Current Science Research Bulletin

e-ISSN: 3050-7103
p-ISSN: 3050-709X

Volume 03 (05), May 2026, pp. 151-157

CrossRef DOI: 10.55677/csrb/08-V03I05Y2026

Available on: https://csrbjournal.org/

Classification of Rural Bank Liquidation Duration Using Random Forest with Hyperparameter Optimization

Fathya Maulana Aziz[1], Triastuti Wuryandari[2*], Rahmila Dapa[3]
[1,2,3] Departement of Statistics, Faculty of Science and Mathematics, Universitas Diponegoro, Semarang, Indonesia

ABSTRACT: Bank liquidation is a process of resolving the obligations of banks whose licenses have been revoked and plays a crucial role in maintaining financial system stability. The duration of Rural Bank liquidation varies depending on financial conditions and asset recovery performance, requiring robust analytical methods for accurate classification. This study aims to classify the liquidation duration of Rural Banks using the Random Forest algorithm with class imbalance handling through the Synthetic Minority Oversampling Technique (SMOTE) and hyperparameter optimization using GridSearchCV. The dataset consists of secondary data of Rural Bank liquidations handled by the Indonesia Deposit Insurance Corporation during the period 2006–2024, comprising 120 observations and 22 financial variables. Data preprocessing includes median imputation for missing values, winsorizing for outlier treatment, and data standardization. The results indicate that the best-performing model is Random Forest with SMOTE and hyperparameter tuning, achieving an accuracy of 83.3%, precision of 75.0%, recall of 93.8%, and F1-score of 83.3% for the problematic class. Feature importance and SHapley Additive exPlanations (SHAP) analysis reveal that Asset Liquidation Estimate, Total Gross Assets, and Guarantee Claim Value are the most influential variables. These findings provide important implications for data driven decision making in bank resolution policies.

KEYWORDS: Liquidation Duration, Random Forest, SMOTE, Hyperparameter Tuning, Feature Importance, SHAP.

1. INTRODUCTION

Rural Banks play an important role in providing financial services to small and medium communities, particularly in rural areas. However, these institutions are relatively vulnerable due to their limited scale of operations and capital structure. When a Rural Bank experiences irrecoverable financial distress, license revocation and liquidation become necessary resolution mechanisms, with the liquidation process in Indonesia being carried out by the Indonesia Deposit Insurance Corporation in accordance with Law Number 24 of 2004. In practice, the duration of Rural Bank liquidation varies significantly, reflecting differences in financial conditions, asset-liability structures, asset realization effectiveness, and recovery rates. With the advancement of technology and the availability of historical data, machine learning methods particularly Random Forest as a tree-based ensemble approach are considered capable of capturing nonlinear relationships and complex interactions among financial variables. Although previous studies have explored bank failure prediction using such approaches, research specifically analyzing the determinants of Rural Bank liquidation duration in Indonesia remains limited.

Therefore, this study aims to evaluate the performance of the Random Forest algorithm in classifying Rural Bank liquidation duration before and after the application of the Synthetic Minority Oversampling Technique (SMOTE), assess performance improvement after hyperparameter tuning using GridSearchCV, and identify dominant influencing variables based on feature importance analysis in the best-performing model. The study is limited to secondary data of Rural Banks whose licenses were revoked and whose liquidation processes were completed by Deposit Insurance Corporation from January 2006 to December 2024, focusing on variables relevant to the liquidation process and employing Random Forest classification as the main analytical approach.

Cite the Article: *Aziz, F.M., Wuryandari, T., Dapa, R. (2026). Classification of Rural Bank Liquidation Duration Using Random Forest with Hyperparameter Optimization. Current Science Research Bulletin, 3(5), 151-157.* https://doi.org/10.55677/csrb/08-V03I05Y2026

Publication Date: *May 30, 2026*

2. LITERATURE REVIEW

The liquidation duration of Rural Bank refers to the period from license revocation to the completion of asset settlement and liability resolution. In Indonesia, this process is regulated under Government Regulation Number 25 of 1999 and strengthened by Law Number 24 of 2004. The liquidation process must be completed within a maximum of two years, with possible extensions of up to two additional one-year periods, resulting in a maximum duration of four years.

