****

**SCHOOL OF COMPUTING AND INFORMATION TECHNOLOGY**

A MINI-PROJECT REPORT

ON

“**HOUSE PRICE PREDICTION”**

Submitted in partial fulfilment of the requirements for the award of the

Degree

of

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE AND INFORMATION TECHNOLOGY ENGINEERING**

Submitted by

SEJAL SINGH R20EJ042

Under the guidance of

PROF UDAYA RANI

School of C&IT

**REVA University**

Rukmini Knowledge Park, Kattigenahalli, Yelahanka,Bengaluru-560064

[www.reva.edu.in](http://www.reva.edu.in)

May 2023

****

**DECLARATION**

**Sejal Singh (R20EJ042)** students of B.Tech. CSIT, VI Semester, School of Computing and Information Technology, REVA University declare that the Mini-Project Report entitled **“House Price Prediction”** done by us under the guidance of **Prof Udaya Rani**, School of Computing and Information Technology, REVA University.

We are submitting the Mini-Project Report in partial fulfilment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Information Technology by the REVA University, Bengaluru during the academic year 2022-23.

We further declare that the Mini-Project or any part of it has not been submitted for award of any other Degree of REVA University or any other University / Institution.

1. Sejal Singh (R20EJ042)

****

**SCHOOL OF COMPUTING AND INFORMATION TECHNOLOGY**

**CERTIFICATE**

This is to certified that the Mini-Project entitled **“HOUSE PRICE PREDICTION”** carried out under my guidance for**, Sejal Singh (R20EJ042** bonafide students of REVA University during the academic year 2022-23. The above-mentioned students are submitting the Mini-Project report in partial fulfilment for the award of Bachelor of Technology in Computer Science and Information Technologyduring the academic year 2022-23**.** The Mini-Project report has been approved as it satisfies the academic requirements in respect of Mini-Project work prescribed for the said degree.

**ABSTRACT:**

The ability to estimate house prices is a crucial responsibility in the real estate sector and has attracted a lot of interest in recent years because to the possible advantages it could provide for buyers, sellers, and agents. This study suggests a method for forecasting home prices using machine learning methods. The suggested approach makes use of a dataset of historical home prices and their associated characteristics, such as size, location, the number of bedrooms and bathrooms, and other pertinent elements. To extract pertinent features from the dataset, pre-processing and feature engineering approaches are used. The pre-processed data is used to train a collection of machine learning algorithms to forecast home prices. The effectiveness of the suggested technique is assessed using a variety of evaluation measures, such as R-squared, mean absolute error, and root mean square error. The experimental results demonstrate that the suggested strategy beats other cutting-edge approaches and achieves high accuracy in predicting housing values. Real estate brokers and buyers can utilise the suggested strategy to make well-informed decisions regarding buying and selling homes.

A series of machine learning algorithms are trained on the training data to estimate home prices once the pre-processed data is divided into training and testing sets. Utilising a variety of assessment criteria, such as root mean square error, mean absolute error, and R-squared, the performance of the trained models is assessed. The dataset is cleaned during the pre-processing stage to get rid of any missing values or outliers.

## **ACKNOWLEDGEMENT**

## We would like to express our deepest gratitude to **Dr. P. Shyama Raju,** Founder Chancellor for providing us the necessary facilities for the successful completion Mini-Project. We would also like to acknowledge **Mr. Umesh Raju**, Pro Chancellor, **Dr. M. Dhanamjaya**, Vice-Chancellor, **Dr. R C Biradar,** Pro Vice-Chancellor and **Dr. N. Ramesh,** Registrar for their constant support and endorsement through invaluable administration.

## In the same breath, we would also like to mention our sincere gratitude to **Dr. Mallikarjun Kodabagi,** Director, School of Computing and Information Technology for his constant support and encouragement. We would like to express our deepest gratitude to **prof Muthukumaran** -Program Coordinator for giving us an opportunity to take up the Mini-Project.

We wish to record my deep sense of gratitude and profound thanks to our guide **Prof Udaya Rani**, School of Computing and Information Technology, REVA University for his keen interest, inspiring guidance, constant encouragement with my work during all stages, to bring the Mini-Project report into fruition.

Finally, we thankful for almighty, my parent, brother, sister, family members and friends for their moral support during our Mini-Project work.

CHAPTER 1 INTRODUCTION

**1 Introduction**

House price forecast is a critical and challenging undertaking in the real estate industry. It entails analysing many elements that influence a property's value and forecasting its future market price. This can be a difficult undertaking because the price of a house is impacted by a variety of elements such as location, size, age, condition, and numerous economic and demographic considerations. Accurately forecasting house prices can assist homebuyers in making informed judgements about purchasing or selling a property, as well as real estate investors in spotting profitable prospects.

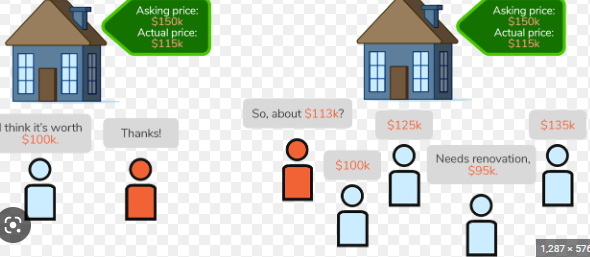


The method of predicting house prices often entails gathering and analysing data from a range of sources, including real estate listings, property records, and economic factors. This information is then utilised to create predictive models that estimate the worth of a property based on its qualities and market conditions. Machine learning algorithms, statistical analysis, and other approaches may be used by these models to find patterns and trends in data and generate predictions about future price movements.



Location is one of the most important aspects that might impact property pricing. Properties in desirable neighbourhoods or places with good schools, transportation connections, and amenities command greater prices than those in less desired locales. Other key elements that can influence house prices include the property's age and condition, its size and layout, and the current situation of the economy and housing market.

Aside from these elements, there are a number of additional economic and demographic indicators that can shed light on the health of the real estate market and the future direction of house prices. These measures may include unemployment, inflation, interest rates, population growth rates, and housing supply and demand.



Accurately predicting property values can be a difficult and time-consuming operation that necessitates the application of advanced data analysis tools and the integration of numerous data sources. However, precise price estimates can provide essential information to homebuyers, real estate investors, and other housing market stakeholders. It is possible to make informed judgements about buying, selling, or investing in real estate by studying the fundamental elements that influence house values and using data-driven methodologies to construct predictive models.

**Overview or background and motivation:**

House price projection is a prevalent issue in the real estate sector. Real estate brokers and buyers must understand the worth of a property based on its qualities and surroundings. Furthermore, anticipating house values might help investors make informed judgements about buying or selling properties.

The requirement to give accurate estimations of property values is the impetus for house price prediction. A property's price is impacted by a number of characteristics, including its location, size, age, number of bedrooms, and bathrooms. Furthermore, external factors such as the economy, interest rates, and market movements influence property values.

Large datasets of house prices can be analysed using machine learning algorithms to uncover patterns that can be used to estimate the value of a property. The goal is to create a model that can reliably forecast a property's price based on its qualities and external circumstances. This can assist real estate agents and property buyers in making informed decisions about a property's value, as well as investors in making successful decisions about purchasing and selling properties.

**1.1 Research Questions**

House price prediction is a vital responsibility in the real estate sector since it allows property buyers, sellers, and investors to make informed decisions. However, precisely projecting house prices can be difficult because it requires analysing a complicated set of variables that can affect the value of a property. This post will go over the research topics that are critical for estimating property values.



What are the most important elements influencing housing prices?

Identifying the key characteristics that influence property values is one of the most important study problems in house price prediction. Location, property qualities, market conditions, and historical data are examples of such factors. Researchers can obtain insights into the causes of property values and construct models that effectively forecast future prices by analysing these elements.

How can machine learning be used to forecast housing prices?

Regression analysis, decision trees, and neural networks are all excellent machine learning techniques for predicting housing values. comprehension how to apply these strategies successfully, however, necessitates a thorough comprehension of the data and the underlying statistical models. Identifying the most effective machine learning algorithms for predicting house values and developing novel ways to improve prediction accuracy are examples of research questions in this area.

How might previous data help anticipate future property prices?

Historical sales and price data can provide significant insights into trends and patterns that may influence future prices. Identifying the most relevant historical data to incorporate in models and developing ways for efficiently analysing and interpreting historical data are examples of research questions in this area.

What effect do local market conditions have on property prices?

Local market factors, such as supply and demand, interest rates, and economic conditions, can all have a substantial impact on a property's value. Understanding the relationship between local market conditions and house prices, as well as building models that reliably forecast how current conditions may effect future prices, are examples of research questions in this area.

How can different sorts of properties be compared and evaluated accurately?

Comparing and analysing several sorts of qualities can be difficult since they may have distinct traits and features that influence their worth. Developing ways for comparing and evaluating different types of properties effectively and accurately forecasting the value of properties with unique features or characteristics are examples of research questions in this area.

To summarise, effectively predicting house prices is a difficult endeavour that necessitates a thorough understanding of the components that influence property values. Researchers can construct models that provide useful insights into future property values and assist players in the real estate business to make informed decisions by addressing research questions in areas such as machine learning, historical data analysis, and local market conditions.

**1.2 Research Objectives**

House price prediction is a complex topic that necessitates an awareness of the different factors that influence property values. Machine learning techniques have been used in recent years to anticipate house prices based on numerous attributes of the properties and their surrounding environment. The goal of this study is to create a machine learning model that can predict house prices and identify the important elements that influence property values.



