PARAMETER
JURNAL MATEMATIKA, STATISTIKA DAN TERAPANNYA

e-ISSN. 2829-6303
Vol. 04, No. 3, December 2025
Page 497-508

Comparison of Random Forest and XGBoost Methods for Predicting Work Accident Claim Reserves

Sri Ayu Anugrah[1*], Sri Dewi Anugrawati[2], Adnan Sauddin[3], Andi Mariani[4]

[1,2,3,4]Department of Mathematics, Universitas Islam Negeri Alauddin Makassar
H.M. Yasin Limpo Street, No. 36 Samata, Gowa, 92113, South Sulawesi, Indonesia

***E-mail Correspondence Author:** *sriayuanugrah06@gmail.com* ✉

Abstract

The potential high claim burden in the work accident insurance sector managed by BPJS Ketenagakerjaan have an impact on the company's financial stability. This encourages insurance companies to provide additional funds to maintain the company's operational sustainability. Thus, preparing future fund reserves is a crucial step in risk and financial management to minimize payment delays, up to the risk of default. This study aims to determine the best method for predicting work accident claim reserves by comparing the Random Forest and XGBoost methods. The result of the analysis shows that the XGBoost method has an outstanding ability to predict work accident claim reserves on BPJS Ketenagakerjaan in the period July 2016 to August 2023, with a MAPE of 5.14% and an accuracy rate of 94.86%.

Keywords: *Claim reserves, random forest, severity claim prediction, work accident claim, XGBoost*

doi: https://doi.org/10.30598/parameterv4i3pp497-508

Research Article ***Open Acces***
Email: jurnalparameter@gmail.com
Homepage: https://ojs3.unpatti.ac.id/index.php/parameter

Submitted: June 2025
Accepted: December 2025

1. INTRODUCTION

Prediction of future conditions based on past experience is expected to provide prediction results with a fairly high accuracy. To get this, the right approach or method is needed. Machine learning is a unique approach that adopts the way humans think by constantly being given information. The more information obtained, the algorithm in machine learning will continue to make updates to decisions through resampling to get a conclusion that is closer to the actual situation.

Several previous studies have used machine learning approaches in predicting, including research conducted by [1] using the Random Forest regression method, [2] with SVM, KNN, and RF methods, [3] using the XGBoost regressor method, and research conducted by [4] using the GBDT, XGBoost, and LightGBM algorithms. From all of these studies, it was found that the use of Random Forest method when compared with machine learning methods such as SVM and KNN, Random Forest will produce a higher accuracy rate. Similarly, when the XGBoost method is compared with the GBDT and LightGBM methods, the XGBoost method produces a higher level of accuracy than the other methods. So researchers want to know which method is more effective in predicting work accident claim reserves.

In the work accident insurance sector as managed by BPJS Ketenagakerjaan, predicting the right claim reserves is very important to manage the potential high claim burden and maintain the company's operational sustainability. Some prior research in predicting claim risk has been done such as the use of the XGBoost method in predicting insurance claims has been conducted by [5]. This research compares several machine learning methods such as Logistic Regression, Random Forest, and XGBoost. The results of the analysis obtained that the XGBoost method provides the best performance with a higher level of accuracy than other methods. Furthere, for the comparison between Random Forest and XGBoost methods has also been studied by [6] to predict employee turnover. Although there are small differences in the prediction results, both methods show strong potential in addressing employee turnover prediction. Furthermore, related to claim reserves research has been conducted by [7]. The results of the analysis show the importance of estimating the amount of claims reserves and considering unexpected claims in an effort to deal with any risks that may occur in the future.

In addition to introduce the use of actual operational data from BPJS Ketenagakerjaan—which has never been empirically examined in previous studies on claim reserving—this research also provides a methodological contribution. The study adapts Random Forest and XGBoost to the unique statistical characteristics of occupational accident insurance claims in Indonesia, which are highly skewed, heavy-tailed, and administratively structured. Several model adjustments, including tailored hyperparameter tuning, feature engineering based on actuarial principles, and a reserving-oriented evaluation framework, offer methodological insights that have not been discussed in earlier literature. Therefore, the contribution of this study lies not only in applying machine learning to a new data context, but also in demonstrating how these algorithms behave, adjust, and perform under real social-insurance conditions, thereby extending methodological understanding of ML-based reserving models.

