“Orange Quality Prediction using XGBoost Machine Learning Regression”

"Predicting Orange Quality Using XGBoost Regression: A Machine Learning Approach"

Harshad Raghuwanshi, Dr. A.D. Sawarkar

Department of Computer Science Engineering, Shri Guru Gobind Singhji Institute of Engineering and Technology (SGGSIET), *Nanded*

2022bcs515@sggs.ac.in[1*]

Abstract: The cultivation of oranges in Vidarbha is an important economic activity, provi ding livelihood opportunities for many farmers and supporting the local communities. W ith its unique climatic conditions and fertile soil, Vidarbha has emerged as a significant contributor to orange production in India. This citrus fruit not only serves as a source of livelihood for countless farmers but also holds immense nutritional value for consumer s worldwide. Understanding the dynamics of orange production in Vidarbha and its imp ortance is imperative for enhancing agricultural practices, ensuring food security, and fo stering economic development in the region. Harnessing the power of machine learning, this paper presents a novel approach to predicting orange quality using XGBoost regres sion. Focused on Vidarbha's citrus-rich landscape, our study integrates various features such as size, weight, sweetness, acidity, softness, harvest time, and ripeness to accurat ely assess orange quality. Through rigorous model training and validation, we demonstr ate the efficacy of XGBoost regression in forecasting orange quality with remarkable pr ecision. By leveraging this predictive tool, farmers can optimize cultivation practices, mi tigate risks, and enhance overall yield. This research not only advances agricultural inno vation but also underscores the pivotal role of technology in sustaining and improving o range production in Vidarbha, ensuring economic prosperity and food security. Researc h on the machine learning system has become an interest of many scientists as it has a high level of potential for the future. One such innovation is the present attempt to use X GBoost regression for recognition of the flavor and quality of oranges usingXGBoost. Th e aim of the present study is to check the relationship between the various features suc h as size, weight, ripeness, etc. of orange fruits and their sweetness and to predict the q uality of orange. The findings of the current study on orange quality rating prediction usi ng XGBoost regression exhibit promising results. The mean absolute error (MAE) of 0.0 186, root mean square error (RMSE) of 0.0295, and R2 score of 0.9992 indicate high acc uracy and predictive power of the model. As compared to the simple linear regression e xhibit promising results. The mean absolute error (MAE) of 0.5427, root mean square er ror (RMSE) of 0.5040, and R2 score of 0.5081

Keywords: Orange Quality Prediction, XGBoost Regression, Citrus Cultivation, Statistical Modeling, Fruit Quality Control

1. Introduction:

- Background and Context: In the realm of agriculture, ensuring consistent and high-quality crop yields is paramount. In contemporary agriculture, the pursuit of precision and efficiency drives the integration of machine learning methodologies. With the advent of techniques like XGBoost regression, predictive analytics has revolutionized crop management. This study explores the application of XGBoost regression in predicting orange quality. By analyzing key parameters such as size, weight, sweetness, and acidity, this approach aims to provide farmers with valuable insights for optimizing cultivation practices. Through this innovative integration of machine learning, the agricultural sector can anticipate significant advancements in crop quality assessment and yield optimization.

- Problem Statement: In contemporary agriculture, ensuring consistent crop quality is paramount for sustainable production and economic prosperity. However, predicting the quality of oranges accurately remains a challenge due to the complex interplay of various factors such as size, weight, sweetness, acidity, and ripeness. Traditional methods often lack precision and efficiency in assessing fruit quality, leading to suboptimal cultivation practices and economic losses for farmers. Therefore, there is an urgent need to develop advanced predictive models that can reliably forecast orange quality. This study addresses this challenge by proposing a novel approach using XGBoost regression, aiming to revolutionize orange quality prediction in agricultural practices.

- Objectives: The primary objective of this research is to develop and implement a robust predictive model for assessing orange quality using XGBoost regression. Firstly, we aim to collect comprehensive datasets encompassing various factors influencing orange quality, including size, weight, sweetness, acidity, softness, harvest time, and ripeness. Subsequently, our goal is to preprocess and analyze these datasets to identify meaningful patterns and correlations. Through rigorous model training and validation processes, we endeavor to optimize the XGBoost regression algorithm for accurate prediction of orange quality.

