The Classification of Ultraviolet Index Using Logistic Regression and Random Forest methods for Predicting Extreme Conditions

Alfin Syarifuddin Syahab*[1,2], Galih Langit Pamungkas[3], and Saif Akmal[4]

[1] Climatology Station of Yogyakarta, Meteorological, Climatological, and Geophysical Agency, Yogyakarta, Indonesia
[2] Department of Information Technology, University of Technology Yogyakarta, Yogyakarta, Indonesia
alfin.syahab@bmkg.go.id
[3] Global Atmosphere Watch Lore Lindu Bariri, Meteorological, Climatological, and Geophysical Agency, Palu, Indonesia
[4] Climatology Station of Bengkulu, Meteorological, Climatological, and Geophysical Agency, Bengkulu, Indonesia

Abstract. The Ultraviolet (UV) index is one of the most important markers for estimating potential exposure to harmful sun radiation. For the purpose of managing public health and environmental monitoring, it is imperative to predict high UV levels accurately. A study was designed and analyzed that used methodologies of logistic regression (LR) and random forest (RF) to forecast extreme UV conditions based on the UV Index. This article presents the findings. This study used logistic regression and random forest algorithms to assess the classification accuracy of high UV scenarios. The historical UV index data was used to train and validate the categorization model. Data gathered in 2022 from radiometer-based UV A and UV B radiation measurements done in Palu City, Central Sulawesi. In this test, the training and testing data sets are randomly divided into 70% and 30% respectively. Accuracy and F1 score are used to assess the model. Based on these findings, while the random forest model achieved an accuracy of 0.997 and an F1 score of 0.947, the logistic regression model achieved an accuracy of 0.958 and an F1 score of 0.996. Logistic regression is better than random forest techniques. These results show that the Random Forest model has better predictive ability than the Logistic regression model in making predictions in extreme and non-extreme condition of ultraviolet index.

Keywords: Classification, Logistic Regression, Prediction, Random Forest, Ultraviolet.

1 Introduction

At the surface of the earth, ultraviolet (UV) radiation travels through the atmosphere, where a great deal of absorption and processing takes place. There are three types of UV radiation: UV-A (315–400 nm), UV-B (280–315 nm), and UV-C (200–280 nm).

A. Putro Suryotomo and H. Cahya Rustamaji (eds.), *Proceedings of the 2023 1st International Conference on Advanced Informatics and Intelligent Information Systems (ICAI3S 2023)*,
Advances in Intelligent Systems Research 181,
https://doi.org/10.2991/978-94-6463-366-5_2

Very little UV-A radiation is absorbed by atmospheric gases, but all UV-C radiation is absorbed by oxygen and ozone, keeping it from reaching the troposphere and the earth's surface. In the UV-B region, ozone absorbs light quickly as wavelength decreases, which results in a significant decrease in surface radiation [1]. For humans, low levels of UV radiation are healthy and necessary for the synthesis of vitamin D. Additionally, an increasing amount of data indicates that environments with high UV radiation levels may increase the risk of infectious diseases, skin cancer, and cataract development [2].

The World Meteorological Organization (WMO), the World Health Organization (WHO), the United States National Weather Service (NWS), and the Environmental Protection Agency (EPA) all adopted the UV Index (UVI) in 1994 after it was initially created in Canada in 1992 [3]. The UVI is a measure of the skin-damaging solar UV radiation that is erythemally weighted and falls on a horizontal surface at the bottom of the atmosphere. It is designed to represent radiation in a simple form, as a single number. The values of the index range from zero upward – the higher the index value, the greater the potential for damage to the skin and eye [2].

One statistical analysis methodology that is frequently used for categorization in prediction models is logistic regression (LR). High algorithm performance is typically attained by this categorization model. When the response variable is binary, binary logistic regression is applied [4]. In addition, an ensemble of decision trees is used in the supervised machine learning method known as random forest, where each tree is trained separately using a different subset of the data. Understanding how the models generate predictions can be aided by using random forests [5]. This paper attempts to propose best method for reliably predicting high UV conditions based on UVA and UVB radiation measured using a radiometer in Palu City, Central Sulawesi. The method makes use of logistic regression and random forest algorithms. The study approach is to first focus on the analysis of UV extreme classifications, then develop a prediction model for extreme and non-extreme values, and then determine the optimal performance of a machine learning algorithm between logistic regression and random forest for UVI data prediction.

