# DATA-DRIVEN FORECASTING OF RICE PRODUCTION IN SRI LANKA USING SARIMAX AND MACHINE LEARNING MODELS

A project report presented by:

**Ashfaq M - (KIC-HNDDS-241-F-007)**

to the

**KANDY INNOVATION CENTER**

**NATIONAL INSTITUTE OF BUSINESS MANAGEMENT**

*In partial fulfillment of the requirement for the award of the diploma of*

**HIGHER NATIONAL DIPLOMA IN DATA SCIENCE**

**NATIONAL INSTITUTE OF BUSINESS MANAGEMENT**

**SRI LANKA**

**2025**

# ABSTRACT

## Rice Production Prediction, Sri Lanka

**Ashfaq M**

NIBM-KIC

ashfaqmohamed422@gmail.com

In Sri Lanka, accurate rice production forecasting is essential for both efficient agricultural planning and food security. This study applies statistical and machine learning models to predict yields based on seasonal trends, climatic, and environmental factors. The dataset spans from 1950 to 2024 including rice cultivating seasons in Sri Lanka namely, Yala and Maha, where all these variables that impact rice production were gathered from the Department of Census and Statistics, Department of Meteorology, and the Central Bank of Sri Lanka. To make a correlation with the variables, Spearman correlation method was used as it handles non-linear relationships very well. Multiple SARIMAX models were performed and a SARIMAX model with the configuration (1,1,1) (1,1,0,2) was developed, incorporating rainfall and harvested acres as exogenous variables to forecast production. This model performed the best among the other models and successfully captured the seasonality of rice production, with a Mean Squared Error (MSE) of 13,143.60, Root Mean Squared Error (RMSE) of 114,65, and a Mean Absolute Percentage Error (MAPE) of 6.67%. In comparison, machine learning models such as Random Forest and XGBoost were trained and tested by considering the input variables as rainfall, sown acres, and harvested acres. The Random Forest model performed well with an $R^2$ score of 0.9134 and a MAPE of 19.25%. However, its error rates were greater than those of the best performed SARIMAX model. A hybrid model that combined Random Forest and SARIMAX was also tested. However, it did not increase accuracy, displaying high error values. Overall, the SARIMAX (1,1,1) (1,1,0,2) model proved to be the most accurate and reliable method for forecasting rice production in Sri Lanka including some external factors that impact the rice production.

**Keywords**: Rice Production, Time Series, Forecasting, SARIMAX model, Machine Learning)

# TABLE OF CONTENTS

Abstract ii

Table of Contents .......... iii

List of Tables .......... iv

List of Figures .......... v

Chapter 01: INTRODUCTION .......... 1

1.1. Background of the stUdy .......... 1

1.2. Justification of the problem .......... 2

1.3. Objectives of the study .......... 3

Chapter 02: LITERATURE REVIEW .......... 4

Chapter 03: METHODOLOGY .......... 10

3.1. Data Collection .......... 10

3.2. Data Preprocessing .......... 11

3.3. Exploratory data analysis .......... 12

3.3.1. Time Series .......... 12

3.3.2. Stationarity check .......... 13

3.3.3. time series Model equations .......... 14

3.3.4. Evaluation metrics equations .......... 16

3.3.5. Machine learning .......... 17

Chapter 04: RESULTS AND DISCUSSION .......... 20

4.1. Exploratory Data Analysis .......... 20

4.1.1. Distribution of Variables .......... 20

4.1.2. Spearman Correlation heatmap .......... 21

4.2. Time Series .......... 22

4.2.1. Time series plot (line plot) for total production .......... 23

4.2.2. Line plot of production for yala vs maha .......... 23

4.2.3. rolling average plot (yala vs Maha) .......... 24

4.2.4. Correlated line plot .......... 24

4.2.5. ACF & pacf plots sfter log transformation .......... 25

4.2.6. first differenced time series plot .......... 26

4.2.7. acf & pacf plots after the first difference .......... 27

4.2.8. Model comparisons .......... 28

4.2.9. predictive accuracy measures .......... 34

4.2.10. Actual vs fitted plot for the best time series model ........ 34
4.2.11. actual, fitted & forecasted plot ........ 36
4.2.12. Forecasted plot ........ 36
4.3. Machine learning ........ 38
4.4. limitations ........ 41
4.4.1. time series ........ 41
4.4.2. hybrid model ........ **Error! Bookmark not defined.**
4.5. Recommendations ........ 45
Chapter 05: CONCLUSION ........ 46
References 47
Appendix 48

## LIST OF TABLES

Table 3-1: Dataset Overview ........ 10
Table 3-2: Summary of the Dataset ........ 12
Table 4-1: Results of SARIMA (1,1,1) (1,1,0,2) ........ 28
Table 4-2: SARIMAX (0,1,1) (0,1,1,2) ........ 30
Table 4-3: Results of SARIMAX (1,1,1) (1,1,0,2) ........ 31
Table 4-4: Actual vs Predicted values ........ 36
Table 4-5: Forecasted values of rice production ........ 37
Table 4-6: Performance metrics for Machine Learning models ........ 39
Table 4-7: SARIMAX (1,1,0) (1,1,0,2) results ........ 43
Table 4-8: SARIMAX (1,1,1) (1,1,0,2), exog variables = rainfall and temperature .. 44
Table 4-9: SARIMAX (0,1,1) (0,1,1,2). Exog variable = temperature ........ 44

# LIST OF FIGURES

Figure 4-1: Distribution of Variables ........................................................ 20

Figure 4-2: Spearman Correlation Heatmap ............................................ 21

Figure 4-3: Line plot for the total production (Yala+Maha) ..................................... 23

Figure 4-4: Line plot of production for Yala and Maha seasons ............................... 23

Figure 4-5: Line plot of production for mostly correlated climatic and environmental variables with production ............................................................................ 24

Figure 4-6: ACF & PACF plots for the log-transformed production ........................ 25

Figure 4-7: First differenced plot of production ....................................................... 26

Figure 4-8: ACF & PACF plots after the first difference .......................................... 27

Figure 4-9: Residual Diagnostic Panel...................................................................... 29

Figure 4-10: Residuals Diagnostic Panel .................................................................. 30

Figure 4-11: Residual Diagnostic Panel.................................................................... 32

Figure 4-12: Residual Diagnostic Panel.................................................................... 33

Figure 4-13: Actual vs Predicted plot........................................................................ 35

Figure 4-14: Box plot of the variables ...................................................................... 38

Figure 4-15: Actual vs Predicted Production for Random Forest Model.................. 39

Figure 4-16: Actual vs Predicted plot for XGBoost Model ...................................... 40

Figure 4-17: Line Plot of the variables ..................................................................... 41

Figure 4-18: Normalized line plot of the variables ................................................... 42

Figure 4-19: Spearman Correlation Heatmap ........................................................... 42

# Chapter 01:

# INTRODUCTION

## 1.1. BACKGROUND OF THE STUDY

An essential pillar of Sri Lanka’s economy, rice cultivation supports the country’s agricultural industry and guarantees food security. Although rice is grown widely as a staple food, a few factors, such as market dynamics, seasonal fluctuations, climatic conditions, and economic policies, affect its yield. Yala and Maha are the two seasons when rice is being cultivated and produced in Sri Lanka. During Maha season the rainfall is high and the production of rice is also somewhat greater in quantity than in Yala season. The complex relationship between these factors makes it difficult to predict production with any degree of accuracy. These relationships are frequently missed by conventional forecasting techniques, producing forecasts that are not trustworthy.

Recent studies have revealed the successful implementation of time series models and machine learning approaches for rice production forecasting. For example, the ARIMA (Auto-Regressive Integrated Moving Average) model was used in the Munasinghe and Peiris (2023) study to historical data from 1951 to 2019 and found ARIMA (5,1,0) suitable for Yala and Maha seasonal data, while 3MA (3-Moving Average) model showed better accuracy with a MAPE value of 9.13%. ARIMA was used in account for long-term trends and seasonal fluctuations in rice yields. Short-term forecasting has revealed the effectiveness of ARIMA, especially when trained on segmented seasonal data from Yala and Maha. This illustrates how well it may be used to highlight the cyclic pattern of Sri Lanka’s rice growing cycles.

Moreover, another study by Sivapathasundaram and Bogahawatte (2013), developed and evaluated ARIMA models to forecast paddy cultivation in Sri Lanka. They have used ARIMA (2,1,0), ARIMA (2,1,1), and ARIMA (2,1,2) to compare the AIC and BIC scores with every model. Among these models, ARIMA (2,1,0) performed well for short-term forecasting of annual rice production in Sri Lanka with a MAPE value of 10.5% which indicates a reasonably good forecasting accuracy.

This study mainly focuses on time-series forecasting, and statistical analysis to improve predictive accuracy. While climatic conditions like temperature, rainfall, and severe weather are important, economic issues like trade policy, inflation, and subsidies influence productivity. By identifying machine learning, hidden patterns in large datasets can be leveraged, time-series models can capture trends and seasonal impacts, and statistical correlations can clarify key influences on rice production. Through strong evaluation, the study aims to develop a solid, data-driven predictive model for effective agricultural planning and policy making, guaranteeing accurate forecasts.

## 1.2. JUSTIFICATION OF THE PROBLEM

Accurately forecasting rice production in Sri Lanka requires a multidimensional approach due to the complex interaction of climatic, economic, and seasonal variables. Traditional models frequently fall short by focusing only on one kind of data or relationship. Time series forecasting models, however, offer strong methods for identifying historical trends, seasonal, and cyclic patterns in rice production. Notably, most studies have not explored the statistical correlation between key climatic factors such as rainfall and temperature along with rice production. This gap shows the need for more thorough forecasting techniques that combine climatic effect analysis and time series models to predict accuracy and promote data-driven agricultural planning.

Economic and seasonal considerations have a significant effect on Sri Lanka's rice production. The two major cultivation seasons, Yala and Maha are heavily dependent on monsoonal rainfall, and variations in rainfall or temperature during these seasons can have a direct impact on yield productions. Furthermore, economic factors also play a crucial role in influencing farmer behavior and production levels. Incomplete modelling is the result of ignoring any kind of influence. A thorough understanding of how the interactions between these seasonal and economic factors are necessary to represent the real-world dynamics impacting rice production.

Correlation analysis is essential for enhancing the prediction accuracy of rice production by balancing the advantages of various modelling techniques. Time series models like SARIMA are effective at identifying seasonal patterns and long-term trends in data, making them ideal for forecasting applications where seasonal and historical behaviors are important, like rice production. Researchers can improve model inputs and forecast accuracy by analyzing statistical correlations between climatic factors and yield data. Measures like MSE, RMSE, and MAPE help assess model accuracy, guaranteeing predictions are not only accurate but also useful for real world decision making in agriculture and food security planning as well.

## 1.3. OBJECTIVES OF THE STUDY

- This study aims to develop a comprehensive rice production forecasting model for Sri Lanka by integrating seasonal patterns, climatic conditions, and economic data. It focuses on the Yala and Maha seasons, identifying key variables such as rainfall and temperature that significantly influence yield. By analyzing these correlations, the study will apply advanced time series models and machine learning techniques that incorporate external climatic factors. The objective is to improve forecasting accuracy and provide more relevant, data-driven insights into agricultural planning.

# Chapter 02:

# LITERATURE REVIEW

A study (Munasinghe, 2023) conducted using historical rice production data from 1951 to 2019 evaluated the performance of ARIMA and 3-Moving Average (3MA) models for forecasting rice yields in Sri Lanka. While the ARIMA (5,1,0) model was selected based on statistical criteria such as the Akaike Information Criterion (AIC), the 3MA model eventually showed better predicting accuracy, achieving a Mean Absolute Percentage Error (MAPE) of 9.13%. This finding underscores the importance of considering both complex and simpler time series models in agricultural forecasting. It highlights that, in certain contexts, simpler models like 3MA may outperform more experienced methods depending on the underlying data patterns. The study also acknowledged the influence of climatic and socioeconomic variables on production trends, suggesting the need for integrating such external factors in future forecasting efforts. This supports the rationale for evaluating multiple time series techniques, as well as incorporating exogenous variables, as adopted in the present study.

