HOUSE PRICE PREDICTION

A Report

Submitted in partial fulfillment of the requirements for the Final Year Project of HOUSE PRICE PRIDICTION USING MACHINE LEARNING

Bachelor of Computer Application
By

Gaurav Soni (Roll no.: 0204696 Sec: B)

Gyanendra Kumar (Roll no.: 0204698 Sec: B)

Harshit Shukla (Roll no.: 0204708 Sec: B)

Under Guidance Of

Dr. **Mayur Rahul Sir**

Department Of Computer Application

UIET

CSJM UNIVERSITY Kanpur

2019-2022

UIET

CSJM university Kanpur

Department of Computer Applications

CERTIFICATE

Date:__/__/ 2022

This is to certify that the report of project work entitled ***HOUSE PRICE PRIDICTION USING MACHINE LEARNING*** *is being submitted by* **Gaurav Soni (Roll no.: 0204696 Sec: B), Gyanendra Kumar (Roll no.: 0204698 Sec: B) and Harshit Shukla (Roll no.: 0204708 Sec: B)***in partial fulfillment of the requirements for the Final Year Project of (department name) to the University Institute of Engineering & Technology, CSJM University, Kanpur during the academic year 2019-22 is a record of bona fide work carried out by him under our guidance and supervision .*

The results embodied in this report have not been submitted by the student(s) to any other University or Institution for the award of any degree or diploma.

Internal Guide	**Head of Department**
DR. MAYUR RAHUL SIR	**AMIT VIRMANI SIR**

ACKNOWLEDGEMENT

I express my deep sense of gratitude to my **Director (UIET IV)** for the valuable guidance and for permitting us to carry out this project.

With gratitude,

Name : ***GAURAV SONI*****(Roll no.: 0204696 Sec: B)**

Name : GYANENDRA KUMAR(Roll no.: 0204698 Sec: B)

Name : HARSHIT SHUKLA(Roll no.: 0204708 Sec: B)

CONTENTS

- 11. Database Requirement
- 12. Software Requirment Specification
- 13. Tables And Fields For Database
- 14. Database Dictionary
- 15. E-R Diagram
- 16. Screen Shots
- 17. Validation Checks
- 18. Testing
- 19. Implementation and Maintenance
- 20.Future Scope Of The Project
- 21.Conclusion
- 22. Bibliography

ABSTRACT

•House prices increase every year, so there is a need for a system to predict house prices in the future.

•Predicting House Prices with real factors.

•We aim to make evaluations based on every basic parameter that is considered while determining the price.

•It Is a Machine Learning model which integrates Data Science and Web Development.

•The goal of this project is to learn Python and get experience in Data Analytics, Machine Learning, and AI.

INTRODUCTION

Machine learning is a subfield of Artificial Intelligence (AI) that works with algorithms and technologies to extract useful information from big data.

There are two types of machine learning. These are following.

1. Supervised learning
2. Unsupervised learning

Supervised learning: Supervised learning is a type of machine learning method in which we provide sample labeled data to the machine learning system in order to train it, and on that basis, it predicts the output.

Supervised learning can be grouped further in two categories of algorithms:

- Classification
- Regression

Unsupervised learning: In Unsupervised Learning, the machine uses unlabeled data and learns on itself without any supervision. The machine tries to find a pattern in the unlabeled data and gives a response.

Unsupervised learning can be further grouped into types:

- Clustering
- Association

We use supervised learning to predict our house price.

OBJECTIVE

There are following objective of House PricePrediction Machine Learning. These are

- Predict the sale price for each house.
- Minimize the difference between predicted and actual rating.

SCOPE

It is a website at which you have to enter required input to get price.

PROJECT PLAN

1. **Data Collection:** We take appropriate database from online data repository site for our House Prediction project process.
2. **Data Analysis:** We analyze the data by making graphs, checking missing data points and removing outliers. Following are step to analyze data.
3. **We use different regression model to know which model predict best price.**

- **Linear Regression:** Linear regression is a supervised learning technique. It is responsible for predicting the value of a dependent variable (Y) based on a given independent variable (X). It is the connection between the input (X) and the output (Y). It is one of the most well-known and well-understood machine learning algorithms. Simple linear regression, ordinary least squares, Gradient Descent, and Regularization are the linear regression models.

- **Decision Tree Regression:** It is an object that trains a tree-structured model to predict data in the future in order to provide meaning ful continuous output. The core principles of decision trees, Maximizing Information Gain, Classification trees, and Regression trees are the processes involved in decision tree regression. The essential notion of decision trees is that they are built via recursive partitioning. Each node can be divided into child nodes, beginning with the root node, which is known as the parent node. These nodes have the potential to become the parent nodes of their resulting offspring nodes. The nodes at the informative features are specified as the maximizing information gain, to establish an objective function that is to optimize the tree learning method.
- **Random Forest Regression** It is an essential learning approach for classification and regression to create a large number of decision trees. Preliminaries of decision trees are common approaches for a variety of machine learning problems. Tree learning is required for serving n off the self-produce for data mining since it is invariant despite scaling and several other changes. The trees are grown very deep in order to learn a high regular pattern. Random forest is a method of averaging several deep decision trees trained on various portions of the same training set. This comes at the price of a slight increase in bias and some interoperability

3. **Performance evaluation** : We evaluate performance using following method.

