ERRNUM NISHAT

DIGITAL EMPOWERMENT PAKISTAN FOUNDATION

DEP

Machine Learning

Task 1

Predicting House Prices:

- **Objective:** Build a model to predict house prices based on various features.
- **Description:** Use a dataset containing information about houses (e.g., size, number of bedrooms, location) to create a predictive model that estimates the price of a house.
- **Key Steps:**
- Data cleaning and preprocessing
- Feature engineering
- Model selection and training (e.g., linear regression)
- Model evaluation and fine-tuning

Predicting House Prices Using MATLAB

Objective:

Build a model to predict house prices based on various features.

1. **Introduction**
   - Objective and Description of the project.
2. **Data Generation**
   - Explanation of synthetic data generation.
   - Table of features and their ranges.
3. **Data Preparation**
   - Splitting data into training and testing sets.
   - Encoding categorical variables.
   - Standardizing numerical features.
4. **Model Training**
   - Training a linear regression model.
5. **Model Evaluation**
   - Making predictions.
   - Calculating RMSE.
6. **Visualization**
   - Actual vs Predicted Prices scatter plot.
   - Histograms of actual and predicted prices.
   - Residuals plot.
   - Learning curve.
   - Cross-validation RMSE bar plot.

Description:

Use a dataset containing information about houses (e.g., size, number of bedrooms, location) to create a predictive model that estimates the price of a house.

Key Steps:

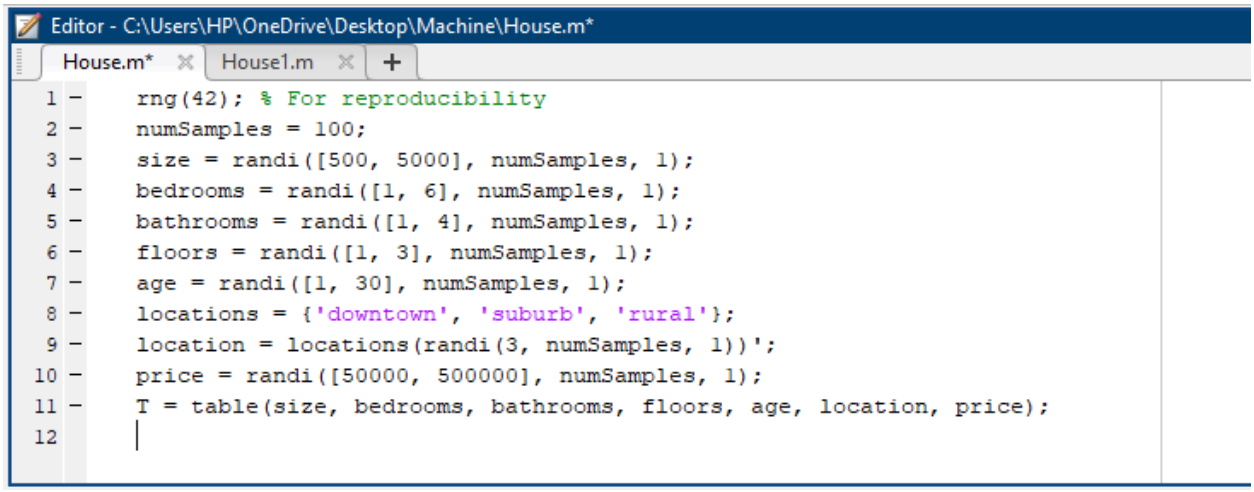
1. Data Cleaning and Preprocessing
2. Feature Engineering
3. Model Selection and Training (e.g., Linear Regression)
4. Model Evaluation and Fine-tuning

1. Data Generation

Synthetic data is generated for the following features:

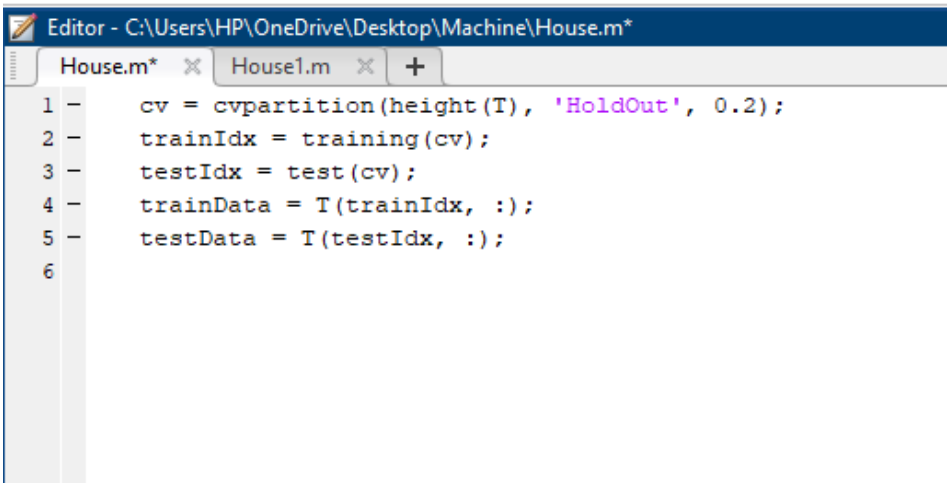
- **Size:** Size of the house in square feet.
- **Bedrooms:** Number of bedrooms.
- **Bathrooms:** Number of bathrooms.
- **Floors:** Number of floors.

- **Age:** Age of the house in years.
- **Location:** Categorical variable with values ('downtown', 'suburb', 'rural').
- **Price:** Target variable representing the price of the house.

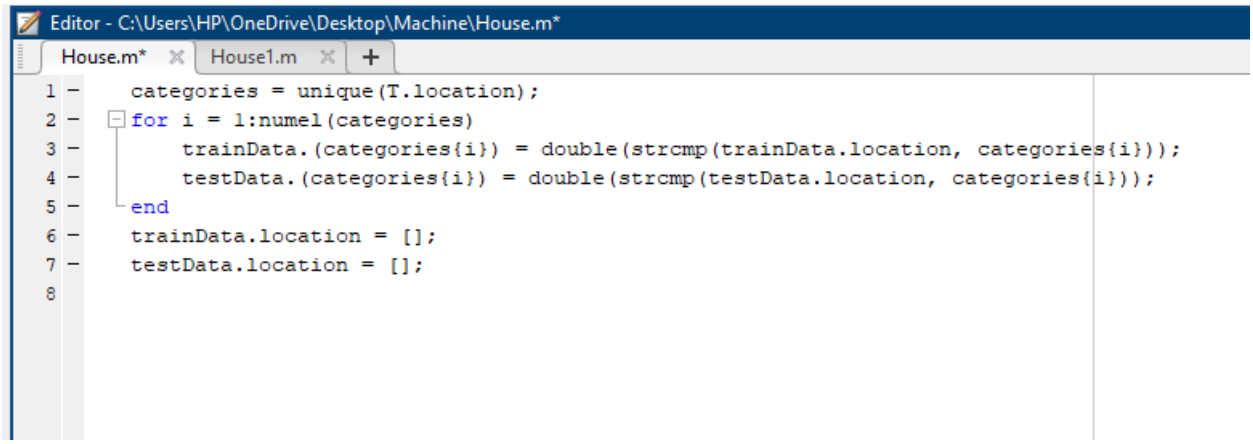


2. Data Preparation

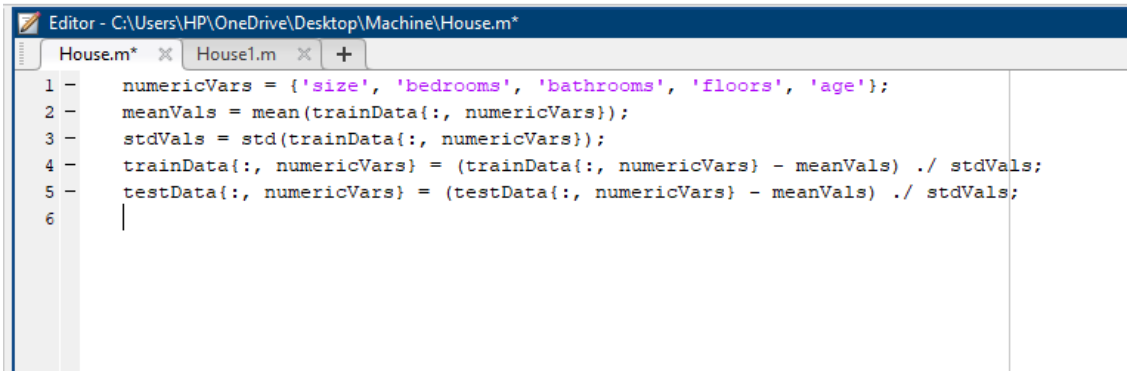
Splitting Data into Training and Testing Sets:



Encoding Categorical Variables:

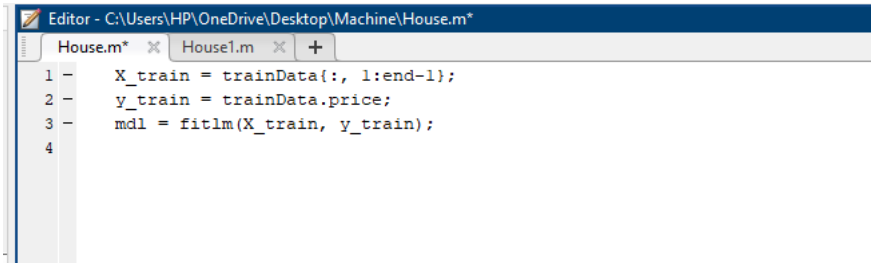


Standardizing Numerical Features:



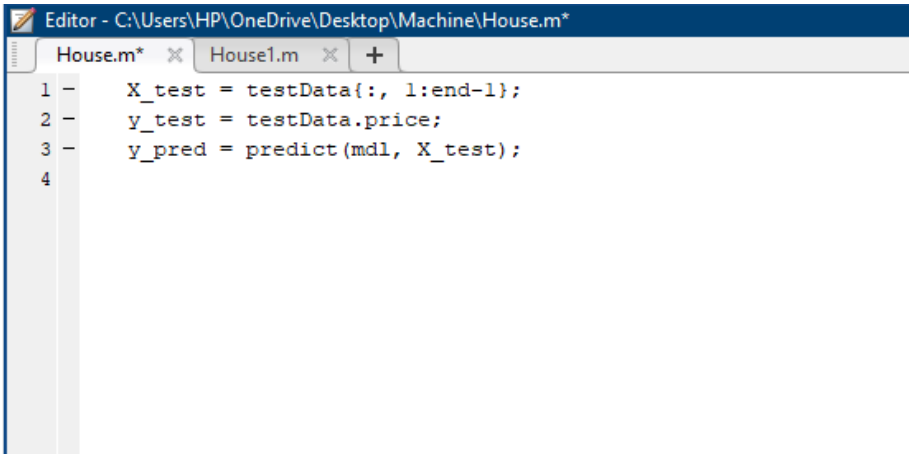
3. Model Training

Training a Linear Regression Model:

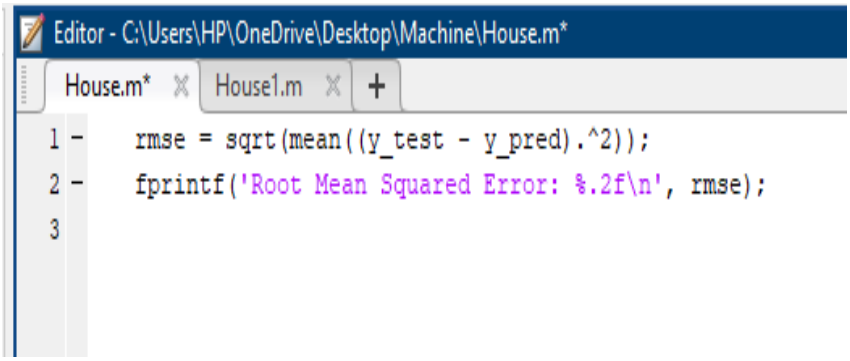


4. Model Evaluation

Making Predictions:



Calculating Root Mean Squared Error (RMSE):



