**HOUSE PRICE PREDICTION USING ML**

***Mini Project Report***

***Submitted in partial fulfilment of the***

***Requirements for the award of the Degree of***

**BACHELOR OF ENGINEERING**

IN

**INFORMATION TECHNOLOGY**

BY

**<E KAVYA><1602-19-737-019>**

**<V SIDDHARTHA><1602-19-737-040>**

**<K UPENDER><1602-19-737-051>**

****

**Department of Information Technology**

**Vasavi College of Engineering (Autonomous)**

**(Affiliated to Osmania University)**

**Ibrahimbagh, Hyderabad-31**

****

**DECLARATION BY THE CANDIDATE**

We, E KAVYA,V SIDDHARTH and K UPENDER , bearing hall ticket numbers, 1602-19-737-019,1602-19-737-040 and 1602-19-737-051 , hereby declare that the project report entitled ”HOUSE PRICE PREDICTION” Department of Information Technology, Vasavi College of Engineering, Hyderabad, is submitted in partial fulfilment of the requirement for the award of the degree of Bachelor of Engineering in Information Technology

        This is a record of bonafide work carried out by me and the results embodied in this project report have not been submitted to any other university or institute for the award of any other degree or diploma.

**E Kavya  1602-19-737-019**

**V Siddhartha  1602-19-737-040**

**K Upender 1602-19-737-051**

**(Faculty In-Charge)                 (Head,Dept of IT)**

**BONAFIDE CERTIFICATE**

This is to certify that the project entitled **"House price prediction"** being submitted by **E KAVYA, V SIDDHARTHA and K UPENDER ,** bearing hall ticket numbers, **1602-17-737-019, 1602-19-737-040 and 1602-17-737- 051** in partial fulfillment of the requirements for the award of the degree of Bachelor of Engineering in Information Technology is a record of bonafide work carried out by him/her under my guidance.

**Mrs Leelavathi Dr. K. Ram MohanRao**

**Faculty In-charge Professor & HOD**

**Department of IT**

# ACKNOWLEDGEMENT

It gives us immense pleasure to thank the **department of INFORMATION TECHNOLOGY**, for introducing the subject **“mini project”** in BE fifth semester that let us learn and explore more features in web development and machine learning.

I would also like to show my appreciation to our honourable principal, **Dr S V RAMANA** sir, for supporting us and our beloved mini project lecturer, **Mrs LEELAVATHI** mam, for letting us properly understand the process of doing the mini project and for providing insight and expertise that greatly assisted the project.

My parents were my first teachers and they have provided me with such a great exposure that has helped me bloom. My family and friends will always be loved for sticking by me through thick and thin. THANK YOU!

# TABLE OF CONTENTS

1.**Abstract**

**2.Introduction**

**3.Related Work**

3.1 Related Questions

3.2 Limitations

3.3 Linear Regression

3.4 Lasso Regression

3.5 Ridge Regression

**4.Proposed Work**

4.1 Experiment

4.2 Architecture

4.3 Use Case Diagram

4.4 Implementation

4.4.1 Login

4.4.2 Registration

4.4.3 Algorithms used for Prediction

**5.Results**

**6.Discussion and Future Work**

**7.References**

# 

**1.ABSTRACT**

House Price Index (HPI) is commonly used to estimate the changes in housing price. Since housing price is strongly correlated to other factors such as location, area, population, it requires other information apart from HPI to predict individual housing price. An accurate prediction on the house price is important to prospective homeowners, developers, investors, appraisers, tax assessors and other real estate market participants, such as, mortgage lenders and insurers. Traditional house price prediction is based on cost and sale price comparison lacking an accepted standard and a certification process. Therefore, the availability of a house price prediction model helps fill up an important information gap and improve the efficiency of the real estate market.

This project will cover all the information about how we can use previous data to predict future value to sell a house. Suppose you want to sell your house in a particular place but you don’t know the perfect price for the house to sell, our project can predict a price for you. When someone wants to buy a house, the buyer can have an idea about the house price, which is predicted from data in our project.

