Jurnal Statistika Dan Komputasi (STATKOM), Vol. 5 No. 1 (June 2026)
ISSN : 2963-038X (print), ISSN : 2963-0398
DOI : https://doi.org/10.32665/statkom.v5i1.6455

Stability of SHAP-Based Feature Importance Ranking under Class Imbalance, Feature Correlation, and Feature Dimensionality

Amri Luthfi Najih[1], Bagus Sartono[2], Septian Rahardiantoro[3]

[1,2,3] School of Data Science, Mathematics, and Informatics, IPB University
E-mail: mr.najih@gmail.com

Submitted: April 15, 2026 ***Revised:*** June 19, 2026 ***Accepted:*** June 19, 2026

Abstract

Background: *Gradient boosting machine learning models, such as XGBoost and LightGBM, are widely used because of their high predictive performance. However, their complexity requires interpretation methods. SHAP is often used to explain feature contributions, but the stability of its interpretations may be affected by data characteristics and the model used.*
Objective: *This study evaluates the stability of SHAP-based feature importance rankings in XGBoost and LightGBM under various data conditions.*
Methods: *The study used controlled simulation and empirical BPJS Kesehatan claims data. In the simulation, 64 datasets were generated from combinations of minority class proportion, feature correlation, and number of features. The models were fitted repeatedly, and ranking stability was evaluated using Sequential Rank Agreement (SRA), where smaller values indicate more stable rankings.*
Results: *Higher feature correlation and extreme class imbalance reduced feature ranking stability. In the BPJS Kesehatan data, imbalance handling methods improved both model performance and interpretation stability. LightGBM with ADASYN produced a smaller SRA value (0.5350) than the model without imbalance handling (2.2534).*
Conclusion: *Feature correlation and class imbalance play an important role in determining the stability of SHAP interpretations. SMOTE and ADASYN improve predictive performance and increase the stability of feature importance rankings.*

Keywords : *Feature Importance, LightGBM, SHAP, SRA, XGBoost.*

Abstrak

Latar Belakang: Model *machine learning* berbasis *gradient boosting* seperti XGBoost dan LightGBM banyak digunakan karena memiliki kinerja prediksi tinggi, tetapi bersifat kompleks sehingga memerlukan metode interpretasi. SHAP sering digunakan untuk menjelaskan kontribusi fitur, namun stabilitas hasil interpretasinya dapat dipengaruhi karakteristik data dan model.
Tujuan: Penelitian ini mengevaluasi stabilitas peringkat kepentingan fitur berbasis SHAP pada XGBoost dan LightGBM dalam berbagai kondisi data.
Metode: Penelitian menggunakan simulasi data terkontrol dan data empiris klaim BPJS Kesehatan. Pada simulasi, 64 dataset dibangkitkan dari kombinasi proporsi kelas minoritas, korelasi antar fitur, dan jumlah fitur. Pemodelan dilakukan berulang, lalu stabilitas peringkat fitur dievaluasi menggunakan *Sequential Rank Agreement* (SRA), dengan nilai lebih kecil menunjukkan peringkat lebih stabil.
Hasil: Peningkatan korelasi antar fitur dan ketidakseimbangan kelas ekstrem menurunkan stabilitas peringkat fitur. Pada data BPJS Kesehatan, penanganan ketidakseimbangan data meningkatkan kinerja model dan stabilitas interpretasi. LightGBM dengan ADASYN menghasilkan SRA lebih kecil (0.5350) dibandingkan model tanpa penanganan (2.2534).
Kesimpulan: Korelasi antar fitur dan ketidakseimbangan kelas penting dalam menentukan stabilitas interpretasi SHAP. SMOTE dan ADASYN tidak hanya meningkatkan kemampuan prediksi, tetapi juga stabilitas peringkat kepentingan fitur.

Kata kunci: Kepentingan Fitur, LightGBM, SHAP, SRA, XGBoost.

INTRODUCTION

Machine learning has rapidly evolved and is widely used due to its ability to generate accurate predictions. One of the most commonly applied approaches is ensemble learning, particularly gradient boosting, which has demonstrated strong predictive performance. Gradient boosting–based models such as XGBoost and LightGBM have become popular due to their computational efficiency and their ability to handle large-scale data and complex data structures (Alshboul et al., 2024; Zhao et al., 2024; Zhang & Zhao, 2026; Imani et al., 2026).

Despite their strong predictive performance, XGBoost and LightGBM are categorized as black-box models, as they do not explicitly reveal how individual features contribute to the final predictions (Molnar, 2025). This limitation may reduce trust in the model outcomes, highlighting the importance of model interpretability (Asif et al., 2025). Sartono et al. (2025) demonstrate that the proposed MIMO-LSTM model is capable of detecting market downturn signals one day prior to the event, indicating its ability to capture complex patterns in multivariate data. However, this capability would be more effective if complemented by methods that can explain the contribution of each feature in generating predictions. On the other hand, Rahardiantoro et al. (2024) and Aswi et al. (2025) analyze factors associated with stunting using spatio-temporal modeling approaches, including Bayesian methods to examine socioeconomic factors. Although these studies successfully identify influential factors, they do not explicitly explain the contribution of each feature within the model. Therefore, interpretable machine learning methods such as SHAP (SHapley Additive Explanations) can be used to provide a theoretically grounded and fair explanation of feature contributions (Lundberg & Lee, 2017). However, before being widely applied, it is important to evaluate the stability of these interpretations under various data conditions.

However, SHAP-based interpretations are not always stable. The resulting feature rankings highly affected by the adopted ML model (Salih et al., 2025). Most existing studies have primarily focused on improving predictive performance or applying SHAP for model explanation (Noviandy et al., 2025; Mohammed et al., 2025; Salih et al., 2025). Dharmawan et al. (2022) attempted to address this issue, but their analysis was limited to class imbalance. Therefore, this study evaluates the stability of SHAP-based feature importance rankings under these three data conditions. Stability is measured using Sequential Rank Agreement (SRA), which quantifies agreement among ranked feature lists across repeated model runs at different ranking depths (Ekstrøm et al., 2019). Lower SRA values indicate more stable rankings

The BPJS Kesehatan claims dataset is used as an empirical case study because it represents a real-world dataset with extreme class an imbalance. Out of 586,733 participants, only 7,826 individuals (1.33%) belong to the mental health INACBGs group, while 578,907 participants (98.67%) do not belong to this group. Such imbalance may affect not only predictive performance but also the stability of model interpretation, particularly in SHAP-based feature importance rankings. Therefore, the BPJS Kesehatan data complement the controlled simulation by providing an empirical setting to evaluate whether SHAP-based interpretations remain stable in complex and non-ideal real-world data. The findings are expected to provide practical insights into the use of interpretable machine learning in highly imbalanced healthcare classification problems.

METHOD

Research Design

This study employs a quantitative approach using both simulated data and secondary data. The objective of the study is to evaluate the stability of feature ranking generated by SHAP in XGBoost and LightGBM models.