2. RESEARCH METHODOLOGY

This research is applied research in the form of applying machine learning theory [8], [9] in forecasting claim reserves for work accident insurance. The statistical methods and machine learning algorithms involved in the data analysis process are as follows:

2.1 Data Pre-processing

Pre-processing is the initial stage performed before analyzing data by first identifying outliers to produce high-quality and relieble analysis [9], [10]. A data can be said to be outliers, if the observation value is smaller than $Q_1 - 1.5 \times IQR$ or larger than $Q_3 + 1.5 \times IQR$ [11].

2.2 Logarithmic Transformation

Data transformation is a technique used to change the values of variables in data analysis, one such technique is logarithmic transformation which is often used in statistical analysis, including regression [12].

$$Y = log(X) \tag{1}$$

This logarithmic transformation is used for positive data with a very wide range of values. In other words, the logarithm transformation cannot be directly applied to 0 or negative values and needs adjustment for data that has values less than 10 [13].

2.3 Random Forest

Random Forest is a supervised learning algorithm that consists of a collection of structured decision trees. In the case of regression, the final prediction of the Random Forest is obtained by calculating the average prediction of all the decision trees built.

$$\hat{y}_i = \frac{1}{N_{tree}} \sum_{k=1}^{N_{tree}} \hat{y}_k \tag{2}$$

$\hat{y}_i$ is defined as the prediction result, N_{tree} is the total tree in the Random Forest, and $\hat{y}_k$ is the prediction result at the k-th tree [14].

The model's ability to generalize to new data is measured using the mean-squared generalization error: $E_{X,Y}\left(Y - h(X)\right)^2$, which reflects the model's performance on unseen data [15]. In addition, during the tree building process, a random selection of features will be made to increase the diversity between trees in the Random Forest starting by selecting the variable m from a number of independent variables p, provided that $m \leq p$. In the case of regression [16], the number of randomly selected features in each split $(mtry)$ is calculated using the following Equation (3).

$$mtry = \frac{p}{3} \tag{3}$$

2.4 XGBoost

XGBoost is a gradient boosting-based algorithm that efficiently builds tree models. In regression, each tree predicts continuous values by predicting the target variable and calculating the residuals from the previous trees [17].

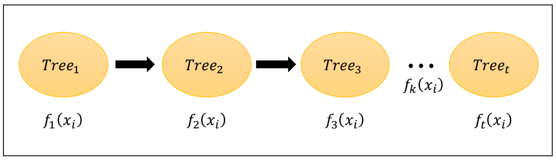

Figure 1. XGBoost Illustration

Source: [18]

From Figure 1, it is shown that $f_t(x_i)$ represents a tree model where the predicted value at t is modeled as $\hat{y}_i^{(t)}$ with the following Equation (4).

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) \tag{4}$$

The objective function has two main components, namely the loss function and the regularization value. In general, the objective function can be formulated as follows.

$$Obj^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{5}$$

with:

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

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

In XGBoost, the model complexity can be defined by the number of leaves (T) and weights (w) of each tree. Regularization is set via the function $\Omega(f_t) = \gamma T + 1/2\, \lambda \sum_{j=1}^{T} w_j^2$, where γ and λ are parameters controlling the model complexity. These parameters allow setting a balance between the model's ability to capture data patterns and the risk of overfitting. In general, **Equation (5)** can be rewritten into a simpler form namely:

$$Obj^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2 \tag{6}$$

In the **Equation (6)**, g_i is the gradient or first derivative of the loss function on the i-th data, while h_i is the Hessian value or second derivative of the loss function on the i-th data. The parameter T indicates the number of leaf nodes in the decision tree, w_j is the weight associated with the j-th leaf node, while γ and λ are regularization parameters that control the complexity of the model.

Furthermore, by grouping the samples by leaf nodes, the function can be rewritten in a form that involves the contribution of each node. If $I_j = \{i | q(x_i) = j\}$ is the set of samples that belong to the j-th leaf node, then the **Equation (7)** can be restructured into:

$$Obj^{(t)} = \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T \tag{7}$$

In this case, the sum of the gradients at each leaf node is expressed as $G_j = \sum_{i \in I_j} g_i$, while the total hessian value at that node is $H_j = \sum_{i \in I_j} h_i$. Based on this definition, the objective function can further be:

$$Obj^{(t)} \approx \sum_{j=1}^{T} \left[G_j w_j + \frac{1}{2} (H_j + \lambda) w_j^2 \right] + \gamma T \tag{8}$$

