HOUSE PRICE PREDICTION USING MACHINE LEARNING

A Project Report

Submitted in partial fulfilment of the

Requirements for the award of the Degree

of

MASTER OF SCIENCE (INFORMATION TECHNOLOGY)

By

PRATIK P BAMBULKAR

Under the esteemed guidance of

Prof. Dhanraj Jadhav

DEPARTMENT OF INFORMATION TECHNOLOGY

ML DAHANUKAR COLLEGE OF COMMERCE

(Affiliated to University of Mumbai)

DIXIT ROAD, VILE PARLE (EAST), PIN CODE: 400057 MAHARASHTRA

2020-2021

ABSTRACT

Presently, there are manual specialists who examine highlights of a real estate market and gives the expectations, however in some cases it will be correct or wrong along these lines, it makes an enormous issue for the organization, their mistake rates increments, so for this issue there is an answer by utilizing Machine Learning models to build their benefits. We need to examine the information appropriately, understanding the highlights and afterward carrying out different models on it and choosing the best one for the task which will give us precise cost for the houses in that specific zone. It will help the land organization to acquire benefits by utilizing this House Price Prediction ML model.

ACKNOWLEDGEMENT

M.L. Dahanukar College of Commerce has provided me with this wonderful opportunity to provide me with this wonderful opportunity to learn and practically implement all that I have studied in my academic years. I have followed a systematic linear and rational approach while dealing with the software project.

The project has benefited from the many useful suggestions provided to us by numerous individuals and hence we take the opportunity to thank all of them.

We would like to thank Prof.Archana Talekar Our head of department in MASTER OF SCIENCE (INFORMATION TECHNOLOGY) for providing us with enough lab facilities and giving us morale support.

We would also like to thank Prof.Dhanraj Jadhav whose guidance and experience has helped us to overcome many complexities in the development.

Lastly, we would like to thank all those who directly or indirectly helped us in the completion of the project.

Thanking you,

Pratik P Bambulkar

DECLARATION

I hereby declare that the project entitled, "**House Price Prediction using Machine Learning**" done at **place where the project is done**, has not been in any case duplicated to submit to any other university for the award of any degree. To the best of my knowledge other than me, no one has submitted to any other university.

The project is done in partial fulfilment of the requirements for the award of degree of **MASTER OF SCIENCE (INFORMATION TECHNOLOGY)** to be submitted as final semester project as part of our curriculum.

Pratik Bambulkar

TABLE OF CONTENTS

SYNOPSIS

TOPIC OF THE PROJECT:

House Price Predictions Using Machine Learning Algorithms.

OBJECTIVE AND SCOPE:

Currently, there are manual experts who analyze features of a housing market and gives the predictions, but sometimes it will be right or wrong because of this, it creates a huge problem for the company, their error rates increases, so for this problem there is a solution by using Machine Learning models to increase their profits. We need to study the data properly, understanding the features and then implementing various models on it and selecting the best one for the project which will give us accurate price for the houses in that particular area. It will help the real estate company to gain more profits by using this House Price Prediction ML model.

PROCESS DESCRIPTION:

We need a system to predict the prices of house in the future, as the house prices are increasing every year. So this house price prediction will help the company to determine the selling price of the house and also help the customer to get the right price of the house to purchase.

For the real estate, House Price Forecasting is a really important topic. To derive the useful knowledge from historical data of property markets is attempts by literature. Machine

learning techniques are applied to analyze historical property transactions to discover useful models for house buyers and sellers. The main goal is first to analyze the data properly understanding the real estate or company requirements that are the features of the house holding like the number of bedrooms, number of bathrooms, study room, etc. Then, afterwards understanding their needs that means the given task which is predicting the house prices in given area or classifying the houses in two different classes like high price houses and low price houses. So in this project the main focus is predicting the price of house in any given area with the help of Machine Learning Algorithms.

In this Housing Price Predictions the requirements are given to us by the company or real estate that means features are available to us which means it is a Supervised Learning.

Supervised Learning in which the example inputs is provided by the computer then they are labelled with the desired outputs. The main goal of this algorithm is able to 'learn' by comparing its actual output with the desired output to find out the errors in it, and modify the model accordingly. It uses the patterns to predict the labelled values on additional unlabelled data.

Having all the features and now knowing the task that is to predict the price of houses we are going to used Regression Algorithm. It predict the output values based on input features from the data fed in the system and it is used to predict the continuous value. It is a go-to methodology algorithm which builds a model on the features of training data and using the model to predict the value for new data.

It is a batch learning technique which means we already had a data on which we will make model on it.

Steps Involved:-

1) Importing the required packages into our python environment
2) Importing the house price data and do some EDA on it.
3) Data Visualization on the house price data
4) Feature Selection & Data Split
5) Modelling the data using the algorithms.
6) Evaluating the built model using the evaluation metrics.

RESOURCES AND LIMITATIONS:

Hardware: Minimum System Requirements.

RAM: 4 GB.

System Type: 64-bit operating system, x64-based processor.

Software:

Language: Python.

Tools: PyCharm, Jupyter Notebook.

Limitations:

If the customer wants to predict the future price of the house, so it is not possible because there is a risk, as the prices of that area increases continuously. So the customers tend to hire a broker or agent to minimize this error, but in this again the cost of the process increases.

CONCLUSION:

It will provide us the accurate prices of the houses in that particular area with the help of using accurate and robust machine learning models which will help the real estate company to know the right prices of that houses and to make profits for their company. By using this model it will help them to reduce the human error.

CHAPTER 1

INTRODUCTION

1.1 Background:

In today's world the accuracy is really important. House Price Prediction is important to Real Estate Company as well as for the customers who are going to buy the house because everyone should get the actual house price, in terms of which the company will get the actual price and profits. Every single organization in today's real estate business is operating fruitfully to achieve a competitive edge over alternative competitors. There is a need to simplify the process for a normal human being while providing the best results. This project proposes a system that predicts house prices using a regression machine learning algorithm. In case you're going to sell a house, you have to recognize what sticker price to put on it. What's more, a PC calculation can give you a precise gauge!

A precise expectation on the house cost is imperative to forthcoming property holders, engineers, financial backers, appraisers, charge assessors and other housing market members, for example, contract moneylenders and guarantors. Customary house value expectation depends on cost and deal value correlation lacking of an acknowledged norm and an affirmation cycle. Hence, the accessibility of a house value expectation model assists load up with increasing a significant data hole and improve the proficiency of the housing market.

Request that a home purchaser depict their fantasy house, and they likely will not start with the stature of the cellar roof or the closeness to an east-west railroad. However, this playground competition's dataset demonstrates considerably more impacts value arrangements than the quantity of rooms or a white-picket fence. However, this dataset related to the project proves that may have more effects on the housing price than the number

of bedrooms or floors. Additionally, I need to foresee the sensible lodging cost with these parts of the houses by utilizing this dataset.

1.2 Objectives:

1. Predict the sale price for each house.

2. Minimize the difference between predicted and actual rating (RMSE/MSE).

Regression is a machine learning apparatus that encourages you to make expectations by taking in – from the current measurable information – the connections between your target parameter and a lot of different independent parameters. As per this definition, a house's cost relies upon parameters, for example, the number of rooms, living region, area, and so forth. On the off chance that we apply counterfeit figuring out how to these parameters, we can compute house valuations in a given land region.

We are using the housing dataset which is taken UCI machine learning repository. We will analyse the data, explore the features and then implement different models on it and with the help of RMSE, having less error will be selected as a model.

1.3 Purpose, Scope and Applicability:

1.3.1 Purpose:

Machine Learning is a sub-field of computerized reasoning (AI) that gives frameworks the capacity to consequently take in and improve for a fact without being unequivocally modified.

For the way toward learning (model fitting) we need to have accessible a few perceptions or information (otherwise called tests or models) to investigate possible fundamental examples, covered up in our information. These learned patterns are nothing more that some functions or decision boundaries. Machine Learning has been utilized for quite a long time to bring to the table picture acknowledgment, spam location, characteristic discourse appreciation, item suggestions, and clinical findings. Today, ML calculations can help us upgrade network protection, guarantee public security, and improve clinical results. Machine Learning frameworks can likewise make client support better and vehicles more secure. In case you will sell a house, you need to understand what sticker price to put on it. Furthermore, a PC calculation can give you a precise gauge!

Machine learning algorithms are usually categorized as supervised or unsupervised.

- ***Supervised learning*** is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labelled training data consisting of a set of training examples. In supervised learning, every model is a couple comprising of an information object (normally a vector) and an ideal yield esteem (likewise called the supervisory signal).
- ***Unsupervised learning*** is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labelled responses. The most well-known unsupervised learning technique is bunch examination, which is utilized for exploratory information investigation to discover covered up examples or gathering in information. The clusters are displayed utilizing a proportion of comparability which is characterized upon measurements like Euclidean or probabilistic distance.
- ***Reinforcement learning (RL)*** is an area of machine learning concerned with how intelligent agents ought to take actions in an environment in order to maximize the

notion of cumulative reward.). Reinforcement Learning taking in contrasts from supervised learning in not requiring marked info/yield sets be introduced, and in not requiring imperfect activities to be expressly revised. Rather the emphasis is on finding a harmony between investigation (of unfamiliar region) and misuse (of current information).

In this project, we will develop and evaluate the performance and the predictive power of a model trained and tested on data collected. Once we get a good fit, we will use this model to predict the monetary value for house. A model like this would be entirely important for a realtor who could utilize the data gave in a regular basis.

Steps for House Price Predictions:

- Getting the Data and Previous Pre-process.
- Data Exploration.
- Feature Observation.
- Exploratory Data Analysis.
- Developing a Model.
- Analysing Model's Performance.
- Evaluating Model's Performance.
- Cross-Validation.

1.3.2 Scope:

While getting the accurate prices of houses will create a big profit for the real estate company. Manually getting insights for the house price will create a huge loss for the company. Land is quite possibly the most universally perceived areas. Lodging, Retail, Hospitality and Commercial are four sub-areas of it. The development of this area can be

seen by the development in Corporate Environment and need for Office Spaces just as metropolitan and Semi-Urban Accommodations.

Nowadays, e-education and e-learning is highly influenced. Everything is moving from manual to robotized frameworks. The target of this venture is to foresee the house costs to limit the issues looked by the client. The current technique is that the client moves toward a realtor to deal with his/her ventures and recommend appropriate bequests for his speculations. But this method is risky as the agent might predict wrong estates and thus leading to loss of the customer's investments. The manual strategy which is at present utilized in the market is out dated and has high danger. In order to conquer this shortcoming, there is a requirement for a refreshed and mechanized framework. Machine Learning algorithms can be used to help investors to invest in an appropriate estate according to their mentioned requirements. Also the new system will be cost and time efficient. This will have simple operations. The proposed system works on Regression Algorithm.

The scope consists of what the issues are solved and function created for solving it.

- *Issues:*
    - Sometimes it creates problem to manually predict price for a given house.
    - To get lot of insights in which time can be consumed more to get actual results.
    - Not best results are provided.
- *Functions:*
    - First of all it give accurate results by using machine learning models.
    - House price prediction is based on cost and sale price comparison lacking of an accepted standard and a certification process. Consequently, the accessibility of a house value expectation model assists load up with

increasing a significant data hole and improve the productivity of the housing market.

- Trends and Patterns Are Identified with Ease.
- Machine Learning Improves Over Time.
- Machine Learning Lets You Adapt Without Human Intervention.
- Enables Automation.

1.3.3 Applicability:

The housing market is quite possibly the most urgent parts of any public economy. Henceforth, perceptions of the housing market and precise expectations of land costs are useful for land purchasers and merchants just as financial trained professionals. In any case, land estimating is a confounded and troublesome assignment inferable from numerous immediate and aberrant factors that unavoidably impact the exactness of forecasts. When all is said in done, factors impacting land costs could be quantitative or qualitative. The quantitative factors conceivably incorporate macroeconomic components, business cycles, and land ascribes. The macroeconomic elements contain joblessness rates, share record, current record of a country, modern creation, and total national output. Properties of land, for instance, incorporates past deal costs, land zone, long periods of developments, floor space, surface territory, number of floors and building conditions. The qualitative elements allude to subject inclinations of leaders, like perspectives, building styles, and living climate. Nonetheless, a few challenges emerge in information assortment for qualitative elements. For qualitative elements, now and then these information are experiencing absence of estimations. Information of qualitative factors some of the time are experiencing absence of estimations. Hence, qualitative components are difficult to gauge. Consequently, this

investigation didn't take qualitative elements affecting land costs into contemplations and utilized quantitative information.

Individuals hoping to purchase another home will in general be more traditionalist with their spending plans and market systems. The current framework includes estimation of house costs without the vital expectation about future market patterns and cost increment. The objective of this undertaking is to anticipate the productive house estimating for land clients as for their spending plans and needs. By examining past market patterns and value ranges. These days, e-instruction and e-learning is profoundly impacted. Everything is moving from manual to computerized frameworks. The current technique is that the client moves toward a realtor to deal with his/her ventures and recommend appropriate domains for his speculations. Yet, this strategy is hazardous as the specialist would foresee wrong homes and subsequently prompting loss of the client's ventures. So with the help of machine learning we can solve this problem.

1.4 Achievements:

In House price prediction the dataset is taken from UCI Machine Learning, after analysing the data properly the datasets which present in the Boston House data I understand that which features are more important to find the house price. Doing the EDA Exploratory Data Analysis in which it is a way of visualizing, summarizing and interpreting the information that is hidden in rows and column format. Looking for correlations that is to find the outliers and determine which factors is most influencing that is the features affecting on house prices. In this predicting prices on a dataset which comes under in supervised learning. According to supervised learning we will used regression algorithms which will help us to

find the house price predictions. In this main goal achieved is that how we can do EDA, feature scaling, model selection, training-testing approach.

1.5 Organization of Report:

House costs increment consistently, so there is a requirement for a framework to anticipate house costs later on. House price prediction can help the developer determine the selling price of a house and can help the customer to arrange the right time to purchase a house.

Chapter 2:

There is Literature review in which various researcher's had work on the thesis of House Price Prediction. It explains about which model is used and on which dataset they had made the prediction model.

Chapter 3:

The Planning of the project is done in this chapter which dataset is used in this project, model, Programming Language, Flowchart, measure performance, Machine learning libraries.

Chapter 4:

In this chapter the system is design that is the designing is described in detail with features and operations.

CHAPTER 2

REVIEW OF LITERATURE

In our literature survey we have investigated various researches on this particular domain some of them are as follows:-

According to researcher's **Nikita Malik, Vidhu Gaba, Priyansh (2020) [1]**. They have done research on **Employing Machine Learning for House Price Prediction**. AI (ML) targets creating self-learning calculations utilizing datasets, to such an extent that the machine can had the option to project future movement dependent on the past information. It has displayed noteworthy improvements over the course of the years with the fast expansion away limit and preparing force of PCs, and accomplished high importance with its capacity to create frameworks utilized routinely in industry, training, and somewhere else. Taking the instance of one such application in land, in this paper, ML has been utilized to help anticipate the costs of houses by gaining from an example dataset comprising of characteristics which impact this expense. Initially, the central ideas of ML, alongside its applications are investigated. Then, concerning house value forecast, highlight appraisal, learning procedures and the libraries utilized for execution of the framework utilizing Python are examined. In view of the acquired expectation results, execution of the ML approach is assessed and ends are drawn.

According to research conducted by **Darshil Shah, Harshad Rajput, Jay Chheda. (2020) [2]** on topic **House Price Prediction Using Machine Learning and RPA**. They are using CatBoost calculation alongside Robotic Process Automation for continuous information extraction. In this day and age, everybody wants for a house that suits their way of life and gives conveniences as per their requirements. House costs continue changing as

often as possible which demonstrates that house costs are frequently misrepresented. There are numerous elements that must be contemplated for anticipating house costs like area, number of rooms, cover zone, how old the property is? Furthermore, other essential neighbourhood conveniences. Robotic Process Automation includes the utilization of programming robots to computerize the undertakings of information extraction while ML calculation is utilized to foresee house costs regarding the dataset.

The research is conducted by **Dr. M. Thamarai, Dr. S P. Malarvizhi (2020) [3]** on **House Price Prediction Modeling Using Machine Learning.** They had work on Decision Tree Classification, decision tree regression and multiple linear regression and is actualized utilizing Scikit-Learn Machine Learning Tool.ML is seeing its development all the more quickly in this decade. Numerous applications and calculations advance in Machine Learning every day. One such application found in diaries is house value expectation. House costs are expanding each year which has required the displaying of house value expectation. They built models, which assist the clients for buying a house appropriate for their need. Proposed work utilizes the ascribes or highlights of the houses, for example, number of rooms accessible in the house, age of the house, voyaging office from the area, school office accessible close by the houses and shopping centres accessible close by the house area. House accessibility dependent on wanted highlights of the house and house value forecast are displayed in the proposed work and the model is built for an unassuming community in West Godavari region of Andhra Pradesh.

According to researcher's **Ahmad Abdulal, Nawar Aghi (2020) [4]** on topic **House Price Prediction.** This investigation proposes a presentation correlation between ML Regression algorithms and Artificial Neural Network (ANN). The Regression algorithm utilized in this examination are Multiple Linear, Least Absolute Selection Operator (Lasso), Ridge, Random Forest. Also, they had investigated to examine the connection between

factors to decide the main factors that influence house costs in Malmö, Sweden. There are two datasets utilized in this examination which called public and local. They contain house costs from Ames, Iowa, United States and Malmö, Sweden, individually. The precision of the forecast is assessed by checking the root square and root mean square error scores of the preparation model. The test is performed subsequent to applying the necessary pre-preparing techniques and parting the information into two sections. Notwithstanding, one section will be utilized in the training and the other in the test stage. They have additionally introduced a binning system that improved the exactness of the models. This postulation endeavours to show that Lasso gives the best score among different calculations when utilizing the public dataset in training. The correlation graphs show the variables' level of dependency. In addition, the empirical results show that crime, deposit, lending, and repo rates influence the house prices negatively. Where inflation, year and unemployment rate sway the house costs emphatically.

According to the research conducted by **Alisha Kuvalekar, Shivani Manchewar, Sidhika Mahadik. (2020) [5]** on **House Price Forecasting using Machine Learning.** The housing market is a champion among the most engaged with respect to estimating and continues to vacillate. It is one of the excellent fields to apply the thoughts of ML on the best way to upgrade and anticipate the expenses with high exactness. The goal of the paper is the expectation of the market estimation of a land property. This framework helps locate a beginning cost for a property dependent on the topographical factors. By separating past market examples and worth ranges, and coming progressions future costs will be expected. This assessment intends to foresee house costs in Mumbai city with Decision tree regressor. It will assist customers with placing assets into an inheritance without moving towards a specialist. The consequence of this exploration demonstrated that the Decision tree regressor gives an exactness of 89%.

According to the researcher's **Winky K.O. Ho , Bo-Sin Tang & Siu Wai Wong (2020) [6]**. They had research **on Predicting property prices with machine learning algorithms**. This investigation utilizes three ML calculations including, Support vector machine (SVM), Random Forest (RF) and Gradient boosting machine (GBM) in the examination of property costs. They applied these strategies to analyze an information test of around 40,000 lodging exchanges in a time of more than 18 years in Hong Kong, and afterward thinks about the consequences of these calculations. Regarding predictive power, RF and GBM have accomplished better execution when contrasted with SVM. The three presentation measurements including mean squared mistake (MSE), root mean squared error (RMSE) and mean absolute percentage error (MAPE) related with these two calculations likewise unambiguously outflank those of SVM. In any case, their investigation had discovered that SVM is as yet a helpful calculation in information fitting since it can deliver sensibly precise expectations inside a tight time requirement. Their decision is that ML offers a promising, elective strategy in property valuation and examination research particularly according to property value forecast.

The research is conducted by **Sayan Putatunda (2019) [7]** on **PropTech for Proactive Pricing of Houses in Classified Advertisements in the Indian Real Estate Market**. Property Technology (PropTech) is the following huge thing that will upset the housing market. These days, there has been utilizations of Machine Learning (ML) and Artificial Intelligence (AI) in practically all the spaces however for quite a while the land business was very delayed in embracing information science and AI for critical thinking and improving their cycles. Notwithstanding, things are changing very quickly as there has been seen a great deal of selection of AI and ML in the US and European housing markets. Yet, the Indian housing market needs to get up to speed a great deal. This paper proposes an ML approach for tackling the house value forecast issue in the grouped promotions. This

examination centres on the Indian housing market. They apply progressed Machine learning calculations, for example, Random Forest, Gradient boosting and Artificial Neural Network on a genuine world dataset .They locate that the Random forest technique is the best entertainer regarding expectation precision.

This research is conducted **by Maharshi Modi, Ayush Sharma, Dr. P. Madhavan.(2020) [8]** on **Applied Research on House Price Prediction Using Diverse Machine Learning Techniques.** With the roaring civilization and always changing business sector prerequisites, it is crucial for realize the market floats. Today expectation of house costs as indicated by the patterns is the central quintessence of the examination. It is basic for a person to comprehend the business drifts so he can set up his budgetary necessities as per his prerequisites. Land is a steadily developing undertaking with an extending society. For a financial backer, it is fundamental to fathom the business floats, which can help him to guarantee in the correct manner and increase his business throughput. Once in a while customers get trick by the lie market rate set up the specialist because of which the land business is less clear nowadays. With an uptick in persuade of the dataset, it's feasible for a scientist to build up a model with high precision. The past model with diminished precision and overfitting of information decreases the proficiency, while this proposed framework resolves such issues and furnishes a superior and improved model with a rich UI. The chief expectation of their plan is to build up an extensive model that is profitable for a business society just as a person, which is the fundamental stub of this examination. This plan is expected to help a customer by decreasing his hands on work also remove his time and cash. They had used ML models like Extra Tree, Support Vector Machine, K Nearest Neighbor, Naive Bayes, Logistic Regression, Stochastic Gradient Descent, and they are coupled by actualizing the stacking strategy.

According to Research by **Neelam Shinde, Kiran Gawande. (2018) [9]** on **Valuation of House Prices using Predictive Techniques**. In this paper, they are foreseeing the deal cost of the houses utilizing different ML calculations. Lodging deals cost are dictated by various factors like territory of the property, area of the house, material utilized for development, age of the property, number of rooms and carports, etc. This paper utilizes ML calculations to construct the expectation model for houses. Here, ML calculations, for example, logistic regression and support vector regression, Lasso Regression technique and Decision Tree are utilized to fabricate a prescient model. They have thought about lodging information of 3000 properties. Calculated Regression, SVM, Lasso Regression and Decision Tree show the R-squared estimation of 0.98, 0.96, 0.81 and 0.99 individually. Further, they have thought about these calculations dependent on boundaries like MAE, MSE, RMSE and precision.

According to researcher's **Ping-Feng Pai and Wen-Chang Wang. (2020) [10]** on topic **Using Machine Learning Models and Actual Transaction Data for Predicting Real Estate Prices.** Land value forecast is significant for the foundation of land strategies and can help land proprietors and specialists settle on instructive choices. The point of this investigation is to utilize genuine exchange information and ML models to foresee costs of land. The real exchange information contain qualities and exchange costs of land that separately fill in as autonomous factors and ward factors for ML models. They had utilized four ML models-in particular, least squares support vector regression (LSSVR), classification and regression tree (CART), general regression neural networks (GRNN), and backpropagation neural networks (BPNN), to conjecture land costs. Also, genetic algorithms were utilized to choose boundaries of ML models. Mathematical outcomes demonstrated that the least squares support vector regression beats the other three ML models as far as determining exactness. Moreover, guaging results produced by the least squares support

vector regression are better than past related investigations of land value expectation regarding the normal total rate mistake. Consequently, the ML based model is a significant and doable approach to estimate land costs, and the least squares support vector regression can give moderately serious and good outcomes.

According to research by **Aswin Sivam Ravikumar. (2017) [11]** on **Real Estate Price Prediction Using Machine Learning.** The beneath record presents the execution of value expectation project for the housing markets and lodging. Numerous algorithms are utilized here to adequately expand the accuracy rate. They have done this task and actualized the calculations like hedonic regression, artificial neural networks, AdaBoost, J48 tree which is considered as the best models in the value expectation. These are considered as the base models and by the help of advanced data mining tools algorithms like a random forest, gradient boosted trees, multi-layer perceptron and ensemble learning models are utilized and expectation exactness is accomplished in a higher rate. The outcomes and assessment of these models utilizing the ML and progressed information data devices used like Weka, Rapid Miner will have the more impact in the value expectation.

According to researcher's **Chao Xue, Yongfeng Ju, Shuguang Li, Qilong Zhou and Qingqing Liu. (2020) [12]** on topic **Research on Accurate House Price Analysis by Using GIS Technology and Transport Accessibility.** House value forecast is urgent for the foundation of land arrangements and can help land proprietors and specialists settle on instructive choices. They examine this information and ML models to anticipate costs of land. The genuine exchange information contain characteristics and exchange costs of land that separately fill in as free factors and ward factors for ML models. This examination utilized four ML models-specifically, least squares support vector regression (LSSVR), classification and regression tree (CART), general regression neural networks (GRNN), and backpropagation neural networks (BPNN), to gauge land costs. Hereditary calculations were

utilized to choose boundaries of ML models. Mathematical outcomes demonstrated that the least squares support vector regression beats the other three ML models regarding determining precision. Moreover, anticipating results produced by the least squares support vector regression are better than past related investigations of land value expectation regarding the normal outright rate blunder. Along these lines, the ML based model is a considerable and practical approach to gauge land costs, and the least squares support vector regression which gives generally serious and acceptable outcomes.

According to Research by **Puneet Tiwari, Varun Singh Thakur (2020) [13]** on Review on **House Price Prediction through Regression Techniques.** The wide and predictable land qualities are much of the time recorded independently from the enquiring cost and the general portrayal. In this manner, these attributes or the highlights are separately recorded in a readied coordinated manner, to such an extent that they can be easily looked at across the whole scope of forthcoming houses. However, every house has its own features highlights, like a specific view, balcony 1 or 2, stopping region, Kids Park or kind of sink the merchants can give a précis of all the significant depiction of the house. In this manner the given land highlights can be estimated by the plausible purchasers, however it is by all accounts almost difficult to make accessible a mechanized assessment on all highlights or factors because of the gigantic assortment. This is also obvious the recent way: house merchants need to define an assessment of the value dependent on its qualities or highlights in comparability to the current market cost of related houses using the Machine Learning or the theory work a robotized framework is to make to foresee the house cost.

According to research by **Thuraiya Mohd, Suraya Masrom, Noraini Johari. (2019) [14]** on **Machine Learning Housing Price Prediction in Petaling Jaya, Selangor, Malaysia.** This paper exhibits the use of ML algorithms in the expectation of housing selling costs on genuine dataset gathered from the Petaling Jaya region, Selangor, Malaysia. The

Machine Learning forecast of lodging selling cost in Malaysia is scant. This paper gives a concise audit of the current ML calculations for the expectation issue and presents the attributes of the gathered datasets with various gatherings of highlight determination. The discoveries show that utilizing insignificant highlights from the dataset can diminish the precision of the prediction models.

According to research conducted by **Anurag Sinha (2020) [15]** on topic **Utilization of Machine Learning Models in Real Estate House Price Prediction.** Machine Learning take an interest a huge part in each and every zone of innovation according to the present situation. They say that each period of our lives is encircled by the execution of new time innovations like Hospitality the executives, Railway, Transportation, and Health care, Industry and so on. ML has been utilized for some areas since past many years like picture handling, design acknowledgment, clinical finding, and prescient examination, item proposal. House costs changes each year, so it is compulsory for a design to predict house costs later on. House value expectation can help in fixing and in this way foreseeing house costs and client can assess it. Their intension is to foresee house costs utilizing a few ML methods. House cost of specific area does relies upon different components like lot size, rooms, restrooms, area, drawing room, material utilized in house , insides, stopping territory and for the most part on square feet per region. Their intension behind proposing this paper is to utilize distinctive ML procedures for anticipating the cost dependent on these measurements. The calculation utilized in this investigation is Data refining, OLS regression, Classification, Clustering, correlation matrix.

According to researcher's **Akshay Babu, Dr. Anjana S Chandran.(2019) [16]** research on **Literature Review on Real Estate Value Prediction Using Machine Learning.** The housing market is quite possibly the most serious as far as valuing and same will in general fluctuate essentially dependent on various components; gauging property cost is a

significant module in dynamic for both the purchasers and financial backers in supporting spending designation, discovering property discovering tricks and deciding appropriate arrangements thus it gets one of the excellent fields to apply the ideas of ML to upgrade and foresee the costs with high exactness. Accordingly, in this paper, they present different significant highlights to utilize while anticipating lodging costs with great exactness. They used regression models, utilizing different highlights to have lower Residual Sum of Squares error. While utilizing highlights in a regression model some component designing feature engineering is needed for better forecast. Frequently a bunch of highlights multiple regression or polynomial regression (applying a different arrangement of forces in the highlights) is utilized for improving model fit. For these models are relied upon to be powerless towards over fitting edge relapse is utilized to diminish it. In this way, it coordinates to the best utilization of regression models notwithstanding different methods to streamline the outcome.

According to researcher's **Gaikwad Purva Chandrakant, Ganjave Pratiksha Namdev, Gorade Pooja Subhash, S. S. Gore (2019) [17]** on topic **Implementation of House Price Prediction Model Using Image Processing and Machine Learning.** While buying the house, the cost of house is the primary factor which is considered by individuals. The valuing of house not just relies upon the size of the property and no. of rooms, yet in addition on the areas like vehicle office, banks, schools or universities, shops and so forth. At the point when an individual purchases a home, they think about underlying highlights, working availability, neighbourhood administrations. Subsequently, a house price forecast framework is concocted to improve assessment of house costs. This framework presents a House Price Prediction utilizing Image Processing and Machine Learning. In this framework would give correlation of costs of house at specific area for clients. It would likewise give relative evaluating rates to manufacturer so he can gauges his development spending plan to rival different manufacturers at that region. The satellite pictures have been utilized to

imagine impression of neighbourhood. The Image Processing will have applied to satellite pictures and Machine Learning Algorithm Convolutional Neural Network (CNN) and Linear Regression is utilized for assessment of house evaluating. The venture is reason to anticipate cost of houses at specific region to individuals and builders.

According to research by **Zhou, Yichen (2020) [18]** on **Housing Sale Price Prediction Using Machine Learning Algorithms.** In this postulation, he investigate how prescient displaying can be applied in lodging deal value expectation by examining the lodging dataset and use ML models. As a matter of fact, he attempt four distinct models, to be specific, linear regression, lasso regression, random forest and xgboost. Furthermore, as the information have 79 illustrative factors with many missing qualities, he invest a lot of energy managing the information. He done data analysis, feature engineering before model fitting. And afterward utilizing rmse and R-squared to gauge the model exhibition. After he attempt four unique models, he get a few outcomes. With respect to the primary model - linear regression, it doesn't meet the suspicion of uniformity of the differences. Along these lines it can't utilize the linear model as the applicant of our last model. At that point he attempt lasso regression, however the RMSE and R-squared looks not very great. At that point he attempt Random Forest. The R squared in this model of preparing set is excellent, yet in the test set the R squared is moderately low, which may show the RF model is a tad overfitting. At long last he attempt the fourth model - xgboost. The entirety of the aftereffects of this xgboost model appear to be excellent. Thusly, he will utilize this xgboost model as his last model to foresee the lodging cost. The xgboost model additionally shows which factors effectively affect deal cost.

This research is conducted by **Thuraiya Mohd, Syafiqah Jamil, Suraya Masrom (2020) [19]** on **Machine learning building price prediction with green building determinant.** In the period of Industrial 4.0, numerous pressing issues in the ventures can be

viably addressed with man-made reasoning methods, including ML. Planning a compelling ML model for expectation and characterization issues is a continuous undertaking. Other than that, time and skill are significant variables that are expected to tailor the model to a particular issue, for example, the green structure lodging issue. Green structure is known as a possible way to deal with increment the productivity of the structure. To the most amazing aspect our insight, there is still no execution of ML model on GB valuation factors for building value forecast contrasted with ordinary structure improvement. This paper gives a report of an observational investigation that model structure value expectation dependent on green structure and other regular determinants. This investigations utilized five regular ML algorithms specifically Linear Regression, Decision Tree, Random Forest, Ridge and Lasso tried on a bunch of genuine structure datasets that covered Kuala Lumpur District, Malaysia. The outcome showed that the Random Forest calculation beats the other four calculations on the tried dataset and the green structure determinant has contributed some encouraging impacts to the model.

According to researcher's by **Mr. Rushikesh Naikare, Mr. Girish Gahandule, Mr. Akash Dumbre, Mr. Kaushal Agrawal (2019) [20]** on **House Planning and Price Prediction System using Machine Learning.** The lodging area has climb as it is the one of the fundamental need. Lodging the fundamental space of land. In the significant metropolitan urban communities and the urban areas with numerous renowned Educational organizations and IT Parks have sensible cost increment in lodging. Home purchasing plans can crashes the family's monetary arranging and different objectives. Presently a day's home cost changing quickly as per different boundaries. The purchaser gets befuddled in picking his fantasy home as contrast in value making it trying. Both the purchaser and vender ought to fulfil so they don't overestimate or disparage cost. So to fabricate the stage where purchaser can discover home as per its requirements and amicable to its monetary condition. House value forecast on

various boundaries is our objective. Doing that it will utilize regression calculations utilizing ML on dataset so it can separate highlights from dataset. Aftereffect of this methodology give greatest effectiveness and least mistakes. They additionally propose to decide the plane for house building.