Population and Sample

For the simulation data, a population of 600,000 observations was generated. For the BPJS healthcare claims data, the population consists of all BPJS Kesehatan service data in Indonesia, while the sample used in this study is the dataset provided and published by BPJS Kesehatan.

Sampling Techniques

The simulation data were generated under predefined and controlled conditions based on combinations of minority class proportion, feature correlation, and feature dimensionality. From each simulated population, 100,000 observations were selected using simple random sampling and then divided into training and testing sets using stratified sampling with a 70:30 ratio. For the BPJS Kesehatan data, the sample was determined by the availability of anonymized secondary claims data released by BPJS Kesehatan for 2018–2023. After data cleaning, records with complete values for the selected features and target variable were retained for analysis. The cleaned dataset was repeatedly split into training and testing sets using stratified sampling with a 70:30 ratio.

Research Subjects

This study utilizes two data sources: simulated data and secondary data from BPJS Kesehatan. The simulation data consist of 64 datasets generated from combinations of three different data conditions, namely the proportion of minority class (0,01;0,1;0,3;0,5), feature correlation (0;0,2;0,5;0,95), and the number of features (7;15;25;35). Table 1: The following are the 64 datasets generated through the simulation.

Table 1. Simulation Data.

Data Set	Data conditions		
	Proportion of minority class	Feature correlation	Number of features
1	0,01	0	7
2	0,01	0	15
3	0,01	0	25
4	0,01	0	35
5	0,01	0,2	7
...	...	...	...
63	0,5	0,95	25
64	0,5	0,95	35

The secondary data used in this study are subject to the availability of datasets released by BPJS Kesehatan for the period 2018–2023. Two primary data sources are utilized, namely participant data and advanced healthcare facility data. A total of 22 variables are employed as features, while 1 variable serves as the target. These features are grouped into four main categories: participant demographics, membership status as a proxy for socioeconomic conditions, healthcare access, and healthcare utilization history.

Based on data type, the features were divided into numerical and categorical features. The numerical features consisted of age, number of family members with a history of mental health clinic visits, total number of family members, number of outpatient visits, number of inpatient visits, number of anesthesiology clinic visits, number of neurology clinic visits, number of nutrition clinic visits, number of internal medicine clinic visits, number of cardiology clinic visits, number of pulmonology clinic visits, number of medical rehabilitation clinic visits, number of diabetes diagnoses, and number of hypertension diagnoses. The categorical features consisted of marital status, family relationship, gender, membership status, treatment class, participant segment, type of primary healthcare facility, and type of advanced referral healthcare facility. The target variable was mental disorder status, coded

as 1 if the participant belonged to the mental health INACBGs group and 0 otherwise.

Data Analysis Techniques

This study applies two approaches to imbalanced data. In the simulation data, class imbalance is maintained as is, without applying oversampling or undersampling techniques, as the primary objective is to evaluate the stability of SHAP under pure imbalance conditions. The application of balancing techniques such as SMOTE would alter the minority class proportion and eliminate the original data characteristics that are intended to be evaluated. This approach follows several previous studies that retained the original class imbalance to evaluate its impact on model behavior. Bracke et al. (2019), for example, used a dataset with a minority class proportion of 2.5% to analyze the effect of imbalance on Shapley values in Gradient Boosting models without applying imbalance handling techniques. Other studies using SHAP in imbalanced data contexts also do not apply oversampling, undersampling, or other balancing methods (Bussmann et al., 2021; Óskarsdóttir & Bravo, 2021; Liu et al., 2022).

In contrast, for the BPJS dataset, imbalance handling methods were applied using Synthetic Minority Oversampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN). SMOTE generates synthetic minority-class observations by interpolating between existing minority samples and their nearest neighbors, thereby expanding the minority class distribution more evenly. ADASYN also generates synthetic minority samples, but it adaptively assigns more synthetic observations to minority samples that are harder to classify, allowing the model to focus more on difficult regions of the feature space (Zhu et al., 2024). The procedures were implemented in Python using the *SMOTE* and *ADASYN* functions from the *imblearn* package.

Both SMOTE and ADASYN were applied only to the training data after the train–test split, while the testing data were kept in their original class distribution to provide an unbiased evaluation of model performance. This procedure was used to prevent data leakage and avoid overly optimistic performance estimates. This study design clearly distinguishes between evaluating the stability of SHAP-based interpretation in simulated data and applying imbalance handling in empirical data.

The machine learning models used in this study are Extreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LightGBM). XGBoost employs the following objective function (Chen & Guestrin, 2016):

$$y_i^{(t)} = y_i^{(t-1)} + f_t t(xi) \quad (1)$$

where $y_i^{(t)}$ is the updated prediction for the i observation at iteration t, $y_i^{(t-1)}$ is the prediction from the previous iteration, and $f_t t(xi)$ represents the residual learned by the decision tree at iteration t. In other words, each iteration adds a new tree to correct the errors from previous predictions. The objective function is defined as:

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, y_i^{(t-1)} + f_t(x_i)) + \sum_{t=1}^{T} \Omega(f_t) \quad (2)$$

where $l(y_i, y_i^{(t-1)} + f_t(x_i))$ is the loss function measuring the difference between predictions and true values, $\Omega(f_t)$ is the regularization term controlling model complexity, and T is the number of trees. The regularization function is defined as:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum w_j^2 \quad (3)$$

where T represents the number of leaf nodes, w_j denotes the weight of each leaf node, γ controls model complexity (larger values result in fewer tree splits), and, λ is the regularization parameter used to prevent overfitting.

XGBoost constructs decision trees using

a level-wise approach, meaning trees grow layer by layer, where all nodes at the same level are expanded before moving to the next level. This approach is more stable but slower compared to the leaf-wise strategy used by LightGBM.

LightGBM has a similar objective function to XGBoost but differs in its implementation. It adopts a leaf-wise growth strategy, where the leaf that yields the largest reduction in loss is expanded first. This approach allows LightGBM to build trees more efficiently compared to the level-wise method used in XGBoost.

For XGBoost, the main hyperparameters were set to n_estimators = 100, max_depth = 6, and learning_rate = 0.1, with log-loss used as the evaluation metric. For LightGBM, n_estimators was set to 100, while the remaining parameters followed the default settings of the LightGBM package. These fixed hyperparameter settings were used to ensure comparability across simulation scenarios and repeated model runs, rather than to optimize predictive performance. The analysis was conducted in Python using the *XGBoost*, *LightGBM*, *SHAP*, *scikit-learn*, and *imbalanced-learn packages*. Random seeds were fixed to ensure reproducibility.

