BAREKENG: Journal of Mathematics and Its Applications
September 2026 Volume 20 Issue 3 Page 2535-2548
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol20iss3pp2535-2548

# COMPARATIVE STUDY OF LIGHTGBM, CATBOOST, AND RANDOM FOREST IN MODELING PUBLIC COMPLAINTS CLASSIFICATION

**Oktaviyani Daswati** [1*], **Hari Wijayanto** [2], **Farit Mochamad Afendi** [3]

[1,2,3]*Department of Statistics and Data Science, School of Data Science, Mathematics, and Informatics, IPB University*
*Jln. Meranti W22, L4 Kampus IPB Dramaga, Bogor, 16680, Indonesia*

*Corresponding author's e-mail: * **vi.daswati@gmail.com***

***Article Info***

***Article History:***

*Received: 11th September 2025*
*Revised: 1st December 2025*
*Accepted: 17th March 2026*
*Published: 8th April 2026*

***Keywords:***

*CatBoost;*
*Classification;*
*LightGBM*
*Public complaints;*
*Random Forest.*

***ABSTRACT***

*Public complaints data on maladministration in Indonesia is a dataset with high-cardinality categorical variables and imbalanced category distributions, posing significant challenges for conventional machine learning algorithms. To address this issue, this study aims to evaluate and compare the performance of three widely used classification algorithms (LightGBM, CatBoost, and Random Forest) on actual public complaint data that has never been analysed using machine learning methods. Hyperparameter tuning was applied to obtain optimal configurations and ensure robust performance. Analysis was conducted using 30 repeated simulations with accuracy and sensitivity as the primary metrics. ANOVA followed by Tukey HSD was used to explicitly determine whether there were differences in performance between models at a 95% confidence level. The results show that LightGBM performed best with an accuracy of 74.50% and a sensitivity of 76.70%, followed by CatBoost with an accuracy of 74.12% and a sensitivity of 75.54%, while Random Forest lagged far behind. Statistical tests confirmed significant performance differences between the three models. This study is not without limitations. Only three classification algorithms were evaluated, encoding strategies were not systematically compared, and the hyperparameter search space was restricted, meaning broader model exploration may yield improved performance. Nonetheless, the study provides originality and value by representing the first empirical application of machine learning to Indonesian public complaint data on maladministration, demonstrating how algorithm selection directly affects predictive outcomes when handling complex categorical structures. The findings offer practical insights for government agencies, highlighting how data-driven models can support policy design, strengthen transparency, and improve the quality of public services.*



***How to cite this article:***

O. Daswati, H. Wijayanto, F.M. Afendi, "COMPARATIVE STUDY OF LIGHTGBM, CATBOOST, AND RANDOM FOREST IN MODELING PUBLIC COMPLAINTS CLASSIFICATION", *BAREKENG: J. Math. & App.*, vol. 20, no. 3, pp. 2535-2548, Sep, 2026.


Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

**Research Article · Open Access**

## 1. INTRODUCTION

Statistical modeling is a fundamental approach in data analysis for representing relationships between variables, identifying patterns, and generating predictions that support data-driven decision-making [1]. Advances in modern statistics have expanded the modeling paradigm from classical parametric models to machine learning techniques that are more flexible in handling large, complex, structured data [2]. In the context of data science, prominent methodological challenges arise when data is dominated by categorical variables with high cardinality and imbalanced class distributions, as commonly found in social data [3], [4].

Public complaints data is a form of social data with such structural complexity. In Indonesia, the Ombudsman of the Republic of Indonesia manages thousands of public reports related to alleged maladministration, most of which are represented by categorical variables, such as report type, source of complaint, and handling status. Conceptually, public complaint data modeling plays an important role not only as a predictive tool but also as an analytical instrument for uncovering systemic patterns in public services, supporting the prioritization of handling, and improving the effectiveness of public service oversight. A number of studies show that the application of machine learning can improve the efficiency and accuracy of public complaint classification. In the JakLapor system in Jakarta, the Random Forest algorithm is reported to be capable of achieving an accuracy rate of up to 90% [5], while research on the LAKSA platform in Tangerang City shows the success of various classification algorithms, including k-Nearest Neighbors, Random Forest, Support Vector Machine, and AdaBoost, in supporting the acceleration of public complaint resolution [6].

