Rainfall prediction using Extreme Gradient Boosting

by Wiwien Hadikurniawati

Journal of Physics: Conference Series

Rainfall prediction using Extreme Gradient Boosting

Rainfall prediction using Extreme Gradient Boosting

M T Anwar[1,*], E Winarno[1], W Hadikurniawati[1] and M Novita[2]

[1] Faculty of Information Technology, Universitas Stikubank, Jl. Tri Lomba Juang No 1 Semarang 50241, Central Java, Indonesia
[2] Faculty of Engineering and Informatics, Universitas PGRI Semarang, Jl. Sidodadi-Timur No.24 Semarang, Central Java 50232, Indonesia

*taufiq@edu.unisbank.ac.id

Abstract. Rainfall greatly affects human life in various sectors including agriculture, transportation, etc. and also can affect natural disasters such as drought, floods, and landslides. This situation prompts us to build an accurate rainfall prediction model so that prescriptive measures can be made. Previous research on rainfall prediction uses models that have their limitations and thus produce poor performance. This study aims to build a multivariate rainfall prediction model using the best performing technique to date namely the Extreme Gradient Boosting. This model is built based on 7 years of historical weather data collected by the weather station. The result had demonstrated that the model is capable of producing accurate predictions for daily rainfall estimates with training RMSE of 2.7 mm and the testing MAE of 8.8 mm.

1. Introduction

Rainfall greatly affects human life in various sectors including agriculture, transportation, etc. and also can affect natural disasters such as drought, floods, and landslides. Thus, rainfall prediction models are needed to assist decision making and management in these various needs. Research on rainfall prediction can use models based on univariate time-series analysis such as Auto-Regressive Integrated Moving Average (ARIMA) [1] and Exponential Smoothing [2]. Univariate time-series models can be useful if the factors affecting the objective variable are not well understood. Research [3] has shown that the two main factors that influence rain are minimum temperature and average relative humidity. Time series models can also be performed using multivariate analysis, for example by using Vector Auto-Regression (VAR) [4,5]. However, VAR is built based on the assumption of a linear relationship between the determinant attributes and the objective attributes, whereas research [3] found that the relationship between the variable affecting rainfall and rainfall itself is non-linear. Based on these findings, this study tries to build a predictive model to estimate the amount of rainfall with a multivariate approach with a non-linear relationship. One of the Data Mining methods that can handle these needs and has the best performance today is Extreme Gradient Boosting (XGBoost). Meanwhile, the latest research using the Machine Learning (ML) approach was conducted using several methods such as Artificial Neural Network (ANN), Support Vector Machine (SVM), and Particle Swarm Optimization - Adaptive Neuro-Fuzzy Inference System (PSO-ANFIS) [6]. It should be noted that apart from estimates based on Earth-based data, rainfall estimates can also use remote-sensing data [7].

2. Methods

Daily weather data were obtained from the Indonesian Meteorology, Climatology, and Geophysics Agency (BMKG) for Tanjung Mas, Semarang City, Indonesia with 11 attributes. Of the 11 attributes, only 8 attributes are used as shown in Table 1 with the attributes RR (rainfall) being the value to be predicted by the model. The training dataset consists of daily weather data from 2013 to 2019, whereas the testing data is the daily weather data in 2020. In the training phase, entries with missing values are omitted. Experiments were carried out using RStudio version 1.2.5001, R version 3.6.1, and XGBoost package version 1.1.1.1.

Table 1. The attributes of the weather data.

Attribute	Data type	Description
Tn	Numeric	Minimum temperature
Tx	Numeric	Maximum temperature
Tavg	Numeric	Average temperature
RH_avg	Numeric	Average Humidity (%)
ss	Numeric	Sun exposure time (hours)
ff_x	Numeric	Maximum wind speed (m/s)
ff_avg	Numeric	Average wind speed (m/s)
RR	Numeric	Rainfall (mm)

2.1. Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is an ensemble learning method. Sometimes, relying on the results of a single machine learning model such as J48 may not be enough. Although J48 is good enough and had been used in several problems such as wildfire modeling [8] and rain modeling [3]. Ensemble learning combines multiple learners to get a more powerful prediction. In this case, XGBoost uses boosting. Multiple trees are created sequentially in a way that each one of the next trees tries to reduce the errors from the previous tree.

XGBoost was first released in 2014 and had been implemented in Python, R [9] packages, etc. XGBoost is very popular and wins numerous Kaggle competitions. Currently, XGBoost has been used for various purposes such as prediction of crude oil prices [10], diagnosis of chronic kidney disease [11], prediction of accidents [12], prediction of employees changing jobs [13], prediction of material particulates in the atmosphere [14], and intrusion detection [15]. However, until now there has been no research using XGBoost for rainfall prediction. So this research is the first research to use XGBoost for rainfall prediction. XGBoost itself can handle both classification and regression tasks. Although XGBoost has a good performance, it has a drawback that is the possibility of overfitting. This can be handled by experimenting with modeling parameters.

