STUDENT EXAM SCORE PREDICTION PROJECT

Forecasting Exam Results Using Academical and Behavioral Features

JUNE 1, 2025

ALEX YEPUR
alexyepur@gmail.com

Contents

1. Executive Summary

This report presents a comprehensive analysis of the **Student Academic and Behavioral Dataset** with the goal of predicting **exam scores** using machine learning techniques.

The primary objectives of this project were to:

- Explore and understand the key features influencing student performance;
- Engineer relevant features from raw data to improve model accuracy;
- Benchmark multiple regression models to identify the best predictive approach;
- Provide actionable insights based on the analysis.

The dataset contains **1,000** records and **16** features, including academic metrics such as **exam scores, attendance rates, after school self-study hours** and behavioral indicators such as **leisure time, sleep time, exercise frequency.**

After thorough exploratory data analysis and preprocessing steps — including handling missing data and feature scaling — several regression models were trained and evaluated. The best-performing model achieved an **R^2 score of 0.9**, demonstrating strong predictive power by explaining **90%** of the variance in the data.

This project offers a detailed understanding of the key factors and their respective impacts on student academic performance. The findings can assist educational institutions in enhancing educational quality, monitoring emerging negative trends among students, and diagnosing the root causes behind these patterns.

2. Introduction

Project Motivation

This project focuses on analyzing a dataset containing academic and behavioral information about students to predict exam scores. The motivation behind this study stems from the rapidly increasing volume of data and information that demands proficient skills to effectively process, analyze, and leverage in a timely manner. As new discoveries and technologies emerge daily, the effectiveness of student learning directly impacts our ability to adopt and utilize these innovations.

Objectives

The outcomes of this investigation should be viewed not merely as improvements in student grades, but as **part of a comprehensive strategy** aimed at enhancing the **overall quality of education**. While grades represent final results, they often fail to reveal the underlying causes of student performance. The core objective of this project is to identify the most influential factors driving student success*. Only through precise measurement and understanding of these root causes can educational institutions make informed decisions, allocate resources efficiently, and, as a result, ultimately improve learning outcomes.*

Target Audience

The target audience for this analysis includes educational institutions, administrators, curriculum developers, teachers, policymakers, and analysts interested in leveraging data-driven methods to enhance student success and identify students at risk of underperforming. Additionally, students seeking to improve their academic performance and learning strategies can benefit from these insights. *Understanding the key factors equips all stakeholders with clear insights into the underlying causes and facilitates the development of targeted, effective strategies for improvement.*

3. Project Objective

The primary goal of this project is to develop a machine learning pipeline capable of predicting students' exam scores based on **academic and behavioral features**. These features capture general patterns in student behavior and academic habits, including study time, class attendance, sleep duration, lifestyle factors, and leisure activities.

At the core of the solution is a **modular and extensible pipeline**, designed to ensure reproducibility, scalability, and efficient experimentation. The pipeline supports robust and maintainable modeling by:

- **Preventing data leakage** through a clear separation of preprocessing steps applied during training versus evaluation;
- **Applying feature engineering techniques** to improve the expressiveness and informativeness of input variables;
- **Automatically selecting and applying appropriate transformations** to different feature types using a dynamic ColumnTransformer;
- **Handling near-raw input data**, with manual review applied only to missing values and duplicate records to preserve data integrity and volume.

The task is framed as a **regression problem**, as the target variable — students' exam scores — is a continuous numerical outcome. To address this, a diverse set of regression models will be tested, including:

- Linear Regression (both with and without regularization),
- Decision Trees,
- Random Forests, and
- Gradient Boosting models (e.g., XGBoost).

This **diverse modeling suite** was selected to reflect different algorithmic strengths and biases. Each model processes data differently, by comparing them, we can evaluate which algorithm is best suited to the structure and characteristics of this dataset.

To evaluate model performance from multiple perspectives, the following metrics will be used:

- **R^2 Score** and **Explained Variance** — to assess how well variance in the target is captured;
- **Mean Squared Error (MSE)** and **Root Mean Squared Error (RMSE)** — to penalize large prediction errors;

- **Mean Absolute Error (MAE)**, **Median Absolute Error**, and **Maximum Error** — to provide additional insight and robustness in evaluation.

The ultimate objective is not only to produce accurate predictions but also to deliver interpretable insights that can support early intervention strategies and improve educational decision-making. Therefore, among models with comparable predictive performance, preference will be given to those that provide clearer and more meaningful feature interpretability.

4. Data Overview and Analysis

4.1 Dataset Description

The dataset used in this project consists of **1,000** student records collected from **Kaggle repository**. It includes **15 basic academic and behavioral features** along with **one target variable**, the exam score.

Data composition and features:

- **student_id** (object): Unique identifier for each student.
- **age** (int64): Age of the student in years.
- **gender** (object): Gender category (e.g., male, female, other).
- **study_hours_per_day** (float64): Hours spent on extracurricular study daily.
- **social_media_hours** (float64): Daily hours spent on social media.
- **netflix_hours** (float64): Daily hours spent watching Netflix or similar streaming.
- **part_time_job** (object): Whether the student has a part-time job (yes/no).
- **attendance_percentage** (float64): Percentage of classes attended.
- **sleep_hours** (float64): Average hours of sleep per day.
- **diet_quality** (object): Self-reported diet quality in three categories: fair, poor, good.
- **exercise_frequency** (int64): Number of exercise sessions per week.
- **parental_education_level** (object): Highest education level of parents, categorized as high school, bachelor, or master. Contains missing values (~9.1%).
- **internet_quality** (object): Quality of internet access, categorized as poor, average, or good.
- **mental_health_rating** (int64): Mental health rating on a scale from 1 to 10.
- **extracurricular_participation** (object): Participation in extracurricular activities (yes/no).
- **exam_score** (float64): Target variable, final exam score ranging from 0 to 100.