In methodological literature, decision tree-based algorithms are widely used to handle data with a predominance of categorical variables. LightGBM was developed to improve computational efficiency through a leaf-based tree growth strategy, with implications for differences in information gain between nodes [7]. CatBoost was designed to handle categorical variables directly through a sequential boosting mechanism [8]. Meanwhile, Random Forest utilizes a bagging approach to improve prediction stability and has been widely used in social studies, including predicting school dropouts [9], modeling consumer loyalty [10], and classifying household poverty status [11], It is a robust, relatively easy-to-interpret method [12]. Comparative studies also show that LightGBM and CatBoost often produce higher AUC and F1 values than classical algorithms on complex social data [13].

Although these three algorithms have been widely applied in various social data contexts, to date, there has been no comprehensive study directly comparing the performance of LightGBM, CatBoost, and Random Forest on public complaints data managed by the Ombudsman of the Republic of Indonesia. This dataset is national administrative data with unique characteristics that have never been systematically modeled using machine learning. Therefore, this study aims to compare the performance of LightGBM, CatBoost, and Random Forest in classifying public complaints related to alleged maladministration. The focus of the study is to evaluate classification performance and determine the most accurate and reliable algorithm, so that the results can serve as a basis for developing an automated complaint classification system within the Ombudsman of the Republic of Indonesia and for strengthening data-driven decision-making in public service oversight.

## 2. RESEARCH METHODS

In the classification process, each public complaint data is transformed into a set of explanatory variables $X = (x_1, x_2, \dots, x_n)$, where the explanatory variables are categorical attributes. The response variable is to assign each complaint to a specific category based on whether maladministration findings are present. Mathematically, the classification model attempts to learn a function $f(X)$ that map the explanatory variables $X$ into the most likely class label. The decision is made by selecting the category with the highest estimated probability, expressed as $\hat{y} = \arg max_c \, P(y = c|X)$. To evaluate the model's reliability, the predicted complaint categories are compared with the actual complaint categories using performance metrics. Accuracy measures the proportion of correct predictions, while sensitivity emphasizes the model's ability to correctly identify specific complaint classes. These mathematical formulations provide a systematic foundation for developing and validating the classification model in public complaint analysis.

### 2.1 Data

The data used in this study are public complaints in Indonesia regarding alleged maladministration at the Ombudsman of the Republic of Indonesia in 2023 and 2024. The data is sourced from an internal database extracted on 28 February 2025. The response variable in this data is a binary class with the categories of there are maladministration findings and no maladministration findings.

Table 1. List of Variables in The Dataset

| Variable | Variable Description | Type of Variable | Characteristics |
|---|---|---|---|
| $Y$ | Findings of maladministration | Categorical | Biner |
| $x_1$ | Group of agencies reported | Categorical | 20 level |
| $x_2$ | Type of area reported | Categorical | 2 level |
| $x_3$ | Substance | Categorical | 38 level |
| $x_4$ | Type of allegation | Categorical | 11 level |
| $x_5$ | Classification of report | Categorical | 3 level |
| $x_6$ | Reporter | Categorical | 2 level |
| $x_7$ | Group of reporter | Categorical | 5 level |
| $x_8$ | Submission method | Categorical | 11 level |
| $x_9$ | Request for confidentiality | Categorical | 2 level |
| $x_{10}$ | Inspection office | Categorical | 35 level |

All variables in the dataset do not contain empty values because every attribute in the Ombudsman database system must be filled in when recording reports, so no empty values appear in the data extraction process. However, some variables with high cardinality, such as the reported institution group (20 levels), substance (38 levels), and inspection office (35 levels), contain categories with very low frequencies. These rare categories are retained because they are considered to represent important empirical conditions and can significantly affect classification patterns, especially in model analyses sensitive to category-level distributions.

### 2.2 LightGBM

LightGBM uses several techniques to improve efficiency and accuracy, namely Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) [14]. GOSS works by discarding most data with small gradients and using the remaining data to calculate information gain, enabling faster computation because observations with larger gradients contribute more significantly to split decisions. EFB reduces dimensionality by combining mutually exclusive features into a single feature bundle, thereby lowering the number of variables processed during training. LightGBM also incorporates native handling of categorical variables through Gradient-based One-Hot Encoding (GHOE), in which categorical values are internally converted into integer representations and optimally grouped based on histogram binning to maximise split gain—avoiding full one-hot expansion and preventing feature explosion in high-cardinality settings. This algorithm uses a leaf-by-leaf tree growth strategy, meaning decision trees are expanded by selecting the leaf that yields the highest gain at each iteration, thereby accelerating training and enabling more optimal split selection.

$$\mathcal{L} = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k). \tag{1}$$

