TIME SERIES MODELLING FOR FORECASTING PRICE AND ARRIVAL OF JOWAR CROP IN MAHARASHTRA MARKETS

A RESEARCH PROJECT SUBMITTED TO

KARMAVEER BHAURAO PATIL UNIVERSITY, SATARA

FOR THE COURSE OF

M.SC.

IN

STATISTICS

BY

Ambupe Pooja Sanjay

Yashavantrav Chavan Institute of Science, Satara(Autonomous)

Lead college of Karmaveer Bhaurao Patil University,Satara

2024-25

Abstract:

In recent years, accurate forecasting of crop prices and arrivals has become increasingly crucial for supporting farmers and stabilizing agricultural markets. This study focuses on predicting the monthly arrival and modal price of jowar (sorghum) in Maharashtra’s markets using dynamic time series models. By employing statistical and machine learning technique specifically Exponential Smoothing, ARIMA, LSTM, hybrid LSTM-ARIMA, Random Forest and XGBoost the research examines the influence of variables such as rainfall, temperature, seasonality, and lagged arrivals on market dynamics. The dataset includes historical data on jowar arrivals and prices from 2010 to 2023 was obtained from Agmarknet website. Each model's performance is evaluated through metrics like Mean Absolute Percentage Error (MAPE) and ensuring reliable comparisons for forecasting accuracy. This research delivers a forecasting framework that supports Maharashtra’s jowar farmers in making informed market decisions, promotes efficient interventions, and ultimately contributes to the resilience of the agricultural economy. This research delivers a forecasting framework that supports Maharashtra’s jowar farmers in making informed market decisions, promotes efficient interventions, and ultimately contributes to the resilience of the agricultural economy.

Keywords: Agricultural Markets, Jowar Crop, Time series

Introduction:

Time series analysis is a branch of statistical and data analysis that focuses on understanding and modelling data points collected or recorded at specific time intervals. Unlike traditional data analysis, where observations are assumed to be independent, time series data is inherently ordered, and each observation is typically dependent on previous observations. This sequential nature of time series data introduces unique challenges and opportunities for analysis. By recognizing and modelling the inherent patterns in time series data, analysts and researchers can gain valuable insights that support better decision-making and planning in various domains. In time series forecasting, the results are the predicted outputs from the trained models. There are many forecasting models available like LSTM and ARIMA models. ARIMA forecasts temporal dependencies using only historical values. These models help to gain better insights into the data and predict future trends.

Agriculture is a critical sector that plays a significant role in the economy, especially in countries where a large portion of the population depends on farming for their livelihood. Maharashtra is, a state in India that plays a vital role in the country’s Agriculture sector. Agriculture is the backbone of Maharashtra's economy. The state is known for its diverse agro-climatic conditions, which allow for the cultivation of a wide variety of crops. Maharashtra's agriculture is characterized by a wide range of crops, including cereals, pulses, oilseeds, fruits, vegetables, and cash crops like cotton and sugarcane. Jowar (sorghum) is one of the most important cereal crops in the world and is one of the four major food grains of our country. It is a staple food for millions of poor rural people in Asian and African countries. The origin of sorghum is generally believed to be around present-day Ethiopia or East Central Africa. Sorghum was taken from East Africa to India during the first millennium. India is one of the major producing countries. In the year 2005-06, Maharashtra occupied the highest position in the production of Jawar with 3.90 million tons of production (51.11%).

Crop prediction in Maharashtra is a critical area of research and policy planning, given the state's diverse agro-climatic zones, varied cropping patterns, and significant reliance on agriculture. Accurate crop predictions are essential for ensuring food security and enhancing the livelihoods of millions of farmers in the state. By leveraging advanced technologies and integrating region-specific data, accurate crop predictions can help address the challenges faced by the agricultural sector in the state. This paper aims to leverage dynamic time series modeling to better understand and predict the behaviors of jawar crop arrivals and prices in Maharashtra's markets.

Objectives:

1. To analyze historical trends and seasonal patterns in jawar crop arrivals and prices in Maharashtra's markets
2. To develop and implement dynamic forecasting models such as ARIMA, LSTM, hybrid LSTM-ARIMA, and ensemble models (Random Forest, XGBoost) for accurately predicting jawar crop arrivals and prices.
3. To evaluate and compare the performance of different forecasting models.
4. To provide useful insights for farmers and policymakers by predicting future trends in jawar crop arrivals and prices.

Literature Review:

This literature indicates that while traditional models remain valuable for certain applications, neural network-based approaches like LSTM and CNN often provide better accuracy and adaptability, making them suitable for predicting sales in dynamic and seasonal markets, such as the furniture industry. This study aims to build on these insights by comparing various forecasting techniques to identify the most effective method for predicting future sales in the retail sector**.[2]**

The literature on price forecasting for brinjal (eggplant) highlights the importance of accurate predictions for aiding farmers, traders, and consumers in decision-making. Traditional statistical models like ARIMA have been commonly used but are limited by their assumptions and inability to capture non-linear patterns. These studies indicate that ML models outperform traditional methods in handling complex time series data, capturing sudden price fluctuations, and providing more accurate forecasts.[**4]**

The literature on time series forecasting in agricultural markets, indicates the importance of accurate prediction models. Studies have applied models such as the Autoregressive Integrated Moving Average (ARIMA) and Seasonal ARIMA (SARIMA) for stationary time series with clear seasonality and trends. Advanced models like Prophet and Simple Recurrent Neural Networks (RNNs) are also used for non-stationary or complex nonlinear time series. These models help predict market dynamics, which is crucial for stakeholders such as farmers and traders**.[3]**

This study indicates that price prediction and forecasting in agriculture highlight the importance of selecting the right algorithm for accurate predictions. Traditional models are used, particularly when seasonality influences price variations. Advanced machine learning models, including ensemble methods like stacking, have shown promise in improving prediction accuracy by combining multiple models. Recent research suggests that hybrid and ensemble models often outperform individual models, particularly for complex and nonlinear datasets**.[1]**

