Journal of Theoretical and Applied Information Technology
30th April 2022. Vol.100. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

OPTIMIZING REGRESSION ALGORITHM PERFORMANCE FOR WEAK RAINFALL DATASET PREDICTION VIA ENSEMBLE MACHINE LEARNING MODELS

PRABOWO WAHYU SUDARNO[1], AHMAD ASHARI[2], MARDHANI RIASETIAWAN[3]
[1, 2, 3]Department of Computer Science and Electronics, Universitas Gadjah Mada, Yogyakarta, Indonesia
Email: [1]prabowowahyu45@mail.ugm.ac.id, [2]ashari@ugm.ac.id, [3]mardhani@ugm.ac.id

ABSTRACT

A flood is a natural disaster that cannot be stopped, but preventive measures can be taken to deal with it. The factors that cause flooding can be predicted using machine learning, one of which is by predicting rainfall. But in reality, rainfall data has many shortcomings, such as missing values and the appearance of outliers that can affect model performance. Therefore, we propose an ensemble stacking method to deal with this problem. The performance value of the Multilayer Perceptron algorithm without Stacking is 10.128 for MSE and1.5696 for MAE. The performance value of the XGBoost algorithm without stacking is 9.2548 for MSE and 1.4427 for MAE. While the performance value of combining the Multilayer Perceptron and XGBoost algorithm with Stacking resulted in an MSE value of 9.2377 and an MAE value of 1.4396. The results show that the ensemble method with stacking can be a solution to improve algorithm performance on weak datasets to predict rainfall value. The novelty of this paper is as follows: machine learning ensembles can handle the weak rainfall dataset to give a better result.

Keywords: *Ensemble Machine Learning, Stacking, MLP, XGB, Rainfall*

1. INTRODUCTION

Floods are natural disasters that are often encountered in various countries. Research on floods continues, as natural disasters cannot be stopped, but preventive measures can be taken to deal with them. This is a challenge for researchers to continue to find the best model that can reduce flood risk. when humans can take preventive steps in dealing with floods, damage and losses will be minimized. Methods for flood congestion analyzing require a continuous enhancement, majorly in the present context of progressive changes in climate change that result in an incremented susceptibility to floods observed in diverse locations globally [1]. One of the modeling steps that can be used is in the field of machine learning is the regression techniques. In this study, we will predict the amount of rainfall, which is one of the factors causing flooding.

In recent years, machine learning algorithms provided an optimal approach to real issues. In the era of technological advances, machine learning can analyze the occurrence of floods in a specific area using a machine learning algorithm [2]. One of the methods is using Regression. Regression techniques in machine learning are frequently used for many purposes, including student performance prediction, classification of diseases, and much more. Specifically, regression is a processes that analyze a pattern to describe classes or future trends in a dataset [3].

Many regression algorithms are often used in research related to machine learning. One of the powerful machine learning regression methods is Extreme Gradient Boosting (XGBoost). This model had been used in several problems such as rain modelling [4]. In this case, XGBoost create multiple trees sequentially in a way that each one of the next trees tries to reduce the errors from the previous tree [4].

Another powerful method is Multilayer Perceptron (MLP). This model had been used in several problems such as flood prediction and other complex hydrogeological models due to these characteristics [5]. The characteristic of MLP is nonlinear activation and a high number of layers. MLP models were reported to be more efficient compared to several traditional machine learning methods [5].

However, most machine learning algorithms have weaknesses when it comes to handling weak

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

datasets [3]. In this study, weak dataset means a condition that causes the dataset to be incomplete or have many missing values. In addition, the dataset also has a non-standard range. Missing values are a common problem present in data from various sources. When building machine learning classifiers, incomplete data creates a risk of drawing invalid conclusions and producing biased models. This could lead to a tremendous impact on many business sectors or even human lives.

