A Comparison of Bagging and Boosting on Classification Data: Case Study on Rainfall Data in Sultan Syarif Kasim II Meteorological Station in Pekanbaru

A Adnan¹, A M Yolanda² and F Natasya³

Statistics Study Program, Department of Mathematics, Universitas Riau
Jl. HR. Soebrantas, Km. 12.5, Pekanbaru, 28293, Indonesia
arisman.adnan@lecturer.unri.ac.id

Abstract. A frequent way for classification data is using a machine learning algorithm alongside ensemble methods like bagging and boosting. In earlier studies, these two algorithms have shown to be very accurate. The aim of this research is to discover performance of bagging and boosting to classify rainfall data obtained at the Sultan Syarif Kasim II Meteorological Station in Pekanbaru from 1 January 2018 until 31 July 2021. Rainfall data are classified into two categories: rainy and non-rainy. The parameters are average temperature, average humidity, sunshine duration, wind direction at maximum speed, and average wind speed. For comparison, this study developed Stochastic Gradient Boosting with Gradient Boosting Modelling and C5.0 from boosting, as well as Bagged Classification and Regression Tree (CART) and Random Forest from bagging. In order to generate reliable conclusions, each algorithm is run 30 times with repeated cross validation. The result demonstrates that Stochastic Gradient Boosting with Gradient Boosting Modelling is the best algorithm based on average accuracy.

1. Introduction
Finding a high-performing machine learning algorithm can take some time. This is due to the fact that applied machine learning is full of trial and error process. An ensemble method, which combines multiple different models, is one way to improve accuracy. Bagging and boosting are the two most used strategies for integrating predictions from different models. Building multiple models with the same type using distinct subsamples of the training dataset is known as bagging. Boosting is developing numerous models, usually of the same type, that successively learn to correct the prediction mistakes of the previous model in the loop [1].

For data classification, both bagging and boosting are highly recommended. Classification of a new observation with the ensemble method is done by combining algorithms from several ways. This method is often called committee-based learning or the multiple classifier system. In contrast to other classification approaches that only use one classification algorithm, the ensemble method builds a set of algorithms and combines the results to make decisions [2].

The use of the ensemble method helps us to obtain a more representative model than using only one algorithm. From this point of view, the ensemble method may enlarge the scope of effective hypotheses and expand the space of functions that can be represented. Another reason is
computational process, in the sense that the use of multiple algorithms increases the ability to provide optimal results. [3].

Based on the description above, a study was conducted to compare the accuracy of the two ensemble methods, bagging and boosting. Classification modelling is applied to daily rainfall data and its parameters. The results of this study are expected to be able to recommend a description of the classification accuracy using the boosting and bagging methods.

There are several preliminary studies related to boosting. According to Febriantono et al [4] that the C5.0 algorithm is performed better than C4.5 and ID3 algorithms in solving multiclass unbalanced data issues, with performance percentages of 40.91 percent, 40.24 percent, and 19.23 percent for C5.0, C4.5, and ID3, respectively. In another study, Osman et al [5] have used machine learning algorithm to predict groundwater levels in Selangor, Malaysia as a case study, and showed that the Extreme gradient boosting technique produced excellent predictions compared to Artificial Neural Network and Support Vector Regression on that case [5]. For bagging algorithm, Bagged Classification and Regression Tree produce high accuracy in some previous study. The accuracy of the CART algorithm with bagging approach in classification data periods is higher than the CART algorithm alone [6]. Besides, random forest also has good prediction model on classification. For example in classification of rice-plant growth phase using supervised random forest method, the accuracy is 70.91 percent for more than one temporal feature of vegetation index [7]. Due to the fact that bagging and boosting are strongly suggested, these algorithms are implemented as a new procedure to classifying rainfall data as rainy or non-rainy. The constructed models could be used by appropriate agencies as a prediction technique in the future.

2. Research Method

2.1. Bagging
Bagged Classification and Regression Tree and Random Forest are two popular machine learning algorithms in bagging method. Although it is a specific implementation, CART (Classification and Regression Trees) is sometimes used as a generic acronym for the decision tree. The Gini diversity index is used by CART to deal with continuous variables [8]. It creates a series of subtrees, calculates the cost of misclassification for each one using cross-validation, and selects the one with the lowest cost. An ensemble technique is created by combining this algorithm with bagging to increase accuracy, called Bagged Classification and Regression Tree.

Random forest is a sophisticated method for analysing many different classification trees. It is used to increase the overall accuracy of a singular classification tree technique by minimizing bias in a dataset from harming a single tree by spreading the bias across a multiple number of trees. This algorithm allows trees to vote on the final classification based on their accuracy score from validation data [9].

