Predicting Exam Scores using Supervised Regression

Franrey Anthony S. Saycon

January 2026

Problem Understanding & Framing

Educational institutions aim to improve student performance and reduce failure rates. The objective of this study is to **develop a supervised regression model that predicts students' exam scores** using the available dataset to predict future academic risks.

Success Criteria

- Supervised regression performance thresholds
- Metrics:
  - $R^2 \geq 0.70$
  - RMSE ≤ 10

Dataset Overview

- Source: Kaggle Education Dataset
- 20,000 student records
- Mixed numerical and categorical features
- Behavioral, academic, and lifestyle factors

Numerical Features

- Student ID
- Age
- Study hours
- Class Attendance
- Sleep hours
- Exam score (target)

Numerical 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: Summary Statistics for Numerical Variables in the Dataset

Categorical Features

- Gender
- Course
- Internet access
- Sleep quality
- Study method
- Facility rating
- Exam difficulty

Categorical Features

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

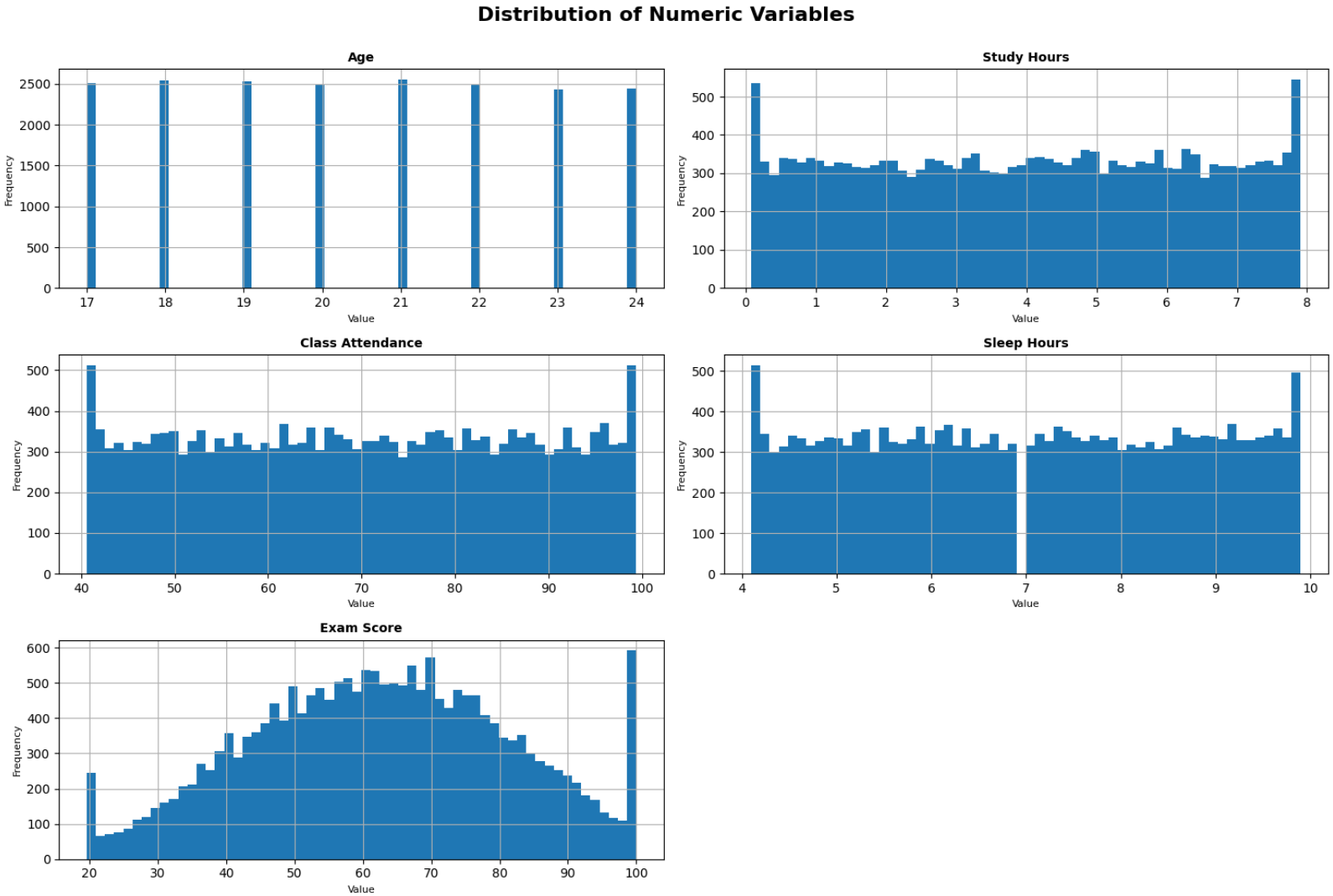
Table: Categorical Variables and Their Categories in the Dataset

