UDC 330
DOI: 10.15587/1729-4061.2024.315066

# BIG DATA ANALYTICS FOR SEASONAL CROP PATTERNS: INTEGRATING MACHINE LEARNING TECHNIQUES

**Roni Yunis**
*Corresponding author*
Assistant Professor
Department of Information Systems*
E-mail: roni@mikroskil.ac.id

**Arwin Halim**
Assistant Professor
Department of Informatics*

**Irpan Adiputra Pardosi**
Assistant Professor
Department of Informatics*
*Universitas Mikroskil
Thamrin str., 140, Medan, Indonesia, 20212

***This study addresses the challenge of predicting rice growing season lengths, crucial for agricultural planning in tropical regions. Climate variability and season timing create uncertainties in decision-making, and while machine learning is widely used in agriculture, a gap persists in integrating spatial-temporal data for accurate season length prediction and region-specific pattern analysis influenced by rainfall. Using a combination of Random Forest algorithms with hyperparameter optimization (grid search), and clustering techniques such as PCA, K-Means, and Hierarchical Clustering, this study analyzes key features such as the start of the season (SOS), end of the season (EOS), and their significance indicators (sig_sos and sig_eos). The findings reveal a strong correlation (0.98) between SOS and EOS, with an optimal growing season ranging from day 93 to day 207 (113.82 days). The Random Forest model, optimized with Grid Search, achieved a MSE of 28.9474 and an $R^2$ of 0.8636, showing an outstanding predictive result. SHAP and LIME analyses identified sos and eos as the most influential predictors, while cluster analysis highlighted three distinct growing season groups characterized by variations in rainfall and seasonal stability. These results underscore the importance of understanding localized agricultural conditions and provide actionable insights for optimizing planting schedules, resource allocation, and climate adaptation strategies. By integrating advanced machine learning techniques with spatial-temporal data, this study establishes a foundation for improving agricultural resilience and sustainability in the face of climate variability***

***Keywords: seasonal crop patterns, random forest, grid search, SHAP, LIME, cluster analysis, predictive model, climate variability, local agricultural, model accuracy***

*Received 10.10.2024*
*Received in revised form 29.11.2024*
*Accepted 13.12.2024*
*Published 30.12.2024*

***How to Cite:** Yunis, R., Halim, A., Pardosi, I. A. (2024). Big data analytics for seasonal crop patterns: integrating machine learning techniques. Eastern-European Journal of Enterprise Technologies, 6 (4 (132)), 46–56. https://doi.org/10.15587/1729-4061.2024.315066*

## 1. Introduction

Integrating digital innovations into agricultural practices is critical for solving the challenges in the planning and production of food crops. Applying big data analytics and machine learning is key to improving agricultural efficiency and productivity. In the agricultural sector, seasonal patterns significantly influence productivity, especially in the cultivation of rice crops. Climate variability and changes in the length of the growing season create uncertainties that can impact food security [1, 2]. Some studies also suggest that climate change can increase the risk of crop failure and reduce agricultural yields, affecting global food security [3, 4]. This challenge underscores the importance of research aimed at enhancing agricultural resilience and sustainability.

Big data comprises information from various sources, such as weather data, harvest history, and climate conditions. The insights from this data can be used to make better agricultural management decisions [5]. Machine learning can be used to process this big data and identify hidden patterns that can help farmers determine optimal planting and harvesting times, reducing the risk of crop failure due to climate anomalies [6]. Previous studies demonstrate that utilizing machine learning for agricultural prediction and weather patterns improves the precision of crop cultivation planning. However, these studies often lack a localized approach to address specific regional challenges, particularly in tropical countries like Indonesia.

Several studies have identified how machine learning can improve the efficiency and effectiveness of agricultural management. For example, research [7] highlights the impact of machine learning in agriculture, showing that applying this technology can improve agricultural yields and resource optimization. A review of machine learning for crop yield prediction and nitrogen status estimation also found that this technique can provide better accuracy in crop cultivation planning [8].

Moreover, the research [9] highlighted that machine learning could be utilized in agriculture to analyze weather data and help farmers make better planting decisions. This study shows that by leveraging big data and machine learning, farmers can identify weather patterns that affect crop productivity, improving the accuracy of crop cultivation planning in various regions. Nevertheless, there is still a gap in understanding how these technologies can be specifically applied to analyze planting season patterns in regions with diverse growing conditions, such as Indonesia. Addressing this gap is essential to improve resource allocation and adaptation strategies in agriculture.

Therefore, studies that are devoted to integrating machine learning and big data analytics specifically for understanding and improving seasonal planting patterns in regions with diverse agricultural conditions, such as Indonesia, are of significant scientific relevance. These studies are essential not only for enhancing local agricultural productivity but also for developing localized strategies that address climate variability and resource optimization. Research in this area contributes to the broader goal of improving food security, resilience, and sustainability in agriculture, especially in tropical regions where unique challenges exist.

## 2. Literature review and problem statement

Big data analytics and machine learning have emerged as transformative tools in agriculture, especially in understanding and optimizing growing season patterns. These technologies improve the decision-making process, increasing agricultural productivity and sustainability. As agricultural practices evolve, the digital transformation of agriculture has also become a focal point for researchers and practitioners. Climate variability significantly complicates the prediction of growing season patterns, especially in tropical regions that experience multiple growing seasons annually. Research indicates that climate change is leading to alterations in the length and timing of growing seasons. For instance, [10] highlights that rising temperatures are expected to lengthen the growing season in Turkiye, emphasizing the importance of understanding local climate systems and their responses to global climate drivers. However, an aspect that remains unexplored is how these climate projections can be applied to local agricultural practices, particularly in optimizing planting and harvesting schedule for specific crops in regions with diverse climatic conditions. This part has not been studied in depth because large-scale climate models often lack the resolution necessary for region-specific, localized agricultural decision-making, and the integration of such models with real-time, local data on crops and weather conditions presents significant challenges.

Similarly, research [11] discuss how various climatic variables, including growing degree days, influence agricultural land use in the Nordic region, suggesting that a broader range of climatic factors must be considered in modelling agricultural responses. This underscores the necessity for models that can adapt to dynamic weather patterns, particularly in regions with complex seasonal dynamics. However, this aspect has not been studied in depth because the focus has primarily been on broad climatic variables, and incorporating complex local factors requires detailed, granular data as well as more comprehensive modelling approaches that account for regional variations and specific agricultural conditions.

