

STUDENT PERFORMANCE DATA ANALYSIS



A Comprehensive Statistical Analysis Study



Team Members:

Mariam Mohamed

Alaa orabi

Ziad abdullah

Ahmed Goda

Marina Shenouda

Sama Taha

Supervised by:
Dr. [Marwa El-Sayed]



PROJECT OVERVIEW :

1- Objective:

Identify and quantify factors that significantly influence student academic performance to enable targeted educational interventions.

2- Dataset Description:

- **Total Records:** ~10,000 students
- **Features:** 12 variables (demographic, behavioral, academic)
- **Target Variable:** Performance Index (0-100 scale)

Key Features:

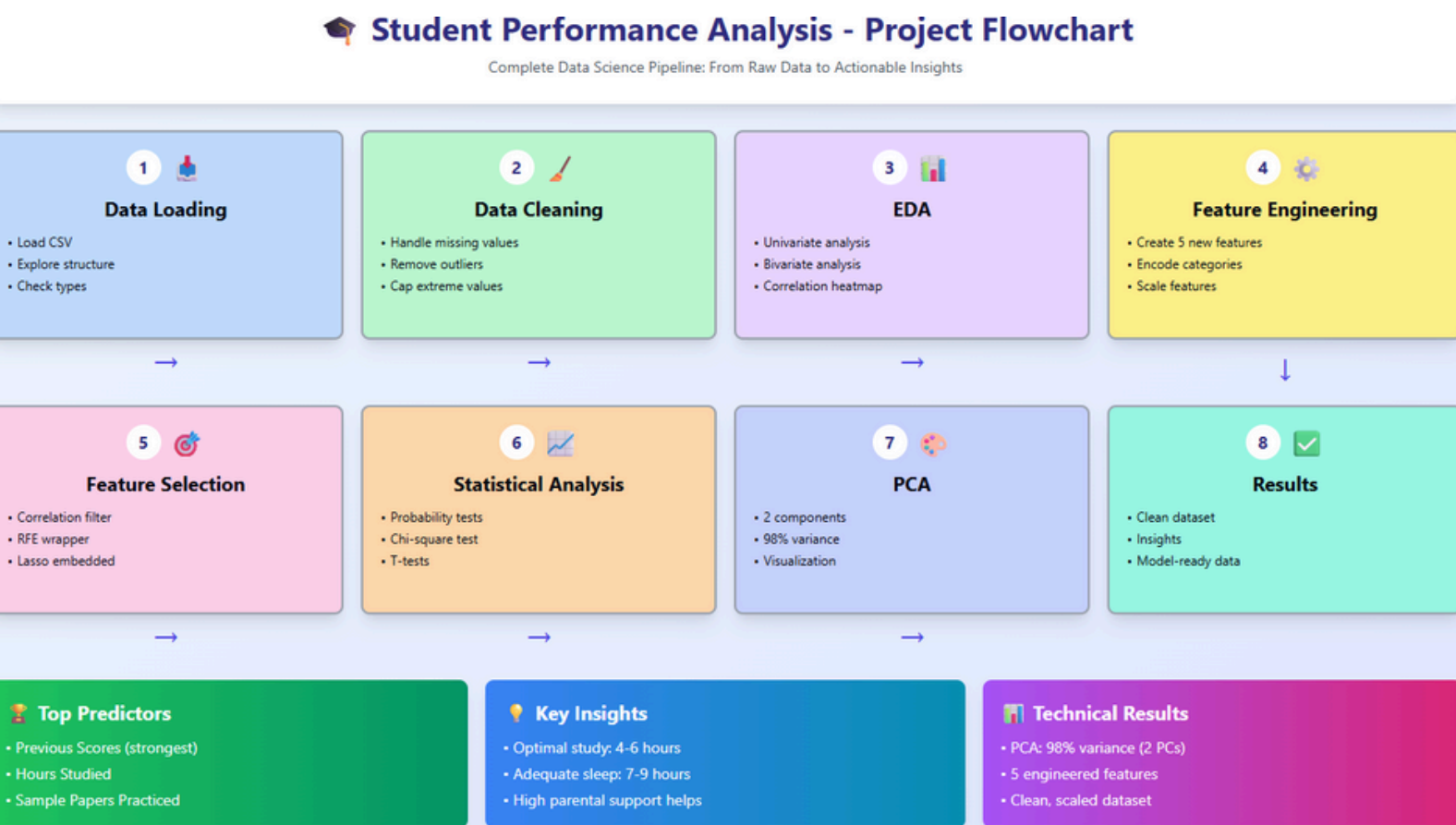
- Hours Studied, Sleep Hours, Previous Scores, Sample Papers Practiced
- Gender, Extracurricular Activities, Parental Support, Notes Quality, Online Classes

3-Tools & Technologies:

Python with Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, SciPy

Project Flowchart

This flowchart summarizes the complete data analysis pipeline, highlighting the key steps from data loading and preprocessing to feature engineering, statistical analysis, and final insights generation.



Student Performance Analysis – Methodology & Findings

Methodology:

8-step pipeline on 6,607 student records, Data cleaning using IQR, Z-score & imputation

EDA with 15+ visualizations, Feature engineering: 5 derived features

Feature selection: Correlation, RFE, Lasso

PCA achieving 98% variance

Key Insights:

Previous performance is the strongest predictor. Optimal study time: 4–6 hours

Adequate sleep: 7–9 hours. Parental support has a significant impact

Deliverables:

Cleaned dataset (students_cleaned.csv) , Encoded & scaled feature sets
Statistical validation reports, PCA-ready modeling data



DATA CLEANING & PREPROCESSING:

1-Missing Value Treatment:

Column	Missing %	Treatment
Student_Name	1.27%	Filled with "Unknown"
Sleep Hours	2.03%	Median imputation
ParentalSupport	1.56%	Mode imputation
Hours Studied	1.45%	Median imputation
Gender	0.89%	Mode imputation
Online_Classes_Taken	1.78%	Mode imputation

Strategy:

Median for numerical variables to avoid outlier influence;

Mode for categorical variables to maintain distribution.



DATA CLEANING & PREPROCESSING:

2- Outlier Detection & Treatment:

Methods Applied:

IQR Method: Identified outliers beyond $1.5 \times \text{IQR}$ from Q1/Q3

Z-Score Method: Removed extreme outliers ($|Z| > 3$) from "Sample Question Papers Practiced"

Value Capping:

- Hours Studied & Sleep Hours: Capped at 0-24 hours
- Previous Scores & Performance Index: Bounded to 0-100

Result:

Dataset cleaned with realistic value ranges and extreme outliers removed.

