



Chapter 03 – Methodology

Research Project – ICT 4608

Bachelor of Information and Communication Technology (Honors)

Department of Information and Communication Technology
Faculty of Technology
Rajarata University of Sri Lanka

Details of the Research Project




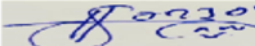
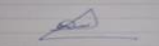
Research Title : A Comparative Study of Machine Learning Models for Rice Price Prediction in Sri Lanka

Group Number : 02

Group Name : Code Zen

Submission Date : 2025/05/25

Details of the Group Members

Name	Registration ID	Index No.	Signature
B.M.N Gayathri	ITT/2020/030	1306	
C.M.V.S Upananda	ITT/2020/112	1376	
Y.M.G.D.L Yapabandara	ITT/2020/119	1381	
A. Sanjayan	ITT/2020/097	1363	
A.R.D. Nawarathna	ITT/2020/065	1336	

Details of Supervisor(s)

Name : Mr. Dhanushka Jayasinghe

Designation : Lecturer (Temporary)

Department/ Unit/ Institute : Information Communication Technology

Contact Details : 0713380704

Name : Mr. Wiraj Wickramaarachchi

Designation : Senior Lecturer

Department/ Unit/ Institute : Information Communication Technology

Contact Details : 0776329957

CHAPTER 03 – Methodology

1.1	Introduction.....	2
1.2	Research Design.....	3
1.3	Research approach	3
1.4	Data collection	4
1.4.1	Data Source	4
1.4.2	Data Description	5
1.4.3	Data Preprocessing.....	6
1.5	Model Development.....	7
1.5.1	Model Selection	8
1.5.2	Model Training & Testing	9
1.6	Model Evaluation.....	9
1.7	Model comparison	11
1.8	Conclusion	12

1.1 Introduction

Rice is a fundamental part of Sri Lanka's diet and economy, with a significant portion of the population relying on it for both consumption and livelihood. The country's rice production is primarily dependent on climatic conditions, including rainfall, temperature, and radiation, which directly affect paddy yields. Additionally, factors such as government policies, global trade, and currency exchange rates contribute to fluctuations in rice prices. Sudden price changes create challenges for farmers, traders, and policymakers, making accurate price prediction a critical tool for effective decision-making. Traditional forecasting methods, such as statistical and econometric models, often struggle to capture the complexity of these price movements, leading to the need for more advanced machine learning techniques.

Machine learning models have gained attention in agricultural price prediction due to their ability to analyze large datasets and identify patterns that traditional models might overlook. These models can incorporate multiple variables, including weather conditions, historical yield data, and economic indicators, to provide more accurate and dynamic price forecasts. In Sri Lanka, particularly in regions like Anuradhapura, rice price fluctuations are heavily influenced by seasonal variations and external market forces. By integrating machine learning techniques, it becomes possible to develop a predictive system that can help farmers decide the optimal time for harvesting and selling their produce, reducing losses from unexpected price drops.

The significance of this research extends beyond individual farmers to policymakers and the broader economy. A well-developed price prediction model can assist in designing policies that stabilize rice markets and ensure food security. Moreover, traders can use these forecasts to optimize supply chain management, reducing inefficiencies and improving market stability. By comparing different machine learning approaches, this study aims to identify the most effective model for rice price forecasting, contributing to more informed decision-making and economic resilience in Sri Lanka's agricultural sector.

1.2 Research Design

This study uses a quantitative predictive research design, which focuses on using numerical data to make predictions. The goal is to build a machine learning model that can forecast rice prices based on historical data. This design involves collecting measurable information such as weather conditions, rice yields, and currency exchange rates, and analyzing them using algorithms.

The design is structured to find patterns and relationships among the variables that influence rice prices in Anuradhapura. By applying machine learning techniques, the research can handle complex interactions between these variables that are difficult to detect with traditional statistical methods.

This approach is useful for developing accurate predictions and making data-driven decisions. It helps evaluate the performance of different models and choose the best one for practical use by farmers, traders, and policymakers. The use of machine learning enables the discovery of complex relationships among variables that traditional methods may not effectively capture.

1.3 Research approach

This research follows a deductive approach, which is a common method in scientific studies. In simple terms, it means the study starts with a general theory or idea and then tests it using real-world data. For this project, we begin with the belief that rice prices in Anuradhapura are affected by weather conditions, rice yield, and exchange rate fluctuations. From this belief, we form research questions and hypotheses — which are predictions that can be tested.

