

**The Superior University Lahore**

**Faculty of Computer Science & Information**

**Technology**

**Name: Alishba Haroon**

**Roll No: BSAI-116-4C**

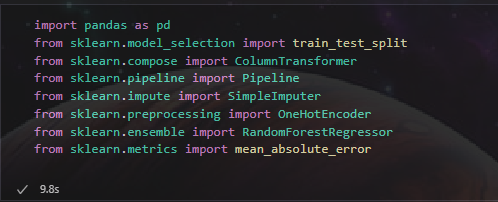
**Date: 27 Feb 2025**

**Subject: PAI LAB**

**House Price Prediction**

This Python program builds a house price prediction model using Random Forest Regression. The workflow follows these key steps:

**1. Importing Necessary Libraries:**

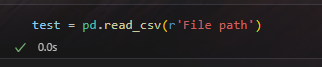


* **pandas:** For data handling and manipulation.
* **train\_test\_split:** Splits the dataset into training and testing subsets.
* **ColumnTransformer:** Applies different transformations to numerical and categorical columns.
* **Pipeline:** Automates preprocessing and model training steps.
* **SimpleImputer:** Handles missing values.
* **OneHotEncoder:** Converts categorical variables into numeric format.
* **RandomForestRegressor:** The machine learning model used for prediction.
* **mean\_absolute\_error (MAE):** Evaluates the model’s performance.

**2. Loading the Data:**

Loads the training and testing datasets from CSV files.





**3. train.head():**

* Displays the first **five rows** of the training dataset (train.csv).
* Helps in getting a quick overview of the structure and values.





**4. test.head():**



**5. train.tail():**

* Displays the last five rows of the training dataset.
* Useful for checking missing values or inconsistencies at the end of the dataset.

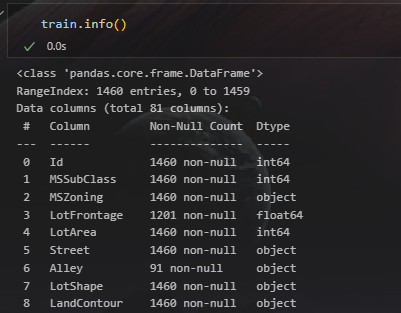


**6. test.tail():**

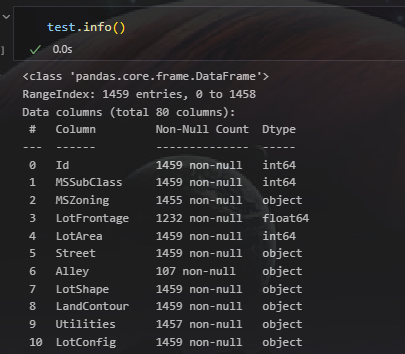


**7. train.info():**

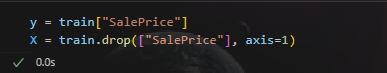
* Provides a summary of the dataset, including:
* Number of non-null values in each column.
* Data types of each column.
* Helps in identifying missing values and selecting appropriate preprocessing techniques.



**8. train.info():**



**9. Splitting Features and Target Variable:**

****

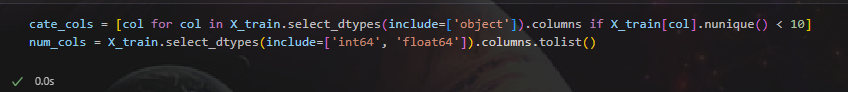
* y (target): Stores the house sale prices.
* X (features): Stores the rest of the dataset, excluding the target variable.

**10. Splitting Data into Training and Testing Sets:**

****

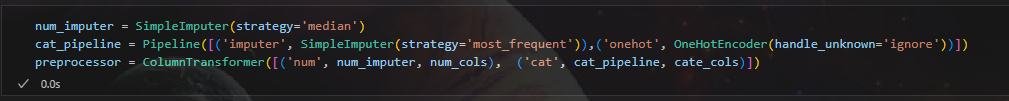
* 80% of the data is used for training (X\_train, y\_train).
* 20% is reserved for testing (X\_test, y\_test).
* random\_state=42 ensures reproducibility.