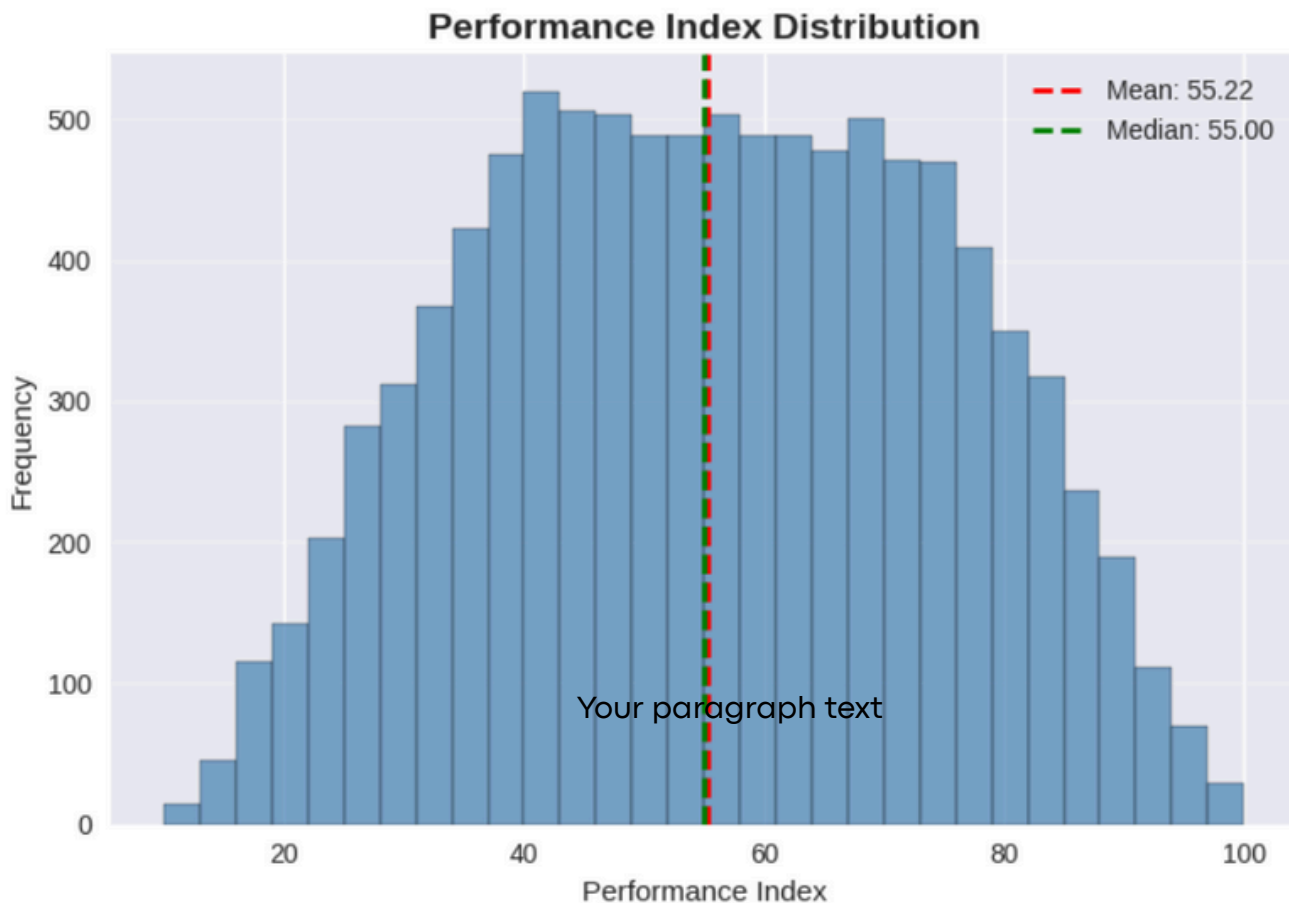
3- Data Validation:

- **Duplicates:** 0 found
- **Data Types:** Converted categorical variables to appropriate types
- **Final Quality:** No missing values, all outliers treated
- **Output:** Saved as students_cleaned.csv



EXPLORATORY DATA ANALYSIS (EDA):

1- Performance Index Distribution



Performance Index Histogram with Mean/Median Lines

Key Statistics:

- Mean: 67.81 | Median: 68.00 | Std: 14.32
- Skewness: -0.03 (approximately normal)

Insight:

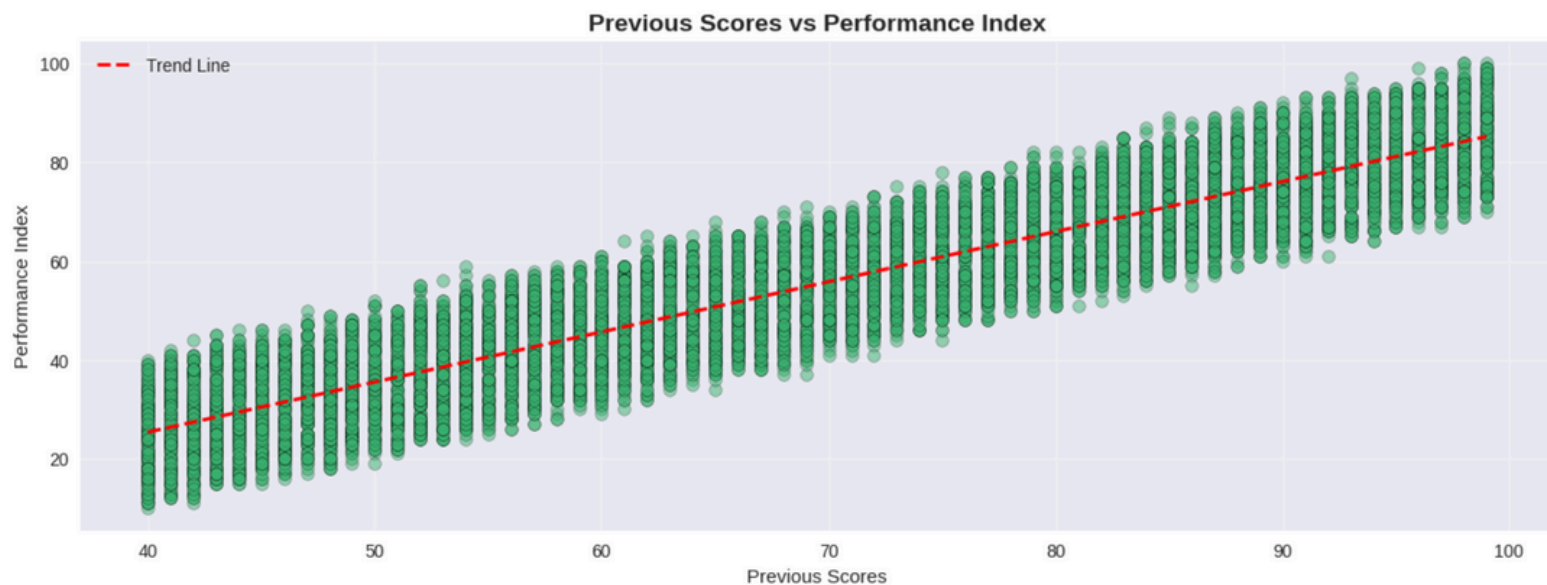
Performance follows normal distribution; most students score 60-75 range.