2.2. Boosting
The Boosting has many algorithms, including are C5.0 and Stochastic Gradient Boosting. In comparison to traditional boosting, stochastic gradient boosting appears to be a significant improvement. It integrates the iteration of very small trees, random sampling from the training data at each training cycle, slow learning via very small model releases through each training cycle, selective rejection of training data based on model residuals, and the ability to use a variety of objective functions to produce a system that has performed admirably in a real-world applications [10]. Stochastic Gradient Boosting is performed with Gradient Boosting Modelling (GBM). Gradient boosting is a method of fixing an infinite-dimensional optimization problem that generates a model in the form of linear combinations of decision trees [11].

The C5.0 algorithm is a decision tree-specific data mining algorithm, and it is a development over the ID3 and C4.5 algorithms [12]. Information gain will be used to process the attribute selection in this C5.0 algorithm. This is a classification algorithm that takes less time, consumes less memory, and
includes features such feature selection, cross validation, pruning error reduction, and model complexity [13]. The stochastic gradient boosting machines deviated from the adaboost, but C5.0 operates similarly. Weights are determined after the first tree is built, and further iterations result in weighted trees or rulesets. The size of subsequent trees (or rulesets) is limited to about the same as the first model. The final prediction is based on simple \( p \)-class averages. The final outcome is a simple average of the class probabilities generated by each tree or ruleset.

3. Data Description

Data used in this study have been collected from Meteorological, Climatological, and Geophysical Agency (BMKG Indonesia). Daily rainfall with some parameters were collected during the period January 1, 2018 and July 31, 2021 [14]. The parameters are average temperature, average humidity, sunshine duration, wind direction at maximum speed, and average wind speed. The analysis was carried out using R Program. The steps in the research process are as follows:

1. Collecting rainfall data and its parameter from website
2. Handling Missing Value
3. Data exploration
4. Setting up training data with 30 times the number of resamples and repeated cross validation.
5. Classification data using Stochastic Gradient Boosting with Gradient Boosting Modelling, C5.0, Bagged Classification and Regression Tree (CART), and Random Forest.
6. Comparing the average accuracy and kappa of each algorithm to find the best one.
7. Conclusion

4. Result and Discussion

Due to a lack of documentation or measurement, some missing values exist in the rainfall data and its parameters employed in this study. As a result, the pre-processing data is required to obtain reliable prediction findings. The interpolation is used to address missing values in this study.

| Observation          | Average Temperature | Average Humidity | Rainfall | Sunshine Duration | Wind direction at maximum speed | Average Wind Speed |
|----------------------|---------------------|------------------|----------|-------------------|-------------------------------|--------------------|
| 06-03-2018           | 27.7                | 82               | 77.3     | 0.50              | 340                           | 1                  |
| 07-03-2018           | 26.5                | 85               | 50.5     |                   | 340                           | 2                  |
| 08-03-2018           | 27.0                | 81               |          | 6.30              | 340                           | 1                  |
| 09-03-2018           | 28.0                | 77               | 0        | 5.50              | 50                            | 2                  |
| 10-03-2018           | 25.1                | 92               |          | 8.00              | 90                            | 2                  |
| 11-03-2018           | 27.6                | 78               | 2.80     |                   | 60                            | 1                  |
| 12-03-2018           | 26.6                | 85               | 13.2     |                   | 360                           | 2                  |
| 13-03-2018           | 26.3                | 85               | 0.20     | 1.60              | 50                            | 1                  |
| 14-03-2018           | 27.6                | 80               |          | 2.10              | 80                            | 1                  |

There appears to be some missing values (blank rows at each category) in Table 1, otherwise known as not applicable or not available (NA) data. To deal with it, interpolation is used to generate results that impute NA with real values by interpolation using approx from Package ‘imputeTS’ in R Program [15]. The interpolation results are shown in Table 2.

Replacement of missing values is also called imputation. Values from an approximation interpolation are used to fill in missing values. In second row, observation in sunshine duration is missing, value imputation by interpolation with 3.4. It is calculated by dividing the sum of 0.5 (first row) to 6.3 (third row) by 2. All remaining missing data is then processed using the same procedure.
Table 2. Rainfall data using interpolation for missing values

