ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 1, March 2024

ENHANCING RICE PRODUCTION PREDICTION: A COMPARATIVE MACHINE LEARNING ANALYSIS OF CLIMATE VARIABLES

Roni Yunis[1], Sudarto[2], Irpan Adiputra Pardosi[3]

[1,2,3] Informatics Faculty, Universitas Mikroskil, Medan, Indonesia

email: roni@mikroskil.ac.id[1], sudarto@mikroskil.ac.id[2], irpan@mikroskil.ac.id[3]

Abstract

This study aims to enhance rice production prediction through a comparative analysis of machine learning models utilizing climate variables. Eight models were assessed on a predetermined dataset, with Support Vector Regression (SVR) emerging as the top performer. Following the identification of significant climate variables influencing rice production, the models underwent evaluation using two hyperparameter approaches: random search and manual tuning. SVR outperformed other models, achieving impressive metrics with MAE 0.180, MSE 0.186, RMSE 0.431, and an exceptionally low MAPE of 0.020. Key factors influencing rice production included productivity and area, along with humidity, rainfall, temperature, wind velocity, and sunshine duration. Favorable conditions for rice output encompassed low humidity, moderate rainfall, increased wind speed, and prolonged sunshine, while rainfall and temperature exhibited minimal impact. The success of random search emphasizes the importance of effective hyperparameter tuning. This research provides valuable insights for enhancing rice production prediction.

Keywords : machine learning, comparative analysis, SVR, random search, rice production prediction

Received: 09-12-2023 | **Revised:** 03-03-2024 | **Accepted:** 06-03-2024
DOI: https://doi.org/10.23887/janapati.v13i1.71527

INTRODUCTION

Agriculture, a cornerstone of the global economy, plays a vital role in sustaining food production, with a particular emphasis on rice cultivation to meet worldwide nutritional demands[1][2]. As technological advancements, particularly in machine learning, gain prominence, their integration into agricultural practices becomes pivotal for enhancing productivity. Accurate prediction of rice yield serves as a crucial tool for governments, researchers, and farmers to formulate effective agricultural policies. The impact of climate change poses a significant challenge to rice cultivation, manifesting correlations between climate variability and rice output, influencing temperature, rainfall, and growing season patterns[3][4].

To achieve precise rice production forecasts, the development of a proficient prediction model capable of handling climate data as input is imperative. Machine learning, a subset of artificial intelligence, has shown promise in interpreting extensive data sets to provide more accurate predictions. Despite various machine learning techniques explored for rice production prediction using climatic data, a comprehensive comparison of these algorithms is yet to be undertaken. This study aims to fill this gap by conducting a thorough comparative analysis of different machine learning techniques employed in rice production prediction, specifically considering climatic data. Recent research underscores the increasing interest in utilizing machine learning, such as Random Forest and Artificial Neural Networks (ANN), in predicting rice harvests and rice yield[5][1][6]. Emphasizing climatic data has been shown to enhance prediction accuracy. Regionally localized models have also proven more effective, underscoring the importance of incorporating local climatic variables in prediction model construction[7][8][9].

This study addresses the existing gap by conducting a comprehensive comparative examination of various machine learning methods for predicting rice production in conjunction with climate data. Data on rice production and meteorological factors, including temperature, humidity, and rainfall, were collected from reputable sources in ten Indonesian provinces on Sumatra Island.

The model's performance was evaluated using a Random Forest Regressor, Support Vector Regressor, Decision Tree, XGBoost, Gradient Boosting Machine, K-Nearest Neighbors, and Artificial Neural Network. Variable significance evaluation and

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 1, March 2024

data preparation are critical aspects considered in this study. Model efficacy was assessed using performance metrics such as Mean Absolute Percentage Error (MAPE) and Mean Absolute Error (MAE). The study aims to provide a reliable decision-making model for stakeholders, including farmers, with practical implications. By illuminating the merits of various machine learning algorithms, it also contributes to the development of sustainable agricultural policies. Anticipated outcomes include advancements in the field through a thorough comparison of machine learning approaches, offering valuable insights to inform the development of sustainable agricultural policies, particularly in the domain of rice production forecasts.

METHOD

The analysis procedure and approach are explained in Figure 1.

Data Collection and Study Area

Based on the availability of data on rice production in Indonesia, the dataset used to support this study was gathered. In this instance, data collection was restricted to the island of Sumatra, which has ten provinces: Aceh, North Sumatra, West Sumatra, Riau, Riau Islands, Jambi, South Sumatra, Bengkulu, Lampung, and Bangka Belitung Islands (Figure 2). Complete rice production data for the province from 2006 to 2015 may be found at:

- https://www.bps.go.id and
- https://tanamanpangan.pertanian.go.id.

Climate variables from:

- https://dataonline.bmkg.go.id,
- https://climateserv.servirglobal.net/, and
- https://www.ncei.noaa.gov/

were collected throughout the same period. Three distinct sources of data were combined based on Figure 2. Eleven variables are generated from the integrated data, which includes statistics on climate, and rice production. The eleven factors in research are: year, province, commodity, output, productivity, harvest area, humidity, wind velocity, temperature, sunshine duration, and rainfall. The total data used in this study is 700 rows/records.

Data Preprocessing and Exploratory Data Analysis

Data pre-processing involves steps such as missing value handling, outlier removal and transformations or feature selection to prepare data for analysis. Meanwhile, Exploratory Data Analysis (EDA) involves using descriptive statistics and data visualisations such as histograms and correlation matrices to understand patterns, relationships, and trends in the dataset. Both are critical to validating data, ensuring integrity, and providing key insights before proceeding to further analysis or modelling.

Data Partition

To train and assess machine learning models, data partitioning is splitting a dataset into training and test sets, usually via randomization. By ensuring that models are assessed on untested data, this procedure helps to avoid overfitting and offers a trustworthy indicator of how well a model generalizes.

