COMPARATIVE STUDY OF SUPERVISED LEARNING ALGORITHMS ON RAINFALL PREDICTION USING NEX-GDDP-CMIP6 DATASET

Ratih Prasetya, Adhi Harmoko S, Nelly F Riama

OBJECTIVES

This study aimed to focus on comparation study on these 10 regression ML models to select the best model based on model testing and evaluations metrics. The exploratory data analysis also applied for better understanding on the dataset.

INTRODUCTION

- The uncertainty of precipitation prediction is complex problem, merely in predicting non-linear trend in precipitation data
- Previous studies have shown that precipitation extreme has increased in Indonesia and likely continue.
- One statistical method that is widely used for rainfall estimation involves machine learning.
- Recently, NASA Earth Exchange Global Daily Downscaled Projections (NEX-GDDP-CMIP6) dataset has been released in September 2022 and last updated in August 2023.
- There is need of improving the accuracy of rainfall estimation involving the recent global downscaled GCM in Tropics

DATA

Year 2014, EC-Earth Model from NEX-GDDP-CMIP6:

- precipitation (pr),
- mean temperature (tas),
- maximum temperature (tasmax),
- minimum temperature (tasmin)
- relative humidity (hurs)

METHODS

Methodologies for rainfall prediction based on ML Algorithms

- **Exploratory Data Analysis (EDA)**
- **Models Comparation** of 10 (ten) regression algorithms using 70:30 for train and test dataset, utilizing MSE, MAE and RMSE.

RESULTS:EDA

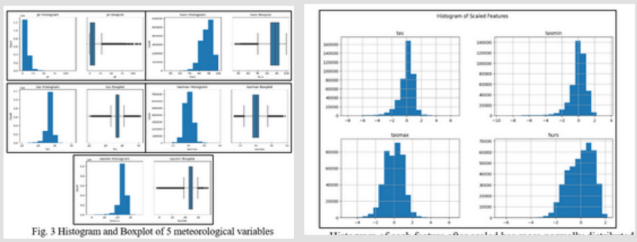
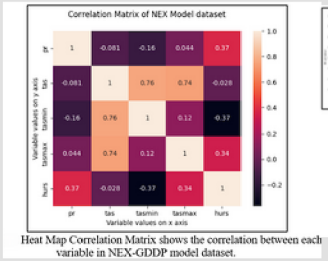


Fig. 3 Histogram and Boxplot of 5 meteorological variables



Heat Map Correlation Matrix shows the correlation between each variable in NEX-GDDP model dataset.

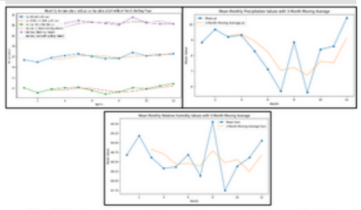


Fig. 5 Moving Average of temperature, precipitation, and relative humidity

RESULTS:VISUALIZATION

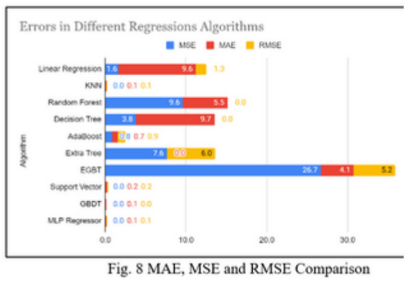


Fig. 8 MAE, MSE and RMSE Comparison

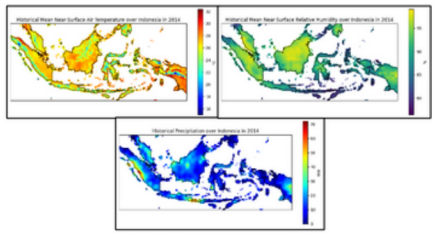


Fig. 6 Spatial Visualization of Temperature, Precipitation and Relative Humidity in NEX-GDDP Dataset

CONCLUSION

- The comparative study revealed that Gradient Boosting Decision Tree, KNN, MLP Regressor, and Support Vector Regressor performed exceptionally well, achieving nearly zero RMSE error values.
- Conversely, XGBoost, Linear Regression, and AdaBoost exhibited inferior performance in terms of RMSE.
- While most models demonstrated a strong fit with R-Squared values around 0.9, XGBoost lagged behind significantly with a score of only 0.49.
- In summary, based on metrics such as MAE, MSE, RMSE, and R-Squared, XGBoost emerged as the poorest performer in the training model.
- Future enhancements could employ surface observations data for validation purposes could enhance the reliability of machine learning predictions.

ratih.prasetya@bmkg.go.id