In Eq. (1), $\mathcal{L}$ denotes the objective function minimized by LightGBM. The first term, $\sum_{i=1}^{n} \ell(y_i, \hat{y}_i)$, represents the empirical loss function that measures the discrepancy between the true class label $y_i$ and the predicted value $\hat{y}_i$ for the $i$-th observation. In binary classification problems, this loss function is commonly specified as the logistic loss. The second term, $\sum_{k=1}^{K} \Omega(f_k)$, is a regularization component that penalizes model complexity by controlling the structure of each decision tree $f_k$, such as the number of leaves and the magnitude of leaf weights. This regularization term plays a crucial role in preventing overfitting and improving the model's generalization performance. By minimizing this objective function iteratively using gradient-based optimization, LightGBM constructs an ensemble of decision trees that balances predictive accuracy and model complexity.

### 2.3 CatBoost

CatBoost is an algorithm specifically designed to handle data with categorical features without requiring manual encoding during preprocessing [15]. CatBoost uses a gradient descent approach to minimise the loss function and employs a computational scheme with multiple permutations of the training dataset called ordered boosting, calculating statistics for categorical explanatory variables during training in the tree structure formation stage. CatBoost uses symmetric trees, where each split is performed uniformly across all branches to improve machine learning efficiency [8]. The expected value of the loss function minimised during decision tree construction in Catboost, as shown below.

$$\mathcal{L} = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k). \tag{2}$$

Similar to other Gradient Boosting Decision Tree (GBDT) algorithms, CatBoost minimizes an objective function consisting of two main components, namely the loss function and model regularization, as stated in Eq. (2). In this equation, $\mathcal{L}$ represents the objective function optimized by the model. The first component, $\sum_{i=1}^{n} \ell(y_i, \hat{y}_i)$, represents the loss function that measures the difference between the actual label $y_i$ and the predicted value $\hat{y}_i$ for each observation $i$. In binary classification problems, this loss function is generally a logistic loss. The second component, $\sum_{k=1}^{K} \Omega(f_k)$, is a regularization term that serves to control the complexity of each decision tree $f_k$, thereby helping to prevent overfitting and improve the model's generalization ability. Although mathematically the CatBoost objective function is no different from the general GBDT framework, CatBoost's main advantage lies in its training mechanism, specifically in its use of ordered boosting and direct handling of categorical variables. The ordered boosting approach is designed to reduce prediction bias due to target leakage by ensuring that residual estimates in each iteration depend only on previously available data. Thus, CatBoost can maintain the same objective formulation as GBDT in general while improving model stability and accuracy when applied to data with a predominance of categorical variables.

### 2.4 Random Forest

Random forest builds several independent decision trees from subsets of training data [16]. Random Forest assigns equal weight to each tree and combines predictions through majority voting, while selection of splits in its trees generally uses the Gini impurity criterion. This algorithm forms many decision trees from randomly selected subsets of data. Classification results are calculated from the majority value. In the gradual tree-building stage, random forests use the information gain provided by the largest impurity reduction.

In most machine learning libraries, Random Forest cannot process categorical data directly because the underlying decision tree algorithm expects numerical input. This limitation arises because classic CART trees compare values using numerical thresholds, so categorical data must be converted to numerical representations before they can be used. Therefore, categorical features need to be encoded, such as via one-hot, ordinal, or target encoding, before being fed into a Random Forest. One-hot encoding is often used, but it can lead to a dimensionality explosion for variables with high cardinality, increasing model complexity and the risk of overfitting.

$$\Delta Gini(t) = Gini(t) - \sum_{i \epsilon \{left, right\}} \frac{n_j}{n_t} . Gini(j). \tag{3}$$

In Eq. (3), $\Delta Gini(t)$ represents the reduction in Gini impurity achieved when a parent node $t$ is split into two child nodes, namely the left and right nodes. The term $Gini(t)$ denotes the impurity of the parent node, while $Gini(j)$ is the impurity of child node $j$, weighted by the proportion of observations in that node. The optimal split is determined by maximizing $\Delta Gini(t)$, which corresponds to the greatest decrease in node impurity after the split. In Random Forest, this splitting criterion is applied independently within each decision tree, where each tree is trained on a bootstrap sample of the data and considers only a randomly selected subset of features at each split. This combination of Gini-based splitting, bootstrap aggregation, and random feature selection reduces correlation among trees, resulting in lower variance and improved stability of the ensemble classifier.

