Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer
Vol. 23, No. 3, July 2024, pp. 519∼528
ISSN: 2476-9843, accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: 10.30812/matrik.v23i3.3891

Optimizing Rain Prediction Model Using Random Forest and Grid Search Cross-Validation for Agriculture Sector

Ahmad Fatoni Dwi Putra , Muhamad Nizam Azmi , Satria Utama , I Gede Putu Wirarama Wedashwara Wirawan , Heri Wijayanto
Universitas Mataram, Mataram, Indonesia

Article Info

Article history:

Received February 21, 2024
Revised Juny 08, 2024
Accepted June 18, 2024

Keywords:

Agriculture
Cross-Validation
Grid Search
Rain Prediction
Random Forest

ABSTRACT

Agriculture is an important sector for all countries, especially Indonesia. However, agriculture is very dependent on rainfall. Erratic rainfall causes agricultural yields to be less than optimal. So it is necessary to use Machine Learning to predict rainfall. Several studies have conducted research related to rainfall prediction using machine learning algorithms and then compared which algorithm is the best. **The purpose** of this research is to increase the value of the R-squared rain prediction model by tuning parameters using Grid Search Cross-Validation. **The method** used in this research is Random Forest by tuning parameters to increase the R-squared value of the rain prediction model. The research process was divided into dataset collection, Data analysis, Preprocessing, Modeling, and optimization. The data used was obtained from the West Nusa Tenggara Meteorology, Climatology, and Geophysics Agency covering the period 2000 to 2023. **The results** of this research show that using parameter tuning using Grid Search Cross-Validation produces an R-squared of 0.0268 or an increase of 79,91%. These results have improved from those that initially only got an R-squared value from random forest of 0.1334. **This research concludes** that by tuning parameters using Grid Search Cross-Validation, you can increase R-Sqared from the previous model obtained from random forest. Further research can be carried out using other prediction algorithms, adding features, and adding data from other regions.

Corresponding Author:

Heri Wijayanto, +6285132614226,
Faculty of Engineering, Magister of Information Technology Program,
Universitas Mataram, Mataram, Indonesia,
Email: heri@unram.ac.id.

How to Cite:
A. Putra, M. Azmi, A. Husna RS, H. Wijayanto, and S. Utama, ”Optimizing Rain Prediction Model Using Random Forest and Grid Search Cross-Validation for Agriculture Sector”, *MATRIK: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer*, Vol. 23, No. 3, pp. 519-530, July, 2024.

1. INTRODUCTION

Agriculture is an important sector for all countries, especially Indonesia, but agriculture is very dependent on weather conditions [1]. Good rainfall predictions will influence the agricultural sector such as irrigation schedules, fertilizer use, and plant protection from extreme weather events, as well as in making planting schedules [2]. Traditional rainfall predictions often require assistance in processing large and complex data [3]. Therefore, it is necessary to use machine learning to make it easier to process this data more accurately [4, 5]. The random forest algorithm is a machine learning algorithm that can be used to obtain a rain prediction model [6, 7]. However, random forest is an old algorithm. Therefore, it is necessary to tune parameters using Grid Search Cross-Validation to get a more optimal rain prediction model [8].

The study conducted by [9] carried out optimization of Support Vector Machine (SVM) and Gradient Boosting Models by tuning parameters using Grid Search Cross-Validation. This research discusses detecting fake job postings with increasing accuracy by 0.17% compared to previous research. The study conducted by [10] conducted research related to heart disease prediction using SVM, Adaboost, Logistic Regression, Naive Bayes, and Random Forest. After performing hyperparameter tuning using Grid Search, Cross-Validation showed that random forest with parameter hyper-tuning produces the best model with an accuracy of 99.98%. The study conducted by [11] discusses customer feedback sentiment prediction using grid search tuning of hyperparameters in random forest. By tuning the hyperparameters, the number of maximum trees in the forest, and the depth of trees, accuracy can be increased from the initial 84.53% to 90.02%. The study conducted by [12] using random forest to Weather prediction was conducted in Tamil Nadu, India, utilizing global solar radiation (GSR) data in MJ/m2/day and wind speed in m/s. The obtained results yielded an R2 score of 0.97 and a Mean Squared Error (MSE) of 0.75. The researchers further [13] compared three machine learning methods: multiple linear regression, random forest regression, and replicated neural networks. The data used in the analysis consisted of daily rainfall data from the BMKG of Semarang City. The findings indicated that the Neural Network algorithm outperformed the others with an error rate of 13.055 for root mean squared error (RMSE) and 6.621 for mean absolute error (MAE). The study conducted by [14] for Forecasting Severe Weather utilized nine years of data from April 2003 to April 2012. The data was obtained from NOAA's Second-Generation Global Ensemble Forecast System Reforecast (GEFS/R). The researcher compared the prediction results from the Storm Prediction Center's (SPC's) convective outlook with SPC and random forest. The findings indicated that the SPC and random forest combination yielded significantly better results than SPC alone. The study conducted by [15] demonstrates that random forests can generate probabilistic severe weather hazard forecasts from numerical weather prediction (NWP) ensemble data. However, using two methods, the author attempted to compare how predictors should be generated from NWP ensembles. The data utilized in this study was derived from the convection-allowing High-Resolution Ensemble Forecast System, version 2.1 (HREFv2.1). The first method used predictors from individual ensemble members (IM), while the second used ensemble mean (EM) predictors at multiple spatial points. The results show that EM Random Forests (RFs) outperform identically constructed IM RFs for all risks, likely because EM predictors have less noise.

