Forecasting House Prices accurately Using Smart Regression

Techniques in Data Science

**Student Name:** S.Abinaya

**Register Number:** 422223243003

**Institution:** Surya group of Institutions

**Department:** B.Tech Artificial Intelligence and Data Science

**Date of Submission:** 13-05-2025

**Github Repository Link:<https://github.com/abinaya-209/abinaya.git>**

# 1. Problem Statement

Accurately forecasting house prices is crucial for buyers, sellers, investors, and real estate companies. This project aims to create robust regression models that leverage advanced data science techniques to predict housing prices based on multiple factors such as location, size, condition, and amenities.

# 2. Project Objectives

* Develop regression models that can predict house prices with high accuracy.
* Compare traditional regression methods with smart techniques like ensemble methods and regularization.
* Ensure real-world applicability and interpretability of results.

# Project Workflow (Flowchart )

DATA COLLECTION

PREPROCESSING

EDA

FEATURE ENGINEERING

MODEL TRAINING

EVALUATION

VISUALIZATION

REPORTING

# Data Description

* Dataset Source: Kaggle (e.g., Ames Housing Dataset )
* Type: Structured Tabular Data
* Records: ~2,900; Features: ~ 80
* Supervised learning with target variable: SalePrice

# Data Preprocessing

* Imputed missing values (mean, median, mode )
* Removed duplicates and outliers
* Encoded categorical variables (One-Hot Encoding )
* Scaled features using StandardScaler

# Exploratory Data Analysis (EDA )

EDA is commonly performed using tools like Python (with libraries such as Pandas, Matplotlib,

Seaborn) or R. It helps in forming hypotheses, cleaning the data, and guiding further

modeling or hypothesis testing. Visualized distributions, correlations, and key

relationships Identified top influencing factors like overall quality, living area, and neighborhood

Detected multicollinearity among numeric features

# Feature Engineering

Created new features like price per square foot . Binned year variables to capture age categories.

Applied polynomial and interaction terms .Optional: Used PCA to reduce dimensionality

# Model Building

* Linear Regression (Baseline )
* Ridge and Lasso Regression (to reduce overfitting )
* Random Forest Regressor and XGBoost (smart regression techniques )
* Evaluation metrics: RMSE, MAE, R²

# Visualization of Results & Insights

* Feature importance (Random Forest, XGBoost )
* Residual plots for error analysis
* RMSE comparison chart
* Interpretable summaries of model outputs

# Tools and Technologies Used

* **Language:** Python
* **IDE**: Jupyter Notebook
* **Libraries:** pandas, numpy, matplotlib, seaborn, scikit-learn, xgboost
* **Visualization:** Plotly, Seaborn

# Team Members and Contributions

* Data Cleaning & Preprocessing: Aashara Begam.H
* EDA & Feature Engineering: Abinaya.S
* Modeling: Heefa Sharin.M
* Reporting & Documentation: Archana.KS