Random Forest is an ensemble learning method that combines multiple decision trees to improve prediction accuracy and stability (Breiman, 2001). It operates using the bagging principle, where multiple models are built from bootstrap samples and aggregated to produce the final prediction (Hastie et al., 2009). The general steps include bootstrap sampling, random feature selection, tree construction, and aggregation of predictions.

SMOTE is an oversampling technique that generates synthetic minority class samples through interpolation between nearest neighbors (Chawla et al., 2002), defined as:

$$\mathrm{x}_{\mathrm{new}} = \mathrm{x}_{\mathrm{i}} + \delta \cdot (\mathrm{x}_{\mathrm{zi}} - \mathrm{x}_{\mathrm{i}}) \quad (1)$$

where:

$\mathrm{x}_{\mathrm{new}}$: new data after SMOTE
x_{i} : minority class data
x_{zi} : nearest neighbor from x_{i} within the same class
$\delta \sim \mathrm{U}\ (0,1)$: random numbers from uniform distribution between 0 and 1

Hyperparameter tuning plays a crucial role in controlling model complexity and balancing bias variance trade offs (Probst et al., 2019). Key parameters include the number of trees, tree depth, minimum samples for splitting and leaves, and the number of features considered. Several key hyperparameters in the Random Forest algorithm include: (i) n_estimators, which represents the number of decision trees in the model; (ii) max_depth, which defines the maximum depth of each tree; (iii) min_samples_split, which specifies the minimum number of samples required to split an internal node; (iv) min_samples_leaf, which indicates the minimum number of samples required at a leaf node; and (v) max_features, which determines the number of features considered at each split.

2.1. Feature Importance

Feature Importance measures the contribution of each variable based on the average reduction in Gini impurity:

$$\mathrm{FI}_{\mathrm{j}} = \frac{1}{\mathrm{T}} \sum_{\mathrm{t}=1}^{\mathrm{T}} \sum_{\mathrm{n} \in \mathrm{N}_{\mathrm{j,t}}} \Delta \mathrm{Gini}_{\mathrm{n,t}} \quad (2)$$

where:

FI_{j} : *feature importance* of the *j*-th feature
$\mathrm{N}_{\mathrm{j,t}}$: set of nodes in tree *t* that use feature *j*
$\Delta\mathrm{Gini}_{\mathrm{n,t}}$: decrease in Gini Impurity at node *n* in tree *t*

2.2. SHapley Additive exPlanations (SHAP)

SHAP is a model interpretation method based on game theory introduced by Scott M. Lundberg and Su-In Lee (2017), which is used to explain the contribution of each feature to the prediction output of a machine learning model. This method adopts the concept of Shapley values, where each feature is considered a “player” contributing to the model output by accounting for all possible combinations of feature subsets, resulting in an additive and consistent interpretation. Mathematically, SHAP is defined as:

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i \quad (3)$$

where:

ϕ_0: base value pr the average model output
ϕ_i: SHAP value contribution of the *i*-th feature
M: total number of features in the model

2.3. Evaluation Classification Models

Model evaluation in classification aims to measure how well a model correctly classifies data into predefined classes. In this study, the performance of the Random Forest model is evaluated using a confusion matrix and several derived performance metrics, namely accuracy, precision, recall, and F1-score. The use of multiple evaluation metrics is essential because each metric captures different aspects of model performance, particularly in imbalanced data conditions.

A confusion matrix is an evaluation table used to compare actual class labels with predicted class labels generated by a classification model. It provides a detailed view of model performance by showing correct and incorrect predictions for each class. The confusion matrix consists of four main components: True Positive (TP), which represents correctly predicted positive instances; False Negative (FN), which represents positive instances incorrectly predicted as negative; False Positive (FP), which represents negative instances incorrectly predicted as positive; and True Negative (TN), which represents correctly predicted negative instances.

Model performance evaluation is conducted to assess how well the classification model correctly predicts the target classes. In this study, the performance of the Random Forest model is evaluated using a confusion matrix along with several derived metrics, namely accuracy, precision, recall, and F1-score, to provide a comprehensive assessment, especially under imbalanced data conditions. The confusion matrix compares actual and predicted classes and consists of four components: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}.$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}},$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}.$$

$$\text{F1-score} = \frac{2\times(\text{Precision}\times\text{Recall})}{\text{Precision}+\text{Recall}}.$$

The use of these multiple evaluation metrics ensures a more reliable and balanced interpretation of model performance, particularly when handling class imbalance.

