palAI & mAIs: Comparative Analysis of Machine Learning Algorithms for Forecasting Palay and Corn Production in Region VI (Western Visayas)

Kzlyr Shaira Manejo, Gliezel Ann Pajarilla, Ma. Christina Kane Vito, and Ara Abigail Ambita

University of the Philippines Visayas
Miagao, Iloilo, Philippines
kmanejo@up.edu.ph, gtpajarilla@up.edu.ph, mbvito@up.edu.ph

Abstract. The agricultural sector continues to be an important component of the Philippine economy, with rice and corn being staple crops for millions. However, climate change and other environmental factors severely affect crop production, posing significant challenges to farmers. This study aims to forecast palay (rice) and corn production in Region VI (Western Visayas) using various machine learning algorithms. The six models considered were trained and tested on historical data, including production volume and area harvested. These six models are Extreme Gradient Boosting (XGBoost), Random Forest, Linear Regression, K-Nearest Neighbor (KNN), Support Vector Regression (SVR), and Artificial Neural Network (ANN). Model performance was measured with metrics such as Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), and Symmetric Mean Absolute Percentage Error (SMAPE) Among the models, XGBoost persistently showed the highest predictive accuracy. It has the lowest Mean Absolute Error of 1217.25 with a Root Mean Squared Error of 2225.08 and a relatively lower Mean Absolute Percentage Error of 14.20% with a Symmetric Mean Absolute Percentage Error of 7.16%. Therefore, the study indicates that XGBoost has the potential to contribute to accurate crop production forecasting in Region VI. These findings will be of great value in agricultural planning, supporting decisions that enhance productivity and sustainability in the region's agricultural sector.

Keywords: Agriculture · crop production · machine learning · time series analysis

[1]

[1] Github Repository: https://github.com/kazeulo/Project_197.git
Dataset Link: https://psa.gov.ph/

1 Introduction

Agriculture has been an important key driver in the Philippine economy as it has served as a major source of livelihood for around 36% of the total employed population working in this sector. Rice is the most important staple food in the Philippines as it provides almost half the calorie requirements of the population, and corn ranks second as a staple crop in the country. A prominent feature of crop farming in the country from the 1960s to 2010s is the dominance of the same five traditional crops: palay, corn, coconut, sugarcane, and banana [4]. Rice supports about 2.5 million households, including 2.1 million farmers, 110,000 workers for post-farm activities, and 320,000 for ancillary activities [25]. Additionally, corn farming is essential for 600,000 farm households and benefits other sectors such as transport services, traders, processors, and agricultural input suppliers who directly gain from the corn industry production, marketing, processing, and distribution activities [8].

The agricultural sector faces several challenges that can undermine efforts in crop production, resulting in food insecurity [3]. According to the World Health Program (2022), one out of ten households in the Philippines, including those relying on agricultural livelihoods, is food insecure [6]. Through the years, farming communities base agricultural decision-making and crop prediction mainly on traditional practices and historical data. However, this approach is becoming increasingly challenging due to changing precipitation patterns and extreme weather events associated with climate change, which have the potential to negatively affect crop production through environmental factors such as heatwaves, droughts, and heavy rainfall [2]. Given these issues, it is essential to enhance our understanding of the requirements for optimal crop growth in farming endeavors.

Addressing the problems within the agricultural sector is crucial, particularly by harnessing practices to improve crop productivity. With continuous technological advancements, various machine learning algorithms can be applied to predict crop production. This project aims to utilize machine learning for time series forecasting, offering a practical solution to predict crop outcomes based on historical data [31]. It specifically focuses on employing machine learning to address crop production prediction challenges in Western Visayas, utilizing widely used methods such as Linear Regression, Support Vector Regression, Random Forest Regression, k-nearest neighbors, XGBoost, and Artificial Neural Networks. Furthermore, the application of time series forecasting—which involves analyzing a collection of time-dependent observations collected at regular intervals (daily, monthly, quarterly, and yearly)—will be an essential component of this project.

This project aims to promote sustainable agricultural practices in Western Visayas by utilizing machine learning to provide accurate crop production forecasts. Analyzing patterns and forecasting yields will empower farmers to adapt to environmental challenges, optimize resource allocation, and ensure stable and resilient agricultural production.

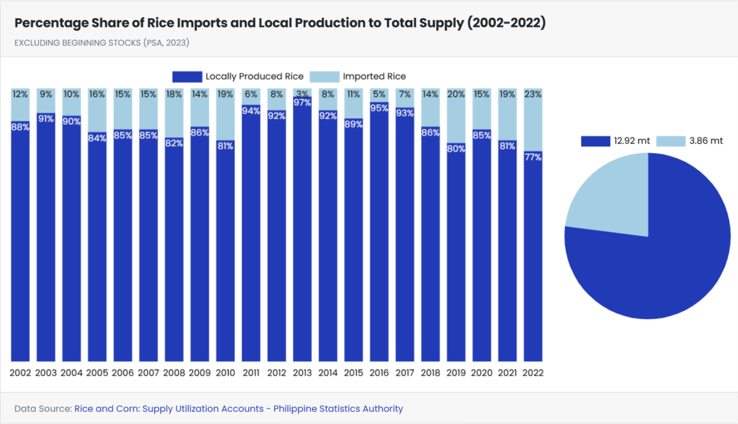
2 Objectives

The primary objective of this study is to evaluate various machine learning algorithms specifically for forecasting the production of palay and corn across different provinces of Western Visayas, Philippines. Specifically, this study aims to:

1. Utilize machine learning models to predict future crop production based on historical data, namely linear regression, support vector regression, random forest regression, k-nearest neighbor, XGBoost, and Artificial Neural Network.
2. Evaluate the accuracy and effectiveness of the different machine learning models in forecasting crop production, using appropriate performance metrics

3 Related Literature Review

3.1 Rice and Corn as Staple Crops


Percentage Share of Rice Imports and Local Production to Total Supply (2002-2022)

EXCLUDING BEGINNING STOCKS (PSA, 2023)

Data Source: Rice and Corn: Supply Utilization Accounts - Philippine Statistics Authority

Fig. 1. Source: `https://www.philrice.gov.ph/ricelytics/consumption`

The Philippines is considered an agricultural country with a large population of Filipinos living in rural areas and supporting themselves through agricultural activities. A 2024 report showed that a quarter of the employed Filipinos work in the agricultural sector comprising four sub-sectors: farming, fisheries, livestock, and forestry [1]. Agriculture is the backbone of the Philippine economy. The main crops usually cultivated for local consumption are rice, corn, and sweet potatoes The Philippines is the world's eighth largest producer with an estimated crop of 19.9 million tonnes of rough rice production in 2023-24 [27].

With the crops cultivated by Filipino farmers, rice and corn are primary crops that are produced in most of the agricultural areas in the Philippine geographic territory [30]. Rice is an important staple in the Philippines and a food constant for millions of Filipinos [21]. Rice contributes approximately 20% to Philippine agriculture's Gross Value Added (GVA). It supports 2.5 million households, including 2.1 million farmers, 110,000 workers for post-farm activities, and 320,000 engaged in ancillary activities [25]. In 2023, According to the Philippines Statistics Authority, the Philippines produced approximately 20.06 million metric tons of unmilled rice, or palay, marking the highest production volume within the recorded period [21].