The literature on time series forecasting highlights its critical role in various domains such as electricity demand, cloud workload, weather, sales, and business costs. Research has evolved from traditional statistical methods like ARIMA to advanced machine learning and deep learning approaches, including CNNs, LSTMs, and their hybrid forms like CBLSTM. These modern techniques offer improved accuracy and flexibility compared to older methods. The paper surveys these forecasting methods, comparing their effectiveness across different types of time series and applications.**[10]**

Methodology:

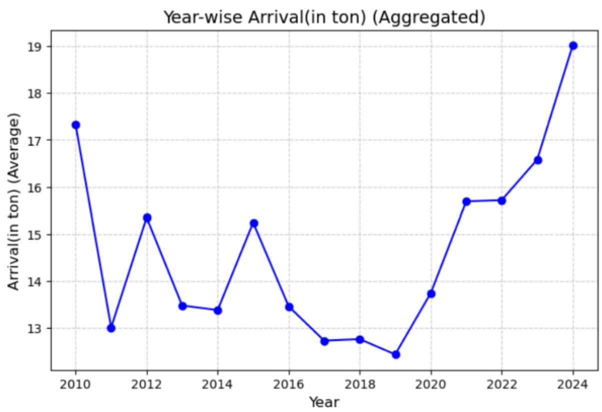
Secondary data on Jowar crop price and arrival was obtained from the Agmarknet website, a reliable source of agricultural market data in India. The dataset included monthly data from multiple agricultural markets in Maharashtra, covering price (in INR per quintal) and arrival (in tons) of Jowar. The data period spans 2010 to 2024, allowing for comprehensive trend analysis over recent years.

For predicting Jowar prices and arrivals, we focused target variables as modal price (INR/quintal) and arrival (in tons) and independent variables as rainfall, temperature, rain days, and time-based variables (month and year). Seasonal categories (e.g., summer, monsoon, winter) were generated from the month data to capture any cyclical effects in Jowar arrivals and prices. The preliminary aim of this study is to develop time series forecasting models to predict the modal price and arrival of jowar crop, leveraging historical data and multivariate features such as rainfall, temperature, rainy days, and lagged arrival.

Statistical Analysis:

- **Graphical Representation:**

1) Year wise Price(per quintal) and Arrival(in ton)



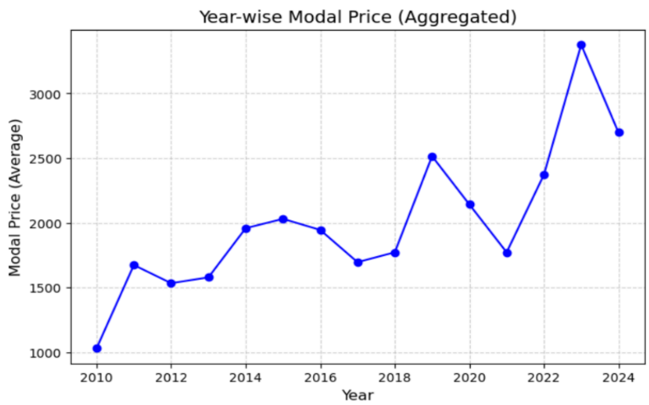
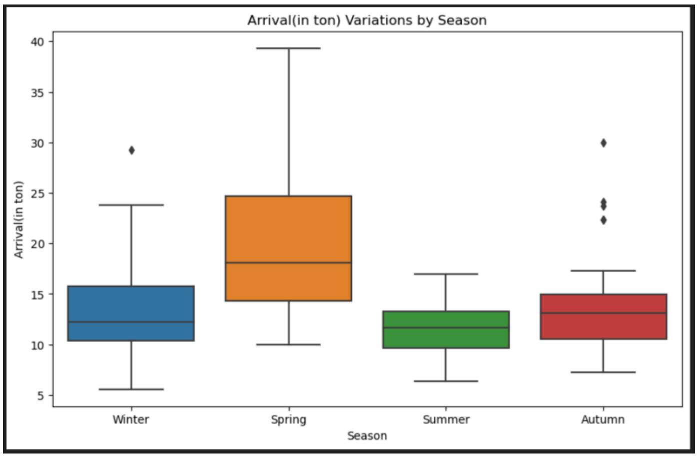


Fig 1

2) Variation of Arrival and Price by Season:



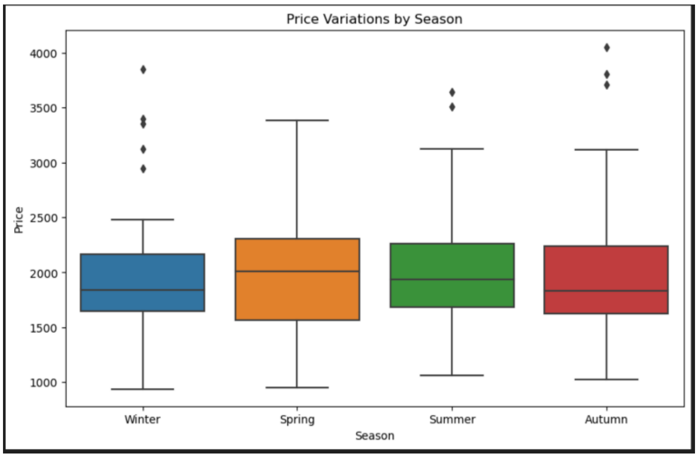


Fig 2 **Fig 3**

- **Modelling:**

1) Exponential Smoothing :

Exponential smoothing is a baseline method of time series forecasting that assigns exponentially decreasing weights to past observations. This approach gives more importance to recent data. It handles both trend and seasonality, either in additive or multiplicative form. In this project ,The Holt-Winters seasonal method was used due to the observed trend and seasonal patterns in the price and arrival data. The models' parameters were optimized to minimize error metrics such as Mean Absolute Percentage Error (MAPE).

