A PROJECT REPORT

on

"Bengaluru House Price Predictor"

Submitted to

KIIT Deemed to be University

In Partial Fulfilment of the Requirement for the Award of

BACHELOR'S DEGREE IN
INFORMATION TECHNOLOGY

BY

NIHAR RANJAN SAHOO	21052165
HEMA MALIK	21052156
MD. AFAQUE AKHTAR	21051484
SUJAL SINGH	21052927

UNDER THE GUIDANCE OF
Dr. Manas Ranjan Nayak

SCHOOL OF COMPUTER ENGINEERING
KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY
BHUBANESWAR, ODISHA - 751024

Acknowledgements

We would like to express our sincere gratitude to our guide and professor, **Dr. Manas Ranjan Nayak**, who gave us this opportunity to implement this project and also guided us in the process in understanding the basic underlying principle of this project which helped us to complete the same. We used this medium to acknowledge and appreciate him as he contributed immensely and graciously for the achievement of this project work.

NIHAR RANJAN SAHOO
HEMA MALIK
MD. AFAQUE AKHTAR
SUJAL SINGH

ABSTRACT

The housing market plays a crucial role in the economy, impacting individuals, businesses, and governments alike. Predicting house prices accurately can provide valuable insights for homeowners, buyers, sellers, and policymakers.

In this project, we develop a machine learning model to predict house prices based on various features such as the number of bedrooms, bathrooms, size of the house, and zip code.

Leveraging a comprehensive dataset consisting of historical housing data from Bengaluru, we employ techniques such as data preprocessing, feature engineering, and model training using linear regression with regularization (Lasso and Ridge) to achieve accurate predictions.

Additionally, a web application is developed to provide a user-friendly interface for users to input property details and receive predicted prices in real-time.

Keywords: Supervised Machine Learning, Regularization, Linear Regression, Pipeline

Contents

1.Introduction

The project aims to develop a machine learning model capable of predicting residential property prices based on a dataset containing housing data. With the real estate market's complexity, accurate price estimation is crucial for informed decision-making. Leveraging machine learning techniques and historical housing data, this project seeks to create a predictive model that provides precise price estimates based on property attributes like bedrooms, bathrooms, size, and location. This report outlines the project's methodology, implementation details, results, and recommendations for future work, highlighting its significance in facilitating informed decision-making in the real estate market.

2.Basic Concepts/ Literature Review

In this section, we provide an overview of the fundamental concepts and techniques used in the Bengaluru House Price Prediction project. These concepts are essential for understanding the methodology and implementation of the project:

2.1. Machine Learning

Machine learning is a subset of artificial intelligence (AI) that focuses on developing algorithms and models capable of learning from data to make predictions or decisions without being explicitly programmed. It encompasses various techniques including supervised learning where models are trained on labeled data, where each example is associated with a target variable, including algorithms like linear regression, decision trees, and support vector machines.

2.2 Regression Analysis

Regression analysis is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (features). The goal is to predict the value of the dependent variable based on the values of the independent variables. Types of regression techniques include polynomial regression, ridge regression, and lasso regression, which introduce regularization to prevent overfitting and improve model generalization.

2.3 Feature Engineering

Feature engineering is the process of transforming raw data into a format that is suitable for machine learning algorithms. It includes handling missing values, encoding categorical variables, scaling numerical features, and creating interaction terms or polynomial features.

2.4 Web Development with Flask

Flask is a lightweight and flexible web framework for Python, designed to create web applications quickly and efficiently. It provides tools and libraries for routing, handling requests and responses, template rendering, and interacting with databases. With Flask, developers can build dynamic web applications with minimal boilerplate code and easily integrate machine learning models for real-time predictions.

3.Problem Statement / Requirement Specifications

The House Price Prediction project aims to address the need for a reliable and accurate solution to predict residential property prices in India. The problem statement revolves around the challenges faced by homebuyers, sellers, and real estate professionals in estimating property prices due to the complex and dynamic nature of the real estate market. The lack of transparency and accessibility to relevant data further complicates the decision-making process in property transactions.