On the other hand, corn is second to rice as the staple food crop grown in 9.67% of crop area [8]; and comprises several strategic value chains in food, feeds, and multi-industries involving farmers and industry players of various scales (Department of Agriculture, n.d.). The Department of Agriculture reports that approximately 14 million Filipinos favor white corn as their primary staple over rice, while yellow corn is primarily utilized for livestock feed. Additionally, corn is processed into various products such as cornstarch, corn syrup, oil, and gluten. used as the main ingredient for several delicacies [24].

With the numerous benefits of the Philippines being an agricultural country, agricultural sustainability must be ensured to have sufficient supply and affordably-priced food for future generations.

3.2 Crop Yield Prediction Using Machine Learning

Agriculture faces many challenges today, including climate change, soil degradation, water scarcity, pest invasions, and natural disasters. These factors not only threaten food security but also impact the livelihoods of farmers worldwide.

In this context, accurate crop yield prediction has become increasingly essential. Crop yield, defined as the amount of produce harvested per unit area of cultivated land, serves as a key indicator of agricultural productivity citeducker2022. Predicting crop yields enables farmers to effectively plan their strategies, allocate resources, and schedule harvests [31]. For policymakers and agricultural stakeholders, understanding potential yields is vital to ensure food security and availability [15].

Traditionally, crop yield prediction relied on farmers' experiences and historical data. While this method provided a basic understanding of expected yields, it was inherently limited by subjective interpretations and failed to take into account the variability of local conditions.

With the proliferation of advanced technologies, machine learning has emerged as a tool for automating such processes. In 2022, [12] conducted a review of various crop yield prediction methods, specifically examining fruit harvesting data from the past decade. The authors emphasize the shift from labor-intensive manual and traditional techniques to automated yield monitoring systems. The study explores various vegetation indices applicable to different fruit types, environments, and data types and how machine learning models are becoming increasingly popular due to their capacity to increase prediction accuracy. In many instances [23, 26], machine learning has also been proven to outperform conventional statistical methods, particularly in capturing complex relationships within data.

3.3 Related Works

Various studies have demonstrated the effectiveness of machine learning in enhancing crop yield prediction. For instance, [22] employed several machine-learning algorithms for predicting crop yields. The results of their study showed Random Forest Regression significantly outperformed other models, achieving the highest prediction accuracy.

Similarly, in a study published by [20], they proposed several machine learning algorithms, namely; Artificial Neural Network, Support Vector Regression, K-Nearest Neighbor, and Random Forest (RF), to enhance the prediction accuracy for crop yield. The researchers utilized an agricultural dataset which consists of 745 samples, with 70% randomly selected for training the model and the remaining 30% set aside for testing to assess the model's predictive ability. The results show that the Random Forest (RF) algorithm achieves the highest accuracy, as demonstrated by its error analysis metrics, across all distinct feature subsets when trained on the same agricultural data.

In another study by [14], the researchers examined seven widely used machine learning algorithms for three crops: soybean and maize in Iowa and Illinois, and paddy rice in South Korea. They conducted six time-series simulations for each crop, covering different month ranges from 2003 to 2016. The time-series data included a land cover map with geographical and temporal resolutions at the county level, vegetation indices from MODIS, crop yield statistics, and weather information. Among the methods tested, the Support Vector Machine (SVM) achieved the lowest average RMSE, demonstrating the highest accuracy compared to LSTM, CNN, SSAE, DT, RF, and ANN.

Moreover, [31] conducted a comprehensive analysis of studies about machine learning-based crop yield prediction. Among the 50 papers they evaluated, they found that the most used models for crop prediction are random forest, neural networks, linear regression, and gradient-boosting trees. However, there is no specific conclusion of the best model performance in their study.

These studies collectively illustrate the growing importance of machine learning in agriculture, showcasing its potential for crop yield prediction.

4 Methodology

4.1 Procedure Summary

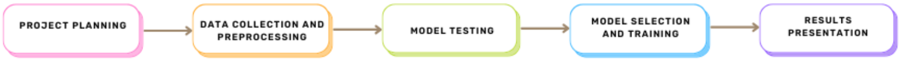

PROJECT PLANNING
DATA COLLECTION AND PREPROCESSING
MODEL TESTING
MODEL SELECTION AND TRAINING
RESULTS PRESENTATION

Fig. 2. The Project Procedure Summary Flowchart

The proposed forecasting of crop production in Region VI utilized machine learning models to predict the volumes of production for palay and corn across different provinces of Western Visayas, Philippines. The project procedure was divided into five phases: Project Planning, Data Collection and Preprocessing, Model Selection and Training, Model Testing, and Results Presentation.

The Project Planning phase involves establishing the scope, objectives, and goals for predicting palay and corn production across different quarters and provinces of Western Visayas. Following this is Data Collection and Preprocessing, involving acquiring the two datasets from the Philippine Statistics Authority (PSA). Furthermore, it involved a series of preprocessing techniques to ensure consistency of information, which included handling missing values, normalization/scaling, and structuring the dataset to prepare them for machine learning operations.

The machine learning algorithms used in this project include SVR, Random Forest Regression, Linear Regression, KNN, XGBoost, and ANN. During the Model selection and Training phase, these algorithms were trained on the prepared dataset that was divided into training and testing sets. Hyperparameter tuning was conducted using GridSearchCV with a 3-fold cross-validation and the performance of each model was evaluated based on the negative mean squared error metric.

During the Model Testing phase, performance metrics such as Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Squared

Error (RMSE), and Symmetric Mean Absolute Percentage Error (SMAPE) were computed to evaluate the accuracy and predictive ability of each model. The predictions of the models were also compared with the actual production values to visually assess their forecasting ability. Finally, in the Results Presentation phase, the performance of the models, significant observations, and findings regarding crop production prediction were presented.

4.2 Dataset Description

For this project, the publicly available crop production dataset from the Philippine Statistics Authority (PSA) that contains historical data on the volume of production for palay and corn in Western Visayas from the year 1987 to 2024 was utilized. The PSA gathers crop data as a part of the Crops Production Survey (CrPS) and the Palay Production Survey, which is a quarterly survey that aims to generate the basic production statistics and data for crops, including the palay and corn in the whole Philippines, that can be segregated into regions and provinces [32].

The dataset was divided into two categories: area harvested and production volume. Each of the datasets includes the variables crop type, province, quarter, and year. These variables were then merged into a single comprehensive dataset. A detailed description of these variables is presented in Table 1. The data is publicly available and can be accessed on the Philippine Statistics Authority website (https://psa.gov.ph/).

Variable Name	Description
Date	Refers to the date the data was collected
Year	Year of data collection from 1987 - 2024 (present)
Quarter	Divided into four quarters in a year
Crop Type	Type of crop being produced: Irrigated Palay, Rainfed Palay, Palay, White Corn, Yellow Corn, Corn
Province	Data broken down into provinces in Western Visayas: Iloilo, Capiz, Guimaras, Aklan, Antique
Area	Refers to the size of the land in hectares
Production	Measured in tons and refers to the volume of crops being produced