| Observation | Average Temperature | Average Humidity | Rainfall | Sunshine Duration | Wind direction at maximum speed | Average Wind Speed |
|-------------|---------------------|------------------|----------|-------------------|---------------------------------|-------------------|
| 06-03-2018  | 27.7                | 82               | 77.3     | 0.50              | 340                             | 1                 |
| 07-03-2018  | 26.5                | 85               | 50.5     | 3.40              | 340                             | 2                 |
| 08-03-2018  | 27.0                | 81               | 25.25    | 6.30              | 340                             | 1                 |
| 09-03-2018  | 28.0                | 77               | 0        | 5.50              | 50                              | 2                 |
| 10-03-2018  | 25.1                | 92               | 1.40     | 8.00              | 90                              | 2                 |
| 11-03-2018  | 27.6                | 78               | 2.80     | 5.80              | 60                              | 1                 |
| 12-03-2018  | 26.6                | 85               | 13.2     | 3.73              | 360                             | 2                 |
| 13-03-2018  | 26.3                | 85               | 0.20     | 1.60              | 50                              | 1                 |
| 14-03-2018  | 27.6                | 80               | 0.1      | 2.10              | 80                              | 1                 |

Although this imputation is not faultless, it can provide a comprehensive summary of rainfall data for further research. In earlier studies, when it came to substituting missing data, the linear interpolation method outperformed the mean method [16]. Furthermore, daily rainfall is a time series data, and the imputeTS package focuses in imputation of time series data. It is important to highlight that the lot of the imputation strategies described in the studies are case specific and perform effectively in response to variations. To find the optimal imputation technique to improve accuracy, a detailed analysis such as a data distribution analysis or a missing data pattern analysis would be needed [17].

The next phase is to classify the rainfall data once there are no more missing values. This study is part of research roadmap on accuracy of machine learning algorithm. In this part, the algorithm is executed on rainfall data with two classes: rainy and non-rainy. If the daily rainfall is equal to 0 millimeter, the data is classified as non-rainy, while the rest is labelled as rainy. There were 608 rainy days between January 1, 2018 and July 31, 2021. The time series plot average temperature of is presented in Figure 1.

![Time series plot of average temperature](image_url)

**Figure 1.** Time series plot of average temperature

The longest sunshine duration is about 9.8 hours. According to Figure 1, the average temperature in Pekanbaru is between 23 and 31°C, with a minimum of 23.4°C and a maximum of 30.50°C. This is
consistent with Pekanbaru's status as one of Indonesia's hottest cities. Another key component that impacts rainfall is average humidity which presented in Figure 2.

![Figure 2. Time series plot of average humidity](image)

The minimum and highest average humidity (see Figure 2) are 65 and 98 %, respectively. For wind direction at maximum speed, it reached 360° during this period time. Nevertheless, the maximum average wind speed in this area is 4 m/s. Average temperature, average humidity, sunshine duration, wind direction at maximum speed, and average wind speed have an impact on the heavy rain simultaneously (see Figure 3).

![Figure 3. Time series plot of daily rainfall](image)
The highest rainfall in Pekanbaru during this period is 131 mm. According to record from the Sultan Syarif Kasim II Meteorological Station, light rainfall (0.5 – 20 mm/day) occur frequently, followed by moderate rainfall (20 – 50 mm/day). Very heavy rainfall (100 – 150 mm/day) is rare in this region, as shown in Figure 3.

Rainy days are counted when the precipitation rate is equal to or more than 0.5; otherwise, it is classified as non-rainy. This dataset is used as a case study to determine which of two ensemble methods is the most effective. This research will go through boosting and bagging in detail, demonstrating how these machine learning methods can improve the accuracy of models on large datasets. The mlbench [18], caret [19], and caretEnsemble [20] packages are used to do this study. To achieve high accuracy, each algorithm is built 30 times (number of resamples) using repeated cross validation.

| Accuracy | Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
|----------|---------|--------------|--------|------|--------------|---------|
| C5.0     | 0.618   | 0.677        | 0.693  | 0.696| 0.714        | 0.771   |
| GBM      | 0.649   | 0.674        | 0.701  | 0.703| 0.725        | 0.779   |

Boosting algorithms explored in this study are C5.0 and Stochastic Gradient Boosting with Gradient Boosting Modelling (GBM). GBM method creates a more accurate model, as seen in Table 3, with an accuracy of 0.701 (average accuracy/mean). During this trial, the maximum accuracy is around 0.779 (also GBM). In this scenario, C5.0 is faster than GBM in terms of processing time. This supports the claim that the C5.0 method works pretty good but is considerably easier to understand and use [1]. Traditional Gradient boosting methods require screening all data instances for each feature to determine the information gain of all feasible split points. When dealing with large amounts of data, this makes these methods extremely time consuming [21]. Regardless, the technique is not difficult to apply and produces good results.

