House Price Prediction Using Machine Learning

1st SACHIN KUMAR.S
Post Graduate Student
dept. of Mathematics, SAS
VIT University
Vellore, India
sachinkumar.s2022@vitstudent.ac.in

2nd VIGNESHWARAN.P
Post Graduate Student
dept. of Mathematics, SAS
VIT University
Vellore, India
vigneshwaran.p2022@vitstudent.ac.in

3rd KIRAN KUMAR.S
Post Graduate Student
dept. of Mathematics, SAS
VIT University
Vellore, India
kirankumar.s2022@vitstudent.ac.in

Under the guidance of
DR. JAYALAKSHMI M
Assistant Professor
dept. of Mathematics, SAS
VIT University
Vellore, India
m.jayalakshmi@vit.ac.in

***Abstract*—The use of predictive technologies to estimate the sale price of residences in large cities is getting harder and harder. The sale price of real estate in cities like Bangalore is affected by a number of interconnected factors, including Area, location, and the property's amenities are significant factors that could affect a home's price. This system tries to create a prediction model for the price evaluation based on the price-affecting factors. In this study, we employ a variety of prediction methods, including gradient boosting regression, random forest regression, decision trees, lassos, and linear regression. All of the strategies presented have been used to predict house prices on the dataset to see which is most successful.The primary goal of this essay is to assist readers in selecting the best sort of home and neighbourhood for their needs. The model will forecast the price using price-influencing criteria including square footage, the number of bedrooms, bathrooms, and most importantly location.**

***Index Terms*—Machine Learning, Linear Regression, Lasso, Decision Tree, Random Forest Regression**

I. Introduction

Machine learning is a trending technology that we use either directly or indirectly in our daily operations. Every business is moving toward automation, and over the past few years, machine learning has become increasingly important in fields like health care, banking and finance, traffic prediction, image and speech recognition, etc. Today's real estate market is unique among others because it is primarily concerned with pricing. People are searching to purchase a new residence and are considering their budget, top priorities, and market research. For a very long time, the cost of the houses was manually determined, which was subject to human error. Therefore, the primary goal of our study is to accurately anticipate the property price in a specific location. We try to predict the efficient house pricing for customers by considering many factors with respect to their budgets and their priorities. So, we are creating a house price prediction model. By using different Machine Learning algorithms like Linear Regression, Lasso, Decision Tree, Random Forest Regression Gradient Boosting Regression. Data is the heart of Machine Learning without data we can't train model. Machine learning uses previous data and by using them predict the new data by using the models. 80% of the known dataset has been used for the training the models and remaining 20% is used for the testing the models. This model will help the people to know about the pricing of the houses in their priorities without moving towards a broker. One hot encoding, feature engineering, dimensionality reduction, outlier elimination, and k fold cross validation are just a few of the approaches used in this work. This demonstrates that the field of machine learning is necessary for the burgeoning field of housing price prediction. The study's findings show that the linear regression model has the highest degree of accuracy. [1].

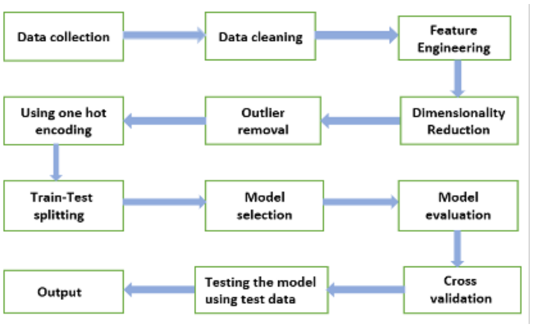


Fig. 1. Research Flow Diagram.

II. Literature survey

There are many systems are proposed for house price prediction.

The idea for House Resale Price Prediction Using Classification Algorithms was put forth by P. Durganjali and M. Vani Pujitha. They used a variety of classification methods in this study, including Linear Regression, Decision

Tree, Naive Bayes, and Random Forest. They also used the AdaBoost technique to help weaker learners become stronger. They added that the physical characteristics, geographic location, and economic variables are the elements influencing the resale price. [4].

The idea for using data mining to predict house prices was put out by Ankit Mohakar and Shreyash Mane. In this study, they analysed the present house price and projected the future price of the property based on the user's criteria, using linear regression methods to estimate pricing. [5].