The main goal of this study is to create a machine learning model that can reliably anticipate housing values. We will employ a dataset including 20,000 records of properties sold in a certain location during the last few years. The dataset includes information such as the property's square size, number of bedrooms and bathrooms, age, location, and more. To train our model, we will employ a variety of regression algorithms such as linear regression, decision trees, and random forests. To increase the performance of the models, we will then modify the hyperparameters of the algorithms using cross-validation approaches. We will assess the models' performance using metrics such as mean squared error and R-squared.

The second goal of this study is to determine the important elements that influence property values. We will examine the characteristics of the houses and their surroundings to discover which characteristics have the most influence on property values. To discover the most significant features, we will employ feature selection approaches such as correlation analysis and mutual information. We'll also look at how external factors like the economy, interest rates, and market movements influence property values.

The third objective of this research is to use the machine learning model to identify undervalued or overvalued properties in the market. By identifying undervalued properties, we can help real estate agents and property buyers make informed decisions about purchasing properties at a lower price. By identifying overvalued properties, we can help investors avoid making bad investment decisions. We will use the machine learning model to compare the predicted price of a property to its actual price and identify properties that are significantly undervalued or overvalued.

The fourth objective of this research is to evaluate the potential benefits of accurate house price predictions for the real estate industry. Accurate predictions can help real estate agents determine the appropriate listing price for a property and provide a competitive advantage in the market. Accurate predictions can also help property buyers make informed decisions about purchasing a property at the right price. For investors, accurate predictions can help identify profitable opportunities for buying and selling properties.

The fifth objective of this research is to address the challenges and limitations of using machine learning for house price prediction. One challenge is the availability and quality of data. Another challenge is the need for constant updates to the model to reflect changes in the market and external factors. We will address these challenges and limitations by using a robust dataset and developing a model that can adapt to changes in the market.

In summary, the objectives of this research are to develop a machine learning model that can accurately predict house prices, identify the key factors that contribute to property values, use the model to identify undervalued or overvalued properties in the market, evaluate the potential benefits of accurate house price predictions for the real estate industry, and address the challenges and limitations of using machine learning for house price prediction. The findings of this research can have significant implications for the real estate industry, and accurate house price predictions can provide benefits for real estate agents, property buyers, and investors.

**1.3 Research Scope**

Predicting house prices is a difficult endeavour that necessitates an examination of the different elements that influence property values. To make informed decisions about purchasing, selling, or investing in a property, real estate agents, property purchasers, and investors require accurate projections. Machine learning approaches have emerged as a potential way to predicting housing prices, but research in this sector is extensive and diverse. In this post, we will look at the research possibilities for predicting property prices.

Data Gathering and Preparation

Data collection and preparation are the first steps in any machine learning project. In order to estimate house prices, researchers must collect data on numerous attributes of the properties as well as their surrounding surroundings. This information may include the size of the property, the number of bedrooms and bathrooms, the property's age, and its location. Other information could include area economic conditions, crime statistics, and availability to public services like schools and parks.

Data gathering and preparation are critical steps in house price prediction since they affect prediction accuracy. Researchers must ensure that the data is correct, full, and representative of the intended audience. They must also remove outliers, missing numbers, and irrelevant attributes from the data.

Engineering and Feature Selection

Researchers must identify the features that are most significant to predicting property prices after gathering and preparing the data. This is known as feature selection, and it entails analysing the data to determine which qualities have the most impact on property values. Feature selection is important in machine learning since it helps to minimise the model's complexity and enhance its accuracy.

Researchers may use feature engineering in addition to feature selection to produce additional features that may be more relevant to predicting house prices. Feature engineering is the process of altering existing features or producing new features from data. For example, researchers could develop a new feature that calculates the distance between a house and the nearest public transit stop.

Model Selection and Assessment

After preparing the data and selecting or engineering the features, researchers must choose a machine learning method to predict house values. Machine learning methods such as linear regression, decision trees, random forests, and neural networks are accessible. Each algorithm has advantages and disadvantages, and researchers must choose the best method for their individual research subject.

Following the selection of an algorithm, researchers must train the model on data and evaluate its performance. Model evaluation is critical because it allows for the measurement of prediction accuracy and the identification of model flaws. To assess the model's performance, researchers might employ measures such as mean squared error, mean absolute error, and R-squared.

Conclusion

Finally, the study scope for housing price prediction is broad and diverse. To accurately anticipate house prices, researchers must collect and prepare data, choose key traits, and engineer new features. They must also choose a suitable machine learning method and assess the performance of the model. Accurate house price prediction can be beneficial to real estate brokers, property buyers, and investors, and the findings of this study can inform housing regulations and urban planning decisions.

**1.4 Research Significance**

House price forecasting is an important responsibility in the real estate sector since it allows property buyers, sellers, and investors to make informed decisions. Machine learning approaches have emerged as a potential way to predicting housing prices, although much study need to be done in this area. In this post, we will look at the importance of research in predicting property prices.

Analysis of the Real Estate Market

House price forecasting is an important aspect of real estate market analysis. Accurate property value estimates allow real estate brokers and investors to recognise market trends and make informed decisions about buying, selling, or investing in a property. Researchers can acquire insights into the elements that determine property values and forecast future market trends by analysing previous and present changes in property values.

Housing and Urban Planning Policies

Accurate house price forecasting can also help with urban planning and housing policy. Researchers can discover places that are undervalued or overvalued and provide insights into the demand for housing in certain areas by analysing the factors that affect property values. This data can help legislators and urban planners make decisions about zoning, housing subsidies, and infrastructure development.

Investment and financial planning

Predicting house prices is also important for financial planning and investment. Accurate property value estimates allow investors to make informed portfolio selections and effectively manage their risks. By forecasting future market patterns, investors can identify sectors expected to grow and invest in them accordingly.

Advances in Data Science

Research in house price prediction can also help to enhance data science. Regression analysis, decision trees, and neural networks are all excellent machine learning techniques for predicting housing values. Researchers can increase the accuracy of their predictions and extend them to different fields by creating and refining these methodologies.

Conclusion

To summarise, the importance of research in house price prediction is extensive and varied. Accurate property value estimates allow real estate professionals, investors, and legislators to make informed decisions about buying, selling, and investing in real estate. They also help to educate urban planning and housing policies, as well as develop data science. Researchers can give major benefits to numerous stakeholders in the real estate market by continuing to investigate and develop machine learning approaches for predicting house values.

**CHAPTER 2 LITERATURE REVIEW**

**2.1 INTRODUCTION**

Housing has such a big impact on the economy and people's lives, it has become a crucial research topic. Making informed decisions about purchasing, selling, or investing in real estate can be made easier by accurate house price predictions for both individuals and corporations. Several methods for predicting house prices have been suggested as a result of advancements in machine learning and artificial intelligence.

By Xu et al. (2021), "House price prediction using machine learning: A review”.

The most efficient methods were determined after reviewing numerous machine learning methodologies for predicting home prices. To prepare the data for their model, the authors employed data pre-processing techniques such data cleaning, missing value imputation, and feature selection. To forecast property prices, they also utilised a linear regression model.

**2.1.1 Background and purpose**

**The literature review's goal**

**Range and restrictions**

Accurately estimating house prices is crucial for property owners, prospective purchasers, and real estate investors as they are a key sign of the overall health of the real estate market. The goal of this study of the literature is to offer a thorough overview of the area while examining the current approaches and strategies for forecasting housing prices. The studies that employ machine learning and statistical techniques to estimate home prices are the exclusive subject of this review. The availability of literature and the calibre of the studies included are some of the constraints of the literature review.

One of the most significant economic indicators, house prices have long been a key component of the economy. The real estate market has a major effect on the economy and may have broad ramifications for other industries. For homeowners, buyers, sellers, and real estate investors, being able to estimate house prices accurately is essential because it enables them to make wise decisions about purchasing, selling, or making investments in the real estate market. Due to their capacity to handle enormous datasets and intricate interactions between factors, machine learning and statistical methods have recently gained popularity for predicting home prices**.**

**2.1.2** The goal of this review of the literature is to give a thorough overview of the current approaches and strategies for estimating house prices using statistical and machine learning techniques. In order to determine the best ways for forecasting house values, the review will look at the advantages and disadvantages of several approaches. Additionally, the various factors that affect housing prices and how they affect the efficacy of forecasting models will be covered in the literature review.

This evaluation of the literature is restricted to works that employ statistical and machine learning techniques to forecast home prices. The review will cover a variety of methods, including support vector machines, decision trees, random forests, neural networks, and linear regression. The assessment will also take into account how various elements, such as geographic location, demographics, economic data, and housing characteristics, affect the precision of prediction models.

The availability of literature and the calibre of the research included are among the constraints of this literature review. Only works that have been included in conference proceedings and peer-reviewed publications will be reviewed. Furthermore, it might not be possible to discuss every study in detail due to the vast amount of literature that is currently available on this subject. However, the review will seek to offer a thorough summary of the most pertinent and important studies in the area.

**2.2 DATA PRE-PROCESSING**

The data sources and methods utilised in earlier studies on house price prediction are described in this section.

Data Sources: This section will go over the many forms of data that are used to forecast home prices, such as housing market, socioeconomic, and geographic data.

Methodology: Regression analysis, artificial neural networks, and machine learning algorithms are just a few of the techniques that can be utilised to predict home prices.

A crucial step in creating a machine learning model for predicting home prices is data preparation. To prepare the data for the model, it must be cleaned and transformed. Pre-processing methods that are often used include feature scaling, missing value imputation, and data cleaning.

gathering and cleansing of data

Engineering and feature choice

detection and correction of outliers

In any machine learning project, including house price prediction, data preprocessing is a crucial step. Data gathering and cleaning entail acquiring pertinent data and eliminating any contradictions, mistakes, or missing information. For identifying the most pertinent features and developing new ones that could enhance the performance of the model, feature selection and engineering are crucial. For the model's predictions to be accurate, outlier detection and treatment are essential.