Data quality notes:

- The dataset contains no duplicate records.
- Missing values are present only in the parental_education_level feature, affecting approximately 9.1% of entries (91 records).

This dataset provides a comprehensive yet manageable set of features reflecting student habits, lifestyle, and academic background, suitable for building predictive models aimed at exam score estimation.

4.2 Initial Exploration

Missing Values, Duplicates and Data Imputation

Exploratory analysis begins with assessing missing values and duplicate records. The dataset is mostly complete, with the exception of **parental_education_level**, where approximately **9%** of the values were missing. These were imputed using a placeholder **'unknown'**, which was encoded as the lowest ordinal category (0) during mapping. Other known categories were mapped to progressively higher integers (1, 2, 3), preserving ordinal relationships and enabling model interpretability. No other missing or duplicate values were present in the dataset.

Categorical Feature Distributions

Categorical variables exhibit reasonably diverse distributions:

- **Gender:** The distribution is nearly balanced between male and female students, with approximately 5% identifying as *other*. While this diversity provides modeling stability, the small size of the *other* group may reduce statistical reliability in its coefficient estimation.
- **Part-time Job:** Only 21% of students report having a part-time job, indicating that the majority are not combining studies with employment. This could suggest that the working subgroup may face time constraints or increased stress levels, possibly impacting academic outcomes.
- **Parental Education Level:** The distribution is as follows: High School (~39%), Bachelor's (~35%), Master's (~17%), and Unknown (remaining). This variable serves as a proxy for educational support at home and may strongly influence performance. In general, students with more educated parents are expected to have access to more academic resources and encouragement.
- **Diet Quality:** 38% of students report a *good* diet, 43% *fair*, and 19% *poor*. While self-reported, dietary quality may reflect underlying lifestyle habits affecting energy levels, cognitive performance, and academic stamina. It can be considered a behavioral signal rather than just a health measure.
- **Internet Quality:** About 39% rate their connection as *average*, 45% as *good*, and 16% as *poor*. Given the growing reliance on online learning and digital study tools, limited internet quality could act as a bottleneck in preparation and academic efficiency, making this feature highly relevant. However, on the other hand, high-speed and reliable internet access might also encourage students to spend more time on non-academic online activities — such as social

media, streaming platforms, and gaming — potentially diverting time and attention away from study-related tasks.

- **Extracurricular Participation:** Only 31% of students engage in extracurricular activities. This low involvement could signal high academic load or low institutional engagement. It may correlate with time management skills, motivation, or general well-being — all of which are known to affect academic success.

