Celebal Assignment Week-5

Project Objective:

The goal of this project is to **predict the sale prices of houses** using a **Random Forest Regressor**. The model is trained on historical housing data and predicts prices for unseen properties in a test dataset. This is a **supervised machine learning regression problem**.

Dataset:

- **train.csv**: Contains features and the target column SalePrice (house price).
- **test.csv**: Contains the same features (excluding SalePrice) for which predictions are to be made.
- **house_price_predictions.csv**: Output file with predicted prices for test data.

Key Steps:

Importing Libraries

Essential Python libraries such as `pandas`, `numpy`, `matplotlib`, `seaborn`, and `scikit-learn` are imported for data manipulation, visualization, and machine learning.

Loading and Preparing Data

- The `Id` column is dropped as it's not useful for prediction.
- Features (`X_train`) and target (`y_train`) are separated from the training dataset.
- The `Id`s from the test dataset are saved for the final output.

Combining Data for Preprocessing

- Training and test data are combined to ensure consistent preprocessing.

Data Cleaning (Imputation)

- **Numerical Features**: Missing values are filled with the **median**.
- **Categorical Features**: Missing values are filled with the constant `'Missing'`.

Encoding Categorical Features

- Categorical columns are **OneHotEncoded** using `ColumnTransformer`.
- Numeric columns are **passed through** without modification.

Model Building

- A **Random Forest Regressor** (with 10 trees and random_state=0) is used to train on the processed data.
- The model is fit on the training set and used to predict house prices for the test set.

Prediction and Output

- The predicted prices are saved in a file house_price_predictions.csv with columns Id and SalePrice.

Visualization

- A simple **line plot** of Id Vs Predicted SalePrice is generated to visualize prediction trends.

Predicted salesprice values: