Indo-MathEdu Intellectuals Journal

p-ISSN: 2808 – 5604
e-ISSN: 2808 – 5078
Volume. 4, No. 3, 2023

EXTREME GRADENT BOOSTING METHOD FORECASTING RAINFALL IN LEMBANG DISTRICT, WEST JAVA PROVINCE

Salma Azzahra Putri[1], Gumgum Darmawan[2], Restu Arisanti[3], Chrysentia Clarissa Clorinda[4]
[1, 2, 3, 4]Universitas Padjadjaran, Jl. Raya Bandung Sumedang KM.21, Sumedang, Jawa Barat, Indonesia
Email: salma20013@mail.unpad.ac.id

Article History

Received: 22-11-2023

Revision: 26-11-2023

Accepted: 28-11-2023

Published: 29-11-2023

Abstract. Lembang is a notable regional tourism destination that bears considerable significance within the urban area of Bandung. Lembang is widely recognized for its flourishing agricultural sector, which supports a significant community of farmers engaged in the cultivation of fruits, vegetables, and ornamental plants, in addition to its intrinsic scenic beauty. Therefore, the acquisition of precipitation data is of considerable significance for individuals live in the area to maintain their economic endeavors. This study employs daily historical data from the period of 2018 to 2021, wherein approximately 70% of the data is categorized as sparse. This discourse aims to examine the utilization of the Extreme Gradient Boosting (XGboost) technique for predicting rainfall in the Lembang region, specifically emphasizing its effectiveness in handling limited data. The findings indicate that the model, when trained and tested using a 7:3 data split ratio, achieved a mean absolute error (MAE) of 1.834 for training and 4.473 for testing. Additionally, the root mean square error (RMSE) was calculated to be 3.319 for training and 7.637 for testing. The optimal hyperparameters consist of a learning rate of 0.005, a max_depth value of 10, and the utilization of 300 decision trees as n_estimators. The model effectively captures the pattern of sparse time series data and non-rainy days data, as evidenced by its low error metrics. However, it slightly underestimates the rainfall rate on the days with intense precipitation.

Keywords: Lembang, Rainfall, XGBoost, Sparse Data

Abstrak. Lembang merupakan tujuan wisata regional yang memiliki arti penting di wilayah perkotaan Bandung. Lembang dikenal luas karena sektor pertaniannya yang berkembang pesat, yang mendukung komunitas petani yang terlibat dalam budidaya buah-buahan, sayuran, dan tanaman hias, di samping keindahan alamnya yang intrinsik. Oleh karena itu, perolehan data curah hujan sangat penting bagi individu yang tinggal di daerah tersebut untuk mempertahankan usaha ekonomi mereka. Penelitian ini menggunakan data historis harian dari periode 2018 hingga 2021, di mana sekitar 70% dari data tersebut dikategorikan jarang. Diskusi ini bertujuan untuk mengkaji pemanfaatan teknik Extreme Gradient Boosting (XGboost) untuk memprediksi curah hujan di wilayah Lembang, secara khusus menekankan pada keefektifannya dalam menangani data yang terbatas. Hasil penelitian menunjukkan bahwa model yang dilatih dan diuji dengan menggunakan rasio pembagian data 7:3 menghasilkan mean absolute error (MAE) sebesar 1.834 untuk pelatihan dan 4.473 untuk pengujian. Selain itu, root mean square error (RMSE) dihitung sebesar 3,319 untuk pelatihan dan 7,637 untuk pengujian. Hyperparameter yang optimal terdiri dari learning rate sebesar 0.005, nilai max_depth sebesar 10, dan penggunaan 300 pohon keputusan sebagai n_estimator. Model ini secara efektif menangkap pola data deret waktu yang jarang dan data hari tidak hujan, yang dibuktikan dengan metrik kesalahannya yang rendah. Namun, model ini sedikit meremehkan tingkat curah hujan pada hari.

Kata Kunci: Lembang, Rainfall, XGBoost, Sparse Data.