Some gaps have not been resolved by previous research. These gaps are found in the R-squared value produced, then the amount of data used to carry out training and testing. This research utilizes daily rainfall data from BMKG NTB collected from January 2000 to May 2023. However, the data obtained cannot be directly utilized due to various outliers and unrecorded data. Therefore, pre-processing is required to prepare the data for training and testing purposes. Then, in this research, Grid Search Cross-validation was used to optimize the model produced by random forest.

The difference between this research and previous research is in the amount of data, the quantity, and the variability of the data. This research aims to create an optimum model of rain prediction. After obtaining an optimal model, it can be used to predict rain so that farmers can make irrigation schedules, fertilizer usage, planting schedules, and the protection of crops from extreme weather events. This research produces an optimum rain prediction model after combining Random Forest with Gird Search Cross-Validation with an R-squared value of 0.0268. This research's contribution lies in improving the rain prediction model, which initially only used the random forest algorithm. We tuned parameters using Grid Search Cross-Validation to obtain a more optimal model. The results showed that the initial R-squared value was only 0.1334, and after tuning parameters, it improved to 0.0268. Therefore, it can be seen that the use of Grid Search Cross-Validation to optimize the Random Forest model increased the R-squared by 79.91%. Based on the problems outlined, the research writing format is as follows: The first chapter summarizes the study's history, reviews relevant studies, and describes the study's contributions. Chapter two discusses the study approach, the dataset, and the preprocessing methods. The third chapter addresses the findings and analysis. The final chapter summarizes the findings from this study.

2. RESEARCH METHOD

This research employs quantitative methodology to enhance the accuracy of the rainfall prediction models using random forest and optimization through Grid Search Cross-Validation. Figure 1 proposed process in this research includes data collection, preprocessing, modeling, and visualization. The study utilizes data from the Meteorology, Climatology, and Geophysics Agency (BMKG) in the West Nusa Tenggara region from January 2000 to May 2023.

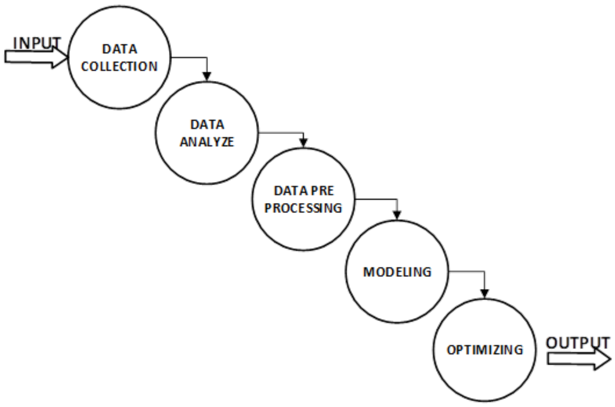

Figure 1. Research flow

2.1. Dataset Collection

Dataset Collection is one of the collection, measurement, and analysis stages. Rainfall data was collected from the Meteorology, Climatology, and Geophysics Agency (BMKG) website using the link:www.bmkg.go.id. The data was collected in the West Nusa Tenggara region between January 2000 and May 2023. The parameters for rainfall data include rainfall intensity, average humidity, duration of sunlight, average wind speed, temperature, and wind direction.

2.2. Data Analyses

The data analysis process uses various visualizations to analyze the data before preprocessing. The obtained data is then visualized in the form of graphs. This process aims to facilitate researchers in identifying abnormal data or unrecorded data, providing a clearer overview of the dataset.

