Predicting Exam Scores using Supervised Regressions

Franrey Anthony S. Saycon

January 2026

1 Problem Understanding & Framing

1.1 Problem Statement

Educational institutions aim to improve student performance and reduce failure rates, yet they often lack early indicators of academic risk.

The objective of this study is to **develop a supervised regression model that predicts students' exam scores** based on relevant historical and behavioral features. By learning the relationship between input variables and continuous exam score outcomes, the model seeks to provide accurate score predictions prior to the actual examination. These predictions can enable educators and administrators to identify at-risk students early and implement timely academic interventions.

1.2 Success Criteria

A supervised regression model is considered successful if it achieves a high coefficient of determination ($\mathbf{R^2}$) and low prediction error, as measured by metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE), on both training and unseen test data. We will focus on $\mathbf{R^2}$ and RMSE for this study.

Metric	**Target Value**
R-squared (R^2)	≥ 0.70
Root Mean Squared Error (RMSE)	≤ 10

Table 1: Target Success Metrics

We aim for a R^2 of at least 0.7 and an RMSE of no more than 10 to ensure that the regression model captures the majority of the variability in the exam scores while maintaining reasonably low prediction errors. An $R^2 \geq 0.70$ indicates that 70% of the variation in scores can be explained by the input features of the model, making it reliable for identifying patterns and trends. At the same time, an RMSE ≤ 10 ensures that predicted scores typically deviate from actual scores by no more than 10 points, which is small enough to support actionable decisions, such as early interventions for students at risk of underperformance. Together, these thresholds balance predictive accuracy.

2 Data Collection & Understanding

2.1 Dataset Description

Kaggle Education Dataset: https://www.kaggle.com/datasets/kundanbedmutha/exam-score-prediction-dataset?resource=download

This dataset provides a representation of various factors that contribute to student exam performance. It contains 20,000 records, each describing the academic behavior, study habits, lifestyle routines, and exam conditions of the student. Together, these variables help to understand how different aspects of a student's daily life and learning environment influence their exam results.

2.2 Features Summary

The dataset includes a mix of categorical and numerical features that describe student demographics, academic behaviors, and lifestyle factors. Categorical variables capture information such as gender, course enrollment, internet access, sleep quality, study methods, facility ratings, and perceived exam difficulty. Numerical variables include age, study hours, class attendance, sleep hours, and exam scores.

Table 2 and Table 3 summarizes these features.

Variable	Type	Min	Max	Mean	Unique
age	int64	17.00	24.00	20.47	8
study_hours	float64	0.08	7.91	4.01	784
class_attendance	float64	40.60	99.40	70.02	589
sleep_hours	float64	4.10	9.90	7.01	59
exam_score	float64	19.60	100.00	62.51	805

Table 2: Summary Statistics for Numerical Variables in the Dataset

Variable	Categories
Gender	male, other, female
Course	diploma, bca, b.sc, b.tech, bba, ba, b.com
Internet Access	yes, no
Sleep Quality	poor, average, good
Study Method	coaching, online videos, mixed, self-study, group study
Facility Rating	low, medium, high
Exam Difficulty	hard, moderate, easy

Table 3: Categorical Variables and Their Categories in the Dataset

3 EDA + Feature Engineering Report

3.1 Exploratory Data Analysis (EDA)

The dataset contains no missing or null values across all columns, as shown in Table 4, which simplifies preprocessing. Outlier detection was performed on all numerical columns using the interquartile range (IQR) method, and no significant outliers were found, indicating a clean distribution of data, as shown in Figure 1.

The `student_id` column will be dropped from the analysis as it is not relevant for prediction. The dataset is now ready for regression modeling, as the features are complete, free of significant outliers, and suitable for training predictive models. One hot encoding will be

introduced to the categorical features with drop first as well.