How to Cite: Putri, S. A., Darmawan, G., Arisanti, R., & Clorinda, C. C. (2023). Extreme Gradent Boosting Method Forecasting Rainfall in Lembang District, West Java Province. *Indo-MathEdu Intellectuals Journal,* 4 (3), 1959-1972. http://doi.org/10.54373/imeij.v4i3.452

INTRODUCTION

Rainfall, the amount of rain falling on the Earth's surface within a specific time period, usually measured in millimetres (mm), is a crucial factor in determining the climate and environmental patterns of a region (Trenberth et al., 2013). An accurate understanding of rainfall is key to planning and managing various aspects of life in a region, such as agricultural irrigation, flood mitigation, water resource management, and more. Therefore, monitoring and forecasting rainfall are essential efforts in maintaining sustainability and the well-being of the community in the Lembang region.

Rainfall in the Lembang region exhibits considerable fluctuations. Based on data recorded by the Meteorological, Climatological, and Geophysical Agency, it is observed that in early 2020, the rainfall in Lembang reached 83 mm. High rainfall is also observed in several areas across Indonesia. Deputy for Climatology, Herizal, asserted that this phenomenon is a result of the strengthening of the Asian monsoon flow and the convergence of the intertropical monsoon north of Java (Movanita, 2020). It is further noticeable that the frequency of rainfall throughout the period from 2020 to 2021 occurred quite frequently with a moderate intensity. Presently, there is a perceived shift towards increasingly unpredictable weather, often leading to destructive consequences. Hence, rainfall forecasting, especially in critical regions like Lembang, is essential to enable relevant authorities to undertake preventive measures concerning such occurrences (Li et al., 2021).

There are several benefits of rainfall forecasting from various perspectives. Firstly, in the agricultural sector, farmers can plan planting and harvesting times more efficiently based on rainfall predictions (Mishra & Singh, 2010; Adisa,O., 2019). This can reduce crop losses due to excessive or insufficient rainfall. Secondly, in the tourism sector, Lembang is known as a popular tourist destination with its natural beauty. Accurate weather forecasting will assist tourism site managers and tourists in planning their visits better and avoiding adverse weather conditions (Gad et al., 2020; Jorge-Gonzales, 2020). This can enhance the tourist experience and promote the tourism sector, which is crucial for the local economy. Thirdly, in terms of water resource management, rainfall forecasting can assist in planning the distribution of clean water and flood control (Huang et al., 2018; Nguyen,D., 2020). This is essential in maintaining an adequate supply of clean water for domestic and industrial purposes.

The advantage of using machine learning in rainfall forecasting in the Lembang region lies in its ability to process and analyze vast and complex weather data quickly. (Schultz, M., 2021). Machine learning's ability to identify patterns and trends that are difficult to identify by conventional methods makes it a valuable tool in predicting rainfall (Pathan, M., 2022).

Notably, the advantage of employing machine learning and deep learning methodologies in precipitation forecasting lies in their adeptness at handling large volumes of data. This capability allows for the amalgamation of diverse data sources such as satellite imagery, radar data, and ground-based observations, offering a more comprehensive and precise depiction of atmospheric conditions and rainfall likelihood. Furthermore, the training of machine learning algorithms enables the generation of highly refined predictions at a granular geographical and temporal scale (Latif et al., 2023).

Machine learning techniques have significantly improved rainfall rate predictions. Latif et al. (2023) evaluated machine learning and remote sensing methodologies in rainfall prediction, finding exceptional capabilities. Zhao et al. (2021) introduced an hourly rainfall forecast model, demonstrating their ability for precise temporal predictions. Ko et al. (2020) developed a method for hydrological applications, using machine learning to rectify quantitative precipitation forecasts. Tang et al. (2022) introduced a novel approach for medium- and long-term precipitation forecasting, integrating data augmentation methods and machine learning algorithms. Pontoh et al. (2022) investigated the correlation between Bandung's rainfall forecasts and Niño 3.4 using a nonlinear autoregressive exogenous neural network. These studies highlight the adaptability and effectiveness of machine learning methodologies across various forecasting timeframes and environmental contexts.

Extensive research corroborates the effectiveness of machine learning algorithms in handling large historical datasets and predicting sparse time series data. Woillard et al. (2021) demonstrated the utility of machine learning techniques in predicting medication exposure, while Rufaida et al. (2020) specifically leveraged gradient-boosting decision trees to enhance accuracy in indoor radio environment maps. Gupta and Batra (2017) emphasized the necessity for algorithms adept at identifying patterns within datasets characterized by limited observations. Furthermore, Kourentzes (2013) utilized neural networks as predictive tools for intermittent demand, a common challenge in time series data with sparse data points.