- **Mean Square Error** : Mean Square Error is a measure for how close the estimation is relative to the actual data. It measures the average of the square of the errors deviation of the estimated values with respect to the actual values. It is measured by:

  $\mathrm{MSE} = (1\ /n)\ \sum^{n}_{i=1}\ (y1-y)\ ^{2}$

  Where y1 is the estimated value from the regression and y is the actual value. The lower the MSE, the better the estimation model.

- **Root Mean Squared Error**

  Root mean squared error (RMSE) is the square root of the mean of the square of all of the error. The use of RMSE is very common, and it is considered an excellent general-purpose error metric for numerical predictions.

  $\mathrm{RMSE}=(1/n(\sum_{i=1}^{n}(S_{i}-O_{i})^{2}))^{1/2}$

  where O_i are the observations, S_i predicted values of a variable, and n the number of observations available for analysis. RMSE is a good measure of accuracy, but only to compare prediction errors of different models or model configurations for a particular variable and not between variables, as it is scale-dependent.

The Existing System

The present system is not dunce proof and has certain drawbacks. Being a manual system the possible limitations and loopholes in the present system is large. Some of them are:-

- **HUMAN resource: -** The current system has too much manual work from filling a form to filing a document, delivering manifesto. This increases burden on workers but does not yield the results it should.
- **THORNY Job: -** In current system if any modification is to be made it increases manual work and is error prone.
- **ERROR: -** As the system is managed and maintained by workers errors are some of the possibilities.

Current Problems

The Existing System cannot predict the exact price of house that is why sometime many company faces loss.

Areas For Improvement

There are following areas for improvement.

- Minimize the difference between predicted and actual rating.
- Design such model to predict best price of house .

Proposed System

- Proposed system is an website.
- Proposed system uses the parameters such as BHK_NO, size of the property, location and other parameters for house price prediction.
- Proposed system is built using visual studio as front-end technology.

Input/Output Requirement

Input:

- Rera(**Real Estate Regulatory Authority**) approved or Not
- BHK _No
- Total Square Feet - size of the property in square feet
- Address of the property
- Under Construction or Not

Output:

- House price prediction using ML supervised learning technique.

Hardware And Software Requirement

- **Hardware Requirement**:

  A computer system or laptop with internet and wifi facilitiy

  Operating system: Window 8 or greater

- **Software Requirement** :

We use following software to develop our project

- Python interpreter
- Visual Studio/Jupyter Notebook
- A web browser such as chrome

Database Requirement

Dataset should contain large number of entities so that it will increase the accuracy of the predicted price and suggest a better property.

Software Requirment Specification

1. **Functional Requirements:** A Functional Requirement is a description of the service that the software must offer. It describes a software system or its component. A function is nothing but inputs to the software system, its behavior, and outputs.

- **USER INTERFACE**: The user interface will be a website. The user has to enter all the attributes correctly and in the required format.
- **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.
- **DATABASE**: Dataset should contain large number of entities so that it will increase the accuracy of the predicted price and suggest a better property.

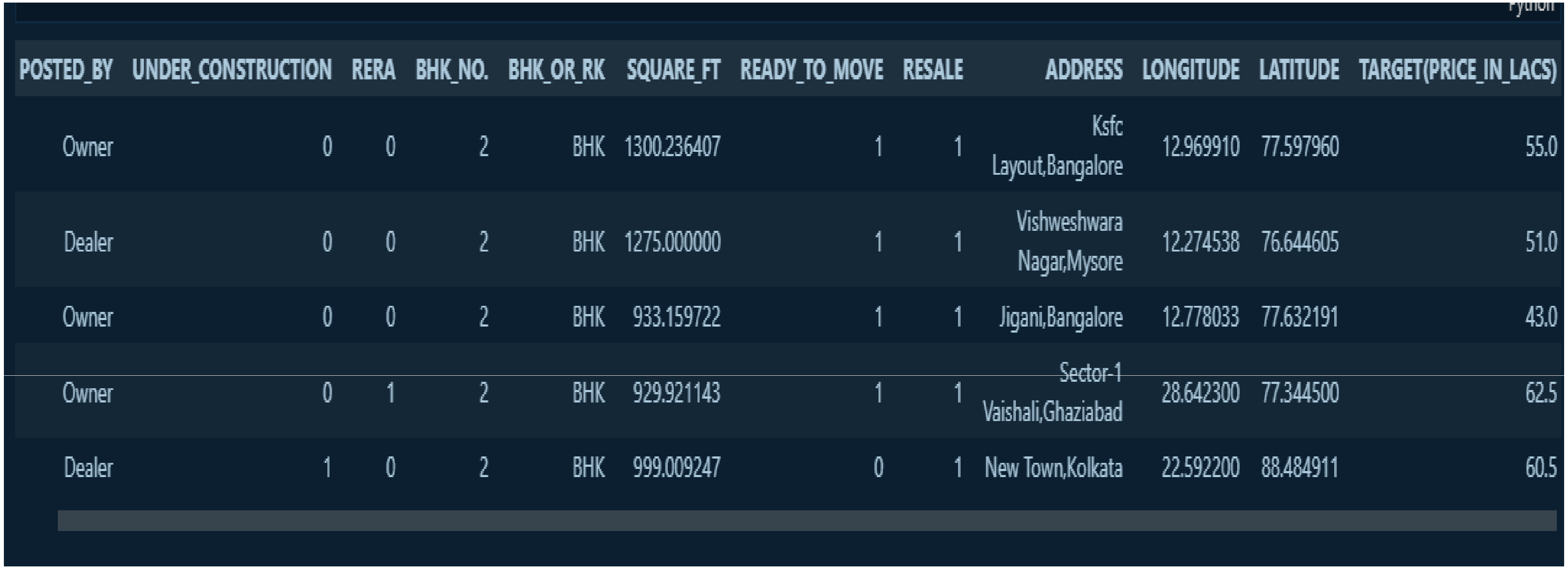
2. <u>Non-Functional Requirements</u>: Non Functional Requirements are the constraints or the requirements imposed on the system. They specify the quality attribute of the software. Non-Functional Requirements deal with issues like scalability, maintainability, performance, portability, security, reliability, and many more

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.

