Dataset:

https://www.kaggle.com/datasets/spscientist/students-performance-in-exams

TASK 0 – Dataset Preparation

1. Download and unzip the dataset.
2. Load the StudentsPerformance.csv file using pandas.
3. Print the following:
    - The shape of the dataset
    - The column names
    - The first 5 rows of the dataset
    - Describe
    - Info
4. Check and report the number of missing values per column.

TASK 1 – Exploratory Data Analysis (EDA)

Goal: Understand the dataset structure, target behavior, and feature relationships.

1. Select **math score** as the target variable.
2. Analyze the distribution of the target variable:
    - Plot the distribution of math score.
    - Identify whether outliers are present.
3. Examine the relationship between:
    - math score and reading score
    - math score and writing score
4. Analyze how categorical features (gender, lunch, test preparation course) affect math score.
5. Identify and justify which features appear most informative for predicting the target.

TASK 2 – Feature Engineering & Preprocessing

Goal: Prepare the data for modeling.

1. Separate numerical and categorical features and into X and y.
2. For numerical features:
    - Choose an appropriate missing value handling strategy.
    - Decide whether feature scaling is required and justify your choice.
3. For categorical features:
    - Select an appropriate encoding technique.
    - Explain why this encoding method is suitable.
4. Combine all preprocessing steps into a single pipeline.

TASK 3 – Train-Test Split

Goal: Ensure reliable model evaluation.

1. Split the dataset into training and testing sets using an 80/20 ratio.
2. Use random_state = 42.
3. Create X_train, X_test, y_train, and y_test.
4. Explain what data leakage is and how it can be avoided at this stage.

TASK 4 – Regression Model

Goal: Build a regression model to predict a continuous target.

1. Train a Linear Regression model using the prepared data.
2. Ensure preprocessing and the regression model are combined into a single pipeline.
3. Fit the model using the training data.
4. Generate predictions on the test data.

TASK 5 – Model Evaluation

Goal: Evaluate model performance.

1. Compute the following evaluation metrics:
    - Mean Absolute Error (MAE)
    - Root Mean Squared Error (RMSE)
    - R^2 score
2. Discuss which metric is most appropriate for this problem and why.
3. Determine whether the model is underfitting or overfitting and justify your answer.

TASK 6 – Regularization

1. Train a Ridge Regression model.
2. Train a Lasso Regression model.
3. Compare Linear Regression, Ridge, and Lasso models.
4. Discuss which model is more stable and why.

TASK 7 – Cross-Validation

1. Apply 5-fold cross-validation to the model.
2. Compare cross-validation results with the test set results.
3. Explain why cross-validation provides a more reliable performance estimate.

TASK 8 – Learning Curve

1. Generate a learning curve for the model.
2. Analyze the behavior of training and validation errors.
3. Discuss whether adding more data is likely to improve performance.
4. Suggest the next steps to improve the model.