| Accuracy | Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
|----------|---------|--------------|--------|------|--------------|---------|
| Treebag  | 0.595   | 0.634        | 0.653  | 0.656| 0.672        | 0.738   |
| RF       | 0.626   | 0.649        | 0.670  | 0.675| 0.692        | 0.754   |

Bagged Classification and Regression Tree (Treebag) and Random Forest are two bagging algorithms built in this work (RF). Table 4 presents that random forest generates a model which is more accurate, with an accuracy of 0.675 (average accuracy). Table 5 also reveals that bagging algorithm achieves the highest accuracy of 0.754 using Random Forest. On the other hand, treebag is faster than random forest in terms of compute time. This is due to the Random Forest complexity, which is directly proportional to the number of trees, hence increases overall running times [22].

| Accuracy | Average Kappa |
|----------|---------------|
| C5.0     | 0.379         |
| GBM      | 0.395         |
| Treebag  | 0.290         |
| RF       | 0.339         |
Using rainfall as a case study, accuracy is still below 80% (0.8). However, prediction utilizing C5.0, Stochastic Gradient Boosting, Bagged CART, and Random Forest is quite good. For data with a huge number of missing values, the results are pretty respectable. Stochastic Gradient Boosting with Gradient Boosting Modelling is the best algorithm with an average accuracy of 0.703. Apart from accuracy, kappa can be used as a metric in terms of choosing the best algorithms. The highest kappa, 0.395, was also achieved using GBM, can be seen in Table 5.

5. Conclusion
Based on the average accuracy of all algorithms, there is adequate evidence to conclude that the Stochastic Gradient Boosting with Gradient Boosting Modelling is the best in this case study. This algorithm is also more applicable compared to others algorithms. In the future, researchers can use more advanced algorithms to combine multiple approaches, such as stacking algorithms.

References
[1] Yobero C 2018 Determining Creditworthiness for Loan Applications Using C5.0 Decision Trees RPubs by RStudio
[2] Zhou Z-H 2012 Ensemble Methods: Foundations and Algorithms (New York (US): Chapman & Hall/CRC Press)
[3] Valentini G and Re M 2001 Ensemble Methods: A review Data Mining and Machine Learning for Astronomical Applications (Florida: CRC Press)
[4] Febriantono M A, Pramono S H, Rahmadwati and Naghdy G 2020 Classification of multiclass imbalanced data using cost-sensitive decision tree C5.0 JAES Int. J. Artif. Intell. 9 65–72
[5] Osman A I A, Ahmed A N, Chow M F, Huang Y F and El-shafie A 2021 Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia Ain Shams Eng. J 12 1545–56
[6] Arrahimi A R, Ihsan M K, Kartini D, Faisal M R and Indriani F 2019 Teknik Bagging Dan Boosting Pada Algoritma CART Untuk Klasifikasi Masa Studi Mahasiswa J. Sains dan Inform 5 21–30
[7] Tosida E T, Andria F, Wahyudin I, Widianto R, Ganda M and Lathif R R 2021 A hybrid data mining model for Indonesian telematics SMEs empowerment IORA-ICOR2018 IOP Conf. Ser. Mater. Sci. Eng 567
[8] BMKG 2021 Rainfall Data Rainfall Data https://dataonline.bmkg.go.id/
[9] Moritz S 2021 Package ‘imputeTS’: Time Series Missing Value Imputation Version 3.2. Repository CRAN
[10] Noor N M, Abdullah M M A B, Yahaya A S and Azam N 2014 Comparison of Linear Interpolation Method and Mean Method to Replace the Missing Values in Environmental Data Set Mater. Sci. Forum
[17] Osman M S, Abu-Mahfouz A M and Page P 2018 A Survey on Data Imputation Techniques: Water Distribution System as a Use Case IEEE Access 6 63279–91

[18] Leisch F 2021 Package ‘ mlbench ’: Machine Learning Benchmark Problems Version 2.1-3 Repository CRAN

[19] Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, Mayer Z, Kenkel B, R Core Team, Benesty M, Lescarbeau R, Ziem A, Scrucca L, Tang Y, Candan C and HunT T 2021 Package ‘ caret ’: Classification and Regression Training Version 6.0-88 Repository CRAN

[20] Deane-mayer Z A and Knowles J E 2019 Package ‘ caretEnsemble ’: Ensembles of Caret Models Version 2.0.1 Repository CRAN

[21] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q and Liu T 2017 LightGBM: A Highly Efficient Gradient Boosting Decision Tree 31st Conference on Neural Information Processing Systems (NIPS 2017) (CA, USA: NIPS)

[22] Yates D and Islam M Z 2021 FastForest: Increasing Random Forest Processing Speed While Maintaining Accuracy Inf. Sci. (Ny) 557 130–52