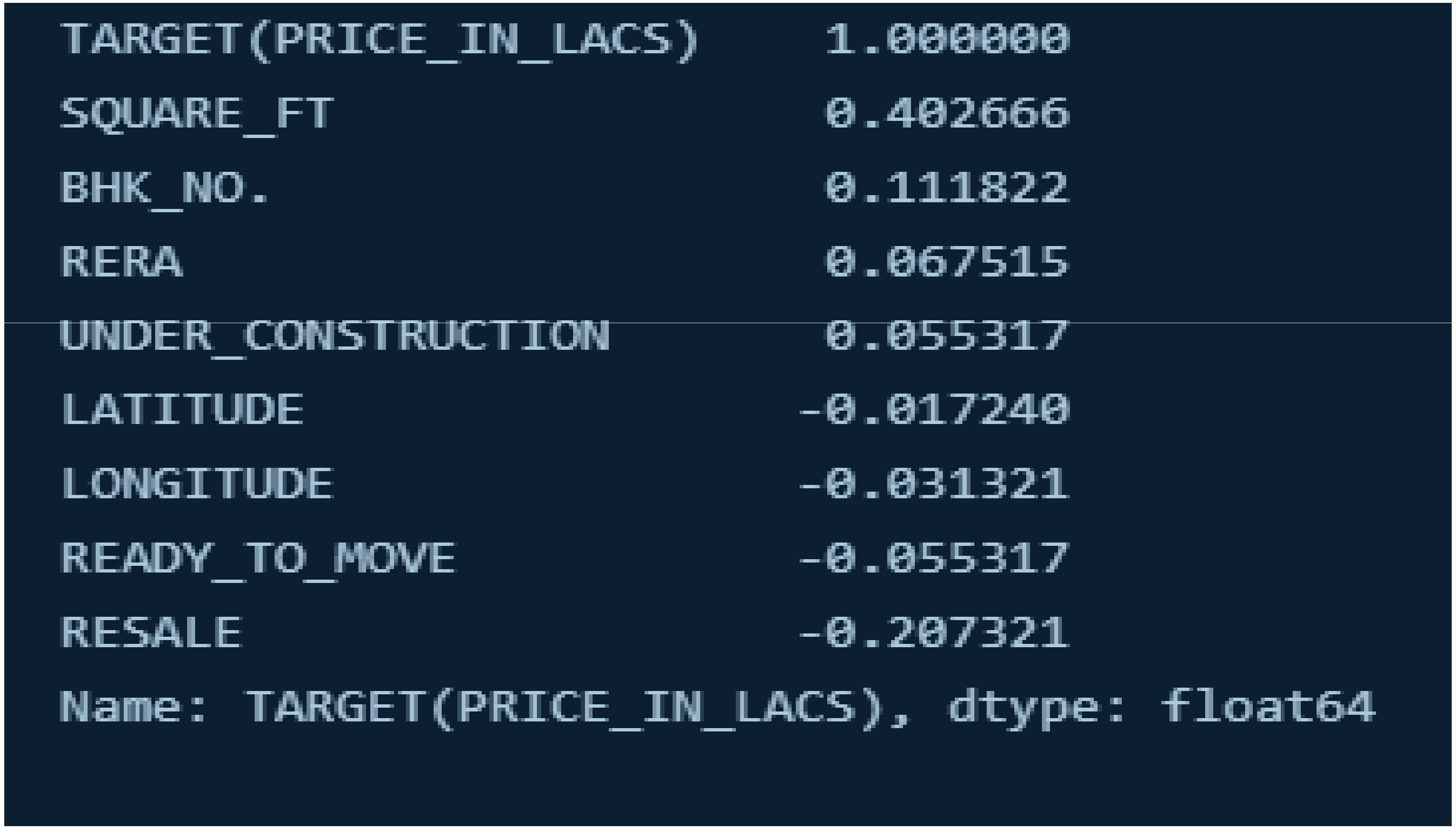
Tables And Fields For Database



	POSTED_BY	UNDER_CONSTRUCTION	RERA	BHK_NO.	BHK_OR_RK	SQUARE_FT	READY_TO_MOVE	RESALE	ADDRESS	LONGITUDE	LATITUDE	TARGET(PRICE_IN_LACS)
	Owner	0	0	2	BHK	1300.236407	1	1	Ksfc Layout,Bangalore	12.969910	77.597960	55.0
	Dealer	0	0	2	BHK	1275.000000	1	1	Vishweshwara Nagar,Mysore	12.274538	76.644605	51.0
	Owner	0	0	2	BHK	933.159722	1	1	Jigani,Bangalore	12.778033	77.632191	43.0
	Owner	0	1	2	BHK	929.921143	1	1	Sector-1 Vaishali,Ghaziabad	28.642300	77.344500	62.5
	Dealer	1	0	2	BHK	999.009247	0	1	New Town,Kolkata	22.592200	88.484911	60.5

