Documentation for 1-A

Question -1

Assignment 1-A : Regression without regularization

Problem Statement

- The given dataset is a synthetic dataset consisting of 1,000 data points, each having one feature variable and one continuous target variable.

Group Members

Group Member	ID
Teerth Vasant Patel	2021A7PS2090H
Manthan Patel	2021A7PS2691H
Shrey C Paunwala	2021A7PS2808H

Documentation for 1-A

Question 1-A

Task 1: Data Preprocessing

a. Load the Shared Dataset into a Pandas DataFrame
Description:

- Utilize Pandas to import the dataset, storing it in a DataFrame called "data."

b. Normalise the feature variable

- Normalize the feature variables and update the "data" DataFrame accordingly.

c. Shuffle the Dataset and Split into Training and Testing Sets
Description:

- The dataset undergoes shuffling to introduce randomness and ensure reproducibility.
- It is subsequently partitioned into training and testing subsets, following an 80-20 ratio, while maintaining a fixed random seed of 42 to ensure consistency.
- *The resulting training dataset is denoted as "train_data," and the testing dataset is denoted as "test_data."*

Task 2: Polynomial Regression

a. Build Polynomial Regression Models
Description:

- Within this task, we construct polynomial regression models encompassing degrees spanning from 1 to 9. The primary objective is to make predictions of the target variable based on the input feature variable. The overarching aim is to ascertain the optimal polynomial degree that effectively captures the underlying data patterns.
- To achieve this, we employ the "polynomial_features" function to perform the transformation of input features into polynomial features, adhering to the designated degree.

b. Apply Batch Gradient Descent
Description:

Documentation for 1-A

- Within this task, our approach involves the application of batch gradient descent for the training of polynomial regression models. We rely on the "batch_gradient_descent" function to efficiently fine-tune the model's parameters (coefficients) and minimize the associated cost function.

- Notably, the "batch_gradient_descent" function accepts input features X, the target variable y, initial parameters denoted as theta, the learning rate denoted as α, and the number of iterations as its input parameters.

- Following the execution of batch gradient descent, the resultant theta values represent the optimized parameters for the polynomial regression model.

c. Train Each Model for 500 Iterations
Description:

- In this task, we train each polynomial regression model through 500 iterations of batch gradient descent. Our objective is to identify the optimal learning rate and its corresponding polynomial degree for achieving superior performance in polynomial regression. This determination is based on the mean squared error (MSE) computed on the test data.

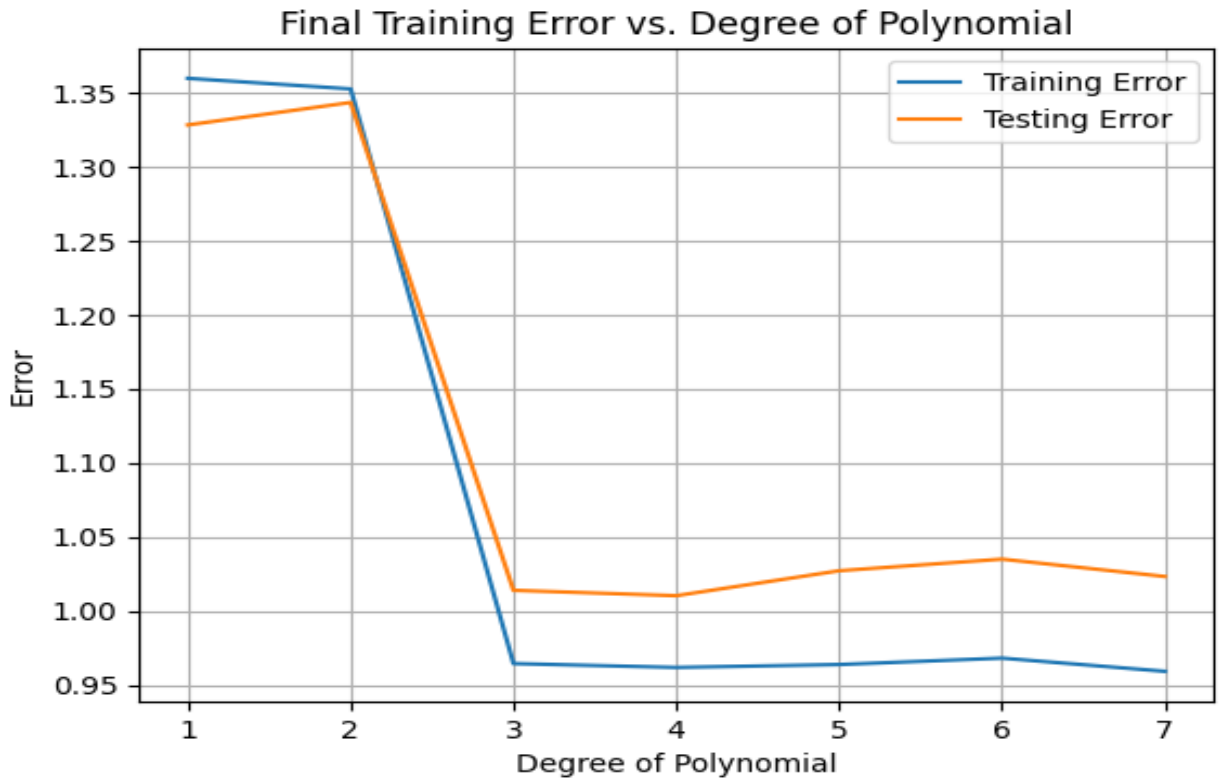
- The learning rate is a pivotal hyperparameter within gradient descent algorithms, and extensive experimentation and analysis have established 0.01 as the most effective value.

- For each polynomial degree within the range of 1 to 9, we train the polynomial regression model utilizing the aforementioned optimal learning rate of 0.01.