(Bogahawatte, 2013) conducted a thorough time series analysis to forecast paddy production in Sri Lanka, utilizing data from 1952 to 2010. Their study identified the ARIMA (2,1,0) model as the most suitable, achieving forecasting accuracy with a Mean Absolute Percentage Error (MAPE) of 10.5%. The research demonstrated that ARIMA models are effective in capturing long-term production trends and can serve as valuable tools for policy planning, especially regarding import strategies. However, while the model successfully captured overall trends, it did not incorporate external variables such as climatic or economic factors. This limitation highlights the need for more dynamic models that can integrate exogenous influences affecting rice production. The current study addresses this gap by employing SARIMAX models that incorporate rainfall and harvested acres as external regressors, thereby enhancing forecasting precision and enabling a more holistic understanding of the factors impacting rice yield in Sri Lanka.

A recent study (Napagoda, 2021) analyzing paddy production trends in Sri Lanka from 1952 to 2020 introduced a seasonal approach by developing separate ARIMA models for the Yala and Maha cultivation periods. Using the Mann-Kendall and Cox-Stuart trend tests, the study confirmed increasing production trends in both seasons, with the Maha season showing a more pronounced upward slope. The best-fit ARIMA (2,1,1) model for the Yala season and ARIMA (2,1,0) for the Maha season were validated using AIC, RMSE, and MAPE, indicating the utility of season-specific modeling in improving forecast accuracy. This approach highlights the importance of disaggregating data by cultivation season; a concept incorporated in the present study using seasonal time series models like SARIMA and SARIMAX. By explicitly modeling seasonal dynamics and integrating exogenous climatic variables such as rainfall and harvested acres, this study aims to advance forecasting precision beyond traditional univariate ARIMA frameworks.

The Crop Forecast Report (Crop Forecast Maha, 2019/2020), published by the Department of Agriculture, provides vital empirical data on the actual sown extent, production estimates, and climate-related damages for paddy, other field crops (OFCs), and vegetables across Sri Lanka. The report indicates that 606,702 hectares of paddy had been sown by November 2019-74% of the seasonal target—projecting a national paddy production of 2.40 million metric tons. It also highlights how regional variances and natural disasters, such as the 27,750-metric-ton production loss in Batticaloa due to flooding, can significantly affect yield forecasts. This real-time data is highly relevant to modeling efforts, supporting the inclusion of seasonal**,** climatic**, and** regional cultivation trends in forecasting systems. The current study incorporates such dynamics by using exogenous climatic variables (e.g., rainfall) and seasonal components (Yala and Maha) in SARIMAX modeling to develop a more responsive and adaptive rice production forecasting framework.

Sri Lanka's traditional rice farming practices (Dharmasena, 2010), rooted in over 5,000 years of history, represent a sustainable and ecologically attuned agricultural system. Ancient Sinhalese societies engineered sophisticated irrigation infrastructure, including tanks, canals, and reservoirs, forming what is now known as the “tank-village” system. These systems were strategically designed to manage water resources efficiently across diverse landscapes, combining lowland paddy fields, home gardens, and upland chena cultivation. Traditional methods like *bethma govithena* (water sharing during droughts) and *kekulam govithena* (dry sowing) showcased adaptive strategies to variable rainfall. Studies have shown that traditional farming reduces soil salinity, increases nutrient availability, and supports biodiversity while being more cost-effective than modern conventional methods. Incorporating this perspective into rice production forecasting allows for a deeper understanding of the agricultural landscape, highlighting that effective predictive models must account for not just climatic and economic variables, but also the legacy of sustainable farming practices shaped by Sri Lanka's cultural and environmental heritage.

A recent study (Sherin Kularathne, 2024) explored the predictive relationship between macroeconomic indicators and rice production in Sri Lanka from 1960 to 2020 using various machine learning models, including Linear Regression, Support Vector Machines (SVM), Ensemble Methods, and Gaussian Process Regression (GPR). Key economic variables such as GDP, inflation, manufacturing output, imports, and arable land area were analyzed to assess their impact on production levels. The findings revealed that models like GPR and SVM performed well, while Decision Trees showed relatively poor accuracy. This research highlights the potential of machine learning techniques in capturing complex, non-linear relationships between economic factors and agricultural output. The present study employs time series models such as SARIMAX to incorporate exogenous variables specifically rainfall and harvested area alongside machine learning methods like Random Forest and XGBoost, enhancing predictive accuracy by addressing both seasonal and economic influences on rice production in Sri Lanka.

A study (Piyal Ekanayake, 2022) employing Gaussian Process Regression (GPR) investigated the relationship between climatic variables and paddy yield across Sri Lanka, using weather data from 2009 to 2019 and yield data from 2004 to 2019. Independent variables included rainfall, temperature, humidity, wind speed, evaporation, and sunshine hours, while paddy yield served as the target variable. GPR was implemented using various kernel functions Matern 5/2, Squared Exponential, Rational Quadratic, and Exponential and evaluated through robust statistical metrics such as MAPE, MSE, and the Nash number. The results demonstrated that GPR outperformed traditional statistical models, SVM, and Artificial Neural Networks in terms of predictive accuracy. This study reinforces the importance of modeling non-linear climatic influences on rice production. The current work adopts a similar data-centric approach by incorporating exogenous climatic features like rainfall into time series models such as SARIMAX, as well as evaluating machine learning models including Random Forest and XGBoost to capture climatic impacts on rice yield in the Sri Lankan context.

A recent study (Windhya Rankothge, 2021) developed crop-weather models for paddy yield prediction in Sri Lanka using multiple regression techniques and machine learning algorithms, with a focus on identifying the most influential weather variables. Using data from 2009 to 2019 across seven major paddy-producing districts, the study analyzed Yala and Maha seasons separately to capture seasonal variability. Random Forest (RF) outperformed Multiple Linear Regression (MLR) and Power Regression (PR), producing the lowest RMSE, MAE, and MAPE values. The model identified minimum relative humidity and maximum temperature as the most significant weather indicators influencing yield, followed by wind speed, evaporation, and rainfall. This work confirms the suitability of RF for handling complex, nonlinear interactions among climatic variables. In the present study, RF is similarly employed to evaluate the impact of key weather and agricultural features—such as rainfall and harvested area on rice production in Sri Lanka, affirming its effectiveness as a robust predictive tool.

A recent study (Sangarasekara1, 2024) applied machine learning techniques, particularly Random Forest (RF), to predict rice production in Sri Lanka using weather data from the Yala and Maha seasons between 1982 and 2019. By focusing on two major districts, Kurunegala and Anuradhapura, the researchers demonstrated that RF outperformed other models such as Linear Regression, Support Vector Regression (SVR), k-Nearest Neighbors (KNN), and Multilayer Perceptron (MLP), particularly in capturing yield variability across seasonal and regional conditions. The study analyzes the value of feature engineering, showing that RF could achieve comparable prediction accuracy using a reduced feature set, thus increasing model efficiency without compromising performance. These findings validate the application of RF in rice yield forecasting and support its use in the current study, where weather-related variables such as rainfall and harvested area are utilized to predict rice production trends. The ability of RF to model complex interactions and non-linear relationships makes it a powerful tool for agricultural decision-making in the context of climate variability and food security.

A recent econometric analysis (B. K. D. J. R. Samarasinghe, 2025) examined the symmetric and asymmetric impacts of climate change on rice production in Sri Lanka using ARDL and NARDL models, with annual data from 1952 to 2022. The study revealed that temperature and cultivated land exert long-term effects on rice yield, while rainfall impacts are asymmetric—positive rainfall changes tend to reduce yields, whereas negative rainfall variations (possibly indicative of less extreme or more optimal conditions) enhance production. The findings underscore the complex, non-linear nature of climate-rice interactions, highlighting the importance of considering both the magnitude and direction of climatic changes. This research supports the use of advanced modeling techniques that account for external climatic variables in rice yield forecasting. In the present study, rainfall and harvested areas are similarly used as exogenous regressors in SARIMAX models, reflecting a data-driven approach to understanding seasonal and climatic dependencies in Sri Lanka's rice production landscape.

A study (DHINAKARAN SAKTHIPRIYA1, 2024) conducted in the Madurai district of Tamil Nadu, India, developed several machine learning models to predict paddy yield using daily weather variables, soil nutrients, and water availability during the cropping period. Among the models tested—Linear Regression (LR), Support Vector Regression (SVR), Random Forest Regression (RFR), and CatBoost Regression (CBR)—the hybrid CBR model incorporating Variance Inflation Factor (CBR-VIF) achieved the highest predictive accuracy across all regional divisions, with normalized RMSE values as low as 0.56%. The use of VIF helped reduce multicollinearity among features, enhancing the interpretability and reliability of model predictions. Although the present study does not employ CBR or VIF-based hybridization, it draws on similar principles by using Random Forest to evaluate the effects of weather-related variables such as rainfall and harvested area on rice production. This reinforces the role of machine learning in agricultural modeling, where careful feature selection and model optimization are critical for achieving accurate and scalable yield forecasts.

The study (Yoshino & Kurniadi, 2023) conducted in Indonesia demonstrates the effectiveness of using machine learning models such as Linear Regression, Support Vector Machine, and XGBoost Regression for rice production forecasting. Among the models tested, XGBoost achieved superior accuracy across all evaluation metrics (e.g., RMSE = 0.015, MAE = 0.008, $R^2 = 0.997$), highlighting its strength in modeling complex, non-linear relationships in relatively short-term datasets (2018–2022). This demonstrates the suitability of ensemble techniques like XGBoost for agricultural prediction tasks, particularly in contexts where policy is a concern, like food sustainability and import planning.

# Chapter 03:

# METHODOLOGY

## 3.1. DATA COLLECTION

The dataset consists of agricultural, economic, and climatic data for Sri Lanka spanning from 1950 to 2024. It includes both Yala and Maha rice cultivation seasons, capturing important factors associated with rice production and general economic and climatic conditions.

The variables in the dataset offer a thorough understanding of the economic and climatic factors affecting rice production over a 74-year period.

| **Variable** | **Definition** | **Source** |
|---|---|---|
| Season | Specifies whether the record pertains to the Yala season or Maha season | Department of Census and Statistics, Sri Lanka |
| Sown (*000) Acres | The total area (in acres) where rice has been planted | Department of Census and Statistics, Sri Lanka |
| Harvested (*000) Acres | The actual land area (in acres) from which rice was successfully harvested | Department of Census and Statistics, Sri Lanka |
| Rainfall (mm) | Annual rainfall during a specific season either Yala or Maha | Department of Meteorology |
| Temperature (Celsius) | Annual temperature during a specific season either Yala or Maha | Department of Meteorology |
| Production (*000 Mt.) | The total rice production in metric tons | Department of Census and Statistics, Sri Lanka |
| GDP ($B) | GDP of Sri Lanka in billion dollars | Central Bank of Sri Lanka |
| Inflation (%) | Inflation rate of Sri Lanka in percentage | Central Bank of Sri Lanka |

Table 3-1: Dataset Overview

## 3.2. DATA PREPROCESSING

The dataset spans from 1950 to 2024 and includes various economic and climatic indicators and rice production data from Sri Lanka. These indicators consist of Sown paddies in Acres, Harvested paddies in acres, Rainfall in millimeters, Temperature in Celsius, Production of rice in metric tons, country's GDP in billions of dollars, and Inflation rate as a percentage. This dataset allows for a reliable method of projecting rice production by integrating agronomic, economic, and climatic variables. Rainfall and temperature are included to enable correlation analysis with production, where production is the target variable, an area that was frequently ignored in earlier studies. To determine the main factors influencing Sri Lankan rice production, its organized format allows the use of statistical analysis and time series models such as ARIMA, SARIMA, and SARIMAX.

Before conducting analysis, the dataset was examined for missing values. The results confirmed that there were no null values across any of the variables. As a result, statistical and times series models could be applied directly, guaranteeing data completeness and removing the requirement for imputation or data cleaning associated with missing values.

And this study applied some machine learning models such as XGBoost and Random Forest to compare the accuracies with time series models. To apply machine learning models, the data should be cleaned and well structured. Since the data has no missing value and is well structured, the data was then normalized and split into 7:3 ratio.

