ILKOM Jurnal Ilmiah Vol. 16, No. 2, August 2024, pp. 115-124
Accredited 2nd by RISTEKBRIN No. 200/M/KPT/2020; E-ISSN 2548-7779 | P-ISSN 2087-1716

Research Article **Open Access (CC–BY-SA)**

Prediction and Analysis of Rice Production and Yields Using Ensemble Learning Techniques

Yudha Islami Sulistya[a,1*]; Aina Musdholifah[b,2,]; Chrissandy Sapulete[b,3,]; Elsi Titasari Br Bangun[b,4]; Hizbullah Hamda[b,5]; Sarah Anjani[b,6] Abednego Dwi Septiadi[a,7];

[a] *Department of Software Engineering, Faculty of Informatics, Institut Teknologi Telkom Purwokerto, Indonesia*
[b] *Department of Computer Science and Electronics, FMIPA, Universitas Gadjah Mada, Indonesia*
[1] *yudhaislami@telkomuniversity.ac.id;* [2] *aina_m@ugm.ac.id;* [3] *chrissandysapulete@mail.ugm.ac.id;* [4] *elsititasaribrbangun@mail.ugm.ac.id;* [5] *hizbullah.hamda@mail.ugm.ac.id;* [6] *sarahanjani@mail.ugm.ac.id;* [7] *abednego@telkomuniversity.ac.id*
** Corresponding author*

Article history: Received December 11, 2023; Revised January 28, 2024; Accepted August 04, 2024; Available online August 20, 2024

Abstract

This research focuses on predicting and analyzing rice production and yield throughout the world using ensemble learning techniques. The study applies and compares three methods: linear regression, ARIMA, and ensemble learning, to predict rice harvest yields. The results show that ensemble learning techniques significantly improve prediction performance. For instance, the ensemble model for predicting area harvested, combining Model 6 (linear regression) and Model 10 (ARIMA), achieved 0.9604 of coefficient of determination outperforming the individual models. Similarly, for predicting yield, the ensemble model combining Model 4 (linear regression) and Model 9 (ARIMA) achieved 0.9949 of coefficient of determination indicating superior prediction accuracy. For predicting production, the ensemble model combining Model 2 (linear regression) and Model 8 (ARIMA) achieved 0.9864 of coefficient of determination. These results demonstrate the effectiveness of ensemble learning in enhancing prediction accuracy with lower MSE and RMSE values. By analyzing various factors influencing rice yields, this research provides valuable insights for increasing rice production and yield, supporting efforts to improve the efficiency and effectiveness of rice farming, and contributing to achieving the United Nations Sustainable Development Goals (SDGs).

Keywords: Agricultural Efficiency; ARIMA; Ensemble Learning; Prediction; Rice Production; Rice Yield

Introduction

In the current era of globalization, accurate predictions regarding rice production and yield have important relevance to sustainable development. Referring to the Sustainable Development Goals (SDGs) of the United Nations, the second goal underlines the importance of food security. This study aims to predict and analyze rice production and yield throughout the world using ensemble learning techniques. The research uses three methods: linear regression, Auto-Regressive Integrated Moving Average (ARIMA), and ensemble learning with ensemble selection techniques. These methods were applied and compared to select the most effective model for predicting rice yields.

Accurate predictions of rice production and yield are important in the context of sustainable development. Based on the Sustainable Development Goals (SDGs) of the United Nations, the second goal is "End hunger, achieve food security and improved nutrition, and promote sustainable agriculture". In this regard, efficient and effective research on rice yields is essential to achieving global food security. The research uses a dataset consisting of 21613 rows and 14 columns from crops and livestock products. This data includes various factors that may influence rice yields, and it was used to train and test the prediction model in this study. The ensemble selection method used in this research is a combination of linear regression and ARIMA. Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables [1]. On the other hand, ARIMA is a forecasting model used for time series data that combines autoregression, differencing, and moving average [2].

In research [3] used a linear regression model, two ensemble models (a Random Forest Regressor and a Booster Regressor) to predict solar irradiation. The linear regression method was also used by [4] to predict house prices in Thailand. Another thing was done by [5] using linear regression with several other basic algorithms to predict rainfall based on the climate in India. Apart from that, there is also multiple linear regression (MLR), which is used by [6] to predict basketball winning presentations.

The ARIMA forecasting algorithm is used in research [7] to predict carbon trading prices, [8] to predict crude oil prices, [9] to predict hepatitis E, and [10] for stock price predictions.

https://doi.org/10.33096/ilkom.v1xix.xxx.x-x

From several studies that have been carried out using the linear regression method and the ARIMA model, it is hoped that by combining the two techniques through the selection ensemble method, it can increase the accuracy of predictions and provide new insights to increase rice production and yields throughout the world.

Methodology

This research was conducted using an experimental approach. In this section, we will describe the experimental steps taken and the methods used to achieve the research objectives.