In this study, we try to handle weak dataset and try to make the dataset better using machine learning processing with an ensemble approach. Ensemble methods are meta-algorithms that combines weak base estimators into stronger classifiers. Ensemble method is chosen because it has been proven that it produces more accurate results than when a single model is used to solve the same problem. Most of the researchers used heterogeneous ensemble approach in their work because it provides better performance [6].

Based on the research that has used ensemble machine learning regarding the handling weak dataset, two approaches can be applied, at the algorithmic level and data level. Several drawbacks may occur in the data level approach, namely the risk of data duplication and missing out information in the dataset. This problem will also affect the performance of the regression algorithm [7]. To tackle these drawbacks, researchers often change or correct the skewed distribution of classes in the dataset using resampling and data synthesis techniques. Many studies explained handling weak datasets, in some of these studies using several approaches [8]. They prove that the application of resampling techniques or data level approaches to deal with weak datasets can improve the performance of algorithms.

At the algorithmic level approach, how the operation of the existing algorithm is adjusted to make the algorithm more propitious in analyzing minorities, or in other terms, modification or ensemble of several algorithms is carried out [9].

From previous research, it is known that most machine learning algorithms have weaknesses when it comes to handling weak datasets [3]. The statement described before is a research question in this study to be completed, because the rainfall dataset obtained in this study is a dataset that is classified as a weak dataset.

One method that can be used to handle weak datasets is using the ensemble machine learning. Ensemble machine learning can be used to make a better prediction. Ensemble machine learning executes the learning model by constructing and combining multiple learners. This approach gives better prediction results than using a single algorithm [10].

This research aims to prove that machine learning ensemble methods can improve weak datasets with better performance results for predicting rainfall, which can later be useful in flood mitigation.

The remaining sections of this research is arranged as follows: Section 2 describes about the dataset; Section 3 describes the methods used in this research; Section 4 describes and discuss about the findings; Section 5 shows the conclusion of this research.

2. DATASETS

In this section, we will discuss where the source of the dataset comes from, what features will be used and how the initial process of processing the data will be.

The dataset in this study is different from the previously mentioned studies. This research uses the Australian Weather dataset, which is taken from Kaggle. The dataset has 15 feature attributes and 145.460 instances. After checking, we decided that the dataset is weak, so it will be used as the object of this paper's research. The features of the dataset and its description are tabulated in Table 1.

2.1 Data Preprocessing

In this study, data preprocessing includes filling in null data and correcting outliers errors in the data. Some Null data must be filled so that the data can still be used. Null data will be filled with the predicted value from K-Nearest Neighbor (KNN) imputer method.

We also correct data that has outliers. Figure 1 shows the distribution of data from each column in the dataset. It can be seen from Figure 1 that there are several outliers in the dataset, for example in the Rainfall and Humidity9am variables. It is necessary to correct outliers to improve performance of the model.

Table 1: Dataset Dictionary

Data	Description
MinTemp	Minimum Temperature

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

	(Celsius)
MaxTemp	Maximum Temperature (Celsius)
Rainfall	Record data of Rainfall data in a day (mm)
Sunshine	How long bright sunshine in the day (Hours)
WindGustSpeed	The speed (km/h) of the strongest wind in a day
WindSpeed9am	9 AM Wind speed averaged over ten minutes (km/hr)
WindSpeed3pm	3 PM Wind speed averaged over ten minutes (km/hr)
Humidity9am	9 AM Humidity (percent)
Humidity3pm	3 PM Humidity (percent)
Cloud9am	Fraction of sky obscured by cloud at 9 am (Oktas)
Cloud3pm	Fraction of sky obscured by cloud at 3 pm (Oktas)
Pressure9am	Atmospheric pressure average sea level at 9 AM (hpa)
Pressure3pm	Atmospheric pressure average sea level at 3 PM (hpa)
Temp9am	The temperature at 9 AM ©
Temp3pm	The temperature at 3 PM ©

2.2 Train-Validation Split

The train-validation split technique is used so that the performance of the model can be evaluated unbiasedly. The purpose of this technique is that the model only learns from the train section and does not see or learn from validation data during the training process. Thus, the model's ability is not biased when tested with validation data. In our research, the training data and testing data are divided with 80%:20% ratio.

