JURNAL SITEKNIK
Sistem Informasi, Teknik dan Teknologi Terapan
Volume 1, No.2, Juni 2024, hal. 90-101
E-ISSN: 3032-3991 | P-ISSN: xxxx.xxxx

Comparison Accuracy of C4.5 Algorithm and K-Nearest Neighbors for Rainfall Classification

Muhammad Fauzan Nasrullah [1]⊠, Rd. Rohmat Saedudin[2], Faqih Hamami[3]

[1] S2 Sistem Informasi, Telkom University, Indonesia
[2] S2 Sistem Informasi, Telkom University, Indonesia
[3] S2 Sistem Informasi, Telkom University, Indonesia

Abstract

Indonesia has a predominantly tropical climate, hence Indonesia experiences limited temperature variations, but has diverse rainfall variations. The variability of rainfall is also inseparable from the impact it has on various aspects of human life and business activities. Therefore, rainfall information is an important aspect in decision making. However, of course, there are stages and methods needed to carry out the analysis process. Therefore, this study looked for the best method between C4.5 and K-Nearest Neighbors which included algorithms in data mining to classify rainfall data. Both algorithms are used to build classification models based on relevant attribute attributes. Then, testing and evaluating both models using various metrics such as Accuracy, Precision, Recall and F1-Score were carried out. In this study also applied Hyperparameter Tuning with the RandomizeSearchCV method to get the best parameters to get maximum accuracy values. The results showed good accuracy values for both algorithms, in the sense that both algorithms were able to classify rainfall based on Indonesia's climate well. Based on the accuracy values obtained with the default parameters of both algorithms, C4.5 produces a higher accuracy value of 81.42%, while K-Nearest Neighbors is only 78.10%. However, after using the best parameters resulting from the application of RandomizedSearchCV Hyperparameter Tuning, a significant change in accuracy value occurred in K-Nearest Neighbors which was found to be 83.37%, while C4.5 increased to 82.56%.

Riwayat Artikel :	**Kata Kunci :**	***Keywords:***
Diserahkan :01-06 2024 Direvisi :06-06-2024 Diterima :10-06-2024	*Curah hujan, iklim Indonesia, Klasifikasi, C4.5, K-Tetangga Terterdekat, Penambangan Data*	*Rainfall, Climate Indonesia, Classificiation, C4.5, K-Nearest Neighbor, Data Mining*

Corresponding Author :
Muhammad Fauzan Nasrullah
S2 Sistem Informasi, Telkom University, Indonesia
Jl. Telekomunikasi. 1, Terusan Buahbatu - Bojongsoang, Telkom University, Sukapura, Kec. Dayeuhkolot, Kabupaten Bandung, Jawa Barat 40257
Email : rdrohmat@telkomuniversity.ac.id

INTRODUCTION

A form of ecosystem imbalance that occurs on earth or global warming (global warming) occurs due to the process of increasing the average temperature of the atmosphere, sea, and land on earth. In addition to directly affecting the increase in surface temperature, the phenomenon of global warming also results in changes in climate patterns, which affect human life. Climate can be explained as weather conditions within an area over a longer period [1]. Data on climate or weather conditions is very crucial and inseparable information from various human activities, especially in sectors such as agriculture, plantations, forestry, transportation, irrigation, environmental protection, mining, disaster mitigation efforts, and other sectors [2]. As the problem raised in this study, rainfall has various significant impacts on climate variability. According to Sipayung [3], the climatic characteristics of an area can also be seen from the variation of rainy weather. Rainfall variability is also inseparable from the impacts given such as causing several natural disasters such as floods and droughts. This makes the surrounding community feel such a significant impact from the variability of rainfall.

