Predicting House Prices: A Comparative Analysis of Regression Models

Einas Nasasra ,212961916 , einasns@ac.sce.ac.il.
Haneen Abu Salook , 324813039, haneeab@ac.sce.ac.il .

1. General Problem Description

Accurately predicting house prices is a critical aspect of the real estate industry. It helps buyers, sellers, and investors make informed decisions. By analyzing various features such as location, size, and condition of houses, machine learning models can be used to estimate property values effectively.This project uses the **House Prices - Advanced Regression Techniques** dataset from Kaggle. The dataset contains information on house characteristics and their corresponding sale prices. By applying regression algorithms, we aim to predict house prices and identify the most significant factors influencing them.

2. Why You Chose This Project

This project was chosen due to its practical relevance in the real estate domain. Predicting house prices has a direct impact on economic decisions and investments. Additionally, this project aligns with our interest in machine learning and its applications in solving real-world problems. It provides an opportunity to enhance technical skills by comparing multiple regression models.

3. The Designed Implementation

- **Data Loading and Exploration:** We loaded the dataset and explored its structure, focusing on the distributions of numerical features and their correlations. Histograms and heatmaps highlighted patterns and relationships in the data.
- **Data Preprocessing:** Missing values were filled using median or mode imputation, categorical variables were one-hot encoded, and numerical features were scaled for consistency.
- **Feature Engineering:** New features like "price per square foot" were created to enhance model performance. Relevant features were selected based on their correlation with the target and importance in model training.
- **Model Training:** Linear Regression, Ridge Regression, and Random Forest Regressor models were trained to predict house prices effectively.
- **Model Evaluation:** The models were evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared. Residual plots and feature importance analysis provided further insights.
- **Comparative Analysis:** Model performances were compared through evaluation metrics and visualizations. The Tuned Random Forest model emerged as the most accurate and insightful for predicting house prices.

4 . General Description of the Experiments

- **Experiment 1**: Evaluate baseline model performance without feature engineering.
- **Experiment 2**: Assess the impact of feature engineering and scaling.
- **Experiment 3**: Compare hyperparameter-tuned models to identify the optimal model.
- **Experiment 4**: Analyze feature importance and identify key drivers of house prices.