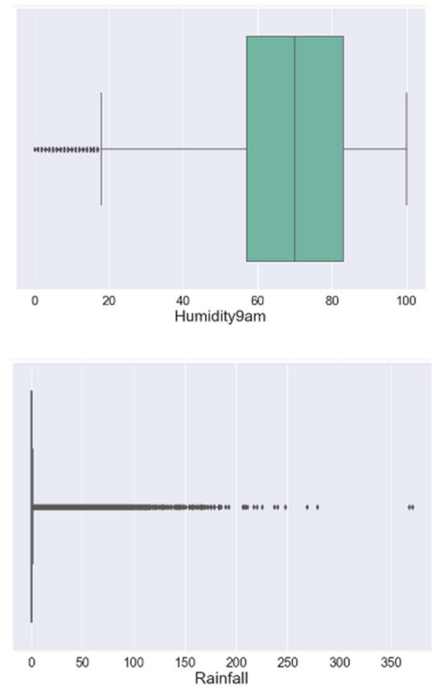

Figure 1: Outlier Data

3. METHOD

In this section will be describe the method used in this research, explain step of the methods, describe the strength or the weakness of method, explain how measurements were made and what calculations were performed.

3.1 Stacking Method

In ensemble machine learning, the multiple base learners are trained altogether and combine their predictions into a single result. It improves the robustness over a single model. Ensemble machine learning uses voting in classification cases or average in regression cases and it will give better result [11]. One of the ensemble machine learning approaches is Stacking.

Stacking is one of the powerful methods in ensemble machine learning that combines several machine learning algorithms through meta-learning for solving classification and regression problems. The purpose of stacking is to predict the dataset from the base model in the previous level as input variables and then combining the models on the next level. Stacking make predictions that have better performance than any single model in the ensemble [12].

Stacking is different from bagging and boosting. Stacking often takes into heterogeneous

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

weak model learners, whereas bagging and boosting consider homogenous weak model learners. Pseudocode of ensemble stacking can be written as follows [13].

Algorithm 1 Stacking

Input:
Dataset $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_M, y_M)\}$
Base learning algorithm L_t $(t = 1, 2, \ldots, T)$
Meta learning algorithm L

Process:
For $t = 1, 2, \ldots, T$:
  $h_t = L(D_t)$
End

$D' = \emptyset$
For $m = 1, 2, \ldots, M$:
  For $t = 1, 2, \ldots, T$:
    $z_{it} = h_t(x_i)$
  End
  $D' = D' \cup \{((z_{it}, z_{it}, \ldots, z_{it}), y_i)\}$
End

$h' = L(D')$

Output:
$H(x) = h'(h_1(x), h_2(x), \ldots, h_T(x))$

The advantage of stacking is the capabilities of combining multiple models that perform well on a regression task to better produce better results compared to the single model without an ensemble. In addition, stacking method also improves the prediction accuracy of the model. However, it has its own disadvantage.

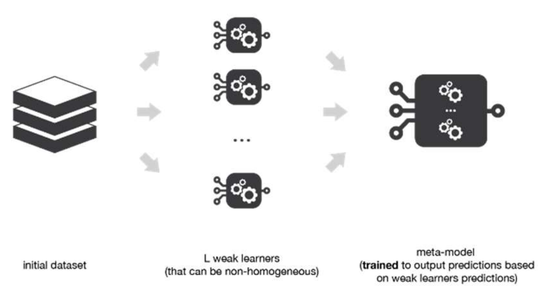

Figure 2: Stacking Ensemble

The method will take longer computation time, since we train the entire dataset with each classifier individually. Figure 2 visualise the stacking method [14].

3.1.1 Base Regressor and Meta Regressor

The ensemble stacking method uses multiple base regressors in the learning process. There are two stages in stacking learning. Stage 1, each base regressor is trained using the same dataset data to produce their respective prediction results. Step 2, the meta regressor retrieves prediction result from the base regressor as their input to determine which class the test data most likely is.