EDA Summary

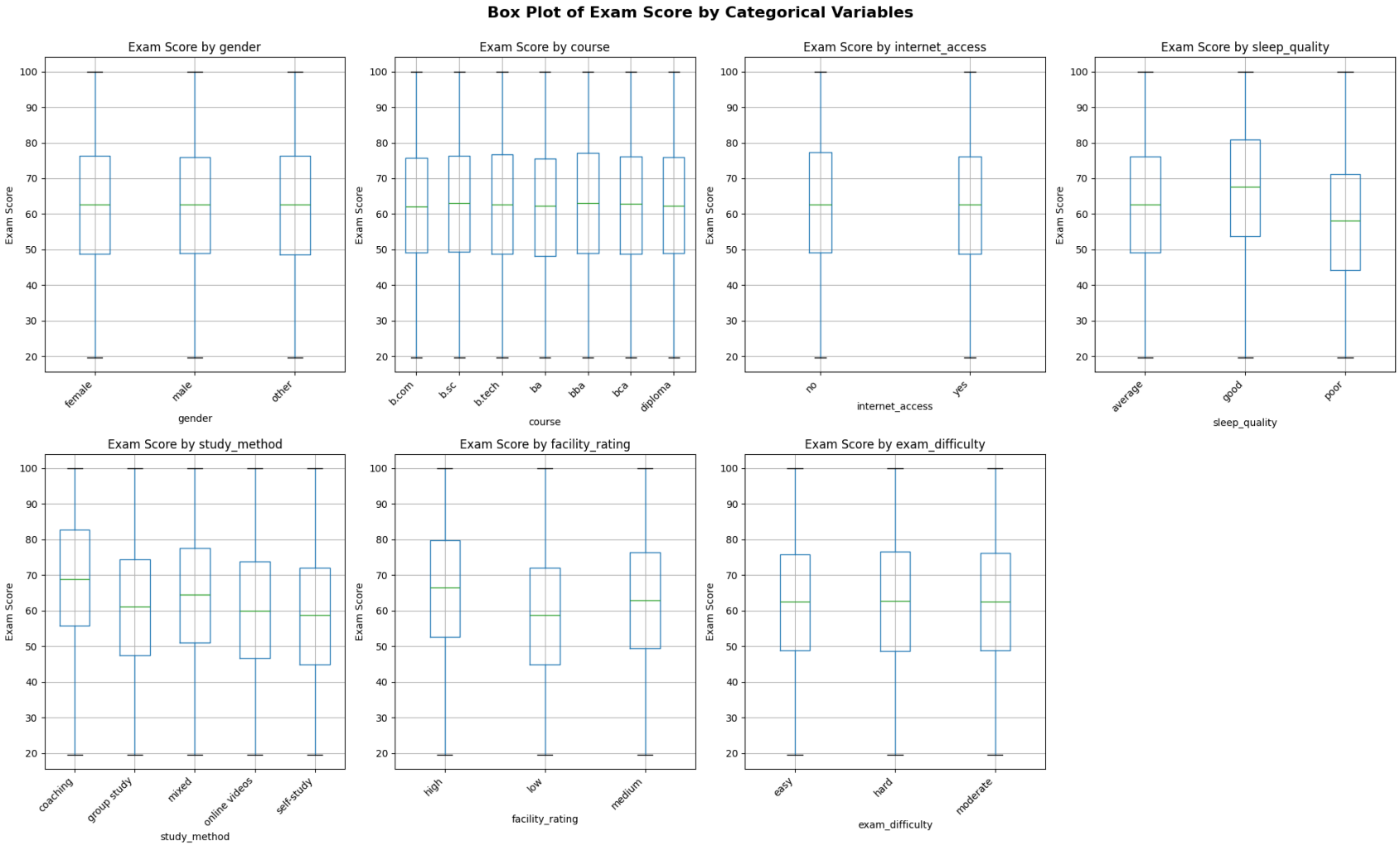
No missing or null values

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

