CROP YIELD FORECASTING

Ankita Sarkar, B.Sc. (Hons.) Statistics, 2nd Year, St. Xavier's College (Autonomous), Kolkata

Aparajita Basu, B.Sc. (Hons.) Statistics, 2nd Year, St. Xavier's College (Autonomous), Kolkata

Anchal Bhattacharya, B.Sc. (Hons.) Statistics, 2nd Year, St. Xavier's College (Autonomous), Kolkata

Project Guide Name: Koulika Paul

Period of Internship: 14th Jan 2025 - 30th April 2025

Report submitted to: IDEAS – Institute of Data Engineering, Analytics and Science Foundation, ISI Kolkata

1. Abstract

This project involved a comprehensive time series analysis of crop production. The project was done in two parts. In the first part, we chose to analyze two key crops of West Bengal — **rice**, the principal food crop, and **jute**, an important cash crop — using time series techniques. Using time series decomposition, we extracted trend and seasonal components and applied the Augmented Dickey-Fuller test to assess stationarity. ARIMA, SARIMA and Exponential Smoothing Models were used for forecasting rice and jute production in West Bengal, achieving high forecasting accuracy. In the second part, a custom Streamlit web application was developed to enable dynamic time series analysis, allowing users to upload datasets, perform stationarity checks, visualize patterns, and generate forecasts using AutoARIMA and Exponential Smoothing. The interface made time series techniques accessible for practical agricultural analysis. Through this dual approach of detailed statistical analysis and real-time interactive visualization, the project demonstrated the power of data-driven insights for agriculture. Our findings highlight the importance of predictive analytics in supporting policy-making. Overall, the project successfully combined theoretical foundations with practical application to forecast agricultural production trends.

2. Introduction

Agriculture remains the backbone of the Indian economy, and accurate forecasting of crop production is crucial for ensuring food security, and formulating policy decisions. West Bengal, with its diverse agricultural practices, produces a wide variety of crops, among which two of the major crops are rice (a staple food crop) and jute (a major cash crop). Recognizing their significance, our project focused on a detailed time series analysis of rice and jute production data in West Bengal, spanning the period from 1997 to 2019. Raw, unit-level data on agricultural production statistics of various Indian states was collected from https://www.data.gov.in/catalog/district-wise-season-wise-crop-production-statistics-0 and cleaned to make it suitable for the project. We conducted a thorough background study covering time series concepts like stationarity, decomposition, exponential smoothing, ARIMA, and SARIMA models. Technologies employed included Python and R for data analysis and modeling, Streamlit for building an interactive web interface, and various libraries like statsmodels, pmdarima, and matplotlib. Our procedure involved pre-processing the data, performing exploratory analysis, model fitting, evaluation using error metrics, and visualization of forecasts. Additionally, we created a Streamlit application to allow users to perform real-time time series forecasting on any dataset they upload. The project was undertaken to bridge theoretical concepts with practical applications, making predictive analytics accessible and meaningful for agricultural planning.

3. Project Objective

- To analyze historical production patterns of rice and jute, two major crops of West Bengal, using time series methods.

- To forecast future crop production using ARIMA, SARIMA, and Exponential Smoothing techniques and evaluate model accuracy.
- To build an interactive Streamlit application for dynamic visualization, model fitting, and forecasting based on user-uploaded datasets.
- To illustrate the importance of stationarity and seasonal effects in time series forecasting and apply relevant tests like the Augmented Dickey-Fuller (ADF) test.
- To show that time series modeling can help in better agricultural planning, policy formulation, and resource allocation.

4. Methodology

Data Collection and Processing:

The data used for our analysis of crop statistics was collected from the website https://www.data.gov.in/catalog/district-wise-season-wise-crop-production-statistics-0. The data is attached in the Appendix. It contained district wise production data for various crops from various states of India. We chose West Bengal for our analysis. Two crops, rice and jute, were analysed. The data was pre-processed using R for the analysis. Since the data was available at the district level, we aggregated it season-wise and year-wise for West Bengal.

- For rice, we considered three main seasons — **Autumn, Summer, and Winter** — and computed a **weighted mean** of production across districts for each season of each year, using the area under cultivation as weights.

- For jute, which is a single-season crop (Kharif), the data was already annual. So, we computed the **weighted mean** of production across districts for each year, using the area under cultivation as weights.

Data Analysis Methods:

All the steps are done using R and Python and the codes are attached in the *Appendix*.

- **Data Understanding:** We started the time series analysis by plotting line diagrams and checking time series characteristics like autocorrelation, and decomposition components (trend, seasonality, and residuals).
- **Stationarity Check:**We identified stationary and non-stationary data to determine necessary pre-processing steps using the ADF test. Stationarity ensures consistent statistical properties over time.
- **Decomposition:** We then segmented time series data into its components (trend and seasonality, if present) for a clearer understanding of underlying patterns.
- **Autocorrelation and ACF/PACF:** For pre-processing the data we have used autocorrelation functions to observe relationships within the data, and partial autocorrelation to measure the direct effect of a variable on its lags. We have constructed ACF and PACF plots respectively to understand the same.

- **Exponential Smoothing:** Then we have also applied exponential smoothing to handle noise and forecast future observations by giving weighted importance to recent data points.
- **AR and MA Models:** For analysing the data we have developed auto-regressive models (AR) that use past values as predictors, and moving average models (MA) that adjust based on past errors.
- **ARIMA:** Then, we have built Auto-Regressive Integrated Moving Average Models to handle non-stationary data and make precise predictions.
- **SARIMA:** Enhanced analysis was done with Seasonal ARIMA for datasets containing repeating seasonal patterns.

Developing Streamlit Interface:

To make the analysis interactive, we developed an interactive using Streamlit that allows users to upload time series datasets (in CSV format), select relevant columns, and perform a range of analyses without requiring direct coding. The code for the interface is attached in the *Appendix*. Users can visualize the time series plots, decompose the series into trend, seasonality, and residual components, and examine autocorrelation and partial autocorrelation plots (ACF and PACF). The interface also provides stationarity tests (ADF Test) and automatically fits forecasting models using AutoARIMA and Exponential Smoothing methods (Simple, Holt, Holt-Winters). Forecasting results are presented both graphically and numerically, with model accuracy metrics like MAE and MAPE reported. While the development of this interface marks a key milestone in our internship, it is still at a very preliminary stage and can be refined even further to enhance accuracy of results.