**2.INTRODUCTION**

Machine learning is a subfield of Artificial Intelligence (AI) that works with algorithms and technologies to extract useful information from data. Machine learning methods are appropriate in big data since attempting to manually process vast volumes of data would be impossible without the support of machines. Machine learning in computer science attempts to solve problems algorithmically rather than purely mathematically. Therefore, it is based on creating algorithms that permit the machine to learn. However, there are two general groups in machine learning which are supervised and unsupervised. Supervised is where the program gets trained on pre-determined set to be able to predict when a new data is given. Unsupervised is where the program tries to find the relationship and the hidden pattern between the data

Several Machine Learning algorithms are used to solve problems in the real world today. However, some of them give better performance in certain circumstances, as stated in the No Free Lunch Theorem . Thus, this thesis attempts to use regression algorithms and artificial neural network (ANN) to compare their performance when it comes to predicting values of a given dataset. The performance will be measured upon predicting house prices since the prediction in many regression algorithms relies not only on a specific feature but on an unknown number of attributes that result in the value to be predicted. House prices depend on an individual house specification. Houses have a variant number of features that may not have the same cost due to its location. For instance, a big house may have a higher price if it is located in desirable rich area than being placed in a poor neighbourhood. The data used in the experiment will be handled by using a combination of pre-processing methods to improve the prediction accuracy. In addition, some factors will be added to the local dataset in order to study the relationship between these factors and the sale price in Malmö.

**SOFTWARE REQUIREMENTS**

**TOOLS:** Python 3, Jupiter notebook, pycharm

**LIBRARIES:**

Pandas – for handling structured data

Scikit Learn – for machine learning

NumPy– for linear algebra and mathematics

Seaborn– for data visualization

**DATA SET:** Kaggle**,** sklearn

**FRONT END:** html and css

**BACK END:** php

**3.RELATED WORK**

**3.1 Related Questions**

The study answers the following research questions: - Research question 1: Which machine learning algorithm performs better and has the most accurate result in house price prediction? And why? - Research question 2: What are the factors that have affected house prices in Malmö over the years?

**3.2 Limitations**

The local data will be requested from the Svensk mäklarstatistik . The request contains a list of features, that matches the public dataset's features, that is desired to be available when the data is sent. There is no guarantee that the data will be available in time nor contains the exact requested list of features. Thus, there might be a risk that the access will be denied or delayed. If so, the study will be accomplished based only on the public dataset. Moreover, this study will not cover all regression algorithms; instead, it is focused on the chosen algorithm, starting from the basic regression techniques to the advanced ones. Likewise, the artificial neural network that has many techniques and a wide area and several training methods that do not fit in this study.

**3.3.Linear Regression**

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

An example is let’s say you are running a sales promotion and expecting a certain number of count of customers to be increased now what you can do is you can look the previous promotions and plot if over on the chart when you run it and then try to see whether there is an increment into the number of customers whenever you rate the promotions and with the help of the previous historical data you try to figure it out or you try to estimate what will be the count or what will be the estimated count for my current promotion this will give you an idea to do the planning in a much better way about how many numbers of stalls maybe you need or how many increase number of employees you need to serve the customer. Here the idea is to estimate the future value based on the historical data by learning the behaviour or patterns from the historical data.

**3.4. Lasso Regression**

Least Absolute Shrinkage and Selection Operator (Lasso) is an L1-norm regularised regression technique that was formulated by Robert Tibshirani in 1996 . Lasso is a powerful technique that performs regularisation and feature selection. Lasso introduces a bias term, but instead of squaring the slope like Ridge regression, the absolute value of the slope is added as a penalty term. Lasso is defined as: 𝐿 = 𝑀𝑖𝑛(𝑠𝑢𝑚 𝑜𝑓 𝑠𝑞𝑢𝑎𝑟𝑒𝑑 𝑟𝑒𝑠𝑖𝑑𝑢𝑎𝑙𝑠 + 𝛼 ∗ |𝑠𝑙𝑜𝑝𝑒|) Where 𝑀𝑖𝑛(𝑠𝑢𝑚 𝑜𝑓 𝑠𝑞𝑢𝑎𝑟𝑒𝑑 𝑟𝑒𝑠𝑖𝑑𝑢𝑎𝑙𝑠) is the Least Squared Error, and 𝛼 ∗ |𝑠𝑙𝑜𝑝𝑒| is the penalty term. However, alpha 𝑎 is the tuning parameter which controls the strength of the penalty term. In other words, the tuning parameter is the value of shrinkage. |𝑠𝑙𝑜𝑝𝑒| is the sum of the absolute value of the coefficients . Cross-validation is a technique that is used to compare different machine learning algorithms in order to observe how these methods will perform in practice. Cross-validation method divides the data into blocks. Each block at a time will be used for testing by the algorithm, and the other blocks will be used for training the model. In the end, the results will be summarised, and the block that performs best will be chosen as a testing block . However, 𝛼 is determined 4 by using cross-validation. When 𝛼 = 0, Lasso becomes Least Squared Error, and when 𝛼 ≠ 0, the magnitudes are considered, and that leads to zero coefficients. However, there is a reverse relationship between alpha 𝑎 and the upper bound of the sum of the coefficients 𝑡. When 𝑡 → ∞, the tuning parameter 𝑎 = 0. Vice versa when 𝑡 = 0 the coefficients shrink to zero and 𝑎 → ∞ . Therefore, Lasso helps to assign zero weights to most redundant or irrelevant features in order to enhance the prediction accuracy and interpretability of the regression model. Throughout the process of features selection, the variables that still have non-zero coefficients after the shrinking process are selected to be part of the regression model . Therefore, Lasso is powerful when it comes to feature selection and reducing the overfitting.