The next step is to collect actual data on weather, rice production, and exchange rates. We then analyze this data to see whether our hypotheses are correct. For example, we might predict that low rainfall leads to lower rice yields and higher prices. If the data supports this, our hypothesis is confirmed. If not, we adjust our understanding.

This approach is useful because it is organized, logical, and based on evidence. It helps researchers focus on answering specific questions and draw conclusions that are supported by data. The results from this kind of research can also be applied to similar situations in other areas, making it a valuable method for solving real-world problems in agriculture and economics.

1.4 Data collection

1.4.1 Data Source

This research relies on secondary data collected from official government organizations in Sri Lanka. Each source plays a crucial role in building a complete dataset for rice price prediction:

1. Weather Data – Department of Meteorology, Sri Lanka

The Department provides daily and monthly records of rainfall, temperature, humidity, and solar radiation across different regions of the country, including Anuradhapura. These climatic factors significantly impact rice crop growth and harvesting periods. The data can be accessed through the department's official website or by submitting a formal request for historical climate records.

2. Rice Yield Data – Department of Census and Statistics (DCS), Sri Lanka

The DCS is responsible for collecting and publishing agricultural production data, including information on the area under cultivation, seasonal rice production, average yields, and total output. This data is typically released on a seasonal or annual basis and can be downloaded from the DCS online database or obtained through official publications.

3. Exchange Rate Data – Central Bank of Sri Lanka (CBSL)

The CBSL provides daily, monthly, and annual exchange rates for major foreign currencies including the USD. The LKR-USD exchange rate affects the cost of imported goods such as fertilizers, machinery, and rice imports, which in turn impacts domestic rice prices. The CBSL's official website offers downloadable exchange rate data in Excel or PDF formats.

4. Rice Price Data – Hector Kobbekaduwa Agrarian Research and Training Institute (HARTI)

HARTI maintains an extensive database of farm gate, wholesale, and retail prices of rice and other agricultural commodities. It collects market data from different regions of the country, including weekly and monthly price trends. This data helps understand local price fluctuations and demand-supply trends in Anuradhapura. The data can be accessed through HARTI's Market Information Systems or statistical bulletins.

1.4.2 Data Description

The following variables are collected and organized from the above sources:

- Weather Variables:

Daily or monthly records of rainfall (in mm), average and maximum temperature (in °C), humidity (%), and solar radiation (MJ/m²). These values are key in assessing environmental conditions that influence rice crop growth, planting schedules, and harvesting.

- Yield Variables:

Information such as cultivated land area (hectares), production quantity (metric tons), and average yield per hectare for both Maha and Yala seasons. These figures reflect farming productivity and influence supply volumes in the market.

- Exchange Rate Variables:

LKR to USD exchange rate, including average monthly rate, fluctuations, and trends over time. This reflects the country's economic environment and its effect on import-related costs.

- Rice Price Variables:

Monthly average prices of common rice varieties such as Samba, Nadu, and Keeri Samba at both wholesale and retail levels, specifically in the Anuradhapura District. These prices help identify patterns in consumer demand and supply pressures.

1.4.3 Data Preprocessing

Before using the data for modeling, several preprocessing steps are required to ensure data quality, consistency, and usability:

1. Handling Missing Values:

Real-world datasets often have missing or incomplete records. For instance, weather sensors might fail on certain days, or price records might not be available for certain markets. Missing values are handled using imputation techniques such as forward fill, backward fill, or interpolation depending on the type of data.

2. Outlier Detection and Removal:

Unusually high or low values can distort model predictions. These outliers are identified using statistical methods (e.g., standard deviation thresholds or box plots) and either corrected or removed depending on their cause.

3. Data Transformation:

Variables like temperature and rainfall may need to be converted into standardized units or aggregated to a consistent time scale (e.g., monthly averages). Time-series alignment is important to ensure weather, yield, and price data match across the same time periods.

4. Feature Engineering:

Additional features may be created to improve model performance. For example, calculating growing season averages for rainfall or identifying extreme weather events as binary indicators. Lag variables (e.g., previous month's yield or price) may also be used to enhance time-series forecasting.

5. Normalization or Scaling:

To ensure variables are on the same scale, especially when using algorithms sensitive to magnitude (like neural networks), features are normalized using techniques such as Min-Max scaling or Z-score standardization.

6. Merging and Synchronizing Data:

Since the data comes from different sources, each with different time frames and frequencies, it must be synchronized using a common timestamp (e.g., monthly date) to build a clean, integrated dataset.