5. Data Analysis and Results

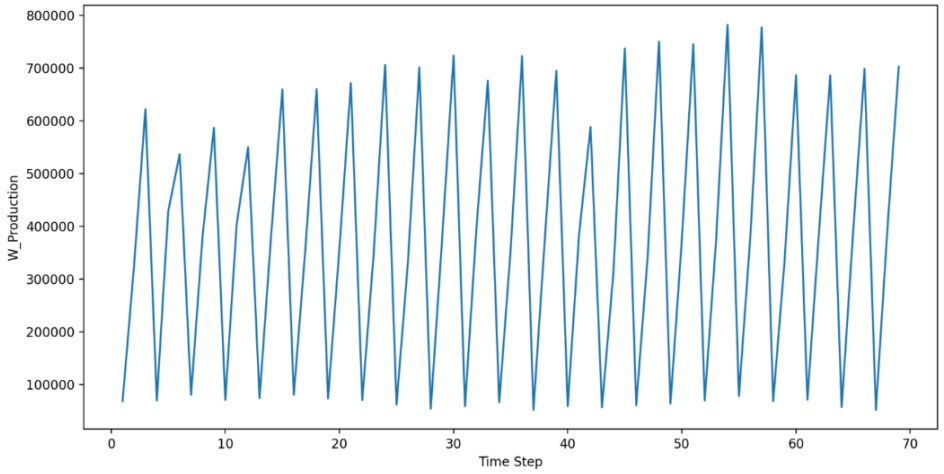
5.1. Analysis of Crop Yield

First, analysis was done independently using Python and R for the two crops selected by us - Rice and Jute.

Rice: Rice is a staple cereal grain cultivated globally. Sowing times vary by region and variety, often coinciding with the rainy season (e.g., June-July). Harvesting typically occurs 3-6 months after sowing, depending on the variety and growing conditions (e.g., October-November). The data is available from 1997 to 2019 for **autumn, summer and winter** seasons. The line diagram is given below.

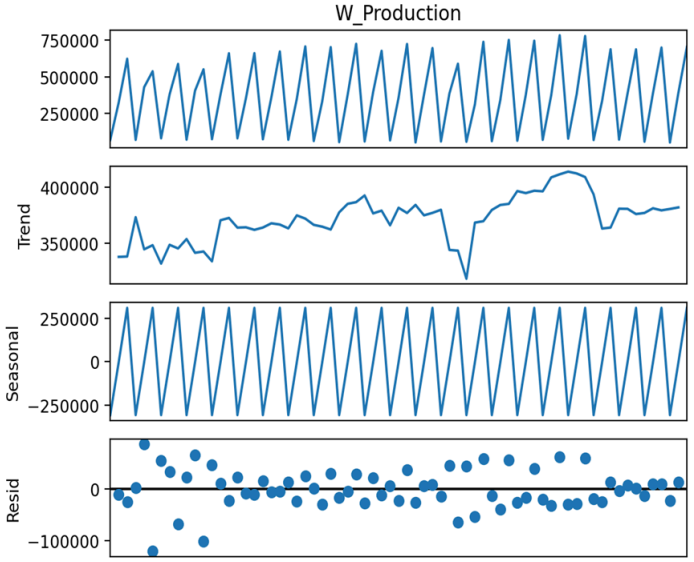
Since each year has 3 seasons, time steps have been considered instead of years. For example, Autumn 1997 has been taken as Round 1, Summer 1997 as Round 2, Winter 1997 as Round 3, Autumn 1998 as Round 4, Summer 1998 as Round 5 and so on, following this pattern.

Line Diagram showing Rice Production over the Years



Rice production exhibits substantial year-to-year variability. A recurring pattern of high production followed by lows suggests potential seasonal or economic influences. After the **initial dip** in the **late 1990s**, production generally shows a trend of recovery with **higher peaks** in the later years. The year **2016** recorded the **highest** rice production within the observed period. Despite a dip after 2016, the production in 2019 remains at a relatively high level.

Now we get the decomposition plot.



Observed Line Chart (Top panel): This panel displays the original rice yield data over time, showing periodic fluctuations. The presence of cycles suggests recurring patterns influenced by seasonal agricultural factors. The data exhibits strong periodic behavior, confirming the influence of recurring seasonal trends in rice production.

Trend (2nd Panel): This panel highlights the underlying long-term movement in rice yield, revealing whether production is increasing, decreasing, or remaining stable. It smooths out short-term fluctuations to show structural changes in agricultural output. Here the trend line provides insight into overarching shifts in production, helping to identify long-term sustainability and efficiency in rice farming.

Seasonality (3rd Panel): The repeating patterns in this panel indicate a consistent annual cycle in rice yield. Seasonality is a significant factor in rice yield variation, meaning future production can be predicted based on these cyclical patterns.

Residual (4th Panel): This panel captures variations not explained by trend or seasonality, representing irregular fluctuations due to external factors. The residuals appear moderate, suggesting that the decomposition effectively captured most of the structured variation in rice yield.