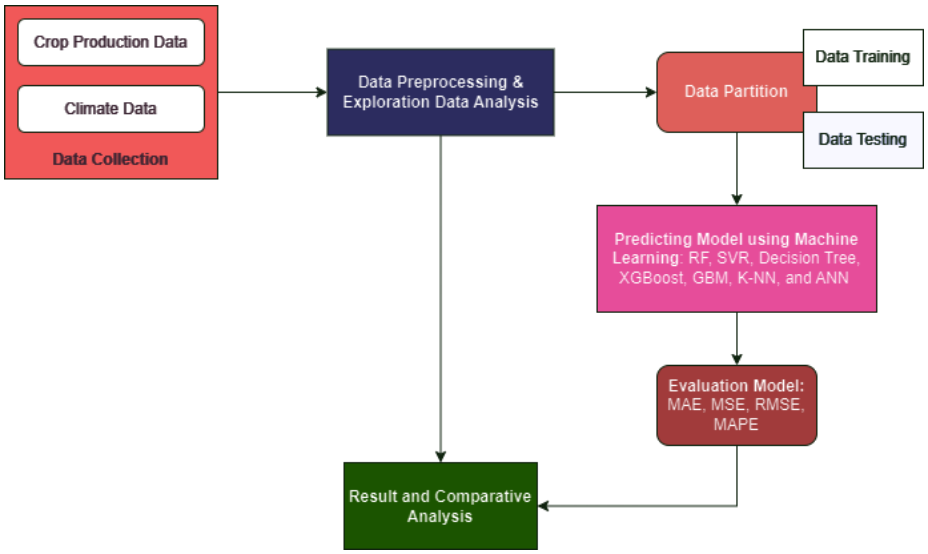


Figure 1. Research Methodology

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 1, March 2024

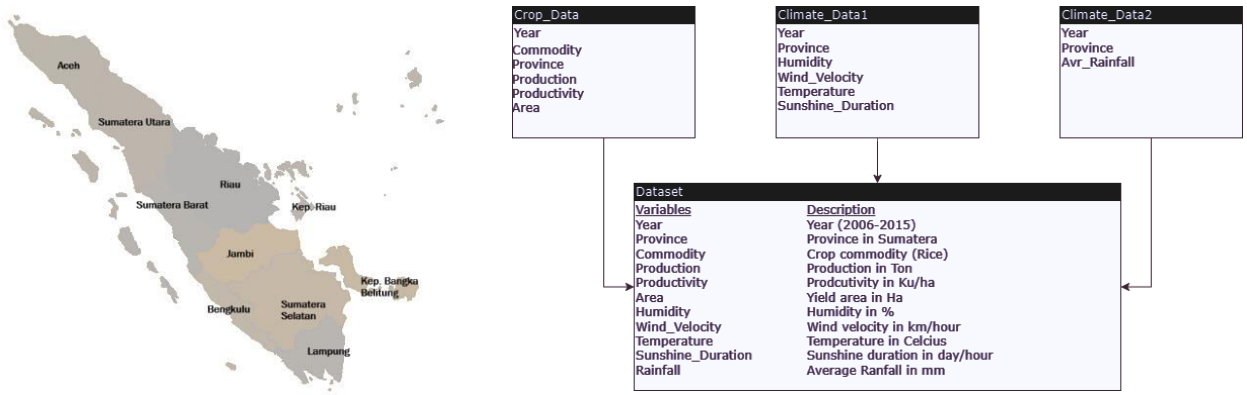


Figure 2. Location Map of The Study and Variables

A careful compromise is struck between a sufficient training set and a reliable evaluation of the test set when determining the size of the subsets. For training data, the percentage size is 70%, and for testing data, it is 30%. A crucial stage in creating a very successful model based on some experimental data is training [10]. This ratio is a widely used conversion that performs well in a variety of scenarios because overfitting occurs when a model is trained with an excessive amount of data.

Predicting Model Using Machine Learning

To predict rice production given climate data, this study uses eight machine learning models: GLM, RFR, XGBoost, SVR, Decision Tree Regression, GBM, KNN, and ANN. In order to gain a comprehensive grasp of the intricate relationship between rice production and climate, they are used in situations requiring linear interpretation, managing complexity and non-linear patterns, spatial interpretation and dependency, high complexity, and iterative improvement.

Random Forest Regression

Decision trees are used in Random Forest Regression, a machine learning technique, to do regression tasks. It creates multiple decision trees and averages their projections to reach the desired outcome. This method helps to increase the accuracy of the regression model and decrease overfitting.[1]. Numerous fields, including economics, healthcare, and environmental research, frequently use Random Forest Regression. It has been applied to climate modeling and ecological forecasting in environmental studies. [11].

All things considered, Random Forest Regression is an effective tool for regression problems, offering robustness against overfitting and precise predictions. The Random Forest Regression (RFR) method starts with creating a forest of uncorrelated trees. Each tree is developed using a randomized subset of predictor variables. The trees are averaged together after being allowed to grow to their greatest potential without any pruning[12].

$$y = \frac{1}{N}\sum_{i=1}^{N} f_i\,(x) \qquad (1)$$

where x is the feature vector of the newly predicted observation, f_i is the regression function of tree i, N is the number of decision trees, and y is the prediction's output value.

Support Vector Regressor

The machine learning method known as Support Vector Regression (SVR) applies the fundamental concepts of support vector machines to regression issues. It is frequently used to create decision boundaries in non-linear space by learning from training data[13][14]. The formulation for liner support vector regression that was used in this investigation is [15].

$$y = \sum_{n=1}^{N} (\alpha_n - \alpha_n^*)\,(x_n\,.x) + b \qquad (2)$$

Where x_n and $x_n.n$ represent the dot product and α_n and α_n^* are nonnegative multipliers matching to each observation

Decision Treen Regression

Decision Tree Regression is one of the most potent machine learning algorithms because it can accurately depict intricate, non-linear relationships between input and output variables[16][17]. Using particular rules or criteria, the program separates the data into subsets in this regression approach. With the following formulation, Decision Tree Regression

is quite similar to the Classification Tree used for Regression.[18].

$$y = f(x, \theta), O \epsilon R \quad (3)$$

where x is the new observation, y is the output that corresponds to the set of real numbers R, $f(.)$ is the regression function, and θ is the regression function's parameter set.

XGBoost Regression

A well-liked gradient-boosting framework called XGBoost uses a tree-based learning mechanism. With millions of examples and features, it can handle large-scale datasets because to its scalable and efficient design. Its ability to handle categorical or numeric characteristics directly, without the need for one-hot encoding or other preprocessing procedures, is one of its important strengths[11]. The following formula can be used to describe the XGBoost regression model.

$$y = \sum_{i=1}^{N} f_i(x; \theta_i) \quad (4)$$

Where f_i is the regression function of tree i, x is the feature vector of the new predicted observation, N is the number of decision trees, y is the prediction's output value, and θ is tree i's parameter.

