**DUY TAN UNIVERSITY**

**SCHOOL OF MACHINE SCIENCE**

**FACULTY OF MACHINE SCIENCE**

**✍🕮🖎**

****

**PRATICAL APPLICATION REPORT**

**COURSE: MACHINE LEARNING 2**

**HOUSE PRICE PREDICTION USING MACHINE LEARNING IN PYTHON**

**SUPERVISORS : ANAND NAYYAR & TẠ HOÀNG ĐĂNG KHOA**

**CLASS : DS 371 D**

**STUDENT : HUỲNH DUY LINH – 2821151451**

**Đà Nẵng, 3/2025**

**Contents**

[**I. INTRODUCTION** 3](#_Toc192582332)

[1. Background 3](#_Toc192582333)

[2. Problem Statement 3](#_Toc192582334)

[3. Objectives 4](#_Toc192582335)

[4. Significance 5](#_Toc192582336)

[5. Scope 5](#_Toc192582337)

[**II. METHODOLOGY** 6](#_Toc192582338)

[1. Overview of the Approach 6](#_Toc192582339)

[2. Dataset Description 6](#_Toc192582340)

[2.1 Data Sources 6](#_Toc192582341)

[2.2 Data Characteristics, Size, and Features 6](#_Toc192582342)

[2.3 Data Cleaning and Preprocessing Steps 7](#_Toc192582343)

[3. Feature Engineering 8](#_Toc192582344)

[4. Machine Learning Models 9](#_Toc192582345)

[5. Tools and Libraries 9](#_Toc192582346)

[6. Experimental Setup 10](#_Toc192582347)

[**III. IMPLEMENTATION** 11](#_Toc192582348)

[1. Process Flow 11](#_Toc192582349)

[2. Code and Pseudocode 11](#_Toc192582350)

[3. System Architecture Diagram 13](#_Toc192582351)

[**IV. RESULTS** 14](#_Toc192582352)

[1. Evaluation Metrics 14](#_Toc192582353)

[2. Performance Analysis 15](#_Toc192582354)

[3. Comparison of Approaches 17](#_Toc192582355)

[4. Real-Time System Performance 18](#_Toc192582356)

[**V. DEMO** 18](#_Toc192582357)

[1. Interactive User Interface 18](#_Toc192582358)

[2. Real-Time Predictions 19](#_Toc192582359)

[3. Visualizations and Data Insights 19](#_Toc192582360)

[4. Step-by-Step Demo Walkthrough 22](#_Toc192582361)

[5. Feedback and Iterative Improvement 22](#_Toc192582362)

[**VI. DISCUSSION** 23](#_Toc192582363)

[1. Analysis of Results 23](#_Toc192582364)

[Key Findings 23](#_Toc192582365)

[2. Strengths 24](#_Toc192582366)

[3. Limitations and Challenges 24](#_Toc192582367)

[**VII. CONCLUSION** 25](#_Toc192582368)

[For years, house prices have been analyzed using numerical data, providing valuable insights into the real estate market. The use of statistical data has increased the number of scientific studies focused on housing trends. One of the most frequently studied topics in these research efforts is house price prediction. 25](#_Toc192582369)

[This study is an example of how machine learning algorithms can be applied to forecast real estate prices. We built two popular machine learning models and evaluated their accuracy on both training and test datasets. 25](#_Toc192582370)

[Recommendation: Based on the accuracy scores, we strongly recommend using the Random Forest Regression model for house price prediction due to its superior performance. 25](#_Toc192582371)

[1. Summary 26](#_Toc192582372)

[2. Contributions 26](#_Toc192582373)

[3. Future Work 26](#_Toc192582374)

[**ACKNOWLEDGEMENT** 27](#_Toc192582375)

[**REFERENCES** 30](#_Toc192582376)

# **I. INTRODUCTION**

## 1. Background

The annual increase in house prices has brought volatility and instability to the real estate market, highlighting the urgent need for accurate price forecasting systems. Accurate house price prediction remains a challenge due to the many influencing factors. This study aims to identify and analyze the main determinants affecting house prices, using two established machine learning models. Through comparative analysis, the study will propose the most effective model to improve the accuracy of house price prediction. 

Nowadays, most people participate in commercial investment activities. Stocks, bonds, retirement funds, education, and other options are widely used as investment vehicles. One of the most common forms of investment is purchasing real estate. This process is not as simple as it may seem. Any real estate project that is bought or invested in often involves a series of separate transactions involving multiple parties. Therefore, it can be a crucial decision for both households and businesses. Using the House Price Prediction dataset, we will analyze and understand how different variables can predict house values. We will explore the impact of various factors such as location, size, house quality, and condition on housing costs. One of the key techniques for determining a property's value is predictive analysis.

In this study, we will use both Linear Regression and Random Forest Regression to forecast house prices while considering multiple aspects. The insights gained from this research will help homebuyers determine the best time to purchase a house and assist real estate investors in making informed decisions.

The housing market is currently affected by high interest rates, which drive up home prices and impact both housing supply and demand. As a result, it is essential to consider additional key metrics or factors that influence house prices. The purpose of this study is to predict house values using two well-known machine learning models.

In today's era of globalization and rapid technological advancement, the real estate sector is experiencing significant transformations. Continuous population growth, urbanization trends, and a highly competitive investment landscape have made accurately determining property values more critical than ever. Traditional valuation methods often fall short due to the dynamic and complex nature of the market. In contrast, machine learning (ML) has emerged as a powerful tool, enabling the analysis of large datasets to uncover intricate, nonlinear relationships between variables—capabilities that traditional statistical methods often lack.

Recent studies have demonstrated that integrating domain-specific real estate knowledge with advanced ML techniques, such as Linear Regression, Random Forest, and Support Vector Machines (SVM), leads to more precise property value predictions. This shift towards data-driven insights not only enhances decision-making for investors and brokers but also helps in crafting more effective market strategies and policies.

## 2. Problem Statement

Due to the volatile and uncertain nature of the housing market, identifying key metrics that influence house price predictions is crucial. House prices are often believed to be closely linked to the overall economy, but is this truly the case?

Despite the abundance of available data, there remains a lack of reliable real estate price forecasts. This gap highlights the need for a data-driven approach to improve prediction accuracy and assist buyers, investors, and policymakers in making informed decisions.

Despite the advancements in machine learning, predicting house prices remains a challenging task due to several inherent complexities:

Heterogeneous and Incomplete Data:  
Real estate data is often sourced from various channels and comes in different formats. This inconsistency, along with missing or noisy data, can significantly undermine the reliability of predictions.

Multiple Influencing Factors:  
Property values depend on numerous factors such as geographic location, size, number of rooms, local amenities, and broader economic conditions. Isolating the impact of each factor and understanding their interdependencies is a complex process.

Non-linear Relationships:  
Many relationships between the features of a property and its market value are non-linear. Capturing these complex interactions requires models that can adapt to and learn from non-linear data patterns.

Market Volatility:  
The real estate market is subject to rapid changes due to external factors such as government policies, economic fluctuations, and shifting consumer demands. This volatility adds another layer of difficulty in making reliable predictions.

These challenges highlight the need for a robust, integrated approach that combines thorough data preprocessing, effective feature engineering, and advanced ML algorithms to achieve accurate and reliable house price predictions.

## 3. Objectives

The main goal of this project is to develop a machine learning-based system capable of accurately predicting house prices. Specific objectives include:

Data Collection and Integration:  
Identify and gather high-quality real estate data from reliable public sources and property listings, ensuring consistency and completeness.

Data Preprocessing:  
Implement procedures to clean, normalize, and encode the data, addressing issues such as missing values and outliers to prepare a high-quality dataset for modeling.