Documentation for 1-A

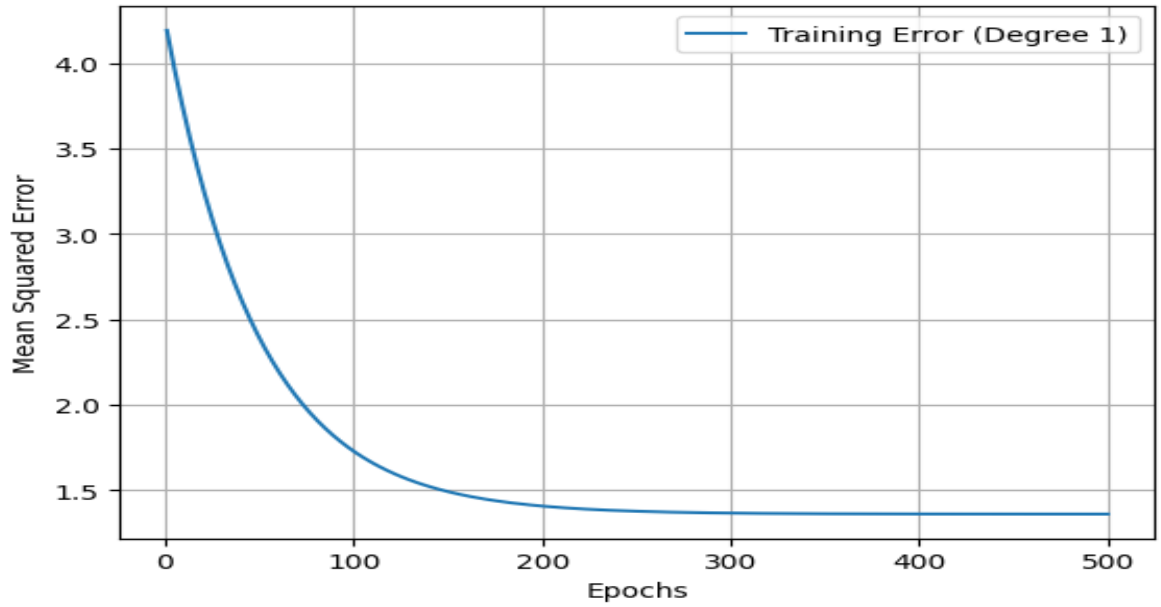
Task 3: Graph Plotting

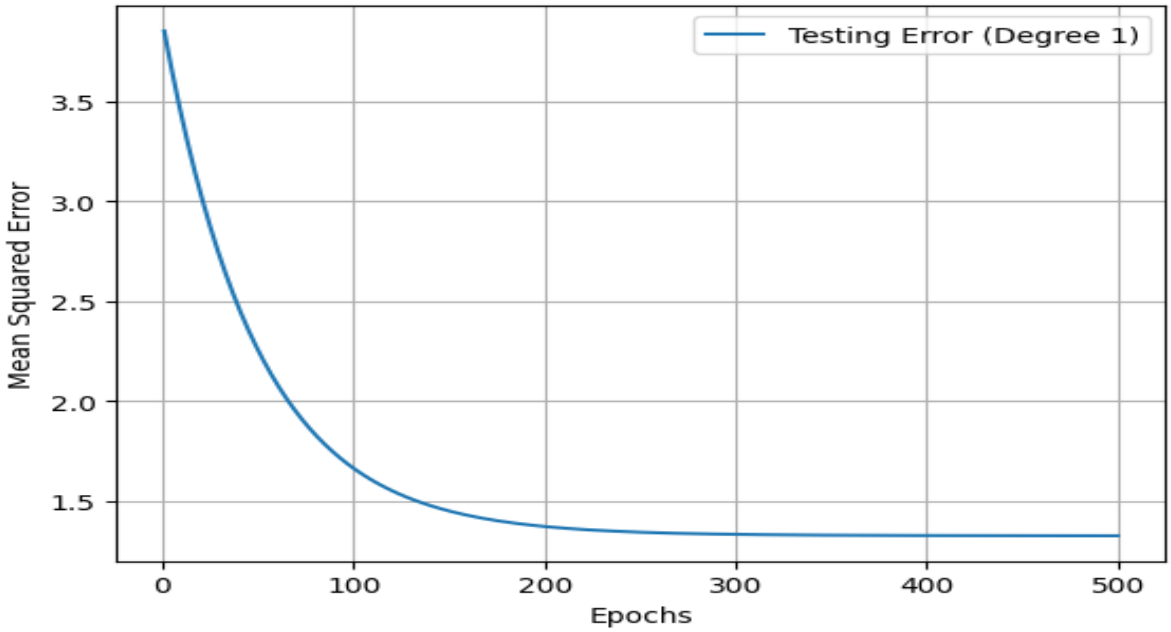
a. Plot 1 - Final Training and Testing Errors vs. Degree of Polynomial



