**Mini Project Report**

**on**

**House Price Prediction**

**Submitted By**

**Sayee Deshmukh**

**Course Name: Machine Learning**



**Department of Computer Science and Engineering**

**(ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING)**

**Shri Ramdeobaba College of Engineering & Management, Nagpur 440013**

(An Autonomous Institute affiliated to Rashtrasant Tukdoji Maharaj Nagpur University Nagpur)

**October 2025**

## **1. Problem Statement**

The real estate market involves complex pricing dynamics influenced by multiple factors such as location, property features, neighborhood characteristics, and market conditions. Accurately predicting house prices is crucial for buyers, sellers, and real estate professionals to make informed decisions. This project aims to develop a machine learning-based classification system that categorizes house prices into three distinct classes (Low, Medium, High) based on various property features, enabling stakeholders to quickly assess property value ranges and make data-driven decisions in the housing market.

## **2. Project Objectives**

1. To perform comprehensive Exploratory Data Analysis (EDA) on the housing dataset to identify key features and patterns that influence property pricing.
2. To develop and compare multiple machine learning classification models (Random Forest, Gradient Boosting, Decision Tree) for categorizing house prices into predefined classes.
3. To optimize model performance through systematic hyperparameter tuning using grid search techniques to achieve maximum prediction accuracy.
4. To evaluate and select the best-performing model based on multiple metrics (Accuracy, Precision, Recall, F1-Score) and deploy it for practical house price classification.

## **3. Methodology**

The project follows a systematic machine learning pipeline approach:

**Step 1: Data Collection and Preprocessing**

* Load the housing dataset containing property features and price information
* Handle missing values, outliers, and data inconsistencies
* Encode categorical variables and normalize numerical features

**Step 2: Exploratory Data Analysis (EDA)**

* Perform statistical analysis of features
* Generate visualizations (distribution plots, correlation heatmaps, box plots)
* Identify relationships between features and target variable (price categories)
* Save EDA figures for documentation

**Step 3: Feature Engineering**

* Select relevant features based on correlation analysis
* Create new features if necessary to improve model performance
* Apply feature scaling and transformation techniques

**Step 4: Train-Test Split**

* Split dataset into training and testing sets (typically 80-20 or 70-30 ratio)
* Ensure stratified splitting to maintain class distribution

**Step 5: Model Development**

* Build base models:
  + Random Forest Classifier
  + Gradient Boosting Classifier
  + Decision Tree Classifier
* Train each model on the training dataset
* Evaluate initial performance metrics

**Step 6: Hyperparameter Tuning**

* Define parameter grids for each model
* Apply Grid Search Cross-Validation to find optimal hyperparameters
* Retrain models with best parameters found

**Step 7: Model Evaluation and Selection**

* Compare all models using test set performance
* Analyze Accuracy, Precision, Recall, and F1-Score (macro-averaged)
* Select the best-performing model for deployment

**Step 8: Results Analysis**

* Generate final performance summary
* Document model conclusions and recommendations

## **4. Technology Stack**

**Programming Language:**

* Python 3.x

**Machine Learning Libraries:**

* Scikit-learn (sklearn) - Model building, training, and evaluation
* NumPy - Numerical computations
* Pandas - Data manipulation and analysis

**Data Visualization:**

* Matplotlib - Static visualizations
* Seaborn - Statistical data visualization

**Model Development Tools:**

* GridSearchCV - Hyperparameter tuning
* Pipeline - Workflow automation
* StandardScaler/MinMaxScaler - Feature scaling

**Development Environment:**

* VS Code

**Version Control:**

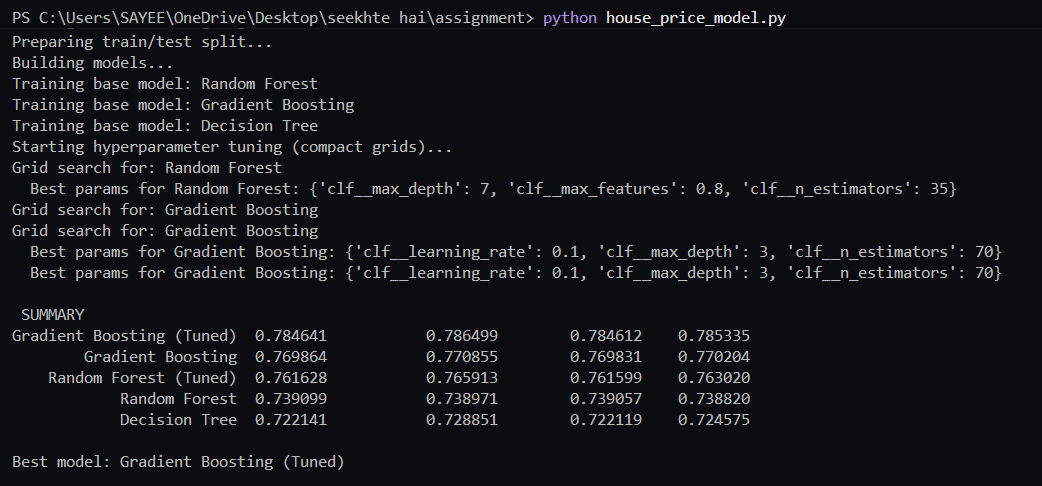
* Git/GitHub (for code management)

