A Mini Project Report

On

HOUSE PRICE PREDICTION

Submitted in partial fulfillment of the requirement for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE & BUSINESS SYSTEMS

Submitted by:

Student Name 1: LAKSHYA SHARMA	**University Roll No.: A122**
Student Name 2: AASHIR SHAIKH	**University Roll No.: A106**
Student Name 3: SATYAM KARDE	**University Roll No.: A116**

School of Technology Management & Engineering
SVKM's NMIMS Deemed to be University
Navi Mumbai
November 2022

CANDIDATE’S DECLARATION

We hereby certify that the work which is being presented in the Synopsis entitled **“House Price Prediction”** in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineering **(Computer Science and Business Systems)** in the Department of Computer Science and Engineering of the School of Technology Management & Engineering, SVKM’s NMIMS Deemed to be University, Navi Mumbai shall be carried out by the undersigned under the supervision of **Guide Name, Designation**, Department of Computer Science and Engineering, School of Technology Management & Engineering, SVKM’s NMIMS Deemed to be University, Navi Mumbai.

Lakshya Sharma	A122	Signature
Aashir Shaikh	A106	Signature
Satyam Karde	A116	Signature

The above-mentioned students have worked under the supervision of the undersigned on the

“HOUSE PRICE PREDICTION”

Signature

Examiner

Internal Evaluation (By DPRC Committee)

Status of the Report: Accepted / Rejected

Any Comments:

Name of the Committee Members: **Signature with Date**

1.

2.

3.

Table of Contents

Chapter 1

Introduction and Problem Statement

1.1 Introduction

The real estate market is a vital component of economic activity, influencing housing decisions, investments, and serving as a key economic and social indicator. Among its many facets, the determination of housing prices profoundly influences the choices made by homebuyers, sellers, and investors. The accuracy of these house price predictions is of utmost importance, affecting not only individual property decisions but also guiding the actions of policymakers, financial institutions, and real estate professionals navigating this dynamic landscape [1].

Traditionally, forecasting house prices relied on the expertise of real estate agents, historical data, and a limited set of influencing factors. However, in today's era marked by rapid urbanization, changing demographics, and dynamic market dynamics, these traditional methods often lack the precision and reliability necessary for informed property-related decisions [2]. As a result, there is a growing demand for data-driven approaches that harness modern technology to improve the accuracy and accessibility of house price predictions.

This project focuses on house price prediction through the application of Machine Learning (ML) techniques. Machine Learning has emerged as a powerful tool for uncovering complex patterns and relationships within extensive datasets. By utilizing ML algorithms such as Naïve Bayesian, Random Forest, Support Vector Machines (SVM), and Gradient Boosting, this study aims to develop a robust and efficient property price prediction model [3].

Our research is founded on the identification of key factors influencing house prices, with particular attention to the impact of location and property size. Location, often deemed the cornerstone of real estate, encompasses factors like proximity to amenities, crime rates, school quality, and transportation options, all of which substantially affect property values [4]. Similarly, property size, including attributes such as the number of bedrooms, bathrooms, and square footage, plays a fundamental role in determining market prices.

The dynamic nature of the real estate market presents unique challenges for accurate price predictions. Fluctuations in supply and demand, shifts in economic conditions, and changes in buyer preferences necessitate the development of models capable of adapting to these real-time dynamics. Furthermore, as urbanization continues to drive migration and reshape housing markets, the traditional methods relying on real estate brokers fall short in providing timely and precise property valuations [5].

In response to these challenges, our proposed model aims to serve as a valuable tool for homeowners, prospective buyers, investors, and policymakers, facilitating informed decisions in the ever-evolving real estate landscape. This Research will contain literature review conducted to get familiarized with the existing solutions, Proposed methodology, Project Plan, Results and Analysis and finally conclusion on the model chosen and its future prospects.

1.2 Problem Statement

The central issue addressed in this report pertains to the pressing need for precise forecasts of residential property prices. This necessity is a direct result of the intricate nature of the real estate market, characterized by its frequent price fluctuations. This challenging and ever-changing market environment presents a significant hurdle for a range of stakeholders, including prospective buyers, sellers, and investors.