Table 1. Description of the Crop Production Dataset

4.3 Statistical Analysis

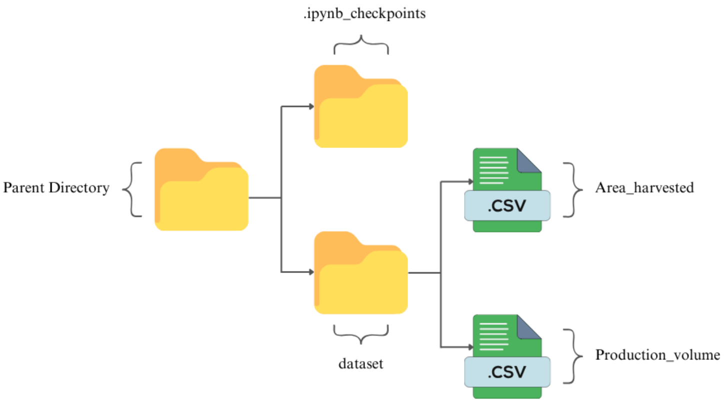


Fig. 3. Dataset Directory Structure in GitHub Repository

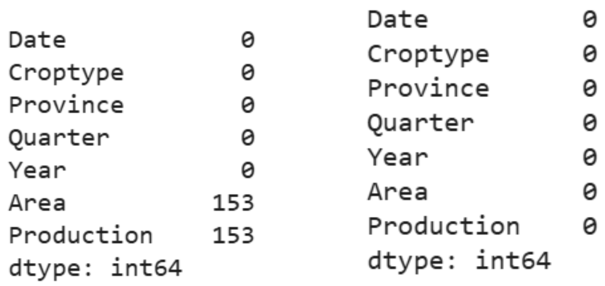
The project was carried out in Google Colab, a cloud-based environment that allows collaborative coding. It provides a Jupyter Notebook interface with access to GPUs, which enables faster model training and evaluation [18]. Libraries such as pandas, numpy, scikit-learn, matplotlib, and seaborn were imported and used for data handling, preprocessing, visualization, and model implementation.

The dataset was uploaded to the GitHub repository, which also provides easy access when working with the Google Colab. To read the dataset within the environment of Colab, the raw URLs of the CVS files on GitHub were passed as arguments to the pd.read_csv() function in pandas. Thus, pandas can directly have access to and download its contents from GitHub and create the pandas DataFrames for further processing and analysis.

4.4 Preprocessing

Data processing is a vital part of machine learning which pertains to the preparation of the raw data collected into a format that is suitable for training and machine learning operations [29]. The preprocessing of the dataset began with loading and merging the production and area harvested data. The datasets were reshaped, merged, and cleaned by adding a 'Date' column, filling in missing values, and removing inaccurate data points.

Data accuracy was improved by removing data points for Guimaras and Iloilo before 1994, reflecting the historical inclusion of Guimaras within Iloilo (see Figure 5). Zero values (see Figure 4(a)) were replaced with NaN (Not a Number), and rows containing NaN values were subsequently dropped (see Figure 4(b)). Outliers in the 'Production' variable were identified and eliminated using the interquartile range (IQR) method.



```
Date            0
Croptype        0
Province        0
Quarter         0
Year            0
Area          153
Production    153
dtype: int64
```

(a) Before

```
Date          0
Croptype      0
Province      0
Quarter       0
Year          0
Area          0
Production    0
dtype: int64
```

(b) After

Fig. 4. Zero Values Before and After Data Cleaning

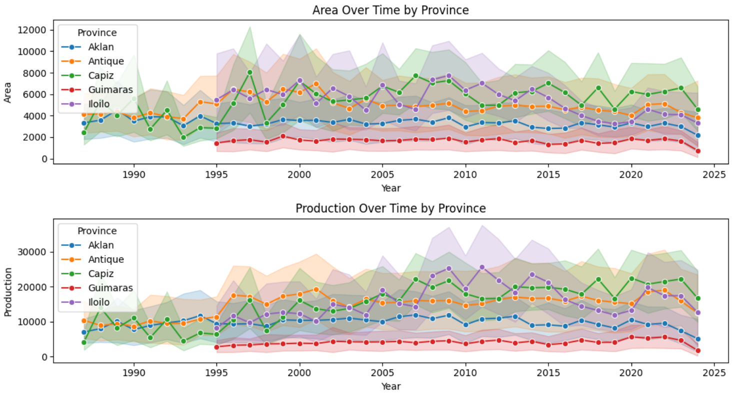


Fig. 5. Dataset Imputation

For feature engineering, categorical variables were transformed using one-hot encoding, and a log-transformed production variable to address skewness. The dataset was sorted by the 'Date' column in ascending order to maintain time-related and chronological ordering. The data was then scaled using StandardScaler, which standardizes data by centering it around zero with a standard deviation of one. This process improves model convergence and performance [28].

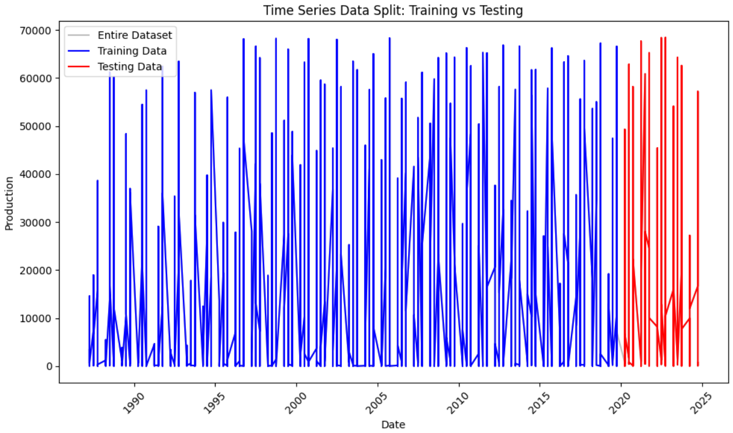


Fig. 6. Splitting the Dataset Into Train and Test

Finally, the dataset was then split into training and testing sets to prepare it for model training and evaluation. The train-test split was performed based on the ‘Date’ column using a specific cutoff date, where the training data includes data points with a ‘Date’ before ‘2020-04-01’, and the testing data includes data points with a ‘Date’ on or after ‘2020-04-01’ (see Figure 6).

4.5 Machine Learning Model Development

This forecasting project investigated the performance of six well-known and widely used machine learning models for regression analysis: Linear Regression, Support Vector Regression, Random Forest, Extreme Gradient Boosting, K-nearest neighbor, and Artificial Neural Network.

Before training the models, the dataset was preprocessed and split into two parts: training data from Quarter 1 of 1984 to Quarter 4 of 2019 and testing data from Quarter 1 of 2020 to Quarter 2 of 2024. Performance without finding the optimal parameters was suboptimal; thus, to improve it, GridSearchCV was applied to fine-tune each model and obtain the best possible results. The performance of the models was then evaluated on the test sets and compared to the actual and predicted values, supported by specific visualizations. The search was performed with 3-fold cross-validation and negative mean squared error as the evaluation metric.