House Price Prediction Using Machine Learning was proposed by Dr. Vinayak A. Bharadi, Anand G. Rawool, Dattaray V. Rogye, Sainath G. Rane, and others. They do this by employing machine learning techniques to develop a model, such as linear regression, decision tree regression, K-means regression, and random forest regression, the latter of which provides them the highest level of accuracy. [1].

Arshiya shaikh, R.Vinayaki, G.siddhanth, Panindra varma proposed the House Price Prediction Using Multivariate Analysis. In this they performed Multiple Linear Regression for estimating the house price based on area in square feet and number of bed rooms [6].

G.Gayathri Priya proposed the House Price Prediction Using Machine Learning Techniques. In this paper she used the various like Linear regression, Random forest regression, Polynomial regression, Robust regression, Lasso regression etc., and determine which is the most effective and accurately estimating the selling price of the house [7].

Using a variety of classification methods, including Naive Bayes, logistic regression, SVM classification, and Random Forest classification, as well as regression techniques, including Lasso, Ridge, SVM regression, and Random Forest regression, Hujia Yu and Jiafu Wu proposed Real Estate Price Prediction in 2016. Furthermore, the PCA technique is used to increase prediction accuracy. With an accuracy of 0.6740, SVC with a linear kernel is the model that performs the best for classification problems. With PCA pretreatment, this accuracy was enhanced to 0.6913, making it the best among all other algorithms with PCA preparation. The best-performing model for a regression problem is SVR with a gaussian kernel, which has an rmse of 0.5271. However, due to SVR's large dimensionality, visualisation is challenging. However, the lasso regression model can offer insights into certain qualities, which is useful in helping us understand the relationships between the features of a house and its selling prices [2].

Quang Truong and Minh Nguyen proposed house price prediction via improved machine learning techniques. in the paper they applied both traditional and advanced machine learning methods to their model to get optimum results for predicted house prices, particularly they applied the lowest root mean squared logarithmic error(RMSLE)on the test dataset, which belongs to the stack generalization method. other than that they used random forest, extreme gradient boosting, hybrid regression, and stacked generalization machine learning algorithms for getting results [12].

Machine learning-based house price prediction modelling was proposed by Dr. M. Tamarai and S. P. Malarvizhi in 2019. They used Scikit-learning to accomplish their work, which includes decision tree classification, decision tree regression, and multiple linear regression. [13].

Under the direction of the International Journal of Innovative Technology and Exploring Engineering, G.Naga Sathish, ch. V. Ragavendran, and m.d. Sugnana Rao (2019) completed their work on the housing price prediction project utilising machine learning ideas (IJITEE). For testing the data set and obtaining findings, they used numerous regression techniques in this research, including linear and lasso. [14].

III. Proposed System

In this proposed system "House Price Prediction Using Machine Learning" we mainly focus on the predicting the house price of the Bengaluru city area using various Machine learning models like Linear regression, Lasso regression, Decision tree regression, Random forest Regression. The dataset we used is the secondary data collected from Kaggle which is stored in a .csv format in which the attribute were area_type, availability, location, size, total_sqft, bath, balcony, price etc., Next, we cleaned the data, remove outliers, used one hot encoding, and split the data into 80% for training purpose and then remaining 20% for testing purpose. We majorly used three machine learning libraries to solve they are 'pandas', 'numpy', 'sklearn'. Pandas are used for load the .csv file into Jupiter notebook and used to clean data as well as manipulate the data. Numpy has the useful mathematical function and methods. Sklearn is the most useful library for Machine learning, real analysis, train test splitting, importing regressions contains the various inbuilt functions. The models are trained by various machine learning algorithms and their respective results are analysed for accuracy of the models. Among them Linear regression model gives tha maximum accuracy. Finally, the trained models are applied to predict the house price for the particular area. The trained model are saved by using 'pickle'.

IV. Methodology

A. Algorithms

Numerous Machine learning algorithms were researched while creating this model some of them are Linear regression, Lasso regression, Decision tree, Random forest regression. Out of this, Linear regression provides the most accurate house price prediction.