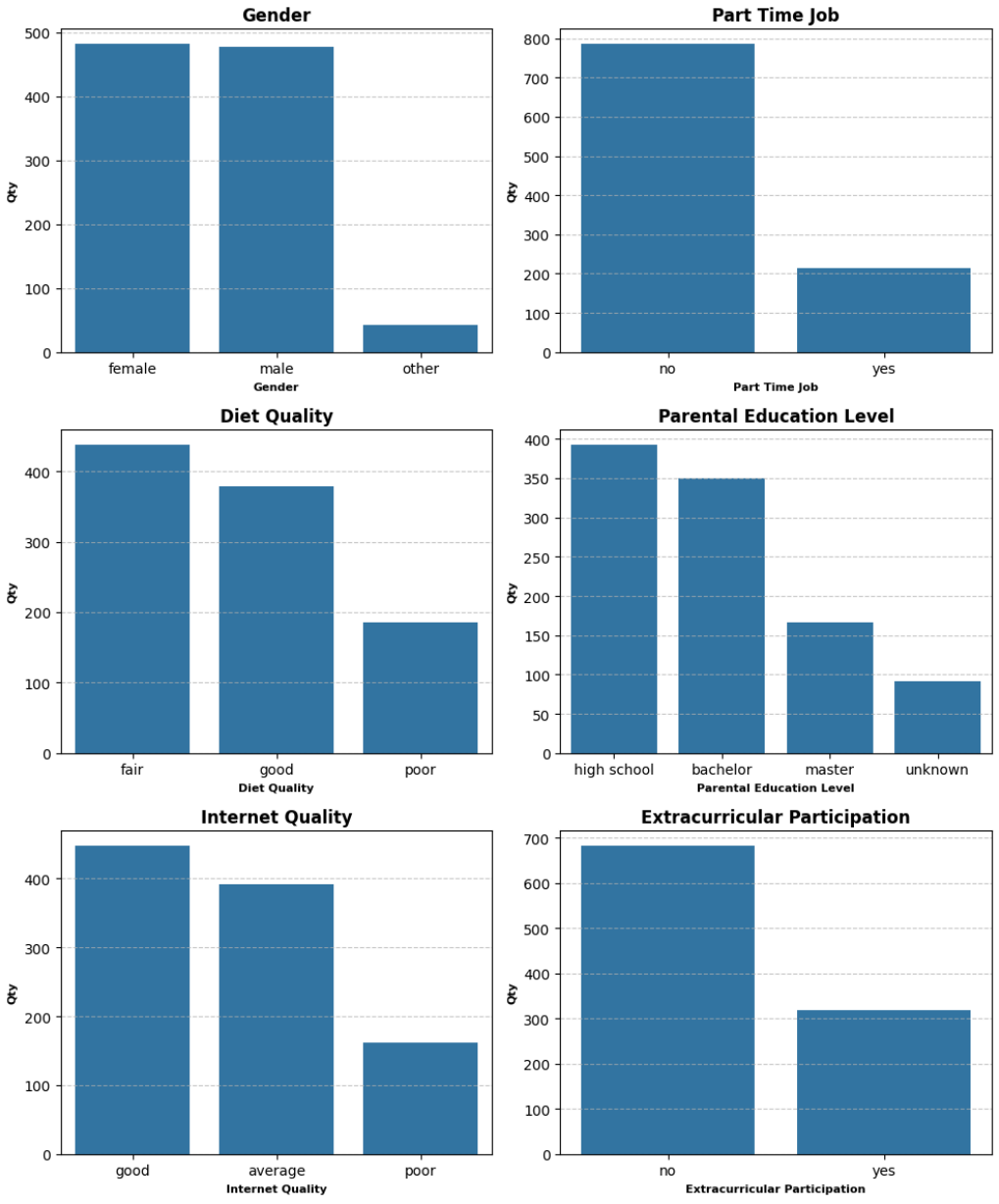

Figure 4.1 Categorical Features Distribution among Students.

Categorical variables in the dataset capture both socio-economic context and behavioral patterns. Their distributions not only describe the sample composition but also offer insight into which features may hold explanatory power in predicting academic performance.

Numerical Feature Distributions

- **Age:** Spanning from 17 to 24 years, with a relatively uniform distribution. Slightly higher density observed at ages 23–24.

- **Study Hours per Day:** This feature follows a roughly Gaussian distribution centered around 3.5 hours/day, ranging from 0 to 8.3. While some students report zero daily study hours, the peak around 3–4 hours indicates a general engagement level. A positive association between study time and academic performance is anticipated, making this variable a potentially strong predictor of exam scores.

- **Social Media and Netflix Usage:**
  Daily social media usage ranges from 0 to 7.2 hours, and streaming hours (e.g., Netflix) reach up to 5.4 hours, both exhibiting slight right-skewed distributions. While the majority of students fall within moderate usage levels, a notable minority reports high engagement.
  These features show how much time students spend on leisure activities like social media and streaming. While not directly negative, more time here might mean less time for studying, so the model can use this information to better predict outcomes.

- **Attendance:** Attendance spans from 56% to 100%, forming a bell-shaped distribution slightly skewed right, suggesting that most students maintain strong attendance, though a minority show significant absences. Since attending class helps students learn, attendance is likely an important feature for prediction.

- **Sleep Hours:** Ranging from 3.2 to 10 hours per night, the distribution appears symmetric and bell-shaped, centered around 6.5–7 hours. The reported minimum of 3.2 hours seems either unlikely or significantly underestimated. If the model's performance suffers, removing this outlier could be considered. However, given the standardization of numerical features, this single extreme value is unlikely to have a significant impact on the overall results. Both sleep deprivation and oversleeping may reduce cognitive efficiency. Optimal performance typically aligns with 7–8 hours, suggesting a U-shaped relationship may exist between this feature and exam scores.

- **Exercise Frequency:** Reported from 0 to 6 sessions per week, the distribution is relatively flat, indicating no dominant pattern Physical activity is often associated with better mental health and cognitive function, suggesting a potential indirect influence on academic results.
- **Mental Health Rating:** Ratings range from 1 to 10, with spikes at the extreme values (1 and 10) and around the midpoint (5). This tri-modal pattern suggests that students' self-assessments tend to be polarized. This feature potentially impacts the model, as it reflects the student's mindset and motivation, which can directly influence their learning behavior and academic performance.

