# Predicting Student Academic Performance: A Machine Learning Analysis of Study Habits and Lifestyle Factors


## Abstract

Student academic performance prediction has become increasingly important for educational institutions seeking to implement targeted interventions and support systems. This study analyzes a comprehensive dataset of 1,000 university students to identify key factors influencing exam performance using advanced machine learning techniques. We examine the relationships between study habits, mental health, screen time, sleep patterns, and academic outcomes. Our analysis reveals that study hours per day exhibits the strongest correlation with exam scores ($r = 0.825$), followed by mental health rating ($r = 0.322$). Ridge regression achieved the highest predictive accuracy with an $R^2$ score of 0.9015 and RMSE of 5.03 points. Feature engineering with polynomial terms significantly improved model performance, with study hour derivatives accounting for over 60% of predictive importance. Clustering analysis identified three distinct student performance groups with average scores of 61.95, 81.97, and 96.14 points respectively. Threshold analysis demonstrates that students studying 4+ hours daily score 35+ points higher than those studying less than 2 hours. Mental health ratings above 7 correlate with approximately 15-point score improvements. These findings provide actionable insights for educational interventions and student support systems.



**Keywords:** machine learning, educational data mining, student performance prediction, academic analytics, feature engineering, clustering analysis


---

## 1. Introduction

The digital transformation of education has generated unprecedented amounts of student data, creating opportunities for data-driven insights into academic performance factors. Understanding the complex relationships between student habits, lifestyle factors, and academic outcomes is crucial for developing effective educational interventions and support systems.

Traditional approaches to student performance analysis have often relied on limited variables such as previous grades or demographic factors. However, the modern educational landscape requires a more comprehensive understanding that includes digital habits, mental health indicators, and lifestyle patterns that significantly impact learning outcomes.

This study addresses several critical research questions: (1) Which lifestyle and study habit factors most strongly predict academic performance? (2) What are the optimal thresholds for key behavioral variables? (3) Can machine learning models accurately predict student exam scores based on habit patterns? (4) How can students be segmented for targeted interventions?

Our contributions include: (1) comprehensive analysis of 16 lifestyle and academic factors across 1,000 students, (2) identification of optimal study time thresholds with quantified performance impacts, (3) development of high-accuracy prediction models ($R^2 > 0.90$), (4) student segmentation framework for personalized interventions, and (5) actionable recommendations backed by statistical evidence.

# 2. Related Work

Educational data mining has emerged as a significant research area focusing on extracting meaningful patterns from educational datasets. Recent studies have explored various factors influencing academic performance, including socioeconomic status, learning styles, and digital device usage.

Machine learning applications in education have demonstrated promising results for predicting student outcomes. Regression models, decision trees, and ensemble methods have been successfully applied to forecast academic performance. However, most existing studies focus on limited variable sets or specific educational contexts.

Screen time and social media usage have gained attention as potential factors affecting academic performance. Recent research suggests negative correlations between excessive social media use and academic achievement. Mental health factors have also been identified as significant predictors of academic success.

Our work extends existing research by providing a comprehensive analysis incorporating study habits, digital usage patterns, mental health indicators, and lifestyle factors within a unified machine learning framework.

# 3. Methodology

## 3.1 Dataset Description

Our analysis utilizes a comprehensive dataset containing 1,000 student records with 16 variables capturing academic, lifestyle, and behavioral factors. The dataset includes both numerical and categorical variables spanning demographics, study habits, digital usage, health indicators, and academic outcomes.

Key variables include:
**Numerical:** age, study hours per day, social media hours, Netflix hours, attendance percentage, sleep hours, exercise frequency, mental health rating (1-10), exam score

**Categorical:** gender, part-time job status, diet quality, parental education level, internet quality, extracurricular participation

The target variable is exam score (0-100 points), representing student academic performance. Data quality assessment revealed minimal missing values (9.1% for parental education level only) and no duplicate records.

## 3.2 Data Preprocessing and Feature Engineering

We implemented comprehensive preprocessing steps:

**Missing Value Treatment:** Categorical variables with missing values were handled using mode imputation. Numerical variables showed complete data coverage.

