****

**Superior University Lahore**

**Name: Ahsan Ali**

**Roll no: 061**

**Subject: Programming for AI**

**Task: 1**

**Submitted To: Sir Rasikh**

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

Table of Contents

[1. Introduction 1](#_Toc192280918)

[2. Dataset Description 2](#_Toc192280919)

[2.1 Missing Values 2](#_Toc192280920)

[3. Data Preprocessing 2](#_Toc192280921)

[3.1 Handling Missing Values 2](#_Toc192280922)

[3.2 Feature Encoding 3](#_Toc192280923)

[3.3 Feature Scaling 4](#_Toc192280924)

[4. Model Building 4](#_Toc192280925)

[4.1 Model Selection 4](#_Toc192280926)

[4.2 Model Training Process 4](#_Toc192280927)

[5. Model Evaluation 4](#_Toc192280928)

[6. Predictions & Results 5](#_Toc192280929)

[7. Conclusion 5](#_Toc192280930)

# 1. Introduction

This report provides a comprehensive analysis of the house price prediction dataset. The objective is to analyze housing data, preprocess the dataset, train a machine learning model, make price predictions and evaluate its performance. The house price dataset is used for regression tasks, where the goal is to predict the sale price of a house based on various features such as size, condition and location.

# 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 survival predictions (1459 rows, 80 columns).
* **data\_description.txt**: Provides details on dataset features.

### 2.1 Missing Values

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

* **LotFrontage: 259 missing**
* **Alley: 1369 missing**
* **MasVnrType: 872 missing**
* **MasVnrArea: 8 missing**
* **BsmtOual: 37 missing**
* **BsmtCond: 37 missing**

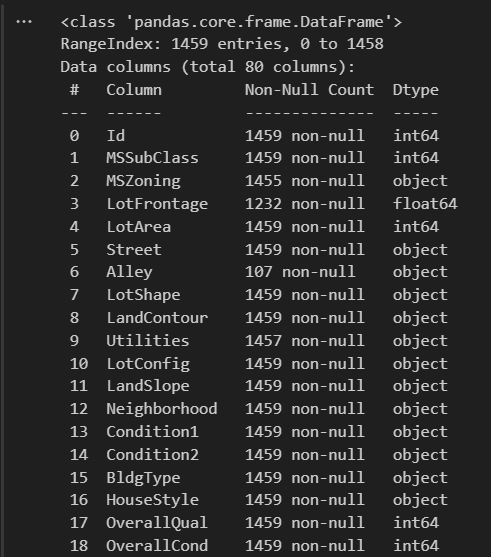
#### **Testing Data (test.csv)**

* **MSzoning: 4 missing**
* **LotFrontage**: 227 missing
* **Alley:** 1352 missing
* **Utilities:** 2 missing
* **Exteriorist1st:** 1 missing
* **Exterior2nd:** 1 missing

# 3. Data Preprocessing

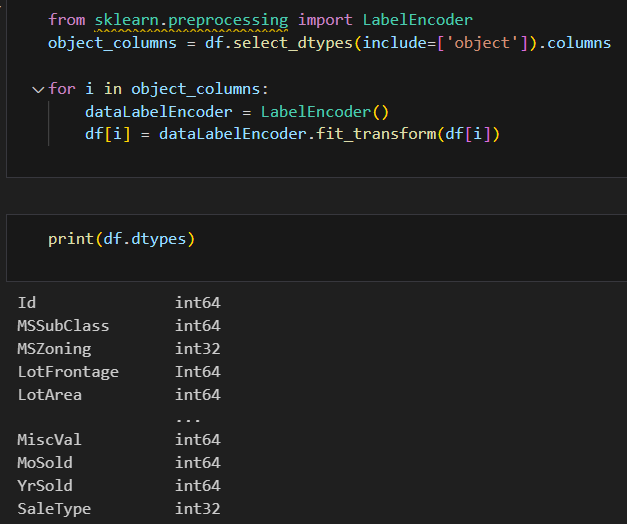
### 3.1 Handling Missing Values

* **MSzoning, LotFrontage, Alley, Utilities, VIP:** Missing values imputed using the mode.