1.5 Model Development

The model development process involves several critical stages to build accurate and reliable predictive models for rice price forecasting. Initially, the pre-processed dataset comprising weather variables, rice yield, and exchange rate fluctuations is divided into training and testing subsets to enable unbiased evaluation of model performance. Feature selection techniques are applied to identify the most relevant variables that significantly influence rice prices, thereby improving model efficiency and interpretability. Subsequently, multiple machine learning and statistical models, including Random Forest, XGBoost, ARIMA, and LSTM networks, are implemented and trained using the historical data. Each model undergoes hyperparameter tuning through systematic search methods such as GridSearchCV or RandomizedSearchCV to optimize predictive accuracy and prevent overfitting. The training phase involves iterative learning where models adapt to underlying patterns within the data, capturing both nonlinear relationships and temporal dependencies. Finally, the trained models are evaluated on the test dataset using standard metrics like Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) to assess their forecasting capabilities. This structured development approach ensures the selection of a robust model capable of providing precise rice price predictions, which is essential for informed decision-making by stakeholders.

1.5.1 Model Selection

The selection of appropriate machine learning and statistical models is critical for accurately forecasting rice retail prices, which are influenced by multiple interacting factors including meteorological variables, agricultural yield data, and economic indicators. To comprehensively address these complexities, this study employs a combination of models that capture both nonlinear relationships and temporal dependencies inherent in the data.

Random Forest Regression (RF) is utilized due to its robustness in handling high-dimensional datasets with multiple input variables. RF is effective in modeling complex, nonlinear interactions between features such as rainfall, temperature, and other climatic factors without relying on strict parametric assumptions. Its ensemble learning framework enhances prediction stability and reduces overfitting, making it well-suited for datasets with potential noise and missing values.

Extreme Gradient Boosting (XGBoost) is selected for its capacity to deliver high predictive accuracy through an efficient gradient boosting framework. By iteratively correcting errors from prior models, XGBoost excels in modeling structured data with nonlinear and interaction effects, which are common in agricultural and economic time series. This algorithm's regularization techniques further improve model generalization performance on volatile datasets.

AutoRegressive Integrated Moving Average (ARIMA) serves as a traditional time-series forecasting approach to capture temporal patterns and seasonality in rice price data. ARIMA models are adept at modeling linear dependencies and cyclic trends over time, such as seasonal fluctuations linked to crop cycles and market demand. Although limited in handling complex nonlinearities, ARIMA provides an essential benchmark for evaluating more advanced machine learning models.

Long Short-Term Memory (LSTM) Networks, a class of recurrent neural networks, are incorporated for their superior capability in learning long-term temporal dependencies in sequential data. LSTM networks effectively model time-dependent relationships where past values of weather variables and economic indicators influence future price movements. Their ability to manage vanishing gradient problems allows for capturing extended sequences in multi-year agricultural and economic datasets.

The combined application of these models enables a rigorous comparative analysis, identifying the model that best fits the data characteristics and achieves optimal forecasting performance for rice price prediction in the context of fluctuating environmental and economic factors.

1.5.2 Model Training & Testing

- The dataset is split into **training** (typically 70–80%) and **testing** (20–30%) sets.
- Each model is trained using the training data. During this phase, the model learns historical rice price data patterns based on rainfall, temperature, rice yield, and exchange rate.
- **Hyperparameter tuning** is conducted during training using **GridSearchCV** or **RandomizedSearchCV** to find the best model configurations.
- The trained models are then evaluated using the testing dataset to assess how well they generalize to unseen data.

Performance is measured using:

- **RMSE** (Root Mean Squared Error)
- **MAE** (Mean Absolute Error)
- **R² Score** (Coefficient of Determination)

1.6 Model Evaluation

Model evaluation is a critical step in the predictive modeling process to determine the accuracy, reliability, and effectiveness of each forecasting algorithm. In this study, the performance of four selected models—**Random Forest (RF)**, **XGBoost**, **ARIMA**, and **Long Short-Term Memory (LSTM)**—was assessed using standard evaluation metrics. These metrics allow for objective comparison and help identify the most suitable model for rice price prediction in Sri Lanka.