3.1 Project Planning

- Data Collection: Gathering a comprehensive dataset of residential properties from Bengaluru, including features such as the number of bedrooms, bathrooms, house size, and location (zip code).

- Data Preprocessing: Cleaning and preprocessing the dataset to handle missing values, outliers, and inconsistencies, and preparing it for model training.

- Model Development: Implementing and training machine learning regression models, including Linear Regression, Lasso Regression, and Ridge Regression, to predict house prices based on the collected features.

- Web Application Development: Developing a user-friendly web application using the Flask framework to provide an interactive interface for users to input property details and receive real-time price predictions.

- Deployment and Testing: Deploying the trained model and web application on a suitable hosting platform for public access. Conducting thorough testing to ensure functionality, reliability, and scalability.

3.2 Project Analysis

After collecting the requirements and conceptualizing the problem statement, it is essential to analyze the project to identify any ambiguities or mistakes. This analysis ensures that the project requirements are well-defined and aligned with the objectives.

3.3 System Design

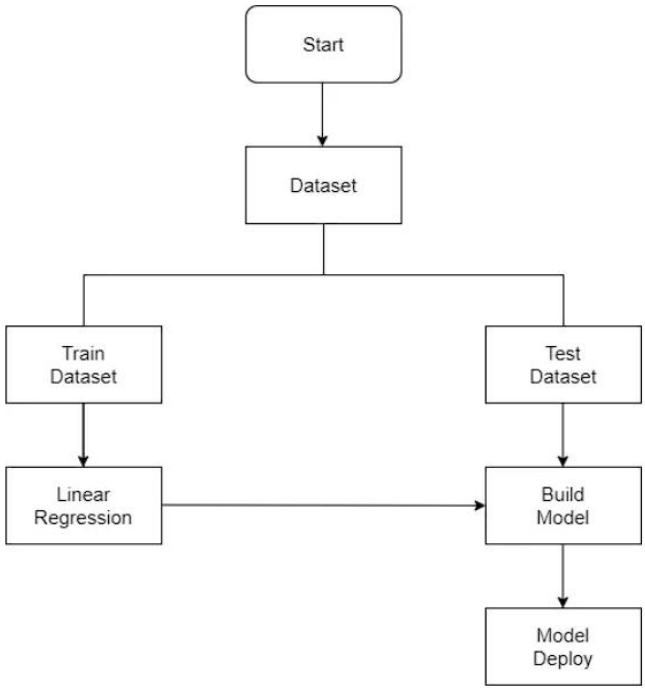
3.3.1 Design Constraints

This project primarily relies on software tools and techniques for data preprocessing, model development, and web application development. The following software and hardware specifications are required:

- **Software**: Python programming language, pandas library for data manipulation, scikit-learn library for machine learning, Flask framework for web development.

- **Hardware**: Standard computing hardware with sufficient processing power and memory to handle data preprocessing, model training, and web application deployment.

3.3.2 System Architecture OR Block Diagram



4.Implementation

This section provides a detailed explanation of the steps involved in building the house price prediction model and deploying it as a web application.

4.1 Methodology OR Proposal:

• **Data Preprocessing:** The first step in our implementation involved preprocessing the dataset to prepare it for model training. We loaded the dataset from the provided CSV file (train.csv) using the pandas library and performed exploratory data analysis (EDA) to understand its structure and contents. This included checking the shape of the dataset, examining the data types of each column, and identifying missing values. After dropping irrelevant columns and handling missing values, we proceeded to calculate the price per square foot.

```
In [1]: import pandas as pd
        import numpy as np
```

```
In [2]: data=pd.read_csv('train.csv')
```

```
In [3]: data.head()
```