According to researcher's **Arshiya Shaikh, R.Vinayaki, G. Siddhanth, Y. Phanindra Varma (2020) [21]** on topic **House Price Prediction using Multi Variate Analysis.** Land is the most un-straightforward industry in our biological system. Lodging costs continue to change all day every day and once in a while are advertised instead of being founded on valuation. Anticipating lodging costs with genuine components is the primary core of their exploration project. Here they intend to make their assessments dependent on each essential boundary that is thought of while deciding the cost. In this paper, they played out numerous multiple regression for assessing house cost dependent on region in square feet and number of bed rooms. Regression is a proportion of the connection between the mean estimation of one variable and relating estimations of different factors. In measurable demonstrating, regression analysis is a bunch of factual cycles for assessing the connections among factors. The Multiple linear regression clarifies the connection between one constant ward variable (y) and at least two free factors (x1, x2, x3...etc). Here they actualized through three modules: Data passage module, is utilized to give the required information to the venture. The Analysis module is utilized to examine and anticipate the house costs, in light of the client needs. The Front-end module is utilized to make the required GUI evaluates for the undertaking.

According to researcher's **Parth Ambalkar, Akash Mane, Tanmay Maity (2019) [22]** on research **House price prediction using various machine learning algorithms.** House cost increments continuously that is the reason there is a need to make such a framework at expectation of house costs. This forecast will help designers realizing the

selling cost of a house. It will likewise help clients to think about which is the ideal chance to purchase a level. In this paper, they will foresee the selling cost of different houses. Selling costs are dictated by different factors like the area of the house, zone of the property, the swelling pace of the current year, Apartment type, month and year of which is need to know the specific cost. They are actualizing different ML algorithms for building a prescient model for houses. They have thought about lodging information of 2000 properties. In this paper, they will look at the calculations based on boundaries like MAE, RMSE, MSE, accuracy.

According to research by **Bindu Sivasankar, Arun P. Ashok, Gouri Madhu, Fousiya S (2020) [23]** on **House Price Prediction.** ML assumes a significant part from past years in picture identification, Spam acknowledgment, ordinary discourse order, item suggestion and clinical conclusion along it gives better client care and more secure car frameworks. This shows that ML is pattern in practically all fields so they attempt to begat up ML in their task for improvement. These days, individuals hoping to purchase another home will in general be more moderate with their financial plans and market procedures. The current frameworks principle weakness is that the figuring of house costs are managed without the essential expectation about future market patterns and cost increment. The objective of this task is to foresee the effective house valuing for land clients concerning their spending plans and needs. In the current paper they examine about the forecast of future lodging costs that is created by ML algorithms. To choose the expectation techniques they look at and investigate different forecast strategies. To foresee the future value, the past market patterns, value ranges and furthermore impending advancement will be dissected. Consistently House costs increment, so there is a requirement for a framework to foresee house costs later on. They make a lodging cost expectation model considering Machine Learning calculation models like Lasso Regression, Ridge Regression, Ada-Boost Regression, XGBoost Regression, Decision Tree Regression, and Random Forest Regression.

House value expectation on an informational collection has been finished by utilizing all the previously mentioned strategies to discover the best among them. The designer and client will be profited by this model on deciding the selling cost of a house and causes the last to organize the correct chance to buy a house.

The research is conducted by **Mr. S. Vijayakumar, Mr. B. Ramkumar, Mr. S. Ranjith, Mr. M. Seenivasan, Mr. G. Siva (2020) [24]** on **topic House Price Prediction Based on Some Economic Factors Using Machine Learning.** House costs expanded step by step, it should have a framework to anticipate the future lodging costs. Expectation of house costs, designers can assist with deciding the selling cost of lodging, the client will assist you with organizing a period for the option to purchase a house. The offer of these houses, for the individuals who purchased, the forecast of the great house costs, better before they make quite possibly the main monetary choices in their lives, to expect what the set them up for. States of being, ideas, and where there are three factors that influence the cost of the house to incorporate. Lodging costs is a significant impression of the economy, the extent of lodging costs is a significant worry for both the purchaser and the merchant. As constant lodging costs, this lasso, ridge, SVM regression, and like a random forest regression, will be anticipated utilizing an assortment of regression method. As individual value range, they are guileless naive Bayes, logistic regression, is expected in the classification method, including SVM classification, and the random forest classification. Likewise, do the PCA to improve the forecast exactness. The objective of this venture is to make a can be assessed precisely regression model and classification model.

According to research by **Hujia Yu, Jiafu Wu (2016) [25]** on **Real Estate Price Prediction with Regression and Classification.** Housing costs are a significant impression of the economy, and housing value ranges are of incredible interest for the two purchasers and venders. In this venture house costs will be anticipated given logical factors that cover

numerous parts of private houses. As consistent house costs, they will be anticipated with different regression strategies including Lasso, Ridge, SVM regression, and Random Forest regression; as individual value ranges, they will be anticipated with classification methods including Naive Bayes, logistic regression, SVM classification, and Random Forest classification. They will likewise perform PCA to improve the forecast exactness. The objective of this undertaking is to make a regression model and classification model that can precisely appraise the cost of the house given the highlights.

According to research by **Jingyi Mu, Fang Wu and Aihua Zhang (2014) [26]** on topic **Housing Value Forecasting Based on Machine Learning Methods.** In the time of huge information, numerous pressing issues to handle on the whole different backgrounds all can be settled by means of enormous information procedure. Contrasted and the Internet, economy, industry, and aviation handle, the use of enormous information nearby design is generally not many. In this paper, based on the real information, the estimations of Boston suburb houses are figure by a few ML strategies. As indicated by the expectations, the public authority and designers can settle on choices about if building up the land on comparing locales. In this paper, Support vector machine (SVM), least squares support vector machine (LSSVM), and partial least squares (PLS) techniques are utilized to conjecture the home estimations. Also, these calculations are contrasted agreeing with the anticipated outcomes. Examination shows that albeit the informational collection exists genuine nonlinearity, the trial result additionally show SVM and LSSVM strategies are better than PLS on managing the issue of nonlinearity. The worldwide ideal arrangement can be found and best anticipating impact can be accomplished by SVM in view of tackling a quadratic programming issue. In this paper, the distinctive calculation efficiencies of the calculations are contrasted concurring with the registering seasons of important calculations.

According to researcher's **Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, Rick Siow Mong Goh (2017) [27]** on research of **A Hybrid Regression Technique for House Prices Prediction**. Normally, House value record addresses the summed up value changes of private lodging. While at a solitary family house cost forecast, it needs more precise technique dependent on the spot, house type, size, construct year, nearby conveniences, and some different variables which could influence house interest and supply. With restricted dataset and information includes, a viable and composite information pre-handling, inventive component designing technique is analyzed in this paper. The paper additionally proposes a hybrid Lasso and Gradient boosting regression model to anticipate singular house cost. The proposed approach has as of late been conveyed as the vital bit for Kaggle Challenge "House Prices: Advanced Regression Techniques". The presentation is promising as their most recent score was positioned top 1% out of all opposition groups and people.

According to research **by Nihar Bhagat, Ankit Mohokar, Shreyash Mane (2016) [28]** on **House Price Forecasting using Data Mining.** Individuals hoping to purchase another home will in general be more moderate with their spending plans and market techniques. The current framework includes figuring of house costs without the fundamental expectation about future market patterns and cost increment. The objective of the paper is to anticipate the proficient house evaluating for land clients concerning their financial plans and needs. By examining past market patterns and value ranges, and furthermore impending improvements future costs will be anticipated. The working of this paper includes a site which acknowledges client's details and afterward consolidates the utilization of different multiple linear regression calculation of data mining. This application will assist clients with putting resources into a bequest without moving toward a specialist. It additionally diminishes the danger engaged with the exchange.

According to Researcher's **Prabha D, Anindhitha A, Archana A, Balaji Narasimhan M.V.L. (2020) [29]** on topic **Predicting House Price Values Using Linear Regression with Ridge Regularization Approach.** The valuation of land is the focal precept of all business. The term valuation is characterized as the logical interaction of compelling the current worth of declare or an organization. In any case, there is wide scope of direction for which valuations are required. Be that as it may, here valuations are accomplished for powerful approach to ascertain the selling cost of an element. To build up a land valuation model which predicts the estimation of a property utilizing the space of Machine Learning. The algorithmic methodology includes ridge regression on top of linear regression approach (Supervised Learning). The selling cost is gauges utilizing by considering different boundaries such as population rate in particular area, distance to roadways, property age etc. The dataset assortment is taken from a standard source to such an extent that 80 boundaries alongside 1000's of test and preparing information are considered for property valuation and separate dataset is considered for testing and preparing a model. For additional improvement of precision, Ridge regularization is applied on top of linear regression so information are regularized with increment in model exactness. Clients who will sell the property can get the exact qualities dependent on this regression prediction. Clients requires no transitional individual (intermediary) to sell in the substance. The python language with its standard libraries are used for model assumptions reliant on dataset regard. Since end-client can't run this model every single time by using python inactive there comes the convenience lab. To beat this just as for amazing use of this model by end-clients a different site page is organized with the objective that customers can honestly pass regards from site to python code and get the specific incentive for the element.

According to researcher's **Bruno Klaus de Aquino Afonso, Luckeciano Carvalho Melo, Willian Dihanster Gomes de Oliveira, Samuel Bruno da Silva Sousa, Lilian**

Berton (2019) [30] on **topic Housing Prices Prediction with a Deep Learning and Random Forest Ensemble.** The advancement of a housing costs forecast model can help a house merchant or a realtor to settle on better-educated choices dependent on house value valuation. A couple of works report the utilization of (ML) algorithms to anticipate the estimations of properties in Brazil. This examination investigates a dataset made out of 12,223,582 lodging notices, gathered from Brazilian sites from 2015 to 2018. Each example includes 24 highlights of five distinctive information types: number, date, string, float, and image. To anticipate the property costs, we troupe two diverse ML designs, in view of Random Forest (RF) and Recurrent Neural Networks (RNN). This examination shows that enhancing the dataset and joining diverse ML approaches can be a superior option at expectation of lodging costs in Brazil.

CHAPTER 3
REQUIREMENTS AND ANALYSIS

3.1 Problem Definition:

Before going in the methodology understanding of the problem is much important. The problem is creating the hypothesis function that may give the prediction of the target value based on the data given as the training part. Then see or analyze the prediction on the testing part of the data. Here the data given is on the house price and its respective features which accommodate the price of the house. Thus to build the machine to learn the data features and predict the price accurate is the challenging task. This will also help the society of the real estate builder to easily predict the price of the land, house, etc. according to their feature with the help of this model. The dataset used in this project comes from the UCI Machine Learning Repository. This data was collected in 1978 and each of the 506 entries represents aggregate information about 14 features of homes from various suburbs located in Boston.

The problem that we are going to solve here is that given a set of features that describe a house in Boston, our machine learning model must predict the house price. To train our machine learning model with Boston housing data, we will be using scikit-learn's Boston dataset.

- The features can be summarized as follows:

- CRIM: This is the per capita crime rate by town.
- ZN: This is the proportion of residential land zoned for lots larger than 25,000 sq.ft.
- INDUS: This is the proportion of non-retail business acres per town.

- CHAS: This is the Charles River dummy variable (this is equal to 1 if tract bounds river; 0 otherwise)
- NOX: This is the nitric oxides concentration (parts per 10 million)
- RM: This is the average number of rooms per dwelling
- AGE: This is the proportion of owner-occupied units built prior to 1940
- DIS: This is the weighted distances to five Boston employment centers
- RAD: This is the index of accessibility to radial highways
- TAX: This is the full-value property-tax rate per $10,000
- PTRATIO: This is the pupil-teacher ratio by town
- B: This is calculated as $1000(Bk — 0.63)^2$, where Bk is the proportion of people of African American descent by town
- LSTAT: This is the percentage lower status of the population
- MEDV: This is the median value of owner-occupied homes in $1000s.

The last variable is considered as the target value; here it is named as MEDV, which is the actual price of the house. The when machine will predict the price it will get matched with the actual value and the mean error will get calculated which will give the accuracy rate of the model. The data set may contain the various detail features of the houses. Now import the data set by the help of the pandas in python platform and analyze the data set. Check all the features of the house related to the dependent target. Analyze and visualize the data by checking the missing values, fill all the missing values by taking median of all the values of that attribute. Change the data which are in categorical form, place the one hot encoder, or the label encoder coding for changing the categorical data

into the numerical data. Change the entire alphabet values of the attribute into the numerical values. Find the appropriate features by the help of heat map and the correlation matrix.

Select the most nearly features to which the label target is truly dependent. Before applying the machine learning regression function to the data, split the data into two parts one is training data and another is the testing data .Apply the machine learning on the training part of the data by the help of the *sklearn* library on python platform. The fetching operation is done by the help of the pandas library function as in the format of .csv file and giving the path where data is stored.

After fetching the data, some cleaning process is applied to the data to make it provide useful information. Thus the missing values is the attribute is checked and clean it out i.e. drop attribute if it is not much useful feature or fill the missing value by taking the median of the all values. After cleaning process is done then the categorical data is get specified and applied the one hot encoder to it. Thus after applying the one hot code encoder the correlation matrix is calculated to select the appropriate features, further the whole data set is divided or split into two parts in 80% in training data and 20% in testing data. Then the further process applying machine learning is processed and three regression technique is applied to the dataset *decision tree, linear regression, random forest.*

3.2 Requirements Specifications:

Requirement analysis gives a minimum requirement that a system should have to make the software to work properly. Usually the requirement specification will be the same as that of the operating system.

PROPER FORECASTING: The system has to properly predict the price of the house according to the input given by the user.

RECOMMENDATION SYSTEM: According to the input given by the user, the recommendation system will recommend the best property.

USER INTERFACE: The user interface will be on a Jupyter lab. The user has to enter all the attributes correctly and in the required format.

PLATFORM INDEPENDENT: The application would be platform independent if all the requirements are installed in the device.

PERFORMANCE: The application should have better accuracy and should provide the information in less time.

CAPACITY: The capacity of the storage should be high so that large amount of data can be stored in order to train the model

3.3 Software and Hardware Requirements:

Hardware details: Minimum System Requirements.

Processor	AMD – E1 1500 APU with Radeon™
Graphics	HD Graphics 1.48GHz
System type	64-bit operating system, x64-based processor.
RAM	4 GB
HDD	180 GB

Table 3.1. Hardware details

Software details: Minimum System Requirements.

Operating System	Windows 8.1
Programming Language	Python
Software tools	PyCharm, Jupyter Notebook
Libraries	Numpy, Scikit-Learn, Matplotlib, Pandas

Table 3.2. Software details

3.4 Preliminary Product Description:

- **In this thesis I had used Python programming language.**
  - Python is an interpreted, high-level and general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.
  - Python is a general-purpose, and high-level programming language which is best known for its efficiency and powerful functions. Python is loved by data scientists because of its ease of use, which makes it more accessible. Python provides data scientists with an extensive amount of tools and packages to build machine learning models. One of its special features is that we can build various machine learning with less-code.
  - To implement my project I have made used of the Python IDE (Integrated Development Environment) named as Jupyter Notebook and PyCharm which is a famous IDE specialized in the field of performing Machine Learning tasks.

❖ **Jupyter Notebook**

- Jupyter is a free, open-source, interactive web tool known as a computational notebook, which researchers can use to combine software code, computational output, explanatory text and multimedia resources in a single document. Computational notebooks have been around for decades, but Jupyter in particular has exploded in popularity over the past couple of years.
- This rapid uptake has been aided by an enthusiastic community of user–developers and a redesigned architecture that allows the notebook to speak dozens of programming languages
- **Jupyter Notebook** provides you with an easy-to-use, interactive data science environment across many programming languages that doesn't only work as an **IDE**, but also as a presentation or education tool.
- Jupyter notebook is an open-source IDE that is used to create **Jupyter documents** that can be created and shared with live codes.

3.5 Conceptual Models:

3.5.1 Flowchart:

- A flowchart is a graphical representations of steps. It was originated from computer science as a tool for representing algorithms and programming logic but had extended to use in all other kinds of processes.
- It shows steps in sequential order and is widely used in presenting the flow of algorithms, workflow or processes. Typically, a flowchart shows the steps as boxes of various kinds, and their order by connecting them with arrows.