This study uses one of the predictive modelling techniques, XGBoost. The primary objective is to examine the benefits of utilising XGBoost in the context of rainfall prediction. In particular, the study highlights the effectiveness of XGBoost in managing big datasets and sparse data, which often contain several zero values. This is achieved by the use of a split-finding technique that is specifically designed to handle sparsity in the data. The utilised dataset encompasses daily records of rainfall in Lembang from 2018 to 2021. This dataset is classified as univariate time series data, with 70% of the observations displaying sparsity. The primary focus of this study will involve the assessment of the XGBoost algorithm's effectiveness in

predicting sparse time series data. The purpose of the following analysis is to investigate the ideal ratio for dividing datasets into training and testing sets, as well as to determine the most effective hyperparameters for achieving accurate predictions of rainfall.

METHODS

Description of the Data Collection and Study Area

The data used in this research is the daily rainfall data for the Lembang district, acquired from the database maintained by the Meteorology, Climatology, and Geophysics Agency, specifically the BHLK database. The data included in this research was collected during the period spanning from January 1, 2018, to December 31, 2021. The study area of this research is Lembang district, Bandung, West Java. October to May is the start of the 6.6-month rainy season for Lembang, during which there is a greater than 44% probability of rainy days. In Lembang, the month characterized by the highest frequency of precipitation is January, during which an average of 22.4 days have rainfall of at least 1 millimeter (Weather Spark, 2019).

Ensemble Learning

In the process of data analysis, various methods often exhibit imperfections, with many displaying shortcomings, notably a lack of stability and susceptibility to overfitting conditions when applied to specific cases that may not align with the inherent characteristics of these methods (Suyanto, 2018). Consequently, a number of researchers have proposed the adoption of ensemble learning techniques. Ensemble learning is a machine learning paradigm in which the algorithm serves as a search for the best predictive solution compared to other algorithms (Gonzales et al., 2020). This is achieved through the utilization of ensemble methods, also known as committee-based learning, employing multiple learning algorithms to attain a predictive solution superior to what can be obtained from any single constituent algorithm (Zhou Zhihua, 2012).

Gradient Boosting

Quoted from the official IBM website (IBM.com), Boosting is an ensemble method that combines multiple weak models to generate a robust model while minimizing training error. Boosting was initially proposed by Robert E. Schapire in 1990 (Schapire, 1990). Gradient Boosting is a machine learning method as a linear additive model consisting of an ensemble of weak prediction models (Zhang et al., 2021). It takes M steps to obtain a complete model F. Equation (3) articulates this statement more explicitly as follows:

$$F_{m+1} = F_m + h_{m+1}(x) \quad (1)$$

The initial model $h_{m+1}(x)$ will be trained to calculate the residual of $y - F_m$ to further predict on $m + 1$ step, instead of directly optimizing the model F_m in the $m + 1$ step. In general, the negative gradient in the object function is used as a residual to learn the initial model $h(x)$.

Extreme Gradient Boosting

XGBoost, introduced by Chen and Guestrin in 2016, is an open-source project implementing Gradient Tree Boosting for efficient and scalable machine learning in diverse paper-based learning problems. The XGBoost algorithm is a robust regression model that uses multiple Classification and Regression Trees (CART) to address regression and classification problems. The structure consists of root nodes, internal nodes, leaf nodes, and branches. The i-th parameter is input and propagated to all CARTs, followed by internal nodes, branch points, and leaf nodes.

XGBoost splits large-scale data into several quantiles to find the best splits. The proposed approach use parallel learning and a weighted quantile sketch technique to partition data into smaller groups and arrange them into tree structures. The technique additionally incorporates feature regularization to address the issue of overfitting. The algorithm for identifying splits, which takes into account the sparsity of the data and the presence of missing values, ensures a high level of robustness. The recursive algorithm iteratively improves the model by picking the direction that yields the highest gain and minimizes the computing cost. The core concept of XGBoost is to continuously add weak trees with different weights to the ensemble (Lingyu, 2021). The trees in the ensemble should be as close as possible to the previous prediction, as indicated in the following Equation (2).

