Project

On

House Price Prediction using Machine Learning

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

Master of Computer Applications

By

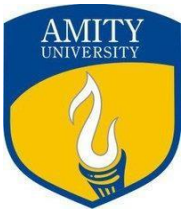
Akshay Sharma

Enroll no. A50500718003

Under the guidance of

Dr. Vikas Thada

Associate Professor



Amity Institute of Information Technology
Amity University Haryana

Oct 2020

Amity Institute of Information Technology

Amity University Haryana

DECLARATION

I, **Akshay Sharma** student of MCA (MASTER OF COMPUTER APPLICATIONS) hereby declare that the report entitled **"House Price Prediction using Machine Learning"** which is submitted to, Amity Institute of Information Technology, Amity University Haryana, in partial fulfillment of the requirement for the award of the degree of Masters of Computer Applications, has not been previously formed the basis for the award of any degree, diploma or other similar title or recognition.

Akshay Sharma
A50500718003
2018-2021

Amity Institute of Information Technology

Amity University Haryana

Certificate

This is to certify that the work in the report entitled **"House Price Prediction using Machine Learning"** by **Akshay Sharma** bearing **A50500718003** is a bonafide record of term paper carried out by him under my supervision and guidance in partial fulfilment of the requirements for the award of the degree of **Master of Computer Applications** in the Department of Computer Applications, Amity Institute of Information Technology, Amity University Haryana. Neither this term paper nor any part of it has been submitted for any degree or academic award elsewhere.

Date: / /2020

Signature

Dr. Vikas Thada

Associate Professor

Head

Department of Computer Science & Engineering

Amity School of Engineering and Technology

Amity University Haryana, Gurgaon

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my supervisor, Dr. Vikas Thada, Associate Professor Department of CSE, Amity Institute of Information Technology, Gurugram, for their consistent guidance and unwavering support, collegiality and mentorship throughout this project. Without their thoughtful reassurance and careful supervision, this thesis would never have taken shape. Their encouragement and belief in me kept me enthusiastic and helped me perceive the work with new perception striving for excellence

I am thankful to my family whose value to me only grows with each passing day. They have helped me fight the highs and the lows and given me the strength to preserver especially through the times when the project seemed to be a mammoth task.

My earnest gratitude to the administration and staff of Amity Institute of Information Technology and all the other contributors to the study who have directly or indirectly supported and helped me.

Last but not the least I thank God Almighty.

Signature of the Student

ABSTRACT

Machine Learning is application of artificial intelligence (AI) that provides system the ability to automatically learn and improve from experience without being explicitly programmed. Machine Learning focuses on the development of computer programs that can access data and use it learn for themselves.

In this project we going to use Jupyter Notebook. Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

Using Jupyter Notebook and python language we Aim to develop this project. The aim of this project is very practical and depends on today's problem faced by peoples. In this project we are taking dataset of Bangalore house and try to make a reliable system through which we can predict the data of house with the help of given data.

***Keywords**: Machine Learning, Python, Jupyter Notebook.*

List of Figures

List of Tables

TABLE OF CONTENTS

Name **Page no.**

Chapter 1

INTRODUCTION

1.1 Machine Learning

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly. But, using the classic algorithms of machine learning, text is considered as a sequence of keywords; instead, an approach based on semantic analysis mimics the human ability to understand the meaning of a text.[2]

Machine learning algorithms are often categorized as supervised, unsupervised or Reinforcement learning.

- **Supervised machine learning**: - It is applied when we have dataset and we try to learn from it and predict according to that. examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.

- **Unsupervised machine learning**: - It is used when the information used to train is neither classified nor labeled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data. The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data.

- **Reinforcement machine learning: -** It is a learning method that interacts with its environment by producing actions and discovers errors or rewards. Trial and

error search and delayed reward are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behavior within a specific context in order to maximize its performance. Simple reward feedback is required for the agent to learn which action is best; this is known as the reinforcement signal.[2]

1.2 Prediction in Machine Learning

"Prediction" refers to the output of an algorithm after it has been trained on a historical dataset and applied to new data when forecasting the likelihood of a particular outcome, such as whether or not a customer will churn in 30 days. The algorithm will generate probable values for an unknown variable for each record in the new data, allowing the model builder to identify what that value will most likely be.