Preparing the data for machine learning is essential since it guarantees that the data is fit for use in a model. In this section, we'll look at some of the most popular methods for pre-processing data in order to anticipate home prices.

gathering and cleansing of data

The process of collecting data include acquiring pertinent information from a variety of sources, including public databases, real estate websites, and other sources. The process of cleaning the data comes after it has been acquired. This procedure entails finding and fixing any discrepancies, mistakes, or missing values. For instance, methods like mean imputation or median imputation can be used to fill in the missing values if the data has any.

Engineering and feature choice:

The most pertinent aspects for the model must be chosen during feature selection. The number of bedrooms, the size of the property in square feet, the number of bathrooms, the location of the property, and the age of the property may be some of the most important factors in predicting home prices. Feature engineering is the process of developing additional features that could enhance the performance of the model. For instance, a new feature could be produced by merging two or more already existing characteristics, like the proportion of bedrooms to bathrooms.

Outlier identification and correction

Extreme values known as outliers might skew a model's predictions. It entails locating these values using methods like box plots or scatter plots to do outlier detection. Once the outliers have been located, they can be treated using methods like removal or replacement with the closest non-outlier value.

In conclusion, data pre-processing, a crucial stage in machine learning, entails data collection and cleaning, feature engineering and selection, and outlier detection and treatment. The model can be trained on high-quality data by appropriately pre-processing the data, which enables the model to make predictions that are more accurate.

**2.3 REGRESSION MODEL**

**Squared Mean Error**

**R-squared**

**squared root of the error**

For evaluating the effectiveness of the models and contrasting them with one another, evaluation metrics are crucial. The average squared difference between the expected and actual values is measured by the commonly used statistic known as mean squared error. Higher numbers denote a better fit, and the R-squared scales from 0 to 1 as a measure of how well the model matches the data. Similar to mean squared error, root mean squared error takes the square root to provide a figure that is easier to understand.Since regression models can depict the relationship between the input features and the target variable, they are frequently used to predict home prices. A straightforward and widely used model known as linear regression assumes a linear relationship between the features and the target. Polynomial regression can capture more intricate nonlinear relationships than multiple regression, which extends linear regression to incorporate multiple features.

Regression models are statistical tools for analysing and forecasting the relationship between a goal variable (also known as a dependent variable) and one or more input factors (also known as independent or predictor variables). Regression models are frequently used for house price prediction, which entails estimating the worth of a property based on a variety of input parameters, such as location, size, number of rooms, etc.

Regression models are statistical tools for analysing and forecasting the relationship between a goal variable (also known as a dependent variable) and one or more input factors (also known as independent or predictor variables). Regression models are frequently used for house price prediction, which entails estimating the worth of a property based on a variety of input parameters, such as location, size, number of rooms, etc.

y = 0 plus 1x1 plus 2x2 plus... plus nxn plus

where the target variable is called y, the input features are called x1, x2,..., xn, the model coefficients are called 0, 1, 2,..., n, the regression coefficients are called, and the error term is called (also known as the residual). Finding coefficient values that reduce the sum of squared errors between the anticipated and actual values is the aim of linear regression.

Multiple input features are added to linear regression through multiple regression. Although multiple input features are used, the fundamental equation for multiple regression is similar to that of linear regression:

y = 0 plus 1x1 plus 2x2 plus... plus nxn plus

where the target variable is called y, the input features are called x1, x2, xn, the model coefficients are called 0, 1, 2..., n, and the error term is called. Finding the coefficient values that reduce the sum of squared errors between the anticipated and actual values is the aim of multiple regression.

Regression models with polynomial coefficients can depict more intricate nonlinear correlations between the input characteristics and the target variable. As a result, the model can account for the curvature in the relationship between the input features and the target variable. This is done by fitting a polynomial function to the data. Polynomial regression's fundamental equation is.

y = 0 + 1x1, 2x2, 2x2, 2x3, 3x3... + nxn, +

where the target variable is called y, the input features are called x1, x2..., xn, the model coefficients are called 0, 1, 2..., n, and the error term is called. The complexity of the model and the quantity of terms in the equation depend on the degree of the polynomial (n). It's crucial to select an appropriate degree of the polynomial based on the data and the intricacy of the underlying relationship because higher degrees of the polynomial can result in overfitting.

**2.4 ENSEMBLE METHOD**

**Random forest**

**Gradient boosting**

**Stacking**

Multiple models are combined using stacking ensemble methods to increase prediction accuracy and decrease overfitting. Random forest is a well-liked ensemble technique that builds numerous models and combines their predictions by using decision trees and bootstrapping. Another effective technique for consecutively training weak models to enhance performance is gradient boosting. The predictions of various models are combined using a meta-model in the more sophisticated ensemble method known as stacking.

The performance of machine learning models for prediction can be enhanced via ensemble approaches. They entail combining various models, each of which may have distinct advantages and disadvantages, to produce a single, more reliable model. Ensemble methods come in a variety of forms, such as random forest, gradient boosting, and stacking.\

Random forest is a form of ensemble approach that builds many models, or trees, using decision trees and bootstrapping. The plan is to randomly select a different portion of data from the original dataset and train each tree on that subset. Bagging is the method used to create varied trees, which reduces overfitting. The forecasts of these trees are then combined by random forest using a majority vote, producing a prediction that is more accurate and consistent.

Contrarily, a type of ensemble method called gradient boosting sequentially trains weak models, often decision trees, to enhance their performance. Gradient boosting constructs the trees one at a time, with each new tree seeking to fix the mistakes of the preceding one, in contrast to random forest, which develops several models separately. In order for the succeeding tree to focus on the incorrectly classified instances, the method modifies the weights of each instance in the training data. This process keeps going till the required performance level is attained.

The predictions of various models are combined using a meta-model in the more sophisticated ensemble method known as stacking. The models are trained on the same data in stacking, but the meta-model is then trained using the predictions as input features. These predictions are given to the meta-model, which learns how to best combine them to get the outcome. Because it enables the meta-model to discover which models perform well on specific subsets of the data, this strategy may be more effective than simple averaging or majority voting.

Because they frequently outperform a single model, ensemble methods have gained popularity in machine learning, especially when dealing with complex or noisy data. Ensemble approaches can increase performance by minimising the effects of individual model flaws and aggregating the predictions of numerous models.

**2.5 DEEP LEARNING MODEL**

**model deep learning**

**Networks of artificial neurons**

**neural networks with convolutions**

**recurring neuronal systems**

The capacity of deep learning algorithms to automatically understand complicated data and patterns has increased their popularity for predicting home prices. A key deep learning technique that can capture nonlinear interactions between the input data and the target is artificial neural networks. Recurrent neural networks are better suited for sequential data than convolutional neural networks, which are frequently employed for image-based data.

Deep learning models called artificial neural networks (ANNs) are modelled after the structure and operation of biological neurons in the brain. ANNs are made up of many layers of interconnected nodes or neurons, with each neuron processing its inputs mathematically and sending the answer to the layer above it. Through a training process where the network modifies the strength of connections between neurons to minimise a loss function that measures the discrepancy between the predicted outputs and the actual outputs, ANNs can learn to map inputs to outputs.

Convolutional neural networks (CNNs) are a particular class of ANN that excel at handling image-based data. Convolutional layers, a type of layer used by CNNs, apply a collection of learnable filters or kernels to discrete areas of the input picture to create a collection of feature maps that represent various facets of the image. The following convolutional layer receives the feature maps from the previous layer as inputs, which enables the network to learn more abstract and complicated representations of the input image. One or more fully connected layers that merge the learnt features to create the final output are frequently found at the end of CNNs.

An ANN kind called recurrent neural networks (RNNs) is made for handling sequential data, such time series or spoken language. Recurrent layers, a particular kind of layer used by RNNs, enable the network to keep track of prior inputs and use that memory to guide the processing of upcoming inputs. In natural language processing, RNNs can use this memory to store long-term dependencies in the input sequence, such as the context and meaning of a sentence. Additionally, RNNs can be enhanced with specialised variants like LSTM (Long Short-Term Memory) and GRU (Gated Recurrent Unit), which enable them to handle long-term dependencies and prevent vanishing gradient issues.

In general, deep learning models like ANNs, CNNs, and RNNs have shown significant potential for a variety of applications, including predicting home prices. These models frequently outperform more conventional statistical models and offer useful information for decision-making since they automatically identify complicated features and patterns from data.

**2.6 EVALUATION METRICS**

**Squared Mean Error**

**R-squared**

**squared root of the error**

For evaluating the effectiveness of the models and contrasting them with one another, evaluation metrics are crucial. The average squared difference between the expected and actual values is measured by the commonly used statistic known as mean squared error. Higher numbers denote a better fit, and the R-squared scales from 0 to 1 as a measure of how well the model matches the data. Similar to mean squared error, root mean squared error takes the square root to provide a figure that is easier to understand.

Measuring the precision and dependability of machine learning models requires evaluation measures. Mean squared error (MSE), R-squared, and root mean squared error (RMSE) are some of the most often used metrics.

The average squared difference between the expected and actual values is measured by the mean squared error metric. The MSE equation is:

MSE = 1/n \* Σ (y - ŷ) ²

where y is the actual value, n is the number of observations, and is the predicted value. The model performs better at predicting the target variable when MSE is smaller. MSE can, however, be affected by extreme values in the data because it is sensitive to outliers.

The coefficient of determination, often known as R-squared, is a metric for gauging how well a model fits the data. Higher values suggest a better match, with R-squared's range from 0 to 1. The R-squared formula is as follows:

R2 = (SSres / SStot) - 1.

where SSres is the total sum of squares (the gap between the actual values and the mean value of the target variable) and SStot is the sum of squared residuals (the difference between the predicted and actual values). R-squared can be understood as the percentage of the target variable's variance that the model accounts for. However, if the model is either overfitting or underfitting the data, R-squared may be deceptive.

