House Price Prediction using Machine Learning

Karan Srivastava,
Department of computer science engineering,
Chandigarh University, Gharuan, Mohali, Punjab

Sahil Verma,
Uttaranchal University, Dehradun
Email: sahilverma@ieee.org

Mohammad Shuaib Khan,
Department of computer science engineering, Chandigarh University, Gharuan, Mohali, Punjab
Email: shuaibkhan.it@gmail.com

Ajay Singh
Uttaranchal University, Dehradun
profajaysingh@uumail.in

Abstract—Real estate values change daily, making it a un- predictable market. So our project is based on using machine learning to predict property prices. We sought to employ machine learning techniques in our study to predict more accurate housing values. We explore and analyze a variety of forecasting strategies that use the linear regression model because of its flexible housing table and probabilistic model selection techniques. Because of its model selection flexibility, we used lasso hindsight as our machine learning model. Machine learning evolved tremendously in recent years. Existing models have a variety of faults, including a lack of security and an inability to keep up with price fluctuations that occur often. It also had a major impact on the field of medicine and types of equipment. Our findings indicate that our approach to solving the problem is likely to be successful and that we can test theories against rising housing prices.

Index Terms—Data Cleaning, Outliers, Grid Search CV, Linear Regression

I. INTRODUCTION

Linear regression is basically the best algorithm known for statistics and plotting.

In statistics and machine learning, linear regression is a well-known and well-understood method. Guided learning is used in which the independent variables help to guide through the goal value. Which is mostly used in forecasting [5]

Predictive modelling basically reduces the error and tries to provide the most accurate solution. To achieve these applied machine learning goals, we'll employ approaches from a variety of domains, including statistics. We have an eclipse linear regression that takes the input variable and creates a relation with the output variable [23], [25].

The main purpose of this project is to create a project that will help people to get the accurate price of the real estate conveniently. We choose linear regression as our model due to its adaptive housing table and probabilistic model selection method. Based on the results that we got from our projects we found that we are able to generate a cost equivalent to what real estate offers

Models while the housing cost estimates, improve real estate policy. [1]

II. RELATED WORK

A great deal of research is being conducted on machine learning-based home price forecast models. Before the construction of this model, a significant amount of research and study was undertaken and studied. In their study, Zulkifli et al [1] developed a prediction model for the n local, structural, neighbourhood, and economic components. He experimented with several machine learning techniques to order to find the best model, such as linear regression, gradient boost, and so on. Truong et al [2] attempted to

create a model for predicting property prices. His approach was hybrid regression. To anticipate the price, the approach used several living rooms and the property's distance from the city center. Varma et al. [3] took a different approach to the topic. Using neural networks and machine learning, he projected the price of a house. The structure is offered as a novel method for identifying housing submarkets. For predicting individual property prices.

According to Madsen [6] nominal mortgage payments and nominal income influence property values in the short run, while acquisition expenses influence house prices in the long run. Holly et al. [7] Looked examined changes in actual property prices in the United States. It looks at how fundamentals like real per capita disposable income and common shocks affect state-level real house prices, as well as how quickly real house prices adapt to macroeconomic and local disturbances. Bork's [8] method tested home price forecast ability across all 50 states utilizing Dynamic Model Averaging and Dynamic Model Selection, which allow for model and parameter changes. Forecasting accuracy improves dramatically when the entire forecasting model is permitted to evolve over time and between places. According to Adair et al. [9], the notion of housing submarkets is inherent in the comparative valuation technique, but there has been little attempt to integrate property market study with the valuation process. Based on price information from Belfast, hedonic estimates predict that submarkets may be spread across a larger geographical area than previously anticipated. The value's ability to evaluate the quality of elements has ramifications for the valuation process. Bin et al. [12] estimate a hedonic pricing function using semi-parametric regression and com- pare the price prediction performance to that of traditional parametric models. This study makes use of a large data set from Pitt County, North Carolina, consisting of 2595 single- family housing unit transactions between July 2000 and June 2002. Information Systems (GIS) is used [28], [29], [30], [31], [32], [33], [34], [35], [36]. The results reveal that semi-Parametric regression outperforms parametric equivalents in both in-sample and out-of-sample price predictions, implying that the semi-parametric model might be effective for home sales price measurement and prediction [26], [27].