2.3. Data Preprocessing

Preprocessing is a form of data mining that replaces aberrant data. This method solves the problem of inadequate, inconsistent, and inaccurate raw data. Collecting low-quality data might lead to low-quality results. Preprocessing data involves five main tasks: cleaning, reducing, scaling, transforming, and dividing [16]. This research involved cleaning and transforming data, considering missing values [17]. When creating operational data, there are two options for dealing with missing numbers. The first step is to delete data samples with missing values, as many data mining algorithms cannot handle these. This approach is used when the class label value does not exist or when the tuple has many attributes with empty values [18]. This approach can only be utilized when the fraction of missing information is not considerable. In the second stage, practice utilizing missing number imputation to replace missing data with inferred values. The most popular way is to fill in missing data using averages, medians, or feature forms. Data conversion aims to transform data into a format suitable for data mining. Data transformations include the square root, logarithm, arcsin, and inverse functions, among others.

2.4. Modelling

Random Forest Regressor algorithm was used in this study using a randomly constructed decision tree showed Figure 1. Random Forest uses training data as input data, and testing is used to test or evaluate the model or output [19]. This algorithm performs the training process on each decision tree. The tree can increase or expand with the amount of data. There are several important compensators in the RF algorithms applied in this study: (1) The algorithm selects a random sample from the available datasets. (2) makes a decision tree for each selected sample and then collects the prediction results from each decision tree. (3) For each prediction result, a selection is made (mode), or the most frequently appearing value, is used for the classification problem, and mean, or the average value, is applied for the regression problem. (4) The algorithm can choose the most selected prediction outcome as the final prediction.

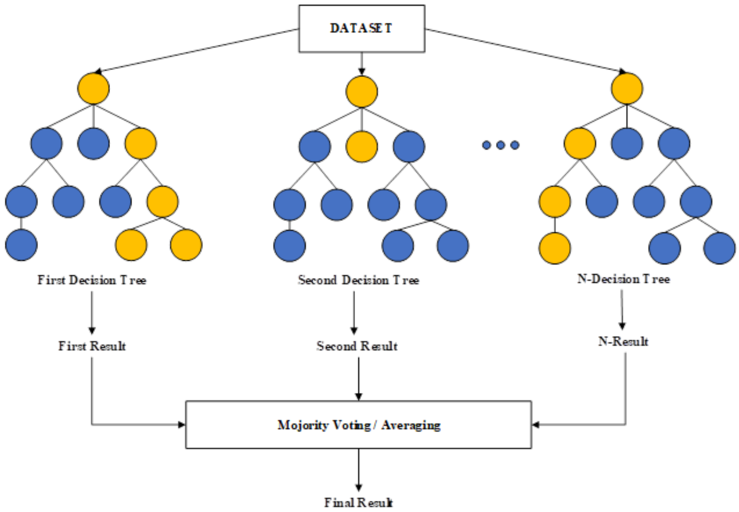

Figure 2. Modeling random forest

2.5. Optimizing

Grid Search Cross-validation (GSCV) is a method for optimizing hyperparameters. It optimizes model performance by identifying the ideal hyperparameter combinations. GSCV trains many models, each with a unique set of hyper-parameters. GSCV aims to train and assess models' performance using cross-validation. Finally, the model with the best results is chosen [20].

3. RESULT AND ANALYSIS

In this particular section, we perform a thorough evaluation of the methodology utilized. The model was created using the Python programming language implemented on the Google Collab platform. To test the model, we use the daily rainfall data acquired from the BMKG website specifically for the West Nusa Tenggara region. In this section, we conduct a performance evaluation of the methodology used. The developed model uses the Google Collab platform with the Python programming language. Testing is done using daily rainfall data obtained from the BMKG website for the West Nusa Tenggara region.

3.1. Dataset Collection

Rainfall data was obtained from the Meteorology, Climatology, and Geophysics Agency (BMKG). Which can be accessed at the URL link: https://drive.google.com/file/d/1z5FViQFnejzNXhzNBcdJL-sYj61hbeUP/view?usp=sharing . The data was gathered in the West Nusa Tenggara area between January 2000 and May 2023. The total amount of gathered data is 8,459. Rainfall data factors include rainfall intensity, average humidity, sunshine duration, average wind speed, temperature, and wind direction.