Table 3-2: (Summary of statistical measures for economic, agricultural, and climatic variables in Sri Lanka (1950-2024).

| **Variable** | **Mean** | **Std Dev** | **Min** | **25%** | **50% (Median)** | **75%** | **Max** |
|---|---|---|---|---|---|---|---|
| Year_New | 1986.75 | 21.58 | 1950 | 1968 | 1987 | 2005 | 2024 |
| *Sown Area (000 Acres) | 1,025.88 | 404.36 | 354 | 684 | 948 | 1,329 | 2,007.98 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| *Harvested Area (000 Acres) | 971.64 | 392.13 | 328 | 661 | 888 | 1,250 | 1,996.28 |
| GDP (Billion Rs) | 32.40 | 119.73 | 0.489 | 1.97 | 6.83 | 28.28 | 1,435.00 |
| Inflation (%) | 8.33 | 6.71 | -1.54 | 3.87 | 7.01 | 11.38 | 49.72 |
| Rainfall (mm) | 1,093.53 | 633.11 | 385 | 477 | 894 | 1,730 | 2,470 |
| Temperature (°C) | 27.06 | 1.29 | 24.48 | 26.32 | 26.76 | 27.35 | 31.08 |
| *Production (000 Mt.) | 1,171.17 | 732.83 | 167 | 534 | 1,049 | 1,671 | 3,197 |

Table 3-2: Summary of the Dataset

## 3.3. EXPLORATORY DATA ANALYSIS

### 3.3.1. TIME SERIES

To begin the time series modelling process, the relationship between rice production and its environmental and climatic factors that impact production were analyzed. The relationship between the variables in the dataset was investigated using Spearman correlation because the dataset was non-linear. Study aims to correlate rainfall and temperature with the target variable production to see how these factors impact the rice production, while the correlation for production and temperature showed a weak negative relation and rainfall and production showed a weak positive correlation. In the line plot of rainfall, temperature, harvested acres, and production that changes over time, rainfall, harvested acres, and production showed an upward trend while, temperature was showing a constant variation without any fluctuations or upward trend.

Therefore, we applied normalization for rainfall, temperature, harvested acres, and production to address scale differences between the variables. The variance was then stabilized, and the relationships were linearized using log transformation. Log transformation was used to reduce the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) scores and to improve the model accuracies as well. Finally, several time series models such as SARIMAX were fitted by taking the target variable as production and exogenous variables as rainfall, temperature, and harvested acres. Still the p-value of temperature in the result was greater than 0.05. So, we had to exclude the temperature, where it already showed a weak negative correlation with production. And temperature doesn't have a great impact on rice production than rainfall. Then the focus was changed to rainfall and harvested acres, two variables that have larger positive relationships with rice production.

### 3.3.2. STATIONARITY CHECK

To evaluate the stationarity of the rice production time series prior to any transformations, the augmented dickey-fuller (ADF) test was conducted. Since the p-value is considerably higher than the conventional significance level of 0.05, the null hypothesis of non-stationarity cannot be rejected. This result indicates that the series is non-stationary in its current form and necessitates transformation or differencing to achieve stationarity before proceeding with time series modeling.

Log transformation: $\boldsymbol{y' = log(y)}$

where 'y' is the original value and $y'$ is the logged-transformed value

### 3.3.3. STATIONARITY CHECK AFTER LOG TRANSFORMATION

After taking the variables rainfall and harvested acres along with production, applied log transformation to the variables directly without normalization as the variables do not need to be normalized. To check stationarity, the Augmented Dickey-Fuller (ADF) test was applied after log-transformation applied to production variable. The ADF test helps identify whether a time series is stationary or has a unit root. Since the p-value remained above 0.05, indicating non-stationarity, applied first differencing after the log transformation for production.

First Differencing: $\boldsymbol{y_t' = y_t - y_{(t-1)}}$,

Where '$\boldsymbol{y_t}$' is the current value at time 't', '$\boldsymbol{y_{(t-1)}}$' is the previous value at time 't', and $\boldsymbol{y_t'}$ is the first-differenced value

After the first difference of production, the p-value resulted in less than 0.05. Then the Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) plots were used to determine the appropriate parameters for ARIMA modelling. SARIMA and SARIMAX models were chosen to capture both trend and seasonality because the data is seasonal and includes two different agricultural seasons (Maha and Yala). SARIMAX used to include harvested acres and rainfall exogenous variables. To find the best accurate model, several model configurations were analyzed.

Such as, SARIMAX (1,1,0) (1,1,0,2), SARIMAX (1,1,1) (1,1,0,2), and SARIMA (1,1,1) (1,1,0,2). Assumptions like Linearity, Stationarity, Independence and no auto correlation of residuals, normally distributed residuals, Homoscedasticity, no perfect multicollinearity in exogenous variables, and Exogenous variables are known and accurate were checked before fitting SARIMAX models. Evaluation metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) were used to determine the final model selection. This ensured that the model selected offered the highest prediction accuracy while managing seasonality and external influences.

### 3.3.4. TIME SERIES MODEL EQUATIONS

1. SARIMA (p, d, q) (P, D, Q, s)

Where:

- $p$ = order of the non-seasonal AR (Auto-Regressive) part
- $d$ = order of the non-seasonal differencing
- $q$ = order of the non-seasonal MA (Moving Average) part

- *P* = order of the seasonal AR part
- *D* = order of the seasonal differencing
- *Q* = order of the seasonal MA part
- *s* = length of the seasonal cycle

$$\text{SARIMA} = \phi_p(B^s)\phi_p(B)(1-B)^d(1-B^S)^D y_t = \Theta_Q(B^s)\theta_q(B)\varepsilon_t$$

Where:

- $y_t$: the observed time series
- $\varepsilon_t$: white noise (error term)
- B: backward shift operator
- $\phi_p(B)$: non-seasonal AR polynomial
- $\theta_q(B)$: non-seasonal MA polynomial
- $\phi_p(B^s)$: seasonal AR polynomial
- $\Theta_Q(B^s)$: seasonal MA polynomial
- $(1-B)^d$: non-seasonal differencing
- $(1-B^S)^D$: seasonal differencing

2. SARIMAX (SARIMA with exogenous variables)

- SARIMAX (p, d, q) (P, D, Q, S) with exogenous variables '$x_t$.

$$\text{SARIMAX} = \phi_p(B^s)\phi_p(B)(1-B)^d(1-B^S)^D y_t = \Theta_Q(B^s)\theta_q(B)\varepsilon_t + \beta x_t$$

Where:

- $y_t$: the observed time series
- $\varepsilon_t$: white noise (error term)
- B: backward shift operator

- $\phi_p(B)$: non-seasonal AR polynomial
- $\theta_q(B)$: non-seasonal MA polynomial
- $\phi_p(B^s)$: seasonal AR polynomial
- $\Theta_Q(B^s)$: seasonal MA polynomial
- $(1-B)^d$: non-seasonal differencing
- $(1-B^S)^D$: seasonal differencing
- $x_t$: vector of exogenous variables at time $'\mathrm{t}'$
- $\beta$: coefficient vector for the exogenous variables

### 3.3.5. EVALUATION METRICS EQUATIONS

3. MSE (Mean Squared Error)

$$\text{MSE} = \frac{1}{n}\Sigma(y_i - \bar{y}_i)^2$$

- Helps detect large prediction errors in the study.
- Highlights the magnitude of the error to help guide model tuning.

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n}\left(\frac{\bar{y}-y_i}{n}\right)^2}$$

- It makes it easy to determine forecast errors in metric tons.
- Makes it possible to compare model performance directly.

5. MAPE (Mean Absolute Percentage Error)

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-\bar{y}_i}{y_i}\right| \times 100\%$$

- It provides a percentage representation of the model's accuracy.
- Useful for communicating findings to those who are not technical (e.g., farmers, policymakers).

Where:

- n = number of observations
- yi = actual value
- y^I = predicted value
- (yi – y^i) = squared difference between actual and predicted values

### 3.3.6. MACHINE LEARNING

In this study, machine learning models such as Random Forest and XGBoost were employed to improve rice production forecasting by capturing complex, non-linear relationships among key variables. These models were trained using features including rainfall, harvested acres, and sown acres as they correlate well with production. Where, highly correlated variables improve model performance, reduce noise, faster training, and improve interpretability. Before modelling, the dataset was examined for outliers across key variables, and they were removed using the Interquartile Range (IQR) method to improve model reliability. Any data points that fall below Q1 – 1.5IQR (Lower Bound) or above Q3 + 1.5IQR (Upper Bound) are regarded as outliers.

$$IQR = Q_3 - Q_1$$

Where:

- $Q_1$= First quartile (25th percentile)
- $Q_3$= Third quartile (75th percentile)

And the categorical variable 'Season' was label encoded to convert it to numerical variables (Yala:0, Maha:1). Following outlier treatment, a correlation analysis was conducted to identify the most influential factors affecting rice production. Based on the results, rainfall (a climatic factor) and harvested acres (an environmental factor) were selected as the primary independent variables due to their strong correlation with the target variable production. Since Random Forest and XGBoost don't require normalization or standardization, this study refused to perform normalization as both are tree-based models, where normalization or scaling doesn't impact tree splits.

Finally, the dataset was split into 70% training and 30% testing subsets, and both Random Forest and XGBoost models were trained to prdict rice production, aiming to capture both linear and non-linear dependencies for improved forecasting accuracy.

1. Random Forest

$$y'RF = \frac{1}{T}\sum_{t=1}^{T} y'_t$$

Where:

- $y'RF$ : Final prediction (regression output)
- T: Total number of trees in the forest
- $y'_t$: Prediction from the decision tree for input 't'

a. Assumptions

- Training samples are assumed to be independently and identically distributed.
- Assuming some features (alone or combined) help predict the target.
- Require enough samples to build diverse, reliable trees.
- Input data should be cleaned where Random Forest doesn't handle missing values natively.
- Doesn't assume normal distribution or require feature scaling.

2. XGBoost

- Input features should have meaningful relationships with the target variable.

- XGBoost does not require normality or linearity in the data.

- Needs enough data to learn complex patterns without overfitting.

- XGBoost can handle missing values internally.

- Does not require feature scaling like normalization or standardization.

3. Equation for R² evaluation metric

$$R^2 = 1 - \frac{\sum_{1=1}^{n} \quad (y_i - y_i')^2}{\sum_{1=1}^{n} \quad (y_i - \bar{y})^2}$$

Where:

- $y_i$= Actual value

- $y_i'$= Predicted value

- $\bar{y}$ = Mean of actual values

- n = Number of observations

# Chapter 04:

# RESULTS AND DISCUSSION

## 4.1. EXPLORATORY DATA ANALYSIS

The figure (Figure 4-1) below shows histograms for the variables in the dataset, giving an overview of how data values are spread across different features. Most of the Sown and Harvested areas are less than 1500 acres, indicating a right skewed distribution. The GDP is heavily skewed, showing low initial values followed by rapid growth in recent years. Rainfall and Inflation also show the right skewness, indicating occasional spikes. The distribution temperature is almost typical, indicating stable climate conditions. Production seems to be multifaceted, indicating variability over time. These patterns help identify skewness, outliers, and transformation needs, for efficient data preprocessing and predictive modelling.