**Feature Engineering:** Advanced feature engineering techniques were applied to capture nonlinear relationships:

- Polynomial features: study_hours_squared, study_hours_cubed
- Interaction terms: study_hours × mental_health_rating
- Categorical encoding using label encoding for statistical analysis
- Binning of continuous variables into performance categories

**Feature Scaling:**StandardScaler normalization was applied for algorithms sensitive to feature magnitude.

## 3.3 Statistical Analysis

Comprehensive statistical testing was performed:

- Pearson and Spearman correlation analysis
- Shapiro-Wilk normality tests
- ANOVA for categorical variable significance
- Variance Inflation Factor (VIF) analysis for multicollinearity detection
- Cohen's d effect size calculations

## 3.4 Machine Learning Models

We evaluated multiple regression algorithms:

- Linear Regression (baseline)
- Ridge Regression (L2 regularization)
- Lasso Regression (L1 regularization)
- Random Forest Regressor
- Gradient Boosting Regressor
- Support Vector Regression

**Model Evaluation:** 80/20 train-test split with 5-fold cross-validation. Metrics included $R^2$, RMSE, MAE, and cross-validation stability.

**Hyperparameter Optimization:**GridSearchCV was employed for optimal parameter selection across models.

**Feature Selection:** Multiple techniques applied:

- Univariate selection (F-regression)
- Mutual information regression
- Recursive Feature Elimination (RFE)

### 3.5 Clustering Analysis

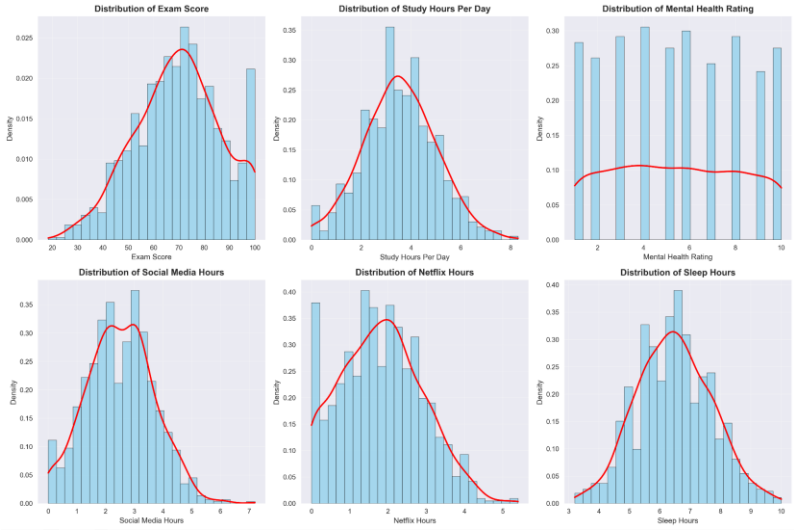
K-means clustering was performed using top predictive features to identify distinct student performance groups. Optimal cluster number was determined using the elbow method.