In July 2022, West Lombok Regency experienced low rainfall, resulting in sever-al regions facing drought problems. Furthermore, based on data compiled by the Center for Disaster Management Operations Control (Pusdalops PB), there were 10 flood events caused by high rainfall in Probolinggo Regency, which occurred since March 2020 according to the Regional Disaster Management Agency (BPBD) in 2022. Of course, such phenomena can hinder various human activities in diverse sectors, and one of the sectors most affected by fluctuations in rainfall is the agricultural sector, because it causes the shifting dynamics of rainy and dry seasons that create an increased risk of crop failure [4]. Based on events that have occurred, climate/weather information is now an important aspect of infor-mation systems [5]. This is because it has an impact on various aspects of human life and business activities, such as agricultural companies that benefit from pre-dicting accurate rainfall to efficiently manage planting schedules, crop mainte-nance, and other agricultural activities [6]. In addition, companies in the field of renewable energy also benefit from rainfall information in optimizing energy production and predicting the potential availability of energy sources [7]. And there are many other companies in various fields that can be helped with climate / weather information for more precise and efficient decision making [8]. There-fore, data or information about climate / weather has a very strategic value for decision making [9]. However, of course, various stages and methods are needed to carry out the process of analyzing the climate of a region, one of which is the data mining process.

Data mining is a series of steps to obtain meaningful information from within a large-scale database warehouse [10]. The concept of data mining can also be described as the process of extracting innovative information from large amounts of data to provide support in the decision-making process. The term data mining is now gaining popularity along with the development of today's all-digital era. However, there are some challenges in implementing data mining today. As data structures tend to be heterogeneous coming from various sources, hence the need for initial processing and merging of data first [11]. Then the development of methods also led to the birth of many algorithms that are complicated and require many parameters to achieve optimal results. Apart from that, of course data min-ing will be increasingly needed currently to come, therefore researchers want to prove that the implementation of data mining can provide a good solution for the classification of rainfall based on climate in Indonesia [12]. However, the many types of algorithms in data mining certainly have their own positive and negative values for each algorithm [13]. Therefore, the study aims to find an algorithm that can classify rainfall with the highest accuracy from two classification algorithm options, namely K-Nearest Neighbors and C4.5.

From some previous research that has been done, it can be said that both algo-rithms are good enough to classify a data, but with differences in datasets, tech-niques and tools carried out may be able to make one of the two algorithms some-times superior and not superior. Therefore, in this study, the K-Fold Cross Vali-dation validation technique was used in validating the two models to see the con-sistency of the two models. In addition, the author also uses Hyperparameter Tuning with the Randomized search CV approach to improve the performance of both models in order to find the most appropriate parameter according to the dataset used. So that in the end the author can compare the performance of the C4.5 algorithm and K-Nearest Neighbors in classifying rainfall based on Indone-sia's climate.

RESEARCH METHOD

Rainfall

Precipitation is the amount or height of rainwater that accumulates over a flat area, without evaporation, infiltration, or flow over a period. The unit commonly used in measuring rainfall is millimeters or inches, but in Indonesia, the unit commonly used for rainfall is millimeters (mm). In general, if there is rainfall of one millimeter, then on a flat surface of one square meter, water will collect as high as one millimeter or equivalent to one liter. According to BMKG there are five rainfall criteria as follows [14].

Table 1. Categories Rainfall

Status	Rainfall Range
Cloudy	0 mm/day

Very Light	<5 mm/day
Light	5 - 10 mm/day
Currently	21 - 50 mm/day
Heavy	51 – 100 mm/day

C4.5 Algorithm

The C4.5 algorithm is used to form the decision tree structure, which is basically a classification and prediction technique that is widely known and used in vari-ous situations. Decision trees have the goal of extracting information from data and revealing hidden relationships between variables or attributes used. The C4.5 algorithm, which is a development of the ID3 algorithm initiated by J. Ross Quin-lan, also belongs to the group of algorithms capable of building decision tree structures [15].

K-Nearest Neighbors Algorithm

K-Nearest Neighbors, commonly abbreviated as KNN, is one algorithm that aims to classify new objects using learning data (neighbors) that have the closest dis-tance to the object to be classified. The calculation of the degree of proximity or farhereness from these neighbors is often done through the Euclidean [16] dis-tance method. KNN also has advantages such as being tough and effective against learning data that has a lot of noise and large size [17].

Confusion Matrix

Confusion matrix is an instrument used to assess the effectiveness of classifica-tion models in the context of data mining or machine learning tasks. It is used to describe a comparison between the results of the model classification with the actual class of the data tested. Confusion matrix is formed by four elements or values, namely True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN), this matrix is used to compute several model performance evaluation metrics, such as accuracy, precision, recall, and F1-score [18].