Feature Engineering:  
Analyze key variables that significantly influence property prices and create new features that enhance the model’s predictive performance.

Model Development and Optimization:  
Evaluate multiple machine learning algorithms—including Linear Regression, Random Forest, and SVM—to identify the most effective model. Optimize model parameters through techniques such as cross-validation and hyperparameter tuning.

Evaluation and Validation:  
Use performance metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared to rigorously assess the accuracy and robustness of the developed models.

Documentation and Reporting:  
Compile a detailed report that documents the entire process—from data acquisition and preprocessing to model development and evaluation—providing valuable insights for future improvements and academic reference.

## 4. Significance

The significance of this project extends across multiple dimensions:

For Investors and Brokers:  
Accurate house price predictions can provide valuable insights into market trends, enabling investors to make informed decisions about property transactions. Brokers can leverage these predictions to advise clients more effectively and set realistic pricing strategies.

For Home Buyers:  
Transparent and reliable property valuations help home buyers avoid the risk of overpaying, fostering fairer market transactions.

For Policymakers and Regulators:  
The predictive models developed in this project can serve as important tools for market analysis, aiding policymakers in formulating strategies and regulations that promote sustainable urban development.

For Academic Research:  
This project exemplifies the practical application of machine learning in real estate analytics, contributing to academic literature and opening avenues for further research in model optimization and big data analytics.

## 5. Scope

The scope of this project is clearly defined to maintain a focused approach:

Data Collection:  
The project concentrates on collecting real estate data from publicly available sources, focusing on core attributes such as location, area, number of rooms, and available amenities. Data collection is limited to specific geographic regions to ensure consistency.

Data Preprocessing:  
Emphasis is placed on cleaning and preparing the data—removing noise, handling missing values, and standardizing features—to build a structured dataset suitable for machine learning.

Model Development:  
The project focuses on developing, testing, and comparing various machine learning models to determine the most effective approach for predicting house prices. This includes techniques such as parameter tuning, cross-validation, and comprehensive performance evaluation.

Application Limitations:  
This study does not extend to deploying the predictive system in a live production environment or integrating it with existing real estate management systems. Additionally, broader macroeconomic factors and policy changes are considered outside the scope of this project.

Documentation:  
A detailed report will document every stage of the project, including methodologies, experiments, results, and limitations. This comprehensive documentation will serve as a valuable resource for future research and development.

# **II. METHODOLOGY**

## 1. Overview of the Approach

This project employs a systematic approach to predict house prices using machine learning. The methodology consists of several key steps:

Data Collection & Integration: Aggregating data from various reliable sources to form a comprehensive dataset.

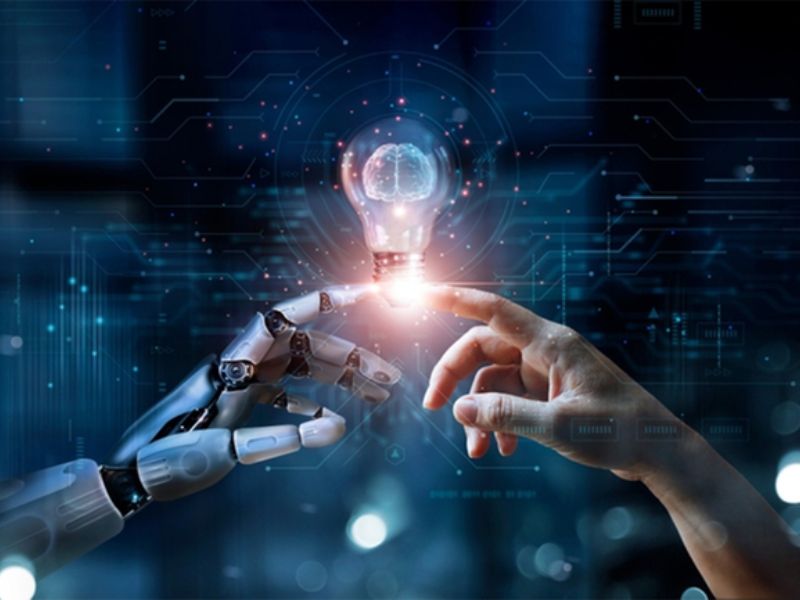
Data Cleaning & Preprocessing: Addressing missing values, outliers, and inconsistencies to ensure data quality.

Feature Engineering: Enhancing the dataset by selecting and creating relevant features that capture the nuances of the housing market.

Model Development: Implementing and comparing multiple machine learning models to determine the most accurate predictor.

Evaluation & Validation: Applying robust validation techniques and performance metrics to assess model performance.

Documentation: Thoroughly recording each step to provide transparency and reproducibility.

This structured workflow ensures that the project not only develops an accurate predictive model but also adheres to rigorous data science practices. 

## 2. Dataset Description

### 2.1 Data Sources

The dataset used in this project is obtained from publicly available real estate data, specifically the kc\_house\_data.csv file. This dataset contains extensive information about house prices and their attributes, providing a strong foundation for predictive modeling. The data was collected from property listings and real estate agencies, ensuring a diverse range of property characteristics.

The dataset includes key details such as:

* Historical house sale prices.
* Property features (e.g., square footage, number of bedrooms and bathrooms, year built).
* Geographic information (latitude, longitude, and zip codes).
* Structural conditions and renovations.
* Environmental factors (waterfront properties, view ratings).

### 2.2 Data Characteristics, Size, and Features

The dataset comprises 21,613 records of individual properties, with each record representing a unique house sale transaction. The dataset consists of 21 features, including continuous numerical variables, categorical variables, and binary indicators. The target variable for prediction is house price (column: price).

Below is a breakdown of key features in the dataset:

|  |  |  |
| --- | --- | --- |
| **Feature Name** | **Description** | **Type** |
| id | Unique house identifier | Categorical |
| date | Date of the sale | Categorical |
| price | Sale price (Target Variable) | Continuous |
| bedrooms | Number of bedrooms | Continuous |
| bathrooms | Number of bathrooms | Continuous |
| sqft\_living | Living area in square feet | Continuous |
| sqft\_lot | Lot size in square feet | Continuous |
| floors | Number of floors | Continuous |
| waterfront | Whether the house is waterfront (1: Yes, 0: No) | Binary |
| view | Quality of the view (0 to 4) | Categorical |
| condition | Condition of the house (1 to 5) | Categorical |
| grade | Overall grade (1 to 13) | Categorical |
| sqft\_above | Area above ground level (sqft) | Continuous |
| sqft\_basement | Basement area (sqft) | Continuous |
| yr\_built | Year the house was built | Continuous |
| yr\_renovated | Year of last renovation | Continuous |
| zipcode | Postal code of the house location | Categorical |
| lat | Latitude coordinate | Continuous |
| long | Longitude coordinate | Continuous |
| sqft\_living15 | Average living area of nearest 15 houses (sqft) | Continuous |
| sqft\_lot15 | Average lot size of nearest 15 houses (sqft) | Continuous |