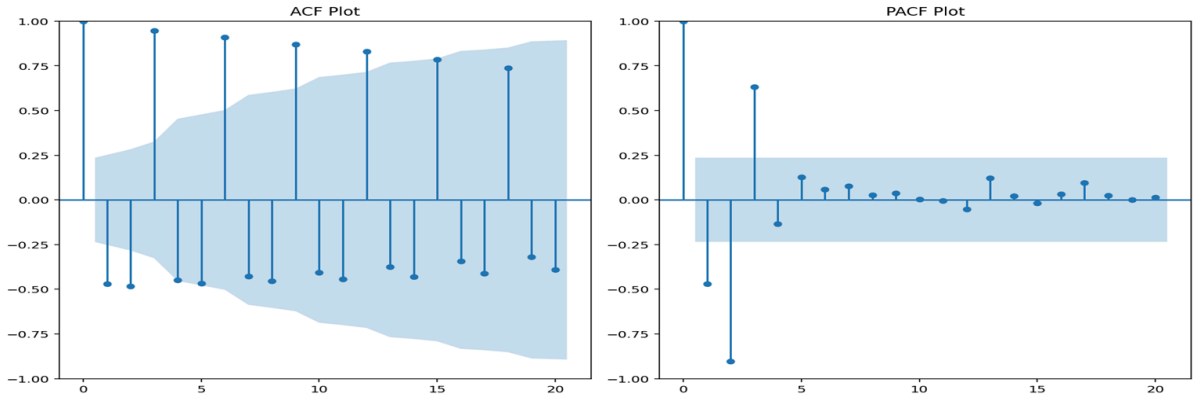
We then apply the **Augmented Dickey-Fuller (ADF) Test** to check the stationarity of the series. The hypotheses are such that,

H_0: the time series is non-stationary
H_1: the time series is stationary

```
ADF Statistic: -3.012059161712291
p-value: 0.033783470435699105
```

We get the p-value as **0.0338 < α=0.05**. So, we **reject** the null hypothesis and say that the data is stationary. So, no differencing is involved. Therefore we should consider **d=0** for ARIMA modeling.

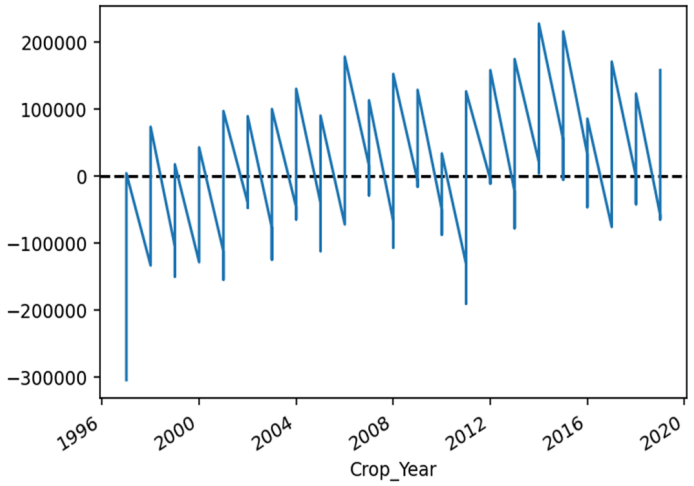
Now, we consider the Autocorrelation and Partial Autocorrelation Plots.



- The ACF plot shows significant spikes at lags 1, 2, 3, and 4 (outside the confidence interval).Based on this, (q = 4) is likely a good choice, indicating that the moving average model considers the previous 4 lags.
- The PACF plot has a significant spike at lag 1 and becomes less prominent afterward (inside the confidence interval for higher lags). This sharp cutoff at lag 1 suggests (p = 1) for the AR component (Autoregressive model), indicating that the current value depends primarily on the immediate prior value.

Therefore we should consider **p=1** and **q=4** for ARIMA modeling.

So, now we plot the residuals against time after fitting the ARIMA(1,0,4) model.

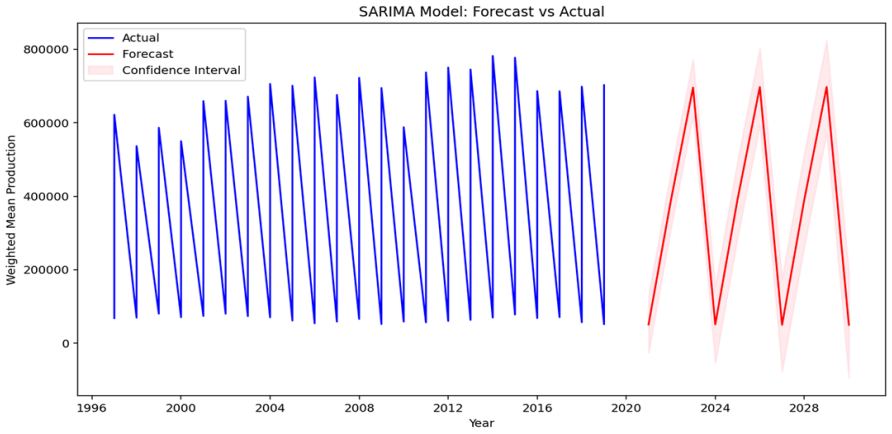


- **Centered around zero:** Residuals generally fluctuate above and below zero, indicating no consistent bias.
- **No obvious patterns:** Lack of clear trends or seasonality in the errors suggests the model captured these components.
- **Relatively consistent variance:** The spread of residuals appears reasonably stable over time.
- **Presence of some outliers:** A few larger deviations indicate some unpredictable events.
- **Overall good fit (tentative):** The model seems to have captured much of the predictable variation.

As there exists a clear seasonality in the data we have calculated the seasonal differencing once with lag=3 (as there are 3 seasons). Now we test if the seasonal patterns are removed from the data or not by seasonal differencing once by **ADF test.**

```
ADF Statistic: -4.025645815550079
p-value: 0.0012820182439288838
```

We get the p-value as 0.00128 < α=0.05. So, we can say that the seasonal patterns are removed from the data.As there are 3 seasons, we should consider **s=3** and we should consider **D=1** for SARIMA modelling.The P and Q remain the same as the p and q of the ARIMA model. We then fit a SARIMA (1,1,4,3) model with the forecasting of the next 10 years (2020-2030).