K-Fold Cross Validation

K-fold cross-validation is a statistical approach used to measure the efficiency of a model or algorithm that has been developed. In the training process, the data set will be separated into two subsets, namely training data and validation data. The model will be instructed using the training data, and its effectiveness will be test-ed with validation data in a series of iterations totaling k-fold [19].

SMOTE

SMOTE, short for Synthetic Minority Oversampling Technique, is a method to deal with the problem of imbalance in the number of samples in the minority class by producing synthetic samples. This technique works by adding synthetic samples so that the number of samples in the minority and majority classes is balanced [20].

Hyperparamater

Hyperparameters are parameters that are not calculated by the model itself during the training process but must be determined by the user before the model is trained. In machine learning and deep learning, algorithms have hyperparameters that affect the way models learn and operate. Proper hyperparameter settings are critical in achieving optimal model performance. A good hyperparameter deter-mination process can involve experimentation and cross-validation to find the hyperparameter combination that produces the best results on never-before-seen data [21-24].

Data Collection

In this study, researchers compared the level of accuracy in two classification algorithms, namely K-Nearest Neighbors and C4.5. Using Indonesian climate data that has been obtained through the data provider site, Kaggle with the address https://www.kaggle.com/datasets/greegtitan/indonesia-climate/. The data used is data with the last 5 years from 2016 to 2020 [25, 26].

Data Identification

At this stage, researchers understand and analyze data that has been successfully obtained about the type of data, data format, and data structure. It also involves selecting data by selecting attributes that are relevant or necessary for subse-quent processing. At this stage, attributes that are irrelevant or do not have a sig-nificant impact on the purpose of classification can be removed from the data [27]. The attributes that will be used for the next data processing process are Tavg, RH_avg, RR, ss and ddd_x, these attributes are selected based on factors that can affect rainfall. The results of data selection based on the attributes used can be seen in table 2.

Table 2. Research Variables

Variable	Nama	Definition
Independent Variables (X)	Tavg	Average temperature (
	RH_avg	Average humidity (%)
	ss	Duration of sunlight (hour)
	Ddd_x	Wind direction with maximum speed (deg)
		Rainfall (mm) categories :
		0 = light
Dependent Variables (Y)	RR	1 = moderate
		2 = heavy
		3 = very heavy

Data Preprocessing

At this stage, researchers carry out the data preparation process before further processing. The purpose of this stage is to convert raw data into data that is more ready to be used or can be processed by computer systems. In the Data Prepro-cessing stage, the stage starts from data cleansing by removing or removing in-complete, inaccurate, or duplicate data from the data used. If there is indeed in-complete or null data, a search will be carried out for the mean and / or median of each column to be included in the null data [28, 29]. There are still null values (NaN) in the dataset, if calculated from each column it produces the number of null values as in table 3.

Table 3. Number of values null every variable

Variable Name	Number of Null Values
Tavg	29355
RH_avg	30286
RR	76390
Ss	21390
Ddd_x	2215
ff_avg	1468

After calculating the number of null values, the researcher then imputed the missing value to overcome data that had no value (null). In this study, missing value imputation is done by replacing or filling in the missing value (null) with the average (mean) and median of each same variable. The results of such processes can be seen in tables 4 and 5.

Table 4. Dataset before data cleansing

Tavg	RH_avg	RR	Ss	Ddd_x	Ff_avg
27.6	81.0	0.0	7.0	110.0	7.0
28.0	84.0	Nan	10.0	70.0	3.0
27.2	87.0	6.5	4.5	110.0	4.0
28.1	78.0	Nan	3.0	260.0	2.0

28.4	81.0	Nan	6.5	260.0	2.0

Table 5. Dataset after data cleansing

Tavg	RH_avg	RR	Ss	Ddd_x	Ff_avg
27.6	81.0	0.0	7.0	110.0	7.0
28.0	84.0	1.6	10.0	70.0	3.0
27.2	87.0	6.5	4.5	110.0	4.0
28.1	78.0	1.6	3.0	260.0	2.0
28.4	81.0	1.6	6.5	260.0	2.0

The next stage is data labeling, aiming to label the dataset used with certain criteria. Labels are given to RR or precipitation attributes categorized based on rainfall probability by BMKG. The data can be seen in table 6. After the data labelling process on the dataset based on certain criteria, the display of data that has been labeled will be as in table 7.