For Arrival(in ton):

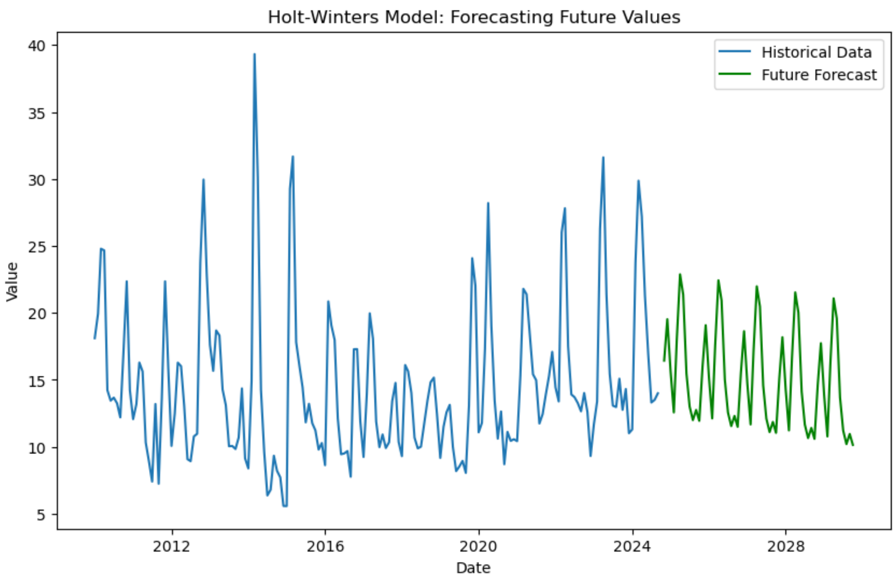


Fig 4

For Price (per quintal):

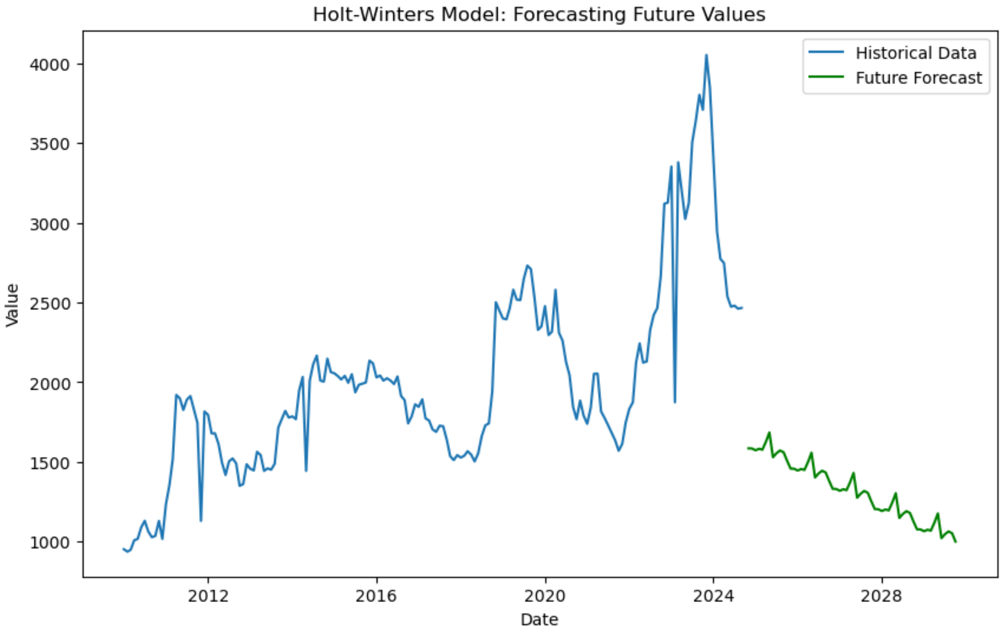


Fig 5

2) <u>ARIMA:</u>

The ARIMA (Auto-Regressive Integrated Moving Average) model is a time series forecasting technique that combines three components: Auto-Regressive (AR), Differencing (I), and Moving Average (MA). Each of these components plays a specific role in capturing patterns and relationships in time series data. The AR component captures the relationship between an observation and a specified number of previous observations (lags). The Integrated (I) part refers to differencing, which is used to make the time series data stationary by removing trends or seasonality. The MA component models the relationship between an observation and a residual error from a moving average of previous error.

For Price (per quintal):

By ADF test, it is concluded that data is not stationary. Hence ,we use differencing to make the data stationary.

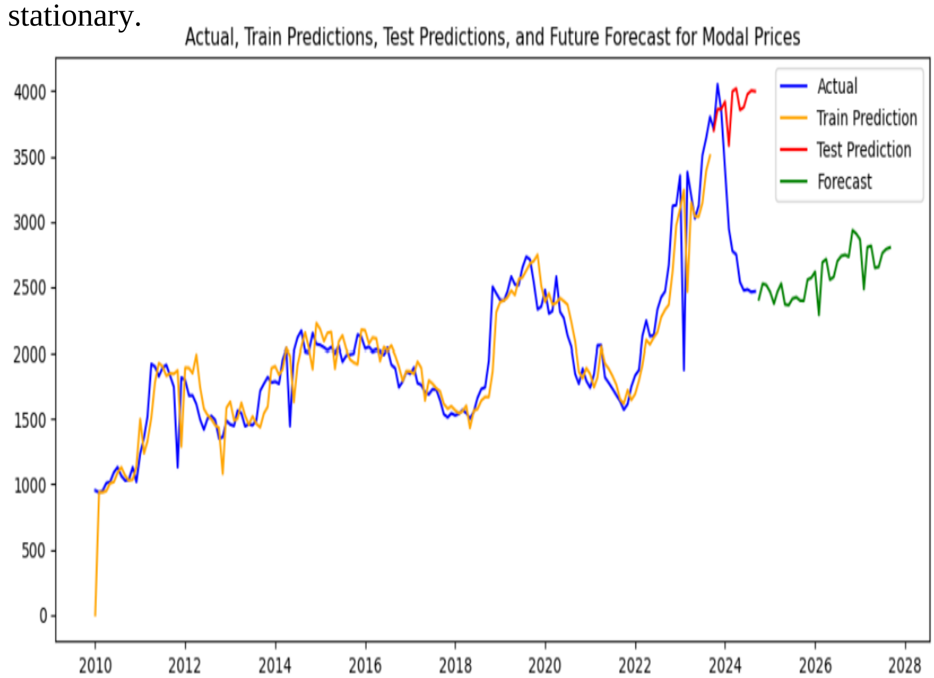


Fig 6

For Arrival (in ton):

By ADF test, it is concluded that data is stationary.