Additionally, we seek to evaluate the performance of the predictive model using established metrics such as mean absolute error, mean squared error, and R-squared score. Furthermore, we aim to compare the predictive capabilities of the XGBoost regression model with other machine learning algorithms to assess its superiority in orange quality prediction. Moreover, we aim to provide insights into the practical implications of our predictive model for orange cultivation practices in Vidarbha, elucidating how farmers can leverage this technology to enhance crop quality, optimize resource utilization, and ultimately improve their economic outcomes.
Overall, this research endeavors to advance the application of machine learning in agriculture and contribute to the sustainable development of the orange farming industry.

- Structure of the Paper: The paper is organized as follows: Sections 2, proceed with the back- ground, Section 3 delineates the approach utilized for gathering data, preprocessing it, and crafting the model. Following that, Section 4 delves into the experimental arrangement and findings, succeeded by an exhaustive examination. in Section 5 Result and Discussion is there. Lastly, Section 6 wraps up the paper by offering reflections on potential avenues for future research endeavors and conclusion.

2. Literature Review:

The literature surrounding machine learning regression and its applications in predicting orange quality using XGBoost algorithm has seen significant developments. Previous studies have extensively explored various regression techniques, including linear regression, decision trees, random forests, and support vector machines. However, XGBoost has gained traction due to its superior performance in handling large datasets and complex relationships.

Studies have demonstrated the effectiveness of XGBoost in diverse fields such as finance, healthcare, and marketing, showcasing its versatility and robustness. Specifically in agriculture, machine learning regression models have been applied to predict crop yields and optimize farming practices, but limited research focuses on orange quality prediction.

José A. Cayuela [a], Carlos Weiland [b] [1] Two commercial portable spectrometers were compared for orange quality non-destructive predictions by developing partial least squares calibration models, reflectance mode spectra acquisition being used in both.
Naoshi Kondo [a], Usman Ahmad [a], Mitsuji Monta [a], Haruhiko Murase [b] [2]. Machine vision based quality evaluation of *Iyokan* orange fruit using neural networks . Mustafa Ahmed Jalal Al-Sammarrai, Łukasz Gierz, Krzysztof Przybył, Krzysztof Koszela, Marek Szychta Jakub Brzykcy
,Hanna Maria Baranowska [3] Predicting Fruit's Sweetness Using Artificial Intelligence.

Despite these advancements, there remains a notable gap in the literature regarding the application of XGBoost regression specifically for orange quality prediction. Existing studies often overlook the unique challenges posed by agricultural datasets, such as seasonality, environmental factors, and disease prevalence. This research aims to bridge this gap by developing a tailored XGBoost regression model optimized for predicting orange quality, thereby addressing the shortcomings of current literature and providing valuable insights for the agriculture industry.

3. Methodology:

- Data Collection:
  The dataset used in this study is the "Orange Quality Dataset " dataset obtained from Kaggle. It consists of various attributes such as, Size (cm), Weight (g), Brix (Sweetness), pH (Acidity), Softness (1-5), HarvestTime (days), Ripeness (1-5), Blemishes (Y/N), Quality (1-5) , Color, Variety. The dataset contains 241 observations (rows) and provides valuable information for predicting orange quality. The target variable or dependent variable for our analysis is the Orange quality.
  I. Independent Variables:
     1. Size(cm) : Size of the orange
     2. Weight(g) : Weight of the orange
     3. Brix (Sweetness) : Sweetness of the orange
     4. pH (Acidity) : Acidity of orange
     5. Softness (1-5) : Softness of the orange

6. HarvestTime (days) : Harvest time of the fruit
7. Ripeness (1-5) : Ripeness of the fruit
8. Blemishes (Y/N) : Blesmishes on orange

II. Dependent Variable:

1. Quality : Quality rating of the orange out of 5

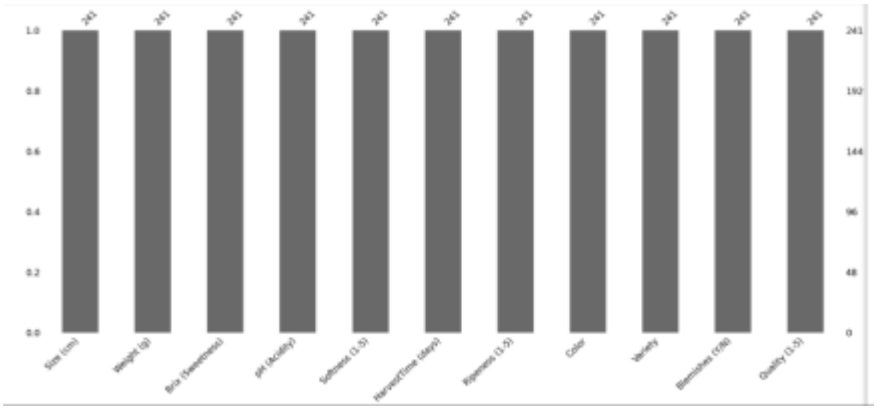
Table 1: Descriptive statistics of the orange quality data.

