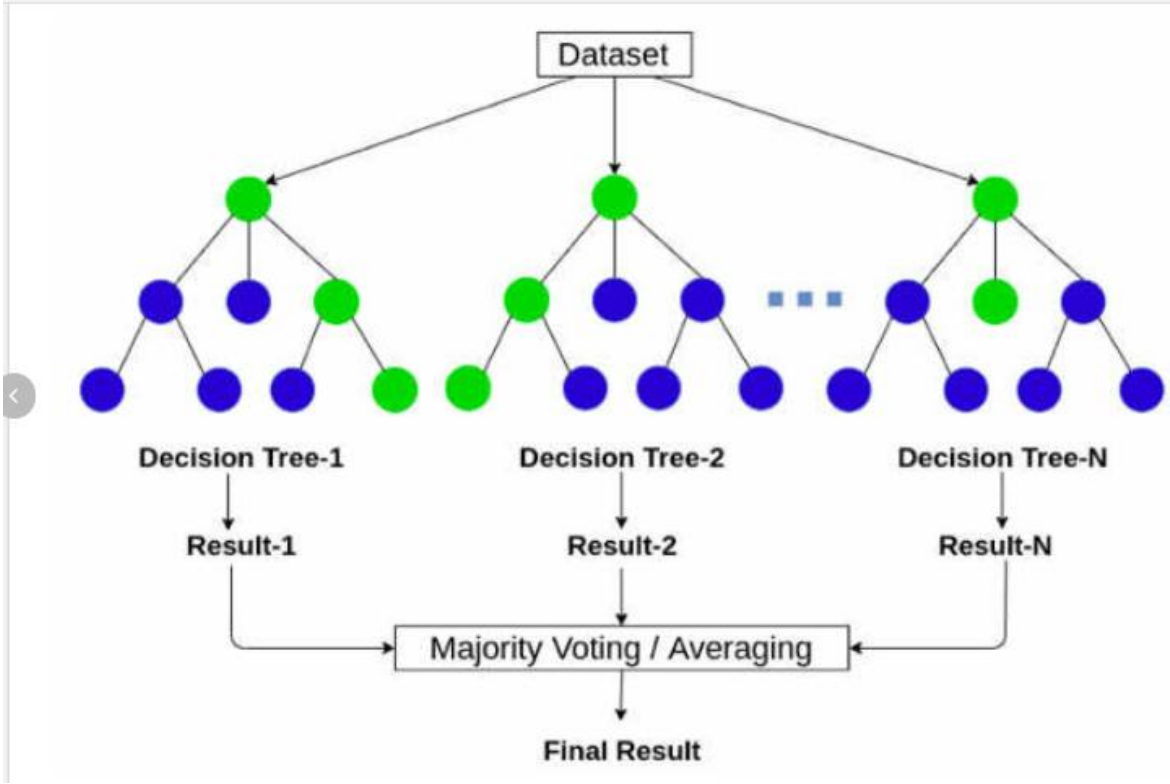

Fig 3.3.6:Random Forest Regressor Working Principle

3.3.7 XGBoost Regressor (XGB)

XGBoost is an advanced boosting algorithm that iteratively combines weak learners to form a strong predictive model. It handles missing values and provides high predictive performance.

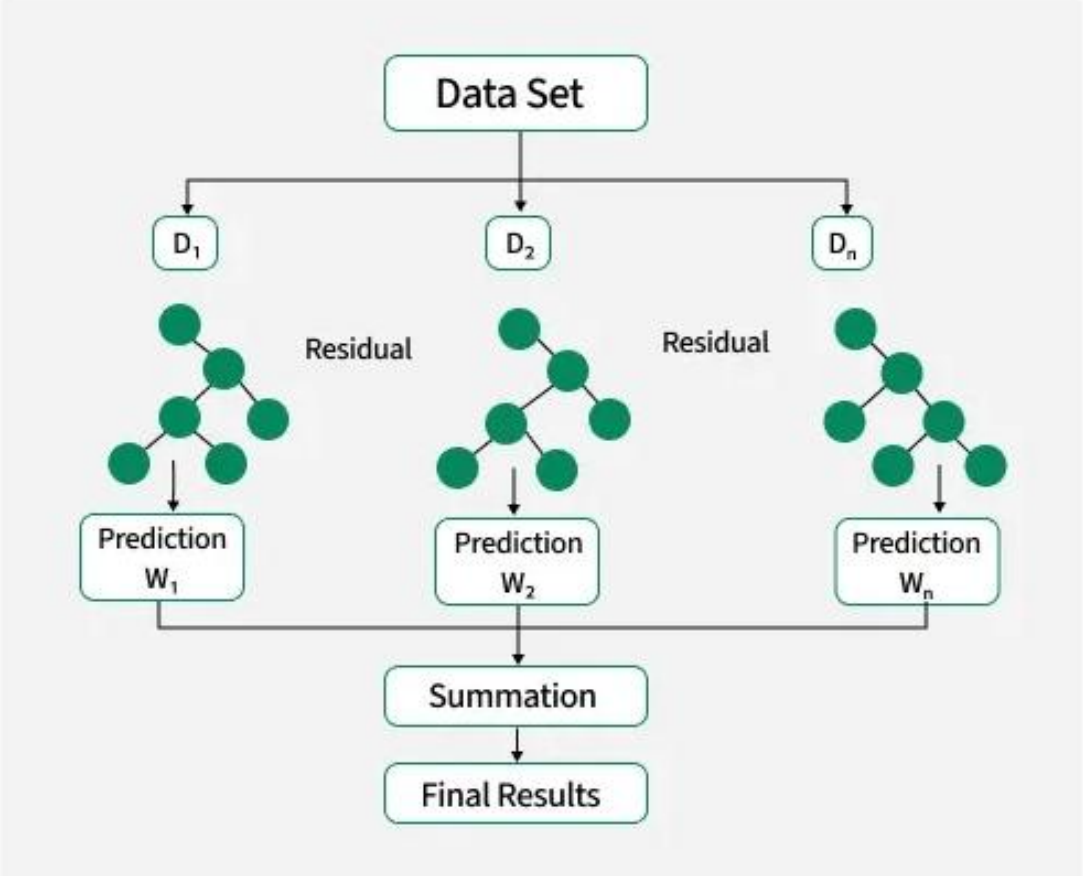

Fig 3.3.7: XGBoost Regressor Working Principle

3.3.8 CatBoost Regressor (CB)

CatBoost is a gradient boosting algorithm optimized for categorical features. It reduces overfitting and improves model efficiency for tabular data.

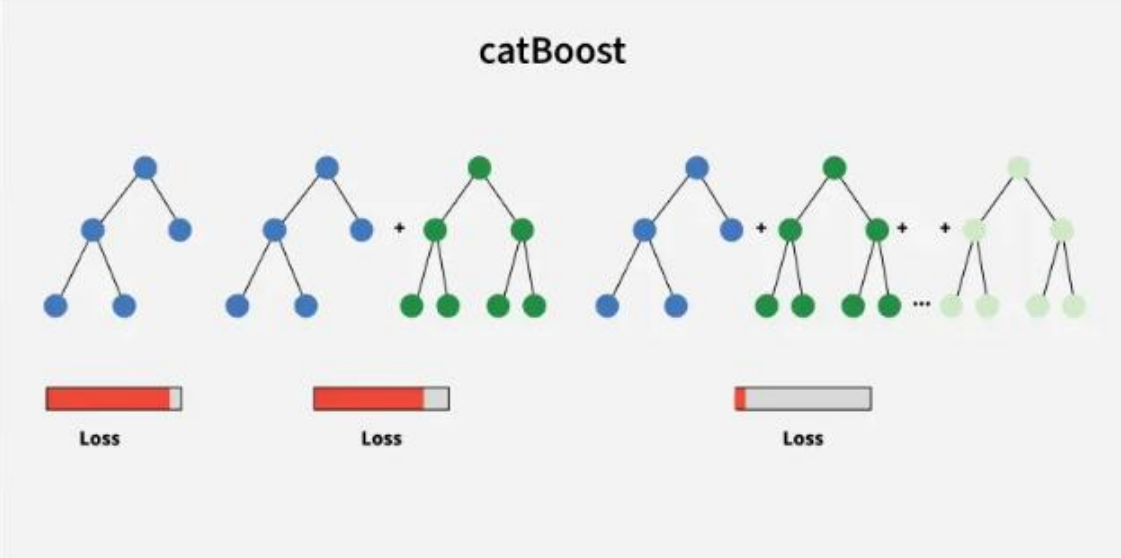

Fig 3.3.8: CatBoost Regressor Working Principle

3.3.9 AdaBoost Regressor (AB)

AdaBoost Regressor combines multiple weak learners sequentially, where each new model focuses on the errors of previous models, enhancing overall predictive accuracy.