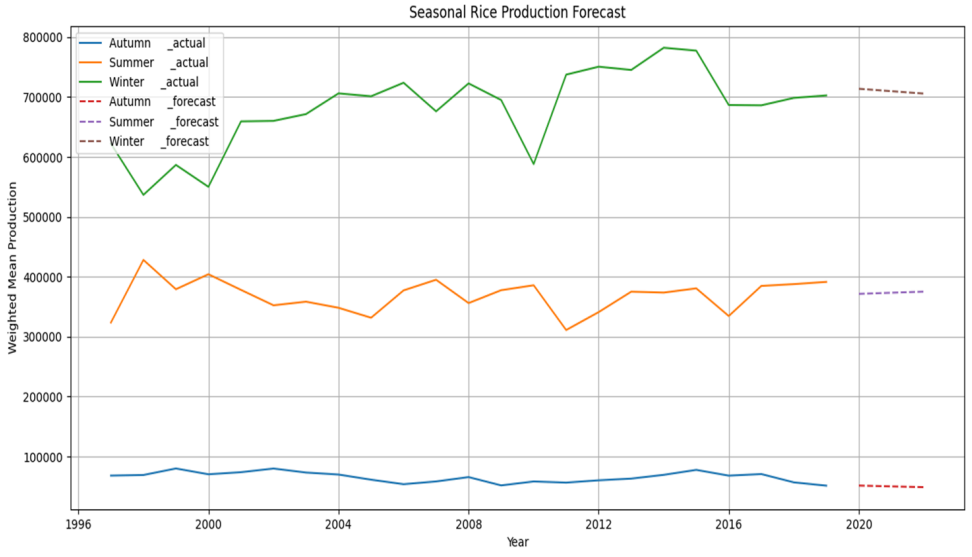
Model predicts continued up-and-down patterns in rice production. The forecast suggests larger production highs and lows in the future. Confidence interval widens over time, indicating less prediction certainty further out. No strong upward or downward long-term trend is evident in the forecast's average level. The model effectively projects the recurring seasonal variations.

Error Metrics:
MAE: 31019.006017904547
RMSE: 33896.95763089267
MAPE: 8.864465271876673

Observations:

1. On average, the forecast deviates by 31,019 units from the actual values. This indicates a relatively small absolute error magnitude, suggesting that the model performs well in terms of precision.
2. The RMSE value of 33,897 reflects the standard deviation of residuals (prediction errors). Since RMSE considers squared errors, it gives more weight to larger errors than MAE. The close values between MAE and RMSE indicate that the forecast errors are fairly consistent and not significantly influenced by extreme deviations.
3. The forecast has a MAPE of 8.86%, which means, on average, the model's predictions deviate by 8.86% from the actual values in percentage terms, so, it can be considered very accurate in many forecasting contexts, highlighting that the SARIMA model is well-fitted and performs excellently in capturing the patterns of the data.

Overall Interpretation:
The error metrics suggest the model is performing robustly, with minimal absolute and percentage errors. The low MAPE, in particular, is a strong indicator of reliable forecasting accuracy.

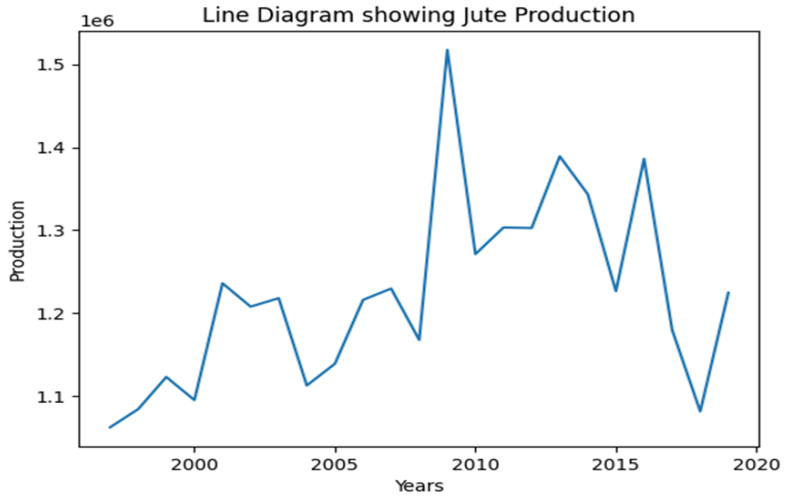


Error Metrics:

	Autumn	Summer	Winter
MAE	5031.03	20913.17	40009.52
RMSE	6550.94	27041.92	45374.70
MAPE	8.10%	5.59%	5.93%

Observations: All seasons show slight continuation of recent trends — no sharp shifts or anomalies. Forecasts are smooth, which is typical for exponential smoothing unless trend or seasonality is explicitly modeled.
Winter (Forecast Line is Brown dashed): Slight decline, from around 715,000 to 705,000. From Error Metrics, MAPE: 5.93% indicates good performance even with the highest absolute errors. Reflects confidence in proportional accuracy.
Autumn (Forecast Line is Red dashed): Nearly flat, forecasting production to stay around 60,000–65,000.MAPE (8.10%) indicates slightly less accuracy, suggesting that autumn data may be more volatile or harder to capture with smoothing.
Summer (Forecast Line is Purple dashed): Mild upward trend, stabilizing near 380,000. MAPE: 5.59% indicates strong accuracy despite higher raw errors. Production values are large, so small percentage deviations are tolerable.

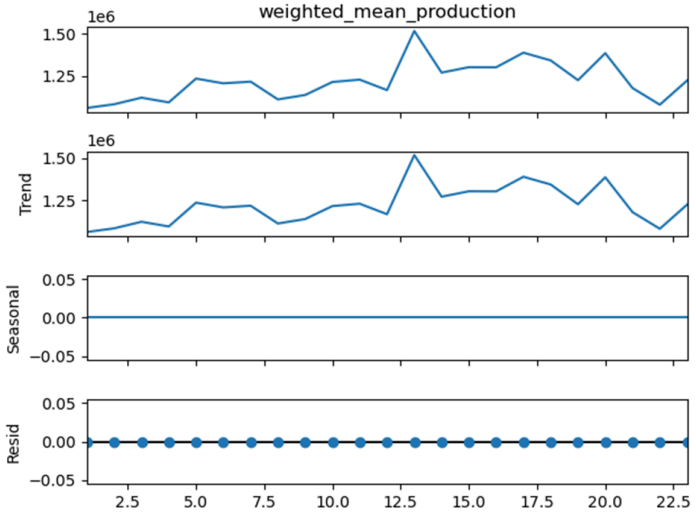
Jute: Jute is a Kharif Crop, grown only once a year, with sowing occurring between March and May and harvesting taking place from June to September. So, the yearly data is available from 1997 to 2019 from the Kharif Season. We plot that in a line diagram.