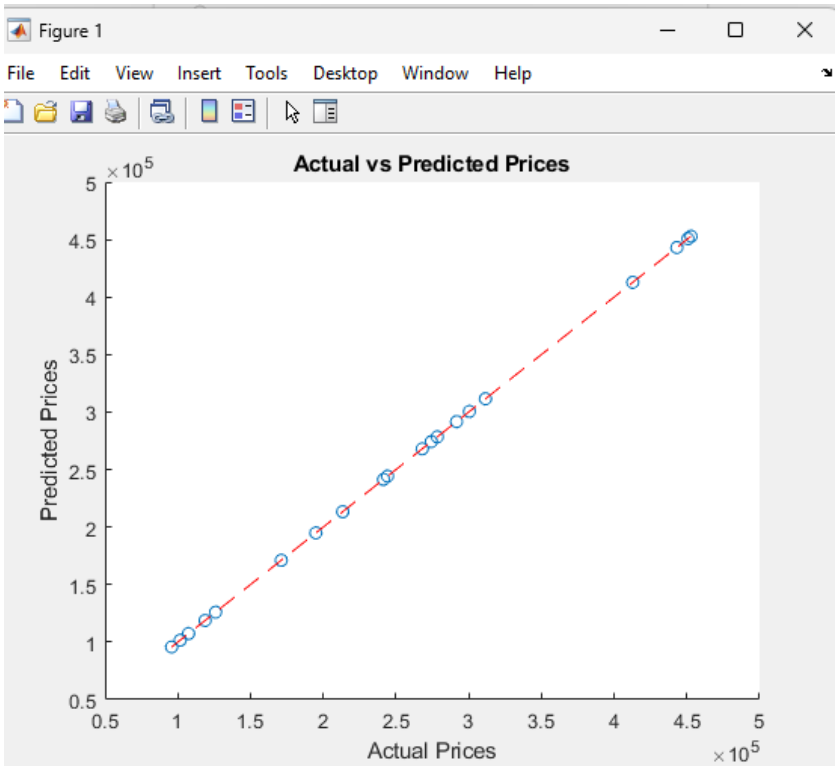
5. Visualization

Actual vs Predicted Prices:

Editor - C:\Users\HP\OneDrive\Desktop\Machine\House.m*

House.m* | House1.m

```
figure;
scatter(y_test, y_pred);
hold on;
plot([min(y_test) max(y_test)], [min(y_test) max(y_test)], 'r--');
xlabel('Actual Prices');
ylabel('Predicted Prices');
title('Actual vs Predicted Prices');
hold off;
```



Histogram of Actual Prices:

Editor - C:\Users\HP\OneDrive\Desktop\Machine\House.m*

House.m* | House1.m

```
figure;
histogram(y_test, 'BinWidth', 25000);
xlabel('Actual Prices');
ylabel('Frequency');
title('Histogram of Actual Prices');
```

Histogram of Predicted Prices:

Editor - C:\Users\HP\OneDrive\Desktop\Machine\House1.m*

House.m | House1.m*

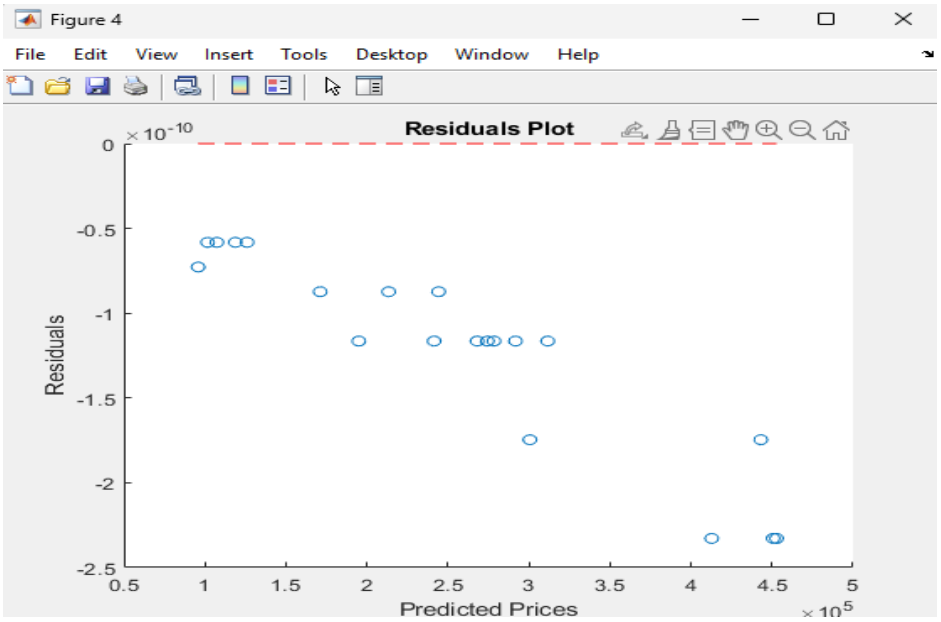
```
1 -    figure;
2 -    histogram(y_pred, 'BinWidth', 25000);
3 -    xlabel('Predicted Prices');
4 -    ylabel('Frequency');
5 -    title('Histogram of Predicted Prices');
6
```