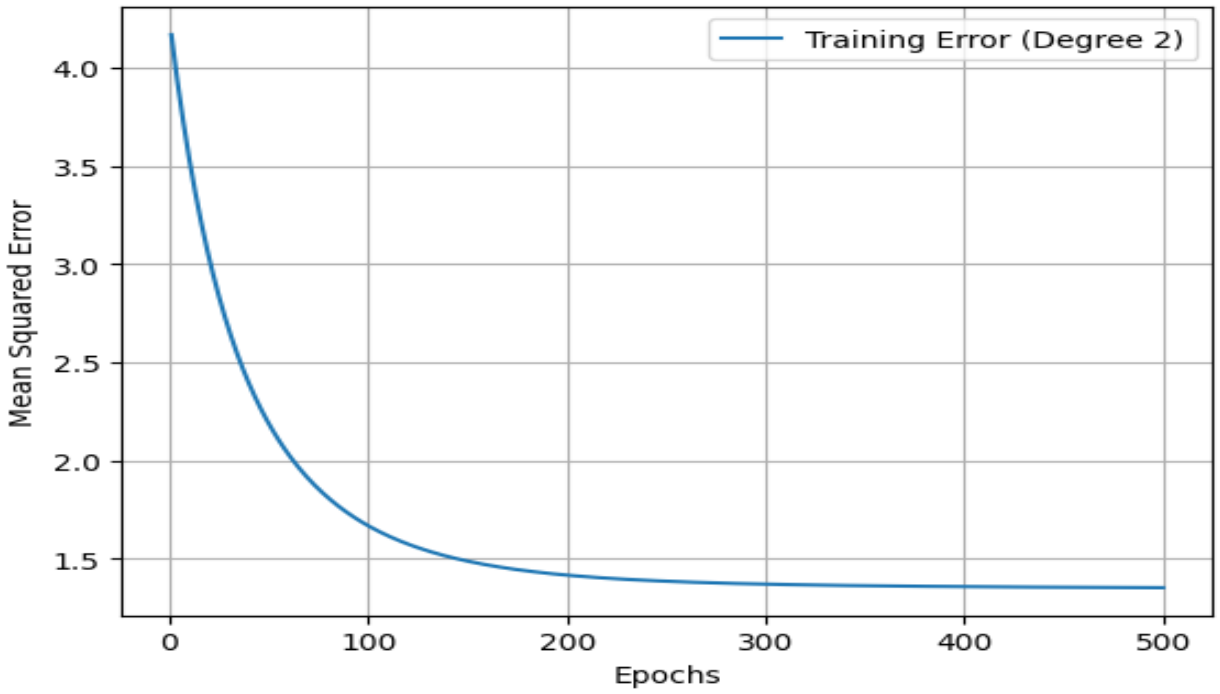
Documentation for 1-A

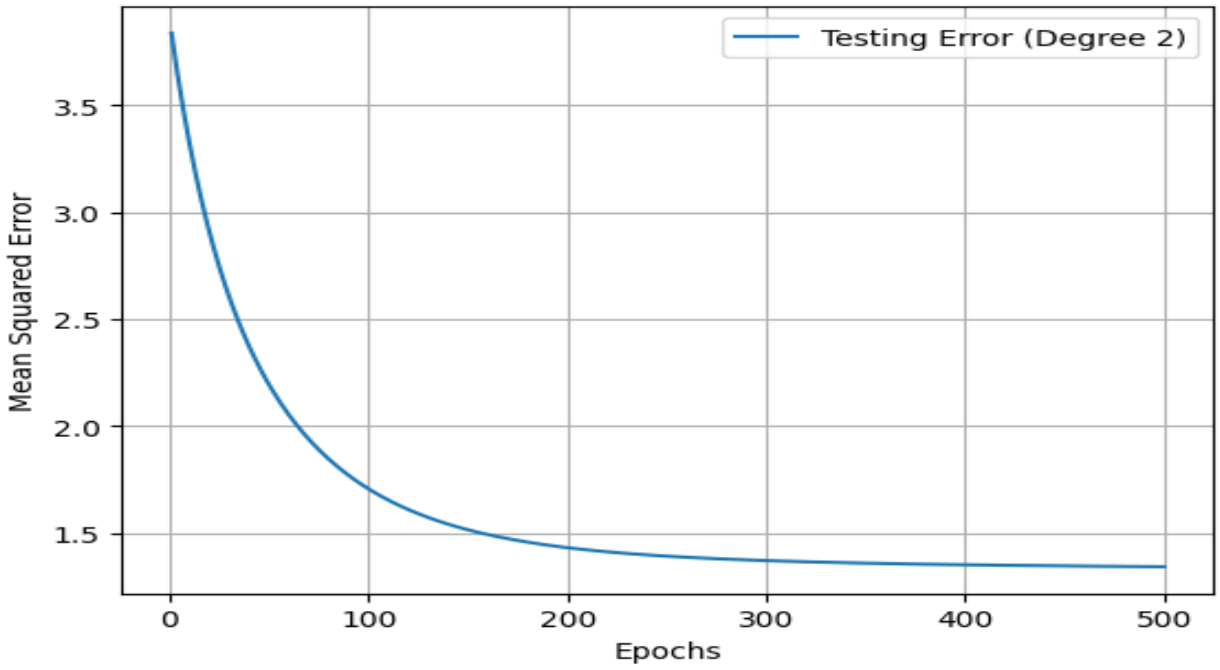
b. Training Error and Testing Error vs. Epochs for All Degrees of Polynomial