Feature importance rankings are computed using SHAP, based on the contribution of each feature to the prediction. SHAP is derived from the Shapley value, a concept originally introduced in cooperative game theory by Shapley (1953). The Shapley value provides a principled way to fairly distribute the total payoff among players based on their marginal contributions across all possible coalitions. In the context of machine learning interpretation, features are treated as players, and their Shapley values represent the average marginal contribution of each feature to the model prediction across different feature coalitions. The Shapley value for feature j is defined as follows (Molnar, 2025):

$$\phi_j(val) = \sum_{S \subseteq \{1,\dots,p\} \setminus \{j\}} \frac{|S|!\,(p-|S|-1)!}{p!} (val(S \cup \{j\}) - val(S)) \quad (4)$$

where ϕ_j represents the contribution of feature j (nilai *shapley*), p is the number of features, S denotes the number of subsets (coalitions), $val(S)$ represents the prediction for subset S without the inclusion of feature j and $val(S \cup \{j\})$= represents the prediction for subset S with feature j included. The Shapley value ϕ_j represents the average contribution of feature j to the model prediction, considering all possible combinations (coalitions) of the other features.

SHAP (SHapley Additive Explanations) is an additive feature attribution method used to explain individual predictions of machine learning models (Lundberg & Lee, 2017). It bridges the concept of Shapley values with additive models. The SHAP model is defined as (Molnar, 2025):

$$g(z') = \phi_0 + \sum_{j=1}^{p} \phi_j z'_j \quad (5)$$

where $g(z')$ is the SHAP explanation model, ϕ_0 is the baseline value (expected value), ϕ_j is the contribution of feature j, and z'_j is a binary indicator (1 if the feature is present, 0 otherwise).

Sequential Rank Agreement (SRA) is a statistical method used to measure the level of agreement between two or more ranked lists generated from feature modeling. SRA is particularly useful for evaluating ranking stability, especially within the top-ranked positions, which typically have the greatest influence on the target prediction (Ekstrøm et al., 2019). SRA has been widely applied in various studies to assess the stability and agreement of rankings.

Ballegeer et al. (2025) applied SRA to evaluate the stability of LIME and SHAP interpretations in instance-dependent cost-sensitive (IDCS) models for credit scoring. Dietz et al. (2022) utilized the SRA concept to compare incomplete ranked lists, enabling the evaluation of ranking agreement even when the items differ

across lists. Sejling et al. (2025) extended the SRA framework by introducing entropy rank agreement, an entropy-based metric that provides a more stable evaluation of ranking agreement in variable selection results obtained from bootstrap data. Meanwhile, Angulo et al. (2022) adapted the fundamental principle of distance-based ranking comparison underlying SRA to develop the Modified Manhattan Distance (MMD), a metric designed to quantitatively assess the similarity between model-generated rankings and expert-defined rankings.

In this study, SRA is used to evaluate the stability of feature rankings generated by the SHAP interpretation method. The following are the steps of the SRA procedure:

1. Suppose there are P features to be ranked, denoted as $X = \{X_1, X_2, \dots, X_P\}$. There are L ranked lists to be compared, each providing a ranking of all features. These lists are denoted as $R = \{R_1, R_2, \dots, R_L\}$, where $R_l: X \rightarrow \{1,2, \dots, P\}$, for $l = 1, \dots, L$. This means that $R_l(X_P)$ represents the rank of feature X_P in the l-th list.
2. Select a depth level $d \in \{1, \dots, P\}$, representing the top-d ranks to be analyzed.
3. Construct the feature set $S(d)$, defined as the set of all features that appear in positions 1 to d in at least one of the L ranked lists:

$$S_L(d) = \left\{ X_P : \frac{1}{L} \sum_{l=1}^{L} 1(R_l(X_P) \leq d) > \varepsilon \right\} \quad (6)$$

with $\varepsilon = 0$, a feature is included in $S(d)$ if it appears within the top-d positions in at least one ranked list.

4. Compute the mean and variance of the rank for each feature X_P across all lists, as a measure of ranking variability:

$$\bar{R}_L(X_P) = \frac{1}{L} \sum_{l=1}^{L} R_l(X_P) \quad (7)$$

$$\hat{A}_L(X_P) = \frac{1}{L-1} \sum_{l=1}^{L} (R_l(X_P) - \bar{R}_L(X_P))^2 \quad (8)$$

5. Compute the SRA value at depth d as the average variance of the ranks for all features in $S(d)$:

$$sra_L(d) = \frac{1}{|S(d)|} \sum_{X_P \in S(d)} \hat{A}_L(X_P) \quad (9)$$

this value represents the magnitude of ranking variability at depth d. A smaller SRA value indicates greater agreement among the ranked lists, particularly within the top-ranked positions.

The steps of the simulation data analysis are as follows:

1. Define the simulation parameters: the proportion of minority class (r) (0,01;0,1;0,3;0,5), feature correlation (π) = 0;0,2;0,5;0,95, and the number of features (p) = 7;15;25;35.
2. Construct the covariance matrix such that all pairs of features have the same correlation π.

$$\sum ij = \begin{cases} \sigma^2 & i = j \\ \pi\sigma^2 & i \neq j \end{cases} \quad (10)$$

3. Each dataset consists of n=600000 numerical features $X = (X_1, X_2, \dots, X_p) \in \mathbb{R}^{n \times p}$ and one binary target variable $Y \in \{0,1\}$ representing two classes.
4. The feature data X are generated from a multivariate normal distribution:

$$X \sim N(0, \Sigma) \quad (11)$$

5. The target variable $Y \in \{0,1\}$ is determined using the following linear function:

$$f(X) = 0.9X1 + 0.7X2 - 0.5X3 + 0.3X6 + 0.01X7 \quad (12)$$

this function represents the true relationship between the features and the target in the simulated data. Based on this function, features X_1, X_2, X_3, X_6, and X_7 are defined as the truly relevant features influencing the target. The magnitude of each coefficient reflects the level of influence, where X_1 has the strongest contribution, followed by X_2, X_3 and X_6, while X_7 has a very small effect.

6. Add noise to the function: $f(X)$

$$score = f(X) + \epsilon;\ \ \epsilon \sim N(0, \sigma^2) \qquad (13)$$

7. Transform the $score$ using the sigmoid function:

$$P = \frac{1}{1 + e^{-score}} \qquad (14)$$

8. Generate the target variable using the Bernoulli distribution:

$$y \sim Bernoulli(P) \qquad (15)$$

9. Draw a sample from the generated dataset, with 70% used as training data and 30% as testing data. Sampling is performed using stratified random sampling for each class.
10. Perform modeling using XGBoost and LightGBM.
11. Compute SHAP values to obtain feature rankings.
12. Repeat steps 9–11 for 100 times.
13. Compute SRA for each combination of data conditions and each repetition.
14. Compare the results across models and data conditions.

The analytical workflow of the simulation data is illustrated in Figure 1.