- There's a **general upward trend** in jute production initially, although it's not strictly linear. Production increased until a **sharp peak in 2009**, which seems to be the highest production value on the chart.

- There are significant **fluctuations** throughout the graph, indicating variability in production. Some sharp drops and rebounds suggest periods of external shocks, like weather, policy changes, etc.
- The **highest production** reaches above 1.5 metric tonnes, and then there's a noticeable decline **followed by volatility** in the subsequent data points.
- Towards the end, after a sharp drop, there's a small **rebound**, which might indicate a recovery or seasonal effect.

We first get the **decomposition plot.**



Observed Line Chart (Top Panel): It shows the original jute production values over time. There is moderate fluctuation with an upward trend until around the 13th year, i.e., 2009, followed by a few drops and recoveries. This aligns with what we saw in the initial line plot.

Trend (2nd Panel): The trend line captures the general increase up to the peak, then a gradual decline. This confirms that production has long-term movement, which needs to be handled in modelling (e.g., through differencing).

Seasonality (3rd Panel): The seasonal component is flat (zero), as expected. This confirms that the data does not have seasonality within the year since jute is an annual Kharif-season crop and the data is available yearly.

Residual (4th Panel): The residuals are very small and relatively constant. This indicates that most of the variance in the data is explained by the trend component, not noise or seasonality.

We then apply the **Augmented Dickey-Fuller (ADF) Test** to check the stationarity of the series. The hypotheses are such that,

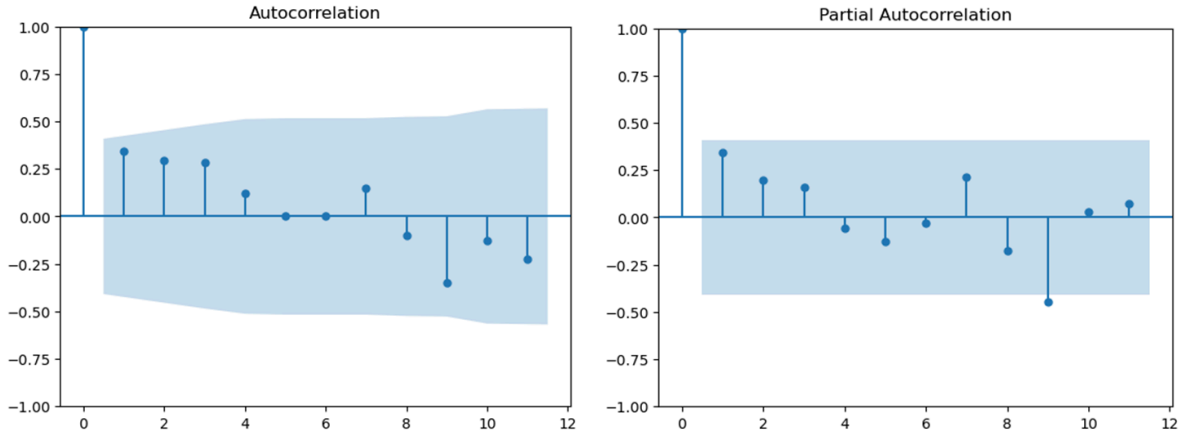
H_0: the time series is non-stationary
H_1: the time series is stationary

```
> adf.test(jute$weighted_mean_production)

        Augmented Dickey-Fuller Test

data:  jute$weighted_mean_production
Dickey-Fuller = -0.75158, Lag order = 2, p-value = 0.954
alternative hypothesis: stationary
```

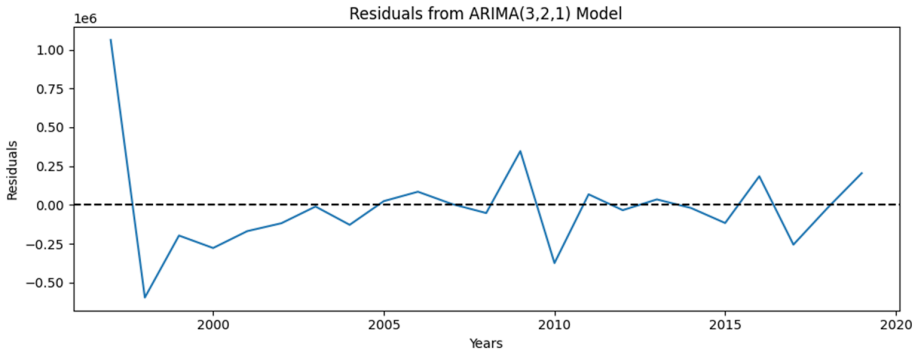
We get the **p-value as 0.954 > α=0.05**. So, we **accept the null hypothesis** and say that the data is **non-stationary**. As seen before, there is a trend in the data which makes it non-stationary. So, the trend needs to be handled by differencing. The following table sums the ADF test results after differencing.

Series	**ADF Statistic**	**p-value**	**Stationarity**
Original Series	-0.75158	0.954	Non-stationary
1st differencing	-3.5619	0.0553	Borderline non-stationary
2nd differencing	-5.4338	<0.01	Stationary

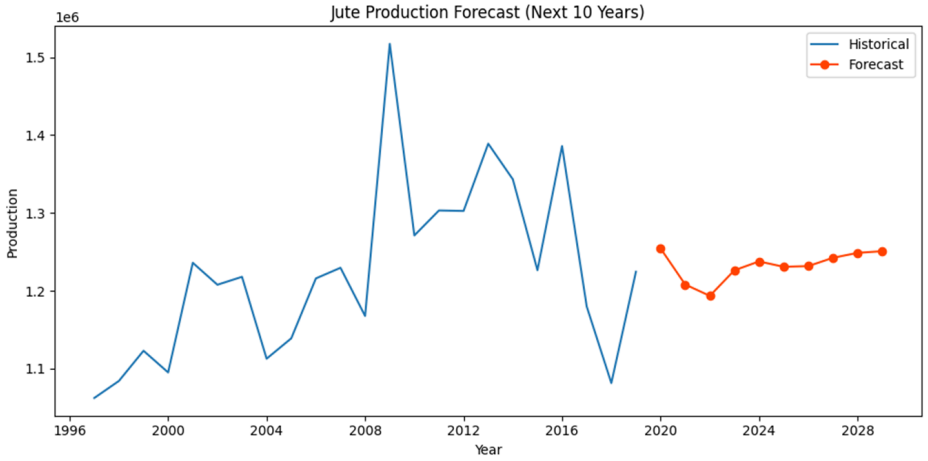
We should consider **d=2** for ARIMA modelling. Now, we consider the Autocorrelation and Partial Autocorrelation Plots.