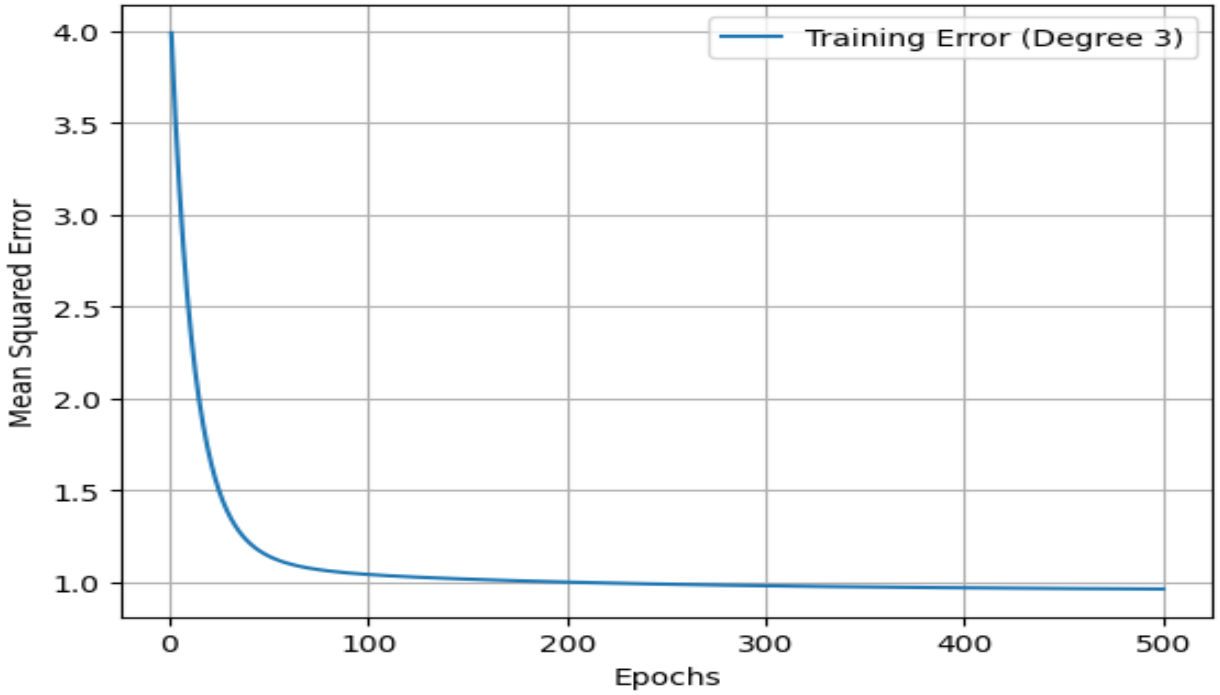


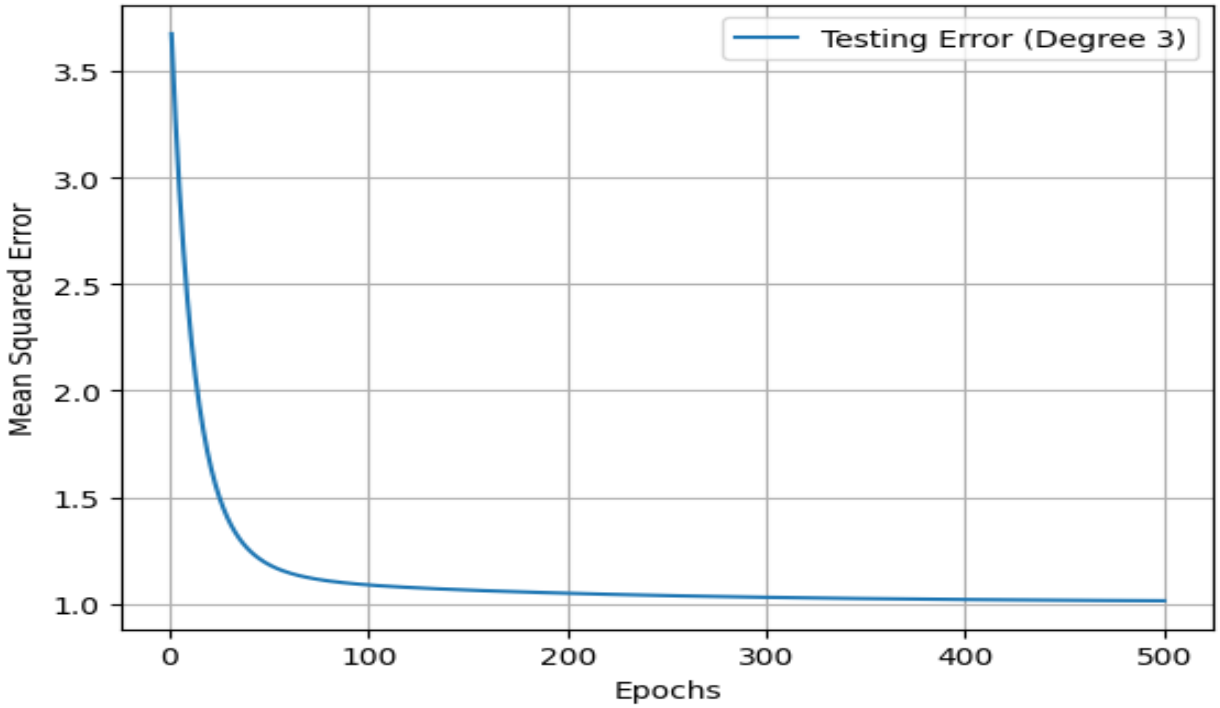
Documentation for 1-A



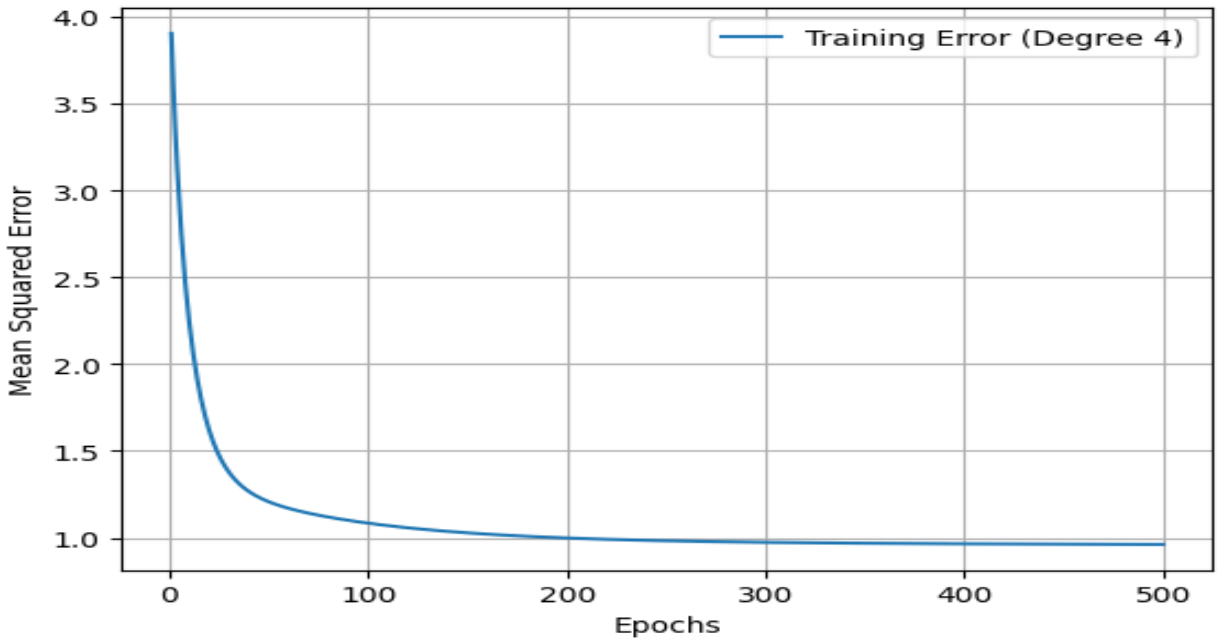


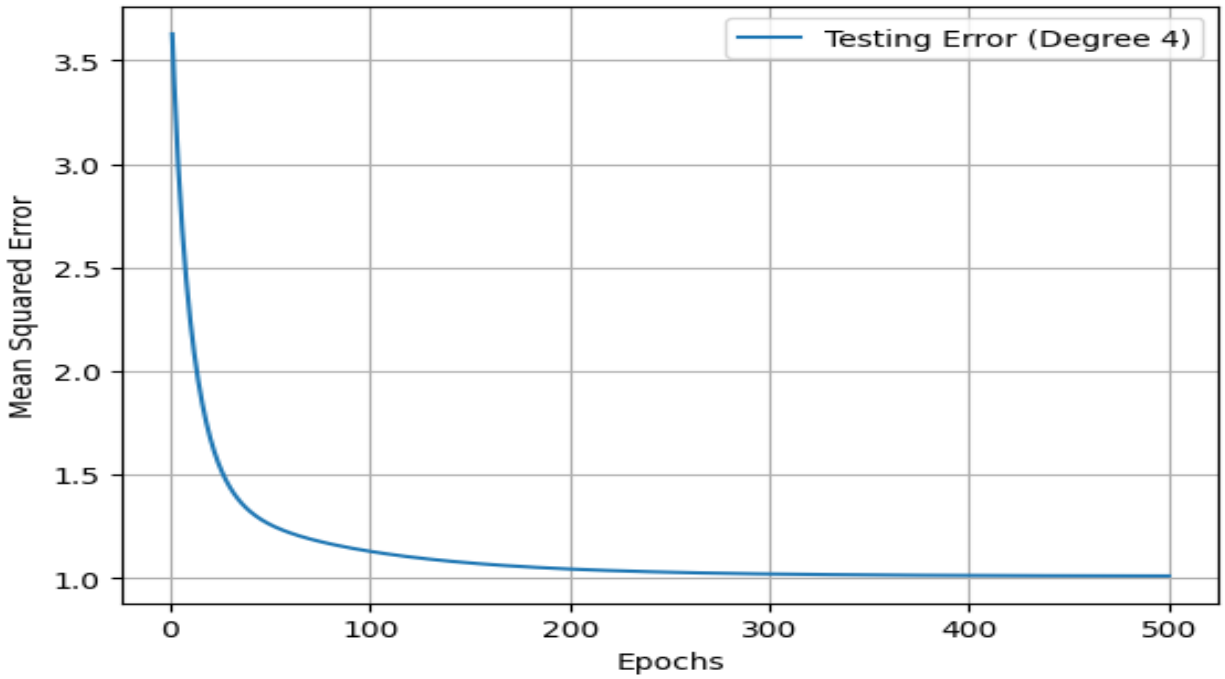