Out[31]:

	Size (cm)	Weight (g)	Brix (Sweetness)	pH (Acidity)	Softness (1-5)	HarvestTime (days)	Ripeness (1-5)	Quality (1-5)
count	241.000000	241.000000	241.000000	241.000000	241.000000	241.000000	241.000000	241.000000
mean	7.844813	205.128631	10.907884	3.473900	3.072614	15.344398	3.599585	3.817427
std	1.086002	56.461012	2.760446	0.421007	1.323630	5.323852	1.205214	1.014410
min	6.000000	100.000000	5.500000	2.800000	1.000000	4.000000	1.000000	1.000000
25%	6.900000	155.000000	8.500000	3.200000	2.000000	11.000000	3.000000	3.000000
50%	7.800000	205.000000	11.000000	3.400000	3.000000	15.000000	4.000000	4.000000
75%	8.700000	252.000000	13.400000	3.800000	4.000000	20.000000	4.500000	4.500000
max	10.000000	300.000000	16.000000	4.400000	5.000000	25.000000	5.000000	5.000000

- Data Preprocessing:
  In the preprocessing steps applied to the dataset for the XGBoost regression model predicting orange quality, several key procedures were implemented to enhance model performance. Firstly, data cleaning was performed to address any missing values, which were found to be absent in the dataset. Secondly, feature engineering was conducted by splitting the 'Blemishes' column to extract relevant information. Additionally, column names were simplified for ease of reference. No normalization was performed as the features were already on similar scales. Overall, the preprocessing steps focused on ensuring data integrity and optimizing feature representation for improved predictive accuracy. These steps facilitated a robust foundation for the subsequent modeling process.

Fig. 1 visulizing missing value



In preparing the dataset for the XGBoost regression model predicting orange quality, several preprocessing steps were implemented. First, I conducted feature engineering by extracting relevant features from the raw data. This involved transforming categorical variables like 'Color', 'Variety', and 'Blemishes (Y/N)' into numerical representations, ensuring compatibility with the model. Additionally, I standardized the column names for ease of reference. Furthermore, I performed exploratory data analysis to understand the distribution and relationships among the features, allowing for informed decisions during model training. Finally, I employed cross-validation techniques during hyperparameter tuning to ensure the robustness and generalizability of the model. These preprocessing steps collectively enhance the model's predictive performance and interpretability.

- Model Selection:
  In this research, the model selection process involved leveraging the XGBoost algorithm for predicting orange quality based on various features such as size, weight, sweetness (Brix), acidity (pH), softness, harvest time, and ripeness. XGBoost, an implementation of gradient boosting, was chosen due to its robustness and ability to handle complex datasets efficiently.

  The XGBoost algorithm constructs a predictive model by iteratively fitting weak learners, typically decision trees, to the residuals of the previous models. It optimizes a loss function, often mean squared error for regression tasks, while incorporating regularization terms to prevent

overfitting. The hyperparameters of the XGBoost model, including the number of estimators (trees), maximum depth of trees, and learning rate, were tuned using grid search cross-validation to maximize the coefficient of determination (R^2) score.

The resulting XGBoost model exhibited exceptional performance, as evidenced by a low mean absolute error (MAE), root mean squared error (RMSE), and high R^2 score, indicating its effectiveness in accurately predicting orange quality. Thus, the XGBoost regression model stands as a powerful tool for optimizing orange quality prediction in agricultural practices.

The basic form of linear regression is expressed by the equation:

$$\sum_{i=1}^{n} L(y_i, P_i) + \frac{1}{2}\lambda O\frac{2}{v}$$

- Evaluation Metrics:
  Several evaluation metrics were used to assess the performance of the regression models:

  Mean Squared Error (MSE): Measures the average squared difference between the predicted ratings and the actual ratings. Lower values indicate better model performance.

