****

**Assessment Report**

on

**“Student Performance Prediction”**

submitted as partial fulfillment for the award of

**BACHELOR OF TECHNOLOGY**

**DEGREE**

SESSION 2024-25

in

**CSE AIML**

By

Tarun Singh (202401100400198)

**Under the supervision of**

“Abhishek Shukla”

**KIET Group of Institutions, Ghaziabad**

Affiliated to

**Dr. A.P.J. Abdul Kalam Technical University, Lucknow**

(Formerly UPTU)

**May, 2025**

**2. Introduction**

**In modern education systems, the ability to analyse and understand student performance is critical for teachers, institutions, and policymakers. With the rise of data science and machine learning, educational data can be processed to extract meaningful patterns that help improve teaching strategies and student outcomes.**

**This project focuses on creating a Python-based tool to analyse student performance using a CSV dataset. The system is designed to run on Google Collab, allowing users to easily upload their dataset through an interactive interface. Once uploaded, the program performs data cleaning, exploratory data analysis (EDA), and visualizations to reveal insights about student scores across various subjects.**

**Furthermore, the analysis includes gender-based comparisons, correlation studies, and performance categorization. By leveraging libraries like Pandas, Seaborn, and Matplotlib, the project demonstrates how simple yet powerful tools can be used to analyse educational data efficiently. This solution aims to be a practical starting point for deeper learning analytics projects**

**c. Methodology**

**The following steps were used to analyze the dataset:**

1. **CSV Upload: The user uploads a dataset file interactively using google.colab.files.upload() in Colab.**
2. **Data Loading: The uploaded file is read into a Pandas DataFrame.**
3. **Data Cleaning: The code checks for missing values and ensures the data is ready for analysis.**
4. **Exploratory Data Analysis (EDA):**
   * **Summary statistics using .describe()**
   * **Visualizations using Matplotlib and Seaborn (e.g., histograms, box plots, heatmaps)**
5. **Categorization: Based on average of scores (Math, Reading, Writing), students are categorized into "Low", "Medium", and "High" performers.**
6. **Additional Analysis: Performance is also compared based on gender to identify any patterns.**

**Code**

**✅ Below is the code implemented for the project. This code was executed and tested in Google Collab.**

**# Import required libraries**

**import pandas as pd**

**import matplotlib.pyplot as plt**

**import seaborn as sns**

**from google.colab import files**

**from io import StringIO**

**# Upload the CSV file**

**print("📁 Please upload your CSV file:")**

**uploaded = files.upload()**

**# Read the uploaded file**

**for filename in uploaded.keys():**

**print(f"\n✅ File '{filename}' uploaded successfully!\n")**

**data = pd.read\_csv(StringIO(uploaded[filename].decode('utf-8')))**

**# Preview the data**

**print("🔍 Preview of the Dataset:")**

**display(data.head())**

**# Dataset info**

**print("\n📌 Dataset Info:")**

**print(data.info())**

**# Summary statistics**

**print("\n📊 Statistical Summary:")**

**display(data.describe())**

**# Check for missing values**

**print("\n❓ Missing Values:")**

**print(data.isnull().sum())**

**# Score distribution**

**score\_columns = ['math score', 'reading score', 'writing score']**

**existing\_scores = [col for col in score\_columns if col in data.columns]**

**if existing\_scores:**

**print("\n📈 Score Distributions:")**

**for col in existing\_scores:**

**plt.figure(figsize=(8, 4))**

**sns.histplot(data[col], bins=20, kde=True, color='skyblue')**

**plt.title(f'Distribution of {col}')**

**plt.xlabel(col)**

**plt.ylabel('Frequency')**

**plt.grid(True)**

**plt.show()**

**# Correlation heatmap**

**print("\n🧠 Correlation Heatmap:")**

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

**sns.heatmap(data.corr(numeric\_only=True), annot=True, cmap='coolwarm')**

**plt.title('Correlation between Numeric Features')**

**plt.show()**

**# Gender-based performance**

**if 'gender' in data.columns and existing\_scores:**

**print("\n📊 Gender-based Score Comparison:")**

**for col in existing\_scores:**

**plt.figure(figsize=(6, 4))**

**sns.boxplot(x='gender', y=col, data=data)**

**plt.title(f'{col.capitalize()} by Gender')**

**plt.show()**

**# Categorization**

**if len(existing\_scores) == 3:**

**data['average score'] = data[existing\_scores].mean(axis=1)**

**data['performance'] = pd.cut(data['average score'],**

**bins=[0, 50, 70, 100],**

**labels=['Low', 'Medium', 'High'])**

**print("\n🎯 Performance Categories:")**

**display(data[['average score', 'performance']].head())**

**plt.figure(figsize=(6, 4))**

**sns.countplot(x='performance', data=data, palette='Set2')**

**plt.title('Performance Level Distribution')**

**plt.show()**

**5. Output / Result**

**The output of this project includes both visual and statistical insights from the dataset:**

* **Data Preview: Displays the top records from the dataset to understand structure.**
* **Summary and Null Check: Ensures the data is clean and ready.**
* **Histograms: Clear distribution of scores in math, reading, and writing.**
* **Correlation Heatmap: Shows strong interrelation between subject scores.**
* **Gender Boxplots: Highlights performance trends across genders.**
* **Performance Binning: Categorizes students into three clear groups: Low, Medium, and High.**
* **Graphs**
* **Heatmaps**
* **Performance classification**

**Any relevant data tables or outputs**

**6. References / Credits**

* **Dataset: Provided by Instructor or from Kaggle (Student Performance Dataset)**
* **Tools Used:**
  + **Pandas: Data analysis and manipulation**
  + **Seaborn & Matplotlib: Data visualization**
  + **Google Collab: Interactive cloud-based Python environment**
* **Acknowledgment:**
  + **Faculty of [CSE AIML]**
  + **Friends/Peers for testing and feedback**
  + **Online community forums for additional help**