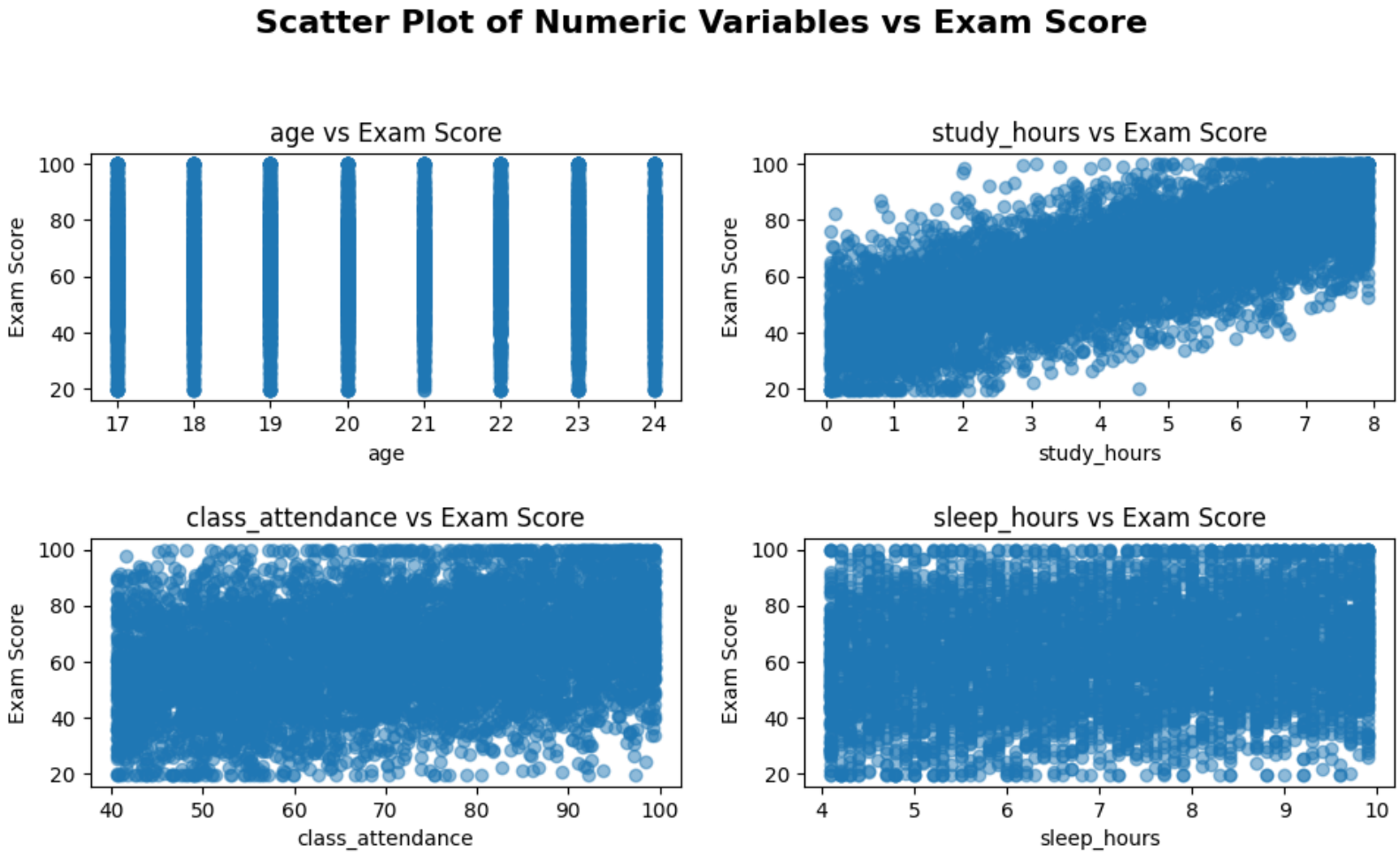
EDA Summary


Distribution of Numeric Variables
Age
Study Hours
Class Attendance
Sleep Hours
Exam Score
Frequency
Value