### 3.2 Feature Encoding

* The target variable **LotFrontage** was mapped as **Yes → 1, No → 0**.
* Categorical variables converted using **Label Encoding**.



### 3.3 Feature Scaling

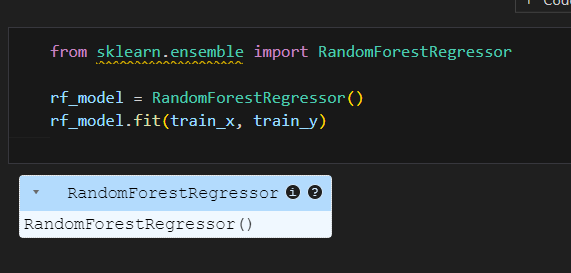
* Numerical features were standardized using **StandardScaler()**.

# 4. Model Building

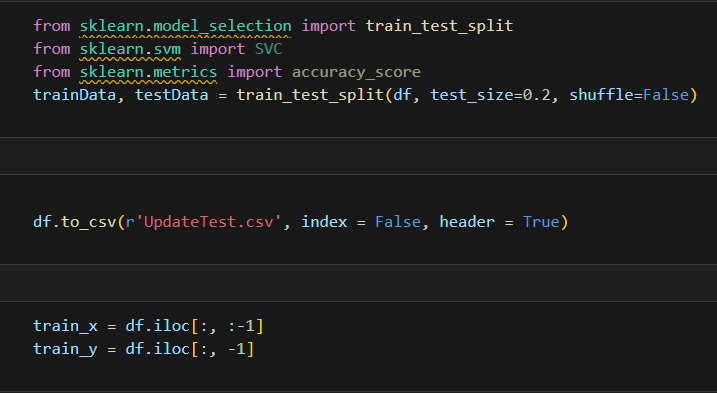
The model training was conducted in **model.ipynb** using machine learning classification techniques.

### 4.1 Model Selection

* **Random Forest Regressor** was chosen.

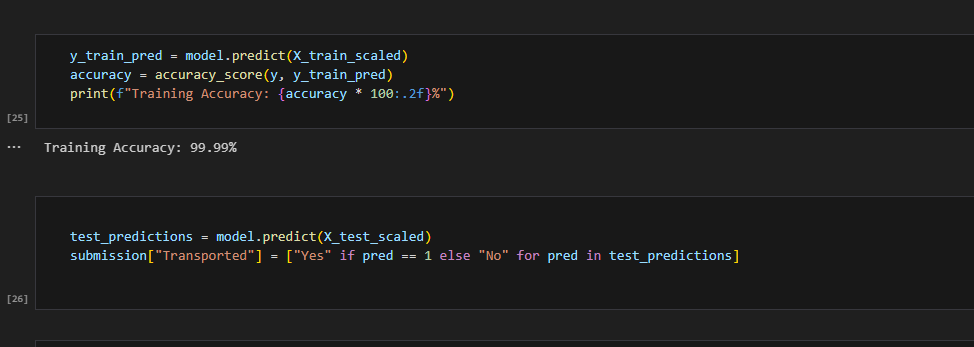


### 4.2 Model Training Process

* **Preprocessing:**
  + **Standard Scaling** applied to numerical features.  
      
    
  + **Label Encoding** applied to categorical variables.
* **Training Pipeline:**
  + Features (X) and target variable (y = LotFrontage) were separated.
  + Model trained using Random Forest Regressor.

# 5. Model Evaluation

The trained model was evaluated using:

* **Accuracy Score:** Measures model performance.
* **Training Accuracy:** Printed after model fitting.  
    
  

# 6. Predictions & Results

* Model used to predict **LotFrontage** 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 evaluation metrics and house price predictions were generated for submission. Future improvements can be made by exploring additional ensemble learning, hyperparameter tuning and feature engineering techniques to further enhance prediction accuracy.