Once the decision tree structure is determined, the optimal solution for the leaf node weights w_j is obtained by performing a direct derivative to the objective function, where:

$$w_j^* = -\frac{G_j}{H_j + \lambda}$$

With G_j as the sum of gradients and H_j as the sum of hessians at the j-th node. If the solution is substituted into **Equation (8)**, the optimal value of the objective function becomes:

$$Obj^* = -\frac{1}{2} \sum_{j=1}^{T} \left(\frac{G_j^2}{H_j + \lambda} \right) + \gamma T \tag{9}$$

Next, XGBoost will calculate the gain value to determine the most optimal feature to use as the separation point. The gain value is calculated using the following **Equation (10)**:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma \tag{10}$$

Based on the above **Equation (10)**, G_L and H_L represent the sum of gradient and hessian in the left subtree, respectively, while G_R and H_R are the gradient and hessian in the right subtree. The values $\frac{G_L^2}{H_L+\lambda}$ and $\frac{G_R^2}{H_R+\lambda}$ indicate the structure scores for the left and right subtree respectively. Meanwhile, the value of $\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}$ represents the structure score when the node is not split (remains intact). In the node splitting process, the highest gain value will be selected to determine the best feature and the most optimal splitting point. This process is repeated for all features until the tree structure is optimally formed **[18]**, **[19]**.

The XGBoost model builds the predicted value of ($\hat{y}_i$) through an incremental boosting approach. The model can be expressed with the following **Equation (11)**:

$$\hat{y}_i = \hat{y}_0 + \eta \sum_{k=1}^{K} f_k(x_i) \tag{11}$$

Based on the above **Equation (11)**, $\hat{y}_0$ is the base prediction. Each $f_k(x_i)$ represents the k-th decision tree, where K is the total number of trees (boosters) built during iteration. Parameter η is the learning rate that controls the contribution of each decision tree f_k in the process of updating the model **[20]**.

2.5 Model Performance Evaluation

2.5.1 Mean Squared Error (MSE)

Mathematically, MSE is calculated by dividing the total sum of the squares of the differences between each predicted and true value by the number of observations n in the dataset. The smaller the value, the better **[21]**, **[22]**.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{12}$$

2.5.2 Coefficient of Determination (R^2)

The coefficient of determination or R^2 measures how well the regression model approximates the actual data **[23]**. In **Equation (13).** The value of R^2 ranges from 0 to 1, where values close to 1 indicate a high level of fit between the model and the empirical data. This reflects that the model effectively explains the variability in the data **[24]**.

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{13}$$

2.5.3 Mean Absolute Percentage Error (MAPE)

Mean Absolute Percentage Error (MAPE) is considered a loss function that measures the error rate in the model evaluation results, thus providing a perspective on how far away the model's predicted value is from its actual value.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{14}$$

$$\text{Accuracy} = 100\% - MAPE \tag{15}$$

Using MAPE, the accuracy of the model can be estimated in terms of the difference between the actual value and the estimated value, expressed as a percentage **[25]**.

Table 1. **Interpretation of MAPE Value**

MAPE (%)	Interpretation
$< 10\%$	Very Good
$10 - 20\%$	Good
$20 - 50\%$	Good Enough
$> 50\%$	Bad

Source: [26]

2.5.4 Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) is calculated by measuring the difference between the actual value and the predicted value, then squaring the difference to avoid negative influences. Next, the sum of the squares of this difference is divided by the total amount of data, then the square root is taken to return the results to the same scale as the original data [27].

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n}} \tag{16}$$

2.5.5 Mean Absolute Deviation (MAD)

Mathematically, MAD uses the absolute value of each difference between the actual and predicted values, thus ensuring that errors are accounted for fairly regardless of the direction of the deviation [28],

$$MAD = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \tag{17}$$

2.5.6 Insurance Claims Reserve

In insurance, a claim is an official demand by a customer to request funds in accordance with the initial contract. Therefore, companies must set aside a certain amount of funds to cover future claims [29].

2.6 Analysis Steps

The steps taken in this research are as follows:

1. Determine descriptive statistics.
2. Performing data pre-processing, such as handling outliers.
3. Perform data transformation.
4. Divide the training data and testing data based on 4 proportions, namely 65:35, 70:30, 75:25, and 80:20. Then, Select training data and testing data based on the highest R^2 value.
5. Calculating the predicted value.
   a. Random Forest
   The steps in Random Forest analysis are as follows:
      1) Determined 10 parameter combinations for the number of estimates,mtry nodesize, and maxnodes.
      2) Build a Random Forest model using the training data.
      3) Determine the best model based on the smallest MSE, RMSE, MAPE, and MAD.
      4) Make predictions using testing data.
   b. XGBoost
   The steps in the XGBoost analysis are as follows:
      1) Determined 10 parameter combinations for the number of estimates, eta, max_depth, gamma, and lambda.
      2) Build XGBoost model using training data.