## 2.5 Model Evaluation

In this study, model evaluation was conducted using a confusion matrix to calculate performance metrics. The confusion matrix provides a quantitative representation of the distribution of model predictions against actual classes, making it an important tool for measuring the performance of classification algorithms [17]. From this matrix, the five main metrics used are accuracy, sensitivity, specificity, precision, and AUC. Accuracy measures the proportion of correct predictions relative to the total data, reflecting the model's overall accuracy. Sensitivity or true positive rate emphasises the model's ability to correctly detect positive classes, which is particularly important when failure to recognise positive cases can have significant consequences for the effectiveness of public complaint handling. Specificity is the proportion of negative classes correctly recognized by the model (true negative rate), indicating the model's ability to avoid false positives. Meanwhile, precision measures the accuracy of positive class predictions, i.e., the proportion of positive predictions that are actually positive cases. These three metrics complement each other: AUC assesses overall discrimination quality, specificity assesses accuracy in the negative class, and precision assesses the reliability of the model's positive predictions. Meanwhile, the Area Under the Curve (AUC) measures the model's ability to distinguish between positive and negative classes, with higher values indicating better separation across different decision thresholds.

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + False\ Negative + True\ Negative}, \tag{4}$$

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative}, \tag{5}$$

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive}, \tag{6}$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}, \tag{7}$$

$$AUC = \sum_{i-1}^{n-1} (False\ Positive\ Rate_{i+1} - False\ Positive\ Rate_i) \times \frac{(True\ Positive\ Rate_{i+1} + True\ Positive\ Rate_i)}{2}. \tag{8}$$

## 2.6 One-Way Repeated Measures ANOVA

Repeated Measures ANOVA separates variance due to treatment effects from variance due to individual differences and residual error, thereby improving the accuracy of the F-statistic. One-way Repeated Measures ANOVA is a statistical method used to test differences in the means of a response variable across several conditions or treatments, with repeated measurements on the same subjects or observation units. This technique accounts for measurement correlation because each subject receives all treatments, thereby increasing the analysis's power compared to a standard one-way ANOVA. This model is commonly used to evaluate the effects of a method, treatment, or model that is repeated across several experiments on the same dataset.

$$Y_{ij}\ \mu + \alpha_j + s_i + \varepsilon_{ij}. \tag{9}$$

In Eq. (9), each observation is represented as Yij, which is the response value for subject i and condition or treatment j. This model separates the total variation into several components. The first component is μ, which describes the overall mean response value without accounting for differences in conditions or subjects. The treatment effect is captured by alfaaaaa, which is the fixed effect for the jth condition or treatment, showing how each condition deviates from the overall mean. In addition, the model includes SIIIII as a random effect for the ith subject, reflecting natural differences that remain consistent across conditions. Finally, Eij is the residual error for subject i in treatment j, representing random variation that cannot be explained by the overall mean, treatment effects, or individual characteristics. This model allows for more accurate analysis because it accounts for the fact that each subject is measured repeatedly under different conditions.

### 2.6 Procedures

The initial stage of the study is pre-processing and data exploration. At this stage, variable formats are adjusted to suit the algorithm used, and data duplication that could interfere with the analysis is checked and removed. Once the data is ready, exploration is conducted to understand the dataset's characteristics, including the distributions of predictor variables, the proportions of each category in categorical variables, the distributions of target classes, and the identification of potential problems such as class imbalance or categories with low frequencies. The results of this exploration form the basis for determining modeling strategies and interpreting model performance.

Once the data is ready, the dataset is divided into two parts: 70% for training and 30% for testing. This separation is intended to allow the model to be trained on a portion of the data and evaluated on data that has never been seen before, enabling objective performance measurement. Data separation is performed using stratified sampling to ensure that the proportions of both classes in the response variable remain consistent between the training data and the test data. By using a stratified split, the class distribution in both subsets reflects the original data pattern, making model performance assessment more reliable and representative.

The next step is to build classification models using LightGBM, CatBoost, and Random Forest. The training process is performed on the training data using 5-fold cross-validation. This technique divides the training data into five parts: four for training and one for validation. The process is repeated until all parts have been used as validation data. Cross-validation helps reduce the risk of overfitting and provides a more stable estimate of model performance. At this stage, the hyperparameter space used is deliberately made relatively narrow and close to the default configuration. The main reason for this decision is to maintain consistency across models and avoid performance inflation from aggressive tuning, ensuring fairer comparisons and avoiding bias towards models that are more sensitive to parameter optimization. In addition, tuning is performed once at the beginning to obtain a stable base configuration; thereafter, these hyperparameter values are kept constant throughout the simulation. This approach ensures that the performance variations observed in each iteration are due solely to data changes resulting from the random division process, not to tuning variations. After the model is trained using this configuration, it is then used to make predictions on the test data. The prediction results are compared with the actual labels to calculate the model's accuracy. To ensure the model's robustness and consistency, the procedure from data division to accuracy calculation is repeated 30 times. This approach allows researchers to obtain a distribution of accuracy values for each model, enabling a more comprehensive analysis of its stability across data splits.