3.2. Data Analyses

Data analysis is a process of comprehending, evaluating, and making inferences from gathered data, known as data analysis. To improve decision-making or forecasting, the main goal of data analysis is to uncover hidden patterns, correlations, or trends in the data. Graphs are then created using the data acquired from BMKG to aid the analysis process. Figure 3 shows the daily air temperature data in the West Nusa Tenggara over the last 23 years. The red line represents the maximum temperature, the blue line indicates the minimum temperature, and the green line shows the average temperature obtained on that day. However, there's a lot of unrecorded data, like maximum and average temperature data. This led to the need for further data preprocessing to get better results.

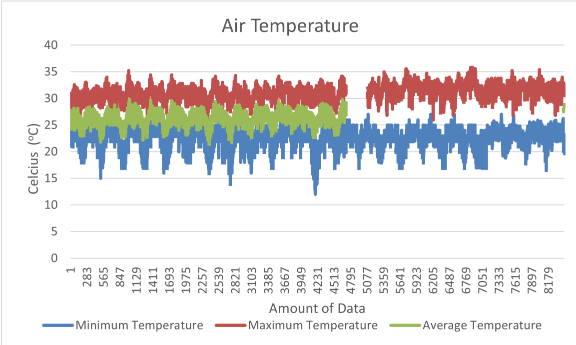

Figure 3. Average minimum and maximum temperature values

Figure 4 shows the average daily air humidity. The average air humidity data in Figure 4 ranges from 65% to 97%. However, some unrecorded data by BMKG render the data unusable. The subsequent step involves preprocessing by converting these empty data into the Not a Number (NaN) format. This process aims to enable the system to read the data. So, the process must be shut down or eliminated so that the data out layer does not interfere with data processing.

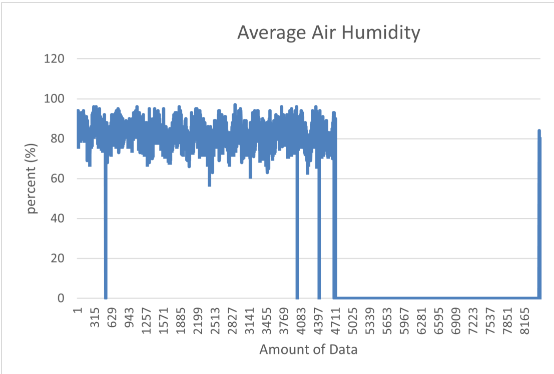

Figure 4. Average air humidity values

Figure 5 shows the daily rainfall in millimeters. Out of 8,459 daily rainfall data points, 624 have a value of 8888. These values of 8888 are categorized as outlier data.The next step is preprocessing, which is formatting these 8888 values in the Not a Number (NaN) format. The system's ability to read the data is the goal of this process. In order to prevent the data out layer from interfering with the data processing, the operation must be stopped or terminated.

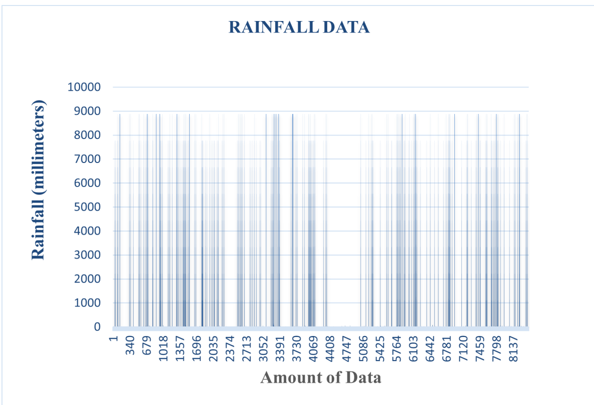

Figure 5. Rainfalls data

Figure 6 illustrates the duration of sunlight in a single day. The obtained data ranges from 0 to 12 hours. This indicates that there are days with sunlight throughout the day and also one day without sunlight. Additionally, there are unrecorded data points, rendering the system unable to read them. These unrecorded data points are then preprocessed by replacing them with NaN.