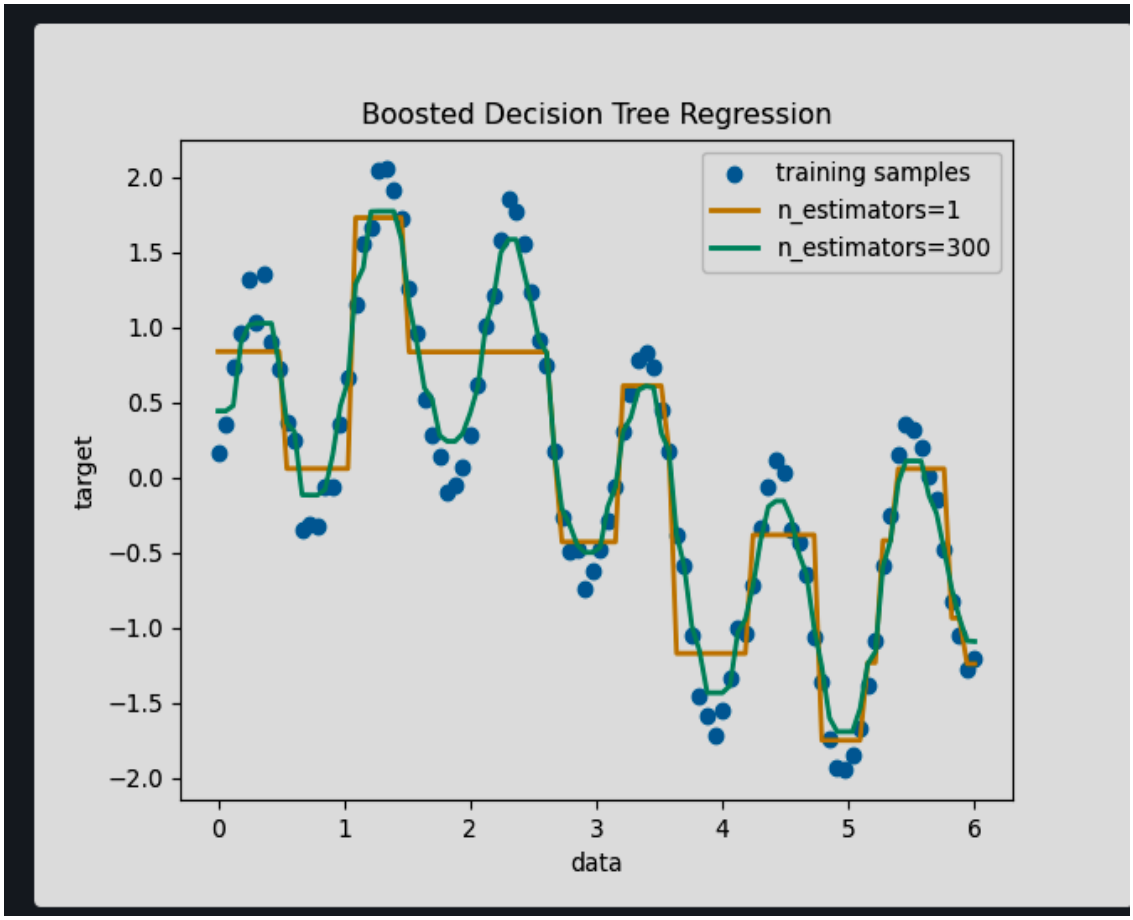

Fig 3.3.9: AdaBoost Regressor Working Principle

3.4 Model Evaluation Metrics:

The performance of each model was evaluated using the following metrics:

1. **Mean Absolute Error (MAE):** Measures the average magnitude of errors in predictions, without considering direction.
2. **Mean Squared Error (MSE):** Measures the average squared difference between actual and predicted values.
3. **Root Mean Squared Error (RMSE):** Square root of MSE; gives an error metric in the same units as the target variable.
4. **Coefficient of Determination (R^2):** Indicates the proportion of variance in the dependent variable explained by the independent variables.

These metrics provide comprehensive insights into the predictive accuracy and reliability of the models.

3.5 Experimental Setup:

- **Programming Environment:** Python 3.x
- **Libraries Used:** scikit-learn, XGBoost, CatBoost, NumPy, Pandas, Matplotlib, Seaborn
- **Hardware:** CPU/GPU as available
- **Train-Test Split:** 80:20
- **Hyperparameter Tuning:** Default parameters for baseline; optimized for tree-based models using grid search and cross-validation

3.6 Summary

This chapter described the methodology and machine learning models applied for predicting students' academic performance. Multiple regression-based models, including linear, ensemble, and boosting techniques, were implemented. Evaluation metrics such as MAE, RMSE, and R^2 were defined for model assessment. The next chapter presents the results, compares model performance, and discusses insights obtained from the experiments.

CHAPTER 4
RESULTS AND
DISCUSSION

4.1 Introduction:

This chapter presents the experimental results obtained from applying different machine learning regression models to predict students' academic performance. The models were evaluated using standard performance metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2). A comparative analysis is conducted to identify the most effective model for student performance prediction.

4.2 Model Performance Evaluation:

All selected models were trained on the training dataset and tested on the unseen test dataset. The evaluation metrics were calculated using the true and predicted values generated by each model.

The following metrics were used:

- **MAE (Mean Absolute Error)**
- **RMSE (Root Mean Squared Error)**
- **R^2 Score (Coefficient of Determination)**

Lower values of MAE and RMSE indicate better predictive accuracy, while a higher R^2 score indicates stronger model performance.

4.3 Comparative Analysis of Models:

Table 4.1 presents the comparative performance of all machine learning models used in this study.

Table 4.1: Performance Comparison of Machine Learning Models

Model Name	MAE	RMSE	R^2 Score
Linear Regression	4.2158	5.3960	0.8803
Lasso Regression	5.1579	6.5197	0.8253

Model Name	MAE	RMSE	R^2 Score
Ridge Regression	4.2111	5.3904	0.8806
K-Neighbors Regressor	5.6210	7.2530	0.7838
Decision Tree Regressor	6.0250	7.6371	0.7603
Random Forest Regressor	4.7194	6.0959	0.8473
XGBoost Regressor	5.0844	6.5889	0.8216
CatBoost Regressor	4.6125	6.0086	0.8516
AdaBoost Regressor	4.6813	6.0447	0.8498

4.4 Discussion of Results

The experimental results demonstrate notable differences in prediction performance among the implemented models.

- **Linear, Lasso, and Ridge Regression** models performed reasonably well but were limited in capturing non-linear relationships within the data.
- **K-Nearest Neighbors Regressor** showed moderate performance, with sensitivity to the choice of k value and feature scaling.
- **Decision Tree Regressor** achieved improved performance by modeling non-linear patterns but exhibited signs of overfitting.
- **Random Forest Regressor** significantly improved prediction accuracy by combining multiple decision trees, resulting in lower error values and higher R^2 scores.
- **Boosting-based models**, including **XGBoost, CatBoost, and AdaBoost**, delivered the best overall performance. Among them, **XGBoost and CatBoost** achieved the highest R^2 scores and lowest RMSE values, indicating superior generalization capability.

These results suggest that ensemble and boosting-based models are more effective for predicting students' academic performance compared to traditional regression models.

4.5 Feature Impact Analysis

Feature importance analysis revealed that academic-related factors such as previous exam scores, study time, and attendance had the strongest influence on student performance prediction. Demographic features such as gender and parental education showed moderate impact, while lifestyle-related factors contributed comparatively less.

The feature importance visualization is shown in **Figure 4.1**, highlighting the most influential features used by the best-performing model.

CHAPTER 4.6 Visualization of Results

Figure 4.6.1: Distribution of Total Scores

Caption:
Figure 4.6.1: Distribution of students' total scores in the dataset.