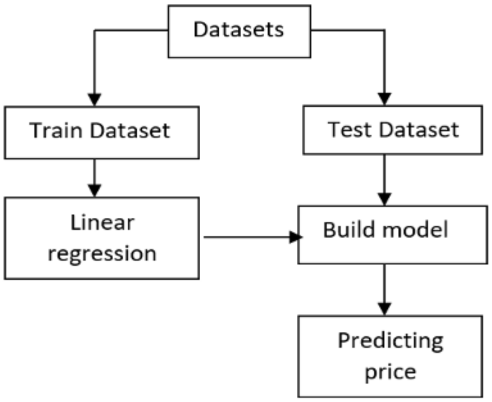


Fig. 2. Flow of development of the model.

B. Linear Regression

In this paper, we mainly analysed simple linear regression. This approach deals with the finding the relation between dependent and independent variables. This approach gives us the good estimation of the house price for a particular area, square feet, number of bedroom and bathroom. Simple linear regression expressed as followed,

$$y = a_0 + a_1 x + e \tag{1}$$

y is the independent variable and x is the dependent variable. a0 is the constant term, is the intercept of the regression line on the vertical axis and a1 is the regression coefficient slope of the regression line. e is the error term or disturbance term.

C. Lasso Regression

Least Absolute Shrinkage and Selection Operation. As the name suggest Lasso employs a "shrinkage" technique in which the determined coefficients are shrunk towards the mean central point. This is also known as penalized regression. This helps to increase the model interpretations and to minimise the number of errors.

$$\sum_{i=1}^{n}(y_i - \sum x_{ij}\beta_j) + \lambda \sum \ |\beta_j| \tag{2}$$

Where,

- λ denotes the quantity of shrinkage.
- λ=0 all parameters are taken into consideration and the estimate is equal to the linear regression.
- λ= ∞ no parameters taken into consideration.
- The bias increases with increase in λ
- Variance increases with decrease in λ

D. Decision Tree

The model is trained in the tree structure using a specific set of data after carefully observing the features of an object. The average value of the dependent variable in a specific leaf node serves as the final prediction. It benefits from being simple to grasp and requires less data cleaning.

E. Random Forest Regressor

Decision tree may have an overfitting problem it can be fixed in this method. It observes features of an attribute and train the model by analysing using the features and predict the price of the house. It produces several randomly selected decision trees from the data, averaging the outcomes to provide a fresh result that frequently produces accurate predictions or classifications.

F. K-Fold Cross Validation

This assessment strategy, which divides our sample data into folds of equal size, is used to determine how effectively our machine learning model can predict the outcome of the unseen data set. The data set was divided into k groups, with one group serving as the test and the other groups serving as the training. The machine learning model was fitted to the training set, and its performance was assessed using the test set. We can quickly determine the model's accuracy by utilising this method.

G. Grid Search CV

By locating the ideal hyperparameters, GRID Search CV is used in machine learning to improve the performance of our machine learning model. GRID SEARCH CV is a library function that is a part of the model selection package of the Sklearn framework. From the list of hyperparameters, this function assists in selecting the best settings. Along with the cross-validation methodology, this procedure is effective. In this project, to find which is more effective among the regression techniques, grid search cv is used.

Si. No.	Model	Score
1	Linear Regression	0.818354
2	Lassos Regression	0.687430
3	Decision Tree	0.751088
4	Random Forest Regression	0.799246

TABLE I
REGRESSION RESULTS

H. One Hot Encoding

It is a process of converting categorical data to binary numbers and this is called crucial part of feature engineering for building best model. The machine learning algorithm cannot understand categorical data, to let the computer understand, we use one hot encoding. In one hot encoding the categorical data are directly assigned to integer values and every integer value is transformed to binary value. In this proposed system we convert the 'location' attribute into binary value to let the machine learning algorithm understand our data to run the predictions [3].

	Feature Vector						Target
ID	F1	F2	F3	F4	...	F^m	classes
01	F$^1_1$	F$^2_1$	F$^3_1$	F$^4_1$	...	F$^m_1$	C1
02	F$^1_2$	F$^2_2$	F$^3_2$	F$^4_2$	...	F$^m_2$	C2
03	F$^1_3$	F$^2_3$	F$^3_3$	F$^4_3$	...	F$^m_3$	C3
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
N	F^1_N	F^2_N	F^3_N	F^4_N	...	F^m_N	Ck

TABLE II
INPUT DATA SET EXAMPLE BEFORE ONE HOT ENCODING.