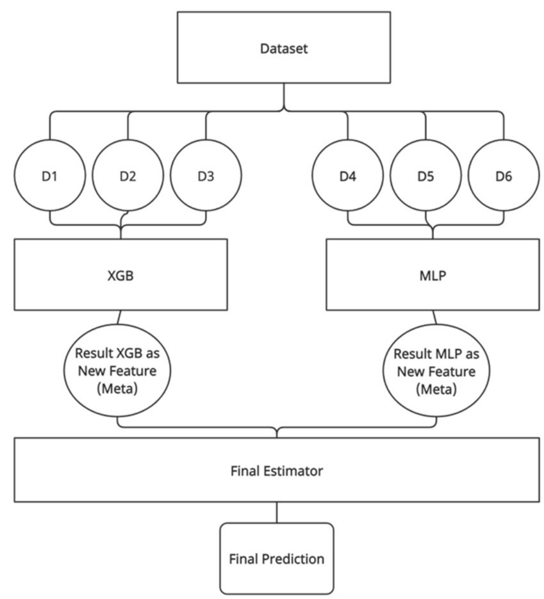

Figure 3: Stacking Research

The ensemble stacking method diagram is visualized in Figure 3. In our research, we use two different base algorithms as the base regressor, namely the Multilayer Perceptron and the Xtreme Gradient Boost algorithm.

3.2 Multilayer Perceptron

Multilayer Perceptron (MLP) is the most widely used Artificial Neural Network architecture, either for solving classification or regression problems. As the name implies, there are three main layers in the MLP, namely the input layer, the hidden layer, and the output layer. Relations between layers in the MLP architecture are as follows: weight U_{ij} for the connection between input layer $x_i$and hidden layer Z_j, weight V_{jk} for the connection between hidden layer $Z_j$andits next hidden layer, and the weight W_{kl} for the connection between the last hidden layer towards output layerY_k.The learning process in MLP aims to find the most optimum synaptic weights for classifying the set of training data and validation data. The process of updating the weights of MLP is done by using the Backpropagation technique. Figure 4 visualizes the Multilayer Perceptron architecture.

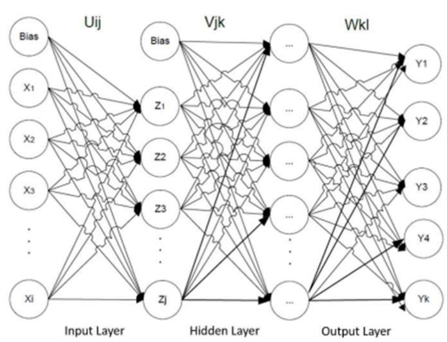

Figure 4: Multilayer Perceptron

We use MLP as our base regressor because it is one of the recommended techniques for predicting floods. This is because MLP improves the conjugate gradient algorithm, as [15] shows that MLP gives high prediction accuracy in identifying rainfall in the Kelantan River. We also tune the MLP hyperparameter, aiming to achieve optimal regression results. Later, the regression performance is evaluated using several well-known statistical measures, namely Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) [16].

3.3 Xtreme Gradient Boosting

Xtreme Gradient Boosting (XGBoost) optimizes poor models to make better prediction result. This algorithm usually uses the decision trees approach, which builds the model in stages and generalizes it by optimizing an arbitrary differentiable loss function [17].

XGBoost is a method that combines boosting with gradient boosting. This method is the first time Friedman introduced the correlation between boosting and optimization to make a Gradient Boosting Machine (GBM). Model built using the boosting method to predict the error of the previous model. This algorithm using gradient descent to reduce errors when building a new model. So, it is called gradient boosting. The ultimate goal of this process is to get the closest function $\hat{f}_x$ to functions constructor f_x by minimizing the value of the loss function $L(y, f_{(x)})$ defined by the equation [18]:

$$\hat{f} = argmin_f E_{x,y} [L(y, f_{(x)})] \quad (1)$$

In the training process, each iteration minimizes the value of the loss function based on the initial function $f_0(x)$. In general, gradient boosting algorithm has the following equation [18]:

$$\{\gamma_m h_m\} = argmin \sum_{m=1}^{M} (L(y_i, f^{(m-1)}(x_i) + \gamma_m h_m(x_i)) \quad (2)$$

Where M is the number of boosting stages, f is the imperfect model, h_m is an estimator, and γ is a pseudo-regularization parameter.