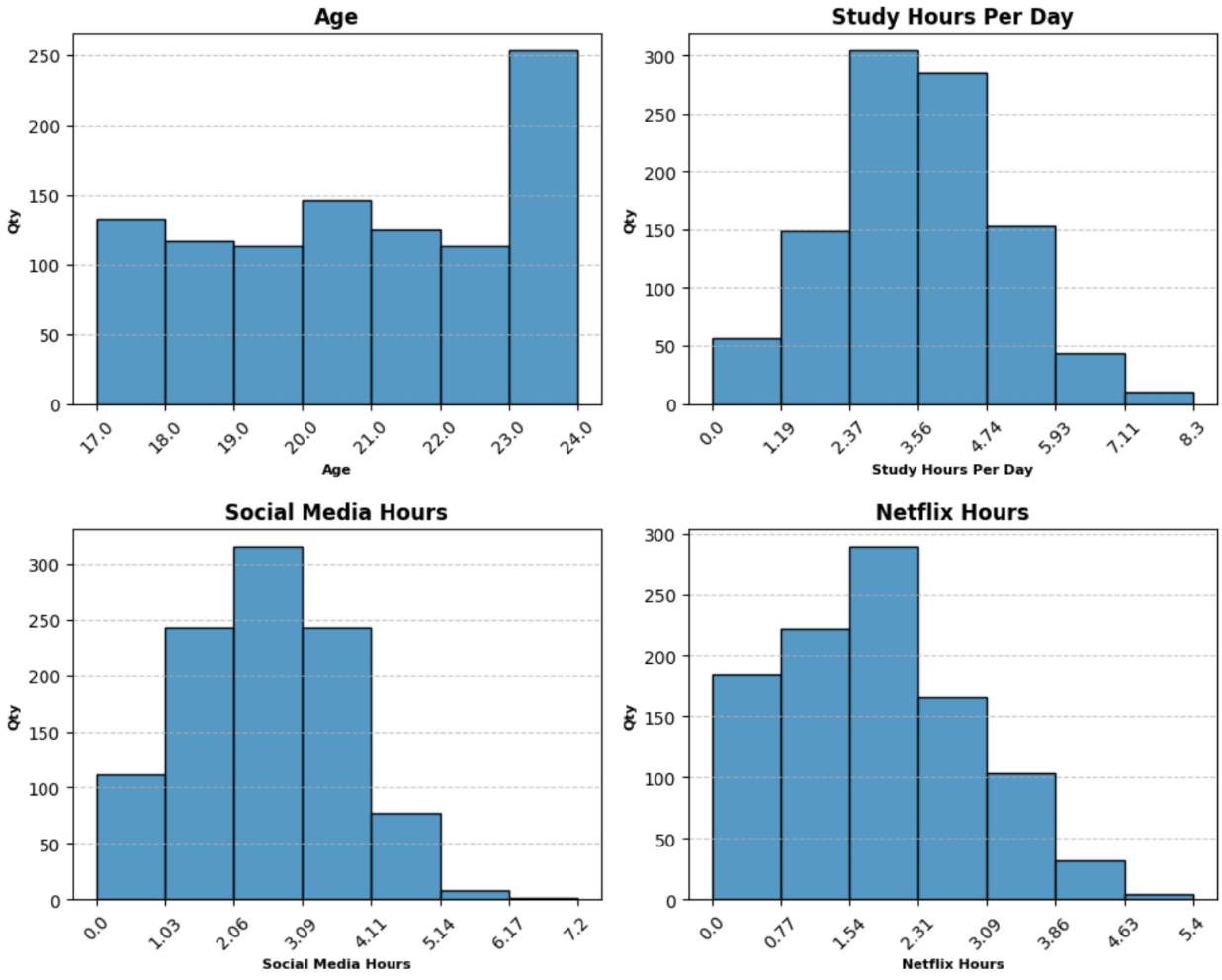

Figure 4.2a Numerical Features Distribution among Students.

