Understanding and Forecasting Student Performance in India

Objective

Analyze student performance data to uncover patterns in academic success across subjects and demographics. Use statistical analysis and machine learning to predict performance and identify interventions for improvement — useful for educators and policymakers alike.

Dataset

- **Source:** [Kaggle - Student Performance Dataset (Math, Reading, Writing Scores)]
- **Link:** https://www.kaggle.com/datasets/spscientist/students-performance-in-exams

Columns include:

- gender, race/ethnicity, parental level of education,
- lunch, test preparation course
- math score, reading score, writing score

Use Cases

- Predict students at risk of underperforming
- Understand how socio-economic and educational backgrounds impact scores
- Recommend interventions (e.g., test prep, tutoring)
- Visualize gaps across gender or ethnicity groups

SECTION A: Python & Data Cleaning

- Load the dataset and inspect the first few rows, datatypes, and null values.
- Check for duplicate rows or invalid data entries.
- Standardize categorical values (e.g., group education levels, rename ethnicities).
- Add derived columns:
    - **Average Score** = (Math + Reading + Writing)/3
    - **Performance Category**: Low, Medium, High based on average score
    - **Preparation Effectiveness**: Compare scores with and without test prep

SECTION B: SQL Operations

- Import the cleaned dataset into a SQL database.
- Write SQL queries to:
    - List top 5 students with highest average scores.
    - Find the average math, reading, and writing scores by gender.
    - Compare average scores of students who completed test prep vs. those who didn't.
    - Count how many students fall into each performance category.
    - Identify which ethnic group has the highest average total score.

SECTION C: Exploratory Data Analysis & Descriptive Statistics

Exploratory Data Analysis:

- Histograms for all three subject scores.
- Box plots comparing scores by gender and parental education level.
- Grouped bar plots of average scores across test preparation and lunch type.
- Heatmap of correlation among numerical features.
- Scatter plot of math vs. reading scores with performance category color-coding.
- Stacked bar chart showing performance category by ethnic group.
- Count plot of number of students in each performance tier by gender.

Descriptive Statistics:

- Calculate mean, median, variance, and standard deviation for all three scores.
- Create summary tables showing average scores per category (e.g., lunch type, test prep).
- Calculate coefficient of variation for math, reading, and writing scores.
- Rank top 3 factors associated with high performance (based on group averages).
- Determine which feature (e.g., lunch type, gender) has the largest score variance.

SECTION D: Tableau Dashboard

- Build an interactive dashboard that includes:
    - Filters for gender, test prep, parental education
    - Line chart comparing average scores by parental education
    - Bar chart of performance tiers across ethnic groups
    - Pie chart showing proportion of students by lunch type or test prep status
    - Score tracker: given a demographic filter, show average subject scores
    - Highlight students with highest total scores per group
    - Dynamic performance heatmap based on math, reading, writing combinations

SECTION E: Machine Learning

Regression Task

- Predict total or average score based on demographic features.
- Models to try: Linear Regression, Random Forest Regressor
- Evaluate using MAE, MSE, R^2

Classification Task

- Classify students into performance buckets (Low, Medium, High).
- Label encode categorical variables.
- Train and compare:
    - Logistic Regression
    - Decision Tree
    - Random Forest
- Display feature importance from tree-based models