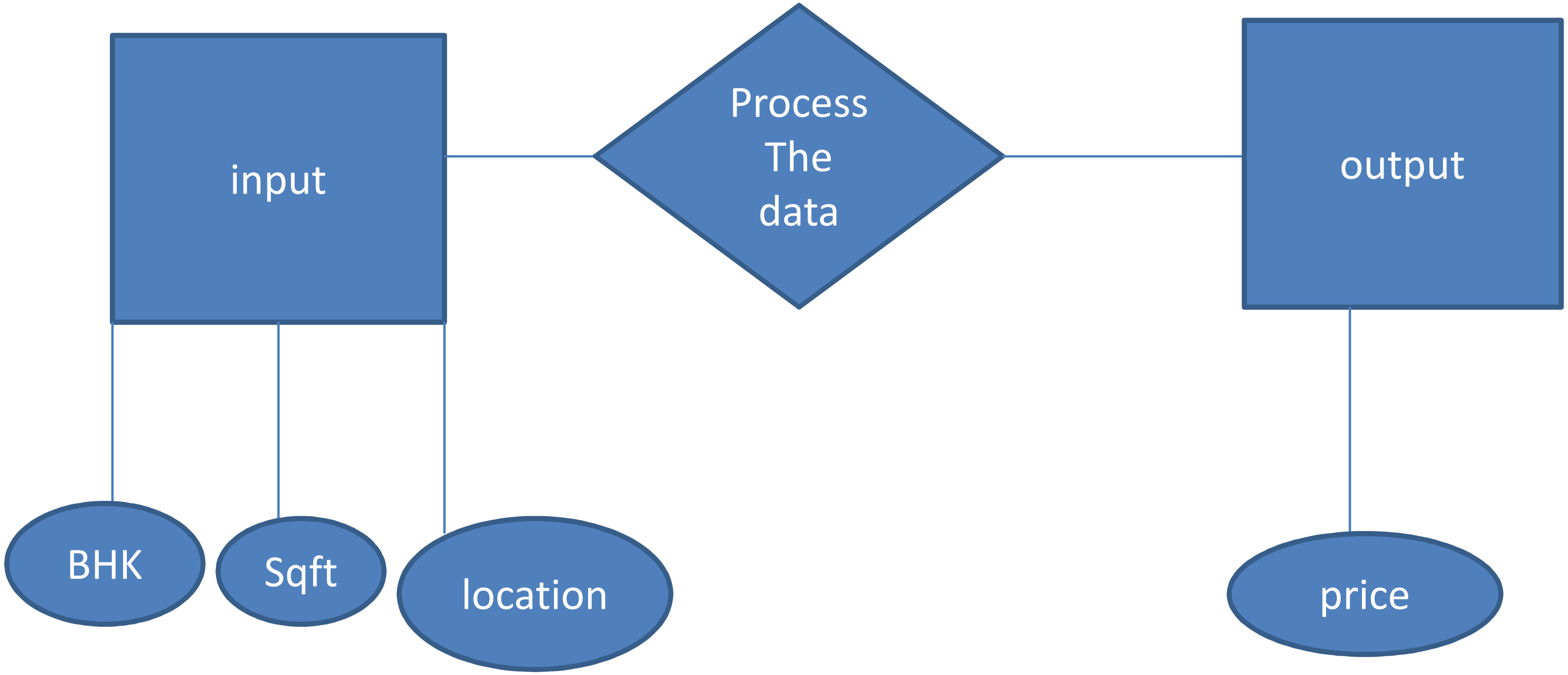
Database Dictionary

COLUMN	DATA TYPE	SIZE OF DATA
POSTED_BY	Object(string)	29460
UNDER_CONSTRUCTION	int	29460
RERA	int	29460
BHK_NO	int	29460
SQUARE_FT	float	29460
READY_TO_MOVE	int	29460
RESALE	int	29460
ADDRESS	Object(string)	29460
LONGITUDE	float	29460
LANGITUDE	float	29460
TARGET(PRICE IN LACS)	float	29460