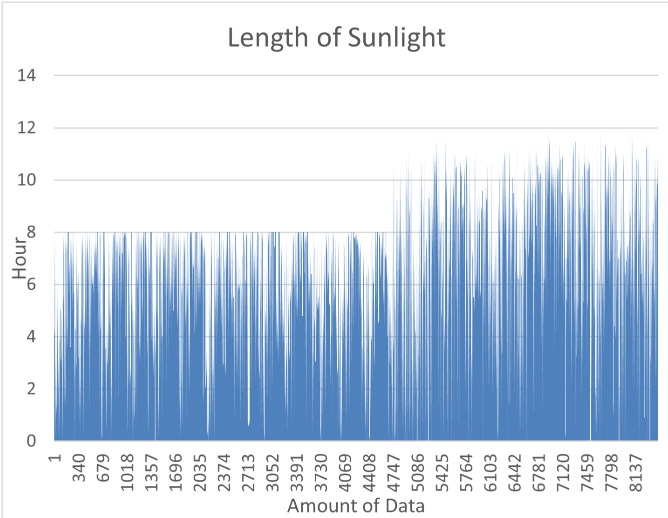

Figure 6. Length of sunlight

Figure 7 depicts the average daily wind speed trend in West Nusa Tenggara. Upon examining Figure 7, data shows an average wind speed reaching 13 m/s. Clearly, such wind speed is classified as very high, as wind speed is considered high when it ranges between 6-8 m/s [21]. Additionally, some data points are unrecorded by the BMKG system. These unrecorded data points are converted to NaN format within the system. The purpose of converting unrecorded data to NaN format is to indicate to the system that the data is either empty or unrecorded.

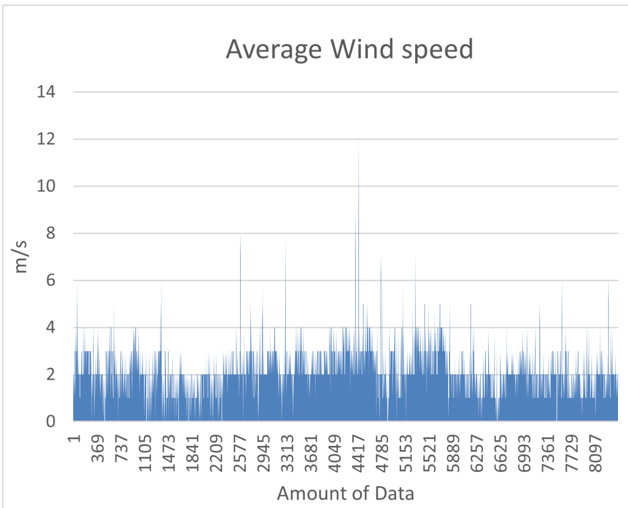

Figure 7. Average wind speed

3.3. Data Preprocessing

The weather data BMKG provides includes temperature, rainfall, wind speed, and others. However, this data must be preprocessed before it can be used in a machine-learning model. Some standard parameters frequently used in rain prediction include. Table

1 is a data collection from the BMKG NTB website from January 2000 to May 2023. Several features in the data collection are often used in rainfall predictions, including the first feature, Rainfall (RR). The amount of rain is measured in millimeters. This is a direct parameter that reflects the amount of water falling in rain. The second feature is Average Humidity (RH_avg). Humidity levels (%) can influence the possibility of rainfall. The third feature is the duration of sunlight (ss) in one day. The duration of sunlight (hours) can provide information about weather conditions and potential rainfall. The fourth feature is Average Wind Speed (ff_avg): Wind speed (m/s) can influence cloud patterns and rainfall. The fifth feature is air temperature, which consists of three categories, including (Tn, Tx, Tavg): Minimum temperature (Tn), maximum (Tx), and average air temperature (Tavg) in Celsius (^{o}C). Temperature changes can affect cloud formation and potential rainfall. The final feature is Wind Direction (ddd_x). Wind direction (degrees) can also influence cloud movement and rainfall.

Table 1. Dataset

Number	Date	Tn	Tx	Tavg	RH_avg	RR	ss	ff_x	ddd_x	ff_avg	ddd_car
1	1/1/2000	25	30,2	27,9	86,0	4,5	5,2	5,0	180,0	2,0	SW
2	2/1/2000	25	30,6	27	82,0	6	4,8	9,0	225,0	2,0	N
3	3/1/2000	24	31	27,7	80,0	5,5	8	7,0	180,0	2,0	SW
4	4/1/2000	24	30,6	26,3	86,0	0	3,2	5,0	225,0	1,0	SW
5	5/1/2000	24	31	25,7	90,0	4	4	12,0	315,0	2,0	S
…	…	…	…	…	…	…	…	…	…	…	…
8457	30/05/2023	19,6	32,2			8888	7,3	3,0	250,0	1,0	SW
8458	31/05/2023	22,2	32,4	28,9	75,0	0	10,4	4,0	190,0	2,0	SE
8459	30/06/2023	23,2	32,1	28,4	80,0	3,5	8,7	3,0	240,0	1,0	S

Table 2 shows the next step is to delete the feature that is not used in the data set. The feature that has been deleted is the date and direction of the windshield feature. Then, change the use of the comma mark (,) on the data sets using the dot mark (.). This is intended to enable the system to read the data already entered.Table 3 shows the process of changing out layers and unrecorded or empty data. Data-out layers and empty are modified using the Not a Number (NaN) format. This is intended to improve the accuracy of rain detection.