Out[3]:

	beds	baths	size	size_units	lot_size	lot_size_units	zip_code	price
0	3	2.5	2590.0	sqft	6000.00	sqft	98144	795000.0
1	4	2.0	2240.0	sqft	0.31	acre	98106	915000.0
2	4	3.0	2040.0	sqft	3783.00	sqft	98107	950000.0
3	4	3.0	3800.0	sqft	5175.00	sqft	98199	1950000.0
4	2	2.0	1042.0	sqft	NaN	NaN	98102	950000.0

saving final dataset to be used

```
In [21]: data.to_csv("final_dataset.csv")
```

```
In [22]: X=data.drop(columns=['price'])
         y=data['price']
```

• **Model Training:** Next, we split the dataset into features (X) and the target variable (y). We then divided the data into training and testing sets using the train_test_split function from the sklearn.model_selection module. For model training, we implemented three regression algorithms: Linear Regression, Lasso Regression, and Ridge Regression.

saving final dataset to be used

```
In [21]: data.to_csv("final_dataset.csv")

In [22]: X=data.drop(columns=['price'])
         y=data['price']

In [23]: from sklearn.model_selection import train_test_split
         from sklearn.linear_model import LinearRegression,Lasso,Ridge
         from sklearn.preprocessing import OneHotEncoder, StandardScaler
         from sklearn.compose import make_column_transformer
         from sklearn.pipeline import make_pipeline
         from sklearn.metrics import r2_score

In [24]: X_train,X_test,y_train,y_test = train_test_split(X,y, test_size=0.2, random_state=0)

In [25]: print(X_train.shape)
         print(y_train.shape)

         (1612, 4)
         (1612,)
```

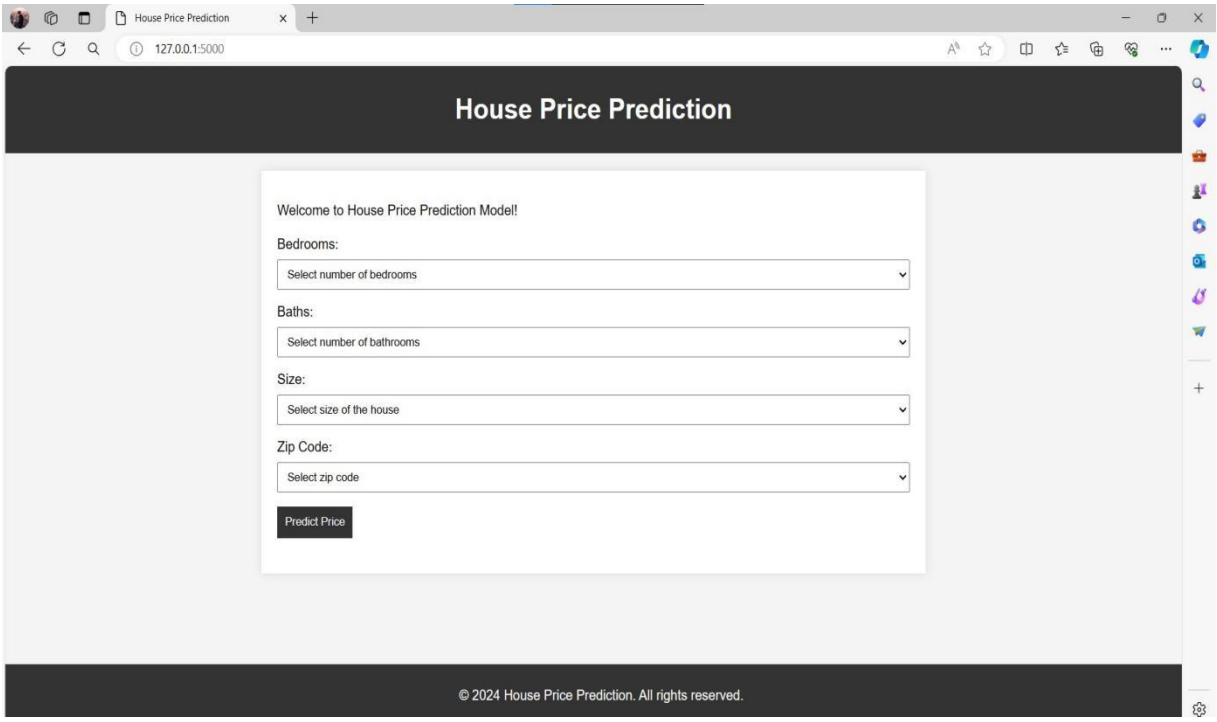
• **Model Evaluation:** After training the models, we evaluated their performance using the coefficient of determination (R-squared) metric. We made predictions on the testing set and calculated the R-squared score for each model.

