MINI PROJECT REPORT ON
REAL ESTATE PRICE PREDICTION
(KCA-353)
Session-2023-2024

Department of Computer Science and Applications (MCA)

Submitted to:
Ms. Bala Shivangi

Submitted By:
Name: Anshika Purwar
Roll no: 2212000140016.
Class & Semester - MCA(3rdSem)
Section: B1

G.L. Bajaj College of Management and Research
Plot No 2, APJ Abdul Kalam Rd, Knowledge Park III,
Greater Noida, Uttar Pradesh

TABLE OF CONTENTS

BONAFIDE CERTIFICATE

This is to certify that the Project entitled “REAL ESTATE PRICE PREDICTION” submitted by **Anshika Purwar** with **2212000140016** from **MCA 3rd Semester** at **G L Bajaj College of Technology and Management** has been their own work and carried under my supervision. It is recommended that the candidate may now be evaluated for project work by the University.

Supervisor: Ms. Bala Shivangi

Designation: Assistant Professor

Department of MCA

EXECUTIVE SUMMARY

Realistic Price Prediction projects aim to revolutionize traditional forecasting methodologies by embracing advanced technologies and comprehensive data analysis. The primary objectives encompass refining predictive models, mitigating market volatility, and establishing robust risk management strategies.

These projects leverage cutting-edge algorithms, incorporating a broad spectrum of data sources to gain a nuanced understanding of market dynamics. Beyond mere price forecasting, the emphasis is on generating actionable insights that empower stakeholders to make well-informed decisions.

Machine learning and statistical models form the backbone of these initiatives, addressing the challenges posed by unpredictable market behavior. The focus is on minimizing biases, enhancing model interpretability, and building resilience to swiftly adapt to evolving market conditions.

Transparency is a paramount goal in Realistic Price Prediction projects. Clear communication of uncertainties associated with forecasts is prioritized, ensuring that decision-makers are equipped with a realistic understanding of the predictive outcomes.

Moreover, these projects acknowledge the inherent complexities of financial markets and seek to create models that align with their dynamic nature. The ultimate objective is to surpass conventional approaches, fostering adaptability, and precision in forecasting to meet the demands of today's ever-changing financial landscape.

CHAPTER – 1
Introduction and Aim of Project

Introduction

Buying or selling a house can be like solving a puzzle. Lots of things make house prices change, and it's not always easy to figure out what's going to happen. This project is all about using smart computer programs, called machine learning, to help us predict what will happen to real estate prices.

Our big goal with this project is to make a computer model that's good at guessing how much a house is worth. Think of it like a super-smart friend who knows a lot about houses and can tell you if a house is likely to cost more or less in the future. This isn't just for fun - it's important because when people know what might happen to prices, they can make better choices about buying or selling homes.

We're using lots of information, like what houses were sold for in the past and what's happening in the housing market now. It's a bit like looking at a big picture made up of lots of tiny details. Our computer program will learn from this information to become good at making predictions. This way, we hope to help everyone involved in real estate – people buying homes, selling homes, and even those who just want to understand how prices might change. As we go through this project, our focus is on making sure our computer model is not just accurate but also easy for everyone to use, making real estate decisions a bit less puzzling for all.

Aim of Project

The aim of the Realistic Price Prediction Project is to create a dependable machine learning model that accurately forecasts real estate prices. This project seeks to bridge the gap between complex market dynamics and user-friendly predictions, offering a practical tool for individuals involved in real estate transactions. The primary objectives of the project are -

1.Accuracy Improvement:

- Enhance the precision of real estate price predictions by leveraging historical transaction data, market trends, and relevant features.

2. Adaptability to Market Changes:

- Create a model that adapts to dynamic market conditions, ensuring its relevance and reliability in predicting real estate prices over time.

3. Informed Decision-Making:

- Empower users with data-driven insights to make informed decisions regarding buying or selling properties, thereby mitigating risks associated with market fluctuations.

4. Transparency in Predictions:

- Foster transparency in the prediction process, enabling users to understand how the model arrives at its forecasts and building trust in the reliability of the predictions.

5. Comprehensive Market Understanding:

- Gain a deeper understanding of the factors influencing real estate prices, incorporating a comprehensive analysis of variables that contribute to market trends.