Gradient Boosting Machine

Ensemble learning is a subset of machine learning algorithms, including Gradient Boosting Machine. Using multiple models, ensemble learning increases prediction performance and accuracy[19][20]. Gradient boosting machines, or GBMs, combine decision trees iteratively via an additive model, lowering the loss function by gradient descent to minimize prediction errors. This improves weak models [21].

$$F_n(x_t) = \sum_{i=1}^{n} f_i(x_t) \quad (5)$$

A decision tree (regression tree) is represented by each $f_i(x_t)$. By estimating the new decision tree $f_{n+1}(x_t)$ using the following equation, the ensemble of trees is constructed successively. In cases where the differentiable loss function $L(.)$.

$$argmin \sum_{t} L(y_t . F_n(x_t) + f_{n+1}(x_t)) \quad (6)$$

K-Nearest Neighbors Regression

A well-liked machine learning approach for classification and regression applications is K-Nearest Neighbors (KNN). Its foundation is the idea of locating a data point's closest neighbors and forecasting information about them based on their value or class[22][23]. In KNN regression, euclidean distances are evaluated first, and then the distance level. By determining the optimal K value, the algorithm determines the values that are closest. It determines the inverse distance average with its neighbors[22].

$$y = \frac{1}{k} \sum_{i=1}^{k} y_i \quad (7)$$

where y_i is the output value of the i nearest neighbor, k is the number of nearest neighbors utilized in the prediction, and y is the expected output value.

Artificial Neural Network

Three layers make up a typical artificial neural network's architecture: input, hidden, and output layers. Backpropagation is a crucial learning technique used during training[23]. Artificial neural networks (ANNs) are efficient for predictive modeling in agriculture and various domains, adept at handling both linear and non-linear correlations in time series data. Comprising interconnected layers, input layer with nodes, output layer of neurons, and hidden layers with one to three layers of neurons it utilizes weighted links to represent numerical values.[6].

$$h_i = \sigma\left(\sum_{j=1}^{N} V_{ij} x_j + T_i^{hid}\right) \quad (7)$$

Where N is the number of input neurons, σ is the activation function, V_{ij} is the weight, x_j is the neuron input, and T_i^{hid} is the threshold term for the hidden neurons.

Hyperparameter Search

This study's machine learning model tuning procedure makes use of both the manual hyperparameter search method and random search. Hyperparameter combinations are chosen at random from a predetermined search space in a random search[24]. Five-fold cross validation was used to evaluate the efficacy of a random search to yield a more dependable estimate of the hyperparameter value. The goal is to find an efficient or ideal configuration by methodically examining a wide range of hyperparameter combinations.[25]. The training data that was previously partitioned is used in the hyperparameter search procedure.

Evaluation Model

During the assessment stage, a comprehensive examination of several regression techniques and performance indicators was provided, providing the user with option in choosing accuracy parameters that are pertinent to them. The common error rate measures used in applied machine learning are described in this section. The average of the absolute difference between the expected and actual values is called the mean absolute error, or MAE. represents the average mistake magnitude without taking the direction of the faults into account.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - y'_i| \quad (8)$$

The average of the squared difference between the expected and actual values is known as the mean squared error, or MSE. increases the weith of bigger errors, making it more susceptible to outliers.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y'_i)^2 \quad (9)$$

The square root of the average squared discrepancies between the expected and actual values is known as the root mean squared error, or RMSE. offers a comprehensible scale similar to the initial data.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y'_i)^2} \quad (10)$$

The average of the absolute percentage discrepancies between the expected and actual values is known as the mean absolute percentage error, or MAPE. shows the typical percentage difference between the expected.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - y'_i}{y_i}\right| \times 100 \quad (11)$$

RESULT AND DISCUSSION

After analyzing collected and preprocessed data, it was found that there are 700 observations and 22 NA's in both food crop and climate datasets. The combined dataset, consisting of rice crop and climate data, comprises 100 rows/records and 8 variables. Figure 3 displays the distribution of preprocessed data. In the exploration stage of data analysis, the findings are as follows: a) North Sumatra has the highest rice production, averaging 749157.70 tons (Figure 4). b) Pearson correlation (Figure 5) indicates a strong, significant negative correlation between wind velocity (-0.532) and production, humidity (-0.183) and production, a weak positive correlation (0.196) between rainfall and production, and a weak negative correlation (-0.183) between humidity and production. Temperature (0.003) and sunshine duration (0.045) show very weak correlations with production. Area (0.914) and productivity (0.785) are strongly positively correlated with production.

In order to lessen the weight of highly correlated variables, ridge regression analysis was performed in this study in order to demonstrate multicollinearity due to the relatively small sample size in the data. Therefore, it is anticipated that more stable estimates would be produced later on in the analytical and model-building process. Based on the regression model for 7 variables selection (predictor variables) that are thought to affect the response variable (production) and cross validation, it was found that the minimum lambda value or MSE evaluation was 0.237, and lambda 1se was 2.663 with the number of predictor variables fixed 7 and the predictor variable coefficient is not zero. So, it can be concluded that these 7 variables contribute significantly and are not eliminated in the model selection. The 7 variables predictor maintained include: area, productivity, humidity, rainfall, temperature, wind velocity, and sunshine duration.

