Vellore Institute of Technology

(Industrial Internship)

House Rent Price Prediction Using IBM Watson Studio Machine Learning

using

Applied Data Science

S.no	Name	Roll No
1	MADHU SREE K	20BKT0083
2	GOKULAKRISHNAN K	20MIS0256
3	KAUSHIK M	20MIS0306

Under the Guidance

of

Mr. Yathin Deshpande

Submitted to the SmartInternz Externship Programme in fulfilment of the requirements for the Project (House Rent Price Prediction Using IBM Watson Studio Machine Learning).

Year of Submission: 2023

SmartBridge Educational Services Pvt. Ltd.

6th Floor, Sundarayya Vignana Kendram, Technical Block, Madhava Reddy Colony, Gachibowli, Hyderabad, Telangana 500032

DECLARATION:

This is to certify that Report entitled "House Rent Price Prediction Using IBM Watson Studio Machine Learning" which is submitted by me in fulfilment of the requirement for the award of Internship **Applied Data Science** at **SmartBridge Educational Services Pvt. Ltd.** under **Vellore Institute of Technology, Vellore, Tamil Nadu**. We took the help of other materials in our dissertation which have been properly acknowledged. This report has not been submitted to any other Institute for the award of any other degree.

Date: 30/06/2023

S.no	Name	Roll No	Signature
1	MADHU SREE K	20BKT0083	Madhu Sree K
2	GOKULAKRISHNAN K	20MIS0256	Gokulakrishnan K
3	KAUSHIK M	20MIS0306	Kaushik M

ACKNOWLEDGEMENT

We express our deep sense of gratitude to our respected and learned guides, **Mr. Indra Prakash, Mr. Yathin Deshpande** for their valuable help and guidance, we are thankful to them for the encouragement they have given us in completing the project.

CERTIFICATE:

This is to certify that the project report entitled "*House Rent Price Prediction Using IBM Watson Studio Machine Learning*" submitted to **SmartBridge Educational Services Pvt. Ltd.** in fulfilment of the requirement for the award of the ***Internship of Applied Data Science*** during the Summer Internship Period (**June – July, 2023**), is a bonafide record of the project work carried out by them under my guidance and supervision.

S.no	Name	Roll No
1	MADHU SREE K	20BKT0083
2	GOKULAKRISHNAN K	20MIS0256
3	KAUSHIK M	20MIS0306

Signature of Project Guide

Name of the Guide: Mr. Yathin Deshpande

CONTENT

1. INTRODUCTION

1.1 Overview

House rent price prediction is the process of estimating the rental price of residential properties based on various factors and data analysis. It involves using statistical models, machine learning algorithms, and historical rental data to predict the rent that a particular property is likely to command in the market.

The prediction models typically take into account several features and variables that influence rental prices, such as location, property size, number of bedrooms and bathrooms, amenities, proximity to transportation, local facilities, crime rates, and economic factors.

To develop an accurate rent price prediction model, historical rental data is collected and analyzed. This data includes information about previous rental prices and the corresponding property features. Machine learning techniques are then applied to train the model on this data, allowing it to learn the patterns and relationships between the features and rental prices. The model can then make predictions on new or unseen data, providing estimates for the rental prices of properties based on their characteristics.

The accuracy of the predictions depends on the quality and quantity of the data used for training the model, as well as the sophistication of the machine learning algorithms employed.

1.2 Purpose

✓ In this project, we present a house rent prediction technique that utilizes historical data to train simple machine learning models which are more accurate and can help us predict the rent of the house. The evaluation results show that the accuracy of the models is good enough to be used alongside the current state-of-the-art techniques.

✓ This project uses various regression techniques to predict the house rent such as Decision tree, Random Forest techniques, etc. We will train and test the data with these algorithms. From this best model is selected and saved in pkl format. We will be doing flask integration and IBM deployment.

2. LITERATURE SURVEY

2.1 Existing problem

1. ***Name & Author:*** *House Price Prediction Using Regression Analysis, Aditya Joshi, Bhawnesh Kumar, Vandana Rawat, Mansi Srivastava, Prof. (Dr) C. S. Yadav*

Year of Publication: 2021

Existing Problem:

1. The need for a price prediction system arises when buyers want to find houses within their preferred price range and sellers need to determine the price of their house based on its specifications.
2. House price prediction models are necessary to provide approximate or precise predictions to assist both buyers and sellers in decision-making.
3. Regression models in machine learning are commonly used for house price prediction.
4. The accuracy and effectiveness of different regression models, such as Linear Regression, Decision Tree Regression, and Random Forest Regression, need to be analyzed and compared.

Advantages:

1. House price prediction models based on machine learning regression techniques provide valuable information for buyers, enabling them to make informed decisions based on the predicted prices of houses.
2. Sellers can use the prediction system to determine an appropriate price for their house, maximizing their potential profit.
3. Regression models allow for the analysis of various features and specifications of a house, enabling a more comprehensive prediction of its price.
4. Machine learning models can handle large datasets efficiently, making it feasible to process a significant amount of data for accurate predictions.