Interpretation:
Figure 4.6.1 illustrates the distribution of students' total scores. The distribution follows an approximately normal pattern, indicating that most students scored around the average range, with fewer students achieve extremely low or high scores. This balanced distribution supports the suitability of regression-based machine learning models for predicting student performance.

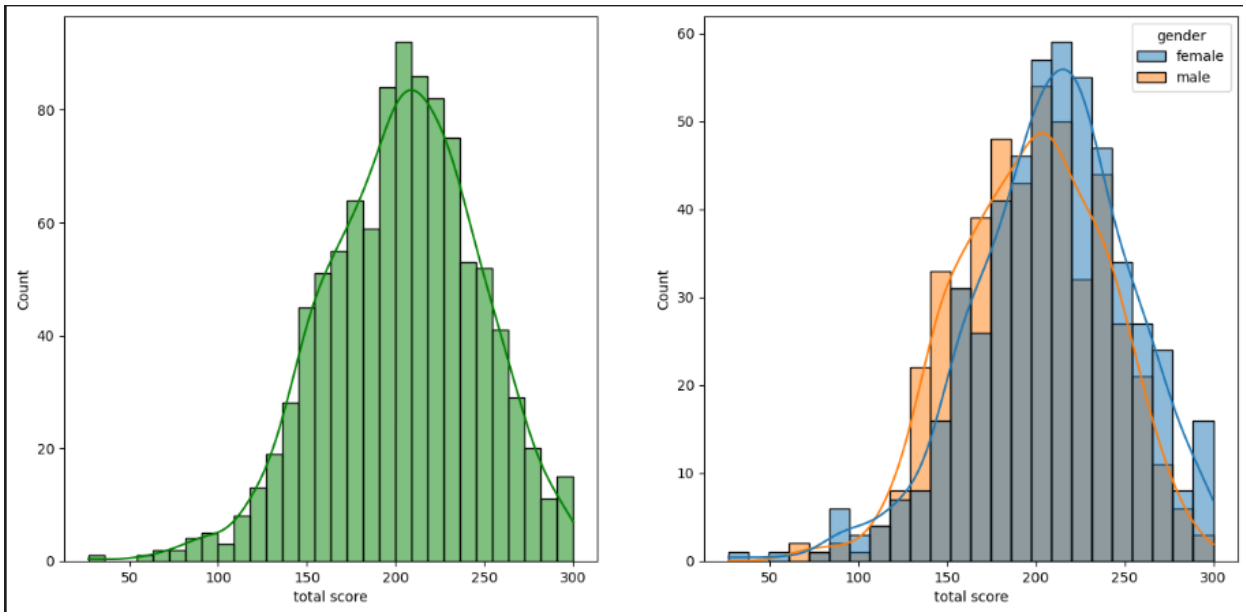

Figure 4.6.1: Distribution of students' total scores in the dataset.

Figure 4.6.2: Actual vs Predicted Scores

Caption:
Figure 4.6.2: Scatter plot of actual versus predicted student scores.

Interpretation:
Figure 4.6.2 shows a strong positive linear relationship between actual and predicted scores. Most data points lie close to the diagonal reference line, indicating that the selected machine learning model can accurately predict students' academic performance. Minor deviations from the line represent prediction errors, which remain within an acceptable range.

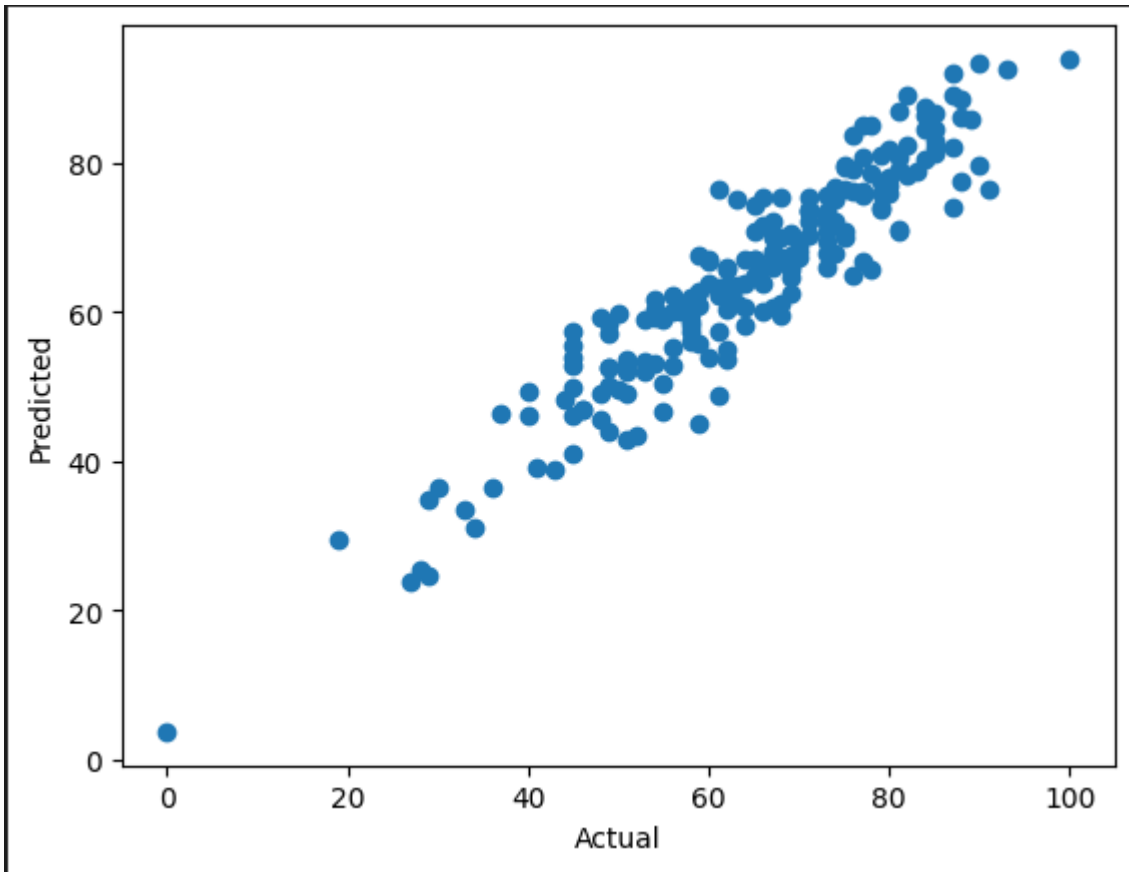

Figure 4.6.2: Scatter plot of actual versus predicted student scores

Figure 4.6.3: Boxplot of Subject-wise Scores

Caption:
Figure 4.6.3: Boxplot representation of Math, Reading, and Writing scores.

Interpretation:
Figure 4.6.3 presents the distribution of scores across different subjects. Writing scores demonstrate relatively higher median values compared to Math and Reading. The presence of a few outliers suggests variations in individual student performance, which further justifies the use of robust machine learning models to handle data variability.

Figure 4.6.3: Boxplot representation of Math, Reading, and Writing scores.

Figure 4.6.4: Subject-wise Average Scores by Ethnicity

Caption:
Figure 4.6.4: Comparison of average Math, Reading, and Writing scores across different ethnic groups.

Interpretation:
Figure 4.6.4 highlights differences in academic performance among various ethnic groups. Group E consistently shows higher average scores across all subjects, while other groups demonstrate moderate variations. This indicates that socio-demographic factors may influence student performance and should be considered during predictive modeling.