Documentation for 1-A



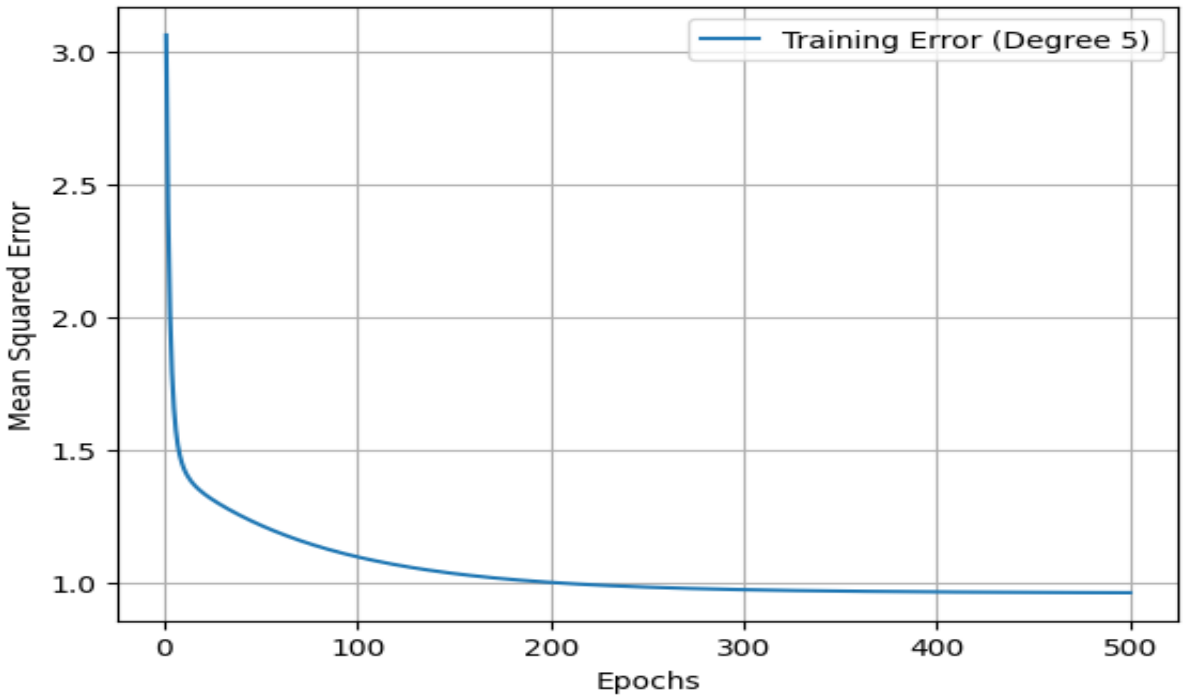


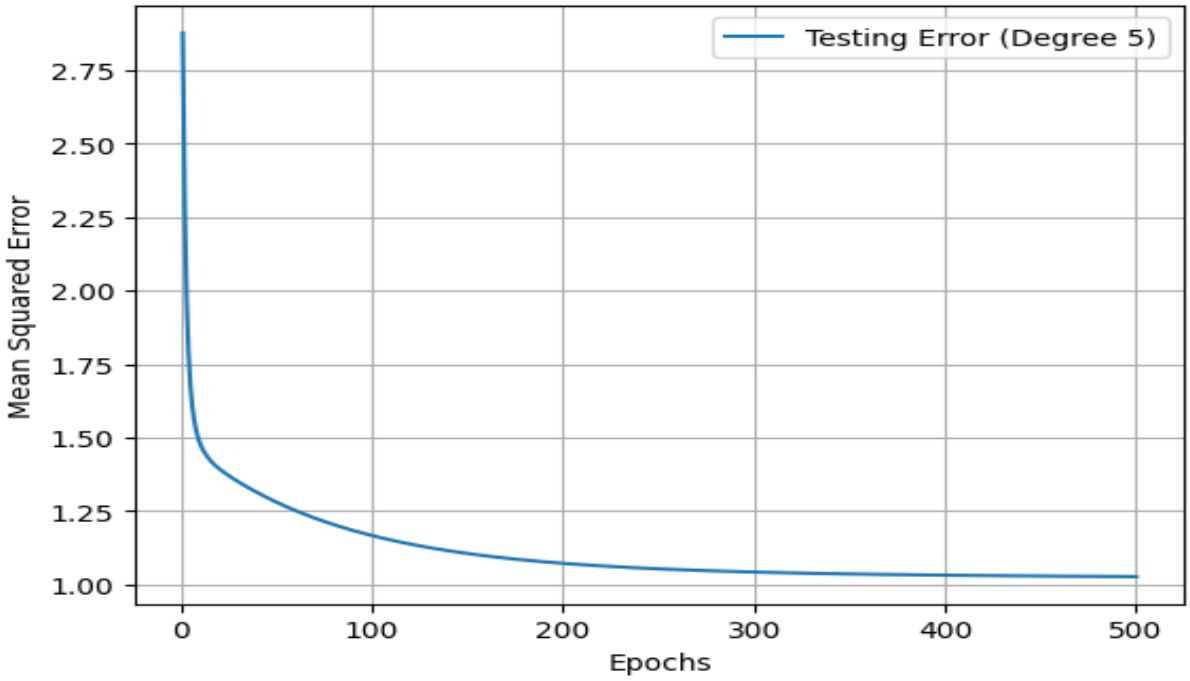
Documentation for 1-A



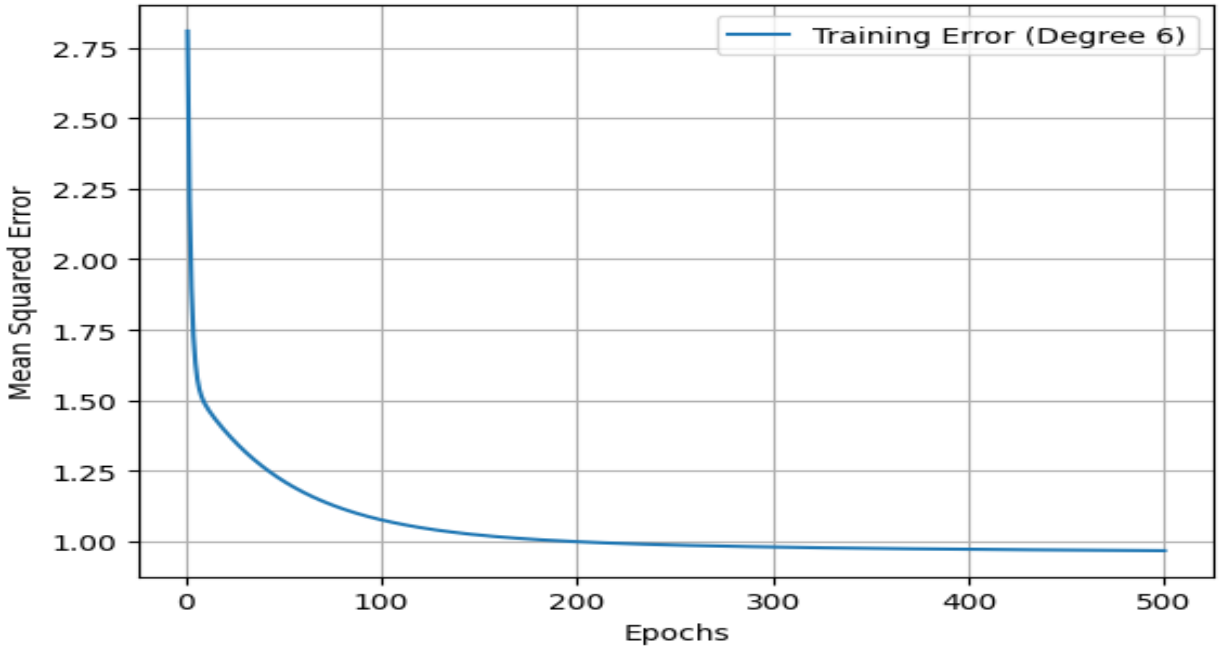