The ACF plot shows a **small spike at lag 1**, though it is still within the confidence interval, and then a slow tapering off. So, we take the **Moving Average process to be of order 1** for our model. So, **q = 1.** In the PACF plot, we see mildly significant spikes up to **lag 3**, though the still values fall within the confidence interval. This suggests a cut-off at **lag 3**, meaning the data likely has an **AR(3)** component. So, **p = 3.** We consider p=3 & q=1 for ARIMA modelling and fit an **ARIMA(3,2,1)** model.
We plot the residuals against time after fitting the ARIMA(3,2,1) model.



We can see **no clear trend in residuals.** This is good as the residuals don't seem to follow systematic upward or downward trends. The values **alternate around 0,** indicating that the model is neither overpredicted nor underpredicted. A sharp spike, in the beginning, can suggest some mild errors committed in the predicted values for the initial years.



The forecast picks up **from 2019** and projects **a generally stable trend** from 2020 to 2029. After a slight **dip around 2021–2022**, production **stabilizes** and **gradually increases**. **No major spikes or crashes** are predicted — the model suggests that jute production is likely to remain relatively **consistent**. The drop immediately after 2019 may be due to sudden changes in recent production values in the original dataset or due to some errors in forecasting. From **2023 onward**, forecast values show a **modest upward trend**, suggesting consistent productivity. There will be **no major growth or decline** — production is **likely to stay within a narrow range (1.2 – 1.25 metric tonnes)**. There can be inaccuracies in prediction as the model assumes **no external shocks or drastic policy/environmental changes**.

Error Metrics:

MAE: 275250.6795822271
RMSE: 306467.149294462
MAPE:23.49357824738806

Interpretation:

For the previous forecasting graph:

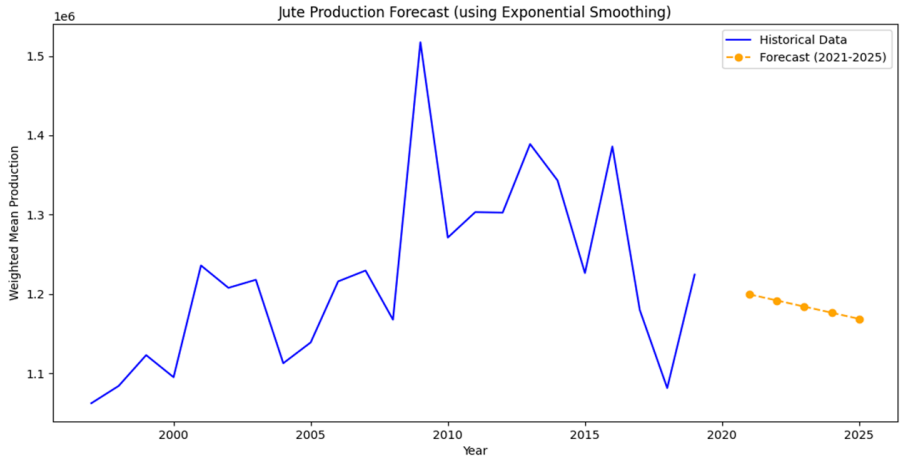
MAE (Mean Absolute Error): The average difference between the actual values and the predicted values is approximately 275,250.68.

RMSE (Root Mean Squared Error): The standard deviation of the residuals (prediction errors) is roughly 306,467.15.

MAPE (Mean Absolute Percentage Error): The average percentage error between the forecasted values and the actual values is **23.49%**, suggesting that the **model has moderate forecasting accuracy.**

Overview:

On average, the model's forecasts deviate by about 23.49% from the true values.



Error Metrics:

MAE: 3.2

RMSE: 3.521363372331802

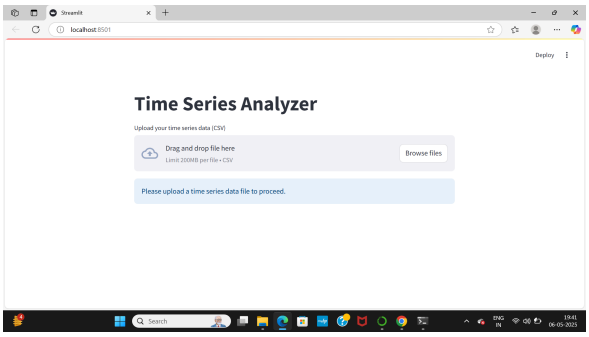
MAPE: 2.7385814185814183

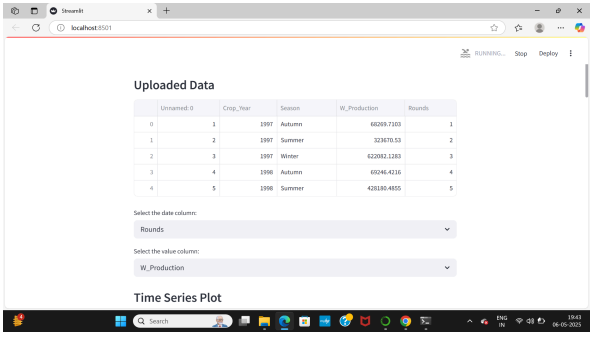
Findings from the graph: The forecast line indicates a possibility of observing a slightly decreasing pattern in jute production in the next 5 years (2021-2025). Also, from the previous forecasting model, we have observed that there is a slightly decreasing pattern just after 2020. So, the results are more or less similar in both cases, indicating a reliable forecast.