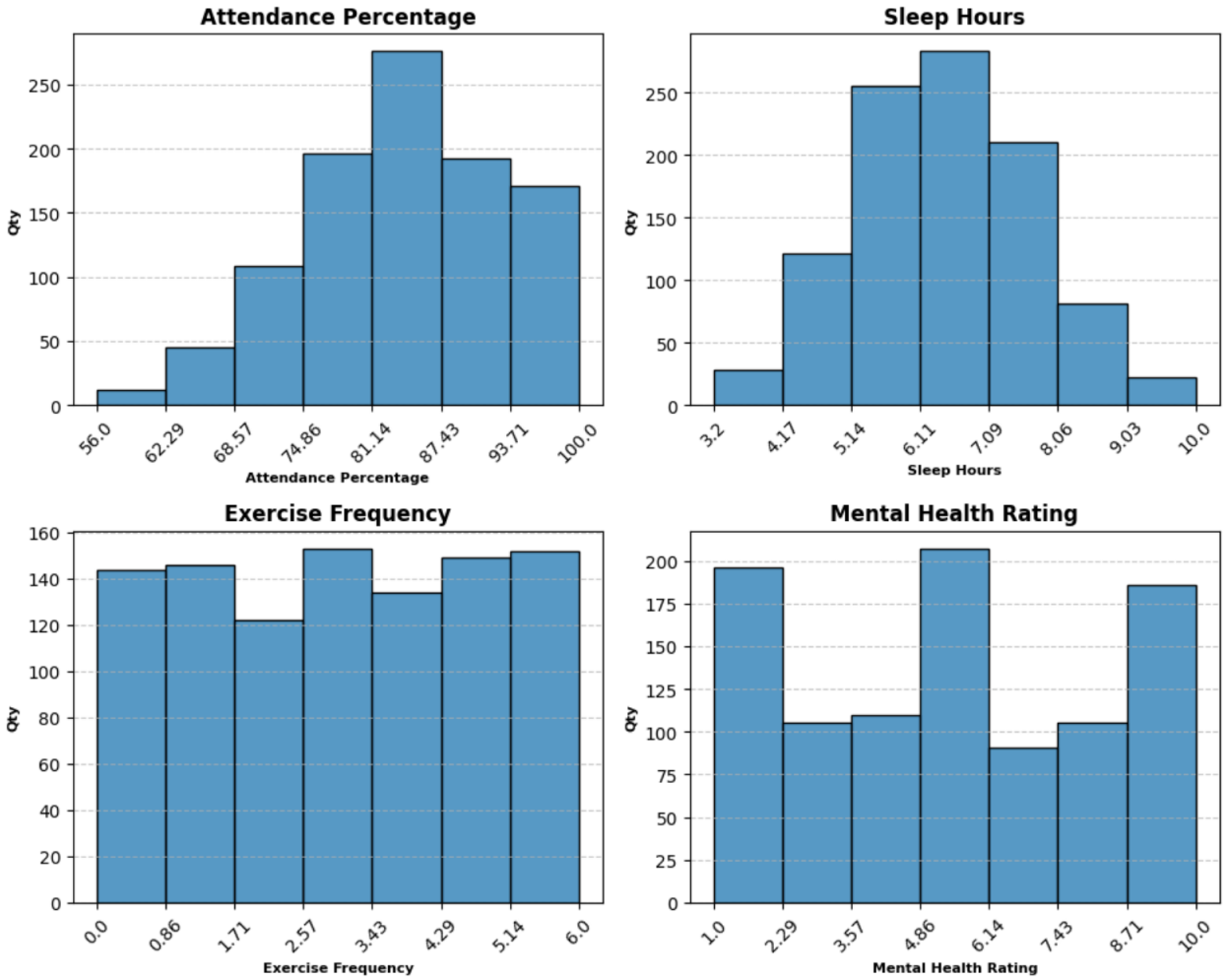

Figure 4.2b Numerical Features Distribution among Students.

These patterns demonstrate that while some features are normally distributed, others have inherent skew, which was assessed but not transformed, as skewness and kurtosis values fell within acceptable ranges and transformation was not necessary.

Target Variable Distribution — Exam Score

The target variable — *exam_score* — exhibits the following characteristics:

- The target variable — *exam_score* — has a mean of approximately 69.6 and a standard deviation of 16.9, reflecting a moderate but meaningful spread across the full range of possible values, from 18.4 to 100.
- The distribution is slightly right-skewed, with a notable concentration of high scores between 95 and 100. This suggests the presence of a subgroup of high-achieving students and indicates that low performance is relatively less common. Clustering near the top end is typical in educational

datasets and reflects an expected trend, rather than an anomaly. It is unlikely to negatively impact model performance.

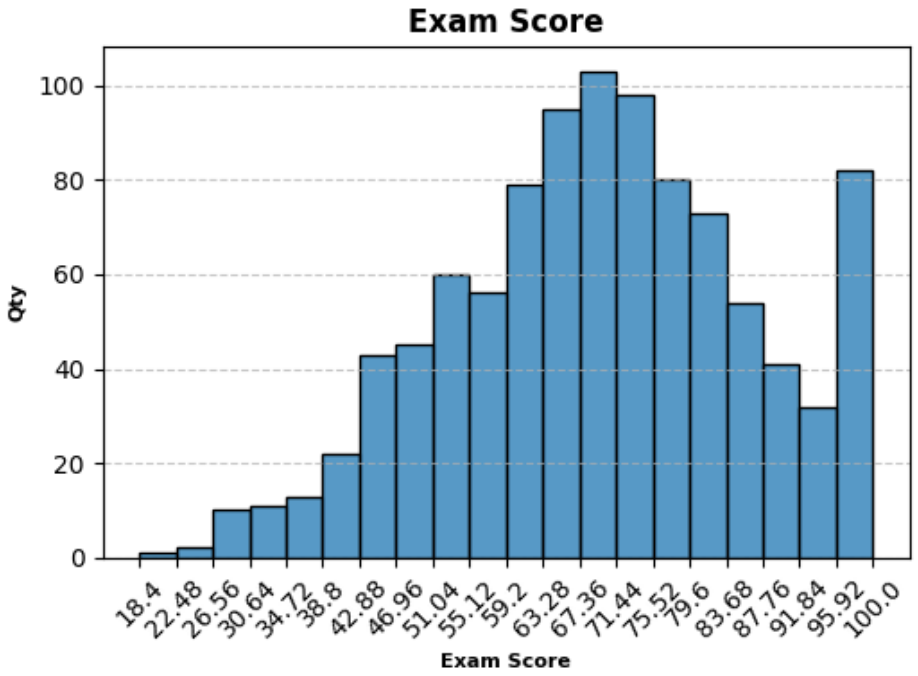

Figure 4.3 Exam Score Histogram.

Overall, the target distribution is well-suited for regression modeling and allows the model to capture both average and exceptional academic outcomes effectively.

Outlier Analysis

Outliers were assessed using:

- Interquartile Range (IQR)
- Z-score thresholds
- Visual inspection of distributions

IQR analysis identified the presence of outliers in several features. However, kurtosis values indicated that these outliers do not significantly distort the tails of the distributions, suggesting a limited impact on overall data shape. Moreover, these extreme values appear to reflect legitimate individual behavior patterns — such as unusually high study time or minimal engagement with streaming services — rather than data entry errors or anomalies. Therefore, none were removed, as all were deemed plausible and meaningful.

5. Data Preprocessing

Effective preprocessing is essential for ensuring that the modeling pipeline correctly interprets categorical and numerical features and applies appropriate transformations to each.

In this project, the preprocessing phase focused on building a modular transformation strategy using ColumnTransformer, allowing us to apply different preprocessing techniques to distinct feature groups based on their data type and semantic meaning.

Categorical variables in the dataset were treated based on their inherent structure:

Ordinal Encoding was applied to features with a natural ranking:

- diet_quality (Poor < Fair < Good)
- parental_education_level (Unknown < High School < Bachelor < Master)
- internet_quality (Poor < Average < Good)

These were mapped manually to ensure the encoded values reflect semantic progression.

One-Hot Encoding was used for nominal categorical features, such as gender, part_time_job, and extracurricular_participation, as these do not exhibit an inherent order and benefit from independent binary indicators.

All continuous variables (e.g., age, study_hours_per_day, attendance_percentage, sleep_hours) were standardized using **Standard Scaler** to zero mean and unit variance. This helps prevent features with larger numerical ranges from dominating the learning process and is considered a standard practice when working with numerical data.