The final stage is to compare the performance of the three methods based on the accuracy and sensitivity of the three classification models. The comparison is done by evaluating the accuracy results of all iterations using analysis of variance. This analysis will also provide insights into the influence of dataset characteristics, such as the proportion of categories in categorical variables and the balance of target classes. The evaluation results provide an overview of which method is most effective in classifying public complaint data.

## 3. RESULTS AND DISCUSSION

### 3.1 Data Exploration

This study began with an exploration of the public complaint dataset. The dataset used in this study contained 9,795 public complaints received and processed by the Ombudsman of the Republic of Indonesia between 2023 and 2024. This data not only reflects the number of complaints received, but also includes information about the substance of the issues and the agencies reported. A preliminary analysis was conducted to gain a comprehensive understanding of the structure of the explanatory variables in the dataset, including variable types, the number of levels in the categorical variables, and the distribution of complaints within each category. This step is important because the diverse data structure, both in terms of the number of categories and the level of distribution imbalance, can affect the quality of classification results in detecting maladministration findings. For example, some variables have only two categories with a relatively balanced distribution, while others have dozens of categories with dominance at only certain levels. By understanding these distribution patterns, researchers can anticipate potential bias due to data imbalance and determine the appropriate data processing strategy before building a classification model. In addition, this descriptive

analysis provides an initial overview of the role of each variable in explaining variations in public complaints, thereby providing a strong foundation for further analysis of maladministration classification.

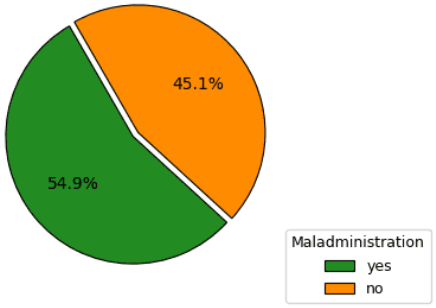


**Figure 1**. Response Variable Class Proportions

Fig. 1 shows the class distribution of the response variable. The response variable is relatively evenly distributed across both response classes, with 54.9% of public complaints reported as cases of maladministration and 45.1% showing no indication of maladministration. This relatively balanced proportion indicates that the data falls into the balanced data category for classification purposes, so no special handling of class imbalance is required before applying machine learning models. However, it is important to note that, although the response variable is balanced, many predictor variables contain highly imbalanced category distributions—some with dominant categories and many rare levels—posing additional modelling challenges, especially for algorithms that do not natively handle high-cardinality categorical features. This condition is an advantage in research because it allows the resulting model to focus on the dataset's patterns and characteristics without additional interventions, such as oversampling or undersampling. In addition, a balanced distribution also increases the reliability of model performance evaluation, because metrics such as accuracy, sensitivity, and AUC can be interpreted more fairly between the two response classes.