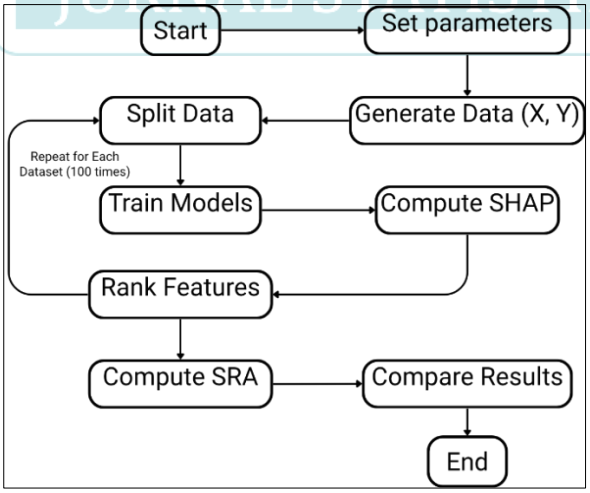

Figure 1. Workflow Data Simulation Analysis

The steps of the analysis for the BPJS Kesehatan secondary data are as follows:

1. Request BPJS healthcare claims sample data for the period 2018–2023 from BPJS Kesehatan.
2. Use 2 primary data sources as references, namely participant data and advanced healthcare facility data.
3. Identify variables to be used as features. Based on previous studies, factors associated with mental health include participant demographics, social and economic conditions, physical health/comorbidities, cultural factors, environmental factors, and access to healthcare services (Septiani et al., 2025; Sari et al., 2025; Pratamawati et al., 2025; Asrullah et al., 2025; Maharani et al., 2025; Su et al., 2023).
4. Aggregate the data according to the selected features.
5. Perform data cleaning and encode categorical variables.
6. Split the data into 70% training data and 30% testing data.
7. Build models using XGBoost and LightGBM. Modeling is conducted both with class imbalance handling (SMOTE and ADASYN) and without any imbalance treatment.
8. Compute SHAP values to obtain feature rankings.
9. Repeat steps 6–8 for 100 iterations.
10. Compute SRA for each combination of model and imbalance handling method.
11. Compare the results across model and imbalance handling combinations.

The analytical workflow of the BPJS Kesehatan data is illustrated in Figure 2.

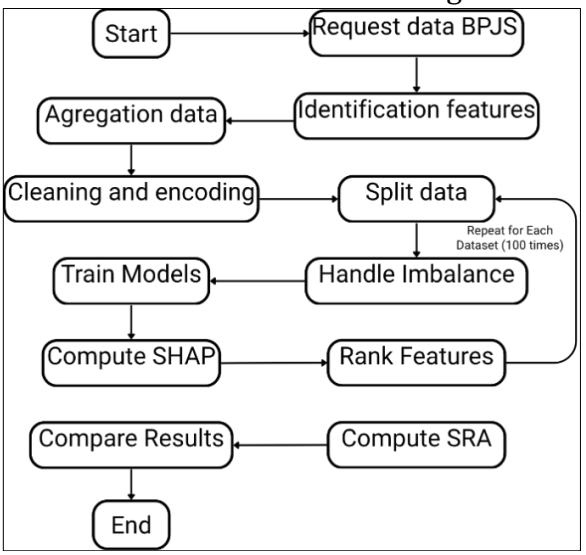

Figure 2. Workflow Data BPJS Kesehatan

RESULTS AND DISCUSSION

Simulation Study to Evaluate SHAP Performance

The distribution of SHAP rankings across all combinations of data conditions in the number of features 15 (Figure 3) reveals a clear separation between truly relevant and weak or irrelevant features. Features X_1, X_2, X_3, and X_6 repeatedly occupied the top ranks, indicating that SHAP was able to reliably recover the underlying data-generating mechanism. This stability reflects the dominance of their coefficients in the ground truth function, suggesting that SHAP effectively captures strong signal features even under varying data conditions.

In contrast, the feature X_7 exhibits substantially higher ranking variability, often overlapping with features that are not included in the true model. Despite being part of the ground truth, its very small coefficient (0.01) results in a negligible contribution to the target variable. Consequently, SHAP struggles to distinguish X_7 from irrelevant features, indicating that features with weak effects are highly sensitive to noise and data perturbations. This finding highlights a critical limitation of SHAP: while it is reliable for identifying strongly influential features, it may produce unstable rankings for features with marginal contributions.

Furthermore, although the top features generally demonstrate stable rankings, the presence of outliers in the global distribution suggests that their rankings are not entirely invariant. These deviations indicate that data characteristics, such as class imbalance, feature correlation, and feature dimensionality, can still introduce instability, even for dominant features.

Overall, these results suggest that the reliability of SHAP-based feature rankings is not solely determined by feature relevance, but also by the strength of feature effects and the underlying data structure. Therefore, evaluating the stability of SHAP interpretations is essential before drawing substantive conclusions from feature importance rankings.

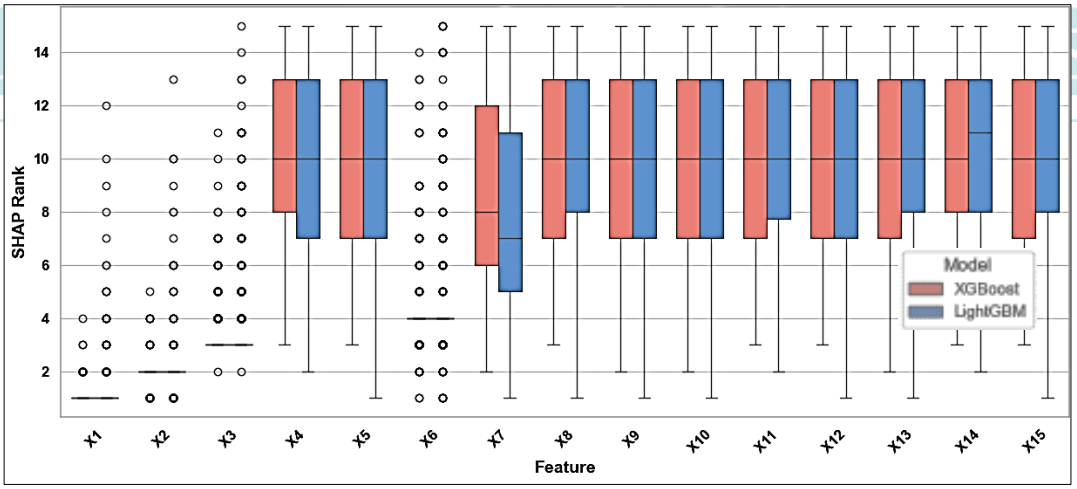

Figure 3. Distribution of SHAP feature rankings across all data condition combinations in number of features 15.

Further analysis is presented in Figure 4, which presents the percentage of correct feature ranking for X_1, X_2, X_3, X_6, and X_7 across 100 repetitions of the LightGBM model. The results indicate that under low to moderate correlation conditions (0, 0.2, and 0.5), SHAP was able to reliably identify the true ranking of the dominant features (X_1, X_2, X_3, and X_6), achieving an accuracy level above 90% across all repetitions. This suggests that when the dependency among features is relatively low, SHAP can accurately capture the underlying signal structure generated during the data generation process.