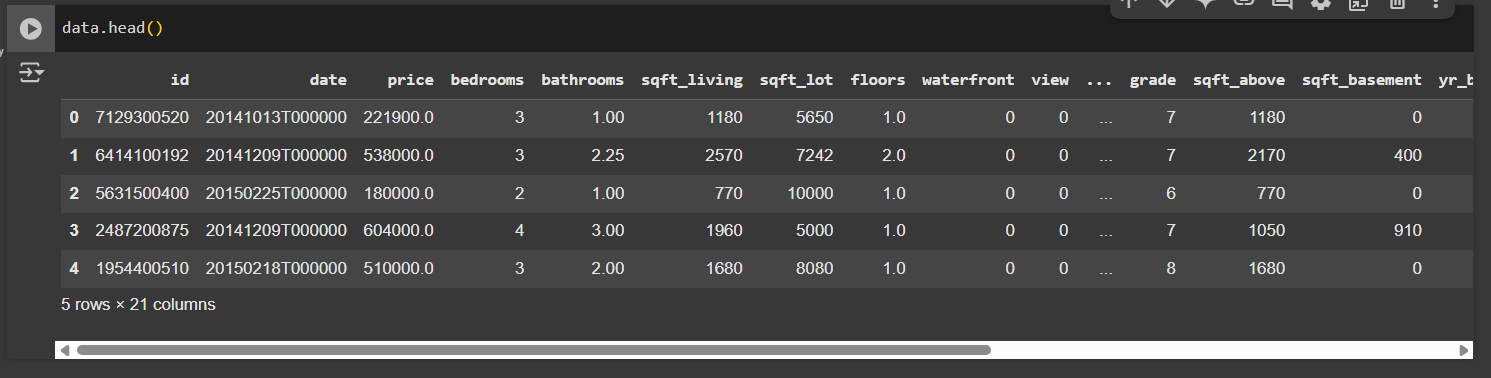
The dataset comprises a substantial number of records, each representing individual properties. Key characteristics include:

Quantitative Features: Numerical data such as property size (in square feet/meters), number of bedrooms and bathrooms, age of the property, and historical price trends.

Qualitative Features: Categorical data including property type, neighborhood, and condition. These are later encoded for model compatibility.

Temporal Aspects: Time-series data that reflects market trends over different periods.

Size & Diversity: The dataset is sufficiently large to allow for both training and testing, ensuring that the model generalizes well to unseen data.

The diversity of features enables the exploration of various relationships that influence property prices. 

### 2.3 Data Cleaning and Preprocessing Steps

Ensuring high-quality data is critical for model accuracy. The preprocessing phase involves:

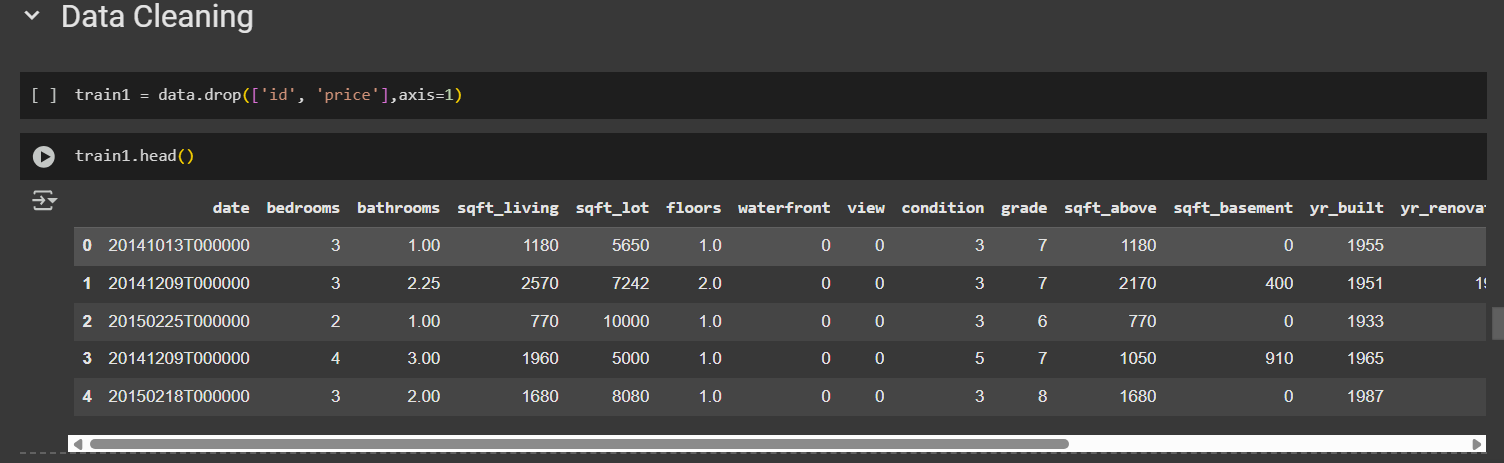
Handling Missing Values: Implementing strategies such as imputation or removal of records with missing or inconsistent data.

Outlier Detection: Identifying and addressing anomalies that may skew the analysis. Techniques like Z-score analysis and box plots are utilized.

Normalization & Scaling: Standardizing numerical features to ensure that they contribute equally to the model. Methods like Min-Max Scaling or Standardization are applied.

Encoding Categorical Variables: Converting non-numeric data into a machine-readable format using techniques such as one-hot encoding.

Data Splitting: Dividing the data into training and testing sets, ensuring that the split maintains the underlying distribution for robust model evaluation.



These steps lay a solid foundation for the subsequent modeling phase, reducing noise and enhancing feature relevance.

## 3. Feature Engineering

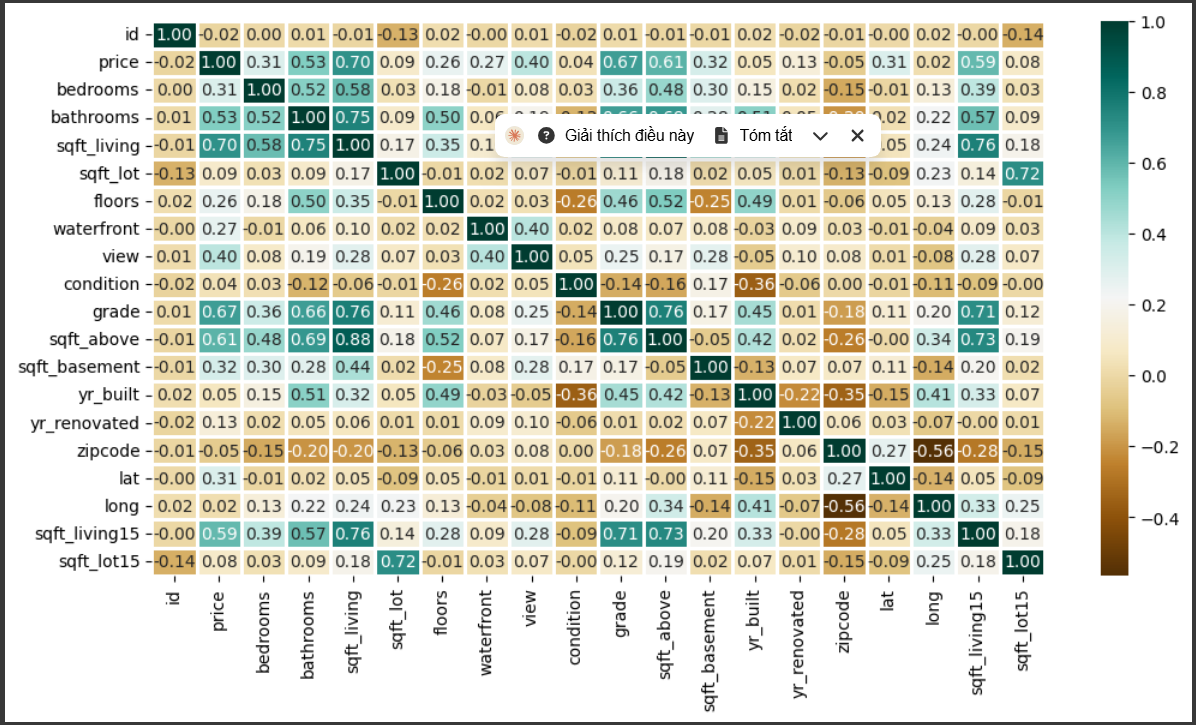
Feature engineering is essential for unlocking the full predictive power of the data. The process involves:

Feature Selection: Identifying the most relevant features through correlation analysis, domain knowledge, and statistical tests.

