Indonesian Review of Physics (IRiP)
Vol.6, No.2, December 2023, pp. 99 - 121
p-ISSN: 2621-3761 | e-ISSN: 2621-2889
DOI: 10.12928/irip.v6i2.8216

Assessing the Efficacy of the UV Index in Predicting Surface UV Radiation: A Comprehensive Analysis Using Statistical and Machine Learning Methods

Edy Ervianto[1], Noveri Lysbetti Marpaung[2]*, Abu Yazid Raisal[3], Sakti Hutabarat[4], Rohana Hassan[5], Ruben Cornelius Siagian[6], Nurhalim[7], Rahyul Amri[8]
[1]Environmental Science Doctoral Program, Postgraduate Program, Riau University, Indonesia
[2,7,8]Department of Electrical Engineering, Faculty of Engineering, Riau University, Indonesia
[3]Falak Science Observatory, Universitas Muhammadiyah Sumatera Utara, Indonesia
[4]Department of Agribusiness, Faculty of Agriculture, Riau University, Indonesia
[5]Institute for Infrastructure Engineering and Sustainable Management (IIESM), Universiti Teknologi MARA, Shah Alam 40450, Malaysia
[6]Department of Physics, Faculty of Mathematics and Natural Sciences, Universitas Negeri Medan, Indonesia
Email: noveri.marpaung@eng.unri.ac.id

Article Info

Article History
Received: Jun 2, 2023
Revision: Sept 21, 2023
Accepted: Dec 23, 2023

Keywords:
UV Index prediction
Public health implications
UV radiation modelling
Predictive analytics
UV protection strategies

ABSTRACT

The study investigated the relationship between the UV Index and measurements of ultraviolet A (UVA) and ultraviolet B (UVB) radiation to evaluate the effectiveness of the UV Index in predicting and understanding UV radiation at the surface. The implications of this study are significant for public health policies and UV protection strategies. This study used a variety of statistical analyses and modelling techniques, including ANOVA, Naive Bayes classification, decision trees, artificial neural networks, support vector machines (SVM), and k-means clustering, to examine relationships and predict UV Index values. ANOVA analysis showed a significant relationship between the UV Index and UVA and UVB measurements. Prediction models such as Naive Bayes classification, decision trees, and artificial neural networks showed variability in their accuracy. Notably, SVM showed a high degree of accuracy in predicting UV Index values, while k-means clustering effectively clustered the data based on similarities in UV Index and UV measurements. These findings confirm that the UV Index is a reliable indicator for predicting and understanding UV radiation levels at the Earth's surface. This research underscores the importance of developing more accurate and precise UV Index prediction models. Further investigation is essential to understand the temporal variations and environmental impacts on the UV Index, as well as the broader implications of UV exposure on public health. This study lays a strong foundation for the development of early warning systems and more effective UV protection strategies, ultimately improving public health outcomes and safety measures against UV radiation.

To cite this article:
E. Ervianto *et al.*, "Assessing the Efficacy of the UV Index in Predicting Surface UV Radiation: A Comprehensive Analysis Using Statistical and Machine Learning Methods," *Indones. Rev. Phys.*, vol. 6, no. 2, pp. 99–121, 2023, doi: 10.12928/irip.v6i2.8216.

I. Introduction

The research originated from a deep concern about the impact of ultraviolet (UV) light exposure on human health and the environment. As global warming increases and the ozone layer depletes, the intensity of UV exposure, particularly UVA and UVB, increases [1]–[3]. Excessive UV rays can cause a range of serious health problems, including a higher risk of skin cancer, faster skin aging, and various eye disorders [4]–[6]. UV rays cause ecosystem damage, damage crops, and degrade water quality [7]–[9]. The problem is becoming increasingly urgent due to worsening global climate change, which exacerbates UV exposure [1], [3], [10]. In an effort to understand and address the negative impacts, research focuses on analyzing the influence of the UV Index on UVA and UVB variables and investigating the relationship between them. The research developed effective mitigation strategies to protect human health and preserve the environment. Research highlights and attempts to address problems caused by excessive UV exposure. Careful and comprehensive analysis, contributing to the assessment of UV-related risks, and assisting in the development of better analysis methods and prediction models. The research is expected to be used by policy makers, researchers, and health practitioners to implement more effective measures to protect human health and the environment from UV hazards.

The study aims to analyze the effect of UV Index on UVA and UVB variables. The research measures the extent to which the UV Index affects the level of UVA and UVB exposure under various environmental conditions. The study determined whether there were significant differences in UVA and UVB exposure based on variations in the UV Index. The research investigated the relationship between UVA and UVB variables. The research examined the correlation between the two variables to understand the interactions that occur. Research identifies factors that may influence the relationship between UVA and UVB. Research evaluating the performance of Naive-Bayes classification models in predicting UV level categories. The research developed a prediction model that can categorize UV levels into various classes, such as "Very Low", "Low", "Medium", "High", and "Very High". The research tested the accuracy and reliability of the Naive-Bayes model in predicting UV level categories based on observational data. The research used the decision tree method to estimate the UV Index value. The research involved creating a decision tree model that could predict UV Index values based on various observed features. The research analyzed how each feature contributed to the prediction of the UV Index value. The research developed an understanding of the impact of UV exposure on health and the environment. Research provides scientific knowledge on how variations in the UV Index, UVA and UVB can affect human health, including the risk of skin cancer and premature aging. Research assesses the environmental impacts of increased UV exposure, including effects on ecosystems and air quality.

Research is expected to provide a range of significant benefits. Research contributes to science and research by deepening scientific understanding of the relationship between the UV Index, UVA and UVB and their impact on human health and the environment. The data and findings from the research are used by other researchers for further research or for policy development relating to UV exposure. This study developed an effective prediction model. The accurate prediction models for UV level categories generated from this research can be used by authorities to provide early warnings to the public regarding the risks of UV exposure. This research provides a decision tree-based analysis tool that can be used to identify the main factors affecting UV Index values. This research will improve public awareness and health policy. Useful information on the dangers of UV exposure and ways to protect themselves from excessive UV exposure generated from this research can help people to be more vigilant. Research can assist governments and health agencies in designing educational programs and policies to reduce health risks associated with UV exposure. The research has applications in environmental management. Data generated from research can be used to monitor and manage the environmental impacts of UV exposure, such as impacts on plants, animals and water quality. Research can support the development of mitigation strategies to protect ecosystems from damage due to increased UV exposure. This research will evaluate the performance of the analysis and classification models used. The research will assess the effectiveness of the ANOVA analysis method and the Naive-Bayes and decision tree classification models in the study. Based on the research findings, the research will provide recommendations for the use of better analysis and classification methods in the future.

The study has some limitations that need to be noted to understand the interpretation of the results. The study used data sourced from a single training dataset. The use of data from a single source may limit the generalizability of the results to larger or different populations [11], [12]. Additional data from multiple sources or different geographical locations can provide more comprehensive results [13]. The variability of the data in the dataset used does not reflect the variation that exists in the wider population, which can affect the accuracy and validity of the model built. The study has limitations regarding the variables studied. The main focus is only on three variables: UV Index, UVA, and UVB. Influential variables, such as atmospheric conditions, pollution levels, or weather conditions, were not included in the analysis. This limits a thorough understanding of the factors that influence the UV Index. The study did not explore in depth the interactions between other variables that may play a role in influencing UVA and UVB, such as geographical or temporal influences. The use of ANOVA and Naive-Bayes classification methods and decision trees are specific method choices. They have limitations in identifying non-linear patterns or complex interactions

p-ISSN: 2621-3761
e-ISSN: 2621-2889

between variables. Other methods such as non-linear regression, artificial neural networks, or ensemble methods can provide different explanations. The decision tree method used in this study is the rpart method, while there are other methods in decision tree modeling such as Random Forest or Gradient Boosting that provide more accurate results and can consider more features and interactions between variables [14]–[16]. The model evaluation showed some limitations. The Naive-Bayes model evaluation showed good performance in the "Very Low" and "Low" categories, but was less effective in the other categories. This suggests that the model has limitations in predicting data with more complex or varied distributions. The study did not provide complete information on the evaluation metrics used to assess model performance, such as precision, recall, F1-score, or AUC-ROC. The study did not consider external factors that could influence the results, such as climate change, environmental policies, or UV protection technologies. The influence of these factors can be significant and needs to be considered in further research. The data used was taken over a period of time and does not reflect seasonal changes or long-term trends. Long-term and seasonal data may provide a more accurate picture of the patterns of UV Index, UVA, and UVB. Results cannot be directly applied to the wider population without further verification. Further studies with various data sets are needed to ensure generalization of the results. The implementation of research results in policies or mitigation measures needs to be further tested to assess their effectiveness in the real world. This includes assessing the costs, benefits and impacts of implementing the research results.

The research fills several gaps in the literature regarding the analysis of the influence of the UV Index on UVA and UVB variables and the use of UV prediction models. Most studies tend to focus on the individual impacts of UVA and UVB on human health and the environment, without comprehensively combining and analyzing the relationship between the UV Index and these two variables [17], [18]. Previous studies have often only examined one type of UV radiation at a time, lacking the ability to provide a complete picture of how UVA and UVB interact and contribute to total UV exposure [19], [20]. Many studies use simple statistical methods that are insufficient to capture the complexity of the relationship between UV Index, UVA, and UVB. There is a need for the use of more sophisticated and diverse analytical methods to understand the deeper dynamics of UV data. Previous research has often not considered alternative analysis and classification methods that may be more effective or provide a different perspective. For example, the use of other machine learning methods such as Random Forest or Support Vector Machine (SVM) for classification and prediction is still rarely applied. Time series analysis and dynamic prediction models that consider temporal changes in UV data are rarely used in existing studies. Many previous studies used limited datasets or data from specific locations, which do not represent global conditions or wider geographical variations. Many studies do not provide an in-depth evaluation of the performance of the prediction models used. Model evaluation is often limited to prediction accuracy without considering other metrics such as precision, recall and F1-score. The lack of cross-validation analysis and model testing with independent data to ensure generalizability of results is also a common weakness in previous studies. A comprehensive evaluation is essential to ensure that the model is reliable and performs well across different data conditions. Existing studies often do not link their scientific findings to practical implications or policies that can be taken to reduce the risk of UV exposure. The lack of clear recommendations on mitigation measures that can be taken by individuals or governments based on research results points to the need for more applicable and public policy-relevant research.

Research offers a significant contribution in deepening the understanding of the relationship between the UV Index, UVA, and UVB. While there have been previous studies on this topic, this research takes a more specific approach by highlighting the complex interactions between the three. Integrating statistical analysis methods such as ANOVA to test the influence of the UV Index on UVA and UVB variables, as well as utilizing predictive models such as Naive-Bayes and decision trees, the research not only confirmed the significant influence of the UV Index on both variables, but also provided a deeper understanding of how these variables influence each other. Performance evaluation of predictive models, such as the one conducted in the study showed the Naive-Bayes model to be well capable of predicting UV Index levels with some accuracy, although it has limitations in certain categories of predictions. The visual representation of classification predictions not only helps in assessing the accuracy of the model, but is also important in the context of risk assessment related to UV exposure for humans and the environment. With increasing attention to the health impacts of UV exposure, the research emphasized the importance of understanding the relative contributions of UVA and UVB to the UV Index separately. Research provides a firmer foundation for the development of more effective and responsive mitigation strategies against health and environmental threats caused by UV light.

II. Theory

The Role and Impact of the UV Index on Public Health

The UV index is an international standardized measure of the intensity of ultraviolet radiation from the sun reaching the Earth's surface [21]–[23]. UV index developed by World Health Organization (WHO), World Meteorological Organization (WMO), United Nations Environment Programme (UNEP), dan International Commission on Non-Ionizing Radiation Protection (ICNIRP) [24]–[26]. The aim is to raise public awareness

p-ISSN: 2621-3761
e-ISSN: 2621-2889

of the dangers of excessive UV radiation, which can cause health problems such as skin cancer and eye damage. Ultraviolet A radiation (UVA) is the part of the UV spectrum that has a wavelength between 320 and 400 nm [27]–[29]. UVA plays a role in skin penetration and can cause skin aging and contribute to the development of skin cancer [30]–[32].

Research by [33] found that the UV Index can be used as a proxy to measure UV radiation exposure, including UVA, although the UV Index is influenced more by UVB. The study emphasized that although the UV Index was originally developed to measure UVB, it also reflects significant UVA exposure. Research conducted by [21] found that variations in the UV Index were significantly related to changes in total UV radiation levels, which includes UVA. The results suggest changes in UV Index values can be used to estimate variations in UVA exposure levels, providing an empirical basis that the UV Index is an effective indicator of the different types of UV radiation reaching the Earth's surface. Research by [34] showed that variations in the UV Index significantly affect UVB exposure in different geographical locations. Studies show that differences in the UV Index can be attributed to differences in UVB exposure at the Earth's surface. This is due to factors such as altitude, latitude, and time of year that affect the intensity of UV reaching the Earth's surface [35]. According to [36], The UV Index is an indicator designed to convey the level of risk of UV exposure that varies by environmental and geographical factors. Inter-group differences in the UV Index reflect variations in UVB exposure resulting from different environmental conditions. Research by [37] showed that the UV Index significantly influences the level of UVB radiation measured at the Earth's surface. Research reveals that variations in UV Index are closely correlated with changes in UVB intensity, with an increase in UV Index indicating an increase in UVB levels. Research by [38] found that the UV Index can be used as a predictive tool to estimate UVB radiation levels. In the study, UVB radiation measurement data showed a strong correlation with the observed UV Index values, supporting the claim that the UV Index is a valid indicator for UVB variations. Research by [39] supports the assertion that the UV Index is a strong indicator for predicting UVB radiation levels. Studies show that UV Index values can be used to estimate the risk of UVB exposure, which is important for public health protection measures. Furthermore, studies by [40] confirms that the UV Index is a reliable tool for monitoring and predicting UVB exposure.

Studies using analysis of variance (ANOVA) to explore the relationship between environmental variables, particularly UV radiation exposure, and health have provided strong evidence of the effectiveness of this method in evaluating their impact. Studies conducted by [41] illustrates the use of ANOVA to analyze the effects of UV radiation on DNA. Research shows that variations in UV exposure can significantly affect the extent of genetic damage. Research conducted by [42], where they used ANOVA to evaluate the relationship between UV exposure levels and skin cancer incidence. The study results showed that variations in UV exposure levels correlated significantly with variations in skin cancer incidence. The findings underscore that ANOVA is not only able to identify significant relationships between environmental exposures and health, but also provides a strong scientific foundation to support the need for further research on environmental factors in public health.

Naive Bayes Model for UV Index Prediction

Naive Bayes is a classification algorithm based on Bayes' Theorem with the assumption of independence between features [43]–[45]. In UV Index prediction, the model uses features such as UV-A radiation, UV-B, and other environmental parameters to estimate UV Index values. The model is highly efficient and fast, and is suitable for large datasets.