### 4.1.1. DISTRIBUTION OF VARIABLES

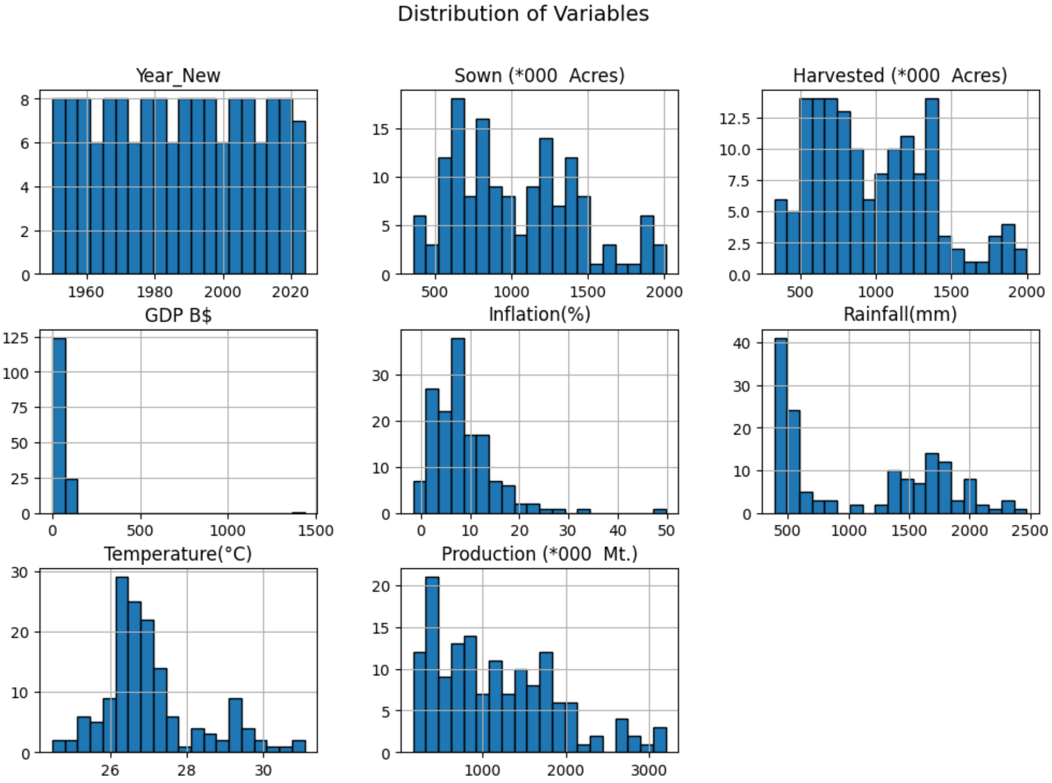


Figure 4-1: Distribution of Variables

A normality test on all numerical variables using Shapiro-Wilk test was done. After iterating over each numerical variable and testing if the data there follows a normal distribution using the ‘shapiro()’ function. If the p-value is less than a selected significance value (usually 0.05) suggests that the data in that variable is probably not normally distributed.

```
Year_New: p-value=0.0001
Sown (*000  Acres): p-value=0.0001
Harvested (*000  Acres): p-value=0.0001
GDP B$: p-value=0.0000
Inflation(%): p-value=0.0000
Rainfall(mm): p-value=0.0000
Temperature(°C): p-value=0.0000
Production (*000  Mt.): p-value=0.0000
```

The output represents that the p-values from the Shapiro-Wilk normality test applied to the numerical variables of the dataset. Each variable has a p-value significantly less than 0.05. These low p-values illustrate that the null hypothesis of the data being normally distributed is rejected for every variable. This test is done to choose the best correlation method to make correlations with the target variable ‘Production’. Since it’s a non-linear relationship where none of these numerical variables have a normal distribution. Therefore, choosing ‘Spearman’ correlation method as the best method based on normality test.

### 4.1.2. SPEARMAN CORRELATION HEATMAP

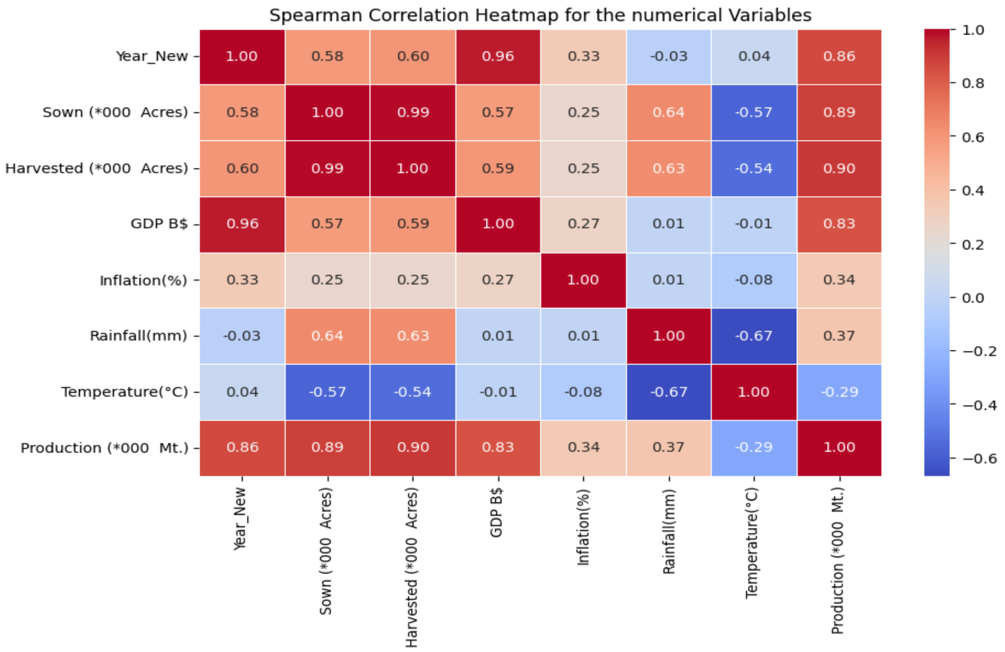


Figure 4-2: Spearman Correlation Heatmap

The heatmap (Fig 4-2) presented is a Spearman correlation matrix that shows the relationships between the numerical variables of the dataset. Spearman correlation is a non-parametric measure that assesses the direction and strength of non-linear correlations between two or more variables. Spearman is appropriate for capturing trends that are not always linear because it concentrates on rank-order relationships rather than linear ones like Pearson correlation does. A perfect positive monotonic relationship is represented by a correlation value of +1, a perfect negative monotonic relationship by a correlation value of -1, and no correlation is represented by a correlation value of zero. We can see a few strong positive relationships from the heatmap. Notably, there is an almost perfect link (0.99) between the Sown and Harvested areas, which is to be expected given that more land is often harvested when it is sown. Likewise, there are strong positive relationships between Rice Production and Harvested area (0.9) and Sown area (0.89).

Rainfall and Temperature have weak relationships with the production, but they go in different directions. Rainfall has a weak positive correlation (0.37), indicating a slight rise in production with additional rainfall. On the other hand, Temperature has a weak negative correlation (-0.29) between temperature and production, suggesting that higher temperatures may slightly reduce production due to heat stress. These findings show that while rainfall can moderately support rice production, rising temperatures could be dangerous underlining the need for climate-resilient farming practices. Since the research aims to make a correlation between climatic factors that impact rice production, it focuses on analyzing how these factors affect and understand how variations in these factors influence yield production.

## 4.2. TIME SERIES

Time Series analysis is a strong statistical method for examining data points collected or recorded at successive time intervals. In this study, time series techniques are used to investigate and understand the seasonal correlations between climatic factors and rice production in Sri Lanka which takes place during Yala and Maha seasons. Since rice is the nation’s staple food and is heavily dependent on seasonal weather patterns, it is crucial to analyze how these climatic factors affect production over time to ensure food security and efficient agricultural planning.

### 4.2.1. TIME SERIES PLOT (LINE PLOT) FOR TOTAL PRODUCTION

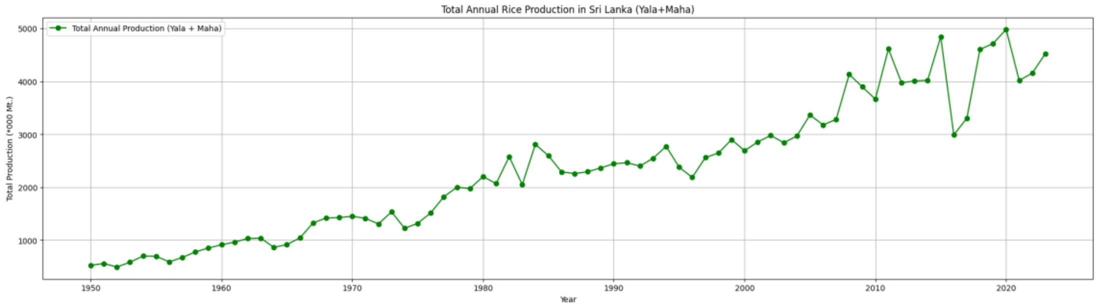


Figure 4-3: Line plot for the total production (Yala+Maha)

The time series plot (Fig 4-3) shows the total annual rice production in Sri Lanka combining both Yala and Maha seasons from 1950 to 2024. There is a clear upward trend in production over this time frame, suggesting that rice production has grown significantly. In the earlier years, especially from 1950 to the early 1970s, production was below 2000 (*000) metric tons, showing very slight annual increase. In the middle of the 1970s, the growth became more apparent, with frequent fluctuations. After 2010, there are clearer increases in the production approaching 5000 (*000) metric tons around 2020. The overall trend indicates that Sri Lanka has increased rice production significantly, even with some drops in certain years. Since the mean and the variance of the plot are not stable, differencing must be applied.

### 4.2.2. LINE PLOT OF PRODUCTION FOR YALA VS MAHA

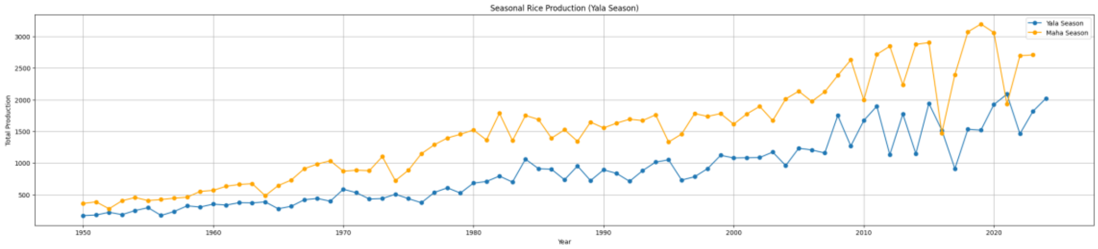
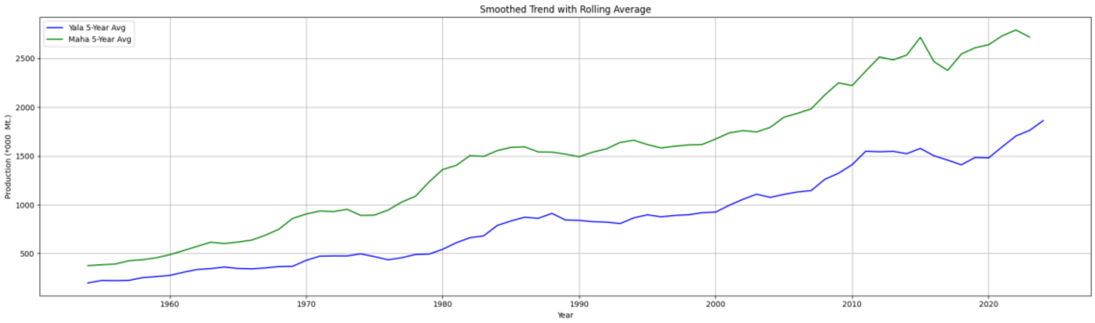


Figure 4-4: Line plot of production for Yala and Maha seasons

The plot (Fig 4-4) illustrates seasonal rice production in Sri Lanka from 1950 to 2024, comparing the Yala season (blue line) and the Maha season (orange line). It clearly shows that the Maha season produces more rice than the Yala season. Over time, both seasons show as increasing tendency in production, reflecting advancements in agricultural technology, policy, or practices. But overall, Maha season production has been more stable in addition to being higher. At the same time Yala season shows more fluctuations and severe drops. Overall, Maha season contributes more to total rice production.

### 4.2.3. ROLLING AVERAGE PLOT (YALA VS MAHA)



- Smooth fluctuations: The 5-year rolling average makes production trends more apparent by reducing short-term ups and downs.

- Seasonal Comparison: The trends of Yala and Maha together allow one to determine which season makes a greater contribution.

- Reduces Noise: Assists in improved long-term planning by filtering out annual variations (such as weather effects).

