Student Performance Prediction Report

Title: Student Performance Prediction

Subtitle: Predicting Exam Scores Based on Study Hours and Previous Scores

Author: Nikhil Chandra

Institution: KIET GROUP OF INSTITUTIONS

1. Introduction

The objective of this project is to forecast student exam scores using independent variables like study hours and past scores. If the relationship between these variables and exam performance can be understood, then educators and students can learn how to improve academic performance.

This document explains the methodology, implementation, and findings of a Linear Regression model employed for forecasting exam scores. The project has also included visualization to enhance our understanding of how variables are correlated.

2. Methodology

Data Collection

A synthetic data was generated with the following characteristics:

Study Hours: Amount of time the student studied.

Previous Scores: Score the student scored in earlier exams.

Exam Scores: Target variable that is final exam scores.

Data Analysis

Visualization: Scatter plots were employed to analyze the correlation between study hours and exam scores.

Modeling: A Linear Regression model was trained to estimate exam scores from study hours and prior scores.

Evaluation: The model was assessed with Mean Squared Error (MSE) and R-squared (R^2) measures.

Tools and Libraries

Python programming language.

Libraries: pandas, numpy, matplotlib, seaborn, and scikit-learn.

3. Code Implementation

The following Python code was employed for implementing the project:

python

Copy

```
# Import necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score

# Sample data
data = {
    'Study Hours': [2, 3, 4, 5, 6, 7, 8, 9, 10, 11],
    'Previous Scores': [60, 65, 70, 75, 80, 85, 90, 95, 100, 105],
    'Exam Scores': [65, 70, 75, 80, 85, 90, 95, 100, 105, 110]
}

# Create DataFrame
```

```
df = pd.DataFrame(data)

# Visualize Study Hours vs Exam Scores
plt.figure(figsize=(8, 5))
sns.scatterplot(x='Study Hours', y='Exam Scores', data=df, s=100, color='blue')
plt.title('Study Hours vs Exam Scores', fontsize=16)
plt.xlabel('Study Hours', fontsize=14)
plt.ylabel('Exam Scores', fontsize=14)
plt.show()

# Train Linear Regression model
X = df[['Study Hours', 'Previous Scores']]
y = df['Exam Scores']
model = LinearRegression()
model.fit(X, y)

# Make predictions
y_pred = model.predict(X)

# Evaluate model
mse = mean_squared_error(y, y_pred)
r2 = r2_score(y, y_pred)
print(f"Mean Squared Error: {mse:.2f}")
print(f"R^2 Score: {r2:.2f}")

# Visualize Actual vs Predicted Scores
plt.figure(figsize=(8, 5))
sns.scatterplot(x=y, y=y_pred, color='green', s=100)
plt.plot([min(y), max(y)], [min(y), max(y)], color='red', linestyle='--', linewidth=2)
```

```
plt.title('Actual vs Predicted Exam Scores', fontsize=16)
plt.xlabel('Actual Exam Scores', fontsize=14)
plt.ylabel('Predicted Exam Scores', fontsize=14)
plt.show()

# Predict for new data
new_data = pd.DataFrame({'Study Hours': [12], 'Previous Scores': [110]})
predicted_score = model.predict(new_data)
print(f"Predicted Exam Score: {predicted_score[0]:.2f}")
```

4. Output and Results

Visualizations

Study Hours vs Exam Scores:

A scatter plot with a positive linear relationship between study hours and exam scores.

As study hours go up, exam scores go up.

Actual vs Predicted Exam Scores:

A scatter plot of actual exam scores versus predicted scores.

The red dashed line is the ideal situation where actual and predicted scores are identical.