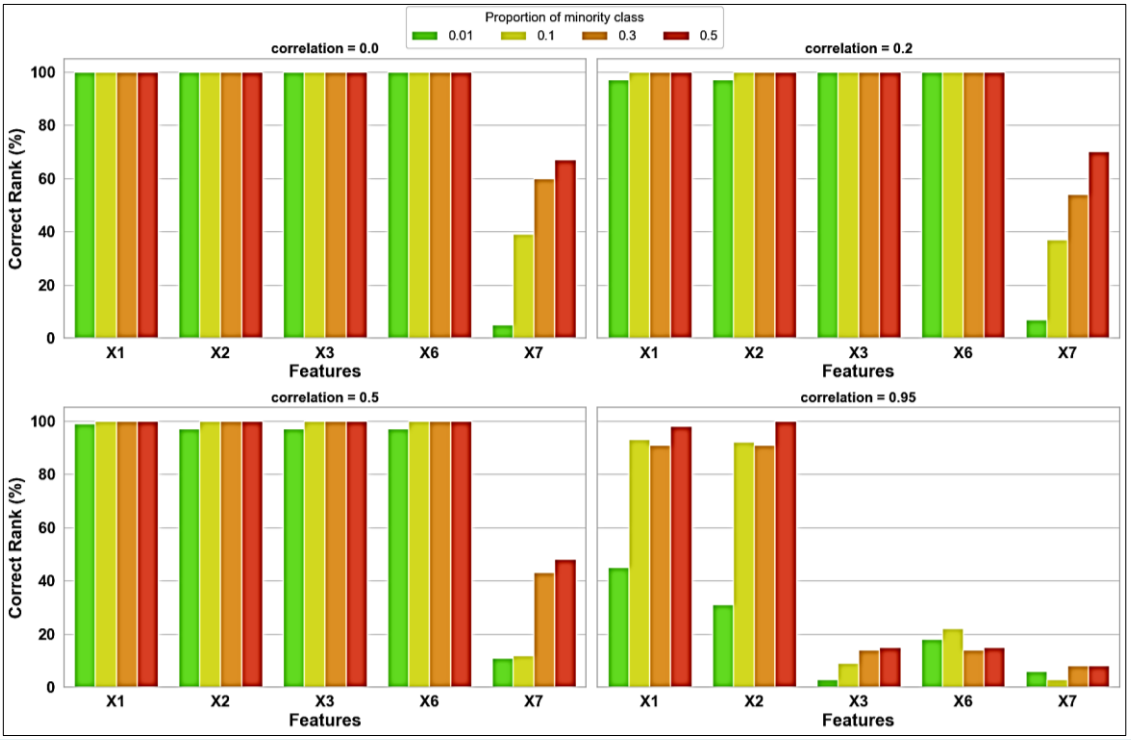

Figure 4. Percentage of correct feature ranking for X_1, X_2, X_3, X_6, and X_7 across 100 repetitions of the LightGBM model under different correlation levels in number of features 15.

However, this condition changes significantly when the correlation becomes very high (0.95). Under this scenario, SHAP begins to struggle in determining the true ranking of all features. The high correlation among features makes it difficult to distinguish the individual contributions of each variable, resulting in less stable rankings that tend to vary across repetitions.

In addition to correlation, the level of class imbalance and the magnitude of feature coefficients also play a crucial role in influencing SHAP performance. When the proportion of the minority class is very small, the model's ability to capture feature contributions decreases. This issue is further exacerbated for features with small coefficients in the data-generating function of the target variable Y, such as X_7. As the minority class proportion decreases and the feature contribution weakens, SHAP becomes increasingly challenged in correctly identifying the true feature rankings. The stability of SHAP-based feature rankings was further evaluated using Sequential Rank Agreement (SRA), as shown in Figure 5. Lower SRA values indicate more stable ranking structures across repeated model runs. The results show that SRA values increased under extreme class imbalance, particularly when the minority class proportions were 0.1 and 0.01, and under very high feature correlation. These findings indicate that SHAP-based rankings become less stable when the data contain both severe class imbalance and strong feature correlation.

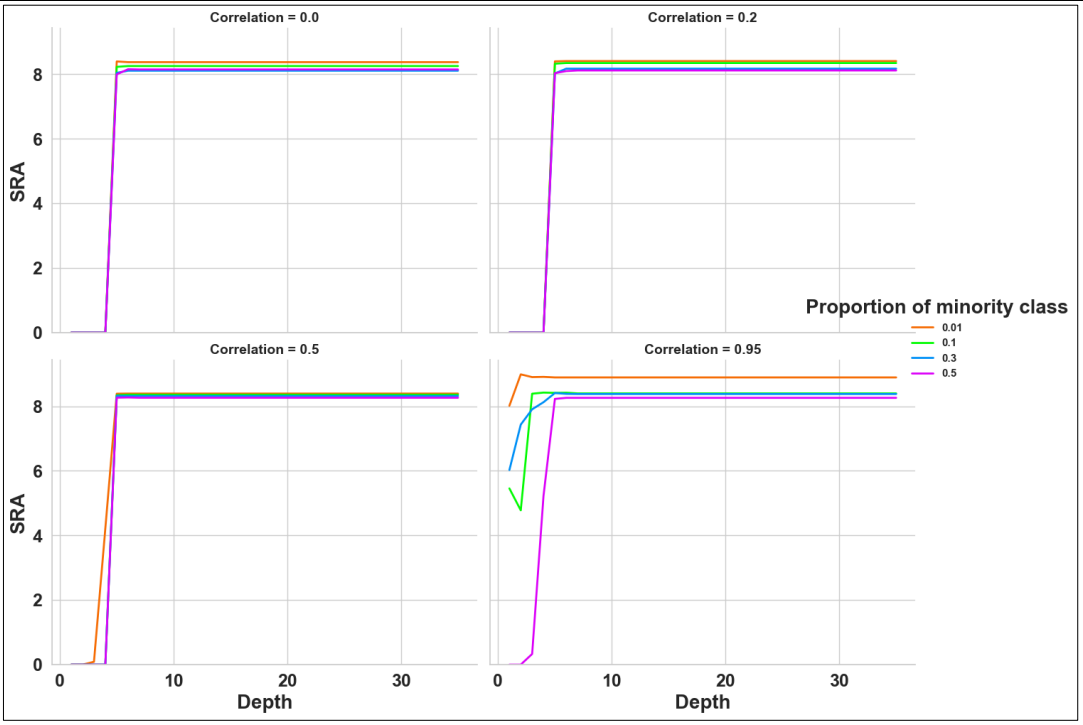

Figure 5. SRA values by ranking depth (top-k) for 35-feature data (LightGBM model).

The SRA values were lower at shallow ranking depths, indicating that SHAP reliably identified the most influential features across repeated model runs. This pattern reflects the dominant effects of the main signal features in the data-generating process. However, SRA values increased as the ranking depth increased because weak or irrelevant features were progressively included in the ranking comparison. Overall, the simulation results suggest that SHAP can reliably identify strongly influential features, but the stability of its feature rankings deteriorates when class imbalance, high feature correlation, and weak feature effects occur simultaneously.