### 4.2.4. CORRELATED LINE PLOT

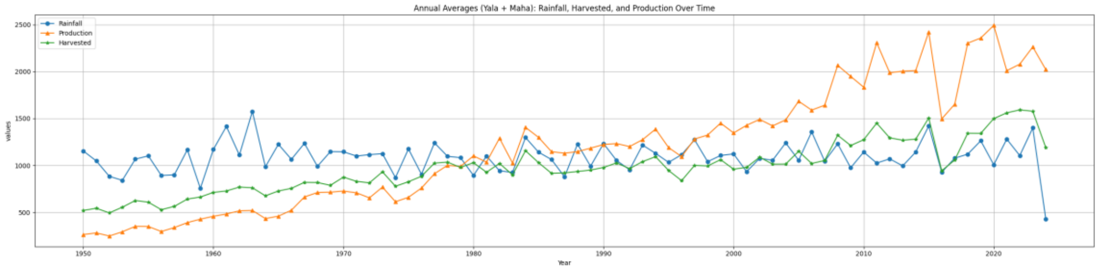


Figure 4-5: Line plot of production for mostly correlated climatic and environmental variables with production

- There is a strong positive correlation between harvested acres and production

- Rainfall shows a weak positive correlation with production.

- There is a little link between rainfall and harvested acres even in years with less rainfall, harvested area increases.

An Augmented Dickey-Fuller (ADF) test was performed to assess the stationarity of the rice production time series. The test yielded an ADF statistic of -0.5605 and a p-value of 0.8797, indicating that the series is non-stationary.

```
ADF Statistic: -0.5605203613523394
p-value: 0.879655633717813
```

As a result, transformation or differencing is required before applying time series forecasting models.

The variables production, rainfall, and harvested acres were subjected to a log transformation to improve model performance and achieve lower AIC and BIC scores. The given ADF statistics and p-value were examined for the log-transformed target variable (production).

```
ADF Statistic: -2.2782633547734594
p-value: 0.1790743275856005
```

### 4.2.5. ACF & PACF PLOTS AFTER LOG TRANSFORMATION

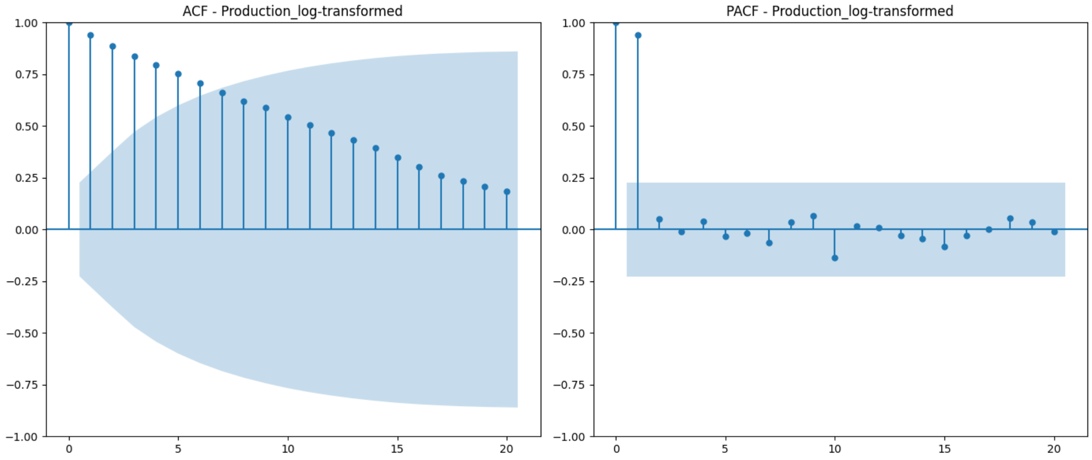


Figure 4-6: ACF & PACF plots for the log-transformed production

4. ACF (Autocorrelation Function):

   - Displays a slow decay in autocorrelations across many lags.
   - It shows that past values continue to affect future values.
   - Implies that AR (Autoregressive) components are present.

5. PACF (Partial Autocorrelation Function):

   - Only the first lag is quite large, with a sharp drop after lag 1.
   - Suggests a close and direct connection with one previous value.

According to the p-value and the ACF and PACF plots obtained above, it concludes that differencing must be made to make the mean and variance stable.

The given ADF statistics and the p-value were obtained after the first difference for the log-transformed production variable.

```
ADF Statistic: -8.508876402754584
p-value: 1.182266626807716e-13
```

Since the p-value is less than 0.05, mean and variance are stationary. Therefore, we can reject the null hypothesis ($H_0$).

### 4.2.6. FIRST DIFFERENCED TIME SERIES PLOT

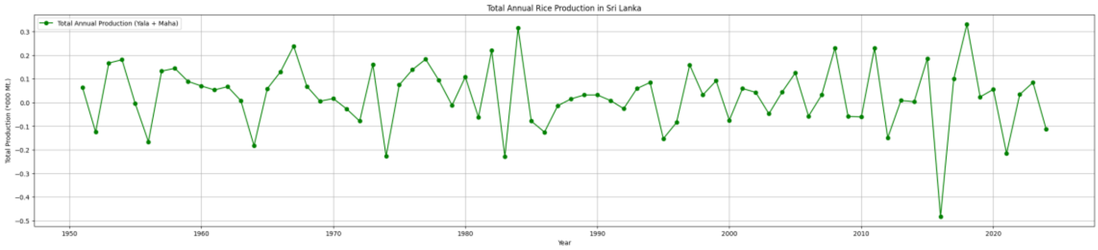


Figure 4-7: First differenced plot of production

- Stationarity is achieved after the first order differencing of the log-transformed series. It removed trends and stabilized variance, preparing the data for SARIMA/SARIMAX modelling.

- The absence of strong recurring seasonal spikes in the plot indicates that seasonal differencing was effective.

The ACF and PACF plots were plotted after the first difference of log-transformed production.

### 4.2.7. ACF & PACF PLOTS AFTER THE FIRST DIFFERENCE

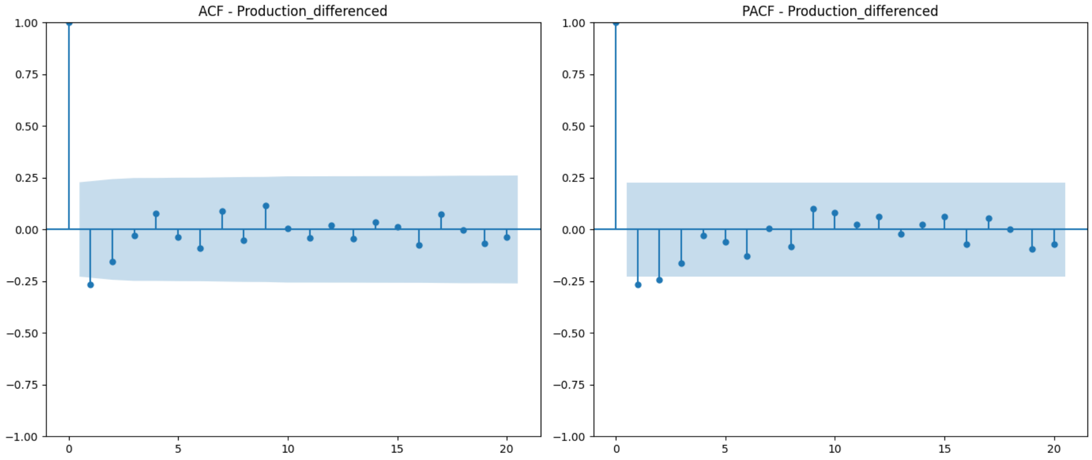


Figure 4-8: ACF & PACF plots after the first difference

6. ACF Plot

   - Strong spike at lag 1, then rapid drop-off.
   - Suggests a Moving Average (MA (1)) component is applicable.

7. PACF Plot

   - Strong spike at lag 1, followed by smaller minor spikes.
   - Suggests the possibility of an AR (1) component.

Since the plots are for the first differencing log-transformed series, and the series is now stationary. Recommended models based on these plots:

- ARIMA (1,1,1): combines both AR (1) and MA (1) components.

- ARIMA variants: ARIMA (1,1,0) and ARIMA (0,1,1,), comparing using AIC/BIC/MSE to select the best-fitting model.

- SARIMA/SARIMAX (Seasonal Models): seasonal order = (1,1,0,2) or (0,1,1,2). Using SARIMAX is the best option among these as we use rainfall and harvested acres as exogenous variables.

### 4.2.8. MODEL COMPARISONS

8. SARIMA (1,1,1) (1,1,0,2)

   - Using SARIMA without exogenous variables, applying only for the target variable production.

| Parameter | Coefficient | Std. Error | z-Value | p-Value | 95% CI (Lower) | 95% CI (Upper) |
|---|---|---|---|---|---|---|
| **AR(1)** | 0.3885 | 0.163 | 2.387 | 0.017 | 0.070 | 0.707 |
| **MA(1)** | -0.9723 | 0.086 | -11.356 | 0.000 | -1.140 | -0.804 |
| **Seasonal AR(2)** | -0.5217 | 0.101 | -5.157 | 0.000 | -0.720 | -0.323 |
| **$\sigma^2$ (Variance)** | 0.0192 | 0.003 | 5.859 | 0.000 | 0.013 | 0.026 |

Table 4-1: Results of SARIMA (1,1,1) (1,1,0,2)

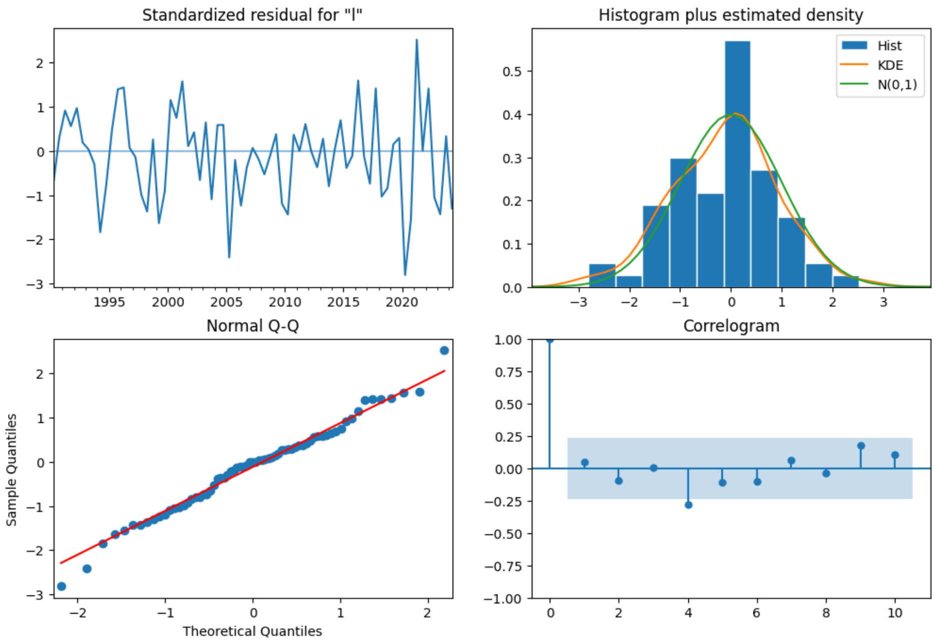


Figure 4-9: Residual Diagnostic Panel

- Standardized Residuals: Residuals fluctuate randomly around zero, indicating no obvious patterns left.

- Histogram + KDE: Residuals follow a near-normal distribution, ligning well with the overlaid normal curve.

- Q-Q Plot: Points lie close to the line, supporting the assumption of normal residuals.

- Correlogram: No specific autocorrelation in residuals, suggesting a good model fit.

