Agriculture Commodities Price Prediction and Forecasting

Surya S, R. Dhamotharan ME...,
DEPARTMENT OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY, KALASALINGAM ACADEMY OF RESEARCH AND EDUCATION
suryasomasundaram4@gmail.com

ABSTRACT Recent days interaction between computer and human is gaining more popularity or momentum, especially in the area of speech recognition. There are many speech recognition systems or applications got developed such as, Amazon Alexa, Cortana, Siri etc. To provide the human like responses, Natural Language Processing techniques such as Natural Language Toolkit for Python can be used for analyzing speech, and responses. In our country, INDIA, agriculture is backbone of economy and major contributor for GDP. However, farmers often, do not get sufficient support or required information in the regional languages. Prediction analysis for farmers in agriculture is not only for crop growing but is essential to develop Crop recommendation system based on price forecasting for agricultural commodities in addition to providing useful advisories for the farmers of any state. Currently, to protect the farmers from price crash or control the inflation, the governments (Central and State) predicting the price for agricultural commodities using short-term arrivals and historical data. However, these methods are not giving enough recommendations for the farmers to decide the storage/sales options with evidence-based explanations. This project implements machine learning algorithms such as multi linear regression, Random Forest and Decision Tree regressor. To achieve the commodity price, by analyze the r2 score. The highest r2 score, the best model could be selected.

INDEX TERMS Crop Recommendation system, Price Prediction, Price Forecasting

I. INTRODUCTION

Technological advancements have fundamentally altered how humans interact with computers, particularly in the field of speech recognition. Applications like Amazon Alexa, Microsoft Cortana, and Apple Siri demonstrate how speech recognition technology can provide consumers with realistic, human-like responses. These systems use Natural Language Processing (NLP) techniques, such as the Natural Language Toolkit (NLTK) for Python, to analyze voice and support meaningful communications between people and gadgets. In a country like India, where agriculture is the backbone of the economy and makes a considerable contribution to GDP, technology advancements have the potential to play a critical role in assisting farmers. Farmers, on the other hand, frequently struggle to obtain critical information and recommendations in their own language, limiting their ability to make informed decisions. Such difficulties underscore the need for technology-driven solutions to close the information gap and empower farmers to improve their agricultural methods. Prediction analysis in agriculture goes beyond crop cultivation to include crop recommendation systems based on price predictions for agricultural commodities. These systems serve as vital advisories for farmers in many states, assisting them in making key crop selection, storage, and sales decisions. Traditionally, government agencies (both central and state) used short-term data on arrivals and historical trends to forecast agricultural commodity prices. Although these methods offer some insights, they frequently fall short of providing full suggestions that allow farmers to make evidence-based decisions about when and how to store or sell their product. The purpose of this study is to review existing research on agricultural prediction models and highlight the strengths and limits of various techniques. By assessing the current landscape, the study hopes to find areas for improvement and future research directions to increase the performance of crop recommendation systems and price forecasting tools.

The difficulties you've mentioned represent an exciting opportunity to use technology to empower farmers and enhance agricultural practices. Here's a breakdown of possible ways and considerations for tackling these issues:

Language Accessibility: Creating speech recognition and natural language processing systems in regional languages can considerably improve accessibility for farmers who are not fluent in English. To effectively interpret and reply to requests in several languages, robust language models must be trained.

Price Forecasting: Advanced predictive analytics and machine learning algorithms can help improve the accuracy of agricultural commodity price forecasts. Market demand, supply chain dynamics, weather patterns, and government regulations can all be factored into predictions to improve their accuracy.

Crop Recommendation System: A crop recommendation system must consider a variety of criteria, including soil quality, climate conditions, market demand, and historical production data. Machine learning techniques can be used to create individualized suggestions for farmers based on their unique conditions and aims.

Advisory Services: Providing timely and relevant advisories to farmers necessitates real-time data collection and processing. Weather predictions, pest and disease outbreak alerts, and market trends can be integrated into advisory systems to assist farmers in making informed crop management, pest control, irrigation scheduling, and marketing decisions.

Evidence-Based Explanations: To increase openness and trust in pricing forecasting and advising systems, recommendations must be supported by evidence. This entails presenting not only the expected results, but also the underlying data, assumptions, and reasoning for the suggestions. Farmers can use interactive visualization tools to better understand and analyze information.

