EasyChair Preprint

№ 10669

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

Alfin Syarifuddin Syahab, Galih Langit Pamungkas and Saif Akmal

August 4, 2023

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

Alfin Syarifuddin Syahab†
Climatology Station of Yogyakarta
Agency for Meteorology,
Climatology, and Geophysics
Yogyakarta, Indonesia
alfin.syahab@bmkg.go.id

Galih Langit Pamungkas
Global Atmosphere Watch Lore
Lindu Bariri
Agency for Meteorology,
Climatology, and Geophysics
Palu, Indonesia
galih.pamungkas@bmkg.go.id

Saif Akmal
Climatology Station of Bengkulu
Agency for Meteorology
Climatology and Geophysics
Bengkulu, Indonesia
saif.akmal@bmkg.go.id

ABSTRACT

The Ultraviolet (UV) Index is a crucial parameter for assessing the risk of harmful solar radiation exposure. Accurate prediction of extreme UV conditions is essential for public health management and environmental monitoring. This article presents a study focused on developing and evaluating a classification model using logistic regression (LR) and random forest (RF) methods to predict extreme UV conditions based on the UV Index. The objective of this research is to assess the performance of logistic regression and random forest algorithms in accurately classifying extreme UV conditions. Historical UV Index are used to train and validate the classification models. Data obtained from measurement of UV A and UV B radiation using a radiometer in Palu City, Central Sulawesi throughout 2022. This test uses split data randomly with a ratio of 70% for training data and 30% for testing data. Performance metrics such as accuracy are employed to evaluate the models. The results indicate that both logistic regression and random forest algorithms show promising performance in classifying extreme UV conditions. The logistic regression model achieves an accuracy of 0.958 while the random forest model achieves an accuracy of 0.997. The random forest algorithm work better than logistic regression. These findings demonstrate the potential of these models for accurately predicting extreme UV conditions. Furthermore, the model contributes to environmental monitoring efforts by providing insights into distribution of extreme UV conditions.

CCS CONCEPTS

• **Applied computing** → **Physical sciences and engineering**; Earth and atmospheric sciences; Environmental sciences • **Computing methodologies** → **Machine learning**; Learning paradigms; Supervised learning by classification • **Information systems** → **Information systems applications**; Data mining

KEYWORDS

Ultraviolet, Classification, Logistic Regression, Random Forest

1. Introduction

The Ultraviolet (UV) radiation at the earth's surface passes through atmosphere, where there is a lot of absorption and process that occurs. UV radiation is classified as UV-A (315-400 nm), UV-B (280-315 nm) and UV-C (200-280 nm). Atmospheric gases absorbs very little UV-A radiation which oxygen and ozone absorbs all UV-C radiation and prevents it from reaching the troposphere and Earth surface. Absorption by ozone increases rapidly with decreasing wavelength in the UV-B range and causes surface radiation to fall off sharply with decreasing wavelength [1]. Small amounts of UV radiation are beneficial for people and essential in the production of vitamin D. UV radiation is also used to treat several diseases, including rickets, psoriasis and eczema. Furthermore, a growing body of evidence suggests that environmental levels of high UV radiation may enhance the risk of infectious diseases; skin cancers and cataracts [2]. The UV Index (UVI) was first developed in Canada in 1992 and adopted by the United States National Weather Service (NWS) and Environmental Protection Agency (EPA) as well as the World Meteorological Organization (WMO) and World Health Organization (WHO) in 1994 [3]. The UVI describes the erythemally weighted skin-damaging solar UV radiation on a horizontal surface at the bottom of the atmosphere [4]. 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]. This article aims to present a novel approach utilizing logistic regression and random forest algorithms for accurately predicting extreme UV conditions based on measurement of UV A and UV B radiation using a radiometer in Palu City, Central Sulawesi. The research findings have significant implications for public health, and environmental management. By leveraging advanced machine learning techniques, the study contributes to the field of UV index prediction and provides valuable insights into mitigating the risks associated with extreme UV exposure.