Model	**Hyperparameters**
Artificial Neural Network (ANN)	- none
Extreme Gradient Boosting (XGBoost)	- learning_rate - max_depth - n_estimators - subsample - reg_alpha - reg_lambda
K-nearest Neighbor (KNN)	- n_neighbors - weights - metric
Linear Regression	- none
Random Forest Regression	- n_estimators - max_depth - min_samples_split - min_samples_leaf - max_features - bootstrap
Support Vector Regression (SVR)	- C - epsilon - kernel - gamma

Table 2. Model Hyperparameters

4.6 Evaluation Metrics

The following metrics were used to evaluate the performance of each model objectively:

- *Mean Absolute Error (MAE)*: This metric will help assess the average magnitude of the errors in predictions which can provide a clear understanding of how far off the predictions are from the actual crop production. The formula for MAE is:
$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1}$$
- *Mean Absolute Percentage Error (MAPE)*: This is the measure of the prediction accuracy in comparison to the actual value. It is measured in terms of percentage. It can easily be interpreted as to what the accuracy of the forecasting is. The formula for MAPE is:
$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \times 100 \tag{2}$$

- *Root Mean Square Error (RMSE)*: This metric will provide insight into the magnitude of errors in predictions by emphasizing larger discrepancies which can be particularly informative for understanding model performance in the context of crop production forecasting. The formula for RMSE is:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

- *Symmetric Mean Absolute Percentage Error (SMAPE*: This measure of prediction accuracy addresses limitations of MAPE, particularly with zero or near-zero values in the actual data. The formula for SMAPE is:

$$\text{SMAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{y}_i - y_i|}{(|\hat{y}_i| - |y_i|)/2} \times 100 \tag{4}$$

5 Results and Discussion

5.1 Dataset Exploration

After completion of a series of preprocessing steps, data exploration was performed to characterize the dataset using data visualizations and statistical methods [16]. To effectively examine the dataset, a correlation matrix was performed to determine the relationship of each feature to the target variable, as shown in Figure 2. A correlation matrix is a graphical representation that helps identify which features are most closely related to the target feature. In this visualization, each feature in the dataset is represented by colors corresponding to correlation values, which allows the determination of the strength of the relationship between features [5].

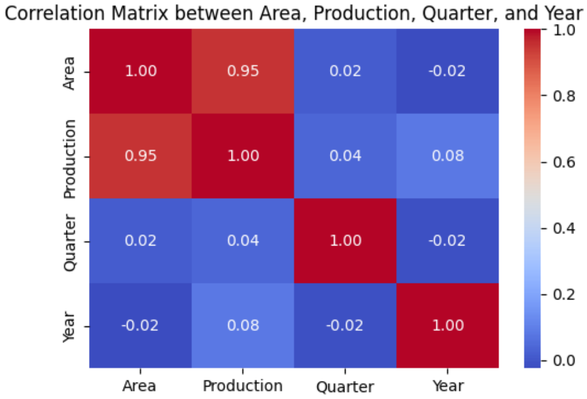


Fig. 7. Heatmap of the Correlation Matrix for the Crop Dataset

The correlation matrix among the variables and their relationship with the target was identified and illustrated in Figure 7. The features that displayed a low positive correlation with each other in the matrix are area and quarter (r = 0.02), indicating feature independence. Additionally, a negative correlation can be observed between area and quarter (r = -0.02), suggesting minimal association. Correlation with the target variable was also examined, where production showed the highest correlation (r = 0.95) with area, indicating that as area increases, production tends to rise. Furthermore, a low positive correlation was observed between production and year (r = 0.08), production and quarter (r = 0.04), indicating a trend that the production slightly increases as both the quarter and year increases.

5.2 Comparison of Model Performance

Table 2 displays the comparison of model performance and shows the results of the machine learning algorithms for time-series forecasting. In this project, Skicit-learn package, which is a library that contains necessary tools for machine learning operations and statistical analysis, was utilized. The models were evaluated based on the value of the Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and Symmetric Mean Absolute Percentage Error (SMAPE).

Model	**RMSE**	**MAPE (%)**	**SMAPE (%)**
XGBoost	2225.081848	14.195011	7.162455
Random Forest	3856.861919	27.171197	11.038082
K-Nearest Neighbor	4271.281769	33.826273	16.806541
ANN	14421.802959	45.868636	19.274818
Support Vector Regression	3666.856666	62.052071	19.274818
Linear Regression	54028.439415	133.043564	36.773868

Table 3. Performance of the Model in Crop Production Prediction

The model that achieved the overall lowest value of RMSE, MAPE, and MAPE is the XGBoost, with an RMSE of 3012.241888, MAPE of 20.562219, and SMAPE of 7.162455. On average, this means that the difference between the actual and the predicted values is approximately 3012.241888 units. Following XGBoost, the second-best model is Random Forest, which has an RMSE of 4096.848756, MAPE of 26.173708, and SMAPE of 11.038082. The third model is the K-Nearest Neighbor, with an RMSE of 4271.281769, MAPE of 33.826273, and SMAPE of 6.806541. The fourth model is the Artificial Neural Networks with an RMSE of 14421.802959, 77.638271 for MAPE, and SMAPE of 19.274818. The fifth model is Support Vector Regression with an RMSE of 3666.856666, MAPE of 62.052071, and SMAPE of 19.274818. Lastly, the lowest-performing model is the Linear Regression with an RMSE of 54028.439415, MAPE of 133.043564, and SMAPE of 36.773868. Based on these results, XGBoost was the best model to forecast crop production since it consistently has the lowest value based on performance metrics.

MAPE Value	SMAPE Value	Predictive Performance Evaluation
<10%	<10%	Highly accurate forecasting
10 - 20%	10 - 20%	Good forecasting
20 - 50%	20 - 50%	Reasonable forecasting
>50%	>50%	Inaccurate forecasting

Table 4. Interpretation of the MAPE and SMAPE evaluation indicators

As displayed in Table 4, only the XGBoost model demonstrated highly accurate forecasting performance with a MAPE of 14.195011 and SMAPE of 7.162455, falling within the range of 10-20%. Followed by the Random Forest with a MAPE of 27.171197 and SMAPE of 11.038082. The K-Nearest Neighbor fell into the range of 20-50%, indicating reasonable forecasting with MAPE of 33.826273 and SMAPE of 16.806541. The ANN shared the same interpretation with Random Forest and KNN with a MAPE of 45.868636 and a SMAPE of 19.274818. However, other models exceeded the range for interpretation, greater than 50%, suggesting that they are not advisable to be used in time-series forecasting because it doesn't result in a reasonable value and predictive power. One of the two models is Support Vector Regression with a MAPE of 62.052071 and a SMAPE of 19.274818. Then, the Linear Regression with a MAPE of 133.043564 and SMAPE of 36.773868.

5.3 Forecasting Results

The figures below show the forecast results of the six models. The plots above display the actual vs predicted values, while the figures below show the predictive errors illustrated in the scatterplot. These visualizations aid in evaluating model performances and highlight how closely the predicted results align with the actual values.