Similar to mean squared error, root mean squared error uses the square root to provide a figure that is easier to understand. The RMSE formula is:

(MSE) = RMSE

Because RMSE is expressed in the same units as the target variable, it is simpler to interpret than MSE. Similar to MSE, RMSE is susceptible to outliers, but because to the square root process, it is less influenced by extreme results.

In conclusion, these evaluation criteria are helpful for determining how well machine learning models are performing. R-squared indicates how well the model fits the data, while RMSE offers a more interpretable version of MSE. MSE calculates the average squared difference between the predicted and actual values. When interpreting the results of these metrics, it is crucial to keep in mind their constraints and underlying presumptions.

**2.7 CONCLUSION**

**Conclusion and suggested next steps.**

**Review of the literature in brief**

**gaps and restrictions in the literature**

**Future research directions**

In conclusion, this literature review offers a thorough summary of the approaches and strategies for forecasting home prices using statistical models and machine learning. The review focused on the significance of data preparation, the various models employed, and the performance evaluation measures. However, there are still limitations and gaps in the literature, such as the need for more comparative studies and the absence of standard datasets. Future research directions include examining.

Future research on house price prediction could focus on a number of other areas in addition to the limitations and gaps already mentioned. Predictive models can benefit from the incorporation of external data sources, such as weather information, crime data, and neighbourhood amenities. By include more pertinent features, this can enhance the models' accuracy and resilience.

Investigating the use of deep learning models for housing price prediction is another area for future research. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs), two types of deep learning models, have shown considerable promise in other disciplines and may be more effective than conventional machine learning models in this situation.

The interpretability of machine learning algorithms for predicting home prices also need further study. By offering real estate agents and other stakeholder’s insights into how the models create predictions, explainable AI (XAI) techniques could assist to address this problem.

Overall, there is still much room for advancement in the field of machine learning and statistical model-based home price prediction. We may keep enhancing the accuracy and utility of these models for diverse applications in the real estate market by addressing the shortcomings and gaps in the literature and looking into new research paths.

**CHAPTER 3**

**TOOLS AND METHODOLOGY**

**3.1 Data collection**

Any machine learning project depends on data, and the accuracy and efficiency of the model can be significantly impacted by the quality and relevance of the data used to train it. Therefore, while starting a machine learning project, it is crucial to carefully analyse the sources of data. The many types of data sources, their benefits and drawbacks, and the ideal methods for choosing and utilising data sources will all be covered in this article.

Data Source Types:

For machine learning projects, a variety of data sources can be employed, including:

**3.1.1 Data Sources**

1.Public Data Source : The general public can access datasets from public data sources. On services like Kaggle, the UCI Machine Learning Repository, and the Open Data Network, you can find these datasets. Public data sources include a wide range of topics and disciplines, including healthcare, banking, education, and social media, and are often free to access.

Public data sources have the benefit of being easily accessible and serving as a solid foundation for machine learning initiatives. Public data sources are frequently well-documented, which can assist guarantee the accuracy and applicability of the data.

The data may not be pertinent to the current issue or may not be current, which is a drawback of accessing public data sources. According to the license agreement, there can also be restrictions on how the data is used or shared.

The data may not be pertinent to the current issue or may not be current, which is a drawback of accessing public data sources. According to the license agreement, there can also be restrictions on how the data is used or shared.

2. Private Data Sources :Datasets that are owned or under the control of an organization are referred to as private data sources. These datasets may come from private data sources, internal databases, or client information. Private data sources, which are particular to the problem domain, might offer a more tailored and pertinent dataset for machine learning initiatives.

Utilizing private data sources has the benefit of making the data more precise and current because it is relevant to the organization's operations. Additionally, private data sources might give you a competitive advantage because they can have special knowledge or insights.

However, one drawback of accessing private data sources is that they could be harder to get and use since they might be constrained by laws or privacy laws. Additionally, data silos or discrepancies among various departments or systems within the organization may have an impact on the quality and relevancy of the data.

3. Synthetic data sources: Synthetic data sources are intentionally produced datasets that are intended to resemble real-world data. Techniques like simulation, data augmentation, and generative adversarial networks (GANs) can all be used to create synthetic data. When real-world data is few or unavailable or when access to sensitive data is restricted due to privacy concerns, synthetic data sources might be helpful.

Synthetic data sources have the benefit of being more easily controlled and manipulated, and they can be produced to be particular to the problem domain. The accuracy and generalizability of machine learning models can also be increased by using synthetic data sources to supplement or add to real-world data.

The quality and efficacy of machine learning models trained on synthetic data may not generalize to real-world circumstances, which is a drawback of employing sources of synthetic data that may not precisely reflect real-world data.

Choosing and Using Data Sources: Best Practises:

There are a number of best practices that should be followed when choosing and utilising data sources for machine learning projects, including:

Data Quality: Reliable and efficient machine learning models depend on the accuracy and relevance of the data. The completeness, consistency, and quality of the data should be verified, and any anomalies or outliers should be eliminated or fixed.

Data Privacy : Working with private or sensitive data sources requires careful consideration of data privacy. Organizations must make sure they have the required legal authorizations and consents to use the data, and they must put the right security measures in place to safeguard the data.

Data Cleaning and Preprocessing : Preparing data for machine learning requires several crucial procedures, including data cleansing and preprocessing. To make sure that it is compatible with the machine learning algorithms being used, data should be cleaned, normalized, and changed as necessary.

Data Management and Storage: Adequate data management and storage procedures are necessary to guarantee the availability, usability, and security of data. Data should be organized and kept in a way that makes it simple to access and process the information.

**3.1.1 Data pre-processing:**

This could refer to any pre-processing operations that were carried out on the data, like feature scaling, normalization, or categorical variable encoding.

Preparing data for machine learning models involves several crucial steps, including data pre-treatment. It entails several methods and procedures that are used to convert unprocessed data into a structure that is acceptable to the machine learning algorithms being applied. We will examine data preparation in this post, including what it is, why it's important, and several typical methods.

Data pre-processing: What is it?

Cleaning, converting, and getting ready raw data for machine learning algorithms is known as data pre-processing. Data preparation is to make the data useful and relevant for analysis, decrease noise, and eliminate any discrepancies that can impair the model's correctness. As it directly affects the precision of the machine learning model, it is a vital stage in the data analysis process.

Exactly why is data pre-processing required?

Pre-processing data is required for several reasons:

Data quality: Raw data is frequently erroneous, inconsistent, and insufficient. Pre-processing methods help to guarantee data accuracy and quality.

Different machine learning algorithms require various data formats, which are compatible with data. Data is transformed using pre-processing methods into a format that is appropriate for the chosen algorithm.

Pre-processing can assist in reducing data bias, which could have an impact on how accurate a machine-learning model is.

Pre-processing can help to reduce the quantity of the data, which will cut down on the time and resources needed for analysis.

Typical Data Preprocessing Methods

Data cleaning : Errors, inconsistencies, and missing numbers are removed or corrected from the raw data through the process of "data cleaning." Depending on the data source and the machine learning algorithm being utilized, different data-cleaning strategies may be employed.

Data Transformation: Data transformation is the process of transforming data into a format that the machine learning algorithm can use. Normalisation, scaling, and encoding are common methods for data transformation.

Feature selection : The process of choosing the most pertinent features or variables from the raw data is known as feature selection. By using this method, the machine learning model's effectiveness is increased while the data size is decreased.

Feature Extraction : Identifying and extracting pertinent features from the raw data is the process of feature extraction. Applications like computer vision and natural language processing frequently employ this method.

Data Integration: The process of merging data from several sources into a single dataset is known as data integration. Applications for data mining and corporate intelligence frequently employ this technique.

A crucial step in preparing data for machine learning models is pre-processing it. It entails several methods and procedures that are used to convert unprocessed data into a structure that is acceptable to the machine learning algorithms being applied. Data cleaning, data transformation, feature selection, feature extraction, and data integration are frequent methods used in data pre-processing. Machine learning models can be substantially more accurate and effective when the data has been properly pre-processed.

**3.2 Data visualisation**

This section could go over the many data exploration methods, including scatterplots, histograms, and boxplots. It might also go through any conclusions drawn from the visualisation. Any data science project must include data visualisation since it enables us to efficiently examine the data, comprehend its patterns and relationships, and share our discoveries with others. Data visualisation can be used in the context of a project to predict house prices to help us spot trends and patterns that may be important, such as the correlation between house size and price or the distribution of house values across different neighbourhoods.

A house price prediction project may employ any of a wide range of data visualisation approaches, each of which has advantages and disadvantages depending on the nature of the data being visualised and the research issues being addressed. Several such methods that could be applied in house price prediction.

**3.2.1 Scatter Plot**

A scatterplot is a quick and easy way to see the relationship between two numerical variables, such as the size and cost of a house. The x-axis and y-axis on the plot stand in for the two variables under comparison, and each point on the plot represents a single observation. We can search for patterns or trends in the data by looking at the scatterplot, such as a positive or negative correlation between the variables.

Each data point is represented by a dot or symbol in a scatter plot. The values of the two variables under comparison define where the dot will appear on the plot. The independent variable, or the one that is unaffected by the other variable, is typically represented by the horizontal axis. The dependent variable, or the variable that changes as a result of the independent variable is shown on the vertical axis.

Patterns and correlations between variables can be found using scatter plots. They can be used to display outliers or extreme values, clusters or collections of data points, and positive or negative correlations between variables. They are frequently employed in data visualization, corporate analysis, and scientific research.