A. Experimental Flow

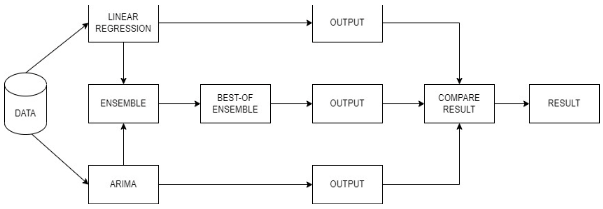

Figure 1. Experimental Flow Chart

In this research, there are several processes that need to be carried out. In **Figure 1**, the flow of this research process is illustrated. The first process involves collecting data from FAOSTAT, an agricultural statistics system developed by the World Food and Agriculture Organization (FAO) (https://www.fao.org/faostat/en/#data/QCL). To ensure data integrity, the dataset was checked for missing values and duplicates. Data preprocessing steps included label encoding for categorical variables, normalization using Min-Max Scaler to adjust values to a common scale, and anomaly detection using the Interquartile Range (IQR) method to identify and remove outlier.

Next, the data is used to make predictions with two methods: linear regression (LR) and ARIMA. These methods are then combined using the best-of-ensemble technique. The performance of the ensemble and the individual models is compared using mean squared error (MSE), root mean squared error (RMSE), and R-squared (R^2) or coefficient of determination metrics. The goal is to achieve the lowest MSE and RMSE and the highest R^2, ensuring the most accurate predictions of rice yields.

B. Dataset Description

The dataset in this research, entitled "Crops and livestock products", is an extensive data set consisting of 21613 rows and 14 columns. This dataset includes various attributes, such as domain code and domain, that provide information about the data category. Geographic attributes are represented by the area code (M49) and area. This dataset also includes Element Code and Element, which refer to specific data elements, as well as Item Code (CPC) and Item, which help in the identification of specific items. The year of data recording is represented by the year code and year.

Each measurement has an assigned unit and value. To provide additional information, there are flags and flag descriptions. The absence of missing data or duplicate data in this dataset ensures the accuracy and reliability of the analysis.

In the context of this research, the most important fields or columns are year, value for harvested area, value for harvest, and value for production. Each of these fields has a key role in the analysis and prediction of rice yields throughout the world.

C. Data Pre-processing

Table 1. Data Set After Remove

No	Area	Element	Year	Unit	Value
1	Afghanistan	Area Harvested	1961	ha	210000
2	Afghanistan	Area Harvested	1962	ha	210000
3	Afghanistan	Area Harvested	1963	ha	210000
…					
21.611	Zimbabwe	Production	2019	tonnes	1134
21.612	Zimbabwe	Production	2020	tonnes	750
21.613	Zimbabwe	Production	2021	tonnes	2908

In **Table 1**, after the deletion process, the dataset remains with five important attributes, namely 'Area', 'Element', 'Year', 'Unit', and 'Value'. The 'Area' attribute refers to the geographic region where the data was collected. 'Element' refers to a specific data element related to production and crop yield. 'Year' indicates the year the data was recorded, while 'Units' specifies the unit of measurement used in the data. Lastly, 'value' describes the value associated with specific items, elements, and areas.

Table 2. Data Set Area Harvested

	Area	Element	Year	Unit	Value
1	Afghanistan	Area Harvested	1961	ha	210000
2	Afghanistan	Area Harvested	1962	ha	210000
3	Afghanistan	Area Harvested	1963	ha	210000
…					
7317	Zimbabwe	Area Harvested	2019	ha	2529
7318	Zimbabwe	Area Harvested	2020	ha	1260
7319	Zimbabwe	Area Harvested	2021	ha	3466

Table 2 shows the simplification process. The "Crops and livestock products" dataset is then further separated based on the 'Element' attribute. As part of this analysis, the dataset was split to form a "Harvested Area Data Set".

This splitting is generated by applying a filter to the 'Element' attribute to select only data related to the 'Harvested Area'. Through this process, the "Harvested Area Data Set" includes 7319 rows of data representing the harvested area in the context of rice production and yield.

Table 3. Data Set Yield

	Area	Element	Year	Unit	Value
1	Afghanistan	Yield	1961	hg/ha	15190
2	Afghanistan	Yield	1962	hg/ha	15190
3	Afghanistan	Yield	1963	hg/ha	15190
…					
6968	Zimbabwe	Yield	2019	hg/ha	15190
6969	Zimbabwe	Yield	2020	hg/ha	15190
6970	Zimbabwe	Yield	2021	hg/ha	15190

In Table 3, after forming the "Harvested Area Data Set", the "Crops and livestock products" dataset is also split based on the 'Element' attribute to form the "Yield Data Set". This separation is achieved by applying a filter to the 'Element' attribute to select only data related to 'Yield'. As a result of this process, the "Yield Data Set" consists of 6970 rows of data that specifically represent yields in the context of rice production.

