Machine Learning Programming Lab 1

GitHub repository link :

Activity 1 (Data Exploration)

Q1. What is the number of rows and columns?
Ans: The dataset has 15 rows and 8 columns.

Q2. What are the data types of each column?
Ans: Some columns contain text (StudentID, Gender, ParentEducation) and some contain numbers (StudyHours, Attendance, PreviousScore, SleepHours, ExamScore).

Q3. Are there any missing values, duplicates, or anomalies?
Ans: There are missing values in StudyHours, Attendance, PreviousScore, and SleepHours, no duplicate rows, and the Gender column has values like “Male” and “female”.

Q4. What is your initial assessment of data quality?
Ans: The data is mostly good, but it needs a little cleaning before analysis.

Activity 2 (Data Cleaning)

Q1. Which columns had missing values?
Ans: The StudyHours, Attendance, SleepHours and PreviousScore columns had missing values.

Q2. How did you handle them (imputation vs removal)?
Ans: The missing values were filled using average values (imputation) instead of deleting rows, to avoid losing data.

Q3. Any duplicates or inconsistent values fixed?
Ans: There were no duplicate rows, but the Gender column had inconsistent values like “Male” and “female”, which were made consistent.

Q4. Any outliers detected and how they were handled?
Ans: 1. StudyHours

Q1 = 7.25, Q3 = 13.5 → IQR = 6.25

Lower limit = -2.12, Upper limit = 22.88

All values are inside this range.

2. Attendance

Q1 = 71.25, Q3 = 91.5 → IQR = 20.25

Lower limit = 40.88, Upper limit = 121.88

All values are inside this range.

3. SleepHours

Q1 = 5.25, Q3 = 7.75 → IQR = 2.5

Lower limit = 1.5, Upper limit = 11.5

All values are inside this range.

4. PreviousScore

Q1 ≈ 60.5, Q3 ≈ 76.5

Calculated limits include all values in the dataset.

5. ExamScore

Q1 = 67.5, Q3 = 85 → IQR = 17.5

Lower limit = 41.25, Upper limit = 111.25

All values are inside this range.

No major outliers were found, so no values were removed.

Q5. The impact of cleaning on data reliability?
Ans: After cleaning, the data became more consistent and reliable for analysis.

Activity 3 (Data Manipulation)

Q1. Give one example of filtering, aggregation, feature creation, and encoding.

Ans: 1. Filtering

We filtered students who have Attendance greater than 80% to analyze only regular students.

2. Grouping (Aggregation)

We grouped the data by Gender and calculated the average ExamScore for each group.

3. Encoding

The Gender column was converted into numbers (for example, Male = 1, Female = 0) so it can be used in calculations and models.

4. Feature Creation

A new feature called StudyEfficiency was created using:
StudyEfficiency = ExamScore / StudyHours
This helps to understand how effective a student's study time is.

Q2. Why are categorical variables encoded into numbers?
Ans: Categorical variables are changed into numbers because computers cannot properly analyze text values.

Q3. What is one new feature created and why is it useful?
Ans: A new feature called study efficiency (ExamScore divided by StudyHours) was created to better understand student performance.

Q4. How do these transformations improve data usability?
Ans: These changes make the data easier to analyze, more consistent, and more useful for getting accurate results.

Activity 4 (Data Modeling & Visualization)

Q1. Which numeric variables were correlated?
Ans: StudyHours, Attendance, and ExamScore were positively correlated with each other.

Q2. What patterns were found between StudyHours, SleepHours, Attendance, and ExamScore?
Ans: Students who studied more and had better attendance usually scored higher, while very low sleep hours sometimes reduced performance.

Q3. Which visualization was most informative and why?
Ans: The scatter plot between Study Hours and ExamScore was the most informative because it clearly showed the relationship between studying and exam results.

Q4. What insights were gained about student performance?
Ans: The data shows that study time and attendance are very important for getting better exam scores.

Activity 5 (Data Preparation for ML)

Q1. What is the purpose of the train-test split?
Ans: It is used to train the model on one part of the data and test it on new data to check how well the model works.

Q2. Why is scaling features important?
Ans: Scaling is important because some columns have bigger values and can affect the model more than others.

Q3. What is the difference between StandardScaler and MinMaxScaler?
Ans: StandardScaler scales data using the mean and standard deviation, while MinMaxScaler scales values between 0 and 1.

Q4. When should one scaler be preferred over the other?
Ans: StandardScaler is better when there are outliers, while MinMaxScaler is better when values need to stay within a fixed range.