The interpretability of machine learning models poses a significant challenge for policymakers and farmers. While machine learning offers advanced predictive capabilities, its complexity often obscures the understanding of how different variables influence outcomes. Research [12] provide a comprehensive review of machine learning applications in agriculture, noting that while these technologies can enhance decision-making, their opaque nature can hinder effective communication of results to stakeholders. Furthermore, [13] emphasizes the importance of understanding the variables affecting crop yield predictions, as this knowledge is crucial for farmers to make informed decisions. The lack of interpretability in machine learning models can thus limit their practical application in agricultural management, necessitating the development of more transparent models that can elucidate the effects of individual variables. While both studies focus on improving predictive accuracy using machine learning (e.g., decision trees, random forests, and deep learning), they primarily rely on historical and general datasets. The lack of exploration into localized, real-time data, such as rainfall or region-specific crop characteristics, may limit the models' applicability in dynamic agricultural environments.

However, the complexity of agricultural data, such as climate variability, soil conditions, and historical crop yields, process significant challenges for effective analysis and interpretation. Therefore, robust models need to be developed to accurately predict growing season patterns and provide actionable insights to farmers or decision-makers [12, 14]. While predicting the length of the growing season is vital for efficient resource allocation, regional clustering is equally important for identifying areas with similar growing season patterns. This approach can optimize agricultural resource management by tailoring strategies to specific regional needs. For instance, the work of [15] on climate change projections in Mozambique highlights the importance of localized adaptation strategies to enhance agricultural resilience. These projections, combined with regional clustering, can identify areas most vulnerable to climate change and its impacts on crop production, enabling more targeted interventions. By aligning resource allocation with the specific climatic and agricultural characteristics of different regions, farmers and decision-makers can optimize strategies for pest control, irrigation, crop selection, and overall land use. This aspect of regional clustering and localized adaptation has not been explored in depth because previous studies often focus on broad trends and general datasets, which fail to capture the regional variations and specific local conditions that are critical for precise agricultural management.

These studies collectively support the necessity of applying big data analytics to seasonal crop patterns by integrating machine learning techniques. Leveraging temporal and spatial data through advanced analytics can address the challenges of heterogeneity, scalability, and region-specific modeling. This approach not only builds on the strengths of prior research but also establishes a robust framework for understanding and optimizing seasonal crop patterns, making it a vital direction for future agricultural innovation. Research [13] illustrates the essential role of big data analytics in understanding and managing seasonal crop patterns. Without the adoption of such technologies, crop losses may increase significantly, emphasizing the necessity of using big data for optimizing agricultural productivity. Similarly, [9] demonstrates how digital technologies, particularly data analysis, facilitate improved crop management by enabling more informed planting and harvesting decisions. Machine learning, as a complementary tool, plays a pivotal role in analyzing large datasets, identifying patterns, and predicting outcomes. These technologies have been successfully applied to enhance crop yield prediction, disease detection, soil management, and irrigation optimization. However, while these studies highlight the importance of big data and machine learning, they often rely on large-scale data and do not fully integrate high-resolution, localized data, such as season length, rainfall, and region-specific agricultural practices. Addressing this gap is crucial for developing models that can adapt to the complexities of regional and seasonal variations, ensuring more precise and actionable insights for local farmers and decision-makers.

All these findings suggest that further studies on scalable, region-specific frameworks for integrating big data analytics with advanced machine learning techniques are essential. By leveraging temporal and spatial data, such frameworks can address challenges related to heterogeneity, scalability, and localization, providing a robust basis for understanding and optimizing seasonal crop patterns. This direction holds significant promise for the future of sustainable, data-driven agricultural innovation. Therefore, it is crucial to conduct research that integrates big data analytics with machine learning to tackle the complexities of agricultural seasonality, optimize resource allocation, and offer actionable insights for farmers, ultimately advancing sustainable agricultural practices.

## 3. The aim and objectives of the study

The aim of the study is to analyze big data with a machine learning approach for food crop prediction, focusing on the prediction and clustering of the length of the growing season of rice crops.

To achieve this aim, the following objectives are accomplished:

– to provide long-term predictions of the growing season to help farmers optimize planting and harvesting schedules based on weather and environmental conditions, and provide an interpretation of the variables that can be explained;

– to cluster areas based on the length of the growing season to support the efficient distribution of resources, such as fertilizers;

– to find changes in the length of the growing season and provide insight into the impact of climate change, especially rainfall, on the planting cycle.

## 4. Materials and methods

### 4. 1. Object and hypothesis of study

The object of this study is the seasonal crop patterns of rice in Asia, focusing on determining the season length (season_length) based on the start of the season (SOS) and end of the season (EOS). This study also analyzes spatial and temporal variations influenced by rainfall and regional characteristics, providing insights into the dynamic agricultural environments of Asian countries. The hypothesis of this study is that the integration of machine learning techniques with spatial-temporal data can improve the accuracy of predicting growing season patterns and clustering regions based on rainfall characteristics. This approach aims to provide actionable insights for agricultural planning and resource allocation.

### 4. 2. Simplifications and assumptions

This study adopts several simplifications to address the complexity of the analysis:

1. The analysis is restricted to the temporal range of 2003–2016, as determined by the availability of RICA data and rainfall data from Google Earth Engines.

2. The model is trained using data from other Asian countries and validated using data from Indonesia to evaluate generalizability across diverse conditions.

3. Principal component analysis (PCA) is applied to reduce the correlation among numerical variables, thereby accelerating the clustering process.

For clarity, the following abbreviations are used in this section: SOS – start of the season; EOS – end of the season; RICA – rice crop calendar Asia; MODIS – moderate resolution imaging spectroradiometer; PCA – principal component analysis; RF – random forest; SHAP – Shapley additive explanations; LIME – local interpretable model-agnostic explanations; MAPE – mean absolute percentage error; MSE – mean squared error.

### 4. 3. Research method

From the review of existing studies, challenges in integrating machine learning with spatial-temporal data for agricultural analysis have been identified. These include handling data heterogeneity, developing scalable predictive models, and providing localized insights for decision-making. This section outlines the methods adopted in this study to address these challenges.

The Fig. 1 illustrates a research method that involved several stages: data collection from the RICA and GEE rainfall dataset, data processing, feature processing, prediction and clustering, hyperparameter tuning, explainable model, and evaluation&validation, all contributing to a comprehensive machine learning workflow.