Research using the Naive Bayes model to predict UV Index values is based on the concept that this algorithm can identify patterns from training data to predict target values in test data [46], [47]. The basic concept of machine learning emphasizes the development of algorithms that allow computers to learn from data and make decisions based on identified patterns [48], [49].

Research [50] shows that the algorithm can be effective in predicting environmental values, including in cases such as UV Index prediction which utilizes UV-A and UV-B radiation data as predictor features. Research conducted by [51], and [52] discussed in detail the strengths and weaknesses of the Naive Bayes model and its applications in various domains, including sensor data-based prediction. They underline that despite its simplicity, Naive Bayes can provide good results in situations where the independence assumption is reasonably close. This concept is supported by the explanation of [53] which outlines that machine learning does not only rely on Naive Bayes as the only algorithm, but also includes various other techniques and models used to analyze and predict environmental data. In the evaluation of prediction model accuracy, as in research, data visualization and summary statistics play an important role. Techniques such as distribution plots, confusion matrix, and evaluation metrics such as accuracy, precision, recall, and F1-score are used to provide a deeper understanding of how Naive Bayes models predict UV Index values. [54] and [55] separately emphasized the importance of effective visualization and summary statistics in conveying information from data analysis, which can help researchers evaluate and improve the performance of prediction models.

Decision Tree and Artificial Neural Networks for UV Index Prediction

A decision tree is a machine learning algorithm used for classification and prediction based on measured features [56], [57]. The algorithm works by dividing the

p-ISSN: 2621-3761
e-ISSN: 2621-2889

dataset into subsets based on the most significant features, forming a tree-like structure with nodes as features and branches as decision rules [58].

The study used decision trees and artificial neural networks to predict and classify UV Index values. Both methods have their own advantages in handling complex and heterogeneous environmental data. The use of decision trees in research is supported by studies conducted by [59]. Which shows that decision trees are very effective in performing classification and prediction due to their ability to handle complex and heterogeneous datasets. Decision trees work by dividing a dataset into smaller subsets based on the most significant features in the dataset. The process forms a tree-like structure, where each node represents a feature and each branch represents a decision rule [58]. At the lowest level of the tree, the leaves, there is a final decision or prediction [60]. Decision trees can decompose large and complex datasets into a form that is easier to understand and interpret.

A decision tree can be used to classify UV Index values based on several measured features. The features include ultraviolet A (UVA) and B (UVB) radiation intensity, time of day, and weather conditions. The decision tree explains how each of these features affects the UV Index value. For example, UVB radiation intensity may be more significant in determining the UV Index value compared to UVA radiation intensity, or certain weather conditions such as clouds or rain may have a large impact on the UV Index value. From the decision tree, it can be shown that on sunny days with high UVB radiation intensity, the UV Index value tends to be high. Conversely, on cloudy days with low UVB radiation intensity, the UV Index value may be lower. Another advantage of decision trees is their ability to handle missing data and handle various types of data, both numerical and categorical [61], [62]. This makes decision trees a flexible and powerful tool in the analysis of environmental data such as the prediction of UV Index values. Research by [63] shows that decision trees can be interpreted easily and provide a clear view of the relationship between input and output variables. Decision tree analysis can help identify the most significant features in influencing the UV Index value.

Artificial neural networks are machine learning models inspired by how the human brain works [64], [65]. The model consists of layers of neurons that are connected and work in parallel to process information. Artificial neural networks are capable of handling large and complex datasets, as well as discovering non-linear patterns that may not be visible with other methods [66], [67]. Study by [68] introduced the concept of backpropagation, which is a key algorithm in training artificial neural networks. Research can develop complex and accurate models to predict the value of the UV Index. Artificial neural networks have the ability to learn from training data and correct the weights of connections between neurons based on prediction errors [69], [70]. This allows neural networks to produce accurate predictions despite variations in the data. Research by [71] shows that artificial neural networks, especially deep learning, have the ability to perform highly accurate predictions in various domains, including the prediction of environmental values such as UV Index.

Support Vector Machine (SVM) in UV Index Prediction

Support Vector Machine (SVM) is one of the machine learning methods often used for classification and regression tasks [72], [73]. SVM works by finding a hyperplane that separates the data into different classes with maximum margin [73], [74]. SVMs are very effective in high-dimensional spaces and in cases where the number of dimensions is greater than the number of samples [75], [76]. SVMs are known to be resistant to overfitting, especially in high-dimensional spaces [76], [77].

The basic concept of a Support Vector Machine (SVM) involves several important elements that make this method effective in classification and regression tasks [73], [78]. One of the key elements is the hyperplane and margin. A hyperplane is a plane in high-dimensional space that is used to separate data sets into different classes [79]. SVM has the main goal of finding a hyperplane that maximizes the distance or margin between the hyperplane and the closest data point of each class [73], [74]. Maximizing the margin, SVM models can improve generalization ability and reduce the possibility of overfitting [80], [81]. SVM uses a technique called kernel trick to handle data that is not linearly separable. There are several types of kernels that are commonly used in SVM. Linear kernels are used for linearly separable data, where a simple hyperplane is sufficient to separate the data [82]. Polynomial kernels map data to a higher dimensional space using polynomials, which allows for the separation of more complex data [83]. Radial Basis Function (RBF) kernel, also known as Gaussian kernel, is one of the most effective kernels for non-linear data [84]. The RBF kernel maps the data to a higher dimensional space using a Gaussian function, thus enabling the separation of more complex and non-linear data [85]. Another important parameter in SVM is the regularization parameter (C) [86]. The parameters control the trade-off between maximizing margin and minimizing classification error [87]. The C parameter determines the extent to which the SVM model will try to separate the data by a large margin while reducing the misclassification of the training data [88]. The C parameter can help SVM models achieve an optimal balance between bias and variance, which is important for good performance on unseen data in advance [89]. In UV Index prediction, basic concepts are used to train SVM models that are able to predict UV Index values with high accuracy. Hyperplane and margin help in separating UV Index data based on environmental parameters, kernel trick allows handling complex and non-linear data [90]. Proper setting of regularization parameters ensures that the SVM model is not only accurate on training data but also has good generalization ability on new data [78].

p-ISSN: 2621-3761
e-ISSN: 2621-2889

One of the most important studies conducted by [51], They used SVM to predict air quality based on environmental parameters such as humidity, temperature, and pollutant concentration. The results showed that SVM has high accuracy in air quality prediction, proving the effectiveness of this method in handling complex environmental data. Study conducted by [91] and [92], applied SVM to predict weather conditions using historical meteorological data. The research showed that SVMs were able to produce accurate and reliable predictions for weather conditions, once again confirming the ability of SVMs in environmental data analysis. Research by [93], [94] discussed the use of various machine learning models, including SVM, to predict UV radiation exposure. The study showed SVM has superior performance in UV radiation prediction compared to other models.

K-means Clustering in UV Index and UV Measurements Analysis

K-means clustering is an unsupervised learning method used to group data into a certain number of clusters based on similarity of features [95], [96]. The algorithm aims to partition n data into k clusters, such that each data belongs to the cluster with the closest centroid, as measured by the Euclidean distance [97], [98].

The process of the K-means clustering algorithm starts with initialization. The desired number of clusters (k) is determined. Determining the number of clusters is a crucial step because it will affect the final result of clustering. After the number of clusters is determined, the next step is to select the initial centroid. The initial centroid selection can be done randomly, but there are other methods such as K-means++ that can be used to select the initial centroid in a smarter way to speed up convergence and improve clustering accuracy [99], [100]. The algorithm proceeds to the cluster assignment stage. Each data in the dataset is measured for its distance to each selected centroid. The most commonly used distance is the Euclidean distance, although other distance metrics can also be used depending on the data. Each data is then assigned to the closest centroid based on the calculated distance. The process groups the data into temporary clusters each centered on its nearest centroid. Once the data is assigned to its respective cluster, the centroid is updated by calculating the average position of all the data belonging to that cluster. The update is done by calculating the average value of each data feature in the cluster. The new centroid then becomes the new center of the cluster. The K-means algorithm then repeats the cluster assignment and centroid update steps. At each iteration, the data is again measured its distance to the new centroid and assigned to the nearest centroid. The centroid is again updated based on the new cluster assignment. The iterative process continues until convergence is achieved, i.e. when there is no significant change in the cluster assignments or centroid positions [101]–[103]. Convergence indicates that the algorithm has found a stable cluster split.

In the study of UV Index and UV measurements, K-means clustering was used to group data based on similarities in UV Index, UVA, and UVB measurements. The process makes it possible to discover hidden patterns in the data that may not be detected through simple descriptive or statistical analysis. Grouping data into meaningful clusters allows for the identification of groups of data that have similar characteristics, such as high or low levels of UV exposure [104]. Data grouped in clusters with high UV Index indicates certain areas or conditions that are susceptible to excessive UV exposure, clusters with low UV Index may indicate safer conditions.

Research by [105] shows how K-means clustering can be used to group environmental data. The study was able to identify statistically significant patterns. Applying the K-means algorithm, researchers were able to cluster geographic areas based on similarities in UV exposure [105]. Clustering helps in understanding the variation of UV exposure across regions and time, which is very important in epidemiological studies related to the impact of UV exposure on public health. The results showed that K-means clustering can identify groups of data with similar characteristics. Studies conducted by [105] and [106] highlights the application of K-means clustering in weather data analysis to predict extreme atmospheric conditions. In the study, K-means was used to cluster complex meteorological data, including UV light intensity. The results show that K-means clustering can effectively cluster weather data, which provides more in-depth information about atmospheric conditions and variations in UV light intensity. The study supports the use of K-means in clustering UV data to understand variations in light exposure and aid in the prediction of extreme weather conditions.

III. Method

Anova Method

The method used in the research is ANOVA (Analysis of Variance), using the `aov` function [107]. The study analyzed the impact of the UV Index variable on the UVA and UVB variables. The procedure begins by performing an ANOVA analysis using the formula `UVA ~ UV Index` on the `data train` dataset. In the analysis, the model consists of two terms namely UV Index and Residual. For each term in the model, the sum of squares was calculated, providing a measure of the total variation attributable to each term [108]. The degrees of freedom (Df) for each term are calculated, representing the number of independent values that can vary [109], [110]. The standard error of the residuals is then determined, which serves as an estimate of the error in the model [111], [112]. To summarize the results, the `summary (anova_result)` function is used. The summary provides comprehensive information regarding the degrees of freedom (Df), sum of squares (Sum Sq), and mean square (Mean Sq) for each term in the model [113], [114]. F values and probabilities

p-ISSN: 2621-3761
e-ISSN: 2621-2889

(Pr(>F)) were calculated to assess the significance of the effect of the UV Index variable on the UVA variable. The study extended this ANOVA test by testing the effect of the UV Index variable on the UVB variable. The test results are presented in summary form which includes degrees of freedom (Df), sum of squares (Sum Sq), and mean square (Mean Sq) for each term in the model.

The study evaluated the influence of UV Index variables on UVA and UVB variables. Using ANOVA analysis, the method provides an understanding of whether there is a significant difference between the groups formed based on the UV Index variable for each of the variables under study. For the UVA variable, the ANOVA analysis is represented by hypotheses, where the null hypothesis (H_0) states that $\mu_1 = \mu_2 = \mu_3 = ... = \mu_k$, indicating that there is no significant difference between the groups formed based on the UV Index variable [115], [116]. The alternative hypothesis (H_1) states that at least one pair of groups has a significant difference, implying that the UV Index variable does have a significant impact on the UVA variable. Similarly, for the UVB variable, ANOVA analysis was conducted with the hypothesis that the null hypothesis (H_0) states $\mu_1 = \mu_2 = \mu_3 = ... = \mu_k$, which indicates that there is no significant difference between the groups formed based on the UV Index variable and the UVB variable [117], [118]. The alternative hypothesis (H_1) states that at least one pair of groups has a significant difference, indicating that the UV Index variable significantly affects the UVB variable.

Naive Bayes Classification Method

The research method used is Naive Bayes classification [119]. In this method, there are two types of data used, namely "predictions" and "data test". The "predictions" data contains the predicted values generated by the model for the target variable being tested. The "data_test" data consists of several observed variables, such as "Year", "Month", "Day", "UVA", "UVB","UV Index", and "Prediction". Summary statistics are given for each of these variables.

To visualize the classification results, a classification plot using the Naive Bayes method was used [120]. This plot shows the distribution of classification results on the test data. The x-axis of the plot shows the actual value of the UV Index variable, while the bars on the plot are filled with colors that represent the prediction results of the Naive Bayes model. The height of each bar on the plot indicates the number of observations or frequency of each prediction category in the test data. This plot provides a visual understanding of how well the Naive Bayes model classifies the test data.

The results of the confusion matrix are also given. Confusion matrix shows the prediction results of the Naive Bayes classification model on the test data [121], [122]. The results show the number of predictions that match the actual class, as well as the prediction error that occurs. The performance of the model in predicting the "Very Low" and "Low" categories based on the available test data is mentioned, as well as the inability of the model to make predictions for the other categories.

Plot displays the proportion of actual classification and predicted classification. The x-axis shows the actual classification, while the y-axis shows the proportion. A different color on the plot indicates the predicted classification. This plot provides visual information about the degree to which the predicted classification matches the actual classification.

The research method used, namely Naive Bayes classification, can be represented in algebraic mathematical formulas [123]. To calculate the probability of the target class (category) based on the observed data:

$$P(C \mid X) = \frac{P(X \mid C) \cdot P(C)}{P(X)} \tag{1}$$

Where P(C | X) is the probability of target class (category) C based on data X, P(X | C) is the probability of data X occurring if target class (category) C is true, P(C) is the probability of target class (category) C as a whole, P(X) is the probability of data X occurring as a whole. To calculate the probability of data X occurring if target class (category) C is true:

$$= P(X \mid C) \tag{2}$$
$$= P(X_1 \mid C) \cdot P(X_2 \mid C) \cdot \ldots \cdot P(X_n \mid C)$$

Where P(X | C) is the probability of data X occurring if target class (category) C is true, P(Xi | C) is the probability of attribute (variable) value Xi in data X if target class (category) C is true, n is the number of attributes (variables) in data X. To calculate the probability of target class (category) C as a whole:

$$P(C) = \frac{\text{number of data with target class C}}{\text{total amount of data}} \tag{3}$$

To calculate the probability of data X occurring overall:

$$P(X) = P(X_1) \times P(X_2) \times \ldots \times P(X_n) \tag{4}$$

Where P(X) is the probability of data X occurring as a whole, P(Xi) is the probability of attribute value (variable) Xi in data X as a whole, and n is the number of attributes (variables) in data X.

