student's Exam Score

The objective of this project is to build a machine learning model to predict student exam scores based on various influential factors. The project utilizes a dataset containing information on study habits, previous academic performance, and other socio-economic factors to predict a student's `Exam_Score`.

Dataset Details

The dataset used is **StudentPerformanceFactors.csv**. It contains **6607 entries** and **20 features**, with the target variable being `Exam_Score`.

Key Features:

- **`Hours_Studied`** (int64): The number of hours a student studied.
- **`Attendance`** (int64): The student's attendance percentage.
- **`Previous_Scores`** (int64): The student's previous academic scores.
- **`Parental_Involvement`** (object): The level of parental involvement (Low, Medium, High).
- **`Motivation_Level`** (object): The student's motivation level (Low, Medium, High)
- **`Exam_Score`** (int64): The final exam score (target variable).

The dataset initially had missing values in the `Teacher_Quality`, `Parental_Education_Level`, and `Distance_from_Home` columns.

Steps Followed

1. Data Loading and Preprocessing

- **Data Loading:** The dataset was loaded into a pandas DataFrame.
- **Handling Missing Values:** Missing values in numerical columns were filled with the **mean** of their respective columns. For categorical columns, missing values were filled with the **mode** (the most frequent value).
- **Dropping Duplicates:** Duplicate rows in the dataset were removed to ensure data integrity.
- **Data Splitting:** The dataset was split into training and testing sets with a **20%** test size, using `train_test_split` from `sklearn.model_selection`.

2. Modeling

A **Linear Regression model** was chosen for this project to predict the student exam scores. The model was trained on the preprocessed training data.

3. Evaluation and Results

The model's performance was evaluated using several key regression metrics:

- **Mean Absolute Error (MAE):** Measures the average absolute difference between the predicted and actual values.
- **Mean Squared Error (MSE):** Measures the average squared difference between the predicted and actual values. It penalizes larger errors more heavily.
- **R^2 Score:** Represents the proportion of the variance in the dependent variable that is predictable from the independent variables. A higher R^2 score indicates a better fit.

Single Feature Model Performance (`Hours_Studied`)

- **MAE:** 1.8344
- **MSE:** 5.0945
- **R^2 Score:** 0.6499

Multiple Features Model Performance

For this model, a combination of features was used: `Hours_Studied`, `Sleep_Hours`, `Attendance`, `Motivation_Level`, and `Previous_Scores`. The model demonstrated improved performance, as expected.

- **MAE:** 1.3562
- **MSE:** 5.3345
- **R^2 Score:** 0.6225

4. Visualizations & Insights

- **Scatter Plot of `Hours_Studied` vs. `Exam_Score`**: This visualization shows a strong **positive linear relationship** between the number of hours studied and the exam score. As the hours studied increase, the exam scores generally also increase.
- **Correlation Heatmap**: The heatmap provides a clear view of the linear correlations between all numerical features. It reveals that `Hours_Studied`, `Previous_Scores`, and `Attendance` have the strongest positive correlation with `Exam_Score`.

Bonus Work

The project includes an extra experiment with a **Multiple Features Model**. By incorporating additional features such as `Sleep_Hours`, `Attendance`, and `Motivation_Level`, the model's predictive capability was enhanced, leading to a better R^2 score and a lower MAE.

Conclusion & Learning Outcomes

This project successfully demonstrates the process of building, training, and evaluating a machine learning model for a regression task. The analysis highlights the importance of data preprocessing, especially handling missing values, and shows how incorporating multiple relevant features can significantly improve a model's performance. The strong correlation between study habits and exam scores underscores the direct impact of these factors on academic success.