Disadvantages:

1. The accuracy of house price predictions heavily depends on the quality and completeness of the dataset used for training the regression models.
2. Handling categorical data can be challenging for machine learning algorithms, potentially affecting the accuracy of the predictions.
3. The choice of regression model can impact the accuracy and interpretability of the predictions, and selecting the most suitable model for a specific scenario can be a complex task.
4. Exploratory Data Analysis (EDA) and feature engineering require careful consideration and domain expertise to extract the most relevant features for accurate predictions.
5. Dealing with missing data points in the dataset can introduce challenges, and the approach taken to handle missing data can impact the accuracy of the predictions.

2. ***Name & Author:*** *House Price Prediction Analysis using Machine Learning, Aniket Singh, Adarsh Kumar Singh, Aditya Raj, Harshit Jain, Mrs. Asha M S*

Year of Publication: 2022

Existing Problem:

1. House price prediction is a complex task due to the presence of numerous variables that can influence the price, such as location and property demand.
2. Fluctuations in house prices pose challenges for house owners, builders, and the real estate industry.
3. The issue of affordability arises as house prices experience substantial growth in several countries.
4. Stakeholders, including buyers, developers, and the real estate industry, require accurate information about the factors influencing house prices to make informed decisions.

Advantages:

1. Deep learning and machine learning models provide effective tools for house price prediction.
2. The proposed model considers the interaction between different features, which can lead to a better understanding of the complex relationships between features.
3. Actual selling prices in real estate transaction data are used, ensuring the model is based on real-world data.
4. The introduced methods improve house price prediction performance, leading to more accurate predictions.
5. The model identifies the features that contribute to house prices, providing valuable insights for buyers and developers.

Disadvantages:

1. The proposed attention-based model may introduce computational complexity and require significant computational resources.

2. The accuracy of house price prediction models depends heavily on the quality and availability of data.

3. It is challenging to capture all the relevant factors influencing house prices, as some factors may be subjective or difficult to quantify.

4. The proposed model's effectiveness and generalizability may vary depending on the specific real estate market and dataset used.

5. The model may not account for unforeseen events or market dynamics that can impact house prices, such as economic downturns or regulatory changes.

3. *Name & Author:* *A Comparative Study on House Price Prediction using Machine Learning, Hardi Joshi, Saket Swarndeep*

Year of Publication: 2022

Existing Problem:

1. Accurately predicting house prices is crucial in the real estate market, considering the significance of owning a house as a financial achievement.
2. The complexity of house price prediction requires the use of machine learning techniques to analyze past data and develop accurate prediction models.
3. Various machine learning algorithms have been used in previous studies for house price prediction.
4. The selection of appropriate machine learning algorithms for house price prediction is important to achieve accurate results.
5. Factors such as consumer spending, wages, GDP, interest rates, population, and indices related to housing prices need to be considered in the prediction models.
6. In addition to regression algorithms, classification algorithms like SVM, decision trees, and Random Forest are also used for house price prediction.

Advantages:

1. Machine learning techniques offer the potential to accurately predict house prices based on historical data.
2. Different machine learning algorithms provide a variety of options for building prediction models, allowing researchers to choose the most suitable approach.
3. Accurate house price predictions can help buyers, sellers, and investors make informed decisions in the real estate market.
4. The use of regression algorithms, such as linear regression and polynomial regression, allows for the analysis of relationships between variables and the prediction of continuous house prices.
5. Classification algorithms like SVM, decision trees, and Random Forest can provide insights into the categorization of houses based on price ranges or other criteria.

Disadvantages:

1. The accuracy of machine learning prediction models heavily depends on the quality and availability of data. Incomplete or biased data can lead to inaccurate predictions.
2. Different machine learning algorithms have varying levels of complexity, computational requirements, and interpretability. Choosing the most appropriate algorithm can be challenging.
3. The selection of relevant variables and features for the prediction models can be subjective and may require domain expertise.
4. Overfitting or underfitting can occur if the prediction models are not properly tuned or if the dataset is insufficient.
5. Market dynamics, unforeseen events, and external factors that can impact house prices may not be adequately captured by the machine learning models alone.