Government Support: Collaboration among government agencies, academic institutions, and technology businesses is critical for creating and implementing creative solutions to help farmers. Governments may play an important role in sponsoring research, giving access to data and infrastructure, and enacting regulations to encourage the use of technology in agriculture.

Continuous Improvement: The creation of agricultural technology solutions is a continual process that necessitates constant monitoring, review, and improvement. Obtaining feedback from farmers, agricultural experts, and stakeholders is critical for finding areas for improvement and iterating on existing systems to better meet end-user demands. By addressing these issues through interdisciplinary collaboration and innovation, we can harness the power of technology to alter agriculture and enhance farmers' livelihoods in India and elsewhere.

MOTIVATIOIN OF THE WORK:

The motivation behind this work stems from the recognition of the critical role that agriculture plays in India's economy and the challenges faced by farmers in accessing timely and relevant information, particularly in their regional languages. Despite the significant contributions of agriculture to the GDP, many farmers struggle to make informed decisions due to the lack of support and information available to them.

- The increasing popularity and advancements in speech recognition and natural language processing technologies offer a promising avenue for addressing these challenges. By leveraging these technologies, it becomes possible to develop innovative solutions such as crop recommendation systems and advisory services tailored to the specific needs of farmers.
- The primary motivation of this study is to bridge the gap between technological advancements and the agricultural sector by exploring the potential of predictive analysis and machine learning in providing actionable insights to farmers. By integrating price forecasting models with evidence-based explanations, farmers can make more informed decisions regarding storage, sales, and other aspects of agricultural management.
- Furthermore, by conducting a comprehensive review of existing research in this area, the study aims to identify the strengths and weaknesses of different models and propose avenues for improvement. This includes exploring the use of advanced analytics techniques, real-time data integration, and user-centric design principles to enhance the effectiveness and usability of agricultural prediction and advisory systems.
- Ultimately, the goal of this work is to empower farmers with the tools and information they need to improve productivity, mitigate risks, and enhance their livelihoods. By leveraging the latest advancements in technology and data analytics, we can contribute to the sustainable development of India's agricultural sector and ensure the well-being of its farmers.

II. RELATED WORK

Feihu Sun et al., on agricultural price prediction has garnered attention due to its importance in sustainable agricultural development. Traditional methods like time series analysis and econometric models have been complemented by intelligent forecasting methods such as machine learning and deep learning techniques. Moreover, combination models that integrate various approaches have shown promise in enhancing prediction accuracy. Emerging trends involve blending structured data (e.g., historical prices) with unstructured data (e.g., news and social media) for comprehensive insights. Researchers face challenges in balancing forecast accuracy and trend precision while exploring optimal model combinations. This literature review underscores the potential of hybrid models and the importance of integrating diverse data sources to improve agricultural price forecasting.

Nhat-Quang Tran., application of machine learning algorithms in agricultural price prediction has become increasingly popular due to their potential to enhance prediction accuracy and adaptability. This review explores recent research on machine learning techniques for forecasting agricultural prices. The importance of agriculture, particularly in developing countries, and the impact of crop price volatility highlight the necessity of improved prediction methods. Various machine learning approaches, such as decision trees, support vector machines, and neural networks, have been investigated for their effectiveness. While these algorithms offer significant 8 promise, challenges remain regarding data quality, model interpretability, and scalability. Further research is needed to optimize these techniques and overcome limitations for more robust and precise agricultural price forecasting.

Zhiyuan Chen., research on automated agricultural commodity price prediction systems utilizing novel machine learning techniques focuses on improving prediction accuracy and addressing challenges in forecasting historical data. Recent studies have shifted from traditional statistical methods to advanced machine learning approaches due to large datasets and the complexity of price fluctuations. Popular algorithms such as ARIMA, SVR, Prophet, XGBoost, and LSTM have been extensively compared using historical data from Malaysia. Findings suggest that the LSTM model, with its ability to handle nonlinearity and long-term dependencies, performs best with an average mean square error of 0.304. While machine learning strategies show promise, careful selection of data and optimization of model parameters remain critical for effective predictions.