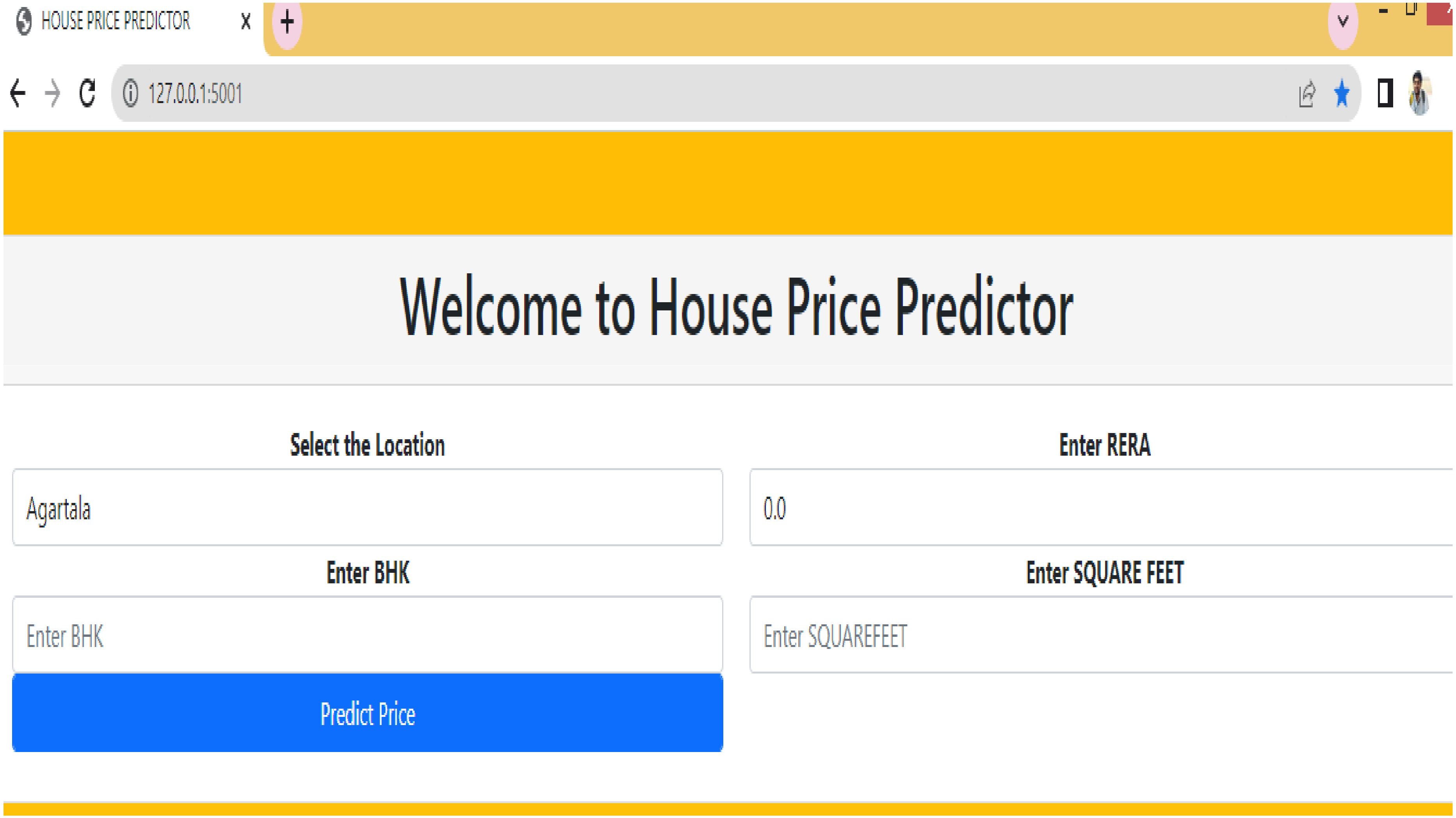


```
TARGET(PRICE_IN_LACS)    1.000000
SQUARE_FT                0.402666
BHK_NO.                  0.111822
RERA                     0.067515
UNDER_CONSTRUCTION       0.055317
LATITUDE                -0.017240
LONGITUDE               -0.031321
READY_TO_MOVE           -0.055317
RESALE                  -0.207321
Name: TARGET(PRICE_IN_LACS), dtype: float64
```