Predictive Modeling Results for Mental Health Cases

The model was evaluated using accuracy, balanced accuracy, sensitivity, and specificity, in line with the approach adopted by Ginasputri et al. (2025), who utilized accuracy and recall (sensitivity) as evaluation metrics. The modeling process, from data splitting to model evaluation, was repeated 100 times to ensure the robustness of the results. As shown in Table 2, both LightGBM and XGBoost maintained very high accuracy across all imbalance handling methods. The accuracy values range from 0.9856 to 0.9912, indicating that the models perform well in overall classification.

However, the high accuracy observed in imbalanced datasets should be interpreted with caution. This metric is inherently dominated by the model's ability to correctly classify the majority class, potentially masking poor performance on the minority class, which is often of greater practical importance. Therefore, relying solely on accuracy may lead to an overly optimistic assessment of model performance.

Furthermore, the comparison of evaluation metrics between training and testing data reveals only marginal differences across all models. This stability suggests that the models do not suffer from overfitting and demonstrate good generalization ability. In other words, the patterns learned from the training data are effectively transferred to unseen data.

Table 2. Mean values of model performance metrics.

Model	Imbalanced handling	Accuracy		Balanced Accuracy		Sensitivity		Specificity	
		train	test	train	test	train	test	train	test
LightGBM	No handling	0.9895	0.9888	0.7107	0.6978	0.4244	0.3989	0.9971	0.9967
	Adasyn	0.9859	0.9857	0.8099	0.8058	0.6291	0.6209	0.9907	0.9906
	Smote	0.9856	0.9854	0.8176	0.8139	0.6450	0.6377	0.9902	0.9901
XGBoost	No handling	0.9912	0.9888	0.7494	0.6991	0.5009	0.4015	0.9978	0.9968
	Adasyn	0.9856	0.9854	0.8084	0.8042	0.6264	0.6181	0.9904	0.9904
	Smote	0.9853	0.9851	0.8174	0.8137	0.6449	0.6375	0.9899	0.9898

The main improvement was observed in sensitivity, which reflects the model's ability to detect the minority class. Under the baseline setting without imbalance handling, sensitivity remained low, with values of 0.3989 for LightGBM and 0.4015 for XGBoost. After applying oversampling methods, sensitivity increased substantially. SMOTE produced the highest sensitivity values, reaching 0.6377 for LightGBM and 0.6375 for XGBoost, while ADASYN yielded slightly lower values of 0.6209 and 0.6181, respectively. This result suggests that SMOTE was more effective in improving minority-class detection. The advantage of SMOTE may be related to its interpolation mechanism, which generates synthetic minority samples between existing minority observations and expands the minority class distribution more evenly. As a result, the model can learn a more general decision boundary for identifying mental disorder cases. Specificity remained high across all model and method combinations, indicating that the improvement in sensitivity was achieved without a substantial loss in majority-class classification performance.

The differences between methods are clearly illustrated in Figure 6, where balanced accuracy under the "basic = no handling" setting is substantially lower than that achieved with oversampling techniques (SMOTE and ADASYN). Overall, these findings demonstrate that oversampling, particularly SMOTE, significantly improves minority class detection without materially compromising overall classification performance. Moreover, the model comparison shows that LightGBM tended to achieve slightly better performance than XGBoost in terms of balanced accuracy and sensitivity, although the differences were marginal.

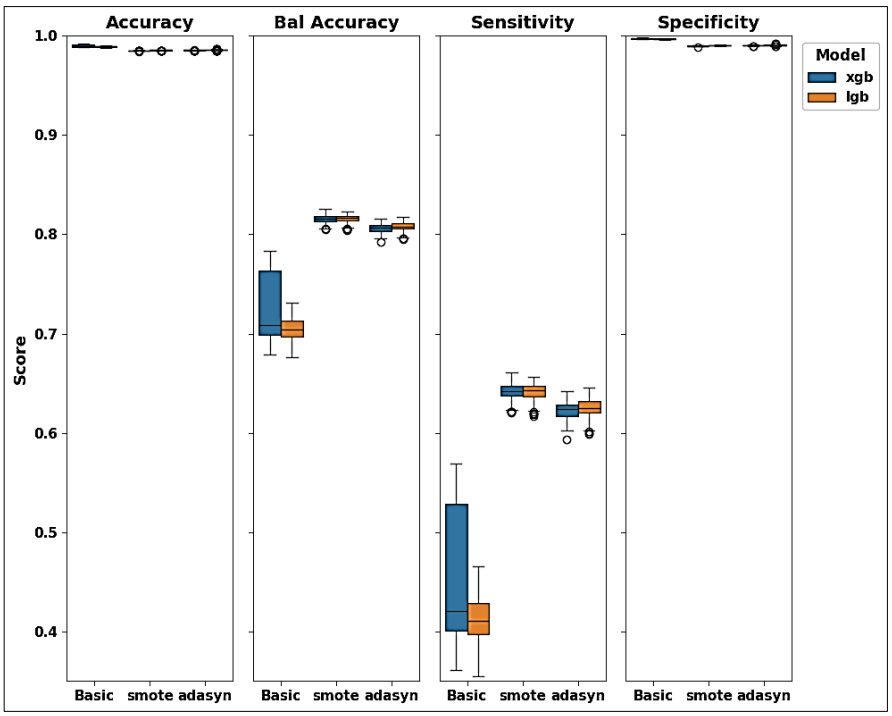

Figure 6. Comparison of model performance across different methods

The SRA results in Figure 7 and Table 3 show that imbalance handling improved the stability of SHAP-based feature importance rankings. Without imbalance handling, the average SRA values were substantially higher, particularly for LightGBM, which reached 2.2534. After applying oversampling, the average SRA values decreased markedly. In LightGBM, ADASYN produced the lowest average SRA value of 0.5350, followed by SMOTE with 0.5950. A similar pattern was observed in XGBoost, where ADASYN and SMOTE reduced the average SRA values to 0.6525 and 0.7661, respectively. These results indicate that imbalance handling not only improved predictive performance but also increased the stability of SHAP-based interpretations.

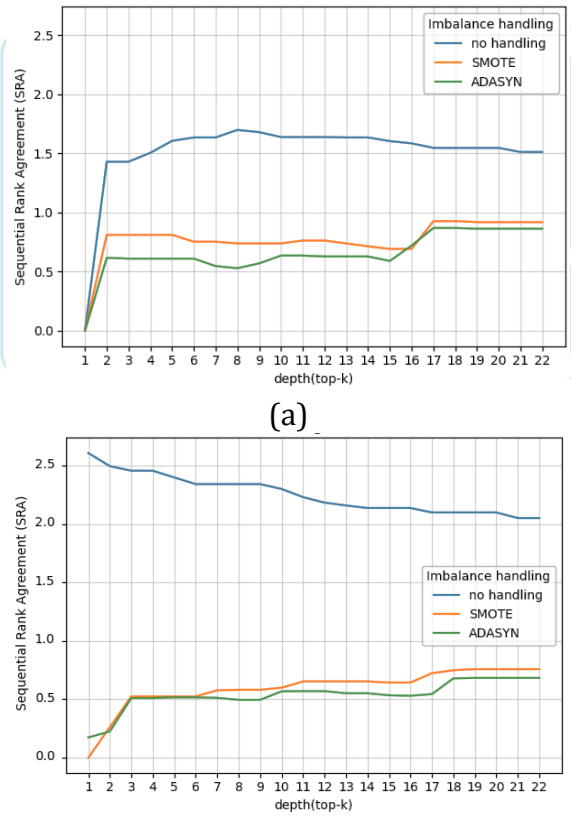

Figure 7. Sequential Rank Agreement (SRA) scores XGBoost (a) and LightGBM (b)

A trade-off was observed between predictive performance and interpretation stability. SMOTE achieved slightly higher balanced accuracy and sensitivity, indicating better minority-class detection. In contrast, ADASYN produced lower SRA values, particularly in the LightGBM model, indicating more stable feature importance rankings across repeated model runs. Therefore, the preferred imbalance handling method depends on the objective of the analysis. If the main goal is to maximize the detection of mental disorder cases, SMOTE may be more appropriate. However, if the goal is to obtain more stable SHAP-based feature rankings, ADASYN provides an advantage.