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 1, March 2024

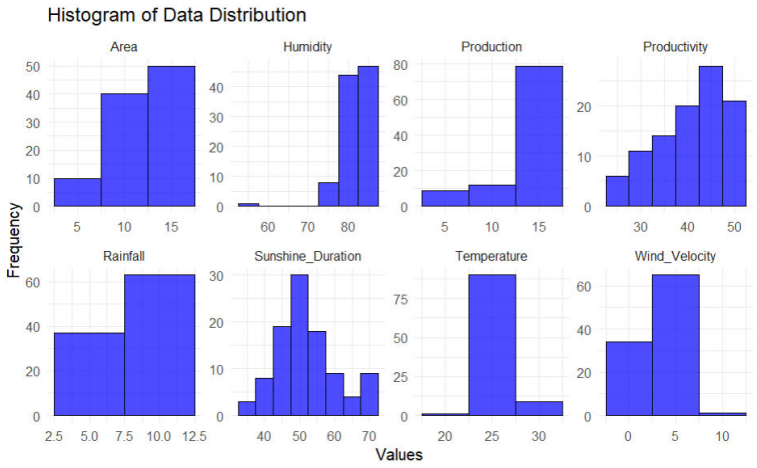


Figure 3. Data Distribution

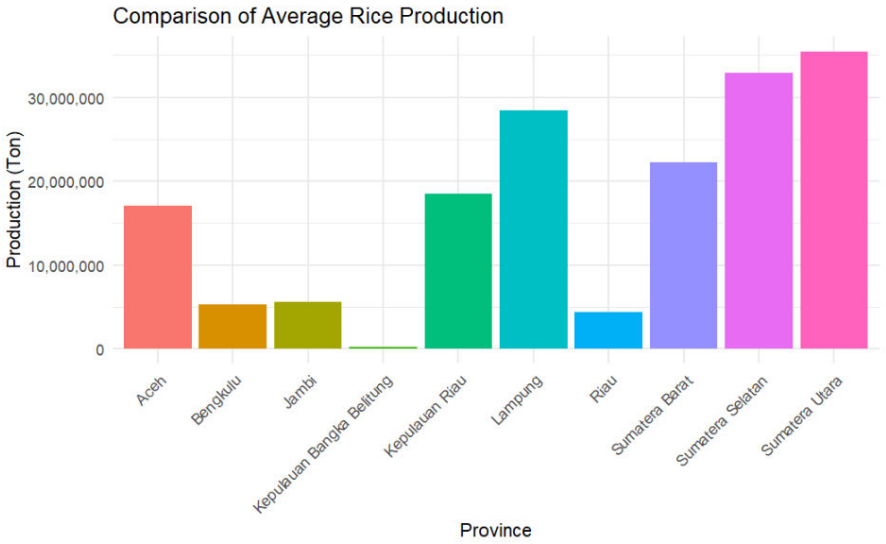
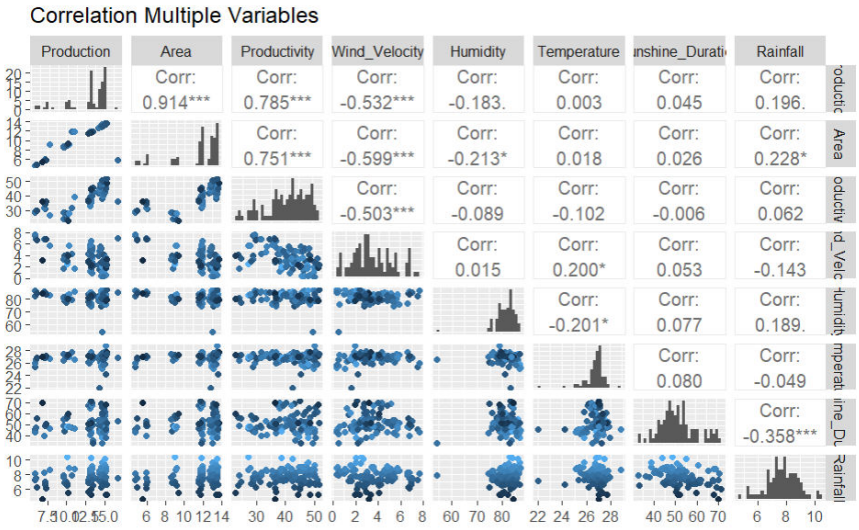


Figure 4. Comparison of Average Rice Production



ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 1, March 2024

Figure 5. Correlation Multiple Variables

The next stage is to use machine learning to create a prediction model after preprocessing and data exploration. Setting up the experimental setup and the necessary hyperparameters is the initial step. Table 1 displays the hyperparameters for each model based on the experimental case. There are two components to the experimental scenario: experiment A and experiment B. Experiment A uses 5-fold cross-validation to evaluate the hyperparameters obtained using the random search technique, while Experiment B tests the manually generated hyperparameters. Anticipated benefits from conducting both experiments include the evaluation of the quality and accuracy of the prediction model created. The goal of this thorough evaluation is to shed light on the ideal ratio between model performance and balance.

The outcomes of the predictions made for rice production for each model tested using prepared experimental situations are described below. The model's ability to predict real values can be thoroughly examined by examining the model performance using four key metrics: MAE, MSE, RMSE, and MAPE. With a greater emphasis on significant errors and consideration of absolute, relative, and squared errors, this analysis offers a thorough picture of the model's prediction quality. Based on the results of performance evaluation from the model with both experiments using testing data 30% in Table 2 and Table 3 it can be explained that: The GLM model with the two experiments carried out has no difference in performance, this can be seen from the comparison plot of prediction results and actual values in Figure 6. With MAE values of 0.438, MSE 0.465, RMSE 0.682, and MAPE 0.049. Hyperparameter tuning on the model has no significant effect, this may be because the model has provided good and stable performance.

For experiments on the RF Regression model, it was found that the performance of models with hyperparameter A was better than the hyperparameter B model. This can be seen from the evaluation metrics MAE 0.522, MSE 0.944, RMSE 0.971, and MAPE 0.067. In Figure 7 it can be seen that by adding n_estimators or the number of decision trees to 200 based on the results of random search hyperparameter tuning results in better model performance. The difference between the MAE values of the two hyperparameter configurations is significant with the mean of the MAE (mean of x) being 0.5475 with a p-value of 0.02963.

It was discovered through SVR model experiments that SVR models using random search or hyperparameter A performed exceptionally well. The values of MAE 0.180, MSE 0.186, RMSE 0.431, and MAPE 0.020 all demonstrate this. In contrast to the radial kernel in hyperparameter B, Figure 8 illustrates how precisely the linear kernel is used in the model. With a very low error rate, the SVR model performs exceptionally well in hyperparameter A for this prediction of rice production while using a linear kernel. A large error rate is caused by the usage of radial kernels in hyperparameter B; this could be because the training model employed a very small quantity of data.

Table 1. Hyperparameter Experiment

Machine Learning/Model	Hyperparameter	Experiment Scenario	
		A	B
GLM	family	Gaussian	Gaussian
	epsilon	1e-8	1e-5
	maxit	100	200
Random Forest	n_estimators	100	200
	max_depth	NULL	NULL
	min_sample_split	2	2
	min_sample_leaf	1	1
	mtry	sqrt	log2
SVR	kernel	linear	radial
	C	1.33548	0.1
	epsilon	0.1	0.01

	gamma	0.01471221	0.1
XGBoost	objective	reg: squarederror	reg: squarederror
	booster	gbtree	gbtree
	nrounds	100	200
	eta	0.3	0.1
GBM	distribution	gaussian	gaussian
	n. tree	100	200
	Interaction.depth	3	5
	shrinkage	0.1	0.01
	cv.fold	5	5
KNN	K	5	7
	weight	Uniform	Distance
Decision Tree	max_depth	NULL	NULL
	min_samples_split	2	2
	min_samples_leaf	1	1
	cp	0.007	0.001
ANN	hidden_neurons	c (5,2)	c (5,2)
	epochs	100	50
	batch_size	32	64
	learning_rate	0.001	0.001
	activation	linear	linear