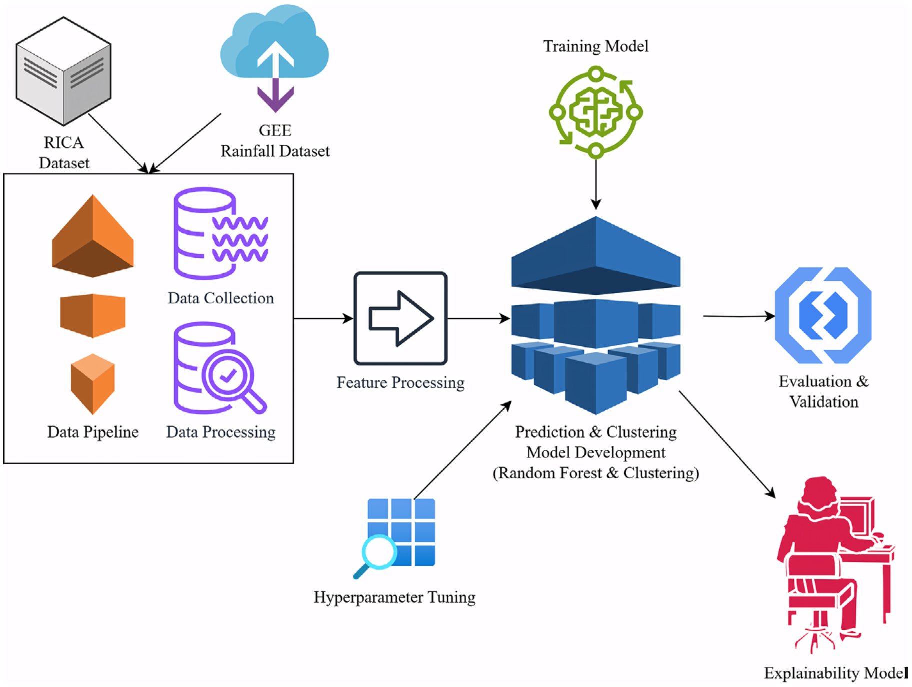


Fig. 1. Research method

The stages can be explained as follows:

1. Data collection.

Rice Crop Calendar Asia (RICA) [16] and rainfall data were taken directly from Google Earth Engines and CMIP6 climate projections [17]. RICA data is only available from 2003 to 2016. Therefore, the rainfall data in Table 1 was adjusted to that time range. This dataset was used in the spatial analysis of hyper-temporal remote sensing information from

MODIS to estimate the start of the season (SOS) and end of the season (EOS) dates for rice crops in Asia. The geometry data and tabular data from the RICA dataset were integrated in this study.

Table 1

Rainfall dataset

| Rainfall data | Shape_Area | Shape_Leng | Mean |
|---|---|---|---|
| 1 | 2.933817 | 30.535271 | 0.000032 |
| 2 | 1.192200 | 16.561909 | 0.000130 |
| 3 | 3.462981 | 23.891844 | 0.000175 |
| ... | ... | ... | ... |
| 28 | 7.285286 | 29.507477 | 0.000156 |
| 29 | 1.375047 | 9.471543 | 0.000107 |
| 30 | 3.721491 | 22.685675 | 0.000079 |

2. Feature processing.

The key feature to be analyzed in this study is the length_season. This feature was calculated by the following formula:

$$\text{length\_season} = \text{EOS} - \text{SOS}, \tag{1}$$

length_season is the length of the growing season, EOS is the end of the growing season, and SOS is the beginning of the growing season. The numerical features in the RICA dataset are shown in Table 2.

Table 2

Numeric features in the RICA dataset

| RICA | pr_season | SOS | EOS | sos_sig | eos_sig | season_length |
|---|---|---|---|---|---|---|
| 1 | 2 | 90.76 | 204.62 | 76.56 | 67.65 | 113.86 |
| 2 | 3 | 229.38 | 346.18 | 157.25 | 165.17 | 116.80 |
| 3 | 3 | 187.97 | 305.15 | 160.61 | 170.87 | 117.18 |
| ... | ... | ... | ... | ... | ... | ... |
| 1262 | 3 | 153.63 | 247.71 | 34.00 | 37.59 | 94.08 |
| 1263 | 3 | 150.14 | 242.59 | 36.42 | 40.11 | 92.45 |
| 1264 | 3 | 186.04 | 309.14 | 8.68 | 28.95 | 123.10 |

3. Prediction and clustering model development.

This study used the random forest (RF) algorithm for the prediction model with cross-country validation to train the model with other Asian countries and evaluate the results with Indonesia's. This approach will provide an idea of whether the model is generalizable in various conditions or if it overfits with data other than test data. The RF model (2) in this study used a regression approach with the following formula:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(x), \tag{2}$$

$B$ is the number of trees, and $T_b(x)$ is the prediction from tree $b$ for input $x$.

The researchers used the hyperparameter tuning grid search and random search to improve the performance of predictive models. Both parameters were compared, and the best result was used to build the predictive model. The RF model was used in this research because it can effectively handle non-linear data. It can reduce overfitting through ensemble learning mechanisms and improves prediction accuracy by the bagging method that links results from various decision trees for better stability and accuracy [18, 19].

Next, this study used the PCA, K-Means (3), and Hierarchical Clustering (4) for the clustering of length_season pattern. The PCA was conducted to eliminate the correlation between variables and accelerate the clustering process [20]:

$$J = \sum_{k=1}^{K}\sum_{x_i \in C_k} \left\| x_i - \mu_k \right\|^2, \tag{3}$$

$$\mu_k = \frac{1}{|C_k|}\sum_{x_i \in C_k} x_i. \tag{4}$$

where the first centroid for the K-means was chosen randomly and the distance of each data $x_i$ to all centroids were calculated and allocated to the nearest cluster (5). This procedure was repeated until there were no significant changes:

$$d_{(x_i, x_j)} = \sqrt{\sum_{k=1}^{p}\left(x_{ik} - x_{jk}\right)^2}, \tag{5}$$

$$d\left(C_i, C_j\right) = \sum_{x \in C_i \cup C_i} \left\| x - \mu_{C_i \cup C_j} \right\|^2. \tag{6}$$

The above formula calculates the distance between samples. It combines two clusters and minimizes the total variation within the cluster (6), where $C_i$ and $C_j$ are two clusters, and μ is the combined cluster centroid.

4. Model explainability.

The SHAP (Shapley Additive Explanations) approach provide a feature contribution value that can later help understand the impact of features on targets in the model (7). Meanwhile, LIME (Local Interpretable Model-agnostic Explanations) provide a local explanation (8) for how the model makes decisions for one specific instance in the model [21, 22]:

$$\varnothing_i = \sum_{S \subseteq 1,\ldots n \backslash i} \frac{\backslash midS|!\left(n - \backslash midS| - 1\right)!}{n!}\left[f\left(S \cup i\right) - f\left(S\right)\right], \tag{7}$$

$$\hat{f} = \arg\frac{\min}{g \backslash inG}\sum_{z' \backslash inZ} \omega\left(x, x'\right)\left(f\left(z'\right) - g\left(z'\right)\right)^2 + \Omega\left(g\right). \tag{8}$$

5. Evaluation and validation.

The prediction model was evaluated using the MSE, R-squared, and MAPE. These tests aim to provide a comprehensive overview of model error, interpretability of data variations, and a better understanding of the model's performance in the context of percentage error [23].