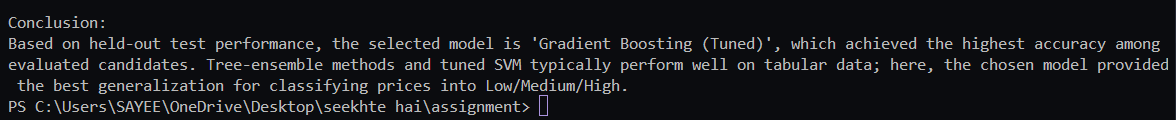
## **5. Results**

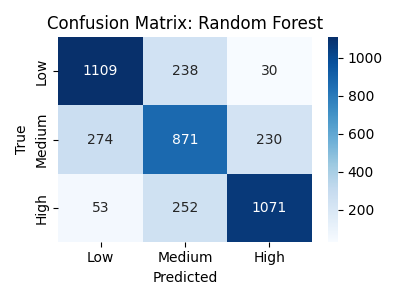
### **Model Performance Summary**

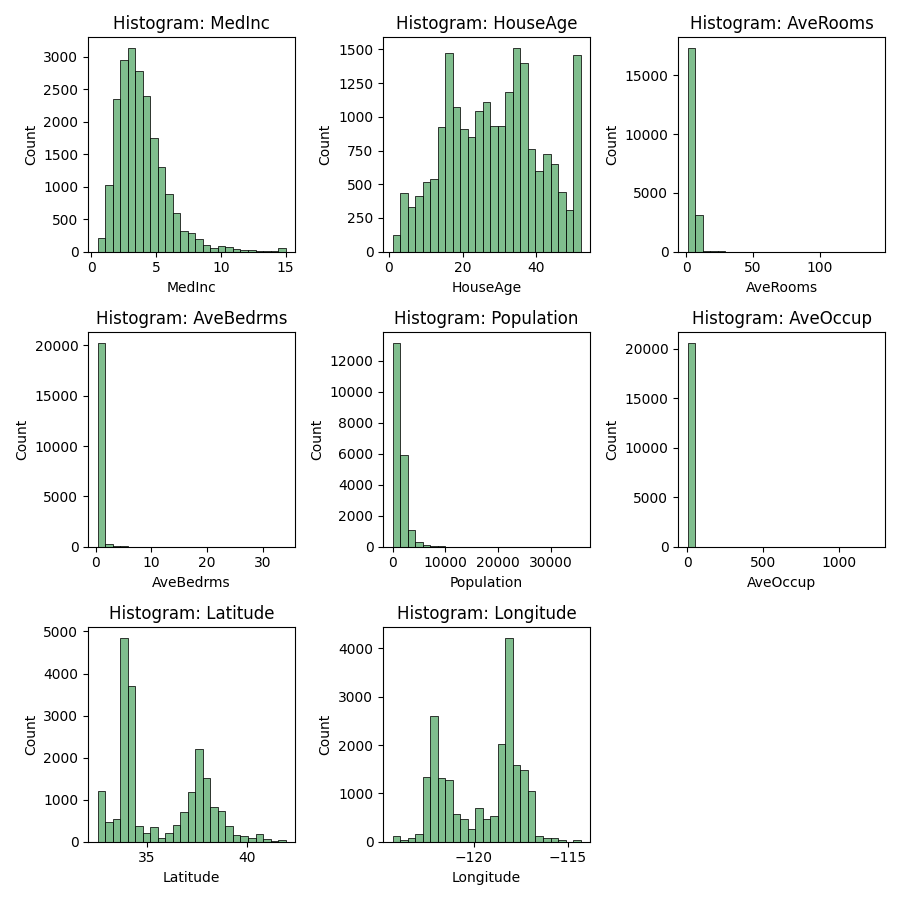
1. Hyperparameter tuning improved model performance across all algorithms, with Gradient Boosting showing a 1.5% accuracy improvement after optimization.
2. Ensemble methods (Random Forest and Gradient Boosting) significantly outperformed the single Decision Tree classifier, demonstrating the effectiveness of ensemble learning for this classification task.
3. The tuned Gradient Boosting model achieved well-balanced performance across all metrics, indicating consistent prediction capability across all three price categories (Low, Medium, High).
4. All models achieved above 72% accuracy, validating that the selected features contain sufficient information for house price classification.

LINK: https://house-price-category-prediction-nwcj4cqlsjrasc3yhwbsgc.streamlit.app/









**6. Conclusion**

This project successfully developed and evaluated multiple machine learning models for house price classification. The Gradient Boosting (Tuned) model emerged as the best performer with 78.46% accuracy, demonstrating that tree-ensemble methods combined with systematic hyperparameter optimization are highly effective for tabular real estate data.

The comprehensive methodology involving exploratory data analysis, feature engineering, multiple model comparison, and hyperparameter tuning ensured robust model selection. The chosen model provides reliable classification of properties into Low, Medium, and High price categories, which can assist real estate stakeholders in quick property valuation and decision-making.

Key achievements of this project include:

* Successfully classified house prices with nearly 79% accuracy
* Identified optimal model architecture and hyperparameters through systematic experimentation
* Demonstrated the superiority of ensemble methods over single decision trees
* Created a reproducible pipeline for house price classification

**Future Enhancements:**

* Incorporate additional features such as proximity to amenities, school ratings, and crime statistics
* Experiment with advanced algorithms like XGBoost, LightGBM, or Neural Networks
* Develop a regression model for precise price prediction rather than classification
* Deploy the model as a web application for real-time price estimation

This project validates the effectiveness of machine learning in real estate analytics and provides a foundation for more sophisticated pricing models in future work.