6. Feature Development and Filtering

During the dataset analysis and transformation process, the following new features were engineered:

- **is_outlier_iqr** and **is_outlier_zscore** flag outliers using two statistical methods, with **is_outlier** marking only those flagged by both for stronger reliability.
- **is_night_owl:** boolean flag set to True if average sleep duration is less than 6 hours per day
- **atleast_18_y_o:** boolean flag indicating if the student is at least 18 years old, to evaluate the potential impact of younger age.

Multicollinearity analysis showed that none of the original base features exhibited significant multicollinearity. However, several of the newly engineered features **not mentioned** above demonstrated strong multicollinearity with their base counterparts and were therefore removed in favor of retaining the original variables.

At this stage, the feature set is considered acceptable and sufficiently comprehensive. After applying all transformations and encoding procedures, the **original 16 features** expanded to **23 derived features**. Given this moderate increase and the relatively low number of features, dimensionality reduction or explicit feature selection on the test model is not currently necessary. Further reduction risks losing valuable information without delivering a significant performance gain.

7. Modeling

At this stage, all preprocessing and transformation steps have been consolidated into a unified pipeline, ensuring that raw data is systematically prepared for modeling. The pipeline includes the following components:

- **Outlier Feature Adder Transformer**
  Adds binary flags indicating whether each observation is an outlier according to both the IQR and Z-score methods. These engineered features help capture distributional anomalies that may correlate with academic performance.

- **Feature Engineering Pipeline**
  A sub-pipeline that creates new derived features, such as is_night_owl (for students sleeping less than 6 hours) and atleast_18_y_o (to flag younger participants). These features aim to introduce domain-informed signals that may enhance model discrimination.

- **Drop High VIF Columns Transformer**
  Automatically removes predefined features exhibiting high multicollinearity. Although implemented programmatically, the list of dropped columns was selected manually based on prior VIF analysis to maintain transparency and modeling control.

- **Array To DF Transformer** applied on Column Transformer
  Wraps the **ColumnTransformer (ct)** and converts its NumPy output back into a pandas.DataFrame, preserving feature names for easier debugging and interpretability.
  Internally, **ct** applies:

  - **OrdinalEncoder** for ordinal categorical variables (e.g., diet_quality, parental_education_level, internet_quality),
  - **OneHotEncoder** for nominal categorical features (e.g., gender, part_time_job),
  - **StandardScaler** for continuous numerical variables.

- **Estimators**
  Since multiple models will be evaluated during the initial benchmarking phase, the final estimator (model) will be passed dynamically as the last step in the pipeline within a loop. This setup allows for efficient model substitution and evaluation, reducing code duplication and ensuring consistent preprocessing across all candidates.

This modular and interpretable structure ensures consistency between training and inference stages, while maintaining a high degree of flexibility for future adjustments or diagnostics.
Importantly, it also safeguards against data leakage by ensuring that all transformers are fitted exclusively on the training data. This strict separation preserves the integrity of the validation set and guarantees a more reliable evaluation of model performance.

7.1 Initial Model Benchmark

Before testing the main regression models, a baseline Dummy Regressor was created to establish a minimum performance threshold. Its results were:

- **MSE:** 263.57
- **MAE:** 12.92
- **RMSE:** 16.23

This baseline serves as a reference point; all subsequent models are expected to significantly outperform these metrics.
Note: R^2 and Explained Variance are not reported for the baseline, as the Dummy Regressor predicts the mean value for all samples, resulting in an R^2 of 0 and zero explained variance by definition.

The selected models represent a range of regression approaches:

Linear Models:

- **Linear Regression:** Basic linear approach.
- **Ridge Regression:** Linear model with L2 regularization to reduce overfitting.
- **Lasso Regression:** Linear with L1 regularization, performs feature selection.
- **ElasticNet Regression:** Combines L1 and L2 regularization for flexibility.

Kernel-Based Model:

- **Support Vector Regressor (SVR):** Captures nonlinear relationships via kernels.

Tree-Based and Ensemble Models:

- **Decision Tree Regressor:** Simple decision tree, prone to overfitting.
- **Random Forest Regressor:** Ensemble of trees to reduce variance and noise.
- **Gradient Boosting Regressor:** Sequentially builds trees to improve accuracy.
- **XGBoost Regressor:** Optimized gradient boosting with regularization and speed improvements.

Models were evaluated using *R^2, Explained Variance, MSE, MAE, RMSE, and Max Error.*

Top 6 models with **$R^2 > 0.8$** and their results:

Table 7.1 Top 6 Models Benchmark Summary.

Model	*R^2*	*Explained Variance*	*MSE*	*Median AE*	*MAE*	*RMSE*	*Max Error*
Linear Regression	0.91	0.91	29.58	3.61	4.33	5.44	15.81
Ridge Regression	0.91	0.91	29.70	3.62	4.34	5.45	15.82
XGBoost Regressor	0.89	0.89	36.22	4.21	4.95	6.02	17.65
Gradient Boosting	0.89	0.89	36.61	4.26	4.91	6.05	18.73
Lasso Regression	0.87	0.87	41.95	4.88	5.27	6.48	18.63
Random Forest Regressor	0.84	0.84	50.96	4.89	5.81	7.14	22.77

The selected models exhibit robust performance, substantially outperforming the baseline. Given that Grid Search with Cross Validation are computationally expensive and may not produce significant gains on less performant models, we exclude those models at this stage and focus hyperparameter optimization solely on the top-performing candidates to ensure efficient use of resources.

7.2 Hyperparameters Tuning and Final Evaluation

Hyperparameter grids were manually defined for each selected top-performing model, based on their specific operational characteristics and taking into account the nature of our dataset. Using a unified structure, each model was incorporated into the preprocessing pipeline, and hyperparameter optimization was performed via RandomizedSearchCV with cross-validation. This approach ensured systematic tuning and robust performance tracking, ultimately returning the most effective parameter sets along with their corresponding R^2 metrics.

For each model, the best hyperparameters were identified and applied prior to the final evaluation. The tuned models were then assessed on the held-out test set to provide an unbiased estimate of their generalization performance.

After evaluation on the held-out test set, the tuned models demonstrated the following performance:

Table 7.2 Top 6 Models Final Evaluation Summary.

Model	*R2 Score*	*Explained Variance*	*MSE*	*Median AE*	*MAE*	*RMSE*	*Max Error*
Lasso	**0.8997**	**0.8998**	**28.60**	**3.54**	**4.18**	**5.35**	**21.42**
Linear Regression	**0.8974**	**0.8976**	**29.24**	**3.43**	**4.20**	**5.41**	**23.85**
Ridge	**0.8974**	**0.8976**	**29.24**	**3.43**	**4.20**	**5.41**	**23.85**
Gradient Boosting Regressor	0.8803	0.8809	34.13	4.12	4.71	5.84	15.46
XGBoost Regressor	0.8683	0.8688	37.53	4.18	4.99	6.13	17.15
Random Forest Regressor	0.8604	0.8611	39.80	4.78	5.14	6.31	17.12

As we can see, **Linear models** exhibit the strongest overall performance in this evaluation, with ensemble methods also delivering robust results albeit with marginally higher error metrics. Notably, unlike the initial benchmark phase, **Lasso regression** emerged as the leading model. Its built-in **L1 regularization** effectively performs feature selection by shrinking less important coefficients toward zero, thereby enhancing model generalization.

While its predictive metrics are comparable to other linear models, Lasso's **coefficient sparsity** offers a clear and interpretable framework for assessing feature importance and their respective impact on the prediction. *For these reasons, Lasso Regression was selected as our final model due to its effective regularization and interpretability, providing a strong balance between performance and feature relevance.*

8. Results Analysis

8.1 Model Deployment Configuration

The final deployment model selected is the **Lasso Regression** configured with:

- **max_iter**=10000
- **selection**='cyclic'
- **fit_intercept**=True
- **alpha**=0.1

8.2 Model Performance Metrics and Interpretation

The model, trained with optimized hyperparameters, was evaluated on a held-out test set. Key performance metrics are summarized below:

Table 8.1 Tunned Lasso Model Performance Metrics.

Metric	**Value**
R^2 Score	**0.8997**
Explained Variance	0.8998
Mean Squared Error (MSE)	**28.60**
Median Absolute Error	3.54
Mean Absolute Error (MAE)	4.18
Root Mean Squared Error (RMSE)	**5.35**
Max Error	**21.42**

An **R^2** score of approximately **0.9** indicates that the model explains **nearly 90%** of the variance in the target variable, which is excellent given the inherent noise and unpredictability in real-world educational data.

An **RMSE** of about **5.35** on a **100-point scale** implies that predictions deviate on average by **±5 points**, an acceptable error margin in the academic context.

The **maximum error** of **21.42** reflects the presence of outliers or difficult-to-predict instances, common in human behavior and performance variability.

The final evaluation clearly demonstrates that the Lasso regression model significantly outperforms the mean-based dummy baseline model. *Overall, the model demonstrates strong generalization capability, suitable for deployment in forecasting student exam outcomes.*

8.3 Feature Importance Analysis

The Lasso regression coefficients provide direct interpretability of feature contributions to the predicted exam scores. The absolute magnitude of each coefficient corresponds to the influence of that feature.

Table 8.2 Feature Importance Coefficients.

	Feature	Coefficient
0	Study Time (hr/day)	13.8
1	Mental Health Score	5.4
2	Social Media Time (hr/day)	-2.9
3	Exercise Frequency (times/week)	2.9
4	Sleep Time (hr/day)	2.3
5	Netflix Time (hr/day)	-2.2
6	Lessons Attendance Rate	1.3
7	Diet Quality	-0.22
8	Has Part Time Job	0.17
9	Age	-0.05
10	Internet Quality	-0.01

As we can observe, the most influential features for predicting a student's exam score include: Study Time, Mental Health Score, Social Media Time, Exercise Frequency, Sleep Time, Netflix Time and Lessons Attendance Rate. These features exhibited the **highest absolute coefficient values** in the Lasso regression model. While other non-zero features still contribute to the prediction, their individual impact is minimal. However, in combination, they may carry additional explanatory power, especially in edge cases where primary feature values are similar.