2. Theoretical Framework

2.1 The UVI Index

The Global Solar UVI is formulated using the International Commission on Illumination (CIE) reference action spectrum for UV-induced erythema on the human skin. It is a measure of the UV radiation that is relevant to and defined for a horizontal surface. The UVI is a unitless quantity defined by the formula:

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

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 equal to 40 m^2/W. [2]

2.2 Logistic Regression

Regression modeling is a popular and useful approach in statistics that is used to explore and describe the relationship between an outcome or dependent/response variable and a set of independent predictors. Logistic regression is concerned with the special situation in regression modeling, where the outcome is of a binary or dichotomous (yes/no) nature [5]. It is a supervised machine learning algorithm developed for learning classification problems when the target variable is categorical. The goal of logistic regression 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 [6]. In particular, the model output in multinomial logistic regression is given by use of a generalization of the logistic function in Equation 2:

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

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. [7].

In 2020, Deng did a research about the influential effect of whether it will rain tomorrow by establishing a logistic regression and decision tree model, and sets up a prediction model to predict whether it will rain tomorrow. The prediction accuracy of logistic regression 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 [8]. 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 [9].

2.3 Random Forest Regression

Besides using the logistic regression method, we also use another machine learning method to compare the accuracy of the resulting predictions. Random forest 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 [10]. The calculation flow of Random Forests is based on Equation 3:

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

where $\boldsymbol{RFfi_i}$ is the importance of feature i calculated from all trees in the Random Forests model, and $\boldsymbol{normfi_{ij}}$ is the normalized feature importance for i in tree j [11].

In 2019, Diez Sierra and del Jesus utilized Random Forest 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 [12]. The random forest method is known for its high prediction accuracy. As an example of research results from Primajaya and Nurina in 2018 about Random Forest 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 [13].

3. Methodology

In the methodology section, the researcher tested the ultraviolet index data classification under extreme conditions using Logistic Regression and Random Forest algorithms. The data used is ground-based UV A and UV B measurement data. This measurement is an operational activity of the Global Atmospheric Watch in Palu, Center of Sulawesi, that observation is conducted by Indonesian Agency for Meteorology, Climatology, and Geophysics (BMKG). The data taken for processing has a span of a year 2022 with one-minute measurement intervals.

The experiment implemented optimally using the Systems Development Life Cycle. The system is worked by the waterfall model of methodology that consist of requirement, design, implementation, verification, and maintenance. This system in systems engineering is the process of developing and updating systems, as well as the models and procedures used to develop those systems. In general, this term refers to a computer or information system [14]. Those steps of system development methodology shown in Figure 1.

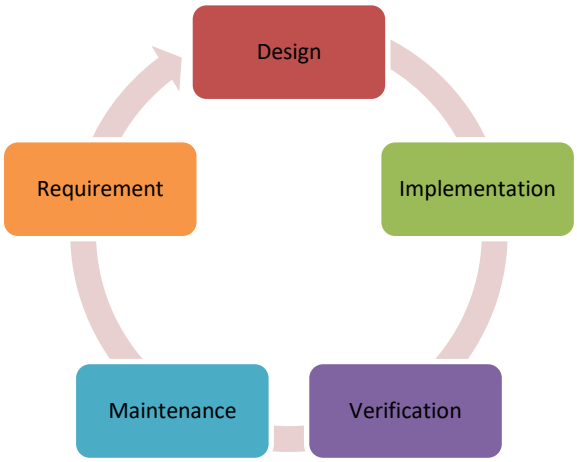

Figure 1: The System Development Life Cycle Methodology

3.1 Requirement

Calculation of the weighting factor is a method of obtaining the UV index obtained from the UV erythermal value of sensor measurements calculated using the UV-A and UV-B weighing factors according to the CIE spectral action function. UV-A penetrates deeper into the surface than UV-B. However, energy is inversely proportional to wavelength. UV-A has lower energy than UV-B. Consequently, the weighing factor for UV-B is higher than for UV-A. The UV B spectrum at 305 nm has a weighing factor of 0.22 while UV A at 325 nm has a weighing value of 0.0029. These variables are associated with the computation of erythermal UV intensity [15]. The formula for calculating the UV index is shown in Equation 4.

$$UV\ Index = \frac{erythermal\ UV\ A + erythermal\ UV\ B}{0.025} \quad (4)$$

The two erythermal UVs are both stated in W/m^2. The denominator of 0.025 W/m^2 is the standard increment number that relates to how much total UV is potentially detrimental to living tissue, or in other words, a rise on one scale of the UV index is equivalent to 25 m^2/W exposure to UV radiation.