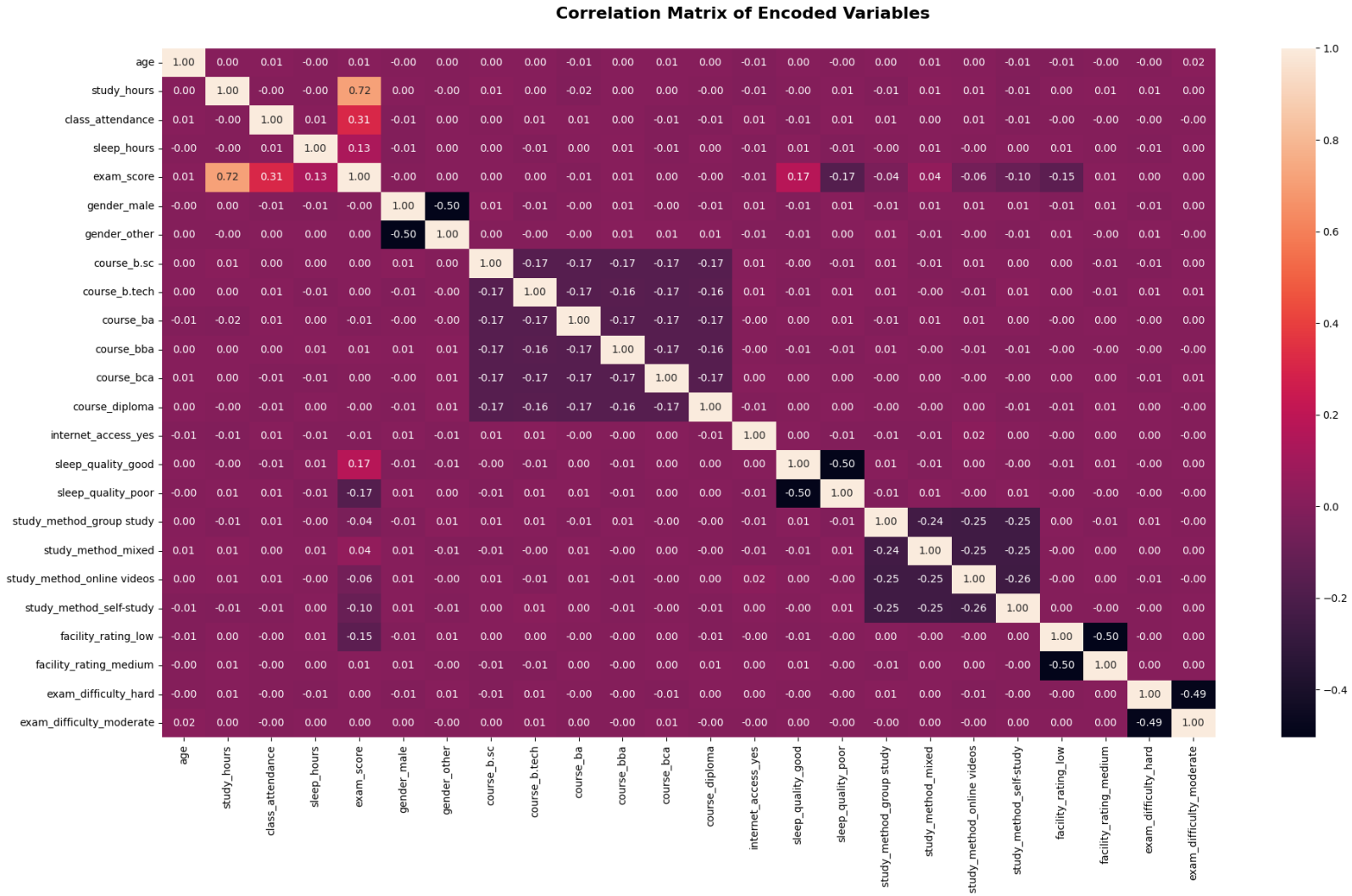
EDA Summary


Box Plot of Exam Score by Categorical Variables
Exam Score by gender
Exam Score by course
Exam Score by internet_access
Exam Score by sleep_quality
Exam Score by study_method
Exam Score by facility_rating
Exam Score by exam_difficulty
Exam Score
gender
female
male
other
course
b.com
b.sc
b.tech
ba
bba
bca
diploma
internet_access
no
yes
sleep_quality
average
good
poor
study_method
coaching
group study
mixed
online videos
self-study
facility_rating
high
low
medium
exam_difficulty
easy
hard
moderate

EDA Summary


Scatter Plot of Numeric Variables vs Exam Score
age vs Exam Score
Exam Score
age
study_hours vs Exam Score
study_hours
class_attendance vs Exam Score
class_attendance
sleep_hours vs Exam Score
sleep_hours