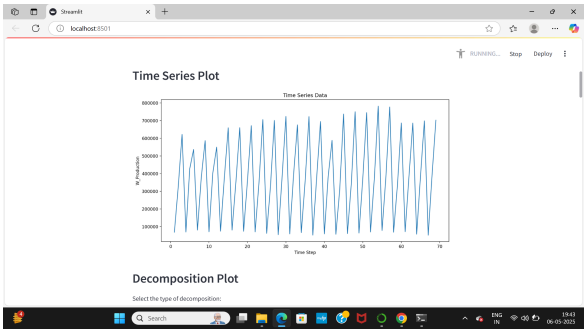
As the MAE value is very low and the small MAPE value reflects high accuracy in forecasting, overall it shows **strong forecasting performance** with very **minimal deviation from the actual values**.

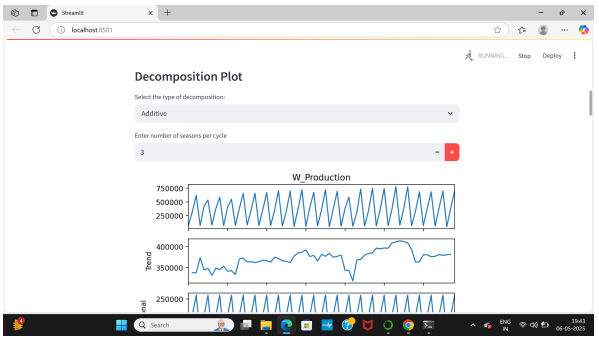
5.2. The Streamlit Interface

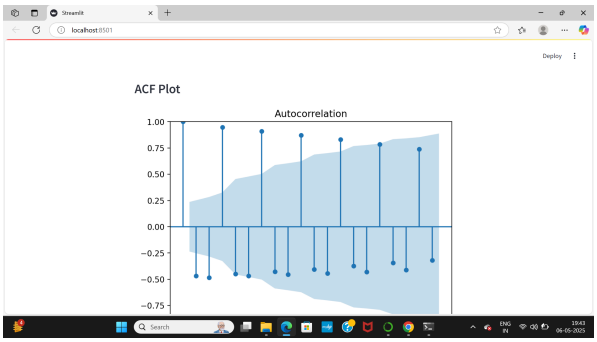
After this analysis, a **Streamlit Interface** was developed integrating different Python codes used for the Data Analysis to create an interactive interface where users can upload any Time Series Data and get plots and analysis corresponding to the data uploaded. The interface is still at a very preliminary stage and can be refined even further to enhance accuracy of results. Screenshots of the interface are attached below.

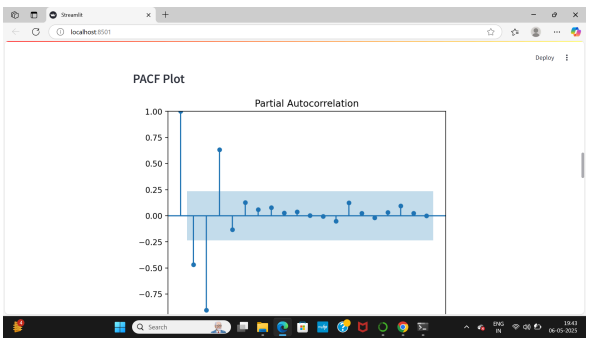

Time Series Analyzer
Upload your time series data (CSV)
Drag and drop file here
Limit 200MB per file • CSV
Browse files
Please upload a time series data file to proceed.

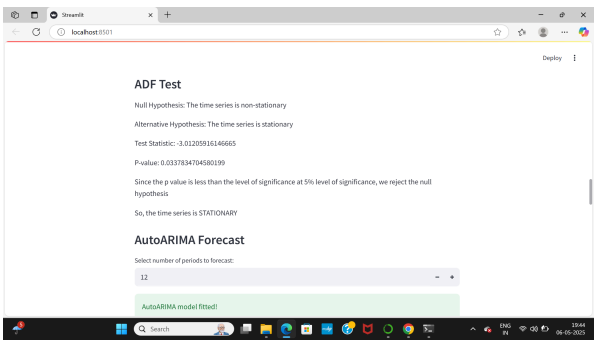

Uploaded Data
Select the date column:
Rounds
Select the value column:
W_Production
Time Series Plot

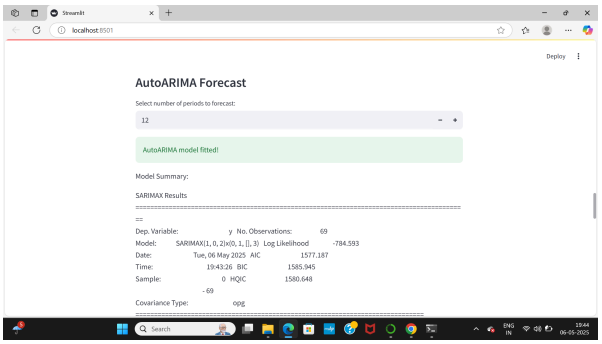

Time Series Plot
Time Series Data
Decomposition Plot
Select the type of decomposition:

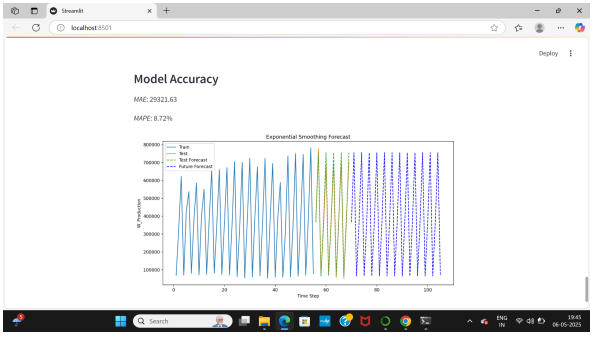

Decomposition Plot
Select the type of decomposition:
Additive
Enter number of seasons per cycle
3
W_Production
Trend