Table 3. Mean SRA values across all ranking depths

Model	Imbalance handling	Mean SRA
LGB	No handling	2.2534
	ADASYN	0.5350
	SMOTE	0.5950
XGB	No handling	1.5085
	ADASYN	0.6525
	SMOTE	0.7661

Because the LightGBM model with SMOTE produced the highest sensitivity in detecting the minority class, SHAP feature importance was evaluated using this model configuration.

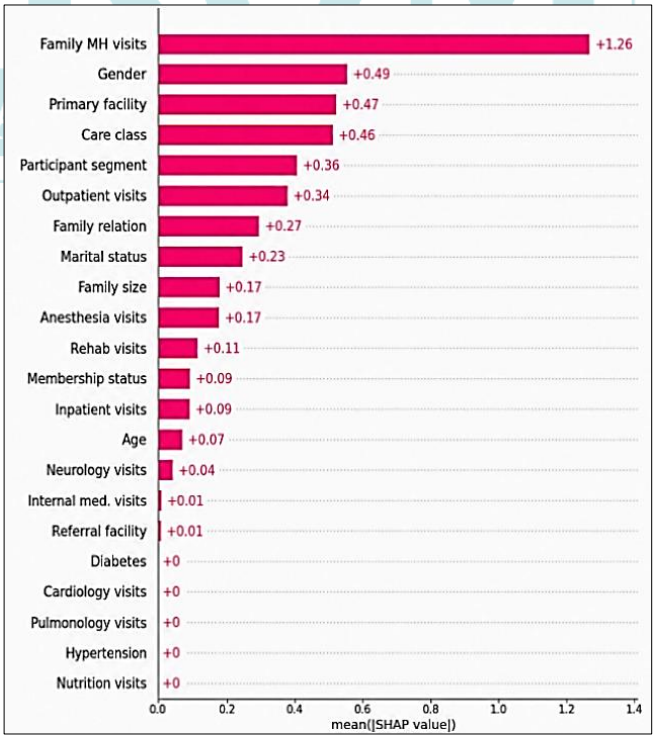

Figure 8. mean absolute SHAP values

Figure 8 presents the mean absolute SHAP values, where a larger value indicates a greater average contribution of a feature to the model prediction. The four most

influential features were family mental-health clinic visit history, gender, primary healthcare facility type, and care class. Family mental-health clinic visit history had the largest contribution, with a mean absolute SHAP value of 1.26, indicating that household-level mental health utilization was the most dominant feature in the classification of mental disorder cases. Gender, primary healthcare facility type, and care class also showed relatively high contributions, with mean absolute SHAP values of 0.49, 0.47, and 0.46, respectively. These results indicate that both individual or household characteristics and healthcare access-related features contributed substantially to the model prediction. However, the SHAP values should be interpreted as predictive feature contributions, not as evidence of causal relationships.

CONCLUSION

Conclusions

This study demonstrates that the stability of SHAP-based feature importance rankings is affected by data characteristics, particularly feature correlation and class imbalance. The simulation results show that SHAP reliably identified the strongly influential features, namely X_1, X_2, X_3, and X_6. However, feature ranking stability decreased under high feature correlation and extreme class imbalance, indicating that SHAP-based interpretations may become unstable when applied to non-ideal data conditions.

The empirical analysis using BPJS Kesehatan claims data showed a similar pattern. Imbalance handling methods improved minority-class detection and increased the stability of SHAP-based feature rankings. LightGBM with SMOTE produced the highest sensitivity, whereas LightGBM with ADASYN produced the lowest SRA value, indicating a trade-off between predictive performance and interpretation stability. These findings suggest that the best modeling strategy may depend on the primary objective of the analysis, whether prioritizing minority-class detection or stable feature interpretation.

Overall, this study provides evidence that SHAP feature importance can be used to support model interpretation, but its results should be interpreted with caution under challenging data conditions. When the data are highly imbalanced or contain strong feature correlations, these issues should be addressed before SHAP-based interpretations are used to support model-based decision making. In addition, the stability of feature importance rankings across repeated model runs should be evaluated to ensure that the identified important features are not merely the result of unstable ranking patterns. In healthcare applications such as mental disorder classification using BPJS Kesehatan data, stable SHAP-based interpretation is important to support reliable decision-making, particularly when the minority class is difficult to detect. However, the simulation component of this study is limited by the linear structure assumed in the data-generating process. Therefore, the findings from the simulation should be interpreted within the context of the simulated conditions used in this study.

Suggestion

Future research should compare SHAP with other interpretability methods, such as LIME or permutation feature importance, to evaluate whether different methods produce similar stability patterns under varying data conditions. Additional performance metrics, such as PR-AUC, F1-score, and class-specific precision, should also be considered to provide a more comprehensive evaluation for imbalanced classification problems. Further studies may extend the stability analysis to other healthcare datasets, different disease outcomes, additional machine learning models, and alternative imbalance handling techniques to assess the generalizability of the findings. Future research should also examine

SHAP ranking stability under more complex data-generating mechanisms, including non-linear relationships, interaction effects among features, and different correlation structures.

ACKNOWLEDGMENTS

The author would like to express sincere gratitude to the Tut Wuri Handayani Scholarship Program, Directorate of Resources, Directorate General of Higher Education, Ministry of Higher Education, Science, and Technology, for the support provided during the author's study. The author also gratefully acknowledges the Department of Statistics and Data Science, School of Data Science, Mathematics, and Informatics, IPB University, for the academic environment and support throughout the research process. Special appreciation is extended to Dr. Bagus Sartono, S.Si., M.Si., and Dr. Septian Rahardiantoro, S.Stat., M.Si., for their valuable guidance, supervision, and constructive suggestions throughout the research process.