These formulas are used in the Naive Bayes method to classify the test data based on the probabilities calculated from the training data [43], [44], [124]. Visualization of the classification results using classification plots and confusion matrix provides a visual understanding of the model's performance in classifying the test data, as well as the degree to which the predictions match the actual classification.

p-ISSN: 2621-3761
e-ISSN: 2621-2889

Decision tree method

The research uses a decision tree modeling method using the rpart algorithm [14], [56]. A decision tree is used to predict UV Index values based on relevant features. The approach involves data partitioning based on important attributes such as UVB, UVA, Month, Year, and Day. Each node in the decision tree represents the partitioning of the data based on the attributes. The decision tree starts with a root node that contains a number of observations and the average value of the UV Index [56], [125], [126]. The data is divided into two branches based on the UVB feature values. The left and right branches show different mean values and Mean Squared Errors (MSE) [127]. The process of division and branching continued at each node, taking into account the complexity of the tree and the importance of the features used [128], [129]. This aided in classifying observations and predicting the UV Indexvalue based on relevant features.

The decision tree involves observing the distribution of data at each node [56], [130]–[132]. Starting from the root node, data is partitioned based on the condition of a particular variable such as UVB. Each node provides information such as the number of observations, deviation value, and predicted value [133], [134]. Deviance measured the discrepancy between the model's predictions and the actual values, with the primary objective being to minimize deviance during the division process [135]. The average value of UV Index and Mean Squared Error (MSE) were used to evaluate the accuracy of the model in predicting the target value. To predict the target value, yhat, based on the decision tree [56], [130], [136]:

$$y_{\text{hat}} = T(x) \tag{5}$$

Where $T(x)$ is a function that generates predictions based on splitting the data at each node in the decision tree. To represent each node in the decision tree:

$$T(x) = \sum yval_i \cdot I(x \in R_i) \tag{6}$$

Where $yval_i$ is the predicted value at node i, R_i is the region defined by the separation condition at node i, and $I(x \in R_i)$ is an indicator function that takes the value 1 if x is in region R_i and 0 otherwise. The splitting at each node in the decision tree is based on a splitting condition that splits the data based on a particular feature value [125], [137]. If splitting is done based on feature F with a splitting boundary c, then the splitting condition can be expressed as:

$$x[F] < c \tag{7}$$

At each node, there is also the mean of the target value y and the Mean Squared Error (MSE):

$$\text{mean} = \sum yval_i \cdot p_i \tag{8}$$

$$\text{MSE} = \sum (y_i - yval_i)^2 \cdot p_i \tag{9}$$

Where y_i is the actual target value, p_i is the proportion of observations falling into the R_i region, and yvali is the predicted value at node i.

Neural Network Method

The method used is the development and training of artificial neural network models [138]–[140]. The model was designed with 5 input variables and 1 output variable, using observed data consisting of responses and covariates for training purposes. The model was built by incorporating appropriate error and activation functions, as well as non-linear outputs [141]–[143].

The data used in the study is organized in a data.frame format, which consists of 6 columns. The neural network model was then used to predict outcomes based on the data [144], [145]. The results showed a tendency for the model to predict the same outcome for most of the test data. However, there were notable differences in the range of values between the predicted and actual values in the test data. The study reported the minimum, average, and maximum values for the predicted and actual results. The difference in the range of values between the predicted and actual results highlights the potential for improving the accuracy of the model [146], [147]. During the construction of the neural network, the number and size of the hidden layers in the model are determined, and the activation functions used for the neurons in the hidden layer and output layer can also be observed [148]–[150]. A graphical representation of the structure and architecture of a neural network helps in understanding how the model works and the flow of information within it, including the connections between neurons [151], [152].

A neural network diagram visualization was created to represent the structure and interconnections among the layers in the model [151], [152]. Each node in the plot represents a neuron or unit in the model, the lines connecting the nodes describe the relationship between neurons in different layers [153], [154]. The plot depicts the weights or parameters that connect the neurons. From the visualization, a better understanding of the flow of information through the network and the interconnections between the neurons can be gained [155], [156].

The method involves developing and training an artificial neural network model using mathematical equations. The artificial neural network model is structured with an input layer consisting of five variables namely x_1, x_2, x_3, x_4, and x_5. These input variables are processed through the first hidden layer, which contains five neurons namely h_1, h_2, h_3, h_4, and h_5. The neurons in the first hidden layer are interconnected with the neurons in the second hidden layer, which consists of three neurons namely h_6, h_7, and h_8. The connection between the first and second hidden layers is determined by a set of weights

p-ISSN: 2621-3761
e-ISSN: 2621-2889

namely w_1, w_2, w_3, w_4, w_5, w_6, w_7, and w_8. The neurons in the second hidden layer are then connected to the output layer, which generates the final output variable, y [148], [157]. The weights connecting the neurons in the second hidden layer are then connected to the output layer [158], [159]. The weights connecting the neurons in the second hidden layer to the output are represented by v_1, v_2, v_3, v_4, v_5, v_6, v_7, and v_8. Through training, the neural network model adjusts these weights to minimize the error between the predicted output and the actual output, thus optimizing its performance in predicting y based on the given input variables [160], [161].

In the development of artificial neural network models, the activation function $\phi(x)$ plays a very important role [143], [149]. The function can be sigmoid, ReLU, or another type of non-linear activation function, depending on the model used [162], [163]. The formula for obtaining the output (y) of the neural network model is:

$$y = \phi(v_1 \cdot h_1 + v_2 \cdot h_2 + v_3 \cdot h_3 + v_4 \cdot h_4 + v_5 \cdot h_5 + v_6 \cdot h_6 + v_7 \cdot h_7 + v_8 \cdot h_8) \quad (10)$$

To calculate the value in the first hidden layer (h_1, h_2, h_3, h_4, h_5), use the formula:

$$h_1 = \phi(w_1 \cdot x_1 + w_2 \cdot x_2 + w_3 \cdot x_3 + w_4 \cdot x_4 + w_5 \cdot x_5)$$
$$h_2 = \phi(w_6 \cdot x_1 + w_7 \cdot x_2 + w_8 \cdot x_3 + w_1 \cdot x_4 + w_2 \cdot x_5)$$
$$h_3 = \phi(w_3 \cdot x_1 + w_4 \cdot x_2 + w_5 \cdot x_3 + w_6 \cdot x_4 + w_7 \cdot x_5)$$
$$h_4 = \phi(w_8 \cdot x_1 + w_1 \cdot x_2 + w_2 \cdot x_3 + w_3 \cdot x_4 + w_4 \cdot x_5)$$
$$h_5 = \phi(w_5 \cdot x_1 + w_6 \cdot x_2 + w_7 \cdot x_3 + w_8 \cdot x_4 + w_1 \cdot x_5)$$

Value in the second hidden layer (h_6, h_7, h_8), the following formula is used:

$$h_6 = \phi(w_2 \cdot h_1 + w_3 \cdot h_2 + w_4 \cdot h_3 + w_5 \cdot h_4 + w_6 \cdot h_5)$$
$$h_7 = \phi(w_7 \cdot h_1 + w_8 \cdot h_2 + w_1 \cdot h_3 + w_2 \cdot h_4 + w_3 \cdot h_5)$$
$$h_8 = \phi(w_4 \cdot h_1 + w_5 \cdot h_2 + w_6 \cdot h_3 + w_7 \cdot h_4 + w_8 \cdot h_5)$$

The weights (w_1, w_2, ..., v_7, v_8) and input values (x_1, x_2, ..., x_5) used in this calculation are obtained through the model training process using the observed data. In the process, the neural network model learns to adjust these weights so that it can make predictions based on the input data [164], [165]. The complex interplay between the activation function and the number of weights allows neural networks to learn and make accurate predictions [143], [149].

After formula implementation, the artificial neural network model was successfully developed and provided significant prediction results. Evaluation of the prediction results showed a tendency to predict similar results for most of the test data used [166]. The difference in the range of values between the predicted and actual values indicates the potential to improve the accuracy of the model in making predictions [167].

Support Vector Machine Method

Research using the Support Vector Machines (SVM) method with predetermined parameters [72], [168]. The SVM model type is eps-regression with radial SVM-Kernel, with cost, gamma, and epsilon parameters [169]. SVM model prediction analysis results on test data. In the resulting plot, the x-axis (horizontal) shows the actual values of the test data, while the y-axis (vertical) shows the predicted values of the SVM model. Each point on the plot represents one test data, reflecting the actual value and the corresponding predicted value. The points on the plot are marked in blue, indicating the mapping between the actual and predicted values. To facilitate the interpretation of the performance of the SVM model in predicting actual values, a diagonal line with red color and a dashed line as a reference line are used. The lines indicate the expected position if the actual and predicted values are similar. If the points on the plot tend to approach the diagonal line, it can be concluded that the SVM model has a high level of accuracy. However, if the points are widely scattered around the plot, it indicates a significant difference between the actual and predicted values generated by the SVM model.

IV. Results and Discussion

The Relationship between UV Index and Ultraviolet a Radiation Measurements using ANOVA

The analysis uses the aov function with the formula UVA ~ UV Index on the train data set. There are two terms in the model, namely UV Index and Residual. Sum of Squares for UV INDEX is 13901.489, Residual is 988.621. The Degree of Freedom for UV Index is 1, and for Residual is 2,811. The standard error value of the Residual is 0.5930407. The results show that there is a relationship between UV Index and UVA. UV Index represents the index of ultraviolet (UV) light on the surface, UVA is the actual UV light measurement. The use of the aov function and formula are used to examine the relationship between the UV index and the actual UV light measurement. In the model, the two terms identified are UV Index and Residual. The Sum of Squares used measures the variation explained by each term in the model. The Sum of Squares of the UV Index indicates the extent to which the variability in the actual UV light measurement can be explained by the measured UV light index. Sum of Squares of Residuals describes the variation that cannot be explained by the UV light index, and is therefore considered the "error" in the model. Degrees of Freedom is the number of values that can vary in a statistical test. In this case, the Degree of Freedom for the UV INDEX is 1, which indicates that the UV light index has one degree of freedom in explaining the variation in the actual UV light

p-ISSN: 2621-3761
e-ISSN: 2621-2889

measurement. As for the Residual, the Degrees of Freedom are 2811, which indicates the number of degrees of freedom remaining after considering the variability explained by the UV INDEX. The standard error value of the Residual illustrates the extent to which the actual UV light measurement values tend to differ from the values estimated by the model. A lower standard error value indicates a higher level of precision in the model.

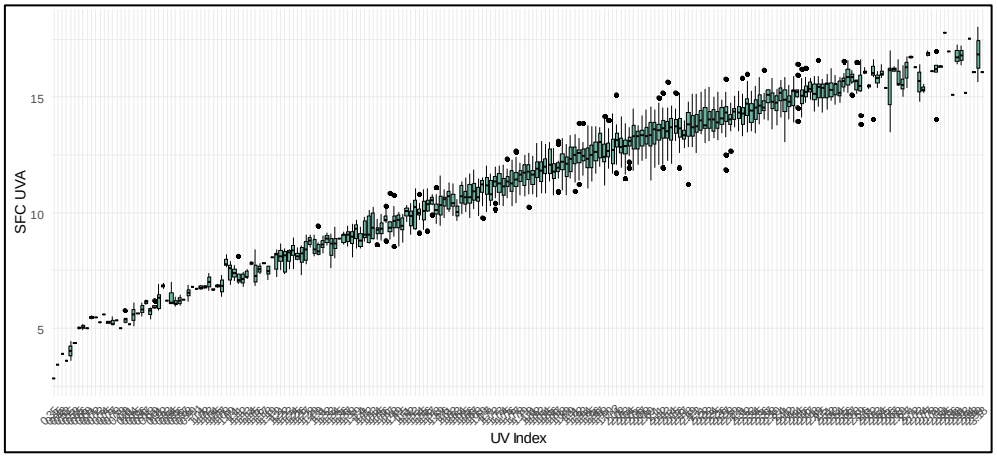

Figure 1. Analysis graph for Anova test of SFC UVA against UV Index

The results of the analysis showed a significant effect between the two variables. To summarize the results, the summary (anova_result) function was used. Two degrees of freedom (Df), with 1 for UV INDEX and 2811 for Residual. The sum of squares (Sum Sq) for the variable UV Index is 13901, Residual is 989. The Mean Sq for UV Index is 13901, while for Residual is 0. The F value is 39527 with a very low probability value (Pr(>F) of <2e-16), indicating the UV INDEX variable has a significant influence on the UVA variable. The significance code (***) indicates a very high level of significance. The UV Index variable has a significant effect on the UVA variable, with a high F value and a very low significance level. This indicates a strong relationship between the two variables in this model. The results show that the UV Index value has a significant influence on the UVA value. There is a strong relationship between the ultraviolet light index and the detected ultraviolet radiation level. This indicates that changes in the ultraviolet light index can have a significant effect on the level of ultraviolet radiation detected on the sky surface.

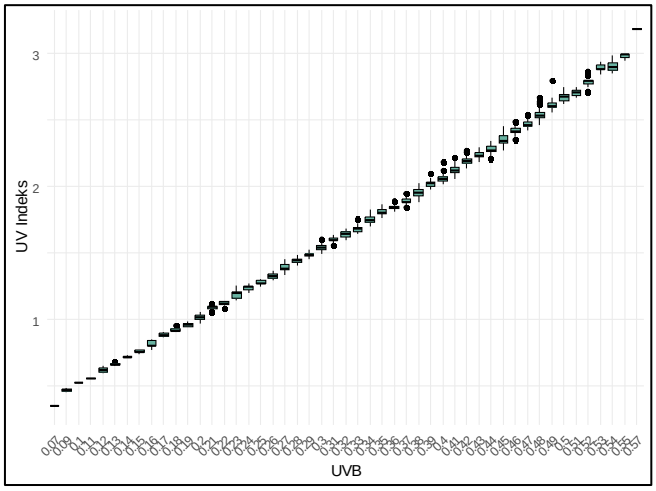

Figure 2. Analysis graph for Anova test of SFC UVB against UV Index

Analysis of Variance (ANOVA) test was conducted to compare the effect of the UV Index variable on the UVB variable. The results showed that there were significant differences among the groups formed based on the UV Index variable on the UVB variable. The ANOVA results showed that the UV Index variable had a sum of squares of 14.724 and a mean square of 14.72, with an F value of 411.821 and a very small p value (<2e-16), indicating a

p-ISSN: 2621-3761
e-ISSN: 2621-2889

significant difference between groups. Further analysis was done by examining the summary of the ANOVA test results. The summary shows that the UV Index variable has 1 degree of freedom and a sum of squares of 14.724, the residuals have 2.811 degrees of freedom with a sum of squares of 0.101. There is a residual standard error of 0.005979357.

The Relationship between UV Index and Ultraviolet B Radiation Measurements using ANOVA

Analysis of variance (ANOVA) showed a significant relationship between the predictor variable "UVA" and the target variable "UVB". This is evident from the very high F value (51.708) and very low p value (<2e-16), indicating a significant difference between the groups distinguished by the predictor variables. In the ANOVA model, the predictor variable "UVA" contributed significantly to the variation in the target variable "UVB". The predictor variables explained most of the variation in the target variable, as indicated by the high Sum of Squares value (14.060) compared to the Sum of Squares value for the Residual (0.764). The evaluation results show that the predictor variable "UVA" has a strong influence on the target variable "UVB" in the dataset. The information can be useful in further modeling and analysis.