3. Results and discussion

3.1. Training

Figure 1 shows the training and the test RMSE for round 1 to 100. As the iterations go, training RMSE gets better but test RMSE gets the best RMSE at about round 5. Beyond this best RMSE, the test error started to rise again. This phenomenon indicates the overfitting tendency of XGBoost as the number of rounds increases.

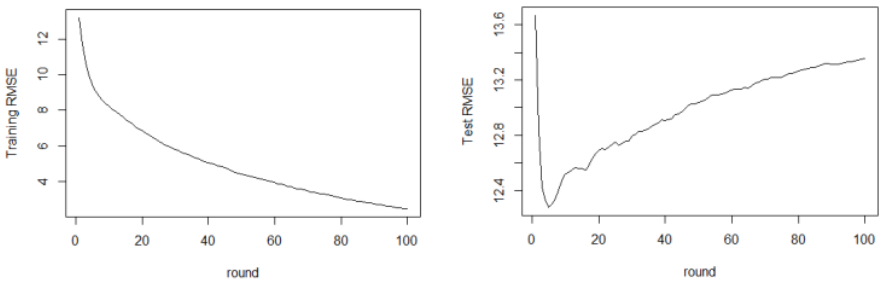

Figure 1. Training and Test RMSE during the iteration.

Table 2 shows that regarding the training error, 10-fold cross-validation produced a slightly lower error than when using the full training dataset which may indicate that the model is prone to overfitting. When the non-rainy data were excluded, the error is higher. This higher error might be caused by the loss of information about the characteristic of non-rainy days. Thus reducing the model's ability to accurately predict the amount of rainfall.

Table 2. Training RMSE using nrounds = 100, nfold = 10, eta = 0.3, max_depth = 6.

Round	Training RMSE			
	Non-rainy data included		Non-rainy data excluded	
	10-fold CV	Full training set	10-fold CV	Full training set
1	13.22+0.25	13.30	22.36+0.39	20.46
50	4.41+0.14	4.75	9.62+0.27	12.93
100	2.46+0.11	2.75	6.92+0.25	10.85

XGBoost is capable of ranking the important attributes that contribute to the model. The ranking of the attributes is shown in Figure 2. It shows that the rainfall prediction model is primarily affected by the average relative humidity (RH_avg) and the minimum temperature (Tn). This result agrees with previous research [3] which used the C4.5 model to predict if a day is rainy or not.

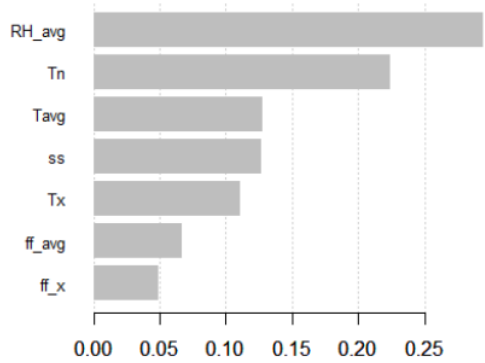

Figure 2. The ranking of attributes importance.

3.2. Testing

When tested against the weather data in 2020, the model gave an MAE of 8.8. Figure 3 shows the scatter plot of the predicted RR against the actual RR on the test dataset. It shows that many of the data are concentrated on near 0 value which is very challenging for the model. The blue line is the linear trend-line with a correlation R of 0.555. This result is lower than recent research [16] which uses Nonlinear Autoregressive Neural Network and has R = 0.9. This call future research to explore the parameter setting (tuning) of XGBoost to improve its prediction ability. Further research can also combine Earth-based data with remote-sensing data to create a more accurate model.

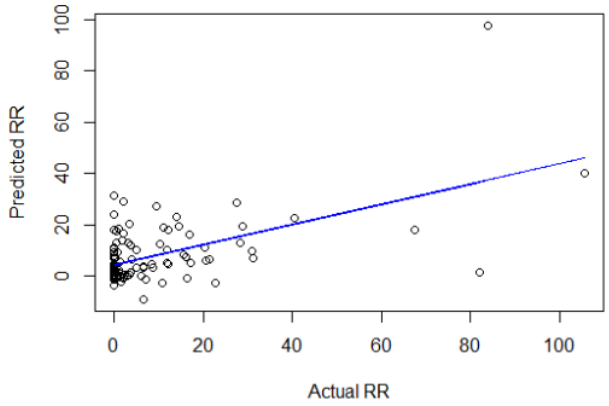

Figure 3. Scatter plot showing the predicted RR against the actual RR on the test dataset.

4. Conclusion

This study proposes a non-linear multivariate rainfall prediction model with XGBoost. This model is built based on 7 years of historical weather data collected by the weather station. The result had demonstrated that the model is capable of producing accurate predictions for daily rainfall estimates with training RMSE of 2.7 mm and the testing MAE of 8.8 mm. The results also show that the factors that most influence rainfall are the average humidity and the minimum temperature. Future research needs to explore the parameter tuning for XGBoost both to increase the accuracy and reduce overfitting, investigate how zero values affect the model accuracy, and add remote-sensing data to enrich the model attributes.