III. PROPOSED MODEL

The manual technique that is presently getting used within the market is out-of-date and has high risk. To overcome this, there is a need for an updated and automated system. Data mining algorithms can help investors by giving them a fair idea of an estate. Our projected model can depend on mining also as a comparison of knowledge. The model can take the identified worth of different homes within the neighborhood as an input parameter. Based on this parameter, even if there is an error in the user's data, the model will be able to predict an acceptable price. Our model will be based on a logistic regression algorithm to provide an accuracy of 75% or more.

A. Algorithm Used

Linear Regression - - the linear Regression is the best algorithm to be used in predicting the house model price. It uses the image of the house and predicts the price according to the image.

Formula Used X = y + z2k

where, x = estimated dependent variable score, y= constant,
z = regression coefficient, and
k = score on the independent variable.

3.2 Technology and Tools used

NumPy and Pandas: In Python, NumPy and Pandas [11] are two extensively used libraries for data cleaning. The required libraries were first imported. We read the CSV file into a Pandas data frame after importing the libraries. After that, the standard missing values are filled, and then the non-standard missing values are filled.

Matplotlib: Matplotlib is a Python package that validates figures. Matplotlib is a program that duplicates graphs and senses. Matlab plotting is a breeze. In general, plotting's follow a similar pattern in each plot. There is a module called pyplotsc in Matplotlib that assists in plotting figures. For executing the charts, we used the Jupyter journal. To call the package module, we 'import matplotlib. pyplot as plt'.

Scikit-learn: Scikit-Learn is a Python s library that performs machine learning, preprocessing, cross-approval, and percep- tual q calculations through a consistent interface[29-33]. It is a simple and effective gadget for information mining and

evaluation. It focuses on various order, relapse, and grouping calculations, such as support vector machines, irregular woodlands, angle boosting and implies. The following activities were taken while using Scikit-learn:

i. Load a dataset
ii. Split the dataset
iii. Train the model

A. Data:

IV. 4. IMPLÉMENTATION

The element of machine learning tasks for which specific consideration ought to be obviously taken is the information. To be sure, the outcomes will be exceptionally impacted by the information dependent on where we discovered them, how they are formatted if they are steady if there is any anomaly, etc. At this progression, numerous inquiries ought to be replied to ensure learning calculations will be productive and precise.

B. Getting the Data: The first issue that came into the picture is to get the data that would create a large enough database since I want to predict the price of the apartment as the real estate agent, so to achieve this I downloaded the dataset from Kaggle To add more entries in the list we have decided to use the web scrapping method so that we can get the more data online from the websites. The main idea behind this is to implement AI in such a way that it gives the result same

C. Cleaning The Data:

Now for the next step, we have performed the cleansing of the dataset. Data cleansing is a way to remove the corrupt data there are some ambient data or false data that need to be removed. so for this, we need to remove the variable having more than 50 % of the missing data hence some variables are removed from the because the data is missing. Below are a few features that have been removed from the dataset.

Apply a range of living rooms i e 1 to 4 The set minimum value for price and area

D. Data Featuring:

i. Add a new feature(integer) for bhk (Bedrooms Hall Kitchen)
ii. Explore the total_sqft feature
iii. Add a new feature called price per square feet
iv. Dimensionality Reduction

E. Outlier Removal:

Outlier refers to the data that does not belong in the specified range. Removing values that have high prices for a specific area is part of this. We'll take the mean price per sqft in the area and eliminate any fields with values much higher than the mean value for that location. Outlier removal was to remove the extreme and non-normal house prices so that our prediction model be more accurate

Fig [6] represents data, after all, large values are removed to make available dataset values lie between 5000 to 15000 per sqft. In scattergraph format.