The ARIMA model is fitted to the data using the selected p, d and q values, combining past values and errors to predict future points.

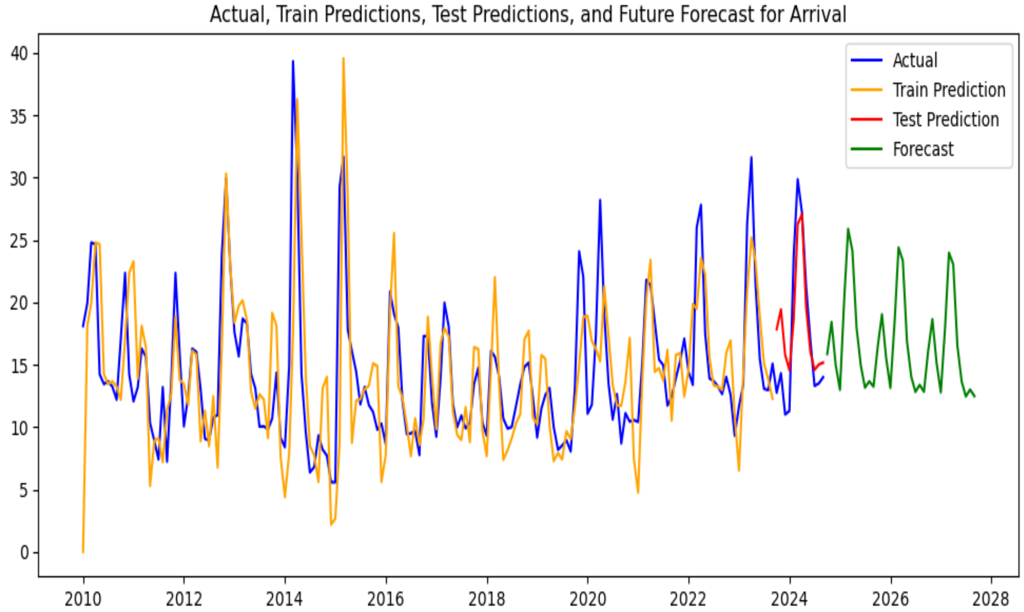


Fig 7

3) **Random forest:**

Random Forest was used to predict two target variables: price and arrival of the jowar crop. The model takes into account several input features like rainfall, temperature, lagged arrival values, modal price and season. It handles both numerical and categorical variables. It builds multiple decision trees, each learning from different parts of the data. By combining the predictions from all these trees, the model becomes more accurate.

For Arrival (in ton):

For forecasting arrival, the input features used as rainfall, temperature, lagged arrival values, modal price and season. The dataset was split into training and testing sets to allow the model to learn patterns on one part of the data and then evaluate on test data. The model was trained on the 80% training set, allowing it to learn the patterns in the data. The remaining 20% was reserved for testing to evaluate the model's performance on unseen data. The model's accuracy was assessed using evaluation metrics like Mean Absolute Percentage Error (MAPE).