Table 6. Categories Rainfall Labels

Rainfall Range	Label	Categories Rainfall
0.5 - 20 mm/day	1	Light
21 - 50 mm/day	2	Keep
51 - 100 mm/day	3	Dense
> 100 mm/day	4	Very Dense

Table 7. Dataset after labelling

Tavg	RH_avg	RR	Ss	Ddd_x	Ff_avg	label
27.6	81.0	0.0	7.0	110.0	7.0	1.0
28.0	84.0	1.6	10.0	70.0	3.0	1.0
27.2	87.0	6.5	4.5	110.0	4.0	1.0
28.1	78.0	1.6	3.0	260.0	2.0	1.0
28.4	81.0	1.6	6.5	260.0	2.0	1.0

In the next stage, the author checks whether the data on the amount of sample data used is balanced or not. The results of the examination found that the amount of data for each class was not balanced. Therefore, the author performs data oversampling, which is a technique in data processing used to overcome the problem of class imbalance in the dataset. In this study, the authors used the SMOTE approach to conduct data oversampling.

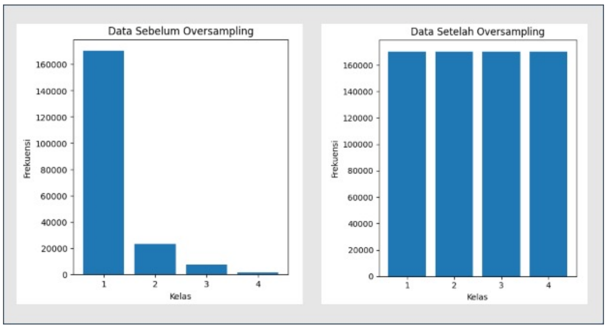

Figure 1. Before and after data count graph oversampling

Modelling and Evaluation

After the data preprocessing process is complete, it can be interpreted that the data is ready to be processed for data modeling. In this study, the author used the C4.5 and K-Nearest Neighbors algorithms [27-30]. The tools used are Google Colab with python programming language. After successfully creating a model, the next stage is to conduct an evaluation stage to test the performance of each algorithm. The evaluation stage is carried out using the Confusion Matrix, Preci-sion, Recall and F1-Score methods.

RESULT AND DISCUSSION

C4.5 Implementation

In implementing the C4.5 algorithm in this study, the author uses the Decision-TreeClassifier function which is part of the tree module in the scikit-learn library which is used to build a decision tree-based classification model. Initial testing of the C4.5 algorithm using an 80:20 splitting data ratio obtained an accuracy result of 81.42%. The test result validation method used in this study is K-Fold Cross Validation. The K-Fold Cross Validation test on the C4.5 algorithm uses a k value of 10 folds and a ratio of 80:20. The validation results can be seen in table 8 and figure 2 below.

Tabel 8. K-fold cross validation c4.5 algorithm

Fold to-	1	2	3	4	5	6	7	8	9	10
Accuracy	0.829	0.822	0.822	0.822	0.822	0.822	0.822	0.822	0.822	0.822
Average Accuracy	82.36%									

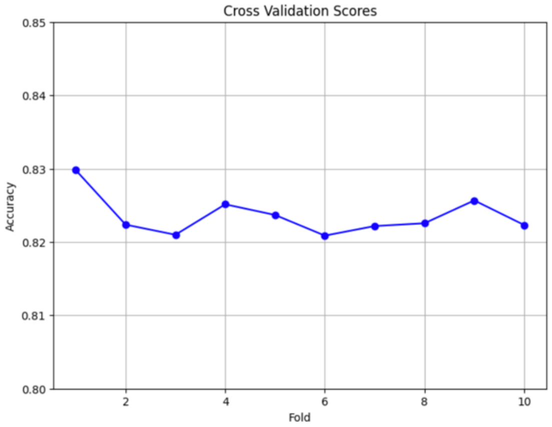

Figure 2. K-fold cross validation c4.5

After testing, the accuracy of K-Fold Cross Validation was obtained on the C4.5 algorithm with an average accuracy of 82.36%. It can also be seen that the trend is relatively stable, in the sense that model performance measured by the K-Fold Cross Validation method shows slight variation between each validation itera-tion. With good stability, it is also very likely that the model is not overfitting (too much in line with the training data) or underfitting (not understanding the patterns in the data enough).