A noteworthy contributing factor that emphasizes the significance of this problem is the ongoing trend of urbanization. As an increasing number of individuals make the transition from rural to urban areas in pursuit of enhanced opportunities and improved living conditions, they frequently encounter unfamiliar terrain, leading to feelings of insecurity and vulnerability. This group, in particular, requires access to dependable price predictions to facilitate well-informed decisions regarding property transactions.

In the conventional approach to property transactions, individuals often place their trust in real estate brokers to navigate them through the process. However, this conventional method doesn't consistently deliver precise price estimates. Relying on brokers introduces an element of subjectivity, and there exists the potential risk of encountering deceptive activities or misrepresentation.

Consequently, the central issue under consideration in this report pertains to the development of a model that is both accurate and trustworthy for predicting property prices. This model will make use of various input factors, including location and size, to provide stakeholders with a reliable tool to aid them in making informed decisions in the dynamic and occasionally uncertain real estate market. The establishment of such a model is of utmost importance in enhancing transparency and ensuring dependable property price forecasts, thus benefiting individuals, the real estate sector, and the broader economy.

Chapter 2

Literature Survey

In this Literature survey we have gone through similar works in the area of House Price Prediction.

Numerous factors, including location and neighborhood, have an impact on house prices. However, these elements must be chosen carefully because they have a significant impact on the anticipated price of a home. After reading many articles on machine learning-based house price prediction and examining historical data on Kaggle, we discovered that the three most important criteria influencing house price are location, size, and number of rooms.

Among models, there were various instances of using classifying models like SVM and Bayesian for prediction and getting boosted accuracy. One such study by Y. Chen, R. Xue and Y. Zhang [6] examined Linear Regression, Bayesian models, Support Vector Machines (SVM), BP neural networks and Deep Neural Networks (DNN) and the research findings indicate that Bayesian models, BP neural networks, and SVM are more effective in predicting house prices. However, suboptimal model training results showed that SVM despite having a very high r-squared value also produced the largest errors also on a small dataset.

Another paper by T. Danh Phan [7] highlighted the use of tuned SVM, demonstrating competitive predictive performance in his paper, reported overfitting issues on data reduction with Principal Component Analysis (PCA), an unsupervised approach. Thus it is clear that our dataset which closely resembles the one used by Danh Phan will require supervised feature selection to achieve lower errors and higher accuracies.

Additionally, along with T. Dahn Phan [7] and Y. Chen, R. Xue and Y. Zhang [6], G. K. Kumar, D. M. Rani, N. Koppula and S. Ashraf [8] have also noted that deep learning models don’t work well with house price prediction as they require huge amounts of data, extensive data training and have low interpretability in a model designed around finding how predictors affect house prices. Thus, as noted in one of the above research papers [8], traditional models like gradient boosting algorithms are the most efficient and least error driven regression technique when provided with historical data. However inadequate information regarding data preprocessing and feature selection may potentially mean the existence of bias in above results. Similarly Tree Regression techniques, like Random Forest [9], etc. are other such traditional techniques suitable for house price prediction.

Chapter 3

Objectives

- The primary objective of this project is to develop and fine-tune predictive models to achieve higher levels of accuracy in forecasting house prices.
- Another key goal is to identify the most relevant features that exert a strong influence on house prices. Additionally, we aim to explore the creation of new features that could provide additional predictive power, enhancing the overall model's performance.
- Our project aims to evaluate and compare various regression algorithms, encompassing a diverse range of approaches, to pinpoint the most suitable model for accurately predicting house prices based on our dataset.
- Lastly, we plan to implement a user-friendly interface that incorporates user feedback. This interface will enable users to input specific property details and receive predicted house prices, making the predictive tool accessible and practical for a wide range of users.

Chapter 4

Hardware and Software Requirements

4.1 Hardware Requirements:

The Hardware Interfaces Required are -

Ram: Minimum 8GB or higher

GPU: 4GB dedicated

Processor: Intel Pentium 4 or higher

HDD: 10GB or higher

Monitor: 15” or 17” color monitor

Mouse: Scroll or Optical Mouse or Touchpad

Keyboard: Standard 110 keys keyboard

4.2 Software Requirements:

Operating System: Windows, Mac, Linux

Software Development Kit: Jupyter Notebook

Python3 installed with flask, scikit-learn.

