Student Performance Prediction using Linear Regression

Created by: Dulhara Lakshan (2025.03.14)

1. Introduction

This project aims to predict student performance based on various factors such as study habits, participation, and stress levels using **Linear Regression**. The model is trained on a dataset containing student records and predicts final grades based on multiple independent variables.

2. Dataset Overview

The dataset contains student-related attributes, including:

- **Demographic Information**: Age, Gender
- **Study Habits**: Study Hours per Week, Online Courses Completed
- **Engagement**: Participation in Discussions, Use of Educational Technology
- **Performance Metrics**: Assignment Completion Rate, Exam Score, Attendance Rate
- **Lifestyle Factors**: Time Spent on Social Media, Sleep Hours per Night
- **Psychological Factors**: Preferred Learning Style, Self-Reported Stress Level
- **Target Variable**: Final_Grade (Encoded as a categorical variable)

3. Data Preprocessing

3.1 Handling Missing Values

The dataset was checked for missing values using df.isnull().sum(). No missing values were found.

3.2 Encoding Categorical Features

Categorical variables were encoded as follows:

- **Binary Encoding**:
  - Gender: Female → 0, Male → 1
  - Participation_in_Discussions: No → 0, Yes → 1
  - Use_of_Educational_Tech: No → 0, Yes → 1
- **One-Hot Encoding**:
  - Preferred_Learning_Style: Converted into three columns (Kinesthetic, Reading/Writing, Visual)

- Self_Reported_Stress_Level: Converted into three columns (Low, Medium, High)

- **Encoding Target Variable**:
  - Final_Grade: Categorical values were converted into numerical values using LabelEncoder.

3.3 Feature Selection

The independent variables (**features**) selected for training the model are:

```
FEATURES = [
  "AGE", "GENDER", "STUDY_HOURS_PER_WEEK", "ONLINE_COURSES_COMPLETED",
  "PARTICIPATION_IN_DISCUSSIONS", "ASSIGNMENT_COMPLETION_RATE (%)", "EXAM_SCORE (%)",
  "ATTENDANCE_RATE (%)", "USE_OF_EDUCATIONAL_TECH", "TIME_SPENT_ON_SOCIAL_MEDIA
(HOURS/WEEK)",
  "SLEEP_HOURS_PER_NIGHT", "PREFERRED_LEARNING_STYLE_KINESTHETIC",
  "PREFERRED_LEARNING_STYLE_READING/WRITING", "PREFERRED_LEARNING_STYLE_VISUAL",
  "SELF_REPORTED_STRESS_LEVEL_LOW", "SELF_REPORTED_STRESS_LEVEL_MEDIUM"
]
```

4. Model Training

The dataset was split into training and testing sets (80%-20%) using train_test_split(). The model used for training was **Linear Regression**, implemented as follows:

```
FROM SKLEARN.LINEAR_MODEL IMPORT LINEARREGRESSION
MODEL = LINEARREGRESSION()
MODEL.FIT(X_TRAIN, Y_TRAIN)
```

THE TRAINED MODEL PRODUCED THE FOLLOWING COEFFICIENTS:

```
PRINT(F"INTERCEPT (C): {MODEL.INTERCEPT_}")
PRINT(F"COEFFICIENTS (M): {MODEL.COEF_}")
```

5. Feature Importance

The importance of each feature was visualized using a bar plot of the regression coefficients.

```
plt.figure(figsize=(10, 6))

sns.barplot(x="Coefficient", y="Feature", data=coef_df, palette="coolwarm")

plt.axvline(0, color="black", linewidth=1.2)

plt.title("Feature Importance (Linear Regression Coefficients)")

plt.xlabel("Coefficient Value")

plt.ylabel("Features")

plt.show()
```

6. Model Evaluation

The model was evaluated using **Mean Squared Error (MSE)** and **R^2 Score**.

```
from sklearn.metrics import mean_squared_error, r2_score

y_pred = model.predict(X_test)

mse = mean_squared_error(y_test, y_pred)

r2 = r2_score(y_test, y_pred)

print(f"Mean Squared Error: {mse:.2f}")

print(f"R² Score: {r2:.4f}")
```

A scatter plot was used to compare actual vs. predicted values:

```
plt.scatter(y_test, y_pred, color='blue')

plt.xlabel("Actual Scores")

plt.ylabel("Predicted Scores")

plt.title("Actual vs Predicted Student Scores")

plt.show()
```

7. Individual Feature Analysis

Scatter plots were generated to visualize how each independent variable impacts the final grade.

for feature in features:

```
plt.figure(figsize=(6, 4))
sns.scatterplot(x=df[feature], y=df["Final_Grade"], alpha=0.5, color="blue")
plt.xlabel(feature)
plt.ylabel("Final_Grade")
plt.title(f"Final Grade vs {feature}")
plt.show()
```

8. Making Predictions

A new student’s data was used to predict their final grade.

```
new_student = np.array([[18, 1, 40, 10, 1, 85, 75, 90, 1, 15, 7, 0, 1, 0, 0, 1]])
predicted_grade = model.predict(new_student)
print(f"Predicted Final Grade: {predicted_grade[0]}")
```