The word "prediction" can be misleading. In some cases, it really does mean that you are predicting a future outcome, such as when you're using machine learning to determine the next best action in a marketing campaign. Other times, though, the "prediction" has to do with, for example, whether or not a transaction that already occurred was fraudulent. In that case, the transaction already happened, but you're making an educated guess about whether or not it was legitimate, allowing you to take the appropriate action.[3]

1.3 Dataset

A data set is an ordered collection of data. While handling the data, the data set can be a bunch of tables, schema and other objects. The data are essentially organized to a certain model that helps to process the needed information. The set of data is any permanently saved collection of information which usually contains either case-level, gathered data, or statistical guidance level data. It is important to have good grasp of input data and the various terminology used when describing data. The training data set and test data set are used for different purposes during project and the success of a project depends a lot on them.[4]

- The training data set is the one used to train an algorithm to understand how to apply concepts such as neural networks, to learn and produce results. It includes both input data and the expected output. Training sets make up the majority of the total data, around 80 %. In testing, the models are fit to parameters in a process that is known as adjusting weights.

- The test data set is used to evaluate how well your algorithm was trained with the training data set. In AI projects, we can't use the training data set in the testing stage because the algorithm will already know in advance the expected output which is not our goal. Testing sets represent 20% of the data. The test set is ensured to be the input data grouped together with verified correct outputs, generally by human verification.

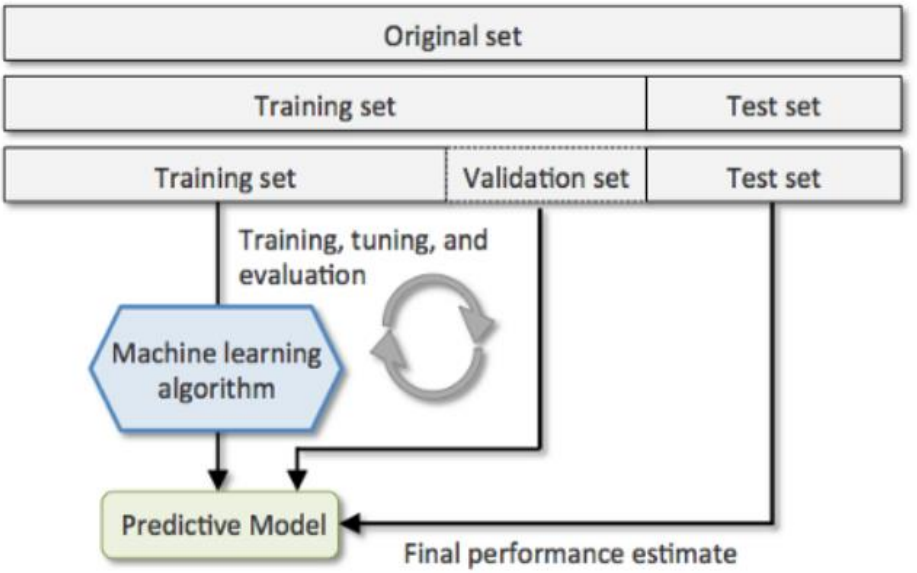


Fig 1: - Cycle of data conversion[4]

Chapter 2

Background Study

2.1 Different techniques given in Research Papers

There are different research papers with different techniques shown in the given table. Every regression problem shows different results and the way of using regression algorithms gives different accuracy rates.

Table 1: - Techniques given in different research papers[5],[6],[7]

S.no.	Paper Title	Algorithm Used	Accuracy Rate
1	House Price Prediction using Machine Learning	Lasso Regression	76.14
		Gradient Boosting Regression	91.27
2	House Price Forecasting Using Machine Learning	Decision tree Regression	89
3	Real Estate Price Prediction Using Machine Learning	Random Forecast	90
		Bagging	70

2.2 Analyzing Market

We are living in a world where nowadays everything is available under our fingers with just one click, and we can have different applications for our daily needs and finding the best home or apartment is one of the main needs of any human being. But at many points sellers and buyers didn't get the correct amount of their property. In India there are only two websites that provide a system for property evaluation but their system fails

in many areas as data is not found or they have estimation of value.

2.2.1 Magicbricks.com

Magic bricks is one the leading websites that provides a common platform for property buyers & sellers to locate properties of interest in India, and source information about all property related issues.

Magic bricks have an option for price estimation of property but at some point it doesn't show the results whereas in some properties it shows high property value. As in the given figure it is estimating the cost of property in Tilak Nagar, New Delhi between 27 lac to 33 lac and try to estimate cost for next 6 months.