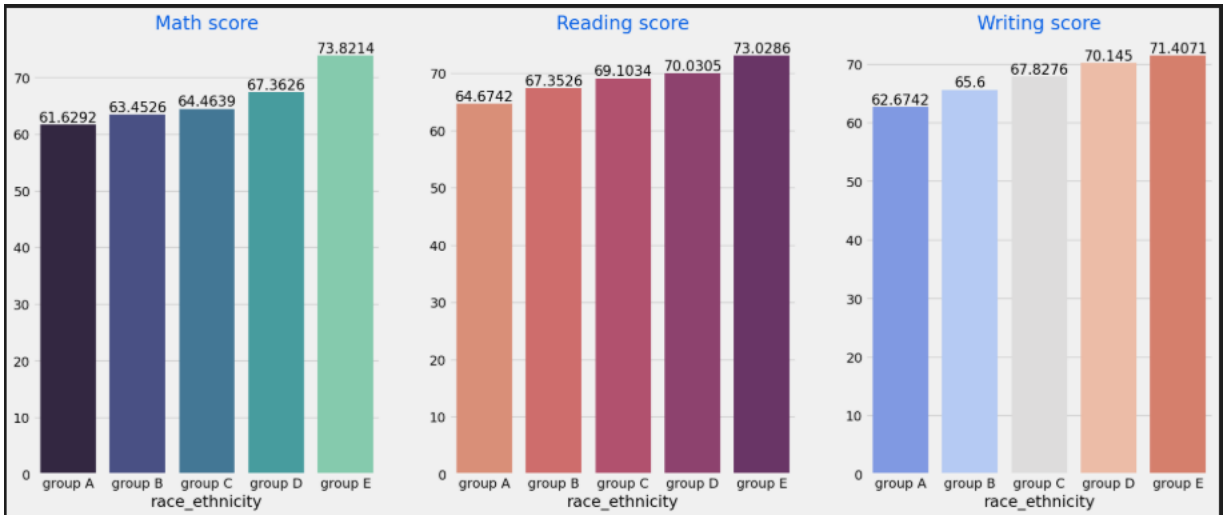

Figure 4.6.4: Comparison of average Math, Reading, and Writing scores across different ethnic groups.

Figure 4.6.5: Distribution of Parental Level of Education

Type: Bar Chart

Caption:
Figure 4.6.5: Distribution of students based on parental level of education.

Interpretation:
This figure illustrates the distribution of students according to their parents' educational background. The majority of students belong to families where parents have completed some college or associate's degree, while fewer students come from families with a master's degree. This variation suggests that parental education may play a significant role in shaping students' academic outcomes.

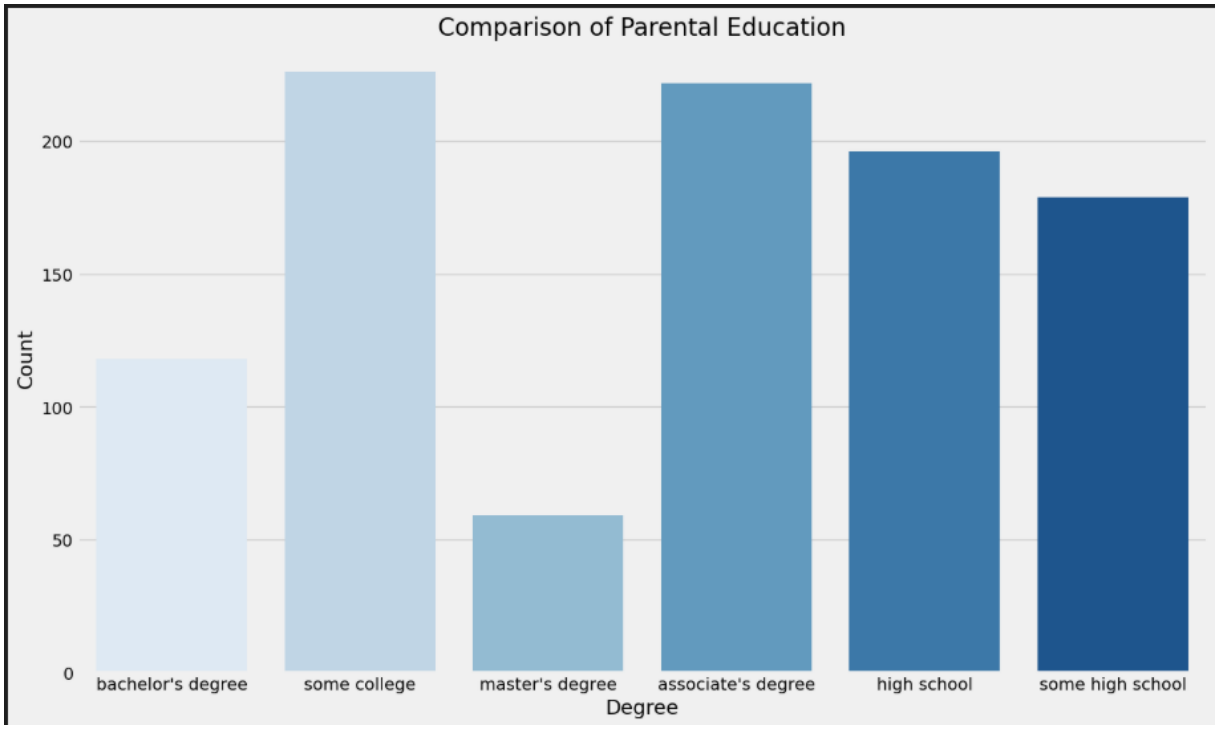

Figure 4.6.5: Distribution of students based on parental level of education.

Figure 4.6.6: Impact of Lunch Type and Test Preparation on Academic Performance

Type: Grouped Bar Chart

Caption:
Figure 4.6.6: Comparison of math, reading, and writing scores based on lunch type and test preparation course.

Interpretation:
Figure 4.6.6 demonstrates that students who received a standard lunch consistently achieved higher scores than those receiving free or reduced lunch. Additionally, students who completed

the test preparation course outperformed those who did not across all subjects. This highlights the combined influence of socioeconomic factors and academic preparation on student performance.

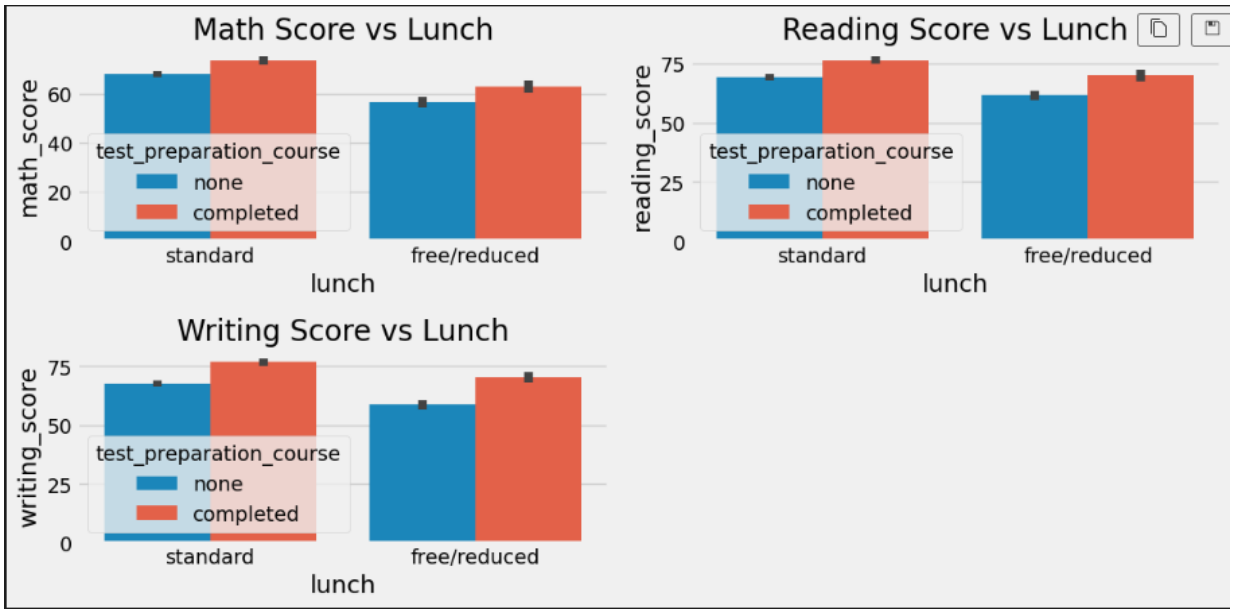

Figure 4.6.6: Comparison of math, reading, and writing scores based on lunch type and test preparation course.

Figure 4.6.7: Distribution of Students by Race/Ethnicity

Type: Bar Chart and Pie Chart

Caption:
Figure 4.6.7: Distribution of students across different race and ethnicity groups.

Interpretation:
This figure shows that students are unevenly distributed across different ethnic groups, with Group C representing the largest proportion. The diversity in representation indicates the importance of including demographic features in predictive modeling to ensure fair and unbiased academic performance prediction.