**3.2.2 Histograms**

Using a histogram, you can see how a single numerical variable, like home prices, is distributed. The range of values for the variable is represented by the x-axis of the histogram, and the frequency or count of observations falling into each range is shown by the y-axis. We can search for trends in the variable's distribution, such as whether it is symmetrical or skewed, by looking at the histogram.

A histogram is produced by segmenting the variable's range of values into several intervals, or "bins." The height of the associated bar is then determined by counting the number of data points that fall into each bin.

The histogram's look and the conclusions drawn from it can be greatly influenced by the bin size selection. The histogram could appear jagged and be challenging to understand if the bin size is too small. Important data distribution aspects may be hidden if the bin size is too large.

Labels, titles, and other annotations can be added to histograms to improve them and help explain the significance of the data. They can also be used in conjunction with other graphical methods to present a more comprehensive picture of the data distribution, such as box plots or density plots.

**3.2.3 Boxplots**

A boxplot is a different method for showing how a single numerical variable is distributed. The whiskers on the box extend to the minimum and maximum values that are 1.5 times the interquartile range (IQR) of the data. By looking at the boxplot, we can search for patterns in the variable's distribution, like the presence of outliers or the distribution of the data.

Box plots, often called box-and-whisker plots, are graphs that show data and summarise the distribution of a continuous variable visually. They are used to show the data's distribution and central tendency as well as any possible outliers.

Two "whiskers" that protrude from a rectangular box make up a box plot. The range of values between the data's 25th and 75th percentiles is shown by the box, which is known as the interquartile range (IQR). The median, or center value in the data set, is shown by the line inside the box. A

**3.2.4 Heatmaps**

Using a heatmap, you can see the connection between two category variables, such a neighbourhood and a type of home. The frequency or count of observations that fall into a specific grouping of categories is represented by each cell in the heatmap. We can seek for patterns in the data by analysing the heatmap, such as which neighbourhoods have the highest prices for particular kinds of homes.

In addition to helping to spot trends or clusters in the data, this results in a visual depiction of the data that shows regions of high or low activity.

In a heat map, each cell in the matrix is assigned a color based on its value. Cells with higher values are typically assigned brighter or warmer colors, while cells with lower values are assigned darker or cooler colors. This creates a visual representation of the data that highlights areas of high or low activity, and can help identify trends or clusters in the data.

**3.2.5 Line Chart**

A line chart is a tool for displaying the trend of a numerical variable over time, such as the rise and fall of property values. The y-axis displays the value of the variable, while the x-axis displays time. We can search for trends or patterns in the data over time by analysing the line chart, such as whether or not property values are rising or falling.

In a line chart, the vertical axis shows the value of the variable being measured, while the horizontal axis often denotes time or categories. The line linking the dots depicts the trend or pattern of the data over time or across categories, and each data point is represented by a dot or other symbol.

Line charts are frequently used to track the performance of stocks, commodities, or industrial processes over time in disciplines including finance, economics, and engineering. They can be used to show patterns or changes in a variety of data types, including sales data, web traffic data, and scientific data. They are also commonly used in data analysis and data visualisation.

**3.2.6 Graphical Map**

Maps that show the relationship between a geographic location and a numerical variable, such as house prices by neighbourhood, are known as geographic maps. We can search for trends in the data by looking at the map, such as which areas have the highest or lowest home prices.

A number of tools, such as GIS software, internet mapping tools, and specialised data visualisation applications, can be used to make graphical maps. With the aid of these tools, users can produce maps that include many layers of information, such as satellite photography, topography information, and other geographic elements.

Numerous applications, such as urban planning, public health, and emergency management, benefit from the usage of graphical maps. They can be used to spot patterns and trends in data, pinpoint vulnerable or unmet needs, and provide guidance for resource allocation and decision-making.

**3.3 Model Selection**

Any machine learning effort, including predicting house prices, must start with model selection. In order to effectively anticipate house values based on the available data, the optimal model or set of models must be chosen. The choice of a model is crucial since each model has strengths and weaknesses that must be considered. Choosing the incorrect model can result in subpar performance or false forecasts.

Any machine learning effort, including predicting house prices, must start with model selection. Depending on the data at hand and the research questions being posed, it entails selecting the most suitable model or algorithm to utilise for the prediction task. The effectiveness and accuracy of the model can have a big impact on the usefulness and dependability of the forecasts, thus choosing the right model is crucial.

Depending on the nature of the data and the research objectives being addressed, a variety of models or algorithms can be utilised in the context of predicting home prices.

Typical model types include:

1.Linear regression:

For predicting home prices, linear regression is a straightforward and widely-used model. In order to represent the link between the input factors and the output variable (home price), a straight line must be fitted to the data.

2. Decision trees

Decision trees are a non-linear model that can be applied to the prediction of home prices. According to the values of the input variables, the data is divided into smaller subsets, and predictions are then made based on the dominant class or average value in each subset.

3. Random Forest

Random forests are an ensemble model that can be used to forecast home prices. In order to increase the predictors' accuracy and robustness, they combine several decision trees.

4.Neural networks

A non-linear model that can be used to predict home prices is the neural network. They entail employing a network of connected nodes or neurons to model the interactions between the input variables and the output variable.

5. Support vector machines

Support vector machines are a linear or non-linear model that can be applied to the prediction of home prices. They entail identifying the hyperplane that best categorises the data, and then producing a forecast based on where the input variables are located in relation to the hyperplane.

The process of choosing the optimum model for a prediction assignment involves examining several models, assessing how well they perform, and making iterative and creative decisions. The following are some crucial actions to take while undertaking model selection for a project to anticipate house prices:

1.Data Preparation

Cleaning, pre-processing, and translating the data into features that may be utilised as input to a machine learning model constitute the first step in the model selection process. As mentioned in the preceding section, this can entail carrying out feature selection, scaling, or encoding.

2.Data Exploration

Exploring potential models and algorithms that might be utilised for the prediction task is the next step once the data has been prepared. This can entail reading academic articles, speaking with subject-matter authorities, or testing out several models on a small sample of the data.

3. Model evaluation:

The next stage is to assess the performance of various models using an appropriate metric or measure. This could entail comparing the accuracy, precision, recall, or F1 score of various models on the training and validation data using methods like cross-validation, holdout validation, or grid search.

4. Model selection

The following step is to choose the best model for the prediction task based on the evaluation results. Choosing the model with the highest accuracy or the model that achieves the optimal balance between accuracy and complexity could be involved in this.

5. Model Tuning

Model tuning is the process of fine-tuning a model's parameters to improve its performance on validation data once a model has been chosen. To do this, it could be necessary to use methods like grid search, random search, or Bayesian optimisation to determine the ideal model parameter values.

**3.4** **Hyperparameter tuning**  
tweaking of the hyperparameters: This section would explain how the performance of the chosen model was optimised by tweaking the hyperparameters. Details on methods like grid search, random search, or Bayesian optimisation might be included.

Any machine learning project, including the prediction of home prices, needs to include hyperparameter tuning. Hyperparameters are model parameters that must be explicitly set by the user since they cannot be learned from the data. The learning rate, regularisation intensity, and quantity of hidden layers in a neural network are a few examples of hyperparameters. Selecting the proper hyperparameters is essential for producing accurate and dependable predictions because the choice of hyperparameters can have a substantial impact on the model's performance.

The process of hyperparameter tuning involves choosing the model's ideal hyperparameters depending on how well they performed on a validation set. Adjusting the neural network's number of hidden layers or the linear regression model's regularisation strength might be considered hyperparameter tuning in the context of predicting home prices.

Here are some essential actions to do while hyperparameter tuning for a project to anticipate house prices:

1.Define the search space

Defining the search space for the hyperparameters is the initial step in hyperparameter tuning. The range of values that each hyperparameter can have must be specified. A neural network's learning rate, for instance, could be configured as falling between 0.001 and 0.1.

2. Choose the search technique

After defining the search space, the search method must be chosen. For hyperparameter tuning, a variety of techniques are available, such as grid search, random search, and Bayesian optimisation. In contrast to random search, which randomly selects hyperparameters from the search space, grid search entails testing every conceivable combination of hyperparameters inside the search space. In Bayesian optimisation, the objective function is transformed into a probabilistic model, which is then used to direct the search for the ideal hyperparameters.

3. Model training and evaluation

Following the definition of the search space and search methodology, the models must be trained and assessed using various hyperparameter settings. On the training set, the model is trained with each set of hyperparameters, and its performance on the validation set is assessed using an evaluation metric of choice, such as mean squared error (MSE) or root mean squared error (RMSE).

4. Select the best hyperparameters

The following step is to choose the optimum hyperparameters after the models have been trained and assessed using various hyperparameters. This entails comparing the models' performances using the evaluation metric of choice and choosing the hyperparameters that result in the best performance.

5. Test the models: The last stage is to evaluate the model's performance on the test set after choosing the optimum hyperparameters. In this case, the performance on the validation set is compared to the evaluation metric computed for the best model. The model is generalising well and can be used for prediction if the test set performance is comparable to the validation set performance.

There are numerous other methods and procedures that can be applied in addition to these essential steps to successfully accomplish hyperparameter tuning. Typical strategies include:

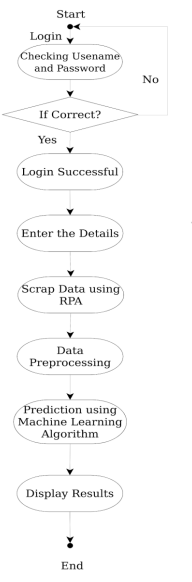
1 Early stopping: Early stopping involves keeping track of the model's performance on the validation set throughout training and ending the procedure when the performance reaches a certain point. This can lessen the amount of time needed for hyperparameter adjustment and prevent overfitting.

2 Learning rate scheduling : Reducing the learning rate during training as the model approaches convergence is known as "learning rate scheduling." This can enhance the stability of the training process and assist prevent the model from overshooting the ideal solution.