Chapter 5

Possible Approach /Architecture/ Algorithms

5.1 Approach:

In this project we are building a model which predicts house prices based on various features.

- First the data is collected from a trustworthy site which is kaggle and explored to understand its composition, look for any missing data, and get knowledge of the feature distribution.
- Uneven and missing data can be dealt with by dropping columns with a high number of missing values, or filling null values in textual columns with unknown data, or imputing mean and modes in numeric data.
- Outliers can be dealt with proper methodology after understanding them and then applying methods like IQR on respective columns which require it.
- Standardizing the data is done using StandardScaler to bring all the features to a similar scale during feature engineering to better train our Gradient-Descent and Forest Regressor based models.
- Gradient-descent, SVM and Random Forest are Regression models which can be used based on our research.
- Label encoding will help our data to fit the models without giving errors or wrong output.
- Approaches taken in evaluating the models include using metrics like R-squared, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).

5.2 Architecture:

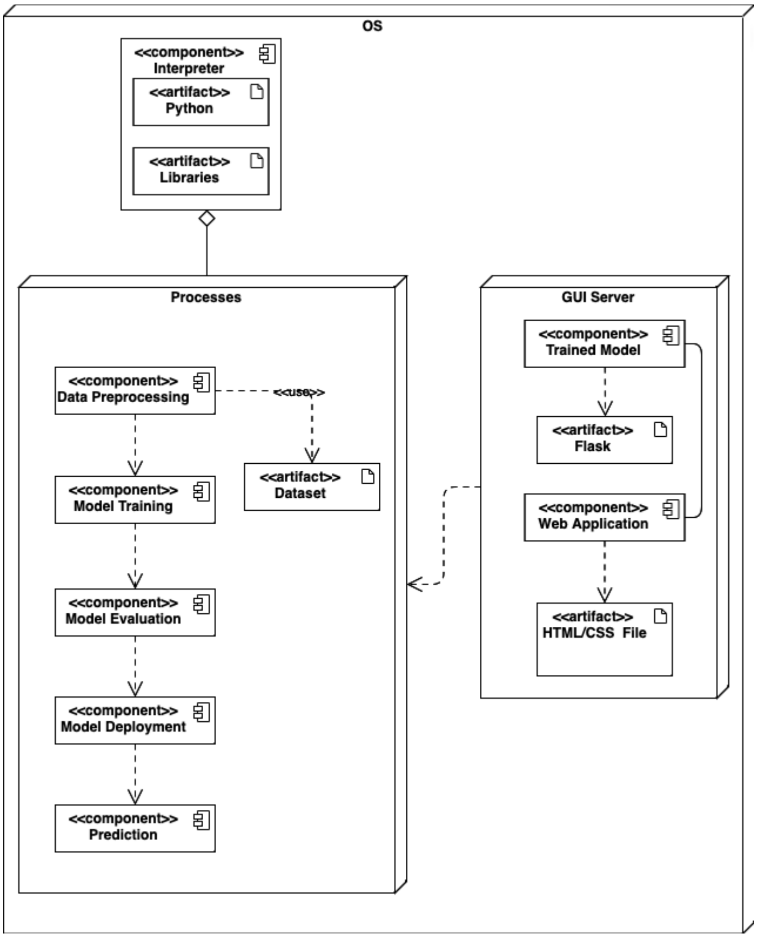

OS
<<component>> Interpreter
<<artifact>> Python
<<artifact>> Libraries
Processes
<<component>> Data Preprocessing
<<use>>
<<artifact>> Dataset
<<component>> Model Training
<<component>> Model Evaluation
<<component>> Model Deployment
<<component>> Prediction
GUI Server
<<component>> Trained Model
<<artifact>> Flask
<<component>> Web Application
<<artifact>> HTML/CSS File

Fig.5.2 Architecture of House Price Prediction System

1. OS Node:

- Operating System (OS): Our house price forecast model's Operating System (OS) serves as its structural backbone. It effectively manages hardware resources, enables software collaboration, and keeps an eye on crucial operations like memory management and job scheduling. This guarantees a stable environment for data processing, model training, and prediction, assuring system reliability.