3) Determine the best model based on the smallest MSE, RMSE, MAPE, and MAD.
4) Make predictions using testing data.

6. Calculate MAPE to compare the accuracy of Random Forest and XGBoost methods based on Equation (14) and Equation (15).

3. RESULTS AND DISCUSSION

The data used in this study are data on work accident insurance claims obtained from the BPJS Ketenagakerjaan Makassar office in the period July 2016 to August 2023 with a total data of 1,177. The data was analyzed using the Random Forest and XGBoost methods to predict work accident claim reserves.

3.1 Descriptive Statistics

An overview of the data can be seen based on its descriptive statistics. Descriptive analysis of the data showed that each variable indicated extreme values with a very wide range of data. Thus, at the data pre-processing stage, outliers were checked for each variable.

Table 2. Descriptive Statistics

Statistical Metrics	Research Variables			
	Claim Amount (IDR)	Claim Frequency	Claim Ratio (IDR)	Maximum Claim (IDR)
Minimum	68,530	1	68,530	68,530
Q1	1,340,080	1	915,872	1,055,960
Median	11,379,841	2	4,596,066	8,891,560
Mean	55,767,052	3	17,795,889	40,740,726
Q3	52,083,738	4	15,443,610	37,187,004
Maximum	6,999,720,285	29	2,333,240,095	6,954,972,736
IQR	50,743,658	3	14,527,738	36,131,044
SD	222,372,132	4	75,457,625	211,626,774

3.2 Data Pre-processing

Checking and handling outliers is done to handle data that has an asymmetrical distribution. From the dataset, the analysis results show that each research variable has outliers.

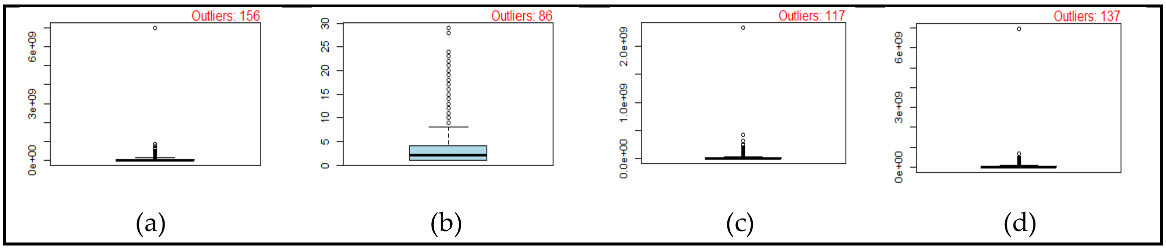

Figure 2. Outliers on Research Variables: (a) Claim Amount, (b) Claim Frequency, (c) Claim Ratio, and (d) Maximum Claim

Figure 2 shows the outlier values for each research variable before handling. In Figure 2(a), the claim amount variable has 156 outliers, which is the total amount of funds paid by the company to workers who filed claims at a certain time. Figure 2(b) shows the claim frequency variable with 86 outliers, which is the number of claims reported to the company every day. Furthermore, Figure 2(c) shows the claim ratio variable, which has 117 outliers, representing the average claim amount for each claim that occurs. Finally, Figure 2(d) illustrates the maximum claim variable with 137 outliers, which is the largest amount paid by the company for a single claim in a day.

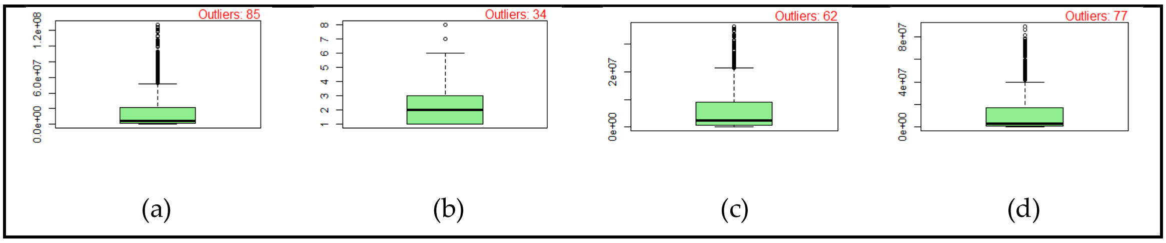

Figure 3. Results of Handling Outliers with IQR on Variables: (a) Claim Amount, (b) Claim Frequency, (c) Claim Ratio, and (d) Maximum Claim

Figure 3 shows the results of reducing the number of outliers using the IQR method. In Figure 3(a), the claim amount was reduced to 85 outliers indicating an improvement in the distribution of total claim funds paid. Figure 3(b) shows that the claim frequency has 34, so that the number of claims reported is well distributed. Figure 3(c), the claim ratio has been minimized to 62 outliers indicating a more representative average claim amount for each claim. Finally, in Figure 3(d) maximum claim had 77 outliers indicating a better distribution of claim funds. As a results, the total data which was initially 1,177 was reduced to 950 after removing the outliers.