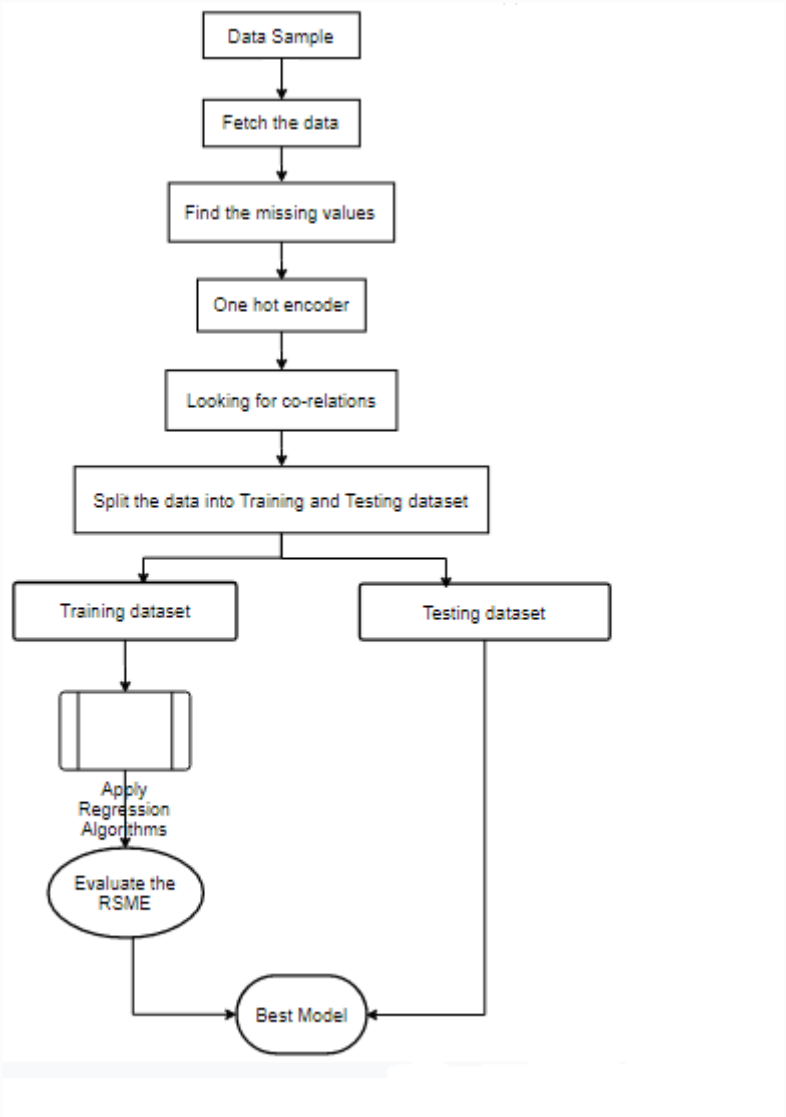


Figure 3.1. Flowchart

3.5.2 Sequence Diagram:

- A sequence diagram simply depicts interaction between objects in a sequential order i.e. the order in which these interactions take place. We can also use the terms event diagrams or event scenarios to refer to a sequence diagram. Sequence diagrams describe how and in what order the objects in a system function.
- Model high-level interaction between active objects in a system.
- Model the interaction between object instances within a collaboration that realizes a use case.
- Model the interaction between objects within a collaboration that realizes an operation.
- Either model generic interactions (showing all possible paths through the interaction) or specific instances of interaction (showing just one path through the interaction).

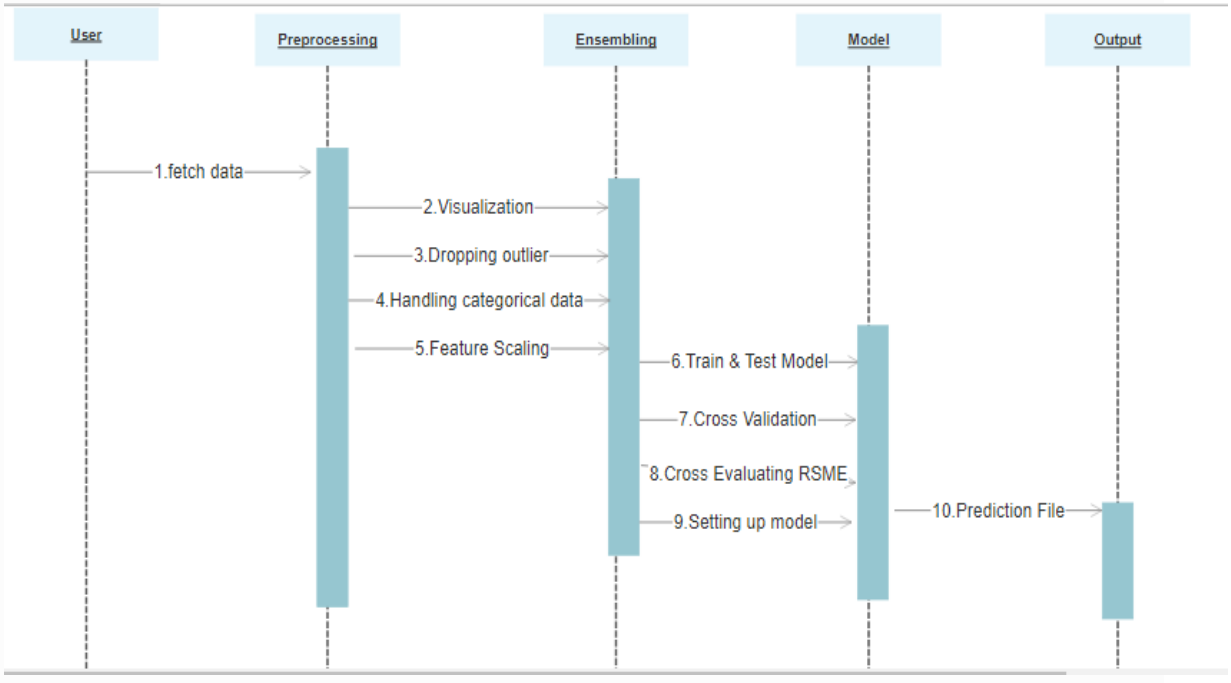


Figure 3.2. Sequence diagram

3.5.3 Activity Diagram:

- An activity diagram is a behavioral diagram i.e. it depicts the behavior of a system.
- An activity diagram portrays the control flow from a start point to a finish point showing the various decision paths that exist while the activity is being executed.
- An activity diagram visually presents a series of actions or flow of control in a system similar to a flowchart or a data flow diagram.
- Activity diagrams are often used in business process modeling. They can also describe the steps in a use case diagram. Activities modeled can be sequential and concurrent.
- Identify pre- and post-conditions (the context) for use cases.
- Model workflows between/within use cases.
- Model complex workflows in operations on objects.
- Model in detail complex activities in a high level activity Diagram.

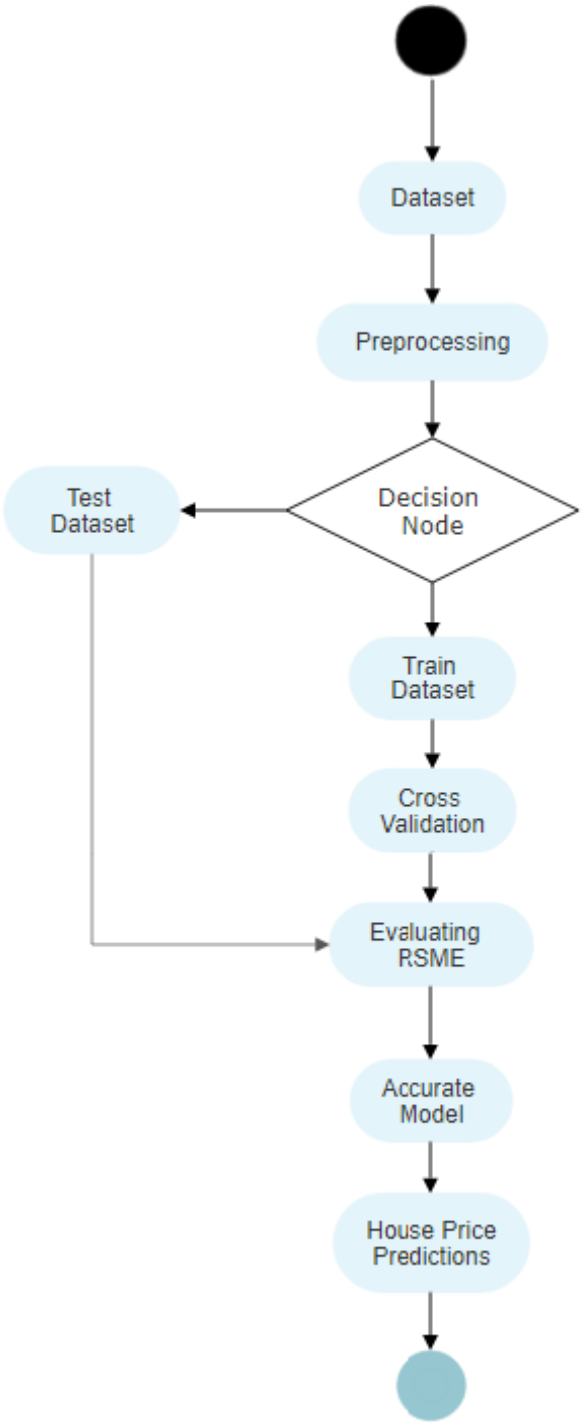


Figure 3.3. Activity diagram

3.5.4 Class Diagram:

- The class diagram is the main building block of object-oriented modeling. It is used for general conceptual modeling of the structure of the application, and for detailed modeling translating the models into programming code.
- Class diagrams can also be used for data modeling.
- Shows static structure of classifiers in a system.
- Diagram provides a basic notation for other structure diagrams prescribed by UML.
- Helpful for developers and other team members too.
- Business Analysts can use class diagrams to model systems from a business perspective.
- A UML class diagram is made up of:
  - A set of classes and
  - A set of relationships between classes.

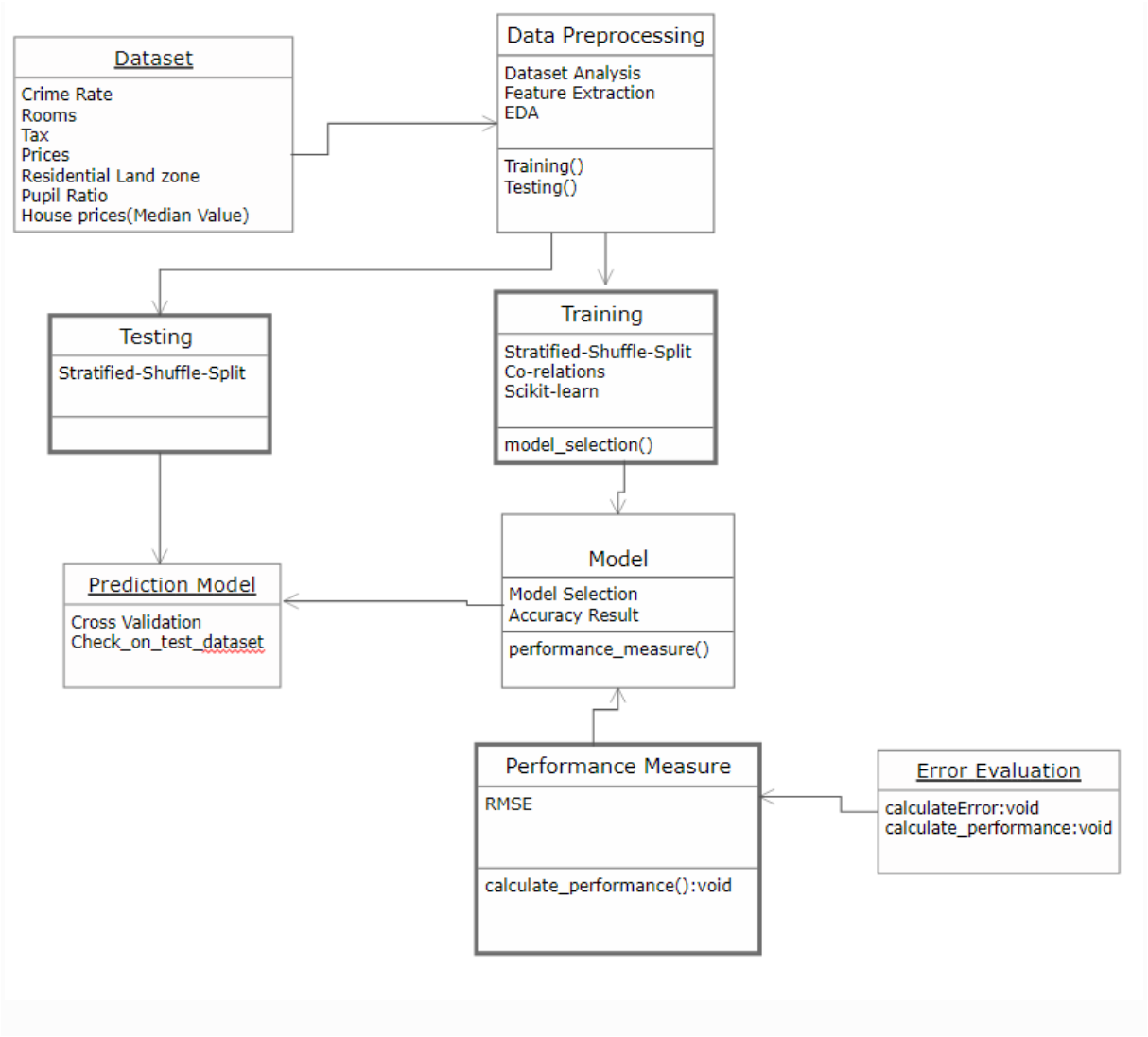


Figure 3.4. Class diagram

CHAPTER 4
PROPOSED DESIGN

4.1 BASIC MODULES:

1- Understanding the Business problem

This means before starting the project we must define the problems thoroughly to which our project is going to provide solutions. Having Clarity about our business problem will help us to move in the proper direction.

Methodology differs from problem to problem also certain Machine learning algorithm works for certain business problem so identification of the business problem at the earliest is a must

2- Get the Data

To train the machine learning model we will require a dataset to work on i.e. to gain insights. As a society, we're generating data at an unprecedented rate. These data can be numeric (temperature, loan amount, customer retention rate), categorical (gender, color, highest degree earned), or even free text (think doctor's notes or opinion surveys). Data collection is the process of gathering and measuring information from countless different sources. In order to use the data we collect to develop practical machine learning solutions, it must be collected and stored in a way that makes sense for the business problem at hand.

3- Data Cleaning

Data cleaning is a critically important step in any machine learning project. In tabular data, there are many different statistical analysis and data visualization techniques you can use to explore your data in order to identify data cleaning operations you may want to perform. Before jumping to the sophisticated methods, there are some very basic data cleaning

operations that you probably should perform on every single machine learning project. These are so basic that they are often overlooked by seasoned machine learning practitioners, yet are so critical that if skipped, models may break or report overly optimistic performance results.

- Identify and remove column variables that only have a single value.
- Identify and consider column variables with very few unique values.
- Identify and remove rows that contain duplicate observations.

4- Discover and visualize the data to gain insights

Data insights refer to the understanding of a particular business phenomenon you are able to achieve by using machine learning and artificial intelligence technology to analyse a dataset. For example, a machine learning model that estimates the likelihood of a customer to churn will reveal what factors drive churn rates, allowing decision-makers to make changes to business strategies and processes. One of the best ways to understand and communicate meaningful insights from data is to use tools that help visualize a model's outcomes and give different ways to explore and understand your data. This translates to real business value in the form of increased ROI on advertising efforts, more accurate loan default predictions, and much more. The clarity of vision from data insights allows users to make better decisions based on increased model interpretability, allowing analysts and other users to explain model outcomes to key stakeholders. Data visualization tools help users understand and explain insights from machine learning model outcomes. Whether it is through simple graphical representations like word clouds or more complicated and flexible data visualization tools like Tableau dashboards, these tools make it easier to understand and communicate the value uncovered by the model and drive better business decision-making.

5- Prepare the data for Machine Learning Algorithms

Data preparation (also referred to as "data pre-processing") is the process of transforming raw data so that data scientists and analysts can run it through Machine learning algorithms to uncover insights or make predictions

The data preparation process can be complicated by issues such as:

1. Missing or incomplete records. It is difficult to get every data point for every record in a dataset. Missing data sometimes appears as empty cells, values (e.g., NULL or N/A), or a particular character, such as a question mark.

2. Outliers or anomalies. Unexpected values often surface in a distribution of values, especially when working with data from unknown sources which lack poor data validation controls.

2. Improperly formatted / structured data. Data sometimes needs to be extracted into a different format or location. A good way to address this is to consult domain experts or join data from other sources.

3. Inconsistent values and non-standardized categorical variables. Often when combining data from multiple sources, we can end up with variations in variables like company names or states. For instance, a state in one system could be "Texas," while in another it could be "TX." Finding all variations and correctly standardizing will greatly improve the model accuracy.