2. Processes Node:

- Processes: Processes are dynamic software entities within the OS architecture in our model for predicting home price changes. They are the heart of our system, cooperating to make sure the process of predicting house prices is simple and effective. They are crucial to the entire operation because of their responsibilities in real-time prediction, model deployment, evaluation, and data preprocessing.

- Data Preprocessing: A crucial part of our method for predicting property prices is data preprocessing. For efficient usage in training and prediction, it entails cleaning and processing raw data. The removal of features, standardization for consistency, and data cleaning to remove anomalies, missing values, and inconsistencies are important activities. Accurate, anomaly-free, and machine-learning ready datasets are ensured by this approach.

- Model Training: The Model Training stage is a turning point in the creation of our method for predicting home prices. Machine learning algorithms start to work in this stage, pulling useful patterns and insights from the painstakingly preprocessed dataset. These algorithms attempt to grasp the complex links between input variables and property values by iteratively learning and adapting while utilizing the prepared data. A variety of model selection, hyperparameter tuning, and cross-validation strategies are used during this phase to enhance the predictive model's precision and generalizability.

- Model Evaluation: After model training, the Model Evaluation procedure acts as the quality assurance stage, and it is a crucial checkpoint in our house price forecast system. Here, the effectiveness of the trained prediction model is thoroughly examined utilizing a wide range of unique measures and methodologies. The goal is to ensure the model's robustness and generalizability in addition to its correctness. We can find any potential flaws, overfitting, or underfitting issues by evaluating the model's performance against a variety of evaluation criteria. This stage is crucial for perfecting the model and ensuring its dependability for practical applications.

- Model Deployment: From development to actual implementation, our focus switches in Model Deployment. To democratize access to the trained predictive model, we frequently provide a web service or API, making sure it's user-friendly, safe, and expandable. In this phase, we turn our predicted insights into a real-time tool that can be used by people, companies, or applications.

- Prediction: Our system for predicting home prices reaches its apex during the Prediction step, where the deployed model assumes the lead in providing up-to-the-minute insights. By bridging the gap between data and useful information, this approach enables consumers to use the prediction model to help them decide. The deployed model uses the knowledge it has gained during training to provide accurate predictions of property prices as new data points are fed into it.

3. GUI Server Node:

- GUI Server: The graphical user interface (GUI) is controlled by the GUI Server, a separate element of our house price forecast system. It acts as a portal for user communication with the prediction model. Through this interface, users may provide data, obtain predictions, and customize their experience, guaranteeing a smooth and user-friendly connection with the system.

- GUI Application: The user-friendly face of our algorithm for predicting home prices is the GUI Application. Users may enter data, change parameters, and get forecasts with ease because of the straightforward interface it provides. Our prediction model's usability is improved by this component's simplified user-system interface, which is accessible to non-technical users and has interactive components and visualization tools.

- Web Browser: As the interface between the user and the GUI application located on the GUI server, the "Web Browser" stands for the client-side portion of our home price forecast system. It offers consumers a flexible and open platform for interacting with our application from any internet-connected device. Users can easily access the GUI programme through a web browser, enter data, and obtain predictions without the need for any installation or complicated setup.

- HTML, CSS, JS Files: The design of our house price forecast system's user interface relies heavily on HTML, CSS, and JS files. While JavaScript provides interaction, enhancing the user interface's responsiveness and usability, HTML organizes and shows information, CSS improves aesthetics and the user experience by defining style and layout.

- Flask: Our house price prediction method's frontend and backend are connected with the Python framework Flask. In addition to managing HTTP requests and answers, it acts as the central hub for data flow, user inputs, and model predictions. Model execution, real-time data processing, and user result display are all made possible by Flask's portability and flexibility.

4. Interpreter Component:

- Interpreter: Our approach for predicting home prices depends on the Interpreter, which executes code and speeds up modeling and data processing. Due to its versatility and popularity in data science and machine learning, Python is our preferred language. For data processing, machine learning activities, and system integration, it makes use of libraries like Pandas, NumPy, and scikit-learn. Effective data preparation, model training, and other crucial activities are made possible as a result.

- Libraries: Our efforts to estimate home prices rely heavily on libraries since they provide pre-written code modules that extend Python's functionality. Python's capabilities are expanded with modules like NumPy for numerical computation, Pandas for data management, and scikit-learn for machine learning. Effective data administration, processing, analysis, and the creation of machine learning algorithms for model testing and assessment are all made possible by these libraries.

