MBA Semester – IV

Research Project – Interim Report

Name	Maria Wilfred Melvin
Project	House Price Prediction
Group	1
Date of Submission	29.08.2024

A study on "House Price Prediction in Urban India using Machine Learning Algorithm"

Research Project submitted to Jain Online (Deemed-to-be University)

In partial fulfillment of the requirements for the award of:

Master of Business Administration

Submitted by:

Maria Wilfred Melvin

USN:

222VMBR02629

Under the guidance of:

Mr. Sharath Srivatsa

(Faculty-JAIN Online)

Jain Online (Deemed-to-be University)

Bangalore

2023-24

DECLARATION

I, Maria Wilfred Melvin, hereby declare that the Research Project Report titled “A Study on House Price Prediction in Urban India using Machine Learning Algorithm” *has been* prepared by me under the guidance of Mr. Sharath Srivatsa. I declare that this Project work is towards the partial fulfillment of the University Regulations for the award of the degree of Master of Business Administration by Jain University, Bengaluru. I have undergone a project for a period of Eight Weeks. I further declare that this Project is based on the original study undertaken by me and has not been submitted for the award of any degree/diploma from any other University / Institution.

Place: Bengaluru

Date: 29.08.2024

Maria Wilfred Melvin
USN: 222VMBR02629

Objectives:

The objective of the study is

- To predict the Indian rental house price by analysing their status of the house
- To predict the Indian rental house price by analysing their negotiability

Scope of the Study:

The purpose of the study is to predict the rental house prices as per the status of the house and its negotiability. It is possible to forecast future prices by examining historical market trends, upcoming developments, and property prices. A common practice in real estate is to list the standard and general characteristics apart from the asking price and general description. These features can be easily compared because each of the attributes is listed separately in an organized manner. Since each home has its own distinct qualities, such as its view or type of sink, House sellers can list all the salient features of the property in the description. Although buyers can take into account real estate features, the great diversity makes it impossible to offer an automated comparison of all variables. Conversely, house sellers must assess the worth of the property on the basis of its features. They have to compare the features with the current market price. It is challenging to forecast a fair market price and rent because of the variety of features. Thus, the information induces the curiosity of buyers to buy the house with its features.

Research Methodology:

Research methods: The present study adopts quantitative research methods. This method quantifies the results on the basis of statistical analysis. The information is interpreted with numbers. It is highly structured and formalized. The determination of results is based on studying a few variables in a large number of entities.

Data set: The Indian Rental House dataset is available on Kaggle and contains information about rental house price details in India (https://www.kaggle.com/datasets/bhavyadhingra00020/india-rental-house-price/data). This dataset contains more than 13913 data points with 16 variables, including house type, house size, location, city, latitude, longitude, price, currency, number of bathrooms and balconies, which are negotiable, price per square foot, verification date, description, security deposit, and status. All the variables served as features of the dataset. This helps to predict the price on the basis of status and negotiability.

As soon as the variables are determined, it is necessary to preprocess the data. It is done with three points. One is to check for missing data points. Variables that contain more than 50% of the missing data are removed from the dataset. Get rid of the whole attribute and set the values closer to the mean. Check the outliers of the features; remove the outliers of the features to increase the performance of the model. Next is normalising the numerical values and encoding the categorical values one at a time. The next step is to explore the data and find the appropriate feature using a heat map. The correlation matrix is performed for all the features to find out the most correlated and least correlated variables. The results help to recognise the interaction between the features and target variables. Next is to select the most important features that will be used for the study. When identifying the features, scaling is carried out to ensure the proper selection of the most appropriate features of the study. The lower the features, the lower the performance of the model. The standard scalar function helps to naturally distribute the data within the function. It is now clustering at about 0 with SD1. Next is model development. It represents the core of the rental price process. The data is divided into testing and training. The ratio of training and testing will be 80:20. The training set of 80% inclusion of target variable. During the training phase, the model is fitted to the training set of data, optimised for hyperparameters, and assessed using suitable metrics like mean squared error and R-squared. The prediction models are trained using training sets that provide a wealth of data to show the relationship between features and different objectives. The model gains knowledge that it uses to make predictions. The model is trained with different machine learning algorithms. The results help predict the model with the highest accuracy.