2 Method

2.1 The UV Index

The International Commission on Illumination (CIE) reference action spectrum for UV-induced erythema on human skin is used to construct the Global Solar UV. It is a measurement of UV radiation specific to and applicable to horizontal surfaces. A unitless quantity, the UV is defined as follows:

$$I_{uv} = k_{er} \cdot \int_{250nm}^{400nm} E_{\lambda} S_{er}(\lambda) d\lambda \quad (1)$$

where E_{λ} is the solar spectral irradiance expressed in W/(m^2.nm) at wavelength (λ) and $d\lambda$ is the wavelength interval used in the summation. S_{er} (λ) is the erythema reference action spectrum and k_{er} is a constant equals to 40 m^2/W.

Vishnu et al. conducted a study in 2020 on the prediction of UV index data, utilizing hourly data for a year. The dataset was obtained from the Open Weather Map API. The

approach makes use of the Gated Recurrent Unit (GRU) and Recurrent Neural Network (RNN) methods. The timestamp model's RMSE assessment results are 0.3606, the univariate model's is 1.3098, and the timestamp model with temperature is 0.3213 [6].

2.2 Logistic Regression

Logistic regression (LR) is concerned with the special situation in regression modeling, where the outcome is of a binary or dichotomous (yes/no) nature [7]. It is a supervised machine learning algorithm developed for learning classification problems when the target variable is categorical. The goal of LR is to map a function from the features of the dataset to the targets to predict the probability that a new example belongs to one of the target classes [8]. In particular, the model output in multinomial LR is given by use of a generalization of the logistic function in (2):

$$P(y = k|x) = \frac{e^{x^T w_k}}{\sum_{j=1}^{K} e^{x^T w_k}} \tag{2}$$

where k is the event class, x is the predictor vector, and w is the vector of regression coefficients. Note that separate coefficient vectors are computed for each event class [9].

In 2020, Deng did a research about the influential effect of whether it will rain tomorrow by establishing a LR and decision tree model, and sets up a prediction model to predict whether it will rain tomorrow. The prediction accuracy of LR and decision tree model is not much different, but the ROC area of logistic regression is slightly higher. Therefore, it is more appropriate to use logistic regression in the actual application of large amounts of data [10]. Another research by Chan et.al in 2018, using logistic regression adopted rainfall depth or maximum rainfall intensity as the hydrological factor to analyze landslide susceptibility. The results indicated that the overall accuracy of predicted events exceeded 80%, and the area under the receiver operating characteristic curve (AUC) closed to 0.8 [11].

2.3 Random Forest

Besides using the LR method, we also use another machine learning method to compare the accuracy of the resulting predictions. Random forest (RF) is a popular machine learning procedure which can be used to develop prediction models. The combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [12]. The calculation flow of RF is based on (3):

$$RFfi_i = \frac{\sum_j normfi_{ij}}{\sum_{j \in all\ features, k \in all\ trees} normfi_{ij}} \tag{3}$$

where $RFfi_i$ is the importance of feature i calculated from all trees in the RF model, and $normfi_{ij}$ is the normalized feature importance for i in tree j [13].

In 2019, Diez Sierra and del Jesus utilized RF method that use atmospheric data and daily rainfall statistics as predictors are evaluated to downscale daily-to-sub daily rainfall statistics on more than 700 hourly rain gauges in Spain. This approach can be

applied for the study of extreme events and for daily-to-sub daily precipitation disaggregation in any location of Spain where daily rainfall data are available [14]. The RF method is known for its high prediction accuracy. As an example of research results from Primajaya and Nurina in 2018 about RF Algorithm for Prediction of Precipitation, implementation of random forest algorithm with 10-fold cross validation resulted in the output with accuracy 99.45%, precision 0.99, recall 0.99, f-measure 0.99, kappa statistic 0.99, MAE 0.09, RMSE 0.14, ROC area 1 [15].

2.4 Research Steps

The experiment examined the use of Random Forest and Logistic Regression algorithms to classify UV index data under extreme and non-extreme settings. UVA and UVB readings taken on land are the data used. This measurement is part of the Meteorology, Climatology and Geophysics Agency's (BMKG) operating program for the Global Atmospheric Watch in Palu, Central Sulawesi. The time period of the data collected for processing is 2022, and measurements are made every minute. There are 508985 lines of data in that dataset. The Systems Development Life Cycle was used to optimize the experiment's implementation. The system development process phases depicted in Fig. 1.