4. Limited or sparse features / attributes. Feature enrichment, or building out the features in our data often requires us to combine datasets from diverse sources. Joining files from different systems is often hampered when there are no easy or exact columns to match the datasets. This then requires the ability to perform fuzzy matching, which could also be based

on combining multiple columns to achieve the match. For instance, combining two datasets on CUSTOMER ID (present in both data datasets) could be easy. Combining a dataset that has separate columns for CUSTOMER FIRST NAME and CUSTOMER LAST NAME with another dataset with a column CUSTOMER FULL NAME, containing "Last name, first name" becomes trickier.

5. The need for techniques such as features engineering. Even if all of the relevant data is available, the data preparation process may require techniques such as feature engineering to generate additional content that will result in more accurate, relevant models.

6- Select a Model and train it

Training a model simply means learning (determining) good values for all the weights and the bias from labelled examples. In supervised learning, a machine learning algorithm builds a model by examining many examples and attempting to find a model that minimizes loss; this process is called empirical risk minimization.

Statistical and mathematical models have multiple purposes, ranging from descriptive to predictive to prescriptive analytics. The goal of developing models in machine learning is to extract insights from data that you can then use to make better business decisions. Algorithmic models tell you which outcome is likely to hold true for your target variable based on your training data. They construct a representation of the relationships and tease out patterns between all the different features in your dataset that you can apply to similar data you collect in the future, allowing you to make decisions based on those patterns and relationships. It is more abstract than an architectural model, but it is the same idea: a distilled representation of a greater picture. Models form the basis of data analysis. Without models, we would be limited to simple computation (1+2=3), without statistical models, we could not

determine relationships between variables, and without machine learning models, we would not be able to uncover relationships and gain insight from historical data.

Algorithms are step-by-step computational procedures for solving a problem, similar to decision-making flowcharts, which are used for information processing, mathematical calculation, and other related operations. Machine learning relies on algorithms to build models that reveal patterns in data, which in turn allow businesses to uncover insights and make predictions to improve operations, better understand customers, and solve other business problems. There are many different algorithms, but most data scientists rely on a small set with which they are familiar.

❖ Machine Learning Libraries in Python:

- ***NumPy*** is a very popular python library for large multi-dimensional array and matrix processing, with the help of a large collection of high-level mathematical functions. It is very useful for fundamental scientific computations in Machine Learning. It is particularly useful for linear algebra, Fourier transform, and random number capabilities. High-end libraries like TensorFlow uses NumPy internally for manipulation of Tensors.
- ***Skikit-learn*** is one of the most popular ML libraries for classical ML algorithms. It is built on top of two basic Python libraries, viz., NumPy and SciPy. Scikit-learn supports most of the supervised and unsupervised learning algorithms. Scikit-learn can also be used for data-mining and data-analysis, which makes it a great tool who is starting out with ML.
- ***Pandas*** is a popular Python library for data analysis. It is not directly related to Machine Learning. As we know that the dataset must be prepared before training. In this case, Pandas comes handy as it was developed specifically for data extraction and preparation. It provides high-level data structures and wide

variety tools for data analysis. It provides many inbuilt methods for groping, combining and filtering data.

- ***Matpoltlib*** is a very popular Python library for data visualization. Like Pandas, it is not directly related to Machine Learning. It particularly comes in handy when a programmer wants to visualize the patterns in the data. It is a 2D plotting library used for creating 2D graphs and plots. A module named pyplot makes it easy for programmers for plotting as it provides features to control line styles, font properties, formatting axes, etc. It provides various kinds of graphs and plots for data visualization, viz., histogram, error charts, bar chats, etc,

4.2 PROCEDURAL DESIGN:

4.2.1 LOGIC DIAGRAMS:

1. ARCHITECTURE DIAGRAM

- An architectural diagram is a diagram of a system that is used to abstract the overall outline of the software system and the relationships, constraints, and boundaries between components.
- It is an important tool as it provides an overall view of the physical deployment of the software system and its evolution roadmap.

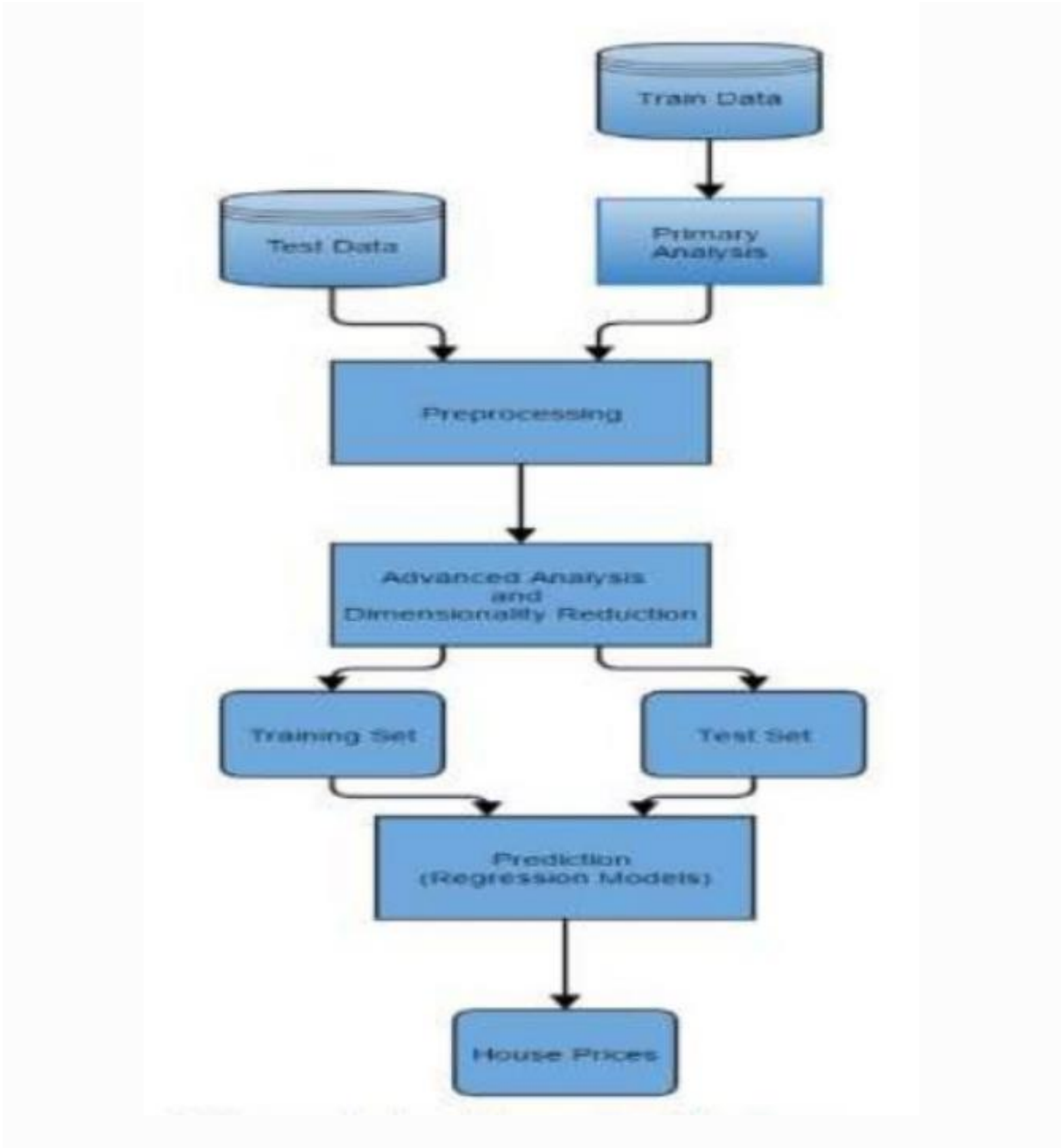


Figure 4.1. Architecture diagram

2. CONTROL FLOW CHART

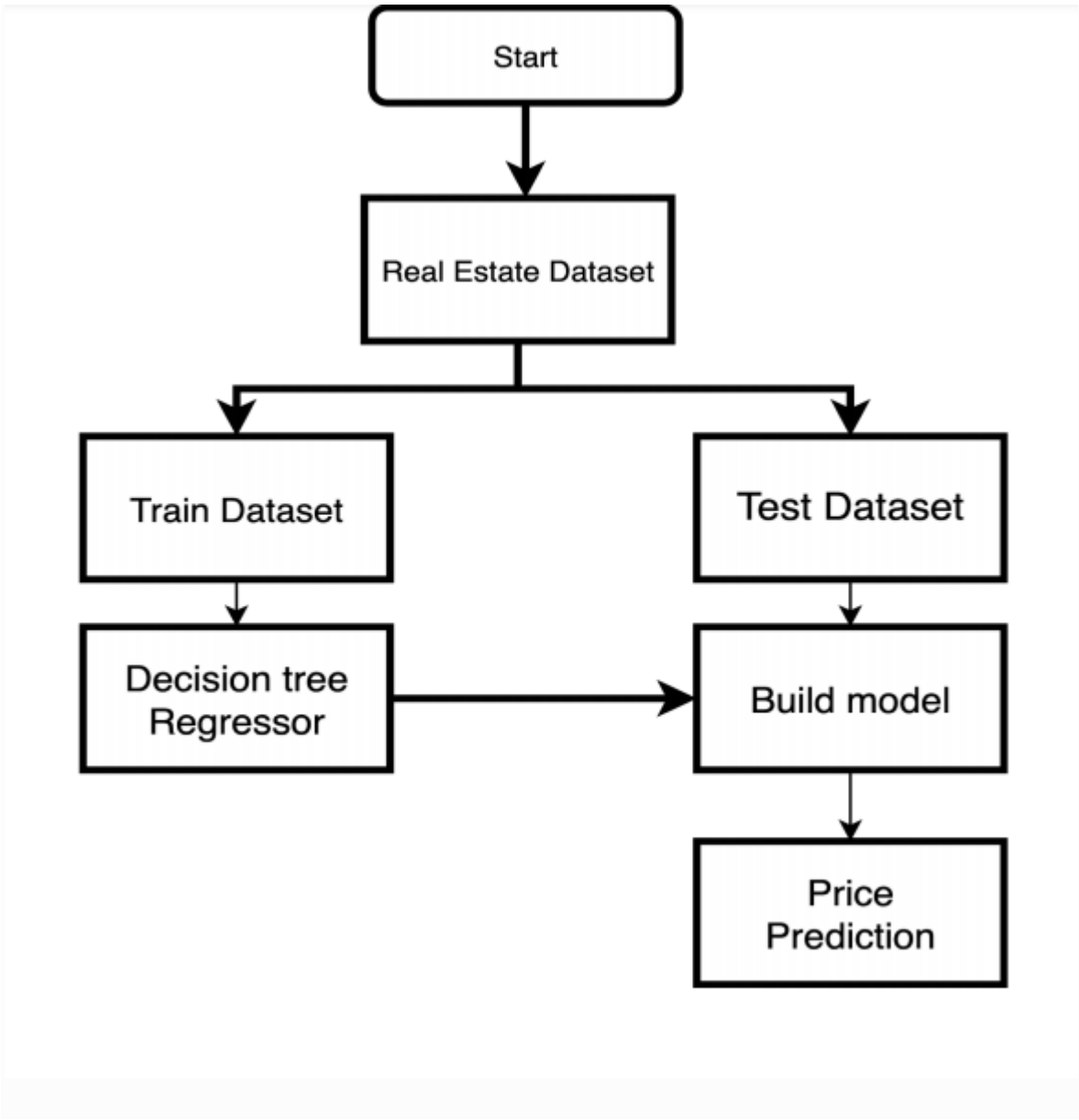


Figure 4.2. Control flowchart

3. PSEUDO CODE:

Step 1: Fetch the data set in appropriate format.

Step 2: Find the missing values in the data set as the cleaning process is done.

Step 3: Apply the one hot encoder to remove the categorical data.

Step 4: Selection of feature by heat map or matrix correlation.

Step 5: Splitting of the dataset into two part training data and the testing data.

Step 6: Apply or fit the regression technique on training data and test it with testing data.

Step 7: Compare the accuracy result.

4.2.2 ALGORITHMS DESIGN:

❖ Machine Learning Model

➢ *Linear Regression*

Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It's used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into categories (e.g. cat, dog). There are two main types:

Simple regression

Simple linear regression uses traditional slope-intercept form, where mm and bb are the variables our algorithm will try to "learn" to produce the most accurate predictions. x represents our input data and y represents our prediction.

$$y = mx + b$$

Multivariable regression

A more complex, multi-variable linear equation might look like this, where w represents the coefficients, or weights, our model will try to learn.

$$f(x, y, z) = w1x + w2y + w3z$$

Cost function-

- The different values for weights or coefficient of lines (a_0, a_1) gives the different line of regression, and the cost function is used to estimate the values of the coefficient for the best fit line.
- Cost function optimizes the regression coefficients or weights. It measures how a linear regression model is performing.

➢ ***Decision Tree***

- Decision Tree is a supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.
- In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
- The decisions or the test are performed on the basis of features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. It is called a decision tree because,

similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

- The complete process can be better understood using the below algorithm:

Step-1: Begin the tree with the root node, says S, which contains the complete dataset.

Step-2: Find the best attribute in the dataset using Attribute Selection Measure (ASM).

Step-3: Divide the S into subsets that contains possible values for the best attributes.

Step-4: Generate the decision tree node, which contains the best attribute.

Step-5: Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

- ***Random Forest***

- Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.
- As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.
- *The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.*

- Below are some points that explain why we should use the Random Forest algorithm:

1. It takes less training time as compared to other algorithms.
2. It predicts output with high accuracy, even for the large dataset it runs efficiently.
3. It can also maintain accuracy when a large proportion of data is missing.

❖ The Working process can be explained in the below steps:

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

❖ Criteria to measure performance

➢ ***RMSE -root mean square error***

- The root-mean-square error (RMSE) represents the square root of the second sample moment of the differences between predicted values and observed values or the quadratic mean of these differences. RMSE is the square root of the average of squared errors. The effect of each error on RMSE is proportional to the size of the squared error; thus, larger errors have a disproportionately large effect on RMSE. Consequently, RMSE is sensitive to outliers.
- The RMSE of an estimator θ^with respect to an estimated parameter $\theta$is defined as the square root of the mean square error:

$$RMSE\ (\theta^)=\sqrt{MSE(\theta^)}=\sqrt{E((\theta^-\theta)2)}$$

- For an unbiased estimator, the RMSD is the square root of the variance, known as the standard deviation.
- The RMSE of predicted values *yi* of a regression's dependent variable *yi*, with variables observed over ntimes, is computed for ndifferent predictions as the square root of the mean of the squares of the deviation:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

4.3 USER INTERFACE DESIGN:

- Reading the data

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2

Table 4.1. Boston house data

- By using describe method we can show count, mean, standard deviation, quantile, min, max.

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	L
count	506.000000	506.000000	506.000000	506.000000	506.000000	501.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.0
mean	3.613524	11.363636	11.136779	0.069170	0.554695	6.284341	68.574901	3.795043	9.549407	408.237154	18.455534	356.674032	12.6
std	8.601545	23.322453	6.860353	0.253994	0.115878	0.705587	28.148861	2.105710	8.707259	168.537116	2.164946	91.294864	7.1
min	0.006320	0.000000	0.460000	0.000000	0.385000	3.561000	2.900000	1.129600	1.000000	187.000000	12.600000	0.320000	1.7
25%	0.082045	0.000000	5.190000	0.000000	0.449000	5.884000	45.025000	2.100175	4.000000	279.000000	17.400000	375.377500	6.9
50%	0.256510	0.000000	9.690000	0.000000	0.538000	6.208000	77.500000	3.207450	5.000000	330.000000	19.050000	391.440000	11.3
75%	3.677082	12.500000	18.100000	0.000000	0.624000	6.625000	94.075000	5.188425	24.000000	666.000000	20.200000	396.225000	16.9
max	88.976200	100.000000	27.740000	1.000000	0.871000	8.780000	100.000000	12.126500	24.000000	711.000000	22.000000	396.900000	37.9

Table 4.2. Describe table

- By using matplotlib we can draw the histogram to identify the differences.