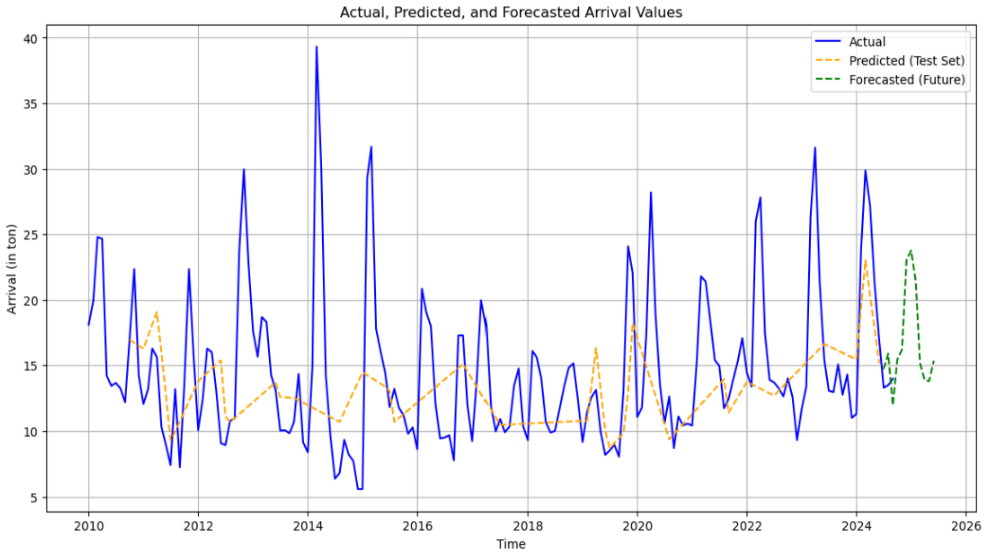


Fig 8

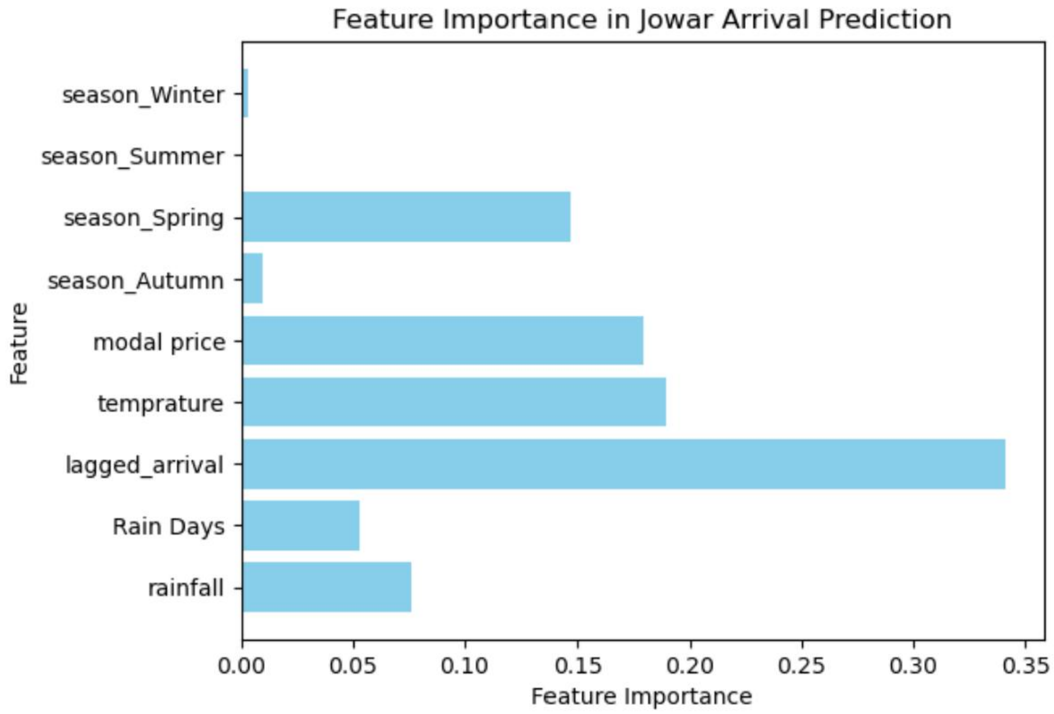


Fig 9

<u>For Price(in quintal):</u>

For forecasting price, the input features used as rainfall, temperature, lagged modal price, season, rain days.

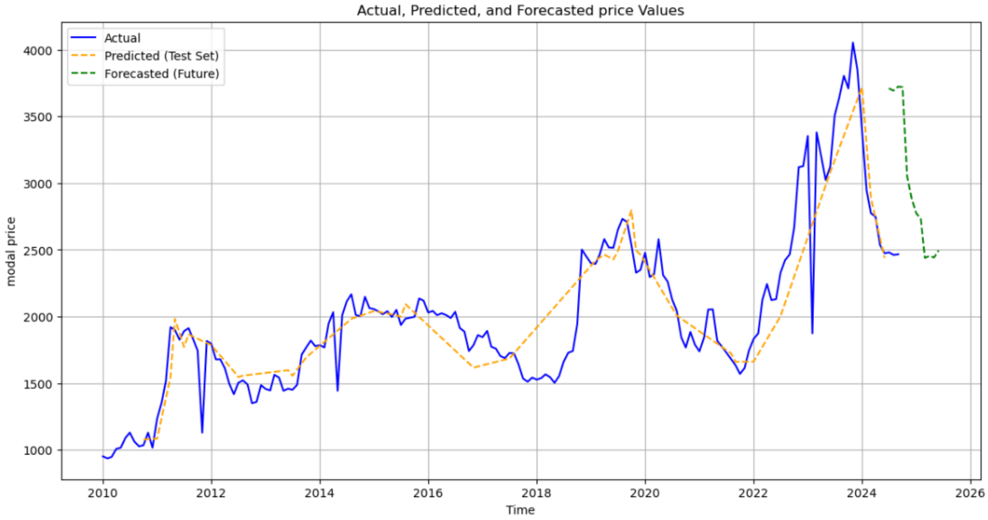


Fig 10

4) XGBoost:

XGBoost is used in this project for forecasting the arrival and price of jowar because it is highly effective at capturing complex patterns and relationships in the data. Given the time series nature of my project, which includes features like rainfall, temperature, lagged arrival values, and modal price, XGBoost can handle both non-linear relationships and interactions between these variables, which traditional models like ARIMA may struggle with. Additionally, XGBoost is robust to overfitting.

For arrival(in ton):

The model was trained on the 80% training set, allowing it to learn the patterns in the data. The remaining 20% was reserved for testing to evaluate the model's performance on unseen data. The model's accuracy was assessed using evaluation metrics like Mean Absolute Percentage Error (MAPE).

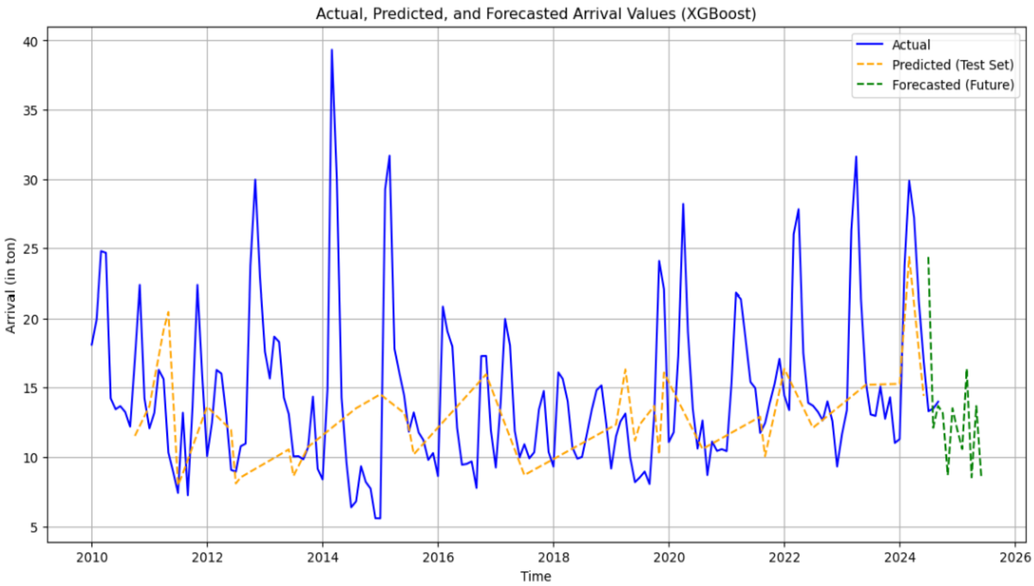


Fig 11

For Price(in quintal):

For forecasting price, the input features used as rainfall, temperature, lagged modal price, season, rain days.