EDA Summary


Correlation Matrix of Encoded Variables
age 1.00 0.00 0.01 -0.00 0.01 -0.00 0.00 0.00 0.00 -0.01 0.00 0.01 0.00 -0.01 0.00 -0.00 0.00 0.01 0.00 -0.01 -0.01 -0.00 -0.00 0.02
study_hours 0.00 1.00 -0.00 -0.00 0.72 0.00 -0.00 0.01 0.00 -0.02 0.00 0.00 -0.00 -0.01 -0.00 0.01 -0.01 0.01 0.01 -0.01 0.00 0.01 0.01 0.00
class_attendance 0.01 -0.00 1.00 0.01 0.31 -0.01 0.00 0.00 0.01 0.01 0.00 -0.01 -0.01 0.01 -0.01 0.01 0.01 0.00 0.01 -0.01 -0.00 -0.00 -0.00 -0.00
sleep_hours -0.00 -0.00 0.01 1.00 0.13 -0.01 0.00 0.00 -0.01 0.00 0.01 -0.01 0.00 -0.01 0.01 -0.01 -0.00 0.01 -0.00 0.00 0.01 0.00 -0.01 0.00
exam_score 0.01 0.72 0.31 0.13 1.00 -0.00 0.00 0.00 0.00 -0.01 0.01 0.00 -0.00 -0.01 0.17 -0.17 -0.04 0.04 -0.06 -0.10 -0.15 0.01 0.00 0.00
gender_male -0.00 0.00 -0.01 -0.01 -0.00 1.00 -0.50 0.01 -0.01 -0.00 0.01 -0.00 -0.01 0.01 -0.01 0.01 -0.01 0.01 0.01 0.01 -0.01 0.01 -0.01 0.00
gender_other 0.00 -0.00 0.00 0.00 0.00 -0.50 1.00 0.00 -0.00 -0.00 0.01 0.01 0.01 -0.01 -0.01 0.00 0.01 -0.01 -0.00 -0.01 0.01 -0.00 0.01 -0.00
course_b.sc 0.00 0.01 0.00 0.00 0.00 0.01 0.00 1.00 -0.17 -0.17 -0.17 -0.17 -0.17 0.01 -0.00 -0.01 0.01 -0.01 0.01 0.00 0.00 -0.01 -0.01 0.00
course_b.tech 0.00 0.00 0.01 -0.01 0.00 -0.01 -0.00 -0.17 1.00 -0.17 -0.16 -0.17 -0.16 0.01 -0.01 0.01 0.01 -0.00 -0.01 0.01 0.00 -0.01 0.01 0.01
course_ba -0.01 -0.02 0.01 0.00 -0.01 -0.00 -0.00 -0.17 -0.17 1.00 -0.17 -0.17 -0.17 -0.00 0.00 0.01 -0.01 0.01 0.01 0.00 -0.00 0.00 -0.00 0.00
course_bba 0.00 0.00 0.00 0.01 0.01 0.01 0.01 -0.17 -0.16 -0.17 1.00 -0.17 -0.16 -0.00 -0.01 -0.01 0.01 -0.00 -0.01 -0.01 -0.00 -0.00 -0.00 -0.00
course_bca 0.01 0.00 -0.01 -0.01 0.00 -0.00 0.01 -0.17 -0.17 -0.17 -0.17 1.00 -0.17 0.00 0.00 0.00 -0.00 0.00 -0.00 -0.00 0.00 0.00 -0.01 0.01
course_diploma 0.00 -0.00 -0.01 0.00 -0.00 -0.01 0.01 -0.17 -0.16 -0.17 -0.16 -0.17 1.00 -0.01 0.00 0.00 -0.00 -0.00 0.00 -0.01 -0.01 0.01 0.00 -0.00
internet_access_yes -0.01 -0.01 0.01 -0.01 -0.01 0.01 -0.01 0.01 0.01 -0.00 -0.00 0.00 -0.01 1.00 0.00 -0.01 -0.01 -0.01 0.02 0.00 -0.00 0.00 0.00 -0.00
sleep_quality_good 0.00 -0.00 -0.01 0.01 0.17 -0.01 -0.01 -0.00 -0.01 0.00 -0.01 0.00 0.00 0.00 1.00 -0.50 0.01 -0.01 0.00 -0.00 -0.01 0.01 -0.00 0.00