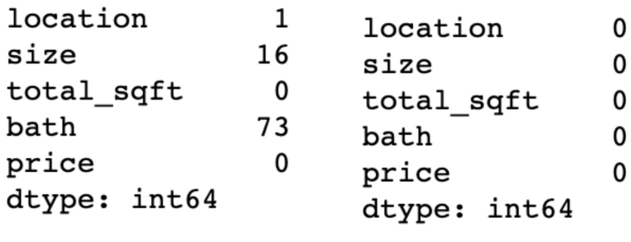
	Feature Vector						Target				
ID	F1	F2	F3	F4	...	F^m	C1	C2	C3	...	Ck
01	F$^1_1$	F$^2_1$	F$^3_1$	F$^4_1$	...	F$^m_1$	1	0	0	...	0
02	F$^1_2$	F$^2_2$	F$^3_2$	F$^4_2$	...	F$^m_2$	0	1	0	...	0
03	F$^1_3$	F$^2_3$	F$^3_3$	F$^4_3$	...	F$^m_3$	0	0	1	...	0
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
N	F^1_N	F^2_N	F^3_N	F^4_N	...	F^m_N	0	0	0	...	1

TABLE III
INPUT DATA SET EXAMPLE AFTER ONE HOT ENCODING.

V. IMPLEMENTATION

A. Data Cleaning

No matter how the data are collected (in face to face, interviews, questionnaires, etc.), there will be some kind of error. While some of the data are correct as they return some changes in the context, others may give back measurement or entry error. These mistakes may happen due to manual entry, poorly designed recording system, or when there is no control over the data imported from external data source. The data cleaning process, varies based on the data set and analytical requirements. For example, a data scientist doing fraud detection analysis on credit card transactions may want to retain outlier values because they could be a fraudulent purchase. The aim of data cleaning is to get an accurate data set to do some analysis or prediction. So that the result will be reliable. In this model, there are many null values in every features. Those null values are dropped from the data.



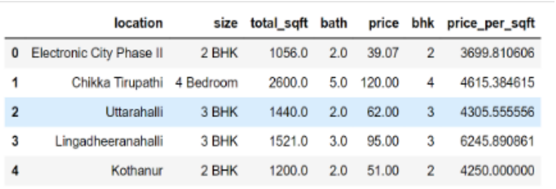
```
location         1
size            16
total_sqft       0
bath            73
price            0
dtype: int64
```

```
location       0
size           0
total_sqft     0
bath           0
price          0
dtype: int64
```

Fig. 3. Null values before and after data cleaning.

B. Feature Engineering

Feature engineering is the main task in preparing the data for machine learning. It is a process of create suitable features from the given information, which will lead to improved predictive performance. This method involves transformation application function such as arithmetic and aggregate operators on available features to generate new features. Transformation helps to make a new feature or convert a non-linear relation between a feature and a target class into linear relation, which is easier to learn. In this model we are creating three new features with the available information. One is bhk feature which has only integer values, which will be easy for prediction. Second one is changing the total square feet feature from a range to a integer or float by calculating the average of minimum and maximum in the range. With the new total square feet feature, we are adding another new feature called price per square feet with price and new total square feet feature [8].



	location	size	total_sqft	bath	price	bhk	price_per_sqft
0	Electronic City Phase II	2 BHK	1056.0	2.0	39.07	2	3699.810606
1	Chikka Tirupathi	4 Bedroom	2600.0	5.0	120.00	4	4615.384615
2	Uttarahalli	3 BHK	1440.0	2.0	62.00	3	4305.555556
3	Lingadheeranahalli	3 BHK	1521.0	3.0	95.00	3	6245.890861
4	Kothanur	2 BHK	1200.0	2.0	51.00	2	4250.000000