Arushi Singh., research on modern agricultural advances has focused on using machine learning algorithms for crop prediction to help farmers make informed decisions on crop cultivation based on climatic conditions and soil nutrients. Popular algorithms such as K-Nearest Neighbor (KNN), Decision Tree, and Random Forest Classifier have been compared in recent studies to evaluate their effectiveness in crop prediction. Different criteria like Gini and Entropy have been used for these evaluations. Results indicate that Random Forest Classifier outperforms the other models, providing the highest accuracy in predictions. This approach assists farmers in selecting appropriate 9 crops, improving productivity, and adapting to environmental challenges while promoting sustainable agriculture.

Banupriya N., Recent research on crop yield prediction in India has shifted focus from complex environmental and agricultural factors to simpler, more accessible data points. This approach aims to facilitate the direct application of predictions by farmers without requiring in-depth understanding of underlying technology. By utilizing basic factors such as state, district, crop type, and season (e.g., Kharif, Rabi), researchers can efficiently gather and analyze data from the Indian Government Repository. Advanced regression techniques like Random Forest, Gradient Boosting, and Decision Tree algorithms have been explored to predict yield, while ensemble algorithms are employed to enhance accuracy and minimize errors. This streamlined approach aids farmers in making informed decisions for improved productivity and sustainability.

III.MACHINE LEARNING ALGORITHMS

MACHINE LEARNING:

Machine learning (ML) is a branch of artificial intelligence (AI) and computer science that focuses on the using data and algorithms to enable AI to imitate the way that humans learn, gradually improving its accuracy.

1.Linear Regression Model/OLS Model

Overview: Ordinary Least Squares (OLS) is a linear regression technique used to estimate the relationship between a dependent variable and one or more independent variables. The primary goal is to find the line (or hyperplane in higher dimensions) that minimizes the sum of the squared differences between the observed and predicted values. OLS is widely employed in statistical modeling, econometrics, and machine learning.

MLR Equation

The multiple linear regression (MLR) equation models the relationship between multiple independent variables(X1,X2,...,Xn)and a dependent variable (Y). The general form of the MLR equation is:

$Y = \beta 0 + \beta 1x1 + \beta 2x2 + \cdots + \beta kxk + \varepsilon$

Where:

- Y is the dependent variable.X1,X2,…,Xn are the independent variables.
- β0 is the y-intercept (constant term).
- β1 ,β2,…,βn are the coefficients that represent the strength and direction of the relationship between the independent variables and the dependent variable.
- ε is the error term, representing the unobserved factors that affect the dependent variable but are not included in the model.
- The goal of MLR is to estimate the coefficients (1,…,β0,β1,…,βn) that minimize the sum of squared differences between the observed and predicted values of the dependent variable.

2.Decision Tree Regression

Decision Tree Regression is a supervised machine learning algorithm used for predicting continuous outcomes. Unlike decision trees in classification, which predict discrete class labels, decision tree regression predicts a numeric target variable. The algorithm works by recursively partitioning the dataset into subsets based on feature conditions, ultimately producing a tree structure where each leaf node corresponds to a predicted numerical value.

Evaluation Metrics in Decision Tree

In the context of Decision Trees, several evaluation metrics are commonly used to assess the performance of the model. These metrics provide insights into how well the decision tree is making predictions compared to the actual outcomes. Here are some key evaluation metrics for Decision Trees:

Accuracy:

- Definition: The ratio of correctly predicted instances to the total number of instances.
- Formula:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

- Consideration: Accuracy is a straightforward metric but may be misleading in imbalanced datasets.

Precision

- Definition: The ratio of correctly predicted positive observations to the total predicted positives.
- Formula:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

- Consideration: Precision focuses on the accuracy of positive predictions and is valuable when the cost of false positives is high.

Recall (Sensitivity or True Positive Rate):

- Definition: The ratio of correctly predicted positive observations to all actual positives.
- Formula:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

- Consideration: Recall emphasizes capturing as many actual positives as possible and is crucial when missing positives is costly.

F1 Score:

- Definition: The harmonic mean of precision and recall, providing a balance between the two metrics.
- Formula:

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- Consideration: F1 Score is useful when there’s a need to balance precision and recall

Confusion Matrix:

- Definition: A table that presents a summary of the model's predictions against the actual outcomes, showing True Positives, True Negatives, False Positives, and False Negatives.