Column	Null Count	Missing Count
student_id	0	0
age	0	0
gender	0	0
course	0	0
study_hours	0	0
class_attendance	0	0
internet_access	0	0
sleep_hours	0	0
sleep_quality	0	0
study_method	0	0
facility_rating	0	0
exam_difficulty	0	0
exam_score	0	0

Table 4: Null and Missing Value Summary for Dataset Columns

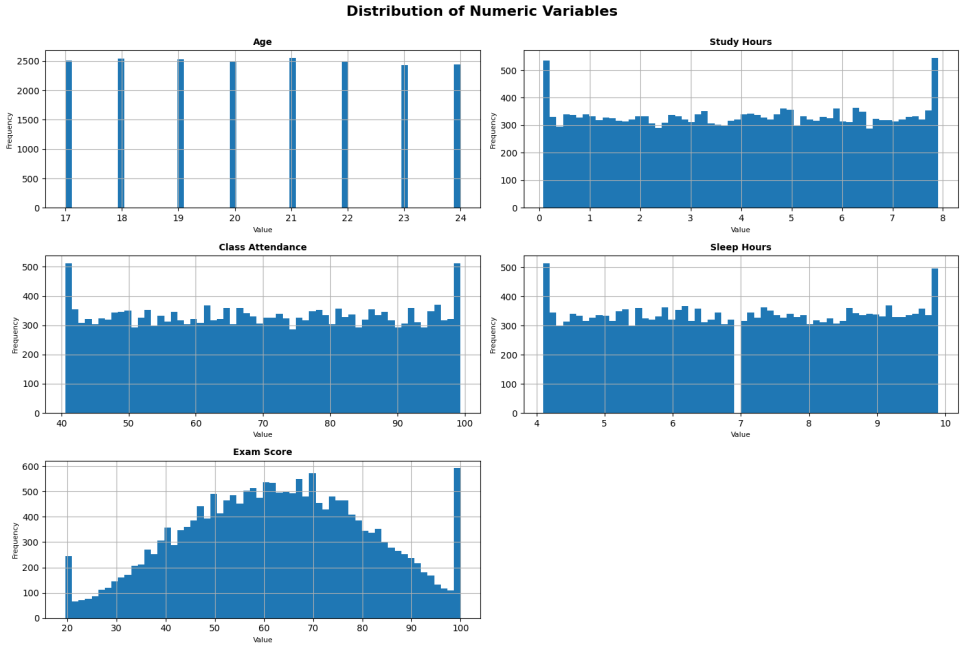

Figure 1: Distribution Charts of Numerical Features

Scatter plots were generated to explore relationships between numerical features and exam scores. A clear linear trend is observed between `study_hours` and `exam_score`, indicating that increased study time is associated with higher scores. In contrast, `age` appears as parallel vertical lines, suggesting little variation in exam scores across different ages. Features such as `class_attendance` and `sleep_hours` do not show clear linear trends with exam scores, implying that their impact on performance may be less direct or more complex. This is shown in Figure 2.

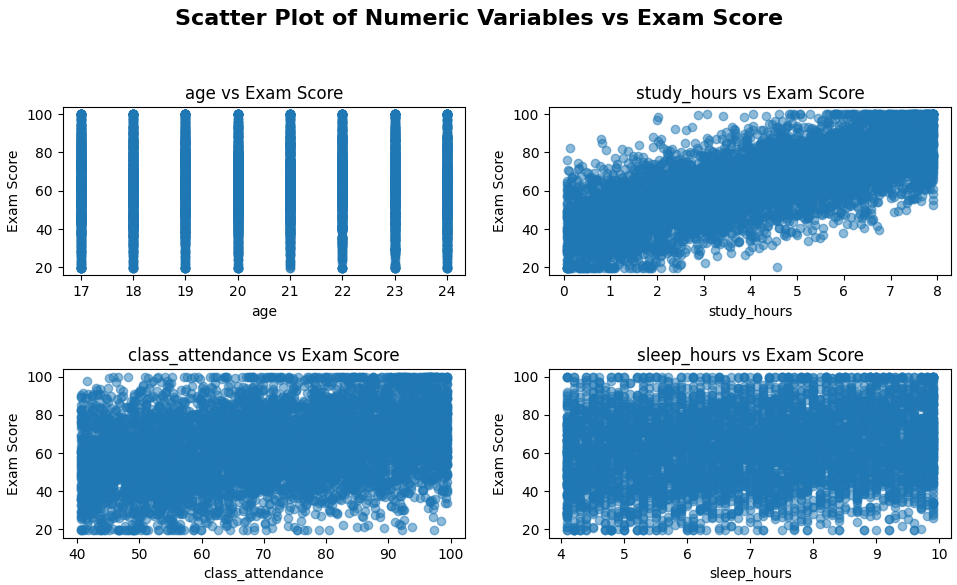

Figure 2: Scatter Plots of Numerical Variables vs Exam Score

Box plots were generated to examine the distribution of exam scores across the different categories of categorical features, as shown in Figure 3. All box plots show that the spread of scores is contained within the whiskers, with no outliers detected, indicating a consistent range of scores within each category.