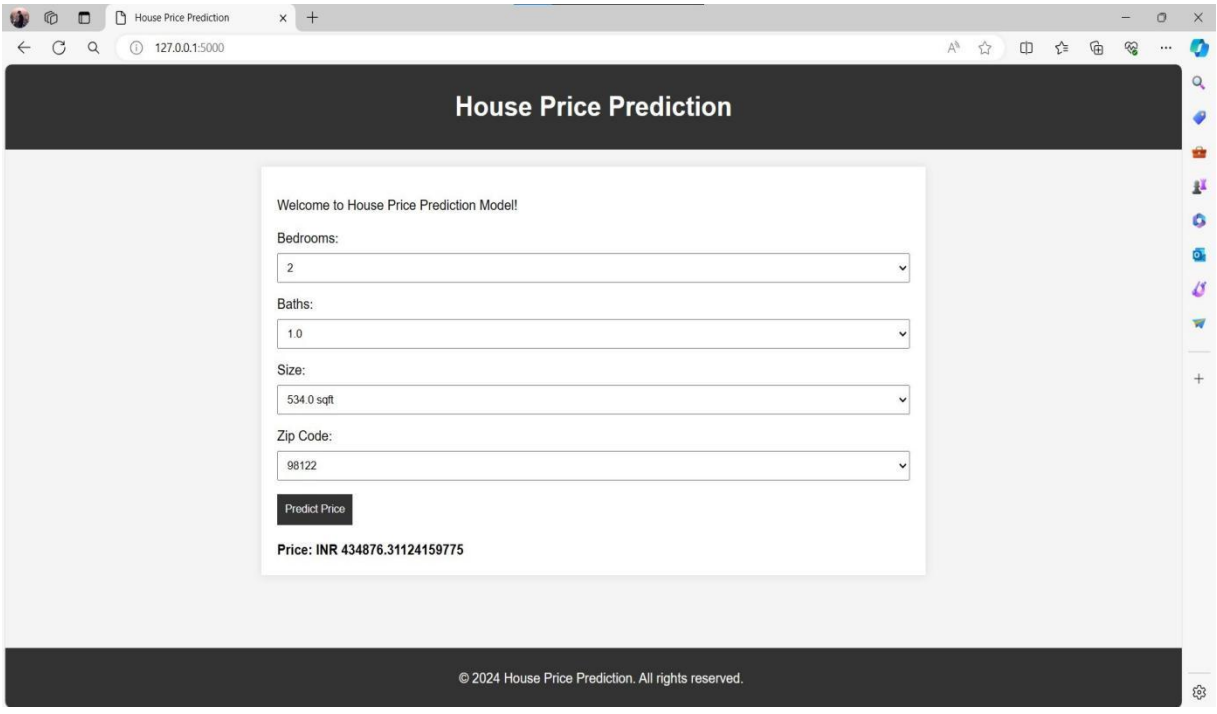
• **Web Application Development:** Finally, we developed a web application using the Flask framework to provide a user-friendly interface for predicting house prices. We implemented a form where users can input property details (e.g., number of bedrooms, bathrooms, size, zip code) and receive real-time price predictions based on the trained model.

```
from flask import Flask, render_template, request
import pandas as pd
import pickle

app = Flask(__name__)
data = pd.read_csv('final_dataset.csv')
pipe = pickle.load(open("RidgeModel.pkl", 'rb'))

@app.route('/')
def index():
    bedrooms = sorted(data['beds'].unique())
    bathrooms = sorted(data['baths'].unique())
    sizes = sorted(data['size'].unique())
    zip_codes = sorted(data['zip_code'].unique())

    return render_template('index.html', bedrooms=bedrooms, bathrooms=bathrooms, sizes=sizes, zip_codes=zip_codes)

@app.route('/predict', methods=['POST'])
def predict():
    bedrooms = request.form.get('beds')
    bathrooms = request.form.get('baths')
    size = request.form.get('size')
    zipcode = request.form.get('zip_code')

    # Convert input data to numeric types
    bedrooms = int(bedrooms)
    bathrooms = float(bathrooms)
    size = float(size)
    zipcode = int(zipcode)

    # Create a DataFrame with the input data
    input_data = pd.DataFrame([[bedrooms, bathrooms, size, zipcode]],
                              columns=['beds', 'baths', 'size', 'zip_code'])

    print("Input Data:")
    print(input_data)
```