3.4 Performance validation

Performance Validation is the step where we evaluate whether the proposed model can perform regression well or not. This process utilizes the Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The formulas of MAE, MSE, and RMSE are shown by Equation 3, Equation 4, and Equation 5 respectively.

$$MSE(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} (y_i - \hat{y}_i)^2 \quad (3)$$

$$MAE(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} |y_i - \hat{y}_i| \quad (4)$$

$$RMSE(y, \hat{y}) = \sqrt{\frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} (y_i - \hat{y}_i)^2} \quad (5)$$

The equation 5 explained that y is the actual label, $\hat{y}$ is the predicted label, and n is the number of data within the dataset.

4. RESULTS AND DISCUSSION

According to [4], XGBoost had been used in several problems and had good results, for

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

example, in rain modelling. It is because the XGBoost creates multiple trees sequentially in a way that each one of the next trees tries to reduce the previous tree's errors. In other research, [5] explained that MLP had been used in several problem such as flood prediction and other complex hydrogeological models. MLP has good result because more efficient. In this research, we found that MLP and XGB methods are not good enough at handling incomplete or have many missing values datasets. Therefore, in this paper try to optimize the handling of weak data sets using ensemble machine learning stacking. We tune the MLP and XGBoost model hyperparameters individually, aiming to produce the best performance model.

Table 2: MLP Parameter Tuning

Parameter	**Value**	**Description**
Learning Rate Init	0.0223	The initial learning rate controls value in updating the weights.
Momentum	0.3999	Gradient descent update value. The value between 0 and 1
Max Iter	1832	How many times each data point will be used.
Validation Fraction	0.6064	The portion of training data as a parameter early stopping.

Table 2 shows the tuning value results for the MLP. The parameters are learning rate init, momentum, max iter and validation fraction. Table 3 shows the tuning value results for the XGboost. The parameters are eta, col sample by tree, max depth, min child weight and subsample.

Table 3: XGBoost Parameter Tuning

Parameter	**Value**	**Description**
Eta	0.0399	The value of step size prevents overfitting.
Col sample By Tree	0.7485	The subsampling ratio column once when constructing every tree.
Max Depth	3832	The value to make the model more complex.
Min Child Weight	13.82	A child must have a minimum sum of instance weight (hessian).
Subsample	0.7071	The training instance subsampling ratio.

Figure 5 shows the performance value of the Multilayer Perceptron. The MLP algorithm without Stacking produced MSE score of 10.128, MAE score of 1.5696, and RMSE score of 3.1824.

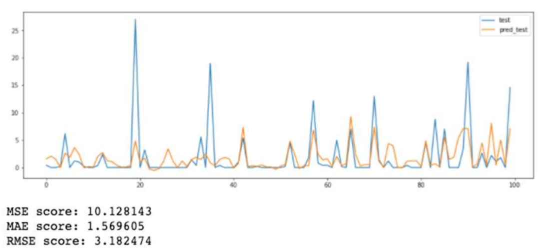

Figure 5: MLP Result

Figure 6 shows the performance value of the XGboost algorithm. The XGBoost algorithm without stacking produced a MSE, MAE, and RMSE score of 9.2548, 1.4427, and 3.0421 respectively.