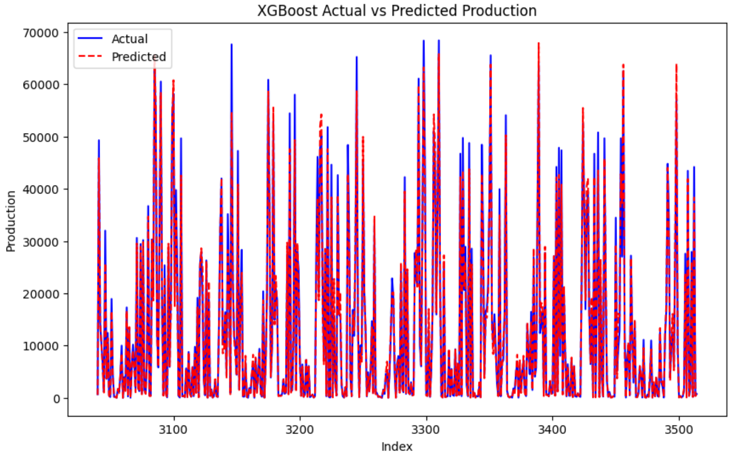


Fig. 8. XGBoost Actual vs. Predicted in the Test Set

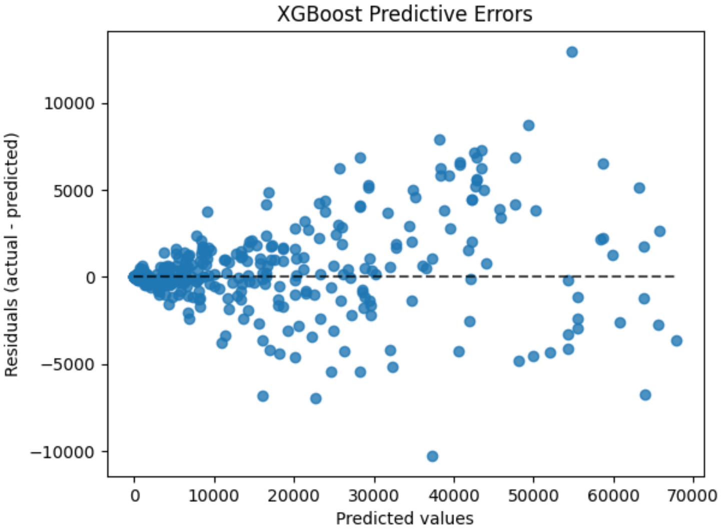


Fig. 9. XGBoost Predictive Errors

The XGBoost model was trained with the following parameters: learning_rate = 0.2, max_depth = 5, n_estimators = 500, reg_alpha = 0.2, reg_lambda = 0.1, and subsample = 0.8. These parameters were chosen to balance the complexity of the model and its ability to generalize, optimizing its performance at predicting.

Figure 8 depicts the actual production values (blue line) and the predicted values produced by the XGBoost model (red dashed line). The predicted values closely track the trend of the actual values, indicating that the model captures the patterns in the data. The close overlap of the two lines signifies that most instances were predicted accurately.

Both actual and predicted values have exhibited strong fluctuations, which align well throughout the index. This means that the model successfully captures variability in the production levels. However, some minor discrepancies are apparent in certain areas, highlighting the potential for further model refinement.

Figure 9 plots the residuals (difference between actual - predicted values) plotted against the predicted values. The residuals exhibit a visible pattern suggesting that most of the points are concentrated around zero on the left, following a funnel shape. This suggests that the model tends to underpredict crop production for many observations.The concentration of residuals on the left and the widening spread for higher predicted values indicate potential weaknesses, such as underfitting for certain values of the target variable.

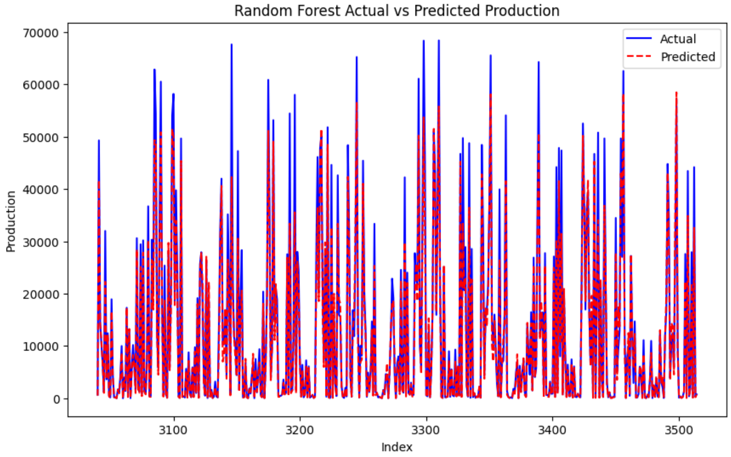


Fig. 10. Random Forest Actual vs. Predicted in the Test Set

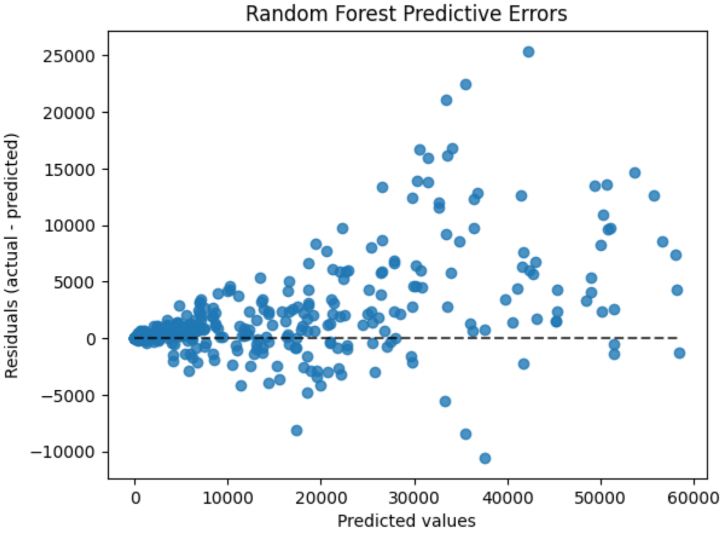


Fig. 11. Random Forest Predictive Errors

The Random Forest Regression model is trained using the following hyperparameters: bootstrap = True, max_depth = None, max_features = 'log2', min_samples_leaf = 1, min_samples_split = 2, and n_estimators = 200. These settings were chosen to ensure the model could create a robust ensemble of decision trees that generalize well across the dataset while minimizing overfitting.

Figure 10 depicts the performance of the Random Forest model, displaying the comparison between actual production values and the predicted values. The close alignment between the predicted and actual production values across the entire test set indicates that the model effectively captures the trend in the production data, demonstrating the strength of the ensemble learning approach to capture the variability in the production levels.

Figure 11 shows residual plots of actual-predicted values against predicted values. Similar to the residual plot of the XGBoost model, the plot also exhibits a funnel shape pattern, where most points are concentrated at 0 on the left and a widening spread to the right can be seen. The presence of the pattern also indicates potential weakness of the model and room for further improvement of the model.

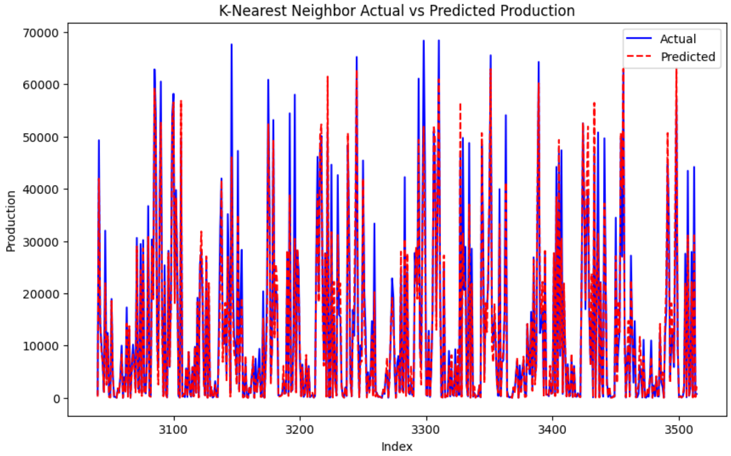


Fig. 12. KNN Actual vs. Predicted in the Test Set

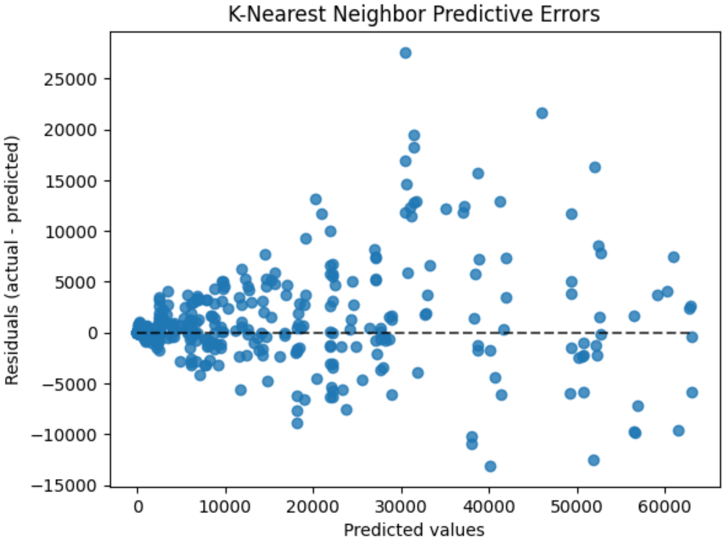


Fig. 13. KNN Predictive Errors

The K-Nearest Neighbors (KNN) Regression model was trained using the optimal parameters through hyperparameter tuning: metric = 'manhattan', n_neighbors = 3, and weights = 'distance'. These settings were selected to enhance the model's ability to predict production values based on giving more importance to closer neighbors and using the Manhattan distance to assess similarity.

