Advance House Price Prediction - Project Report

1. Introduction

Housing prices fluctuate based on various factors, including location, property size, and amenities. The goal of this project is to use Supervised Machine Learning techniques, specifically Linear Regression and other regression models, to predict house prices accurately based on historical data.

2. Dataset Description

The dataset used in this project is USA_Housing.xls, which contains various features influencing house prices. The key attributes include:

- *Avg. Area Income - Average income of residents in the area.*
- *Avg. Area House Age - Average age of houses in the area.*
- *Avg. Area Number of Rooms - Average number of rooms per house.*
- *Avg. Area Number of Bedrooms - Average number of bedrooms per house.*
- *Area Population - Total population of the area.*
- *Price - Target variable (house price).*

3. Exploratory Data Analysis (EDA)

EDA was conducted to understand the dataset and identify key insights:

- *Checked for missing values and duplicates*
- *Visualized correlations using heatmaps and scatter plots*
- *Identified outliers using box plots and handled them using Interquartile Range (IQR)*
- *Analyzed feature distributions to understand data trends*

4. Data Preprocessing

- *Separated features (independent variables) and target variable (price)*
- *Split data into training and test sets (80% training, 20% testing)*
- *Normalized numerical features where necessary*

5. Model Selection & Training

Several regression models were implemented:

- *Linear Regression*
- *Decision Tree Regressor*
- *Random Forest Regressor*
- *AdaBoost Regressor*

Each model was trained on the dataset and evaluated using standard regression metrics.

6. Results & Evaluation

Model performance was assessed using:

- *Mean Absolute Error (MAE)*
- *Mean Squared Error (MSE)*
- *R-squared (R^2) Score*

The best-performing model was Random Forest Regressor, achieving an R^2 score of 0.85, indicating moderate predictive power.

7. Key Insights

- *Avg. Area Income and Avg. Area Number of Rooms were found to have the highest correlation with house prices.*
- *Random Forest Regressor outperformed Linear Regression, achieving a higher R^2 score and lower error rates.*
- *Outliers affected predictions, emphasizing the need for proper data cleaning.*
- *Feature selection significantly improved model performance, as not all variables contributed equally to predictions.*
- *AdaBoost performed poorly compared to other models, indicating that boosting methods might not be ideal for this dataset.*

8. Conclusion & Future Improvements

This project demonstrated how Linear Regression and other models can be used for house price prediction. Future improvements could include:

- *Using advanced models like XGBoost or Neural Networks for better accuracy.*
- *Incorporating additional features such as crime rates, school ratings, and neighborhood quality.*
- *Enhancing data preprocessing to remove more outliers for a cleaner dataset.*

9. References

- *Scikit-Learn Documentation*
- *Pandas & NumPy for Data Analysis*
- *Matplotlib & Seaborn for Visualization*