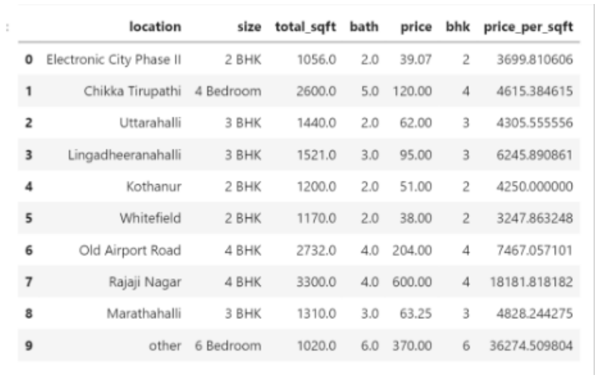
Fig. 4. Data frame after feature engineering.

C. Dimensionality Reduction

The input variables, features, or columns present in a given data is known as dimensionality. The process to reduce these features, variables or columns is called dimensionality reduction. sometimes there are more number of features which is commonly known as the curse of dimensionality, when there is more number of features, it will be tough to visualize and to predict. Most of these features match up so they are not needed. This is where dimensionality reduction technique is used, it is a way of converting the higher dimensions data set into lesser dimensions dataset making sure it provides similar information. In this dataset, there are 1304 locations in which many locations have only one sample. So locations which have lesser than 10 samples are combined together and put in one location called ‘others’ and the average of all those locations price is given as the price of others. After dimensionality reduction process the total number of locations reduced to 1293.

D. Outlier Removal

1) Using Business Logic: A dataset may occasionally contain extreme values that are dissimilar from other data and outside the expected range. These are called outliers. Machine learning models can understand the dataset more after removing these outlier values. An outlier can be identified easily by plotting a scatter plot graph of the data. In this



	location	size	total_sqft	bath	price	bhk	price_per_sqft
0	Electronic City Phase II	2 BHK	1056.0	2.0	39.07	2	3699.810606
1	Chikka Tirupathi	4 Bedroom	2600.0	5.0	120.00	4	4615.384615
2	Uttarahalli	3 BHK	1440.0	2.0	62.00	3	4305.555556
3	Lingadheeranahalli	3 BHK	1521.0	3.0	95.00	3	6245.890861
4	Kothanur	2 BHK	1200.0	2.0	51.00	2	4250.000000
5	Whitefield	2 BHK	1170.0	2.0	38.00	2	3247.863248
6	Old Airport Road	4 BHK	2732.0	4.0	204.00	4	7467.057101
7	Rajaji Nagar	4 BHK	3300.0	4.0	600.00	4	18181.818182
8	Marathahalli	3 BHK	1310.0	3.0	63.25	3	4828.244275
9	other	6 Bedroom	1020.0	6.0	370.00	6	36274.509804

Fig. 5. Data frame after dimensionatily reduction.

model, there are some samples which has many bedrooms and bathrooms, but dose not have sufficient space to have too many bathrooms or bedrooms. So, approximately we take 300 square ft per bedroom and remove the samples which do not satisfies this logic.

2) Using Standard Deviation and Mean: The standard deviation of the sample can be used as a strategy for finding outliers if we are assured that the pattern of values in the sample is distributed normally. The normal distribution has the property that the percentage of values in the sample may be accurately summarised using the standard deviation from the mean. It is said that 68% of data should lie between and one standard deviation. In this model, we are finding mean and one standard deviation for every location and removing anything that lies below mean and one standard deviation, and removing anything lies above mean and one standard deviation. In this dataset, there are many samples in which, the price of 3 bedroom apartment is lesser than 2 bedroom apartment with similar square feet. There may be many reasons for that, but this outlier can affect the accuracy of the prediction. So, for this, we will calculate the mean, standard deviation and count for one and two bedroom apartment and filter out the two bedroom apartments, who's price is less than the mean of one bedroom apartment.

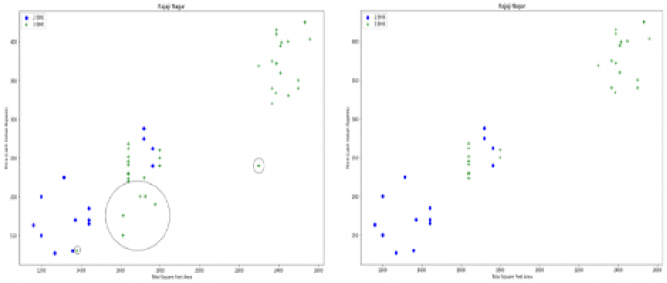


Fig. 6. Before and after outlier removal in RAJAJI NAGAR.