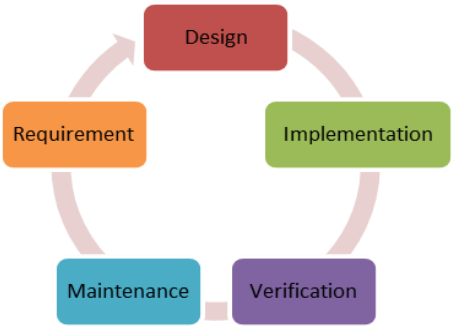

Fig. 1. The system development life cycle.

The waterfall model of approach, which consists of requirements, design, implementation, verification, and maintenance, is used to operate the system. The process of creating and maintaining systems, along with the models and methods utilized in their development, is referred to as this system in systems engineering. This phrase often describes a computer or information system [16].

2.5 Requirement

The UV index is generated from the UV erythemal value of sensor data, which is then computed using the UV-A and UV-B weighing factors in accordance with the CIE spectral action function. This process is known as weighting factor calculation. The weighting factor for the UV B spectrum at 305 nm is 0.22, whereas the weighing value

for the UV A spectrum at 325 nm is 0.0029. These factors are related to the erythemal UV intensity calculation [17]. Equation (4) presents the formula for determining the UV index.

$$UV\ Index = \frac{erythermal\ UV\ A + erythermal\ UV\ B}{0.025} \tag{4}$$

W/m^2 is the unit of measurement for both erythemal UVs. The standard increment number for the amount of total UV that may be harmful to living tissue is 0.025 W/m^2. Put another way, a rise on one UV index scale corresponds to a 25 m^2/W exposure to UV radiation.

2.6 Design

A diagram process gives detailed information on every stage of the process. The design includes the research work flow, which is predicated on the procedures from data collection to a trained prediction model. Fig. 2 illustrates the process diagram for developing a prediction model based on machine learning. The procedure consists of several steps: gathering data, pre-processing, compiling data into test and train datasets, executing the dataset using LR and RF, analyzing the classification accuracy value and F1 score, and producing trained models as well as extreme and non-extreme ultraviolet prediction models as the final product.

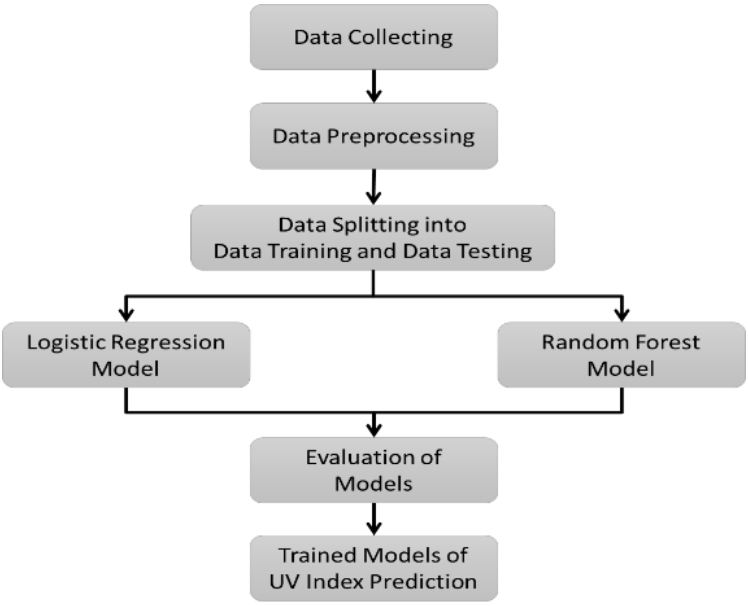

Fig. 2. The diagram process of our work.

2.7 Implementation

The data is split into two categories by the classification algorithm: test data and train data. A random 70:30 data split used in food research can yield the highest accuracy value, reaching 87.9% in the LR algorithm [18]. Additionally, a different study demonstrated that, when evaluating the imbalance of uric acid data sets using a maximum BCR criterion, the RF approach performed the best, with an accuracy of 0.70 [19]. Based on these two studies, it is highly recommended that the training and testing data portions be randomly split into 70:30 in order to provide the best predictions possible.

Python with Pyspark module installed then helps with methods for processing quantitative data utilizing statistics and mathematical equations. Pyspark processes the enormous amount of data and then determines the accuracy value for the data categorization outcomes during the algorithm testing phase. Additionally, it is more appropriate for iterative applications like machine learning and data mining [20]. Binary categorization is the type used. That classifies anything from 0 as not extreme to 1 as extreme. The UV index value more than eleven (>11) is the cutoff point that falls into the extreme category [17].