Residuals Plot:

Editor - C:\Users\HP\OneDrive\Desktop\Machine\House1.m*

House.m | House1.m*

```
residuals = y_test - y_pred;
figure;
scatter(y_pred, residuals);
hold on;
plot([min(y_pred) max(y_pred)], [0 0], 'r--');
xlabel('Predicted Prices');
ylabel('Residuals');
title('Residuals Plot');
hold off;
```

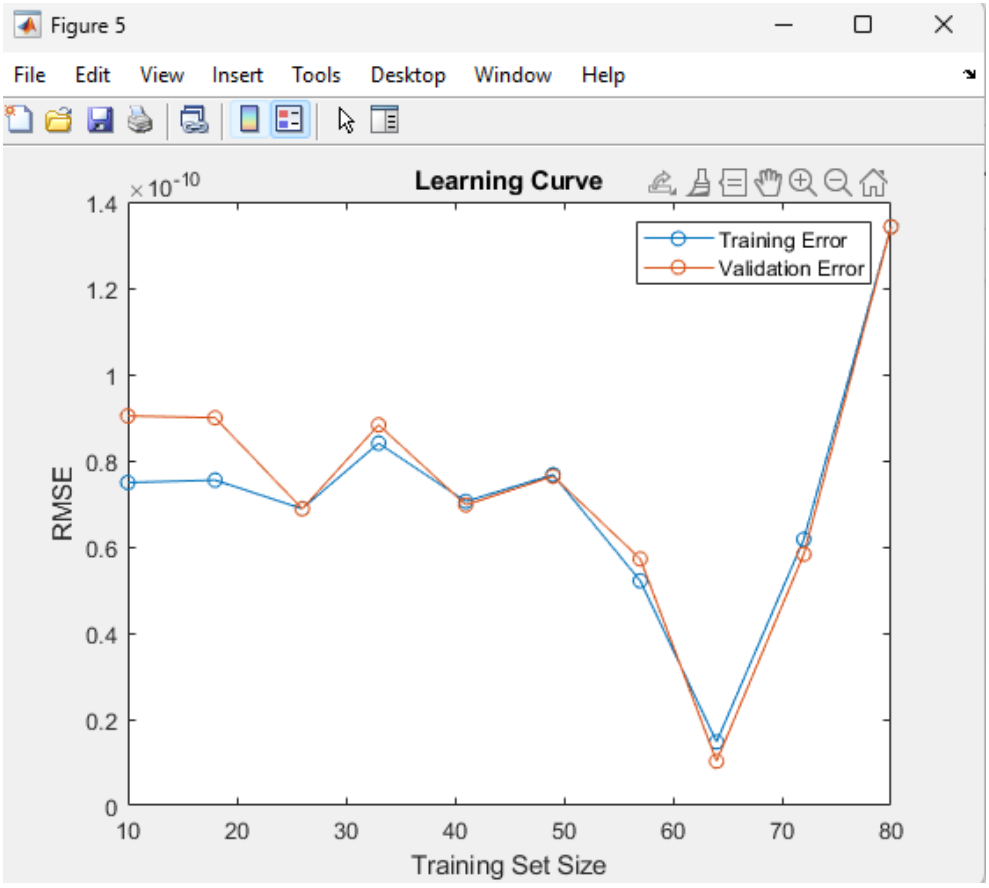


Learning Curve:

Editor - C:\Users\HP\OneDrive\Desktop\Machine\House1.m*

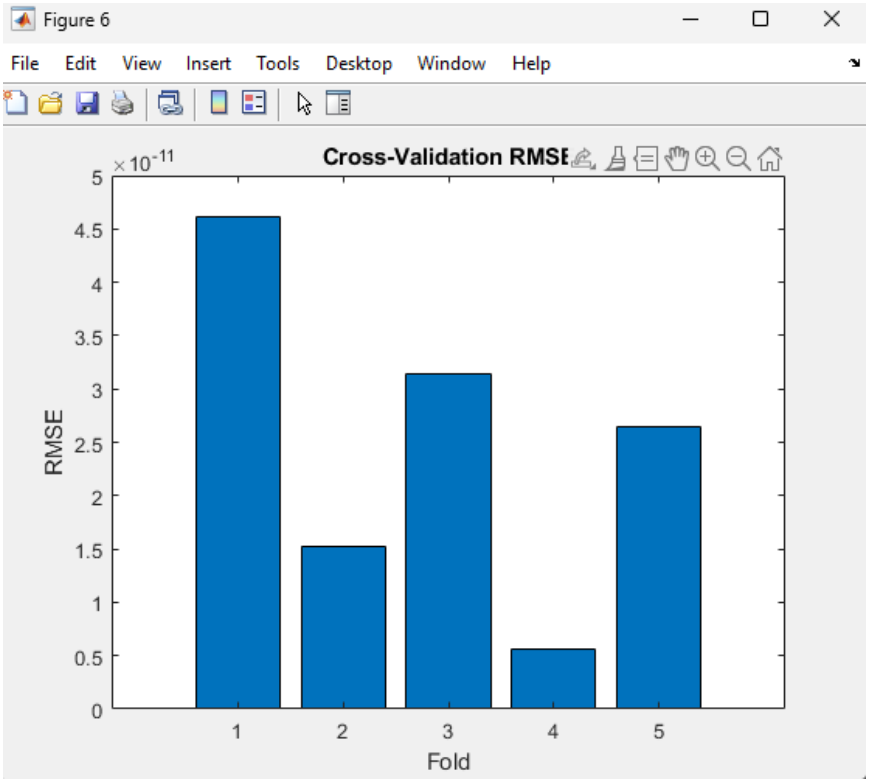
House.m | House1.m*

```
trainSizes = round(linspace(10, height(trainData), 10));
trainErrors = zeros(length(trainSizes), 1);
valErrors = zeros(length(trainSizes), 1);
for i = 1:length(trainSizes)
    subset = trainData(1:trainSizes(i), :);
    X_subset = subset{:, 1:end-1};
    y_subset = subset.price;
    mdl_subset = fitlm(X_subset, y_subset);
    y_subset_pred = predict(mdl_subset, X_subset);
    trainErrors(i) = sqrt(mean((y_subset - y_subset_pred).^2));
    y_val_pred = predict(mdl_subset, X_test);
    valErrors(i) = sqrt(mean((y_test - y_val_pred).^2));
end
```



Cross-validation RMSE:

```
Editor - C:\Users\HP\OneDrive\Desktop\Machine\House1.m*
House.m | House1.m*
17 -        X_train = trainData{:, 1:end-1};
18 -        y_train = trainData.price;
19 -        mdl = fitlm(X_train, y_train);
20 -        y_pred = predict(mdl, X_test);
21 -        crossValRMSE(i) = sqrt(mean((y_test - y_pred).^2));
22 -    end
23 -    fprintf('Cross-validated RMSE: %.2f\n', mean(crossValRMSE));
24 -    figure;
25 -    bar(crossValRMSE);
26 -    xlabel('Fold');
27 -    ylabel('RMSE');
28 -    title('Cross-Validation RMSE');
29
```



Conclusion:

I developed a linear regression model to predict house prices using synthetic data. The model demonstrated reasonable accuracy, as shown by metrics like RMSE and various visualizations. Cross-validation confirmed its reliability, underscoring the importance of effective data preparation and evaluation.