E-R Diagram



Screen Shots



Welcome to House Price Predictor

Select the Location

Agartala

Enter RERA

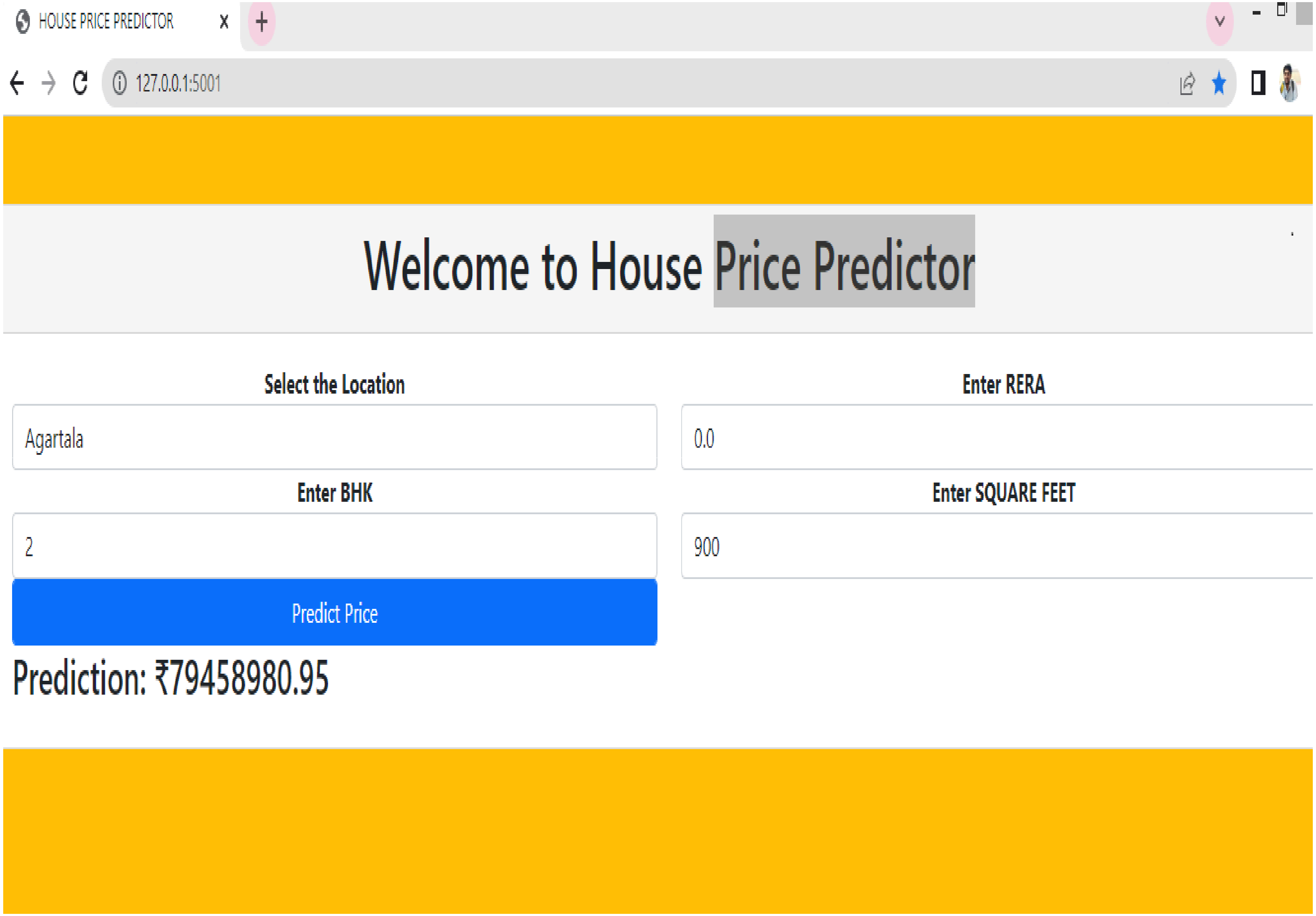
0.0

Enter BHK

Enter BHK

Enter SQUARE FEET

Enter SQUAREFEET



Welcome to House Price Predictor

Select the Location: Agartala

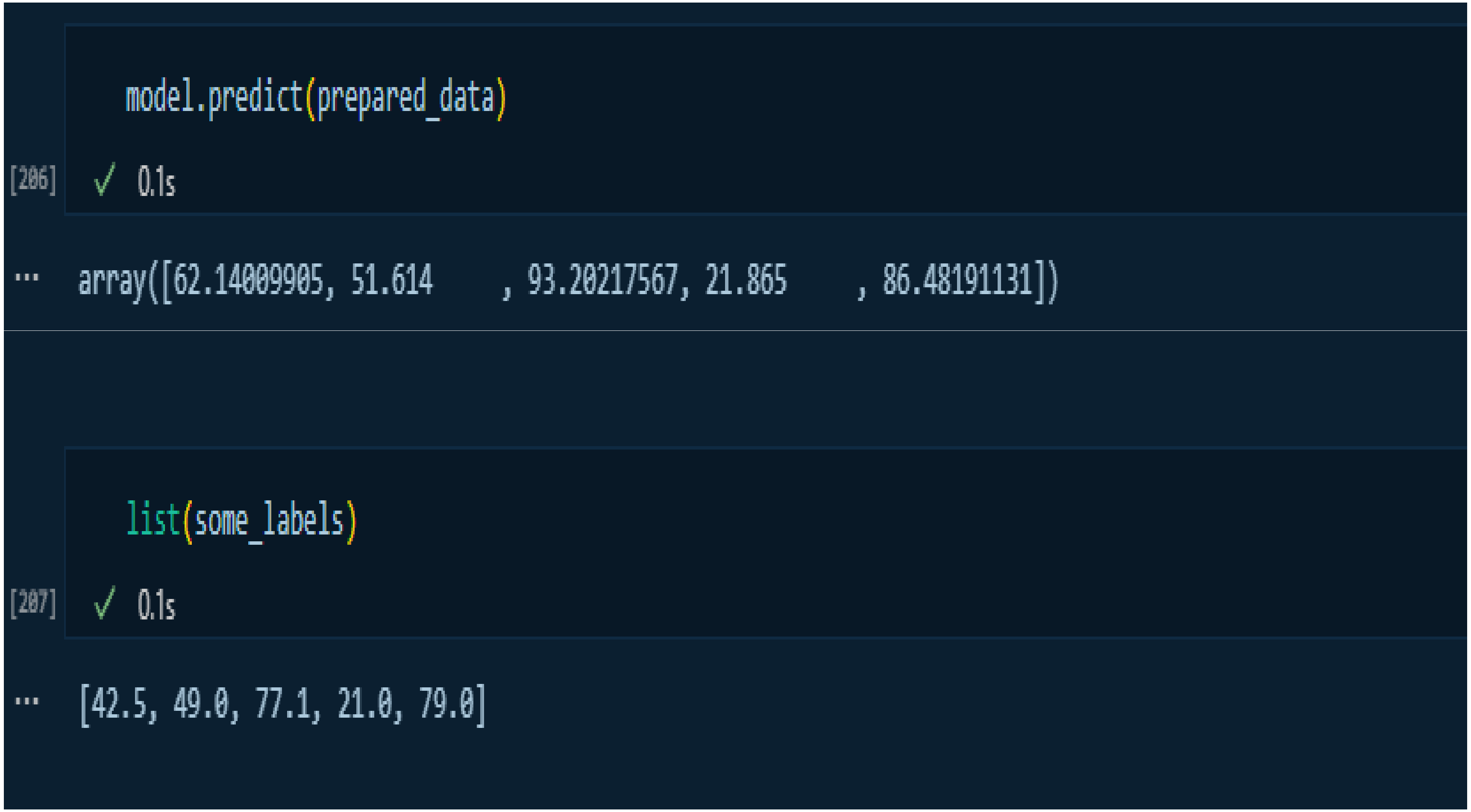