2.8 Verification

A comparison of the number of instances that will be accurately identified and all of the cases that are now open is called accuracy [21]. Performance in classification will be impacted by classification accuracy. Classification accuracy is a performance metric that is frequently used to assess classifier application. The accuracy value formula is described by (5) below.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + TN} \tag{5}$$

The classification accuracy is calculated by dividing the number of correct predictions (True Positive plus True Negative) by the total number of predictions (True Positive plus True Negative + False Positive plus False Negative). In addition, the study uses F-score to evaluate the model. The F1 score penalizes extreme values of either precision or recall because it is the harmonic mean of these two metrics. Equation (6) gives the formula to determine the F-score

$$F1 = \frac{2\,x\,TP}{2\,x\,TP + FP + TN} \tag{6}$$

As a proportionate measure of TP, the F1 score would be high, for instance, if there was a sizable positive class and the classifier was biased in favor of this majority. Even though neither the data nor the relative class distribution have changed, redefining the class labels so that the negative class is the majority and the classifier is biased towards the negative class will result in a low F1 score. The F1-score has a constraint of [0, 1], where 0 denotes no precision and/or recall and 1 denotes the maximum precision and recall values [22].

2.9 Maintenance

This study compares the accuracy testing of extreme UV index data classification using the LR and RF algorithms. According to the test results, LR and RF have a value for accuracy. Splitting data and accuracy values from preprocessing data can be taken into consideration to create a more ideal model for UV index data prediction.

3 Result and Discussion

The outcome of using the algorithms for classification is covered and analyzed in this section. Tables and figures supporting the analyses were displayed in the outcome. The conversation was split up into a number of portions.

3.1 Data Collection

Using UVA and UVB radiometer devices from the Global Atmospheric Watch Station in Palu City, Central Sulawesi, UVA and UVB data were gathered in-situ. There are 508985 lines in the sensor's minutely raw data collection. Following the filtering step, the data showed 254488 lines. Fig. 3 displays the data visualization of the filtered UVA and UVB measurements.

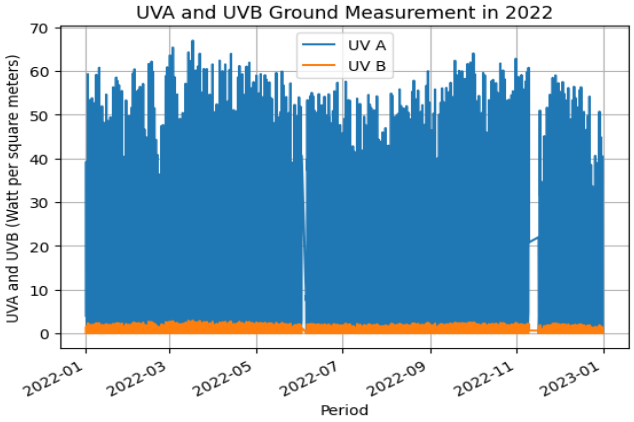

Fig. 3. UVA and UVB ground measurement.

With measurements made every minute, the UV-A and UV-B data are filtered in daily data from 06.00 to 18.00 local time intervals in 2022. The UV observation is added to the sensor measurement database. Local storage is where data are stored. Subsequently, the user gathers data in CSV format.

3.2 Data Preprocessing

The UVA and UVB measurements from the sensor are used to determine the UV index. This represents the UV radiation level in 2022. Fig. 4 shows the UV index visualization graph for the year 2022.

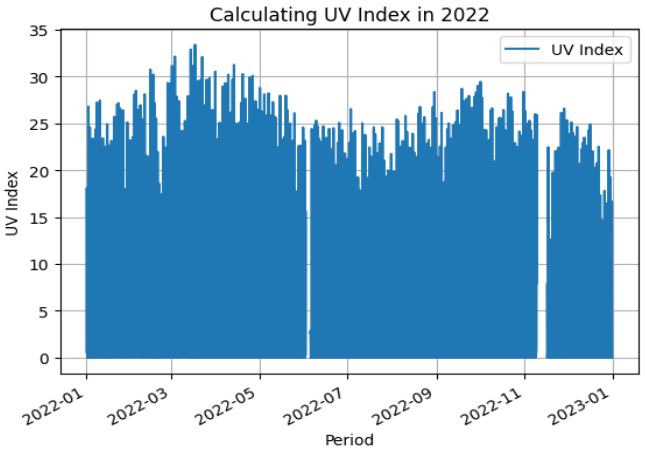

Fig. 4. Calculating UV index in 2022.