3.2 Correlation Analysis

A correlation matrix was computed between all numerical and one-hot encoded categorical features and `exam_score`. Tables 5 and 6 show the top five positive and bottom five negative correlations, respectively. This is shown in Figure 4.

Feature	Correlation with Exam Score
study_hours	0.718
class_attendance	0.309
sleep_quality_good	0.172
sleep_hours	0.133
study_method_mixed	0.045

Table 5: Top 5 Positive Correlations with Exam Score

Feature	Correlation with Exam Score
study_method_group	-0.040
study_method_online_videos	-0.063
study_method_self-study	-0.101
facility_rating_low	-0.146
sleep_quality_poor	-0.172

Table 6: Bottom 5 Correlations with Exam Score

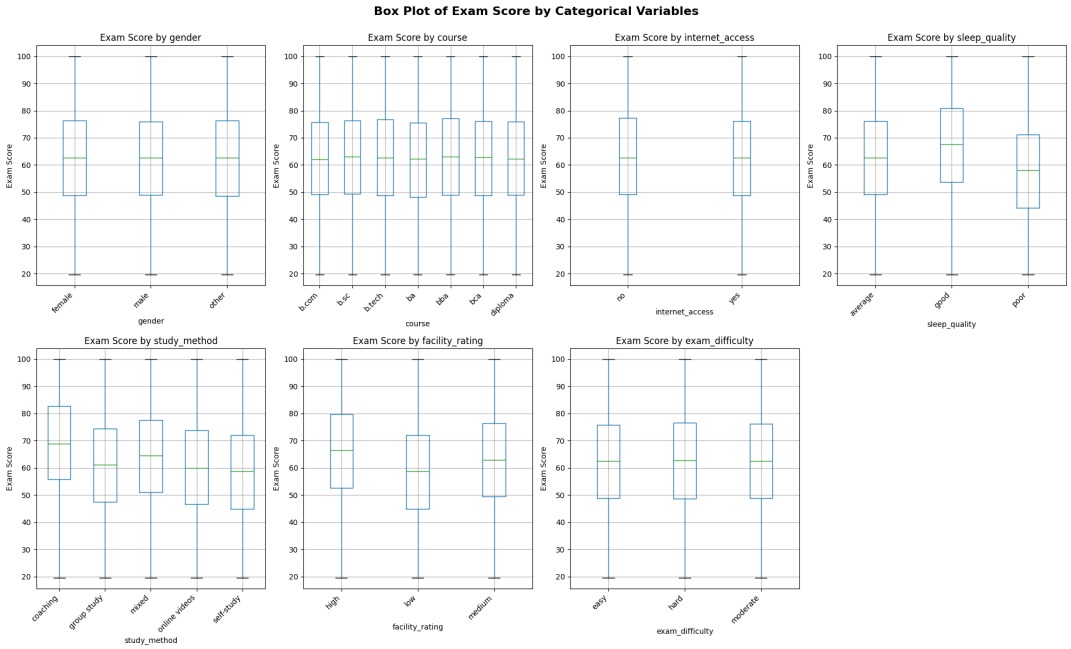

Figure 3: Box Plots of Exam Score by Categorical Variables

These results indicate that `study_hours` and `class_attendance` have the strongest positive influence on exam scores, while factors such as poor sleep quality and low facility ratings are negatively associated.

The top 10 features with the highest absolute correlation with `exam_score` are listed in Table 7. `study_hours` has the strongest relationship ($r = 0.718$), followed by `class_attendance` ($r = 0.309$), indicating that dedicated study time and regular attendance are the most influential factors for exam performance. Sleep quality and facility rating also show moderate correlations, while different study methods exhibit weaker associations. This analysis highlights which features are most relevant for predicting exam scores and provides guidance for feature selection in regression modeling.