Fig [7] represents data and frequency after all large values are removed to make available dataset values lie between 5000 to 15000 per sqft. In bar graph format

4.6 Model Selection: GridSearchCV was used to find the machine learning model that best suited our database. We used the GridSearchCV to test lasso regression, linear regression, and decision tree methods to get the optimal model.

We may conclude from Table [1] that Linear Regression provides the best score based on the given findings. As a result, we pursued it further.

F. 4.7 Training Model:

After choosing the best algorithm based on the accuracy score for our data we will create our model with the chosen algorithm and test it against the dataset for testing. Fig [9] gives us a view of data testing and model training.

V. DATA TABLE VIEW AFTER DATA CLEANSING AND AFTER DATA FEATURING

A. Figures

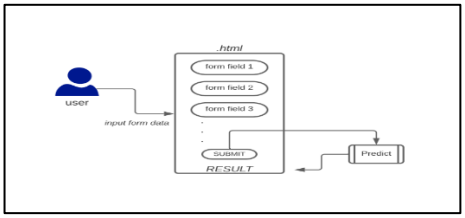

Figure 1. Architectural Flow of data

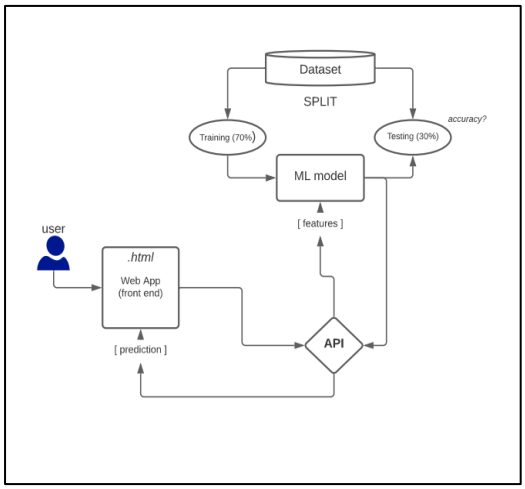

Figure 2. Use Case Diagram of the system

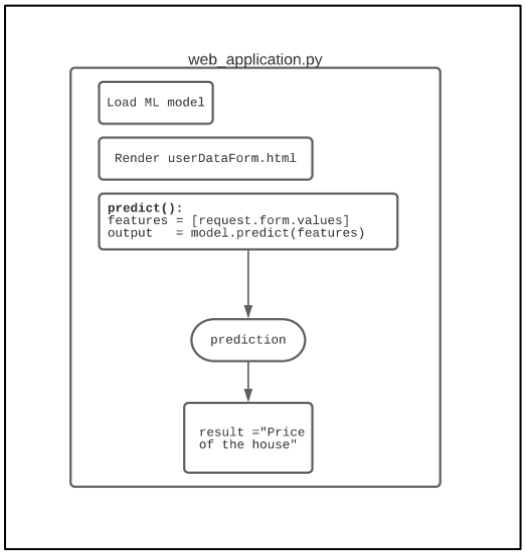

Figure 3. Website flow chart

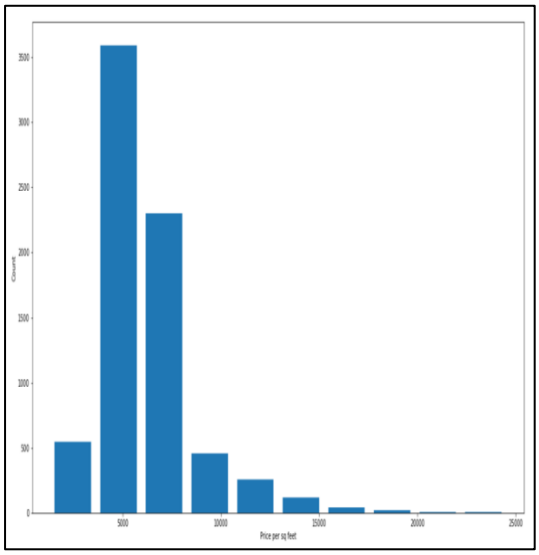

Figure 4. Bar Graph of Data

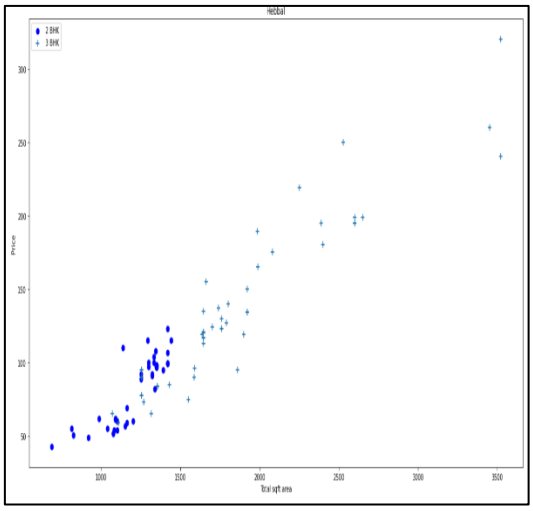

Figure 5. Scatter-plot of Data

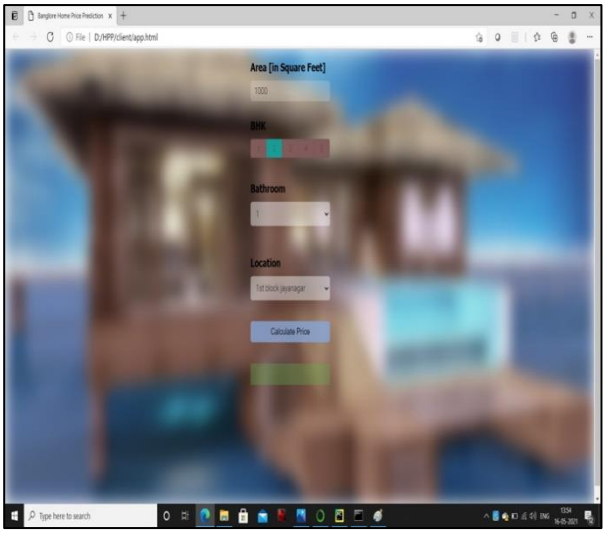

Figure 6. Website form

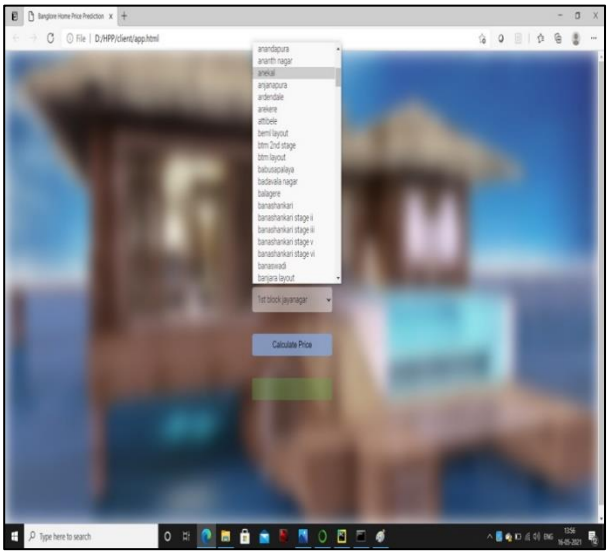

Figure 7. Area selection in website

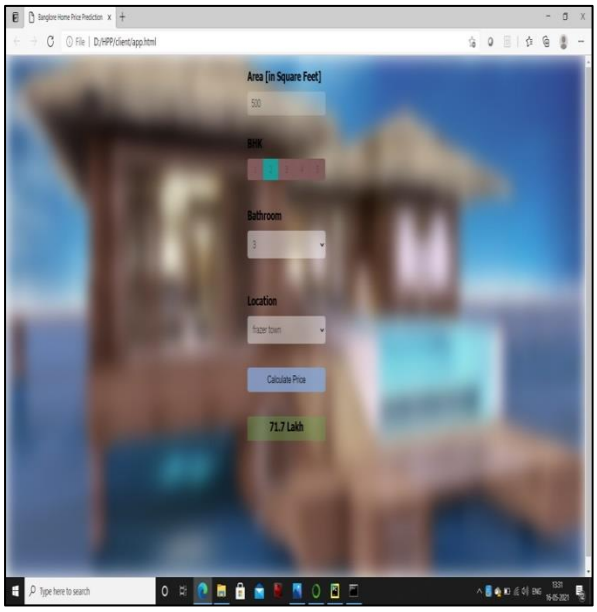

Figure 8. Predicted price

```
In [58]: predict_price('1st Phase JP Nagar', 1000,2,2)
Out[58]: 83.49904677179224

In [59]: predict_price('1st Phase JP Nagar', 1000,3,3)
Out[59]: 86.80519395205835