Table 4. Data Set Production

	Area	Element	Year	Unit	Value
1	Afghanistan	Production	1961	tonnes	319000
2	Afghanistan	Production	1962	tonnes	319000
3	Afghanistan	Production	1963	tonnes	319000
…					
7322	Zimbabwe	Production	2019	tonnes	1134
7323	Zimbabwe	Production	2020	tonnes	750
7324	Zimbabwe	Production	2021	tonnes	2908

Table 4 shows that after forming the "Yield Data Set", the "Crops and livestock products" dataset was continued with a splitting process based on the 'Element' attribute to form the "Production Data Set". This splitting is generated by applying a filter to the 'Element' attribute to select only data related to 'Production'. As a result of this process, the "Production Data Set" includes 7324 rows of data representing production in the context of rice production and yield.

D. Exploratory Data Analysis

Exploratory Data Analysis (EDA) in this research is an important initial stage in the data analysis process. It involves a series of techniques designed to understand the structure, patterns, variability, and anomalies in data.

In the context of this research, EDA is used to understand the features of the "Harvested Area Data Set", "Yield Data Set", and "Production Data Set". This process involves an initial check of the distribution and correlation between various attributes in the dataset. By carrying out EDA, this research gained a better understanding of the nature and structure of the data, which will be further processed using linear regression, ARIMA, and ensemble learning techniques. This allows this research to make more accurate and relevant predictions regarding rice production and yields around the world.

Table 5. Statistik Deskriptif

	Area Harvested (ha)	**Yield (hg/ha)**	**Production (ton)**
Count	7.319	6.970	7.324
Mean	1.483.143,9765	31.037,5858	5.593.586,3722
Stdev	5.674.188,5114	18.380,3937	24.397.166,306
Min	0	968	0
Q1 (25%)	6.461,5	17.092,5	14.092,5
Q2 (50%)	48.400	27.205,5	120.243
Q3 (75%)	327.955,5	41.265,5	981.713,25
Max	46.379.000	113.511	214.429.949,05

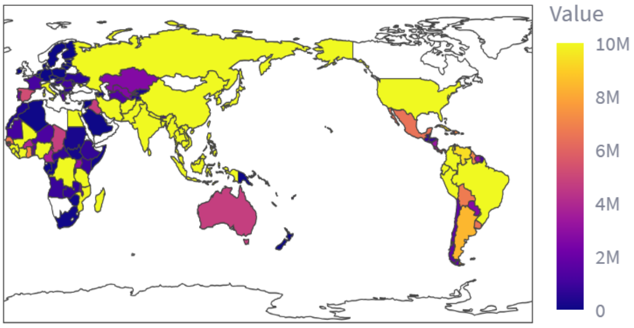

Figure 2. Harvested Area Map Data Visualization

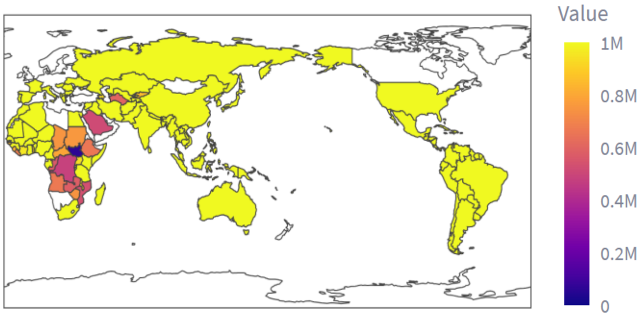

Figure 3. Yield Map Data Visualization

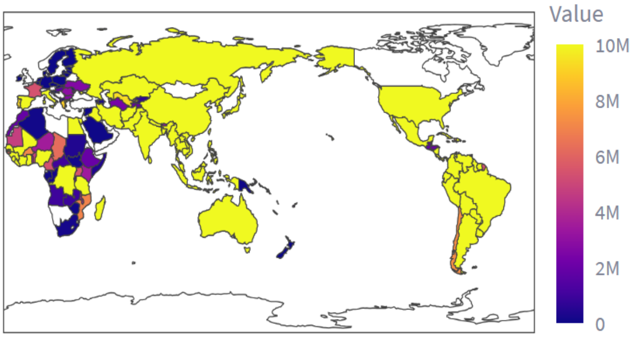

Figure 4. Map Production Data Visualization

Table 5 in this study presents descriptive statistics from the split data, including "Harvested Area Data Set", "Yield Data Set", and "Production Data Set". These descriptive statistics involve calculating various measures of central tendency and dispersion of data, including the number of data entries (count), average (mean), standard deviation (standard deviation), minimum value (min), quartile values (25%, 50%, 75%), and maximum value (max). Through descriptive statistics, this research can explore and understand the distribution of data for each attribute in the three datasets.