3. METHODOLOGY

This study utilizes secondary data obtained from the Indonesia Deposit Insurance Corporation in 2025, with Liquidation Duration (LD) as the dependent variable and its determining factors as independent variables.

Table 1. Variabel Penelitian

No	Variable	Unit
1	Liquidation Duration (LD)	Month
2	Total Gross Assets (TGA)	IDR
3	Total Liabilities (TL)	IDR
4	Equity (EQ)	IDR
5	Guaranteed Claims Value (GCV)	IDR
6	Estimated Asset Liquidation (EAL)	IDR
7	NSL/LANAP/LAPAN Value (NN)	IDR
8	Total Asset Liquidation (TAL)	IDR
9	Liquidation Cost (LC)	IDR
10	LPS Bailout Fund (BF)	IDR
11	Remaining Bailout Obligations (RBO)	IDR
12	Claim Recovery Value (CRV)	IDR
13	Liquidation Proceeds (LP)	IDR
14	Claim Recovery Ratio (CR)	%
15	Recovery Rate (RR)	%
16	NSL/LANAP/LAPAN Achievement (NA)	%
17	Payout Ratio (Account) (PRA)	%
18	Payout Ratio (Nominal) (PRN)	%
19	Net Asset Liquidation Ratio (NALR)	%
20	Audited Non-Performing Assets to LCT Assets (ANA)	%
21	Asset Recovery (AR)	%
22	Fair Value Ratio (FVR)	%
23	Liquidation-to-Cost Ratio (LCR)	%

The data analysis is conducted systematically, starting with descriptive statistical analysis to describe the general characteristics of the data. The target variable is then constructed by classifying liquidation duration into two categories, namely Smooth and Problematic, based on the duration in months. Data cleaning and preprocessing are subsequently performed, including handling missing values using median imputation and treating outliers using the Winsorizing method. The dataset is then split into training and testing sets using a 70:30 train-test split while maintaining class proportions. The training data are standardized to ensure uniform variable scales. Modeling is carried out using Random Forest as the baseline model, followed by hyperparameter tuning to improve performance. Class imbalance is addressed using the SMOTE method to enhance the model's ability to recognize both classes proportionally. Model performance is evaluated using accuracy, precision, recall, F1-score, and confusion matrix to provide a comprehensive assessment. Finally, feature importance and SHAP analyses are conducted on the best-performing model to identify the most influential variables in classifying BPR liquidation duration, followed by drawing conclusions and formulating recommendations based on the findings.

4. RESULT AND DISCUSSION

4.1. Data Preprocessing

Before modeling, data preprocessing is conducted to ensure data quality, including target variable construction, handling missing values, and data standardization. The target variable (liquidation duration) is classified into **Smooth (55.56%)** and **Problematic (44.44%)**, indicating class imbalance. Most variables have low missing values (<20%) and are retained, except BF and RBO, which are excluded due to high missing rates (>70%). The Jarque-Bera test indicates that all variables are not normally distributed (JB > 5.991), leading to the use of median imputation. Outliers are identified across most variables and handled using IQR-based winsorizing. Wide interquartile ranges indicate high variability and data heterogeneity.

4.2. *Splitting Data*

The dataset (120 observations) is split into training (70%, 84 observations) and testing (30%, 36 observations) sets. Each observation includes 22 independent variables. Standardization using StandardScaler is then applied to ensure uniform scaling, resulting in mean values close to zero ($\approx 10^{-16}$) and standard deviations around 1.006. This confirms that no variable dominates the learning process, and the data are ready for modeling.

4.3. *Random Forest Clasification*

Hyperparameter tuning is performed using GridSearch to identify the best model configuration by testing combinations of parameters such as n_estimators (300–900), max_depth (None–30), min_samples_split (2–20), min_samples_leaf (1–8), max_features (sqrt, log2), and bootstrap (True/False).

Table 2. Confusion matrix Random Forest

Without SMOTE			SMOTE		
Prediction	Actual		Prediction	Actual	
	Problematic	Smooth		Problematic	Smooth
Problematic	14	2	Problematic	15	1
Smooth	5	15	Smooth	5	15

The model without SMOTE correctly classifies 29 out of 36 observations, while the model with SMOTE correctly classifies 30 observations, indicating improved performance.