Table 2. Drop Feature

Number	Tn	Tx	Tavg	RH_avg	RR	ss	ff_x	ddd_x	ff_avg
1	25	30.2	27.9	86	4.5	5.2	5	180	2
2	25	30.6	27	82	6	4.8	9	225	2
3	24	31	27.7	80	5.5	8	7	180	2
4	24	30.6	26.3	86	0	3.2	5	225	1
5	24	31	25.7	90	4	4	12	315	2
…	…	…	…	…	…	…	…	…	…
8457	19.6	32.2			8888	7.3	3	250	1
8458	22.2	32.4	28.9	75	0	10.4	4	190	2
8459	23.2	32.1	28.4	80	3.5	8.7	3	240	1

Table 3. Change Empty and Out layer Data

Number	Tn	Tx	Tavg	RH_avg	RR	ss	ff_x	ddd_x	ff_avg
1	25	30.2	27.9	86	4.5	5.2	5	180	2
2	25	30.6	27	82	6	4.8	9	225	2
3	24	31	27.7	80	5.5	8	7	180	2
4	24	30.6	26.3	86	0	3.2	5	225	1
5	24	31	25.7	90	4	4	12	315	2
…	…	…	…	…	…	…	…	…	…
8457	19.6	32.2	NaN	NaN	1032.5183	7.3	3	250	1
8458	22.2	32.4	28.9	75	0	10.4	4	190	2
8459	23.2	32.1	28.4	80	NaN	8.7	3	240	1

3.4. Modeling

The data obtained from preprocessing amounts to 8,459 data points. This data is then divided into two parts, with 80% allocated for training data and 20% for testing data. The coefficient of determination, also known as R-square or R^2, is used to predict and explain future outcomes of a model. R^2 helps determine the goodness of one it for a model. It measures how well the observed results are replicated by the model based on the proportion of the total outcome range explained by the model. The formula for R^2 is given by 1. The best R^2 score is obtained from (1). The Random Forest Regressor achieves an R^2 value of 0.1334.

$$R^2 = \frac{\Sigma_{i=1}^{n} \mid yi - xi \mid^2}{\Sigma_{i=1}^{n} \mid yi - yi \mid^2} \tag{1}$$

3.5. Optimizing

Combining models to obtain the best R^2 value can be done using GSCV. The process involves several steps, including selecting a model, defining the hyperparameter space, splitting the dataset, creating combinations of hyperparameters, looping through hyperparameters, and finally selecting the best model. Using GSCV, an R^2 value of 0.0268 is obtained, as shown in Figure 8. Subsequently, this value is utilized for rain prediction.

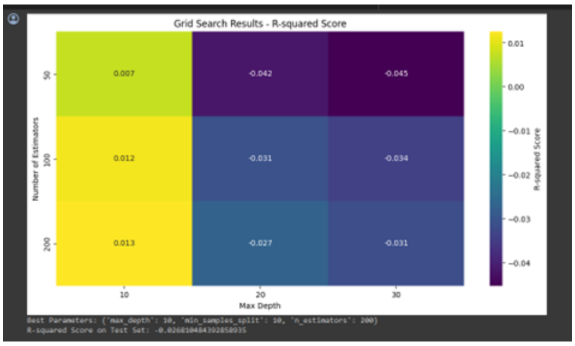

Figure 8. GSCV result

The discussion is based on previous research by [13], which also uses random forests to predict rainfall. This research uses seven features, among them max. temperature in degree Celsius, min.temperature in degree Celsius, surface pressure kpa, percentage relative humidity %, month number, latitude ^{o}(n) of the location, and longitude ^{o}(e) of the location. Then, research conducted by [14] compares three ML method between Multiple linear regression (MLR), random forest regression (RFR), and artificial neural networks (ANN). The data used in this study was collected from January 1, 2018, to March 20, 2021. The data contains eight factors that affect the amount of rainfall. The weather forecast includes minimum, maximum, and general temperatures, humidity, day length, maximum and general wind speeds, and wind direction. The goal is to predict rainfall. In this research, the result of ANN is superior to that of other methods, with an error rate of 13.055 for root mean squared error (RMSE) and 6.621 for mean absolute error (MAE).