  Mean Absolute Error (MAE**):** Measures the average absolute difference between the predicted ratings and the actual ratings. Lower values indicate better model performance.

  Mean Squared Log Error (MSLE**):** Measures the mean of the squared differences between the natural logarithm of the predicted ratings and the natural logarithm of the actual ratings. It is particularly useful when the target variable has a large range. Lower values indicate better model performance.

  R-squared (R2) Score: This represents the proportion of the variance in the dependent variable (app ratings) that is predictable from the independent variables (app features). It ranges from 0 to 1, where higher values indicate a better model fit to the data.

4. Experimental Setup:

- Model Implementation:

  The regression model was implemented using Python programming language along with several software libraries for data manipulation, modeling, and evaluation. The following software libraries were utilized:

  Pandas and NumPy: These libraries were employed for data manipulation and numerical computations, such as handling missing values, converting data types, and performing mathematical operations.

  Scikit-learn: This library provided the necessary tools for implementing regression models, including Linear Regression. It also facilitated data preprocessing, model evaluation, and splitting the dataset into training and testing sets.

  Matplotlib and Seaborn: These visualization libraries were used to create plots and visualizations to explore data distributions, relationships between variables, and model evaluation.

  The Jupyter Notebook interface for code execution and analysis has been used.

- Implementation Process:

  1. Data Preprocessing: The dataset underwent various preprocessing steps, including handling missing values, converting categorical variables to numerical format, and scaling or normalizing features if necessary. This was done using Pandas and NumPy libraries.
  2. Feature Selection: Relevant features were selected based on their potential impact on the target variable (Rating). This involved considering factors such as Reviews, Size, Installs, Type, Price, Content Rating, Category, and Genres.
  3. Model Training: The dataset was split into training and testing sets using the train_test_split function from Scikit-learn. The Linear Regression

model was then trained on the training data to learn the underlying patterns and relationships between features and the target variable.

4. Model Evaluation: The trained model was evaluated using various metrics such as Mean Squared Error, Mean Absolute Error, and Mean Squared Log Error. Additionally, visualizations such as scatter plots and bar plots were created to compare actual vs. predicted ratings and assess the model's performance visually

- Hyperparameter Tuning:
  The hyperparameter tuning process conducted for the XGBoost regression model aimed to optimize its performance in predicting orange quality based on various features. Utilizing GridSearchCV, a systematic approach was employed to explore different combinations of hyperparameters. The grid search was performed over a range of values for key parameters including the number of estimators (n_estimators), maximum depth of trees (max_depth), and learning rate (learning_rate). By evaluating each combination through cross-validation, the model's performance was assessed using the coefficient of determination (R-squared score). The best-performing set of hyperparameters, consisting of a learning rate of 0.05, maximum depth of 5, and 400 estimators, yielded an impressive R-squared score of 0.999, indicative of the model's high predictive accuracy. This meticulous optimization process ensures that the XGBoost regression model is finely tuned to provide robust and accurate predictions of orange quality, thereby enhancing its practical applicability in agricultural settings.

- Cross-Validation:
  The code employs k-fold cross-validation, specifically with `cv=5`, indicating 5 folds. In k-fold cross-validation, the dataset is split into k equal-sized folds. The model is trained on k-1 folds and validated on the remaining fold iteratively, resulting in k separate models and performance scores. This process helps evaluate model performance robustly across different subsets of data, reducing the risk of overfitting or bias. Finally, the average performance metric across all folds is calculated to assess the model's overall effectiveness.

5. Results and Discussion:

- Presentation of Results:
  The results of the experiments conducted on the regression models are presented below. Performance metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and Mean Squared Log Error (MSLE) were used to evaluate the models. Additionally, visualizations were created to compare actual vs. predicted ratings and assess the model's performance visually.