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \qquad f_k \in F \quad (2)$$

In Equation (2), $\hat{y}$ represents the predicted value, and f_k is a regression tree that belongs to F, a set comprising regression trees, with k denoting the number of regression trees in F. The expected prediction value is one that closely approximates the true value y_i without compromising its generalization ability. Furthermore, the formula for calculating the objective function is illustrated in Equation (3).

$$Obj^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{i=1}^{t} \Omega(f_t) + constant \quad (3)$$

In equation (3), the first part of the formula is called as *loss function* $l\left(y_i, \hat{y}_i^{(t)}\right)$ which defines the discrepancy between predicted and actual. The employed loss function can take any form as long as it constitutes a second-order derivative. Meanwhile, $\Omega(f_t)$ denotes regularization, delineating the model's complexity. A smaller $\Omega(f_t)$ signifies reduced model complexity and heightened generalization capabilities. The regularization is expressed in Equation (4) below:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=i}^{T} \omega_j^2 \tag{4}$$

In Equation (4), *T* represents the total number of leaf nodes in each tree, and ω denotes the value representing the leaf node. Throughout the training process of XGBoost, optimization is conducted incrementally. At each step, the objective function is minimized by generating a new tree based on the existing one. Initially, the existing tree is replaced by a constant c in the equation. Subsequently, the Taylor second-order expansion is employed to articulate the loss function, yielding the final objective function as follows:

$$Obj^{(t)} = \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right)\omega_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right)\omega_j^2\right] + \gamma T \tag{5}$$

Where:

$$g_i = \frac{\partial l\left(y_i, \hat{y}_i^{(t-1)}\right)}{\partial \hat{y}_i^{(t-1)}} \tag{6}$$

$$h_i = \frac{\partial^2 l\left(y_i, \hat{y}_i^{(t-1)}\right)}{\partial\left(\hat{y}_i^{(t-1)}\right)^2} \tag{7}$$

g_i and h_i is respectively the first and second derivatives of each data of the error function. Whereas the index I_j represents the sample for each leaf.

$$I_j = \{i | q(x_i) = j\} \tag{8}$$

For a given $q(x_i)$, setting the derivative of ω_j equal to zero allows obtaining the optimal weight $\omega *_j$ from leaf j, where the optimal weight is expressed in Equation (9) as follows:

$$\omega_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda} \tag{9}$$

Therefore, the optimal objective function is obtained as expressed by the following Equation (10).

$$\tilde{L}^{(t)}(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} - \frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} - \gamma \quad (10)$$

Through Equation (10), the optimal value of the objective function *L*, which represents the prediction value displayed at the leaf node, is obtained. To search for the optimal structure for each CART, a greedy algorithm is employed to optimize the regression tree structure (Friedman, 2001).

Sparse Data

Data that is sparsely populated with information or contains only a few relevant features is commonly termed as sparse data (Smith, 2010). This situation is prevalent across various domains, including biomedical science, recommendation systems, text analysis, and many more. Issues might emerge in the occurrence of missing values and sparse data within the collection. XGBoost incorporates a sparsity-aware split-finding method that effectively addresses missing values within its framework. The approach facilitates the creation of a Classification and Regression Tree (CART) on the XGBoost framework, enabling the direct handling of missing variables.

Analytical Workflow

Overall, in this study, the analysis of forecasting daily rainfall rate in Lembang consists of 4 steps that can be summarised as follows (1) supervised learning is a machine learning technique that is not directly applicable to real-time data. Therefore, the time series data must be converted into supervised data using the sliding window approach, (2) data is partitioned into two sets; training and testing. The training dataset is used to analyze data and build models with parameters, while the testing dataset is used to create a model with good performance, (3) then, the XGBoost algorithm will be performed on the training data. To optimize the model, several hyperparameters are considered: n_estimators, max_depth, and learning_rate. Additional algorithms can be used to improve hyperparameter performance, such as GridSearchCV, and (4) the evaluation process to assess the performance of the algorithm used, using the MAE and RMSE methods in this study.

RESULTS

The analysis of Lembang District Rainfall forecasting was performed using the open-source Python software, Google Collaboratory. The best model was chosen based on MAE,

MSE and RMSE. The model with the fewest error size values was selected and thus the best hyperparameter for the model is determined. The selected model will predict the rainfall rate in Lembang District for the next 30 days.