Figure 2, 3, and 4 are visualizations in the form of maps depicting the "harvested area,", "yield,", and "production" of rice in various regions. In each visualization, different colors are used to indicate the value of each element. In this case, a brighter yellow color in an area indicates that the value of the related element (be it harvested area, yield, or production) is high in that area.

Figure 2 focuses on "Harvested Areas", showing areas with extensive rice harvests. Then **Figure 3** depicts the distribution of rice "yield,", giving an idea of which regions have high rice yields. Lastly, **Figure 4** shows the distribution of "production,", visualizing which regions have high rice production.

By presenting data in a visual form like this, this research can more easily see and understand patterns and trends in the data, as well as how these elements are distributed across regions. This visualization also supports drawing conclusions and making predictions about rice production and yields throughout the world.

E. Advanced Data Preprocessing

Data preprocessing is an important step in this research, involving a series of procedures aimed at cleaning and transforming raw data into a format more suitable for further analysis and machine learning models. One of the techniques used in this research is label encoding, which converts categorical variables into a numerical form that can be understood by machine learning algorithms. After the label encoding process, the data is normalized using Min-Max Scaler. This normalization adjusts values in the dataset to a common scale, typically between 0 and 1, ensuring that all features have equal influence in the machine learning model and minimizing bias due to differing scales of the columns.

To address potential anomalies in the dataset, the Interquartile Range (IQR) method is employed. IQR involves calculating the range between the first quartile (Q1) and the third quartile (Q3) and identifying any data points that fall below Q1−1.5IQR or above Q3 + 1.5IQR as outliers. These outliers are then removed to ensure the dataset is clean and free from irrelevant data points that could skew the analysis and predictions. This step is crucial for maintaining the integrity and reliability of the dataset, allowing for more accurate model training and prediction.

Next, the data is summarized by year, ranging from 1961 to 2021, to facilitate the analysis of trends over time and highlight potential patterns. Autocorrelation analysis is then carried out to understand the relationship between numerical predictors and numerical targets, determining the correlation between variables at various points in time. This thorough preprocessing ensures that the data is well-prepared for building effective predictive models using ensemble learning techniques, ultimately leading to more reliable and robust predictions of rice yields.

F. Model Selection for Ensemble Learning

Linear regression is chosen for its simplicity and effectiveness in modeling the relationship between dependent variables (rice production and yield) and independent variables (such as area harvested and year). This method is beneficial due to its ease of interpretation, as it provides clear insights into how independent variables impact the dependent variable through coefficients. Additionally, linear regression is computationally efficient, allowing for quick analysis [11]. Despite its simplicity, linear regression often performs well in capturing linear relationships within the data, as demonstrated by the relatively high R-squared values in the study

The ARIMA model is selected for its capability in handling time series data, making it suitable for predicting rice production and yield based on historical trends. ARIMA is designed specifically for time-dependent data, capturing trends, seasonality, and autocorrelations within the data [12]. It is flexible, able to model a wide range of time series data through its components (AR, I, MA), and provides high forecast accuracy, evidenced by low MSE and high R-squared values. Using ensemble learning to combine linear regression and ARIMA leverages the strengths of both models, leading to improved accuracy, robustness, and error reduction. This approach enhances the overall predictive performance, making the predictions more reliable and accurate compared to using each model individually.

G. Linear Regression

The linear regression method is a statistical tool used to identify the influence of one or several variables on other variables. The advantage of using linear regression is its ability to carry out correlation analysis more accurately, because it is difficult to show the level of change in variables relative to other variables (slope) that can be determined.

With regression analysis, forecasting or estimating the value of the dependent variable based on the value of the independent variable becomes more accurate. In addition, this analysis is used to determine whether the direction of the relationship between the dependent variables is positive or negative, as well as to predict the value of the dependent variable if the value of the independent variable increases or decreases.

In linear regression, the dependent variable is the variable that is to be predicted or explained, while the independent variable is the variable that is used to predict or explain the dependent variable. Linear regression tries to find a linear relationship between the independent variable and the dependent variable by using a straight-line equation that best fits the data [13]

The general form of the linear regression model is as in Equation 1.

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{1}$$

where,

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2}$$

H. ARIMA

The ARIMA method completely ignores independent variables in forecasting, so it is suitable for use with statistical data that are interdependent. This method has several assumptions that must be met, such as autocorrelation, trend, or seasonality. In determining the appropriate ARIMA method, it is necessary to test the assumptions, because a method that is good for one case is not always suitable for other cases. The ARIMA method is divided into 4 groups, namely autoregressive (AR), moving average (MA), autoregressive moving average (ARMA), and ARIMA [14], [15].