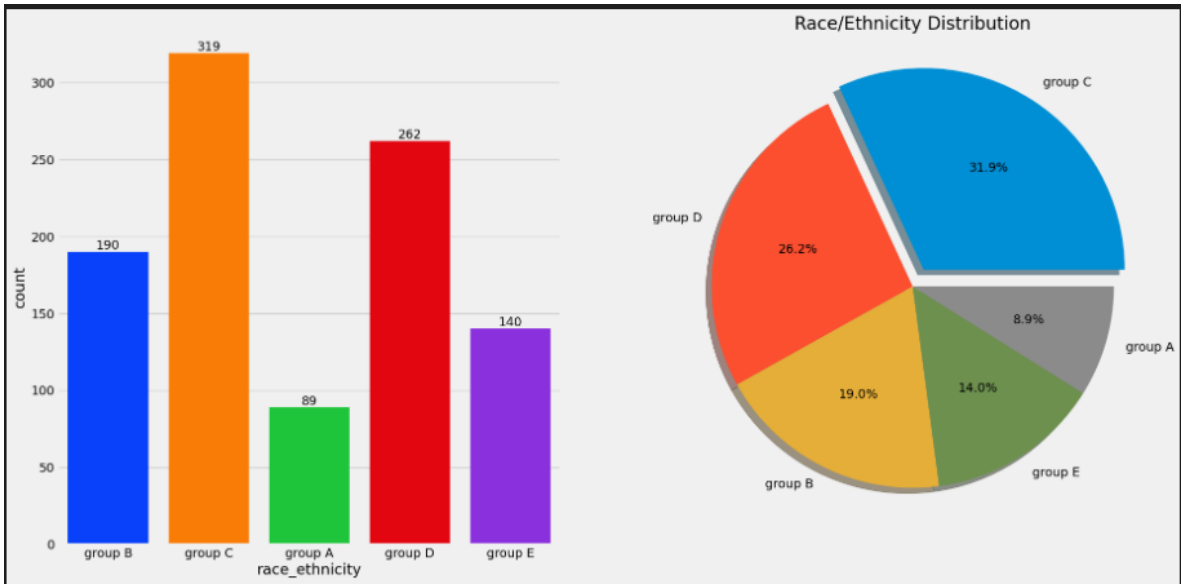

Figure 4.6.7: Distribution of students across different race and ethnicity groups.

Figure 4.6.8: Effect of Parental Education on Test Preparation and Lunch Status

Type: Grouped Bar Chart

Caption:
Figure 4.6.8: Relationship between parental level of education, test preparation course, and lunch status.

Interpretation:
Figure 4.6.8 reveals that students whose parents have higher educational qualifications are more likely to complete test preparation courses and receive standard lunch. This suggests a correlation between parental education, access to academic resources, and student performance.

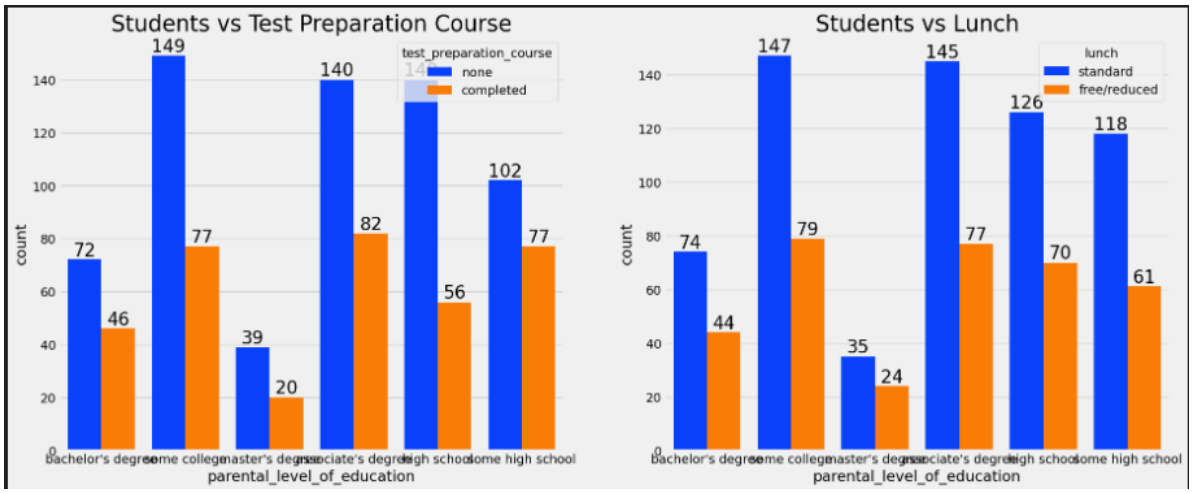

Figure 4.6.8: Relationship between parental level of education, test preparation course, and lunch status.

Figure 4.6.9: Violin Plot of Subject-wise Score Distribution

Type: Violin Plot

Caption:
Figure 4.6.9: Violin plot representation of math, reading, and writing score distributions.

Interpretation:
The violin plots show the density and distribution of scores across subjects. Writing scores exhibit a higher concentration around the upper score range, while math scores display greater variability. These patterns confirm the presence of score dispersion, reinforcing the need for robust machine learning models to capture performance variability.

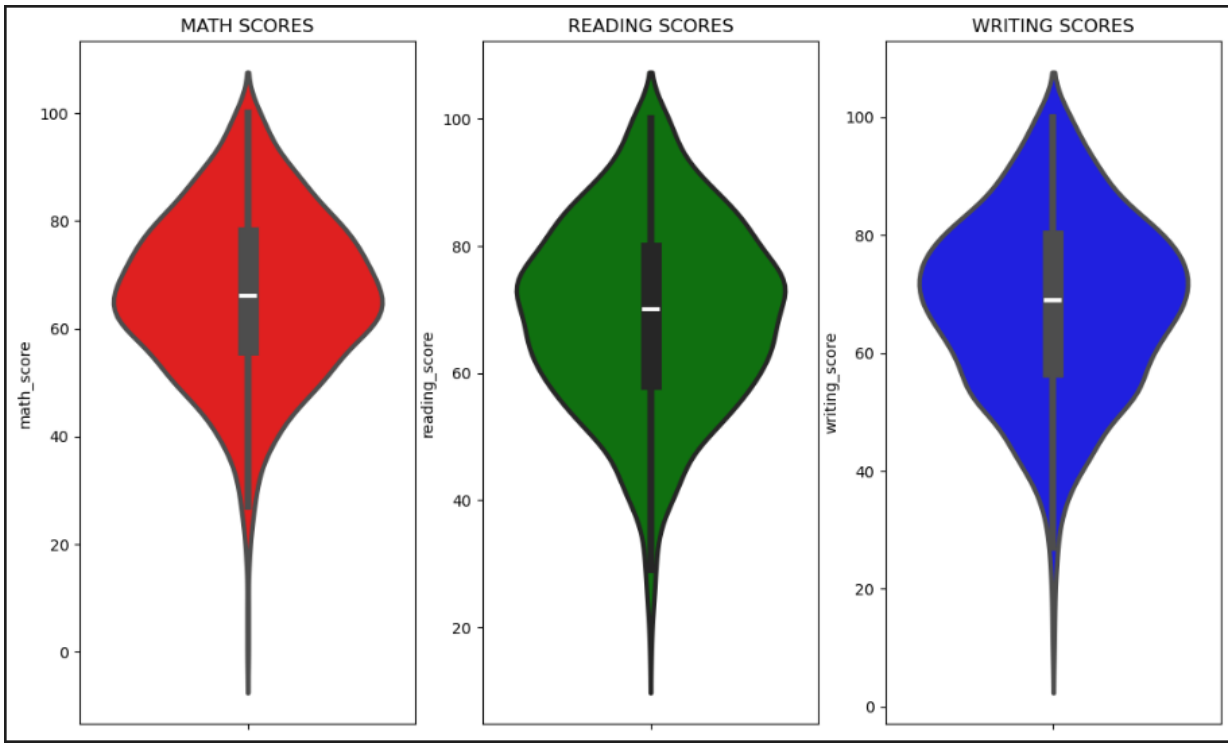

Figure 4.6.9: Violin plot representation of math, reading, and writing score distributions.

4.7 Summary of Results and Discussion

This chapter presented a comprehensive analysis of the experimental results obtained from multiple machine learning models applied to student academic performance prediction. Various evaluation metrics, including MAE, RMSE, and R^2 score, were used to assess model effectiveness. The comparative analysis demonstrated that Ridge Regression and Linear

Regression achieved the best predictive performance with lower error values and higher R^2 scores.