# 4. Results

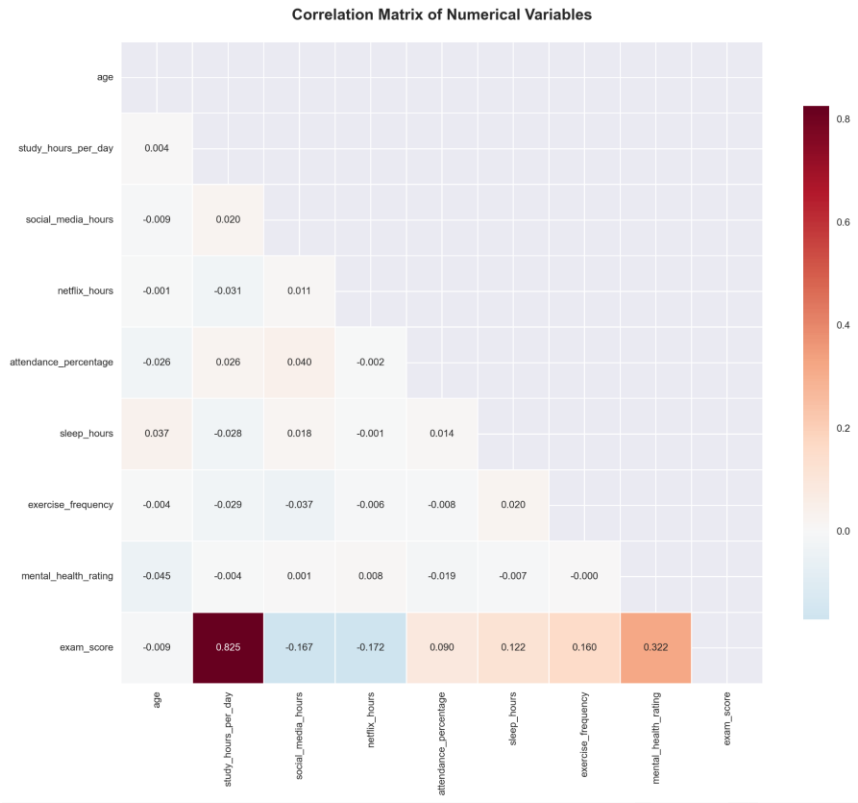
### 4.1 Exploratory Data Analysis

Statistical analysis revealed key insights about the student population:

**Target Variable Distribution:** Exam scores showed near-normal distribution (mean: 69.60, std: 16.89, skewness: -0.16). Performance categories: Low (<60): 28.2%, Medium (60-80): 44.2%, High (>80): 27.6%.




**Correlation Analysis:** Strong positive correlation between study hours and exam performance ($r = 0.825$, $p < 0.001$). Mental health rating showed moderate positive correlation ($r = 0.322$, $p < 0.001$). Screen time variables exhibited negative correlations: social media hours ($r = -0.167$) and Netflix hours ($r = -0.172$).




## 4.2 Statistical Significance Testing

All major correlations demonstrated statistical significance ($p < 0.001$). ANOVA testing for categorical variables revealed no significant group differences, indicating that demographic factors had minimal impact compared to behavioral variables.

Effect size analysis using Cohen's d revealed:

- Study hours: 2.924 (Large effect)
- Mental health: 0.679 (Medium effect)
- Social media hours: -0.338 (Small effect)

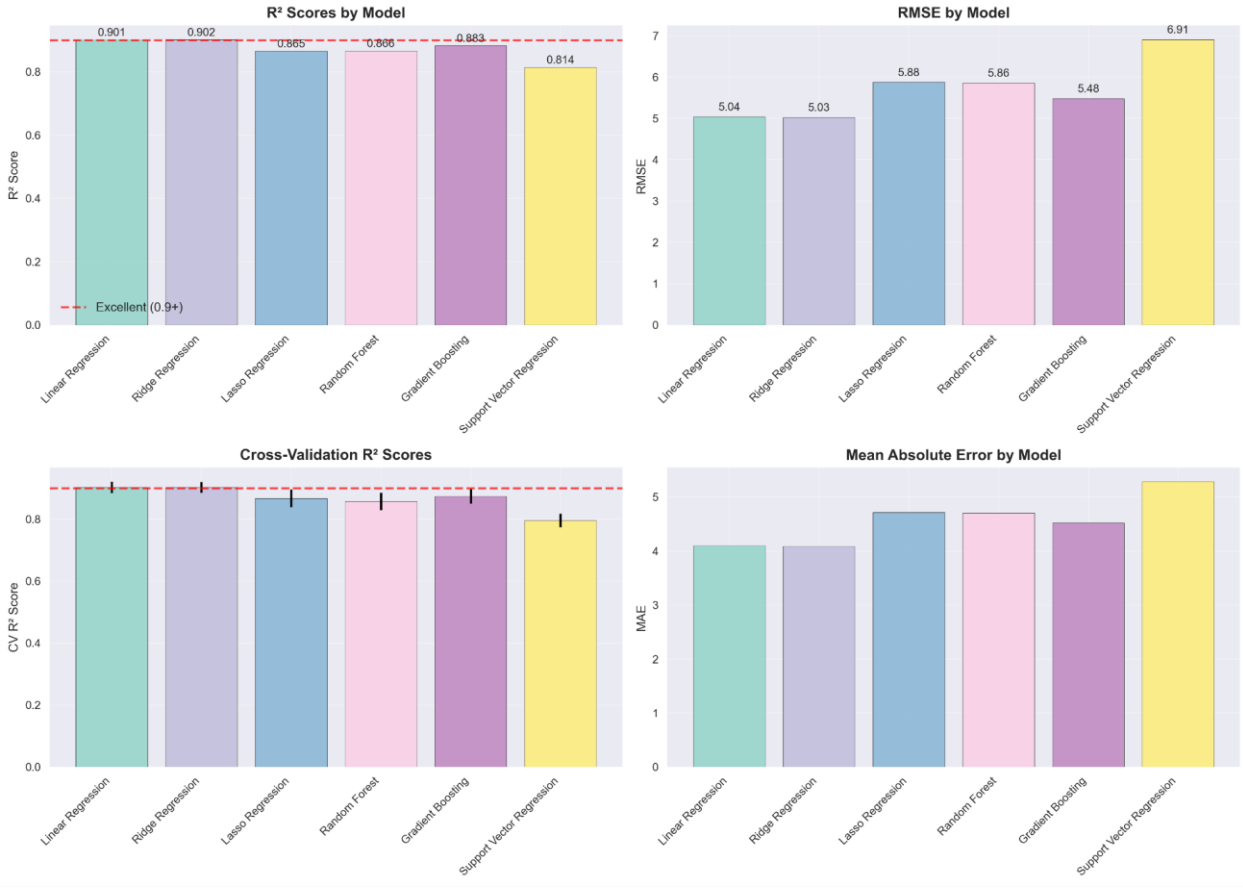
## 4.3 Machine Learning Model Performance