For this test, the data will be ready for the categorization phase. Preprocessing is the process of transforming unprocessed data into datasets that are ready for additional processing. UV A, UV B, UV A erythemal, UV B erythemal, and UV index are all present. Following pre-processing, those randomly select one-line examples from the dataset and present them in Table 1.

Table 1. The result of data preprocessing

UVA	UVB	UVA erythemal	UVB erythemal	UV Index
17.5	0.3	0.05075	0.066	4.67

This data can be used as an input for data processing, producing a dataset with a 70% training data to 30% testing data ratio for training and testing. Local time, UVA, and UV radiation measurements are displayed in W/m2 in the raw data. The acquired data will be used as input for further calculations to produce the UV index, which will be used as input for creating a prediction model.

3.3 Data Splitting

A dataset containing the input column characteristics UVA, UVB, UVA erythemal, UVB erythemal, and UV index is needed for the training data. A label column in the form of extreme values —binary categorization in this case— is then added. The following step involves employing the LR and RF algorithms to randomly train up to 70%

of the dataset. Afterwards, tests employing models that have been trained to yield prediction values, probabilities, and true labels are conducted using the remaining 30% at random.

3.4 Performance of Logistic Regression and Random Forest

This dataset, which is an example of data in the table format with a label attribute in extreme conditions with a value of 1, is provided to test the performance of the classification. An example one-line dataset with an extreme label is displayed in Table 2.

Table 2. Extreme label classification

UVA	UVB	UVA erythemal	UVB erythemal	UV Index	Extreme
33.3	0.9	0.09657	0.198	11.7828	1

In contrast, the dataset indicates that the label is not excessive in other circumstances, with an extreme label value of 0. An example one line dataset with labels that are considered non-extreme is displayed in Table 3.

Table 3. Non-extreme label classification

UVA	UVB	UVA erythemal	UVB erythemal	UV Index	Extreme
16.8	0.6	0.04872	0.132	7.2288	0

Classification findings on LR with a total of 76668 data revealed 3207 incorrectly categorized data and 73461 successfully classified data. and came up with a 0.958 accuracy score. Next, out of 76668 data, 76441 were successfully classified and 227 were misclassified in the classification findings on RF. and came up with a 0.997 accuracy score.

Table 4. Performance of Classification Algorithms

Algorithm	Accuracy	F1 score
Logistic Regression	0.958	0.947
Random Forest	0.997	0.996

The results of the accuracy value-based performance evaluation of classification algorithms are displayed in Table 4. The experiment yielded results for the prediction of extreme UV index in the LR and RF algorithms.

4 Conclusion

The objective of this research is to match the attributes of UV index datasets, which are produced from computed UV sensor values, using the machine learning model

assessment tool. This tool allows users to classify a data set, after which they may use logistic regression and random forest analysis on the data by utilizing extreme thresholds as classification labels. When developing prediction models, we examined the effects of characteristics and results on every machine learning model. To assess the two models' classification performance, accuracy and f-score values are employed. other. Compared to utilizing the logistic regression approach, which has an accuracy value of 0.958 and an F1-score of 0.947, the random forest model generates an accuracy value of 0.997 and an F1-score of 0.996, which can help improve the performance of forecasting UV index values. Future work on this project should focus on data preprocessing for handling empty and outlier-filled sensor data. Since the data has a time attribute, time series prediction algorithms like ARIMA and LSTM-RNN can be used to get factor analyses of past measurement values.