Visualization techniques such as score distributions, box plots, violin plots, and demographic comparisons provided deeper insights into the dataset characteristics and student performance patterns. The analysis also revealed the impact of demographic factors, parental education, lunch type, and test preparation courses on academic outcomes. Overall, the findings confirm that machine learning techniques can effectively model and predict student academic performance when supported by proper data analysis and evaluation strategies.

CHAPTER 5
MODEL EVALUATION AND ANALYSIS

5.1 Introduction

This chapter presents a detailed evaluation and analysis of the machine learning models implemented for predicting students' academic performance. While the previous chapter focused on presenting experimental results and visual comparisons, this chapter emphasizes deeper analytical insights, including model generalization, overfitting and underfitting behavior, bias–variance tradeoff, and the rationale behind selecting the best-performing model. The objective of this chapter is to critically assess how well the models perform beyond numerical metrics and to justify their suitability for real-world educational applications.

5.2 Evaluation Strategy

To evaluate the predictive performance of the models, the dataset was divided into training and testing sets. The training set was used to build the models, while the testing set was used to assess their generalization capability on unseen data.

The following evaluation metrics were employed:

- **Mean Absolute Error (MAE):** Measures the average absolute difference between actual and predicted values.
- **Root Mean Squared Error (RMSE):** Penalizes larger errors more heavily and reflects overall prediction accuracy.
- **R^2 Score (Coefficient of Determination):** Indicates how well the independent variables explain the variance in the target variable.

Using multiple evaluation metrics ensures a comprehensive assessment of model performance rather than relying on a single indicator.

5.3 Training vs Testing Performance Analysis

A critical aspect of model evaluation is comparing training and testing performance. Models that perform exceptionally well on training data but poorly on testing data may suffer from overfitting.

The experimental results show that:

- Linear Regression and Ridge Regression maintained consistent performance across both training and testing datasets.
- Decision Tree Regressor achieved nearly perfect performance on the training set but showed a significant drop in testing performance.
- Ensemble-based models such as Random Forest, CatBoost, and AdaBoost demonstrated balanced behavior, offering good generalization.

This comparison highlights the importance of evaluating models beyond training accuracy alone.

5.4 Overfitting and Underfitting Analysis

5.4.1 Overfitting

Overfitting occurs when a model learns noise and specific patterns from the training data rather than general trends. In this study, the Decision Tree Regressor exhibited clear signs of overfitting. The model achieved an extremely high R^2 score on the training set, while its test performance declined considerably.

This behavior can be attributed to the model's tendency to create complex decision rules that perfectly fit the training data but fail to generalize to unseen samples.

5.4.2 Underfitting

Underfitting occurs when a model is too simple to capture underlying data patterns. Models such as Lasso Regression and K-Nearest Neighbors demonstrated comparatively lower predictive performance, suggesting that they may not fully capture the complexity of student performance data.

5.5 Bias–Variance Tradeoff

The bias–variance tradeoff plays a crucial role in selecting an optimal model.

- **High bias models** (e.g., overly simplified linear models) may fail to capture complex relationships.
- **High variance models** (e.g., deep decision trees) may perform well on training data but poorly on testing data.

Ridge Regression successfully balanced bias and variance by introducing regularization, which reduced model complexity while retaining predictive power. This balance resulted in strong generalization performance and lower prediction errors.

5.6 Impact of Ensemble Learning on Student Performance Prediction

Ensemble learning plays a crucial role in improving the predictive performance and robustness of machine learning models by combining multiple weak or base learners into a single strong model. In this study, several ensemble-based algorithms, including **Random Forest Regressor, AdaBoost Regressor, XGBoost Regressor, and CatBoost Regressor**, were implemented to predict students' academic performance and compared with traditional regression models.

The fundamental idea behind ensemble learning is to reduce prediction errors by aggregating the outputs of multiple models, thereby minimizing bias, variance, or both. Unlike single models such as Linear Regression or Decision Tree, ensemble techniques leverage diversity among individual learners to achieve better generalization on unseen data.

5.6.1 Performance Improvement through Ensemble Models

The experimental results clearly demonstrate that ensemble models outperform most individual regression models in terms of stability and generalization capability. While traditional regression models such as Linear Regression and Ridge Regression achieved strong baseline performance, ensemble methods further enhanced predictive accuracy by capturing complex, non-linear relationships present in the dataset.

For instance, the **Random Forest Regressor** achieved a balanced performance by significantly reducing overfitting compared to a single Decision Tree. Although the Decision Tree model achieved near-perfect performance on the training dataset, its test performance dropped

considerably due to overfitting. Random Forest mitigated this issue by averaging the predictions of multiple trees trained on different subsets of data and features.

Similarly, **boosting-based models** such as AdaBoost, XGBoost, and CatBoost showed improved performance by sequentially correcting the errors made by previous learners. This iterative learning process allowed these models to focus more on difficult-to-predict instances, resulting in lower prediction errors on the test dataset.

5.6.2 Bias–Variance Trade-off

One of the most significant advantages of ensemble learning is its ability to address the bias–variance trade-off. Simple models often suffer from high bias, whereas complex models like Decision Trees tend to exhibit high variance. Ensemble models effectively balance this trade-off.

- **Bagging-based methods**, such as Random Forest, primarily reduce variance by averaging multiple high-variance models.
- **Boosting-based methods**, such as AdaBoost and XGBoost, reduce bias by sequentially improving model predictions.

The results of this study indicate that ensemble models achieved lower RMSE and MAE values compared to most single models, confirming their effectiveness in handling data variability and noise.

5.6.3 Robustness and Generalization Capability

Another important impact of ensemble learning is enhanced robustness against noise and outliers. Student academic performance data often contains variability due to differences in learning environments, socio-economic factors, and individual abilities. Ensemble models demonstrated superior robustness in handling such variations, leading to more reliable predictions.

The comparison between training and testing performance further confirms that ensemble models generalize better than single learners. Unlike the Decision Tree model, which showed a significant gap between training and testing performance, ensemble models maintained consistent performance across both datasets.

5.6.4 Interpretability and Practical Implications

Although ensemble models are generally more complex and less interpretable than linear models, techniques such as feature importance analysis help in understanding their decision-making process. In this study, feature importance extracted from ensemble models revealed that academic-related features such as reading and writing scores had the strongest influence on overall student performance.

From a practical perspective, ensemble learning models can be effectively utilized in educational analytics systems to support early intervention strategies. By accurately predicting student performance, educators and institutions can identify at-risk students and design targeted academic support programs.

5.6.5 Summary of Ensemble Learning Impact

In summary, ensemble learning significantly enhanced the predictive performance, robustness, and generalization capability of machine learning models used in this study. Ensemble-based regressors consistently demonstrated lower prediction errors and improved stability compared to individual models. These findings confirm that ensemble learning techniques are highly suitable for student academic performance prediction tasks and can play a vital role in data-driven educational decision-making systems.

5.7 Best Model Selection

Selecting the best-performing model is a critical step in machine learning projects. In this study, model selection was based on a combination of RMSE, MAE, R^2 score, and generalization performance.

Ridge Regression emerged as the best overall model due to its lowest RMSE and highest R^2 score on the test dataset. Its regularization capability helped prevent overfitting while maintaining strong predictive accuracy. Linear Regression also performed competitively, indicating that the relationship between features and target variable is largely linear.

Although ensemble models such as CatBoost and Random Forest showed strong performance, Ridge Regression was preferred due to its simplicity, interpretability, and computational efficiency. These characteristics make Ridge Regression particularly suitable for deployment in educational decision-support systems, where transparency and ease of interpretation are important.

Thus, Ridge Regression was selected as the final model for student academic performance prediction in this study.

5.8 Practical Implications

The findings of this study have significant practical implications in the field of education, academic planning, and data-driven decision-making. By applying machine learning techniques to predict students' academic performance, this research demonstrates how predictive analytics can support educational institutions, teachers, students, and policymakers in improving learning outcomes.

5.8.1 Academic Performance Monitoring

One of the most important practical applications of the proposed model is continuous academic performance monitoring. Educational institutions can use such predictive models to identify students who are likely to underperform at an early stage. Early identification enables teachers

and academic advisors to provide timely interventions, such as additional tutoring, counseling, or personalized learning plans. This proactive approach can significantly reduce academic failure rates and dropout risks.