**3.5. Ridge Regression**

The Ridge Regression is an L2-norm regularised regression technique that was introduced by Hoerl in 1962 . It is an estimation procedure to manage collinearity without removing variables from the regression model. In multiple linear regression, the multicollinearity is a common problem that leads least square estimation to be unbiased, and its variances are far from the correct value. Therefore, by adding a degree of bias to the regression model, Ridge Regression reduces the standard errors, and it shrinks the least square coefficients towards the origin of the parameter space . Ridge formula is: 𝑅 = 𝑀𝑖𝑛(𝑠𝑢𝑚 𝑜𝑓 𝑠𝑞𝑢𝑎𝑟𝑒𝑑 𝑟𝑒𝑠𝑖𝑑𝑢𝑎𝑙𝑠 + 𝛼 ∗ 𝑠𝑙𝑜𝑝𝑒 2 ) Where 𝑀𝑖𝑛(𝑠𝑢𝑚 𝑜𝑓 𝑠𝑞𝑢𝑎𝑟𝑒𝑑 𝑟𝑒𝑠𝑖𝑑𝑢𝑎𝑙𝑠) is the Least Squared Error, and 𝛼 ∗ 𝑠𝑙𝑜𝑝𝑒 2 is the penalty term that Ridge adds to the Least Squared Error. When Least Squared Error determines the values of parameters, it minimises the sum of squared residuals. However, when Ridge determines the values of parameters, it reduces the sum of squared residuals. It adds a penalty term, where 𝛼 determines the severity of the penalty and the length of the slope. In addition, increasing the 𝛼 makes the slope asymptotically close to zero. Like Lasso, 𝛼 is determined by applying the Cross-validation method. Therefore, Ridge helps to reduce variance by shrinking parameters and make the prediction less sensitive.

**4.PROPOSED WORK**

**4.1. Experiment**

The experiment is done to pre-process the data and evaluate the prediction accuracy of the models. The experiment has multiple stages that are required to get the prediction results. These stages can be defined as: -

**Pre-processing**: both datasets will be checked and pre-processed. These methods have various ways of handling data. Thus, the pre-processing is done on multiple iterations where each time the accuracy will be evaluated with the used combination.

**Data splitting**: dividing the dataset into two parts is essential to train the model with one and use the other in the evaluation. The dataset will be split 75% for training and 25% for testing.

**Evaluation**: the accuracy of both datasets will be evaluated by measuring the R2 and RMSE rate when training the model alongside an evaluation of the actual prices on the test dataset with the prices that are being predicted by the model.

**Performance**: alongside the evaluation metrics, the required time to train the model will be measured to show the algorithm vary in terms of time.

**Correlation**: correlation between the available features and house price will be evaluated using the Pearson Coefficient Correlation to identify whether the features have a negative, positive or zero correlation with the house price

**4.2. Architecture**

Diagram

Description automatically generated

**4.3. Design**

**Use case diagram**

Diagram

Description automatically generated

**4.4. Implementation**

**4.4.1.Login**

The login page is the first page that users see in our modified application. It will provide two text fields - one for entering a login name and one for entering a password. In addition it should have a command button that initiates the register action. If either of the text fields is left blank it is an error that must be reported to the user. If both fields are filled in but there is no record of the user name or the password is incorrect that must also be reported to the user.

**4.4.2.Registration**

The registration page will be displayed when a user clicks on the register command button in the login page. It should provide three text fields - one for entering a login name and two for entering passwords. In addition it will have a command button that initiates the registration action.