**11. Identifying Categorical and Numerical Columns:**

****

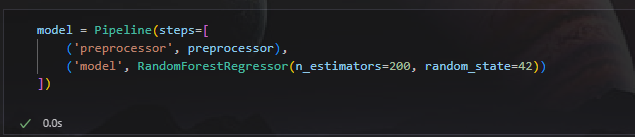
* Categorical columns: Selected if they have a data type of object and contain fewer than 10 unique values.
* Numerical columns: Selected if they are of type int64 or float64.

**12. Data Preprocessing:**

****

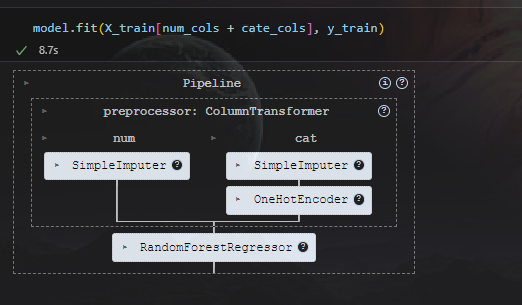
* **Numerical Data:**
* Missing values are replaced with the median using SimpleImputer(strategy='median').
* **Categorical Data:**
* Missing values are replaced with the most frequent category.
* One-hot encoding converts categorical data into numerical form.

**13. Building the Model Pipeline:**

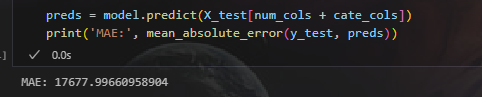
****

* The Pipeline ensures that preprocessing and model training happen in sequence.
* RandomForestRegressor is used with:
* n\_estimators=200: Uses 200 decision trees.
* random\_state=42: Ensures consistent results.

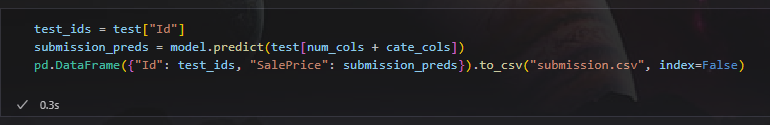
**14. Training the Model:**

****

* The model is trained on the preprocessed training data.

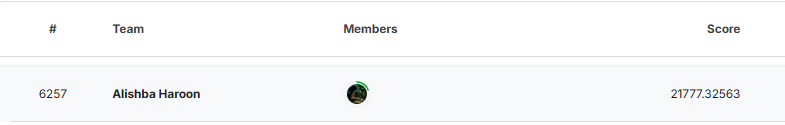
****

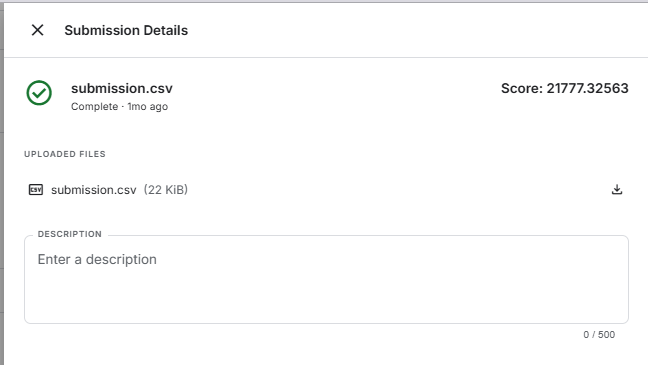
* Predictions are made on the test dataset.
* Mean Absolute Error (MAE) is calculated to evaluate model accuracy.

****

* The trained model makes predictions on the test dataset.
* Results are saved as submission.csv with Id and SalePrice.

**Accuracy:**

****

****