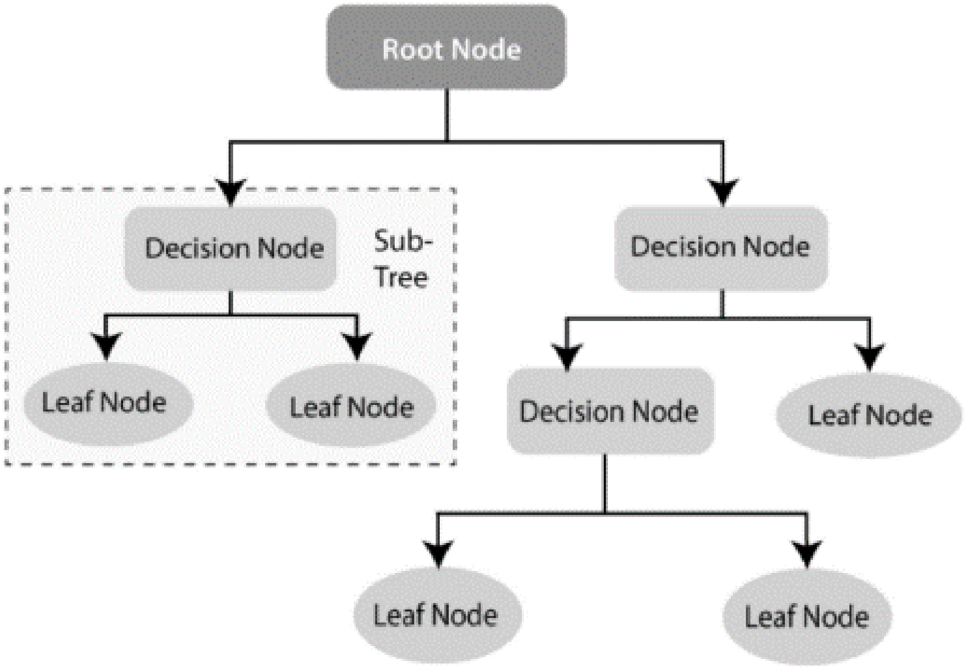
Machine Learning Algorithm:

The present study adopts classification analysis to perform machine learning algorithms. It is also a supervised learning method. Classification analysis can be described as a predictive modelling problem. It is a mathematical mapping of a function (f) as a target from input (x) to output variables (Y). The prediction of class is available for structured or unstructured data. In this analysis, the class label is predicted for the subsequent example. The best example for classification analysis is detecting email spam that can be classified as spam and not spam in the classification problem. The present study adopts the most commonly used methods, including Naïve Bayes, logistic regression, KNN, SVM, Decision Tree, and RF.

- Naïve Bayes algorithm: This relies on the assumption of independence between every pair of features. It is based on Bayes theorem. This is highly suitable for binary classification and multi-class classification in different real-world situations, including text classification, document classification, spam filtering, and so on. When utilizing the NB classifier, this model classifies the noisy occurrences in the data and builds a reliable prediction model. The benefit of this algorithm is that it needs less training data to quickly estimate parameters. This is more efficient than more complex methods. On the other hand, due to its feature independence assumption, the performance of NB classifiers might be affected.

