# TITLE: -

# CREATED BY: -

Name: - Uday Banduni

Branch: - B-Tech (CSEAI)

Course: - Introduction to AI

Institute: - KIET Groups of Institute

University Roll. No.: - 202401100300267

Date: - 11th March, 2025

I worked on this problem statement under the guidance of our AI Teacher Mr. Abhisekh Shukla.

# INTRODUCTION

In this project, we aim to predict student performance based on several factors, primarily study hours , previous exam scores, and final exam score. By using Linear Regression, we model the relationship between these factors and the student's final exam scores.

The main objective of this project is to demonstrate how machine learning techniques can be applied to predict outcomes in an educational setting. We use simple and intuitive tools such as scikit-learn and matplotlib in Python to develop and evaluate the model, while visualizing the relationships between the predictors and the target variable.

# METHODOLOGY

## 1. Dataset

The dataset used in this study contains the following features:

StudentID: A unique identifier for each student (not used in prediction).

StudyHours: The number of hours a student studies.

PreviousScores: The scores obtained in past exams.

FinalExamScore: The actual performance of the student in the final exam (target variable).

## 2. Data Preprocessing

The dataset is loaded from a CSV file.

Unnecessary columns such as Student ID are removed.

The dataset is split into training (80%) and testing (20%) sets.

## 3. Model Selection & Training

A Linear Regression model is selected for predicting student performance.

The model is trained u sing the training dataset.

## 4. Evaluation Metrics

To assess the performance of the model, the following metrics are used:

Mean Absolute Error (MAE): Measures the average absolute difference between actual and predicted values.

Mean Squared Error (MSE): Evaluates the squared differences, penalizing larger errors.

R-Squared Score ($R^2$): Determines how well the model explains the variance in student performance.

# CODE FOR PROBLEM

```
# Import necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_error, r2_score

# Load dataset
file_path = "/mnt/data/student_data.csv"
df = pd.read_csv('student_data.csv')

# Drop StudentID as it's not relevant for prediction
df = df.drop(columns=['StudentID'])

# Feature selection (independent variables)
X = df[['StudyHours', 'PreviousScores']]

# Target variable (dependent variable)
y = df['FinalExamScore']

# Split the dataset into training and testing sets (80% training, 20% testing)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Create and train the Linear Regression model
model = LinearRegression()
model.fit(X_train, y_train)

# Make predictions using the trained model
```

```
y_pred = model.predict(X_test)

# Evaluate the model's performance
mae = mean_absolute_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f'Mean Absolute Error: {mae}')
print(f'R-squared: {r2}')

# Visualize the relationship between study hours and exam scores
plt.figure(figsize=(8, 6))
plt.scatter(df['StudyHours'], df['FinalExamScore'], color='blue')
plt.title('Study Hours vs Final Exam Scores')
plt.xlabel('Study Hours')
plt.ylabel('Final Exam Scores')
plt.grid(True)
plt.show()
# Visualize the relationship between previous scores and final exam scores
plt.figure(figsize=(8, 6))
plt.scatter(df['PreviousScores'], df['FinalExamScore'], color='red')
plt.title('Previous Scores vs Final Exam Scores')
plt.xlabel('Previous Scores')
plt.ylabel('Final Exam Scores')
plt.grid(True)
plt.show()
```

# RESULT OF OUR CODE

The output of the model includes the following evaluation metrics:

Mean Absolute Error: Displays the average error in predictions.

Mean Squared Error: Indicates the variance of prediction errors.

R-squared Score: Shows how well the model fits the data.

Additionally, two scatter plots are generated:

Study Hours vs. Final Exam Scores: To analyze how study time affects final performance.

Previous Scores vs. Final Exam Scores: To observe the impact of past academic performance on final results.

# REFERENCES

References

Scikit-learn documentation: https://scikit-learn.org/

Python Data Science Handbook by Jake VanderPlas

Machine Learning with Python by Andreas Müller and Sarah Guido

Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow by Aurélien Géron

Research papers on student performance prediction using machine learning techniques