- Use: Provides a detailed breakdown of the model's performance and aids in calculating other metrics.

ROC-AUC (Receiver Operating Characteristic-Area Under the Curve):

- Definition: A graphical representation of the trade-off between true positive rate (sensitivity) and false positive rate at various thresholds.
- Use: Measures the model's ability to discriminate between positive and negative instances.

Gini Index (for Decision Trees):

- Definition: A measure of impurity in a node. It assesses how often a randomly chosen element would be incorrectly classified.
- Use: Decision Trees aim to minimize the Gini Index at each split, resulting in a tree that classifies instances more accurately.

3.Random Forest Regression

Random Forest Regression is an ensemble learning technique that extends the concept of Random Forests, originally designed for classification problems, to regression tasks. It is a powerful and flexible algorithm that leverages the strength of multiple decision trees to make more accurate and robust predictions for continuous outcomes.

Key Features and Concepts:

Ensemble of Decision Trees:

- Random Forest Regression is built on an ensemble of decision trees. Multiple decision trees are constructed independently, and their predictions are averaged to obtain a final result.

Bagging (Bootstrap Aggregating):

- Each tree in the Random Forest is trained on a bootstrap sample (randomly selected with replacement) from the original dataset. This helps introduce diversity among the trees.

Random Feature Selection:

- At each node of a decision tree, a random subset of features is considered for splitting. This randomness adds further diversity to the individual trees.

Prediction Aggregation:

- For regression, the predictions of individual trees are averaged to produce the final output. This ensemble approach helps mitigate overfitting and improves generalization.

Handling Missing Values:

- Random Forests can effectively handle missing values in the dataset, reducing the need for extensive data preprocessing.

Robust to overfitting:

- The ensemble nature of Random Forests tends to reduce overfitting, making them less sensitive to noise and outliers in the data.

Versatility:

- Random Forests can be applied to a wide range of regression tasks and are suitable for datasets with a large number of features.

DATA DESCRIPTION:

Number of Data: 62429

Number of attributes: 11

The purpose of this is to identify the study already done in this field and find out the benefits and downsides of different models as well as the future scope of improvement. Attribute information

1.APMC (Agricultural Produce Market Committee):

- Definition: The specific market committee responsible for the regulation and oversight of agricultural trade in a particular area.

- Use: Identifies the market where the data was collected.

2.Commodity:

- Definition: The type of agricultural commodity being traded.
- Use: Specifies the particular crop or product involved in the market transactions.

3.Year:

- Definition: The calendar year when the market transactions took place.
- Use: Provides the temporal dimension for the data.

4.Month:

- Definition: The month during which the market transactions occurred.
- Use: Offers a more granular temporal reference in conjunction with the year.

5.Arrivals_in_qtl (Arrivals in Quintals):

- Definition: The quantity of the commodity brought to the market, measured in quintals.
- Use: Indicates the volume of the commodity traded in the market.

6.Min_price:

- Definition: The minimum price at which the commodity was traded.
- Use: Represents the lowest price observed for the commodity during the specified time.

7.Max_price:

- Definition: The maximum price at which the commodity was traded.
- Use: Represents the highest price observed for the commodity during the specified time.

8.Modal_price:

- Definition: The modal (most frequently occurring) price of the commodity.
- Use: Provides a measure of the central tendency of the commodity prices in the market.

9.Date:

- Definition: The specific date of the market transactions.
- Use: Offers a precise temporal reference for individual market events.

10.District_name:

- Definition: The name of the district where the market is located.
- Use: Specifies the geographical location of the market.

11.State_name:

- Definition: The name of the state where the market is located.
- Use: Specifies the broader geographical region in which the market operates.

A. DATA PREPROCESSING

f.isnull().sum(): this line of code provides a series where the index represents the column names of the Data Frame, and the values represent the count of null values in each column. This is useful for understanding which columns have missing data and the extent of that missing data.
df.info():the df.info() method provides a concise summary of a pandas Data Frame (df).It includes information about the Data Frame’s index and data types of each column, as well as memory usage. It also provides the number of non-null values for each column, helping you understand the data's completeness. This method is useful for quickly understanding the structure of your data, including the types of data each column contains and how much data is missing.