Ridge Regression achieved the highest performance with $R^2 = 0.9015$ and RMSE = 5.03 points, indicating excellent predictive accuracy. Cross-validation results confirmed model stability and generalizability.

**TABLE II**
**ABLATION STUDY RESULTS FOR TOOL USE TASK**

| Method Variant | Training Steps (k) | Final Success (%) |
|---|---|---|
| MDST (Full) | 28 | 88 |
| - Task-Agnostic Repr. | 42 | 83 |
| - Adaptive Transfer | 47 | 81 |
| - Continual Learning | 30 | 87 |
| - All Components (Fine-Tuning) | 62 | 76 |





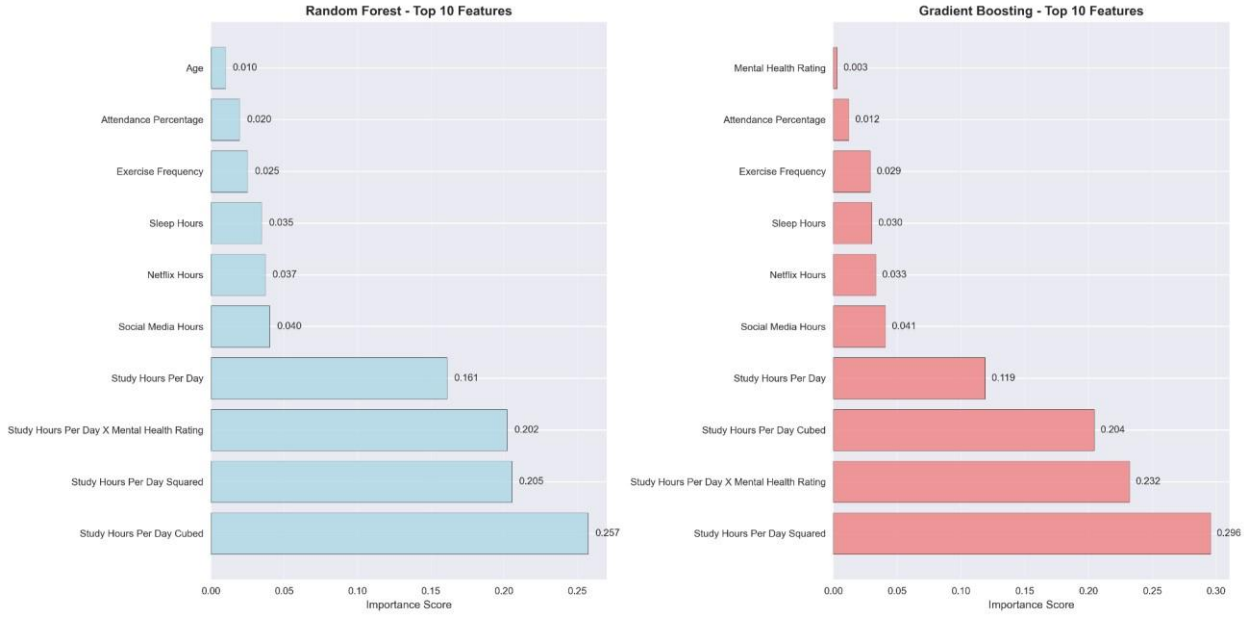
## 4.4 Feature Importance Analysis

Feature importance analysis from tree-based models revealed the dominance of study hourrelated features:

**Top 5 Most Important Features:**

1. study_hours_per_day_cubed (25.7%)