If any of the text fields is left blank it is an error that must be reported to the user. If all fields are filled in but there is already a record for the user name then that must be reported as an error. If the two passwords are different that must also be reported as an error.

**4.4.3. Algorithms for Price Prediction**

**Linear Regression**

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

**Lasso**

Lasso regression is a classification algorithm that uses shrinkage in simple and sparse models(i.e model with fewer parameters). In Shrinkage, data values are shrunk towards a central point like the mean. Lasso regression is a regularized regression algorithm that performs L1 regularization which adds penalty equal to the absolute value of the magnitude of coefficients.

**Ridge**

A Ridge regressor is basically a regularized version of a Linear Regressor. i.e to the original cost function of linear regressor we add a regularized term that forces the learning algorithm to fit the data and helps to keep the weights lower as possible. The regularized term has the parameter ‘alpha’ which controls the regularization of the model i.e helps in reducing the variance of the estimates.

**FRONT END CODE**

(House price predictor window)

<?php

session\_start();

if (!isset($\_SESSION['username'])) {

header("Location: index.php");

}

?>

<!doctype html>

<html lang="en">

<head>

<!-- Required meta tags -->

<meta charset="utf-8">

<meta name="viewport" content="width=device-width, initial-scale=1, shrink-to-fit=no">

<!-- Bootstrap CSS -->

<link rel="stylesheet" href="https://stackpath.bootstrapcdn.com/bootstrap/4.1.3/css/bootstrap.min.css" integrity="sha384-MCw98/SFnGE8fJT3GXwEOngsV7Zt27NXFoaoApmYm81iuXoPkFOJwJ8ERdknLPMO" crossorigin="anonymous">

<title>House Price Predictor!</title>

</head>

<style>

body

{

margin:0;

padding:0;

background:url("house.png");

background-size:cover;

font-family: sans-serif;

}

.card

{

position:absolute;

top:50%;

left:50%;

transform: translate(-50%, -50%);

width:350px;

height:420px;

padding:80px 40px;

box-sizing: border-box;

background:rgba(0,0,0,0.5) ;

}

h1{

margin:0;

padding:0 0 20px;

color:#1E90FF;

text-align:center;

}

</style>

<body class="bg-dark">

<div class="container">

<div class="row">

<div class="card" style="width: 100%; height: 100%;margin-top: 50px">

<div class="card-body">

<div class="card-header" style="text-align:center">

<h1>welcome to house price predictor</h1>

</div>

<form method="post" accept-charset="utf-8">

<div class="row">

<div class="col-md-6 form-group" style="text-align: center;color: #fff">

<label><b>Select the Location:</b></label>

<select class="selectpicker form-control" id="location" name="location" required="1" style=" background:rgba(0,0,0,0) ">

</select>

</div>

<div class="col-md-6 form-group" style="text-align: center;color: #fff;">

<label><b>Enter BHK:</b></label>

<input type="text" class="form-control" id="bhk" name="bhk" placeholder="Enter BHK" style=" background:rgba(0,0,0,0);color: #fff">

</div>

<div class="col-md-6 form-group" style="text-align: center;color: #fff;">

<label><b>Enter no. of bathroomsK:</b></label>

<input type="text" class="form-control" id="bath" name="bath" placeholder="Enter no. of bathrooms " style=" background:rgba(0,0,0,0);color: #fff ">

</div>

<div class="col-md-6 form-group" style="text-align: center;color: #fff;">

<label><b>Enter square feet:</b></label>

<input type="text" class="form-control" id="total\_sqft" name="total\_sqft" placeholder="Enter square feet" style=" background:rgba(0,0,0,0);color: #fff">

</div>

<div class="col-md-12 form-group">

<button class="btn btn-primary form-control" oneclick="send\_data()">Predict price</button>

</div>

</div>

</form>

<br>

<div class="col-md-12" style="text-align: center">

<h1><span id="prediction"></span></h1>

</div>

<a href="logout.php"><font size="80px"><p style="text-align:center">Logout</a>

</div>

</div>

</div>

</div>

<!-- jQuery first, then Popper.js, then Bootstrap JS -->

<script src="https://code.jquery.com/jquery-3.3.1.slim.min.js" integrity="sha384-q8i/X+965DzO0rT7abK41JStQIAqVgRVzpbzo5smXKp4YfRvH+8abtTE1Pi6jizo" crossorigin="anonymous"></script>