sleep_quality_poor -0.00 0.01 0.01 -0.01 -0.17 0.01 0.00 -0.01 0.01 0.01 -0.01 0.00 0.00 -0.01 -0.50 1.00 -0.01 0.01 -0.00 0.01 -0.00 -0.00 -0.00 -0.00
study_method_group study 0.00 -0.01 0.01 -0.00 -0.04 -0.01 0.01 0.01 0.01 -0.01 0.01 -0.00 -0.00 -0.01 0.01 -0.01 1.00 -0.24 -0.25 -0.25 0.00 -0.01 0.01 -0.00
study_method_mixed 0.01 0.01 0.00 0.01 0.04 0.01 -0.01 -0.01 -0.00 0.01 -0.00 0.00 -0.00 -0.01 -0.01 0.01 -0.24 1.00 -0.25 -0.25 -0.00 0.00 0.00 0.00
study_method_online videos 0.00 0.01 0.01 -0.00 -0.06 0.01 -0.00 0.01 -0.01 0.01 -0.01 -0.00 0.00 0.02 0.00 -0.00 -0.25 -0.25 1.00 -0.26 -0.00 0.00 -0.01 -0.00
study_method_self-study -0.01 -0.01 -0.01 0.00 -0.10 0.01 -0.01 0.00 0.01 0.00 -0.01 -0.00 -0.01 0.00 -0.00 0.01 -0.25 -0.25 -0.26 1.00 0.00 -0.00 -0.00 0.00
facility_rating_low -0.01 0.00 -0.00 0.01 -0.15 -0.01 0.01 0.00 0.00 -0.00 -0.00 0.00 -0.01 -0.00 -0.01 -0.00 0.00 -0.00 -0.00 0.00 1.00 -0.50 -0.00 0.00
facility_rating_medium -0.00 0.01 -0.00 0.00 0.01 0.01 -0.00 -0.01 -0.01 0.00 -0.00 0.00 0.01 0.00 0.01 -0.00 -0.01 0.00 0.00 -0.00 -0.50 1.00 0.00 0.00
exam_difficulty_hard -0.00 0.01 -0.00 -0.01 0.00 -0.01 0.01 -0.01 0.01 -0.00 -0.00 -0.01 0.00 0.00 -0.00 -0.00 0.01 0.00 -0.01 -0.00 -0.00 0.00 1.00 -0.49
exam_difficulty_moderate 0.02 0.00 -0.00 0.00 0.00 0.00 -0.00 0.00 0.01 0.00 -0.00 0.01 -0.00 -0.00 0.00 -0.00 -0.00 0.00 -0.00 0.00 0.00 0.00 -0.49 1.00
age study_hours class_attendance sleep_hours exam_score gender_male gender_other course_b.sc course_b.tech course_ba course_bba course_bca course_diploma internet_access_yes sleep_quality_good sleep_quality_poor study_method_group study study_method_mixed study_method_online videos study_method_self-study facility_rating_low facility_rating_medium exam_difficulty_hard exam_difficulty_moderate
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

EDA: Key Observations

- Pre-Processing Results
  - No missing or null values found.
  - No outliers were found using IQR method and box plots.
  - Distribution of numerical features is acceptable.
  - Student ID will be removed from the considerations as it's just an identifier.
  - One hot encoding will be applied to the categorical features with drop first.