Feature	**Absolute Correlation with Exam Score**
study_hours	0.718
class_attendance	0.309
sleep_quality_poor	0.172
sleep_quality_good	0.172
facility_rating_low	0.146
sleep_hours	0.133
study_method_self-study	0.101
study_method_online_videos	0.063
study_method_mixed	0.045
study_method_group_study	0.040

Table 7: Top 10 Features by Absolute Correlation with Exam Score

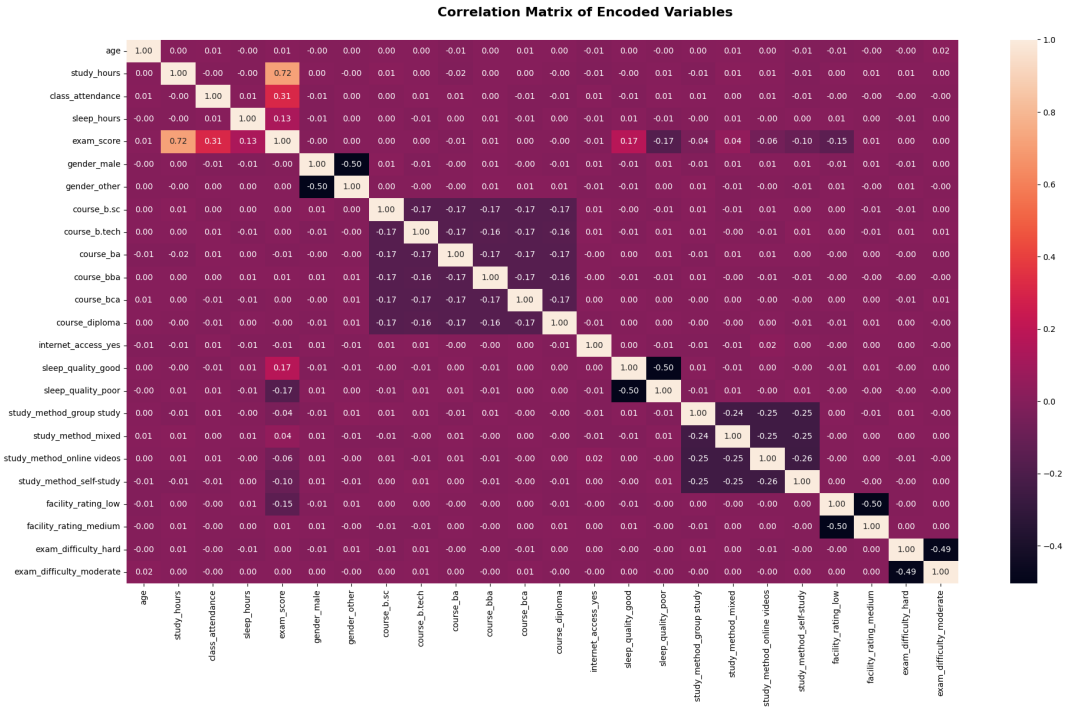

Figure 4: Correlation Matrix for Exam Scores

3.3 Summary

The dataset is of high quality, with no missing values, duplicates, or outliers across all 20,000 records. Numeric features such as study hours, sleep hours, and class attendance show reasonable distributions, while categorical variables—including gender, course, internet access, sleep quality, study method, facility rating, and exam difficulty—capture meaningful differences in performance. Box plots confirm that categorical features are informative, and all categorical variables were one-hot encoded with the first category dropped to avoid multicollinearity. The student ID was removed as it serves only as an identifier, leaving a final feature set of four numeric features plus encoded categorical variables.

Correlation analysis highlights that study hours ($r = 0.72$) and class attendance ($r = 0.31$) are the strongest predictors of exam scores, while good sleep quality and adequate sleep hours show moderate positive associations. Conversely, poor sleep quality, low facility ratings, and self-study methods exhibit moderate negative correlations, indicating that both behavioral and environmental factors influence performance. Mixed or group study methods have slightly better outcomes than self-study or online videos alone, but these effects are relatively small.

Overall, study habits dominate exam performance, while behavioral and environmental factors provide additional predictive value. The top features show minimal multicollinearity, making the dataset well-prepared for regression modeling. These insights justify the use of both numeric and encoded categorical variables to build a robust predictive model for exam scores.