The AR model assumes that current data is influenced by previous period data. This model is called autoregressive because in it, the variable is predicted based on the previous values of the variable itself. The AR method is used to determine the order value of the p coefficient, which indicates the dependence of a value on the closest previous value [16]. The general form of the AR model with order pARp or ARIMA model $(p, 0,0)$ is stated as Equation 3.

$$AR(p) = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} \tag{3}$$

The MA model is used to determine the q-order coefficient value, which explains the movement of the variable from the previous residual value [17]. The general form of the MA model with order q ($MA(q)$) or ARIMA model $(0,0, q)$ is stated as Equation 4.

$$MA(q) = \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} \tag{4}$$

The ARMA model is a combination of the AR and MA models. In this model, it is assumed that the current period data is influenced by the previous period data as well as the residual values from the previous period. The general form of the ARMA model or ARIMA process $(p, 0, q)$ is expressed as Equation 5.

$$ARMA(p, q) \;=\; AR - MA \tag{5}$$

The ARIMA model assumes that the data used must be stationary, which means the average variation of the data is constant. Data that is not stationary must first be changed to become stationary through a differencing process.

Differentiating means calculating changes or differences from observed values. The difference value obtained is then checked to see whether it is stationary or not. If it is still not stationary, then the differencing process is carried out again. The differencing process can be carried out by subtracting the value of one period from the value of the previous period or by using the Equation 6.

$$I(d) = y_t - y_{t-d} \tag{6}$$

According to research [18], and [19], the ARIMA method is a statistical perspective method that is represented by three parameters. First, the AR process takes data from previous periods and retains it. Then, in the integrated (I) process, the data is changed to make the prediction process easier. The general form of the ARIMA model (p, d, q) is expressed as Equation 7.

$$ARIMA(p, d, q) = AR(p) + I(d) + MA(q) \tag{7}$$

I. Ensemble Learning

Ensemble learning is a machine learning technique that combines a number of learning models to produce better predictions than each model independently. The goal is to improve the accuracy, stability, and robustness of predictions.

There are three main types of ensemble methods, namely bagging [20], boosting [21], dan stacking [22]. In bagging, multiple models are built independently, and their predictions are averaged (for regression) or taken by mode (for classification). In boosting, models are built sequentially, with each new model trying to correct the errors made by the previous model. In stacking, the output of several models is used as input to a model (called a 'meta-learner' or 'second-level learner') that makes a final prediction [16].

Ensemble selection, also known as "best-of-ensemble", is another strategy in ensemble learning. In this approach, instead of combining predictions from multiple models, we select the best model from a set of models. These best models are selected based on their performance in a particular validation set or evaluation metric.

In the context of this research, the selection ensemble method is used to combine the results from linear regression and ARIMA. The aim is to obtain the best model from the two methods, thereby producing the most accurate predictions for rice production and yields.

J. Proposed Models

In this research, 7 models are proposed by using linear regression modelling. Model 1 focuses on using "Area Harvested" to predict "Production". In this model, the area harvested is used as an independent variable to predict the amount of rice production. This model operates on the assumption that the area harvested has a direct impact on the amount of production. Model 2 and Model 3 use "year" as an independent variable to predict "production." Model 2 looks at how rice production changes over time globally, while Model 3 looks at how rice production changes over time considering country-specifics. This model can provide important insights into rice production trends and how they change over time.' Model 4 and Model 5, like Model 2 and Model 3, also focus on using "year" for predictions, but they focus on "year" predictions. Model 4 examines changes in rice yields over time globally, while Model 5 analyzes changes in rice yields over time in a specific country context. Finally, Model 6 and Model 7 use "year" to predict "area harvested." Model 6 explores how harvested area changes over time globally, and Model 7 looks at changes in harvested area in the specific context of certain countries. Each of these models allows this research to understand different aspects of rice production and yield and how they change over time. With this understanding, this research seeks to produce more accurate and relevant predictions to support increased rice production and yields in the future.

The ARIMA model is very useful for data that shows time dependencies, and that is why this method was chosen in the context of this research. We proposed 3 models by using ARIMA namely model 8, model 9, and model 10. Model 8 uses year as an independent variable to predict "Value Area Harvested (ha)". In this model, ARIMA is used to predict the area harvested in hectares based on historical data from year to year. Model 9, like Model 8, also uses year as an independent variable, but its focus is on predicting "value yield (hg/ha)". This model uses ARIMA to forecast yield, or yield per hectare (in hectograms per hectare), based on historical trends. Finally, Model 10 also uses year as an independent variable, with the aim of predicting "value production (tons)". This model uses ARIMA to predict rice production in tons based on historical data. Using the ARIMA model, this research can forecast harvested area, yield, and rice production based on historical trends and understand how these variables might change in the future. This can help in better planning and decision-making in the context of rice production and yields.