Data Preprocessing and Setting

When preparing the data for analysis, there are numerous tasks that need to be completed. Given that the dataset is devoid of any missing values, it is unnecessary to apply any type of treatment or additional changes to the data. Despite the fact that 70% of the data in this study is sparse, no treatments will be applied in order to observe the performance of XGBoost in handling sparse time series data. Supervised learning is not ideal for real-time data processing, so it requires several steps to be taken. The first step involves using the lag or time value before the specified time value to analyze data. The sliding window approach is used to analyze real-time data before the temporal value becomes the target variable.

Data is then partitioned into training and testing. The study investigates the impact of different data split ratios on model performance and error rates, focusing on three ratios: 8:2, 7:3, and 6:4. The quantity of training and test data is not predetermined and may vary. The optimal hyperparameter for the model is determined using the GridSearchCV function, which constructs a grid of potential configurations by exploring a predetermined set of hyperparameter values. Cross-validation is used to assess each combination, ultimately choosing the one with the highest performance level. The dataset is randomized and divided into categories, with one designated as the test data group and the remaining for training purposes. The configuration of the hyperparameter range is shown in Table 1.

Table 1. Hyperparameter Range Configuration

Hyperparameter	Range
n_estimators	[50 - 1000]
Max_depth	[5-15]
Learning_rate	[0.005-0.1]

XGBoost Model Performance Result

The best-estimated model was achieved once the best combination of hyperparameters was determined. It was then evaluated on the training and the test dataset, respectively. A dataset of 1,461 sets was analysed, which was divided into training and test parts with a ratio of 7:3. Using the dataset, the depth of the XGBoost decision tree model was adjusted to 10, the learning rate was adjusted to 0.005, and the XGBoost model was trained using an ensemble of 300 regression trees. Table 2 shows the evaluation metrics of the XGBoost model on the

training and testing dataset. As shown in Table 2, the MAE and RMSE value for the training dataset using the 7:3 split ratio is relatively small and the overall prediction results were satisfactory.

The chosen combination of hyperparameters, including a low learning rate (0.005), a deep tree structure with a max depth of 10, and a moderate number of estimators (300), reflects a deliberate emphasis on constructing a robust and accurate predictive model. The decision to employ a low learning rate indicates a cautious training strategy, demonstrating an awareness of the potential pitfalls of overfitting and a commitment to enhancing model generalization. The incorporation of a high max depth reveals an intent to capture intricate patterns within the data, showcasing a nuanced understanding of the underlying complexity. However, a vigilant eye is recommended to monitor for overfitting, particularly in scenarios where the dataset may not be extensive. Notably, the consistency of performance metrics such as MAE, MSE and RMSE between training and testing sets suggests that the model has achieved a balanced fit without exhibiting a pronounced degree of overfitting, validating the efficacy of the chosen hyperparameter configuration.

Table 2. Metrics Evaluation and Hyperparameters

Data Split Ratio	Training	Testing	Hyper Parameter
8:2	MAE: 2.710	MAE: 4.169	Learning rate: 0.005
	RMSE: 4.484	RMSE: 7.162	Max_depth: 5
	MSE: 20.106	MSE: 51.294	n_estimator: 500
7:3	MAE: 1.834	MAE: 4.473	Learning rate: 0.005
	RMSE: 3.319	RMSE: 7.637	Max_depth: 10
	MSE: 11.015	MSE: 58.323	n_estimator: 300
6:4	MAE: 2.483	MAE: 3.943	Learning rate: 0.005
	RMSE: 4.070	RMSE: 6.684	Max_depth: 5
	MSE: 16.564	MSE: 44.675	n_estimator: 500

Daily Rainfall Prediction Result

The data provided in Table 2 and Figure 1 illustrates the accurate depiction of anticipated and projected precipitation levels for every day in January 2022. Additionally, it includes favourable assessment measures such as the Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE).