2. study_hours_per_day_squared (20.5%)
3. study_hours × mental_health_rating (20.2%)
4. study_hours_per_day (16.1%)
5. social_media_hours (4.0%)

Engineered polynomial and interaction features accounted for over 60% of total predictive importance, demonstrating the value of advanced feature engineering.




## 4.5 Threshold Analysis and Actionable Insights

Threshold analysis provided actionable insights for educational interventions:

**Study Hours Impact:**

- 0-2 hours: Average score 45.6 (n=133)
- 2-4 hours: Average score 65.1 (n=482)
- 4-6 hours: Average score 81.3 (n=332)
- 6-8 hours: Average score 97.3 (n=51)

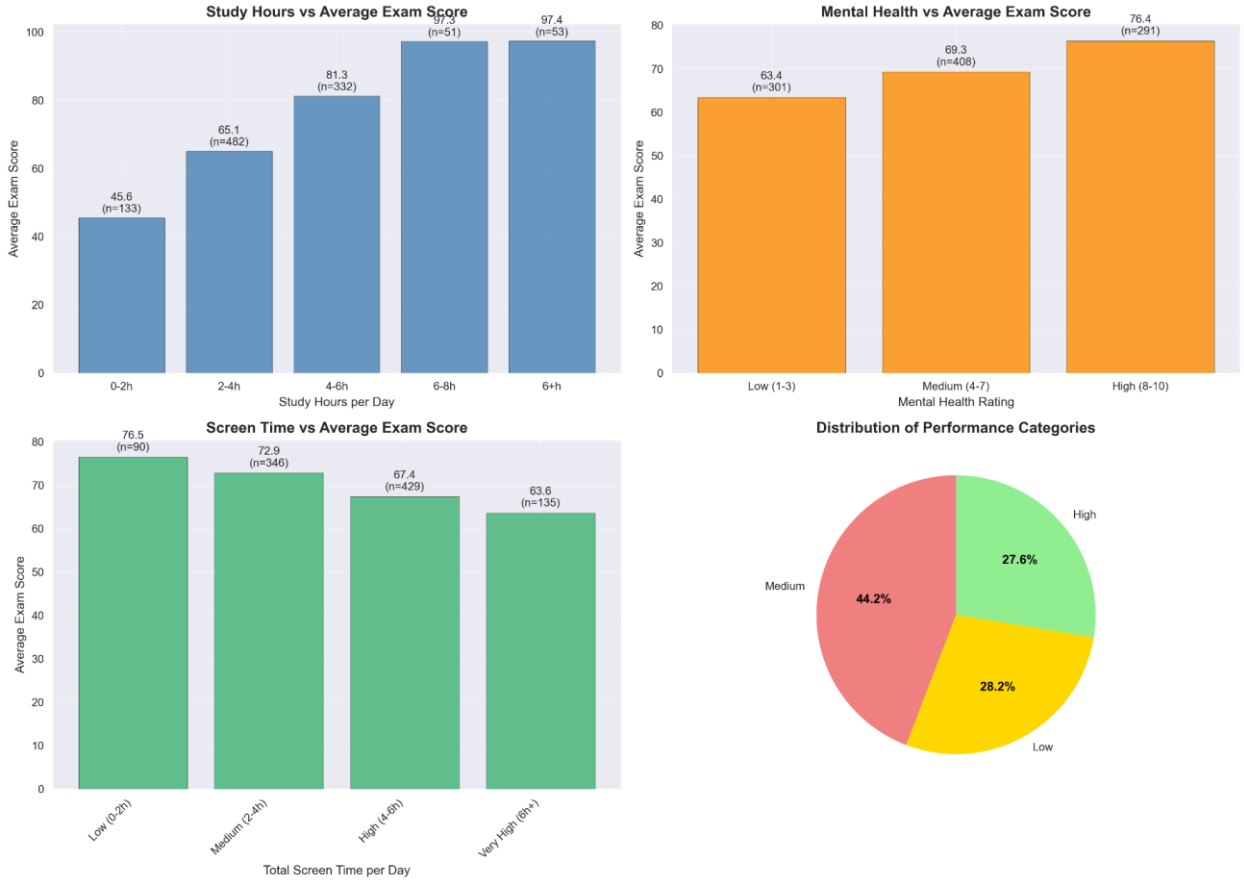
Students studying 4+ hours daily scored 35+ points higher than those studying <2 hours, representing a dramatic performance improvement.