Lastly, in this research, the ensemble method is used to combine the strengths of the linear regression and ARIMA methods. Model 11 uses year as the independent variable to predict "Area Harvested (ha)". In this model, the ensemble method is used to combine predictions from linear regression and ARIMA models to predict the area harvested in hectares. Model 12, like Model 11, also uses year as an independent variable, but its focus is on predicting "Yield (hg/ha)". This model uses an ensemble method to predict yield, or results per hectare (in hectograms per hectare), by combining predictions from linear regression and ARIMA models. Finally, Model 13 uses year as an independent variable with the aim of predicting "production (tons)". This model uses an ensemble method to predict rice production in tons by combining predictions from linear regression and ARIMA models.

Results and Discussion

A. Single Model

Table 6. Single Model

Model	Predictor	Target	MSE	RMSE	R^2
Model 1: Linear Regression	Area Harvested	Production	0.0024	0.0493	0.7396
Model 2: Linear Regression	Year	Production	0.0054	0.0736	0.9488
Model 3: Linear Regression (Indonesia)	Year	Production	0.0008	0.0291	0.8366
Model 4: Linear Regression	Year	Yield	0.0006	0.0243	0.9934
Model 5: Linear Regression (Indonesia)	Year	Yield	0.0013	0.0357	0.8795
Model 6: Linear Regression	Year	Area Harvested	0.0151	0.1228	0.862
Model 7: Linear Regression (Indonesia)	Year	Area Harvested	0.0005	0.0214	0.6032
Model 8: ARIMA	Year	Production	0.0038	0.0616	0.9642
Model 9: ARIMA	Year	Yield	0.0008	0.0285	0.9909
Model 10: ARIMA	Year	Area Harvested	0.0116	0.1078	0.8935

Table 6 shows the single models used. In this research, various single models have been tested to predict aspects of rice production such as area harvested, yield, and production. The prediction quality of each model is evaluated based on mean squared error (MSE), root mean squared error (RMSE), and coefficient of determination R^2.

A linear regression model that uses 'Area Harvested' to predict 'Production' (Model 1), 'Year' to predict 'Production' (Model 2), and 'Year' with the 'Indonesia' filter to predict 'Production' (Model 3) shows quite good results, with values R^2 ranging from 0.7396 to 0.9488. Although this model has a relatively high MSE and RMSE, the high R^2 value indicates that the model has a good degree of fit to the data.

The model that uses 'Year' to predict 'Yield' (Model 4) and 'Year' with the 'Indonesia' filter to predict 'Yield' (Model 5) shows very good results with R^2 values 0.87 and RMSE low, indicating that the model fits the data well and makes accurate predictions.

Models that use 'Year' to predict 'Area Harvested' (Model 6) and 'Year' with the filter 'Indonesia' to predict 'Area Harvested' (Model 7) show good results, although the R^2 value for Model 7 is relatively lower.

Meanwhile, the ARIMA model that uses 'Year' to predict 'Production' (Model 8), 'Yield' (Model 9), and 'Area Harvested' (Model 10) also shows very good results with a value of R^2 0.89, indicating good fit to the data and high prediction accuracy.

B. Ensemble Models

Table 7. Ensemble Models Area Harvested

Models	MSE	RMSE	R^2
ARIMA	0.0116	0.1078	0.8935
Linear Regression	0.0151	0.1228	0.862
Ensemble	0.0043	0.0658	0.9604

Table 8. Ensemble Models Yield

Models	MSE	RMSE	R^2
ARIMA	0.0008	0.0285	0.9909
Linear Regression	0.0006	0.0243	0.9934
Ensemble	0.0005	0.0212	0.9949

Table 9. Ensemble Models Production

Model	MSE	RMSE	R^2
ARIMA	0.0038	0.0616	0.9642
Linear Regression	0.0054	0.0736	0.9488
Ensemble	0.0014	0.038	0.9864

In this research, the ensemble method is used to improve the quality of predictions in predicting area harvested, yield, and production. The ensemble method combines predictions from the linear regression model and the ARIMA model using the best-of-ensemble technique.

Model 6 (linear regression) and Model 10 (ARIMA) to predict area harvested. The results of the ensemble model show a significant improvement in performance with lower MSE and RMSE values and a higher R^2 value (0.9604) compared to the single model.

Model 4 (linear regression) and Model 9 (ARIMA) to predict yield Although both single models have shown good results, the ensemble model managed to produce lower MSE and RMSE values and a higher R^2 value (0.9949), indicating better prediction performance.

Model 2 (linear regression) and Model 8 (ARIMA) to predict production Once again, the ensemble model shows improved performance with lower MSE and RMSE values and a higher R^2 value (0.9864) compared to the single model.

In conclusion, ensemble methods consistently show improvements in prediction performance compared to single models. These results confirm the advantages of the ensemble approach in predicting rice production components and demonstrate its potential in supporting efforts to increase rice production and yield.