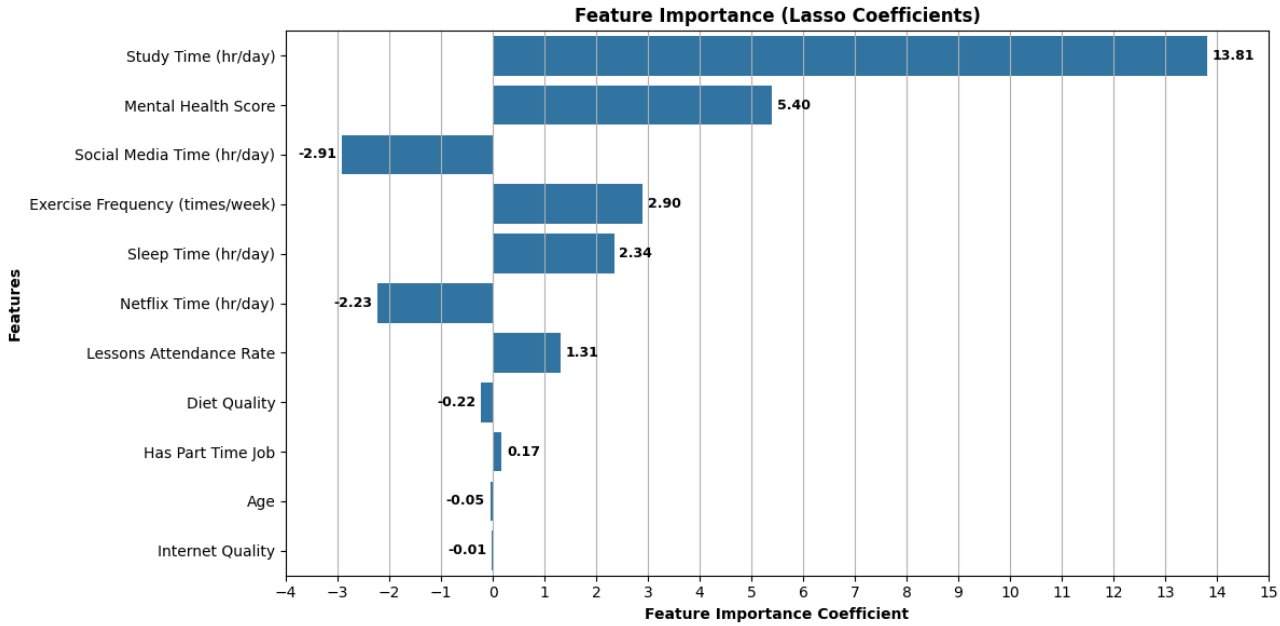

Figure 8.1 Feature Importance Visualization

9. Conclusions and Recommendations

This project successfully developed a robust predictive model for estimating student exam performance using behavioral and contextual data. Through a structured pipeline incorporating preprocessing, feature engineering, model selection, and hyperparameter optimization, we achieved high predictive accuracy. The final model — **Lasso Regression** — demonstrated strong generalization performance on unseen data, substantially outperforming the baseline dummy regressor across all key metrics (e.g. RMSE: 5.35 vs. 16.23).

The most influential predictor in the final model was clearly **Study Time (hours/day)**. This aligns with intuitive expectations — students who dedicate more time to self-study tend to achieve higher academic performance. For educational institutions and advisors, this feature should be a primary focus when designing personalized learning strategies or interventions, as it has the strongest direct correlation with exam success.

The second most significant variable was the **Mental Health Score**. Despite its subjectivity (being self-reported), it proved to be a valuable indicator. A higher score likely reflects greater emotional well-being, motivation, and life satisfaction — factors that can both enhance focus and reduce academic stress. This underlines the importance of supporting student mental health as part of academic success frameworks.

Additionally, positive lifestyle habits such as **adequate sleep** and **regular exercise** also showed meaningful contributions to better outcomes. These findings support the implementation of programs that promote not only academic discipline but also physical wellness and balanced daily routines.

A moderate but consistently **negative effect** was observed from features related to online entertainment, specifically time spent on **social media** and **streaming platforms** (e.g., Netflix). While these factors were not the strongest predictors individually, they become important when usage reaches excessive levels, especially if academic performance begins to decline. It's important to recognize that screen time in itself is not inherently harmful — students need rest and downtime. Therefore, interventions should be context-dependent. If a student maintains good performance despite some leisure activity, no correction may be necessary. However, when digital consumption becomes excessive and coincides with deteriorating results, targeted action is advisable. Many students may not be fully aware of how much time they spend on screens, making simple advisor feedback or reflective conversations effective early interventions.

From a holistic perspective, a well-balanced student lifestyle — one that includes structured study, regular physical activity, and adequate sleep — naturally limits opportunities for excessive screen time. Thus, promoting this balance can be a more sustainable strategy than focusing solely on reducing entertainment usage.

Those interested in applying these findings should note that **data quality and frequency of collection are critical**. For reliable deployment, institutions must implement consistent and accurate data-gathering processes across all relevant sources. The accompanying Jupyter notebook includes a deployment-ready example of the trained Lasso model. It uses only the selected non-zero features and provides exam score predictions with high **confidence level (~90%).** Future improvements could include expanding the dataset, engineering new features, and validating responses more rigorously — all of which may further refine the model. However, the current level of explained variance already demonstrates strong performance and practical applicability **within the context of this dataset** and the behavioral predictors analyzed. Given the **inherently subjective** and **self-reported nature** of many features, achieving nearly 90% explained variance indicates that the model captures meaningful patterns and offers valuable insights for real-world educational settings.

10. Appendices

- **GitHub Repository (Jupyter Notebook & Source Code):**
  https://github.com/alexyepur/student-exam-score-prediction
- **Dataset Source:**
  https://www.kaggle.com/datasets/jayaantanaath/student-habits-vs-academic-performance