In Figure 12, the predicted values align reasonably well with actual values for the moderate production ranges, indicating that the model effectively captures general trends. However, some points deviate significantly, particularly during extreme peaks, suggesting potential limitations of the model under such extreme conditions. Implementing feature scaling or transformations might help address the variance in extreme values. To further improve performance, additional hyperparameter tuning, such as adjusting the number of neighbors or exploring other models, may result in better predictive accuracy.

In Figure 13, a pattern can be observed as it is evident that most residuals are clustered close to zero, indicating that the predictions of the model are quite accurate for many instances. However, residuals spread out at a higher predicted value, forming a funnel shape, suggesting that the model is less effective in predicting these ranges. This could indicate a bias-variance problem, resulting in larger errors for higher production values. This suggests further room for improvement.

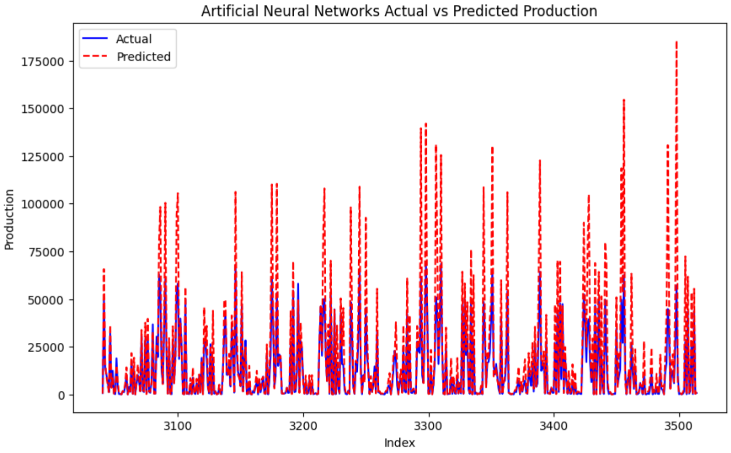


Fig. 14. ANN Actual vs. Predicted in the Test Set

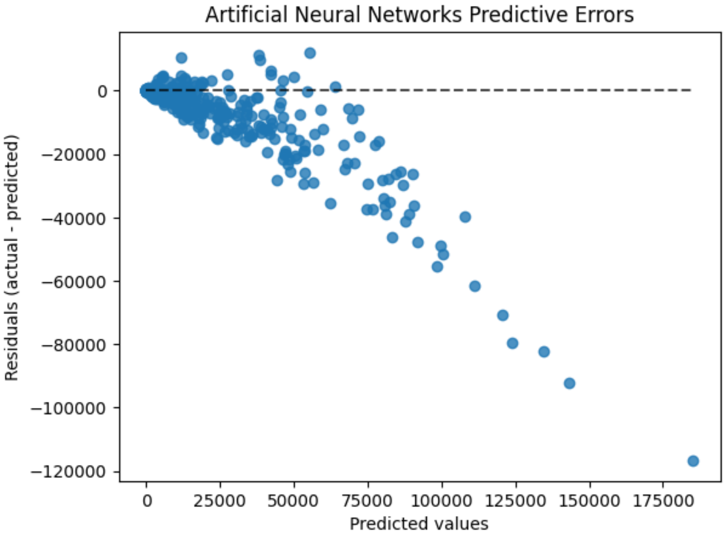


Fig. 15. ANN Predictive Errors

Figure 14 displays significant fluctuations in the actual and predicted production values. Notably, the predicted values often deviate from the actual values, particularly in higher index ranges. This observation suggests that the neural network model struggles to accurately predict the production within the given data range.

Additionally, the downward trend seen in Figure 15, indicates that the residuals become increasingly negative as the predicted values increase. This implies that the model is likely overestimating the actual values at higher predicted levels, and systematically underpredicts the actual values as the predicted values increase.

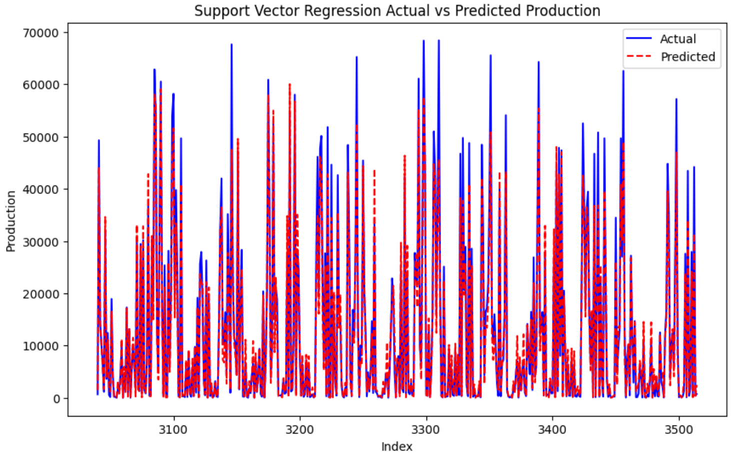


Fig. 16. SVR Actual vs. Predicted in the Test Set

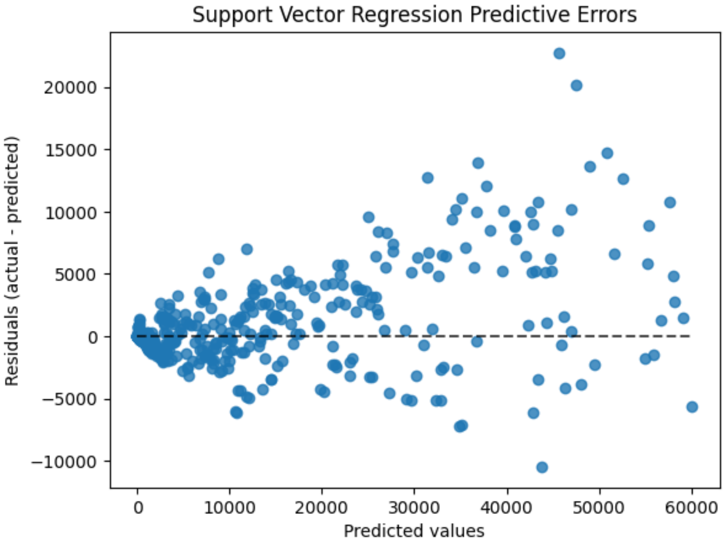


Fig. 17. SVR Predictive Errors

After hyperparameter tuning, the optimal parameters selected for training the Support Vector Regression (SVR) model were as follows: bootstrap = True, max_depth = None, max_features = 'log2', min_samples_leaf = 1, min_samples_split = 2, and n_estimators = 200. These parameters were selected to enhance the model's ability to predict production values by exploiting an ensemble of decision trees, which allows for maximum depth flexibility, logarithmic feature selection, and finely tuned minimum sample splits and leaf sizes.