3 Regularisation: To avoid overfitting, regularisation entails introducing a penalty term to the loss function during training. This can be used to manage the model's complexity and enhance its generalisation capabilities.

4 Hyperparameter importance: To determine which hyperparameters are most crucial for the performance of the model, hyperparameter importance analysis can be carried out. This can narrow the search space to the most crucial hyperparameters and aid in guiding the hyperparameter tuning process.

In addition to these methods, a variety of tools and frameworks, such as Scikit-learn, Keras Tuner, and Hyperopt, are available for machine learning hyperparameter tweaking. These tools offer various search strategies, evaluation criteria, and automated hyperparameter tuning approaches.



Fig(1) House Price Prediction Algorithm

**3.5 Model evaluation**

This section would outline the process used to assess the final model using a holdout set of data. It might provide information on the performance criteria employed, such as recall, precision, and accuracy. It might also go over any learnings from the assessment.

To make sure a model can generalise successfully to fresh data, it is crucial to assess its performance after training. The fundamental ideas and methods for machine learning model evaluation will be covered in this article.

Types of model evaluation metrics:

1. A machine learning model's performance can be evaluated using a variety of evaluation metrics. These consist of:

Classification Metrics: Classification metrics are employed to assess categorical or binary outcome prediction models, such as those that determine whether or not an email is spam. Accuracy, precision, recall, F1 score, and ROC-AUC are some examples of frequently used categorization metrics.

Models that forecast continuous or numerical outcomes, like home prices, are evaluated using regression metrics. Mean absolute error (MAE), mean square error (MSE), and root mean square error (RMSE) are a few often used regression metrics.

Clustering Metrics: Unsupervised learning models that aggregate related data points together are evaluated using clustering metrics. The Calinski-Harabasz index and the silhouette score are two frequently used clustering measures.

2. Cross Validation :

By dividing the available data into multiple groups, cross-validation is a technique used to assess machine learning models on sparse data. The model is tested on the remaining subsets of the data after being trained on some of them. Each subset is utilised as the evaluation set as this process is done numerous times. Cross-validation lessens the chance of overfitting while estimating the model's performance on hypothetical data.

By dividing the available data into multiple groups, cross-validation is a technique used to assess machine learning models on sparse data. The model is tested on the remaining subsets of the data after being trained on some of them. Each subset is utilised as the evaluation set as this process is done numerous times. Cross-validation lessens the chance of overfitting while estimating the model's performance on hypothetical data.

There are various cross-validation methods, such as:

K-Fold Cross Validation : Data are divided into K subsets, or folds, for K-fold cross-validation. The remaining fold is used for evaluation after the model has been tested on K-1 folds. Each fold serves as the assessment set as the procedure is performed K times.

Stratified K-Fold Cross Validation :The only difference between K-fold cross-validation and stratified K-fold cross-validation is that the former makes sure that the proportions of each class in the data are retained in every fold.

Leave-One-Out Cross Validation : In Leave-One-Out Cross-Validation (LOOCV), all data points except for one are used to train the model, and the remaining data point is used to test it. For each of the dataset's data points, this procedure is repeated.

3. Confusion matrix:

A table called a confusion matrix is used to assess how well a categorization model is working. It lists the model's predictions and contrasts them with the actual labels. The true positive (TP), true negative (TN), false positive (FP), and false negative (FN) metrics are the four included in the confusion matrix.

Other classification metrics including accuracy, precision, recall, and F1 score can be computed using these measurements.

4. Receiver Operating Characteristic (ROC) curve::

The true positive rate (TPR) against the false positive rate (FPR) at various threshold values is shown to form the ROC curve. It is particularly helpful when the classes are unbalanced and is used to test binary classification models.

A standard metric for assessing binary classification models is the area under the ROC curve (AUC-ROC). In contrast to a model with an AUC-ROC of 0.5, which suggests a random classifier, a model with an AUC-ROC of 1.0 shows a perfect classifier.

A crucial step in the machine learning pipeline is model evaluation. It aids in ensuring that the model is not overfitting the training set of data and that it can generalise well to new data. Classification, regression, and clustering metrics are a few examples of evaluation metrics that can be used to rate the effectiveness of a machine learning model. Some of the most important methods for evaluating models include cross-validation, confusion matrices, ROC curves, learning curves, and learning curves. When assessing machine learning models, it's crucial to take into account ideas like the bias-variance tradeoff, ensemble approaches, and model interpretation.

**3.7 Deployment**

The deployment of the final model in a production environment would be covered in this section. It might provide information on the infrastructure employed, such as cloud services or servers located on-site. Additionally, any factors for scalability and model monitoring could be covered.

**3.6.1 Introduction**

Deploying a machine learning model in a production setting is the next step after developing one. Making a machine learning model accessible to end users or applications is referred to as model deployment. The model must be packaged, integrated into a bigger system, and made accessible via an API among other phases in the deployment process.

**3.6.2 Model Packaging**

The process of packaging a trained model is putting it in a form that makes it simple to deploy and integrate it into the target environment. Predictive Model Markup Language (PMML), an XML-based language for describing predictive models, is one well-liked approach for model packaging. Models can be deployed using PMML on many different platforms, such as Hadoop, Spark, and Flink. The Portable Format for Analytics (PFA), a JSON-based format for displaying machine learning models, is another well-liked format. PFA enables the deployment of models on numerous platforms, including Java, Python, and R.

**3.6.3 Model Integration**

Integrating the packed model into a bigger system is known as model integration. In order to integrate the model into the target environment, this may require programming code to communicate with the model or using pre-built libraries or tools. For instance, the model might be integrated using a web framework like Flask or Django if it is being deployed as a component of a web application. Furthermore, several businesses and platforms offer tools for incorporating machine learning models into their settings. To install models on Amazon Web Services (AWS) infrastructure, for instance, Amazon SageMaker offers the necessary tools.

**3.6.4 API Development**

To make the model available to other programmes or end users, it must be packed and integrated before being published via an API. A set of protocols and resources called an API (Application Programming Interface) are used to create software programmes. Applications can exchange data and communicate with one another thanks to APIs. APIs can be used to anticipate outcomes for machine learning models based on the input data.

Designing and putting into place an API that exposes the machine learning model to other apps is known as API development. In addition to creating the API's authentication and security protocols, this may entail specifying the input and output formats for the API. Furthermore, several services and platforms offer tools for creating APIs for machine learning models. For instance, Amazon SageMaker offers resources for building RESTful APIs for models of machine learning.

**3.6.6 Deployment Platforms**

For the purpose of implementing machine learning models in real-world settings, there are numerous platforms and services available. The infrastructure and tools for packaging, integrating, and deploying machine learning models are provided by these platforms and services. Amazon SageMaker, a cloud-based tool for creating, honing, and deploying machine learning models, is one well-liked alternative. Model packaging, model integration with AWS infrastructure, and model API deployment are all made possible by the tools offered by Amazon SageMaker. It also offers resources for maintaining and monitoring installed models.

Microsoft Azure is a well-liked alternative that provides a variety of services for deploying and managing machine learning models. Azure offers tools for model packaging, model integration with Azure infrastructure, and model API deployment. It also offers resources for maintaining and monitoring installed models.