4 Model Implementation

To predict student exam scores, we implemented multiple regression models to compare their performance and evaluate the impact of feature selection. The models include vanilla **Linear Regression**, **Ridge Regression**, **Lasso Regression**, **ElasticNet Regression**, and **Random Forest Regression**.

4.1 Model Training and Evaluation

The feature set and target variable were separated, and the data was split into training and testing sets with an 80-20 ratio. Numeric features were scaled for the linear models to improve training stability, while Random Forest Regression used the raw features. Five regression models were trained: Linear Regression, Ridge, Lasso, ElasticNet, and Random Forest. Each model was evaluated on the test set using R^2 and RMSE. Performance was compared against predefined thresholds (R^2 ¿ 0.7 and RMSE ¡ 10) to determine whether each model met the success criteria.

```
# Split features and target
features = df_encoded.drop('exam_score', axis=1)
target = df_encoded['exam_score']

# Train-test split
features_train, features_test, target_train, target_test = train_test_split(
    features, target, test_size=0.2, random_state=RANDOM_STATE
)

# Scale numeric features for linear models
scaler = StandardScaler()
features_train_scaled = scaler.fit_transform(features_train)
features_test_scaled = scaler.transform(features_test)

# Define models
models = {
    'Linear Regression': LinearRegression(),
    'Ridge': Ridge(alpha=1.0, random_state=RANDOM_STATE),
    'Lasso': Lasso(alpha=0.1, random_state=RANDOM_STATE, max_iter=10000),
    'ElasticNet': ElasticNet(
        alpha=0.1,
        l1_ratio=0.5,
        random_state=RANDOM_STATE,
        max_iter=10000
    ),
    'Random Forest': RandomForestRegressor(
        n_estimators=100,
        random_state=RANDOM_STATE,
        n_jobs=-1
    )
}

# Train and evaluate each model
results = []
```

```
for model_name, model in models.items():
    if model_name == 'Random Forest':
        model.fit(features_train, target_train)
        predictions = model.predict(features_test)
    else:
        model.fit(features_train_scaled, target_train)
        predictions = np.clip(model.predict(features_test_scaled), 0, 100)

    r2 = r2_score(target_test, predictions)
    rmse = np.sqrt(mean_squared_error(target_test, predictions))
    mae = mean_absolute_error(target_test, predictions)

    results.append({
        'Model': model_name,
        'R²': r2,
        'RMSE': rmse,
        'MAE': mae,
        'R² Target': 'PASS' if r2 >= 0.7 else 'FAIL',
        'RMSE Target': 'PASS' if rmse <= 10 else 'FAIL'
    })
```

The regression models were evaluated on the test set using R^2 and RMSE. Ridge Regression emerged as the top performer, achieving the highest R^2 and lowest RMSE among all models while vanilla Linear Regression is a very close second. Lasso, and ElasticNet also met both success criteria, with slightly lower performance than Ridge. Random Forest Regression did not meet the predefined thresholds, with an R^2 of 0.684 and RMSE above 10.

Model	R^2	RMSE	R^2 Target	RMSE Target
Ridge	0.728905	9.722083	PASS	PASS
Linear Regression	0.728903	9.722104	PASS	PASS
Lasso	0.728428	9.730623	PASS	PASS
ElasticNet	0.726438	9.766219	PASS	PASS
Random Forest	0.684400	10.489790	FAIL	FAIL

Table 8: Evaluation Metrics for Regression Models

4.2 Feature Selection

Since Ridge and Linear Regression are the top-performing models, the next strategy is to apply Recursive Feature Elimination (RFE) to identify the top 10 informative features. In addition, we will explore filtering features based on the top 10 absolute correlations (Table 7) with exam scores to see if a reduced, highly relevant feature set can further improve R^2 and reduce RMSE for both models.

4.2.1 Top 10 - Absolute Correlated Features

Both Linear Regression and Ridge Regression were evaluated using all features (23) and the top 10 features selected based on absolute correlation with exam scores. Performance slightly declined for both models when using only the top 10 features, with lower R^2 and higher RMSE compared to using all features. Despite this decrease, both models still meet the predefined success criteria, indicating that the most correlated features capture most of the predictive information while simplifying the models.