Table 4. Comparison of Previous Research and Proposed Research

Research	Method	Feature	R^2 Score	RMSE	MAE
Research by [12]	Random Forest	7	0.97	-	-
Research by [13]	MLR	8	-	13,475	7,843
	RFR	8	-	13,403	6,643
	ANN	8	-	13,055	6,621
Proposed research	Random Forest and GSCV	9	0,0268	-	-

The findings of this research were to produce the best rain prediction model with an R-squared value of 0.0268. The results of this research are also supported by research [12], which also examines rain prediction. Compared to regression models and SVM machine learning models, the prediction results of the random forest machine learning model are more accurate with the results of R-Square 0.97. Table 4 is a comparison of previous research with this research. This research has advantages compared to research

[12], because the resulting R-squared value is 0.97 using 7 features, whereas this research produces an R-squared of 0.0268 using 9 features. However, the study [13] only compared these three methods, unlike this study, which optimized the model using Randon Forest and GSCV. Lastly, the research conducted by [14] and [15]; the two previous studies did not explicitly mention R-squared, RMSE, and MAE testing results. However, in their conclusion, they stated that using Random Forest machine learning in weather prediction can increase accuracy compared to not using machine learning methods in certain circumstances.

4. CONCLUSION

In this study, the Random Forest Algorithm was used to make the rain forecast model. The study demonstrates the effectiveness of the Random Forest Algorithm in developing a rain prediction model. Parameter tuningwas performed to determine the best combination of parameters to maximize the R-squared of the rain prediction model. The evaluation test findings show that parameter tuning with Grid Search Cross-Validation improves their modeling rain prediction performance. The result of this research shows that Random Forest before applying parameter adjustment, obtains R-squared 0.1334. After parameter tuning with Grid Search Cross-Validation, R-squared was enhanced by 0.0268 or 79.91%. For further research, we can compare other rain prediction methods, add features, and combine datasets from other regions.

5. ACKNOWLEDGEMENTS

Thanks to the Department of Master of Information Technology, University of Mataram, for supporting and striving to finish this research. We are grateful to the MATRIK publication for publishing our study results. This finding may have implications for future research.

6. DECLARATIONS

AUTHOR CONTIBUTION

Ahmad Fatoni Dwi Putra, the first author, is a data collector and analyst. Muhammad Nizam Azmi, the second author, is the model maker. Satria Utama, Heri Wijayanto, and Asmaul Husna RS, authors third, fourth, and fifth, analyze and help to give a correction of each model result.

FUNDING STATEMENT

This work was self-funded, without external financial support for design, data collection, analysis, or interpretation. The authors covered all expenses themselves.

COMPETING INTEREST

Further research can be directed to make best model either in the same field as this research or different field using big data processing and machine learning.

REFERENCES

[1] R. Meenal, K. Kailash, P. A. Michael, J. J. Joseph, F. T. Josh, and E. Rajasekaran, "Machine learning based smart weather prediction," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 28, no. 1, pp. 508–515, 10 2022, https://doi.org/10.11591/ijeecs.v28.i1.pp508-515.

[2] G. Meti, R. K. G. Krishnegowda, and G. S. Swamy, "Rainfall analysis and prediction using ensemble learning for karnataka state," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 32, no. 2, pp. 1187–1198, 11 2023, https://doi.org/10.11591/ijeecs.v32.i2.pp1187-1198.

[3] T. P. Agyekum, P. Antwi-Agyei, and A. J. Dougill, "The contribution of weather forecast information to agriculture, water, and energy sectors in east and west africa: A systematic review," pp. 1–14, 8 2022, https://doi.org/10.3389/fenvs.2022.935696.

[4] E. Erlin, Y. Desnelita, N. Nasution, L. Suryati, and F. Zoromi, "Dampak smote terhadap kinerja random forest classifier berdasarkan data tidak seimbang," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 3, pp. 677–690, 7 2022, https://doi.org/10.30812/matrik.v21i3.1726.

[5] P. R. Sihombing and I. F. Yuliati, "Penerapan metode machine learning dalam klasifikasi risiko kejadian berat badan lahir rendah

di indonesia," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 20, no. 2, pp. 417–426, 5 2021, https://doi.org/10.30812/matrik.v20i2.1174.