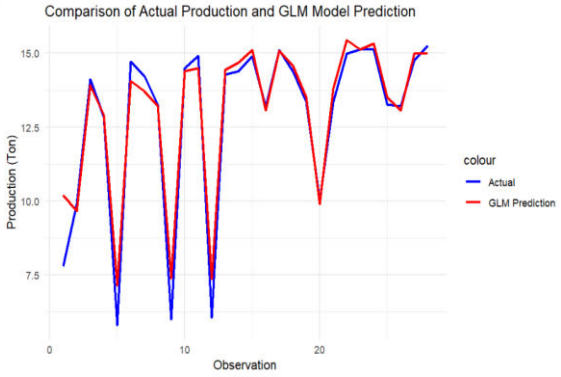
Table 2. Evaluating Metrics with Hyperparameter A

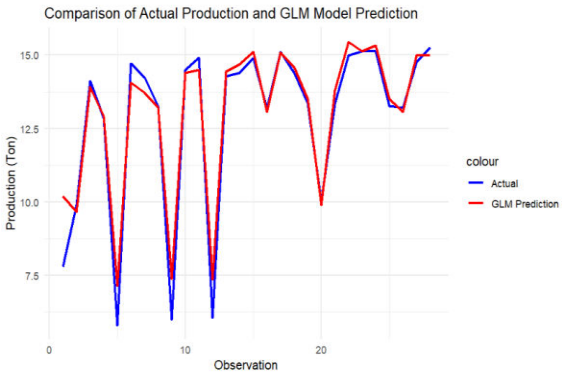
No	Model	MSE	RMSE	MAE	MAPE
1	GLM	0.4658171	0.6825080	0.4280718	0.04938833
2	Random Forest	0.9440180	0.9716059	0.5223369	0.06784828
3	SVR	0.1864071	0.4317489	0.1808325	0.02012520
4	XGBoost	1.8617017	1.3644419	0.4177796	0.06290435
5	GBM	1.1664109	1.0800051	0.6295821	0.07927708
6	KNN	1.2513477	1.1186366	0.7278808	0.08698134
7	ANN	2.4463306	1.5640750	1.1109724	0.12141988
8	Decision Tree	5.0648364	2.2505191	0.8407432	0.11916796

Table 3. Evaluating Metrics with Hyperparameter B

No	Model	MSE	RMSE	MAE	MAPE
1	GLM	0.4658171	0.682508	0.4280718	0.04938833
2	Random Forest	1.2641199	1.124331	0.5732929	0.07653533
3	SVR	5.0526672	2.247814	1.1970930	0.16093730
4	XGBoost	1.8617017	1.364442	0.4177796	0.06290435
5	GBM	2.4422419	1.562767	0.8722005	0.11365346
6	KNN	1.2513477	1.118637	0.7278808	0.08698134
7	ANN	2.0603179	1.435381	1.0334684	0.11184952
8	Decision Tree	4.9920981	2.234300	0.7415292	0.11009584

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 1, March 2024





(Hyperparameter A) (Hyperparameter B)

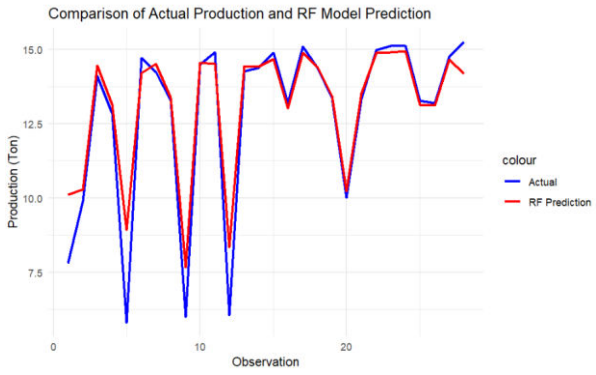
Figure 6. Comparison of GLM Model Results

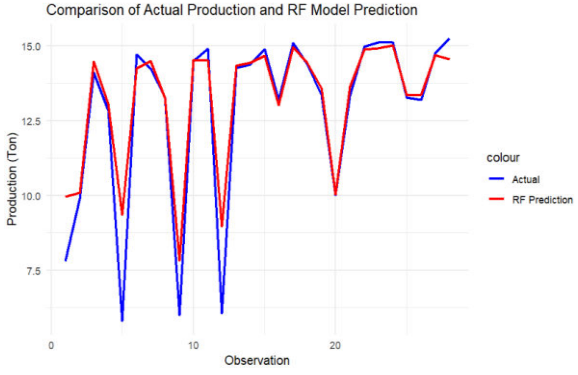
When using hyperparameter B, DTR models perform better than when using hyperparameter A, which is the result of random search. The MAE value of 0.741, MSE of 4.99, RMSE of 2.234, and MAPE of 0.110 all support this. The usage of cp = 0.001 in a DTR model with hyperparameter B exhibits better predictability than cp = 0.007, as seen in Figure 9. The mean of the MAE (mean of x) is 0.7905 with a p-value of 0.03981, indicating a significant difference between the MAE values of the two hyperparameter setups.

The performance of the XGBR model with hyperparameter A is indistinguishable from that of the hyperparameter B model. This fact is evident from the recorded values of MAE 0.417, MSE 1.861, RMSE 1.364, and MAPE 0.062, all of which are identical. Figure 10 visually confirms the absence of any notable disparity between the predicted outcomes produced by the two hyperparameters. It has been determined that employing the same learning rate (eta) value and a large number of iterations (nround) for hyperparameter B does not yield any significant advantages.

Experiments on the GBM model, it was found that the performance of the GBM model with hyperparameter A or random search was good. This can be seen from the value of MAE 0.629, MSE 1.166, RMSE 1.080, and MAPE 0.079. In Figure 11 it can be seen that the increase in n. tree = 200 and shrinkage = 0.01 in hyperparameter B does not provide significant performance.

Experiments on the KNN model, it was found that the performance of the KNN model with hyperparameter A and hyperparameter B was the same. This can be seen from the value of MAE 0.727, MSE 1.251, RMSE 1.118, and MAPE 0.086. In Figure 12 it can be seen that the use of weights = distance in hyperparameter B does not provide significant performance. This can happen because there is no significant non-liner trend in the data, so it does not have a significant impact compared to weight = uniform.