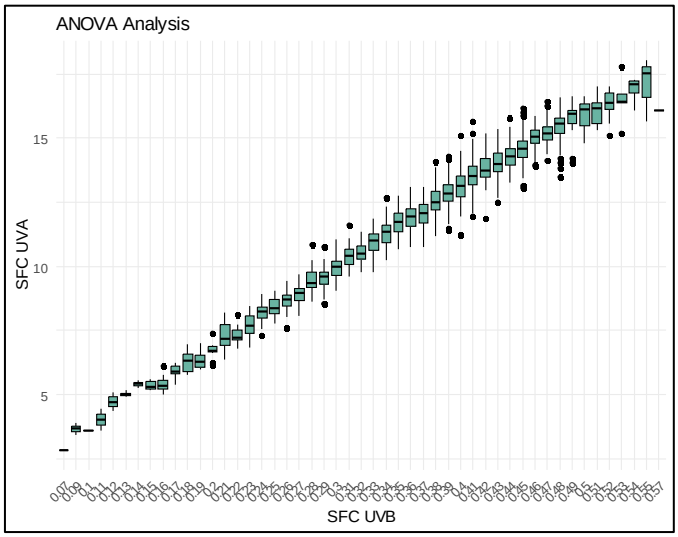

Figure 3. Anova test analysis graph of UVA against UVB

Evaluation and Interpretation of Naive Bayes Classification Predictions

In the research results, there are several prediction values generated by the model. These values show the prediction results for the tested target variables. In the dataset, there are various prediction values such as 2.12, 1.95, 1.73, and others. The sum of each prediction value is also shown. The prediction process involves the use of a pre-studied model. The model uses various methods and algorithms that have been learned from the training data to produce accurate predictions. The model uses machine learning or deep learning techniques to learn patterns and relationships in the training data, so that it can make relevant predictions for new data. At each iteration or trial, the model generates a different prediction value for the target variable under test. It can be seen that there is variation in the predicted values produced by the model for the same target variable. Variations can be caused by several factors such as data complexity, sample size, and the method used in making predictions.

The study examined the results of data known as "test data", which included several observed variables including "Year", "Month", "Day", "UVA", "UVB", "UV Index", and "Prediction". For each variable, comprehensive summary statistics such as minimum value, first quartile, median, mean, third quartile, and maximum value were calculated. In particular, for the variable "Year", the data ranges from a minimum of 2010 to a maximum of 2023, with a median year of 2015. The analysis includes a detailed calculation of certain values in the "Prediction" variable; for example, there are 116 observations where the prediction is 2.12.

The visual representation in the form of a plot illustrates the distribution of classification results across the test data. Here, the x-axis represents the actual value of the UV Index variable, while the bars are color-coded to reflect the predictions generated by the Naive Bayes model. Each different color corresponds to a different prediction category, with the height of each bar indicating the frequency of predictions in the test dataset. This graphical depiction facilitates a direct comparison between predicted and actual values, thereby assessing the performance of the model. Instances where predictions align well with actual UV Index values indicate good model accuracy, whereas significant differences prompt further examination of the model or the quality of the underlying data.

p-ISSN: 2621-3761
e-ISSN: 2621-2889

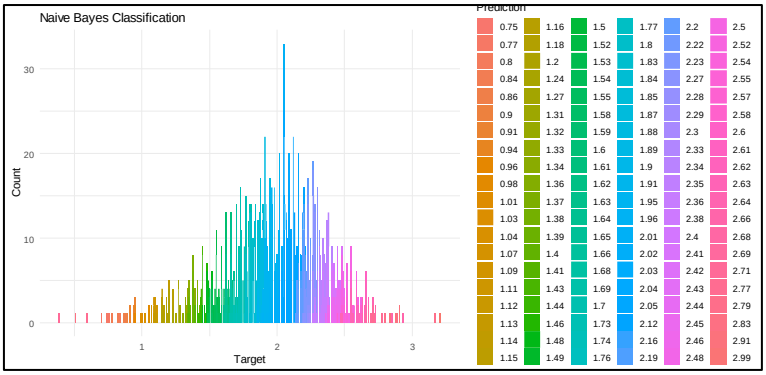

Figure 4. Number of Naive Bayes Classification vs target

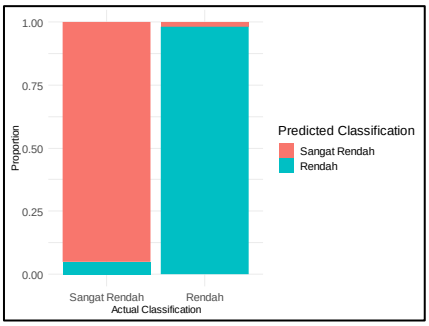

Figure 5. Naive Bayes Classification in Predicted Classification

The confusion matrix results illustrate the performance of the Naive Bayes classification model on the test data set. Specifically, the matrix shows that no predictions were made for the "Medium", "High", "Very High", and "Extremely High" categories, reflecting the absence of corresponding data in the test set. In contrast, the model showed effectiveness in predicting instances categorized as "Very Low" and "Low". Specifically, the model accurately predicted 579 instances as "Very Low" and 588 instances as "Low", matching the actual classes. Some discrepancies were noted, with 29 examples from the "Very Low" category incorrectly predicted as "Low" and 10 examples from the "Low" category incorrectly predicted as "Very Low."

To visually represent the findings, classification plots were created using the ggplot2 library. The plot depicts the distribution of actual and predicted classifications, where the x-axis shows the actual classification categories and the y-axis shows their proportions. Different colors on the plot indicate the predicted classification. Titled "Naive Bayes Classification," the plot uses a minimalist theme with adjusted text size to improve clarity.

Decision tree classification

Results with n = 4019, The first node (root) has a total of 4019 observations with a mean value of 1.957345 and MSE (Mean Squared Error) 0.1529194. The node divides the data into two branches based on the UVB value. If the UVB value is less than 0.365, the observation will be classified on the left branch (node 2) with a mean value of 1.554443 and MSE 0.08087049. If the UVBl value is greater than or equal to 0.365, then the observation will be classified on the right branch (node 3) with an average value of 2.187958 and MSE of 0.0480621. The results provide an understanding of how the measured features affect the UV index values at the atmospheric surface. Using the decision tree, it is possible to classify observations based on relevant features and predict UV index values based on the features.

In a decision tree plot, each node represents a split of data based on the condition of one of the variables. Each node contains information such as the node number, the number of observations included in the node (n), the deviation value, and the predicted value (yval). When interpreting this plot, one can start from the root node, which has 4019 observations. The root node has the first separation based on the UVB variable with a separation threshold of <0.365. The number of observations that meet this condition is 1463, while those that do not is 2556. The root node then branches into two child nodes (left and right) according to the splitting condition. This splitting and branching process continues at each node, where each split is based on the variable that provides the most significant increase in deviance. Deviance measures how much the predictions in the model differ from the actual values, and the goal is to minimize deviance as much as

p-ISSN: 2621-3761
e-ISSN: 2621-2889

possible during the splitting process. At each node, there is also an average value of UV Index and MSE (Mean Squared Error) which measures the model error. The smaller the MSE, the better the model predicts the target value.

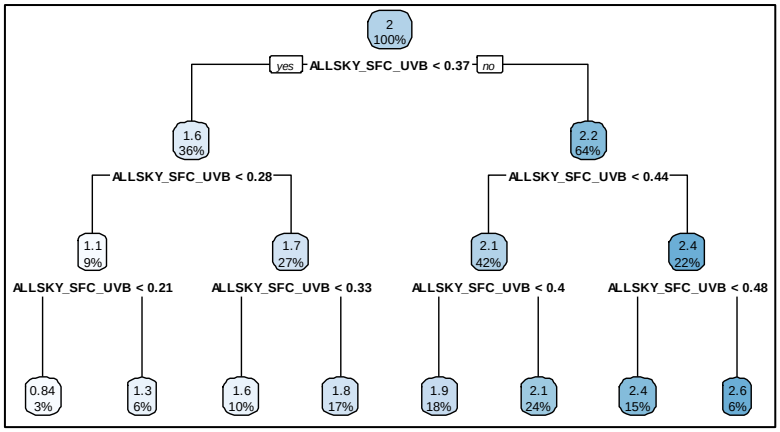

Figure 6. Decision tree analysis

Neural network analysis

The results show that the neural network model that has been built has a structure consisting of 5 input variables and 1 output variable. The model was trained using 2813 observation data as responses, with 14065 observation data as covariates. The model uses appropriate error functions and activation functions, and uses non-linear outputs. The data used consisted of 6 columns in data.frame format. The prediction results from the neural network model show that most of the predictions have a value of 1 [170]. Indicating the model tends to predict the same result for most of the test data. The results show that the predicted and actual values on the test data have different value ranges. The minimum predicted value is 1, while the minimum actual value is 0.390. The average predicted value tends to be close to 1 with an average of about 1.968, while the average actual value is about 1.968. The maximum predicted value and the actual value are also different, with a maximum predicted value of 1 and a maximum actual value of 3,210. The neural network model developed tends to produce the same predictions for most of the test data. However, there is a difference in the range of values between the predicted and actual values. This shows that the model still has the potential to be developed in order to provide predictions that are more accurate and in accordance with the actual values.

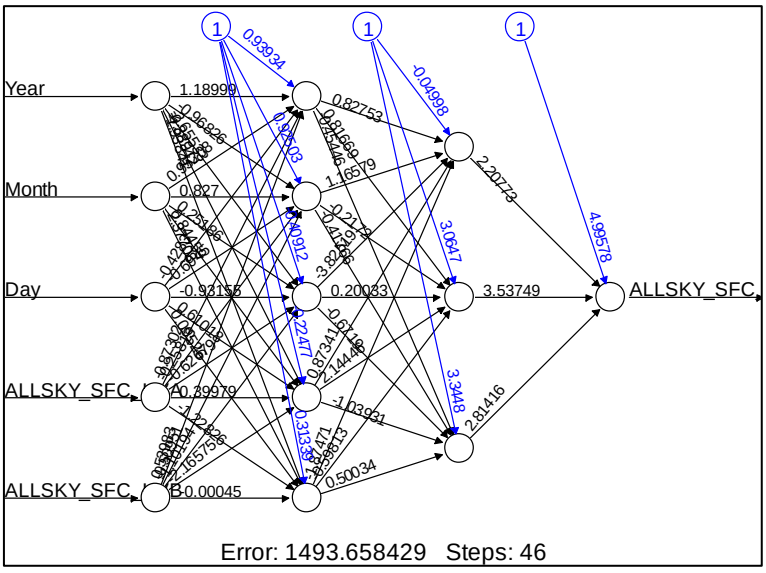

Figure 7. Neural network plot analysis results

p-ISSN: 2621-3761
e-ISSN: 2621-2889

The research produces a plot in the form of a neural network diagram, which illustrates the structure and relationship between the layers in the neural network model built [152]. Each sphere in the plot represents a neuron or unit in the neural network model. The lines connecting these spheres represent the relationships between neurons in different layers. Labeled lines indicate weights or parameters that connect neurons between layers. The number and size of hidden layers in the model is determined by the "hidden" argument during the construction of the network model [148]. There are two hidden layers with 5 and 3 neurons respectively. A sigmoid activation function is used for the neurons in the hidden layer and the output layer [171]. This can be observed from setting linear.output = FALSE during model construction. The plot represents the structure and architecture of the constructed artificial neural network model. The plot provides a visual representation of how information flows through the network and how each neuron is interconnected. This helps in understanding and interpreting the results of the artificial neural network model.

Support vector machine analysis

The research uses the Support Vector Machine (SVM) model with parameters determined including SVM-Type: eps-regression, SVM-Kernel: radial, cost: 1, gamma: 0.2, and epsilon: 0.1. The number of Support Vector used in this model is 508. The prediction results of the SVM model on the test data show a minimum value of 0.5836, a first quartile value of 1.7478, a median value of 1.9943, an average value of 1.9695, a third quartile value of 2.2345, and a maximum value of 2.9193. The SVM model with predetermined parameters can predict the UV index value with an average of 1.9695 on the test data.

SVM model is a method used for classification and regression [172]. SVM works by constructing a hyperplane or dividing surface that separates two different classes of data with a maximum margin [73], [74]. In the case of regression, SVM is used to predict continuous values based on the given data. In this study, an SVM model with epsilon regression (eps-regression) type is used. The SVM kernel used is a radial kernel, which maps the data to a higher dimensional space to facilitate class separation [173], [174]. The cost parameter controls the trade-off between error margin and data labeling error, while the gamma parameter controls how much influence each data sample has in the formation of the hyperplane [175], [176]. The epsilon parameter is used to determine the level of error tolerance in prediction. In this study, the SVM model with predetermined parameters provided prediction results for the UV index value in the test data with an average of 1,969.

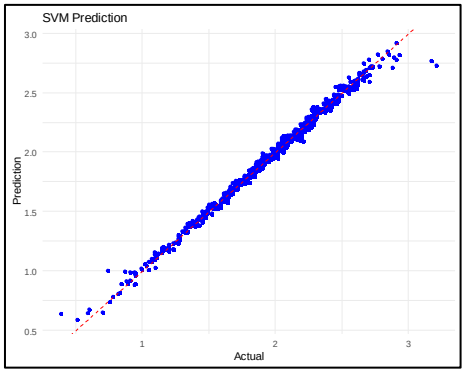

Figure 8. SVM Prediction

In the analysis, a graph has been used where the x-axis represents the actual values of the test data, and the y-axis represents the values predicted by the SVM model for the same data. Each point on the graph signifies one test data, with its position reflecting the actual and predicted values, respectively. The points on this graph are colored blue to highlight the relationship between the actual and predicted values. The graph includes two reference lines, a red diagonal line and a dashed line. The red diagonal line shows the ideal scenario where the dots should be if the actual and predicted values are identical. The dotted line serves as an additional point of comparison. If the points are close to the diagonal line, it indicates that the SVM prediction model performed well in predicting the actual values. Conversely, if the points are widely scattered, this indicates that there is a significant difference between the actual value and the value predicted by the SVM model.

K-Means scattering analysis

The study used the k-means clustering method to analyze the UV, UVA, and UVB Index data, which resulted in the formation of six distinct groups. The analysis revealed a varied distribution in the data set across the groups. Specifically, Group 1 consists of 437 data points, Group 2 includes 180 data points, Group 3 contains 784 data points, Group 4 consists of 1043 data points, Group 5 includes 972 data points, and Group 6 has 603 data points.

p-ISSN: 2621-3761
e-ISSN: 2621-2889

The k-means clustering analysis focuses on several variables, namely clusters, centers, totss (total sum of squares), withinss (sum of squares within clusters), tot.withinss (total sum of squares within clusters), betweens (sum of squares between clusters), size (size of each group), iter (number of iterations), and ifault (exit condition from the algorithm). The k-means clustering method groups data based on similarity or distance [177], [178]. The process starts with the random selection of an initial center for each group. Each data point is then assigned a group label based on its proximity to the nearest center. The group centers are recalculated by averaging the data points within each group. This iterative process continues until the group centers become stable or show minimal change. The k-means clustering method effectively grouped the UV, UVA, and UVB Index data into six groups with different distributions. The volume of data within each group provides an understanding of the relative size of the groups. The variables observed in the analysis provide information on the number of squares within groups, the total number of squares within groups, the number of squares between groups, the size of each group, and other metrics that facilitate a comprehensive understanding of the clustering results.