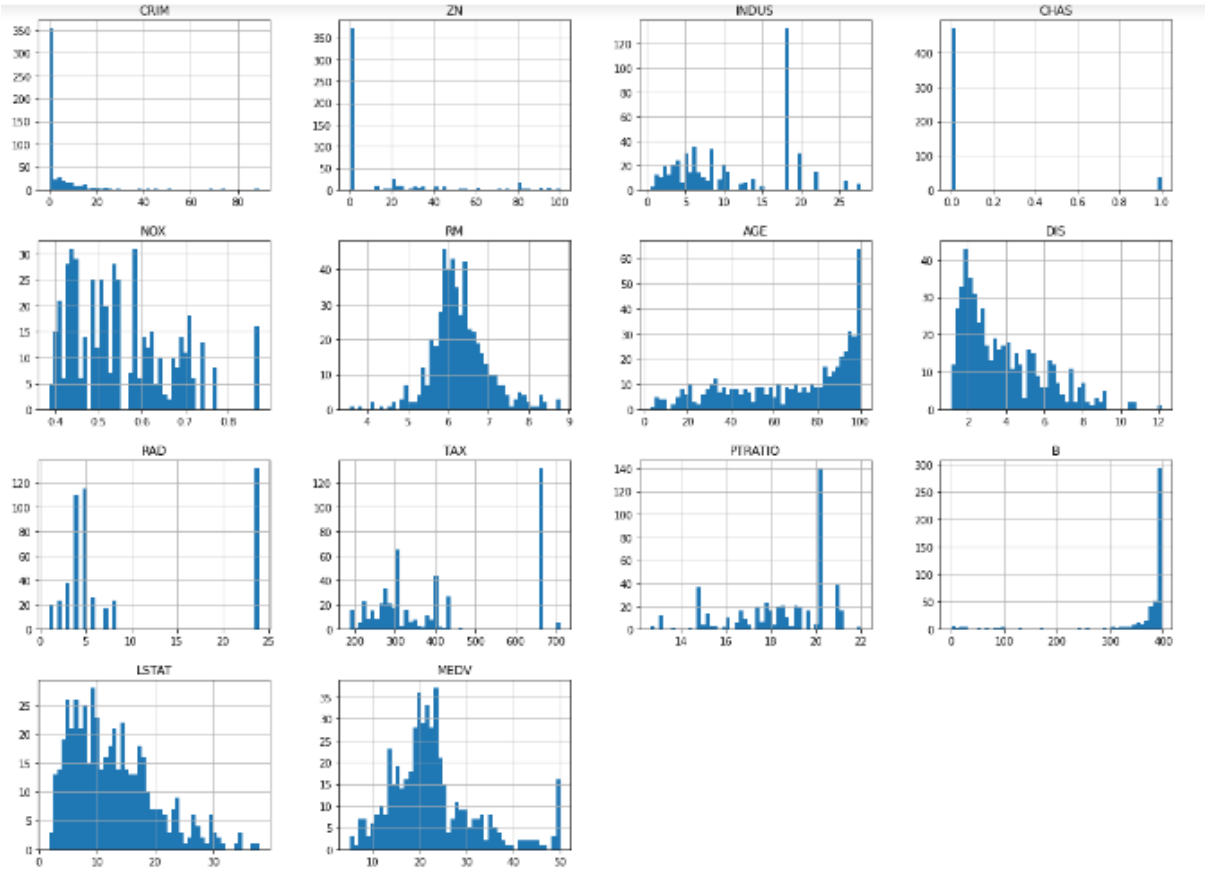


Figure 4.3. Histogram of Boston house data

- Looking for co-relations

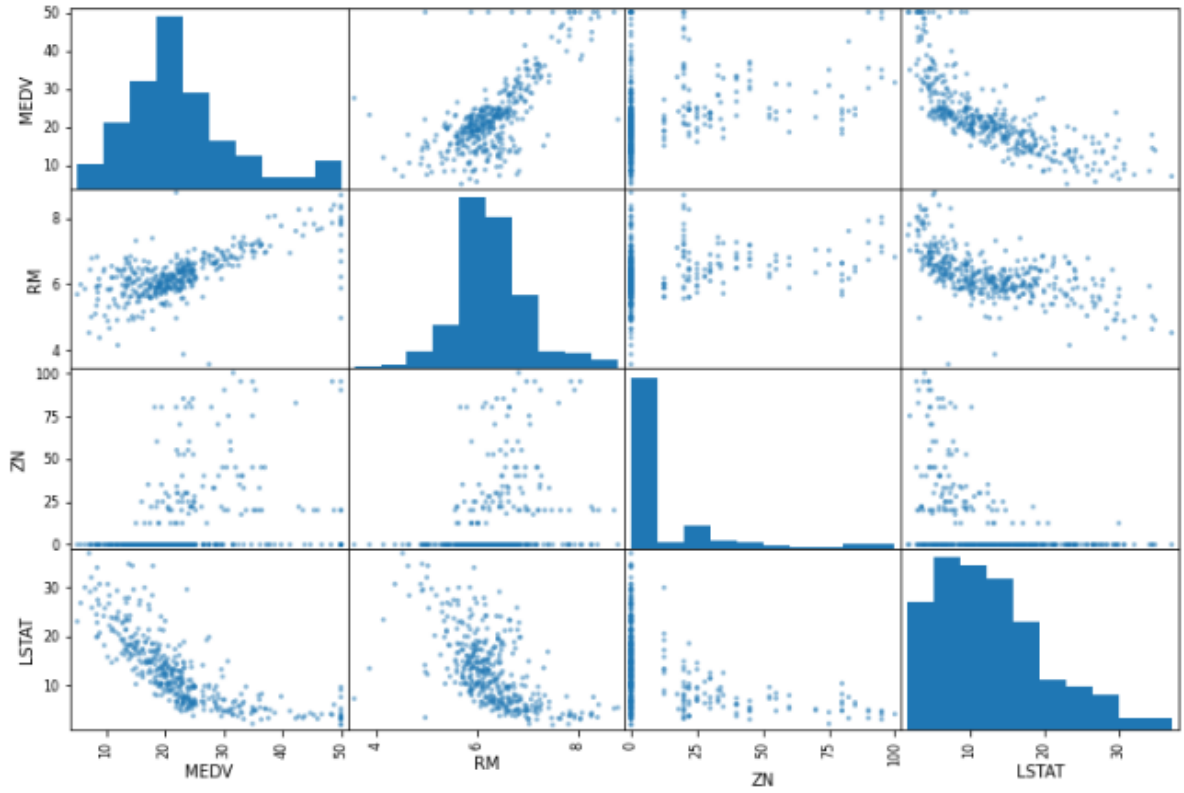


Figure 4.4. Pearson correlation

- The outliers

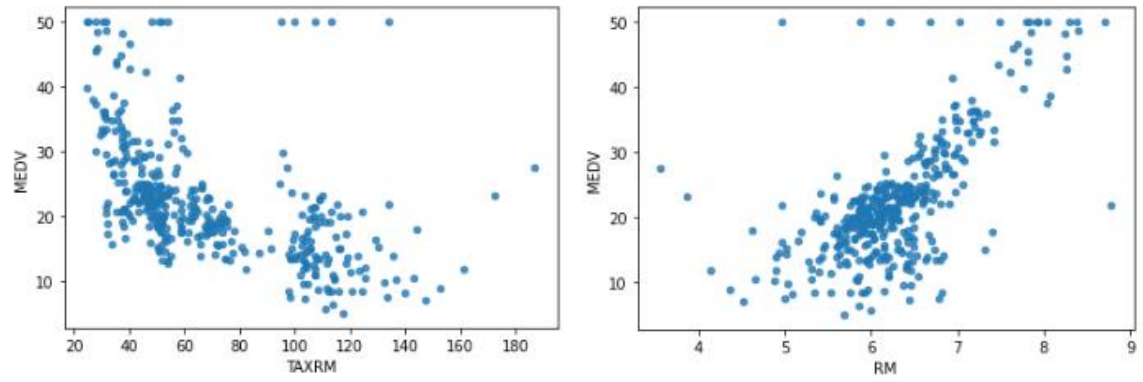


Figure 4.5. Outliers

- Entering dataset features for predicting the house price.

```
In [28]: from joblib import dump, load
         import numpy as np
         model = load('Dragon.joblib')
         features = np.array([[-5.43942006, 4.12628155, -1.6165014, -0.67288841, -1.42262747,
                -11.44443979304, -49.31238772,  7.61111401, -26.0016879 , -0.5778192 ,
                -0.97491834,  0.41164221, -66.86091034]])
         model.predict(features)
Out[28]: array([23.02])
```

Figure 4.6. Predicting Price

4.4 SECURITY ISSUES:

There are security hazards in programming advancement of any sort, including ML. These dangers should be considered during each period of the ML life cycle to make a protected ML system. Security is significant in ML frameworks frequently contain secret data or give an upper hand to the association that they would not need contenders to have the option to get to. A few organizations use ML for security motivations to identify security breaks in different frameworks, so the security of that ML model itself is significant so their security framework can be trusted to get the other framework.

❖ **Machine Learning Security Risks:**

➢ Data confidentiality

Ensuring secret data is as of now troublesome without it being important for an ML system. ML carries extra difficulties to securing secret data, since delicate data is incorporated into the model through preparing. There are successful however unpretentious attacks to separate data from a ML system are a possible danger. To shield the system from this kind of attack, it is important to incorporate security conventions into the model from the start stages in the ML lifecycle.

- Adversarial examples

Adversarial machine learning is a technique employed in the field of machine learning which attempts to fool models through malicious input. This technique can be applied for a variety of reasons, the most common being to attack or cause a malfunction in standard machine learning models.

- Data poisoning

Since data assumes a particularly colossal part in ML security, if an attacker can deliberately control the data utilized by a ML system, it can bargain the whole system. ML designers ought to consider what preparing data an attacker might actually control and how much they could handle it, to focus on forestalling data poisoning.

4.5 TEST CASE DESIGNS:

First of all, you split the data set into three non-covering sets. You utilize a training set to prepare the model. At that point, to assess the presentation of the model, you utilize two sets of information:

Validation set. Having just a training set and a testing set isn't sufficient in the event that you do numerous rounds of hyperparameter-tuning (which is consistently). Furthermore, that can bring about overfitting. To stay away from that, you can choose a little validation informational index to assess a model. Solely after you get most extreme precision on the validation set, you make the testing set come into the game.

Test set (or holdout set). Your model may fit the training dataset totally well. In any case, where are the ensures that it will do similarly well, all things considered. To guarantee that, you select examples for a testing set from your training set — models that the machine hasn't

seen previously. It is critical to stay fair-minded during choice and draw tests indiscriminately. Additionally, you ought not utilize similar set commonly to try not to prepare on your test information. Your test set ought to be sufficiently huge to give measurably significant outcomes and be illustrative of the informational collection overall.

However, similarly as test sets, validation sets "wear out" when utilized over and over. The more occasions you utilize similar information to settle on choices about hyperparameter settings or other model enhancements, the less certain you are that the model will sum up well on new, concealed information. So it is a smart thought to gather more information to 'spruce up' the test set and validation set.

Cross-validation is a model evaluation technique that can be performed even on a limited dataset. The training set is divided into small subsets, and the model is trained and validated on each of these samples.

The most common cross-validation method is called **k-fold cross-validation**. To use it, you need to divide the dataset into k subsets (also called folds) and use them k times. For example, by breaking the dataset into 10 subsets, you will perform a 10-fold cross-validation. Each subset must be used as the validation set at least once.

Evaluate models using metrics: Evaluating the performance of the model using different metrics is integral to every data science project. In this project we had used RMSE and R squared.

RMSE: is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed.

R squared: is the proportion of the variance in the dependent variable that is predictable from the independent variable.

Accuracy is a metric for how much of the predictions the model makes are true. The higher the accuracy is, the better. However, it is not the only important metric when you estimate the performance.

Loss describes the percentage of bad predictions. If the model's prediction is perfect, the loss is zero; otherwise, the loss is greater.

Recall: This metric measures the number of correct predictions, divided by the number of results that should have been predicted correctly. It refers to the percentage of total relevant results correctly classified by your algorithm.

Precision: The precision metric marks how often the model is correct when identifying positive results. For example, how often the model diagnoses cancer to patients who really have cancer

CHAPTER 5

IMPLEMENTATION & EXPERIMENT PERFORMED

5.1 Planning and Scheduling:

5.1.1 Gantt chart:

- A Gantt chart is a type of bar chart that illustrates a project schedule. This chart lists the tasks to be performed on the vertical axis, and time intervals on the horizontal axis. The width of the horizontal bars in the graph shows the duration of each activity. Gantt charts illustrate the start and finish dates of the terminal elements and summary elements of a project. Terminal elements and summary elements constitute the work breakdown structure of the project. Modern Gantt charts also show the dependency (i.e., precedence network) relationships between activities. Gantt charts can be used to show current schedule status using percent-complete shadings and a vertical "TODAY" line.
- Gantt charts are sometimes equated with bar charts.
- Gantt charts are usually created initially using an early start time approach, where each task is scheduled to start immediately when its prerequisites are complete. This method maximizes the float time available for all tasks.

	A	B	C	D	E
1	Task	Start	Finish	Duration	Weeks
2	Feasibility Study	01-10-2020	02-11-2020	32	5
3	Requirement Gather	03-11-2020	13-12-2020	40	6
4	Coding	14-12-2020	20-02-2021	68	10
5	Implementation	20-02-2021	20-03-2021	28	4
6	Testing	21-03-2021	21-04-2021	31	4
7					

Table 5.1. Gantt Table

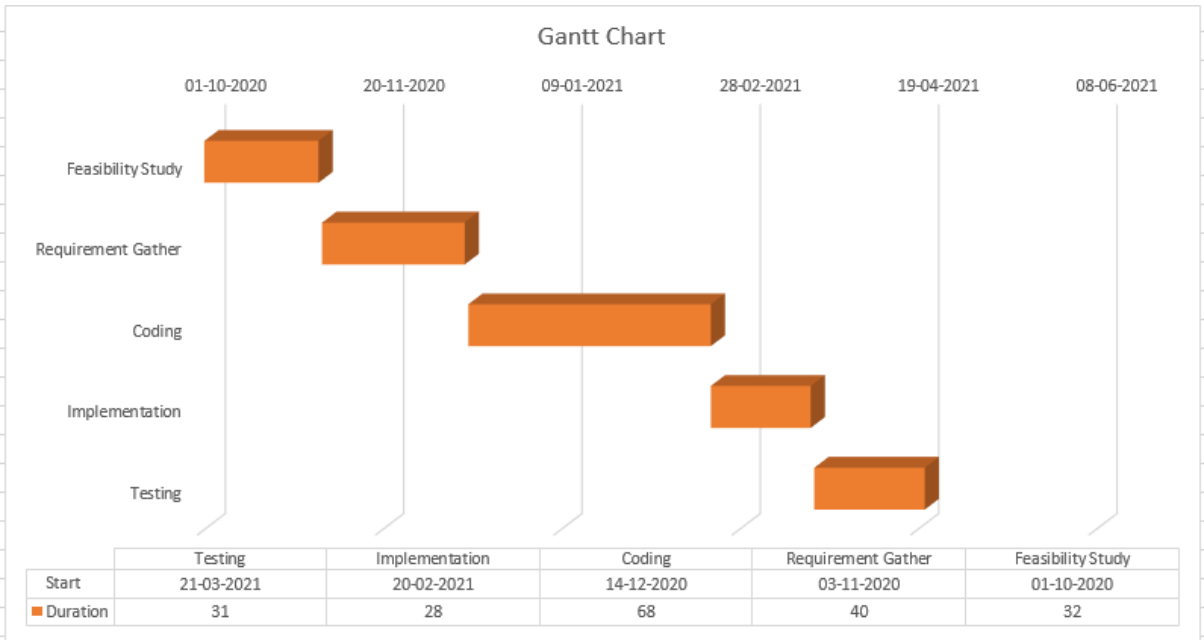


Figure 5.1. Gantt chart

5.1.2 Pert chart:

A PERT chart is a project management tool that provides a graphical representation of a project's timeline. The Program Evaluation Review Technique (PERT) breaks down the individual tasks of a project for analysis. PERT charts are considered preferable to Gantt charts.

Activity	Optimistic	Most Likely	Pessimistic	Expected	SD
A	4	5	6	5	0.33
B	5	6	7	6	0.33
C	9	10	11	10	0.33
D	3	4	5	4	0.33
E	3	4	5	4	0.33

Table 5.2. Pert Table

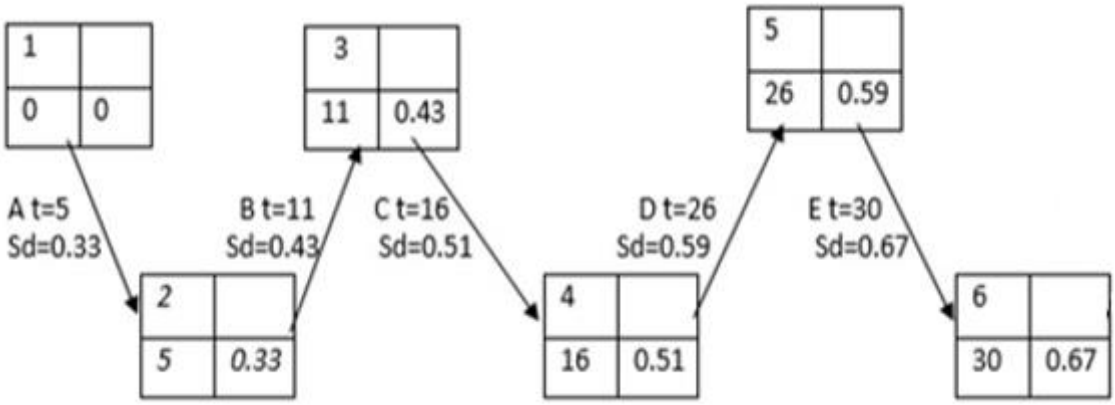


Figure 5.2. Pert chart

5.2 Project Activation:

The problem that we are going to solve here is that given a set of features that describe a house in Boston, our machine learning model must predict the house price. To train our machine learning model with Boston housing data, we will be using scikit-learn's Boston dataset.

