**Student Performance**

**REPORT**

AI & Machine Learning system Engineering ( MSE ) Report

Name: Anand Kr Gupta

Roll No.: 202401100400033

Date: 10/03/2025

Branch-CSEAIML

Sec-A

PROBLEM STATEMENT: Student Performance Prediction – Predict exam scores based on study hours and other factors, visualizing the relationship.

**1. Introduction**

This report analyzes student performance based on study hours, attendance, and previous test scores. The objective is to predict exam scores and understand the relationships between various factors affecting performance. This analysis is crucial for educators and students to optimize study plans and improve outcomes.

**2. Methodology**

The dataset includes 50 entries with the following features:  
- \*\*Study Hours (per day):\*\* Average hours studied by students per day.  
- \*\*Attendance (%):\*\* Attendance percentage of the students.  
- \*\*Previous Test Score (%):\*\* Percentage scored in previous tests.  
- \*\*Exam Score (%):\*\* Percentage scored in the final exam.

Steps undertaken for the analysis:  
1. \*\*Data Exploration:\*\* Displayed the first few rows of the dataset, checked dimensions, and examined data types.  
2. \*\*Data Cleaning:\*\* Handled missing values by either dropping rows or filling them with a specific value (0).  
3. \*\*Data Visualization:\*\* Created histograms and scatter plots to understand the distribution and relationships between variables:  
 - Histograms for attendance, previous test scores, and exam scores.  
 - Scatter plot for study hours vs. exam scores.

**3. Code**  
import pandas as pd # Importing pandas for data manipulation  
import numpy as np # Importing numpy for numerical operations  
import matplotlib.pyplot as plt # Importing matplotlib for plotting  
import seaborn as sns # Importing seaborn for enhanced data visualization  
  
# Load the dataset  
df = pd.read\_csv('/content/Student\_Performance\_Dataset.csv')  
  
# Display the first few rows of the dataset  
print(df.head())  
# Check the dimensions of the dataset  
print(df.shape)  
# Get summary statistics of numerical variables  
print(df.describe())  
# Check the data types of variables  
print(df.dtypes)  
# Check for missing values  
print(df.isnull().sum())  
# Handle missing values  
df.dropna(inplace=True) # Drop rows with missing values  
value = 0 # Value to fill missing data if required  
df.fillna(value, inplace=True)  
  
# Histogram for Attendance  
plt.hist(df['Attendance (%)'], bins=10, color='lightgreen', edgecolor='black')  
plt.xlabel('Attendance (%)')  
plt.ylabel('Frequency')  
plt.title('Distribution of Attendance')  
plt.show()  
# Histogram for Exam Scores  
plt.hist(df['Exam Score (%)'], bins=10, color='dodgerblue', edgecolor='black')  
plt.xlabel('Exam Score (%)')  
plt.ylabel('Frequency')  
plt.title('Distribution of Exam Scores')

plt.show()

# Histogram for Previous Test Scores  
plt.hist(df['Previous Test Score (%)'], bins=10, color='lightcoral', edgecolor='black')  
plt.xlabel('Previous Test Score (%)')  
plt.ylabel('Frequency')  
plt.title('Distribution of Previous Test Scores')  
plt.show()  
  
# Scatter plot for Study Hours vs. Exam Score  
plt.scatter(df['Study Hours (per day)'], df['Exam Score (%)'], color='purple', alpha=0.7)  
plt.xlabel('Study Hours (per day)')  
plt.ylabel('Exam Score (%)')  
plt.title('Study Hours vs. Exam Score')  
plt.show()

**4. Output/Result**

\*\*Key Findings:\*\*  
1. \*\*Data Cleaning:\*\*  
- Missing values were handled by either dropping rows or filling them with 0.  
- Ensured a clean and consistent dataset.  
  
2. \*\*Data Visualization:\*\*  
- \*\*Attendance:\*\* Majority of students maintain moderate to high attendance percentages.  
- \*\*Previous Test Scores:\*\* Distributed over a wide range with a cluster around higher scores.  
- \*\*Exam Scores:\*\* Concentrated towards the higher range, showing good performance.  
- \*\*Scatter Plot (Study Hours vs. Exam Score):\*\* Indicates a positive correlation; students who study more hours tend to score higher in exams.  
  
3. \*\*Summary Statistics:\*\*  
- Statistics such as mean, median, and standard deviation were calculated for numerical variables to summarize the dataset.

A screenshot of a computer

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect. A screenshot of a computer screen

AI-generated content may be incorrect.A screenshot of a graph

AI-generated content may be incorrect.**5. References/Credits**

1. Dataset created for academic purposes.  
2. Libraries used: Pandas, Numpy, Matplotlib, Seaborn.  
3. Google Colab for code execution and visualization.