**Task 1: Student Score Prediction**

**1. Objective**

The goal of this task is to build a machine learning model that can **predict the percentage score of students** based on the number of hours they study.  
This problem demonstrates the concept of **Simple Linear Regression** where we analyze the relationship between study hours (independent variable) and student scores (dependent variable).

**2. Dataset**

* **Source**: "Student Scores Dataset" (commonly available on Kaggle / GitHub).
* **Dataset Description**:  
  The dataset contains **two attributes**:
  + **Hours** → Number of hours a student studies per day.
  + **Scores** → Percentage marks obtained by the student.
* **Dataset Size**: 25 records (small, for demonstration purposes).

**3. Problem Type**

* **Supervised Learning**
* **Regression Problem** (continuous output prediction).
* Algorithm: **Simple Linear Regression**

**4. Approach**

**(a) Data Preprocessing**

1. Load the dataset using Pandas.
2. Check for missing values (dataset is usually clean).
3. Visualize the data using scatter plots to see the relationship between study hours and scores.

**(b) Model Building**

* Split dataset into **Training (80%)** and **Testing (20%)** sets.
* Train a **Linear Regression** model using scikit-learn.
* The regression line equation:

y=m⋅x+cy = m \cdot x + cy=m⋅x+c

where:

* + yyy = predicted score
  + xxx = study hours
  + mmm = slope
  + ccc = intercept

**(c) Model Evaluation**

* Metrics:
  + **Mean Absolute Error (MAE)**
  + **Mean Squared Error (MSE)**
  + **Root Mean Squared Error (RMSE)**
  + **R² Score (Coefficient of Determination)**

**5. Tools & Libraries**

* **Python**
* **Pandas, NumPy** → Data handling
* **Matplotlib, Seaborn** → Visualization
* **Scikit-learn** → Model training and evaluation

**6. Workflow**

1. Import libraries and load dataset.
2. Explore dataset (shape, head, missing values).
3. Visualize Hours vs Scores scatter plot.
4. Split data into training and testing sets.
5. Train a Linear Regression model.
6. Plot the regression line on the data.
7. Make predictions on the test set.
8. Evaluate model performance.

**7. Expected Results**

* A positive linear relationship between study hours and student scores.
* Regression line fitting the data with high accuracy.
* R² Score usually **> 0.95** (excellent fit).
* Example prediction:
  + If a student studies **9.25 hours/day**, the model predicts a score of **~93%**.

**8. Applications**

* Educational institutions can estimate student performance based on study patterns.
* Helps students plan study schedules effectively.
* Can be extended with additional factors like sleep, attendance, or participation to predict performance more accurately.