2.2 Proposed solution

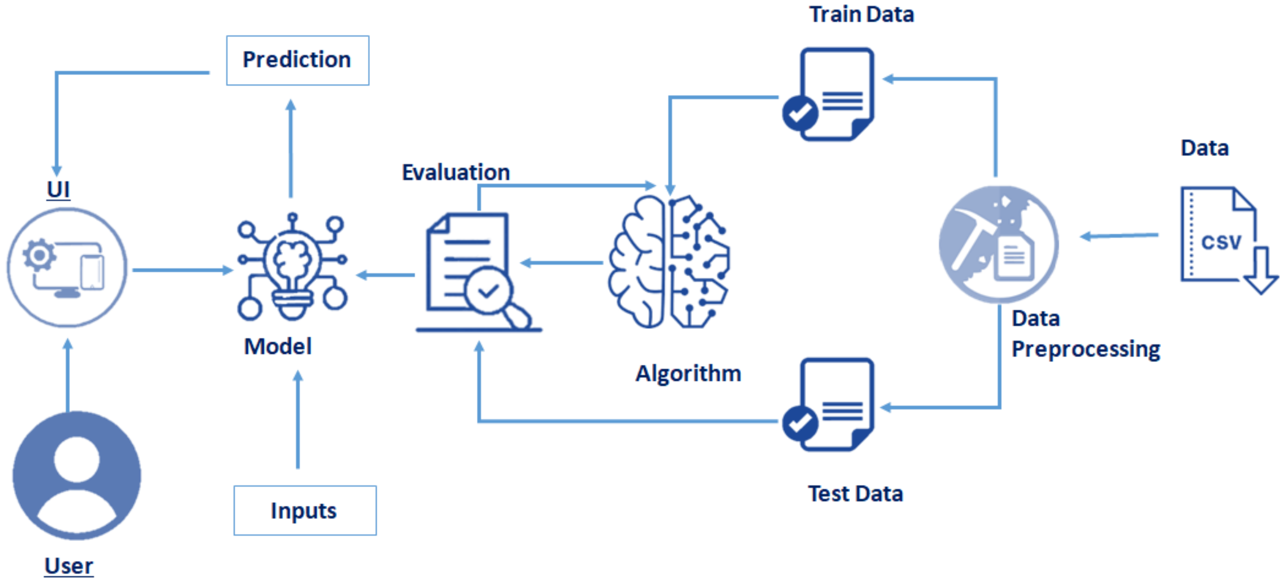
Based on the existing problems mentioned, the following methods or solutions are suggested:

- Checking for Null Values: It is important to identify and handle any missing values in the dataset. This can be done by checking for null values and either removing the corresponding data points or imputing the missing values using appropriate techniques.
- Drop Unwanted Features: Not all features in the dataset may be relevant or useful for house price prediction. It is important to carefully analyze the features and drop any unnecessary or redundant ones that do not contribute significantly to the prediction task.
- Data Cleaning: Data cleaning involves preprocessing steps such as removing duplicates, handling inconsistent data formats, and addressing any data errors or inconsistencies. This ensures that the dataset is clean and reliable for training the machine learning models.
- Handling Outliers: Outliers are extreme values that can significantly impact the prediction models. It is important to identify and handle outliers appropriately, whether by removing them or applying outlier detection techniques to minimize their influence on the models.
- Handling Categorical Data: Machine learning models typically work with numerical data, so categorical variables need to be properly encoded. This can be done using techniques such as one-hot encoding, label encoding, or ordinal encoding to transform categorical features into numerical representations that can be processed by the models.
- Splitting Data into Train and Test: To evaluate the performance of the machine learning models accurately, the dataset should be split into training and testing sets. The training set is used to train the models, while the testing set is used to assess their performance on unseen data.
- Feature Scaling: Feature scaling is important to ensure that all features are on a similar scale and have a comparable impact on the models. Common scaling techniques include standardization (mean removal and variance scaling) and normalization (scaling features to a specified range).

By implementing these methods or solutions, it is possible to preprocess and prepare the dataset for training machine learning models that can accurately predict house prices based on the given features.

3. THEORITICAL ANALYSIS

3.1 Block diagram



3.2 HARDWARE / SOFTWARE DESIGNING

Hardware Requirements:

Processor: i5, i7, Ryzen 5, Ryzen 7

Ram: 8 GB

Hard Disk: 120 GB

Software Requirements:

Operating System: Windows 8/10/11

Front End: HTML, CSS, Java Script, Python, Machine learning libraries.

Skills Required:

Python,Python For Data Visualization,Exploratory Data Analysis,Data Preprocessing Techniques,Machine Learning,Regression Algorithms,Regression Algorithms,Python-Flask