## 5. Result big data analytics in seasonal crop patterns

### 5. 1. Growing season's length analysis

Features such as SOS, EOS, sig_sos, and sig_eos were used in predicting length_season. Fig. 2 presents the correlation heatmap, showing a strong relationship (0.98) between SOS and EOS. The length of the season was calculated as the difference between the SOS and EOS dates. The country-level analysis of length_season in Asia is visualized in Fig. 3, and monthly trends of SOS and EOS are shown in Fig. 4.

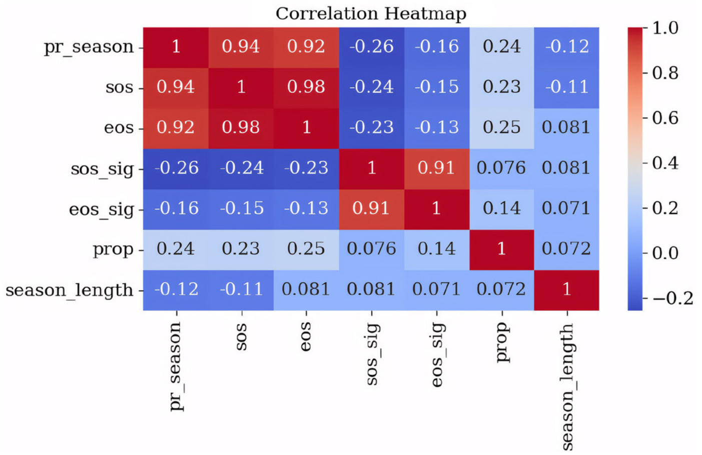


Fig. 2. Correlation heatmap

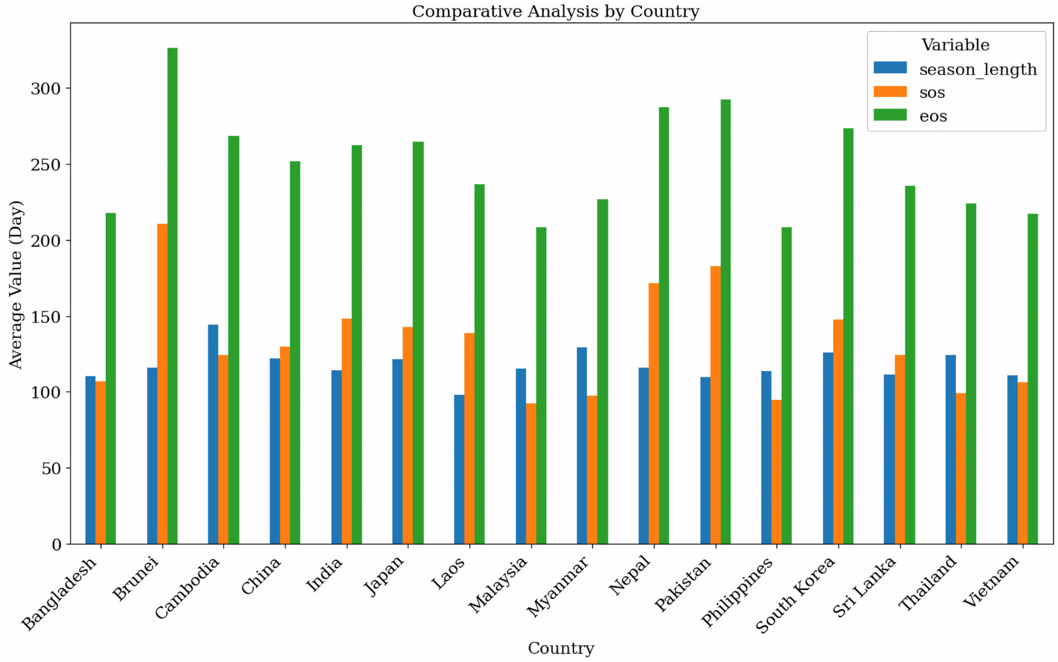
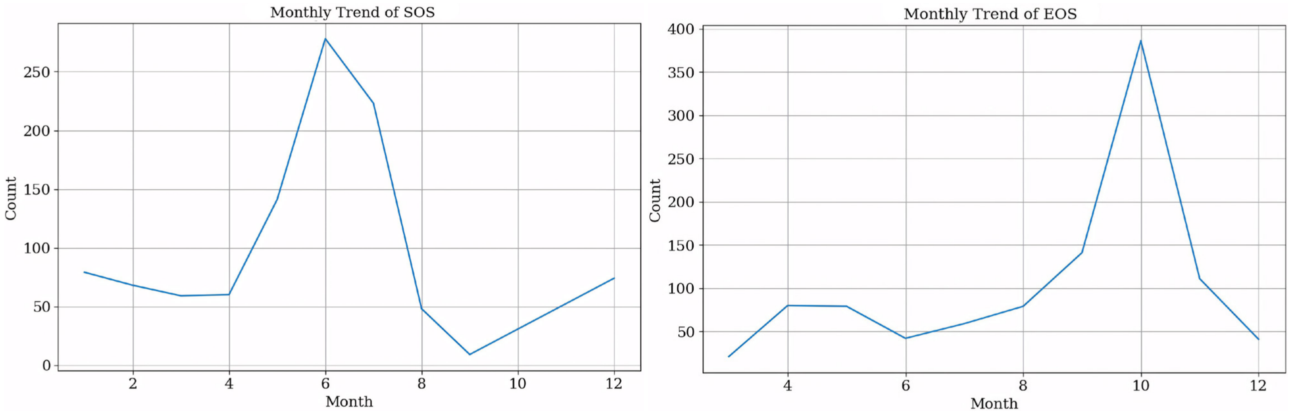


Fig. 3. Comparative analysis of season length



*a* *b*

Fig. 4. Monthly trends: *a* – start of season (SOS); *b* – end of season (EOS)

The season_length analysis results for the Indonesian dataset are shown in Fig. 5. This analysis was used in the cross-country validation experiments to identify the season length in the test area.

The following regions in Indonesia (red zone) have the longest growing seasons: Aceh (161.86 days), South Sulawesi (154.40 days), North Sumatera (151.20 days), East Nusa Tenggara (150.16 days), and Papua (148.14 days).