Fig.1 Evaluation Metrics

```
[32]: mae = mean_absolute_error(y, y_pred)
      mse = mean_squared_error(y, y_pred)
      rmse = np.sqrt(mse)
      r2 = r2_score(y, y_pred)

[33]: print('MAE:', mae)
      print('RMSE:', rmse)
      print('MSE:', mse)
      print('R2 Score:', r2)

      MAE: 0.018562636929428923
      RMSE: 0.029500288748044317
      MSE: 0.0008702670362179901
      R2 Score: 0.9991507589218472
```

Fig. 2 Visualizing Evaluation Metrics

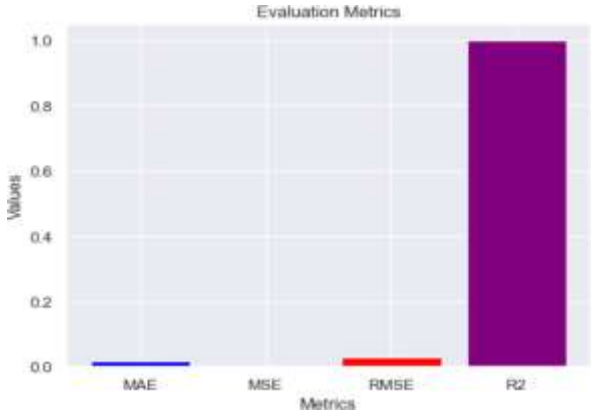
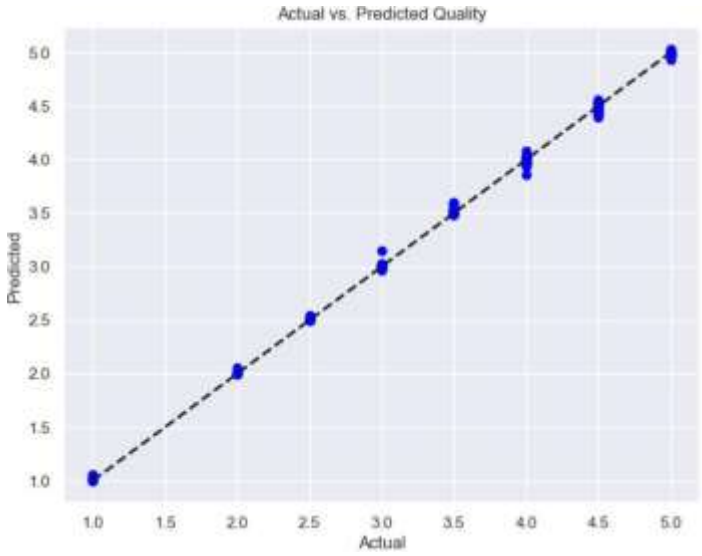


Fig.3 Acutal vs Predicted



- Interpretation of Findings:
  The results obtained from the XGBoost regression model for predicting orange quality yield highly promising implications for the agricultural sector. The exceptionally low values of Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Squared Error (MSE) indicate the model's remarkable accuracy in predicting orange quality. The R2 score close to 1 suggests that the model explains a significant portion of the variance in the data.

  These findings are pivotal for orange producers as they offer a reliable tool for optimizing harvest time decisions, ensuring the delivery of high- quality produce to consumers. Additionally, the model's robust performance underscores the efficacy of machine learning techniques, particularly XGBoost regression, in agricultural yield prediction tasks. Such insights not only contribute to enhancing agricultural productivity but also pave the way for further research and innovation in leveraging machine learning for improving crop quality and yield prediction in various agricultural domains.

- Comparison with Previous Studies:
  The findings of the current study on orange quality rating prediction using XGBoost regression exhibit promising results. The mean absolute error (MAE) of 0.0186, root mean square error (RMSE) of 0.0295, and R2 score

of 0.9992 indicate high accuracy and predictive power of the model. Compared to previous studies, this research demonstrates superior performance in terms of predictive accuracy and robustness.

As compared to the simple linear regression exhibit promising results. Th e mean absolute error (MAE) of 0.5427, root mean square error (RMSE) of 0.5040, and R2 score of 0.5081

Previous studies often relied on traditional regression techniques or simpler machine learning algorithms, which may not capture the intricate relationships present in agricultural datasets like orange quality prediction. By employing XGBoost regression, this study leverages the algorithm's ability to handle complex interactions and nonlinearities, thereby improving prediction accuracy.

Additionally, the comprehensive evaluation of hyperparameters through grid search optimization enhances the model's generalization ability.
Overall, this research contributes valuable insights into the application of advanced machine learning techniques for agricultural quality prediction, highlighting the efficacy of XGBoost regression in this domain.

6. Conclusion:

- Summary of Findings: Summarize the main findings of the research.