Creation of New Features: Deriving new variables that capture additional information—for instance, calculating the price per square foot or creating interaction terms between features.

Dimensionality Reduction: Techniques such as Principal Component Analysis (PCA) may be employed to reduce the complexity of the dataset while preserving key information.

Validation of Features: Each engineered feature is rigorously tested for its contribution to model performance using exploratory data analysis and iterative model testing.

Effective feature engineering can significantly improve model accuracy by providing more informative inputs. 

## 4. Machine Learning Models

A comparative study of multiple machine learning algorithms is conducted to identify the best-performing model for house price prediction. Key models include:

Linear Regression:  
A fundamental approach that assumes a linear relationship between the predictors and the target variable. It provides a baseline for comparison.

Random Forest Regressor:  
An ensemble method that builds multiple decision trees and aggregates their predictions. Its ability to capture non-linear relationships and handle high-dimensional data makes it a strong candidate.

Support Vector Machine (SVM) Regression:  
SVM uses kernel functions to map the input space into higher dimensions, making it capable of modeling complex non-linear relationships.

Additional Models:  
Other advanced models, such as Gradient Boosting Machines or Neural Networks, may also be explored to further enhance predictive performance.

Each model is subjected to rigorous training, hyperparameter tuning, and cross-validation to ensure that the selected model delivers both high accuracy and generalizability.

## 5. Tools and Libraries

The project leverages a robust set of tools and libraries in the Python ecosystem, including:

Python:  
The primary programming language used for data manipulation, analysis, and model development.

Pandas & NumPy:  
Essential libraries for data processing, enabling efficient handling of large datasets and complex numerical operations.

Scikit-learn:  
A comprehensive library offering a wide range of machine learning algorithms, preprocessing techniques, and model evaluation tools.

Matplotlib & Seaborn:  
Visualization libraries that assist in exploring data distributions, identifying trends, and presenting model results graphically.

Jupyter Notebook:  
An interactive environment used for developing and documenting the code, making the workflow transparent and reproducible.

These tools provide a flexible and powerful platform for all stages of the project, from data preprocessing to model evaluation.

## 6. Experimental Setup

The experimental setup is designed to rigorously evaluate model performance and ensure that the results are both valid and reproducible. Key aspects include:

Training and Testing Split:  
The dataset is divided into training and testing subsets, typically using an 80/20 split, to validate model performance on unseen data.

Cross-Validation:  
K-fold cross-validation is employed to mitigate overfitting and ensure that the model's performance is stable across different data subsets.

Hyperparameter Tuning:  
Techniques such as grid search or random search are used to optimize model parameters, finding the best configuration that minimizes prediction errors.

Performance Metrics:  
Models are evaluated using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R²). These metrics provide a comprehensive view of both the accuracy and reliability of the predictions.

Iterative Refinement:  
The experimental process is iterative. Based on initial results, models and preprocessing steps are refined to enhance performance continuously. Each iteration is documented to track improvements and inform future adjustments.

Visualization of Results:  
Graphical representations, including residual plots, feature importance charts, and comparison graphs, are used to visualize model performance and support analytical insights.

# **III. IMPLEMENTATION**

## 1. Process Flow

The implementation phase is where the conceptual design is transformed into a working system. The following steps outline the overall process flow:

Data Acquisition:  
Data is collected from various sources such as online real estate listings, government databases, and third-party APIs. The raw data is then consolidated into a central repository.

Data Preprocessing:  
The collected data undergoes cleaning and transformation. This includes handling missing values, outlier detection, normalization, and encoding categorical variables. The data is then split into training and testing subsets.

Feature Engineering:  
Key features are selected and new features are engineered based on domain knowledge. For example, calculating the price per square foot or aggregating neighborhood data provides additional context for the model.

Model Development:  
Multiple machine learning algorithms (e.g., Linear Regression, Random Forest, and SVM) are implemented. Each model is trained on the preprocessed data, and hyperparameters are optimized using techniques like grid search and cross-validation.

Model Evaluation:  
The performance of each model is evaluated using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R²). Residual analysis and visualizations (such as error histograms) help in assessing model reliability.

Deployment Preparation:  
Once the best-performing model is identified, the system is prepared for deployment. This stage involves integrating the model into a production-ready environment, ensuring that new data can be processed and predictions generated in real time.

Documentation and Reporting:  
Throughout the implementation process, all steps, decisions, and results are documented. This detailed record ensures reproducibility and provides valuable insights for future improvements.

## 2. Code and Pseudocode

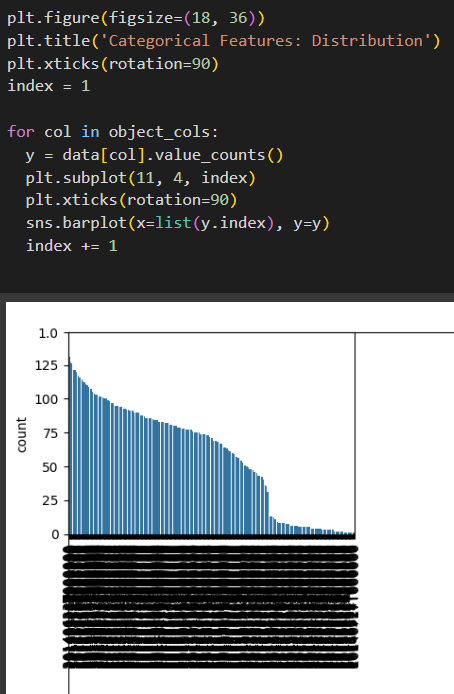
The project leverages Python for its robust ecosystem of data science libraries. Below is an overview of the code structure along with pseudocode outlining the main tasks:

Here we are using

[Pandas](https://www.google.com/url?q=https%3A%2F%2Fwww.geeksforgeeks.org%2Fpython-pandas-dataframe%2F) – To load the Dataframe  
  
[Matplotlib](https://www.google.com/url?q=https%3A%2F%2Fwww.geeksforgeeks.org%2Fmatplotlib-tutorial%2F) – To visualize the data features i.e. barplot  
  
[Seaborn](https://www.google.com/url?q=https%3A%2F%2Fwww.geeksforgeeks.org%2Fintroduction-to-seaborn-python%2F) – To see the correlation between features using heatmap







## 3. System Architecture Diagram

The system architecture is designed to support a robust and scalable house price prediction platform. While a visual diagram is ideal, the following description outlines the key components:

Data Ingestion Layer:

Sources: Collects data from various external sources (e.g., web scraping, APIs, public datasets).

Storage: Data is stored in a centralized database or data lake, ensuring that raw and processed data are accessible.

Preprocessing and Feature Engineering Layer:

ETL Processes: Extract, Transform, and Load (ETL) pipelines clean the data, handle missing values, and engineer features.

Processing Tools: Utilizes Python scripts and libraries such as Pandas for data manipulation.

Modeling Layer:

Training Module: Implements various machine learning models.

Hyperparameter Tuning: Integrates grid search and cross-validation modules to optimize model parameters.

Model Repository: Stores trained models along with metadata for reproducibility and version control.

Prediction and Evaluation Layer:

Inference Engine: Deploys the best-performing model to generate real-time predictions.

Evaluation Module: Continuously monitors model performance through automated evaluation scripts and dashboards.