Table 3. Model Evaluation

Without SMOTE			SMOTE
Evaluation Matrix	Value	Evaluation Matrix	Value
Accuracy	0,805	Accuracy	0,833
Precision (Problematic)	0,737	Precision (Problematic)	0,750
Recall (Problematic)	0,875	Recall (Problematic)	0,938
F1-Score (Problematic)	0,800	F1-Score (Problematic)	0,833

Without SMOTE			SMOTE
Evaluation Matrix	Value	Evaluation Matrix	Value
Precision (Smooth)	0,882	Precision (Smooth)	0,938
Recall (Smooth)	0,750	Recall (Smooth)	0,750
F1-Score (Smooth)	0,811	F1-Score (Smooth)	0,833

Overall, the SMOTE-based model achieves higher accuracy (83.3%) and better balance between precision and recall.

4.4. Model Evaluation

Table 4. Feature Importance

No	Variable	Importance	No	Variable	Importance
1	EPA	0,134887	11	CR	0,040578
2	ANA	0,083097	12	TPA	0,037443
3	TAG	0,081671	13	BL	0,036244
4	NKP	0,075723	14	PRR	0,035681
5	PRN	0,072972	15	PAN	0,034581
6	RR	0,048220	16	RPB	0,031452
7	EK	0,048018	17	CN	0,029849
8	NN	0,047882	18	HL	0,027766
9	TK	0,044793	19	AR	0,024642
10	NPK	0,042439	20	RNW	0,022064

Feature importance shows that EAL is the most influential variable, followed by ANA, TGA, and GCV, indicating the importance of asset liquidation and financial conditions.

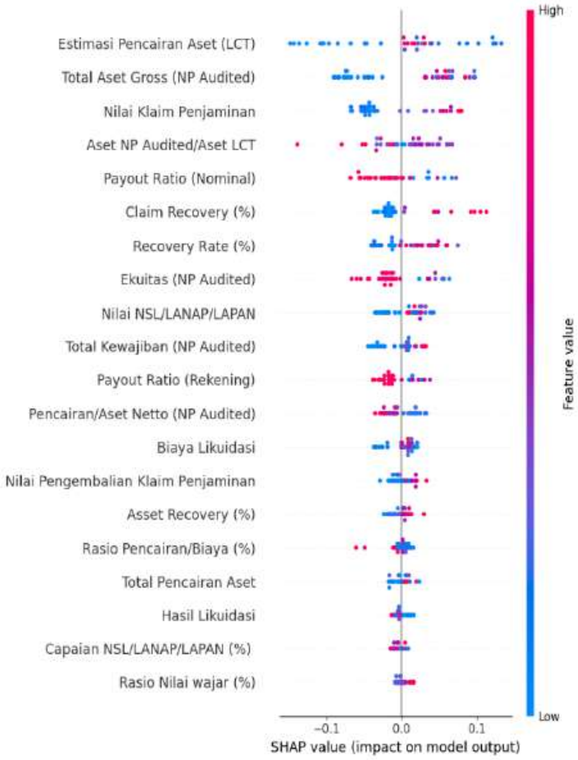

Figure 1. SHAP Value on Both Class

Based on the SHAP visualization, the top five variables—Total Gross Assets (NP Audited), Estimated Asset Liquidation (LCT), Insured Claim Value, Claim Recovery (%), and Payout Ratio (Account)—exhibit the most dominant influence on the model's predictions. In the SHAP plot, color represents the magnitude of the feature values, where red indicates high values and blue indicates low values. Meanwhile, the position of each point along the horizontal axis (SHAP value) reflects the direction of its impact on the model output: positive SHAP values (to the right) push the prediction toward the Problematic class, while negative SHAP values (to the left) push it toward the Smooth (Lancar) class. In general, high feature values (red) tend to appear on the right side, indicating a higher likelihood of being classified as Problematic, whereas low values (blue) are more concentrated on the left side, indicating a tendency toward Smooth classification. High Total Gross Assets suggest greater asset complexity, which can prolong the liquidation process; high Estimated Asset Liquidation reflects challenges in asset realization; and large Insured Claim Values indicate substantial obligations to be settled. Additionally, high values of Claim Recovery (%) and Payout Ratio (Account), although theoretically associated with better recovery and payment performance, tend to signal more complex liquidation cases that require longer resolution times. Overall, these findings indicate that larger asset scale, higher obligations, and more intensive liquidation and payout activities increase the likelihood of a bank being classified as Problematic.