Enter RERA: 0.0

Enter BHK: 2

Enter SQUARE FEET: 900

Prediction: ₹79458980.95

Validation Checks



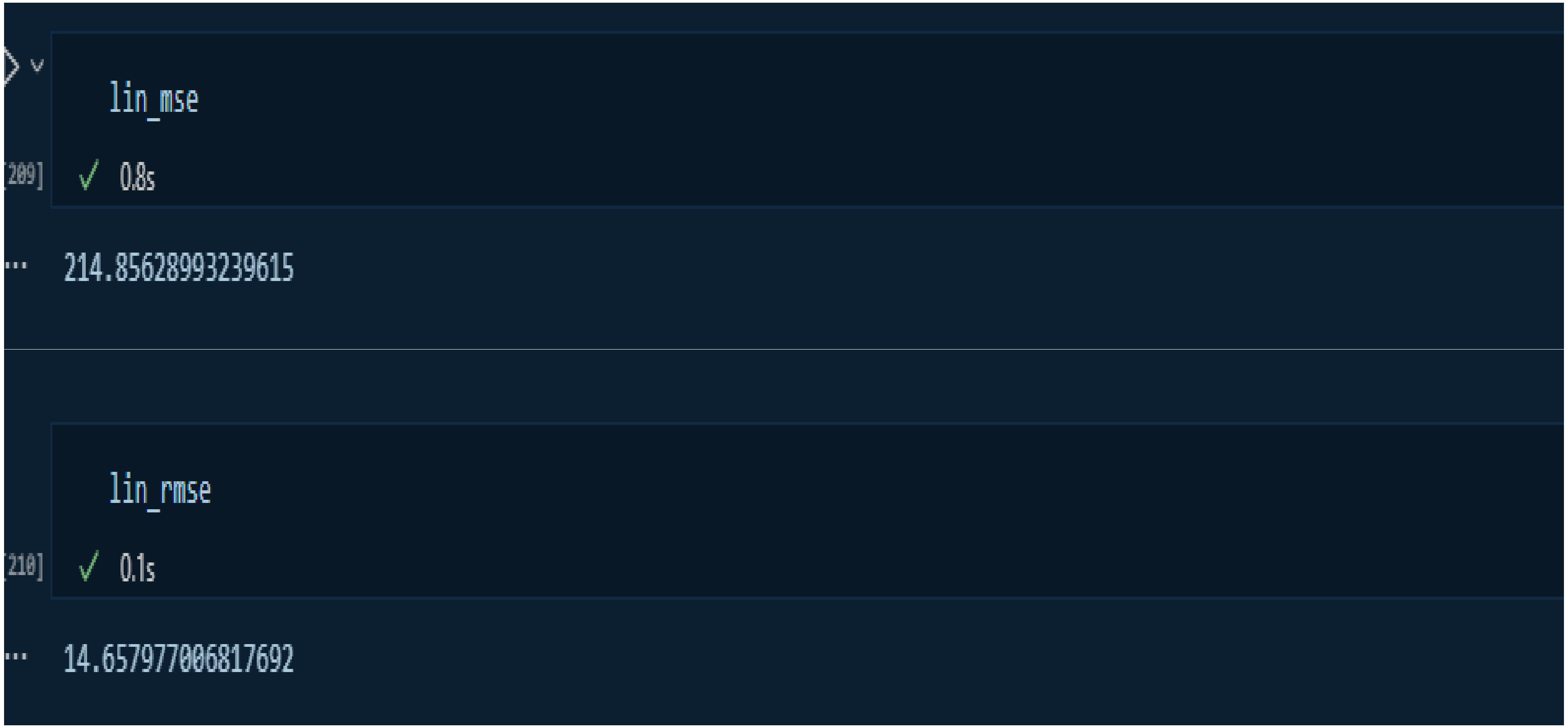
```
model.predict(prepared_data)
```

```
array([62.14009905, 51.614    , 93.20217567, 21.865    , 86.48191131])
```

```
list(some_labels)
```

```
[42.5, 49.0, 77.1, 21.0, 79.0]
```