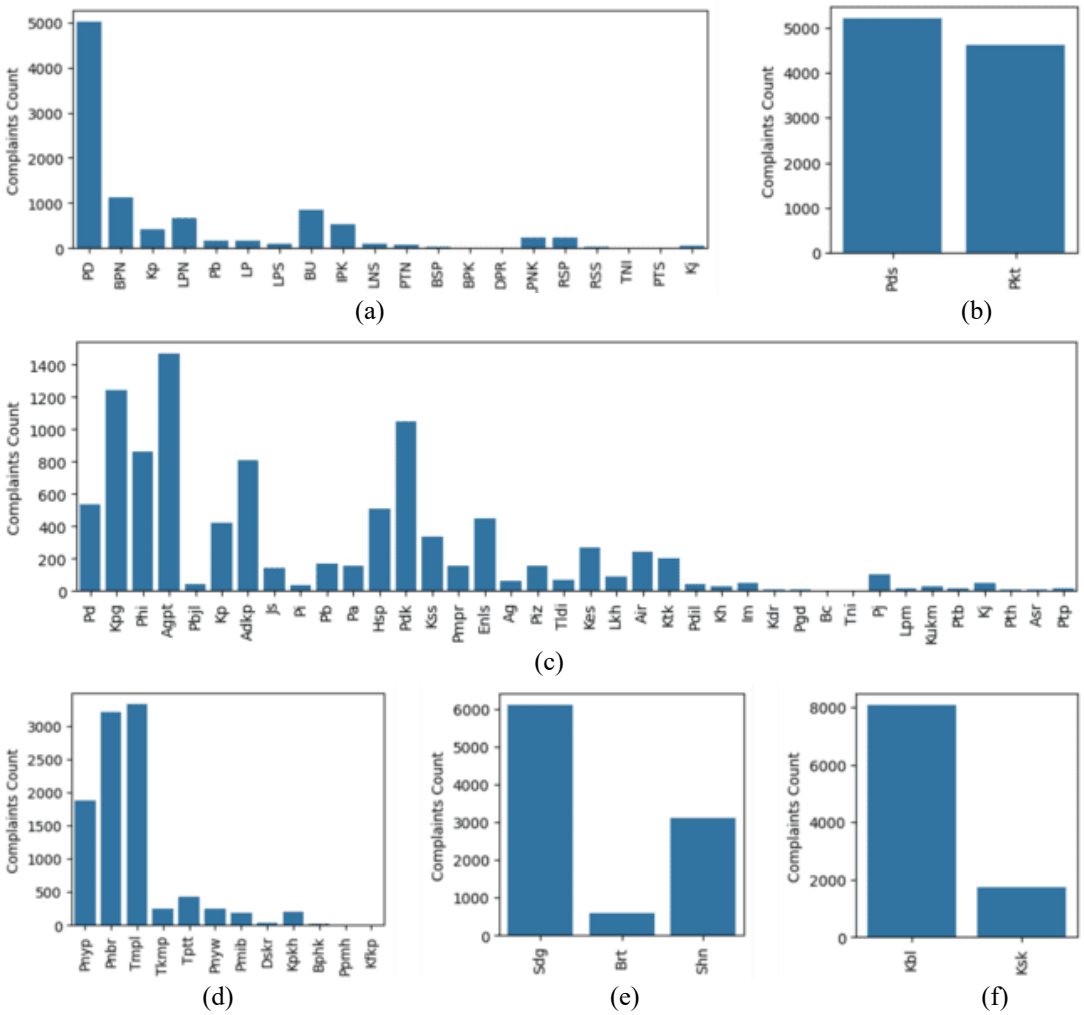


**Figure 2**. Distribution of Explanatory Variables (a) $X_1$, (b) $X_2$, (c) $X_3$, (d) $X_4$, (e) $X_5$, (f) $X_6$

Fig. 2 shows the distribution of complaints across ten categorical variables ($X_1 - X_6$). Variable $X_1$ exhibits a highly imbalanced distribution, with one category accounting for most complaints, while the remaining categories contribute relatively small proportions. Variable $X_2$ consists of two categories with nearly equal frequencies. Variable $X_3$ contains multiple categories with heterogeneous distributions, where a small number of categories account for a large share of complaints, whereas most categories have substantially lower frequencies. A similar pattern is observed for Variable $X_4$, which includes many categories but only a few dominant ones. Variable $X_5$ exhibits a strongly skewed distribution, with one category contributing the majority of complaints, followed by one or two moderately frequent categories and several minor ones. Variable $X_6$ shows a comparable distributional pattern, characterized by a single dominant category and multiple categories with considerably fewer observations.