In [60]: predict_price('Indira Nagar', 1000,2,2)
Out[60]: 181.27815484006857

In [61]: predict_price('Indira Nagar', 1000,3,3)
Out[61]: 184.5843020203347
```

Figure 9. Data Training

VI. RESULT

Using Linear Regression and Decision Tree Regressor, we created an 85 per cent accurate house price forecast model with a functioning User Interface. The price of the property is determined by parameters such as total square footage, number of bedrooms, number of bathrooms, and location. We've designed a basic User Interface for our model that can be used by both buyers and builders who want to build anything in a specific area. Finally, as illustrated in Fig [6], Fig [7], and Fig [8], the model is implemented in a web- based application to give a user-friendly and easy interface. Figure 9 shows a Jupyter notebook with a functional model. The project's future goals include improving the other f model and incorporating more data from various sources.

7. Conclusion

The project's aim is to achieve a human intelligence level house price prediction that can accurately predict the price of the house. It depicts the house price according to the area and uses image processing and machine learning. The experimental result showed that we are able to achieve the linear regression model for house price prediction with an 80% accuracy also we use advanced machine learning libraries of python such as sklearn. We have created a website using HTML and CSS by which one will be able to find the house price add a filter and much more. In the future, we will try to reduce the error margins and update the HTML codes to create a more user-friendly interface