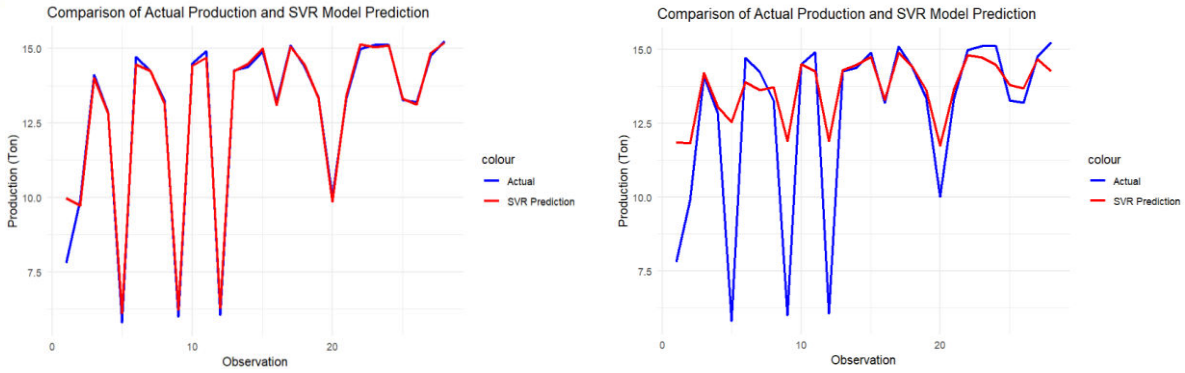




(Hyperparameter A) (Hyperparameter B)

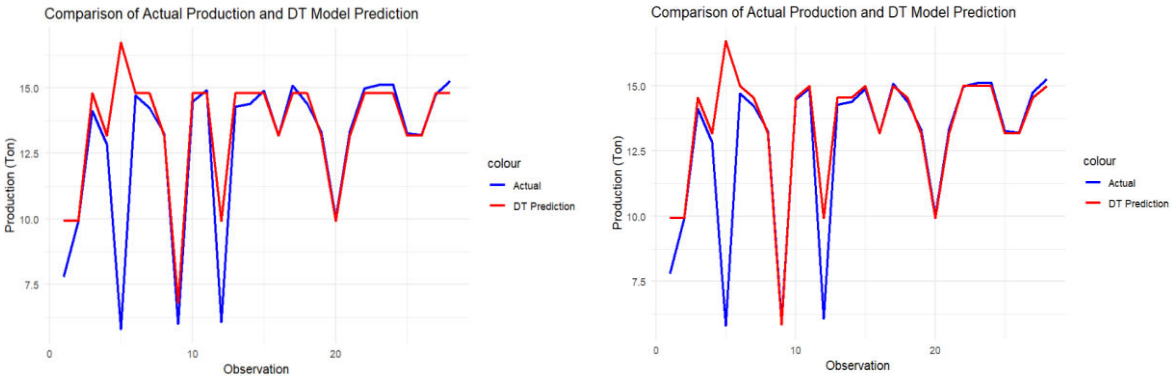
Figure 7. Comparison of RF Model Results

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 1, March 2024



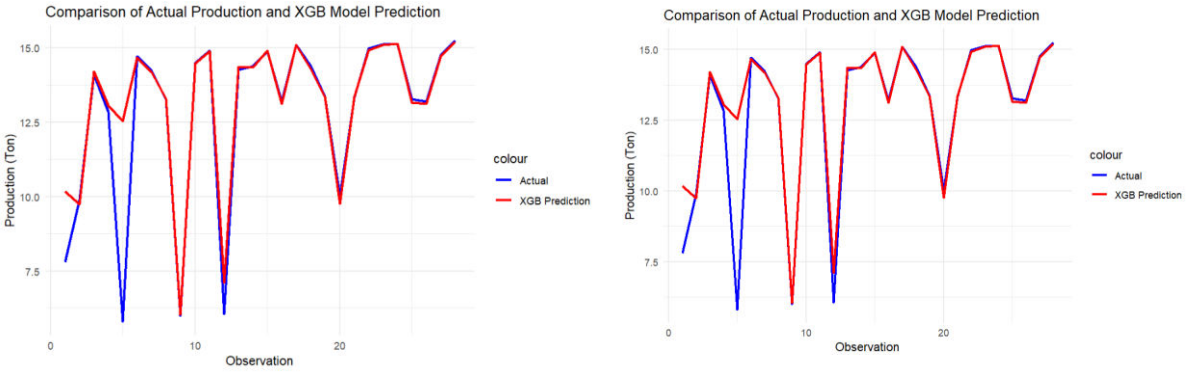
(Hyperparameter A) (Hyperparameter B)

Figure 8. Comparison of SVR Model Results



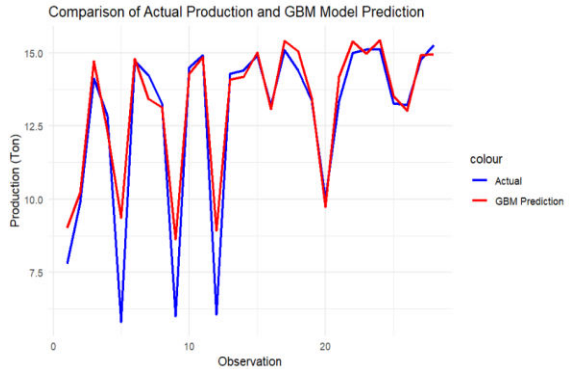
(Hyperparameter A) (Hyperparameter B)

Figure 9. Comparison of DT Model Results

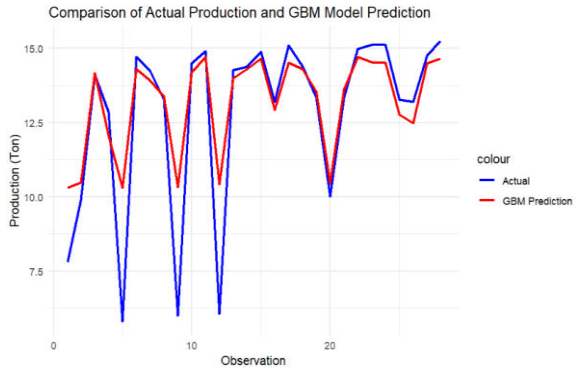


(Hyperparameter A) (Hyperparameter B)

Figure 10. Comparison of XGB Model Results

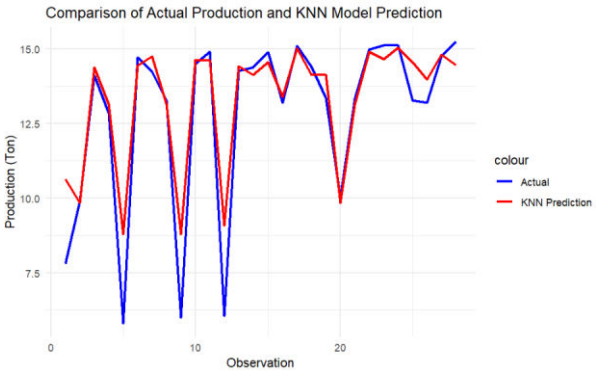


(Hyperparameter A)