5.3 Algorithm:

After properly cleaning the data and dealing with outliers, the next step is applying the algorithms to it. For that, we use supervised learning regression machine learning models such as Random Forest, SVM, and Gradient-Boosting. We can then fit the data to our models for training after label encoding the necessary columns and then normalizing the data.
The following algorithms are used:

5.3.1 Support Vector Machine (SVM)

In SVM, each data point is plotted in an n-dimensional space (n is the number of features) with the value of each feature being the value of a particular coordinate. The classification is done by finding a hyper-plane that differentiates the classes the best.
SVMs or Support Vector Machines are a sort of machine learning algorithm that can perform classification and regression tasks. Vladimir Vapnik and colleagues (Boser et al., 1992; Guyon et al., 1993; Cortes and Vapnik, 1995) developed it at AT&T Bell Laboratories. SVMs, which are based on statistical learning frameworks or VC theory introduced by Vapnik (1982, 1995) and Chervonenkis (1974), are one of the most resilient prediction approaches to use.

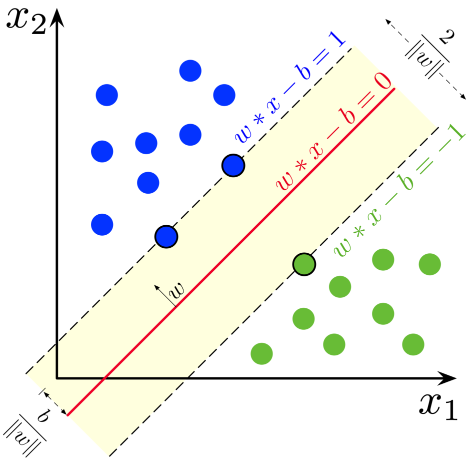


Fig 5.3.1 SVM

In order to successfully separate data points from distinct classes, Support Vector Machines (SVMs) locate a hyperplane within an N-dimensional space. This hyperplane functions as a flat surface, dividing space into two different zones. SVMs seek a hyperplane that maximizes the margin, which is defined as the distance between the hyperplane and the nearest data points from each class. For classifying upcoming data points, this larger margin provides more assurance. SVMs perform particularly well when dealing with multidimensional data and capturing nonlinear relationships. Furthermore, they show resistance to noise and outliers.

5.3.2 Random Forest

Tin Kam Ho introduced the model in 1995, which leverages the random subspace technique. Serving functions such as regression, classification, and more, it is categorized under ensemble learning techniques. During the training phase, the approach involves the development of multiple decision trees. The model delivers the mean or average forecast from these separate trees in regression tasks. Random decision forests mitigate decision trees' tendency to over-adapt to their training data, hence improving their performance.

Random forests are a good model for regression and house price prediction because of their ability to learn complex nonlinear relationships and because they can consider a wide range of elements that influence house pricing, such as the location of the house, its size, the number of bedrooms and bathrooms, and its condition.