BIBLIOGRAPHY

Alshboul, O., Almasabha, G., Shehadeh, A., & Al-Shboul, K. (2024). A comparative study of LightGBM, XGBoost, and GEP models in shear strength management of SFRC-SBWS. *Structures*, 61, 106009. https://doi.org/10.1016/j.istruc.2024.106009

Angulo, E., Romero, F. P., & López-Gómez, J. A. (2021). A comparison of different soft-computing techniques for the evaluation of handball goalkeepers. *Soft Computing*, 26(6), 3045–3058. https://doi.org/10.1007/s00500-021-06440-7

Asif, D., Arif, M. S., & Mukheimer, A. (2025). A data-driven approach with explainable artificial intelligence for customer churn prediction in the telecommunications industry. *Results in Engineering*, 26, 104629. https://doi.org/10.1016/j.rineng.2025.104629

Aswi, A., Rahardiantoro, S., Kurnia, A., Sartono, B., Handayani, D., & Nurwan, N. (2025). Bayesian spatio-temporal conditional autoregressive localized modeling techniques for socioeconomic factors and stunting in Indonesia. *MethodsX*, 15, 103464. https://doi.org/10.1016/j.mex.2025.103464

Ballegeer, M., Bogaert, M., & Benoit, D. F. (2025). Evaluating the stability of model explanations in instance-dependent cost-sensitive credit scoring. *European Journal of Operational Research*, 326(3), 630–640. https://doi.org/10.1016/j.ejor.2025.05.039

Bracke, P., Datta, A., Jung, C., & Sen, S. (2019). Machine Learning Explainability in Finance: An application to default Risk analysis. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3435104

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020). Explainable machine learning in credit risk management. *Computational Economics*, 57(1), 203–216. https://doi.org/10.1007/s10614-020-10042-0

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Dharmawan, H., Sartono, B., Kurnia, A., Hadi, A. F., & Ramadhani, E. (2022). A study of machine learning algorithms to measure the feature importance in class-imbalance data of food insecurity

cases in Indonesia. *Communications in Mathematical Biology and Neuroscience*. https://doi.org/10.28919/cmbn/7636

Dietz, L. W., Sertkan, M., Myftija, S., Palage, S. T., Neidhardt, J., & Wörndl, W. (2022). A Comparative Study of Data-Driven Models for Travel Destination Characterization. *Frontiers in Big Data*, 5, 829939. https://doi.org/10.3389/fdata.2022.829939

Ekstrøm, C. T., Gerds, T. A., & Jensen, A. K. (2018). Sequential rank agreement methods for comparison of ranked lists. *Biostatistics*, 20(4), 582–598. https://doi.org/10.1093/biostatistics/kxy017

Ginasputri, H. N. A., Saputri, A. D., Wahid, S. N., Ningrum, A. F., & Haris, M. A. (2025). Analisis sentimen ulasan pengguna BRIMO terhadap pembaruan fitur aplikasi menggunakan Naive Bayes dengan seleksi fitur Chi-Square. *Jurnal Statistika Dan Komputasi*, 4(2), 56–67. https://doi.org/10.32665/statkom.v4i2.5194

Imani, M., Maroosi, A., Shojaei, S., Heidari, K., Hoseinzadeh, S. M., Daneshi, N., Saber, Z., Sajadi, N., & Mohammadzadeh, M. (2026). Predicting cardiovascular diseases using imbalanced data: An XGBoost-based analysis of the 2022 BRFSS dataset. *American Heart Journal Plus Cardiology Research and Practice*, 63, 100719. https://doi.org/10.1016/j.ahjo.2026.100719

Liu, W., Fan, H., & Xia, M. (2021). Credit scoring based on tree-enhanced gradient boosting decision trees. Expert Systems With Applications, 189, 116034. https://doi.org/10.1016/j.eswa.2021.116034

Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*. https://neurips.cc/Conferences/2017/Schedule?showEvent=9253

Mohammed, A. M. A., Husain, O., Abdulkareem, M., Yunus, N. Z. M., Jamaludin, N., Mutaz, E., Elshafie, H., & Hamdan, M. (2024). Explainable Artificial Intelligence for predicting the compressive strength of soil and ground granulated blast furnace slag mixtures. *Results in Engineering*, 25, 103637. https://doi.org/10.1016/j.rineng.2024.103637

Molnar, Christoph. (2025). Interpretable machine learning : a guide for making black box models explainable (3rd ed.). Christoph Molnar. https://christophm.github.io/interpretable-ml-book

Noviandy, T. R., Maulana, A., Irvanizam, I., Idroes, G. M., Maulydia, N. B., Tallei, T. E., Subianto, M., & Idroes, R. (2025). Interpretable machine learning approach to predict Hepatitis C virus NS5B inhibitor activity using voting-based LightGBM and SHAP. *Intelligent Systems With Applications*, 25, 200481. https://doi.org/10.1016/j.iswa.2025.200481

Óskarsdóttir, M., & Bravo, C. (2021). Multilayer network analysis for improved credit risk prediction. *Omega*, 105, 102520. https://doi.org/10.1016/j.omega.2021.102520

Rahardiantoro, S., Juhanda, A. R. N., Kurnia, A., Aswi, A., Sartono, B., Handayani, D., Soleh, A. M., Yanti, Y., & Cramb, S. (2024). Spatio-temporal modeling to identify factors associated with stunting in Indonesia using a Modified

Generalized Lasso. *Spatial and Spatio-temporal Epidemiology*, 51, 100694. https://doi.org/10.1016/j.sste.2024.100694

Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2024). A perspective on explainable artificial intelligence methods: SHAP and LIME. *Advanced Intelligent Systems*, 7(1). https://doi.org/10.1002/aisy.202400304

Sartono, B., Elenaputri, T. S., Angraini, Y., & Dito, G. A. (2025). Long Short-Term Memory-Based prediction of Indonesian composite stock index returns for early identification of market crises. *Applied Computational Intelligence and Soft Computing*, 2025(1). https://doi.org/10.1155/acis/6174081

Sejling, C., Jensen, A. K., Zhang, J., Loft, S., Andersen, Z. J., Brandt, J., Stayner, L. T., Pedersen, M., & Budtz-Jørgensen, E. (2025). Novel approach for hierarchical family selection of an ambient air pollutant mixture with application to childhood asthma. *Environmetrics*, 36(5). https://doi.org/10.1002/env.70020

Shapley, L. S. (1953). 17. A value for N-Person games. *In Princeton University Press eBooks* (pp. 307–318). https://doi.org/10.1515/9781400881970-018

Zhang, J., & Zhao, Z. (2025). Corporate ESG rating prediction based on XGBoost-SHAP interpretable machine learning model. *Expert Systems With Applications*, 295, 128809. https://doi.org/10.1016/j.eswa.2025.128809

Zhao, C., Yan, Z., Sun, X., & Wu, M. (2024). Enhancing aspect category detection in imbalanced online reviews: An integrated approach using Select-SMOTE and LightGBM. *International Journal of Intelligent Networks*, 5, 364–372. https://doi.org/10.1016/j.ijin.2024.10.002

Zhu, J., Pu, S., He, J., Su, D., Cai, W., Xu, X., & Liu, H. (2024). Processing imbalanced medical data at the data level with assisted-reproduction data as an example. *BioData Mining*, 17(1), 29. https://doi.org/10.1186/s13040-024-00384-y