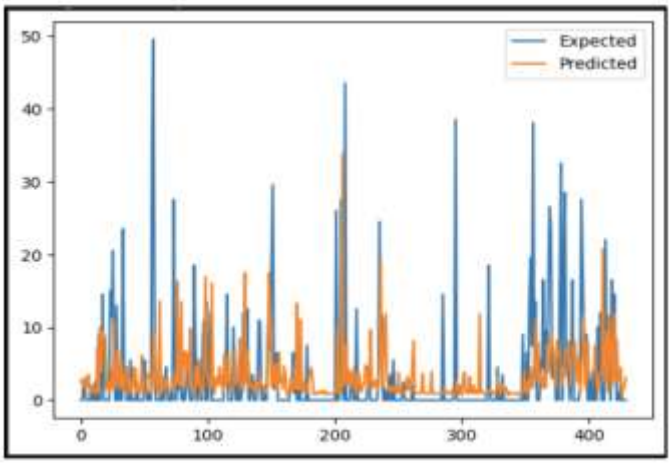

Figure 1. Rainfall rate expected vs predicted for January 2022

Overall, the model exhibits a noteworthy level of accuracy in predicting daily rainfall, as evidenced by continuously low Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) values. The model has exceptional performance in properly forecasting days with limited or no precipitation, highlighting its precision in capturing the sparse patterns present in the time series data. A notable occurrence took place on January 10, 2022, wherein the model demonstrated a high degree of conformity with the anticipated rainfall value of 2.5. This alignment yielded remarkably low Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) values of 0.065.

Notwithstanding these accomplishments, a few challenges take place on days characterized by elevated projected precipitation, specifically on January 11th and January 12th, 2022. The model exhibits a tendency to somewhat underestimate the precipitation levels during these instances, resulting in elevated values for both the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). This observation indicates a possible aspect that could be improved in the process of accurately representing extreme events or intense precipitation, implying a potential for additional optimization in the model's setup.

The efficacy of XGBoost's sparsity-aware split discovery algorithm is apparent in its proficient management of intermittent time series data, specifically in situations characterized by zero or low precipitation. The algorithm's ability to make efficient split decisions, which are specifically designed to accommodate the sparsity features of the data, significantly enhances the accuracy of forecasts in these particular cases.

Table 3. Daily Univariate Model Evaluation for 7:3 splitting

Date	Expected	Predicted	MAE (Pred)	RMSE (Pred)	MSE (Pred)
1/1/2022	0	1.8	1.762	1.762	3.105
1/2/2022	0	1.9	1.866	1.866	3.482
1/3/2022	2.5	1.7	0.837	0.837	0.701
1/4/2022	0	1.5	1.495	1.495	2.235
1/5/2022	0	1.4	1.438	1.438	2.068
1/6/2022	0	0.9	0.881	0.881	0.776
1/7/2022	0	0.8	0.754	0.754	0.569
1/8/2022	0	0.9	0.872	0.872	0.760
1/9/2022	9	6.7	2.261	2.261	5.112
1/10/2022	2.5	2.6	0.065	0.065	0.004
1/11/2022	18	12.8	5.228	5.228	27.332
1/12/2022	21.5	16.2	5.328	5.328	28.388
1/13/2022	6	5.2	0.844	0.844	0.712
1/14/2022	0	1.6	1.573	1.573	2.474
1/15/2022	0	1.5	1.549	1.549	2.399
1/16/2022	1	1.5	0.453	0.453	0.205
1/17/2022	0	1.4	1.352	1.352	1.828
1/18/2022	0	1	1.026	1.026	1.053
1/19/2022	5.5	4.3	1.245	1.245	1.550
1/20/2022	12	8.7	3.309	3.309	10.949
1/21/2022	8.5	7.3	1.236	1.236	1.528
1/22/2022	0	0.7	0.729	0.729	0.531
1/23/2022	0	1.2	1.153	1.153	1.329
1/24/2022	0	0.6	0.602	0.602	0.362
1/25/2022	0	0.7	0.711	0.711	0.506
1/26/2022	0	1	0.986	0.986	0.972
1/27/2022	0	0.6	0.642	0.642	0.412
1/28/2022	0	0.9	0.921	0.921	0.848
1/29/2022	0	0.9	0.895	0.895	0.801
1/30/2022	0	1	0.96	0.96	0.922

DISCUSSION

The results of our investigation underscore the notable precision exhibited by our XGBoost model in predicting daily precipitation levels. Nevertheless, it is imperative to recognize the significant obstacle that was faced during periods of heightened anticipated precipitation, notably on January 11th and January 12th, 2022. In these particular cases, the model exhibited a tendency to underestimate the observed quantities of precipitation, leading to elevated values of MAE, MSE, and RMSE.

Algorithms' inability to generalize when faced with extreme events, such as high precipitation rates, may be hampered by sparse data. The model's capacity to exploit correlations with other factors impacting precipitation is constrained by the univariate character of sparse data. Days with high precipitation may exhibit distinct characteristics or abrupt fluctuations that are not adequately captured in the training dataset, posing a challenge to the model's capacity to appropriately generalize. The presence of algorithmic bias may also contribute to the phenomenon of underestimating. Overcoming these issues may require the implementation of several techniques such as feature engineering, data augmentation, and method assembly.