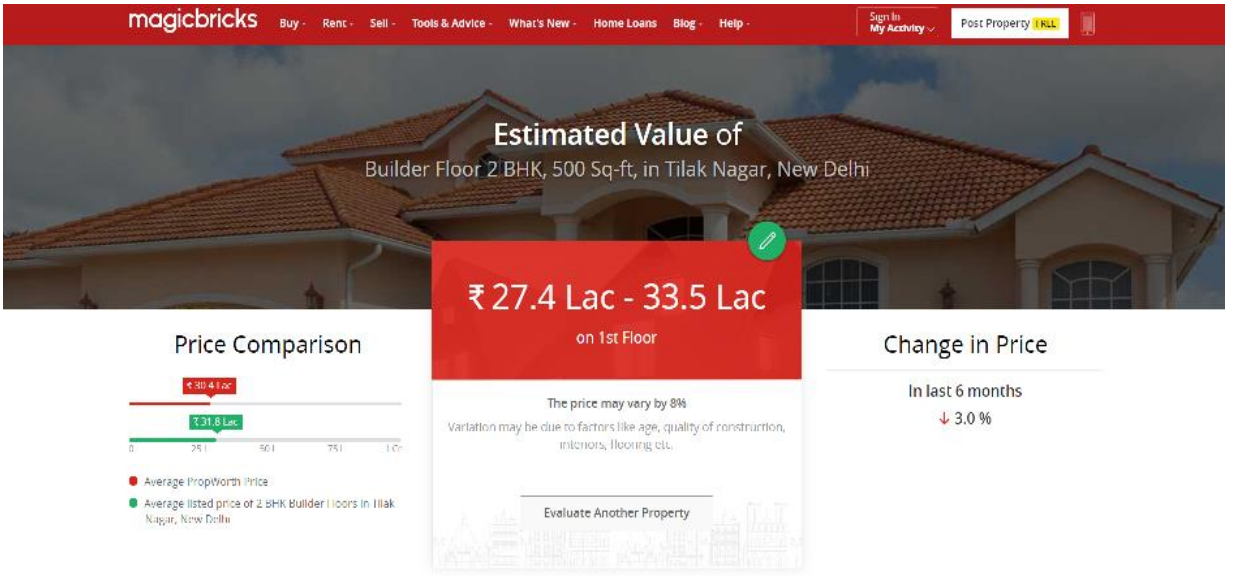


Fig 3: - Price Prediction Algorithm of Magic Bricks[9]

2.2.2 99acres.com

99acres.com is India's No.1 Property portal where It provides its services all over India, they can help the customers to explore real estate experience at different levels. This site also provides property estimation options but its system is not robust as Magic Bricks. I have entered the same details in both the sites to check the reliability of the website but unable to compare. 99acres dataset is very less as compared to Magic Bricks.

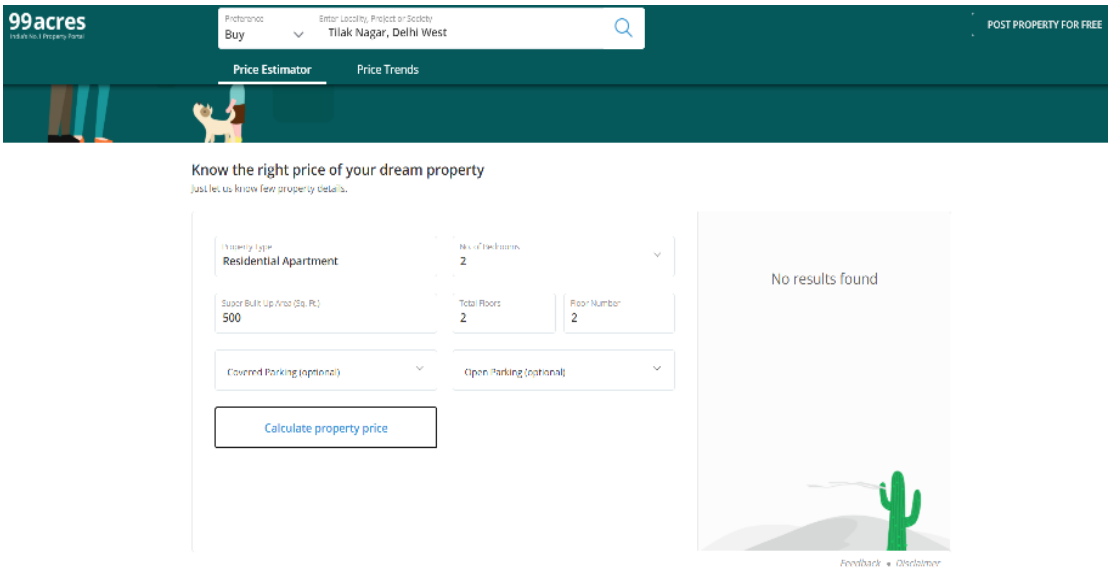


Fig 3: - Price Prediction Algorithm of 99acres[10]

