Project_1: Student Data Analysis

Title: Student Performance Analysis and Prediction using Python

Abstract: The project analyzes student performance on the basis of demographic, behavioral, and academic features. Exploratory Data Analysis (EDA) and Linear Regression were implemented to identify key factors influencing student score. The model predicts average student performance based on lifestyle and study-related variables.

Objectives:

1. To analyze factors influencing students' academic performance
2. To visualize relation between students' habits and marks
3. To build a predictive model which predicts students' average score
4. To test the model performance using MAE and R^2

Dataset Description: The dataset contains student demographic details, behavioral attributes, and subject marks in Mathematics, Science, and Physics.

Key Features Used:

- Study behavior: studytime, failures, absences
- Lifestyle: goout, Walc, activities, internet
- Demographics: sex, age, address
- Academic scores: Maths, Science, Physics

A new feature **Average** was created as the target variable.

Tools & Technologies Used:

- Python
- Pandas
- Numpy
- Matplotlib
- Seaborn
- Scikit-learn
- Google Colab

Project_1: Student Data Analysis

Data Processing: Irrelevant columns were removed for better data insights. Categorical variables were converted to numerical for the purpose of linear regression. Checked for missing values, created average score column and performance categories on the basis of average score column.

Exploratory Data Analysis: Performed key visuals such as heatmap, study time v/s average score, weekly alcohol v/s absences, internet v/s average score, going out v/s weekly alcohol, weekly alcohol v/s health, health v/s study time, sex v/s activities, etc.

Model Building: Linear Regression was used to predict average score. The train-test split was 80-20 and the evaluation metrics were MAE and R^2 .

Results: The MAE was ~2.28 and the R^2 Score was ~0.19. Thus the model shows reasonable prediction error but limited explanatory power, indicating student performance depends on multiple complex factors.

Conclusion: Student performance is influenced by study time, failures, and attendance. While the Linear Regression model provides acceptable error margins, the relatively low R^2 suggests that more advanced models or additional features may improve prediction accuracy.

Improvements: Larger datasets can be used and Random Forest can be applied to improve the accuracy of prediction. An early alarming system can be developed for students on the basis of performance prediction and can be launched as a web app.