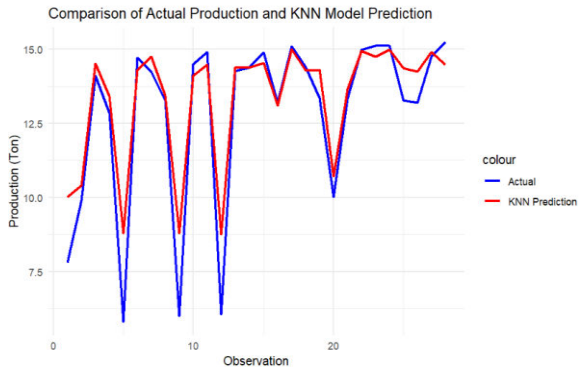


(Hyperparameter B)

Figure 11. Comparison of XGB Model Results



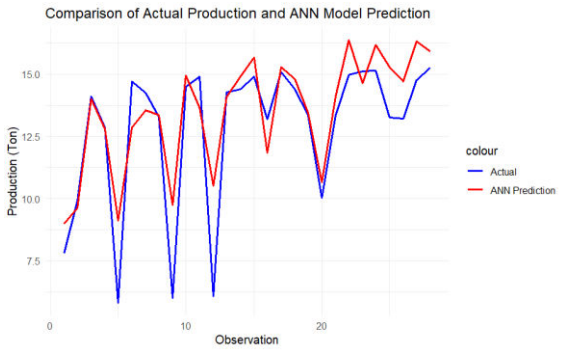
(Hyperparameter A)



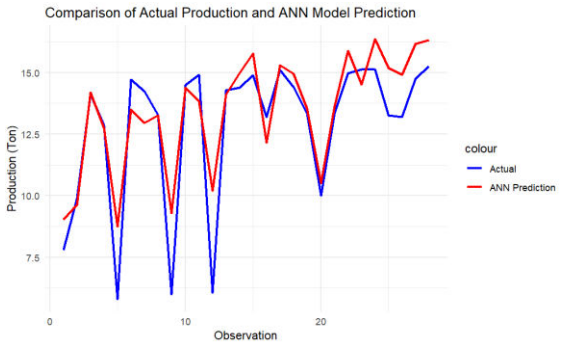
(Hyperparameter B)

Figure 12. Comparison of KNN Model Results

The performance of the ANN model with hyperparameter B is better than hyperparameter A. This can be seen from the values of MAE 1.033, MSE 2.060, RMSE 1.435, and MAPE 0.111. In Figure 13 it can be seen that the use of epoch = 100 in hyperparameter A based on random search does not provide performance improvement. Although statistically the difference from the mean MAE (mean of x) value of the two hyperparameters is 1.0715 with a p-value of 0.02286 significant. This could occur due to the limited amount of data used for model training; perhaps with a larger dataset, performance might differ even with the same epoch.



(Hyperparameter A)



(Hyperparameter B)

Figure 13. Comparison of ANN Model Results

$$production = -0.0458 + 0.8844 * area + 0.1542 * productivity + 0.013 * rainfall - 0.022 * humidity + 0.022 * windvelocity - 0.013 * temperature + 0.021 * sunshineduration \quad (12)$$

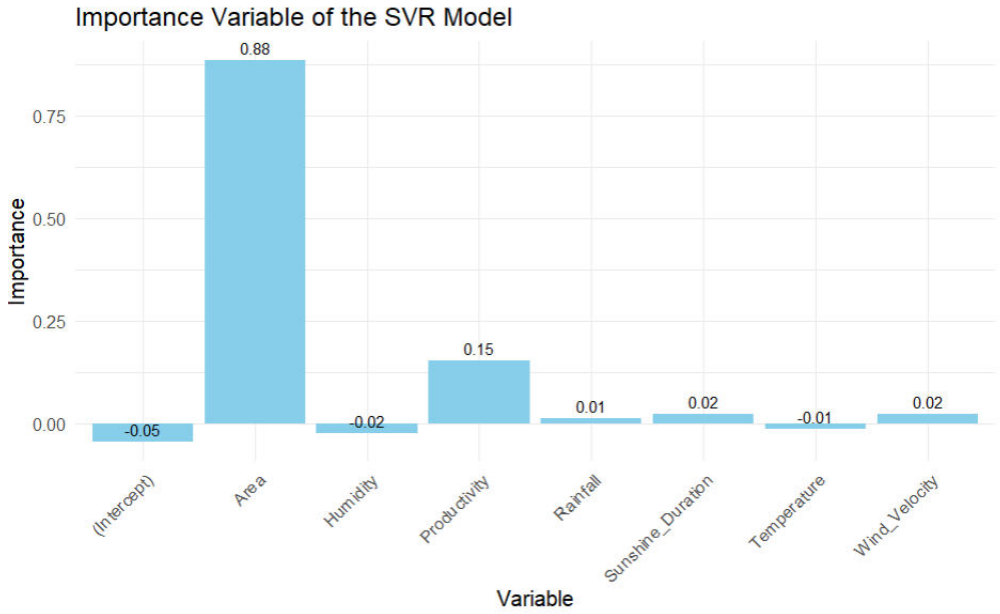


Figure 14. Important Variables of the SVR Model

The outcomes demonstrated that, when compared to seven other models, the SVR model performed exceptionally well in predicting rice production. The model was successful in pinpointing the critical variables in Figure 14 that have a significant impact on rice output, with area and productivity variables being the most influential. The model's regression equation (12) gives a general picture of how the variables interact to effect production. Analysis of climate variables showed that wind speed and sunshine duration had a beneficial effect on output. Variables related to temperature and precipitation have comparatively little effect on output, whereas variables related to humidity have a major detrimental effect. With SVR as the top model, these findings offer insightful information for making decisions about raising rice output.

CONCLUSION

Based on the results of comparison and evaluation on 8 machine learning models tested on rice production datasets and using climate variables, it was found that the Support Vector Regression (SVR) model with hyperparameter random search performed very well with low values for all evaluation metrics. The SVR model has a very low MAPE value of 0.020, MAE value of 0.180, MSE 0.186, and RSME of 0.431. The SVR model successfully identified that area and productivity variables showed a very important and significant influence on production, while climate variables such as sunshine duration, wind velocity and rainfall also played an important role. In planning rice production, close monitoring of these factors is needed, especially for humidity and temperature variables because they have a negative impact even though they are low. SVR models can be an effective tool with good environmental management to support more accurate production predictions.

Future work can carry out transfer learning related to the model that has been generated to predict the production of other types of food crops and develop approaches in the model to adapt to different kinds of tasks and datasets so that it can work better in a new learning context.

