Exploring Real-World Data through Statistical and Predictive Modeling

Project Title:

Predicting and Understanding Student Academic Performance Using Regression, Logistic Analysis, and ANOVA

Finalized Topic:

Statistical and Predictive Modeling using ANOVA, Linear Regression, and Logistic Regression

Problem Statement:

Can we predict and understand student academic performance based on study habits, demographics, and socioeconomic factors, and determine which group differences are statistically significant?

This project aims to uncover what factors most influence student outcomes and whether certain categorical groupings (e.g., gender, parental education, school type) significantly affect academic success.

Why This Topic Matters:

Education is a critical driver of future success. By using statistical models:
- Educators can identify key performance drivers.
- Institutions can design better student support systems.
- Policy-makers can target specific groups needing attention.

Dataset Example:

UCI Student Performance Dataset
https://archive.ics.uci.edu/ml/datasets/Student+Performance

Contains data such as:
- Study time, absences, health, internet access
- Parental education, alcohol consumption
- Grades (G1, G2, G3), final outcome (pass/fail)
- School, gender, and more

Techniques Used:

1. ANOVA (Analysis of Variance)

To compare means of student performance across categories:
- Gender (Male vs Female)
- School (School A vs B)
- Parental education level
- Internet access (Yes/No)

Goal: Determine if group differences are statistically significant.

2. Linear Regression

To predict continuous final grade (G3) using:
- Study time
- Failures
- Family relationship quality
- Free time
- Absences
- Alcohol consumption

Goal: Quantify relationships and predict scores.

3. Logistic Regression

To classify whether a student will pass or fail:
- Convert G3 to binary (Pass if G3 ≥ 10, else Fail)
- Input features: study time, absences, family support, etc.

Goal: Identify risk factors and build a predictive model.

Steps & Deliverables:

1. Data Cleaning & Structuring
2. Exploratory Data Analysis (EDA)
3. Apply ANOVA for group comparisons
4. Develop Regression and Logistic Models
5. Evaluate Model Performance
6. Build a Mini Application (using Streamlit/Flask)
7. Prepare a Comprehensive Report

Tools and Libraries:

- Python: pandas, seaborn, scikit-learn, statsmodels
- Visualization: matplotlib, seaborn

- App: Streamlit or Flask
- Documentation: Jupyter Notebook + PDF report

Timeline:

Week 1: Topic Finalization, Data Exploration
Week 2: Data Cleaning, EDA, ANOVA
Week 3: Regression Modeling
Week 4: Logistic Modeling
Week 5: App Development
Week 6: Final Report & Presentation

Expected Outcomes:

- Identify key predictors of academic performance
- Show statistically significant differences across groups
- Build a working app to predict outcomes from user inputs
- Present insights to educators or policymakers