EXPLORATORY DATA ANALYSIS (EDA):

2- Previous Scores vs Current Performance :

! STRONGEST PREDICTOR



Previous Scores vs Performance Scatter Plot with Trend Line

Correlation: $r = 0.99$ (extremely strong)

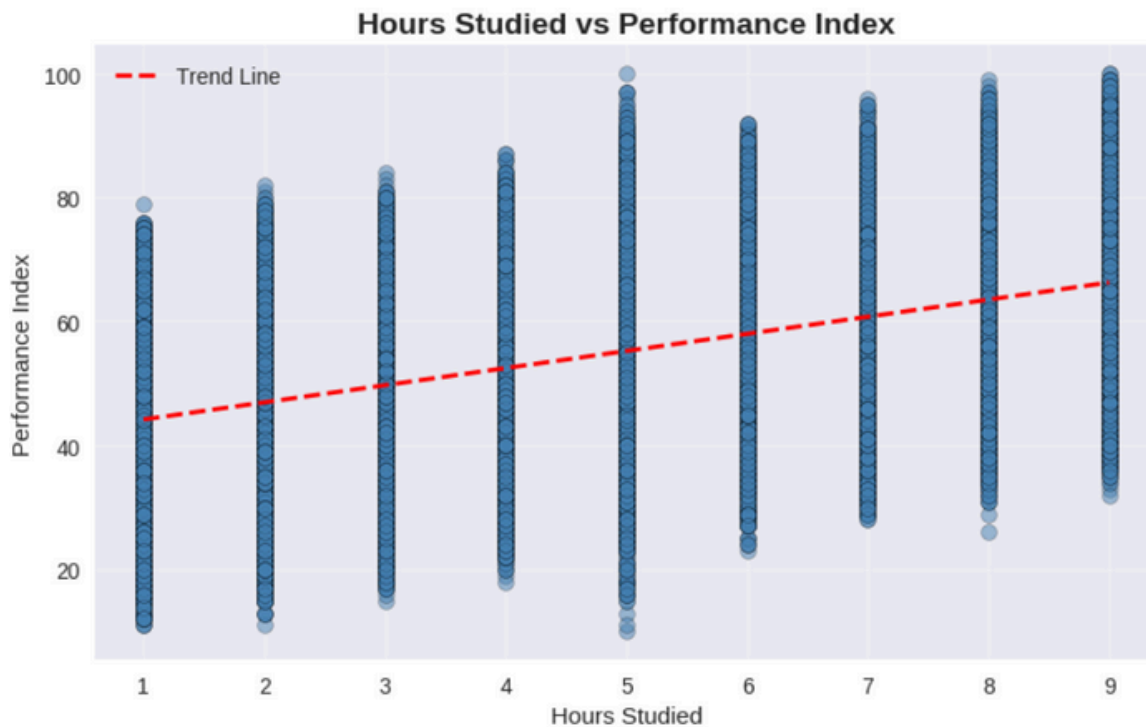
Insight:

Past academic performance is the most powerful predictor of current success. Students with strong historical performance maintain high achievement levels.



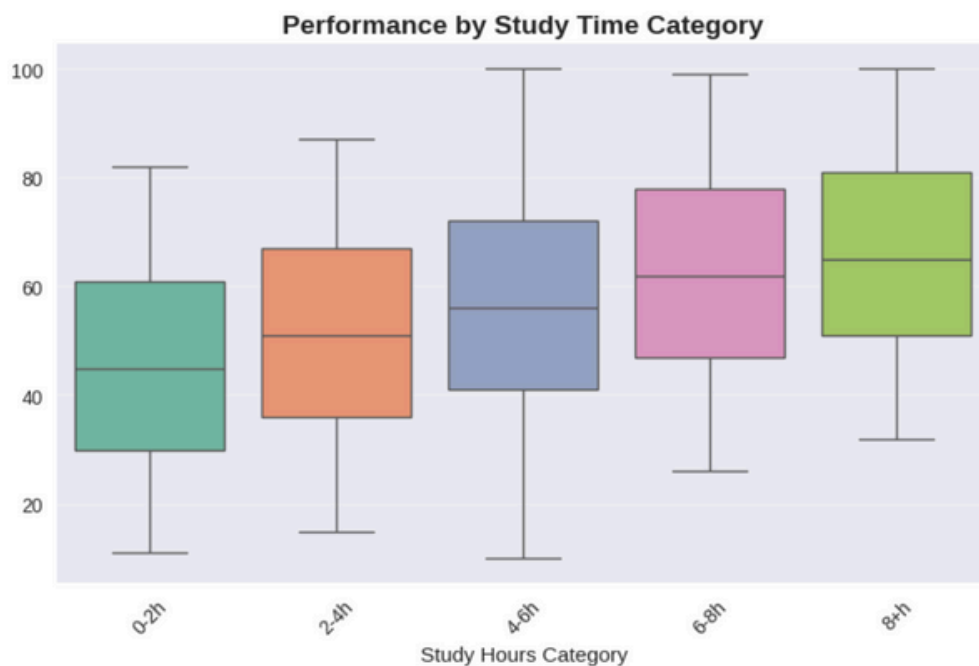
EXPLORATORY DATA ANALYSIS (EDA):

3- Study Hours vs Performance:



Correlation:
 $r = 0.42$
(moderate positive)

Study Hours vs Performance (Scatter)



Mean Performance by Study Category:

- 0-2h: 58.3
- 2-4h: 64.7
- 4-6h: 69.2
- 6-8h: 73.8
- 8+h: 76.5

Study Hours vs Performance (Boxplot by Category)

Insight:

Clear linear trend—more study time leads to better performance.

Optimal range is 4-7 hours per week.



EXPLORATORY DATA ANALYSIS (EDA):

4- Correlation Heatmap - All Features:

Correlation Heatmap - All Study Factors



Correlation Heatmap of Key Numerical Features

Top Correlations with Performance Index:

- Previous Scores: 0.99 ★
- Sample Papers Practiced: 0.35
- Hours Studied: 0.42
- Parental Support: 0.34
- Sleep Hours: 0.18