CONCLUSION

In summary, the research underscores the efficacy of employing the XGBoost algorithm in forecasting univariate time series data of daily precipitation, particularly during days characterized by minimal or absent rainfall. The model effectively captures sparsity patterns within the data, hence bearing substantial consequences for the fields of weather prediction and forecasting. Although there exist difficulties in forecasting significant rainfall, precise daily precipitation predictions have practical implications across multiple industries. In the future, it is crucial to persistently enhance forecast models for extreme weather phenomena through the integration of various data sources, the enhancement of algorithms, and the utilization of creative methodologies. These proposed enhancements would not only facilitate the progress of weather forecasting but also bolster social resilience and readiness in the face of extreme weather phenomena.

REFERENCES

Adisa, O., Botai, J., Adeola, A., Hassen, A., Botai, C., Darkey, D., & Tesfamariam, E. (2019). Application of Artificial Neural Network for Predicting Maize Production in South Africa. Sustainability.

Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: KDD’16 Proceedings of the 22nd SCM SIGKDD international conference on knowledge discovery and data mining, pp 785–794

Feng, S., & Duarte, M. (2018). Graph Autoencoder-Based Unsupervised Feature Selection with Broad and Local Data Structure Preservation. Neurocomputing, 312, 310-323.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, *29*(5), 1189–1232.

Gad, I., & Hosahalli, D. (2020). A comparative study of prediction and classification models on NCDC weather data. International Journal of Computers and Applications, 44, 414 - 425.

González, S., García, S., Ser, J., Rokach, L., & Herrera, F. (2020). A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. Inf. Fusion, 64, 205-237.

Gupta, P., & Batra, S. S. (2017). Sparse short-term time series forecasting models via minimum model complexity. Neurocomputing, 243, 1-11.

Huang, C., Yu, X., & Wu, H. (2018). Flood prediction based on support vector machine with a binary particle swarm optimization. Journal of Hydrology, 556, 347-361.

Hyndman, R.J., & Athanasopoulos, G. (2021) Forecasting: principles and practice, 3rd edition, OTexts: Melbourne, Australia.

Iklim, Cuaca Menurut Bulan, Suhu Rata-Rata Lembang (Indonesia) - Weather Spark. (n.d.). Weather Spark.

Jorge-González, E., González-Dávila, E., Martín-Rivero, R., & Lorenzo-Díaz, D. (2020). Univariate and multivariate forecasting of tourism demand using state-space models. Tourism Economics, 26, 598 - 621.

Ko, C., Jeong, Y., Lee, Y., & Kim, B. (2020). The Development of a Quantitative Precipitation Forecast Correction Technique Based on Machine Learning for Hydrological Applications. Atmosphere, 11, 111.

Kourentzes, N. (2013). Intermittent demand forecasts with neural networks. International Journal of Production Economics, 143(1), 198-206.

Larasati, R. A. (2019, October 18). *Ekspor Kopi Indonesia Turun, Apa Sebabnya?* Money Kompas. Retrieved August 25, 2023.

Latif, S. D., Hazrin, N. A. B., Koo, C. H., Ng, J. L., Chaplot, B., Huang, Y. F., El-Shafie, A., & Ahmed, A. N. (2023, November 1). *Assessing rainfall prediction models: Exploring the advantages of machine learning and remote sensing approaches*. Alexandria Engineering Journal. https://doi.org/10.1016/j.aej.2023.09.060

Li, H., He, Y., Yang, H., Wei, Y., Li, S., & Xu, J. (2021). Rainfall prediction using optimally pruned extreme learning machines. Natural Hazards, 108, 799 - 817.

Mishra, A.K. and Singh, V.P. (2010) A Review of Drought Concepts. Journal of Hydrology, 391, 202-216. https://doi.org/10.1016/j.jhydrol.2010.07.012.

Movanita, A. N. K. (2020, January 3). *BMKG Beberkan Penyebab Tingginya Curah Hujan di Awal 2020*. KOMPAS.com.

Nguyen, D., & Bae, D. (2020). Correcting mean areal precipitation forecasts to improve urban flooding predictions by using long short-term memory network. Journal of Hydrology, 584, 124710.