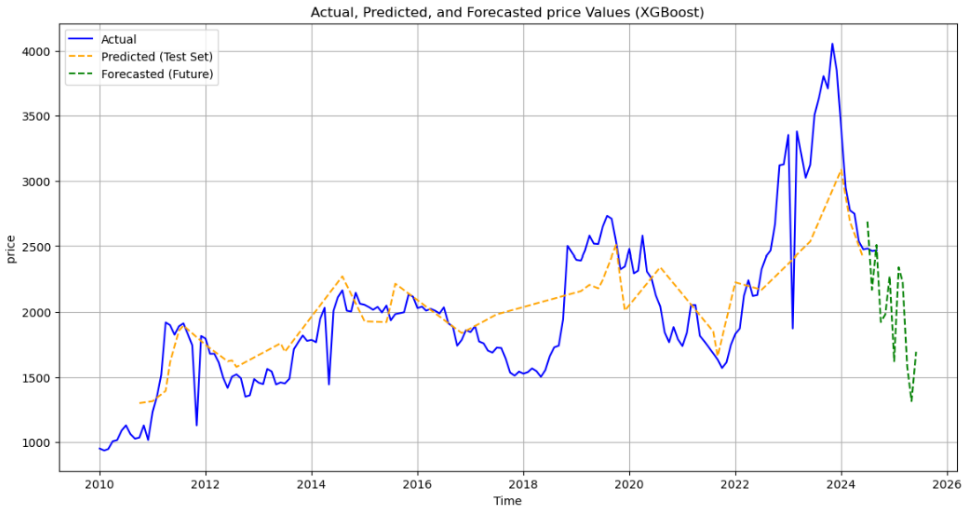


Fig 12

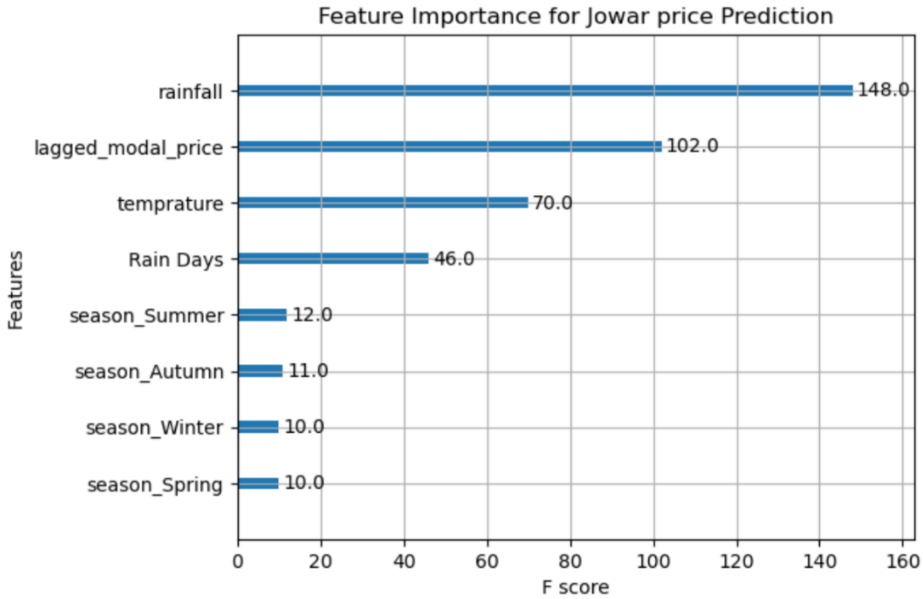


Fig 13

5) <u>ARIMAX-LSTM Hybrid Model:</u>

The hybrid model is used for two approaches as ARIMAX for linear patterns with external factors and LSTM for non-linear relationship to improve the model performance. Residuals are calculated as the difference between actual values and ARIMAX predictions, capturing nonlinear patterns. Residuals from the ARIMAX model, which represent nonlinear components, are used as input for the LSTM model. LSTM learns the complex nonlinear particularly those not captured by ARIMAX.

For Arrival(in ton):

For predicting Arrival of Jowar crop, we use **exogenous variables** as rainfall, temperature, rain days, modal price.

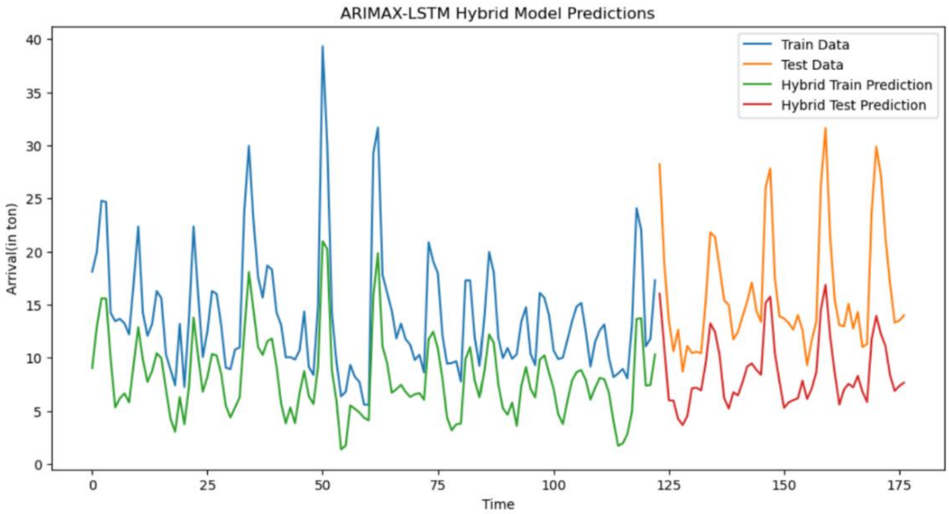


Fig 14

For Price (in quintal):

For predicting Price of Jowar crop, we use **exogenous variables** as rainfall, temperature, rain days, modal arrival(in ton).