In the next stage, the author performs the Hyperparamater Tuning process to im-prove performance and generalization of the model by finding the optimal com-bination of hyperparameters, hoping to produce higher accuracy values from the initial test. In this study, the hyperparameter tuning method used was Random-izedsearch Cross Validation. The results of hyperparameter tuning and accuracy of the C4.5 algorithm after using the results of hyperparameter tuning can be seen in table 9.

Tabel 9. Research Variables

No	Hyperparameters	Values

1	min_samples_split	3
2	Criterion	Entropy
3	max_depth	256
4	min_samples_leaf	1
Accuracy		82.68%

It can be seen in table 10 that from the test results by applying hyperparameter tuning with the Randomizedsearch Cross Validation method on the C4.5 algo-rithm resulted in an increase in accuracy value of 82.68%. Furthermore, the au-thor conducted a performance evaluation process of the C4.5 algorithm using the Confusion Matrix, Precission, Recall and F1-Score methods. The stage is preced-ed by conducting an evaluation with the Confusion Matrix. The results of the Confusion Matrix from the C4.5 algorithm can be seen in table 10.

Tabel 10. Confusion matrix c4.5 algorithm

No	Pred 1	Pred 2	Pred 3	Pred 4
Actual 1	28230	4870	1350	196
Actual 2	4758	24290	3849	983
Actual 3	1557	3802	27959	936
Actual 4	246	736	723	32206

After getting the results from the Confusion Matrix, the bully will calculate the Recall, Precision, and F1-Score values from the C4.5 algorithm. Because the class type in this study is multiclass, therefore calculations are made for each class individually. This helps in evaluating the extent to which the model can accurate-ly classify the data on each class, identify classes that may have classification problems, and compare model performance between different classes. The calcu-lation results with the average of each class can be seen in table 12.

Table 11. C4.5 algorithm evaluation results

Metric	Value
Precision	82%
Recall	82.25%
F1-Score	82.25%

K-Nearest Neighbors Implementation

In the previous stage the author has succeeded in implementing the C4.5 algo-rithm, the next stage the author will do a similar stage to the K-Nearest Neighbors algorithm. Initial testing of the K-Nearest Neighbors algorithm using an 80:20 splitting data comparison ratio obtained an accuracy result of 78.10%. The vali-dation phase is also carried out with K-Fold Cross Validation. K-Fold Cross Vali-dation testing on the K-Nearest Neighbors algorithm uses a k value of 10 folds and a ratio of 80:20. The results of K-fold Cross Validation on the C4.5 algorithm can be seen in table 12 and figure 3 below.

Tabel 8. K-fold cross validation knn algorithm

Fold to-	1	2	3	4	5	6	7	8	9	10
Accuracy	0.836	0.833	0.835	0.835	0.833	0.832	0.831	0.832	0.832	0.834
Average Accuracy	83.37%									

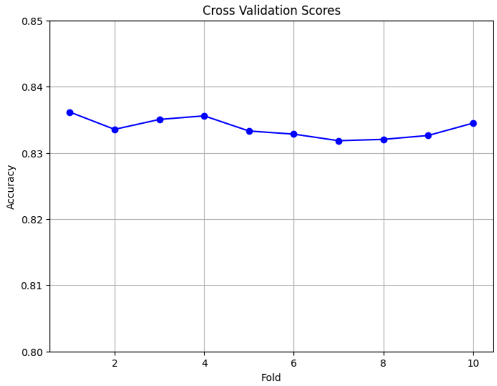

Figure 3. K-fold cross validation c4.5

After testing, the accuracy of K-Fold Cross Validation was obtained on the K-Nearest Neighbors algorithm with an average accuracy of 83.37%. The visualiza-tion results from K-Fold Cross Validation on the K-Nearest Neighbors algorithm also show a stable trend. This means model performance measured by the K-Fold Cross Validation method shows slight variation between each validation iteration.