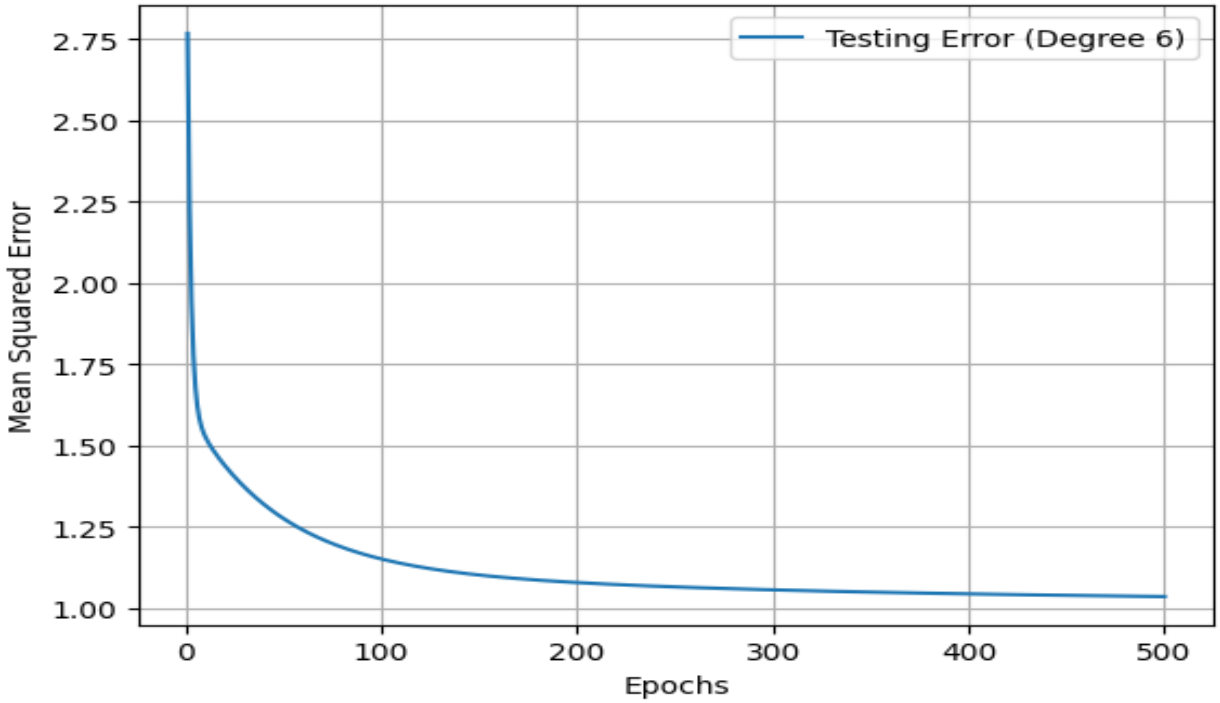
Documentation for 1-A



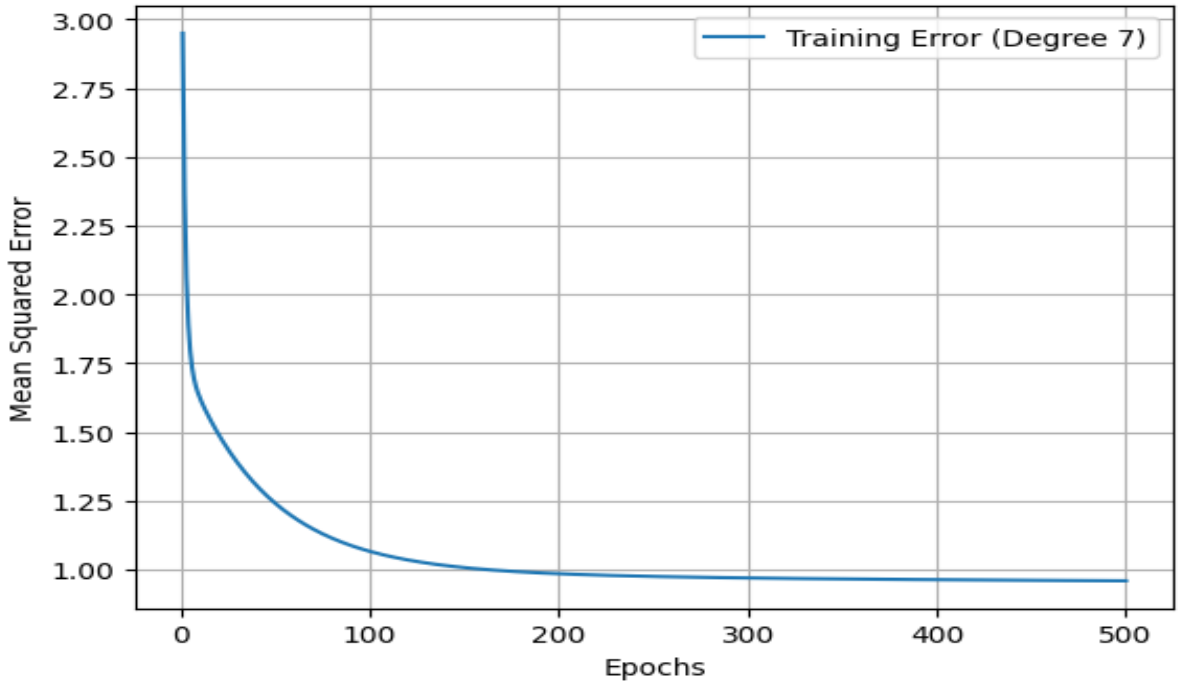


Documentation for 1-A

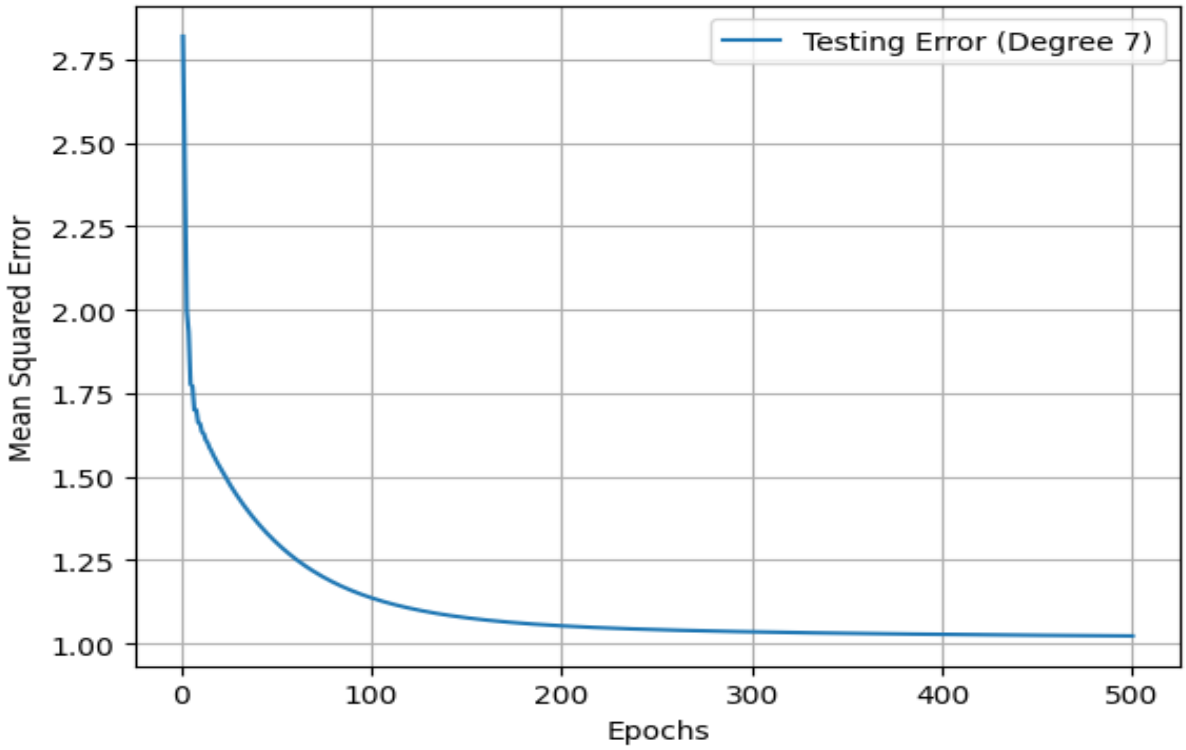