9. SARIMAX (0,1,1) (0,1,1,2)

- Using SARIMAX model by taking the exogenous variables as log-transformed rainfall and harvested acres where they are the most correlated climatic and environmental factors that affect rice production.

| **Parameter** | **Coef** | **Std. Error** | **z-Value** | **p-Value** | **95% CI (Lower)** | **95% CI (Upper)** |
|---|---|---|---|---|---|---|
| **log_rain** | -0.0743 | 0.032 | -2.351 | 0.019 | -0.136 | -0.012 |
| **log_harvested** | 1.0952 | 0.086 | 12.748 | 0.000 | 0.927 | 1.264 |
| **MA(1)** | 0.0702 | 0.153 | 0.459 | 0.646 | -0.229 | 0.370 |
| **Seasonal MA(2)** | -0.9333 | 0.078 | -11.988 | 0.000 | -1.086 | -0.781 |
| **σ² (Variance)** | 0.0038 | 0.001 | 7.058 | 0.000 | 0.003 | 0.005 |

Table 4-2: SARIMAX (0,1,1) (0,1,1,2)

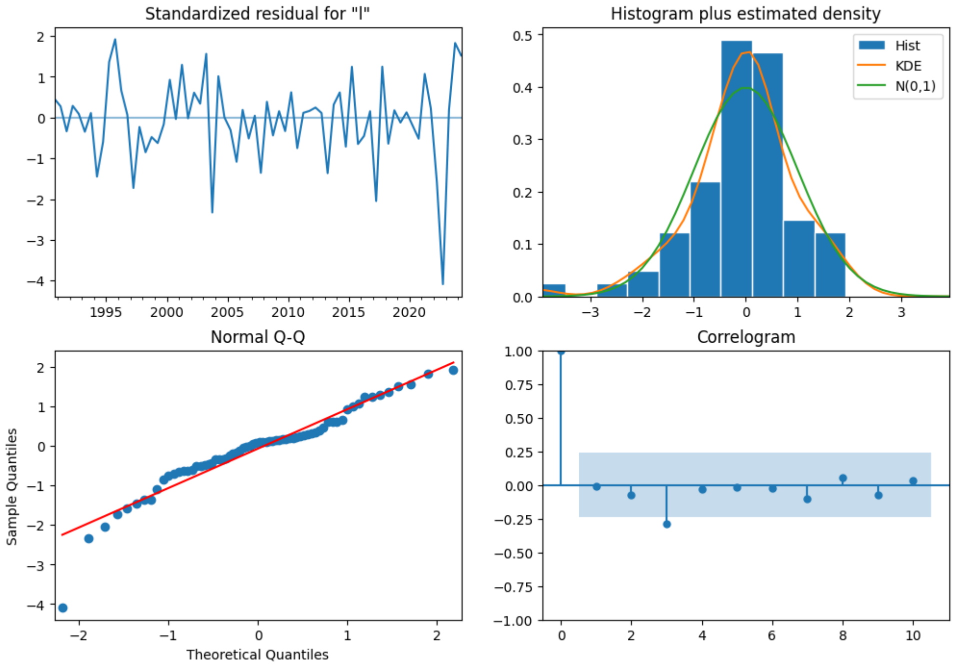


Figure 4-10: Residuals Diagnostic Panel

- Residuals move randomly, showing no clear pattern.

- Histogram**:** Residuals mostly follow a normal (bell-shaped) curve.
- Q-Q plot: Points are close to the line, so residuals are nearly normal.
- Correlogram**:** No strong spikes, residuals are not correlated.

10. SARIMAX (1,1,1) (1,1,0,2)

This model considered rainfall and harvested acres as exogenous variables.

| Parameter | Coef | Std. Error | z-Value | p-Value | 95% CI (Lower) | 95% CI (Upper) |
|---|---|---|---|---|---|---|
| `log_rain` | -0.0799 | 0.033 | -2.410 | 0.016 | -0.145 | -0.015 |
| `log_harvested` | 1.0819 | 0.073 | 14.846 | 0.000 | 0.939 | 1.225 |
| **AR(1)** | -0.6500 | 0.175 | -3.708 | 0.000 | -0.994 | -0.306 |
| **MA(1)** | 0.9296 | 0.122 | 7.638 | 0.000 | 0.691 | 1.168 |
| **Seasonal AR(2)** | -0.5113 | 0.129 | -3.950 | 0.000 | -0.765 | -0.258 |
| **$\sigma^2$ (Variance)** | 0.0047 | 0.001 | 6.740 | 0.000 | 0.003 | 0.006 |

Table 4-3: Results of SARIMAX (1,1,1) (1,1,0,2)

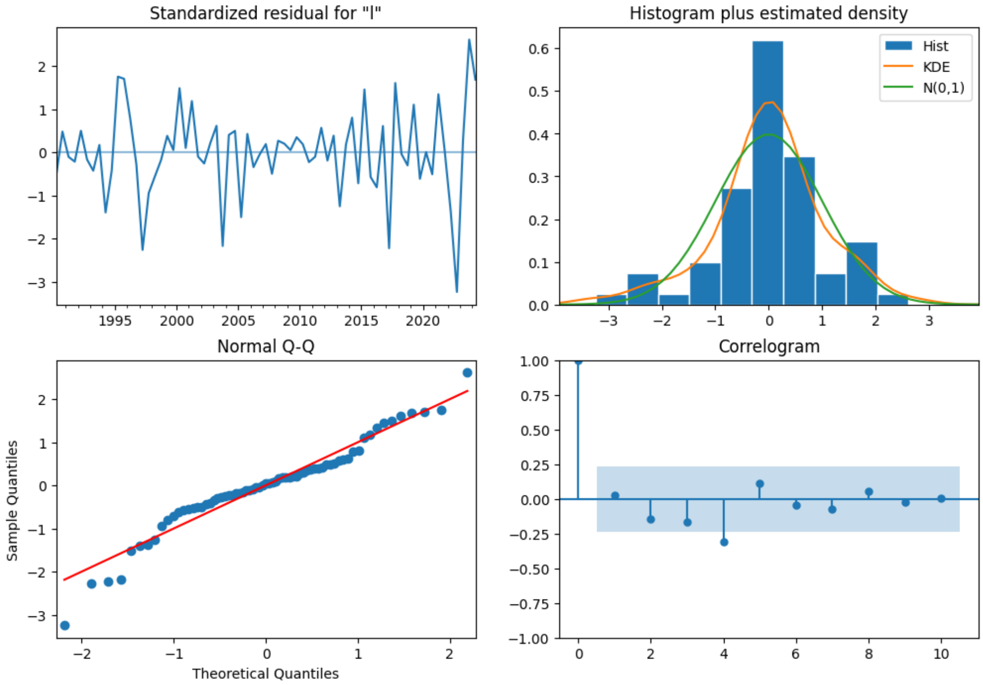


Figure 4-11: Residual Diagnostic Panel

- Constant residuals throughout the time, there is no visible trend or pattern as the residuals fluctuate around zero.
- Histogram shows that residuals have a normal distribution. Green curve closely follows the KDE (estimated density).
- Q-Q plot indicated that the residuals are roughly normally distributed.
- ACF show that there are no significant spikes beyond the confidence level. So, residuals are uncorrelated.

11. SARIMAX (0,1,1) (1,1,0,2)

    - Applying SARIMAX (0,1,1) (1,1,0,2) only by considering the log-transformed rainfall value.

| Parameter | Coef | Std. Error | z-Value | p-Value | 95% CI (Lower) | 95% CI (Upper) |
|---|---|---|---|---|---|---|
| **log_rain** | 0.2312 | 0.100 | 2.320 | 0.020 | 0.036 | 0.426 |
| **MA(1)** | -0.4043 | 0.147 | -2.753 | 0.006 | -0.692 | -0.116 |
| **Seasonal AR(2)** | -0.5531 | 0.113 | -4.875 | 0.000 | -0.775 | -0.331 |
| **σ² (Variance)** | 0.0218 | 0.004 | 5.369 | 0.000 | 0.014 | 0.030 |

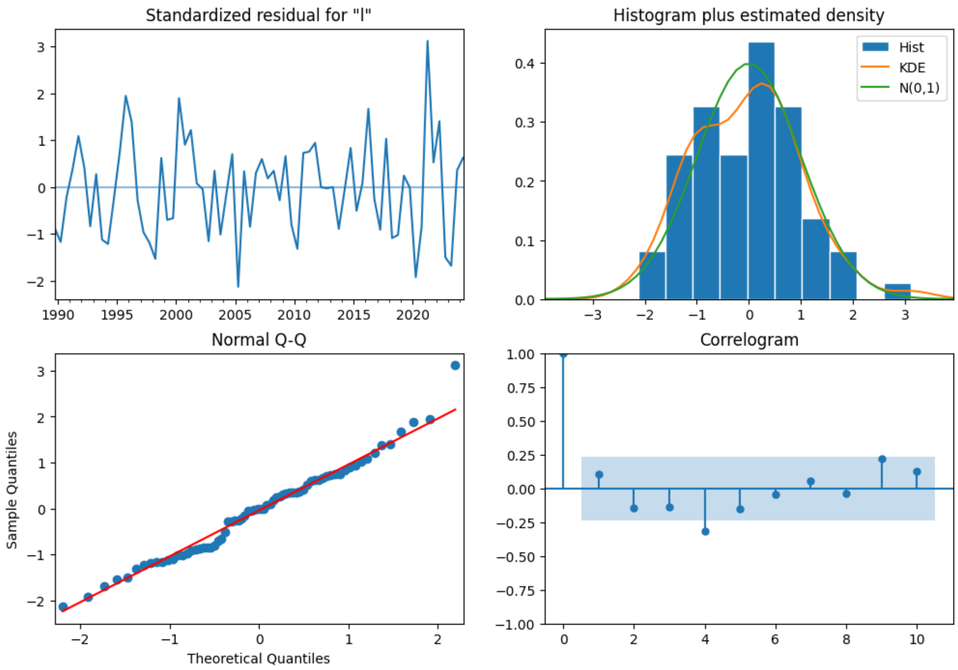


Figure 4-12: Residual Diagnostic Panel

- Residual Plot: Fluctuate around zero, indicating no pattern.
- Histogram + KDE: Residuals are approximately normally distributed.
- Q-Q Plot: Points along with the diagonal, supporting normality.
- Correlogram: No significant lags, residuals are uncorrelated.

### 4.2.9. PREDICTIVE ACCURACY MEASURES

| Model | MSE | RMSE | MAPE | AIC | BIC |
|---|---|---|---|---|---|
| **SARIMA (1,1,1) (1,1,0,2)** | 354229.8846 | 595.1721 | 40.83% | -65.936 | -56.999 |
| **SARIMAX (0,1,1) (0,1,1,2)** | 11470.6143 | 107.1010 | 6.30% | -173.927 | -162.829 |
| **SARIMAX (1,1,1) (1,1,0,2)** | 13143.5962 | 114.6455 | 6.67% | -159.465 | -146.061 |
| **SARIMAX (0,1,1) (1,1,0,2)** | 77490.2601 | 278.3707 | 21.47% | -60.922 | -51.928 |

Multiple time series models like SARIMA and SARIMAX were performed to accurately predict rice production. SARIMAX models included log-transformed rainfall and harvested acres as exogenous variables. Although SARIMAX (0,1,1) (0,1,1,2) showed better metrics (lowest MAPE 6.3% and AIC -173.927), it included an insignificant MA (1) term (p-value = 0.646) where p-value is greater than 0.05, reducing model performance. And it could be due to multicollinearity or overfitting. This means that the moving average component at lag 1 does not contribute meaningfully to modelling the series. On the other hand, SARIMAX (1,1,1) (1,1,0,2) had residuals that behaved well with all significant parameters (p-value < 0.05 for all the variables, MAPE = 6.67%, and AIC = -159.465). Its statistical stability and interpretability made it the best model for forecasting, balancing model validity and predictive accuracy with somewhat higher error metrics.

### 4.2.10. ACTUAL VS FITTED PLOT FOR THE BEST TIME SERIES MODEL

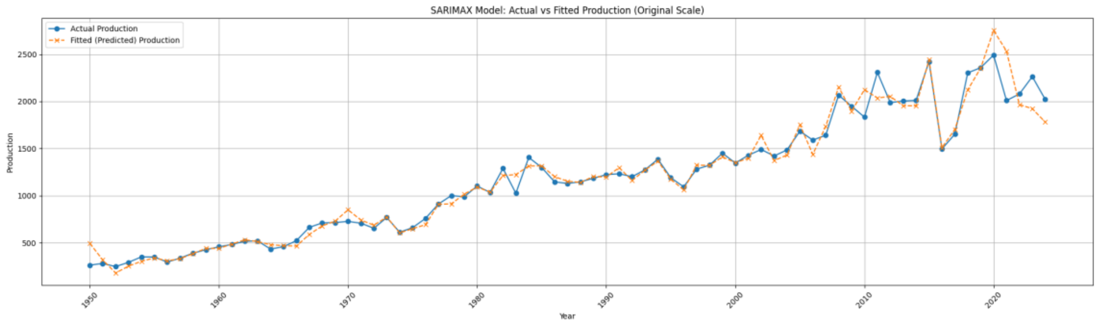


Figure 4-13: Actual vs Predicted plot

- Figure 4-12 shows the actual vs fitted rice production plot combining both Yala and Maha seasons, using the SARIMAX (1,1,1) (1,1,0,2) model.