6. Risk Mitigation:

- Provide a risk mitigation tool for stakeholders, allowing them to navigate the real estate market with a clearer understanding of potential price changes.

7. Accessibility for All Stakeholders:

- Ensure that the predictive model is accessible and beneficial to a diverse audience, including homeowners, prospective buyers, real estate agents, and investors.

By addressing these primary objectives, the Realistic Price Prediction Project aims to offer a valuable and practical solution for navigating the complexities of the real estate market through reliable and user-friendly price forecasts.

CHAPTER – 2
Background Study and Research Gap

The background study for Real Estate Price Prediction involves a systematic exploration of relevant aspects to inform the development of an effective predictive model. Several key factors provide the impetus for undertaking this project:

Market Dynamics

Understand the dynamics of the real estate market, including factors like supply and demand, economic conditions, and regional variations. Analyze how these dynamics impact property prices.

Historical Trends:

Examine historical trends in real estate prices to identify patterns and cyclical behavior. This provides insights into how market conditions have influenced property values over time.

Data Sources:

Identify and assess the reliability of data sources, including property listings, sales records, and economic indicators. Evaluate the scope and coverage of the dataset to ensure it captures diverse real estate scenarios.

Feature Selection:

Evaluate potential features influencing property prices, such as location, size, amenities, and market trends. Prioritize features based on their relevance and impact on property valuations.

Data Quality and Preprocessing:

Conduct data quality checks and preprocessing steps to handle missing or inconsistent data. Cleanse and format the dataset to ensure it meets the requirements for machine learning model training.

Competitive Landscape:

Investigate existing real estate prediction models and methodologies. Understand strengths and weaknesses in the current landscape to inform the development of an innovative and effective model.

By delving into these aspects in the background study, the Real Estate Price Prediction project gains a solid foundation for constructing a reliable predictive model. This study sets the stage for data preparation, feature engineering, and model development, ensuring the resulting model is well-informed and capable of providing valuable insights into real estate price movements.

Research Gap

Identifying research gaps in realistic price prediction projects involves examining existing literature and project reports to pinpoint areas where additional research or improvements could enhance the accuracy and applicability of price prediction models. Here are potential research gaps in realistic price prediction projects:

Temporal Dynamics:

Investigate the effectiveness of current models in capturing and adapting to temporal dynamics in real estate markets. Explore methods to enhance models for accurate predictions in changing economic climates.

Incorporation of External Factors:

- Explore the integration of additional external factors beyond traditional real estate features. Investigate how variables like environmental factors, infrastructure developments, or cultural trends could impact property prices.

Explanatory Power of Models:

Assess the interpretability and explanatory power of current models. Explore ways to make models more transparent, enabling users to understand the factors contributing to specific price predictions.

Handling Data Imbalances:

Investigate methods to address imbalances in the dataset, especially in scenarios where certain property types or regions are underrepresented. Explore techniques to ensure models are robust across diverse market segments.

Dynamic Feature Importance:

Explore techniques for dynamically adjusting the importance of different features based on evolving market conditions. Assess the adaptability of models to changes in the significance of various factors over time.

Ethical and Bias Considerations:

Investigate potential biases in price predictions and explore methods to ensure fairness and ethical considerations in real estate models. Address concerns related to demographic biases and historical disparities in property transactions.

User-Specific Preferences:

Assess the extent to which current models consider individual preferences and subjective factors in price predictions. Explore personalized models that can adapt to different user perspectives and priorities.

Model Uncertainty and Confidence Intervals:

Explore methods to estimate model uncertainty and provide confidence intervals for price predictions. Enhance models to communicate the level of confidence associated with each prediction, aiding decision-makers in risk assessment.