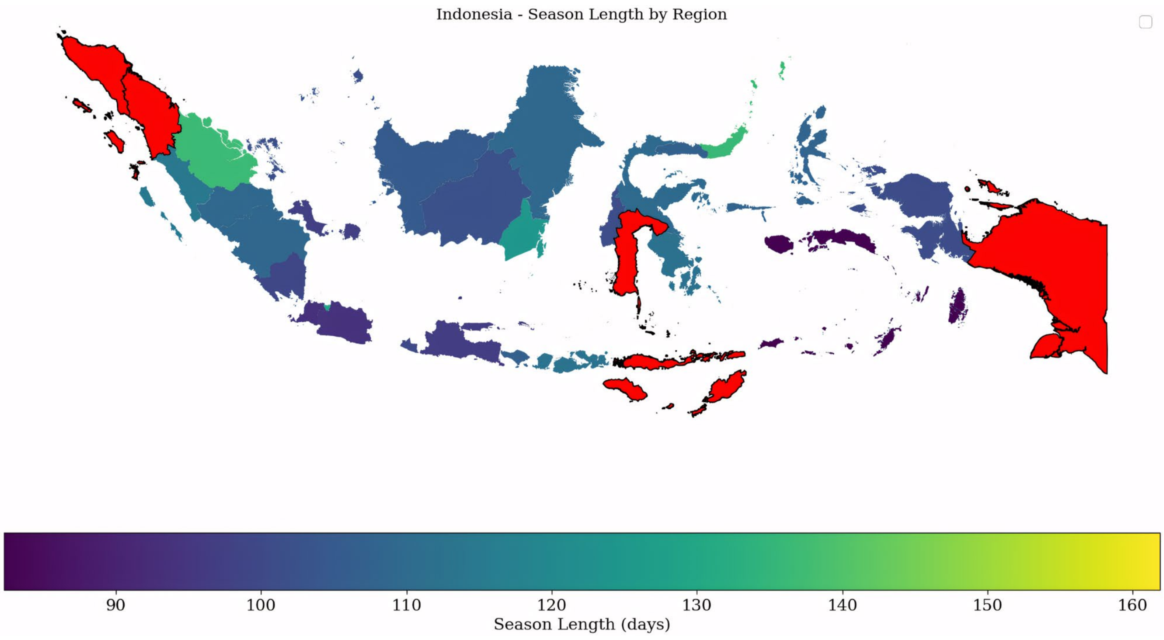


Fig. 5. Season length by region in Indonesia

### 5. 2. Prediction model development with hyperparameter tuning

Hyperparameter tuning was conducted using two approaches (the Grid Search and the Random Search) to optimize the RF model's performance. The Grid Search performs a thorough search by testing all specified combinations of parameters, while Random Search selects a random combination of parameters within a given search space. Table 3 summarizes the hyperparameter tuning results, showing that the RF model with grid search achieved a lower mean squared error (MSE=28.9474) and higher R-squared value ($R^2$=0.8636) compared to random search (MSE=36.9231, $R^2$=0.8260).

Comparative prediction results for both approaches are shown in Fig. 6. Cross-country validation using the Indonesian dataset resulted in an $R^2$ value of 0.7819, MAPE of 5.6579 %, and MSE of 75.6734, as presented in Fig. 7.

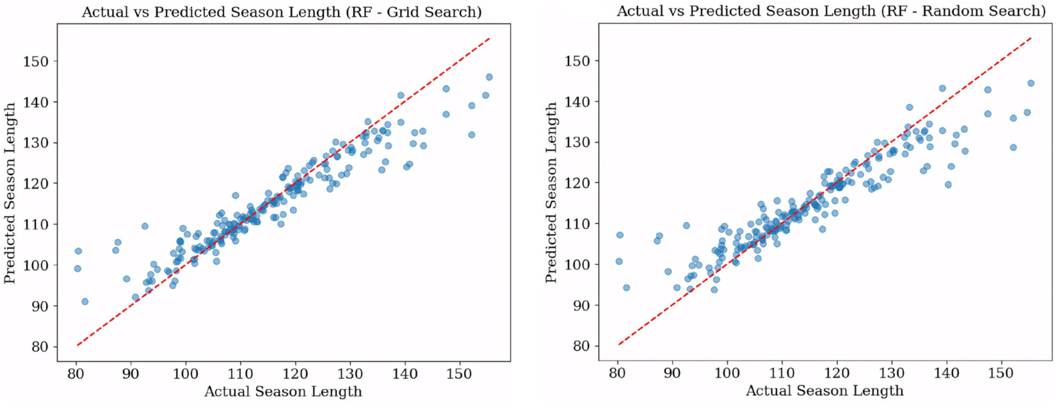
After the prediction model is validated, the next step is to develop an explainable predictive model that humans can easily interpret. The model should easily explain how each feature affects the prediction results and show how to identify the most influential features. The results of the SHAP and LIME methods used to identify these features can be seen in Fig. 8, 9 below.

The SHAP and LIME results show that the most influential features in predicting season_length are EOS and SOS. The features that improve prediction include SOS (11.89), prop (1.99), eos_sig (0.80), sos_sig (0.20), and pr_season (1.99). Meanwhile, the feature that lowers the prediction value is EOS (12.80). Therefore, these results suggest that the greater the EOS value, the lower the prediction results.

Table 3

Hyperparameter tuning result

| Hyperparameter tuning | Best parameter | Mean squared error | R-Squared |
|---|---|---|---|
| Grid search | 'max_depth': 15, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 300 | 28.9474 | 0.8636 |
| Random search | n_estimators': 200, 'min_samples_split': 10, 'min_samples_leaf': 2, 'max_depth': 15 | 36.9231 | 0.8260 |



*a* *b*

Fig. 6. Comparison result: *a* — predicted RF with grid search; *b* — predicted RF with random search

