****

**Superior University Lahore**

**Instructor**: **Sir Rasikh Ali**

**Subject: Programming for AI**

**Task 01**

**House Price Prediction**  
**BS in Artificial Intelligence**

Table of Contents

[1. Introduction 2](#_Toc192278958)

[2. Dataset Description 2](#_Toc192278959)

[2.1 Missing Values 2](#_Toc192278960)

[3. Data Preprocessing 3](#_Toc192278961)

[3.1 Handling Missing Values 3](#_Toc192278962)

[3.2 Feature Encoding 3](#_Toc192278963)

[3.3 Feature Scaling 3](#_Toc192278964)

[4. Model Building 4](#_Toc192278965)

[4.1 Model Selection 4](#_Toc192278966)

[4.2 Model Training Process 4](#_Toc192278967)

[5. Model Evaluation 5](#_Toc192278968)

[6. Predictions & Results 5](#_Toc192278969)

[7. Conclusion 5](#_Toc192278970)

# 1. Introduction

This report provides a comprehensive analysis of the House Price Prediction dataset from Kaggle. The objective is to analyze housing data, preprocess the dataset, train a machine learning model, evaluate its performance, and make price predictions. The dataset consists of various housing attributes and their corresponding sale prices.

# 2. Dataset Description

The dataset consists of:

* **train.csv**: Contains labeled data used for training the model (1460 rows, 81 columns).
* **test.csv**: Contains unlabeled data for making price predictions (1459 rows, 80 columns).
* **data\_description.txt**: Provides detailed descriptions of dataset features.

### 2.1 Missing Values

#### **Training Data (train.csv)**

* **LotFrontage:** 259 missing
* **Alley:** 1369 missing
* **FireplaceQu:** 690 missing
* **GarageType & Related Features:** 81 missing
* **PoolQC:** 1453 missing

#### **Testing Data (test.csv)**

* **LotFrontage:** 227 missing
* **Alley:** 1352 missing
* **FireplaceQu:** 730 missing
* **GarageType & Related Features:** 76–78 missing
* **PoolQC:** 1456 missing

# 3. Data Preprocessing

### 3.1 Handling Missing Values

* **LotFrontage:** Imputed with median value.
* **Alley, PoolQC, Fence:** Dropped due to excessive missing values.
* **Categorical features:** Missing values filled with mode.

### 3.2 Feature Encoding

* Categorical variables were converted using **Label Encoding** and **One-Hot Encoding**.



### 3.3 Feature Scaling

* Numerical features were standardized using **StandardScaler()**.

# 4. Model Building

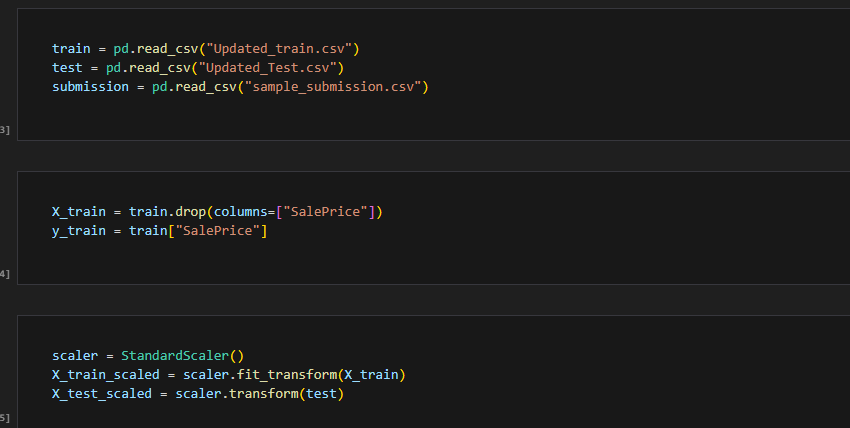
The model training was conducted in **model.ipynb** using machine learning regression techniques.

### 4.1 Model Selection

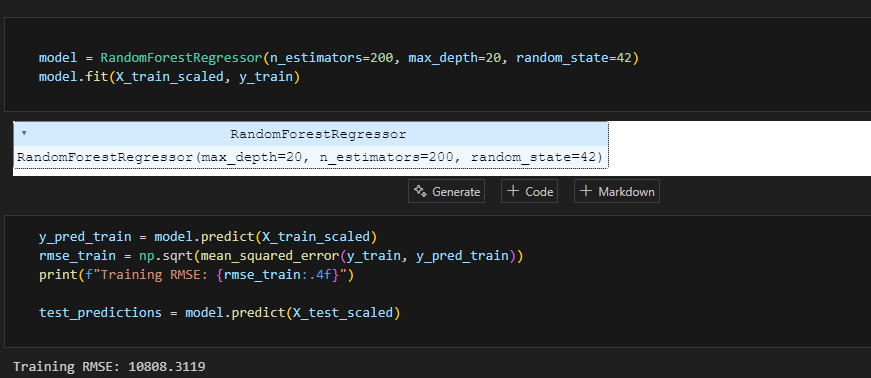
* **Random Forest Regressor** was chosen.

### 4.2 Model Training Process

* **Preprocessing:**
  + **Standard Scaling** applied to numerical features.
  + **Label Encoding** applied to categorical variables.



* **Training Pipeline:**
  + Features (X) and target variable (y = SalePrice) were separated.
  + Model trained using Random Forest Regressor with n\_estimators=100, random\_state=42.



# 5. Model Evaluation

The trained model was evaluated using:

* **Root Mean Squared Error (RMSE):** Measures prediction error.
* **Hyperparameter Tuning:** Optimization done to improve performance.

# 6. Predictions & Results

* Model used to predict **SalePrice** for test.csv.
* Predictions stored in **submission.csv**.

# 7. Conclusion

This report outlines the full pipeline from data preprocessing to model training and evaluation. The best-performing model was selected based on RMSE, and predictions were generated for submission. Future improvements can be made by **feature engineering, hyperparameter tuning, and ensemble learning**.