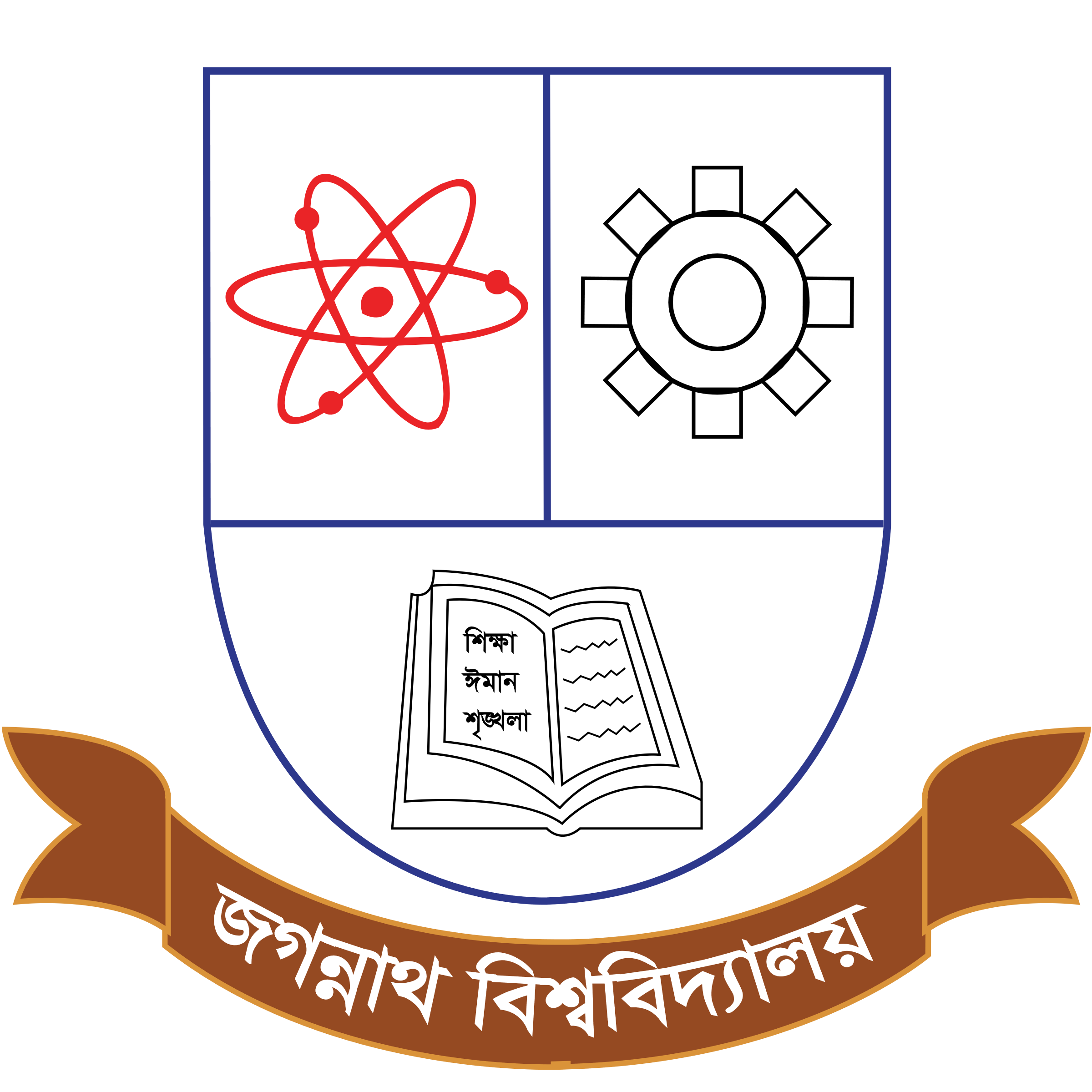
****

**Data Science Lab**

CSEL-42--

|  |
| --- |
| Assignment on Linear Regression |

|  |  |  |
| --- | --- | --- |
| Submitted by | | |
|  |  |  |
| **NISHAT MAHMUD**  ID: B190305003 | **FAHIM HASAN**  ID: B190305029 | **MD. WALIUL ISLAM RAYHAN**  ID: B190305034 |

Submitted To

**MD. MANOWARUL ISLAM, PHD**

Associate Professor

Dept. of CSE

Jagannath University, Dhaka - 1100

Department of Computer Science and Engineering, Jagannath University, Dhaka

Task 1

**Exploring Key Parameters and Model Evaluation**

The task 1 of the assignment performs linear regression on a synthetic dataset to model the relationship between an independent variable 𝑋 and a dependent variable 𝑌. The process includes data visualization, implementation of a linear regression model, evaluation using metrics, and analysis of residuals. The steps are described below:

**1. Dataset Generation**

The dataset is synthetically generated to simulate a linear relationship between 𝑋 (independent variable) and 𝑌 (dependent variable). Noise is added to 𝑌 to mimic real-world variability.

A screen shot of a computer

Description automatically generated

**2. Data Visualization**

The scatter plot provides a visual representation of the relationship between 𝑋 and 𝑌. This helps in understanding the spread and potential correlation between the two variables.