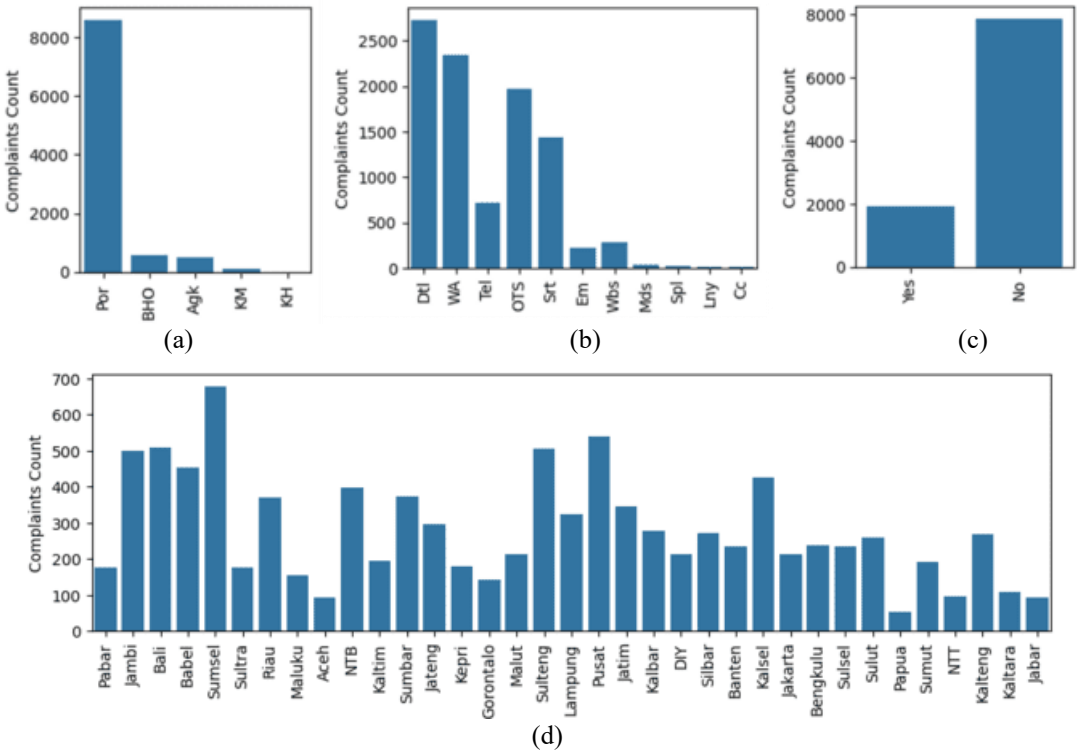


**Figure 3**. Distribution of Explanatory Variables (a) $X_7$, (b) $X_8$, (c) $X_9$, (d) $X_{10}$

Fig. 3 illustrates the distributions of the categorical explanatory variables $X_7 - X_{10}$. Variable $X_7$ exhibits a highly skewed distribution, with one category accounting for the overwhelming majority of complaints, while the remaining categories contribute only negligible counts. Variable $X_8$ shows greater variability across categories, with several categories having moderate frequencies and most others occurring relatively infrequently. In contrast, the binary variable $X_9$ shows a pronounced imbalance between its two categories, with one category having substantially more complaints than the other. Variable $X_{10}$, which represents provincial-level categories, demonstrates a relatively even distribution, with complaint counts spread more uniformly across categories and no single category dominating. Overall, the distributions indicate that most categorical variables exhibit varying degrees of imbalance, a characteristic that may influence analytical results and classification model performance.

Fig. 4 shows the Cramér's V heatmap describing the association structure among the categorical explanatory variables ($X_1 - X_{10}$) and the target variable ($Y$). Most pairwise associations are in the low to moderate range, indicating limited redundancy among predictors and a relatively diverse information set. A small number of variable pairs exhibit moderate-to-strong associations (Cramér's V > 0.5), suggesting partial overlap in information content, which may affect the distribution of feature importance across correlated variables. The associations between individual predictors and the target variable are generally weak to moderate, implying that no single categorical variable dominates the classification outcome. This structure suggests that predictive performance relies on the joint contribution and interactions of multiple variables rather than on isolated effects. From a modeling standpoint, Random Forest is robust to such correlations because it subsamples features at each split, reducing the risk of instability. LightGBM can efficiently handle correlated categorical features through gradient-based tree construction, but may spread split gains across correlated predictors. CatBoost is particularly well-suited to this setting, as its ordered boosting and target-

based encoding help mitigate bias arising from correlated and high-cardinality categorical variables. Overall, the heatmap supports the use of tree-based ensemble models and explains their strong performance on this dataset despite moderate inter-variable associations.

