# Housing Price Prediction Project Report

**Author: Usman Turajo**

**Date: October 17, 2025**

**Location: Rivers State, Nigeria**

## 1. Objective

The goal of this project was to build a predictive model for housing prices using a dataset containing various features of residential properties. The model aims to assist buyers, sellers, and real estate professionals in estimating property values based on structural and locational attributes.

## 2. Dataset Overview

The dataset included 545 entries with the following features:

- **Numerical**: price, area, bedrooms, bathrooms, stories
- **Binary Categorical**: mainroad, guestroom, basement, hotwaterheating, airconditioning, prefarea
- **Multiclass Categorical**: furnishingstatus

## 3. Data Cleaning

- Converted binary categorical variables (yes/no) to 0/1.
- One-hot encoded `furnishingstatus`.
- No missing values were found.
- Scaled numerical features using `StandardScaler`.

## 4. Exploratory Data Analysis (EDA)

- **Price Distribution**: Right-skewed, indicating a few high-priced outliers.
- **Area vs Price**: Strong positive correlation.
- **Correlation Matrix**: area, bathrooms, airconditioning, and prefarea showed strong relationships with price.
- **Boxplots**: Features like mainroad, guestroom, and basement positively influenced price.

## 5. Feature Engineering

- Final feature set included both scaled numerical and encoded categorical variables.
- No multicollinearity issues were observed.

## 6. Model Building

Four models were trained and evaluated:

- **Linear Regression**
- **Decision Tree Regressor**
- **Random Forest Regressor**
- **Gradient Boosting Regressor**

## 7. Model Evaluation

| Model | MAE (₦) | MSE (₦²) | $R^2$ Score |
|---|---|---|---|
| Linear Regression | 970,043.40 | $1.75 \times 10^{12}$ | 0.65 |
| Decision Tree | 1,195,266.06 | $2.64 \times 10^{12}$ | 0.48 |
| Random Forest | 1,018,481.11 | $1.96 \times 10^{12}$ | 0.61 |
| Gradient Boosting | 960,513.57 | $1.69 \times 10^{12}$ | 0.67 |

**Best Model**: Gradient Boosting Regressor

## 8. Insights & Recommendations

- **Area** is the most influential feature in determining price.
- Homes with amenities like **air conditioning**, **basement**, and **guestroom** tend to be priced higher.
- **Gradient Boosting** offers the best predictive performance and should be used for deployment.

## 9. Deployment

The final model was deployed as a RESTful API using **Flask** and hosted on **Render**, a cloud platform for web applications. This deployment enables real-time housing price predictions through HTTP requests, making the model accessible to users and developers.

## Deployment Process

- The trained Gradient Boosting model and preprocessing scaler were serialized using joblib
- A Flask application was created to expose a /predict route that accepts JSON input and returns predicted prices
- A Procfile was added to specify the web process using Gunicorn
- All dependencies were listed in requirements.txt
- The project was pushed to GitHub and connected to Render for automatic deployment.