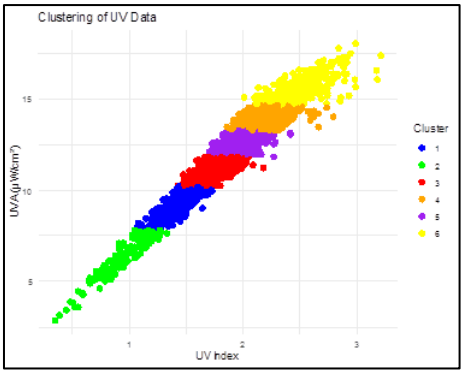

Figure 9. Clustering of UV data

The graph visualizes the results of the k-means clustering analysis on the UV Index, UVA, and UVB data. The x-axis represents the UV Index, while the y-axis shows the UVA value (μW/cm²). Each point on the graph corresponds to one observation, with its position determined by the respective UV Index and UVA values. The dots are colored according to the groups identified by the k-means analysis, as indicated by the color legend to the right of the graph. For example, a blue dot indicates that it belongs to the blue cluster. The graph shows the grouping of the data into six different clusters, allowing the observation of patterns and relationships between the UV Index and UVA values within each cluster.

Discussion

After conducting various statistical analyses and modeling using various methods, this study revealed some significant findings regarding the relationship between the UV Index and measured ultraviolet (UV) radiation levels.

Analysis using ANOVA showed a significant relationship between the UV Index and Ultraviolet A (UVA) radiation measurements. The ANOVA results showed variations in the UV Index significantly explained variations in UVA levels, with high F values and very low p values (<2e-16). This indicates that changes in the UV Index substantially affect the level of UVA radiation detected at the surface. The analysis showed that the ANOVA model had good accuracy in predicting UVA values based on the UV Index, as indicated by the low residual standard error values.

ANOVA analysis of Ultraviolet B (UVB) radiation measurements yielded similar findings. The analysis showed that significant differences existed between the groups formed based on the UV Index variable on UVB levels. The high F value and very low p value confirmed that the UV Index significantly influenced the variation in UVB levels. Results support the use of the UV Index as a robust indicator for predicting UVB radiation levels at the surface.

Classification analysis using Naive Bayes shows variation in predicted values based on the trained model [45], [120]. The model uses machine learning algorithms to identify patterns in the training data and applies those patterns to predict target values in the test data [179], [180]. The visualization and summary statistics of the predictions give an idea of the accuracy of the model in predicting the UV Index values.

Analysis using decision trees and neural networks demonstrated effective approaches in classifying and predicting UV Index values based on relevant features in the data. The decision tree model provides a deep understanding of how the measured features affect the UV Index values, while the neural network demonstrates a complex model structure with the ability to predict the UV Index values well, albeit with variation in the observed predictions [181].

Analysis using Support Vector Machine (SVM) and k-means clustering highlighted the methods' ability to model and cluster data based on the UV Index and associated UV measurements. SVM was successful in

p-ISSN: 2621-3761
e-ISSN: 2621-2889

predicting UV Index values with a high degree of accuracy, while k-means clustering effectively clustered the data into different groups based on their similarity in the context of UV Index and UV measurements.

V. Conclusion

Research theoretically confirmed that the UV Index is an effective indicator in predicting and understanding ultraviolet (UV) radiation levels at the surface. The statistical analyses performed, including ANOVA, Naive Bayes, decision trees, neural networks, SVM, and k-means clustering, consistently showed significant relationships between the UV Index and UV A and UV B measurements. These results provide strong support for the use of the UV Index as an important tool in monitoring and predicting UV exposure. Practically, the research has important implications in the context of public policy and public health. The use of the UV Index can help in informing the public about the risk of high UV exposure, which can contribute to health prevention efforts such as the use of sun protection and planning outdoor activities. The findings also provide a solid basis for the development of early warning systems and more effective UV protection strategies.

For future research, it is recommended to continue the study by considering some additional aspects. The development of UV Index prediction models is a crucial step to improve accuracy and precision in predicting UV radiation levels. Integration of more features and advanced modeling techniques can help in generating models that are more reliable and suited to variations in environmental conditions. It is important to investigate the temporal variations of the UV Index and UV radiation in a broader context, including understanding seasonal changes and long-term trends. Studies will provide a new, deeper understanding of how weather and climate dynamics contribute to UV exposure levels in different regions. Further analysis of environmental effects such as clouds and pollution should also be considered. These factors can significantly affect the distribution and intensity of UV radiation at the surface, so understanding their interactions will aid in the development of more effective mitigation and early warning strategies. Research on the long-term impacts of UV exposure on public health is an important area for further exploration. Studies could include evaluating the risk of skin cancer and other impacts related to UV exposure, providing an important basis for the formulation of sustainable public health policies.

VI. Acknowledgement

This research project was a collaborative effort, and we are grateful for the invaluable contributions and support from various individuals and institutions.

First and foremost, we would like to express our deepest gratitude to our respective institutions for providing the necessary resources and support during this research. We thank the Environmental Science Doctoral Program at Riau University, the Electrical Engineering Department and Agribusiness Department at Riau University, the Falak Science Observatory at Muhammadiyah University of North Sumatra, and the Physics Department at Medan State University for their continuous encouragement and academic support.

Our sincere appreciation goes to the Institute for Infrastructure Engineering and Sustainable Management (IIESM) at Universiti Teknologi MARA for their collaborative contributions that significantly improved the quality of our research.

We are also grateful to the co-authors, for their dedication, expertise and invaluable contributions to this research. Their diverse backgrounds and expertise have enriched this research, resulting in a more comprehensive analysis and understanding of the UV Index and its implications.

We would also like to thank all the technical and administrative staff at our respective institutions for their help and support in facilitating various aspects of this research.

We thank our family and friends for their unwavering support and understanding throughout this journey. Their encouragement and patience have been a source of motivation and strength.

References

[1] G. H. Bernhard *et al.*, "Environmental effects of stratospheric ozone depletion, UV radiation and interactions with climate change: UNEP Environmental Effects Assessment Panel, update 2019," *Photochem. Photobiol. Sci.*, vol. 19, no. 5, pp. 542–584, May 2020, doi: 10.1039/d0pp90011g.

[2] S. A. Umar and S. A. Tasduq, "Ozone Layer Depletion and Emerging Public Health Concerns - An Update on Epidemiological Perspective of the Ambivalent Effects of Ultraviolet Radiation Exposure," *Front. Oncol.*, vol. 12, p. 866733, Mar. 2022, doi: 10.3389/fonc.2022.866733.

[3] P. W. Barnes *et al.*, "Environmental effects of stratospheric ozone depletion, UV radiation, and interactions with climate change: UNEP Environmental Effects Assessment Panel, Update 2021," *Photochem. Photobiol. Sci.*, vol. 21, no. 3, pp. 275–301, Mar. 2022, doi: 10.1007/s43630-022-00176-5.

[4] R. E. Neale *et al.*, "The effects of exposure to solar radiation on human health," *Photochem. Photobiol. Sci.*, vol. 22, no. 5, pp. 1011–1047, Mar. 2023, doi: 10.1007/s43630-023-00375-8.

[5] J. Robinson, R. Begum, and M. Maqbool, *An Introduction to Non-Ionizing Radiation*. BENTHAM SCIENCE PUBLISHERS, 2023. doi: 10.2174/9789815136890123010l.

[6] G. P. Pfeifer, "Mechanisms of UV-induced mutations and skin cancer," *Genome Instab. Dis.*, vol. 1, no. 3, pp. 99–113, May 2020, doi: 10.1007/s42764-020-00009-8.

[7] L. Vanhaelewyn, D. Van Der Straeten, B. De Coninck, and F. Vandenbussche, "Ultraviolet Radiation From a Plant

p-ISSN: 2621-3761
e-ISSN: 2621-2889

Perspective: The Plant-Microorganism Context," *Front. Plant Sci.*, vol. 11, p. 597642, Dec. 2020, doi: 10.3389/fpls.2020.597642.
[8] A. Mavrič Čermelj, A. Golob, K. Vogel-Mikuš, and M. Germ, "Silicon Mitigates Negative Impacts of Drought and UV-B Radiation in Plants," *Plants*, vol. 11, no. 1, p. 91, Dec. 2021, doi: 10.3390/plants11010091.
[9] D. Loconsole and P. Santamaria, "UV Lighting in Horticulture: A Sustainable Tool for Improving Production Quality and Food Safety," *Horticulturae*, vol. 7, no. 1, p. 9, Jan. 2021, doi: 10.3390/horticulturae7010009.
[10] E. R. Parker, "The influence of climate change on skin cancer incidence – A review of the evidence," *Int. J. Women's Dermatology*, vol. 7, no. 1, pp. 17–27, Jan. 2021, doi: 10.1016/j.ijwd.2020.07.003.
[11] I. Degtiar and S. Rose, "A Review of Generalizability and Transportability," *Annu. Rev. Stat. Its Appl.*, vol. 10, no. 1, pp. 501–524, Mar. 2023, doi: 10.1146/annurev-statistics-042522-103837.
[12] H. M. Levitt, "Qualitative generalization, not to the population but to the phenomenon: Reconceptualizing variation in qualitative research.," *Qual. Psychol.*, vol. 8, no. 1, pp. 95–110, Feb. 2021, doi: 10.1037/qup0000184.
[13] Y. Song, J. Wang, Y. Ge, and C. Xu, "An optimal parameters-based geographical detector model enhances geographic characteristics of explanatory variables for spatial heterogeneity analysis: cases with different types of spatial data," *GIScience Remote Sens.*, vol. 57, no. 5, pp. 593–610, Jul. 2020, doi: 10.1080/15481603.2020.1760434.
[14] H. Sulistiani and A. A. Aldino, "Decision Tree C4. 5 Algorithm For Tuition Aid Grant Program Classification (Case Study: Department Of Information System, Universitas Teknokrat Indonesia)," *Edutic - Sci. J. Informatics Educ.*, vol. 7, no. 1, pp. 40–50, Nov. 2020, doi: 10.21107/edutic.v7i1.8849.
[15] M.-J. Jun, "A comparison of a gradient boosting decision tree, random forests, and artificial neural networks to model urban land use changes: the case of the Seoul metropolitan area," *Int. J. Geogr. Inf. Sci.*, vol. 35, no. 11, pp. 2149–2167, Nov. 2021, doi: 10.1080/13658816.2021.1887490.
[16] E. K. Sahin, "Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest," *SN Appl. Sci.*, vol. 2, no. 7, p. 1308, Jul. 2020, doi: 10.1007/s42452-020-3060-1.
[17] H. Liu, Z. Zhang, Z. Tian, and C. Lu, "Exploration for UV Aging Characteristics of Asphalt Binders based on Response Surface Methodology: Insights from the UV Aging Influencing Factors and Their Interactions," *Constr. Build. Mater.*, vol. 347, p. 128460, Sep. 2022, doi: 10.1016/j.conbuildmat.2022.128460.
[18] C.-Y. Wang *et al.*, "Geographical traceability of Eucommia ulmoides leaves using attenuated total reflection Fourier transform infrared and ultraviolet-visible spectroscopy combined with chemometrics and data fusion," *Ind. Crops Prod.*, vol. 160, p. 113090, Feb. 2021, doi: 10.1016/j.indcrop.2020.113090.
[19] N. Yamano *et al.*, "Long-term Effects of 222-nm ultraviolet radiation C Sterilizing Lamps on Mice Susceptible to Ultraviolet Radiation," *Photochem. Photobiol.*, vol. 96, no. 4, pp. 853–862, Jul. 2020, doi: 10.1111/php.13269.
[20] M. M. Delorme *et al.*, "Ultraviolet radiation: An interesting technology to preserve quality and safety of milk and dairy foods," *Trends Food Sci. Technol.*, vol. 102, pp. 146–154, Aug. 2020, doi: 10.1016/j.tifs.2020.06.001.
[21] G. Salvadori, F. Leccese, D. Lista, C. Burattini, and F. Bisegna, "Use of smartphone apps to monitor human exposure to solar radiation: Comparison between predicted and measured UV index values," *Environ. Res.*, vol. 183, p. 109274, Apr. 2020, doi: 10.1016/j.envres.2020.109274.
[22] E. O. Elemo *et al.*, "Ultraviolet Radiation Index over Abuja, Nigeria," *OALib*, vol. 08, no. 09, pp. 1–17, 2021, doi: 10.4236/oalib.1107924.
[23] A. M. Chacko, F. Lagacé, and F. Jafarian, "Ultraviolet index and sun safety: are we all on the same page?," *Br. J. Dermatol.*, vol. 184, no. 6, pp. 1175–1176, Jun. 2021, doi: 10.1111/bjd.19620.
[24] R. Vitt *et al.*, "UV-Index Climatology for Europe Based on Satellite Data," *Atmosphere (Basel).*, vol. 11, no. 7, p. 727, Jul. 2020, doi: 10.3390/atmos11070727.
[25] S. S. Prasad, R. C. Deo, N. Downs, D. Igoe, A. V. Parisi, and J. Soar, "Cloud Affected Solar UV Prediction With Three-Phase Wavelet Hybrid Convolutional Long Short-Term Memory Network Multi-Step Forecast System," *IEEE Access*, vol. 10, pp. 24704–24720, 2022, doi: 10.1109/ACCESS.2022.3153475.
[26] D. P. Igoe, A. V. Parisi, and N. J. Downs, "Cloud segmentation property extraction from total sky image repositories using Python," *Instrum. Sci. Technol.*, vol. 47, no. 5, pp. 522–534, Sep. 2019, doi: 10.1080/10739149.2019.1603996.
[27] R. Bajgar, A. Moukova, N. Chalupnikova, and H. Kolarova, "Differences in the Effects of Broad-Band UVA and Narrow-Band UVB on Epidermal Keratinocytes," *Int. J. Environ. Res. Public Health*, vol. 18, no. 23, p. 12480, Nov. 2021, doi: 10.3390/ijerph182312480.
[28] V. A. Bahamondes Lorca, M. K. McCulloch, Ó. Ávalos-Ovando, A. O. Govorov, F. Rahman, and S. Wu, "Characterization of UVB and UVA-340 Lamps and Determination of Their Effects on ER Stress and DNA Damage," *Photochem. Photobiol.*, vol. 98, no. 5, pp. 1140–1148, Sep. 2022, doi: 10.1111/php.13585.
[29] K. Hedayat, S. Ahmad Nasrollahi, A. Firooz, H. Rastegar, and M. Dadgarnejad, "Comparison of UVA Protection Factor Measurement Protocols," *Clin. Cosmet. Investig. Dermatol.*, vol. Volume 13, pp. 351–358, May 2020, doi: 10.2147/CCID.S244898.
[30] F. Bernerd, T. Passeron, I. Castiel, and C. Marionnet, "The Damaging Effects of Long UVA (UVA1) Rays: A Major Challenge to Preserve Skin Health and Integrity," *Int. J. Mol. Sci.*, vol. 23, no. 15, p. 8243, Jul. 2022, doi: 10.3390/ijms23158243.
[31] K. J. Gromkowska-Kępka, A. Puścion-Jakubik, R. Markiewicz-Żukowska, and K. Socha, "The impact of ultraviolet radiation on skin photoaging — review of in vitro studies," *J. Cosmet. Dermatol.*, vol. 20, no. 11, pp. 3427–3431, Nov. 2021, doi: 10.1111/jocd.14033.
[32] T. M. Ansary, M. R. Hossain, K. Kamiya, M. Komine, and M. Ohtsuki, "Inflammatory Molecules Associated with Ultraviolet Radiation-Mediated Skin Aging," *Int. J. Mol. Sci.*, vol. 22, no. 8, p. 3974, Apr. 2021, doi: 10.3390/ijms22083974.
[33] A. W. Schmalwieser, "Possibilities to estimate the personal UV radiation exposure from ambient UV radiation measurements," *Photochem. Photobiol. Sci.*, vol. 19, no.