8. ACKNOWLEDGEMENT

I want to thank all the people whose assistance was a milestone in the completion of this paper. I wish to express our deepest gratitude to Dr Mohammad Shuaib Khan for his guidance as well as to Er. Sushil Kumar Mishra and Dr Rakesh Kumar for the overall guidance and supervision.

REFERENCES

[1] Sood, M., Verma, S., Panchal, V.K.: Optimal path planning using hybrid bat algorithm and cuckoo search. Int. J. Eng. Technol. 7(4.12), 30–33 (2018).

[2] Kumar, P.; Verma, S. Detection of wormhole attack in VANET. Natl. J. Syst. Inf. Technol. 2017, 10, 71.

[3] Batra, Isha, Sahil Verma, Arun Malik, Kavita, Uttam Ghosh, Joel J. P. C. Rodrigues, Gia Nhu Nguyen, A. S. M. Sanwar Hosen, and Vinayagam Mariappan. 2020. "Hybrid Logical Security Framework for Privacy Preservation in the Green Internet of Things" Sustainability 12, no. 14: 5542.

[4] Gaba S, Verma S. Analysis on Fog Computing Enabled Vehicular Ad hoc Networks. Journal of Computational and Theoretical Nanoscience, 2019, 16(10), pp. 4356–4361.

[5] N. Kaur and S. Verma, "Detection of plant leaf diseases by applying image processing

schemes," Journal of computational and theoretical nanoscience (JCTN), vol. 16, no. 9, pp. 3728–3734, 2019.

[6] A. S. Adair and J. N. S. B. & W, "Hedonic modelling, housing submarkets and residential valuation," Journal of Property Research, vol. 13, no. 1, pp. 67–83, 1996.

[7] C. Fan, Z. Cui, and X. Zhong, "House Prices Prediction with Machine Learning Algorithms," Proceedings of the 2018 10th International Conference on Machine Learning and Computing ICMLC 2018.