A supervised learning method can be used to classify data. The classification algorithm divides the data into two categories: train data and test data. Research on food data using a random 70:30 data split can achieve the best accuracy value reaching 87.9% in the logistic regression algorithm [16]. Also, in another study showed that random forest method performed the best with an accuracy of 0.70 when assessing the imbalance of uric acid data sets based on a maximum BCR criterion [17]. In this case, the distribution of the number of train data and test data in this study is 70:30. The total UV index data utilized as train data in this study was 177820 and test data in this investigation was 76668.

3.2. Design

The stages of the process are detailed in the form of a flowchart. The design contains the flow of research work carried out based on the stages from data collection to a predictive model that has been tested.

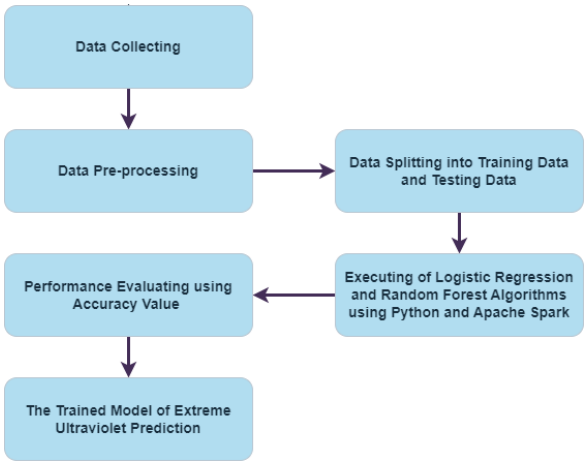

Figure 2: The Flowchart of Data Processing

Figure 2 illustrates a flowchart for developing a machine learning-based recommendation system, which includes several steps such as data collection, pre-processing of data, compiling data into test data and train data, executing of dataset using logistic regression and random forest, then analysis of the accuracy value of the classification results using the algorithm tested, and the final result in the form of a trained model along with prediction model on extreme ultraviolet.

3.3. Implementation

Testing using the LR and RF algorithms is a system development methodology after data pre-processing in analyzing processing data quantitatively. Techniques for processing quantitative data using mathematical equations and statistics are then assisted by Python in Jupyter Notebook that installed Apache Spark pyspark library module. The kind of classification is binary. That has 0 as not extreme category and 1 as extreme category. The threshold that is included in the extreme category is the UV index value above 11 (>11) [15]. This Apache Spark processes the huge size of data then calculates the accuracy value for the results of data classification in the algorithm testing process. Spark is better suited for iterative applications such as data mining and machine learning. The RDD (Resilient Distributed Dataset) in Spark is a fault-tolerant collection of components that may be processed in parallel and allows clients to explicitly store information on compact disk and memory [18].

3.4 Verification

Accuracy is a comparison between cases that will be correctly identified and the total number of existing cases [19]. Classification accuracy will affect classification performance. The performance measure is commonly used to evaluate classifier application is classification accuracy. The ratio of valid predictions (True Positive plus True Negative) divided by the total number of predictions made (True Positive plus True Negative + False Positive plus False Negative) is the classification accuracy. This formula of accuracy value is mentioned by Equation 5 below.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

3.5. Maintenance

Applying the LR and RF algorithms, a comparative analysis of accuracy testing on extreme UV index in data classification. The test results reveal that LR and RF has an accuracy value. From preprocessing data, splitting data and accuracy values can be considered to develop a more optimal model in predicting UV index data.

4. Result

This part discusses the research findings while also offering a full implementation and discussion. Tables, figures, and other forms are used to display the results. The splitting of the discussion into multiple sections.

4.1. Data Collection

Observation of the ultraviolet is recorded to database of sensor measurement. Data are recorded in local storage. Then, data are collected by user in csv format. Data visualization of UVA and UVB from ground measurement showed in Figure. 3.

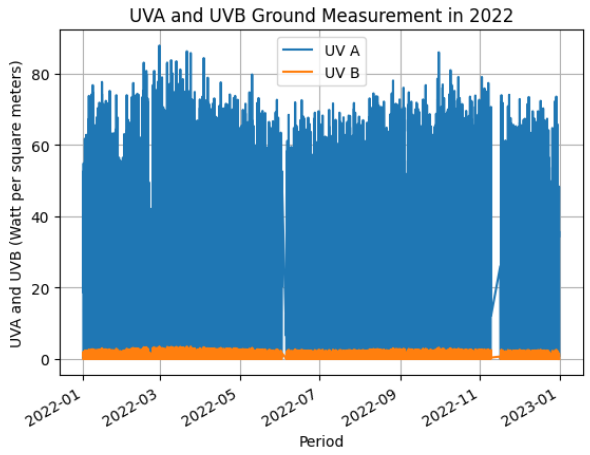

Figure 3: UVA and UVB Ground Measurement

UVA and UVB data were collected in-situ utilizing UVA and UVB radiometer instruments from the Global Atmospheric Watch Observation Station in Palu City, Central Sulawesi. The UV-A and UV-B data are filtered in daily data from 06.00 - 18.00 local time intervals in 2022 with measurements every one-minute record.