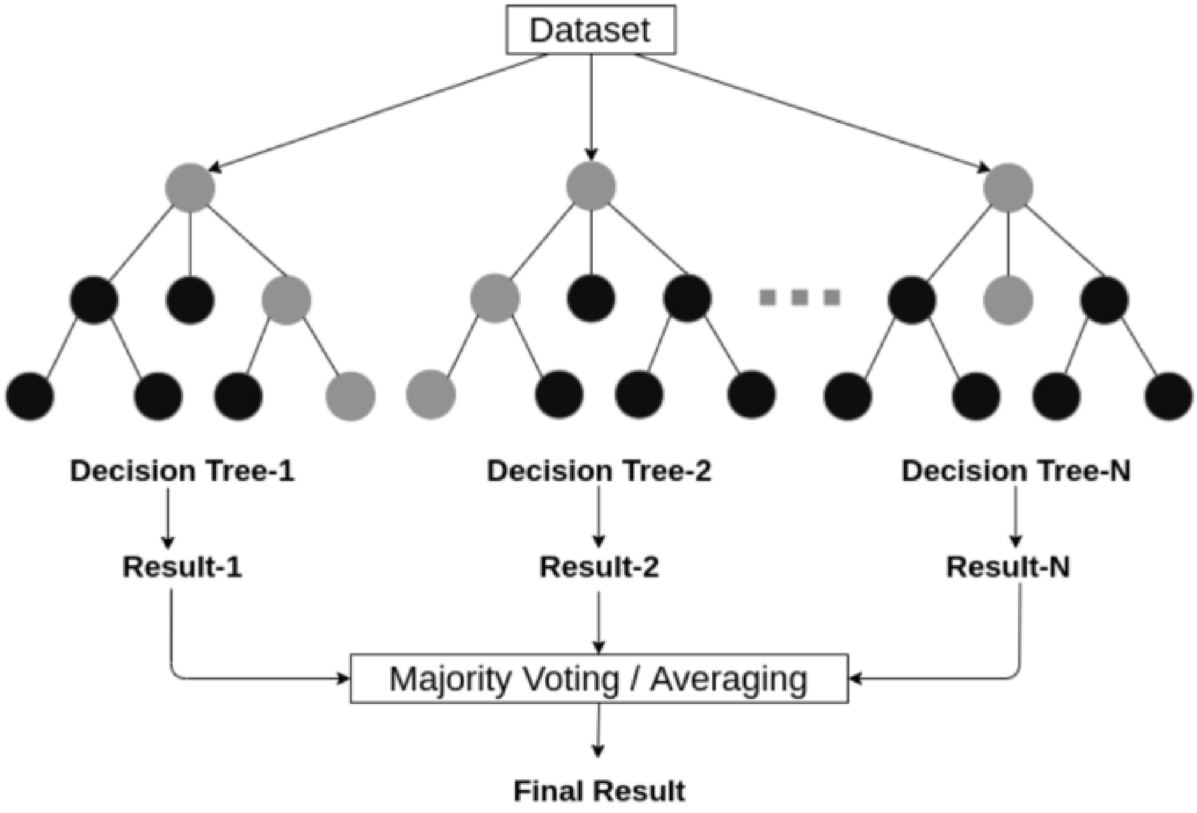
- Logistic regression is a common probabilistic-based statistical model. It is used to solve classification problems in machine learning algorithms. In order to estimate the probabilities in a logistic function, logistic regression is used. It is most effective when a dataset can be divided linearly. It has a tendency to overfit high-dimensional datasets. This can be overcome with the help of regularization techniques. This is highly suitable for classification problems. On the other hand, the drawback of this algorithm is the assumption of linearity between the variables (independent and dependent variables).
- KNN, abbreviated as K-Nearest Neighbors, is a non-generalizing learning method. It is also an instance-based learning method, also known as a lazy learning algorithm. Instead of concentrating on building a general internal model, it stores all instances in n-dimensional space that correspond to the training data. The classification of this algorithm is to determine new data points on the basis of similarity measures. The advantage of this technique is that it is suited for noisy training data. The accuracy of the algorithm relies on the quality of the quality of the data. On the other hand, the most significant problem is determining the optimal number of neighbors.
- SVM is also a classification technique. This is also used for classification, regression, and other tasks. This builds a hyperplane or collection of hyperplanes in high- or infinite-dimensional space. The larger the margin, the lower the classification error. This means that the hyperplane is the furthest from the closest data training points in any class and would achieve a strong separation. This works well, especially in high-dimensional space. It exhibits variable behavior that depends on various mathematical functions, known as the kernel. Some of the popular kernels include linear, radial basis functions, sigmoid, polynomials, and so on. Although the data is noisy due to overlapping target classes, it did not perform well.

- A decision tree is a supervised learning method. This is also suited for classification tasks. Some of the well-known algorithms include CART, C4.5, and ID3.



- The following figure illustrates how to sort the tree down to a few leaf nodes from the root. This is known as an instance, which is also categorized by DT. Instances are classified by examining the attribute defined by the node. Starting at the root node of the tree, work down the branch that corresponds to the attribute values. The most popular criteria include gini for gini impurity and entropy for information gain.

- Random forest is an ensemble classification technique. It is widely used in the field of machine learning. This is suitable for both classification and categorical values. Parallel ensembling is a technique used in this method. This fits multiple decision tree classifiers concurrently. It is illustrated in the subsequent figure. Utilizing sub-samples from various data sets, majority voting or averages are used to determine the outcome. As a result, it can diminish the issue of overfitting and improve prediction accuracy.


Dataset
Decision Tree-1
Decision Tree-2
Decision Tree-N
Result-1
Result-2
Result-N
Majority Voting / Averaging
Final Result