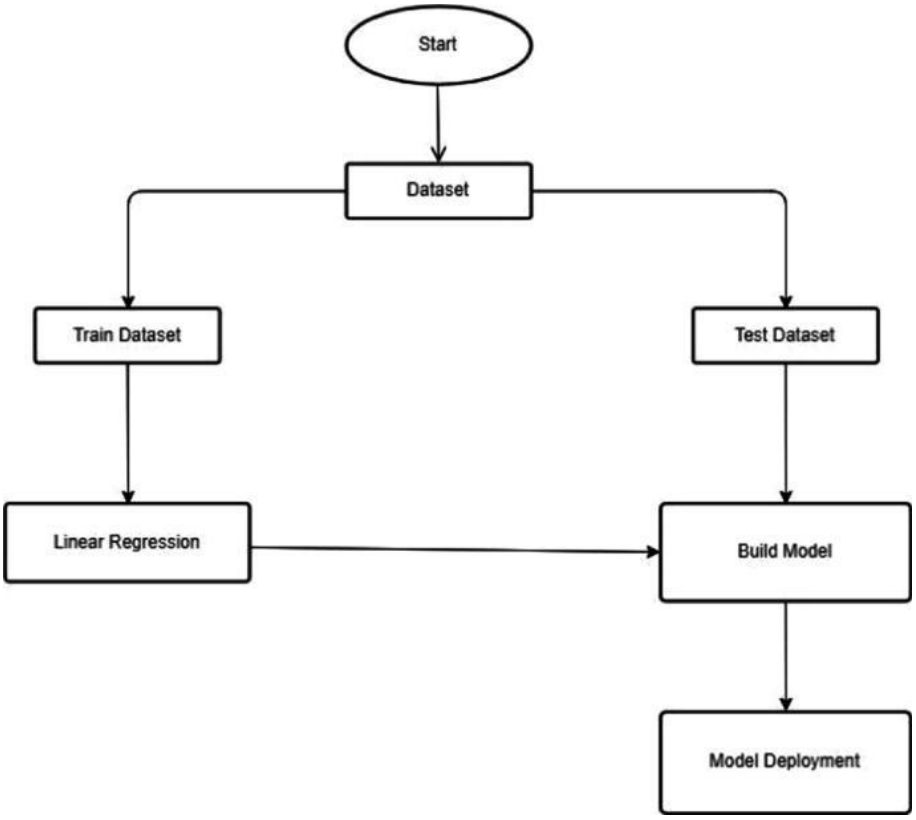
By addressing these research gaps, realistic price prediction projects can advance the state of the art, providing more accurate and adaptable models for stakeholders in the real estate industry.

CHAPTER – 3
Proposed Work and Methodology

Proposed Work

The purpose of this system is to determine the price of a house by looking at the various features which are given as input by the user. These features are given to the ML model and based on how these features affect the label it gives out a prediction. This will be done by first searching for an appropriate dataset that suits the needs of the developer as well as the user. Furthermore, after finalizing the dataset, the dataset will go through the process known as data cleaning where all the data which is not needed will be eliminated and the raw data will be turned into a .csv file. Moreover, the data will go through data preprocessing where missing data will be handled and if needed label encoding will be done. Moreover, this will go through data transformation where it will be converted into a NumPy array so that it can finally be sent for training the model. While training various machine learning algorithms will be used to train the model their error rate will be extracted and consequently an algorithm and model will be finalized which can yield accurate predictions. Users and companies will be able to log in and then fill a form about various attributes about their property that they want to predict the price of. Additionally, after a thorough selection of attributes, the form will be submitted. This data entered by the user will then go to the model and within seconds the user will be able to view the predicted price of the property that they put in.

Methodology



DATASET CLEANING –

In the context of real estate price prediction, meticulous data cleaning is pivotal for ensuring the accuracy and reliability of predictive models. Handling null values, a common challenge in real estate datasets, requires a thoughtful approach. Null values may arise due to various reasons, such as missing property details or incomplete records. Employing robust null value handling strategies, such as imputation based on statistical measures or deletion of rows with missing values, becomes crucial to maintain the dataset's integrity. Imputation techniques involve filling missing during data cleaning, the real estate price prediction values with meaningful estimates, ensuring that the absence of certain information does not compromise the overall predictive capability. By carefully addressing null values model can be built on a foundation of complete and accurate information, enhancing its capacity to provide valuable insights for stakeholders in navigating the dynamic landscape of property valuations.