Model	Features	R^2	RMSE	R^2 Target	RMSE Target
Linear (All)	23	0.728903	9.722104	PASS	PASS
Linear (Top 10)	10	0.722730	9.832182	PASS	PASS
Ridge (All)	23	0.728905	9.722083	PASS	PASS
Ridge (Top 10)	10	0.722731	9.832165	PASS	PASS

Table 9: Comparison of Regression Models Using All Features vs. Top 10 Features

4.2.2 Recursive Feature Elimination

Model	Features	R^2	RMSE	R^2 Target	RMSE Target
Linear (All)	23	0.728903	9.722104	PASS	PASS
Linear (RFE Top 10)	10	0.722730	9.832182	PASS	PASS
Linear (RFE Top 5)	5	0.676485	10.620505	FAIL	FAIL
Ridge (All)	23	0.728905	9.722083	PASS	PASS
Ridge (RFE Top 10)	10	0.722731	9.832165	PASS	PASS
Ridge (RFE Top 5)	5	0.676487	10.620484	FAIL	FAIL

Table 10: Evaluation Metrics for Linear Regression and Ridge Regression Using All Features and RFE-Selected Features

Linear Regression and Ridge Regression were evaluated using all features, as well as the top 10 and top 5 features selected via Recursive Feature Elimination (RFE). In all cases, reducing the number of features did not improve performance. Both models showed a slight decrease in R^2 and an increase in RMSE when using only the top 10 or top 5 features, with the top 5 features falling below the success thresholds. These results indicate that the full feature set provides the best predictive accuracy, and aggressive feature reduction does not benefit model performance.

4.2.3 Summary

Overall, feature selection using either the top 10 or top 5 features—whether by absolute correlation or RFE—did not improve model performance. The full feature set consistently yielded the highest R^2 and lowest RMSE, indicating that retaining all available features is the most effective strategy for predicting exam scores in this dataset.

4.3 Hyperparameter Tuning

Linear Regression does not require hyperparameter tuning, so we focus on optimizing Ridge Regression to investigate whether adjusting the regularization parameter can improve model performance. A range of alpha values was tested: 0.0001, 0.001, 0.01, 0.1, 0.5, 1.0, 2.0, 5.0, 10.0, 20.0, 50.0, 100.0, 200.0, 500.0, and 1000.0. We will be using Grid Search for this.

The baseline model with alpha = 1.0 achieved $R^2 = 0.728905$ and RMSE = 9.722083. The best performance was observed with alpha = 10.0, yielding a negligible improvement (R^2 +0.000010, RMSE -0.000181). Both baseline and tuned models satisfy the predefined R^2 and RMSE success criteria, indicating that the model is largely insensitive to the alpha value within this range.

4.4 Summary

After comparing all models using the full feature set, performing feature selection, and applying hyperparameter tuning, Ridge Regression with alpha = 10.0 emerges as the best-

Model	Alpha	R^2	RMSE	R^2 Target	RMSE Target
Ridge (Baseline)	1.0	0.728905	9.722083	PASS	PASS
Ridge (Tuned)	10.0	0.728915	9.721901	PASS	PASS

Table 11: Performance of Ridge Regression Before and After Hyperparameter Tuning

performing model. It consistently achieved the highest R^2 and lowest RMSE among all variations, outperforming Linear Regression and the reduced feature models. **This indicates that Ridge Regression with alpha 10.0 provides the most accurate and reliable predictions for exam scores in this dataset**.

5 Bias & Fairness Analysis

For bias analysis, we will use the tuned Ridge Regression model (alpha = 10.0), which was identified as the best-performing model after feature selection, hyperparameter tuning, and overall comparison. We will be using two main methods to aid us in this, SHAP and LIME.

5.1 SHapley Additive exPlanations (SHAP)

The SHAP analysis of the tuned Ridge Regression model indicates that `study_hours` and `class_attendance` are the most influential features driving exam score predictions, with mean absolute SHAP values of 11.75 and 5.11, respectively. Behavioral factors such as sleep quality and study methods also contribute to predictions, though to a lesser extent. Demographic and course-related features have minimal impact, suggesting that student behavior and learning habits are the primary determinants of predicted exam performance. These insights provide a clear understanding of the key drivers behind the model's predictions and can guide interventions to improve academic outcomes.