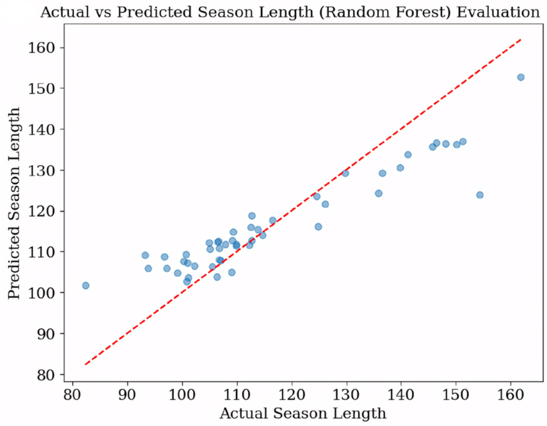


Fig. 7. Cross-country validation result

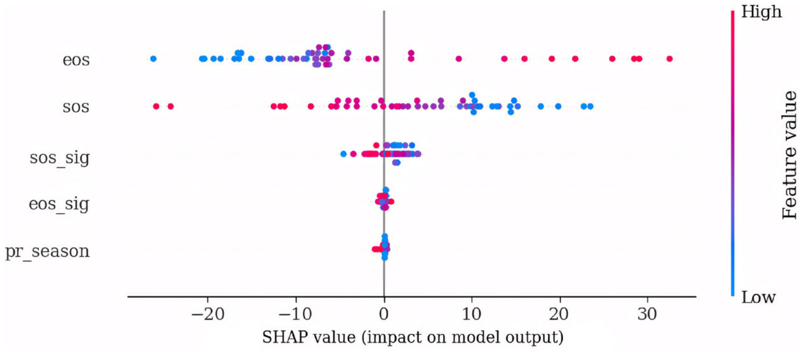
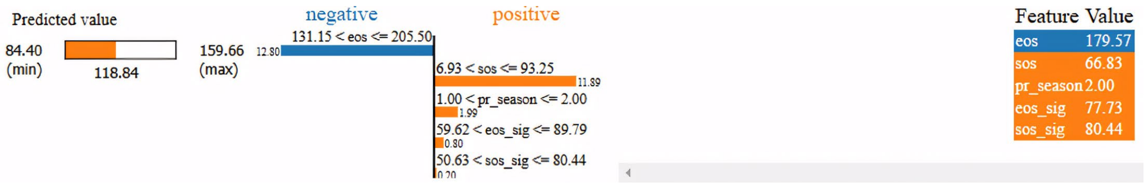


Fig. 8. Shapley additive explanation values



| Feature | Value |
|---|---|
| eos | 179.57 |
| sos | 66.83 |
| pr_season | 2.00 |
| eos_sig | 77.73 |
| sos_sig | 80.44 |

Fig. 9. Local interpretable model-agnostic explanation values

**5. 3. Season length cluster analysis**

The PCA analysis in Fig. 10 exhibits that PC1 is 45.6 % and PC2 is 19.5 % of the total variance in the dataset.

Thus, these two components account for about 65.1 % of the variation in the original data, indicating a good representation of high-dimensional data for these two dimensions.

Clustering analysis using K-Means (Fig. 11, *a*) produced three clusters with a Silhouette Score of 0.5336, while Hierarchical Clustering (Fig. 11, *b*) yielded a slightly improved Silhouette Score of 0.5457.

Table 4 shows the characteristics of each cluster based on several features. This analysis revealed different growing season patterns between clusters.

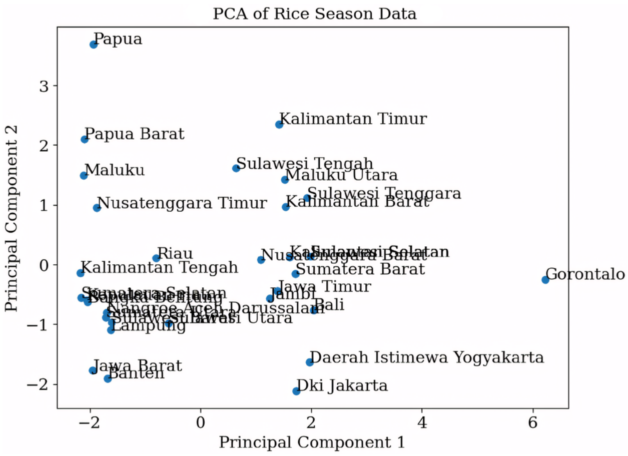
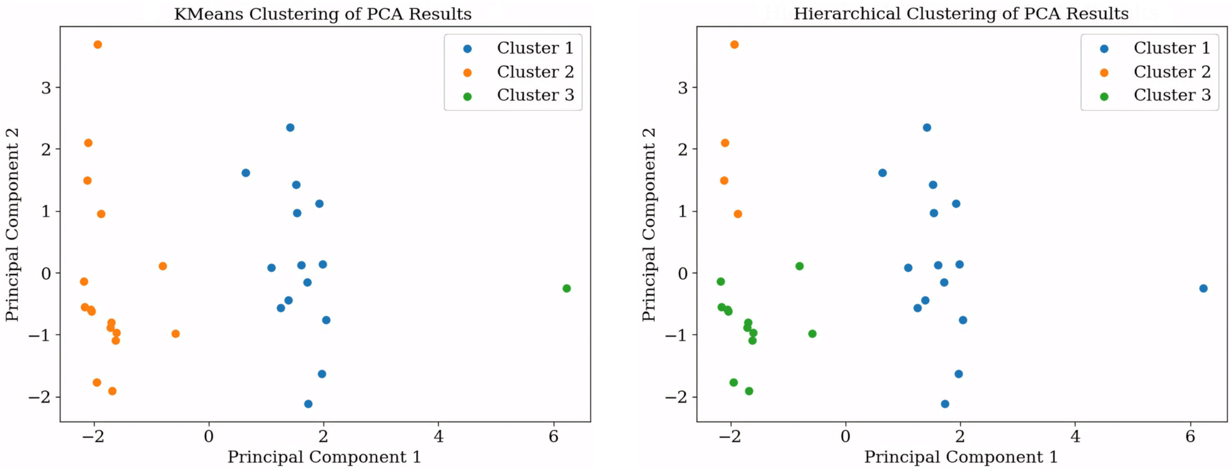


Fig. 10. PCA of rice season length data



*a* *b*

Fig. 11. Results: *a* – K-Means clustering; *b* – hierarchical clustering

Table 4

Clustering results

| Cluster | Season length | SOS | EOS | sos_sig | eos_sig | pr_season | Shape area | Shape length | Mean |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 113,821 | 93.7086 | 207.53 | 85.3993 | 86.5929 | 1.92857 | 4.26464 | 24.2733 | 0.000114 |
| 2 | 118.878 | 0.4831 | 119.36 | 44.4456 | 25.9831 | 1 | 5.54161 | 31.3112 | 0.000113 |
| 3 | 107.09 | 244.69 | 351.78 | 116.43 | 120.1 | 4 | 0.97908 | 7.9623 | 0.000103 |

## 6. Discussion of the big data analysis results for seasonal crop patterns

The findings of this study highlight the critical role of SOS, EOS, sig_sos, and sig_eos in predicting growing season length. Correlation analysis revealed a strong association (0.98) between SOS and EOS, emphasizing their importance in determining season duration. This relationship was consistent across different countries in Asia, as shown in Fig. 3, where regions such as Aceh (161.86 days) and South Sulawesi (154.40 days) exhibited the longest growing seasons. These results underscore the influence of climatic and environmental factors on the timing of SOS and EOS (Fig. 4), reinforcing the need for localized predictions in agricultural management.

The RF model, optimized using Grid Search, demonstrated superior performance compared to Random Search, achieving lower MSE (28.9474) and higher $R^2$ (0.8636), as shown in Table 3. Cross-country validation further validated the model's generalizability, with an $R^2$ of 0.7819 and a low MAPE of 5.6579%. However, the higher MSE (75.6734) indicates that additional features or enhanced data quality may improve predictive accuracy. This demonstrates the model's potential for broader application while also highlighting the need for further refinement to address regional variations effectively.