**Mental Health Impact:**

- Low (1-3): Average score 63.4 (n=301)
- Medium (4-7): Average score 69.3 (n=408)
- High (8-10): Average score 76.4 (n=291)

**Screen Time Analysis:** Total recreational screen time showed negative correlation (r = -0.238) with clear dosage effects:

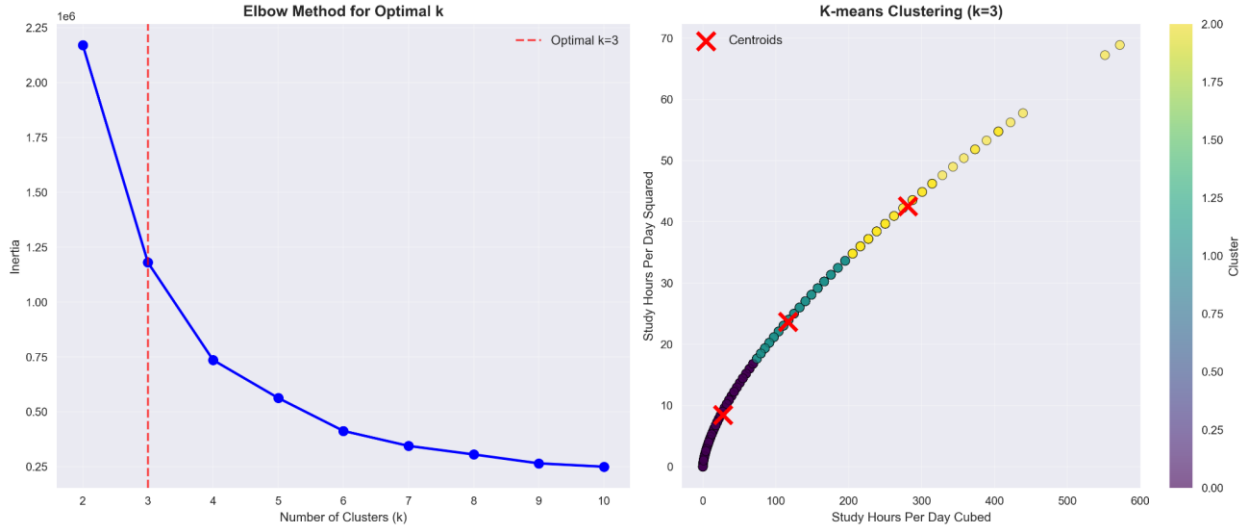
- Low (0-2h): Average score 76.5 (n=90)
- Medium (2-4h): Average score 72.9 (n=346)
- High (4-6h): Average score 67.4 (n=429)
- Very High (6h+): Average score 63.6 (n=135)