FEATURE ENGINEERING FOR NUMERICAL COLUMNS

When you apply MinMaxScaler from sklearn.preprocessing to normalize the numerical columns in your DataFrame (df_num) and then assign the transformed data to a new DataFrame (df_num_mn), you will end up with a DataFrame where the numerical columns have been scaled to a range of 0 to

1. Fit and Transform: By applying the fit_transform method, the MinMaxScaler calculates the minimum and maximum values for each column in the input DataFrame (df_num) and scales the data to a range between 0 and 1.
2. New DataFrame: The result of this transformation is a new DataFrame (df_num_mn) with the same column names as df_num but with the values scaled.

FEATURE ENGINEERING FOR CATEGORICAL COLUMNS

When you use LabelEncoder from sklearn. preprocessing to encode categorical columns in your DataFrame (df_cat), the function transforms each categorical column into a numerical format by assigning integer labels to each unique value in the column. This process is known as label encoding.

1. Label Encoding: For each specified categorical column (e.g., 'APMC', 'Commodity', 'Month', 'district_name', 'state_name'), LabelEncoder assigns a unique integer value to each distinct category.
2. Encoded Columns: The encoded columns replace the original categorical columns with their respective integer labels.

B. MODEL TRAINING

- In the present investigation, an attempt has been made to explore efficient ML algorithms e.g., Support Vector Regression (SVR), Random Forest (RF) and Multiple Linear Regression for forecasting wholesale price of crops in 33 major markets of Maharashtra.
- The superiority of the models is established by means of R2-score, and other accuracy measures such as Mean Error (ME), Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Prediction Error (MAPE).

Splitting-trained Models

The training set is used to train the model, while the testing set is used to evaluate its performance on unseen data. This helps assess how well the model generalizes to new observations.

Once the data is split into training and testing sets, you typically use the training set to train your model (fit the parameters) and the testing set to evaluate its performance by making predictions on the test data and comparing them to the actual values.

It's crucial to ensure that the split is done randomly to avoid any biases in the data. Common ratios for splitting data include 70/30, 80/20, or 90/10, depending on the size of your dataset and the specific requirements of your analysis.

- The dataset has been divided into two – Training and Testing
- The proportion of data allocated to each set is determined by the test_size parameter. In this case, 20% of the data is assigned to the testing set, while 80% is assigned to the training set.

Count Plot for 'min_price' and 'modal_price': It generates two subplots stacked vertically. Each subplot is a count plot that shows the distribution of values for the 'min_price' and 'modal_price' columns from the DataFrame 'df.

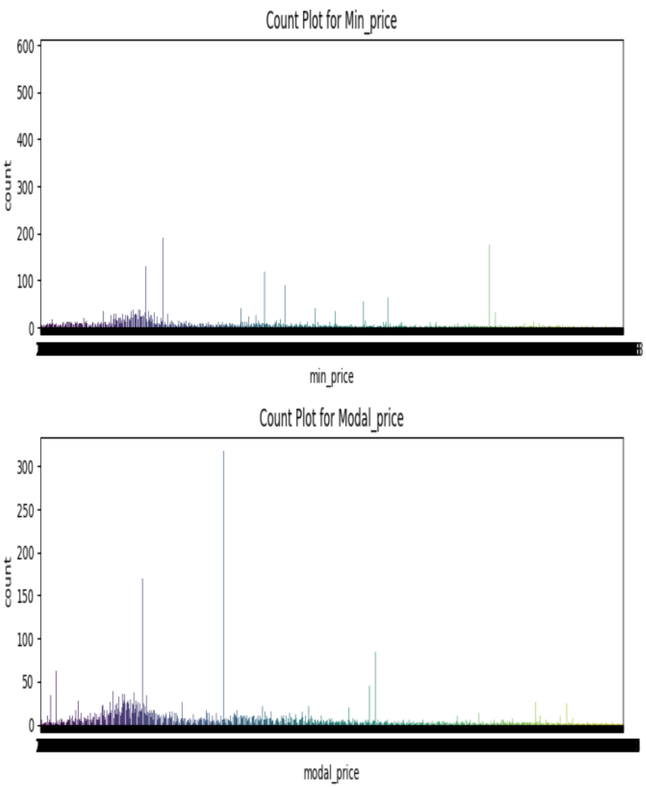


Fig.1 Count Plot

Scatter Plot of 'Min_price' versus 'Max_price':

It generates a scatter plot illustrating the relationship between the'min_price' and'max_price' columns in the DataFrame 'df'.