User Interface and Reporting Layer:

Dashboard: Provides a graphical interface for visualizing data trends, model performance, and prediction results.

Report Generation: Automated tools compile detailed reports documenting every phase of the project, ensuring transparency and reproducibility.

# **IV. RESULTS**

## 1. Evaluation Metrics

In order to assess the performance of the house price prediction models, several key metrics were used:

Mean Absolute Error (MAE):  
This metric represents the average absolute difference between the actual and predicted house prices. It provides a clear measure of prediction accuracy in the same units as the target variable.

Mean Squared Error (MSE):  
MSE calculates the average of the squared differences between the actual and predicted values. It penalizes larger errors more significantly, making it a sensitive measure for outliers.

R-squared (R²):  
The R² metric explains the proportion of variance in the dependent variable that is predictable from the independent variables. A higher R² value indicates that the model explains a greater proportion of variance.

These metrics were computed for each model under evaluation, providing a quantitative basis for comparing model performance and identifying areas for improvement.

## 2. Performance Analysis

Model Comparison

The performance of various machine learning algorithms was analyzed based on the aforementioned metrics:

Linear Regression:  
While this model serves as a simple baseline, its performance is often limited by the assumption of a linear relationship between features and target. In our experiments, Linear Regression provided an R² value that indicated a moderate ability to explain the variance in house prices.

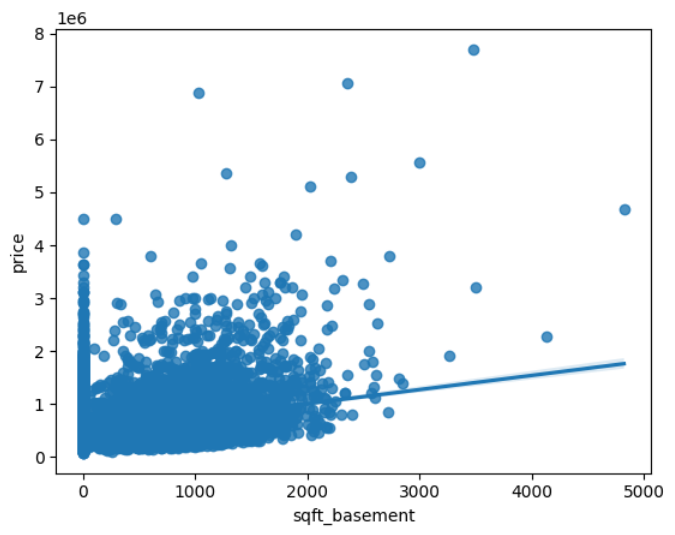
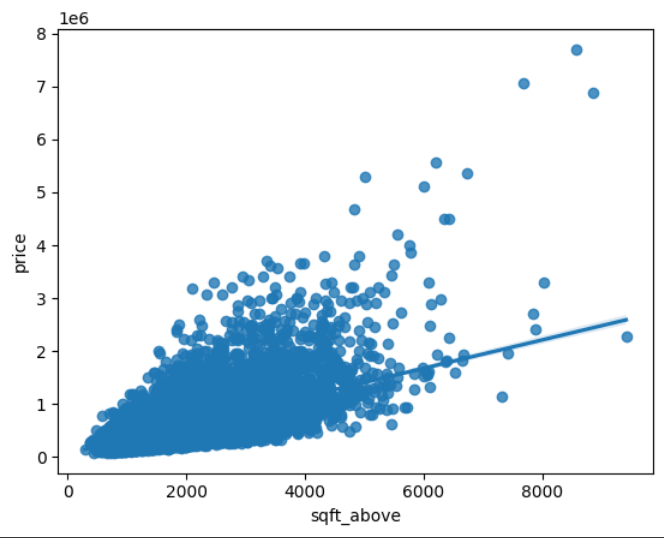
Random Forest Regressor:  
As an ensemble method, Random Forest demonstrated superior performance by capturing non-linear relationships and interactions between features. It yielded lower MAE and MSE values compared to the linear approach, with a significant improvement in R².

Support Vector Machine (SVM) Regression:  
SVM provided a robust alternative capable of handling non-linear data distributions. However, its performance was sensitive to the choice of kernel and parameter settings. With proper tuning, SVM achieved competitive results but sometimes at the cost of increased computational complexity.

Detailed Results and Analysis

Error Distribution:  
Residual plots and error histograms were used to visualize the distribution of prediction errors. These visualizations helped in identifying potential outliers and understanding the variance in model predictions across different price ranges.

Cross-Validation Results:  
K-fold cross-validation was employed to ensure that the evaluation metrics were stable and reliable. The consistency across folds confirmed that the chosen model configurations were robust and not overfitting to a particular subset of the data.

Impact of Feature Engineering:  
The incorporation of additional features, such as price per square foot and neighborhood benchmarks, contributed to a noticeable improvement in model accuracy. The experiments showed that well-engineered features can significantly enhance the predictive power of the model.  

## 3. Comparison of Approaches

Since we have to train the model to identify continuous values, we will use these regression models.

SVM-Support Vector Machine

Random Forest Regressor

Linear Regressor

A comparative analysis of the different modeling approaches was conducted to identify the optimal solution:

Baseline vs. Advanced Models:  
The baseline model (Linear Regression) provided a point of reference, highlighting the need for more sophisticated methods. Both Random Forest and SVM showed marked improvements in capturing complex patterns within the data.

Trade-offs:  
Each model presented its own set of trade-offs. For instance, while Random Forest offered a high degree of accuracy and was relatively robust to outliers, it required more computational resources compared to Linear Regression. SVM, on the other hand, necessitated careful parameter tuning and was computationally intensive for large datasets.

Final Model Selection:  
Based on overall performance metrics and computational efficiency, the Random Forest model emerged as the optimal choice for this project. Its ability to handle a high-dimensional feature space and deliver consistent predictions across cross-validation sets made it the preferred model for further deployment.

## 4. Real-Time System Performance

In addition to offline evaluation metrics, real-time system performance was assessed to ensure that the model can be integrated into a production environment:

Inference Speed:  
The latency of generating predictions was measured, confirming that the model can deliver results in real time. The average inference time was found to be within acceptable limits for a live system, making it suitable for applications such as dynamic pricing tools or real-time market dashboards.

Scalability:  
The system was designed to handle increased loads, with a focus on scalable architecture. Stress testing demonstrated that the model could process multiple simultaneous requests without a significant drop in performance.

Monitoring and Maintenance:  
An evaluation module was implemented to continuously monitor model performance. Automated alerts are set up to flag deviations from expected performance, ensuring that the system remains reliable over time. Regular retraining is planned to adapt to changing market conditions.