10, pp. 1249–1261, Oct. 2020, doi: 10.1039/d0pp00182a.
[34] M. M. Mendes, K. H. Hart, S. A. Lanham-New, and P. B. Botelho, "Exploring the Impact of Individual UVB Radiation Levels on Serum 25-Hydroxyvitamin D in Women Living in High Versus Low Latitudes: A Cross-Sectional Analysis from the D-SOL Study," *Nutrients*, vol. 12, no. 12, p. 3805, Dec. 2020, doi: 10.3390/nu12123805.
[35] R. Jahdi and M. Arabi, "Monitoring summer solar ultraviolet (UV) radiation on the ground level over Ardabil-Sarein, NW Iran," *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.*, vol. X-4/W1-202, pp. 327–333, Jan. 2023, doi: 10.5194/isprs-annals-X-4-W1-2022-327-2023.
[36] J. Turner, D. Igoe, A. V. Parisi, A. J. McGonigle, A. Amar, and L. Wainwright, "A review on the ability of smartphones to detect ultraviolet (UV) radiation and their potential to be used in UV research and for public education purposes," *Sci. Total Environ.*, vol. 706, p. 135873, Mar. 2020, doi: 10.1016/j.scitotenv.2019.135873.
[37] J. Bilbao and A. de Migue, "Erythemal Solar Irradiance, UVER, and UV Index from Ground-Based Data in Central Spain," *Appl. Sci.*, vol. 10, no. 18, p. 6589, Sep. 2020, doi: 10.3390/app10186589.
[38] J. P. Kinney, C. S. Long, and A. C. Geller, "The Ultraviolet Index: A Useful Tool.," *Dermatol. Online J.*, vol. 6, no. 1, Sep. 2000, doi: 10.5070/D35925W4HQ.
[39] N. Hatsusaka *et al.*, "UV Index Does Not Predict Ocular Ultraviolet Exposure," *Transl. Vis. Sci. Technol.*, vol. 10, no. 7, p. 1, Jun. 2021, doi: 10.1167/tvst.10.7.1.
[40] D. R. Roshan, M. Koc, A. Abdallah, L. Martin-Pomares, R. Isaifan, and C. Fountoukis, "UV Index Forecasting under the Influence of Desert Dust: Evaluation against Surface and Satellite-Retrieved Data," *Atmosphere (Basel).*, vol. 11, no. 1, p. 96, Jan. 2020, doi: 10.3390/atmos11010096.
[41] L. Marabini *et al.*, "Effects of Vitis vinifera L. leaves extract on UV radiation damage in human keratinocytes (HaCaT)," *J. Photochem. Photobiol. B Biol.*, vol. 204, p. 111810, Mar. 2020, doi: 10.1016/j.jphotobiol.2020.111810.
[42] E. Kanasuo, H. Siiskonen, S. Haimakainen, J. Komulainen, and I. T. Harvima, "Regular use of vitamin D supplement is associated with fewer melanoma cases compared to non-use: a cross-sectional study in 498 adult subjects at risk of skin cancers," *Melanoma Res.*, vol. 33, no. 2, pp. 126–135, Apr. 2023, doi: 10.1097/CMR.0000000000000870.
[43] E. M. K. Reddy, A. Gurrala, V. B. Hasitha, and K. V. R. Kumar, "Introduction to Naive Bayes and a Review on Its Subtypes with Applications," in *Bayesian Reasoning and Gaussian Processes for Machine Learning Applications*, Boca Raton: Chapman and Hall/CRC, 2022, pp. 1–14. doi: 10.1201/9781003164265-1.
[44] L. M. Sinaga, Sawaluddin, and S. Suwilo, "Analysis of classification and Naïve Bayes algorithm k-nearest neighbor in data mining," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 725, no. 1, p. 012106, Jan. 2020, doi: 10.1088/1757-899X/725/1/012106.
[45] S. S. Bafjaish, "Comparative analysis of Naive Bayesian techniques in health-related for classification task," *J. Soft Comput. Data Min.*, vol. 1, no. 2, pp. 1–10, 2020, [Online]. Available: https://publisher.uthm.edu.my/ojs/index.php/jscdm/article/view/7144
[46] V. Jackins, S. Vimal, M. Kaliappan, and M. Y. Lee, "AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes," *J. Supercomput.*, vol. 77, no. 5, pp. 5198–5219, May 2021, doi: 10.1007/s11227-020-03481-x.
[47] J. Galopo Perez and E. S. Perez, "Predicting Student Program Completion Using Naïve Bayes Classification Algorithm," *Int. J. Mod. Educ. Comput. Sci.*, vol. 13, no. 3, pp. 57–67, Jun. 2021, doi: 10.5815/ijmecs.2021.03.05.
[48] M. M. Taye, "Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions," *Computers*, vol. 12, no. 5, p. 91, Apr. 2023, doi: 10.3390/computers12050091.
[49] R. Pugliese, S. Regondi, and R. Marini, "Machine learning-based approach: global trends, research directions, and regulatory standpoints," *Data Sci. Manag.*, vol. 4, pp. 19–29, Dec. 2021, doi: 10.1016/j.dsm.2021.12.002.
[50] A. Ali, W. Samara, D. Alhaddad, A. Ware, and O. A. Saraereh, "Human Activity and Motion Pattern Recognition within Indoor Environment Using Convolutional Neural Networks Clustering and Naive Bayes Classification Algorithms," *Sensors*, vol. 22, no. 3, p. 1016, Jan. 2022, doi: 10.3390/s22031016.
[51] D. Iskandaryan, F. Ramos, and S. Trilles, "Air Quality Prediction in Smart Cities Using Machine Learning Technologies Based on Sensor Data: A Review," *Appl. Sci.*, vol. 10, no. 7, p. 2401, Apr. 2020, doi: 10.3390/app10072401.
[52] Y. Sun and Y. Ma, "Application of Classification Algorithm Based on Naive Bayes in Data Analysis of Fitness Test," *J. Phys. Conf. Ser.*, vol. 1648, no. 4, p. 042078, Oct. 2020, doi: 10.1088/1742-6596/1648/4/042078.
[53] X. Liu, D. Lu, A. Zhang, Q. Liu, and G. Jiang, "Data-Driven Machine Learning in Environmental Pollution: Gains and Problems," *Environ. Sci. Technol.*, vol. 56, no. 4, pp. 2124–2133, Feb. 2022, doi: 10.1021/acs.est.1c06157.
[54] L. Cui and Z. Liu, "Synergy between research on ensemble perception, data visualization, and statistics education: A tutorial review," *Attention, Perception, Psychophys.*, vol. 83, no. 3, pp. 1290–1311, Apr. 2021, doi: 10.3758/s13414-020-02212-x.
[55] H. Xu, A. Berres, Y. Liu, M. R. Allen-Dumas, and J. Sanyal, "An overview of visualization and visual analytics applications in water resources management," *Environ. Model. Softw.*, vol. 153, p. 105396, Jul. 2022, doi: 10.1016/j.envsoft.2022.105396.
[56] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 01, pp. 20–28, Mar. 2021, doi: 10.38094/jastt20165.
[57] H. Lu and X. Ma, "Hybrid decision tree-based machine learning models for short-term water quality prediction," *Chemosphere*, vol. 249, p. 126169, Jun. 2020, doi: 10.1016/j.chemosphere.2020.126169.
[58] T. Thomas, A. P. Vijayaraghavan, and S. Emmanuel, *Machine Learning Approaches in Cyber Security Analytics*. Singapore: Springer Singapore, 2020. doi: 10.1007/978-981-15-1706-8.
[59] S. Mishra, P. K. Mallick, H. K. Tripathy, A. K. Bhoi, and A. González-Briones, "Performance Evaluation of a Proposed Machine Learning Model for Chronic Disease Datasets Using an Integrated Attribute Evaluator and an Improved Decision Tree Classifier," *Appl. Sci.*, vol. 10, no. 22, p. 8137, Nov. 2020, doi: 10.3390/app10228137.
[60] A. N. Elmachtoub, J. C. N. Liang, and R. McNellis,

"Decision trees for decision-making under the predict-then-optimize framework," in *International conference on machine learning, PMLR*, 2020, pp. 2858–2867. [Online]. Available: https://proceedings.mlr.press/v119/elmachtoub20a.html

[61] A. D. Woods *et al.*, *Missing Data and Multiple Imputation Decision Tree*. 2021. doi: 10.31234/osf.io/mdw5r.

[62] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, "A survey on missing data in machine learning," *J. Big Data*, vol. 8, no. 1, p. 140, Oct. 2021, doi: 10.1186/s40537-021-00516-9.

[63] O. Sagi and L. Rokach, "Explainable decision forest: Transforming a decision forest into an interpretable tree," *Inf. Fusion*, vol. 61, pp. 124–138, Sep. 2020, doi: 10.1016/j.inffus.2020.03.013.

[64] G. R. Yang and X.-J. Wang, "Artificial Neural Networks for Neuroscientists: A Primer," *Neuron*, vol. 107, no. 6, pp. 1048–1070, Sep. 2020, doi: 10.1016/j.neuron.2020.09.005.

[65] A. Mehonic, A. Sebastian, B. Rajendran, O. Simeone, E. Vasilaki, and A. J. Kenyon, "Memristors—From In-Memory Computing, Deep Learning Acceleration, and Spiking Neural Networks to the Future of Neuromorphic and Bio-Inspired Computing," *Adv. Intell. Syst.*, vol. 2, no. 11, p. 2000085, Nov. 2020, doi: 10.1002/aisy.202000085.

[66] A. Thakur, "Fundamentals of Neural Networks," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 9, no. VIII, pp. 407–426, Aug. 2021, doi: 10.22214/ijraset.2021.37362.

[67] M. Rithani, R. P. Kumar, and S. Doss, "A review on big data based on deep neural network approaches," *Artif. Intell. Rev.*, vol. 56, no. 12, pp. 14765–14801, Dec. 2023, doi: 10.1007/s10462-023-10512-5.

[68] T. P. Lillicrap, A. Santoro, L. Marris, C. J. Akerman, and G. Hinton, "Backpropagation and the brain," *Nat. Rev. Neurosci.*, vol. 21, no. 6, pp. 335–346, Jun. 2020, doi: 10.1038/s41583-020-0277-3.

[69] R. Dastres and M. Soori, "Artificial neural network systems," *Int. J. Imaging Robot.*, vol. 21, no. 2, pp. 13–25, 2021, [Online]. Available: https://hal.science/hal-03349542

[70] O. A. Montesinos López, A. Montesinos López, and J. Crossa, "Fundamentals of Artificial Neural Networks and Deep Learning," in *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, Cham: Springer International Publishing, 2022, pp. 379–425. doi: 10.1007/978-3-030-89010-0_10.

[71] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Muller, "Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications," *Proc. IEEE*, vol. 109, no. 3, pp. 247–278, Mar. 2021, doi: 10.1109/JPROC.2021.3060483.

[72] D. A. Pisner and D. M. Schnyer, "Support vector machine," in *Machine Learning*, Elsevier, 2020, pp. 101–121. doi: 10.1016/B978-0-12-815739-8.00006-7.

[73] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, pp. 189–215, Sep. 2020, doi: 10.1016/j.neucom.2019.10.118.

[74] A. Rizwan, N. Iqbal, R. Ahmad, and D.-H. Kim, "WR-SVM Model Based on the Margin Radius Approach for Solving the Minimum Enclosing Ball Problem in Support Vector Machine Classification," *Appl. Sci.*, vol. 11, no. 10, p. 4657, May 2021, doi: 10.3390/app11104657.

[75] D. Hsu, V. Muthukumar, and J. Xu, "On the proliferation of support vectors in high dimensions," in *International Conference on Artificial Intelligence and Statistics, PMLR*, 2021, pp. 91–99. [Online]. Available: https://proceedings.mlr.press/v130/hsu21a.html

[76] N. Ardeshir, C. Sanford, and D. J. Hsu, "Support vector machines and linear regression coincide with very high-dimensional features," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 4907–4918, 2021, [Online]. Available: https://proceedings.neurips.cc/paper/2021/file/26d4b4313a7e5828856bc0791fca39a2-Paper.pdf

[77] P. El Kafrawy, H. Fathi, M. Qaraad, A. K. Kelany, and X. Chen, "An Efficient SVM-Based Feature Selection Model for Cancer Classification Using High-Dimensional Microarray Data," *IEEE Access*, vol. 9, pp. 155353–155369, 2021, doi: 10.1109/ACCESS.2021.3123090.

[78] E. Y. Boateng, J. Otoo, and D. A. Abaye, "Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review," *J. Data Anal. Inf. Process.*, vol. 08, no. 04, pp. 341–357, 2020, doi: 10.4236/jdaip.2020.84020.

[79] M. Hou and C. Kambhampati, "Locally fitting hyperplanes to high-dimensional data," *Neural Comput. Appl.*, vol. 34, no. 11, pp. 8885–8896, Jun. 2022, doi: 10.1007/s00521-022-06909-y.

[80] K. Wang and C. Thrampoulidis, "Binary Classification of Gaussian Mixtures: Abundance of Support Vectors, Benign Overfitting, and Regularization," *SIAM J. Math. Data Sci.*, vol. 4, no. 1, pp. 260–284, Mar. 2022, doi: 10.1137/21M1415121.

[81] V. Vapnik and R. Izmailov, "Reinforced SVM method and memorization mechanisms," *Pattern Recognit.*, vol. 119, p. 108018, Nov. 2021, doi: 10.1016/j.patcog.2021.108018.