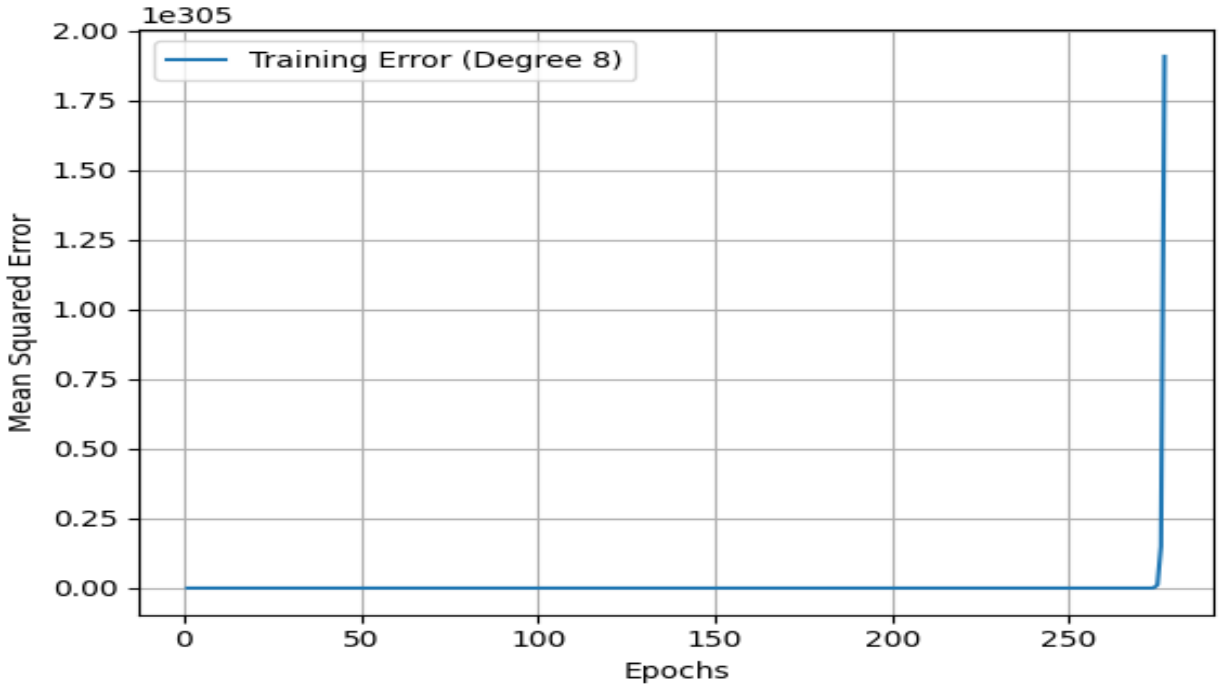
Documentation for 1-A

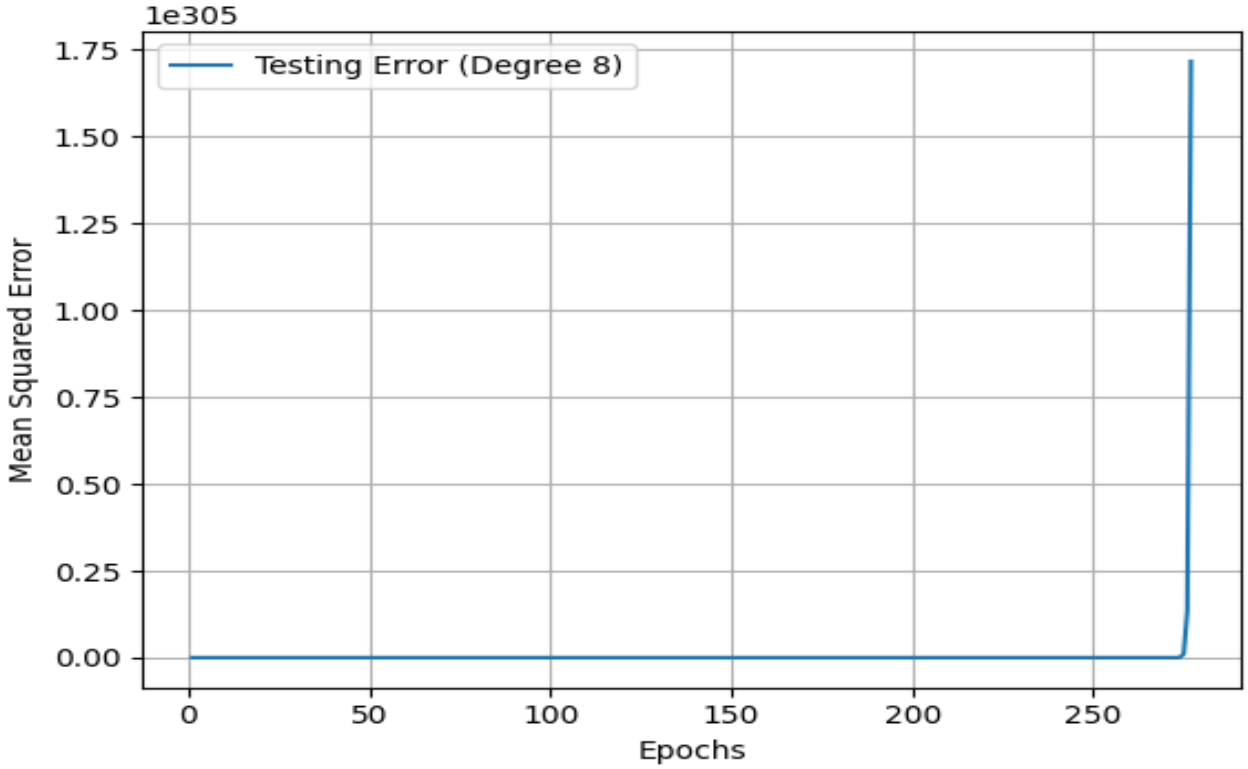


=

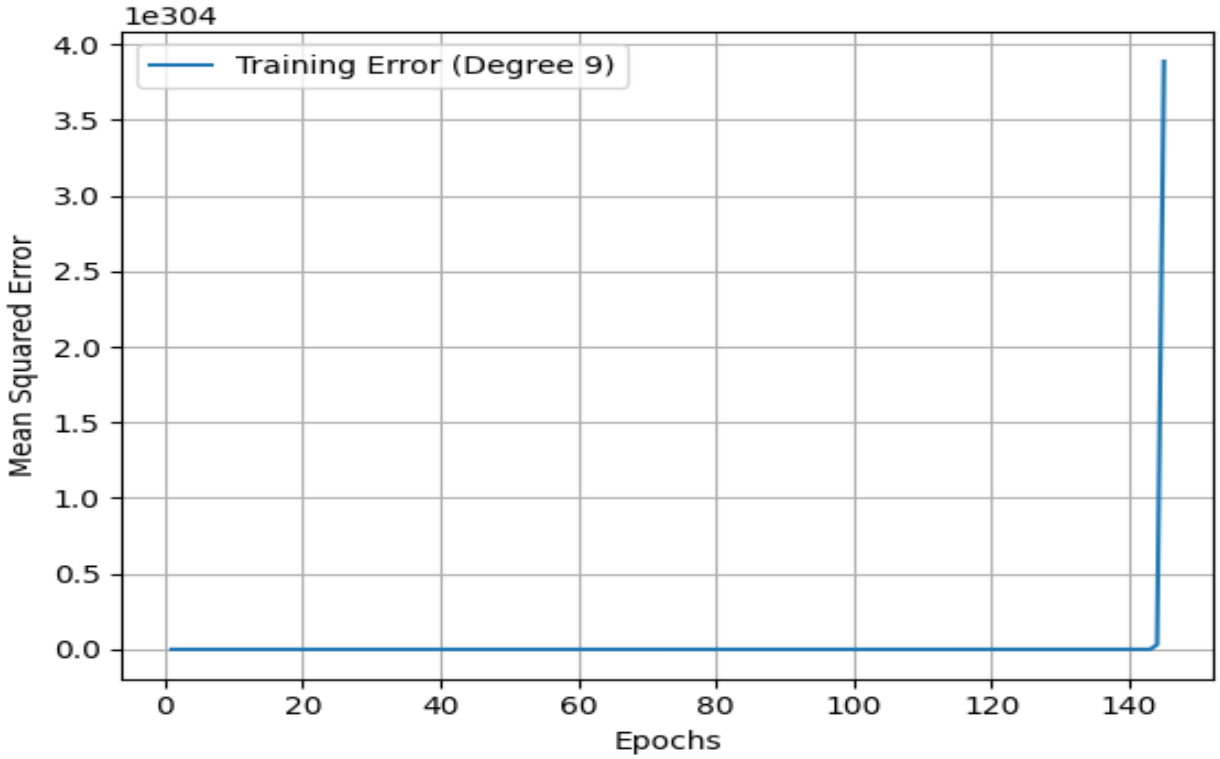