2.3 Objectives

Our objective is to build a Machine Learning model for Predicting House Price in Bangalore.

Chapter 3

Tools and Techniques

3.1 Python Programming Language

Python is a high level, interpreted, general purpose programming language, created by Guido van Rossum and first released in 1991. [1] Python supports multiple programming paradigms such as (particularly, procedural), object-oriented, and functional programming. For the first time python was conceived in the 1980s and later python 2.0, released in 2000, introduced features like list comprehensions and a garbage collection system with reference counting. Python 3.0, released in 2008 was a major revision where most of the python 2 code does not run unmodified on python 3.

3.2 NumPy and Pandas

Similar to NumPy, Pandas is one of the most widely used python libraries in data science. It provides high-performance, easy to use structures and data analysis tools. Unlike NumPy library which provides objects for multi-dimensional arrays, Pandas provides an in-memory 2d table object called Data frame. It is like a spreadsheet with column names and row labels.

Hence, with 2d tables, pandas are capable of providing many additional functionalities like creating pivot tables, computing columns based on other columns and plotting graphs. Pandas can be imported into Python using:

“Import pandas as pd”

Data frames can also be easily exported and imported from CSV, Excel, JSON, HTML and SQL database. Some other essential methods that are present in data frames are:

- **head ():** returns the top 5 rows in the data frame object
- **tail ():** returns the bottom 5 rows in the data frame
- **info ():** prints the summary of the data frame
- **describe ():** gives a nice overview of the main aggregated values over each column

3.3 Matplotlib

Matplotlib is a 2d plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments. Matplotlib can be used in Python scripts, Python and IPython shell, Jupyter Notebook, web application servers and GUI toolkits. The matplotlib Python library, developed by John Hunter and many other contributors, is used to create high-quality graphs, charts, and figures. The library is extensive and capable of changing very minute details of a figure.

3.4 Scikit-learn

Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python. It is licensed under a permissive simplified BSD license and is distributed under many Linux distributions, encouraging academic and commercial use. Extensions or modules for SciPy are conventionally named Scikit. As such, the module provides learning algorithms and is named scikit-learn.

The vision for the library is a level of robustness and support required for use in production systems. This means a deep focus on concerns such as ease of use, code quality, collaboration, documentation and performance. Although the interface is Python, c-libraries are leveraged for performance such as NumPy for arrays and matrix operations, LAPACK, LibSVM and the careful use of cython.

3.5 Jupyter Notebook

The Jupyter Notebook is an open source web application that you can use to create and share documents that contain live code, equations, visualizations, and text. Jupyter Notebook is maintained by the people at Project Jupyter. Jupyter Notebooks are a spin-off project from the IPython project, which used to have an IPython Notebook project itself. The name, Jupyter, comes from the core supported programming languages that it supports: Julia, Python, and R. Jupyter ships with the IPython kernel, which allows you to write your programs in Python, but there are currently over 100 other kernels that you can also use

The next most popular distribution of Python is Anaconda. Anaconda has its own

installer tool called Anaconda that you could use for installing a third-party package. However, Anaconda comes with many scientific libraries preinstalled, including the Jupyter Notebook, so you don’t actually need to do anything other than install Anaconda itself.

Chapter 4

Application Requirements

We need to install the Anaconda Navigator in our system if we want to run without using the internet else we can use Google Colab to run this software online.

Anaconda License: Free use and redistribution under the terms of the End User License Agreement - Anaconda® Individual Edition.

4.1 System Architecture

- Windows- 64-bit x86, 32-bit x86
- MacOS- 64-bit x86
- Linux- 64-bit x86, 64-bit Power8/Power9.
- Minimum 5 GB disk space to download and install.

4.2 Operating System Requirement

Windows 8 or newer, 64-bit macOS 10.13+, or Linux, including Ubuntu, RedHat, CentOS 6+, and others.