<script src="https://cdnjs.cloudflare.com/ajax/libs/popper.js/1.14.3/umd/popper.min.js" integrity="sha384-ZMP7rVo3mIykV+2+9J3UJ46jBk0WLaUAdn689aCwoqbBJiSnjAK/l8WvCWPIPm49" crossorigin="anonymous"></script>

<script src="https://stackpath.bootstrapcdn.com/bootstrap/4.1.3/js/bootstrap.min.js" integrity="sha384-ChfqqxuZUCnJSK3+MXmPNIyE6ZbWh2IMqE241rYiqJxyMiZ6OW/JmZQ5stwEULTy" crossorigin="anonymous"></script>

</body>

</html>

**TESTING**

1.In login register we need to provide valid user details who are already registered

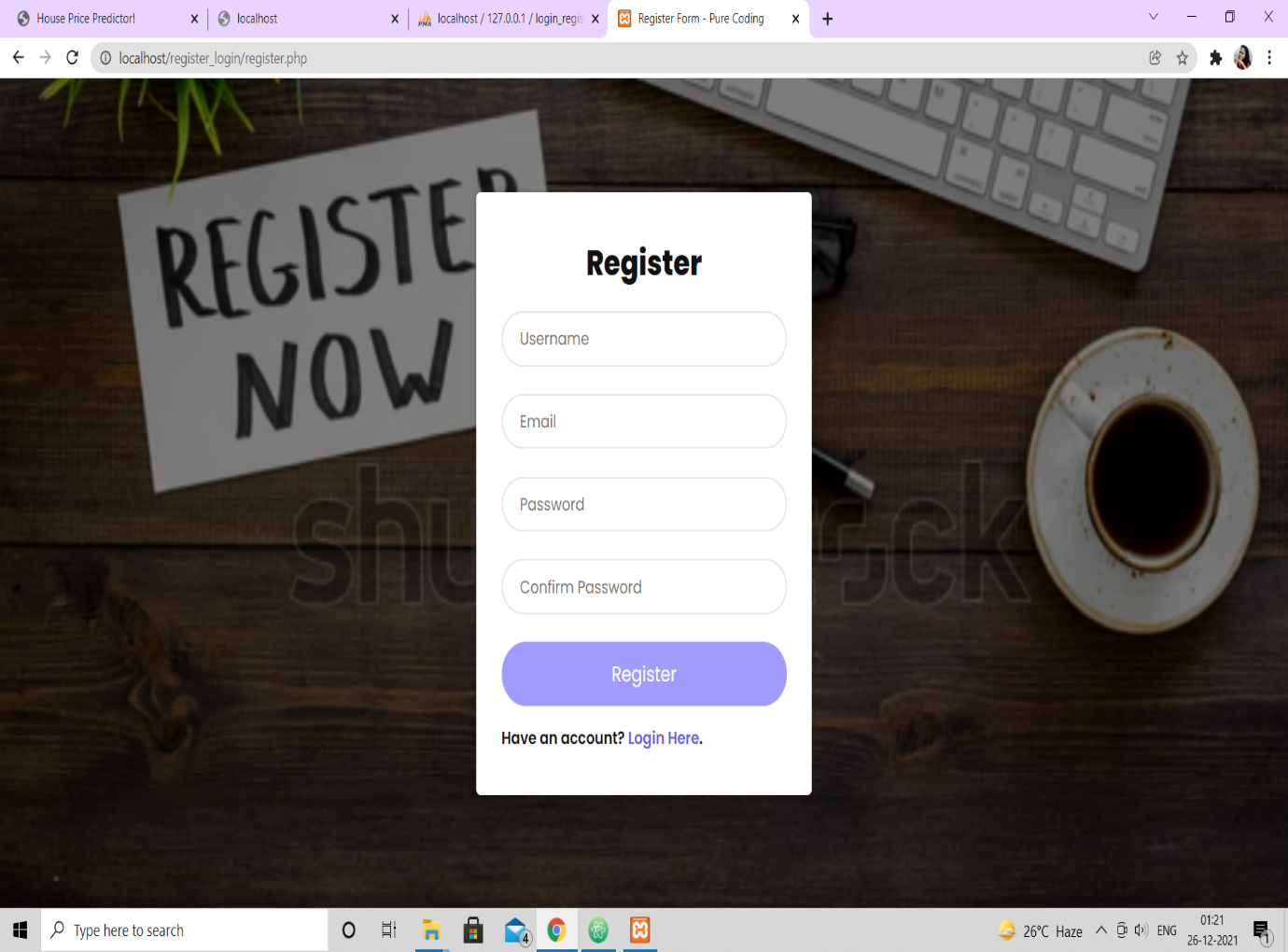
2.We can test the page by giving invalid inputs(Unregistered users)