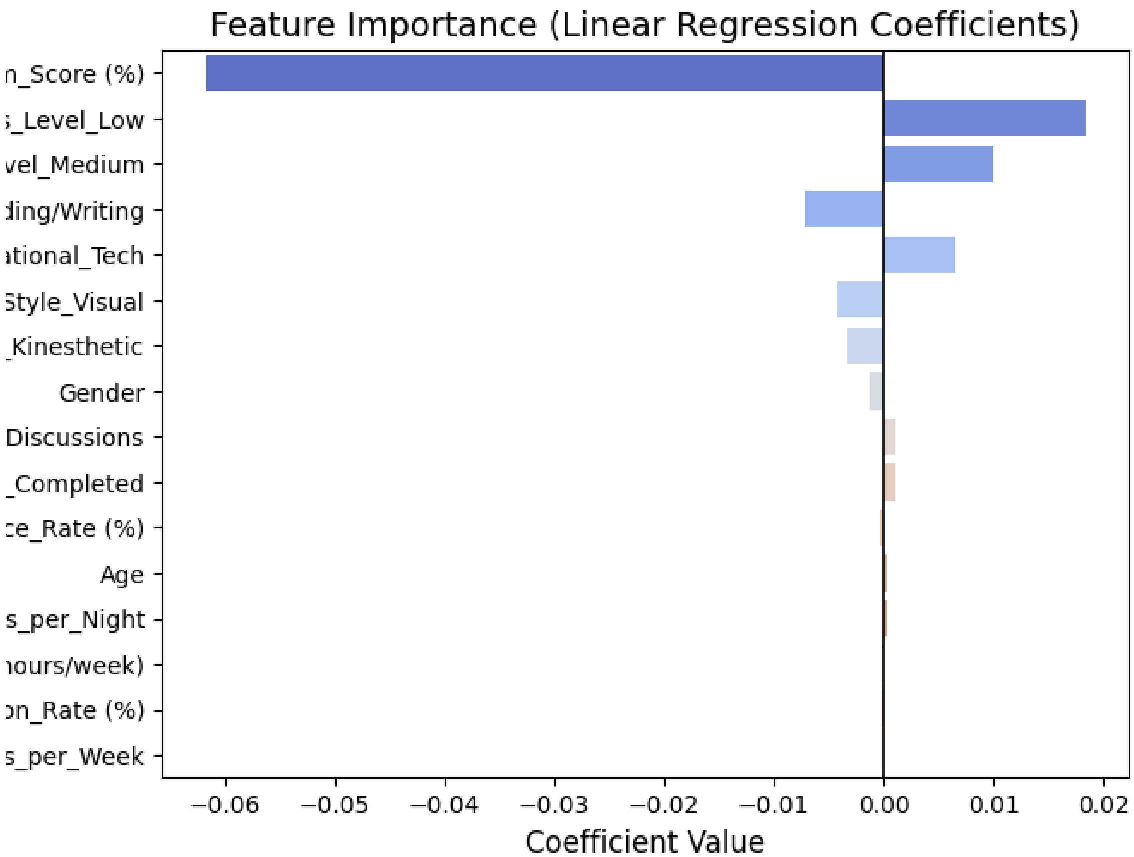
9. Conclusion

This project successfully applied **Linear Regression** to predict student performance based on multiple academic and lifestyle factors. The model demonstrated meaningful relationships between study habits, participation, and final grades. Future improvements could include:

- Testing alternative models (e.g., Decision Trees, Random Forest)
- Expanding feature selection with additional psychological and social factors
- Implementing feature scaling and polynomial regression for better performance.

This project provides a valuable tool for educators and students to identify key factors influencing academic success.

Feature Coefficient



Based on these Linear Regression coefficients, we can interpret the impact of each feature on student grades:

Feature	Coefficient
Age	3.20149716e-04
Gender	-1.33543458e-03
Study_Hours_per_Week	-7.44875421e-05
Online_Courses_Completed	9.93713199e-04
Participation_in_Discussions	1.02266361e-03
Assignment_Completion_Rate (%)	-2.33892195e-04
Exam_Score (%)	-6.17185666e-02
Attendance_Rate (%)	-4.05044953e-04
Use_of_Educational_Tech	6.47230358e-03
Time_Spent_on_Social_Media (hours/week)	-2.50046174e-04
Sleep_Hours_per_Night	2.57071811e-04
Preferred_Learning_Style_Kinesthetic	-3.37131296e-03
Preferred_Learning_Style_Reading/Writing	-7.27618642e-03
Preferred_Learning_Style_Visual	-4.23511487e-03
Self_Reported_Stress_Level_Low	1.83073644e-02
Self_Reported_Stress_Level_Medium	9.88368918e-03

Most Influential Factors:

- Exam Score (%) (-0.0617) – The strongest predictor of final grades (negative coefficient means lower scores significantly decrease final grades).
- Self-Reported Stress Level (Medium) (+0.0099) – Moderate stress slightly improves performance, possibly due to motivation.
- Self-Reported Stress Level (Low) (+0.0183) – Lower stress levels positively influence grades more than medium stress.
- Use of Educational Tech (+0.0064) – Using tech tools has a small but positive effect on student success.

Negligible or Negative Impact:

- Study Hours per Week (-0.00007) – Almost no direct impact, suggesting quality > quantity in study sessions.

- Attendance Rate (-0.0004) – Surprisingly low effect, indicating attendance alone isn't a strong predictor.
- Learning Styles (Kinesthetic: -0.0033, Reading/Writing: -0.0072, Visual: -0.0042) – No learning style has a major influence, meaning adaptability may be key.
- Time Spent on social media (-0.00025) – Slight negative impact, but not as harmful as expected.

Key Takeaways:

- Grades are most impacted by exam performance, stress levels, and educational tech usage.
- Study hours alone don't guarantee higher grades—effective learning strategies matter more.
- Stress management plays a crucial role, as moderate stress may be beneficial.
- Social media usage has minimal impact, contradicting common assumptions.

This analysis can help educators and students focus on what truly improves academic performance.