If your operating system is older than what is currently supported, you can find older versions of the Anaconda installers in our archive that might work for you. See Using Anaconda on older operating systems for version recommendations.

4.3 Google Colab

Colaboratory, or "Colab" for short, allows you to write and execute Python in your browser, whether you're a student, a data scientist or an AI researcher, Colab can make work easier.

- Zero configuration required
- Free access to GPUs
- Easy sharing

Chapter 5

Project Implementation

5.1 Dataset

We have used Kaggle.com for a dataset where **"Bangalore House Price prediction"** dataset is downloaded. This dataset includes area type, Availability, Location, BHK, Bathroom, Society, Balcony, Total Sqft, Price. Each column of the dataset carries 13K data. We are going to predict the price of a house on the basis of location, BHK, Total Sqft. For our analysis, we divide our dataset into two parts training and testing datasets (0.8/0.2).

5.2 Libraries used

Using the Python programming language, we have imported Pandas, Numpy, Seaborn, matplotlib libraries.

5.3 Data Cleaning

This dataset needed some cleanings and modification. Besides, some feature representation should be done. At the initial stage we have area_type, availability, location, size, total_sqft, bath, balcony and price.

- We have dropped 'area_type', 'society', 'balcony', 'availability ' columns as considering that these points will not affect the price detection program.
- Now dropping null values given in the dataset leads to loss of 80 data.
- Some of the features in the dataset are self-reported and they are not the same across subjects. For example, BHK could be either "2 BHK" or "8 Bedroom" or "1 RK". These discrepancies could be fixed manually by shifting the values to the new column of BHK. Now we can see the dataset of BHK.

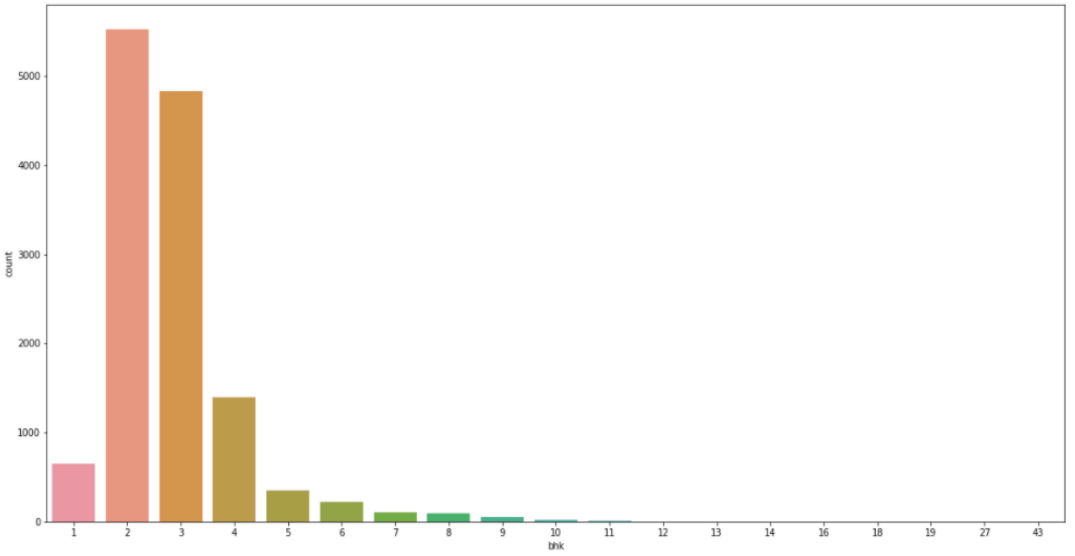


Fig 4: - No. of BHK

- In case of total square feet some places show range in the dataset to eliminate the range we have taken mean of the given dataset.

	location	size	total_sqft	bath	price	bhk
30	Yelahanka	4 BHK	2100 - 2850	4.0	186.000	4
122	Hebbal	4 BHK	3067 - 8156	4.0	477.000	4
137	8th Phase JP Nagar	2 BHK	1042 - 1105	2.0	54.005	2
165	Sarjapur	2 BHK	1145 - 1340	2.0	43.490	2
188	KR Puram	2 BHK	1015 - 1540	2.0	56.800	2
410	Kengeri	1 BHK	34.46Sq. Meter	1.0	18.500	1
549	Hennur Road	2 BHK	1195 - 1440	2.0	63.770	2
648	Arekere	9 Bedroom	4125Perch	9.0	265.000	9
661	Yelahanka	2 BHK	1120 - 1145	2.0	48.130	2
672	Bettahalsoor	4 Bedroom	3090 - 5002	4.0	445.000	4

Fig 5: -Range dataset of Square Feet

- There is a high dimensional problem in location. we have 1293 unique locations and 1052 locations are less than 10. so, in this we have marked “others” as a location that are having less than 10.