In this dataset, each row describes a Boston town or suburb. There are 506 rows and 13 attributes (features) with a target column (price). https://archive.ics.uci.edu/ml/machine-learning-databases/housing

The dataset include features like CRIM, ZN, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, B, LSTAT, and MEDV.

5.3 Project Operation:

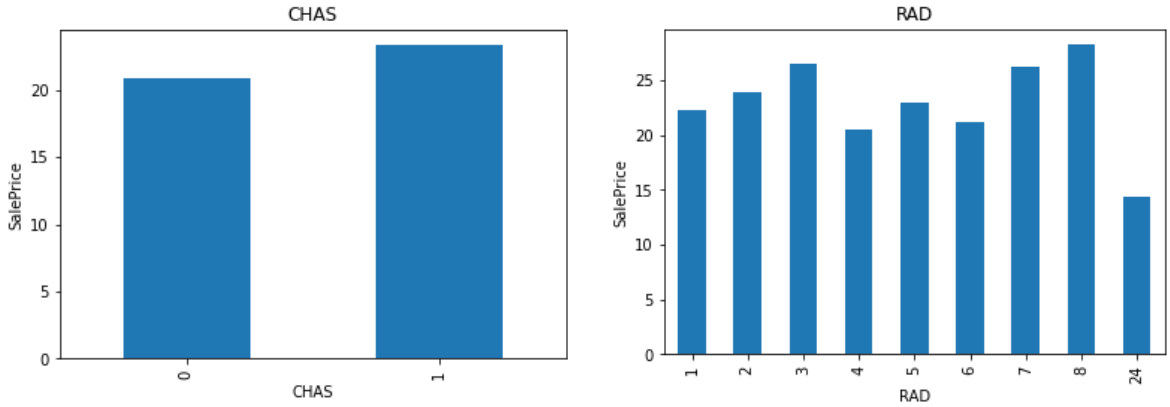
To work on this dataset we had use Jupyter Notebook as our platform. Then in new notebook we load the dataset. Then we start having a first impression by importing various python libraries such as Numpy, Matplotlib, Pandas, Seaborn. Plotting graphs and conducting various operation such as dataframe, shape, dataframe.describe(). This helps us to interact with the data and having a close look at the contents of dataset. Our dataset contains few outliers so to remove that outliers we can apply log transformation. With the help of Pearson correlation we can find the correlation between the dependent and independent features. Then applying the regression model we will get a generalized model for our house price predictions.

5.4 Experimental Methodology:

EDA in which we analysis our data properly check out there any null values are present or not finding the relationship between features. In our dataset there are independent features/variables like CRIM, ZN, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, B, and LSTAT. Basically analysing the relationship between the independent and dependent features. The dependent feature/variable is MEDV which is the house price in median values ($1000).

As we have two discrete variable CHAS and RAD so finding their relationship between House Price and discrete variables.

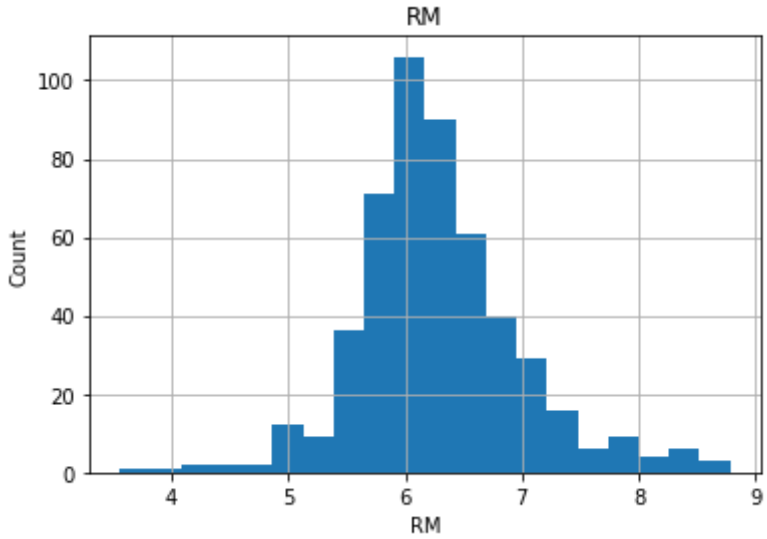
Figure 5.3



As we can see in the figure the CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise). If there is a tract bounds rivers is high the Sales Price is high so this type of relationship is really helpful for our data analysis. In the other RAD index of accessibility to radial highways, the key is the relationship between values - if one RAD value indicates greater highway accessibility than the other (e.g. a score of 5 vs 3), then it is ordinal. Variable comprises a finite set of discrete values with a ranked ordering between values is ordinal value.

The continuous feature in our dataset is CRIM, ZN, INDUS, NOX, RM, AGE, DIS, TAX, PTRATIO, B, LSTAT, and MEDV.

Figure 5.4



By checking the histogram plot in (figure 5.4) all continuous feature some of them are skewed data but in the RM feature is a Gaussian distribution which is also saying that it is one of the main feature in our dataset.

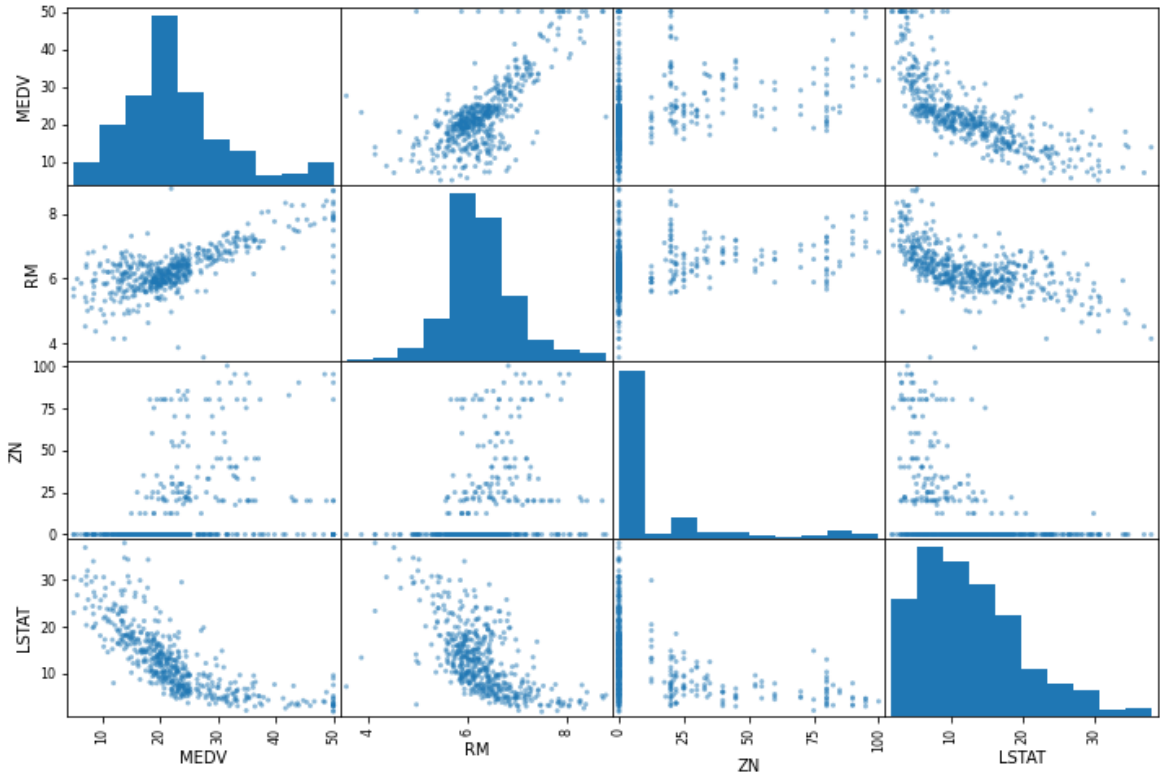
Correlation is also important method in which it describes how much your relation is strong in which we had used Pearson Correlation which ranges from -1 to 1.

Table 5.3

MEDV	**1.000000**
RM	**0.695360**
ZN	**0.360445**
B	**0.333461**
DIS	**0.249929**
CHAS	**0.175260**
AGE	**-0.376955**
RAD	**-0.381626**
CRIM	**-0.388305**
NOX	**-0.427321**
TAX	**-0.468536**
INDUS	**-0.483725**
PTRATIO	**-0.507787**
LSTAT	**-0.737663**

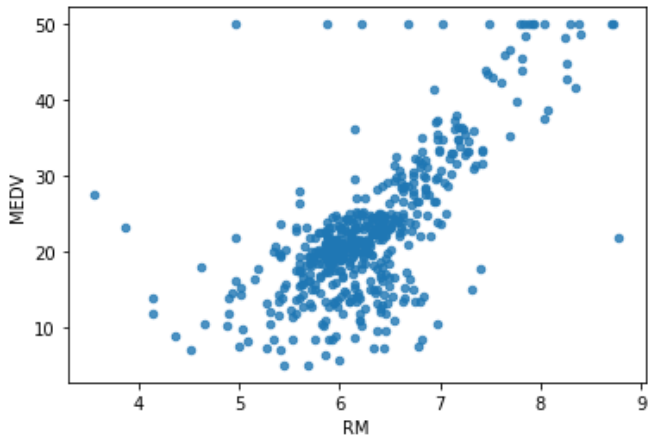
As we can see in the (Table 5.3) RM, ZN, PTRATIO, LSTAT is showing high correlation as there are useful for data analysis which we can plot through scatter matrix which I shown in (Figure 5.5)

Figure 5.5



To improve our models we can remove the outliers by doing log transformation which will reduce the outlier and skewness or also we can do inter-quantile on a Gaussian distribution.

Figure 5.6 Outliers detection



The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model. It is a fast and easy procedure to perform, the results of which allow you to compare the

performance of machine learning algorithms for your predictive modeling problem. Although simple to use and interpret, there are times when the procedure should not be used, such as when you have a small dataset and situations where additional configuration is required, such as when it is used for classification and the dataset is not balanced.

As here we had also used stratified shuffle split which creates a single training/testing set having equally balanced (stratified) classes. Essentially this is what you want with the n_iter=1. You can mention the test-size here same as in train_test_split.

Feature Scaling is of two types Min-Max Scaler and Standard Scaler in this dataset we had used Standard Scaler which is also known as Standardization (Z-Score Normalization).

Then after using various model and from that selecting a generalize model which will help to increase your accuracy. The goal of cross-validation is to test the model's ability to predict new data that was not used in estimating it, in order to flag problems like overfitting or selection bias and to give an insight on how the model will generalize to an independent dataset (i.e., an unknown dataset, for instance from a real problem).

The testing of model can done through cross validation and metrics performance in this project we had used RMSE. Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. We can also use different metrics like R squared, Mean Absolute Error.

CHAPTER 6

RESULTS AND DISCUSSIONS

Source of information while choosing Real Estate property

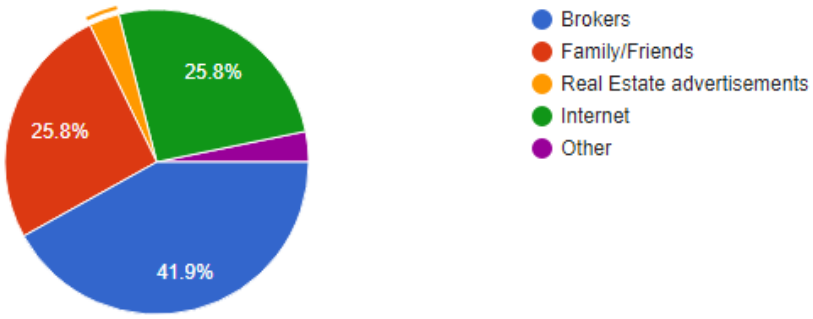
Table 6.1 Source of information while choosing Real Estate property

Total	Brokers	Family/Friends	Real Estate	Internet	Other
31	13	8	1	8	1

Source: Primary data

From the above table regarding choosing Real estate property 41.9% are searching information from Brokers, 25.8% from family, 3.2% from Real Estate advertisements, 25.8% from internet and 3.2% from others.

Figure 6.1 Source of information while choosing Real Estate property



Opinion regarding own a property or not

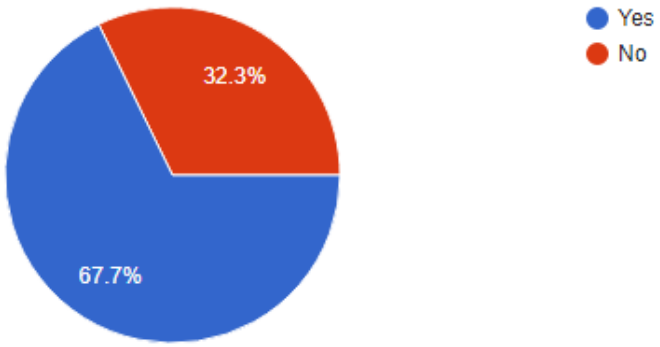
Table 6.2 Opinion regarding own a property or not

Total	Yes	No
31	21	10

Source: Primary Data

From the above table 67.7% of people had own a property and 32.3% of people not own a property.

Figure 6.2 Opinion regarding own a property or not



Opinion regarding Machine Learning is really important for Real Estate Company

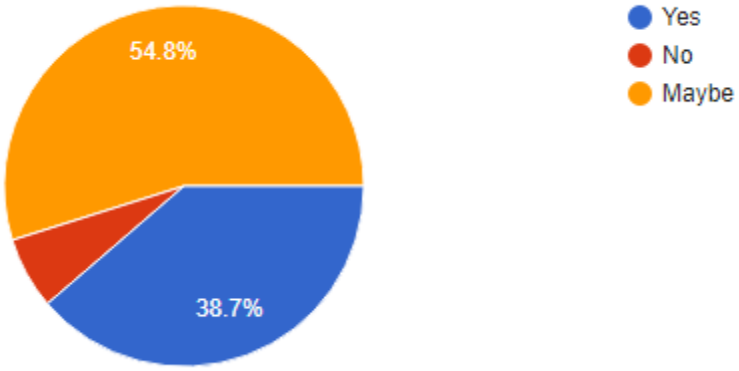
Table 6.3 Opinion regarding Machine Learning is really important for Real Estate Company

Total	Yes	No	Maybe
31	12	2	17

Source: Primary Data

In terms of using machine learning model for predicting house prices 38.7% thinks that is useful but 6.5% thinks it is not and 54.8% people think maybe it is useful.

Figure 6.3 Opinion regarding Machine Learning is really important for Real Estate Company



Opinion regarding price range of the property

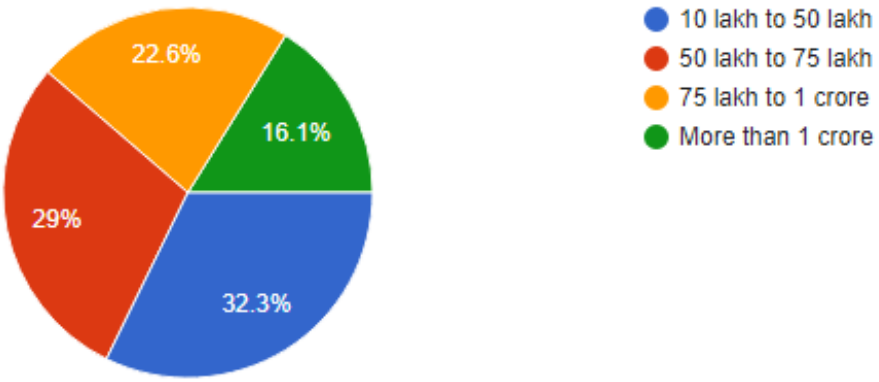
Table 6.4 Opinion regarding price range of the property

Total	10 lakh to 50 lakh	50 lakh to 75 lakh	75 lakh to 1 crore	More than 1 crore
31	10	9	7	5

Source: Primary Data

Most of the people are likely to buy a house between 10 lakh to 50 lakh are 32.3%, then 50 lakh to 75 lakh is 29%, between 75 lakh to 1 crore 22.6% and more than 1 crore is 16.1%.