4.2. Pre-processing Data

The UV index is calculated by UVA and UVB measurement from sensor. This is a measure of the level of UV radiation in 2022. Figure 4 illustrate the graph of visualization of UV index throughout 2022.

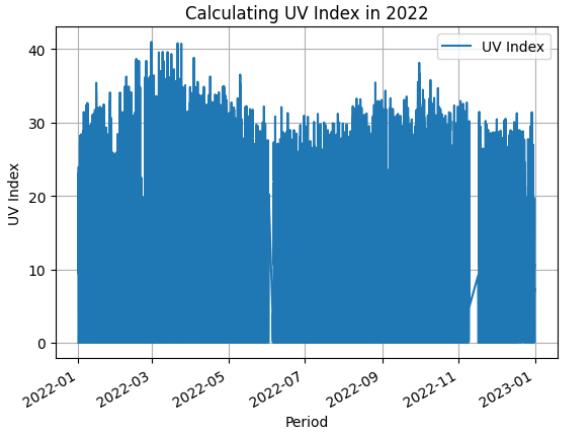

Figure 4: Calculating UV Index in 2022

The data will be prepared for the categorization stage in this test. Pre-processing of data is carried out to convert raw data into datasets that can be used for further processing. It contains of UV A, UV B, UV A erythermal, UV B erythermal, UV index. Those takes five lines examples randomly from the dataset after pre-processing and are displayed in Table 1.

Table 1. The result of pre-processing data

UVA	UVB	UVA erythermal	UVB erythermal	UV Index
17.5	0.3	0.05075	0.066	4.67
21	0.4	0.0609	0.088	4.67
16.6	0.3	0.04814	0.066	5.956
17.5	0.4	0.05075	0.088	4.5656
16.9	0.4	0.04901	0.088	5.55

This data can be utilized as raw material for data processing, resulting in a dataset in training and testing data that has ratio 70% for training data and 30% testing data. Raw data after filtering from measurement equipment collected up to 254.489 lines. The raw data shows local time, UVA, and UV radiation values in W/m^2. The obtained data will be utilized as input for additional computations in generating the UV index as input for developing a prediction model.

4.3. Data Training and Testing

In the training data, a dataset is required with the input column attributes UVA, UVB, UVA erythermal, UVB erythermal, UV index, then adding a label column in the form of extreme values, which in this case uses binary classification. Table 2 shown the example line of the dataset with the extreme attribute as label.

Table 2. Preparing data training and data testing

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

The next stage is to conduct data training as much as 70% of the dataset randomly using the LR and RF algorithms. Then using the other 30% at random to carry out tests with models that have been trained to produce predictive values, probabilities, and true labels.

4.4. Performance of Logistic Regression and Random Forest

Classification results on LR consisting of 76668 showed 73461 correctly classified data and 3207 misclassified data. And obtained an accuracy value of 0.958. Then, classification results on RF consisting of 76668 showed 76441 correctly classified data and 227 misclassified data. And obtained an accuracy value of 0.997.

Table. 3 Performance of Classification Algorithms

Algorithm	Accuracy
Logistic Regression	0.958
Random Forest	0.997

Table 3. shown the result of performance evaluation in classification algorithms based accuracy value. The experiment produced the performance in LR and RF algorithms for extreme UV index predicting.

REFERENCES

[1] V. Fioletov, J. B. Kerr and A. Fergusson, "The UV Index: Definition, Distribution and Factors Affecting It," *Can Journal of Public Health*, vol. 101 (4), pp. 15 - 19 , 2010.

[2] World Health Organization, World Meteorological Organization and United Nations Environment Programme , Global Solar UV Index : a practical guide, World Health Organization, 2002.

[3] C. J. Heckman, K. Liang and R. Mary, "Awareness, understanding, use, and impact of the UV index: A systematic," *Preventive medicine*, vol. 123, pp. 71 - 83, 2019.