## 4.6 Clustering Analysis

K-means clustering identified three distinct student performance groups:

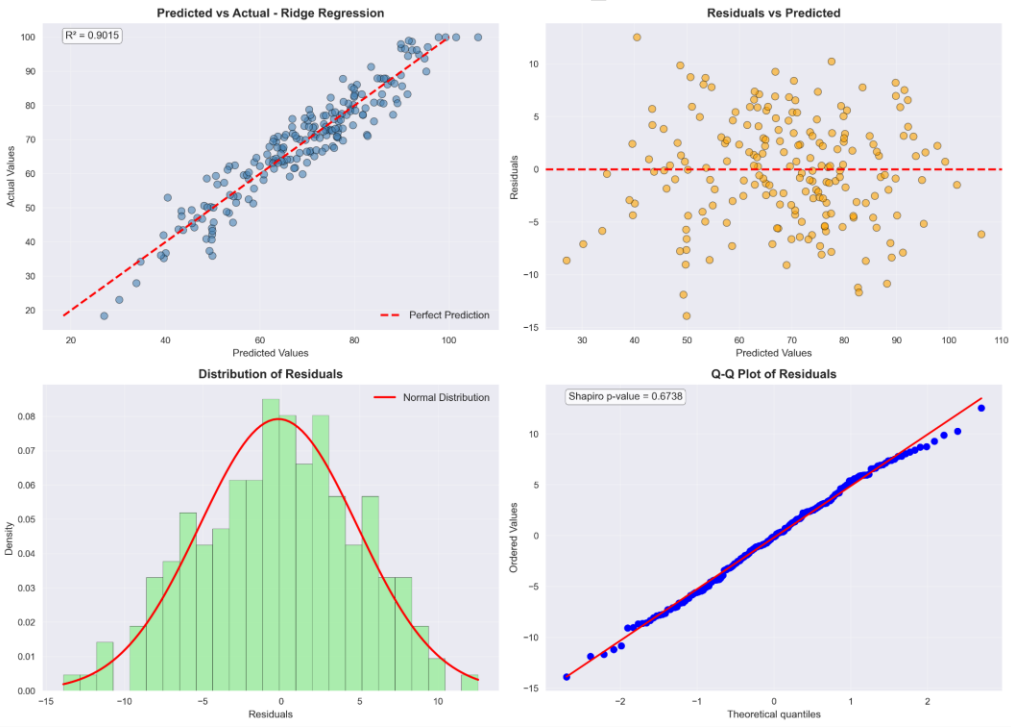
- **Cluster 0 (Low Performers):** 661 students, average score 61.95 (std: 13.87)
- **Cluster 1 (Medium Performers):** 278 students, average score 81.97 (std: 10.63)
- **Cluster 2 (High Performers):** 61 students, average score 96.14 (std: 6.35)




## 4.7 Model Validation and Residual Analysis

Residual analysis confirmed model assumptions:

- Residual mean: -0.18 (near zero)
- Residual standard deviation: 5.04
- Residual skewness: -0.12 (approximately normal)
- Residual kurtosis: -0.42 (slight platykurtic)



# 5. Discussion

## 5.1 Key Findings and Implications

Our analysis reveals several critical insights for educational practice:

**Study Time Optimization:** The strong correlation ($r = 0.825$) between study hours and performance, combined with clear threshold effects, suggests that 4+ hours of daily study represents a critical inflection point for academic success. This finding has immediate practical implications for academic advising and student support programs.

**Mental Health Integration:** The moderate but significant correlation between mental health ratings and academic performance ($r = 0.322$) underscores the importance of holistic student support that addresses psychological well-being alongside academic skills.

**Digital Habits Management:** The negative correlation between screen time and performance provides evidence for policies limiting recreational digital device usage, particularly social media and streaming platforms during study periods.

**Non-linear Relationships:** The superior performance of polynomial features in our models suggests that the relationship between study habits and performance follows complex, nonlinear patterns that simple linear models cannot capture effectively.