Tools like SHAP and LIME further underscored the importance of these features in Fig. 9, with `SOS` positively contributing to predictions and `EOS` having a negative impact in, offering valuable insights into their respective roles. Clustering analysis using PCA, K-Means, and Hierarchical Clustering revealed region-specific growing season patterns, as visualized in Fig. 11. The PCA results indicated that PC1 and PC2 together captured 65.1 % of the variance in the dataset, providing an effective dimensionality reduction. K-Means clustering identified three distinct clusters with a Silhouette Score of 0.5336, while Hierarchical Clustering yielded a slightly higher score of 0.5457. The results in Table 4 are discussed below:

Cluster 1:

– growing season length (season_length_x): 113.82 days. The growing season lasts quite a long time in this region;

– the start of the season (SOS): the growing season starts on day 93, which is quite early;

– the end of the season (EOS): the season ends on day 207, indicating a relatively long growing season;

– the significance of the start and end of the season (sos_sig, eos_sig): the sos significance is 85.4 and 86.6 for eos. These two values indicate that the beginning and end of the growing season significantly influence the season's length;

– number of growing seasons (pr_season): 1.93 growing seasons per year. The region usually experiences about two growing seasons a year;

– average rainfall (mean): 0.000114, indicating a moderate average rainfall in the region;

– shape area and shape length: An area of 4.26 and a length of 24.27 indicates the region covers a large area.

Cluster 2:

– growing season length (season_length_x): 118.88 days. This region has a longer growing season than Cluster 1;

– the start of the season (SOS): the growing season starts very early, around day 0 (almost at the beginning of the year);

– the end of the season (EOS): the season ends early, on day 119;

– the significance of the start and end of the season (sos_sig, eos_sig): the sos significance is 44.4 and 25.98 for eos, indicating a lower influence than Cluster 1;

– number of growing seasons (pr_season): 1 growing season per year. The region experiences only one growing season per year;

– average rainfall (mean): 0.000113, indicating slightly lower rainfall than Cluster 1;

– shape area and shape length: this region has a larger area (5.54) and a longer region length (31.31) than Cluster 1.

Cluster 3:

– growing season length (season_length_x): 107.09 days, the shortest growing season among the three clusters;

– the start of the season (SOS): the growing season starts very late, around day 244;

– the end of the season (EOS): the season ends very late, on day 351;

– the significance of the start and end of the season (sos_sig, eos_sig): the sos significance is 116.43 and 120.10 for eos, indicating that the sos and eos significantly influence growing season length;

– number of growing seasons (pr_season): 4 growing seasons per year. The region experiences four growing seasons per year;

– average rainfall (mean): 0.000103, indicating lower rainfall from Clusters 1 and 2;

– shape area and shape length: this region has a smaller area (0.98) and length (7.96) than other clusters.

Table 4 highlights the unique characteristics of each cluster, such as the longer growing seasons in Cluster 2 and the higher number of growing seasons in Cluster 3. These findings illustrate the variability in seasonal patterns and rainfall across regions, providing valuable insights for resource allocation and crop management strategies.

This method presents unique advantages over prior studies [12]. Unlike earlier research that primarily focused on crop yield predictions, this study emphasizes predicting the length of growing seasons using a RF model coupled with advanced hyperparameter optimization, achieving a commendable R-squared value of 0.8636. Additionally, the use of SHAP and LIME methods addressed the interpretability challenge of black-box machine learning models, as shown in Fig. 8, 9, offering actionable insights into feature importance. Furthermore, the clustering analysis uncovered unique planting season patterns across regions, which were not extensively explored in prior research that focused on generalized predictions [14, 15]. These contributions advance the understanding of seasonal crop dynamics and enhance resource allocation strategies.

Despite its strengths, the study is not without limitations. The findings are geographically confined to Asia, particularly Indonesia, and may not be readily applicable to regions with different climatic or agricultural conditions. Additionally, the research focused on a narrow set of environmental factors, such as rainfall and seasonal stability, excluding other influential variables like soil quality and temperature. Reproducibility poses another challenge, as the model's performance heavily depends on the quality and resolution of the input data, and its stability across diverse datasets cannot be assured without further validation. These limitations highlight the need for broader data inclusion and improved modeling approaches to enhance applicability and robustness.

Nonetheless, this study has several weaknesses that future research can address. Direct validation of the predictive model and clustering results in real-world settings has not been conducted, and field-level testing could significantly improve the practical relevance of the findings. The analysis also lacks granularity, as it relies on country-level data that may overlook microclimatic variations. Future work should consider integrating higher-resolution datasets and additional environmental and crop-specific features. Moreover, the computational expense of Grid Search optimization may hinder scalability for larger datasets, which could be mitigated by exploring alternative optimization techniques, such as Bayesian optimization.

The potential for future developments in this research is vast, although challenges are anticipated. Integrating real-time data from satellite imagery and IoT devices could enhance the model's adaptability but would require resolving technical issues related to data quality and scalability. Expanding the methodology to encompass different crops and regions will necessitate significant adjustments to account for varying environmental and agricultural contexts. Furthermore, developing decision-support systems based on these findings will require intuitive interfaces and real-time data processing capabilities. Investigating the impact of climate change on growing season length represents another valuable avenue for research, though it would require reliable climate models and long-term data availability. These advancements could significantly expand the utility and impact of this research in addressing agricultural challenges.

## 7. Conclusions

1. This research shows that the SOS (start of the season) and EOS (end of the season) features have a significant relationship in predicting the length of a growing season, with a correlation close to 0.98. The season length is determined by the difference between the SOS and EOS; the earlier the SOS (93 days) and the later the EOS (207 days), the longer the season length (113.82 days in Cluster 1). The significance values of sig_sos (85.4) and sig_eos (86.6) indicate that seasonal stability is highly dependent on these variables.

2. The RF model performed better when optimized with the grid search than the random search algorithm. The tuning results showed that the grid search algorithm produces a MSE of 28.9474 and an R-squared of 0.8636, while the random search has an MSE of 36.9231 and R-squared of 0.8260. These results indicate that models optimized with grid search are more accurate at predicting season length due to the use of more estimators (300) and better parameter settings.

3. The cluster analysis using the K-Means and hierarchical clustering methods identified three categories of season length, with Silhouette scores of 0.5336 and 0.5457, respectively, indicating good cluster separation. Cluster 1 has a season length of 113.82 days, with the SOS starting on day 93 and the EOS on day 207. Conversely, Cluster 3 has the shortest season length at 107.09 days, with a late SOS on day 244. This finding confirms the importance of considering local characteristics, such as the growing season patterns in each cluster, in planning more effective agricultural management strategies.