[4] P. Koepke, A. Bais, D. Balis, M. Buchwitz, H. d. C. X. De Backer, P. Eckert, P. Eriksen, D. Gillotay, A. Heikkila, T. Koskela, B. Lapeta, L. Zenobia, L. Jeronimo and M. Bernhard, "Comparison of Models Used for UV Index Calculations," *Photochemistry and Photobiology*, vol. 67 (6), pp. 657 - 662, 1998.

[5] A. Das, Logistic Regression. In: Maggino, F. (eds) Encyclopedia of Quality of Life and Well-Being Research, Springer, Cham. https://doi.org/10.1007/978-3-319-69909-7_1689-2, 2021.

[6] E. Bisong, Logistic Regression. In: Building Machine Learning and Deep Learning Models on Google Cloud Platform, Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-4470-8_20, 2019.

[7] D. S. Wilks, Statistical methods in the atmospheric sciences (Vol. 100), Academic Press, 2011.

[8] F. Deng, "Research on the applicability of weather forecast model—based on logistic regression and decision tree," *Journal of Physics: Conference Series*, vol. 1678, no. https://doi.org/10.1088/1742-6596/1678/1/012110, 2020.

[9] H. Chuan Chan , P. An Chen and J. Tai Lee, "Rainfall-Induced Landslide Susceptibility Using a Rainfall–Runoff Model and Logistic Regression," *Water*, vol. 1354, no. 10 http://dx.doi.org/10.3390/w10101354, 2018.

[10] L. Breiman, "Random Forests," *Machine Learning*, no. 45 https://doi.org/10.1023/A:1010933404324, pp. 5-32, 2001.

[11] S. Ronaghan, "The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark," Towards Data Science, 12 May 2018. [Online]. Available: https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3. [Accessed 17 July 2023].

[12] J. Diez-Sierra and M. del Jesus, "Subdaily Rainfall Estimation through Daily Rainfall Downscaling Using Random Forests in Spain," *Water*, vol. 11, no. 1 https://doi.org/10.3390/w11010125, p. 125, 2019.

[13] A. Primajaya and B. Nurina Sari, "Random Forest Algorithm for Prediction of Precipitation," *Indonesian Journal of Artificial Intelligence and Data Mining*, vol. 1, no. 1 http://dx.doi.org/10.24014/ijaidm.v1i1.4903, 2018.

[14] D. Nurcahya, H. Nurfauziah and H. Dwiatmodjo, "Comparison of Waterfall Models and Prototyping Models of Meeting Management Information System," *Jurnal Mantik*, vol. vol. 6 no. 2, pp. 1934-1939, 2022.

[15] A. S. Syahab and Rianto, "Comparison of Machine Learning Algorithms for Classification of Ultraviolet Index," *Jurnal Teknologi Informasi dan Pendidikan*, vol. vol. 15 no. 2, no. https://doi.org/10.24036/tip.v15i2, pp. 132-146, 2022.

[16] B. Vrigazova, "The Proportion for Splitting Data into Training and Test Set for the Bootstrap in Classification Problems," *Business Systems Research*, vol. Vol. 12 No. 1 , no. DOI: https://doi.org/10.2478/bsrj-2021-0015, pp. 228-242 , 2021 .

[17] S. Lee, E. K. Choe and B. Park, "Exploration of Machine Learning for Hyperuricemia Prediction Models Based on Basic Health Checkup Tests," *Journal of Clinical Medicine*, Vols. 8, 172, no. https://doi.org/10.3390%2Fjcm8020172, pp. 1-10, 2019.

[18] R. Bandi, J. Amudhavel and R. Karthik, "Machine Learning with PySpark - Review," *ndonesian Journal of Electrical Engineering and Computer Science*, Vols. vol. 12, no. 1, no. DOI: 10.11591/ijeecs.v12.i1., pp. 102-106, 2018.

[19] Nofenky and D. B. Rarasati, "Recommendation for Classification of News Categories Using Support Vector Machine Algorithm with SVD," *Ultimatics: Jurnal Teknik Informatika*, vol. vol. 13 no. 2, pp. 72-80, 2021.

[20] T. E. Arijaje, T. V. Omotosho, S. A. Akinwumi, O. O. Ometan and O. O. Fashade, "Analysis of Solar Ultraviolet radiation Index over Nigeria, West Africa," *5th International Conference on Science and Sustainable Development*, vol. 993, pp. 1-11, 2022.