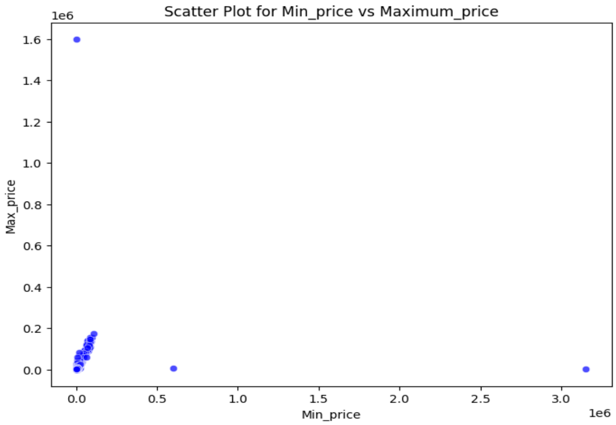


Fig.2 Scatter Plot

Crosstab:

It generates a cross-tabulation of the columns 'APMC' and 'Commodity' in the DataFrame 'df'. This gives a frequency distribution of the occurrences of different commodities in various
APMCs.

Correlation Plot:

It computes the correlation matrix for each numeric column in the DataFrame 'df'. It generates a heatmap visualisation of the correlation matrix, demonstrating the relationship between various numerical parameters.

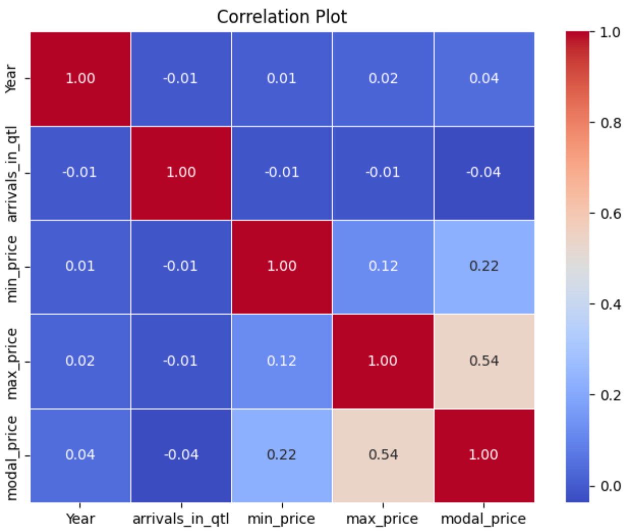
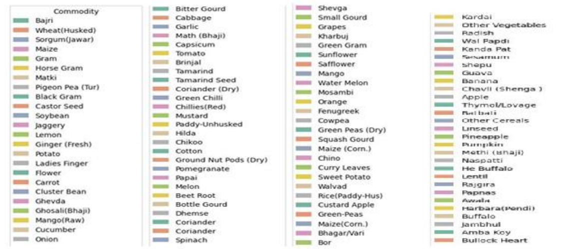


Fig.3 Correlation Plot

Bar Chart for 'Month' and 'Commodity':

It generates a bar chart depicting the number of commodities for each month.
The hue parameter is used to differentiate commodities based on color.



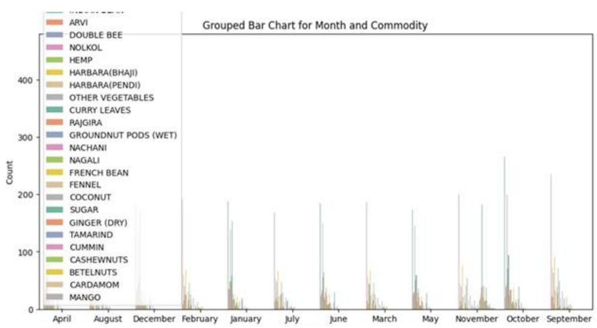


Fig.4 Bar Chart for Month and Commodity

Bar Chart for 'Commodity':

It generates a basic bar chart depicting the count of each commodity.