The bootstrap technique was implemented to improve the model's stability and overall performance.

The predicted values in figure 16 closely align with the actual values, indicating that the model effectively captures the primary patterns in the data. However, noticeable discrepancies can be observed, especially during production spikes, where predictions do not fully reflect these extremes, or when the actual values surpass the predicted values. This may suggest problems in handling extreme values or highly dynamic changes in the production series.

Figure 17. shows a wide spread of residual values, both positive and negative, across the range of predicted values. This spread indicates that the SVR model may not be making highly accurate predictions, as considerable error exists in its outputs. Furthermore, the pattern observed in the scatterplot does not exhibit a clear linear or nonlinear trend, indicating that the model struggles to capture the underlying relationship between the input features and the target variable.

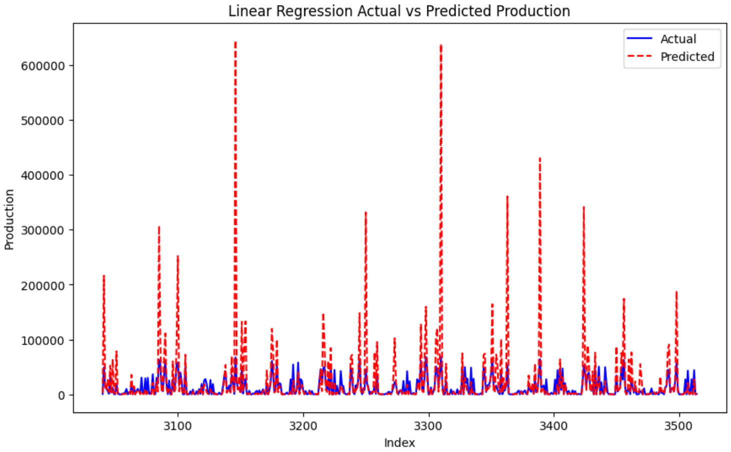


Fig. 18. Linear Regression Actual vs. Predicted in the Test Set

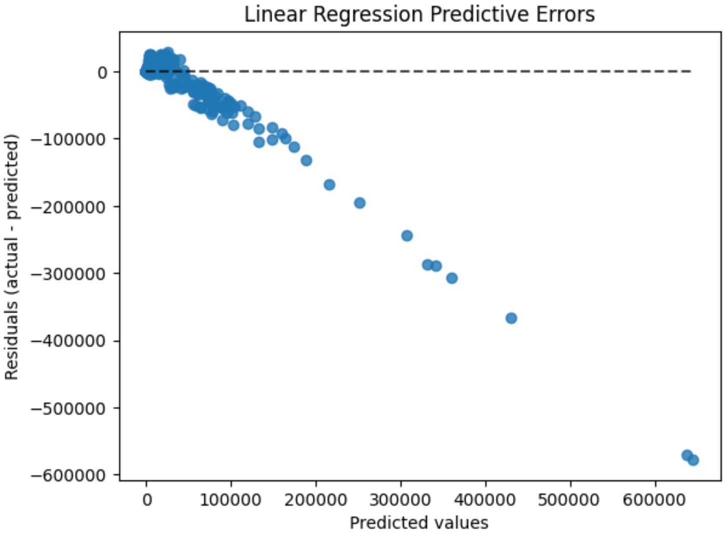


Fig. 19. Linear Regression Predictive Errors

In Figure 18, the actual production values (represented by blue line) and the predicted production values (represented by red line) generally exhibit a similar rising and falling pattern throughout the period. However, there are significant discrepancies between the actual and predicted values at various time points. The predicted values either exceed or fall short of the actual production values at different time points in the plot. Additionally, the actual production values also exhibit greater volatility, with more pronounced fluctuations compared to relatively smoother predicted values. The graph also highlights certain instances where the actual production value deviates significantly from the predictions made by the model.

The plot in Figure 19. trends downward, indicating that as the predicted values increase, the residual errors tend to become more negative. This suggests that the Linear Regression model tends to underestimate the target variable, particularly at higher predicted values. The scatter points are primarily concentrated in the lower-left quadrant of the plot, with a few outliers appearing at the higher end of the predicted values. This pattern indicates that the Linear Regression model fails to make accurate predictions at the higher end of the target variable range.

6 Conclusion

Crop production is crucial for minimizing the risk to the food system in the country. One major area for improvement in the agricultural system in the Philippines is the need for more reliable and up-to-date information on supply, demand, stocks, and export availability. Additionally, there is a national-level need for consistent, accurate, and timely agricultural market data and forecasts. This project aims to compare machine learning models and determine the best model to anticipate supply shortages and surpluses, which can help reduce the risk of food insecurity. By adopting a data-driven approach, this project seeks to improve agricultural forecasting of the Philippine staple foods such as rice and corn.

This paper implemented six widely-used machine learning algorithms to predict palay and corn production in Region VI, Western Visayas, Philippines. The models used were Extreme Gradient Boosting (XGB), Random Forest (RF), Linear Regression (LR), Support Vector Regression (SVR), K-nearest Neighbor (KNN), and Artificial Neural Network (ANN). Among these models, the XGBoost performed the best based on its RMSE, MAPE, and SMAPE values, followed closely by the Random Forest, which ranked second in overall performance.

The results indicate that the performance of the models ranges from reasonable to less than ideal, with the RMSE ranging from a minimum of 2225.09, which is relatively acceptable given the wide range of values in the data, to maximum of 54028.44, which is significantly off from the actual values. To improve the predictive power, it would be advisable to aim for a MAPE value of at least 10%. Moreover, it can also be observed that the residual plots across all the different machine learning models used in this study revealed patterns that suggest potential misfit or underfitting. This occurred despite performing necessary preprocessing and data transformations, including removal of outliers, log transformation, standard scaling, indicating that more robust data preprocessing and transformations are indeed. These issues across all models can be attributed to several factors, some of which are data limitations, residual outliers, and possibility that the data transformation techniques employed were not sufficient to capture the complexities of the data.

Therefore, exploring other models and methods, particularly those tailored for forecasting, such as LSTM, Prophet, and ARIMA, is recommended. Future studies could integrate predictive modeling for specific timelines, such as quarterly forecasts and projections for crop production in years like 2025. Additionally, for better and more meaningful predictions, adding variables such as soil type, weather conditions, irrigation methods, and local agricultural practices and conditions, has a possibility to improve prediction accuracy. Also, study how the accuracy of the model's predictions and analysis can vary depending on the data and variable settings, and determine which approach is superior in all cases. This approach would provide valuable insights; however, this paper primarily focuses on assessing the forecasting ability of various machine learning models for time-series forecasting.