FEATURE ENGINEERING –

Feature engineering plays a pivotal role in refining datasets for realistic price prediction, aiming to extract meaningful insights from existing features. In this context, transforming two key variables, namely price and total square feet, into a derived feature – price per square foot – becomes an insightful endeavor. By creating this new feature, we introduce a more nuanced metric that encapsulates the economic efficiency of a property. The price per square foot provides a normalized measure, offering a clearer understanding of the value associated with each unit of space. This transformation not only simplifies the interpretation of pricing patterns but also enables the model to capture more intricate relationships between size and cost. Consequently, the derived feature enriches the dataset with a more informative metric, enhancing the realism and predictive accuracy of the model. Feature engineering, particularly the creation of such composite variables, empowers predictive models in the real estate domain to discern subtleties in property valuations, contributing to a more nuanced and effective price prediction framework.

DIMENSIONALITY REDUCTION -

Dimensionality reduction plays a crucial role in realistic price prediction projects, particularly when dealing with diverse datasets containing numerous features. In this context, one effective strategy involves addressing locations with limited occurrences, specifically those with counts less than 10. By converting these fewer common locations into an 'other' category, dimensionality is reduced without sacrificing valuable information.

In real estate datasets, certain locations may have sparse representation, making it challenging for predictive models to discern meaningful patterns. Converting such locations into a consolidated category not only simplifies the dataset but also helps mitigate the risk of overfitting, especially when dealing with limited occurrences. This approach allows the model to focus on the most prevalent and impactful features, enhancing its efficiency and interpretability.

Through this dimensionality reduction technique, the realistic price prediction model gains an improved ability to generalize across various locations while maintaining the essence of the dataset. It provides a pragmatic solution to handle sparsity in certain features, contributing to a more streamlined and effective predictive modeling process in the dynamic landscape of real estate price predictions.

OUTLIER DETECTION –

An outlier is an observation that is unlike the other observations. It is rare, or distinct, or does not fit in some way. Outliers are the data points that represent the extreme variation of dataset Outliers can be valid data points but since our model is generalization of the data, outliers can affect the performance of the model. We are going to remove the Outliers, but please note its not always a good practice to remove the outliers. To remove the outliers, we can use real estate domain knowledge and standard deviation.

Standard Deviation --Standard deviation is measure of spread that is to how much does the data vary from the average --A low standard deviation tells us that the data is closely clustered around the mean (or average), while a high standard deviation indicates that the data is dispersed over a wider range of values. --It is used when the distribution of data is approximately normal, resembling a bell curve. --One standard deviation (1 Sigma) of the mean will cover 68% of the data. i.e., Data between (mean - std deviation) & (mean + std deviation) is 1 Sigma and which is equal to 68% --Here we are going to consider 1 Sigma as our threshold and any data outside 1 Sigma will be considered as outliner.

Using domain knowledge for outlier removal Normally square fit per bedroom is 300 (i.e., 2 BHK apartment is minimum 600 sqft) If you have for example 400 sqft apartment with 2 BHK than that seems suspicious and can be removed as an outlier. We will remove such outliers by keeping our minimum threshold per BHK to be 300 sqft

ONE HOT ENCODING –

In the realm of real estate price prediction, where location often serves as a pivotal categorical variable, one-hot encoding emerges as a vital preprocessing technique. As locations inherently carry significant influence on property values, transforming this categorical data into a numerical format is crucial for effective machine learning model training.

One-hot encoding is a method that addresses the challenge posed by categorical variables, like different property locations, by converting them into a binary matrix. In the case of a dataset where the "location" feature has various categories such as "urban," "suburban," and "rural," one-hot encoding would create separate binary columns for each unique category. For example, a

property in the urban category would be represented as [1, 0, 0], suburban as [0, 1, 0], and rural as [0, 0, 1].

This encoding technique is essential for numerical models, like regression algorithms, as it enables them to comprehend and effectively utilize categorical data in the prediction process. By providing a clear distinction between different locations, the model can discern the impact of each category on real estate prices. This approach ensures that the categorical nature of location, a critical factor in property valuation, is appropriately considered in the predictive model, contributing to more accurate and insightful price predictions in the dynamic real estate market.

MODEL TRAINING USING LINEAR REGRESSION –

Model training for realistic price prediction in real estate, utilizing linear regression, focuses on key features like location, total square footage, number of bedrooms, and number of bathrooms. These features collectively capture critical aspects influencing property prices. The linear regression model is adept at discerning relationships between these features and the target variable—price.