[8] S. Mraschka and V. Mirjalili, Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow. Birmingham: Packt Publishing, 2017.

[9] O. Bin, "A prediction comparison of housing sales prices by parametric versus semi-parametric regressions," Journal of Housing Economics, vol. 13, pp. 68–84, 2004.

[10] N. Z. Jhanjhi, S. N. Brohi, N. A. Malik, and M. Humayun, "Proposing a hybrid rpl protocol for rank and wormhole attack mitigation using machine learning," 2020 2nd International Conference on Computer and Information Sciences (ICCIS), pp. 1–6, 2020.

[11] K. Hussain, S. J. Hussain, N. Jhanjhi, and M. Humayun, "SYN Flood Attack Detection based on Bayes Estimator (SFADBE) For MANET," in 2019 International Conference on Computer and Information Sciences (ICCIS), 2019, pp. 1–4.

[12] I. A. Shah, Q. Sial, N. Z. Jhanjhi, and L. Gaur, "Use Cases for Digital Twin," Digital Twins and Healthcare: Trends, Techniques, and Challenges, pp. 102–118, 2023.

[13] Y. Feng and K. Jones, "Comparing multilevel modelling and artificial neural networks in house price prediction," 2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM), pp. 108–114, 2015.

[14] N. Kalra, "House Price Prediction using Machine Learning in Python," International Journal of Advanced Engineering Research and Applications, 2021.

[15] A. M. Daniel, S. Srivastava, V. Dr, and Anbarasu, "' HEART DISEASE PREDICTION SYSTEM USING MACHINE LEARNING," Interna- tional Journal of Creative Research Thoughts, vol. 9, no. 5, pp. 2320– 2882, 2021.

[16] J. Schmidhuber, "Multi-column deep neural networks for image classifi- cation," Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR, pp. 3642–3649.

[17] . T. Kauko, P. Hooimeijer, and J. Hakfoort, "Capturing housing market segmentation: An alternative approach based on neural network model- ing," Housing Studies, vol. 17, pp. 875–894, 2002.

[18] R. Khan, J. Teo, M. A. Jan, S. Verma, R. Alturki and A. Ghani, "A Trustworthy, Reliable, and Lightweight Privacy and Data Integrity Approach for the Internet of Things," in IEEE Transactions on Industrial Informatics, vol. 19, no. 1, pp. 511-518, Jan. 2023, doi: 10.1109/TII.2022.3179728.

[19] Ramisetty, S.; Anand, D.; Verma, S.; Alaboudi, A.A. SC-MCHMP: Score-Based Cluster Level Hybrid Multi-Channel MAC Protocol for Wireless Sensor Network. In Information Security Handbook; CRC Press: Boca Raton, FL, USA, 2022; pp. 1–18.

[20] Kaur, N.; Devendran; Verma, S.; Kavita; Jhanjhi, N. De-Noising Diseased Plant Leaf Image. In Proceedings of the 2022 2nd International Conference on Computing and Information Technology (ICCIT), Tabuk, Saudi Arabia, 25–27 January 2022; pp. 130–137.

[21] Rani G, Oza MG, Dhaka VS, Pradhan N, Verma S, Rodrigues JJ (2020) Applying deep learning for genome detection of coronavirus. Res Sq. https://doi.org/10.21203/rs.3.rs-93564/v1R. Singla, N. Kaur, D. Koundal, and A. Bharadwaj, pp. 1–40, 2021.

[22] Adeyemo, V. E., Abdullah, A., JhanJhi, N. Z., Supramaniam, M., & Balogun, A. O. (2019). Ensemble and deep-learning methods for two-class and multi-attack anomaly intrusion detection: an

empirical study. International Journal of Advanced Computer Science and Applications, 10(9).,

[23] R. Agarwal, and A. G. Thomas, "Performance comparison of deep cnn models for detecting driver's distraction," Materials & Continua, vol. 68, no. 3, pp. 4109–4124, 2021.