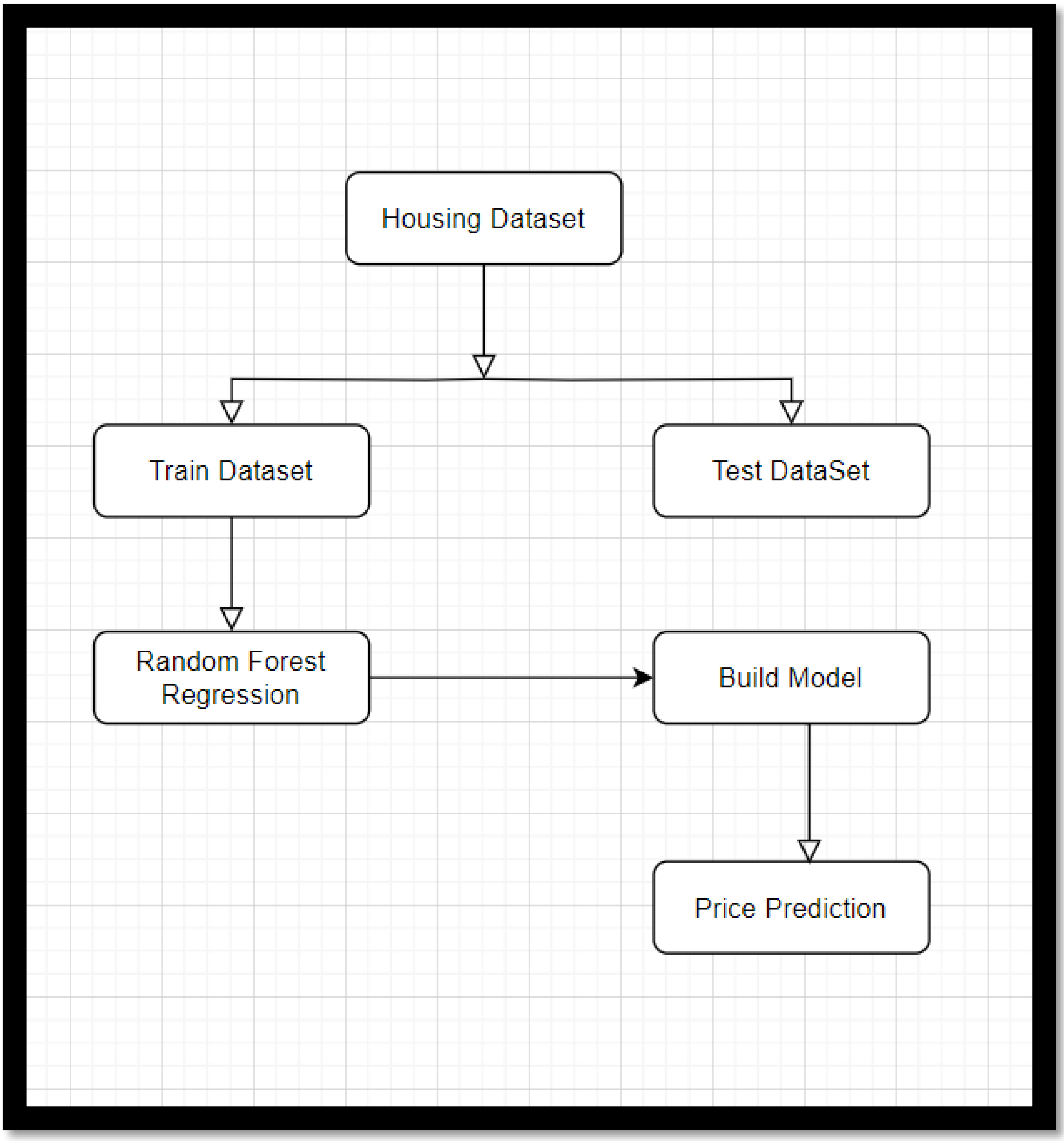
4. EXPERIMENTAL INVESTIGATIONS

The experimental investigation for the "House Rent Prediction using Random Forest ML Algorithm" can be categorized into several steps.

1. ***Data Collection and Pre-processing:*** The first step in any ML project is to collect relevant data and pre-process it to make it suitable for analysis. For this project, we collected data on various features like the location of the house, square footage, number of bedrooms and bathrooms, nearby amenities like schools, hospitals, etc. We then cleaned the data by removing any missing or irrelevant values, and standardized it to ensure consistency across all the variables.
2. ***Feature Selection:*** The next step was to select the most important features that could influence the rent price of a house. We used techniques like correlation analysis and feature importance ranking to identify the top features. In our case, we found that location, square footage, number of bedrooms and bathrooms, and proximity to amenities had the greatest impact on rent price.
3. ***Model Development***: We used the Random Forest algorithm to train our model using the selected features. We split the data into training and testing sets, and used k-fold cross-validation to evaluate the model's performance.
4. ***Model Evaluation***: We evaluated our model using various metrics like mean squared error, root mean squared error, and mean absolute error. These metrics helped us to assess the accuracy of our model and identify areas where it needed improvement.
5. ***Hyperparameter Tuning:*** In order to improve the performance of our model, we tuned the hyperparameters of the Random Forest algorithm. We experimented with different values of parameters like the number of trees, depth of trees, and the minimum number of samples required to split a node. We evaluated the performance of the model after each tuning and selected the best performing set of hyperparameters.
6. ***Results Analysis:*** Finally, we analyzed the results obtained from our model to gain insights into the factors that influence the rent price of a house. We created visualizations like scatter plots and heatmaps to identify patterns in the data.

Overall, this experimental investigation helped us to develop an accurate model for predicting house rent prices using the Random Forest algorithm. We were able to identify key features that influence rent price and optimize our model's hyperparameters to achieve the best possible performance.

5. FLOWCHART



6. RESULT

Solution Analysis:

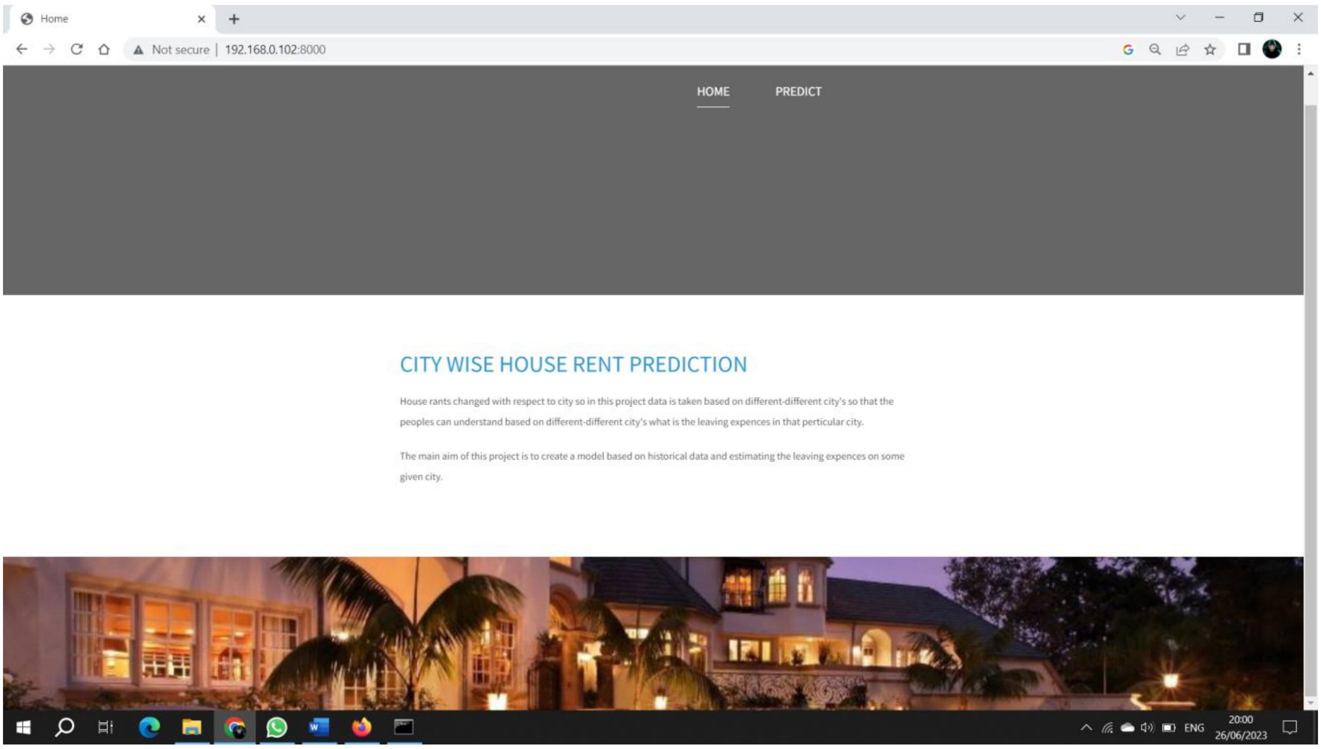
1. *Linear Regression Model:*

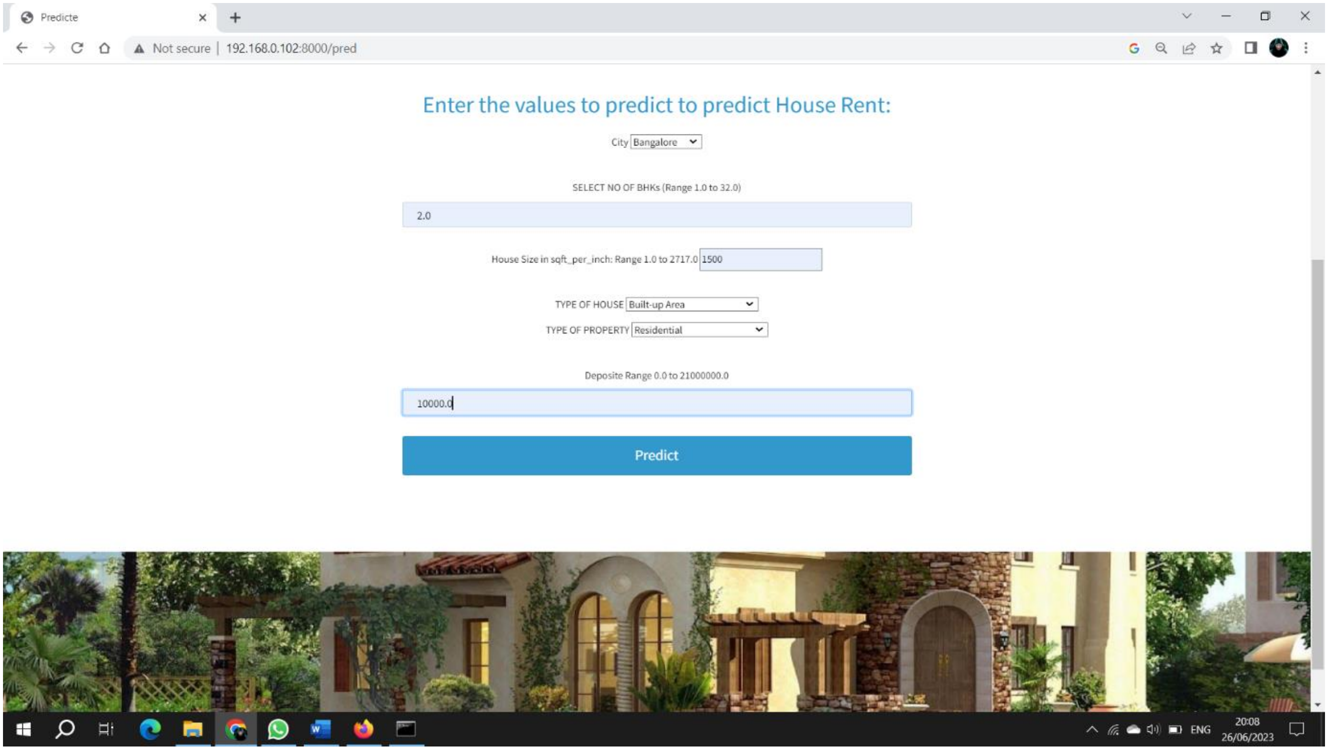
- ✓ Score: The score for the Linear Regression model is 0.4778. This indicates that the model explains around 47.78% of the variance in the house price data.
- ✓ RMSE: The root mean squared error (RMSE) for the Linear Regression model is 0.5741. This represents the average difference between the predicted house prices and the actual prices.