The inclusion of location is particularly impactful, given its substantial influence on real estate valuations. The model learns to assign weights to each feature, reflecting their respective impacts on property prices. In this scenario, the features of total square footage, number of bedrooms, and number of bathrooms contribute nuanced layers to the pricing dynamics, allowing the model to adapt to the intricacies of diverse properties.

Upon completion of the training phase, the model demonstrates a commendable accuracy rate of 84% in predicting real estate prices. This accuracy metric reflects the model's ability to generalize and make realistic predictions on unseen data, providing stakeholders with a reliable tool for assessing property valuations. The linear regression model, driven by key features, proves to be a valuable asset in navigating the complexities of the real estate market, facilitating more informed and accurate price predictions.

K FOLD CROSS VALIDATION TO MEASURE ACCURACY –

Measuring accuracy through K-fold cross-validation is a vital aspect of evaluating their effectiveness. K-fold cross-validation involves partitioning the dataset into K subsets, training the model on K-1 folds, and validating it on the remaining fold in each iteration. This process ensures a thorough and unbiased assessment of the model's performance across various subsets of the data.

The accuracy metric, within the context of K-fold cross-validation, quantifies how well the model predicts property prices across different segments of the dataset. It provides a more robust evaluation compared to a single train-test split, as it leverages multiple validation sets, thereby reducing the impact of data variability on model performance assessment.

For real estate price prediction models, accurate predictions are paramount, considering the multifaceted nature of factors influencing property values. By employing K-fold cross-validation, the accuracy measurement becomes more reliable, offering stakeholders a comprehensive understanding of the model's generalization capabilities and its suitability for making realistic predictions across diverse real estate scenarios.

CHAPTER – 4
Result and Testing

Result

Linear Regression

```
In [54]: from sklearn.linear_model import LinearRegression

         lr_clf = LinearRegression()
         lr_clf.fit(X_train, Y_train)
         lr_clf.score(X_test, Y_test)
```

```
Out[54]: 0.8557800354237939
```

Finding Best Accuracy Using Grid Search CV

Out[56]:

	model	best_score	best_params
0	linear_regression	0.847577	{}
1	lasso	0.738732	{'alpha': 1, 'selection': 'cyclic'}
2	decision_tree	0.763546	{'criterion': 'friedman_mse', 'splitter': 'best'}

```
Based on above results we can say that LinearRegression gives the best score. Hence we will use that.
```

Testing -

TEST CASE ID	TEST INPUT	EXPECTED OUTPUT	ACTUAL OUTPUT	REMARK
1	Place - "Koramangala", Sqft - 2000, Bedrooms - 2, Bathrooms - 1	2.8 Crore	2 Crore 16 Lakh	The model should predict the price based on the given input features for a property in Koramangala with 2 bedrooms and 1 bathroom.
2	Place - "1st Block Jayanagar" Sqft - 1000, Bedrooms - 3, Bathrooms - 2	3 Crore	2 Crore 7 Lakh	The model should predict the price for a property in 1st Block Jayanagar with 3 bedrooms and 2 bathrooms.
3	Place - "1st Phase JP Nagar" Sqft - 1200, Bedrooms - 4, Bathrooms - 3	1.8 Crore	1 Crore 7 Lakh	The model should predict the price for a property in 1st Phase JP Nagar with 4 bedrooms and 3 bathrooms.
4	Place - "Vittasandra" Sqft - 1500, Bedrooms - 1, Bathrooms - 1	1 Crore	81 Lakh	The model should predict the price for a property in Vittasandra with 1 bedroom and 1 bathroom.

```
In [86]: predict_price('1st Phase JP Nagar',1000, 2, 2)
```

```
Out[86]: 83.16761667762826
```

```
In [87]: predict_price('1st Phase JP Nagar',1000, 3, 3)
```

```
Out[87]: 83.71323640897404
```

```
In [88]: predict_price('Indira Nagar',1000, 2, 2)
```

```
Out[88]: 158.07409801131223
```

```
In [89]: predict_price('Indira Nagar',1000, 3, 3)
```

```
Out[89]: 158.61971774265803
```