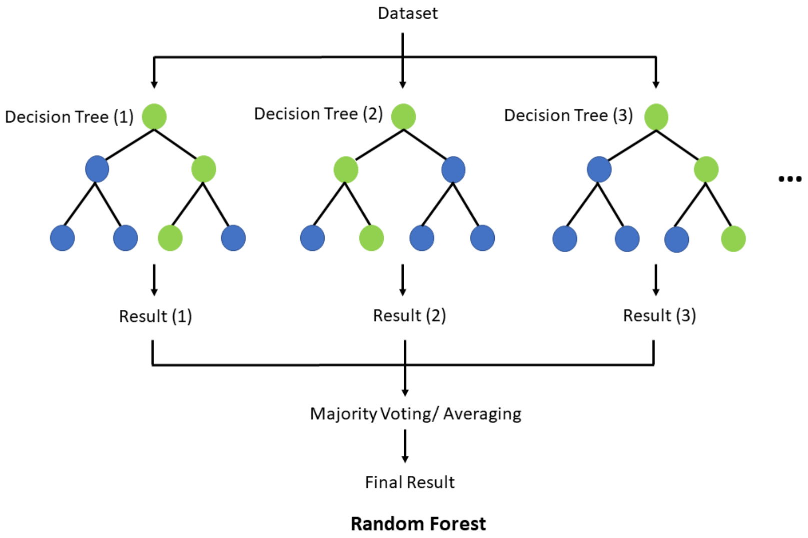


Fig 5.3.2 Random Forest

5.3.3 Gradient Boosting

Gradient boosting is an effective machine learning technique that may be used for both regression and classification tasks. This method creates a prediction model by combining a number of weak prediction models. These "weak" models, distinguished by their few data assumptions, are frequently expressed as simple decision trees. This ensemble technique improves the final model's predicted accuracy and robustness.

Gradient boosting works by iteratively building weak learners and adding them to the ensemble. All weak learners are trained to fix the mistakes made by previous weak learners. This procedure keeps going until the ensemble is able to use the training data to make accurate predictions.

Gradient boosting is a good algorithm for regression because it can learn complex nonlinear relationships between the input and output variables. It is also good at handling noisy data and outliers.

5.3.4 IQR (Interquartile Range)

The Interquartile Range (IQR) method for outlier detection entails several steps. It commences with the computation of the first and third quartiles, denoted as Q1 and Q3, from the dataset. Next, it identifies any data points that lie outside the bounds defined by Q1 - 1.5 * IQR and Q3 + 1.5 * IQR, with IQR representing the spread between Q1 and Q3. Any data point falling beyond this range is categorized as an outlier. This method proves useful for pinpointing unusual or anomalous values within a dataset.

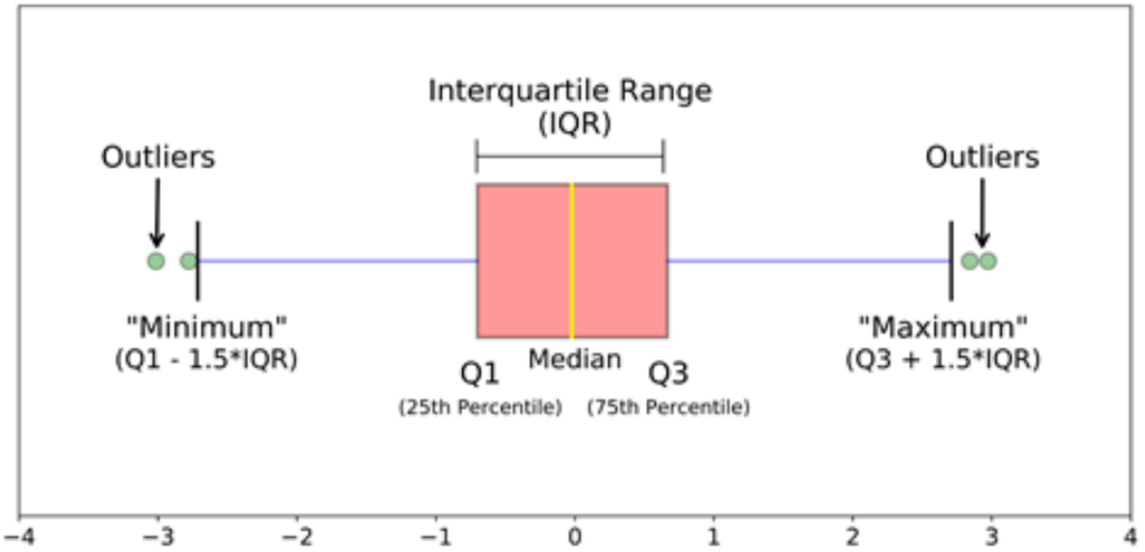
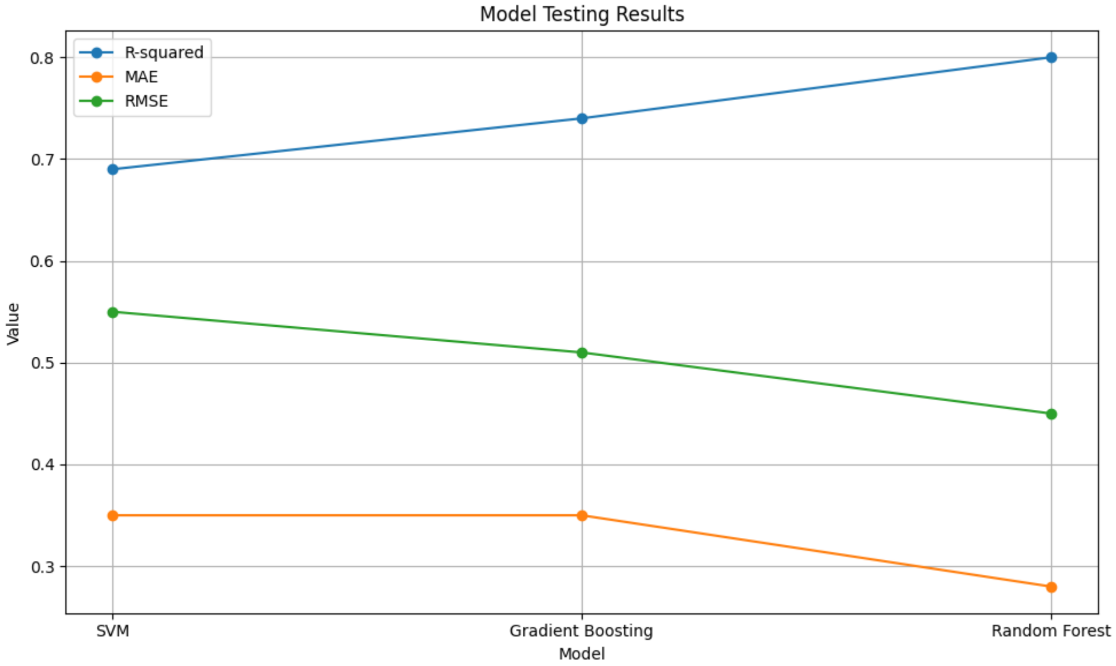