## Conflict of interest

The authors declare that they have no conflict of interest regarding this research, whether financial, personal, authorship, or otherwise, that could affect the research, and its results presented in this paper.

## Financing

The authors of this study received a Research Grant for the 2024 fiscal year from the Ministry of Higher Education, Science, and Technology, Indonesia. With the following Grant No: 103/E5/PG.02.00.PL/2024, and 050/LL1/AL0403/2024.

## Data availability

This study uses Data Underlying the Research on RICA: Rice Crop Calendar for ASIA, which is available at 4TU Research Data and can be accessed at: https://data.4tu.nl/articles/dataset/Data_underlying_the_research_on_RICA_Rice_Crop_Calendar_for_Asia/13468929.

## Use of artificial intelligence

The authors confirm that they did not use artificial intelligence technologies when creating the current work.

## Acknowledgments

We are grateful to DRTPM Kemendiktisintek for fulfilling the provisions of the research contract by giving the regular fundamental grant for the fiscal year 2024.

We are grateful for the assistance provided during the research process by the University of Mikroskil Institute for Research and Community Service (LPPM).

## References

1. El Bilali, H., Henri Nestor Bassole, I., Dambo, L., Berjan, S. (2020). Climate change and food security. The Journal "Agriculture and Forestry," 66 (3). https://doi.org/10.17707/agricultforest.66.3.16
2. Molotoks, A., Smith, P., Dawson, T. P. (2020). Impacts of land use, population, and climate change on global food security. Food and Energy Security, 10 (1). https://doi.org/10.1002/fes3.261
3. Favas, C., Cresta, C., Whelan, E., Smith, K., Manger, M. S., Chandrasenage, D. et al. (2024). Exploring food system resilience to the global polycrisis in six Asian countries. Frontiers in Nutrition, 11. https://doi.org/10.3389/fnut.2024.1347186
4. Adesete, A. A., Olanubi, O. E., Dauda, R. O. (2022). Climate change and food security in selected Sub-Saharan African Countries. Environment, Development and Sustainability, 25 (12), 14623–14641. https://doi.org/10.1007/s10668-022-02681-0
5. Javadi, A., Ghahremanzadeh, M., Sassi, M., Javanbakht, O., Hayati, B. (2022). Economic evaluation of the climate changes on food security in Iran: application of CGE model. Theoretical and Applied Climatology, 151 (1-2), 567–585. https://doi.org/10.1007/s00704-022-04289-w
6. Sirsat, M. S., Mendes-Moreira, J., Ferreira, C., Cunha, M. (2019). Machine Learning predictive model of grapevine yield based on agroclimatic patterns. Engineering in Agriculture, Environment and Food, 12 (4), 443–450. https://doi.org/10.1016/j.eaef.2019.07.003
7. Apat, S. K., Mishra, J., Srujan Raju, K., Padhy, N. (2022). State of the Art of Ensemble Learning Approach for Crop Prediction. Next Generation of Internet of Things, 675–685. https://doi.org/10.1007/978-981-19-1412-6_58
8. Chlingaryan, A., Sukkarieh, S., Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. Computers and Electronics in Agriculture, 151, 61–69. https://doi.org/10.1016/j.compag.2018.05.012
9. Sharma, A., Jain, A., Gupta, P., Chowdary, V. (2021). Machine Learning Applications for Precision Agriculture: A Comprehensive Review. IEEE Access, 9, 4843–4873. https://doi.org/10.1109/access.2020.3048415
10. Çeliktopuz, E. (2024). A Detailed Examination of Türkiye's Projected Precipitation and Growth Season Trends under Climate Change Condition. Black Sea Journal of Agriculture, 7 (3), 215–223. https://doi.org/10.47115/bsagriculture.1416956
11. Nainggolan, D., Abay, A. T., Christensen, J. H., Termansen, M. (2023). The impact of climate change on crop mix shift in the Nordic region. Scientific Reports, 13 (1). https://doi.org/10.1038/s41598-023-29249-w
12. Benos, L., Tagarakis, A. C., Dolias, G., Berruto, R., Kateris, D., Bochtis, D. (2021). Machine Learning in Agriculture: A Comprehensive Updated Review. Sensors, 21 (11), 3758. https://doi.org/10.3390/s21113758
13. Sharma, P., Dadheech, P., Aneja, N., Aneja, S. (2023). Predicting Agriculture Yields Based on Machine Learning Using Regression and Deep Learning. IEEE Access, 11, 111255–111264. https://doi.org/10.1109/access.2023.3321861
14. Ip, R. H. L., Ang, L.-M., Seng, K. P., Broster, J. C., Pratley, J. E. (2018). Big data and machine learning for crop protection. Computers and Electronics in Agriculture, 151, 376–383. https://doi.org/10.1016/j.compag.2018.06.008
15. Mavume, A. F., Banze, B. E., Macie, O. A., Queface, A. J. (2021). Analysis of Climate Change Projections for Mozambique under the Representative Concentration Pathways. Atmosphere, 12 (5), 588. https://doi.org/10.3390/atmos12050588
16. Mishra, B., Busetto, L., Boschetti, M., Laborte, A. G., Nelson, A. (2023). Data underlying the research on RICA: Rice Crop Calendar for Asia. 4TU.ResearchData. https://doi.org/10.4121/13468929
17. CMIP6 Climate Projections (2021). Climate Data Store. https://doi.org/10.24381/cds.c866074c
18. Guo, X., Hao, P. (2021). Using a Random Forest Model to Predict the Location of Potential Damage on Asphalt Pavement. Applied Sciences, 11 (21), 10396. https://doi.org/10.3390/app112110396

19. Fernandez-Gonzalez, P., Bielza, C., Larranaga, P. (2019). Random Forests for Regression as a Weighted Sum of k-Potential Nearest Neighbors. IEEE Access, 7, 25660–25672. https://doi.org/10.1109/access.2019.2900755
20. Zaib, R., Ourabah, O. (2023). Large Scale Data Using K-Means. Mesopotamian Journal of Big Data, 2023, 36–45. https://doi.org/10.58496/mjbd/2023/006
21. Shajalal, M., Boden, A., Stevens, G. (2024). ForecastExplainer: Explainable household energy demand forecasting by approximating shapley values using DeepLIFT. Technological Forecasting and Social Change, 206, 123588. https://doi.org/10.1016/j.techfore.2024.123588
22. Bhandary, A., Dobariya, V., Yenduri, G., Jhaveri, R. H., Gochhait, S., Benedetto, F. (2024). Enhancing Household Energy Consumption Predictions Through Explainable AI Frameworks. IEEE Access, 12, 36764–36777. https://doi.org/10.1109/access.2024.3373552
23. Chicco, D., Warrens, M. J., Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science, 7, e623. https://doi.org/10.7717/peerj-cs.623