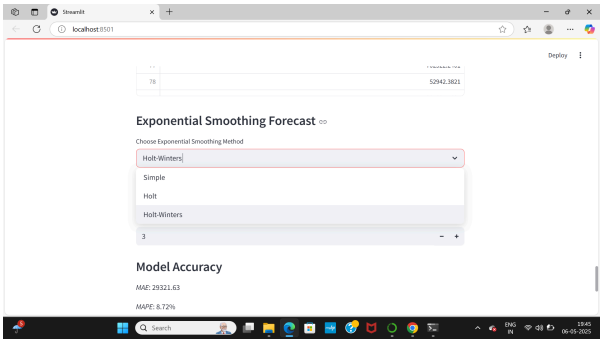

ACF Plot
Autocorrelation

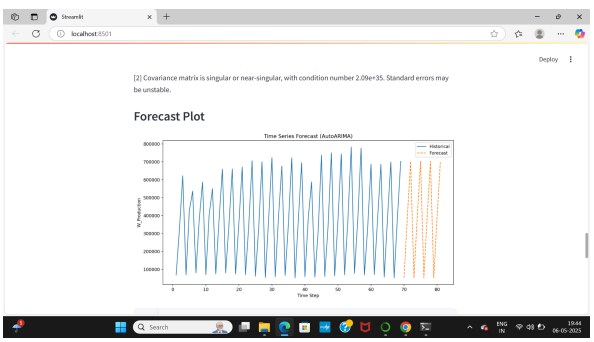

PACF Plot
Partial Autocorrelation

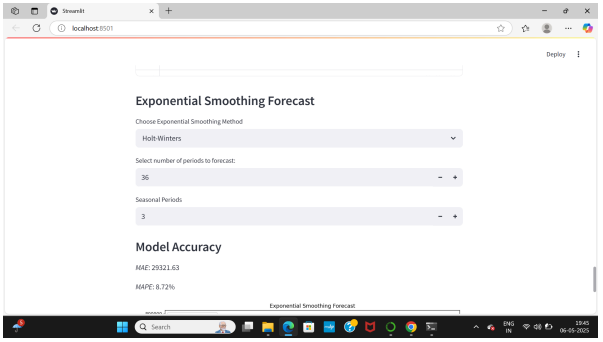

ADF Test
Null Hypothesis: The time series is non-stationary
Alternative Hypothesis: The time series is stationary
Test Statistic: -3.012059161466 65
P-value: 0.033783470458 0199
Since the p value is less than the level of significance at 5% level of significance, we reject the null hypothesis
So, the time series is STATIONARY
AutoARIMA Forecast
Select number of periods to forecast:
12
AutoARIMA model fitted!


AutoARIMA Forecast
Select number of periods to forecast:
12
AutoARIMA model fitted!
Model Summary:
SARIMAX Results
Dep. Variable: y No. Observations: 69
Model: SARIMAX(1, 0, 2)x(0, 1, [], 3) Log Likelihood -784.593
Date: Tue, 06 May 2025 AIC 1577.187
Time: 19:43:26 BIC 1585.945
Sample: 0 HQIC 1580.648
- 69
Covariance Type: opg


Model Accuracy
MAE: 29321.63
MAPE: 8.72%
Exponential Smoothing Forecast
Train
Test
Test Forecast
Future Forecast


Exponential Smoothing Forecast
Choose Exponential Smoothing Method
Holt-Winters
Simple
Holt
Holt-Winters
3
Model Accuracy
MAE: 29321.63
MAPE: 8.72%





6. Conclusion

This project provided a comprehensive exploration of crop yield forecasting using time series analysis, specifically focusing on rice and jute production in West Bengal. Through rigorous statistical modeling and real-time visualization, we demonstrated how data-driven techniques can enhance the understanding and forecasting of agricultural trends. Though we focussed only on two crops since the initial part of our analysis, through the Streamlit Interface developed in the later part, users can do the analysis for multiple other crops as well. Our application of time series decomposition, stationarity tests (ADF), and forecasting models such as ARIMA, SARIMA, and Exponential Smoothing revealed meaningful patterns in crop production data. For jute, we found the data to be non-stationary, requiring second-order differencing. The ARIMA(3,2,1) model was found to provide reliable forecasts, with error metrics (MAPE = 23.49%) indicating moderate accuracy. However, the exponential smoothing model yielded even better performance with a significantly lower MAPE of 2.74%, highlighting the importance of model selection and evaluation. The rice data, segmented by season, also displayed distinct trends and seasonal variations, which were effectively captured using seasonal models. Forecasts indicated relatively stable future production, with minor fluctuations, underscoring the potential of predictive analytics in agricultural planning.

A key achievement of the project was the development of a Streamlit web application, enabling users to perform time series analysis interactively. As mentioned earlier, the Interface is still at a very nascent stage and can be made more accurate and perfect for analysis through more advanced coding and time. If advanced work is done on the interface, it may serve as a tool to bridge the gap between complex statistical modeling and user-friendly interfaces, making complex agricultural analysis a lot simpler.

In the **future**, new tools and metrics can be developed and used that take into consideration external factors such as weather patterns, irrigation coverage, and market prices which may enhance model accuracy and predictive power. While the current Streamlit application demonstrates core functionalities, it was developed within the practical limits of our current programming knowledge and the time available during the internship. With further refinement using advanced coding skills and more knowledge about the agricultural scenario in the country, added features, such as real-time data integration, model comparison tools, and enhanced user experience, can give the platform the potential to bridge the gap between advanced data analytics and the agricultural sector’s real-world needs.

7. APPENDICES

You may create separate Appendix for the following:

1. **References:** Forecasting: Principles and Practice (2nd ed)
2. **Code extract:** There were several codes and hence they have all been uploaded to https://github.com/ankitasarkar-07/analysis-of-crop-statistics
3. **Document Link:** https://github.com/ankitasarkar-07/analysis-of-crop-statistics