Table 13. Confusion Matrix Knn Algorithm

No	Hyperparameters	Values
1	weights	Uniform
2	n_neighbors	1
3	Metric	Minkowski
4	leaf_size	41
Accuracy		83.94%

Furthermore, the author also did the same thing by tuning hyperparameters on the K-Nearest Neighbors algorithm, and getting parameter results and accuracy re-sults as in table 13 above. After testing by applying hyperparameter tuning of the Randomizedsearch Cross Validation method on the K-Nearest Neighbors algo-rithm to a ratio of 80:20, it was proven that the accuracy value had increased significantly from the accuracy value in the previous initial test. The increase in accuracy resulted in a value of 83.94%.

Furthermore, the author also conducted a performance evaluation process of the K-Nearest Neighbors algorithm using the Confusion Matrix, Precission, Recall and F1-Score methods. The stage is preceded by conducting an evaluation with the Confusion Matrix. The results of the Confusion Matrix from the K-Nearest Neighbors algorithm can be seen in table 14.

Tabel 14. Knn algorithm evaluation results

No	Pred 1	Pred 2	Pred 3	Pred 4
Actual 1	23560	6741	3046	896
Actual 2	3662	27051	2367	800
Actual 3	1369	1783	30511	591
Actual 4	176	281	175	33279

After getting the results of the Confusion Matrix, the author will calculate the Recall, Precision, and F1-Score values from the K-Nearest Neighbors algorithm. Because the class type in this study is multiclass, therefore calculations are made for each class individually. This helps in evaluating the extent to which the mod-el can accurately classify the data on each class, identify classes that may have classification problems, and compare model performance between different clas-ses. The calculation results with the average of each class can be seen in table 15.

Table 11. Knn algorithm evaluation results

Metric	Value
Precision	83.50%
Recall	83.40%
F1-Score	83%

Evaluation of Algorithm Comparison

Table 17 displays the comparison results of the two algorithms, namely the C4.5 algorithm and the K-Nearest Neighbors algorithm. First, a comparison was made on the initial accuracy value, in this comparison the C4.5 algorithm was superior with an accuracy value of 81.42% compared to the accuracy value obtained by the K-Nearest Neighbors algorithm of 78.10%. After that the second comparison was made on the K-Fold Cross Validation value of the two algorithms, in this comparison the K-Nearest Neighbors algorithm excelled with a value of 83.37% and 82.56% for the C4.5 algorithm Furthermore, the third comparison was made on the accuracy value after applying Hyperparameter Tuning with the CV Ran-domizedsearch method, using the best parameter, the accuracy value on the algo-rithm K-Nearest Neighbors experienced a significant increase that exceeded the initial accuracy value and was greater than the C4.5 algorithm. The accuracy val-ue obtained was 83.94% for the K-Nearest Neighbors algorithm, while for the C4.5 algorithm it was 82.68%. Furthermore, the author also compared the evaluation value of the model of the two algorithms, where the Precision value for the K-Nearest Neighbors algorithm obtained 83.25%, while for the C4.5 algorithm obtained 82%. Furthermore, the Recall value for the K-Nearest Neighbors algo-rithm obtained a figure of 83.50%, while for the C4.5 algorithm it obtained a fig-ure of 82.25%. While the F1-Score value for the K-Nearest Neighbors algorithm obtained 83% and 82.25% for the C4.5 algorithm. From the results of the compar-ison of the three model evaluations on both algorithms, the K-Nearest Neighbors algorithm produces better accuracy values than the C4.5 algorithm.

With the accuracy results obtained, it is proven as in previous studies that the two algorithms, namely the C4.5 algorithm and K-Nearest Neighbors are indeed able to classify rainfall data well, this is because the accuracy value obtained is above 80%. Furthermore, the addition of the use of K-Fold Cross Validation can also ensure that both models are consistent and do not experience underfitting and overfitting. Then the addition of the use of Hyperparameter Tuning can also contribute to efforts to improve the performance of both models to obtain maximum accuracy values. That way, the results that have been obtained by adding the K-Folde Cross Validation and Hyperparameter Tuning methods to this study can increase the level of accuracy of the results and minimize the error value.