3.3 Logarithmic Transformation

An enormous range in the variables of claim amount, claim ratio, and maximum claim causes imbalance in the data. These three variables have a much larger scale than the claim frequency variable. Therefore, a logarithmic transformation was performed to narrow the data range to be comparable to the claim frequency variable.

3.4 Division of Training and Testing Data

In evaluating the effect of data division variations on model performance, training and testing data were divided. At this stage, the training and testing data is divided into 4 different proportions, namely 65:35, 70:30, 75:25, and 80:20 which are then evaluated based on the highest R^2 value. The training data obtained is used to train models in Random Forest and XGBoost, while the testing data is used to predict work accident claim reserves.

Table 4. Devision

Proportion	Training	Testing	Random Forest (R^2)	XGBoost (R^2)
65:35	617	333	0.9526	0.9894
70:30	665	285	0.9405	0.9902
75:25	712	238	0.9480	0.9900
80:20	**760**	**190**	**0.9590**	**0.9940**

3.5 Prediction Using Random Forest

Before predicting using Random Forest, 10 combinations of parameters are initialized to build a model using training data, as shown in the following table.

Table 5. **Random Forest Parameter Initialization**

Combination of Parameters	Estimated Amount	Mtry	Nodesize	Maxnodes
1	100	1	5	30
2	200	1	10	50
3	300	1	15	70
4	400	1	5	50
5	500	1	10	30
6	100	1	15	50
7	200	1	5	70
8	30 0	1	10	30
9	400	1	15	50
10	500	1	5	70

Evaluation metrics such as MSE, RMSE, MAPE, and MAD are applied to determine the best parameters used to predict claim reserves. Each evaluation metric is calculated based on residuals, which is the difference between the actual and predicted values. The residuals form the basis for calculating the four evaluation metrics for each iteration, where the values obtained are used to assess the overall performance of the model. Models with smaller metric values show better performance.

Table 6. **Model Evaluation Results with Random Forest**

Combination of Parameters	MSE	RMSE	MAPE	MAD
10	0.023125	0.152069	0.733082%	0.109426

The analysis results show that the 10th parameter combination is the best parameter for predicting claim reserves with an actual value of IDR 2,561,770,331 while the predicted value is IDR 2,327,850,389.

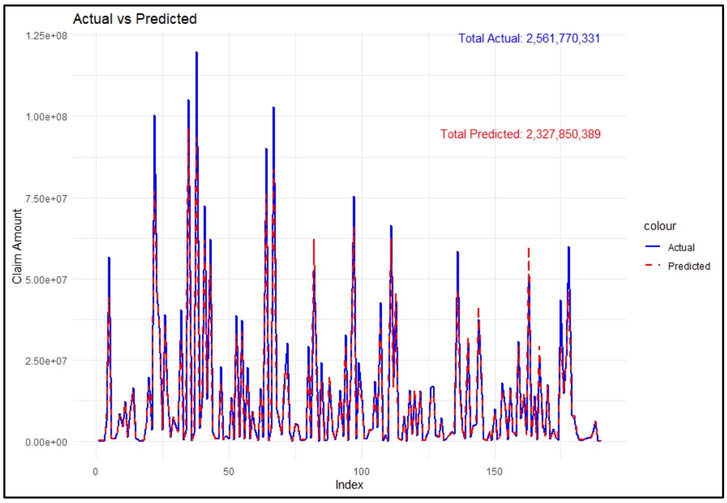

Figure 4. **Comparison of Actual and Predicted Values in Random Forest**

3.6 Prediction Using XGBoost

Before making predictions using XGBoost, 10 parameter combinations are initialized to build a model using the training data, as shown in the following table.