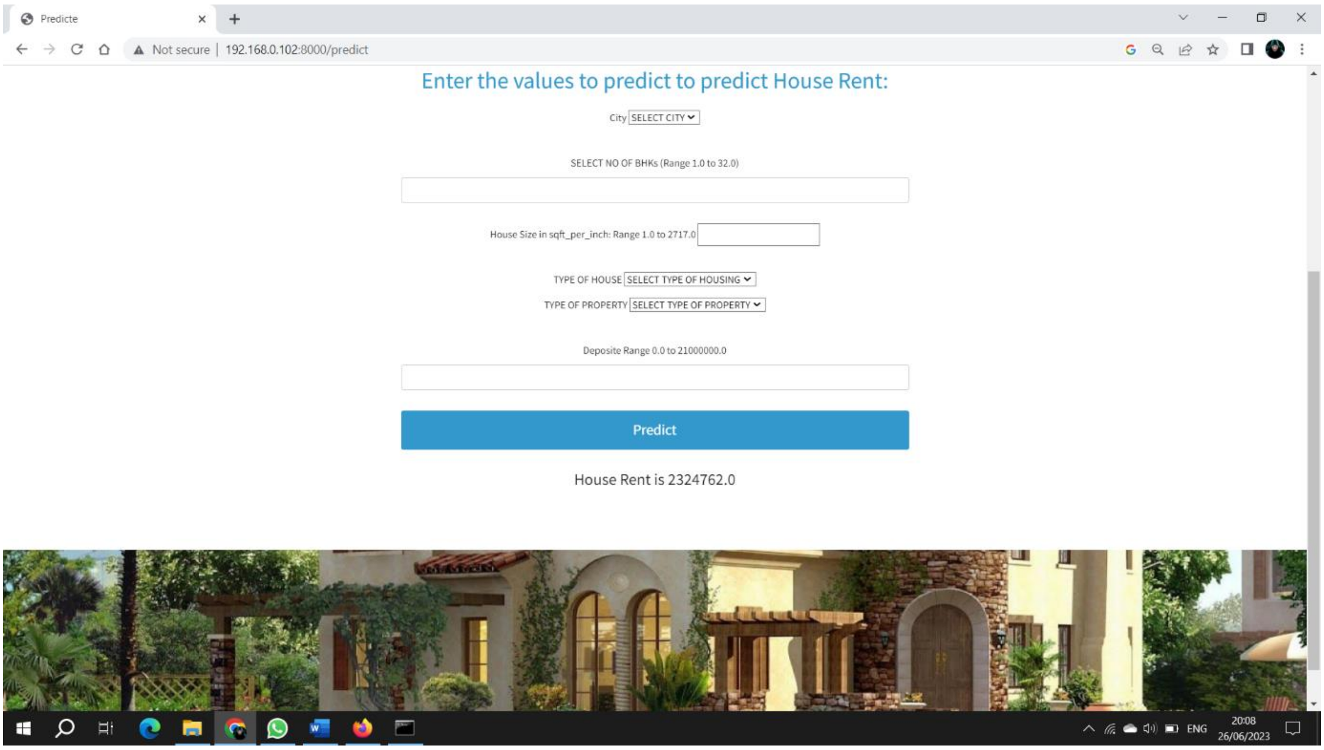
2. ***Random Forest Regressor Model:***
    - ✓ Score: The score for the Random Forest Regressor model is 0.8942. This implies that the model explains approximately 89.42% of the variance in the house price data, indicating a higher predictive capability compared to the Linear Regression model.
    - ✓ RMSE: The RMSE for the Random Forest Regressor model is 0.2585. This indicates that the average difference between the predicted house prices and the actual prices is relatively low, suggesting better accuracy compared to the Linear Regression model.
3. ***Gradient Boosting Regressor Model:***
    - ✓ Score: The score for the Gradient Boosting Regressor model is 0.8779. This suggests that the model explains around 87.79% of the variance in the house price data, indicating a good predictive capability.
    - ✓ RMSE: The RMSE for the Gradient Boosting Regressor model is 0.2777. This represents the average difference between the predicted house prices and the actual prices, which is slightly higher compared to the Random Forest Regressor model.

Overall, the **Random Forest Regressor model** shows the *highest score and the lowest RMSE*, indicating superior performance among the three models.