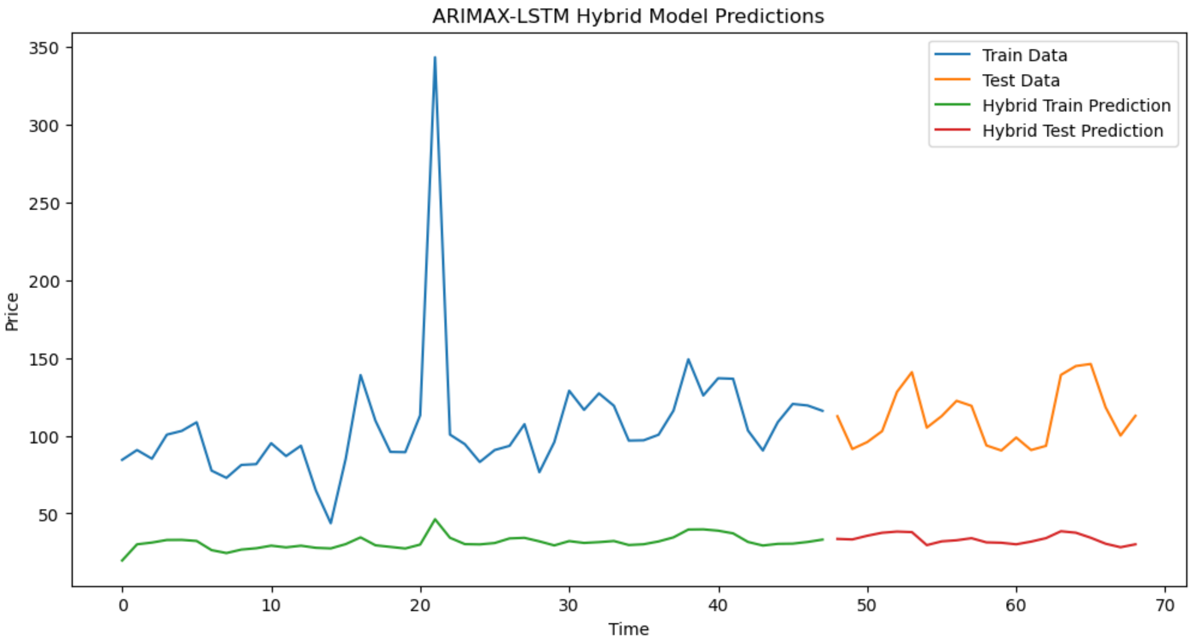


Fig 15

6) <u>**Multivariate LSTM**</u>:

Long Short-Term Memory is type of RNN designed to learn long-term relationship in sequential data. In this study, multivariate LSTM model was employed to predict **modal price** and **arrival**, considering multiple influencing factors.The data was split into 80% for training and 20% for testing. The model was trained to minimize loss (e.g., Mean Squared Error or Mean Absolute Percentage Error) using an optimization algorithm like Adam. The model predicts modal price and arrival simultaneously, utilizing the relationships between features and target variables across time.

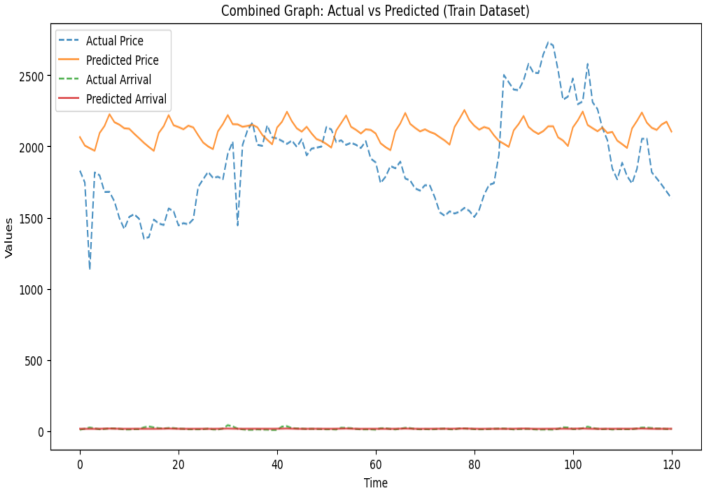


Fig 16

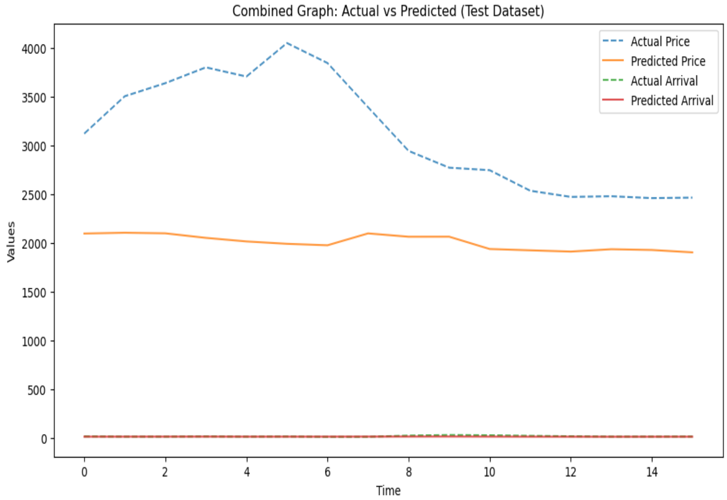


Fig 17

<u>**Conclusion:**</u>

Model Name	**Price**		**Arrival**	
	Train MAPE	Test MAPE	Train MAPE	Test MAPE
Exponential smoothing	18.93%	21.02%	5.63%	42.59%
Arimma	7%	35%	23%	18%
Random forest	4.38%	5.61%	11%	25%
Xgboost	9.78%	10.02%	18.27%	27.88%
Hybrid (ARIMAX+LSTM)	28%	19%	44%	46%
Multivariate LSTM	Price and arrival			
	25.40%	26.71%		

Table 1.1

Exponential Smoothing showed moderate accuracy for price prediction, with a small difference between train and test MAPE but may not suitable for arrival prediction producing high overfitting. ARIMA performed well for price prediction in the training data but showed significant overfitting with small test accuracy may not suitable for price forecasting. the discrepancy indicates that model may not be the most suitable for arrival prediction, as it performs well on the training data hence it is reasonably good for arrival. Random Forest demonstrated excellent performance for price prediction, with low MAPEs, but showed signs of overfitting for arrival prediction.