Figure 6.4 Opinion regarding price range of the property



Opinion regarding property to be based on which region

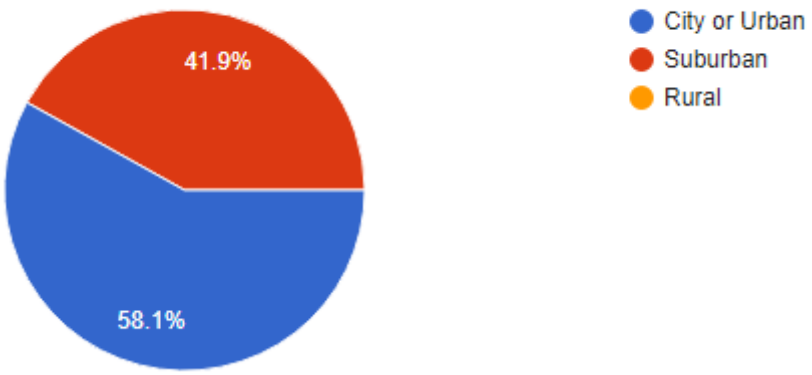
Table 6.5 Opinion regarding property to be based on which region

Total	City or urban	Sub-urban	Rural
31	18	13	0

Source: Primary Data

No one is like to choose rural side region but most of the people are likely to choose City/Urban side area which is 58.1% and 41.9% people are like to choose Sub-urban region.

Figure 6.5 Opinion regarding property to be based on which region



Opinion regarding consulted a Real Estate Company/Agent

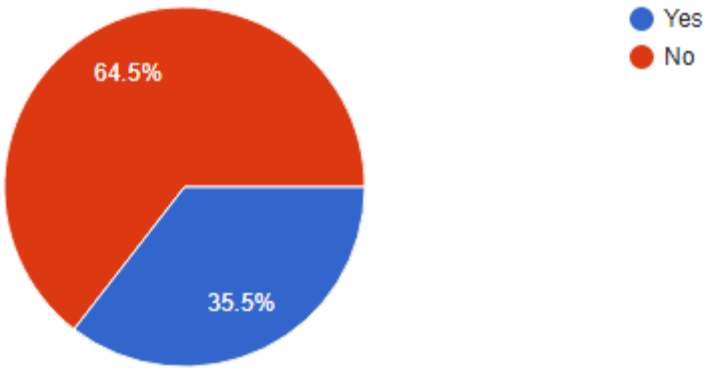
Table 6.6 Opinion regarding consulted a Real Estate Company/Agent

Total	Yes	No
31	20	11

Source: Primary Data

According to the table 64.5% people consulted a Real estate company/Agent and 35.5% had not consulted an agent.

Figure 6.6 Opinion regarding consulted a Real Estate Company/Agent



Opinion regarding on which Real estate property to invest

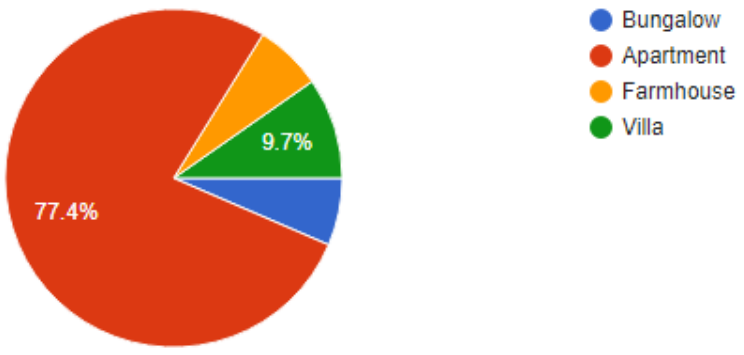
Table 6.7 Opinion regarding on which Real estate property to invest

Total	Bungalow	Apartment	Farmhouse	Villa
31	2	24	2	3

Source: Primary Data

About 77.4% are likely to invest in Apartment and 9.7% like to invest in Villa. In Bungalow and Farmhouse there are 6.5% which is like to invest.

Figure 6.7 Opinion regarding on which Real estate property to invest



Opinion regarding making decision immediately if you get right property

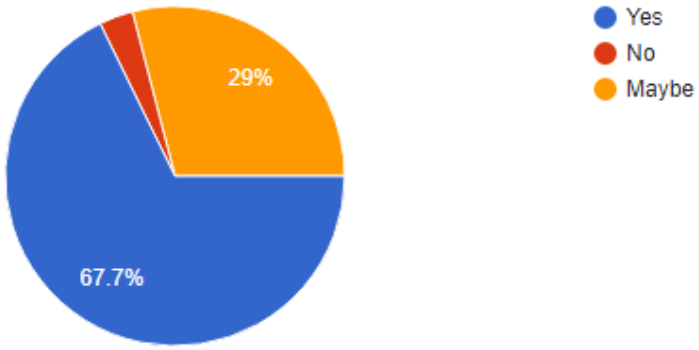
Table 6.8 Opinion regarding making decision immediately if you get right property

Total	Yes	No	Maybe
31	21	1	9

Source: Primary Data

According to the table 67.7% people are likely to make decisions immediately when they get right property, 3.2% thinks that they can't take decisions quickly and 29% thinks it is maybe to make decisions immediately.

Figure 6.8 Opinion regarding making decision immediately if you get right property



CHAPTER 7
CONCLUSION AND FUTURE WORK

7.1 CONCLUSION:

In this paper, an outline of the idea of ML alongside its different applications is examined. Taking the example dataset for houses, and thinking about its different features, the costs for houses have been anticipated by predicting ML model techniques like regression, for predicting the price of estate using prior data. In addition, our models also helped identify which characteristics of housing were most strongly associated with price and could explain most of the price variation.

As the Random Forest Regressor gives the highest accuracy which is having best RMSE score for our dataset. The CRIM, INDUS, NOX, AGE, TAX, PTRATIO, LSTAT, RAD have a weak negative influence on house prices whereas ZN, B, DIS, CHAS have a weak positive influence.

Ml driven predictions are easily comprehensible and significant from a data analysis of point. When correctly implemented a high rate of accuracy can be achieved, and thus ML techniques find applications across a wide range of fields. Algorithms are distinguished based on various metrics, for instance, accuracy, precision and specificity.

7.2 FUTURE SCOPE:

This application can be easily implemented under various situations we can add new features as when we required to increase the accuracy. Reusability is possible as and when require in this project.

Extensibility: Furthermore, we were able to improve our models' prediction accuracy by accounting for the impact of spatial location. We were able to identify most of the residential areas. There may be some more places that have housing complexes or multi-story apartments that are located in commercial areas. The used pre-processing methods do help in the prediction accuracy. However, experimenting with different combinations of pre-processing methods to achieve better prediction accuracy. There are several other models available that can be implemented for prediction. Data given as input to such model should be compatible with the tool used and the operators involved in the process. Also, more number of data sets can be used to increase the accuracy of the model.

Reusability: Reusability is possible as and when require in this project. As we can used the created pipline code for various data analysis. Reducing the amount of code also simplifies understanding which increases the likelihood that the code is correct. We follow both type of reusability sharing of newly written code within a project and reuse of previously written code on new projects

CHAPTER 8

APPENDICES

Appendix I: List of table

Appendix II: List of figures

Appendix III: List of Abbreviation

Abbreviation	Meaning	Pg. No.
RMSE	Root Mean Squared Error	14,15,67,75
MSE	Mean Squared Error	14
AI	Artificial Intelligence	15
CRIM	This is the per capita crime rate by town	43
ZN	This is the proportion of residential land zoned for lots larger than 25,000 sq.ft.	43
INDUS	This is the proportion of non-retail business acres per town.	43
CHAS	This is the Charles River dummy variable (this is equal to 1 if tract bounds river; 0 otherwise)	43
NOX	This is the nitric oxides concentration (parts per 10 million)	43
RM	This is the average number of rooms per dwelling	43
AGE	This is the proportion of owner-occupied units built prior to 1940	43
DIS	This is the weighted distances to five Boston employment centers	43
RAD	This is the index of accessibility to radial highways	43
TAX	This is the full-value property-tax rate per $10,000	43
PTRATIO	This is the pupil-teacher ratio by town	43

B	This is calculated as $1000(Bk — 0.63)^2$, where Bk is the proportion of people of African American descent by town	43
LSTAT	This is the percentage lower status of the population	43
MEDV	This is the median value of owner-occupied homes in $1000s.	44
UML	Unified Modeling Language	53
ASM	Attribute Selection Measure	65

CHAPTER 9

REFERENCES

A. Journal References:

1. Nikita Malik, Vidhu Gaba, Priyansh (2020) [1]. *"Employing Machine Learning for House Price Prediction"*, https://www.researchgate.net/publication/344099603_Employing_Machine_Learning_for_House_Price_Prediction
2. Darshil Shah, Harshad Rajput, Jay Chheda. (2020) [2]. *"House Price Prediction Using Machine Learning and RPA"*, https://www.irjet.net/archives/V7/i3/IRJET-V7I31123.pdf
3. Dr. M. Thamarai, Dr. S P. Malarvizhi (2020) [3]. *"House Price Prediction Modeling Using Machine Learning"*, http://www.mecs-press.org/ijieeb/ijieeb-v12-n2/IJIEEB-V12-N2-3.pdf
4. Ahmad Abdulal, Nawar Aghi (2020) [4]. "*House Price Prediction*", https://www.diva-portal.org/smash/get/diva2:1456610/FULLTEXT01.pdf
5. Alisha Kuvalekar, Shivani Manchewar, Sidhika Mahadik. (2020) [5]. "*House Price Forecasting using Machine Learning"*, https://cutt.ly/7zwWbUr
6. Winky K.O. Ho , Bo-Sin Tang & Siu Wai Wong (2020) [6]. "*Predicting property prices with machine learning algorithms"*, https://www.tandfonline.com/doi/pdf/10.1080/09599916.2020.1832558?needAccess=true
7. Sayan Putatunda (2019) [7]. *"PropTech for Proactive Pricing of Houses in Classified Advertisements in the Indian Real Estate Market*", https://arxiv.org/ftp/arxiv/papers/1904/1904.05328.pdf

8. Maharshi Modi, Ayush Sharma, Dr. P. Madhavan.(2020) [8]. "*Applied Research on House Price Prediction Using Diverse Machine Learning Techniques*", http://www.ijstr.org/final-print/apr2020/Applied-Research-On-House-Price-Prediction-Using-Diverse-Machine-Learning-Techniques.pdf
9. Neelam Shinde, Kiran Gawande. (2018) [9]. *"Valuation of House Prices using Predictive Techniques*", http://www.iraj.in/journal/journal_file/journal_pdf/12-477-153396274234-40.pdf
10. Ping-Feng Pai and Wen-Chang Wang. (2020) [10]. "*Using Machine Learning Models and Actual Transaction Data for Predicting Real Estate Prices",* https://www.mdpi.com/2076-3417/10/17/5832/pdf
11. Aswin Sivam Ravikumar. (2017) [11]. "*Real Estate Price Prediction Using Machine Learning*", http://norma.ncirl.ie/3096/1/aswinsivamravikumar.pdf
12. Chao Xue, Yongfeng Ju, Shuguang Li, Qilong Zhou and Qingqing Liu. (2020) [12]. "*Research on Accurate House Price Analysis by Using GIS Technology and Transport Accessibility*", https://www.mdpi.com/2073-8994/12/8/1329/pdf
13. Puneet Tiwari, Varun Singh Thakur (2020) [13]. "*Review on House Price Prediction through Regression Techniques*", https://www.ijspr.com/citations/v73n1/IJSPR_7301_30604.pdf
14. Thuraiya Mohd, Suraya Masrom, Noraini Johari. (2019) [14]. *"Machine Learning Housing Price Prediction in Petaling Jaya, Selangor, Malaysia"*, https://www.ijrte.org/wp-content/uploads/papers/v8i2S11/B10840982S1119.pdf
15. Anurag Sinha (2020) [15]. *"Utilization of Machine Learning Models in Real Estate House Price Prediction"*, https://amity.edu/UserFiles/aijem/119Paper%203.pdf

16. Akshay Babu, Dr. Anjana S Chandran.(2019) [16]. "*Literature Review on Real Estate Value Prediction Using Machine Learning",* http://ijcsma.com/publications/march2019/V7I302.pdf

17. Gaikwad Purva Chandrakant, Ganjave Pratiksha Namdev, Gorade Pooja Subhash, S. S. Gore (2019) [17]. "*Implementation of House Price Prediction Model Using Image Processing and Machine Learning",* https://www.ijresm.com/Vol.2_2019/Vol2_Iss11_November19/IJRESM_V2_I11_105.pdf

18. Zhou, Yichen (2020) [18]. *"Housing Sale Price Prediction Using Machine Learning Algorithms*", https://escholarship.org/uc/item/3ft2m7z5

19. Thuraiya Mohd, Syafiqah Jamil, Suraya Masrom (2020) [19]. "*Machine learning building price prediction with green building determinant",* http://ijai.iaescore.com/index.php/IJAI/article/download/20491/pdf

20. Mr. Rushikesh Naikare, Mr. Girish Gahandule, Mr. Akash Dumbre, Mr. Kaushal Agrawal (2019) [20]. "*House Planning and Price Prediction System using Machine Learning",* http://www.ierjournal.org/pupload/vol3iss3/House%20Planning%20and%20Price%20Prediction%20System%20using%20Machine%20Learning.pdf

21. Arshiya Shaikh, R.Vinayaki, G. Siddhanth, Y. Phanindra Varma (2020) [21]. "*House Price Prediction using Multi Variate Analysis",* https://ijcrt.org/papers/IJCRT2002194.pdf

22. Parth Ambalkar, Akash Mane, Tanmay Maity (2019) [22]. "*House price prediction using various machine learning algorithms",* https://www.ijariit.com/manuscripts/v5i4/V5I4-1321.pdf

23. Bindu Sivasankar, Arun P. Ashok, Gouri Madhu, Fousiya S (2020) [23]. "*House Price Prediction",* https://ijcseonline.org/pub_paper/16-IJCSE-08250-47.pdf

24. by Mr. S. Vijayakumar, Mr. B. Ramkumar, Mr. S. Ranjith, Mr. M. Seenivasan, Mr. G. Siva (2020) [24]. *"House Price Prediction Based on Some Economic Factors Using Machine Learning"*, https://irjmets.com/rootaccess/forms/uploads/house-price-prediction-based-on-some-economic-factors-using-machine-learning.pdf

25. Hujia Yu, Jiafu Wu (2016) [25]. "*Real Estate Price Prediction with Regression and Classification",* http://cs229.stanford.edu/proj2016/report/WuYu_HousingPrice_report.pdf

26. Jingyi Mu, Fang Wu and Aihua Zhang (2014) [26*]. "Housing Value Forecasting Based on Machine Learning Methods",* https://downloads.hindawi.com/journals/aaa/2014/648047.pdf

27. Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, Rick Siow Mong Goh (2017) [27]. *"A Hybrid Regression Technique for House Prices Prediction",* https://www.researchgate.net/publication/323135322_A_hybrid_regression_technique_for_house_prices_prediction

28. Nihar Bhagat, Ankit Mohokar, Shreyash Mane (2016) [28]. *"House Price Forecasting using Data Mining",* https://www.ijcaonline.org/archives/volume152/number2/bhagat-2016-ijca-911775.pdf

29. Prabha D, Anindhitha A, Archana A, Balaji Narasimhan M.V.L. (2020) [29]. *"Predicting House Price Values Using Linear Regression with Ridge Regularization Approach"*, http://sersc.org/journals/index.php/IJAST/article/view/18069/9175\

30. Bruno Klaus de Aquino Afonso, Luckeciano Carvalho Melo, Willian Dihanster Gomes de Oliveira, Samuel Bruno da Silva Sousa, Lilian Berton (2019) [30].

"Housing Prices Prediction with a Deep Learning and Random Forest Ensemble", https://www.researchgate.net/publication/335527230_Housing_Prices_Prediction_with_a_Deep_Learning_and_Random_Forest_Ensemble

B. Book Reference:

1. MACHINE LEARNING: The Art and Science of Algorithms that ... (n.d.). Retrieved from http://dsd.future-lab.cn/members/2015nlp/Peter_Flach_Machine_Learning._The_Art_and_Scienc(BookZZ.org).pdf
2. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). An introduction to statistical learning: With applications in R. New York: Springer.

C. Website Reference:

1. Patil, P. (2018, May 23). What is Exploratory Data Analysis? Retrieved from https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15
2. Brownlee, J. (2020, August 26). Train-Test Split for Evaluating Machine Learning Algorithms. Retrieved from https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms
3. Sklearn.model_selection.StratifiedShuffleSplit¶. (n.d.). Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedShuffleSplit.html
4. Brownlee, J. (2020, August 27). How to Use StandardScaler and MinMaxScaler Transforms in Python. Retrieved from https://machinelearningmastery.com/standardscaler-and-minmaxscaler-transforms-in-python

5. Brownlee, J. (2020, August 02). A Gentle Introduction to k-fold Cross-Validation. Retrieved from https://machinelearningmastery.com/k-fold-cross-validation
6. Machine Learning - Performance Metrics. (n.d.). Retrieved from https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_algorithms_performance_metrics.htm