Table 7. **XGBoost Parameter Initialization**

Combination of Parameters	Estimated Amount	Eta	Max_depth	Gamma	Lambda
1	100	0.01	3	0	1
2	200	0.1	6	0.1	2
3	300	0.3	9	0.3	5
4	400	0.01	6	0.1	1
5	500	0.1	3	0	2
6	100	0.3	9	0.3	5
7	200	0.01	6	0	1
8	300	0.1	3	0.1	2
9	400	0.3	9	0.1	5
10	500	0.01	6	0	2

Evaluation metrics such as MSE, RMSE, MAPE, and MAD are applied to determine the best parameters used to predict claim reserves. These metrics are calculated based on residuals, which is the difference between the actual and predicted values in both training and testing data. Models with smaller metric values performed better.

Table 8. **Model Evaluation Results with XGBoost**

Model Evaluaion	MSE	RMSE	MAPE	MAD
Training	0.000749	0.027365	0.127100%	0.019388
Testing	0.005395	0.073451	0.335230%	0.051085

Based on the analysis results, the 5th parameter is the best parameter for predicting claim reserves in both training and testing data with an actual value of IDR 2,561,770,331 while the predicted value is IDR 2,549,143,406.

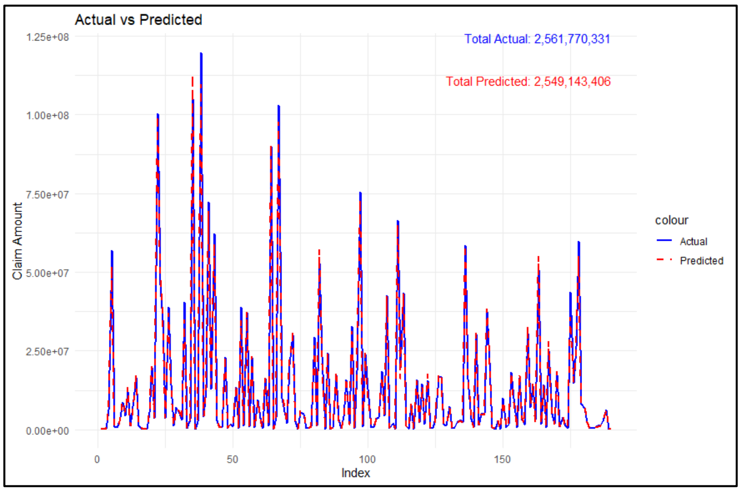

Figure 5. **Comparison of Actual and Predicted Values on XGBoost**

3.7 Calculate MAPE to Compare the Accuracy of Random Forest and XGBoost Methods

The final stage is to calculate the prediction accuracy using the Mean Absolute Percentage Error (MAPE) value to provide a representative picture of how far the model prediction is from the actual value.

Table 9. **Method Comparison Results**

Methods	MAPE (%)	Accuracy (%)
Random Forest	13.39 %	86.61 %
XGBoost	5.14 %	94.86 %

The analysis results obtained a MAPE value for Random Forest of 13.39% while XGBoost is 5.14% with an accuracy of 86.61% for Random Forest and 94.86% for XGBoost, respectively. This shows that the XGBoost method has excellent capabilities in predicting claim reserves when compared to the Random Forest. The results of the study are also consistent with research conducted by [17] which predicted the selling price of cayenne pepper based on daily climate, where XGBoost was also the best method with a MAPE of 9.96% when compared to the K-Nearest Neighbor (KNN) and Random Forest.

4. CONCLUSIONS

Based on the results of the analysis and discussion, it can be concluded that the pattern of claims between actual and predicted values shows a good level of similarity, although there are small differences at some data points. However, the results of the comparison of the two methods obtained an accuracy rate for Random Forest of 86.61% while XGBoost amounted to 94.86%. This indicates that the XGBoost method is better at predicting claim reserves than the Random Forest method. Further research should consider using other algorithms, such as advenced ensembles or deep learning, to improve the accuracy and robustness of the model from the current results. In addition of more up-to-date claim data can strengthen the performance and generalization capabilities of the model. Exploration of different feature engineering strategies can also provide a better understanding of claim reserve behavior.

Acknowledgments

The author would like to thank BPJS Ketenagakerjaan Makassar for its support and for providing the data used in this study.

Funding Information

Authors state no funding involved.

Author Contributions Statement

Sri Ayu Anugrah: Conceptualization, methodology, data analysis, formal analysis, investigation, data curation, writing-original draft, visualization. Sri Dewi Anugrawati: Supervision, methodology, data analysis, validation, writing-review & editing. Adnan Sauddin: Supervision, methodology, data analysis, validation, writing-review & editing. Andi Mariani: Validation, writing-review & editing. All authors discussed the results and contributed to the final manuscript.