References

1. V. Fioletov, J. B. Kerr, and A. Fergusson, "Author Affiliations Science and Technology Branch," *Can J Public Heal.*, vol. 101, no. 4, pp. 5–9, 2010.
2. WHO, "Global Solar UV Index A Practical Guide," *World Heal. Organ.*, p. 18, 2002.
3. C. J. Heckman, K. Liang, and M. Riley, *Awareness, understanding, use, and impact of the UV index: A systematic review of over two decades of international research*, vol. 123. 2019. doi: 10.1016/j.ypmed.2019.03.004.
4. Y. Dani and M. A. Ginting, "Classification of Predicting Customer Ad Clicks Using Logistic Regression and k-Nearest Neighbors," *Int. J. Informatics Vis.*, vol. 7, no. 1, pp. 98–104, 2023, doi: 10.30630/joiv.7.1.1017.
5. S. Kapsiani and B. J. Howlin, "Random forest classification for predicting lifespan-extending chemical compounds," *Sci. Rep.*, vol. 11, no. 1, pp. 1–13, 2021, doi: 10.1038/s41598-021-93070-6.
6. M. Vishnu, Sri; V,Tarun Srivatsa; Meleet, "ULTRAVIOLET INDEX ANALYSIS AND FORECASTING USING DEEP LEARNING METHODOLOGIES FOR BENGALURU CITY," *EPRA Int. J. Multidiscip. Res.*, vol. 8, no. 7, pp. 375–379, 2022.
7. A. Das, "Logistic Regression," in *Encyclopedia of Quality of Life and Well-Being Research*, Cham: Springer International Publishing, 2021, pp. 1–2. doi: 10.1007/978-3-319-69909-7_1689-2.
8. E. Bisong, "Logistic Regression," in *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, Berkeley, CA: Apress, 2019, pp. 243–250. doi: 10.1007/978-1-4842-4470-8_20.
9. D. S. Wilks, *Statistical methods in the atmospheric sciences (Vol. 100)*. Academic Press, 2011.
10. F. Deng, "Research on the Applicability of Weather Forecast Model-Based on Logistic Regression and Decision Tree," *J. Phys. Conf. Ser.*, vol. 1678, no. 1, pp. 0–7, 2020, doi: 10.1088/1742-6596/1678/1/012110.
11. H. C. Chan, P. A. Chen, and J. T. Lee, "Rainfall-induced landslide susceptibility using a rainfall-runoff model and logistic regression," *Water (Switzerland)*, vol. 10, no. 10, pp. 3–12, 2018, doi: 10.3390/w10101354.
12. L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

13. S. Ronaghan, “The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark,” Towards Data Science. [Online]. Available: https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3
14. J. Diez-Sierra and M. del Jesus, “Subdaily rainfall estimation through daily rainfall downscaling using random forests in Spain,” *Water (Switzerland)*, vol. 11, no. 1, 2019, doi: 10.3390/w11010125.
15. A. Primajaya and B. N. Sari, “Random Forest Algorithm for Prediction of Precipitation,” *Indones. J. Artif. Intell. Data Min.*, vol. 1, no. 1, p. 27, 2018, doi: 10.24014/ijaidm.v1i1.4903.
16. D. Nurcahya, H. Nurfauziah, and H. Dwiatmodjo, “Comparison of Waterfall Models and Prototyping Models of Meeting Management Information Systems,” *J. Mantik*, vol. 6, no. 2, pp. 1934–1939, 2022, [Online]. Available: https://iocscience.org/ejournal/index.php/mantik/article/view/2677/2150
17. A. S. Syahab, A. Widiyanto, L. R. Anuggilarso, and A. B. Wijaya, “Comparison of Machine Learning Algorithms for Classification of Ultraviolet Index,” *J. Teknol. Inf. dan Pendidik.*, vol. 15, no. 2, pp. 132–146, 2023, doi: 10.24036/jtip.v15i2.692.
18. B. Vrigazova, “The Proportion for Splitting Data into Training and Test Set for the Bootstrap in Classification Problems,” *Bus. Syst. Res.*, vol. 12, no. 1, pp. 228–242, 2021, doi: 10.2478/bsrj-2021-0015.
19. S. Lee, E. K. Choe, and B. Park, “Exploration of machine learning for hyperuricemia prediction models based on basic health checkup tests,” *J. Clin. Med.*, vol. 8, no. 2, 2019, doi: 10.3390/jcm8020172.
20. R. Bandi, J. Amudhavel, and R. Karthik, “Machine learning with PySpark – Review,” *Indones. J. Electr. Eng. Comput. Sci.*, vol. 12, no. 1, pp. 102–106, 2018, doi: 10.11591/ijeecs.v12.i1.pp102-106.
21. Nofenky and R. D. Bhisetya, “Recommendation for Classification of News Categories Using Support Vector Machine Algorithm with SVD,” *Ultim. J. Tek. Inform.*, vol. 13, no. 2, pp. 72–80, 2022, doi: 10.31937/ti.v13i2.1854.
22. S. A. Hicks *et al.*, “On evaluation metrics for medical applications of artificial intelligence,” *Sci. Rep.*, vol. 12, no. 1, pp. 1–9, 2022, doi: 10.1038/s41598-022-09954-8.