Pathan, M., Nag, A., & Dev, S. (2022). Efficient Rainfall Prediction Using a Dimensionality Reduction Method. IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium, 6737-6740. https://doi.org/10.1109/IGARSS46834.2022.9884849.

Pontoh, R.S.; Toharudin, T.; Ruchjana, B.N.; Sijabat, N.; Puspita, M.D. Bandung Rainfall Forecast and Its Relationship with Niño 3.4 Using Nonlinear Autoregressive Exogenous Neural Network. Atmosphere 2022, 13, 302.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Ben Taieb, S., Bergmeir, C., Bessa, R. J., Bijak, J., Boylan, J. E., Browell, J., Carnevale, C., Castle, J. L., Cirillo, P., Clements, M. P., Cordeiro, C., Cyrino Oliveira, F. L., De Baets, S., Dokumentov, A., . . . Ziel, F. (2022, July). Forecasting: theory and practice. *International Journal of Forecasting*, *38*(3), 705–871.

Rufaida, S., Leu, J., Su, K., Haniz, A., & Takada, J. (2020). Construction of an indoor radio environment map using gradient boosting decision tree. Wireless Networks, 26, 6215 - 6236.

Santika, E. F. (2023, January 19). *Jawa Barat, Provinsi dengan Nilai Ekspor Terbesar pada 2022*. Databoks. Retrieved August 21, 2023

Schapire , R. (1990). A Brief Introduction to Boosting. IJCAI'99: Proceedings of the 16th international joint conference on Artificial intelligence, 1999 (pp. Pages 1401–1406). 180 Park Avenue, Room A279, Florham Park, NJ 07932, USA: AT&T Labs, Shannon Laboratory.

Schultz, M., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L., Mozaffari, A., & Stadtler, S. (2021). Can deep learning beat numerical weather prediction?. Philosophical transactions. Series A, Mathematical, physical, and engineering sciences, 379. https://doi.org/10.1098/rsta.2020.0097.

Tang, T., Jiao, D., Chen, T., & Gui, G. (2022). Medium- and Long-Term Precipitation Forecasting Method Based on Data Augmentation and Machine Learning Algorithms. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15, 1000-1011.

Trenberth, K. E., Dai, A., Van Der Schrier, G., Jones, P. D., Barichivich, J., Briffa, K. R., & Sheffield, J. (2013). Global warming and changes in drought. Nature Climate Change, 4(1), 17-22.

Smith, M. R. (2010). Sparse modeling for image and vision processing. Now Publishers Inc.

Suyanto. (2018). Machine Learning Tingkat Dasar dan Lanjut . Bandung: Informatika Bandung.

Syahrani, I. (2019). Analisis Perbandingan Teknik Ensemble Secara Boosting (XGBOOST) Dan Bagging (RANDOM FOREST) Pada Klasifikasi Kategori Sambatan Sekuens DNA. Sekolah Pasca Sarjana Institut Pertanian Bogor.

Wang, L., & Li, R. (2020). Learning Low-Dimensional Latent Graph Structures: A Density Estimation Approach. IEEE Transactions on Neural Networks and Learning Systems, 31, 1098-1112.

Woillard, J., Labriffe, M., Prémaud, A., & Marquet, P. (2021). Estimation of drug exposure by machine learning based on simulations from published pharmacokinetic models: the example of tacrolimus.. Pharmacological research, 105578 .

Zhang, L. (2021). Time Series Forecast of Sales Volume based on XGBoost. *Journal of Physics: Conference Series*, 9.

Zhao, Q., Liu, Y., Yao, W., & Yao, Y. (2021). Hourly Rainfall Forecast Model Using Supervised Learning Algorithm. IEEE Transactions on Geoscience and Remote Sensing, 60, 1-9.

Zheng F & Zhong S. 2011. Time series forecasting using a hybrid RBF neural network andAR model based on binomial smoothing. World Academy of Science. Eng Technol75:1471- 1475

Zhou Zhihua (2012). Ensemble Methods: Foundations and Algorithms. Chapman and Hall/CRC.

Zou, M., Jiang, G., Qin, H., Liu, C., & Li, L. (2022). Optimized XGBoost Model with Small Dataset for Predicting Relative Density of Ti-6Al-4V Parts Manufactured by Selective Laser Melting. *Materials*, *15*(15).