

MINI PROJECT REPORT

EXPLORING STUDENT ACHIEVEMENT

TRENDS

Name: Nishtha Gupta

Section: CSE-V

Enrollment Number: 240562

Objective-

The primary objective of this project was to analyze the factors influencing students' academic performance and to build a predictive model for mathematics scores. The analysis focused on understanding how demographic and socio-economic attributes—specifically gender, parental level of education, lunch type, and test preparation course completion—correlate with student scores in math, reading, and writing.

Data Source-

It was analyzed using the dataset of the "StudentsPerformance.csv" which was obtained via Kaggle. The data will include 1,000 records of individual students and the following eight attributes: Demographic (gender, race/ethnicity), Socio-economic (parental education, lunch type), Academic Preparation (test prep course), and Performance Scores (math, reading, writing).

Methods used-

Data Cleaning and Preparation:

1. Loading Data: The data was loaded and verified to have no missing values.

2. Dropping columns: The race / ethnicity column was omitted to narrow the analysis on those factors that can be improved and to avoid any possible bias.

3. Identifying outliers: The Interquartile Range (IQR) technique was used to identify outliers in the columns of scores but these were kept because they reflected true variations of students scores as opposed to data errors.

4. Encoding: The following categorical variables were encoded to be used in modeling- Nominal variables (gender, lunch and test preparation course) were coded using One-Hot Encoding.

The ordinal variable parental level of education was encoded using Label Encoding in order to maintain the order that exists in the variable.

5. Scaling: Numerical features (scores and encoded parental education) were scaled using StandardScaler to normalize their range(0-1) and ensure equal contribution during modeling.

Exploratory Data Analysis (EDA):

- **Univariate and Bivariate Analysis:** Histograms, boxplots, violin plots and a correlation heatmap were created to examine how the data were distributed and how the variables related to each other. This was an important step in determining the initial patterns and making hypotheses.

Hypothesis Testing:

- The independent samples t-test was performed so as to prove whether there was a statistically significant difference in average scores in math between the students who did the test preparation those who did not.
- Null Hypothesis (H_0): No difference between the mean scores.
- Alternative Hypothesis (H_1): High average scores are better among students who have already done the preparation.

Predictive Modeling:

- A **Linear Regression** model was developed for prediction math score.
- Predictive features were reading score, writing score and the encoded categorical variables.
- Standard metrics, included in the evaluation of the model performance, were R-squared (R^2), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE).

Key insights and findings

The assessment provided some practical lessons about the performance of students:

- **Holistic Academic Ability:** There was a very strong positive correlation between scores of math, reading, and writing ($R > 0.80$), with the highest relationship existing between reading and writing ($R = 0.95$). It means that the students who have succeeded in their performance in one subject have a high probability of success in other subjects.
- **Impact of Socio-economic Status:** The lunch variable, which acts as a proxy of socio-economic status, indicated that there was a large gap in performance. Students that took a standard lunch always scored high in all subjects compared to those who had free/reduced lunch.
- **Effect of Test Preparation:** The t-test gave a p-value of about 1.04×10^{-8} and is substantially below the 0.05 alpha value. This resulted in the rejection of the null hypothesis which gave a strong statistical evidence that a positive impact of enrolling in a course on test preparation has a positive and significant effect on math scores.
- **Gender-Based Performance Differences:** The performance in the subjects varied between genders. There was a minor difference in scores between male students and female students with males scoring a bit higher in math and female students showing an apparent higher score in reading and writing.
- **Influence of Parental Education:** There was a weak yet significant positive correlation, such that the higher the level of education of parents the higher the performance of students in math.

AI TOOLS USED- Gemini(5%)