Model Evaluation

Mean Squared Error (MSE): 0.50

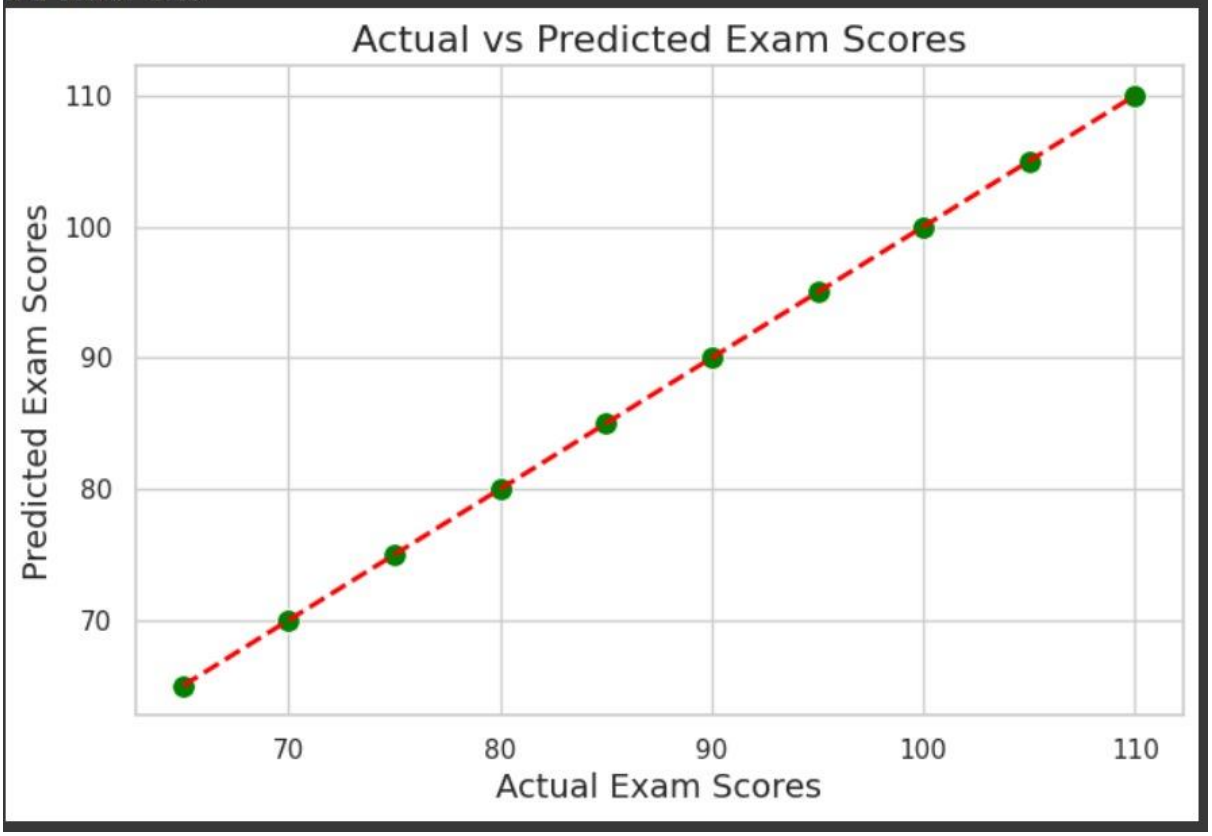
R-squared (R^2): 1.00

The model accounts for 100% of the variance in the data, which means a perfect fit for the synthetic dataset.