Table 11. C4.5 algorithm evaluation results

	C4.5	Value
Accuracy	81.42%	78.10%
K-Fold Cross Validation	82.56%	83.37%
Turning Hyperparameters (accuracy)	82.68%	83.94%
Precision	82%	83.50%
Recall	82.25%	83.50%
F1-Score	82.25%	83%

CONCLUSIONS AND SUGGESTIONS

Conclusion

Based on the overall results of the tests carried out, the two algorithms can be said to be good algorithms in classifying rainfall based on Indonesia's climate. This is because all the end results of both algorithms produce numbers above 80%. However, one thing that can be of concern is the initial test, where before using Hy-perparameter Tuning on the C4.5 algorithm produced an accuracy value higher than the accuracy value of the K-Nearest Neighbors algorithm. This can mean that by using the default parameters of both algorithms, the C4.5 algorithm is superior by producing higher accuracy values. But after using Hyperparameter Tuning, the K-Nearest Neighbors algorithm has a higher number than the C4.5 algorithm in terms of accuracy, K-Fold Cross Validation, Precision, Recall and F1-Score.

REFERENCE

S. Prawirowardoyo, Meteorology. Bandung: ITB, 1996.

N. Sunarmi et al., “Analisis Faktor Unsur Cuaca terhadap Perubahan Iklim di Kabupat-en Pasuruan pada Tahun 2021 dengan Metode Principal Component Analysis,” New-ton-Maxwell Journal of Physics, vol. 3, no. 2, Oct. 2022, [Online]. Available: https://www.ejournal.unib.ac.id/index.php/nmj

S. B. Sipayung, “Dampak Variabilitas Iklim Terhadap Produksi Pangan di Sumatera,” vol. 2, Jun. 2005.

E. Aldrian, “Sistem Peringatan Dini Menghadapi Iklim Ekstrem,” vol. 10, no. 2, Dec. 2016.

H. A. Tambunan and D. Saputra, “Rancang Bangun Aplikasi Prediksi Cuaca Berbasis Android,” Jurnal Bisantara Informatika (JBI), vol. 6, no. 2, 2022.

S. Chodijah, “Strategi Komunikasi Penyampaikan Informasi Iklim Stasiun Klimatologi Sampali Medan Dalam Upaya Meminimalkan Kegagalan Panen Padi Sawah Akibat Iklim Ekstrim,” Persepsi: Communication Journal, vol. 1, no. 1, pp. 55–69, Nov. 2018, doi: 10.30596/persepsi. v1i1.2506.

J. H. Yousif, H. A. Al-Balushi, H. A. Kazem, and M. T. Chaichan, “Analysis and fore-casting of weather conditions in Oman for renewable energy applications,” Case Stud-ies in Thermal Engineering, vol. 13, p. 100355, Mar. 2019, doi: 10.1016/J.CSITE.2018.11.006.

B. Poernomo, R. Dewi, and I. Sari, “Penerapan Data Mining untuk Prakiraan Cuaca di Kota Malang Menggunakan Algoritma Iterative Dichotomiser Tree (ID3),” JOUTICLA, vol. 3, no. 2, 2017.

Irmayani, “Penerapan Algoritma CART Klasisifikasi Sosial Ekonomi Masyarakat Ke-lurahan Amessangeng,” Jurnal Ilmiah Information Technology d’Computare, vol. 10, Jul. 2020.

J. Han and M. Kamber, “Designing Data-Intensive Web Applications,” 2006.

P. Meilina, “Penerapan Data Mining dengan Metode Klasifikasi Menggunakan Decision Tree dan Regresi,” Jurnal Teknologi Universitas Muhammadiyah Jakarta, vol. 7, no. 1, 2015.

R. Purba, “Data Mining: Masa Lalu, Sekarang dan Masa Mendatang,” vol. 13, no. 1, 2012.

S. Anastassia Amellia Kharis and A. Haqqi Anna Zili, “Learning Analytics dan Educa-tional Data Mining pada Data Pendidikan,” Jurnal Riset Pembelajaran Matematika Sekolah, vol. 6, 2022.

Safitra, M. F., & Abdurrahman, L. (2023). Open-up International Market Opportunities: Using the OSINT Crawling and Analyzing Method. SEIKO: Journal of Management & Business, 6(1), 923-931.

Safitra, M. F., Lubis, M., & Widjajarto, A. (2023, March). Security Vulnerability Analy-sis using Penetration Testing Execution Standard (PTES): Case Study of Government's Website. In Proceedings of the 2023 6th International Conference on Electronics, Communications and Control Engineering (pp. 139-145).