Conflict Of Interest Statement

Authors state no conflict of interest.

Data Availability

This study uses data from BPJS Ketenagakerjaan Makassar, which is not publicly available. Relevant processed data can be obtained from the corresponding author upon reasonable request.

REFERENCES

[1] G. Iannace, G. Ciaburro, and A. Trematerra, “Wind Turbine Noise Prediction Using Random Forest Regression,” *Machines*, vol. 7, no. 4, p. 69, Nov. 2019, doi: 10.3390/machines7040069.

[2] S. Adi and Atik Wintarti, “Komparasi Metode Support Vector Machine (SVM), K-Nearest Neighbors (KNN), dan Random Forest (RF) untuk Prediksi Penyakit Gagal Jantung,” *MATHunesa J. Ilm. Mat.*, vol. 10, no. 2, pp. 258–268, July 2022, doi: 10.26740/mathunesa.v10n2.p258-268.

[3] S. Jeganathan, A. R. Lakshminarayanan, N. Ramachandran, and G. B. Tunze, "Predicting Academic Performance of Immigrant Students Using XGBoost Regressor," *Int. J. Inf. Technol. Web Eng.*, vol. 17, no. 1, pp. 1–19, June 2022, doi: 10.4018/IJITWE.304052.

[4] W. Liang, S. Luo, G. Zhao, and H. Wu, "Predicting Hard Rock Pillar Stability Using GBDT, XGBoost, and LightGBM Algorithms," *Mathematics*, vol. 8, no. 5, p. 765, May 2020, doi: 10.3390/math8050765.

[5] A. Hermawan, N. R. Jayanti, A. P. Rahmadian, M. H. Bayhaqi, A. Afdhal, and K. Aurelia, "Prediksi Klaim Asuransi Perjalanan Menggunakan Machine Learning untuk Optimasi Manajemen Risiko," *SABER J. Tek. Inform. Sains Dan Ilmu Komun.*, vol. 3, no. 2, pp. 09–20, 2025, doi: 10.59841/saber.v3i2.2476.

[6] M. B. Setiawan and A. Rahmatulloh, "Analisis Perbandingan Model Random Forest dan XGBoost dalam Memprediksi Turnover Karyawan," *Just IT J. Sist. Inf. Teknol. Inf. Dan Komput.*, vol. 15, no. 2, pp. 393–400, 2025, doi: 10.24853/justit.15.2.393%2520–%2520400.

[7] A. M. R. Yoisangaji, S. M. Pelu, and J. Wijaya, "ESTIMATION OF CLAIM RESERVES USING THE CHAIN LADDER METHOD," *BAREKENG J. Ilmu Mat. Dan Terap.*, vol. 18, no. 4, pp. 2083–2092, 2024, doi: 10.30598/barekengvol18iss4pp2083-2092.

[8] I. Cholissodin, Sutrisno, A. A. Soebroto, U. Hasanah, and Y. I. Febiola, *AI, Machine Learning & Deep Learning (Teori & Implementasi)*. Malang: FILKOM UB, 2020.

[9] Z. Setiawan *et al.*, *Buku Ajar Data Mining*. PT. Sonpedia Publishing Indonesia, 2023.

[10] F. Elfaladonna, I. G. T. Isa, D. Sartika, Yusniarti, and A. M. Putra, *Buku Ajar Dasar Exploratory Data Analysis (EDA)*. Bojong: PT Nasya Expanding Management, 2024.

[11] P. R. Sihombing, Suryadiningrat, D. A. Sunarjo, and Y. P. A. C. Yuda, "Identifikasi Data Outlier (Pencilan) dan Kenormalan Data Pada Data Univariat serta Alternatif Penyelesaiannya," *J. Ekon. Dan Stat. Indones.*, vol. 2, no. 3, pp. 307–316, 2022, doi: 10.11594/jesi.02.03.07.

[12] P. W. Rahayu *et al.*, *Buku Ajar Data Mining*. Jambi: PT. Sonpedia Publishing Indonesia, 2024.

[13] H. Sudarwati and M. H. Natsir, *Statistika dan Rancangan Percobaan (Penerapan dalam Bidang Peternakan)*. UB Press, 2019.

[14] E. S. Lestari and I. Astuti, "Penerapan Random Forest Regression Untuk Memprediksi Harga Jual Rumah Dan Cosine Similarity Untuk Rekomendasi Rumah Pada Provinsi Jawa Barat," *J. Ilm. FIFO*, vol. 14, no. 2, p. 131, Nov. 2022, doi: 10.22441/fifo.2022.v14i2.003.

