**STUDENT PERFORMANCE FACTOR ANALYSIS**

**Project Submitted to APSSDC**



**Submitted By**

      CHINTA SURYA HARSHITHA -22P31A4276

            RAJANA PAVANI -22P31A42B3

     PADALA ESWARA SRI SUBBAREDDY -23MH5A4903

KANKANALA CHARAN ESHWAR -23MH5A0318

     SANAGALA UPENDRA – 23NG5A1206

Under the Guidance of

**K. MEENAKSHI**

**M. RUTHUMMA**

           May 2025

**ABSTRACT**

Student academic performance has long been a focal point in educational research, as it reflects not only individual learning outcomes but also the overall effectiveness of the educational system. However, student performance is not determined solely by classroom instruction or cognitive ability; rather, it is influenced by a complex interplay of multiple external and internal factors. This study explores a wide range of variables—including attendance, parental support, access to internet, participation in extracurricular activities, sleep duration, motivation, tutoring, family income, type of school, peer influence, gender, and exam scores—to assess their impact on student performance.

The research utilizes a structured dataset containing student-related information and applies statistical techniques such as correlation analysis, multiple regression, and machine learning-based feature importance ranking to identify the most significant predictors of academic achievement. The study also employs visualization tools to represent relationships and trends in a clear and interpretable manner, enhancing the usability of the insights for educators and decision-makers.

Findings from the analysis suggest that variables like regular attendance, strong parental involvement, adequate sleep, and consistent motivation significantly correlate with higher academic performance. In contrast, factors such as gender or school type showed minimal direct impact in the dataset studied. These insights emphasize the importance of a supportive home environment and healthy personal habits in shaping student success.

**INTRODUCTION**

In today’s data-driven world, academic success is no longer viewed through the narrow lens of classroom learning alone. Student performance has emerged as a multidimensional concept that reflects a combination of academic ability, psychological well-being, socio-economic background, lifestyle habits, learning environments, and emotional support systems. Understanding what drives student performance is essential not only for improving individual academic results but also for building a more inclusive and effective education system.

Modern education systems around the globe face increasing pressure to cater to diverse student populations with varying abilities, learning styles, and life circumstances. Traditional evaluation methods that focus solely on grades or examination results often fail to capture the broader context that influences a student's academic journey. As a result, there is a growing need to investigate and understand the multiple internal and external factors that contribute to student success.

This project is designed to address that need by conducting a comprehensive analysis of key factors influencing student performance. Variables such as attendance, parental involvement, internet access, participation in extracurricular activities, sleep duration, motivation levels, tutoring, family income, school type, peer influence, gender, and exam scores are considered in this study. These factors are not examined in isolation but rather as part of a larger system in which they interact and influence one another.

By leveraging data analytics and machine learning techniques, this study aims to uncover patterns, correlations, and hidden insights that may otherwise go unnoticed. The outcomes of this research can help identify which students may require additional academic or emotional support and which strategies are most effective in enhancing student outcomes.

* Moreover, the findings are intended to benefit multiple stakeholders:
* Educators can use the insights to personalize teaching strategies.
* Parents can better understand their role in academic success.
* Policymakers can make informed decisions on educational reforms and resource distribution.

In essence, this project not only contributes to the growing field of educational data science but also promotes a shift from a one-size-fits-all model of education to a more equitable and student-centric approach. It highlights that academic performance is not merely a result of intelligence or teaching quality but a reflection of a student's entire ecosystem.

**SYSTEM REQUIRMENTS**

**Software Tools:**

Python (with Pandas, NumPy, Scikit-learn, Seaborn, Matplotlib)

R (optional for statistical testing)

Tableau or Power BI (for visualizations)

Jupyter Notebook or Google Colab

**Hardware:**

Minimum 8GB RAM

Intel i5 / AMD Ryzen processor

10–15 GB storage

**Methodology:**

1. Data Preprocessing

Remove null/missing data

Encode categorical values

Normalize numerical data

2. Exploratory Data Analysis

Use charts, graphs, and summary statistics to identify patterns

3. Statistical Analysis

Apply correlation tests and multiple regression analysis

4. Machine Learning Models

Train models like Decision Tree, Random Forest to identify feature importance

5. Model Validation

Use cross-validation, confusion matrix, and R² score for evaluation

**Data Collection**:

A structured dataset including:

Attendance

Parental support

Internet access

Extracurricular activity participation

Sleep hours

Motivation score

Tutoring support

Family income

School type

Peer influence

Gender

Exam scores

**LIBRARIES**

| Library | Purpose | | |
| --- | --- | --- | --- |
| pandas | Data loading, manipulation, and analysis | | |
| numpy | Numerical computations and array operations | | |
| matplotlib.pyplot | Basic plotting and visualizations | | |
| seaborn | Statistical data visualization (advanced plots) | | |
| sklearn.preprocessing | | Data preprocessing |

**ENVIRONMENT**

1. Programming Language

* Python (most likely version 3.8 or above)

2. Development Environment (IDE/Editor)

* Jupyter Notebook (via Anaconda Navigator)
* OR VS Code / PyCharm (if you used a standalone Python setup)

3. Package Manager

* pip or conda (for installing libraries like pandas, seaborn, scikit-learn)

4.Platform

* Windows 10/11 (based on previous messages and screenshots)
* Possibly using Anaconda Distribution, which simplifies environment and library management

5. Environment Setup Tool

* Either:
  + Virtual environment via venv or virtualenv
  + OR Conda environment via Anaconda (conda create -n your\_env\_name python=3.8)

**ARCHITECTURE**

**1. Data Collection**

This is the starting point where you acquire the dataset. The student performance factors dataset may be collected from:

* CSV/Excel files (often available on Kaggle)
* *Goal:* Gather raw data with all relevant course details (title, price, subject, subscribers, reviews, etc.).

**2. Data Preprocessing**

Before analysis, the data must be cleaned:

* Remove duplicates and irrelevant columns.
* Handle missing values (fill, drop, or interpolate).
* Convert data types (e.g., sleeping hours as float).
* Format text (e.g., lowercase titles, removing symbols).

*Goal:* Prepare the dataset for analysis by making it consistent and usable.

**3. Exploratory Data Analysis (EDA)**

This step helps you understand the data better using:

* Summary statistics (mean, median, count, etc.).
* Univariate and bivariate analysis.
* Checking distributions, correlations, outliers.

*Goal:* Find patterns, trends, and anomalies in the data.

**4. Feature Engineering**

You can derive new features from existing ones to improve insights:

* Course price buckets (Free, Low, Medium, High).
* Duration categories (Short, Medium, Long).
* Engagement ratio (reviews/subscribers).
* Course level conversion (e.g., Beginner = 1, Intermediate = 2).

*Goal:* Enhance data to extract deeper insights.

**5. Data Visualization**

This stage uses visual tools to represent data meaningfully:

* Bar charts (most popular categories)
* Pie charts (free vs paid course ratio)
* Box plots (price variation across subjects)
* Scatter plots (duration vs subscribers)

*Goal:* Make insights easier to understand and interpret.

**6. Reporting / Export**

Finally, present the results using:

* Dashboards (in Tableau, Power BI, or Streamlit)
* Static reports (PDFs, Excel files)
* Jupyter Notebooks with markdown summaries

*Goal:* Communicate findings clearly to stakeholders or users.

**ADVANTAGES**

1. Improved Student Outcomes – Identify students at risk and support them early.

2. Personalized Learning – Adjust teaching based on student profiles.

3. Data-Driven Decisions – Inform school and policy planning.

4. Early Risk Detection – Detect declining performance patterns.

5. Scalability – The framework can be applied to other schools or regions

**DISADVANTAGES**

1. Data Dependency – Inaccurate or missing data can lead to incorrect conclusions.

2. Generalizability – Results may not apply to all regions or age groups.

3. Oversimplification – Emotional and social factors may not be fully captured.

4. Bias in Self-Reported Data – Variables like motivation or sleep may be misreported.

**CONCLUSION**

This study concludes that student performance is affected by a complex combination of behavioral, social, academic, and environmental factors. Attendance, motivation, parental involvement, and sleep patterns emerged as the most influential predictors. Understanding these dynamics can guide educators, parents, and administrators to implement effective interventions, personalize education, and improve student outcomes.

Moving forward, expanding this model with more emotional, psychological, and longitudinal data can enhance accuracy and inclusivity. Education must evolve to support the whole student, not just the academic side, and data-driven insights like these play a critical role in making that shift possible

**REFERENCE**

Data Sets:

<https://www.kaggle.com/datasets/lainguyn123/student-performance-factors/code>

Learning:

<https://www.youtube.com/playlist?list=PLukqfSOJEZcsieLvTnExy7py9Pp9Jgnnr>