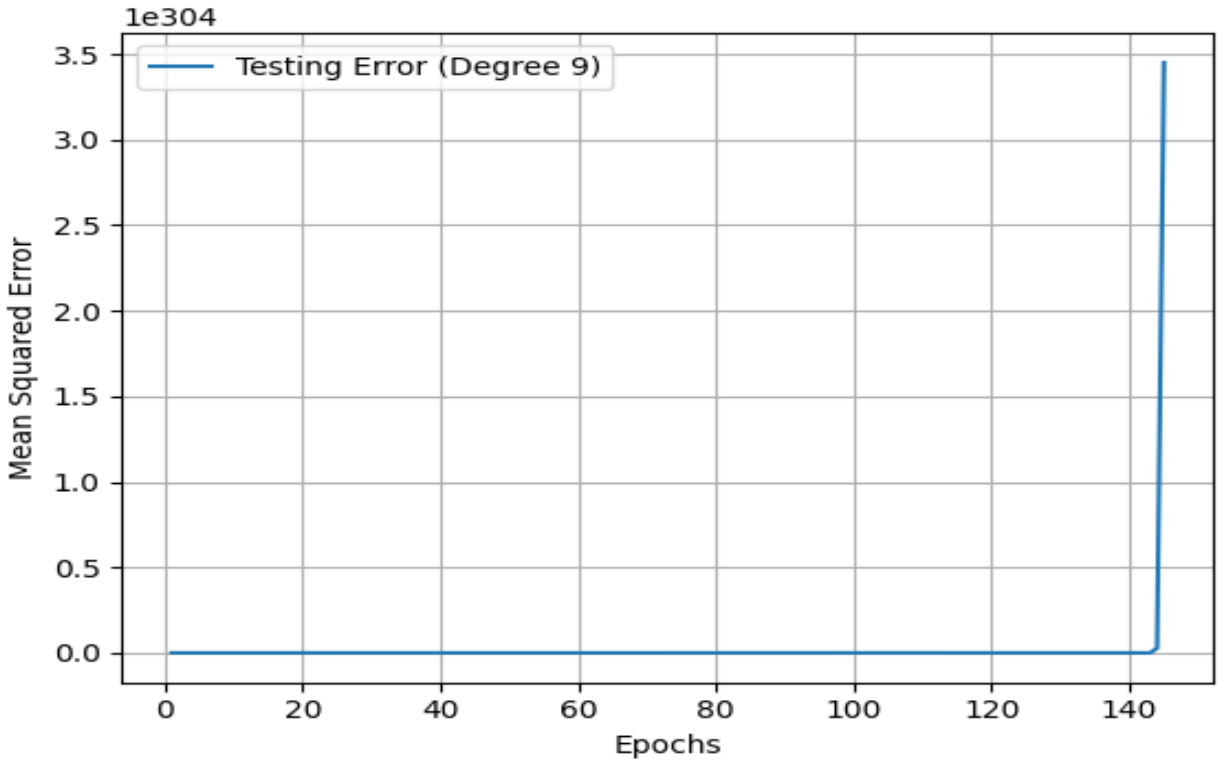
Documentation for 1-A





Documentation for 1-A





Documentation for 1-A

Task 4: Comparative Analysis

Best Learning Rate: 0.01
Best Degree: 4