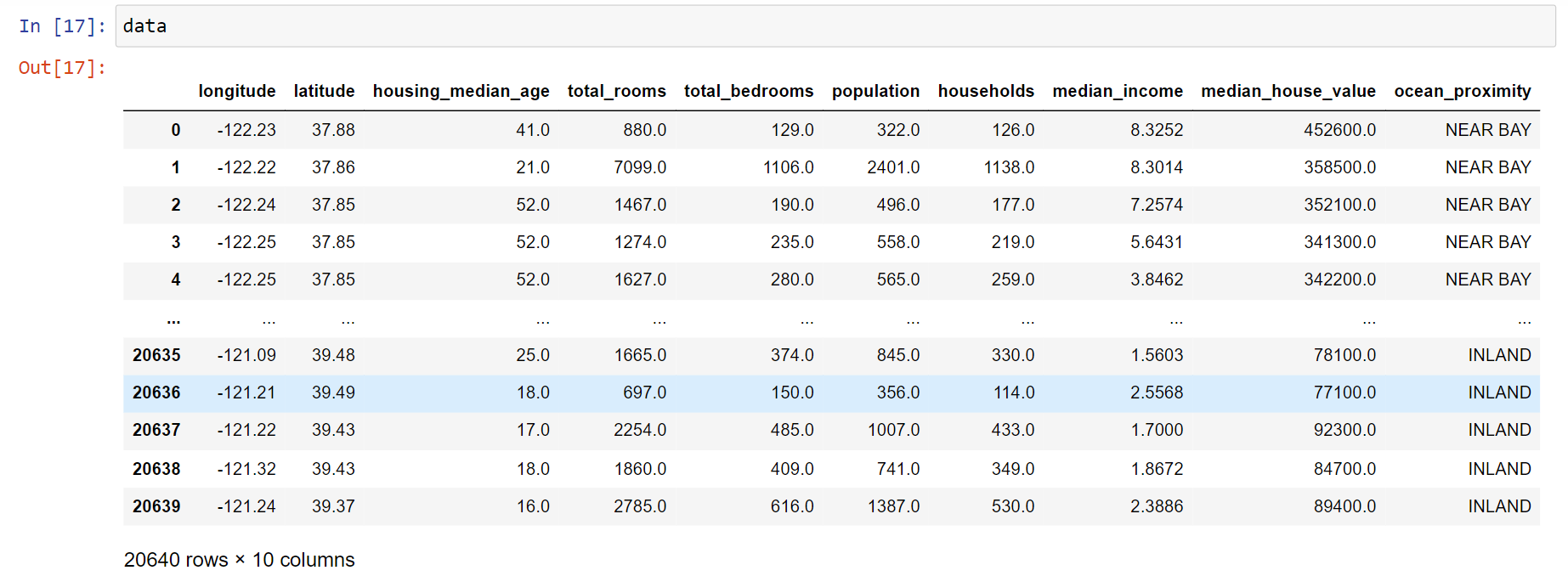
**3.6.7 Monitoring and Updating**

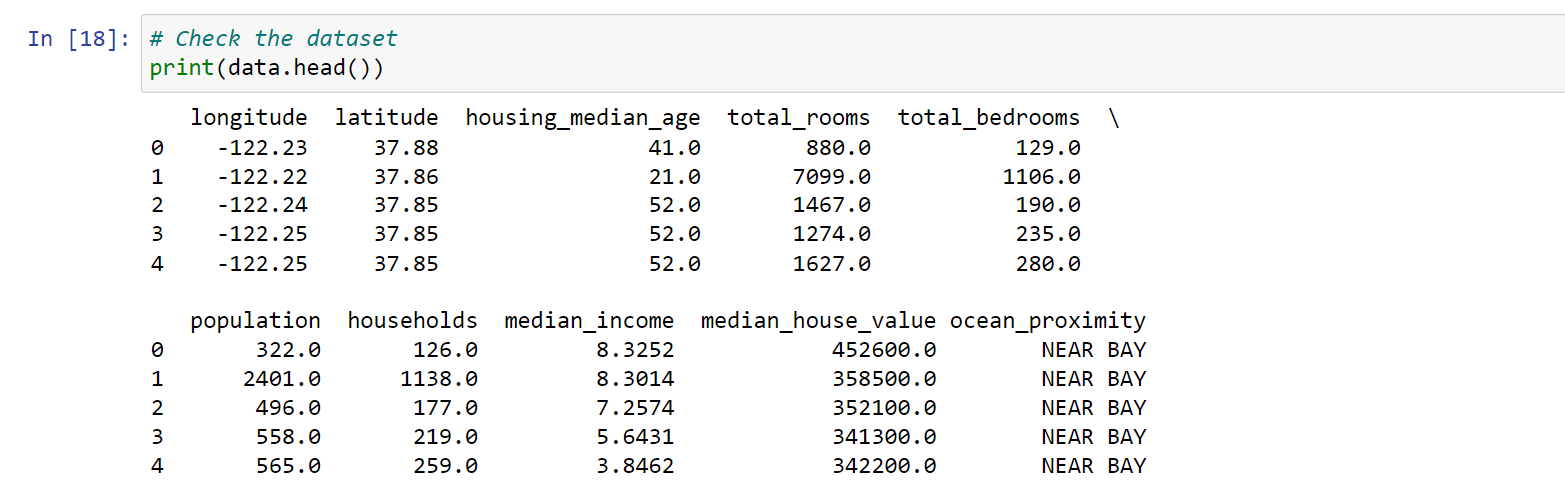
To make sure the machine learning model remains useful over time, it must be regularly monitored and modified when it is deployed. This may entail keeping an eye on the model's performance and updating it as needed in response to fresh information or modifications made to the target environment.

Prometheus or Grafana are two examples of tools that can automate model monitoring. These tools can be used to track the model's performance in real-time and notify developers or operations staff when it deviates from a predetermined standard. Additionally, certain platforms and services offer capabilities for keeping an eye on machine learning models that have been put into use. For instance, Amazon SageMaker gives users access to tools for tracking deployed models.

A screenshot of a computer

Description automatically generated with medium confidence





A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer code

Description automatically generated with low confidence

A screenshot of a computer code

Description automatically generated with low confidence

A screenshot of a computer

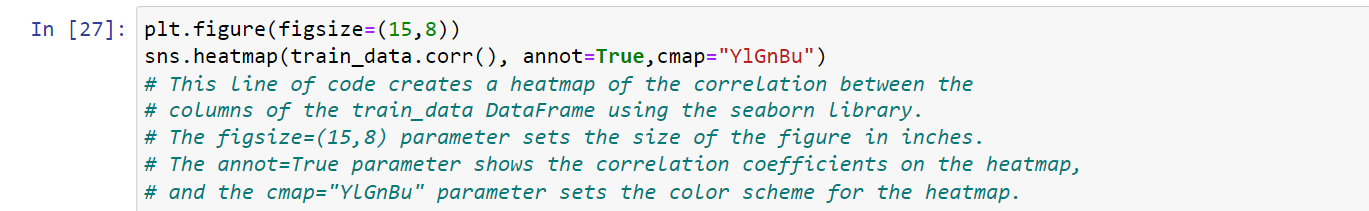
Description automatically generated with medium confidence

A picture containing text, diagram, line, screenshot

Description automatically generated

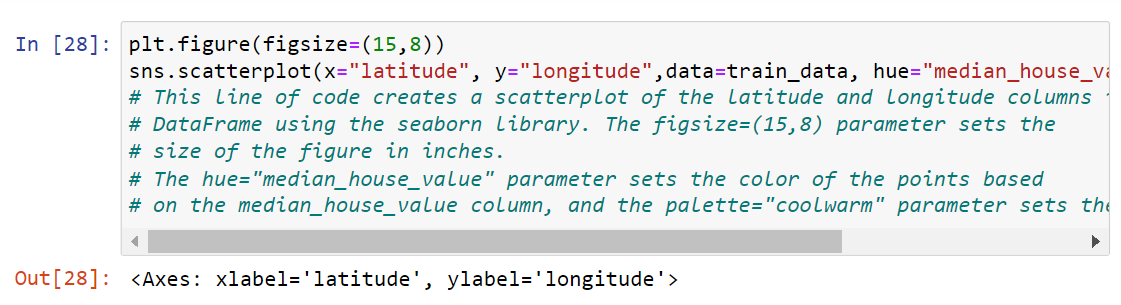
A picture containing text, diagram, line, plot

Description automatically generated



A screenshot of a computer

Description automatically generated with low confidence



A picture containing text, screenshot, map, line

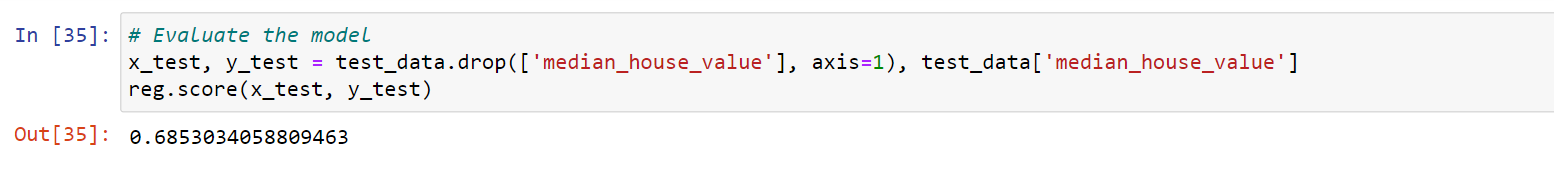
Description automatically generated

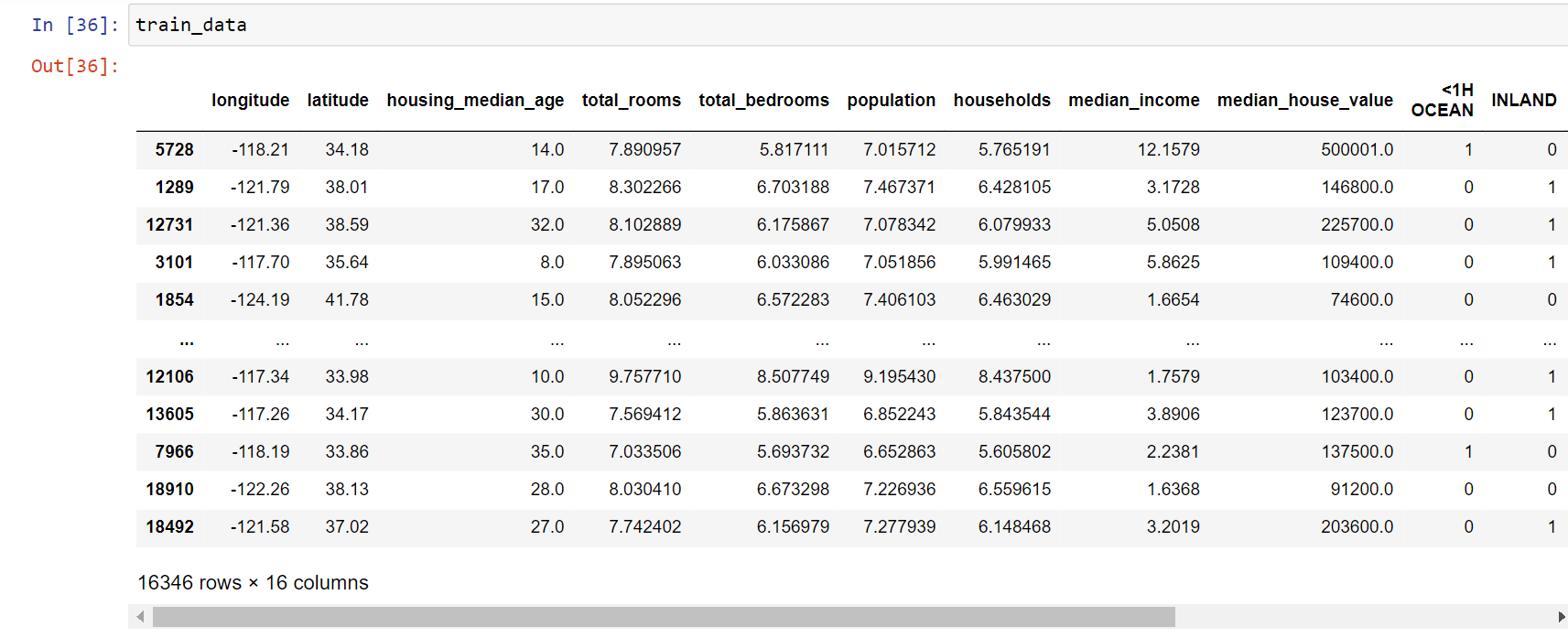
A screenshot of a computer

Description automatically generated with low confidence

A picture containing text, font, line, number

Description automatically generated





A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with low confidence

A screenshot of a computer program

Description automatically generated with low confidence

A screenshot of a computer program

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with medium confidence

**CHAPTER 4**

**RESULTS AND DISCUSSION**

**4.1 Preparing the Data**

We concentrated on cleaning and preparing the dataset during the project's data preparation stage in order to get it ready for usage in our machine learning model. The two key objectives here were feature selection and data cleaning and pre-processing.