[24] A. Almusaylim, Z. Jhanjhi, N. Z. Alhumam, and A, "Detection and mitigation of RPL rank and version number attacks in the internet of things: SRPL-RP," Sensors, vol. 20, no. 21, pp. 5997–5997, 2020.

[25] I. A. Shah, Q. Sial, N. Z. Jhanjhi, and L. Gaur, "The Role of the IoT and Digital Twin in the Healthcare Digitalization Process: IoT and Digital Twin in the Healthcare Digitalization Process," Digital Twins and Healthcare: Trends, Techniques, and Challenges, pp. 20–34, 2023.

[26] A. Srivastava and S. Verma, "Analysis of Quality of Service in VANET," Materials Science and Engineering Conference Series, vol. 993, no. 1, pp. 12 061–12 061, 2020.

[27] Rupesh Chaudhari , Ritik Gad , Pranav Gawali , Mangesh Gite , Dr. A. B. Pawa, Hate Speech Detection on Social Media Using Machine Learning Algorithms, Journal of Cognitive Human-Computer Interaction, Vol. 2 , No. 2 , (2022) : 56-59 (Doi : https://doi.org/10.54216/JCHCI.020203)

[28] M. Sumithra , Kiruthika.S , Nithya S , Poornima B , DharanyaS, Enhancement Of Cloud User Data Access Security Entrusted to AI Face Recognition Techniques, Journal of Cognitive Human-Computer Interaction, Vol. 2 , No. 2 , (2022) : 60-64 (Doi : https://doi.org/10.54216/JCHCI.020204)

[29] Ramgude AkshayDili , K. Vengatesa , Kunal Joshi , Chaitanya Tekane, Counterfeit Product Detection System Using One-Time QR code, Journal of Cognitive Human-Computer Interaction, Vol. 2 , No. 2 , (2022) : 65-71 (Doi : https://doi.org/10.54216/JCHCI.020205)

[30] K.Vengatesan , Raghvendra Vijay Naidu , Kunal Ganesh Joshi , ChaitanyaSantosh Tekane , Siddhant Ravindra Gore, Machine Learning Based Product Price Inference Using Price Elasticity of Demand Approach, Journal of Cognitive Human-Computer Interaction, Vol. 3 , No. 1 , (2022) : 08-16 (Doi : https://doi.org/10.54216/JCHCI.030101)

[31] Hariharan E.K.S , Bharath M , MageshwaranS, Effectiveness and Impact of Online Education on School Students - A Study With Reference to Chennai City, Journal of Cognitive Human-Computer Interaction, Vol. 3 , No. 1 , (2022) : 17-23 (Doi : https://doi.org/10.54216/JCHCI.030102)

[32] Kavita, pp. 2278–0181, 2012.

[33] P. Kumar, "Detection of Wormhole Attack in VANET," National Journal of System and Information Technology, vol. 10, pp. 71–71, 2017.

[34] V. Dogra, "Understanding of Data Preprocessing for Dimensionality Reduction Using Feature Selection Techniques in Text Classification," Lecture Notes in Networks and Systems, vol. 248, pp. 2021–2021.

[35] P. Rani, "Mitigation of black hole attacks using firefly and artificial neural network," Neural Comput & Applic, 2022.

[36] S. Ghosh and A. Singh, "Svm and knn based cnn architectures for plant classification," Computers, Materials & Continua, vol. 71, no. 3, pp. 4257–4274, 2022.

TABLE I

1. GIVES US A VIEW OF DATA AFTER CLEANING.

Locationsize	Total_sqft	price Electronic
City Phase2	2 BHK 1056	39.07K Chikka
Tirupathi	4 BHK 2600	120k
Uttarahalli	3BHK 1440	62k

TABLE II

2. GIVES US A VIEW OF DATA AFTER CLEANING AND FEATURING

Locationsize	Total_sqft	price Hebbal
4 BHK	3067-8156	477.0 k
Sarjapur 2 BHK	1145-1340	43.4 k
Kangeri	1BHK 34.6sq. meter	18.5k