XGBoost become effective for price forecasting with low MAPE and it displayed overfitting for arrival prediction. The Hybrid model (ARIMAX+LSTM) had high MAPEs for both price and arrival predictions indicating less suitable for both arrival and price. Finally, the Multivariate LSTM model which is consistent and stable. But it had higher MAPEs compared to other models. suggesting that it's best suited for problems requiring multivariate forecasting, but may not deliver the most accurate predictions. This model is particularly suitable when capturing the relationships between multiple variables, such as temperature, rainfall, arrival, and price, is essential.

Overall, **Random Forest** and **XGBoost** proved to be the most effective for price prediction, while **ARIMA** showed somehow better for arrival prediction. The **Multivariate LSTM** model, although stable, had relatively high error and might not be the best choice for precise predictions. But it could be useful for handling multiple variables simultaneously. The **Hybrid model** (ARIMAX+LSTM) was generally less effective across both targets.

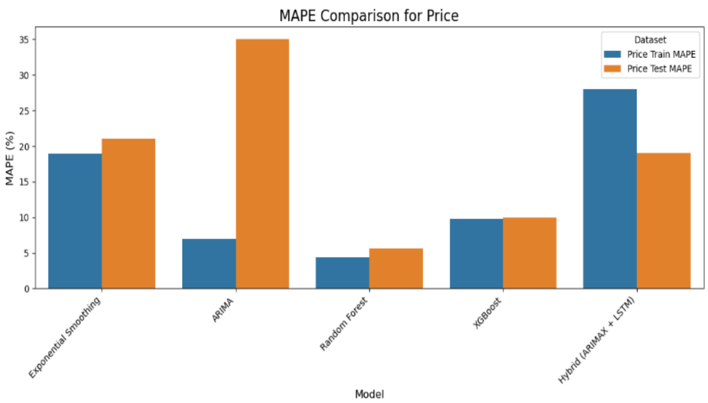


Fig 15

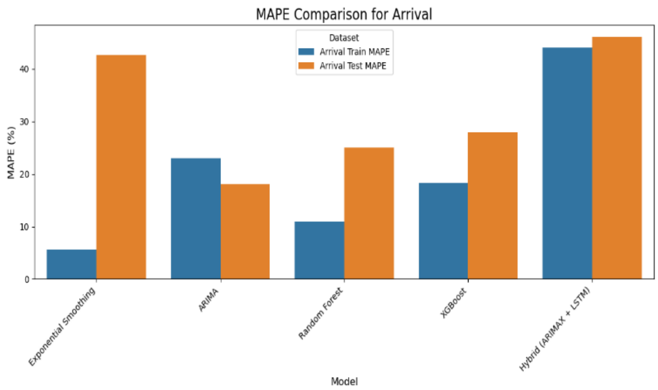


Fig 16

<u>Limitations:</u>

The limitations of the study were,

1) The study is limited to Jowar crops, making it less generalizable to other crops without significant modifications.
2) Complex seasonal patterns or irregular trends may not be fully captured by the time series models used.
3) Expanding the model to larger datasets or real-time systems may require additional computational optimization.

<u>Future Scope:</u>

This study can help market participants to plan for price and arrival. This could reduce the impact of price shocks on both farmers and consumers. It can provide ongoing updates to the agricultural community, enhancing the ability to manage risk effectively. It can help to farmers to predict future prices so they can plan properly to sell their crop. The scope and advantage of this study lie in its ability to provide actionable insights for farmers, traders, policymakers, and other stakeholders in the agricultural sector. It can help to enhance agricultural decision-making by providing accurate, data-driven forecasts for crop arrivals and prices. Its decision-support tool for farmers, advising them on the best time to plant, harvest, and sell their crops based on predicted arrival times and pricing trends.

Reference:

[1] Girish Hegde, Vishwanath R. Hulipalled, Jay B. Simha .Data driven algorithm selection to predict agriculture commodities price. International Journal of Electrical and Computer Engineering (IJECE)Vol. 13, No. 4, August 2023, pp. 4671~4682 ISSN: 2088-8708, DOI: 10.11591/ijece.v13i4.pp4671-4682

[2] Khulood Albeladi, Bassam Zafar, Ahmed Mueen. Time Series Forecasting using LSTM and ARIMA. IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 14, No. 1, 2023

[3] Samapriya Trivedi, Shalini Jaiswal, Ashok Kumar, and Shambhavi Mishra. Time series forecasting for market prices and arrival quantities of potatoes in Lucknow. Journal of Applied Statistics & Machine Learning Volume 2; Number 1; 2023 : pp. 79-96

[4] Ranjit Kumar PaulID, Md. YeasinID, Pramod Kumar, Prabhakar Kumar. Machine learning techniques for forecasting agricultural prices: A case of brinjal in Odisha, India. ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India.

[5] Sourav Kumar Purohit,Sibarama Panigrahi ,Prabira Kumar Sethy and Santi Kumari Behera. Time Series Forecasting of Price of Agricultural Products Using Hybrid Methods. Applied artificial intelligence 2021,VOL-35, https://doi.org/10.1080/08839514.2021.1981659

[6] Chu, C.W., Zhang, G.P.: "A comparative study of linear and Nonlinear models aggregate retail sales forecasting,"International Journal of production economics 86(3), 217{231(2003).

[7] H V Ramachandra, Balaraju G, Rajashekhar A,Harish Patil ,"Machine Learning application for Black Friday Sales Prediction Framework ."2021 International Conference on Emerging Smart Computing and Informatics (ESCI).

[8] Ravikant Suryawanshi, Sandeep Musale, Dr. Snehal Bhosale. Comparative Analysis of use of Machine Learning Algorithm for Prediction of Sales. J. Electrical Systems 20-3s (2024): 851-863.

[9] Tanujit Chakraborty, Ashis Kumar Chakraborty, Zubia Mansoor, hybrid regression model for water quality prediction, OPSEARCH https://doi.org/10.1007/s12597-019-00386-z.

[10] Salvatore Mancha Gonzales, Hasnain Iftikhar, and Javier Linkolk Lopez-Gonzales. Analysis and forecasting of electricity prices using an improved time series ensemble approach an application to the Peruvian electricity market. AIMS Mathematics, 9(8): 21952–21971.DOI: 10.3934/math.20241067.