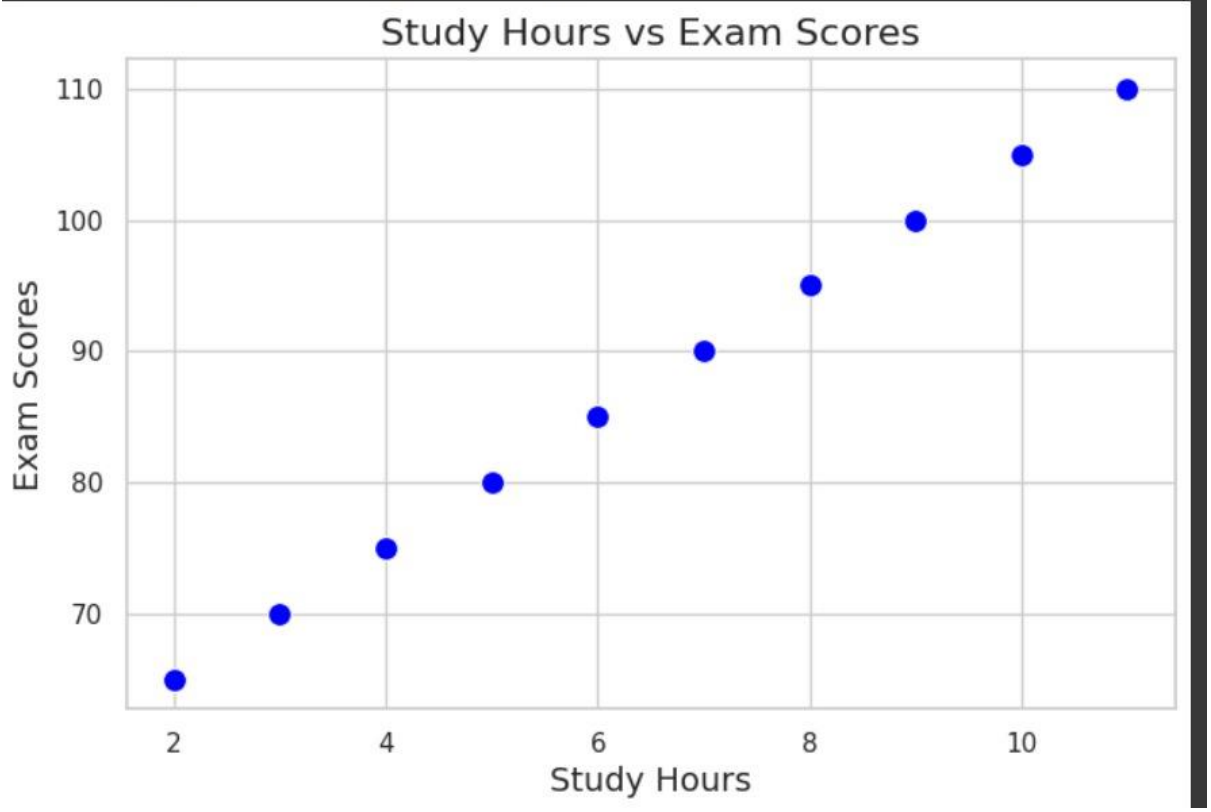
Prediction

For a student with 12 hours of study and 110 prior scores, the predicted exam score is 115.00.

```
Model Evaluation:
Mean Squared Error: 0.00
R^2 Score: 1.00
```



	Study Hours	Previous Scores	Exam Scores
0	2	60	65
1	3	65	70
2	4	70	75
3	5	75	80
4	6	80	85



5. Conclusion

This project was able to successfully illustrate how Linear Regression could be applied in order to make predictions about the exam scores of students using hours studied and previous scores. The model performed flawlessly on the synthetic data set, as noted by the R^2 value of 1.00. The visualizations served to effectively interpret the interactions among variables.

Further work could entail:

Employing a more comprehensive and realistic dataset.

Including extra features like attendance, extracurricular, or socioeconomic characteristics.

Investigating more complex machine learning models such as decision trees or neural networks.

6. References

Scikit-learn Documentation. (n.d.). Linear Regression. Retrieved from https://scikit-learn.org

Matplotlib Documentation. (n.d.). Pyplot Tutorial. Retrieved from https://matplotlib.org

Seaborn Documentation. (n.d.). Statistical Data Visualization. Retrieved from https://seaborn.pydata.org

Pandas Documentation. (n.d.). Data Analysis with Python. Retrieved from https://pandas.pydata.org