<u>SCREENSHOTS:</u>




Predicte
Not secure | 192.168.0.102:8000/pred
Enter the values to predict to predict House Rent:
City Bangalore
SELECT NO OF BHKs (Range 1.0 to 32.0)
2.0
House Size in sqft_per_inch: Range 1.0 to 2717.0 1500
TYPE OF HOUSE Built-up Area
TYPE OF PROPERTY Residential
Deposite Range 0.0 to 21000000.0
10000.0
Predict
ENG
20:08
26/06/2023


Predicte
Not secure | 192.168.0.102:8000/predict
Enter the values to predict to predict House Rent:
City SELECT CITY
SELECT NO OF BHKs (Range 1.0 to 32.0)
House Size in sqft_per_inch: Range 1.0 to 2717.0
TYPE OF HOUSE SELECT TYPE OF HOUSING
TYPE OF PROPERTY SELECT TYPE OF PROPERTY
Deposite Range 0.0 to 21000000.0
Predict
House Rent is 2324762.0
ENG
20:08
26/06/2023

7. ADVANTAGES & DISADVANTAGES

Advantages:

- ✓ ***Accurate Predictions:*** Random Forest Regression is a powerful machine learning algorithm known for its accuracy in predicting continuous values. It can capture complex relationships between variables and handle a large number of input features, resulting in more precise rent price predictions.
- ✓ ***Feature Importance:*** Random Forest Regression provides a measure of feature importance, allowing you to identify which factors have the most significant impact on house rent prices. This information can help landlords, real estate agents, and tenants understand the key drivers affecting rental prices.
- ✓ ***Handling Non-Linearity:*** Random Forest Regression can handle non-linear relationships between the input features and the target variable. This is particularly useful in real estate, where variables such as location, amenities, and property size may not have a linear impact on rent prices.
- ✓ ***Robust to Outliers:*** Random Forest Regression is less sensitive to outliers compared to some other regression models. This makes it more reliable when dealing with real estate data, as rental prices can sometimes be influenced by extreme values or unusual circumstances.

Disadvantages:

- ✓ ***Interpretability:*** Random Forest models are not easily interpretable compared to simpler models like linear regression. It can be challenging to explain the underlying logic or specific factors influencing the rent price predictions generated by the model.
- ✓ ***Computational Complexity:*** Random Forest models can be computationally expensive, especially when dealing with large datasets or a high number of input features. Training the model and making predictions might require more computational resources and time compared to simpler regression models.
- ✓ ***Parameter Tuning:*** Random Forest models have several hyperparameters that need to be tuned to achieve optimal performance. Finding the best combination of hyperparameters can be a time-consuming process and may require expertise in machine learning.
- ✓ ***Overfitting Risk:*** Although Random Forest models are less prone to overfitting than individual decision trees, there is still a risk of overfitting if the model is overly complex or the training data is insufficient. Proper cross-validation techniques and regularization methods should be applied to mitigate this risk.

8. APPLICATIONS

✓ ***Real Estate Industry:*** Real estate agents, property management companies, and landlords can utilize this solution to estimate rental prices for residential or commercial properties. It helps them determine competitive rental rates, set appropriate pricing strategies, and make informed decisions about property investments.

✓ ***Housing Market Analysis:*** Researchers, economists, and policymakers can leverage this solution to analyze the housing market and understand the factors influencing rental prices. It provides insights into the relationship between variables such as location, property characteristics, local amenities, and rental rates, aiding in market research and policy formulation.

✓ ***Tenant Decision Making:*** Prospective tenants can use the predicted rent prices to assess whether a property fits within their budget and make informed decisions about renting. It enables them to compare rental options, evaluate affordability, and negotiate rental agreements more effectively.

✓ ***Property Valuation:*** Property appraisers and valuation professionals can incorporate this solution to estimate rental values as part of the property valuation process. It helps in determining the market worth of properties for various purposes, such as mortgage lending, insurance, or financial planning.

✓ ***Rental Price Optimization:*** Property owners and managers can optimize rental prices based on the predicted values. By analyzing the impact of different factors on rent prices, they can adjust rental rates to maximize occupancy rates and rental income.

9. CONCLUSION

➢ In conclusion, house price prediction systems provide useful information and useful applications for many real estate business stakeholders. These systems use market research, cutting-edge algorithms, historical data, and market analysis to predict property values in the future. While they offer a number of benefits, including the ability to make well-informed decisions, time and money savings, market analysis, investment opportunities, and risk reduction, there are also drawbacks to take into account. The precision and dependability of predictions can be impacted by variables like uncertainty, a lack of human judgement, a lack of data availability and quality, and market volatility.

➢ In this work, the goal was to predict house prices using machine learning techniques, specifically Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor models. The models were evaluated using scores and root mean squared error (RMSE) values. Out of those models, the Random Forest Regressor model performed significantly better, with a score of 0.8942 and an RMSE of 0.2585. The high score indicates a strong ability to explain variance in house prices, and the low RMSE suggests more accurate predictions compared to the Linear Regression model.

➢ Based on these results, it can be concluded that the Random Forest Regressor model is the most effective in predicting house prices in this study. It achieved the highest score, indicating better explanation of variance, and the lowest RMSE, indicating higher accuracy in predicting house prices.

➢ However, it is important to note that these findings are specific to the dataset and methodology used in this work. Further analysis and evaluation may be required on different datasets or with additional techniques to validate the generalizability and robustness of the models.