Acknowledgement

We thank the Meteorology, Climatology, and Geophysical Agency (BMKG) for providing the weather data.

References

[1] Verma A P and Chakraborty B S 2020 Performance Estimation of ARIMA Model for Orographic Rainfall Region *2020 URSI Regional Conference on Radio Science (URSI-RCRS)* pp 1–4

[2] Hartomo K D, Prasetyo S Y J, Anwar M T and Purnomo H D 2019 Rainfall Prediction Model Using Exponential Smoothing Seasonal Planting Index (ESSPI) For Determination of Crop Planting Pattern *Computational Intelligence in the Internet of Things* (IGI Global) pp 234–55

[3] Anwar M T, Nugrohadi S, Tantriyati V and Windarni V A 2020 Rain Prediction Using Rule-Based Machine Learning Approach *Adv. Sustain. Sci. Eng. Technol.* **2**

[4] Ramli I, Rusdiana S, Basri H, Munawar A A and others 2019 Predicted Rainfall and discharge Using Vector Autoregressive Models in Water Resources Management in the High Hill

Takengon *IOP Conference Series: Earth and Environmental Science* vol 273 p 12009
[5] Sivajothi R and Karthikeyan K 2019 Forecasting of Rainfall, Average Temperature, Vapor Pressure and Cloud Cover Using Vector Autoregression Model *J. Comput. Theor. Nanosci.* **16** 1862–9
[6] Pham B T, Le L M, Le T-T, Bui K-T T, Le V M, Ly H-B and Prakash I 2020 Development of advanced artificial intelligence models for daily rainfall prediction *Atmos. Res.* **237** 104845
[7] Parida B R, Behera S N, Bakimchandra O, Pandey A C and Singh N 2017 Evaluation of satellite-derived rainfall estimates for an extreme rainfall event over Uttarakhand, Western Himalayas *Hydrology* **4** 22
[8] Anwar M T, Pumomo H D, Prasetyo S Y J and Hartomo K D 2018 Decision Tree Learning Approach To Wildfire Modeling on Peat and Non-Peat Land in Riau Province *2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS)* (IEEE) pp 409–15
[9] Chen T, He T, Benesty M, Khotilovich V and Tang Y 2015 Xgboost: extreme gradient boosting *R Packag. version 0.4-2* 1–4
[10] Gumus M and Kiran M S 2017 Crude oil price forecasting using XGBoost *2017 International Conference on Computer Science and Engineering (UBMK)* pp 1100–3
[11] Ogunleye A A and Qing-Guo W 2019 XGBoost model for chronic kidney disease diagnosis *IEEE/ACM Trans. Comput. Biol. Bioinforma.*
[12] Shi X, Li Q, Qi Y, Huang T and Li J 2017 An accident prediction approach based on XGBoost *2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)* pp 1–7
[13] Jain R and Nayyar A 2018 Predicting employee attrition using xgboost machine learning approach *2018 International Conference on System Modeling & Advancement in Research Trends (SMART)* pp 113–20
[14] Pan B 2018 Application of XGBoost algorithm in hourly PM2. 5 concentration prediction *IOP Conference Series: Earth and Environmental Science* vol 113 p 12127
[15] Dhaliwal S S, Nahid A-A and Abbas R 2018 Effective intrusion detection system using XGBoost *Information* **9** 149
[16] Le V M, Pham B T, Le T-T, Ly H-B and Le L M 2020 Daily Rainfall Prediction Using Nonlinear Autoregressive Neural Network *Micro-Electronics and Telecommunication Engineering* (Springer) pp 213–21

Rainfall prediction using Extreme Gradient Boosting

ORIGINALITY REPORT

11%	6%	9%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Muchamad Taufiq Anwar, Wiwien Hadikurniawati, Edy Winarno, Wahyu Widiyatmoko. "Performance Comparison of Data Mining Techniques for Rain Prediction Models in Indonesia", 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), 2020 Publication	3%
2	www.science.gov Internet Source	2%
3	Submitted to Asia Pacific University College of Technology and Innovation (UCTI) Student Paper	1%
4	Submitted to Liverpool John Moores University Student Paper	1%
5	Malcolm R. Burns, David J. Faurot. "An econometric forecasting model of revenues from urban parking facilities", Journal of Economics and Business, 1992	1%

	Publication	
6	Shweta Loonkar, Dhirendra S. Mishra. "Defect Classification for Silk Fabric Based on Four DFT Sector Features", 2019 IEEE Conference on Information and Communication Technology, 2019 Publication	1 %
7	authors.library.caltech.edu Internet Source	1 %
8	d-nb.info Internet Source	1 %
9	docksci.com Internet Source	<1 %
10	iieta.org Internet Source	<1 %
11	www.mdpi.com Internet Source	<1 %
12	"Artificial Intelligence and Technologies", Springer Science and Business Media LLC, 2022 Publication	<1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		