5.4 Outlier Detection and Removal

While analyzing the data we have to see all the aspects of the data so that faulty data should be removed. The fact is that to build 1BHK house we required a minimum 300

Square feet area. So, if we look at total_sqft and BHK they are directly dependent on each other. After analyzing we came to know that there are few datasets that don't fulfill these criteria and acting as faulty dataset. To overcome this problem, we negate this type of dataset as the result of it we left with (12502,7).

5.5 Training and Testing

The observations in the training set form the experience that the algorithm uses to learn. In supervised learning problems, each observation consists of an observed output variable and one or more observed input variables. The test set is a set of observations used to evaluate the performance of the model using some performance metric. It is important that no observations from the training set are included in the test set. If the test set does contain examples from the training set, it will be difficult to assess whether the algorithm has learned to generalize from the training set or has simply memorized it.

For training and testing we have used Linear Regression to see the outcome. We have divided the dataset into 2 parts for training and testing (80:20) to evaluate the price of the house. If we see the result of the training set result it goes to 80+ %.

5.6 Algorithms Used

In this project we have tested linear regression, lasso and decision tree testing and try to find which regression testing is more suitable for this model. In the given list of regression testing. Linear regression testing performs best with 81.83% accuracy. (as shown in the given table).

Table 2: - Testing Algorithms Scoreboard

Serial no.	Model	Best Score
1	Linear Regression	81.8354
2	Lasso	68.7464
3	Decision Tree	72.8360

.

5.7 Result After Analysis

To find the price of the house we have to pass location, Square feet, BHK and Bathroom in the statement. As we can see in the given picture it gives the cost in lakhs.

```
predict_price('1st Phase JP Nagar',1000, 2, 2)
```

```
83.49904677167721
```

```
predict_price('Vishwapriya Layout',500, 3, 3)
```

```
8.959304741499203
```

```
predict_price('Yelahanka New Town',1500, 5, 5)
```

```
106.63788461103364
```

Fig 6: - Price of houses in Lakhs

Chapter 6
CONCLUSION AND FUTURE SCOPE

6.1 Conclusion

This research aimed to build a system that can predict Price of houses in Bangalore. The importance of this type of project is to help the buyers and sellers so that they can deal in the correct price. Firstly, we clean data and different methods. After that we try to apply 3 regression models. Median of Means Estimate seemed to have a better prediction. Later we use Linear regression and predict the price of different houses. These prediction models need to achieve high accuracy so the sample data is divided into 80% training data and 20% test data respectively.

6.2 Future Scope

There is a lot of variables affects house prices. If data are available with a lot specification we can introduce more features, for example balcony, security, interior design, how old property and other details then we can more accurate result in this field. This application should be free for all from a reliable source.

REFERENCES

[1] https://www.kaggle.com/aniketyadav1/bangalore

[2] https://expertsystem.com/machine-learning-definition/

[3]https://www.datarobot.com/wiki/prediction/#:~:text=What%20does%20Prediction%20mean%20in,will%20churn%20in%2030%20days.

[4] https://towardsdatascience.com/how-to-build-a-data-set-for-your-machine-learning-project-5b3b871881ac#:~:text=A%20data%20set%20is%20a%20collection%20of%20data.&text=In%20Machine%20Learning%20projects%2C%20we,model%20for%20performing%20various%20actions.

[5] International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-9, July 2019

[6]https://poseidon01.ssrn.com/delivery.php?ID=672090112009085083108120089021081076109025046003043075006116078007098121064109119095098106127035013015098002064070122102030104051055086041049116124078026081012094113036087084028003121103114097000107069115118121114031022070015000087106064098127028001117&EXT=pdf

[7] http://trap.ncirl.ie/3096/1/aswinsivamravikumar.pdf

[8] https://colab.research.google.com/noteboo

[9]https://www.magicbricks.com/propworth/New-Delhi/

[10] https://www.99acres.com/price-estimator