Fig 5.3.4 IQR

5.3.5 Result



References

[1] S. Sharma, D. Arora, G. Shankar, P. Sharma and V. Motwani, "House Price Prediction using Machine Learning Algorithm," 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2023, pp. 982-986, doi: 10.1109/ICCMC56507.2023.10084197.

[2] A. Varma, A. Sarma, S. Doshi and R. Nair, "House Price Prediction Using Machine Learning and Neural Networks," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 2018, pp. 1936-1939, doi: 10.1109/ICICCT.2018.8473231.

[3] A. Gupta, S. K. Dargar and A. Dargar, "House Prices Prediction Using Machine Learning Regression Models," 2022 IEEE 2nd International Conference on Mobile Networks and Wireless Communications (ICMNWC), Tumkur, Karnataka, India, 2022, pp. 1-5, doi: 10.1109/ICMNWC56175.2022.10031728.

[4] M. Jain, H. Rajput, N. Garg and P. Chawla, "Prediction of House Pricing using Machine Learning with Python," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020, pp. 570-574, doi: 10.1109/ICESC48915.2020.9155839.

[5] S. P. Sreeja, V. Asha, B. Saju, N. CP, P. O. Prakash and A. K. Singh, "Real Estate Price Prediction using Machine Learning," 2023 Third International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), Bhilai, India, 2023, pp. 1-7, doi: 10.1109/ICAECT57570.2023.10117910.

[6] Y. Chen, R. Xue and Y. Zhang, "House price prediction based on machine learning and deep learning methods," 2021 International Conference on Electronic Information Engineering and Computer Science (EIECS), Changchun, China, 2021, pp. 699-702, doi:10.1109/EIECS53707.2021.9587907.

[7] T. D. Phan, "Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia," 2018 International Conference on Machine Learning and Data Engineering (iCMLDE), Sydney, NSW, Australia, 2018, pp. 35-42, doi:10.1109/iCMLDE.2018.00017.

[8] G. K. Kumar, D. M. Rani, N. Koppula and S. Ashraf, "Prediction of House Price Using Machine Learning Algorithms," 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2021, pp. 1268-1271, doi: 10.1109/ICOEI51242.2021.9452820.

[9] R. Dwivedi, R. Gupta and P. K. Pal, "House Price Prediction using regression techniques," 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2022, pp. 165-170, doi:10.1109/ICAC3N56670.2022.10074128.