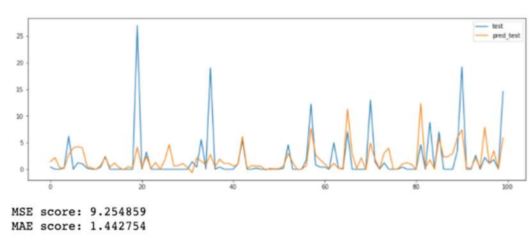

Figure 6. XGB Result

Tuned MLP and XGB models are then added to the stacking method as base learners. Combining the Multilayer Perceptron and

Journal of Theoretical and Applied Information Technology
30th April 2022. Vol.100. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

XGBoost algorithm with Stacking ensemble method results in a MSE, MAE, and RMSE score of 9.2377, 1.4396, and 3.0393 respectively. The stacking result is visualized in Figure 7.

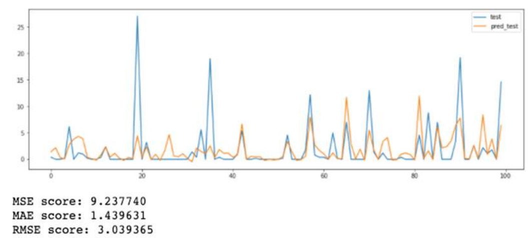

Figure 7: Stacking Result

Figure 5, Figure 6, and Figure 7 show that the MSE, MAE, and RMSE generated by the ensemble stacking method, produce better evaluation values than the method without stacking.

In this study, we compared our proposed method with existing research, namely research [4] using XGBoost and research [5] using MLP. This comparison will measure our proposed method to prove ML ensemble is better at handling weak datasets than single machine learning.

The result is shown in Table 4 explain the application of the ensemble method with the addition of stacking can optimize the performance of the regression algorithm on weak datasets. Although in studies [4] and [5] the use of XGBoost and MLP yielded good results, they produced poor results when tested in our dataset. The use of stacking methods with XGBoost and MLP produces better results, implying that the use of stacking when processing weak datasets is very helpful in improving the model's capabilities.

Table 4: Final Result Comparison

Algo	**MSE**	**MAE**	**RMSE**
MLP	10.128	1.5696	3.1824
XGB	9.2548	1.4427	3.0421
Stacking (MLP+XGB)	**9.2377**	**1.4396**	**3.0393**

4. CONCLUSION

Weak data sets are something that is often encountered in data mining, especially in regression task. Based on the research that has been done, this research successfully prove that the ensemble machine learning method helps improving weak datasets such as the Rainfall dataset, which can later be implemented for flood mitigation plan. We also proved that the ensemble stacking method provides a better result compared to methods without the addition of stacking.

The performance value of the Multilayer Perceptron algorithm without Stacking is 10.128 for MSE, 1.5696 for MAE and 3.1824 for RMSE. The performance value of the XGBoost algorithm without stacking MSE is 9.2548 for MSE, 1.4427 for MAE, and 3.0421 for RMSE. While the performance combining of Multilayer Perceptron and XGBoost algorithm with Stacking the resulting in MSE, MAE, and RMSE score of 9.2377, 1.4396, and 3.0393, respectively.

This paper provides the following novelty: the ensemble method with stacking can be a solution to improve algorithm performance on weak datasets to predict rainfall value. We recommend future research try applying other ensemble methods such as bagging and boosting.

5. CONFLICT OF INTEREST

The authors declare no conflict of interest.

6. AUTHOR CONTRIBUTIONS

The authors confirm contribution to the paper as follows: study conception, developed the theoretical formalism and performed the numerical simulations. Prabowo Wahyu Sudarno performed the analytic calculations in the code, writing, review and editing. Ahmad Ashari, and Mardhani Riasetiawan supervised the conceptualization, methodology, validation, and analysis to the final version of the manuscript. All authors reviewed the results and approved the final version of the manuscript.