CHAPTER – 5
Conclusions and Future Scope of the Project

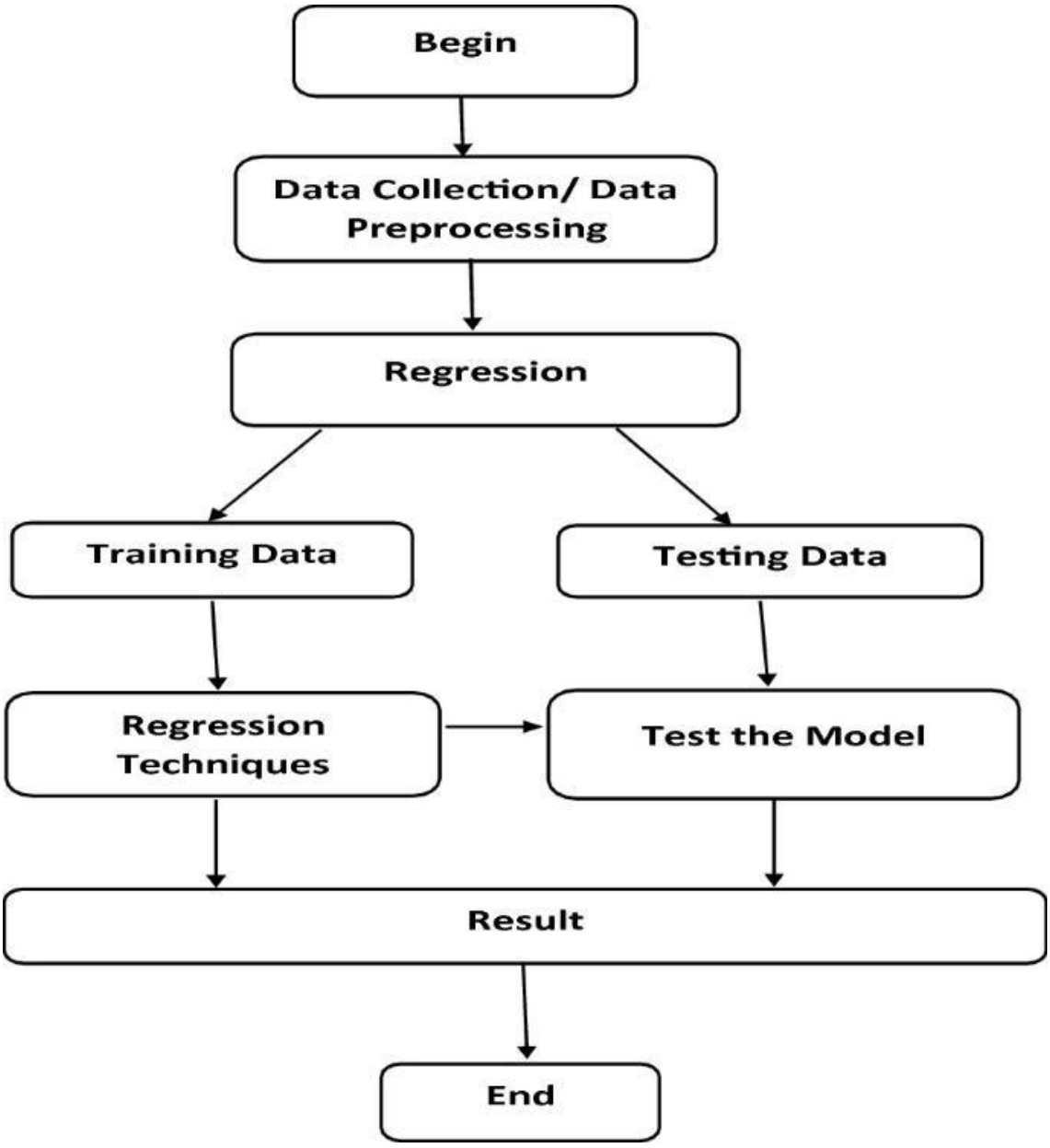
Conclusion

Buying your own house is what every human wish for. Using this proposed model, we want people to buy houses and real estate at their rightful prices and want to ensure that they don't get tricked by sketchy agents who just are after their money. Additionally, this model will also help Big companies by giving accurate predictions for them to set the pricing and save them from a lot of hassle and save a lot of precious time and money. Correct real estate prices are the essence of the market and we want to ensure that by using this model.