3.For prediction we need to give location,Square feet,BHK and number of bathrooms

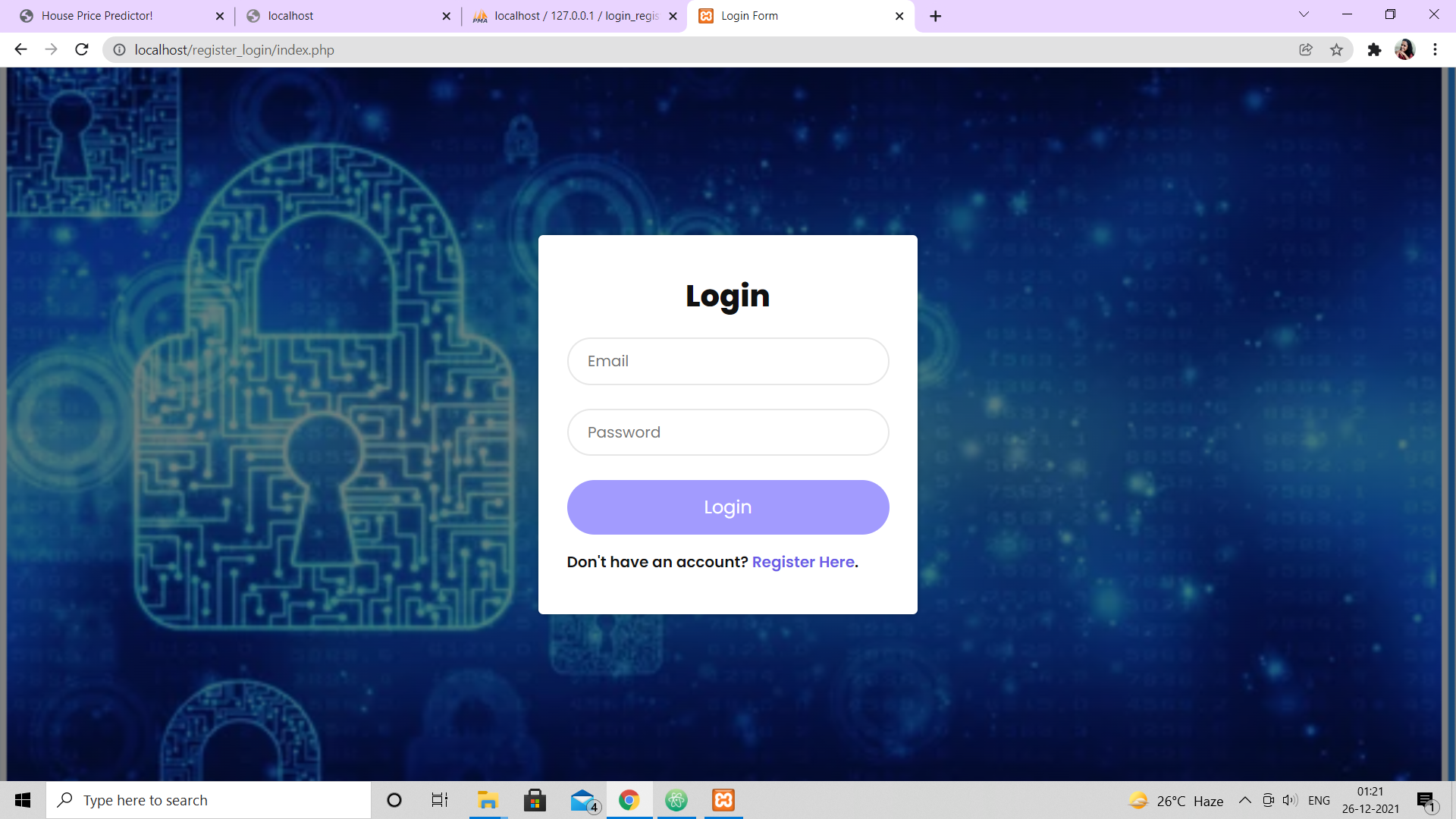
4.By the Accuracy score of the price of the House we can judge the price