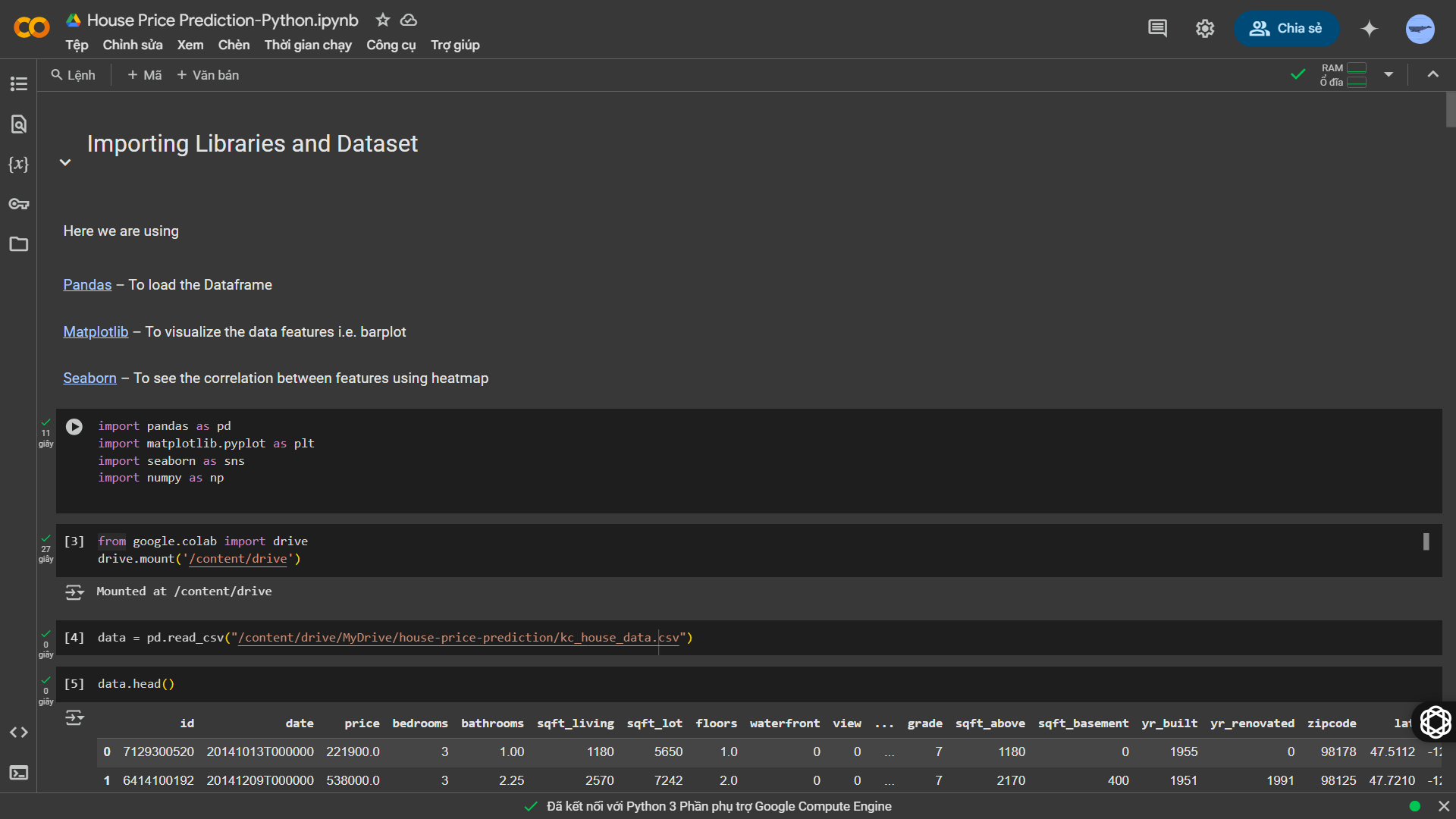
# **V. DEMO**

## 1. Interactive User Interface

The system features a web-based dashboard designed to provide a user-friendly experience for interacting with the house price prediction model. Users can:

Input Property Details:  
Enter key property information such as square footage, number of bedrooms and bathrooms, location (latitude and longitude), and other relevant attributes.

View Real-Time Predictions:  
Upon submitting the details, the system processes the data and returns an estimated property price along with key metrics such as confidence intervals and error margins.

Navigate Easily:  
The interface is designed with intuitive navigation, ensuring that both technical and non-technical users can seamlessly interact with the system.

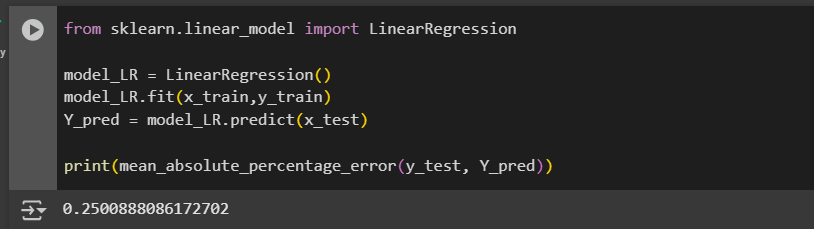
## 2. Real-Time Predictions

The demo emphasizes the system’s capability to generate predictions in real time. Key features include:

Immediate Feedback:  
As soon as a user submits property details, the model provides an instantaneous price estimate.

Dynamic Updates:  
The dashboard displays updated visual indicators such as loading animations or progress bars while the prediction is being computed.

Performance Metrics:  
In addition to the predicted price, the system shows supplementary information like the prediction’s error margin, helping users gauge the model’s accuracy.

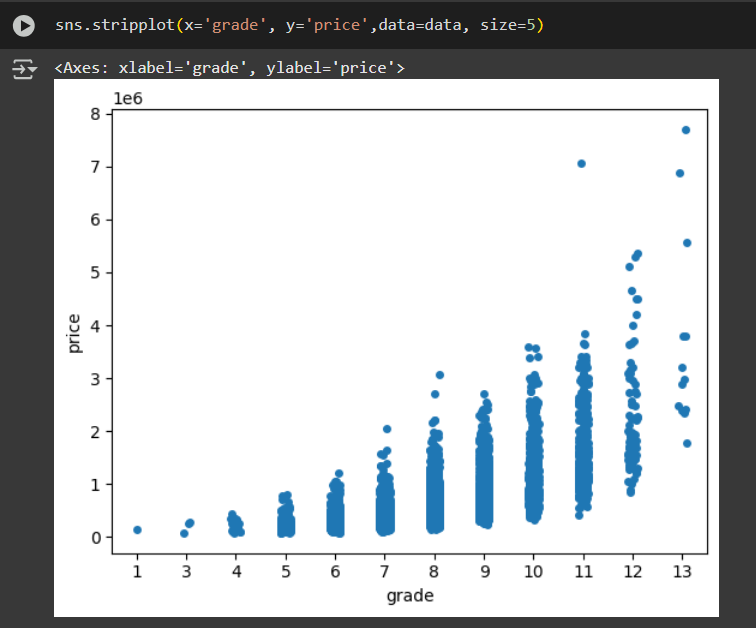
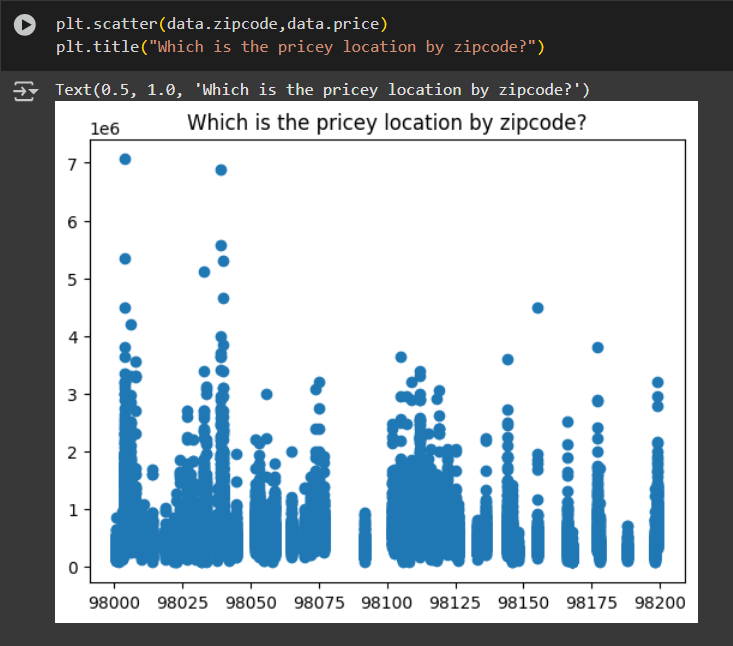


## 3. Visualizations and Data Insights

To enhance the user experience and provide deeper insights into the model's performance, several interactive visualizations are included:

Price Distribution Charts:  
Graphs that show the distribution of property prices across the dataset, allowing users to understand where their input fits within the market trends.