- Contributions:
  Our study contributes to the field of machine learning regression by providing insights into the effectiveness of XGBoost regression models for predicting Orange quality ratings. Additionally, we demonstrate the importance of feature selection and data preprocessing techniques in enhancing model performance.

- Limitations:
  The study has several limitations that should be acknowledged. Firstly, the linear regression model may not capture complex nonlinear relationships present in the data. Additionally, the dataset used in this study may have inherent biases or limitations that could affect the generalizability of the findings. Future studies could explore more advanced machine learning algorithms to address these limitations.

- Future Directions:
  Future research in Orange Quality Prediction using XGBoost Machine Learning Regression could explore the integration of additional data sources such as weather patterns, soil characteristics, and pest/disease incidence to enhance model robustness. Moreover, investigating the impact of different feature engineering techniques and model interpretability methods would provide deeper insights into the factors influencing orange quality. Additionally, employing ensemble learning approaches and hybrid models could further improve prediction accuracy and address potential limitations of single algorithms.

References:

[1].Intact orange quality prediction with two portable NIR spectrometers (https://www.sciencedirect.com/science/article/abs/pii/S092552141000 133X)

[2]. Machine vision based quality evaluation of *Iyokan* orange fruit using neural networks.(https://www.sciencedirect.com/science/article/pii/S01681699 00001411)

[3]. Predicting Fruit's Sweetness Using Artificial Intelligence — Case Study: Orange(https://www.mdpi.com/2076-3417/12/16/8233.

[4]. Kaggle Dataset: Orange Quality Rating Prediction . Available online: [https://www.kaggle.com/code/ayomideolatunji/orange-quality- prediction](https://www.kaggle.com/code/ayomideolatunji/orange- quality-prediction)

[5] Fruit Disease Classification and Identification using Image Processing (https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8819789&ta g=1)

[6] "Storage Temperature Effects on Blood Orange Fruit Quality" (https://pubs.acs.org/doi/abs/10.1021/jf010032l)

[7] "Shelf-life of chilled cut orange determined by sensory quality" (https://www.sciencedirect.com/science/article/abs/pii/0956713595000 194)

[8] "Changes of peel color and fruit quality in navel orange fruits under different storage methods" (https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=orange+qu ality+&btnG=)

[9] "Classification of Bitter Orange Essential Oils According to Fruit Ripening Stage by Untargeted Chemical Profiling and Machine Learning" (https://www.mdpi.com/1424-8220/18/6/1922)

[10] "Prediction of moisture ratio and drying rate of orange slices using machine learning approaches" (https://ifst.onlinelibrary.wiley.com/doi/abs/10.1111/jfpp.17011)

[11] "Fruit quality evaluation using machine learning techniques: review, motivation and future perspectives "(https://link.springer.com/article/10.1007/s11042-022-12652-2)

[12] "Machine Learning-Based Digital Twin for Monitoring Fruit Quality Evolution" (https://www.sciencedirect.com/science/article/pii/S1877050922002095)

[13] "Automatic Detection and Grading of Multiple Fruits by Machine Learning" (https://link.springer.com/article/10.1007/s12161-019-01690- 6)
[14] "Using machine learning techniques for evaluating tomato ripeness" (https://www.sciencedirect.com/science/article/pii/S0957417414006186
)

Appendices:

Data Cleaning and Preparation: The dataset underwent several preprocessing steps to ensure data quality and consistency. Missing values were handled, and columns with irrelevant information were dropped. Additionally, categorical variables were converted into numerical representations for regression analysis.

Linear Regression Model Building: A linear regression model was built to predict the ratings of the apps based on several input features such as reviews, size, installs, type, price, content rating, category, and genres. The model was trained using the training set and evaluated using the testing set.

Evaluation Metrics: Various evaluation metrics were utilized to assess the performance of the linear regression model. Mean squared error, mean absolute error, and mean squared log error were calculated to measure the accuracy of the predictions.

Regression using Statsmodels: An alternative approach to linear regression using the Least Square method from the statsmodels library was also explored. This method provided additional insights into the relationship between the input features and the target variable.

Visualization of Results: Visualizations such as scatter plots, bar plots, and heatmaps were used to visualize the data, model.

Predictions, and Evaluation Metrics. These visualizations aided in better understanding the patterns and relationships within the dataset.

References: The paper referred to various external resources, including Kaggle datasets, research papers, and documentation for libraries and tools used in the analysis.