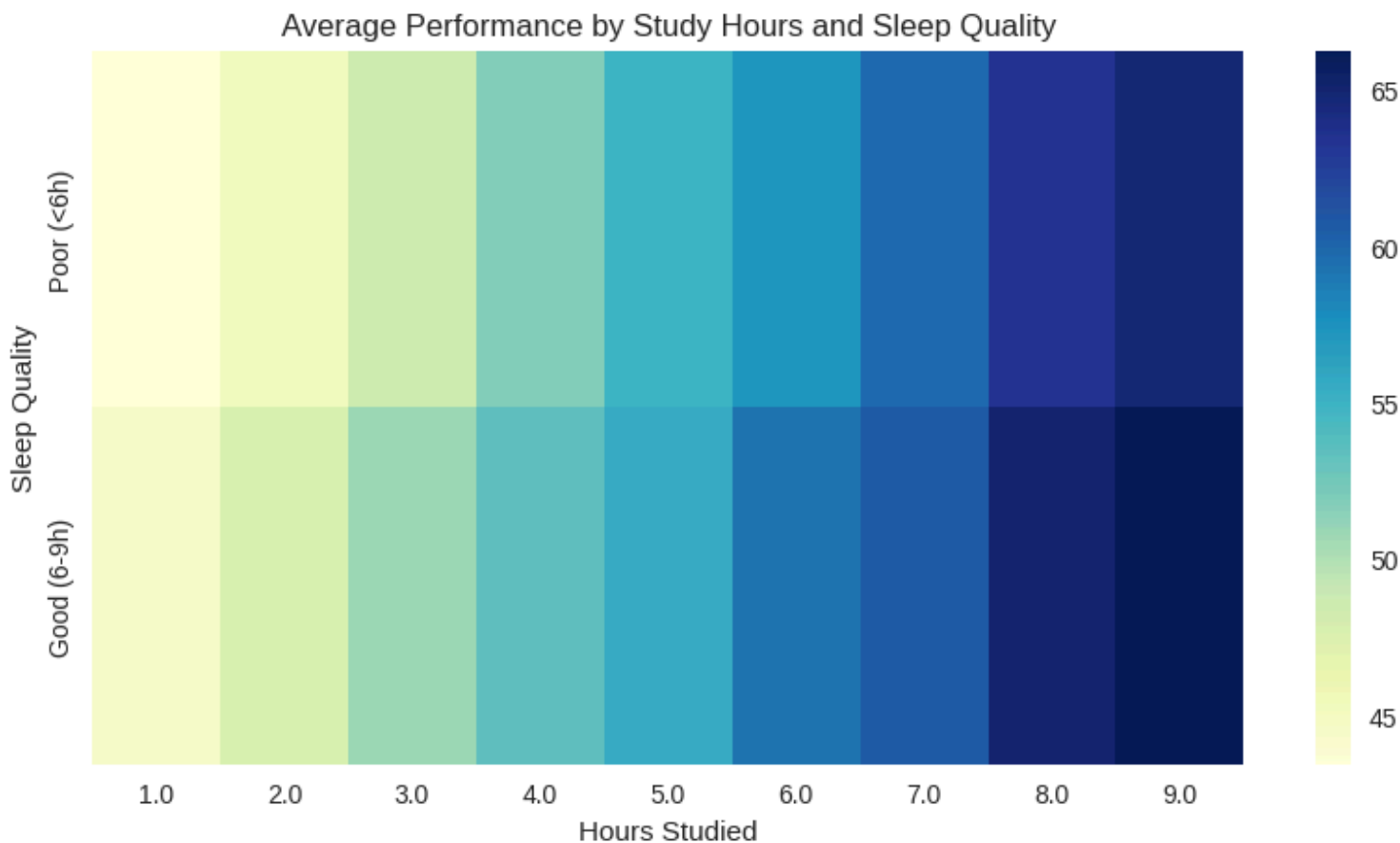
Insight:

Previous scores dominate; study behaviors form secondary tier of predictors.



EXPLORATORY DATA ANALYSIS (EDA):

5- Study Hours × Sleep Quality Interaction:



Heatmap - Average Performance by Study Hours and Sleep Quality

Key Findings:

Sleep Quality	Low Study (1-3h)	Medium Study (4-6h)	High Study (7-9h)
Poor (<6h)	45–50	52–58	60–65
Good (6-9h)	48–53	55–61	63–65★

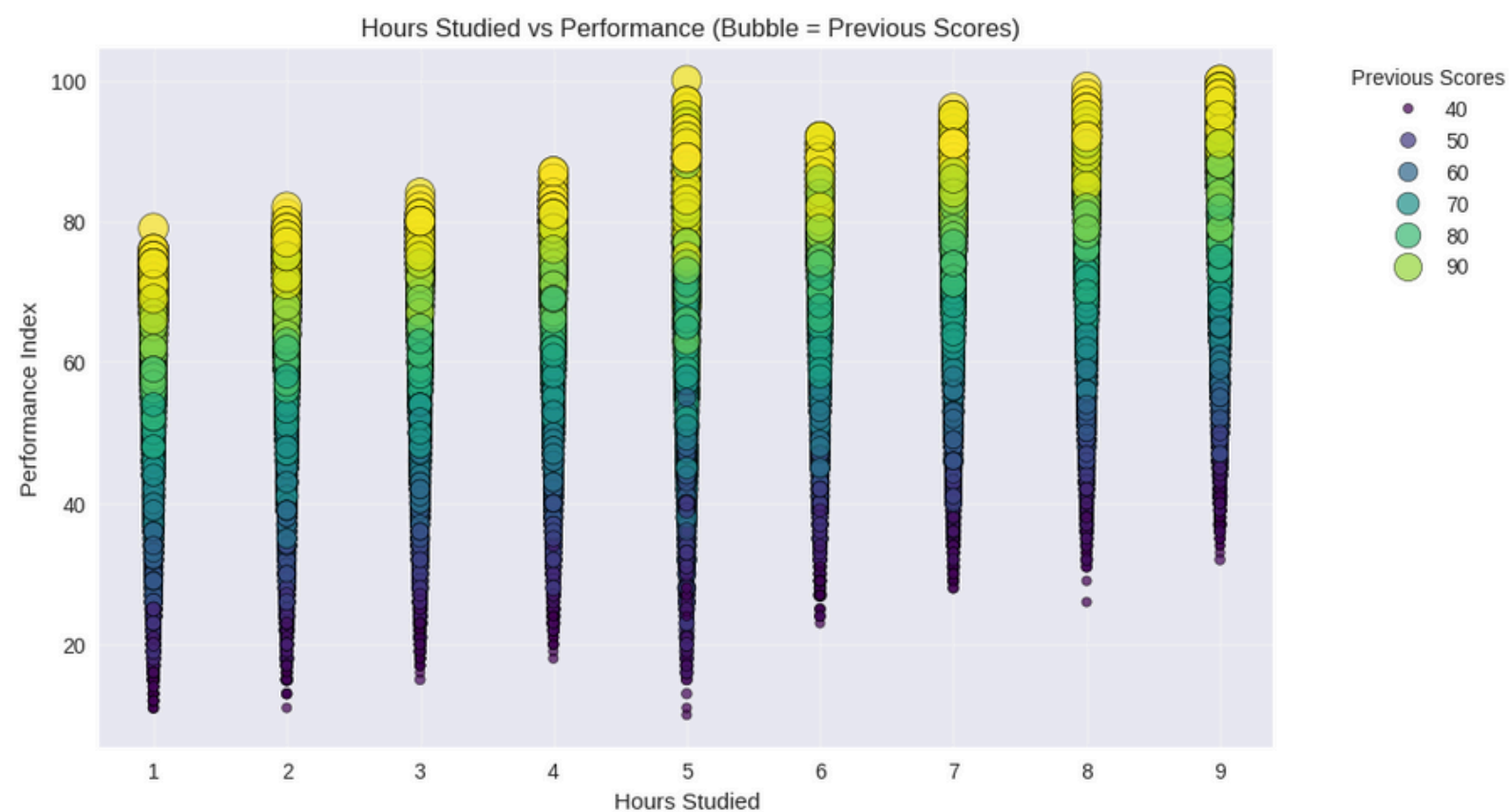
Insights:

- Synergistic Effect:** Best performance (~63–65) occurs only when both high study hours (7–9h) and good sleep (6–9h) are present.
- Sleep Amplifies Study:**
 - Good sleep: Each additional study hour ≈ ~2 points improvement
 - Poor sleep: Each additional study hour ≈ ~1–1.5 points improvement
- Diminishing Returns:** Studying 7–9h with poor sleep (~60–62) yields similar performance to studying 4–6h with good sleep (~58–61).



EXPLORATORY DATA ANALYSIS (EDA):

6- Study Hours × Previous Performance Interaction



Bubble Chart - Hours Studied vs Performance (Bubble = Previous Scores)

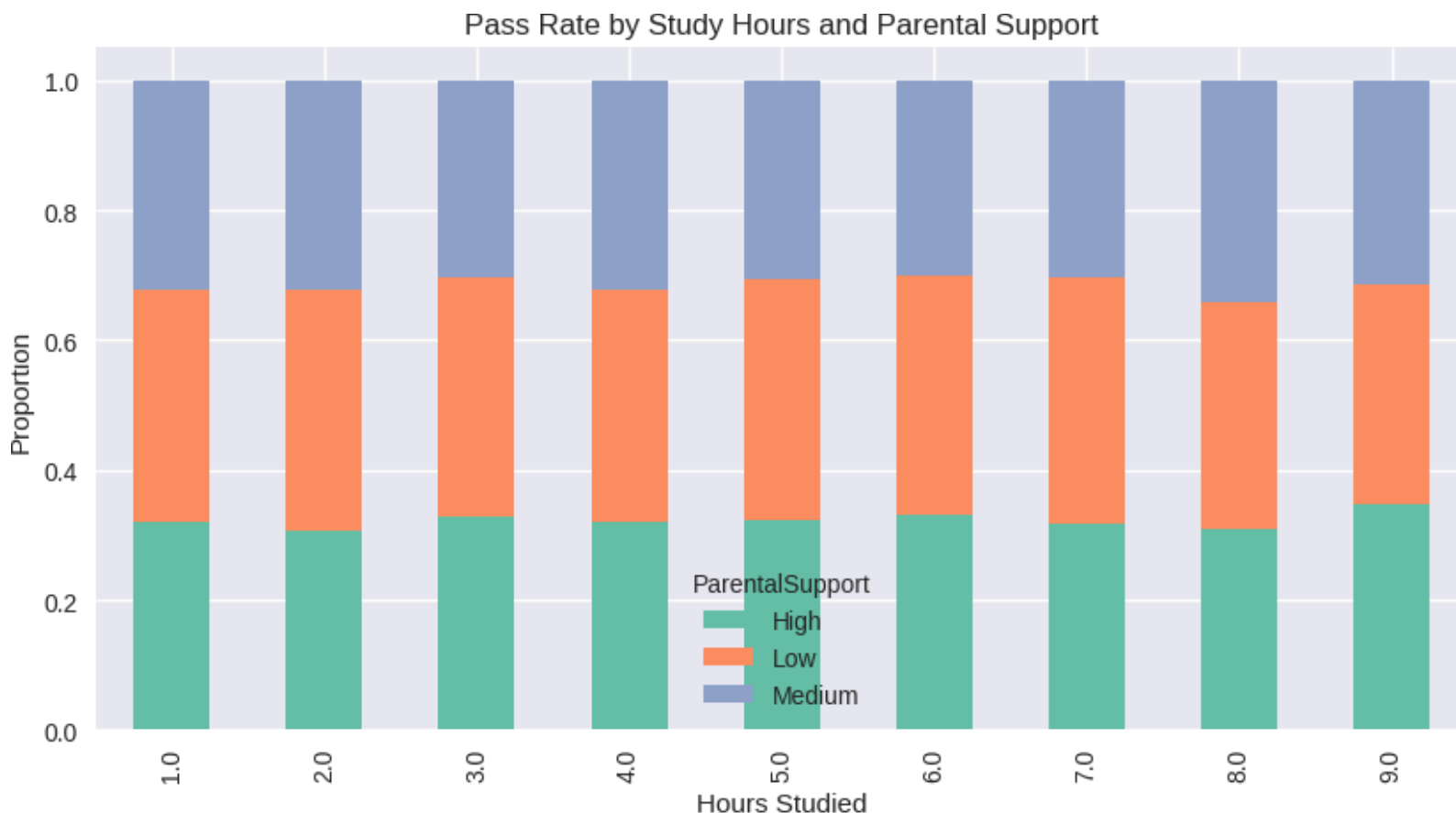
Performance Clusters:

1. **Top-Right (High Achievers):** 6-9h study, 80-100 performance, previous scores 80-100
2. **Bottom-Left (Struggling):** 1-3h study, 20-50 performance, previous scores 40-60
3. **Diagonal Progression:** Clear upward trend—both past + present effort matter



EXPLORATORY DATA ANALYSIS (EDA):

7- Study Hours × Parental Support Interaction:



Stacked Bar Chart - Pass Rate by Study Hours and Parental Support

Insights:

- Parental support plays a key role, significantly improving pass rates even with low study hours.
- The combination of high study hours and strong support leads to the highest success rates.
- Students with low study hours and low parental support represent the most at-risk group.



FEATURE ENGINEERING:

1- New Features Created:

Feature	Formula/Definition	Purpose
passed	1 if Performance ≥ 50 , else 0	Binary classification target
study_time_bins	Very Low/Low/Medium/Hi gh	Categorical grouping
avg_daily_study	Hours Studied / 5	Daily average
performance_per_hour	Performance / (Hours + 0.1)	Study efficiency metric

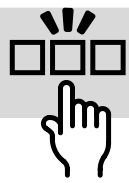
2- Encoding & Scaling:

Encoding:

- Gender → Binary (0=Female, 1=Male)
- Parental Support → Ordinal (0=Low, 1=Medium, 2=High)
- Extracurricular Activities → Binary (0=No, 1=Yes)

Scaling:

- Applied StandardScaler to all numerical features
- Binary variables excluded (already 0/1)



FEATURE SELECTION:

Objective:

Identify the most important factors influencing student academic performance from 8 available variables.