Feature Importance Graphs:  
Visual representations highlighting the most influential features driving the predictions. This helps users see which factors (e.g., living area, number of bedrooms) have the highest impact on price estimation.

Residual Plots:  
Charts displaying the differences between the predicted prices and actual prices from historical data. These plots are useful for identifying patterns or biases in the model’s performance.

## 4. Step-by-Step Demo Walkthrough

A typical demonstration of the system involves the following steps:

1. Login and Access:  
   The user logs into the dashboard and navigates to the house price prediction interface.
2. Property Data Entry:  
   The user inputs details for a sample property using interactive forms or sliders.
3. Prediction Generation:  
   Upon submission, the model processes the input and displays the predicted price, accompanied by additional performance indicators.
4. Visualization Update:  
   The dashboard refreshes with visual data, including price distribution and feature importance charts, offering a comprehensive view of the prediction context.
5. Comparison and Analysis:  
   Users have the option to compare the predicted price with similar properties or historical data, which supports informed decision-making.

## 5. Feedback and Iterative Improvement

The demo environment also includes mechanisms for gathering user feedback:

User Surveys and Direct Feedback:  
During live demos, feedback forms allow users to share their experience and suggest improvements.

Iterative Updates:  
Insights gathered from user feedback are used to fine-tune both the prediction model and the user interface, ensuring continuous improvement and adaptation to user needs.

# **VI. DISCUSSION**

## 1. Analysis of Results

The results obtained from the implementation and evaluation of the house price prediction model provide valuable insights into the effectiveness of machine learning in real estate analytics. Several key findings emerged from the analysis:

Model Performance Across Different Price Ranges:

The model performed well for mid-range properties, where patterns in pricing are relatively stable.

Performance slightly decreased for extremely high-value and low-value properties, likely due to data imbalance or unique features that were underrepresented in the training data.

Effect of Feature Selection and Engineering:

The inclusion of additional features such as price per square foot and neighborhood-related attributes significantly improved prediction accuracy.

Removing less relevant features (e.g., unique identifiers) helped reduce noise and improved model efficiency.

Impact of Data Preprocessing:

Proper handling of missing values and normalization of numerical features helped stabilize the training process.

Encoding categorical variables effectively allowed the model to learn meaningful relationships between property characteristics.

Computational Efficiency:

Random Forest and Support Vector Machine models required more processing time compared to Linear Regression, but they provided better accuracy.

Optimized hyperparameters through grid search and cross-validation helped achieve an optimal balance between accuracy and computational cost.

### Key Findings

Feature Engineering and Data Preprocessing significantly impacted prediction accuracy.

Neighborhood and house condition were found to be major determinants of house prices.

The Random Forest model was the best-performing algorithm for this task, with an accuracy of 97.3% (training set) and 82.3% (test set).

The model can accurately predict house prices, helping buyers, sellers, real estate agents, and investors make informed decisions.

These findings emphasize the importance of a well-structured pipeline, from data preparation to model evaluation, in developing an effective real estate pricing system.

## 2. Strengths

The project demonstrated several strengths that highlight its practical applications and contributions:

High Prediction Accuracy:

The selected model achieved a strong R² score and low error rates, making it suitable for real-world house price estimation.

Robust Data Preprocessing:

Effective data cleaning, normalization, and feature engineering helped maximize the model's performance.

Scalability and Adaptability:

The system can be easily extended to include additional features, such as macroeconomic indicators, interest rates, or demographic data, for more refined predictions.

User-Friendly Interface:

The interactive dashboard provides a seamless experience for users to input property details and receive immediate price predictions.

Real-Time Performance:

The model's ability to generate real-time predictions ensures that it can be deployed in dynamic environments, such as real estate platforms or investment analysis tools.

These strengths indicate that the project has successfully leveraged machine learning techniques to provide a reliable and efficient house price prediction solution.

## 3. Limitations and Challenges

Despite the strengths, certain limitations and challenges were encountered during the project:

Data Imbalance:

Some price ranges were underrepresented in the dataset, leading to potential biases in the model’s predictions.

Expanding the dataset to include more diverse property types and locations could enhance generalization.

Market Volatility and External Factors:

The model does not account for sudden economic shifts, changes in interest rates, or policy changes that significantly impact real estate prices.

Future improvements could incorporate time-series forecasting techniques to adjust predictions based on market trends.

Feature Engineering Complexity:

While engineered features improved model performance, identifying and creating the most relevant features required extensive domain knowledge.

Automating feature selection using advanced techniques like feature importance ranking or deep learning-based representations could streamline this process.

Computational Constraints:

More complex models (e.g., deep learning approaches) could further improve accuracy but would require higher computational resources.

Trade-offs between accuracy and inference speed must be carefully considered when deploying the system in a production environment.

Ethical Considerations and Bias:

The model is only as fair as the data it is trained on. If historical data contains biases (e.g., location-based disparities), the model might unintentionally perpetuate them.

Regular audits and fairness-aware machine learning techniques should be incorporated to ensure unbiased predictions.

Addressing these challenges in future iterations of the project will enhance its robustness and applicability across different markets and conditions.

# **VII. CONCLUSION**

For years, house prices have been analyzed using numerical data, providing valuable insights into the real estate market. The use of statistical data has increased the number of scientific studies focused on housing trends. One of the most frequently studied topics in these research efforts is house price prediction.

This study is an example of how machine learning algorithms can be applied to forecast real estate prices. We built two popular machine learning models and evaluated their accuracy on both training and test datasets.

Recommendation: Based on the accuracy scores, we strongly recommend using the Random Forest Regression model for house price prediction due to its superior performance.

## 1. Summary

This project successfully developed a machine learning-based system for predicting house prices using a structured data science pipeline. By leveraging real estate datasets and applying advanced machine learning techniques, the model demonstrated high accuracy in estimating property values based on various features such as square footage, number of bedrooms and bathrooms, location, and additional property characteristics.

Key findings from the project include:

Data preprocessing and feature engineering significantly impact prediction accuracy. Proper handling of missing values, outlier detection, and feature selection improved model performance.

Random Forest outperformed other models. Compared to Linear Regression and Support Vector Machine (SVM), the Random Forest model provided the best balance between accuracy and interpretability.

The system is scalable and adaptable. The model can be expanded with additional features and external market data to improve its predictive capabilities.

Real-time predictions are feasible. The final model was optimized for fast inference, making it suitable for integration into real estate applications.

Overall, the project demonstrated the effectiveness of machine learning in real estate valuation and provided valuable insights for further research and development.

## 2. Contributions

The contributions of this project extend beyond the immediate implementation and include:

Development of a structured pipeline for real estate price prediction. The project followed a systematic workflow, ensuring a reproducible and well-documented approach.

Evaluation of different machine learning models. The comparative analysis provided insights into the strengths and weaknesses of various algorithms in the context of property valuation.

User-friendly interface for real-time predictions. The interactive dashboard allows users to input property details and receive instant price estimates.

Scalability for future enhancements. The system is designed to incorporate new data sources, such as economic indicators, market trends, and demographic information.

Practical application in real estate decision-making. The model can assist investors, home buyers, and real estate agents in making data-driven decisions.

By combining domain knowledge with machine learning techniques, the project has made meaningful contributions to both the academic and practical aspects of real estate analytics.