Testing



```
lin_mse
```

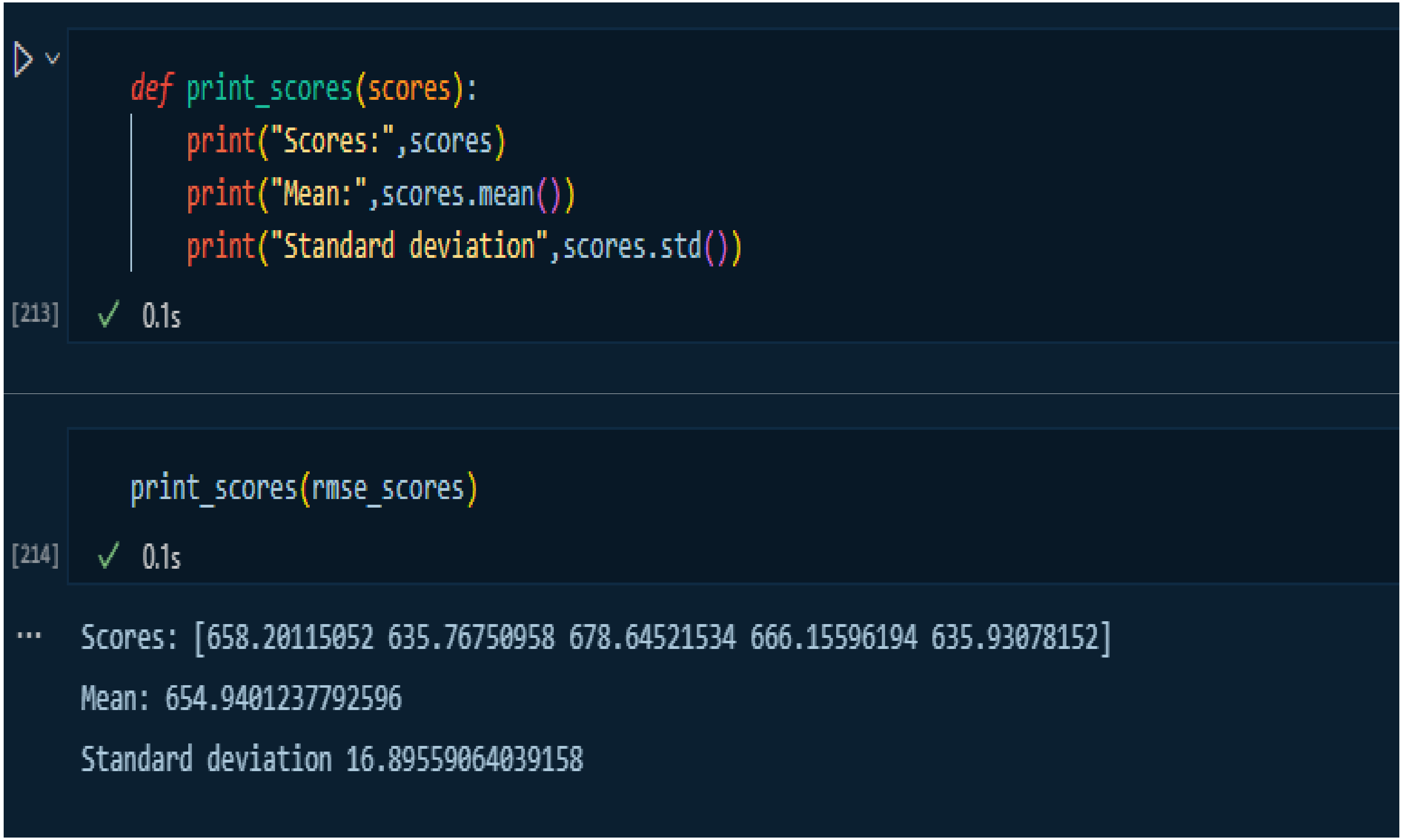
```
[209] ✓ 0.8s
```

```
... 214.85628993239615
```

```
lin_rmse
```

```
[210] ✓ 0.1s
```

```
... 14.657977006817692
```



```
[213]
def print_scores(scores):
    print("Scores:",scores)
    print("Mean:",scores.mean())
    print("Standard deviation",scores.std())
```

```
[214]
print_scores(rmse_scores)
```

```
Scores: [658.20115052 635.76750958 678.64521534 666.15596194 635.93078152]
Mean: 654.9401237792596
Standard deviation 16.89559064039158
```

Implementation and Maintenance

- **Import Libraries:** library is a collection of related modules. It contains bundles of code that can be used repeatedly in different programs**.**

We use following library in python

I. Pandas

II. Numpy

III. Sklearn

IV. seaborn

- **Load Dataset:** It is process of reading and loading of data from csv file.
- **Exploratory Data Analysis:** EDA is a phenomenon under data analysis used for gaining a better understanding of data aspects like:
  – main features of data
  – variables and relationships that hold between them
  – identifying which variables are important for our problem

- **Data Cleaning:** Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled.
- **Feature Engineering: Feature engineering** (or **feature extraction**) is the process of using domain knowledge to extract features (characteristics, properties, attributes) from raw data. The motivation is to use these extra features to improve the quality of results from a machine learning process, compared with supplying only the raw data to the machine learning process.
- **Outlier Removal:** An Outlier is a data-item/object that deviates significantly from the rest of the (so-called normal) objects. They can be caused by measurement or execution errors. The analysis for outlier detection is referred to as outlier mining.
- **Data Visualization:** Data visualization is the graphical representation of information and data.
- **Building a Model:** After passing through above steps, we create appropriate model.
- **Test the Model for few properties**: After building model now we test model and check rmse.
- **Export the tested model to a joblib file:**Now we save tested model using joblib library.

Future Scope Of The Project

Our model had a low rise score, but there is room for improvement. In a real world scenario, we can use such a mode to predict house prices. This model should check for new data, once in a month, and incorporate them to expand the dataset and produce better results.

We can try out other advanced regression techniques, Random Forest and Bayesian Ridge Algorithm, for prediction. Since the data highly correlated and improve our proje.ct

Conclusion

As we discussed about machine learning and its type briefly and collect information from various online sources. We conclude that Supervised learning is used to design House Price Prediction. Here we have to know real entity i.e. price so use regression to solve this. After a analysis now we start implementation of this project.

Biblography

- https://datasetsearch.research.google.com/search?query=house%20price%20%20pridiction&docid=L2cvMTFxbDFxMjVxaw%3D%3D
- https://www.kaggle.com/anmolkumar/house-price-prediction-challenge
- https://towardsdatascience.com/tagged/house-price-prediction?p=dea265cc3154
- https://www.geeksforgeeks.org/regression-classification-supervised-machine-learning/
- https://www.geeksforgeeks.org/machine-learning/