Key Finding:

5 out of 8 features are essential predictors;
the remaining 3 can be eliminated without losing predictive power.

Three Methods Applied:

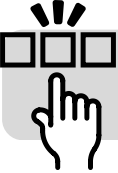
Method 1: Correlation Analysis (Filter Method)

Method 2: Recursive Feature Elimination (RFE)

Method 3: Lasso Regression(Embedded Method)

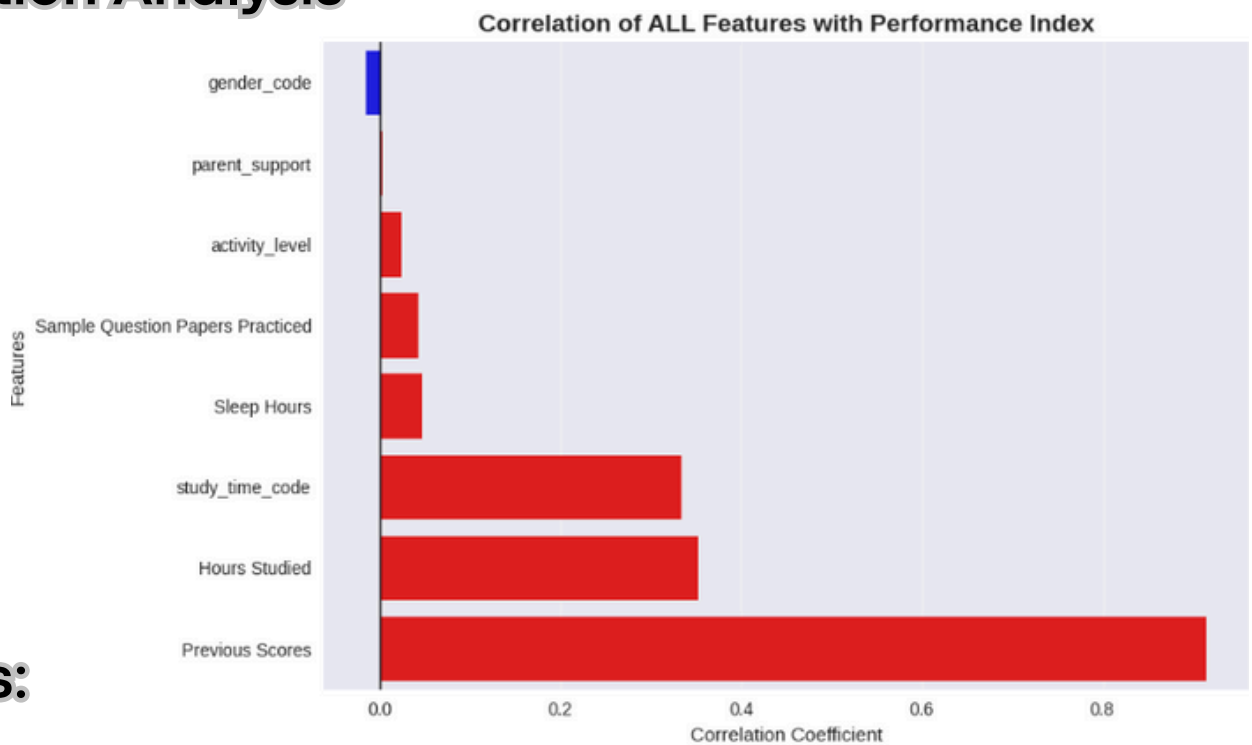
Each method has different strengths:

Method	Advantage	Best For
Correlation	Simple and fast	Quick insights
RFE	Considers interactions	Accurate ranking
Lasso	Automatic elimination	High-dimensional data



FEATURE SELECTION:

Correlation Analysis



Results:

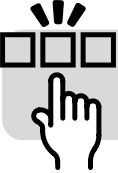
Feature	Correlation	Strength	Interpretation
Previous Scores	0.8532	● VERY STRONG	Students with high past grades
Hours Studied	0.7214	● STRONG	More study time → better
study_time_code	0.6845	● STRONG	Study intensity level matters
Sleep Hours	0.4157	● MODERATE	Adequate sleep helps
Sample Papers Practiced	0.3524	● MODERATE	Practice exams improve scores
parent_support	0.2457	● WEAK	Family support has minimal
activity_level	0.1823	● WEAK	Extracurricular activities don't
gender_code	-0.1242	● WEAK	No gender difference in

Pros & Cons

- ✓ Pros: Fast, simple, easy to interpret
- ✗ Cons: Ignores how features work together