5. CONCLUSION

The Random Forest algorithm effectively classifies BPR liquidation duration, where the optimized model without SMOTE achieves 80.5% accuracy but shows bias toward the majority class, while the application of SMOTE improves performance balance, particularly in detecting the minority class. Feature Importance and SHAP analyses indicate that financial variables such as Total Assets, Liabilities, Equity, Asset Liquidation, Claims, and Costs play dominant roles. Larger asset and liability values and negative equity tend to prolong liquidation, while stronger asset recovery accelerates it. The nonlinear and heterogeneous nature of the data further emphasizes the importance of these variables. Future research is recommended to use broader datasets and develop decision-support tools incorporating policy and qualitative analysis.

REFERENCES

1. Azel F. Azmi, & Apriade Voutama. (2024). *Prediksi churn nasabah bank menggunakan klasifikasi Random Forest dan Decision Tree dengan evaluasi confusion matrix*. Komputa: Jurnal Ilmiah Komputer dan Informatika, 13(1).
2. Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. Expert Systems with Applications, 83, 405–417.
3. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
4. Brown, C. O., & Dinc, I. S. (2011). Too many to fail? Evidence of regulatory forbearance when the banking sector is weak. Review of Financial Studies, 24(4), 1378–1405.
5. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.
6. DeYoung, R., & Torna, G. (2013). Nontraditional banking activities and bank failures during the financial crisis. Journal of Financial Intermediation, 22(3), 397–421.
7. Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong but many are useful: Variable importance for black-box models. Journal of Machine Learning Research, 20(177), 1–81.
8. Granja, J., Matvos, G., & Seru, A. (2017). Selling failed banks. Journal of Finance, 72(4), 1723–1784.
9. Hadad, M. D., Santoso, W., Mardanugraha, E., & Ilyas, D. (2011). Analisis risiko dan kinerja Bank Perkreditan Rakyat di Indonesia. Jakarta: Bank Indonesia.
10. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. New York, NY: Springer.
11. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284.
12. Husna Afanyn Khoirunissa, Amanda Rizky Widyaningrum, & Annisa Priliya Ayu Maharani. (2021). Comparison of Random Forest, Logistic Regression, and Multilayer Perceptron Methods on Classification of Bank Customer Account Closure. *Indonesian Journal of Applied Statistics, 4*(1).
13. Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance, 34(11), 2767–2787.
14. Kou, G., Peng, Y., & Wang, G. (2014). Evaluation of classification algorithms for financial risk prediction. Knowledge-Based Systems, 54, 243–256.

15. Kraus, M., & Feuerriegel, S. (2017). Decision support from financial disclosures with deep neural networks and transfer learning. Decision Support Systems, 104, 38–48.
16. Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. New York, NY: Springer.
17. Laeven, L., & Valencia, F. (2018). Systemic banking crises revisited (IMF Working Paper No. 18/206). International Monetary Fund.
18. Lembaga Penjamin Simpanan. (2024). Laporan kelembagaan Lembaga Penjamin Simpanan Triwulan IV Tahun 2024. Jakarta: Lembaga Penjamin Simpanan.
19. Lembaga Penjamin Simpanan. (2024). Laporan tahunan Lembaga Penjamin Simpanan tahun 2024. Jakarta: Lembaga Penjamin Simpanan.
20. Lembaga Penjamin Simpanan. (2024). Peraturan Lembaga Penjamin Simpanan Nomor 2 Tahun 2024 tentang rencana resolusi bagi bank umum. Jakarta: Lembaga Penjamin Simpanan.
21. Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring. European Journal of Operational Research, 247(1), 124–136.
22. Philipp Probst, Marvin N. Wright, & Anne-Laure Boulesteix. (2019). Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery, 9*(3), e1301.
23. Zhang, J., Li, X., & Wang, Y. (2024). Impact of an improved random forest-based financial risk model on corporate decision-making. Decision Analytics Journal, 12, 100–118.