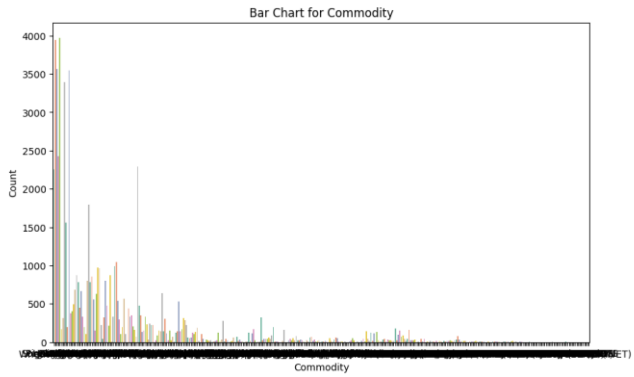


Fig.4 Bar Chart for Commodity

IV. CONCLUSION AND FUTURE WORKS

In order to empower Indian farmers and improve agricultural decision-making, this research examined a large dataset containing market transactions, crop specifics, and pricing information. Exploratory data analysis provided insights that guided the creation of crop recommendation and price forecasting models, addressing important agricultural concerns. The findings highlighted geographical and temporal diversity in agricultural prices, emphasizing the need for localized knowledge. Moving forward, increasing model sophistication using advanced techniques like ensemble learning and including external elements like weather patterns and government policies would improve model accuracy. Deploying user-friendly tools and encouraging community interaction will guarantee that offered solutions are realistic and sustainable, eventually benefiting farmers and increasing agricultural prosperity.

REFERENCES

[1] https://www.india.gov.in/official-website-commission-agricultural-costs-and-pricesN. Abiwinanda, M. Hanif, S. T. Hesaputra, A. Handayani, and T. R. Mengko, ‘‘skin cancer tumor classification using convolutional neural network,’’ in Proc. World Congr. Med. Phys. Biomed. Eng., vol. 68, 2018, pp. 183–189, doi: 10.1007/978-981-10-9035-6_33.

[2] www.maths.qmul.ac.uk/ bb/TSChapter72.pdf -accessed on May,16 2019

[3] Rehman U, Bahar F, Saad A, Bhat M, Singh L, Mahdi S, Mir MS, Bashir M, Bhat TA, Palmo T (2022) Trends, challenges and prospects of underutilized food crops under temperate ecologies: a review.

[4] Liwen, L.; Shixin, C.; Dabin, Z.; Boting, Z. A Multi-Time Scales Combination Strategy for Pork Price Forecasting. J. Syst. Sci. Math. Sci. 2021, 41, 2829–2842.

[5] Zhao, L.; Zeng, G.; Wang, W.; Zhang, Z. Forecasting oil price using web-based sentiment analysis. Energies 2019, 12, 4291. [CrossRef]

[6] Asma Hasifa Nurcahyono, Fhira Nhita,Deni Saepudin and Annisa Aditsania,“Price Prediction of Chili in Bandung Regency Using Support Vector Machine (SVM) Optimized with an Adaptive Neuro- Fuzzy Inference System (ANFIS),” 7th International Conference on Information and Communication Technology (ICoICT), 2019.

[7] Dabin Zhang, Shanying Chen , Liwen Ling, And Qiang Xia “Forecasting Agricultural Commodity Prices Using Model Selection Framework With Time Series Features and Forecast Horizons,” IEEE Access February 14, 2020 .

[8] Darekar, Ashwini and B Datarkar, Snehal(December 2016),”Onion Price Forecasting on Kolhapur Market of Western Maharashtra”International Journal of Information Research and Review.

[9] V.K. Verma, P. Kumar, Singh. S.P and H. Singh(July 2016), ”Use of ARIMA modeling in forecasting coriander prices for Rajasthan,International J. Seed Spices 6(2),40-45.

[10] R. Lombardo, J. Flaherty(January 24-27 2000),”Modelling Private New Housing Starts in Australia”,Pacific-Rim Real Estate Society Conference, University of Technology Sydney(UTS)