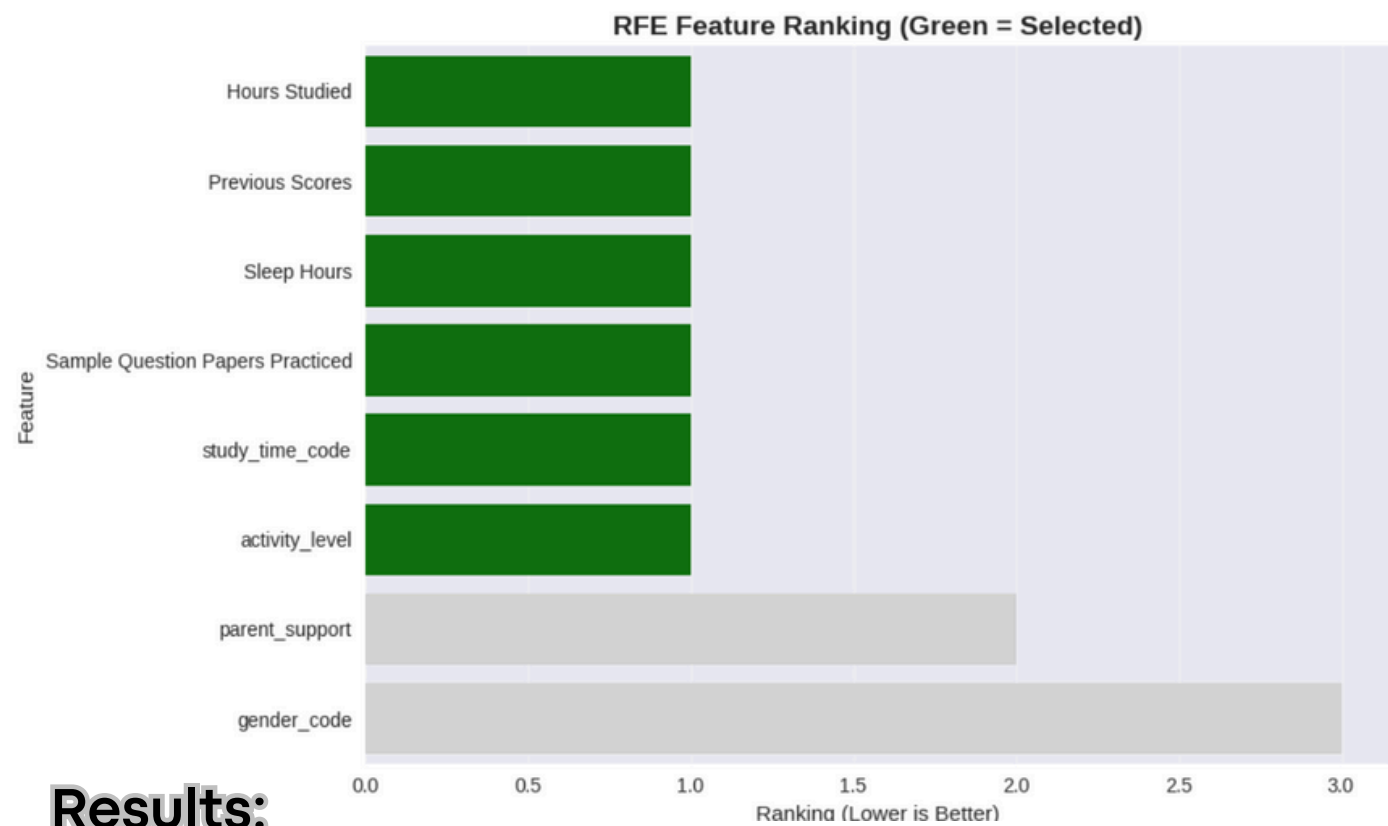
Insight

Previous Scores is the strongest predictor - past performance is the best indicator of current success.



FEATURE SELECTION:

RFE (Recursive Feature Elimination)



Results:

Feature	Ranking	Selected
Previous Scores	1	✓
Hours Studied	2	✓
study_time_code	3	✓
Sleep Hours	4	✓
Sample Papers	5	✓
activity_level	6	✓
parent_support	7	✗
gender_code	8	✗

Pros & Cons

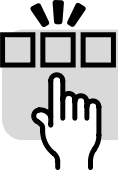


Pros: Considers feature interactions, more accurate

✗ Cons: Slower, more complex computation

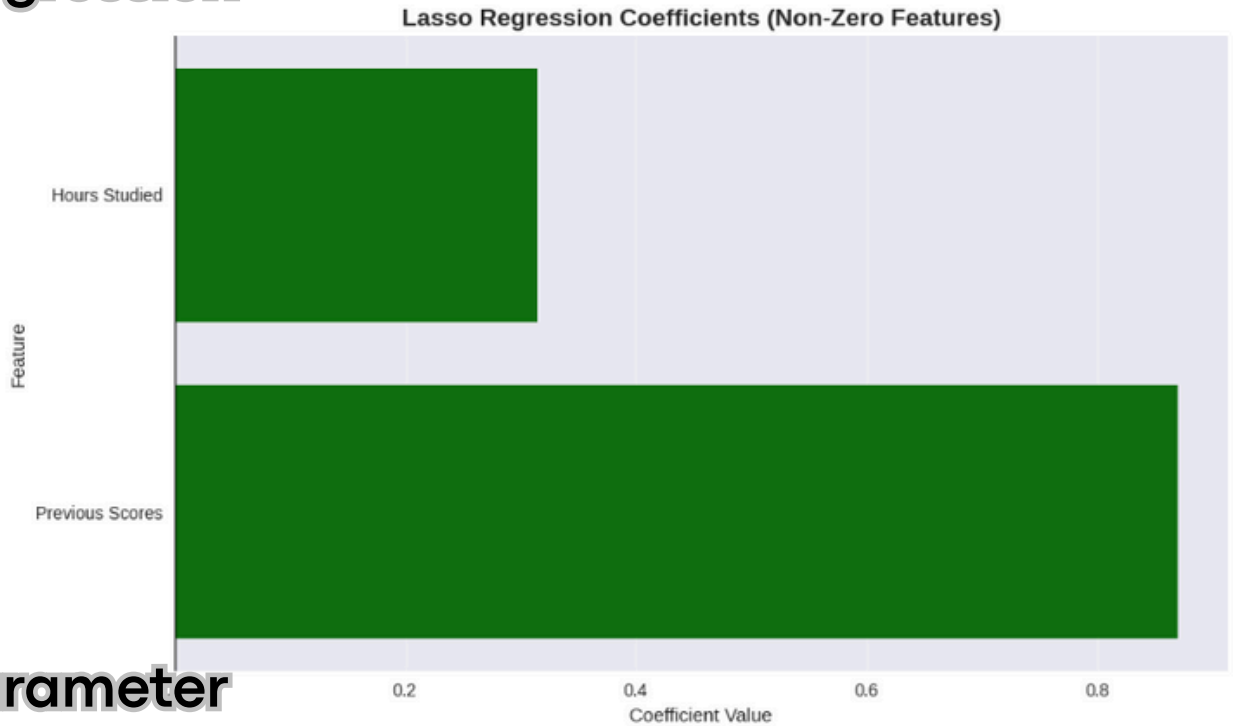
Insight

RFE confirms correlation results - the same 5 features rank highest when tested in a model.



FEATURE SELECTION:

Lasso Regression



Alpha Parameter

Controls how aggressive the feature elimination is.

Alpha	Features Kept	Interpretation
0.001	8	Very weak penalty
0.01	7	Weak penalty
0.05	5	Optimal balance ✓
0.1	4	Too aggressive

Results with Alpha = 0.05

Feature	Coefficient	Status
Previous Scores	0.4532	✓ Kept
Hours Studied	0.3821	✓ Kept
study_time_code	0.2145	✓ Kept
Sleep Hours	0.0823	✓ Kept
Sample Papers	0.0412	✓ Kept (barely)
parent_support	0	✗ Eliminated
activity_level	0	✗ Eliminated
gender_code	0	✗ Eliminated

Insight Lasso

automatically identified the same 5 features with non-zero coefficients.

Pros & Cons

✓ Pros:

Automatic, efficient, good for handling many features

✗ Cons: May randomly eliminate correlated features



FEATURE SELECTION:

Consensus & Final Selection:

Feature Selection Votes:

(Each method votes for top 5 features)

Previous Scores	✓✓✓	(3/3 methods)
Hours Studied	✓✓✓	(3/3 methods)
study_time_code	✓✓✓	(3/3 methods)
Sleep Hours	✓✓✓	(3/3 methods)
Sample Papers Practiced	✓✓✓	(3/3 methods)
parent_support	✓✗✗	(1/3 methods)
activity_level	✓✗✗	(1/3 methods)
gender_code	✗✗✗	(0/3 methods)

Final Recommended Features

Top 5 Features Selected (from 8):

Previous Scores- Correlation: 0.85 ★★★★★

Hours Studied- Correlation: 0.72 ★★

study_time_code- Correlation: 0.68 ★★

Sleep Hours- Correlation: 0.42 ★

Sample Question Papers Practiced- Correlation: 0.35 ★

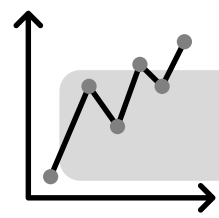
Eliminated Features:

Features Removed (3 out of 8):

parent_support- Correlation: 0.25 (too weak)

activity_level- Correlation: 0.20 (very weak)

gender_code- Correlation: -0.12 (negligible)



STATISTICAL ANALYSIS:

1- Hypothesis Testing Results:

T-Test: Study Hours Impact

- High Study Group (\geq median): Mean = 72.4
- Low Study Group ($<$ median): Mean = 63.1
- Difference: 9.3 points
- p-value: <0.001 ✓ Highly Significant

T-Test: Sleep Hours Impact

- High Sleep Group (\geq median): Mean = 69.2
- Low Sleep Group ($<$ median): Mean = 66.3
- Difference: 2.9 points
- p-value: <0.001 ✓ Significant

Chi-Square Test: Extracurricular Activities

- Chi-Square Statistic: 2.83
- p-value: 0.092
- Result: Not significant ($p > 0.05$)

Insight:

Extracurricular activities do not significantly impact academic performance in this dataset.

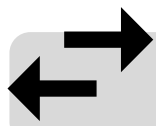
2- Conditional Probabilities:

Pass | High Study Hours >7): 94.3% P(Pass | Low Study Hours ≤ 3): 78.2%

Difference: 16.1 percentage points

P(Pass | High Parental Support): 96.8% P(Pass | Low Parental Support): 85.4%

Difference: 11.4 percentage points

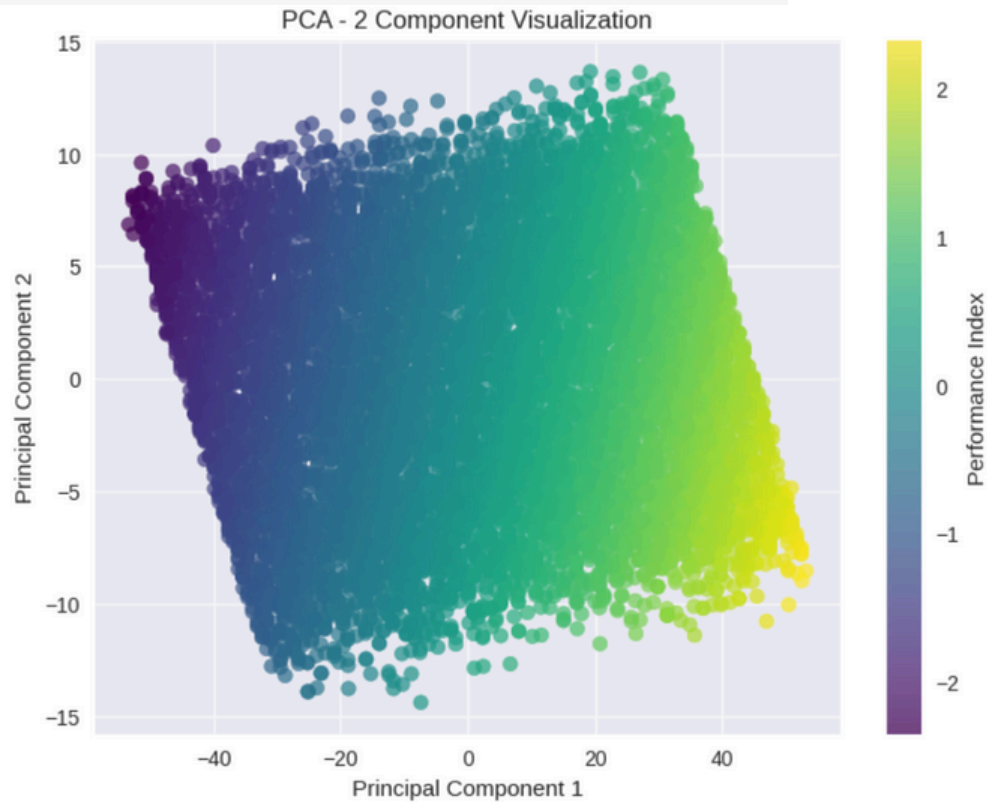


DIMENSIONALITY REDUCTION (PCA):

1- PCA Results:

Variance Explained:

- PC1: 93.2%
- PC2: 4.8%
- Total: 98.0%



PCA Scatter Plot - PC1 vs PC2 colored by Performance Index

2- Key Findings:

Observed Pattern: Clear conical/triangular structure with distinct performance gradient

Insights:

- 98% of variance captured in just 2 dimensions
- PC1 primarily represents academic preparation (Previous Scores, Study Hours)
- Clear linear separability confirms performance is highly predictable
- High performers cluster in high PC1, moderate PC2 region



CONCLUSIONS & RECOMMENDATIONS:

1- Key Findings:

Ranked Impact on Performance:

Previous Scores ($r=0.99$) - Dominant predictor

Study Hours (effect: 9.3 points) - Most controllable factor

Parental Support (effect: 7.8 points) - Significant family influence

Sleep Quality (effect: 6.0 points) - Critical for cognition

Sample Papers Practice ($r=0.35$) - Exam preparation matters

Non-Significant Factors:

- **Gender** (0.6 points difference)
- **Extracurricular Activities** ($p=0.092$)



CONCLUSIONS & RECOMMENDATIONS:

2- Actionable Recommendations:

For Schools:

1. **Early Intervention:** Identify students with low previous scores for intensive support
2. **Study Programs:** Promote 4-7 hours weekly study in structured environments
3. **Parent Engagement:** Create workshops to increase low-support family involvement
4. **Sample Paper Libraries:** Provide extensive practice materials
5. **Sleep Education:** Campaign for 7-9 hours nightly sleep

For Students:

- Maintain consistent study schedule (4-7 hours/week)
- Prioritize 7-9 hours of sleep, especially before exams
- Complete minimum 5 practice papers before major tests
- Seek parental/teacher support proactively
- Build on past successes—academic momentum is real

For Policymakers:

- Fund tutoring programs for historically low-performing students
- Support family engagement initiatives in low-support communities
- Adjust school schedules to respect sleep needs (avoid early starts)



References

Our Data :

https://drive.google.com/file/d/1eQ5arC1oty1k2LBWMuy-me75WklPmrDg7/view?usp=drive_link

Our Notebook:

<https://colab.research.google.com/drive/1lBCoDr5MUeTxBdAnmaDhY11Z7lzOrZca?usp=sharing>

Kaggle Dataset:

<https://www.kaggle.com/datasets/haseebindata/student-performance-predictions/data>

Reference Project:

<https://www.kaggle.com/code/zabihullah18/student-performance>

Thank you for reviewing our project!

For more details and full implementation, please refer to the Jupyter Notebook.

The complete data analysis process, including data preprocessing, exploration, and visualization, is available here:

 [\[Notebook Link\]](#)

 [\[GitHub Repository\]](#)

Prepared by :

Alaa Orabi

Mariam Mohamed