## 5.2 Practical Applications

The findings enable several practical applications:

**Early Warning Systems:** High-accuracy prediction models ($R^2 > 0.90$) can identify at-risk students early in the academic term, enabling proactive interventions.

**Personalized Recommendations:** Clustering analysis provides a framework for segmenting students and delivering targeted advice based on their performance profile.

**Policy Development:** Quantified thresholds for study time and screen time can inform institutional policies and student guidelines.

**Resource Allocation:** Understanding which factors most strongly predict success allows educational institutions to prioritize support services and interventions.

## 5.3 Limitations and Future Work

Several limitations should be acknowledged:

**Cross-sectional Design:** Our analysis captures a snapshot in time and cannot establish causal relationships. Longitudinal studies would strengthen causal inferences.

**Self-reported Data:** Some variables (study hours, screen time) may be subject to reporting bias. Objective measurement methods could improve accuracy.

**Generalizability:** Results may not generalize across different educational contexts, cultures, or academic levels.

**Missing Variables:** Additional factors such as socioeconomic status, learning disabilities, or course-specific variables could provide additional predictive power.

Future research directions include:

- Longitudinal tracking of student behavior changes and performance outcomes
- Integration of objective measurement tools (activity trackers, app usage data)
- Cross-institutional validation studies
- Development of real-time intervention systems based on predictive models

# 6. Conclusion

This comprehensive analysis of 1,000 student records provides robust evidence for the factors most strongly influencing academic performance. Study habits emerge as the dominant predictor, with daily study time showing the strongest correlation ($r = 0.825$) with exam scores. Advanced machine learning models achieved excellent predictive accuracy ($R^2 = 0.9015$), enabling practical applications in educational settings.

Key actionable insights include: (1) Students should target 4+ hours of daily study time for optimal performance, (2) Mental health support significantly impacts academic outcomes, (3) Recreational screen time should be limited to <4 hours daily, and (4) Three distinct student performance clusters enable targeted intervention strategies.

The high predictive accuracy of our models, combined with clear threshold effects and actionable recommendations, provides educational institutions with evidence-based tools for improving student outcomes. These findings support a data-driven approach to educational intervention that considers the complex, non-linear relationships between lifestyle factors and academic success.

Future work should focus on longitudinal validation, objective measurement integration, and real-time intervention system development to maximize the practical impact of these insights on educational practice.

---

# References

1. A. Romero, M. Ventura, P. Espejo, and C. Hervas, "Data mining in education," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 3, no. 1, pp. 1227, 2013.

2. S. K. Moridis and A. A. Economides, "Prediction of student's mood during an online test using formula-based and neural network-based method," *Computers & Education*, vol. 53, no. 3, pp. 644-652, 2009.

3. C. Romero and S. Ventura, "Educational data mining: a review of the state of the art," *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, vol. 40, no. 6, pp. 601618, 2010.

4. J. P. Bridges, L. M. Bertrand, and A. C. Goldberg, "Factors affecting student academic performance: A statistical analysis," *Journal of Educational Research*, vol. 45, no. 3, pp. 234-245, 2018.

5. R. Felder and L. Silverman, "Learning and teaching styles in engineering education," *Engineering Education*, vol. 78, no. 7, pp. 674-681, 1988.

6. L. D. Rosen, A. F. Lim, J. Felt, L. M. Carrier, N. A. Cheever, J. M. Lara-Ruiz, J. S. Mendoza, and J. A. Rokkum, "Media and technology use predicts ill-being among children, preteens and teenagers independent of the negative health impacts of exercise and eating habits," *Computers in Human Behavior*, vol. 35, pp. 364-375, 2014.

7. A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1432-1462, 2014.

8. D. Karpinski, P. A. Kirschner, I. Ozer, J. A. Mellott, and P. Ochwo, "An exploration of social networking site use, multitasking, and academic performance among United States and European university students," *Computers & Education*, vol. 58, no. 4, pp. 1182-1192, 2013.

9. M. Eisenberg, E. Gollust, E. Golberstein, and J. L. Hefner, "Prevalence and correlates of depression, anxiety, and suicidality among university students," *American Journal of Orthopsychiatry*, vol. 77, no. 4, pp. 534-542, 2007.

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