[6] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *Stata Journal*, vol. 20, no. 1, pp. 3–29, 3 2020, https://doi.org/10.1177/1536867X20909688.

[7] H. Yao, X. Li, H. Pang, L. Sheng, and W. Wang, "Application of random forest algorithm in hail forecasting over shandong peninsula," *Atmospheric Research*, vol. 244, pp. 1–39, 11 2020, https://doi.org/10.1016/j.atmosres.2020.105093.

[8] V. Shalamov, V. Efimova, and A. Filchenkov, "Faster hyperparameter optimization via finding minimal regions in random forest regressor," vol. 212, no. 3. Elsevier B.V., 2022, pp. 378–386, https://doi.org/10.1016/j.procs.2022.11.022.

[9] R. Rofik, R. A. Hakim, J. Unjung, B. Prasetiyo, and M. A. Muslim, "Optimization of svm and gradient boosting models using gridsearchcv in detecting fake job postings," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 23, no. 2, pp. 419–430, 3 2024, https://doi.org/10.30812/matrik.v23i2.3566.

[10] S. Rasheed, G. K. Kumar, D. M. Rani, M. V. Kantipudi, and M. Anila, "Heart disease prediction using gridsearchcv and random forest," *EAI Endorsed Transactions on Pervasive Health and Technology*, vol. 10, no. 1, pp. 1–8, 1 2024, https://doi.org/10.4108/eetpht.10.5523.

[11] S. George and B. Sumathi, "Grid search tuning of hyperparameters in random forest classifier for customer feedback sentiment prediction," pp. 173–178, 2020, https://doi.org/10.14569/IJACSA.2020.0110920. [Online]. Available: www.ijacsa.thesai.org

[12] R. Meenal, P. A. Michael, D. Pamela, and E. Rajasekaran, "Weather prediction using random forest machine learning model," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, no. 2, pp. 1208–1215, 5 2021, https://doi.org/10.11591/ijeecs.v22.i2.pp1208-1215.

[13] C. D. Usman, A. P. Widodo, K. Adi, and R. Gernowo, "Rainfall prediction model in semarang city using machine learning," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 30, no. 2, pp. 1224–1231, 5 2023, https://doi.org/10.11591/ijeecs.v30.i2.pp1224-1231.

[14] A. J. Hill, G. R. Herman, and R. S. Schumacher, "Forecasting severe weather with random forests," *Monthly Weather Review*, vol. 148, no. 5, pp. 2135–2161, 5 2020, https://doi.org/10.1175/MWR-D-19-0344.1.

[15] E. D. Loken, A. J. Clark, and A. McGovern, "Comparing and interpreting differently designed random forests for next-day severe weather hazard prediction," *Weather and Forecasting*, vol. 37, no. 6, pp. 871–899, 6 2022, https://doi.org/10.1175/waf-d-21-0138.1.

[16] S. Efendi and P. Sihombing, "Sentiment analysis of food order tweets to find out demographic customer profile using svm," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 3, pp. 583–594, 7 2022, https://doi.org/10.30812/matrik.v21i3.1898.

[17] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data," pp. 1–17, 3 2021, https://doi.org/10.3389/fenrg.2021.652801.

[18] Y. Shaikh, V. Parvati, and S. R. Biradar, "Early disease prediction algorithm for hypertension-based diseases using data aware algorithms," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 27, no. 2, pp. 1100–1108, 8 2022, https://doi.org/10.11591/ijeecs.v27.i2.pp1100-1108.

[19] R. Siddabathuni, S. Palanivel, and G. L. N. Murthy, "Alzheimer image registration using hybrid random forest and deep regression network algorithm," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 33, no. 2, pp. 824–831, 2 2024, https://doi.org/10.11591/ijeecs.v33.i2.pp824-831.

[20] M. Adnan, A. A. S. Alarood, M. I. Uddin, and I. ur Rehman, "Utilizing grid search cross-validation with adaptive boosting for augmenting performance of machine learning models," *PeerJ Computer Science*, vol. 8, no. 1, pp. 1–29, 2022, https://doi.org/10.7717/PEERJ-CS.803.

[21] N. W. Hesty, D. G. Cendrawati, A. Aminuddin, B. Pranoto, and A. Fahim, "Estimasi potensi energi angin indonesia menggunakan model weather research and forecast - four dimension data assimiliation (wrf-fdda)," *Jurnal Sains Dirgantara*, vol. 19, no. 2, pp. 11–20, 2022, https://doi.org/DOI:10.30536/j.jsd.2022.v19.a3614.