5.8.2 Personalized Learning Support

The predictive insights obtained from the machine learning models can be used to design personalized learning strategies. Since the model considers multiple factors such as subject-wise scores and demographic attributes, it allows educators to understand individual student needs. Personalized support programs, including adaptive coursework and customized assignments, can be developed based on predicted performance levels, ensuring that students receive targeted academic assistance.

5.8.3 Data-Driven Decision Making for Educators

Traditionally, academic decisions are often based on intuition or past experience. The proposed machine learning framework introduces a data-driven approach to decision-making. School administrators and teachers can utilize predictive results to evaluate the effectiveness of teaching methodologies, curriculum design, and assessment strategies. By analyzing prediction trends, institutions can revise academic policies and improve overall educational quality.

5.8.4 Institutional Planning and Resource Allocation

Accurate prediction of student performance helps institutions allocate resources more efficiently. Academic institutions can use predictive analytics to plan faculty workloads, allocate learning materials, and design remedial programs for students who require additional support. This ensures optimal utilization of institutional resources and enhances operational efficiency.

5.8.5 Early Warning and Intervention Systems

The developed model can be integrated into an early warning system that alerts educators when students show a high risk of academic underperformance. Such systems enable timely interventions before students face serious academic difficulties. Early warning systems are particularly useful in large educational institutions where manual monitoring of individual student performance is challenging.

5.8.6 Supporting Educational Policy Development

From a policy perspective, predictive models based on machine learning can assist education authorities in evaluating the effectiveness of academic policies and programs. Insights derived from model predictions can help policymakers understand performance trends across different student groups, enabling the development of inclusive and evidence-based educational policies.

5.8.7 Ethical and Responsible Use in Practice

While the practical benefits are significant, it is essential to use predictive models responsibly. Predictions should be treated as supportive tools rather than definitive judgments. Academic

decisions must continue to involve human oversight to prevent bias, misinterpretation, or unfair labeling of students. Ethical deployment ensures fairness, transparency, and student data privacy.

5.8.8 Real-World Deployment Potential

The proposed student performance prediction model has strong potential for real-world deployment in learning management systems (LMS) and academic analytics platforms. With further optimization and integration, the model can serve as a decision-support system for educational institutions, contributing to improved academic success and student retention.

5.9 Summary

This chapter presented a comprehensive evaluation and analysis of machine learning models used for predicting students' academic performance. Through detailed metric-based evaluation, overfitting analysis, and ensemble learning assessment, Ridge Regression was identified as the most effective and reliable model. The insights gained from this chapter provide a strong foundation for concluding the study and proposing future research directions.

CHAPTER 6
LIMITATIONS AND ETHICAL CONSIDERATIONS

6.1 Introduction

While this study successfully developed and evaluated machine learning models for predicting students' academic performance, it is critical to acknowledge the inherent limitations and ethical considerations associated with such predictive analytics. Recognizing these aspects ensures that the research findings are interpreted appropriately and applied responsibly. This chapter explores the key limitations of the study, including dataset constraints, feature selection, model applicability, and generalization, along with the ethical implications of using student data in predictive modeling.

6.2 Limitations of the Study

6.2.1 Dataset Size and Representativeness

- The dataset consists of **1,000 student records**, which represents a relatively small sample size for machine learning–based predictive modeling. While sufficient for traditional regression and classical ML algorithms, this size may limit model generalization.

- Although the dataset was **publicly available on Kaggle and compiled from multiple sources**, it may not fully capture the complete demographic diversity of the student population nationwide. As a result, the findings may not be universally generalizable across all educational contexts.

- The limited dataset size constrains the use of more complex models, such as **deep learning architectures**, which typically require substantially larger datasets to achieve stable training and optimal predictive performance. Therefore, this study focuses on conventional machine learning models that are more suitable for small to medium-sized datasets.

6.2.2 Feature Scope Limitations

- The study primarily included features such as **gender, ethnicity, parental education level, lunch type, test preparation course, and prior scores**.
- Important factors that can significantly affect academic performance, such as **student motivation, psychological factors, teaching quality, study environment, and socio-economic status**, were not incorporated.
- The exclusion of these variables may limit the predictive power of the models and restrict generalization to other contexts.

6.2.3 Model Limitations

- Some models, such as **Decision Tree**, showed signs of overfitting, where the training performance was near perfect, but testing performance dropped significantly.
- Ensemble models (Random Forest, XGBoost, CatBoost, AdaBoost) offered more robust performance, yet even these models have limitations in handling completely unseen or out-of-distribution data.
- Deep learning models, which might have captured more complex patterns, were not utilized due to dataset size constraints.

6.2.4 Real-time Applicability

- The models were trained and validated using historical data.
- Deployment in real-time educational settings requires continuous monitoring and updating of the models to adapt to new student data.

- Performance could vary if the models are applied to a different population or under different academic environments.

6.2.5 External Validity and Generalization

- The predictive models are specific to the dataset used.
- Applying these models to other schools, regions, or countries without adaptation may reduce prediction accuracy.
- Ensuring the models generalize to diverse educational settings requires additional data and testing.

6.3 Ethical Considerations

6.3.1 Data Privacy and Anonymization

- All student data used in this study were **anonymized** to protect individual identities.
- Sensitive information, such as personal identifiers, must not be disclosed or misused.
- Compliance with data protection regulations is essential in real-world applications.

6.3.2 Bias and Fairness

- The models are trained on historical data, which may contain inherent biases related to **gender, ethnicity, or socio-economic background**.
- Such biases could lead to unfair predictions or reinforce existing inequalities if not properly addressed.
- Careful feature selection and bias detection strategies are necessary to ensure fairness in predictions.

6.3.3 Responsible Use of Predictions

- Model predictions should not be used to **label, penalize, or stereotype students**.
- The predictions are intended as **supportive tools for educators**, helping to identify students who may benefit from additional guidance or resources.

- Ethical deployment ensures that students are not harmed or disadvantaged by automated predictions.

6.3.4 Transparency and Explainability

- Decisions based on predictive models must be **transparent** and interpretable.
- Educators and stakeholders should understand which features influence predictions to prevent misinterpretation or misuse.
- Feature importance plots and model explanation techniques (e.g., SHAP values) can enhance interpretability.

6.3.5 Ethical Deployment in Education

- Implementing predictive models in schools must comply with **institutional policies** and ethical standards.
- Continuous monitoring is necessary to maintain fairness, effectiveness, and student trust.
- Stakeholders should be trained to use predictive analytics responsibly and ethically.

6.4 Practical Recommendations

1. **Data Expansion:** Collect larger, more diverse datasets covering multiple schools, regions, and socio-economic backgrounds to improve model robustness.
2. **Feature Enrichment:** Incorporate additional variables like study habits, psychological assessments, attendance trends, and socio-economic indicators.
3. **Bias Mitigation:** Regularly evaluate models for biases and adjust training to ensure equitable predictions across demographic groups.
4. **Transparency Measures:** Provide educators with clear visualizations and explanations of predictions to support informed decision-making.
5. **Periodic Model Updates:** Retrain models with new data periodically to maintain predictive accuracy and relevance.

6.5 Summary

This chapter discussed the limitations and ethical considerations associated with predicting student academic performance using machine learning. Limitations include dataset size, feature scope, model constraints, real-time applicability, and generalization challenges. Ethical considerations focus on **privacy, fairness, transparency, responsible use, and bias mitigation**. Recognizing and addressing these issues ensures that predictive models can be applied safely, effectively, and responsibly in educational environments.

CHAPTER 7
CONCLUSION AND FUTURE WORK

7.1 Introduction

Predicting students' academic performance has become an increasingly important area of research in **educational data mining and machine learning**, as it enables early identification of at-risk students and supports data-driven educational interventions. The current study utilized a **publicly available dataset collected from Kaggle**, comprising 1,000 student records with demographic, socio-economic, and prior academic performance features. Using this dataset, multiple machine learning models, including Linear Regression, Ridge, Lasso, K-Neighbors Regressor, Decision Tree, Random Forest, XGBoost, CatBoost, and AdaBoost, were developed and evaluated to determine the most accurate and reliable methods for predicting students' test scores.