The SHAP analysis for the first three students reveals how key features influenced their predicted exam scores.

Student 1 had an actual score of 58.10, while the model predicted 75.94. The primary driver of the higher prediction was `study_hours`, which increased the score by 18.76 points. Attendance and poor sleep quality partially offset this increase, but overall, the model overestimated the student's performance (see Table 13).

Student 2 scored 64.40, while the model predicted 57.86. The main negative contributors were `study_hours` and `study_method_self-study`, which together reduced the predicted score by over 16 points. Attendance, sleep quality, and facility rating partially mitigated this reduction, but the model underestimated the student's performance (see Table 14).

Student 3 had an actual score of 46.60, with a predicted score of 34.62. The largest negative impact came from `study_hours` (-22.47) and `study_method_online videos` (-6.91), while adequate sleep and facility rating slightly increased the predicted score. Overall, the model underestimated performance due to low study hours and suboptimal study methods (see Table 15).

Conclusions: Across these examples, the model predictions are highly sensitive to `study_hours` and other behavioral factors such as attendance, study methods, and sleep quality. Demographic and course-related features had minimal influence. While the model sometimes overestimates (Student 1) or underestimates (Students 2 and 3), the feature contri-

Feature	Mean —SHAP Value—
study_hours	11.751869
class_attendance	5.106994
facility_rating_low	3.455381
study_method_online videos	2.922703
study_method_self-study	2.776959
study_method_group study	2.396610
sleep_quality_poor	2.238471
sleep_hours	2.152742
sleep_quality_good	1.960646
facility_rating_medium	1.703405
study_method_mixed	1.487935
gender_other	0.083001
course_ba	0.043449
course_b.tech	0.038228
course_bba	0.033427
course_b.sc	0.030553
gender_male	0.028933
exam_difficulty_hard	0.023575
exam_difficulty_moderate	0.023224
course_bca	0.020719
internet_access_yes	0.011870
age	0.008969
course_diploma	0.004744

Table 12: Mean Absolute SHAP Values for Features in the Tuned Ridge Regression Model

butions align intuitively with known drivers of academic performance, providing clear insights into which behaviors most impact predicted exam scores.

Actual Score	58.10
Predicted Score	75.94
Top 5 Contributing Features	**Effect on Score**
study_hours	+18.76
class_attendance	-6.20
sleep_quality_poor	-2.96
facility_rating_low	+2.58
facility_rating_medium	-2.51

Table 13: Prediction explanation for Student 1

5.2 Local Interpretable Model-agnostic Explanations (LIME)

The LIME analysis indicates that `study_hours`, `class_attendance`, `facility_ratings`, `sleep_quality`, and `study_methods` are the most influential features in predicting exam scores. For the first three students, `study_hours` consistently had the largest impact, either increasing or decreasing the predicted scores depending on the individual's behavior. Other factors such as class attendance and group study methods also contributed significantly to the predicted scores (see Tables 17–19).

Actual Score	64.40
Predicted Score	57.86
Top 5 Contributing Features	**Effect on Score**
study_hours	-9.00
study_method_self-study	-7.86
class_attendance	+4.44
sleep_quality_good	+3.18
facility_rating_low	+2.58

Table 14: Prediction explanation for Student 2

Actual Score	46.60
Predicted Score	34.62
Top 5 Contributing Features	**Effect on Score**
study_hours	-22.47
study_method_online videos	-6.91
sleep_hours	+3.43
sleep_quality_poor	-2.96
facility_rating_low	+2.58

Table 15: Prediction explanation for Student 3

Comparison with SHAP shows a high degree of agreement: nine out of the top ten LIME features overlap with SHAP's top features, yielding a 90% agreement rate. Shared features—including `study_hours`, `class_attendance`, `facility_ratings`, `sleep_hours`, `sleep_quality`, and `study_methods`—remain consistently important, while only `study_method_mixed` and `sleep_quality_good` differ in ranking. LIME weights are generally slightly larger than the corresponding SHAP values, but the overall ranking of influential features is consistent (see Table 16).