- The model captures both short-term fluctuations and long-term growth patterns effectively.

- Particularly, throughout the years, the fitted line closely matches the pbserved data.

- Overall, the plot illustrates the model's excellent prediction capabilities and capacity to reproduce past patterns in rice production.

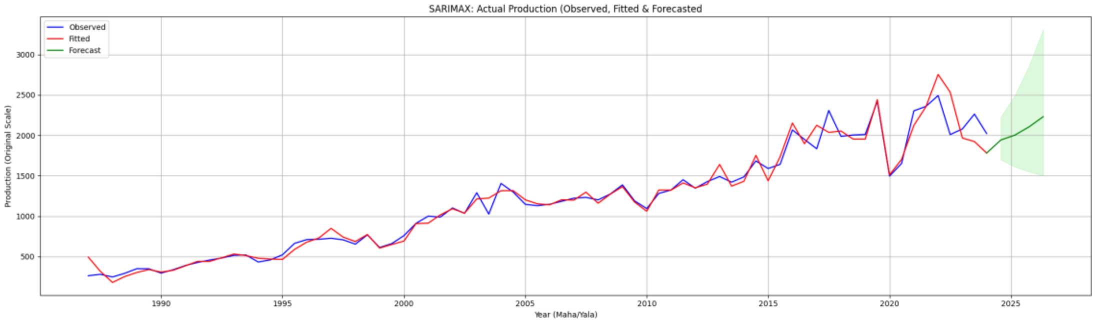
- The table below shows actual vs predicted values from 2016 to 2024.

| **Season** | **Year** | **Actual Production** | **Predicted Production** |
|---|---|---|---|
| Yala | 2020 | 1924 | 1986.17 |
| Maha | 2020/21 | 3061 | 1871.18 |
| Yala | 2021 | 2088 | 1453.87 |
| Maha | 2021/22 | 1931 | 2182.27 |
| Yala | 2022 | 1462 | 2026.28 |

| | | | |
|---|---|---|---|
| Maha | 2022/23 | 2696 | 2507.53 |
| Yala | 2023 | 1817 | 2664.44 |
| Maha | 2023/24 | 2709 | 2146.10 |
| Yala | 2024 | 2022 | 1844.58 |

Table 4-4: Actual vs Predicted values

### 4.2.11. ACTUAL, FITTED & FORECASTED PLOT



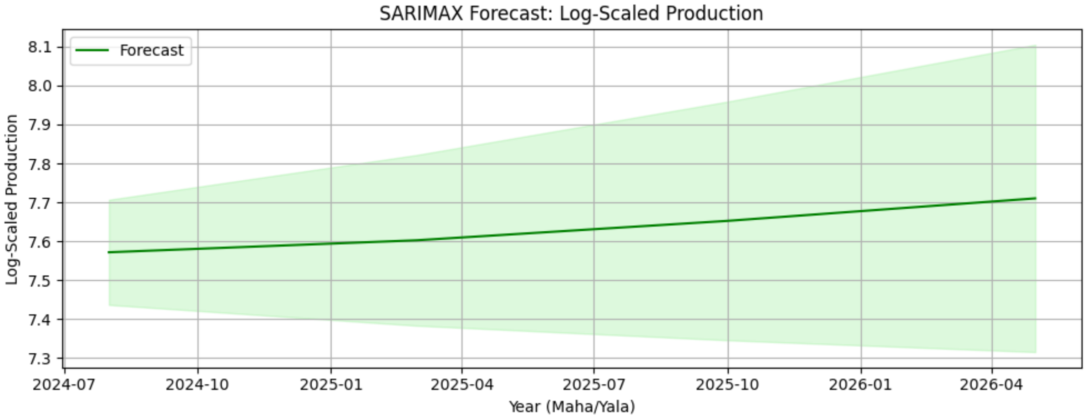
- Model fit: The SARIMAX model accurately captures past seasonal output trends, as seen by the red (fitted) and blue (observed) lines.

- Forecasting: By predicting production from the end of 2025 Maha to 2027 Yala, the forecasted line (green) helps predict future outputs by using historical trends.

- The shaded area in green color shows a margin of uncertainty, accounting for possible variations due to the correlation of real-world factors like weather and environment.

- Seasonal cycles: The model's efficiency in seasonal analysis has appeared, which corresponds to the alternating Maha and Yala seasons.

### 4.2.12. FORECASTED PLOT



- The green line in the forecasting plot shows the predicted rice production values over time, specifically from mod-2024 to mid-2026.

- The plot begins with the Maha season of 2024/2025, which usually begins in August. The model continues to forecast production through successive Maha seasons, spanning 2025/2026.

- The forecast also includes the 2025-2026 Yala seasons as well. These forecasts offer a thorough analysis of Sri Lanka's two main rice growing seasons, which helps in predicting patterns of production throughout the year.

| Season | Year | Forecasted Production Values (Original Scale): |
|---|---|---|
| Maha | 2024/2025 | 1942.53 |
| Yala | 2025 | 2003.28 |
| Maha | 2025/2026 | 2105.59 |
| Yala | 2026 | 2230.89 |

Table 4-5: Forecasted values of rice production

There is a high degree of consistency between the past data and the forecasted production values. The forecasted values exhibit a smoother and gradually increasing trend, which suggests improved stability in future forecasts. When seasonal variations and relevant exogenous factors like rainfall and harvested acres are included in the forecasting model, this pattern improves the forecasts' dependability.

## 4.3. MACHINE LEARNING

### 4.3.1. BOXPLOT OF THE VARIABLES

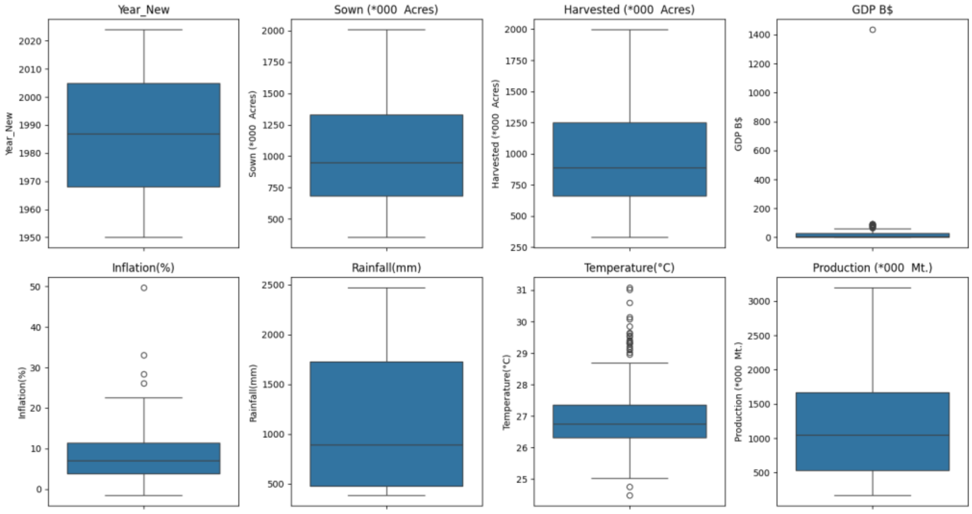


Figure 4-14: Box plot of the variables

- The box plot (Figure 4-13) Helps identify the median and variability in key variables, helping in selecting stable vs volatile predictors for modelling.

- Variables like GDP have several extreme outliers, inflation has multiple high outliers, and temperature has few high outliers.

- Understanding the shape of data distributions (symmetric vs skewed) supports choosing relevant statistical or machine learning models for analysis.

Based on boxplot visualizations, it was discovered that some variables, including GDP, Inflation, and Temperature, contained outliers during the initial exploratory data analysis.

```
Outliers by column:
GDP B$: 28 outliers
Inflation(%): 4 outliers
Temperature(°C): 23 outliers
```

These outliers may break the findings and harm the analysis's accuracy. Therefore, these outliers have been removed using the Inter Quartile Range (IQR) method.

```
Original shape: (149, 9)

Cleaned shape: (103, 9)
```

The results show how well the models can capture complex patterns and highlight key elements affecting rice production.

| **Models** | **R² Score** | **RMSE** | **MAE** | **MAPE** |
|---|---|---|---|---|
| Random Forest | 0.9134 | 210.33 | 160.85 | 19.25% |
| XGBoost | 0.8789 | 248.81 | 181.93 | 21.96% |

Table 4-6: Performance metrics for Machine Learning models

### 4.3.2. ACTUAL VS PREDICTED PRODUTION

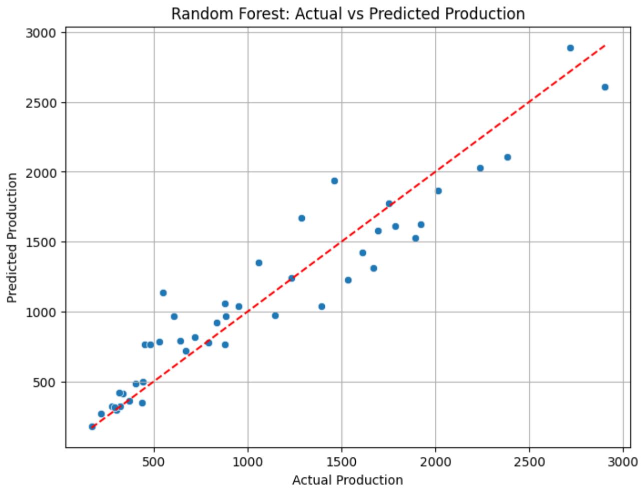


Figure 4-15: Actual vs Predicted Production for Random Forest Model

- According to the figure (Figure 4-14), most of the points are close to the red line, which means the model's predictions closely match the actual production values. This indicates that the model is performing well.
- Only a few production values are slightly off from the model's accurate prediction.
- The random forest model is appropriate for forecasting rice production because it performs well with a variety of data types, including rainfall, harvested acres, and sown acres.

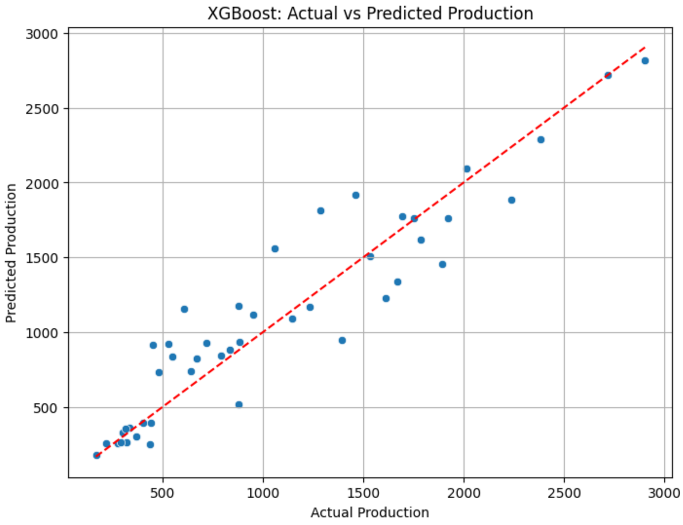


Figure 4-16: Actual vs Predicted plot for XGBoost Model

- Based on the above plot (Figure 4-15), most of the points are near the red line, indicating that the model correctly predicted the values.
- There are a few points that deviate slightly from the line, especially in the middle range, but no major errors are seen.
- Overall, the XGBoost model also does well and can be used to forecast rice production in this study.