[82] F. Nie, W. Zhu, and X. Li, "Decision Tree SVM: An extension of linear SVM for non-linear classification," *Neurocomputing*, vol. 401, pp. 153–159, Aug. 2020, doi: 10.1016/j.neucom.2019.10.051.

[83] S. Shojae Chaeikar, A. A. Manaf, A. A. Alarood, and M. Zamani, "PFW: Polygonal Fuzzy Weighted—An SVM Kernel for the Classification of Overlapping Data Groups," *Electronics*, vol. 9, no. 4, p. 615, Apr. 2020, doi: 10.3390/electronics9040615.

[84] F. Farshadmoghadam, H. D. Azodi, and M. R. Yaghouti, "An Efficient Alternative Kernel of Gaussian Radial Basis Function for Solving Nonlinear Integro-Differential Equations," *Iran. J. Sci. Technol. Trans. A Sci.*, vol. 46, no. 3, pp. 869–881, Jun. 2022, doi: 10.1007/s40995-022-01286-6.

[85] K. Thurnhofer-Hemsi, E. López-Rubio, M. A. Molina-Cabello, and K. Najarian, "Radial basis function kernel optimization for Support Vector Machine classifiers," Jul. 2020, [Online]. Available: http://arxiv.org/abs/2007.08233

[86] I. S. Al-Mejibli, J. K. Alwan, and D. H. Abd, "The effect of gamma value on support vector machine performance with different kernels," *Int. J. Electr. Comput. Eng.*, vol. 10, no. 5, p. 5497, Oct. 2020, doi: 10.11591/ijece.v10i5.pp5497-5506.

[87] Q. Gu, Y. Chang, X. Li, Z. Chang, and Z. Feng, "A novel F-SVM based on FOA for improving SVM performance," *Expert Syst. Appl.*, vol. 165, p. 113713, Mar. 2021, doi: 10.1016/j.eswa.2020.113713.

[88] R.-C. Chen, C. Dewi, S.-W. Huang, and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods," *J. Big Data*, vol. 7, no. 1, p. 52, Dec. 2020, doi: 10.1186/s40537-020-00327-4.

[89] N. Rtayli and N. Enneya, "Enhanced credit card fraud detection based on SVM-recursive feature elimination and hyper-parameters optimization," *J. Inf. Secur. Appl.*, vol. 55, p. 102596, Dec. 2020, doi: 10.1016/j.jisa.2020.102596.
[90] T. Wang, L. Zhang, and W. Hu, "Bridging deep and multiple kernel learning: A review," *Inf. Fusion*, vol. 67, pp. 3–13, Mar. 2021, doi: 10.1016/j.inffus.2020.10.002.
[91] M. A. Deif, A. A. A. Solyman, M. H. Alsharif, S. Jung, and E. Hwang, "A Hybrid Multi-Objective Optimizer-Based SVM Model for Enhancing Numerical Weather Prediction: A Study for the Seoul Metropolitan Area," *Sustainability*, vol. 14, no. 1, p. 296, Dec. 2021, doi: 10.3390/su14010296.
[92] B. Azari, K. Hassan, J. Pierce, and S. Ebrahimi, "Evaluation of Machine Learning Methods Application in Temperature Prediction," *Comput. Res. Prog. Appl. Sci. Eng.*, vol. 8, no. 1, pp. 1–12, 2022, doi: 10.52547/crpase.8.1.2747.
[93] N. Sultana, "Predicting sun protection measures against skin diseases using machine learning approaches," *J. Cosmet. Dermatol.*, vol. 21, no. 2, pp. 758–769, Feb. 2022, doi: 10.1111/jocd.14120.
[94] M. Jafari Gukeh, S. Moitra, A. N. Ibrahim, S. Derrible, and C. M. Megaridis, "Machine Learning Prediction of TiO 2 - Coating Wettability Tuned via UV Exposure," *ACS Appl. Mater. Interfaces*, vol. 13, no. 38, pp. 46171–46179, Sep. 2021, doi: 10.1021/acsami.1c13262.
[95] K. P. Sinaga and M.-S. Yang, "Unsupervised K-Means Clustering Algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.
[96] K. P. Sinaga, I. Hussain, and M.-S. Yang, "Entropy K-Means Clustering With Feature Reduction Under Unknown Number of Clusters," *IEEE Access*, vol. 9, pp. 67736–67751, 2021, doi: 10.1109/ACCESS.2021.3077622.
[97] E. U. Oti, M. O. Olusola, F. C. Eze, and S. U. Enogwe, "Comprehensive Review of K-Means Clustering Algorithms," *Int. J. Adv. Sci. Res. Eng.*, vol. 07, no. 08, pp. 64–69, 2021, doi: 10.31695/IJASRE.2021.34050.
[98] R. Qaddoura, H. Faris, and I. Aljarah, "An efficient clustering algorithm based on the k-nearest neighbors with an indexing ratio," *Int. J. Mach. Learn. Cybern.*, vol. 11, no. 3, pp. 675–714, Mar. 2020, doi: 10.1007/s13042-019-01027-z.
[99] S. Daoudi, C. M. A. Zouaoui, M. C. El-Mezouar, and N. Taleb, "Parallelization of the K-Means++ Clustering Algorithm," *Ingénierie des systèmes d Inf.*, vol. 26, no. 1, pp. 59–66, Feb. 2021, doi: 10.18280/isi.260106.
[100] V. Romanuke, "Random centroid initialization for improving centroid-based clustering," *Decis. Mak. Appl. Manag. Eng.*, vol. 6, no. 2, pp. 734–746, Dec. 2023, doi: 10.31181/dmame622023742.
[101] A. K. Abdalameer, M. Alswaitti, A. A. Alsudani, and N. A. M. Isa, "A new validity clustering index-based on finding new centroid positions using the mean of clustered data to determine the optimum number of clusters," *Expert Syst. Appl.*, vol. 191, p. 116329, Apr. 2022, doi: 10.1016/j.eswa.2021.116329.
[102] Y. Li, X. Zhou, J. Gu, K. Guo, and W. Deng, "A Novel K-Means Clustering Method for Locating Urban Hotspots Based on Hybrid Heuristic Initialization," *Appl. Sci.*, vol. 12, no. 16, p. 8047, Aug. 2022, doi: 10.3390/app12168047.
[103] A. F. R. Guarda, N. M. M. Rodrigues, and F. Pereira, "Constant Size Point Cloud Clustering: A Compact, Non-Overlapping Solution," *IEEE Trans. Multimed.*, vol. 23, pp. 77–91, 2021, doi: 10.1109/TMM.2020.2974325.
[104] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Inf. Sci. (Ny).*, vol. 622, pp. 178–210, Apr. 2023, doi: 10.1016/j.ins.2022.11.139.
[105] P. Govender and V. Sivakumar, "Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019)," *Atmos. Pollut. Res.*, vol. 11, no. 1, pp. 40–56, Jan. 2020, doi: 10.1016/j.apr.2019.09.009.
[106] W. Zeng, Y. Jiang, Z. Huo, and K. Hu, "Clustering Analysis of Extreme Temperature Based on K-means Algorithm," 2020, pp. 523–533. doi: 10.1007/978-3-030-57881-7_46.
[107] P. Stoker, G. Tian, and J. Y. Kim, "Analysis of Variance (ANOVA)," in *Basic Quantitative Research Methods for Urban Planners*, Routledge, 2020, pp. 197–219. doi: 10.4324/9780429325021-11.
[108] K. G. Burkart *et al.*, "Estimating the cause-specific relative risks of non-optimal temperature on daily mortality: a two-part modelling approach applied to the Global Burden of Disease Study," *Lancet*, vol. 398, no. 10301, pp. 685–697, Aug. 2021, doi: 10.1016/S0140-6736(21)01700-1.
[109] G. Arnqvist, "Mixed Models Offer No Freedom from Degrees of Freedom," *Trends Ecol. Evol.*, vol. 35, no. 4, pp. 329–335, Apr. 2020, doi: 10.1016/j.tree.2019.12.004.
[110] D. Shi, C. DiStefano, A. Maydeu-Olivares, and T. Lee, "Evaluating SEM Model Fit with Small Degrees of Freedom," *Multivariate Behav. Res.*, vol. 57, no. 2–3, pp. 179–207, May 2022, doi: 10.1080/00273171.2020.1868965.
[111] G. Pavlov, A. Maydeu-Olivares, and D. Shi, "Using the Standardized Root Mean Squared Residual (SRMR) to Assess Exact Fit in Structural Equation Models," *Educ. Psychol. Meas.*, vol. 81, no. 1, pp. 110–130, Feb. 2021, doi: 10.1177/0013164420926231.
[112] M. A. Mansournia, M. Nazemipour, A. I. Naimi, G. S. Collins, and M. J. Campbell, "Reflection on modern methods: demystifying robust standard errors for epidemiologists," *Int. J. Epidemiol.*, vol. 50, no. 1, pp. 346–351, Mar. 2021, doi: 10.1093/ije/dyaa260.
[113] D. S. K. Karunasingha, "Root mean square error or mean absolute error? Use their ratio as well," *Inf. Sci. (Ny).*, vol. 585, pp. 609–629, Mar. 2022, doi: 10.1016/j.ins.2021.11.036.
[114] T. Kunz and T. Laepple, "Frequency-Dependent Estimation of Effective Spatial Degrees of Freedom," *J. Clim.*, vol. 34, no. 18, pp. 7373–7388, Sep. 2021, doi: 10.1175/JCLI-D-20-0228.1.
[115] L. Tognetti *et al.*, "UVA-1 phototherapy as adjuvant treatment for eosinophilic fasciitis: in vitro and in vivo functional characterization," *Int. J. Dermatol.*, vol. 61, no. 6, pp. 718–726, Jun. 2022, doi: 10.1111/ijd.16003.
[116] B. O. Saguie, R. L. Martins, A. de S. da Fonseca, B. Romana-Souza, and A. Monte-Alto-Costa, "An ex vivo model of human skin photoaging induced by UVA radiation compatible with summer exposure in Brazil," *J. Photochem. Photobiol. B Biol.*, vol. 221, p. 112255, Aug. 2021, doi: 10.1016/j.jphotobiol.2021.112255.
[117] N. Ibrahim and A. B. Abdullahi, "Analysis of Variance (ANOVA) Randomized Block Design (RBD) to Test the Variability of Three Different Types of Fertilizers (NPK, UREA and SSP) on Millet Production," *African J. Agric.*

Sci. Food Res., vol. 9, no. 1, pp. 1–10, 2023, [Online]. Available: https://publications.afropolitanjournals.com/index.php/ajasfr/article/view/326
[118] A. A. T., "Analysis of Variance: The Fundamental Concepts and Application with R," *Int. J. Math. Comput. Res.*, vol. 09, no. 10, pp. 2408–2422, Oct. 2021, doi: 10.47191/ijmcr/v9i10.04.
[119] S. Farhana, "Classification of Academic Performance for University Research Evaluation by Implementing Modified Naive Bayes Algorithm," *Procedia Comput. Sci.*, vol. 194, pp. 224–228, 2021, doi: 10.1016/j.procs.2021.10.077.
[120] K. Maswadi, N. A. Ghani, S. Hamid, and M. B. Rasheed, "Human activity classification using Decision Tree and Naïve Bayes classifiers," *Multimed. Tools Appl.*, vol. 80, no. 14, pp. 21709–21726, Jun. 2021, doi: 10.1007/s11042-020-10447-x.
[121] Y. Ying, T. N. Mursitama, Shidarta, and Lohansen, "Effectiveness of the News Text Classification Test Using the Naïve Bayes' Classification Text Mining Method," *J. Phys. Conf. Ser.*, vol. 1764, no. 1, p. 012105, Feb. 2021, doi: 10.1088/1742-6596/1764/1/012105.
[122] N. Wijaya, "Evaluation of Naïve Bayes and Chi-Square performance for Classification of Occupancy House," *Int. J. Informatics Comput.*, vol. 1, no. 2, p. 46, Feb. 2020, doi: 10.35842/ijicom.v1i2.20.
[123] S. Jayachitra and A. Prasanth, "Multi-Feature Analysis for Automated Brain Stroke Classification Using Weighted Gaussian Naïve Bayes Classifier," *J. Circuits, Syst. Comput.*, vol. 30, no. 10, p. 2150178, Aug. 2021, doi: 10.1142/S0218126621501784.
[124] T. N. Viet, H. Le Minh, L. C. Hieu, and T. H. Anh, "The Naïve Bayes algorithm for learning data analytics," *Indian J. Comput. Sci. Eng.*, vol. 12, no. 4, pp. 1038–1043, Aug. 2021, doi: 10.21817/indjcse/2021/v12i4/211204191.
[125] S. Tangirala, "Evaluating the Impact of GINI Index and Information Gain on Classification using Decision Tree Classifier Algorithm*," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 2, pp. 612–619, 2020, doi: 10.14569/IJACSA.2020.0110277.
[126] M. Bansal, A. Goyal, and A. Choudhary, "A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning," *Decis. Anal. J.*, vol. 3, p. 100071, Jun. 2022, doi: 10.1016/j.dajour.2022.100071.
[127] T. O. Hodson, T. M. Over, and S. S. Foks, "Mean Squared Error, Deconstructed," *J. Adv. Model. Earth Syst.*, vol. 13, no. 12, p. e2021MS002681, Dec. 2021, doi: 10.1029/2021MS002681.
[128] X. Zhu and S. Tang, "A Branch-and-Bound Algorithm for Building Optimal Data Gathering Tree in Wireless Sensor Networks," *INFORMS J. Comput.*, vol. 33, no. 4, pp. 1446–1460, Feb. 2021, doi: 10.1287/ijoc.2020.1012.
[129] G. Zarpellon, J. Jo, A. Lodi, and Y. Bengio, "Parameterizing Branch-and-Bound Search Trees to Learn Branching Policies," *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 5, pp. 3931–3939, May 2021, doi: 10.1609/aaai.v35i5.16512.
[130] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, "Tree-Based Methods," 2023, pp. 331–366. doi: 10.1007/978-3-031-38747-0_8.
[131] D. Kumar and N. A. Priyanka, "Decision tree classifier: a detailed survey," *Int. J. Inf. Decis. Sci.*, vol. 12, no. 3, p. 246, 2020, doi: 10.1504/IJIDS.2020.10029122.
[132] M. M. Ghiasi, S. Zendehboudi, and A. A. Mohsenipour, "Decision tree-based diagnosis of coronary artery disease: CART model," *Comput. Methods Programs Biomed.*, vol. 192, p. 105400, Aug. 2020, doi: 10.1016/j.cmpb.2020.105400.
[133] A. Singh, V. Kotiyal, S. Sharma, J. Nagar, and C.-C. Lee, "A Machine Learning Approach to Predict the Average Localization Error With Applications to Wireless Sensor Networks," *IEEE Access*, vol. 8, pp. 208253–208263, 2020, doi: 10.1109/ACCESS.2020.3038645.
[134] I. Ahmad, M. U. Akhtar, S. Noor, and A. Shahnaz, "Missing Link Prediction using Common Neighbor and Centrality based Parameterized Algorithm," *Sci. Rep.*, vol. 10, no. 1, p. 364, Jan. 2020, doi: 10.1038/s41598-019-57304-y.
[135] B. Albanna, R. Heeks, A. Pawelke, J. Boy, J. Handl, and A. Gluecker, "Data-powered positive deviance: Combining traditional and non-traditional data to identify and characterise development-related outperformers," *Dev. Eng.*, vol. 7, p. 100090, 2022, doi: 10.1016/j.deveng.2021.100090.
[136] A. Shehadeh, O. Alshboul, R. E. Al Mamlook, and O. Hamedat, "Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression," *Autom. Constr.*, vol. 129, p. 103827, Sep. 2021, doi: 10.1016/j.autcon.2021.103827.
[137] Y. Wang *et al.*, "Visualizing Element Migration over Bifunctional Metal-Zeolite Catalysts and its Impact on Catalysis," *Angew. Chemie Int. Ed.*, vol. 60, no. 32, pp. 17735–17743, Aug. 2021, doi: 10.1002/anie.202107264.
[138] Y. Chen, L. Song, Y. Liu, L. Yang, and D. Li, "A Review of the Artificial Neural Network Models for Water Quality Prediction," *Appl. Sci.*, vol. 10, no. 17, p. 5776, Aug. 2020, doi: 10.3390/app10175776.
[139] Q. Liu, M. F. Iqbal, J. Yang, X. Lu, P. Zhang, and M. Rauf, "Prediction of chloride diffusivity in concrete using artificial neural network: Modelling and performance evaluation," *Constr. Build. Mater.*, vol. 268, p. 121082, Jan. 2021, doi: 10.1016/j.conbuildmat.2020.121082.
[140] J. Jawad, A. H. Hawari, and S. Javaid Zaidi, "Artificial neural network modeling of wastewater treatment and desalination using membrane processes: A review," *Chem. Eng. J.*, vol. 419, p. 129540, Sep. 2021, doi: 10.1016/j.cej.2021.129540.
[141] D.-H. Lee, Y.-T. Kim, and S.-R. Lee, "Shallow Landslide Susceptibility Models Based on Artificial Neural Networks Considering the Factor Selection Method and Various Non-Linear Activation Functions," *Remote Sens.*, vol. 12, no. 7, p. 1194, Apr. 2020, doi: 10.3390/rs12071194.
[142] A. A. Alkhouly, A. Mohammed, and H. A. Hefny, "Improving the Performance of Deep Neural Networks Using Two Proposed Activation Functions," *IEEE Access*, vol. 9, pp. 82249–82271, 2021, doi: 10.1109/ACCESS.2021.3085855.
[143] Y. Wang, Y. Li, Y. Song, and X. Rong, "The Influence of the Activation Function in a Convolution Neural Network Model of Facial Expression Recognition," *Appl. Sci.*, vol. 10, no. 5, p. 1897, Mar. 2020, doi: 10.3390/app10051897.
[144] S. Namasudra, S. Dhamodharavadhani, and R. Rathipriya, "Nonlinear Neural Network Based Forecasting Model for Predicting COVID-19 Cases," *Neural Process. Lett.*, vol. 55, no. 1, pp. 171–191, Feb. 2023, doi: 10.1007/s11063-