**RESULTS**

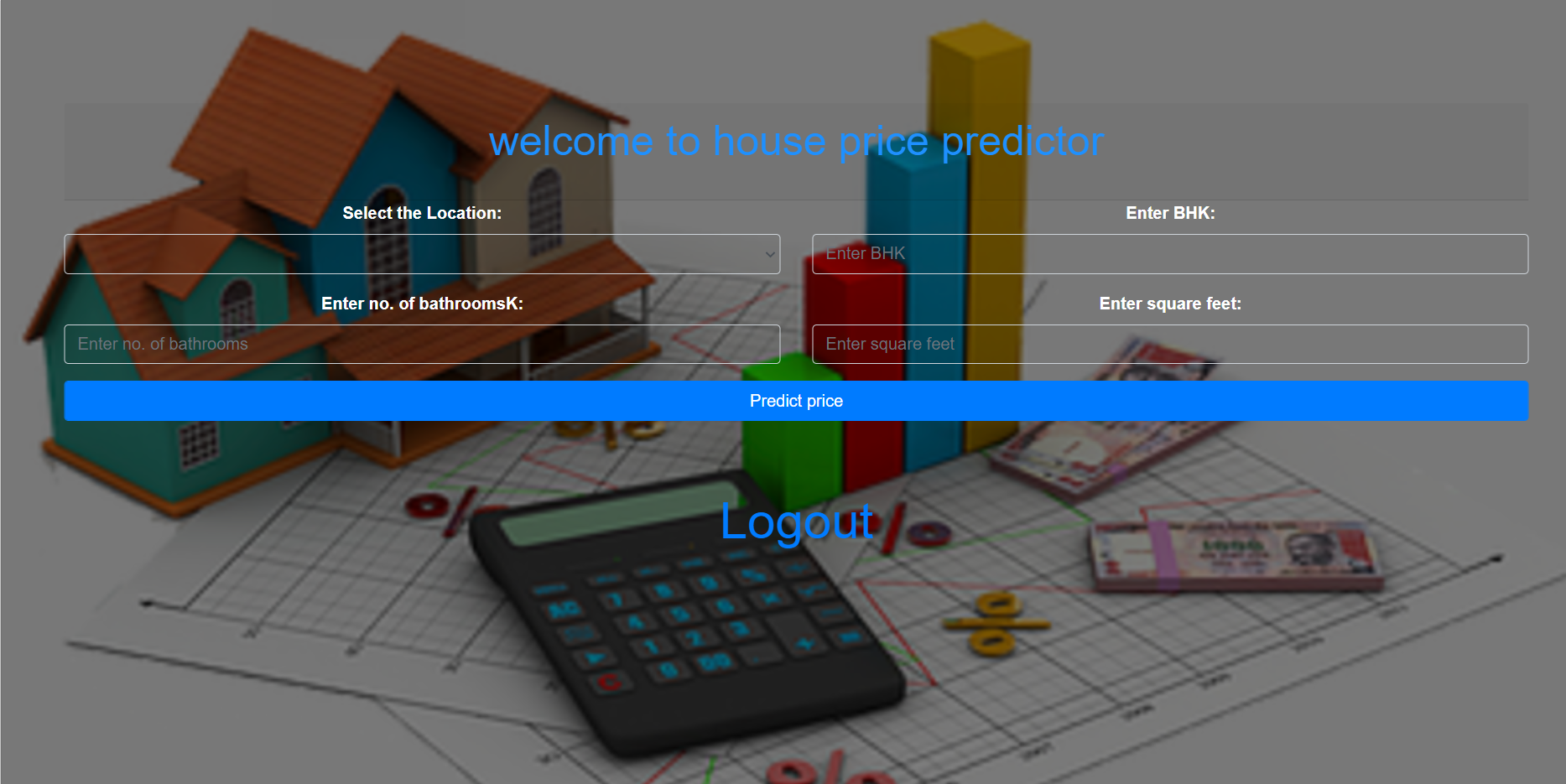
**Register web page:**

****

**Login web page:**

****

**House Price Predictor web page:**

****

**REFERENCES**

1)<https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview>

 2)https://scikit-learn.org/stable/datasets/toy\_dataset.html?highlight=house%20price%20prediction

 3)https://www.analyticsvidhya.com/blog/2021/06/linear-regression-in-machine-learning/

 4)https://www.youtube.com/watch?v=ukzFI9rgwfU

 5)https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

 6)https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/