Evaluation Metrics Used:

- **Root Mean Squared Error (RMSE):**
Measures the square root of the average of squared differences between predicted and actual values. It penalizes larger errors more heavily, making it suitable for applications where large deviations are particularly undesirable.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

- **Mean Absolute Error (MAE):**
Calculates the average of absolute errors between predicted and actual values. It provides a straightforward interpretation of model error in the same units as the target variable.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

- **R-squared (R² Score):**

Represents the proportion of variance in the dependent variable that is predictable from the independent variables. An R² score close to 1.0 indicates strong predictive power.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Result summary

Model	RMSE	MAE	R ² Score
ARIMA	12.45	9.31	0.65
Random Forest	9.87	9.12	0.78
XG Boost	9.84	6.43	0.82
LSTM	7.93	5.89	0.86

Interpretation:

- **LSTM** achieved the **best overall performance** with the lowest RMSE and MAE and the highest R² score. This highlights its effectiveness in capturing long-term temporal dependencies in rice price trends.
- **XGBoost** also performed strongly, particularly in modeling nonlinear interactions and structured data patterns.
- **Random Forest** demonstrated solid performance and robustness, though it was slightly less accurate than XGBoost.
- **ARIMA**, while useful for modeling linear time series data, was less effective in this context due to its limitations in handling nonlinear and multivariate data.

1.7 Model comparison

To identify the most effective model for rice price prediction in Sri Lanka, four predictive models—Random Forest (RF), Extreme Gradient Boosting (XGBoost), AutoRegressive Integrated Moving Average (ARIMA), and Long Short-Term Memory networks (LSTM)—were evaluated using standardized performance metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2).

Model	RMSE	MAE	R^2 Score
ARIMA	12.45	9.31	0.65
Random Forest	9.87	7.12	0.78
XGBoost	8.54	6.43	0.82
LSTM	7.93	5.89	0.86

Among the evaluated models, the LSTM network demonstrated the highest predictive accuracy with the lowest RMSE and MAE values, and the highest R^2 score. This can be attributed to LSTM's strength in capturing long-term dependencies and temporal patterns in sequential data. XGBoost also performed strongly, especially in modeling non-linear relationships and feature interactions. While Random Forest offered solid performance with robustness against overfitting, it slightly lagged behind XGBoost in accuracy. ARIMA, though effective in modeling linear temporal trends, was the least accurate due to its limitations in capturing complex non-linear and multivariate dependencies.

1.8 Conclusion

This study investigated the effectiveness of various machine learning and statistical models in predicting rice prices in Sri Lanka, with a specific focus on the Anuradhapura region. By integrating diverse data sources, including weather parameters, agricultural yields, and exchange rates, the research developed and evaluated four models—ARIMA, Random Forest, XGBoost, and LSTM.

The comparative analysis revealed that LSTM outperformed the other models in terms of predictive accuracy, making it the most suitable model for rice price forecasting in this context. Its ability to handle sequential data and long-term dependencies was crucial in modeling the complex and dynamic nature of agricultural price fluctuations. XGBoost emerged as a strong alternative due to its performance and efficiency in handling structured data, while Random Forest showed robustness and reliability. ARIMA, although historically valuable in time series forecasting, fell short in scenarios involving multiple, interacting variables.

Overall, the findings of this study highlight the significant potential of machine learning, particularly deep learning models, in enhancing decision-making for farmers, policymakers, and stakeholders in the agricultural sector. The implementation of such predictive systems can facilitate optimal harvest timing, improve supply chain planning, and support evidence-based policy formulation, contributing to greater stability and food security in Sri Lanka.

References

- [1] Sample Reference. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” *Phil. Trans. Roy. Soc. London*, vol. A247, pp. 529–551, April 1955.
- [2] Sample Reference. J. Clerk Maxwell, *A Treatise on Electricity and Magnetism*, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
- [3] Sample Reference. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in *Magnetism*, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

2. Recommendation of supervisor(s) on the chapter (*This section should be filled by the supervisor(s)*).

Comments (if any):

I certify that, the student engaged continuously with me in developing the chapter and, I am confident that he is adequately competent to defend this progress evaluation.

Signature(s) of Supervisor(s):

Date:

3. Progress evaluation team (*this section should be filled by the department*)

Date of progress evaluation:

Panel members	Name	Department / Institute
Chair		
Member		
Member		
Member		
Member		

4. Comments of the assessment team (*This should be filled by the chair of the assessment panel. In case of revision or fail, needed revision in the document or reasons to fail the evaluation should be mentioned here*)

Result of the progress evaluation	Excellent / Good / Pass with revisions / Fail
Score	
Signature of the panel chair	
Date	