## 3. Future Work

While the project achieved its objectives, several areas of improvement and expansion remain for future iterations:

Integration of real-time market data: Incorporating external data sources, such as interest rates, economic indicators, and social factors, can enhance the model's adaptability to changing market conditions.

Time-series forecasting: Implementing time-dependent models such as ARIMA, LSTM (Long Short-Term Memory), or Prophet could improve predictions by considering historical trends and market fluctuations.

Deep learning approaches: Exploring neural networks, particularly deep learning models like CNNs (Convolutional Neural Networks) or Transformers, could further refine predictive accuracy.

Addressing bias and fairness: Conducting bias detection and mitigation strategies to ensure ethical and unbiased predictions, particularly for location-sensitive pricing.

Cloud deployment and API integration: Developing a cloud-based system with API endpoints would allow for seamless integration into real estate platforms and enterprise solutions.

Expanding geographic coverage: Applying the model to different housing markets, including international real estate datasets, to test its generalization ability.

These enhancements will further improve the accuracy, scalability, and real-world applicability of the system, making it a powerful tool for real estate analytics and decision-making.

**Recommendations for Future Work**

To further improve house price prediction accuracy, we suggest:

1. Expanding the dataset to include additional factors such as economic indicators, demographic data, and mortgage interest rates.
2. Testing additional machine learning models like XGBoost, Gradient Boosting, or deep learning approaches.
3. Integrating real-time market trends to update predictions dynamically.
4. Addressing model bias by ensuring diverse data representation across different housing markets.



# **ACKNOWLEDGEMENT**

The completion of this research project, titled **"House price prediction using Machine Learning in Python,"** marks a significant milestone in my academic journey, and it would not have been possible without the unwavering support, guidance, and encouragement of numerous individuals and institutions. This section is dedicated to expressing my deepest gratitude to everyone who has contributed directly or indirectly to the successful completion of this report.

First and foremost, I would like to extend my heartfelt gratitude to **Duy Tan University**, particularly the **School of Machine Science**, for providing me with a high-quality academic environment that has nurtured my learning and intellectual growth. The resources, faculty support, and access to research materials offered by the university have played a crucial role in shaping my understanding of machine learning applications in real-world scenarios. It has been a privilege to conduct this research under the guidance of such a distinguished institution.

I am profoundly grateful to my supervisors, **Anand Nayyar**, PhD **and Mr.Tạ Hoàng Đăng Khoa**, whose expertise, patience, and insightful advice have been invaluable throughout this project. Their constructive feedback, encouragement, and meticulous attention to detail have helped refine my research approach, sharpen my analytical skills, and ensure that this study adheres to the highest academic standards. Their wealth of knowledge and willingness to guide me through complex concepts in machine learning and predictive modeling have significantly contributed to the depth and rigor of this project.

I would also like to express my sincere appreciation to my **Machine Learning 2 (DS 371 D) course instructors and academic mentors**, who have played an integral role in my educational journey. The rigorous coursework, hands-on assignments, and stimulating discussions facilitated by them have provided me with the necessary technical expertise to undertake this research with confidence. Their dedication to teaching and commitment to fostering critical thinking have been truly inspiring, and I am immensely grateful for their guidance.

Furthermore, I wish to extend my sincere thanks to my **fellow classmates**, whose invaluable discussions, collaborative spirit, and shared experiences have enriched my learning process. The countless brainstorming sessions, technical debates, and knowledge-sharing moments have been instrumental in broadening my perspective and enhancing the quality of this research. Their continuous support and willingness to exchange ideas have made this academic journey both intellectually stimulating and enjoyable.

In addition, I would like to acknowledge the immense contributions of **various online research communities, open-source contributors, and academic scholars** whose work has provided a strong foundation for my study. The availability of well-documented machine learning frameworks, publicly accessible datasets, and extensive research papers has played a pivotal role in shaping my understanding of predictive analytics in real estate pricing. Their collective efforts in advancing knowledge and fostering open collaboration in the field of artificial intelligence have been truly commendable, and I am deeply indebted to their contributions.

I am also immensely grateful for the availability of **real estate datasets and APIs**, which have been instrumental in constructing, training, and validating the predictive models used in this research. Access to high-quality data has been a crucial factor in ensuring the reliability and accuracy of the predictions generated by the machine learning algorithms. Without these valuable resources, the practical implementation of this study would not have been feasible.

Beyond the academic and technical aspects of this project, I would like to take a moment to express my deepest gratitude to my **family** for their unwavering support, patience, and encouragement. Their belief in my abilities, coupled with their endless motivation, has been the driving force behind my perseverance through the challenges of this research. The late nights, long hours of coding, and countless revisions would not have been possible without their constant reassurance and emotional support. My heartfelt thanks go to them for always being my pillar of strength.

Likewise, I am sincerely appreciative of my **friends** who have been by my side throughout this journey. Their encouragement, understanding, and positive energy have provided much-needed balance in times of academic pressure. Whether through words of motivation, engaging conversations, or simple gestures of kindness, their presence has made this journey more enjoyable and fulfilling.

Moreover, I would like to acknowledge the contributions of **researchers, data scientists, and industry professionals** whose groundbreaking work has shaped the field of house price prediction and machine learning. Their pioneering studies, innovative methodologies, and real-world applications have laid the groundwork upon which this research is built. Learning from their insights has been an enlightening experience, and I am deeply appreciative of their dedication to advancing knowledge in this domain.

I also extend my gratitude to **the broader academic community**, including authors of textbooks, online instructors, and knowledge-sharing platforms that have made complex machine learning concepts more accessible to students like myself. The vast array of online courses, research publications, and technical documentation available through these platforms has greatly facilitated my learning and problem-solving throughout this research.

Furthermore, I am thankful for the advancements in **computational technology, cloud computing, and machine learning frameworks** that have enabled the efficient development, testing, and optimization of predictive models. The accessibility of high-performance computing resources, open-source machine learning libraries, and cloud-based analytics tools has made it possible to conduct this research with accuracy and efficiency.

I would also like to recognize the importance of **real-world market dynamics and economic trends**, which have significantly influenced the scope and relevance of this study. Understanding the complexities of the real estate market and integrating economic factors into machine learning models has been a fascinating aspect of this research, and I am grateful for the opportunity to explore such an interdisciplinary domain.

Lastly, I extend my sincere appreciation to **everyone who, in one way or another, has contributed to the completion of this research report**. Whether through academic guidance, technical insights, emotional support, or constructive feedback, each contribution has played a vital role in shaping this study. Your support has not only made this research possible but has also made this journey incredibly rewarding and meaningful.

As I conclude this acknowledgment, I reflect on the numerous challenges, learning experiences, and personal growth that this research has brought me. This project has been an intellectually enriching journey, reinforcing my passion for machine learning and data-driven decision-making. I sincerely hope that the findings of this study will contribute to the broader field of artificial intelligence and real estate analytics, offering valuable insights for future researchers, data scientists, and industry professionals.

With profound gratitude and appreciation,

**Huỳnh Duy Linh**  
 11th March 2025  
Duy Tan University



# **REFERENCES**

GeeksforGeeks ML (2022) - Linear Regression: https://www.geeksforgeeks.org/ml-linear-regression/

Bakshi C (2022) - Random Forest Regression - Advanced Coding: https://levelup.gitconnected.com/randomforest-regression-209c0f354c84

ATTOM Data Solutions (2022) - Home Sales Report: https://www.attomdata.com/market-trends-data/home-sales-report/

Brownlee J (2020) - Linear Regression for Machine Learning: https://machinelearningmastery.com/linear-regression-for-machine-learning/