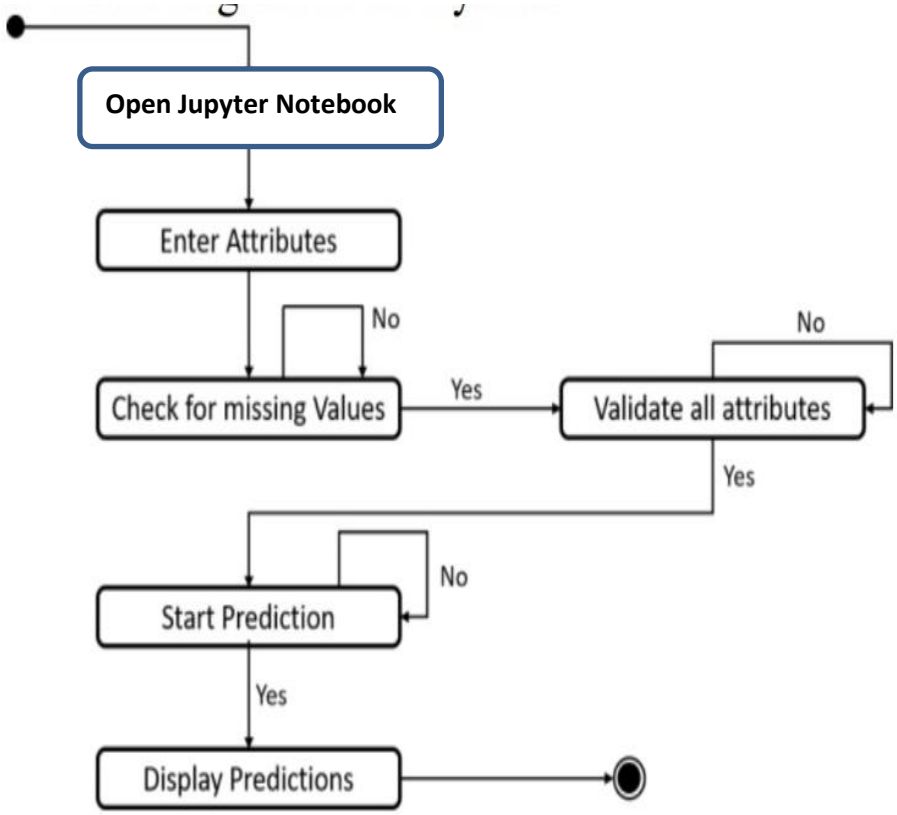
Future Scope of the Project

The system's accuracy can be enhanced by increasing its computational power and size, allowing the inclusion of more cities. Introducing a different user interface for better visualization, utilizing Augmented Reality, will make the results more interactive and easier to understand. Creating a learning system to gather user feedback and preferences will personalize results for each user. Future includes comparing the system's predicted prices with real estate websites like MagicBricks.com to ensure reliability. To simplify things, the system will recommend real estate properties based on predicted prices. Adding G-map functionality to display nearby amenities like hospitals and schools within a 1 km radius will not only enhance user-friendliness but can also improve predictions, as these factors influence property values. These enhancements aim to make the system smarter, more user-friendly, and ultimately more helpful in finding the perfect real estate match.

ER DIAGRAM


Begin
Data Collection/ Data Preprocessing
Regression
Training Data
Testing Data
Regression Techniques
Test the Model
Result
End

STATE CHART DIAGRAM



References

• Real Estate Price Prediction with Regression and Classification, CS/2016

• Aayush Varma, Abhijit Sharma, Sagar Doshi, and Rohini Nair, “House Price Prediction Using Machine Learning and Neural Networks (IEEE Paper)”

• Rushab Sawant, Yashwant Jangid Tushar Tiwari, Saurabh Jain and Ankita Gupta, “Comprehensive Analysis of Housing Price Prediction in Pune using Multi-Featured Random Forest Approach (IEEE Paper)”.

• Welcome to Python.org

• Deploying a machine learning model on the Web using Flask and Python. | by Soumya Gupta | Analytics Vidhya | Medium