The majority of the price lies between 0 to 10000, the dataset is in normal distribution.

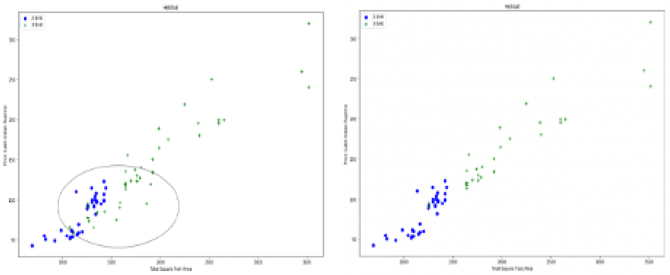


Fig. 7. Before and after outlier removal in HEBBAL.

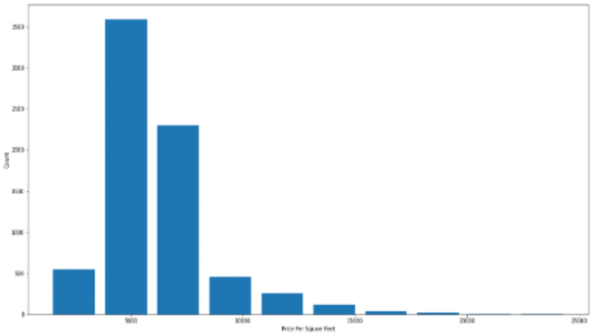


Fig. 8. Distribution of price per sqft.

In this dataset, there are some samples which have more than 10 bathrooms. If the samples have same number of bedrooms, the bathroom count is correct, it is not an outlier. So, what we do is, if the bathroom exceeds the bedroom by 2 or more numbers, those samples are considered as outliers and removed from the dataset.

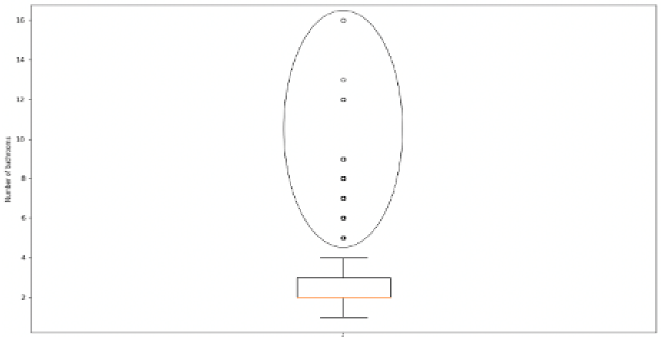


Fig. 9. Outliers in Bathroom feature

E. Train-Test Splitting

The performance of machine learning algorithms used for prediction-based applications is estimated using the train-test split technique. This approach enables us to evaluate the outcomes of our own machine learning model to those of other machines because it is a simple and quick technique to carry out. By default, 30% of the real data are included in the test set, and the remaining 70% are included in the training set. The model selection module in Scikit-learn, popularly known as sklearn, which is the most beneficial library for machine learning, has the splitter function. In this project, we divided the real data into two sets: a training set with 80% of the data and a test set with 20% of the data. The effective predictive model is fitted using the train set. Predictions are only made of the test set.

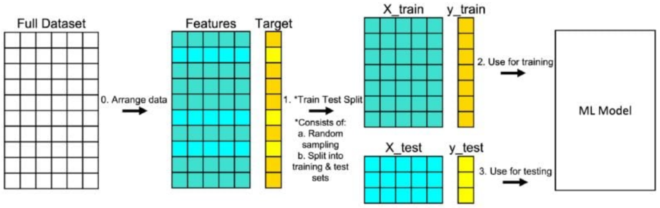


Fig. 10. Train test split procedure.