References

1. Balita, C. (2024, March 26). Agriculture in the Philippines. https://www.statista.com/topics/5744/agriculture-industry-in-the-philippines/#topicOverview
2. Baltazar, R. G. (2024). Forecasting the Impact of Climate Change on Rice Crop Yields under RCP4.5 and RCP8.5 Scenarios in Central Luzon, Philippines, Using Machine Learning Algorithms. *International Journal of Agriculture and Natural Resources*, 51(1). https://doi.org/10.7764/ijanr.v51i1.2494
3. Bordey, F., Collado, W., Sandoval, R., & Espenido, R. (n.d.). ASSESSMENT OF CLIMATE CHANGE IMPACTS ON CROP YIELDS IN THE PHILIPPINES. https://openknowledge.fao.org/server/api/core/bitstreams/105ba894-c865-44e6-a635-c2eb383a02b3/content
4. Briones, R. (2021). Philippine agriculture: Current state, challenges, and ways forward (pp. 2021–2033). https://pidswebs.pids.gov.ph/CDN/PUBLICATIONS/pidspn2112.pdf?fbclid=IwAR3RvKz1Ya2J13utihtCI2SqYcj46HrKQbtt2sx1Tyb7IG65elnjz2Gd2ks
5. BUYRUKOĞLU, S., & AKBAŞ, A. (2022). Machine Learning based Early Prediction of Type 2 Diabetes: A New Hybrid Feature Selection Approach using Correlation Matrix with Heatmap and SFS. Balkan Journal of Electrical and Computer Engineering, 10(2), 110–117. https://doi.org/10.17694/bajece.973129
6. Cruz, M. (2022, December 12). WFP Philippines – Food Security Monitoring - October 2022 | World Food Programme. https://www.wfp.org/publications/wfp-philippines-food-security-monitoring-october-2022
7. Department of Agriculture. (n.d.). YELLOW CORN INDUSTRY. https://www.da.gov.ph/wp-content/uploads/2023/05/Philippine-Yellow-Corn-Industry-Roadmap.pdf
8. Department of Agriculture MIMAROPA. (n.d.). Corn Program. https://mimaropa.da.gov.ph/about/programs-and-projects/corn
9. Dogello, J., & Cagasan, U. (2021). A Review on the Status of Crop Production Innovations of the Philippines. *Eurasian Journal of Agricultural Research*, 5(2), 130–136. https://dergipark.org.tr/tr/download/article-file/1886248
10. Ducker, J. (2022, December 13). Importance of Accurate Yield Predictions in Agriculture. https://www.azolifesciences.com/article/Importance-of-Accurate-Yield-Predictions-in-Agriculture.aspx
11. Exconde, O. (n.d.). CORN IN THE PHILIPPINES: ITS PRODUCTION AND RESEARCH ACTIVITIES WITH EMPHASIS ON DOWNY MILDEW. https://www.jircas.go.jp/sites/default/files/publication/tars/tars8-_21-30.pdf
12. He, L., Fang, W., Zhao, G., Wu, Z., Fu, L., Li, R., Majeed, Y., & Dhupia, J. (2022, April). Fruit yield prediction and estimation in orchards: A state-of-the-art comprehensive review for both direct and indirect methods. Computers and Electronics in Agriculture, 195, 106812. https://doi.org/10.1016/j.compag.2022.106812
13. Horie, T., Yajima, M., & Nakagawa, H. (1992). Yield forecasting. *Agricultural Systems*, 40(1-3), 211–236. https://doi.org/10.1016/0308-521x(92)90022-g
14. Ju, S., Lim, H., Ma, J. W., Kim, S., Lee, K., Zhao, S., & Heo, J. (2021). Optimal county-level crop yield prediction using MODIS-based variables and weather data: A comparative study on machine learning models. *Agricultural and Forest Meteorology*, 307, 108530. https://doi.org/10.1016/j.agrformet.2021.108530
15. Khaki, S., & Wang, L. (2019). Crop Yield Prediction Using Deep Neural Networks. *Frontiers in Plant Science*, 10. https://doi.org/10.3389/fpls.2019.00621

16. Konopka, B. M., Lwow, F., Owczarz, M., & Łaczmański, Ł. (2018). Exploratory data analysis of a clinical study group: Development of a procedure for exploring multidimensional data. PLoS ONE, 13(8), e0201950. https://doi.org/10.1371/journal.pone.0201950
17. Lagrazon, G. G., & Tan, J. B. (2023). A Comparative Analysis of the Machine Learning Model for Crop Yield Prediction in Quezon Province, Philippines. https://doi.org/10.1109/csnt57126.2023.10134593
18. Levy, M. (2023, June 23). Getting Started with Google Colab for Deep Learning: A Step-by-Step Guide. Dataquest. https://www.dataquest.io/blog/getting-started-with-google-colab-for-deep-learning/
19. M. Uma Maheswari, & Ramani, R. (2023). A Comparative Study of Agricultural Crop Yield Prediction Using Machine Learning Techniques. https://doi.org/10.1109/icaccs57279.2023.10112854
20. M., G., & R., B. (2019). Performance Evaluation of Best Feature Subsets for Crop Yield Prediction Using Machine Learning Algorithms. *Applied Artificial Intelligence*, 33(7), 621–642. https://doi.org/10.1080/08839514.2019.1592343
21. Mamiit, R. J., Yanagida, J., & Miura, T. (2021). Productivity Hot Spots and Cold Spots: Setting Geographic Priorities for Achieving Food Production Targets. *Frontiers in Sustainable Food Systems*, 5. https://doi.org/10.3389/fsufs.2021.727484
22. Nigam, A., Garg, S., Agrawal, A., & Agrawal, P. (2019). Crop Yield Prediction Using Machine Learning Algorithms. 2019 Fifth International Conference on Image Information Processing (ICIIP). https://doi.org/10.1109/iciip47207.2019.8985951
23. Pantazi, X. E., Moshou, D., Alexandridis, T., Whetton, R. L.,0 & Mouazen, A. M. (2016). Wheat yield prediction using machine learning and advanced sensing techniques. *Computers and Electronics in Agriculture*, 121, 57–65. https://doi.org/10.1016/j.compag.2015.11.018
24. Philippines, N. M. of the, & Philippines, N. M. of the. (2022, September 2). Rice and Corn Week: Filipino Corn Snacks. *National Museum*. https://www.nationalmuseum.gov.ph/2022/09/02/rice-and-corn-week-filipino-corn-snacks/
25. Propper, C., Hardy, L., Howard, B., Flor, R. J., & Singleton, G. (2020). Role of farmer knowledge in agroecosystem science: rice farming and amphibians in the Philippines rice farming and amphibians in the Philippines on JSTOR. *Jstor.org*, 14(2). https://doi.org/10.2307/27316198
26. Rehman, T. U., Mahmud, Md. S., Chang, Y. K., Jin, J., & Shin, J. (2019). Current and future applications of statistical machine learning algorithms for agricultural machine vision systems. *Computers and Electronics in Agriculture*, 156, 585–605. https://doi.org/10.1016/j.compag.2018.12.006
27. Reidy, S. (2023, August 18). Focus on the Philippines | World Grain. https://www.world-grain.com/articles/18922-focus-on-the-philippines
28. Shaibu, S. (2024, October 15). Normalization vs. Standardization: How to Know the Difference. Datacamp.com; DataCamp. https://www.datacamp.com/tutorial/normalization-vs-standardization
29. Singh, D., & Singh, B. (2019). Investigating the impact of data normalization on classification performance. Applied Soft Computing, 97, 105524. https://doi.org/10.1016/j.asoc.2019.105524
30. Urrutia, J. D., Diaz, J. L. B., & Mingo, F. L. T. (2017). Forecasting the Quarterly Production of Rice and Corn in the Philippines: A Time Series Analysis. *Journal*

of Physics: Conference Series, 820, 012007. https://doi.org/10.1088/1742-6596/820/1/012007

31. van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, 177, 105709. https://doi.org/10.1016/j.compag.2020.105709
32. Philippine Statistics Authority. (2023). Philippine Statistics Authority | Republic of the Philippines. Psa.gov.ph. https://psa.gov.ph/