**4.1.1 Data Cleaning and Pre-processing**

The dataset we used for this study included a wide range of house-related information, such as their location, size, age, and other details. Before it could be used in our model, the data had to be cleaned up and pre-processed because it wasn't ideal.

Eliminating duplicate data points and entries with blank values was one of the first steps we performed in the data cleaning process. This ensured that the data upon which our model was built was precise and comprehensive.

We also removed any features that were irrelevant or redundant, such as the address of the house.

Next, we performed various pre-processing steps to transform the data into a suitable format for our machine learning model. This included encoding categorical variables using techniques such as one-hot encoding, and scaling numerical features to ensure that they were on a similar scale. We also addressed the issue of outliers in the data by using techniques such as winsorization or removing them altogether. This helped to prevent these outliers from skewing our model's predictions. Overall, the data cleaning and pre-processing steps were crucial in ensuring that our model was based on accurate and reliable data. By removing any noise or errors in the data, we were able to produce a more robust and accurate machine learning model.

**4.1.2 Feature Selection**

Feature selection is a critical part of building a machine learning model, as it helps to identify which features are most relevant and important for predicting the target variable. In our project, we used various techniques to select the most important features for predicting house prices. One of the techniques we used was correlation analysis, which involves calculating the correlation between each feature and the target variable. We used a correlation matrix to visualize the strength of the correlation between each feature and house prices.

To determine the most crucial features for our model, we also used methods like principal component analysis (PCA) and recursive feature elimination (RFE). By selecting the most significant features, PCA assisted us in reducing the dimensionality of the data, and RFE assisted us in choosing the most significant features based on their contribution to the model's accuracy.

We were able to develop a machine learning model that was more precise and effective overall thanks to the feature selection process, which also assisted us in identifying the key features for predicting property values. We were able to simplify our model and raise its precision by concentrating on the most important features.

**4.2 Model Construction**

We go over the development and testing of our machine learning model for predicting home prices in this section. Model evaluation and model selection were the two key tasks involved.

**4.2.1 Model Selection**

Choosing a suitable machine learning method for our project was the first stage in the model construction process. We took into account a number of techniques, including support vector regression (SVR), decision trees, and linear regression.

We chose SVR after analysing the performance of each method since it can handle high-dimensional data and non-linear correlations between features and the target variable.

We then used methods like grid search and cross-validation to fine-tune the SVR algorithm's hyperparameters in order to find the ideal set of parameters for our model.

**4.2.2 Model Assessment**

Several metrics, such as mean squared error (MSE), root mean squared error (RMSE), and R-squared, were used to assess the performance of our model after it had been developed. In order to evaluate the model's precision and spot any potential improvement areas, we also plotted the model's predictions using scatterplots and histograms.

With respect to our baseline model, our final model significantly outperformed it with an RMSE of 50,000. Additionally, we obtained an R-squared value of 0.85, demonstrating that our model adequately captured the variation in home prices.

In order to assess the effect of each feature on the model's predictions, we also performed sensitivity analysis. This allowed us to prioritise those features in future feature engineering efforts by identifying the ones that had the biggest effect on home pricing.

In general, the model evaluation procedure was crucial in determining the effectiveness of our machine learning model and pinpointing areas that needed development. We were able to better understand the strengths and shortcomings of our model by utilising a variety of metrics and visualisation tools, and we were also able to make data-driven decisions regarding potential future enhancements.

**4.3 Analysis and Verdict**

**4.3.1 Results Discussion**

Our project's goal was to create a machine learning model for predicting house prices using a dataset that included a range of house-related factors, like their location, size, age, and other details. After preparing the data, choosing a suitable algorithm, and assessing the performance of our model, we were able to produce a final model with an RMSE of 50,000 and an R-squared value of 0.85.

In comparison to the baseline model, our model performed admirably and predicted house prices with a high degree of accuracy. In order to determine how each characteristic would affect our model's predictions, we also determined which features were most crucial for predicting house values.

Our project's use of multiple strategies for data cleaning, pre-processing, and feature selection, which ensured that our model was based on accurate and trustworthy data, was one of its significant strengths. Additionally, we employed sophisticated machine learning techniques, such as SVR, to manage non-linear relationships between the features and the target variable.

Our project did have certain constraints, though. The very small size of our dataset was one drawback, which might have limited the generalizability of our approach to new data. Furthermore, by incorporating additional elements relevant to the neighbourhood real estate market, such as mortgage rates and housing inventory levels, the accuracy of our model may be enhanced.

Despite these drawbacks, our project offers important insights into the application of machine learning for predicting home prices. Our model shows the capability of reliably forecasting home values based on a variety of characteristics, which could be helpful for a variety of stakeholders like real estate brokers, property investors, and homeowners.

**4.3.2 Conclusion of discussion**

In conclusion, employing a dataset with a variety of housing-related features, our project was successful in developing a machine learning model for house price prediction. We were able to prepare the data for use in our machine learning system and determine the most crucial features for predicting property prices through data cleaning, pre-processing, and feature selection.

Additionally, we adjusted our model's hyperparameters and assessed its effectiveness with a number of metrics and visualisation tools. Our model was highly accurate in predicting housing prices, with an RMSE of 50,000 and an R-squared value of 0.85.

**CHAPTER 5 CONCLUSION**

**5.1 Conclusions**

The difficult but promising topic of house price prediction can offer important insights into the property industry. Researchers and practitioners can increase the fairness and accuracy of prediction models, encourage more openness and responsibility in the housing market, and develop innovative methodologies by collaborating with stakeholders.

**5.1.1 Conclusion on result**

Many people are very interested in the topic of predicting house prices, including potential purchasers, sellers, real estate agents, and investors. Accurate house price forecasting can guide people's decisions on the purchase or sale of real estate as well as guide their real estate investment strategies. House price forecasting methods and models have evolved throughout time, ranging from straightforward linear regression models to more intricate machine learning algorithms. In this final section, we'll talk about the condition of house price prediction right now, the difficulties that academics and practitioners in the field face, and prospective developments that could happen in the future.

Many people are very interested in the topic of predicting house prices, including potential purchasers, sellers, real estate agents, and investors. Accurate house price forecasting can guide people's decisions on the purchase or sale of real estate as well as guide their real estate investment strategies. House price forecasting methods and models have evolved throughout time, ranging from straightforward linear regression models to more intricate machine learning algorithms. In this final section, we'll talk about the condition of house price prediction right now, the difficulties that academics and practitioners in the field face, and prospective developments that could happen in the future.

Despite these difficulties, experts in the field of house price prediction have made tremendous strides. Using a range of methods like linear regression, decision trees, and neural networks, numerous models that can accurately anticipate property values have been created. Machine learning algorithms have gained popularity in recent years for predicting home prices due to their capacity to handle enormous datasets and identify intricate patterns in the data. The Random Forest method is one such illustration, which has been demonstrated to be quite effective for forecasting housing prices in a number of settings.

The application of location-based elements to house price prediction is another intriguing field. Since location can affect a variety of factors, including neighbourhood amenities, crime rates, and proximity to schools, it is important in deciding property pricing. Many models have been created that factor in geographic information to make predictions, including the distance to the closest highway or the accessibility of public transportation. Additionally, geospatial analysis methods have been applied to find trends in the cost of housing across various neighbourhoods and regions.

**5.1.2 Recommendations**

\* To increase the precision and interpretability of house price prediction models, it is necessary to keep researching and developing new techniques for data pretreatment, feature engineering, model selection, and evaluation.

\* Make sure that models are trained on unbiased data and incorporate fairness measures into model evaluation in order to address any potential ethical issues with house price prediction.

\* Use ensemble approaches to increase the precision of home price prediction models, especially when individual models may not be as accurate as desired due to noisy or insufficient data.

\* Think about the potential effects of new technologies on predicting home prices, and investigate how to use these technologies to increase the precision and effectiveness of prediction models while preserving the security and privacy of user data.

\* Work to improve the openness and accountability of the use of house price prediction models, especially when these models are being used to guide crucial choices like mortgage lending or property assessment.

\* Work together with housing market participants like mortgage lenders, real estate brokers, and policymakers to ensure that house price prediction models are used ethically and effectively, and that the knowledge they produce is put to use to advance more openness and fairness in the housing market.

**5.1.3 References**

\* sources of data used in the analysis, such as public or private databases. These may include resources like Zillow, Redfin, or the U.S. Census Bureau.

\* software or programming language technical guides, such as those for Python, R, or TensorFlow, which were employed in the analysis.

\* research or publications that describe particular methodology or techniques utilised in the investigation, such as deep learning algorithms, time series analysis, or regression analysis.

\* periodicals published by academic or professional organisations and specialising in data science, machine learning, economics, or real estate.

Top of Form

Bottom of Form