Based on the performance metrics and actual vs predicted plots of both the models. The Random Forest model performs better than the XGBoost model in all key areas.

## 4.4. LIMITATIONS

### 4.4.1. TIME SERIES

Since this study mainly focuses on analyzing the correlation between rainfall, temperature, harvested acres and production, where production is the target variable. By identifying the statistical relationships between these variables, we can clearly understand how fluctuations in climate conditions impact production.

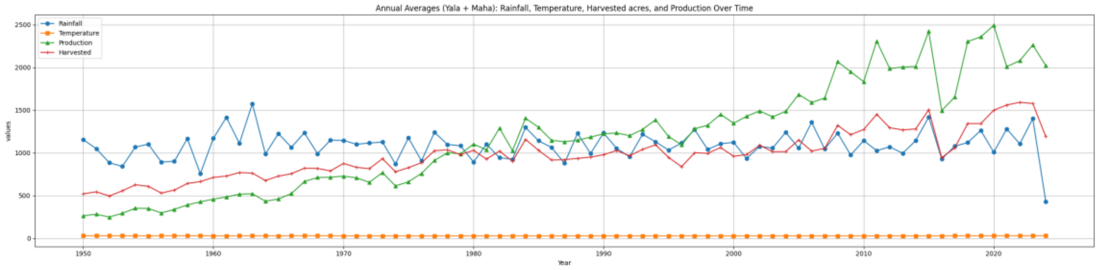


Figure 4-17: Line Plot of the variables

- This plot represents the annual averages of rainfall, temperature, harvested acres, and rice production in Sri Lanka (1950-2024), covering both Yala and Maha seasons.

- Rainfall and harvested acres show moderate fluctuations; temperature appears stable over time.

- Temperature seems flat due to scale mismatch, making it hard to analyze its relationship with other variables.

- Normalize or standardize variables for accurate pattern identification and relevant comparison in future visualizations.

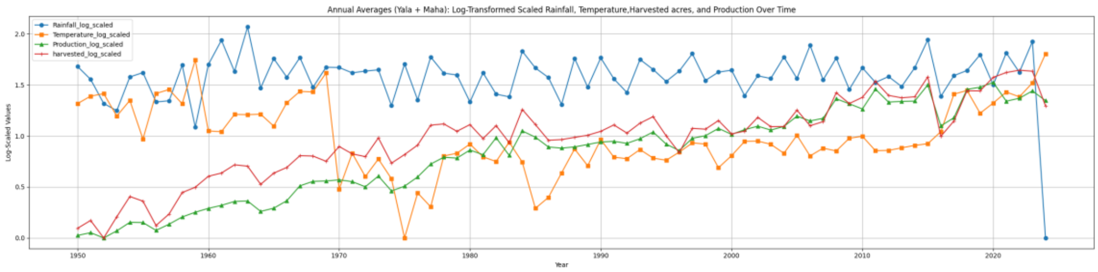


Figure 4-18: Normalized line plot of the variables

- All variables are now on a comparable scale after log transformation and normalization, making comparisons easy.

- Rainfall still varies significantly from year to year, and temperature swings are much more apparent than they were previously.

- Production and harvested acres exhibit comparable trends, pointing to a potential positive link.

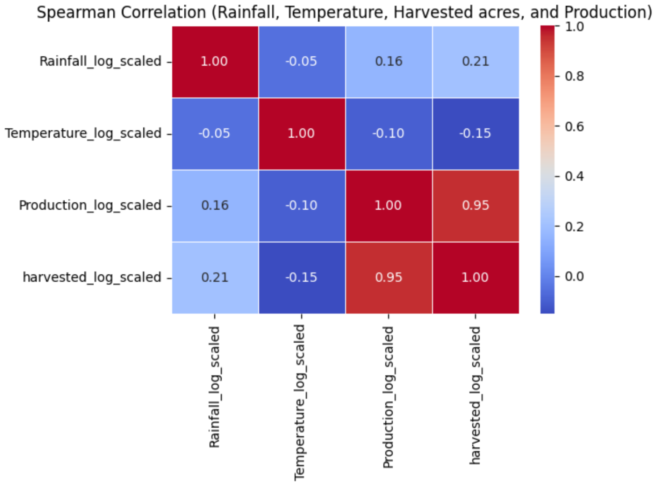


Figure 4-19: Spearman Correlation Heatmap

- According to the correlation heatmap, the strongest relation with rice production is the harvested area.

- Rainfall makes a positive relation but an insignificant contribution.

- There is a weak correlation between temperature and production.

Multiple SARIMAX models were tried to capture the p-value of the variable temperature in the results of the models:

| **Variable** | **Coefficient** | **Std. Error** | **z-score** | **P-value** | **95% Confidence Interval** |
|---|---|---|---|---|---|
| Rainfall_log_scaled | -0.0402 | 0.023 | -1.710 | 0.087 | [-0.086, 0.006] |
| Temperature_log_scaled | 0.0330 | 0.025 | 1.311 | 0.190 | [-0.016, 0.082] |
| Harvested_log_scaled | 0.6307 | 0.039 | 16.350 | 0.000 | [0.555, 0.706] |
| ar.L1 | -0.0511 | 0.124 | -0.411 | 0.681 | [-0.294, 0.192] |
| ar.S.L2 | -0.7599 | 0.094 | -8.115 | 0.000 | [-0.943, -0.576] |
| sigma² | 0.0028 | 0.000 | 7.787 | 0.000 | [0.002, 0.003] |

Table 4-7: SARIMAX (1,1,0) (1,1,0,2) results

| **Variable** | **Coefficient** | **Std. Error** | **z-score** | **P-value** | **95% Confidence Interval** |
|---|---|---|---|---|---|
| Rainfall_log_scaled | 0.1274 | 0.064 | 1.983 | 0.047 | [0.001, 0.253] |
| Temperature_log_scaled | 0.0362 | 0.064 | 0.565 | 0.572 | [-0.089, 0.162] |
| ar.L1 | 0.3221 | 0.164 | 1.963 | 0.050 | [0.000, 0.644] |
| ma.L1 | -1.0000 | 81.437 | -0.012 | 0.990 | [-160.613, 158.613] |
| ar.S.L2 | -0.5006 | 0.102 | -4.891 | 0.000 | [-0.701, -0.300] |

| | | | | | |
|---|---|---|---|---|---|
| sigma² | 0.0093 | 0.756 | 0.012 | 0.990 | [-1.473, 1.491] |

Table 4-8: SARIMAX (1,1,1) (1,1,0,2), exog variables = rainfall and temperature

| **Variable** | **Coefficient** | **Std. Error** | **z-score** | **P-value** | **95% Confidence Interval** |
|---|---|---|---|---|---|
| Temperature_log_scaled | -0.0016 | 0.063 | -0.026 | 0.979 | [-0.125, 0.122] |
| ma.L1 | -0.5779 | 0.104 | -5.544 | 0.000 | [-0.782, -0.374] |
| ma.S.L2 | -1.0000 | 104.991 | -0.010 | 0.992 | [-206.778, 204.778] |
| sigma² | 0.0086 | 0.904 | 0.010 | 0.992 | [-1.763, 1.780] |
| **Variable** | **Coefficient** | **Std. Error** | **z-score** | **P-value** | **95% Confidence Interval** |
| Temperature_log_scaled | -0.0016 | 0.063 | -0.026 | 0.979 | [-0.125, 0.122] |

Table 4-9: SARIMAX (0,1,1) (0,1,1,2). Exog variable = temperature

Although temperature was initially considered as a relevant factor that impacts rice production, exploratory correlation analysis showed weak relationships between temperature and other important variables. When included in the SARIMAX models, the temperature variable produced a p-value greater than 0.05. indicating that it was not statistically significant. This implies that temperature might not have a significant impact on short-term production forecasts in this dataset.

## 4.5. RECOMMENDATIONS

- Prioritize SARIMAX for seasonal forecasting: Use SARIMAX models for forecasting rice production when data shows strong seasonality and clear external influences like climatic and environmental factors.

- Utilize Machine Learning for Exploratory Analysis: While seasonal statistical models should not be used in place of machine learning models like Random Forest, they can be used to investigate feature importance and nonlinear relationships.

- Avoid Hybrid Models: Combining models (like SARIMAX + RF) should only be done if the residuals exhibit a predictable structure. If residuals seem erratic or noisy, steer clear of hybridization as it could impair model performance.

- Add More Exogenous Variables: It could be beneficial for future models to include additional variables like temperature, fertilizer use, the prevalence of pests and diseases, or economic indicators (such as paddy prices and subsidies).

- Create Easy-to-Use Resources for Farmers and Planners: To facilitate proactive planning, incorporate the forecasting model into a dashboard or decision-support system for agriculture officers and legislators.

# Chapter 05:

# CONCLUSION

This study forecasted Sri Lanka's rice production using both statistical and machine learning techniques. The SARIMAX method was utilized for time series modelling, and Random Forest and XGBoost were employed to capture complex non-linear patterns. Additionally, a hybrid model that combines SARIIMAX and Random Forest was explored.

Overall, the most accurate results were obtained using the SARIMAX model (1,1,1) (1,1,0,2), which included production as the target variable, and harvested acres and rainfall as exogenous variables. It achieved a Mean Squared Error (MSE) of 13,143.60, Root Mean Squared Error (RMSE) of 114.65, and a Mean Absolute Percentage Error (MAPE) of 6.67%. These metrics show that the seasonal and climatic effects on rice production are accurately and consistently captured. In line with past production trends, the SARIMAX forecasts also showed a steady upward trend.

On the other hand, the Random Forest model, which used rainfall, harvested acres, and sown acres as features, achieved an $R^2$ score of 0.9134, with an RMSE of 210.33, MAE of 160.85, and MAPE of 19.25%. In terms of prediction errors, particularly the MAPE, which is crucial in real-world agricultural forecasting, SARIMAX outperformed the Random Forest model, especially in explaining variance with a high $R^2$, it was less accurate than SARIMAX in terms of prediction error, particularly the MAPE, which is critical in practical agricultural forecasting.

Overall, SARIMAX (1,1,1) (1,1,0,2) provided the most accurate and consistent forecasts, making it the model that performed the best in this investigation. The hybrid model did not produce the anticipated gains, even though Random Forest also showed strong explanatory power. These results imply that statistical models such as SARIMAX can perform better than advanced machine learning models for forecasting rice production with seasonal and climatic influences, particularly when the seasonal structure is dominant in the data.

# REFERENCES

B. K. D. J. R. Samarasinghe, Y. Z. (2025). Climate change and rice production in Sri Lanka.

Bogahawatte, S. a. (2013). *Forecasting of Paddy Production in Sri Lanka: A Time Series Analysis* .

(2019/2020). *Crop Forecast Maha.* Department of Agriculture, Peradeniya.

Dharmasena, D. P. (2010). Traditional Rice Farming in Sri Lanka .

DHINAKARAN SAKTHIPRIYA1, a. T. (2024). Weather based paddy yield prediction using machine learning regression algorithms.

Munasinghe, ,. B. (2023). Time Series Data Analysis on Rice Production in Sri Lanka .

Napagoda, M. a. (2021). Trend Analysis and Forecasting for Paddy Production in Sri Lanka .

Piyal Ekanayake, L. W. (2022). *Development of Crop-Weather Models Using Gaussian Process Regression for the Prediction of Paddy Yield in Sri Lanka* .

Sangarasekara1, A. A. (2024). *Advancing food sustainability: a case study on improving rice yield prediction in Sri Lanka using weather-based, feature-engineered machine learning models.*

Sherin Kularathne, N. M. (2024). Impact of economic indicators on rice: A Machine learninig approach.

Windhya Rankothge, R. (2021). Machine Learning Modelling of the Relationship between Weather and Paddy Yield in Sri Lanka.

Yoshino, E., & Kurniadi, F. I. (2023). Forecasting Rice Production in Indonesia using Regression Techniques: A Comparative Analysis of Support Vector Machine, Linear Regression, and XGBoost Regression.

# APPENDIX

The Python scripts used for data preprocessing, modelling, and evaluation, along with Excel file (Dataset), are available in the following OneDrive link:

Rice Production Prediction in Sri Lanka