VI. Conclusion

The paper entitled "House Price Prediction Using Machine Learning" has proposed to predict house price in Bengaluru city based on various factors that impact the price such as square feet, number of bedrooms, number of bathrooms on given data. This article used several prediction techniques like linear regression, lasso, random forest, decision tree to discover the precise price of the houses, out of them linear regression model delivers the more accurateness of the price. We can see that from Table I on page 3, linear regression gives 81.83% accuracy which is the highest when compared to other regression techniques. Work is carry out using Scikit-Learn machine learning tool. It aids people to buy house in budget and lessen loss of money. our analysis we set value of RMSE as 2.9131889. In this paper, deployment using Flask and automated result file generation are currently being worked on.

VII. Future Scope

Future data can be structured with additional features like flooring type, lift and parking availability, tax on furnishings, and air quality to set it apart from existing prediction systems. We can create the price prediction model utilising sophisticated machine learning techniques and a different country's housing data set.

References

[1] A. G. Rawool, D. V. Rogye, S. G. Rane, and A. Vinayk, "House price prediction using machine learning," Int. J. Res. Appl. Sci. Eng. Technol, vol. 9, pp. 686–692, 2021.

[2] H. Yu and J. Wu, "Real estate price prediction with regression and classification," CS229 (Machine Learning) Final Project Reports, 2016.

[3] A. Y. Hussein, P. Falcarin, and A. T. Sadiq, "Enhancement performance of random forest algorithm via one hot encoding for iot ids," Periodicals of Engineering and Natural Sciences (PEN), vol. 9, no. 3, pp. 579-591, 2021.

[4] P. Durganjali and M. V. Pujitha, "House Resale Price Prediction Using Classification Algorithms," 2019 International Conference on Smart Structures and Systems (ICSSS), 2019, pp. 1-4, doi: 10.1109/ICSSS.2019.8882842.

[5] Nihar Bhagat, Ankit Mohokar and Shreyash Mane. "House Price Forecasting using Data Mining". International Journal of Computer Applications 152(2):23-26, October 2016

[6] Arshiya Shaikh,R.Vinayaki,G.Siddhanth,Y.Phanindra varma, "HOUSE PRICE PREDICTION USING MULTI VARIATE ANALYSIS", International Journal of Creative Research Thoughts (IJCRT), ISSN:2320-2882, Volume.8, Issue 2, pp.1676-1679, February 2020

[7] Priya, G.. (2021). "House Price Prediction using Machine Learning Techniques". International Journal for Research in Applied Science and Engineering Technology. 9. 3645-3650. 10.22214/ijraset.2021.35831.

[8] Nargesian, Fatemeh & Samulowitz, Horst & Khurana, Udayan & Khalil, Elias Turaga, Deepak. (2017). "Learning Feature Engineering for Classification". 2529-2535. 10.24963/ijcai.2017/352.

[9] Van Der Maaten, L., Postma, E. and Van den Herik, J., 2009. "Dimensionality reduction: a comparative". J Mach Learn Res, 10(66-71), p.13.

[10] F. Wang, Y. Zou, H. Zhang and H. Shi, "House Price Prediction Approach based on Deep Learning and ARIMA Model," 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT), 2019, pp. 303-307, doi: 10.1109/ICCSNT47585.2019.8962443.

[11] A. G. Sarip, M. B. Hafez, and M. N. Daud, "Application Of Fuzzy Regression Model For Real Estate Price Prediction", MJCS, vol. 29, no. 1, pp. 15–27, Mar. 2016.

[12] Truong, Quang, et al. "Housing price prediction via improved machine learning techniques." Procedia Computer Science 174 (2020): 433-442.

[13] Thamarai, M., and S. P. Malarvizhi. "House Price Prediction Modeling Using Machine Learning." International Journal of Information Engineering Electronic Business 12.2 (2020).

[14] Satish, G. Naga, Ch V. Raghavendran, MD Sugnana Rao, and Ch Srinivasulu. "House price prediction using machine learning." Journal of Innovative Technology and Exploring Engineering 8, no. 9 (2019): 717-722.

[15] Ravikumar, Aswin Sivam. "Real estate price prediction using machine learning". Diss. Dublin, National College of Ireland, 2017.

[16] Ng, Aaron, and Marc Deisenroth. "Machine learning for a London housing price prediction mobile application." Imperial College London (2015).