C. Impact, Limitations, and Future Directions

The findings of this research have significant implications for agricultural practices and policy. By providing more accurate predictions of rice yields, farmers can make better-informed decisions about planting and harvesting, optimizing their use of resources and increasing efficiency. Policymakers can utilize these predictions to develop more effective strategies for food security, ensuring that supply meets demand and reducing the risk of shortages. Furthermore, the improved prediction models can help in the allocation of subsidies and support to regions that need it the most, fostering a more equitable and productive agricultural sector. The insights gained from this study also encourage the adoption of advanced data analytics and machine learning techniques in agriculture, driving innovation and sustainability in the industry.

However, there are several limitations to this research. The dataset, while extensive, may not capture all variables influencing rice production, such as climate change effects, soil conditions, and pest infestations. The reliance on historical data means the models may not always predict future trends accurately due to unforeseen changes in agricultural practices or environmental conditions. Future research should focus on incorporating additional variables and exploring more advanced machine learning techniques, such as deep learning, to improve prediction accuracy. Expanding the dataset to include more recent data and different geographical regions can enhance the model's robustness and generalizability. Collaboration with agronomists and policymakers will be crucial to ensure that the models developed are practical and can be effectively implemented to support agricultural decision-making and policy development.

Conclusion

This research has succeeded in achieving its goal of predicting and analyzing rice production and yield throughout the world using ensemble learning techniques. This study successfully evaluates three prediction approaches, namely linear regression, ARIMA, and ensemble learning, in the context of global rice production and finds that the ensemble learning method offers significant improvements in prediction performance compared to other approaches.

Linear regression and ARIMA methods have been used effectively in this research to model the relationship between rice production factors and crop yields, as well as to predict future trends based on historical data. Although both approaches show good results, the ensemble approach, which combines the strengths of these two methods, succeeds in producing a more accurate and robust prediction model.

This research also shows the importance of accurate modeling and predictions in the context of food security and sustainable development. By using a dataset that includes various factors that may influence rice yields, this research not only succeeded in producing a robust prediction model but also provided valuable insights that can be used to increase rice production and yield.

Based on the results of this study, ensemble methods consistently show improvements in prediction performance compared to single models. In particular, the ensemble of linear regression and ARIMA models for predicting area harvested, yield, and production shows a decrease in MSE and RMSE and an increase in R^2values, indicating an increase in prediction quality.

The conclusion of this research is that the ensemble learning technique is a highly effective approach to predicting rice production and yield. The results show that ensemble methods provide significant improvements in prediction performance compared to single models, with lower MSE and RMSE values and higher R^2 values. These findings are important for agriculture and the achievement of Sustainable Development Goals (SDGs), as accurate predictions can support better resource management and food security strategies. Recommendations based on these findings include the adoption of ensemble learning techniques in agricultural forecasting to enhance prediction accuracy and robustness. Additionally, policymakers should consider these advanced predictive models when developing agricultural policies to ensure efficient and sustainable rice production.

References

[1] M. Lutfi, S. P. Agustin, and I. Nurma Yulita, “LQ45 Stock Price Prediction Using Linear Regression Algorithm, Smo Regression, and Random Forest,” in International Conference on Artificial Intelligence and Big Data Analytics, ICAIBDA 2021, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 267–271. doi: 10.1109/ICAIBDA53487.2021.9689749.

[2] M. Skariah and C. D. Suriyakala, “Forecasting reservoir inflow combining Exponential smoothing, ARIMA, and LSTM models,” Arabian Journal of Geosciences, vol. 15, no. 14, pp. 1–11, Jul. 2022, doi: 10.1007/s12517-022-10564-x.

[3] M. Kamble, S. Ghosh, and P. Patel, “Solar Irradiance Prediction using meteorological data by ensemble models,” in Proceedings of the International Database Engineering and Applications Symposium, IDEAS, 2020.

[4] G. Srirutchataboon, S. Prasertthum, E. Chuangsuwanich, P. N. Pratanwanich, and C. Ratanamahatana, “Stacking Ensemble Learning for Housing Price Prediction: a Case Study in Thailand,” in 13th International Conference Knowledge and Smart Technology, Institute of Electrical and Electronics Engineers Inc., Jan. 2021, pp. 73–77. doi: 10.1109/KST51265.2021.9415771.

[5] P. G. Jaiswal et al., “A Stacking Ensemble Learning Model for Rainfall Prediction based on Indian Climate,” in 6th International Conference on Information Systems and Computer Networks, ISCON 2023, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 1–6. doi: 10.1109/ISCON57294.2023.10112077.

[6] D. Sikka and D. Rajeswari, “Basketball Win Percentage Prediction using Ensemble-based Machine Learning,” in 6th International Conference on Electronics, Communication and Aerospace Technology, ICECA 2022 - Proceedings, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 885–890. doi: 10.1109/ICECA55336.2022.10009313.