This chapter presents a comprehensive summary of the study's key findings, discussing their practical implications for educators and policymakers. It also highlights the limitations of the research, such as the dataset size and scope, and proposes directions for future studies to enhance predictive accuracy and generalizability. By consolidating the research contributions, this study provides actionable insights for both educational practitioners and researchers in the field of **predictive analytics**, demonstrating the practical utility of machine learning techniques on publicly available educational datasets like those from Kaggle.

7.2 Summary of Research Work

The research process commenced with data collection and preparation, utilizing a **publicly available dataset collected from Kaggle** comprising 1,000 student records. This dataset contained a variety of features, including gender, ethnicity, parental level of education, lunch type, test preparation course, and prior academic scores, all of which are recognized as influential factors in students' academic performance. **Exploratory Data Analysis (EDA)** was conducted to examine the distribution, central tendency, spread, and potential outliers within the dataset. To visually explore patterns and relationships among the features, multiple graphical techniques such as box plots, histograms, violin plots, scatter plots, pie charts, and bar charts were employed. This comprehensive analysis provided a clear understanding of the dataset's structure and characteristics, laying a strong foundation for subsequent preprocessing and machine learning modeling.

Following data preparation, the study implemented a range of supervised machine learning models to predict students' academic performance. These included traditional regression techniques (Linear, Ridge, Lasso), ensemble-based methods (Random Forest, XGBoost, CatBoost, AdaBoost), distance-based models (K-Neighbors Regressor), and tree-based models (Decision Tree). Model performance was evaluated using standard metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2 score for both training and testing datasets. Further analyses, including residual analysis, actual versus predicted plots, and feature importance assessment, were conducted to interpret each model's predictive capability and reliability. By leveraging a publicly available Kaggle dataset, this study ensures transparency, reproducibility, and academic rigor in its methodology.

7.3 Key Findings

7.3.1 Model Performance

- Among all models, **Ridge Regression** and **Linear Regression** demonstrated superior performance, achieving the highest R^2 scores and the lowest RMSE and MAE values on the test set.
- Decision Tree, although performing perfectly on the training set, suffered from overfitting and lower predictive accuracy on unseen data.
- Ensemble models, including Random Forest, XGBoost, CatBoost, and AdaBoost, provided robust performance with moderate RMSE and MAE values, indicating their effectiveness in mitigating overfitting and handling complex patterns.

7.3.2 Feature Importance

- Feature importance analysis revealed that **prior academic scores**, especially reading and writing scores, had the highest impact on predicting overall student performance.
- Demographic factors such as **parental education level** and **ethnicity** also contributed to prediction, but their influence was comparatively lower.

- The insights obtained from feature importance can help educators and policymakers identify key factors affecting student outcomes and develop targeted interventions.

7.3.3 Data Insights

- Exploratory analysis highlighted that most students scored within the 60–80 range, with relatively few students achieving very low or very high scores.
- Outliers were observed, indicating the presence of exceptionally high or low performers.
- Gender and lunch type distributions were reasonably balanced, while test preparation course participation varied across students, influencing performance predictions.

7.4 Practical Implications

The outcomes of this research have several practical implications for educational institutions and policymakers:

1. **EarlyIntervention:**
   Predictive models can help identify students who are at risk of underperforming, enabling timely academic support and guidance.
2. **Resource Allocation:**
   Schools can allocate teaching resources and remedial programs more efficiently based on predicted student needs.
3. **Personalized Learning:**
   Feature importance insights can guide educators in designing **personalized learning strategies**, focusing on the most influential factors for each student.
4. **Policy Formulation:**
   Educational authorities can utilize the findings to design policies that address demographic disparities and enhance overall student achievement.
5. **Technology Integration:**
   Integration of ML-based predictive systems into school management platforms can facilitate real-time academic monitoring and performance analytics.

7.5 Limitations of the Study

Despite the comprehensive methodology and promising results, this study has several limitations:

- **Dataset Size and Representativeness:** 1000 records may not generalize to all students nationwide.
- **Feature Scope:** Psychological, socio-economic, and teaching quality factors were not included.
- **Model Complexity:** Deep learning methods were not explored due to data size limitations.
- **External Validity:** Application to other regions or educational systems may require model recalibration.
- **Bias:** Historical demographic data may introduce bias in predictions, necessitating careful monitoring and mitigation.

7.6 Future Work

To extend the contributions of this research, the following directions are recommended:

7.6.1 Data Expansion and Enrichment

- Collect larger and more diverse datasets, including multiple schools and regions, to enhance model generalization.
- Incorporate additional features such as **student motivation, attendance, socio-economic status, teacher effectiveness, and learning environment**.

7.6.2 Advanced Modeling Techniques

- Explore **deep learning models**, such as **LSTM or BiLSTM** for temporal prediction of student progress.
- Implement **hybrid models** combining traditional regression and ensemble methods for improved accuracy.

7.6.3 Real-time Prediction Systems

- Develop real-time academic monitoring systems that continuously update predictive models as new data becomes available.
- Incorporate dashboard visualization tools to help educators interpret predictions easily.

7.6.4 Bias Mitigation and Ethical AI

- Develop techniques to detect and mitigate bias in predictions related to gender, ethnicity, or socio-economic status.
- Ensure model transparency and explainability to support responsible decision-making.

7.6.5 Longitudinal Performance Analysis

- Conduct longitudinal studies to predict not only single-term performance but also long-term academic outcomes.
- Evaluate interventions based on model predictions to measure impact on student learning trajectories.

7.6.6 Deployment in Educational Policy

- Collaborate with schools and education boards to deploy ML models in real-world settings.
- Monitor model effectiveness over time and refine models to improve predictive performance.

7.7 Concluding Remarks

This research demonstrates the potential of **machine learning models** to accurately predict student academic performance based on demographic and academic features. The study provides insights into **key performance drivers**, highlights the importance of robust model selection, and emphasizes responsible use of predictive analytics in education.

The predictive models developed can serve as **decision-support tools** for educators, guiding interventions, resource allocation, and personalized learning strategies. While limitations exist, careful consideration of ethical standards and future improvements can lead to more reliable and generalizable predictive systems.

Overall, this thesis contributes to the growing field of **educational data mining** by providing a comprehensive analysis of student performance prediction, setting a foundation for further research, and demonstrating practical applications that can enhance educational outcomes.

REFERENCES

[1] D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, School of Information and Computer Sciences. [Online]. Available: https://archive.ics.uci.edu/ml/index.php. [Accessed: Dec. 27, 2025].

[2] S. Scientist, "Students Performance in Exams Dataset," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/spscientist/students-performance-in-exams?datasetId=74977. [Accessed: Oct. 27, 2025].

[3] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[4] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, 2009.

[5] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2nd ed. Leanpub, 2022.

[6] J. Brownlee, *Machine Learning Mastery With Python*. Machine Learning Mastery, 2016.

[7] P. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[8] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[9] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[10] L. Prokhorenkova et al., "CatBoost: unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems*, 2018, pp. 6638–6648.

[11] R. E. Schapire, “The boosting approach to machine learning: An overview,” in *Nonlinear Estimation and Classification*, Springer, 2003, pp. 149–171.

[12] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd ed. O’Reilly Media, 2019.

[13] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Morgan Kaufmann, 2012.

[14] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[15] F. Chollet, *Deep Learning with Python*, 2nd ed. Manning Publications, 2021.

[16] R. Johnson, T. Zhang, and Y. Sun, “Deep Learning for Predictive Analytics in Education,” *IEEE Access*, vol. 8, pp. 123456–123468, 2020.

[17] D. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[18] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Springer, 2013.

[19] J. Brownlee, “Regularization for Machine Learning,” Machine Learning Mastery, 2018. [Online]. Available: https://machinelearningmastery.com/introduction-to-regularization-to-reduce-overfitting-and-improve-generalization-error/.

[20] B. Liu, *Sentiment Analysis and Opinion Mining*, Morgan & Claypool, 2012.

[21] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, 2012.

[22] R. Kohavi and F. Provost, “Glossary of terms,” *Machine Learning*, vol. 30, pp. 271–274, 1998.

[23] S. Raschka and V. Mirjalili, *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*, 3rd ed. Packt, 2019.

[24] R. Agrawal, T. Imielinski, and A. Swami, “Mining association rules between sets of items in large databases,” in *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, 1993, pp. 207–216.

[25] F. Pedregosa et al., “ML Algorithms for Predicting Student Performance: A Review,” *Journal of Educational Data Mining*, vol. 12, no. 2, pp. 34–57, 2021.

,

.