021-10495-w.
[145] R. Pal, A. A. Sekh, S. Kar, and D. K. Prasad, "Neural Network Based Country Wise Risk Prediction of COVID-19," *Appl. Sci.*, vol. 10, no. 18, p. 6448, Sep. 2020, doi: 10.3390/app10186448.
[146] J. Elliott *et al.*, "Predictive Accuracy of a Polygenic Risk Score–Enhanced Prediction Model vs a Clinical Risk Score for Coronary Artery Disease," *JAMA*, vol. 323, no. 7, p. 636, Feb. 2020, doi: 10.1001/jama.2019.22241.
[147] R. Valavi, G. Guillera-Arroita, J. J. Lahoz-Monfort, and J. Elith, "Predictive performance of presence-only species distribution models: a benchmark study with reproducible code," *Ecol. Monogr.*, vol. 92, no. 1, p. e01486, Feb. 2022, doi: 10.1002/ecm.1486.
[148] Z. Shen, H. Yang, and S. Zhang, "Neural network approximation: Three hidden layers are enough," *Neural Networks*, vol. 141, pp. 160–173, Sep. 2021, doi: 10.1016/j.neunet.2021.04.011.
[149] T. Szandała, "Review and Comparison of Commonly Used Activation Functions for Deep Neural Networks," 2021, pp. 203–224. doi: 10.1007/978-981-15-5495-7_11.
[150] S. H. Bhojani and N. Bhatt, "Wheat crop yield prediction using new activation functions in neural network," *Neural Comput. Appl.*, vol. 32, no. 17, pp. 13941–13951, Sep. 2020, doi: 10.1007/s00521-020-04797-8.
[151] J. You, K. H. J. Leskovec, and S. Xie, "Graph structure of neural networks," in *International Conference on Machine Learning*, *PMLR*, 2020, pp. 10881–10891. doi: 10.48550/arXiv.2007.06559.
[152] M. M. Taye, "Theoretical Understanding of Convolutional Neural Network: Concepts, Architectures, Applications, Future Directions," *Computation*, vol. 11, no. 3, p. 52, Mar. 2023, doi: 10.3390/computation11030052.
[153] J. Senk *et al.*, "Connectivity concepts in neuronal network modeling," *PLOS Comput. Biol.*, vol. 18, no. 9, p. e1010086, Sep. 2022, doi: 10.1371/journal.pcbi.1010086.
[154] K. Lenk, B. Genocchi, M. T. Barros, and J. A. K. Hyttinen, "Larger Connection Radius Increases Hub Astrocyte Number in a 3-D Neuron–Astrocyte Network Model," *IEEE Trans. Mol. Biol. Multi-Scale Commun.*, vol. 7, no. 2, pp. 83–88, Jun. 2021, doi: 10.1109/TMBMC.2021.3054890.
[155] L. Luo, "Architectures of neuronal circuits," *Science (80-.).*, vol. 373, no. 6559, p. eabg7285, Sep. 2021, doi: 10.1126/science.abg7285.
[156] S. Chung and L. F. Abbott, "Neural population geometry: An approach for understanding biological and artificial neural networks," *Curr. Opin. Neurobiol.*, vol. 70, pp. 137–144, Oct. 2021, doi: 10.1016/j.conb.2021.10.010.
[157] L. Das, A. Sivaram, and V. Venkatasubramanian, "Hidden representations in deep neural networks: Part 2. Regression problems," *Comput. Chem. Eng.*, vol. 139, p. 106895, Aug. 2020, doi: 10.1016/j.compchemeng.2020.106895.
[158] M. Uzair and N. Jamil, "Effects of Hidden Layers on the Efficiency of Neural networks," in *2020 IEEE 23rd International Multitopic Conference (INMIC)*, Nov. 2020, pp. 1–6. doi: 10.1109/INMIC50486.2020.9318195.
[159] H. M. D. Kabir *et al.*, "SpinalNet: Deep Neural Network With Gradual Input," *IEEE Trans. Artif. Intell.*, vol. 4, no. 5, pp. 1165–1177, Oct. 2023, doi: 10.1109/TAI.2022.3185179.
[160] M. G. M. Abdolrasol *et al.*, "Artificial Neural Networks Based Optimization Techniques: A Review," *Electronics*, vol. 10, no. 21, p. 2689, Nov. 2021, doi: 10.3390/electronics10212689.
[161] A. A. Movassagh *et al.*, "Artificial neural networks training algorithm integrating invasive weed optimization with differential evolutionary model," *J. Ambient Intell. Humaniz. Comput.*, vol. 14, no. 5, pp. 6017–6025, May 2023, doi: 10.1007/s12652-020-02623-6.
[162] S. Kiliçarslan and M. Celik, "RSigELU: A nonlinear activation function for deep neural networks," *Expert Syst. Appl.*, vol. 174, p. 114805, Jul. 2021, doi: 10.1016/j.eswa.2021.114805.
[163] N. Kulathunga, N. R. Ranasinghe, D. Vrinceanu, Z. Kinsman, L. Huang, and Y. Wang, "Effects of the Nonlinearity in Activation Functions on the Performance of Deep Learning Models," Oct. 2020, doi: 10.48550/arXiv.2010.07359.
[164] F. Emmert-Streib, Z. Yang, H. Feng, S. Tripathi, and M. Dehmer, "An Introductory Review of Deep Learning for Prediction Models With Big Data," *Front. Artif. Intell.*, vol. 3, p. 4, Feb. 2020, doi: 10.3389/frai.2020.00004.
[165] M. N. Fekri, H. Patel, K. Grolinger, and V. Sharma, "Deep learning for load forecasting with smart meter data: Online Adaptive Recurrent Neural Network," *Appl. Energy*, vol. 282, p. 116177, Jan. 2021, doi: 10.1016/j.apenergy.2020.116177.
[166] J. L. Zambrano, J. A. L. Torralbo, and C. R. Morales, "Early prediction of student learning performance through data mining: A systematic review," *Psicothema*, 2021, doi: 10.7334/psicothema2021.62.
[167] H. Meyer and E. Pebesma, "Predicting into unknown space? Estimating the area of applicability of spatial prediction models," *Methods Ecol. Evol.*, vol. 12, no. 9, pp. 1620–1633, Sep. 2021, doi: 10.1111/2041-210X.13650.
[168] J. Marulanda-Durango, A. Escobar-Mejía, A. Alzate-Gómez, and M. Álvarez-López, "A Support Vector machine-Based method for parameter estimation of an electric arc furnace model," *Electr. Power Syst. Res.*, vol. 196, p. 107228, Jul. 2021, doi: 10.1016/j.epsr.2021.107228.
[169] G. Kavithaa and S. Bhuvaneswari, "Kernel Linkage Support Vector Regression For Stock Market Index Prediction And Analysis," *Turkish J. Comput. Math. Educ.*, vol. 12, no. 12, pp. 3289–3300, 2021, doi: 10.17762/turcomat.v12i12.8008.
[170] N. Karunanithi, D. Whitley, and Y. K. Malaiya, "Using neural networks in reliability prediction," *IEEE Softw.*, vol. 9, no. 4, pp. 53–59, Jul. 1992, doi: 10.1109/52.143107.
[171] Y. Koçak and G. Üstündağ Şiray, "New activation functions for single layer feedforward neural network," *Expert Syst. Appl.*, vol. 164, p. 113977, Feb. 2021, doi: 10.1016/j.eswa.2020.113977.
[172] M. A. Chandra and S. S. Bedi, "Survey on SVM and their application in image classification," *Int. J. Inf. Technol.*, vol. 13, no. 5, pp. 1–11, Oct. 2021, doi: 10.1007/s41870-017-0080-1.
[173] J. Shao, X. Liu, and W. He, "Kernel Based Data-Adaptive Support Vector Machines for Multi-Class Classification," *Mathematics*, vol. 9, no. 9, p. 936, Apr. 2021, doi: 10.3390/math9090936.
[174] L. Mohan, J. Pant, P. Suyal, and A. Kumar, "Support Vector Machine Accuracy Improvement with Classification," in *2020 12th International Conference on Computational Intelligence and Communication Networks (CICN)*, Sep. 2020, pp. 477–481. doi: 10.1109/CICN49253.2020.9242572.

[175] R. Guido, M. C. Groccia, and D. Conforti, "A hyper-parameter tuning approach for cost-sensitive support vector machine classifiers," *Soft Comput.*, vol. 27, no. 18, pp. 12863–12881, Sep. 2023, doi: 10.1007/s00500-022-06768-8.
[176] N. Tran, J.-G. Schneider, I. Weber, and A. K. Qin, "Hyper-parameter optimization in classification: To-do or not-to-do," *Pattern Recognit.*, vol. 103, p. 107245, Jul. 2020, doi: 10.1016/j.patcog.2020.107245.
[177] T. M. Ghazal *et al.*, "Performances of K-Means Clustering Algorithm with Different Distance Metrics," *Intell. Autom. Soft Comput.*, vol. 29, no. 3, pp. 735–742, 2021, doi: 10.32604/iasc.2021.019067.
[178] D. M. SAPUTRA, D. SAPUTRA, and L. D. OSWARI, "Effect of Distance Metrics in Determining K-Value in K-Means Clustering Using Elbow and Silhouette Method," in *Proceedings of the Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019)*, 2020. doi: 10.2991/aisr.k.200424.051.
[179] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Comput. Sci.*, vol. 2, no. 3, p. 160, May 2021, doi: 10.1007/s42979-021-00592-x.
[180] A. Chattopadhyay, P. Hassanzadeh, and S. Pasha, "Predicting clustered weather patterns: A test case for applications of convolutional neural networks to spatio-temporal climate data," *Sci. Rep.*, vol. 10, no. 1, p. 1317, Jan. 2020, doi: 10.1038/s41598-020-57897-9.
[181] A. A. M. Ahmed, M. H. Ahmed, S. K. Saha, O. Ahmed, and A. Sutradhar, "Optimization algorithms as training approach with hybrid deep learning methods to develop an ultraviolet index forecasting model," *Stoch. Environ. Res. Risk Assess.*, vol. 36, no. 10, pp. 3011–3039, Oct. 2022, doi: 10.1007/s00477-022-02177-3.

Declarations

Author contribution : Edy Ervianto led the conceptual framework of the research, designed the methodology, curated the data, drafted the original manuscript, and supervised the research team. Noveri Lysbetti Marpaung managed the project, conducted formal data analysis, reviewed and edited the manuscript, and acted as correspondence author. Abu Yazid Raisal was responsible for data analysis, creating visualizations, and developing the software used in this study. Sakti Hutabarat contributed to the validation of the results, provided resources, and participated in reviewing and editing the manuscript. Rohana Hassan oversaw the progress of the study, and participated in reviewing and editing the manuscript. Ruben Cornelius Siagian was involved in the investigation, data curation, and creation of visual representations of the data. Nurhalim assisted in data collection and contributed to the development of the software used in the analysis. Rahyul Amri performed statistical analysis and participated in reviewing and editing the manuscript.

Funding statement : This research did not receive any funding.

Conflict of interest : Both authors declare that they have no competing interests.

Additional information : The study evaluated the correlation between the UV Index and UVA and UVB radiation levels to assess the effectiveness of the UV Index in predicting surface UV radiation. Using statistical methods such as ANOVA and various predictive modeling techniques (Naive Bayes, decision trees, artificial neural networks, SVM, and k-means clustering), the study found that the UV Index is a reliable predictor of UV radiation. SVM showed the highest prediction accuracy, while k-means was effective in clustering UV data. Results support the development of better early warning systems and UV protection strategies, with significant implications for public health and protection against harmful UV radiation.