[7] F. Meng and R. Dou, “Prophet-LSTM-BP Ensemble Carbon Trading Price Prediction Model,” Comput Econ, 2023, doi: 10.1007/s10614-023-10384-5.

[8] P. Theerthagiri and A. U. Ruby, “Seasonal learning based ARIMA algorithm for prediction of Brent oil Price trends,” Multimed Tools Appl, vol. 82, no. 16, pp. 24485–24504, Jul. 2023, doi: 10.1007/s11042-023-14819-x.

[9] Y. Guo, Y. Feng, F. Qu, L. Zhang, B. Yan, and J. Lv, “Prediction of hepatitis E using machine learning models,” PLoS One, vol. 15, no. 9, pp. 1–12, Sep. 2020, doi: 10.1371/journal.pone.0237750.

[10] A. Durgapal and V. Vimal, “Prediction of Stock Price Using Statistical and Ensemble learning Models: A Comparative Study,” in 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering, UPCON 2021, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/UPCON52273.2021.9667644.

[11] M. Lutfi, S. P. Agustin, and I. Nurma Yulita, “LQ45 Stock Price Prediction Using Linear Regression Algorithm, Smo Regression, and Random Forest,” in International Conference on Artificial Intelligence and Big Data Analytics, ICAIBDA 2021, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 267–271. doi: 10.1109/ICAIBDA53487.2021.9689749.

[12] P. Theerthagiri and A. U. Ruby, “Seasonal learning based ARIMA algorithm for prediction of Brent oil Price trends,” Multimed Tools Appl, vol. 82, no. 16, pp. 24485–24504, Jul. 2023, doi: 10.1007/s11042-023-14819-x.

[13] S. I. Busari and T. K. Samson, “Modelling and forecasting new cases of Covid-19 in Nigeria: Comparison of regression, ARIMA and machine learning models,” Sci Afr, vol. 18, pp. 1–9, Nov. 2022, doi: 10.1016/j.sciaf.2022.e01404.

[14] D. Nanthiya, S. B. Gopal, S. Balakumar, M. Harisankar, and S. P. Midhun, “Gold Price Prediction using ARIMA model,” in 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN) , Institute of Electrical and Electronics Engineers (IEEE), Jun. 2023, pp. 1–6. doi: 10.1109/vitecon58111.2023.10157017.

[15] S. Panigrahi, R. M. Pattanayak, P. K. Sethy, and S. K. Behera, “Forecasting of Sunspot Time Series Using a Hybridization of ARIMA, ETS and SVM Methods,” Sol Phys, vol. 296, no. 1, Jan. 2021, doi: 10.1007/s11207-020-01757-2.

[16] R. K. Jagait, M. N. Fekri, K. Grolinger, and S. Mir, “Load Forecasting Under Concept Drift: Online Ensemble Learning With Recurrent Neural Network and ARIMA,” IEEE Access, vol. 9, pp. 98992–99008, 2021, doi: 10.1109/ACCESS.2021.3095420.

[17] M. Singh, A. K. Jakhar, A. Juneja, and S. Pandey, “Machine learning based framework for cryptocurrency price prediction,” in 3rd International Conference on Secure Cyber Computing and Communications, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 31–36. doi: 10.1109/ICSCCC58608.2023.10176572.

[18] Y. Wang et al., “Estimating the COVID-19 prevalence and mortality using a novel data-driven hybrid model based on ensemble empirical mode decomposition,” Sci Rep, vol. 11, no. 1, pp. 1–17, Dec. 2021, doi: 10.1038/s41598-021-00948-6.

[19] C. H. Chien, A. J. C. Trappey, and C. C. Wang, “ARIMA-AdaBoost hybrid approach for product quality prediction in advanced transformer manufacturing,” Advanced Engineering Informatics, vol. 57, pp. 1–11, Aug. 2023, doi: 10.1016/j.aei.2023.102055.

[20] Y. Gong and P. Zhang, “Commodity Price Analysis and Prediction Based on Ensemble Learning,” in 2nd International Conference on Networking Systems of AI, INSAI 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 199–204. doi: 10.1109/INSAI56792.2022.00045.

[21] I. Sardar, M. A. Akbar, V. Leiva, A. Alsanad, and P. Mishra, “Machine learning and automatic ARIMA/Prophet models-based forecasting of COVID-19: methodology, evaluation, and case study in SAARC countries,” Stochastic Environmental Research and Risk Assessment, vol. 37, no. 1, pp. 345–359, Jan. 2023, doi: 10.1007/s00477-022-02307-x.

[22] A. Swaraj, K. Verma, A. Kaur, G. Singh, A. Kumar, and L. Melo de Sales, “Implementation of stacking based ARIMA model for prediction of Covid-19 cases in India,” J Biomed Inform, vol. 121, pp. 1–11, Sep. 2021, doi: 10.1016/j.jbi.2021.103887.