Rafiq Amaliyah, “Aplikasi Klasifikasi Citra Kerusakan Aspal Menggunakan Matlab 2013A,” Universitas Gunadarma, 2014.

R. Amalyah, “Aplikasi Klasifikasi Citra Kerusakan Aspal Menggunakan Matlab 2013A,” Universitas Gunadarma, 2014.

M. Jordan, J. Kleinberg, and B. Schölkopf, “Pattern Recognition and Machine Learn-ing.”

Safitra, M. F., Lubis, M., & Kurniawan, M. T. (2023, March). Cyber Resilience: Re-search Opportunities. In Proceedings of the 2023 6th International Conference on Elec-tronics, Communications and Control Engineering (pp. 99-104).

Safitra, M. F., Lubis, M., & Fakhrurroja, H. (2023). Counterattacking Cyber Threats: A Framework for the Future of Cybersecurity. Sustainability, 15(18), 13369.

I. Goodfellow, Y. Bengio, and A. Courville, “Deep Learning.”

Maulana, F., Fajri, H., Safitra, M. F., & Lubis, M. (2023, August). Unmasking log4j’s Vulnerability: Protecting Systems against Exploitation through Ethical Hacking and Cyberlaw Perspectives. In 2023 9th International Conference on Computer and Communication Engineering (ICCCE) (pp. 311-316). IEEE.

Sutoyo, E., Yanto, I. T. R., Saedudin, R. R., & Herawan, T. (2017). A soft set-based co-occurrence for clustering web user transactions. TELKOMNIKA (Telecommunication Computing Electronics and Control), 15(3), 1344-1353.

Jacob, D. W., Fudzee, M. F. M., Salamat, M. A., Saedudin, R., Abdullah, Z., & Hera-wan, T. (2017). Mining significant association rules from on information and system quality of indonesian e-government dataset. In Recent Advances on Soft Computing and Data Mining: The Second International Conference on Soft Computing and Data Mining (SCDM-2016), Bandung, Indonesia, August 18-20, 2016, Proceedings Second (pp. 608-618). Springer International Publishing.

Zunaidi, W. H. A. W., Saedudin, R. R., Shah, Z. A., Kasim, S., Seah, C. S., & Abdu-rohman, M. (2018). Performances analysis of heart disease dataset using different data mining classifications. International Journal on Advanced Science, Engineering, and In-formation Technology, 8(6), 2677-2682.

Yanto, I. T. R., Saedudin, R. R., Lashari, S. A., & Haviluddin. (2018). A numerical classification technique based on fuzzy soft set using hamming distance. In Recent Ad-vances on Soft Computing and Data Mining: Proceedings of the Third International Conference on Soft Computing and Data Mining (SCDM 2018), Johor, Malaysia, Feb-ruary 06-07, 2018 (pp. 252-260). Springer International Publishing.

Jacob, D. W., Fudzee, M. F. M., Salamat, M. A., Saedudin, R. R., Yanto, I. T. R., & Herawan, T. (2017). An application of rough set theory for clustering performance ex-pectancy of Indonesian e-government dataset. In Recent Advances on Soft Computing and Data Mining: The Second International Conference on Soft Computing and Data Mining (SCDM-2016), Bandung, Indonesia, August 18-20, 2016, Proceedings Second (pp. 638-646). Springer International Publishing.

Seah, C. S., Kasim, S., Fudzee, M. F., Mohamad, M. S., Saedudin, R. R., Hassan, R., ... & Atan, R. (2018). An effective pre-processing phase for gene expression classifica-tion. Indonesian Journal of Electrical Engineering and Computer Science, 11(3), 1223.

Darmawan, M. F., Jamahir, N. I., Saedudin, R. R., & Kasim, S. (2018). Comparison be-tween ANN and multiple linear regression models for prediction of warranty cost. In-ternational Journal of Integrated Engineering, 10(6).

Saedudin, R. R., Sutoyo, E., Kasim, S., Mahdin, H., & Yanto, I. T. R. (2017, October). Attribute selection on student performance dataset using maximum dependency attrib-ute. In 2017 5th International Conference on Electrical, Electronics and Information Engineering (ICEEIE) (pp. 176-179). IEEE.