[15] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[16] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. Stanford, California: Springer, 2008.

[17] M. Ardian, S. Khomsah, and R. Pandiya, "Perbandingan Model Regresi Untuk Memprediksi Harga Jual Cabai Rawit Berdasarkan Iklim Harian," *J. Jupit.*, vol. 16, no. 2, pp. 549–560, 2024, doi: 10.5281/zenodo.13208156.

[18] A. C. Nugraha and M. I. Irawan, "Komparasi Deteksi Kecurangan pada Data Klaim Asuransi Pelayanan Kesehatan Menggunakan Metode Support Vector Machine (SVM) dan Extreme Gradient Boosting (XGBoost)," *J. Sains Dan Seni ITS*, vol. 12, no. 1, pp. A40–A46, May 2023, doi: 10.12962/j23373520.v12i1.107032.

[19] P. Zhang, Y. Jia, and Y. Shang, "Research and Application of XGBoost in Imbalanced Data," *Int. J. Distrib. Sens. Netw.*, vol. 18, no. 6, pp. 1–10, 2022, doi: 10.1177/15501329221106935.

[20] M. Bowers, "XGBoost Explained," Random Realizations. [Online]. Available: https://randomrealizations.com/posts/xgboost-explained/

[21] L. R. Amalia, W. Ramdhan, and W. M. Kifti, "Penerapan Metode Trend Moment Untuk Memprediksi Jumlah Pertumbuhan Penduduk," *Build. Inform. Technol. Sci. BITS*, vol. 3, no. 4, pp. 566–573, 2023, doi: 10.47065/bits.v3i4.1396.

[22] H. Budiman, "Analisis Dan Perbandingan Akurasi Model Prediksi Rentet Waktu Support Vector Machines Dengan Support Vector Machines Particle Swarm Optimization Untuk Arus Lalu Lintas Jangka Pendek," *Syst. Inf. Syst. Inform. J.*, vol. 2, no. 1, pp. 19–24, 2016, doi: 10.29080/systemic.v2i1.103.

[23] B. I. Sanny and R. K. Dewi, "Pengaruh Net Interest Margin (NIM) Terhadap Return on Asset (ROA) Pada PT Bank Pembangunan Daerah Jawa Barat Dan Banten Tbk Periode 2013-2017," *J. E-Bis Ekon.-Bisnis*, vol. 4, no. 1, pp. 78–87, 2020, doi: 10.37339/e-bis.v4i1.239.

[24] D. Siagian and Sugiarto, *Metode Statistika Untuk Bisnis dan Ekonomi*. Jakarta: PT Gramedia Pustaka Utama, 2000.

[25] S. Mulani, "MAPE – Mean Absolute Percentage Error in Python," AskPython. [Online]. Available: https://www.askpython.com/python/examples/mape-mean-absolute-percentage-error

[26] I. Nurvianti, B. D. Setiawan, and Fitra Abdurrachman Bachtiar, "Perbandingan Peramalan Jumlah Penumpang Keberangkatan Kereta Api di DKI Jakarta Menggunakan Metode Double Exponential Smoothing dan Triple Exponential Smoothing," *J. Pengemb. Teknol. Inf. Dan Ilmu Komput.*, vol. 3, no. 6, pp. 5257–5263, 2019.

[27] H. Nurfaidah and W. Abidin, “Penerapan Metode Single Moving Average Dalam Peramalan Curah Hujan Kota Makassar,” *J. MSA Mat. Dan Stat. Serta Apl.*, vol. 11, no. 2, pp. 134–139, 2023, doi: 10.24252/msa.v11i2.45815.

[28] A. E. Armi, A. H. Kridalaksana, and Z. Arifin, “Peramalan Angka Inflasi Kota Samarinda Menggunakan Metode Double Exponential Smoothing (Studi Kasus : Badan Pusat Statistik Kota Samarinda),” *Inform. Mulawarman J. Ilm. Ilmu Komput.*, vol. 14, no. 1, pp. 21–26, 2019, doi: 10.30872/jim.v14i1.1252.

[29] Y. Hikmah and I. R. Hikmah, “Perhitungan Cadangan Klaim dengan Metode Chain Ladder Menggunakan Excel dan RStudio,” *MAp Math. Appl. J.*, vol. 4, no. 2, pp. 122–131, Dec. 2022, doi: doi.org/10.15548/map.v4i2.4837.