7. ACKNOWLEDGEMENT

Researchers acknowledge the Directorate of Research and Community Service, Deputy for Strengthening Research and Development, Ministry of Research, Technology/National Research and Innovation Agency of the Republic of Indonesia in the PMDSU program.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

REFERENCES

[1] M. M. Jazzar, "Flood congestion simulation and prediction using IOT wireless networks on dynamic streets routes," *J. Theor. Appl. Inf. Technol.*, vol. 99, no. 10, 2021.
[2] A. Katti, K. V. Ashish, A. Loke, dan K. Bade, "A pluvial flood detection model using machine learning techniques and simulate the flow of water," 2020, doi: 10.1109/ICCES48766.2020.09137946.
[3] T. M. Christian dan M. Ayub, "Exploration of classification using NBTree for predicting students' performance," 2014, doi: 10.1109/ICODSE.2014.7062654.
[4] M. T. Anwar, E. Winarno, W. Hadikurniawati, dan M. Novita, "Rainfall prediction using Extreme Gradient Boosting," in *Journal of Physics: Conference Series*, 2021, vol. 1869, no. 1, doi: 10.1088/1742-6596/1869/1/012078.
[5] V. V. Ramalingam, R. Mishra, dan S. Parashari, "Prediction of flood by rainfall using MLP classifier of neural network model," *Int. J. Adv. Sci. Technol.*, vol. 29, no. 6, 2020.
[6] N. M. Baba, M. Makhtar, S. A. Fadzli, dan M. K. Awang, "Current issues in ensemble methods and its applications," *Journal of Theoretical and Applied Information Technology*, vol. 81, no. 2. 2015.
[7] K. Khadijah dan P. S. Sasongko, "The Comparison of Imbalanced Data Handling Method in Software Defect Prediction," *Kinet. Game Technol. Inf. Syst. Comput. Network, Comput. Electron. Control*, 2020, doi: 10.22219/kinetik.v5i3.1049.
[8] R. I. Rashu, N. Haq, dan R. M. Rahman, "Data mining approaches to predict final grade by overcoming class imbalance problem," 2003, doi: 10.1109/ICCITechn.2014.7073095.
[9] S. S. Dongre, "RARE CLASS PROBLEM IN DATA MINING: REVIEW," *Int. J. Adv. Res. Comput. Sci.*, 2017, doi: 10.26483/ijarcs.v8i7.4530.
[10] T. Zhang dan J. Li, "Credit Risk Control Algorithm Based on Stacking Ensemble Learning," 2021, doi: 10.1109/ICPECA51329.2021.9362514.
[11] A. Fouad, H. M. Moftah, dan H. A. Hefny, "Brain diagnoses detection using whale optimization algorithm based on ensemble learning classifier," *Int. J. Intell. Eng. Syst.*, vol. 13, no. 2, 2020, doi: 10.22266/ijies2020.0430.05.
[12] B. Pavlyshenko, "Using Stacking Approaches for Machine Learning Models," 2018, doi: 10.1109/DSMP.2018.8478522.
[13] O. O. Petinrin dan F. Saeed, "Stacked ensemble for bioactive molecule prediction," *IEEE Access*, 2019, doi: 10.1109/ACCESS.2019.2945422.
[14] Joseph Rocca, "Ensemble methods: bagging, boosting and stacking," *Towar. Data Sci.*, 2019.
[15] K. Jaafar, N. Ismail, M. Tajjudin, R. Adnan, dan M. H. F. Rahiman, "Identification of significant rainfall stations in Kelantan River using Z-score for Multi-Layer Perceptron (MLP) model development," 2017, doi: 10.1109/I2CACIS.2016.7885306.
[16] K. Blix, "Machine Learning Classification, Feature Ranking and Regression for Water Quality Parameters Retrieval in Various Optical Water Types from Hyper-Spectral Observations," 2020, doi: 10.1109/IGARSS39084.2020.9324717.
[17] K. D. Kankanamge, Y. R. Witharanage, C. S. Withanage, M. Hansini, D. Lakmal, dan U. Thayasivam, "Taxi Trip Travel Time Prediction with Isolated XGBoost Regression," 2019, doi: 10.1109/MERCon.2019.8818915.
[18] T. Chen dan C. Guestrin, "XGBoost: A scalable tree boosting system," 2016, doi: 10.1145/2939672.2939785.