4.2 Result Analysis:

The results of the experiment were analyzed in terms of model performance metrics and the functionality of the web application. Evaluation metrics such as R-squared score, MSE, and MAE were used to assess the accuracy of the machine learning models. Additionally, screenshots of the web application were captured to showcase its functionality and user interface.





5.Standards Adopted

5.1 Design Standards:

For the project design, the following recommended practices and standards are adopted:

- **Software Design:** Design principles outlined by IEEE and ISO standards are followed to ensure robustness, scalability, and maintainability of the software architecture.
- **UML Diagrams**: Unified Modeling Language (UML) diagrams, including use case diagrams, class diagrams, and sequence diagrams, are employed to visualize system architecture and design.

5.2 Coding Standards:

This project adheres to the following coding standards:

- **Conciseness:** Wrote concise codes with as few lines as possible to enhance readability and maintainability.
- **Naming Conventions:** Followed appropriate naming conventions for variables, functions, and classes to ensure code clarity and consistency.
- **Code Structure:** Segmented blocks of code into paragraphs to improve code organization and readability.
- **Indentation:** Used indentation to clearly mark the beginning and end of control structures, enhancing code readability and structure.
- **Modularization:** Avoided lengthy functions and adhered to the principle of modularity, ensuring that each function performs a single, well-defined task.

Conclusion

In conclusion, the House Price Prediction project successfully demonstrates the integration of machine learning techniques and web development to address the need for accurate property price predictions. Through efficient data preprocessing and the implementation of machine learning regression models, the system provides users with real-time price predictions based on property features. The development process adhered to industry standards and best practices, resulting in a reliable and user-friendly web application.

Future Scope

Moving forward, there are several opportunities for expanding the Bengaluru House Price Prediction project. Enhancements could include incorporating additional features and datasets to improve prediction accuracy, exploring advanced machine learning algorithms for better modeling, integrating geospatial analysis for spatial insights, improving user experience through interactive features, and scaling up deployment infrastructure for increased accessibility and performance. By pursuing these avenues, the project can continue to evolve and provide valuable insights for real estate stakeholders and property buyers in Bengaluru.

References

[1] https://www.kaggle.com/datasets/amitabhajoy/bengaluru-house-price-data

[2] https://www.geeksforgeeks.org/house-price-prediction-using-machine-learning-in-python/

[3] https://stackoverflow.com/

[4] https://github.com/suchit-insaid/Python-Machine-Learning-Projects

[5] https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML

[6] https://www.datacamp.com/blog/what-is-machine-learning

[7] https://en.wikipedia.org/wiki/Linear_regression

Bengaluru House Price P

ORIGINALITY REPORT

21%	18%	4%	21%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Queen Mary and Westfield College Student Paper	8%
2	Submitted to KIIT University Student Paper	7%
3	www.worldleadershipacademy.live Internet Source	6%

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		