EDA: Key Observations

- Correlation Analysis (Numerical Features)
  - Study Hours is the top candidate for a strong linear relationship with exam scores.
  - Class Attendance showed minor signs.
  - Sleep hours showed little to no signs.
  - Age doesn't show a linear relationship at all, being the least candidate for regression.

EDA: Key Observations

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: Top 5 Positive Correlations with Exam Score

EDA: Key Observations

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: Bottom 5 Correlations with Exam Score

EDA: Key Observations

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: Top 10 Features by Absolute Correlation with Exam Score

Model Implementations

- Linear Regression
- Ridge Regression
- Lasso Regression
- ElasticNet Regression
- Random Forest Regression

Training & Evaluation Setup

- 80–20 train-test split
- Feature scaling in linear models
- Metrics:
  - R^2
  - RMSE

Model Performance Comparison

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: Evaluation Metrics for Regression Models

Model Implementation - Summary

- Performance Comparison
  - Ridge outperforms all other methods.
  - Linear Regression although second only has a marginal difference with Ridge.
  - Random Forest fails our success criteria.
- Conclusion: Ridge and Linear will be our main methods to be hopefully further optimized through feature selection and hyper parameter tuning.

Feature Selection

- Methods:
  - Absolute correlation filtering
  - Recursive Feature Elimination (RFE)

Feature Selection - Top 10 ABS Correlations

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

Feature Selection - RFE

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

Feature Selection - Summary

Feature selection using top 10 absolute correlation features, RFE top 5, and RFE top 10 - did not reveal any improvements on our success metrics for both methods.

Hyperparameter Tuning

- Linear Regression:
  - No tuning required
- Ridge Regression:
  - Grid search over alpha
  - Best alpha $= 10.0$
- Marginal but consistent improvement

Hyperparameter Tuning

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: Performance of Ridge Regression Before and After Hyperparameter Tuning

Final Model Selection

- Tuned Ridge Regression (Alpha 10.0)
- Highest R^2
- Lowest RMSE

Bias & Fairness Analysis

- Methods used:
  - SHAP
  - LIME

SHAP Insights

- Most influential features (descending order of MEAN abs(SHAP values)):
  - Study hours
  - Class attendance
  - Facility Rating
  - Study Method
  - Sleep Quality
- Least influential features:
  - Gender
  - Course
  - Exam Difficulty
  - Age
  - Internet Access

SHAP Insights

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

Table: Mean absolute SHAP values for top 10 contributing features

LIME Insights

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: Mean absolute LIME values for top 10 contributing features

Bias & Fairness Analysis — Summary

- Model predictions driven by non-sensitive, performance-related features
- Most influential factors:
  - Study hours
  - Class attendance
  - Study methods
  - Sleep patterns
  - Facility quality

- 9 out of top 10 features overlap between SHAP and LIME
- No dominant contribution from sensitive attributes (e.g. gender)
- No evidence of proxy bias or reliance on protected characteristics
- Explanations grounded in educationally relevant and actionable inputs
- No additional bias mitigation required beyond standard validation

Conclusion

- Multiple supervised regression models were evaluated for predicting student exam scores
- Tuned Ridge Regression ($\alpha = 10.0$) achieved the best balance of accuracy and generalization
- Success criteria were satisfied:
  - $R^2 \geq 0.70$
  - RMSE ≤ 10
- Feature selection (correlation-based and RFE) did not improve performance
- Full feature set provided the most informative representation of student performance

Conclusion

- Study hours emerged as the strongest predictor of exam scores across all analyses
- Behavioral and environmental factors dominated predictions:
  - Class attendance
  - Study methods
  - Sleep patterns
  - Facility quality
- SHAP and LIME explanations showed strong agreement and high interpretability
- No evidence of bias from sensitive or demographic-related features
- Tuned Ridge Regression is a reliable, interpretable, and ethically sound model for educational performance analysis for this dataset