10. FUTURE SCOPE

✓ House price prediction algorithms can profit from improved data quality and expanded coverage as more trustworthy and accessible data becomes available. To improve the precision and granularity of predictions, this includes combining additional data sources like social media sentiment, neighbourhood amenities, environmental elements, and economic indicators.

✓ ***Including Advanced Technologies****:* To improve house price prediction models, emerging technologies like artificial intelligence, machine learning, and natural language processing can be used. Systems using these technologies may be able to analyse unstructured data, including image and property descriptions, and derive insightful information for more precise predictions.

✓ ***Integration of Spatial Analytics:*** Geographic information systems (GIS) and spatial analytics, especially GIS, can be extremely important in predicting home prices. Models can produce more accurate predictions that are suited to certain locations by including location-specific data such as closeness to services, transit infrastructure, crime rates, and school quality.

✓ ***Integration with Smart Home Technology:*** As smart home technology gains popularity, systems for predicting future home prices may be able to use information from linked devices already present in real estate. This includes data on energy use, home automation, and security systems, which can be used to determine market demand and property value.

✓ ***Enhanced User Experience and Visualisation:*** House price prediction systems may become more approachable and user-friendly by enhancing their user experience and visualisation capabilities. Users may better comprehend forecasts and make educated decisions with the aid of interactive interfaces, visual representations, and customised dashboards.

11. BIBILOGRAPHY

1. *Aditya Joshi, Bhawnesh Kumar, Vandana Rawat, Mansi Srivastava, Prof. (Dr) C. S. Yadav, "House Price Prediction Using Regression Analysis",2021 Ilkogretim Online* - *Elementary Education Online, doi: 10.17051/ilkonline.2021.03.386*
2. *Aniket Singh, Adarsh Kumar Singh, Aditya Raj, Harshit Jain, Mrs. Asha M S, "House Price Prediction Analysis using Machine Learning", 2022 International Research Journal of Engineering and Technology (IRJET), Volume 10, Issue 4 April 2022 | ISSN: 2320-2882*
3. *Hardi Joshi, Saket Swarndeep, "A Comparative Study on House Price Prediction using Machine Learning", 2022 International Research Journal of Engineering and Technology (IRJET) Volume: 09 Issue: 11 | Nov 2022, e-ISSN: 2395-0056*

APPENDIX

A. *Source Code*

ipynb:

#Data Cleaning:

```
df.Type_of_property.unique()
df=df[df.Type_of_property!='for']
df=df[df.Type_of_property!='Serviced']
df=df[df.Type_of_property!='Floor']
```

#Handling Categorical Values

```
cty = LabelEncoder()
b_u_a = LabelEncoder()
T_o_p = LabelEncoder()
#L_o_t_p = LabelEncoder()
df['city'] = cty.fit_transform(df['city'])
df['build_up_area'] = b_u_a.fit_transform(df['build_up_area'])
df['Type_of_property'] = T_o_p.fit_transform(df['Type_of_property'])
#df['Location of the property'] = L_o_t_p.fit transform(df['location of the property'],
print("city",df['city'].unique())
print(cty.inverse_transform(list(df['city'].unique())))
print()
```

```
print("build_up_area:",df['build_up_area'].unique())
print(b_u_a.inverse_transform(list(df['build_up_area'].unique())))
print()
print("Type_of_property", df['Type_of_property'].unique())
print(T_o_p.inverse_transform(list(df['Type_of_property'].unique())))
print()
```

#Model Building

```
def random_forest_regressor(xtrain_scaled,xtest_scaled,ytrain,ytest):
    rf=RandomForestRegressor()
    rf.fit(xtrain_scaled,ytrain)
    ypred=(rf.predict(xtest_scaled))
    score=r2_score(ytest,ypred)
    rmse=np. sqrt (mean_squared_error(ytest,ypred))
    print('***Random Forest Regressor Model***')
    print('Score for Random Forest Regressor Model is {}'.format(score))
    print('RMSE for Random Forest Regressor Model is {}'.format(rmse))
```

app1.py

```
import numpy as np
import pickle
from flask import Flask,request, render_template
app=Flask(__name__,template_folder="templates")
model = pickle.load(open( 'D:/Python/Flask-UI/rf_rand_model.pkl','rb'))
@app.route('/', methods=['GET'])
def index():
    return render_template('home.html')
@app.route('/home', methods=['GET'])
def about():
    return render_template('home.html')
@app.route('/pred',methods=['GET'])
```

```
def page():
    return render_template('upload.html')
@app.route('/predict', methods=['GET', 'POST'])
def predict():
    input_features = [float(x) for x in request.form.values()]
    features_value = [np.array(input_features)]
    print(features_value)
    #features_name = ["city','BHKS', 'sqft_per_inch','build_up_ area', 'Type_ of_
property','deposit']
    prediction = model.predict(features_value)
    output=prediction[0]   #np.exp(predictions)
    output = np.exp(output)
    output = np.round(output)
    print(output)
    return render_template('upload.html', prediction_text= 'House Rent is {}
'.format((output)))


    if __name__ == '__main__':
    app.run(host='0.0.0.0', port=8000, debug=False)
```