ACKNOWLEDGMENT

We are grateful to DRTPM Kemendikbudristek for fulfilling the provisions of the research contract by giving the regular fundamental grant for the fiscal year 2023. We are grateful for the assistance provided during the research process by the Universitas Mikroskil Institute for Research and Community Service (LPPM).

REFERENCES

[1] E. A. Jiya, U. Illiyasu, and M. Akinyemi, “Rice Yield Forecasting: A Comparative Analysis of Multiple Machine Learning Algorithms,” *J. Inf. Syst. Informatics*, vol. 5, no. 2, pp. 785–799, 2023.

[2] N. Chergui and M. T. Kechadi, "Data analytics for crop management: a big data view," *J. Big Data*, vol. 9, no. 1, 2022.
[3] H. Khan and S. Ali, "Impact of Climate Variability on Rice Productivity in Pakistan," pp. 1–27, 2023.
[4] X. Feng, H. Tian, J. Cong, and C. Zhao, "A method review of the climate change impact on crop yield," *Front. For. Glob. Chang.*, vol. 6, no. July, pp. 1–7, 2023.
[5] M. Kuradusenge *et al.*, "Crop Yield Prediction Using Machine Learning Models: Case of Irish Potato and Maize," *Agric.*, vol. 13, no. 1, 2023.
[6] A. Satpathi *et al.*, "Comparative Analysis of Statistical and Machine Learning Techniques for Rice Yield Forecasting for Chhattisgarh, India," *Sustain.*, vol. 15, no. 3, pp. 1–18, 2023.
[7] M. A. Gedik and T. Günel, "The impact of climate change on edible food production: a panel data analysis," *Acta Agric. Scand. Sect. B Soil Plant Sci.*, vol. 71, no. 5, pp. 318–323, 2021.
[8] R. B. Pickson, G. He, and E. Boateng, "The impacts of climate change and smallholder farmers' adaptive capacities on rice production in Chengdu, China: Macro-micro analysis," *Environ. Res. Commun.*, vol. 4, no. 7, 2022.
[9] I. Shahid and L. Venturi, Bittar, Antonio, "Analysis of Climate Changes and Its Impact on The Yield of Major Food Crops and Food Security in Pakistan," *Rev. Appl. Manag. Soc. Sci.*, vol. 19, no. 5, pp. 465–502, 2023.
[10] M. H. Al-Adhaileh and T. H. H. Aldhyani, "Artificial intelligence framework for modeling and predicting crop yield to enhance food security in Saudi Arabia," *PeerJ Comput. Sci.*, vol. 8, pp. 1–24, 2022.
[11] B. Rajeswari, Lalitha, S. Muheeth, Abdul, S. Naazleen, Vaseem, T. Kumar, Pavan, and V. Phanindraamouli, "Crop and Fertilizer Recommendation System," *Int. J. Food Nutr. Sci.*, vol. 11, no. 12, pp. 1675–1685, 2023.
[12] I. Ouedraogo, P. Defourny, and M. Vanclooster, "Application of random forest regression and comparison of its performance to multiple linear regression in modeling groundwater nitrate concentration at the African continent scale," *Hydrogeol. J.*, vol. 27, no. 3, pp. 1081–1098, 2019.
[13] L. Wickramasinghe, R. Weliwatta, P. Ekanayake, and J. Jayasinghe, "Modeling the Relationship between Rice Yield and Climate Variables Using Statistical and Machine Learning Techniques," *J. Math.*, vol. 2021, 2021.
[14] S. V. Joshua *et al.*, "Crop Yield Prediction Using Machine Learning Approaches on a Wide Spectrum," *Comput. Mater. Contin.*, vol. 72, no. 3, pp. 5663–5679, 2022.
[15] E. Khosla, R. Dharavath, and R. Priya, "Crop yield prediction using aggregated rainfall-based modular artificial neural networks and support vector regression," *Environ. Dev. Sustain.*, vol. 22, no. 6, pp. 5687–5708, 2020.
[16] S. Al-Eidi, F. Amsaad, O. Darwish, Y. Tashtoush, A. Alqahtani, and N. Niveshitha, "Comparative Analysis Study for Air Quality Prediction in Smart Cities Using Regression Techniques," *IEEE Access*, vol. 11, no. September, pp. 115140–115149, 2023.
[17] D. N. Sharma and S. I. M. Iqbal, "Applying Decision Tree Algorithm Classification and Regression Tree (CART) Algorithm to Gini Techniques Binary Splits," *Int. J. Eng. Adv. Technol.*, vol. 12, no. 5, pp. 77–81, 2023.
[18] K. Sevvanthi, S. Ganapathy, P. Penumadu, and K. T. Harichandrakumar, "Comparing the predictive performance of a decision tree with logistic regression for oral cavity cancer mortality: A retrospective study," *Cancer Res. Stat. Treat.*, vol. 6, no. 1, pp. 103–110, 2023.
[19] O. Saidani, L. J. Menzli, A. Ksibi, N. Alturki, and A. S. Alluhaidan, "Predicting Student Employability Through the Internship Context Using Gradient Boosting Models," *IEEE Access*, vol. 10, pp. 46472–46489, 2022.
[20] F. Yamamoto, S. Ozawa, and L. Wang, "eFL-Boost: Efficient Federated Learning for Gradient Boosting Decision Trees," *IEEE Access*, vol. 10, pp. 43954–43963, 2022.
[21] U. Singh, M. Rizwan, M. Alaraj, and I. Alsaidan, "A machine learning-based gradient boosting regression approach for wind power production forecasting: A step towards smart grid environments," *Energies*, vol. 14, no. 16, 2021.
[22] T. Siddique, D. Barua, Z. Ferdous, and A. Chakrabarty, "Automated farming prediction," *2017 Intell. Syst. Conf. IntelliSys 2017*, vol. 2018-Janua, no. September, pp. 757–763, 2017.
[23] S. Kumar, M. K. Sanyal, and A. Naskar, "Prediction of Rice Production in India Using Artificial Neural Network with Genetic Algorithm," *2020 Int. Conf.*

Comput. Sci. Eng. Appl. ICCSEA 2020, 2020.

[24] A. Javeed, S. Zhou, L. Yongjian, I. Qasim, and A. Noor, "An Intelligent Learning System based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection," *IEEE Access*, vol. PP, p. 1, 2019.

[25] Y. A. Ali, E. M. Awwad, and M. Al-razgan, "Hyperparameter Search for Machine Learning Algorithms for Optimizing the Computational Complexity," 2023.