A computer screen shot of a computer code

Description automatically generated

A graph with blue dots

Description automatically generated

**3. Linear Regression Implementation**

The linear regression model is implemented using the LinearRegression class from the scikit-learn library. The model is trained on the dataset, and the slope (𝑚) and intercept (𝑐) of the regression line are extracted.

A white background with black and white clouds

Description automatically generated with medium confidence

**4. Coefficient of Determination (𝑅2)**

The coefficient of determination (𝑅2) is calculated to evaluate the goodness-of-fit of the regression model. It indicates the proportion of variance in 𝑌 that can be explained by 𝑋.

A white background with black and white clouds

Description automatically generated with medium confidence

**5. Predictions and Visualization**

The regression line is overlaid on the scatter plot of the original data to visualize the linear fit of the model.

A computer code with text

Description automatically generated with medium confidence

A graph with a red line

Description automatically generated

**6. Residual Analysis**

Residuals (the differences between actual and predicted 𝑌 values) are plotted to check for patterns. A random distribution of residuals around zero suggests that the model assumptions are valid.

A computer code with text

Description automatically generated with medium confidence

A graph with blue dots and red line

Description automatically generated

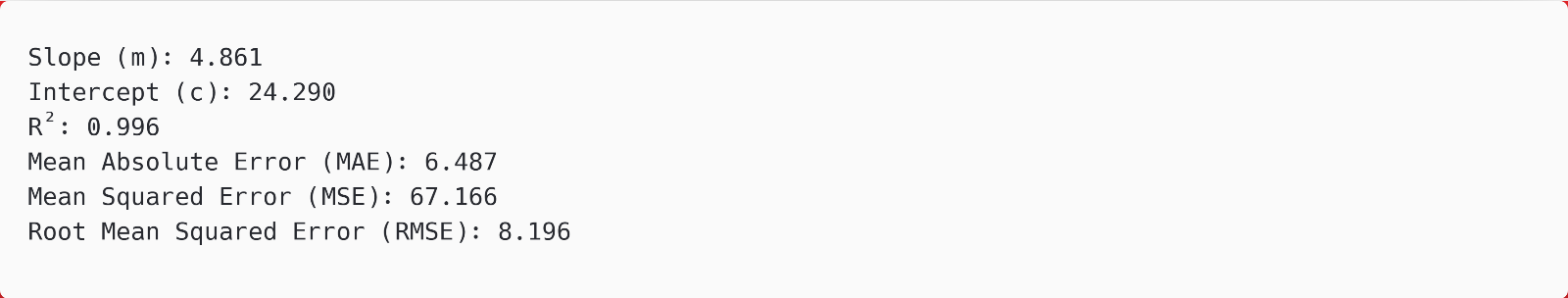
**7. Model Evaluation Metrics**

Three key metrics are calculated to evaluate the performance of the model:

* **Mean Absolute Error (MAE):** The average of absolute errors between actual and predicted values.
* **Mean Squared Error (MSE):** The average of squared errors.
* **Root Mean Squared Error (RMSE):** The square root of the MSE, providing a measure of error in the same units as the dependent variable.

A computer screen shot of a computer code

Description automatically generated



The linear regression model successfully fits the synthetic data, as evidenced by the high 𝑅2 value and low error metrics. The residual analysis confirms that the model assumptions hold true, with residuals distributed randomly around zero.

Task 2

**Gradient Descent**

The task 2 of the assignment implements a linear regression model using gradient descent to optimize the parameters (𝜃0, 𝜃1). The process includes dataset preparation, gradient descent implementation, cost function computation, visualization of results, and model evaluation. Each step is explained in detail below:

**1. Dataset Preparation**

A synthetic dataset is generated to simulate a linear relationship between 𝑋 (independent variable) and 𝑌 (dependent variable). Noise is added to 𝑌 to mimic real-world scenarios.

A computer screen shot of a computer

Description automatically generated

**2. Data Visualization**

A scatter plot is created to visualize the relationship between 𝑋 and 𝑌, providing insights into the data distribution.

A computer screen shot of a computer code

Description automatically generated

A graph with blue dots

Description automatically generated

**3. Cost Function Definition**

The Mean Squared Error (MSE) is defined as the cost function to measure the performance of the regression model. It quantifies the difference between actual and predicted values.

A screenshot of a computer code

Description automatically generated

**4. Gradient Descent Implementation**

The gradient descent algorithm is implemented to iteratively update the parameters 𝜃0 and 𝜃1 by minimizing the cost function.

A screenshot of a computer code

Description automatically generated

**5. Perform Gradient Descent**

The gradient descent function is called with the initial parameters, learning rate, and number of iterations. This optimizes the parameters 𝜃0 and 𝜃1.

A screenshot of a computer

Description automatically generated

**6. Predictions and Regression Line Visualization**

The optimized parameters are used to predict 𝑌 values, and the regression line is plotted over the original scatter plot to visualize the fit.

A computer screen shot of a code

Description automatically generated

A graph with a red line

Description automatically generated

**7. Model Evaluation**

The model is evaluated using performance metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Coefficient of Determination (𝑅2).

A computer code with text

Description automatically generated with medium confidence

A close-up of numbers

Description automatically generated

The gradient descent algorithm successfully optimized the parameters 𝜃0 and 𝜃1, resulting in a well-fitted regression model. The low error metrics and high 𝑅2 value indicate a strong linear relationship between 𝑋 and 𝑌.