Overall, the LIME analysis confirms that behavioral and environmental factors dominate exam performance predictions. The individual-level insights align closely with SHAP results, supporting the reliability of the feature importance conclusions. These findings highlight areas where targeted interventions, such as improving study habits, attendance, and facility quality, could most effectively enhance student outcomes.

Feature	**Mean —LIME—**
study_hours	18.256393
study_method_self-study	7.304888
study_method_online videos	6.985439
study_method_group study	6.571487
class_attendance	5.863197
facility_rating_low	5.396517
sleep_hours	3.993542
study_method_mixed	3.832491
sleep_quality_poor	3.466520
facility_rating_medium	2.929569

Table 16: Mean absolute LIME values for top 10 contributing features

Actual Score	58.10
Predicted Score	75.94
Top 5 Contributing Features	**Impact on Score**
$study\_hours > 0.86$	+23.48
$class\_attendance \leq -0.87$	-10.64
$study\_method\_self - study \leq -0.51$	+7.60
$study\_method\_online\ videos \leq -0.51$	+7.11
$study\_method\_group\ study \leq -0.49$	+6.85

Table 17: LIME individual explanation for Student 1

Actual Score	64.40
Predicted Score	57.86
Top 5 Contributing Features	**Impact on Score**
$0.87 <$ study_hours ≤ 0.01	-7.64
$study\_method\_online\ videos \leq -0.51$	+6.82
$study\_method\_self - study > -0.51$	-6.74
$study\_method\_group\ study \leq -0.49$	+6.51
$facility\_rating\_low \leq -0.71$	+5.66

Table 18: LIME individual explanation for Student 2

Actual Score	46.60
Predicted Score	34.62
Top 5 Contributing Features	**Impact on Score**
$study\_hours \leq -0.87$	-23.65
$study\_method\_self - study \leq -0.51$	+7.57
$study\_method\_online\ videos > -0.51$	-7.03
$study\_method\_group\ study \leq -0.49$	+6.36
$facility\_rating\_low \leq -0.71$	+5.11

Table 19: LIME individual explanation for Student 3

5.3 Summary

The bias analysis, informed by both SHAP and LIME results, indicates that the model's predictions are driven by non-sensitive, performance-related features. Across both methodologies, the most influential factors consistently relate to academic behavior and learning conditions, such as study hours, class attendance, study methods, sleep patterns, and facility quality. These variables are directly connected to educational engagement and outcomes rather than personal or sensitive characteristics.

Importantly, none of the dominant features correspond to inherently sensitive attributes (e.g., demographic or identity-related factors like gender). The absence of such features among the top contributors suggests that the model does not rely on proxies for sensitive issues when generating predictions. This reduces the likelihood of unintended bias arising from protected or socially sensitive characteristics.

Overall, the agreement between SHAP and LIME strengthens confidence that the model's decision-making process is grounded in contextually appropriate and ethically acceptable inputs. As a result, no additional bias mitigation steps are required beyond standard validation, as the explanatory analyses already demonstrate that predictions are not influenced by sen-

sitive or protected factors.

6 Conclusion

This study evaluated multiple regression approaches for predicting student exam scores, including linear, regularized, and ensemble models. After systematic comparison, tuned Ridge Regression emerged as the most reliable model, achieving the best balance between predictive accuracy and generalization performance. Feature selection techniques such as correlation-based filtering and Recursive Feature Elimination did not improve model performance, indicating that the full feature set provided the most informative representation of student performance. Hyperparameter tuning further confirmed the robustness of the Ridge model, with only marginal but consistent improvements observed.

Across all analyses, study hours was identified as the single most influential feature affecting exam scores. Both SHAP and LIME consistently ranked study hours as the top contributor, followed by class attendance, study methods, sleep-related factors, and facility quality. These findings highlight that behavioral and environmental factors dominate academic performance predictions, while no sensitive or demographic-related attributes played a meaningful role in the model's decisions.

Finally, interpretability analyses using SHAP and LIME showed strong agreement, reinforcing confidence in the model's explanations at both global and individual levels. The consistency between these methods demonstrates that the model's predictions are transparent, stable, and ethically grounded. Overall, the results suggest that targeted interventions focusing on improving study habits, attendance, learning environments, and sleep quality are likely to have the greatest impact on student outcomes, making the tuned Ridge Regression model a suitable and trustworthy tool for educational performance analysis.