🏠 House Price Prediction

End-to-End Machine Learning Project Report

1. Introduction

House price prediction is a classic regression problem in data science with real-world business applications in real estate, banking, and urban planning.
This project aims to build an **accurate and interpretable machine learning system** to predict house prices using structured data.

The project follows an **industry-standard machine learning lifecycle**, covering:

- Data understanding
- Preprocessing
- Model building
- Evaluation
- Interpretation
- Deployment readiness

2. Objectives

The main objectives of this project are:

1. Understand the dataset through Exploratory Data Analysis (EDA)
2. Handle missing values and categorical variables properly
3. Build multiple regression models
4. Compare models using robust evaluation metrics
5. Improve performance using ensemble methods and hyperparameter tuning
6. Interpret model predictions and identify key price drivers
7. Prepare the model for deployment

3. Dataset Description

- **Dataset Name:** house_prices.csv
- **Target Variable:** Price
- **Feature Types:**
  - Numerical features (area, rooms, age, etc.)
  - Categorical features (location, neighborhood, quality, etc.)

Challenges in Data

- Missing values
- Mixed data types
- Non-linear relationships between features and target

---

4. Exploratory Data Analysis (EDA)

EDA was performed using **Matplotlib only**, avoiding seaborn due to environment limitations.

Key EDA Steps:

- Dataset shape and structure inspection
- Statistical summary of numerical features
- Missing value analysis
- Distribution of house prices
- Scatter plots of numerical features vs price
- Correlation heatmap

Insights from EDA:

- House prices are right-skewed
- Living area and quality show strong correlation with price
- Some features show non-linear relationships with price

---

5. Data Preprocessing

To ensure clean and reusable preprocessing, **Pipeline and ColumnTransformer** were used.

Preprocessing Steps:

- Numerical features:
  - Mean imputation
  - Standard scaling
- Categorical features:
  - Mode imputation
  - One-hot encoding

This approach prevents data leakage and ensures consistency during training and prediction.

---

6. Model Building

Baseline Models:

- Linear Regression
- Ridge Regression
- Lasso Regression

Advanced Models:

- Polynomial Regression
- Decision Tree Regressor
- Random Forest Regressor
- Gradient Boosting Regressor
- XGBoost (optional, safely skipped if unavailable)

Each model was trained using the same preprocessing pipeline for fair comparison.

---

7. Model Evaluation

Evaluation Metrics Used:

- Mean Absolute Error (MAE)

- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- R² Score
- 5-Fold Cross-Validation RMSE

Evaluation results were saved in:

model_comparison_results.csv

Observations:

- Linear models provided interpretability but lower accuracy
- Polynomial regression improved fit but increased complexity
- Tree-based models captured non-linear patterns effectively
- Random Forest and Gradient Boosting achieved the best performance

---

8. Hyperparameter Tuning

Hyperparameter tuning was performed using **GridSearchCV** for Gradient Boosting.

Tuned Parameters:

- Number of estimators
- Learning rate
- Tree depth

This resulted in improved generalization and reduced error.

---

9. Model Interpretation & Explainability

9.1 Linear Model Coefficients

- Used to understand direction and strength of feature impact
- Saved as:

linear_model_coefficients.csv

9.2 Feature Importance (Random Forest)

- Identified top contributors to house price
- Saved as:

random_forest_feature_importance.csv

9.3 Permutation Importance

- Model-agnostic validation of feature influence
- Saved as:

permutation_importance.csv

9.4 Residual Analysis

- Residual vs predicted plots
- Residual distribution analysis
- Confirmed absence of major bias

9.5 Learning Curves

- Demonstrated good bias-variance balance
- Indicated potential improvement with more data

---

10. Key Insights

- Larger living area significantly increases house prices
- Overall quality and location are major price drivers
- Additional bathrooms and garage space add value
- Age impacts price negatively but less than expected
- Ensemble models outperform linear models in accuracy

---

11. Final Model Selection

Selected Model: Random Forest / Gradient Boosting
Reason:

- Best RMSE and R^2 score

- Robust to non-linear relationships
- Stable performance across cross-validation

The final model was saved as:

house_price_model.pkl

---

12. Deployment Readiness

The trained model is ready for:

- Streamlit web application
- Flask REST API
- Integration into larger systems

---

13. Limitations

- Dataset size limits generalization
- XGBoost and SHAP could not be installed due to environment constraints
- External economic factors are not included

---

14. Future Improvements

- Add more data for better generalization
- Integrate SHAP for advanced explainability
- Build Streamlit dashboard for user interaction
- Automate retraining and monitoring

---

15. Conclusion

This project successfully demonstrates an **end-to-end machine learning workflow**, combining accuracy, interpretability, and best practices.
It reflects real-world data science skills and is suitable for **academic submission, internship evaluation, and professional portfolio**.

16. Interview-Ready Summary

“I built a complete house price prediction system, performed EDA, preprocessing, model comparison, hyperparameter tuning, interpretation, and prepared the model for deployment using industry-standard pipelines.”