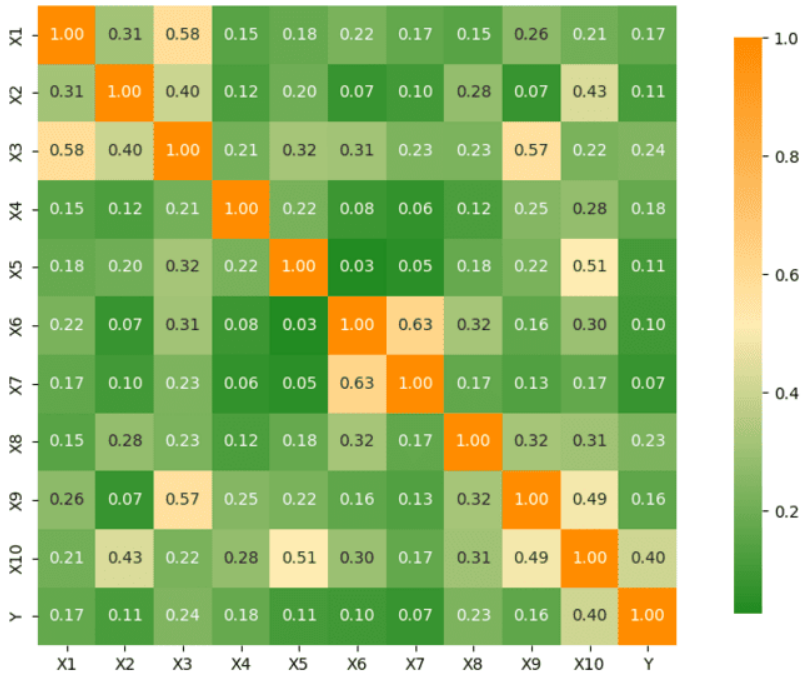


**Figure 4**. Heatmap Cramér's V

The structure of the public complaint dataset used in this study, based on the exploration conducted, consists of categorical explanatory variables, with some variables having many levels. Almost all explanatory variables have a highly imbalanced category distribution. Based on the relationship patterns among variables, the trend in this public complaint data is closer to a non-linear one due to strong interactions among several variables. Therefore, classification methods that are capable of capturing the complexity of interactions and non-linear patterns are very important for obtaining accurate analysis results. Decision tree-based models such as Random Forest, LightGBM, and CatBoost are more suitable because they can effectively handle categorical variables with many levels and accommodate non-linear interactions between variables.

### 3.2 Models

Machine learning modelling was performed by testing models with hyperparameter adjustment. Hyperparameter adjustment was performed to find the best model. The adjustment method is known as hyperparameter tuning. The search space for the best hyperparameters is shown in Table 2.

**Table 2**. The Search Space for Hyperparameters Tuning

| LightGBM | | CatBoost | | Random Forest | |
|---|---|---|---|---|---|
| n_estimators | = 200, 300 | iterations | = 200, 300 | n_estimators | = 400, 500 |
| learning_rate | = 0.05, 0.1 | learning_rate | = 0.05, 0.1 | max_depth | = 8, 10 |
| max_depth | = 6, 8 | depth | = 6, 8 | min_samples_split | = 20, 40 |
| subsample | = 0.8, 0.9 | rsm | = 0.8, 0.9 | max_features | = 0.7, 0.9 |

The best hyperparameter search was conducted using a grid search. For the LightGBM model, the best-performing configuration was n_estimators = 200, learning_rate = 0.1, max_depth = 8, and subsample = 0.8. For the CatBoost model, the parameters were iterations = 300, learning_rate = 0.05, depth = 8, and rsm = 0.9. Meanwhile, for the Random Forest model, the parameters were n_estimators = 500, max_depth = 10, min_samples_split = 20, and max_features = 0.7.

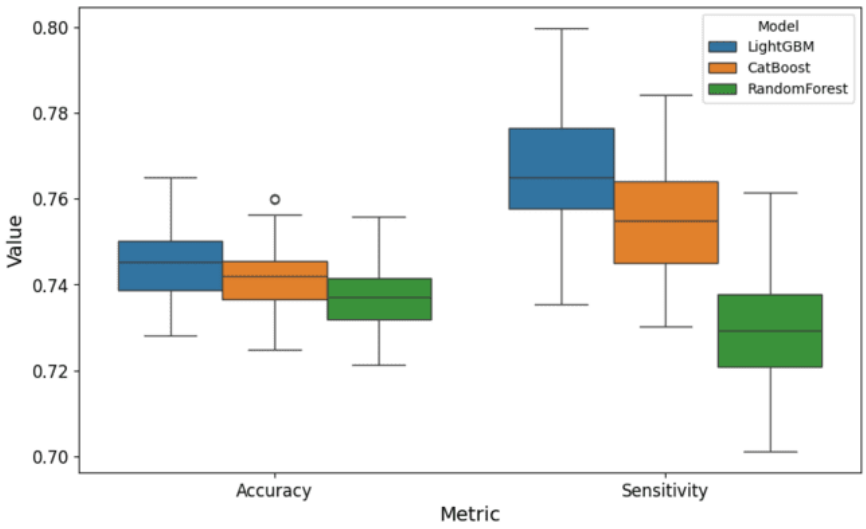


**Figure 5**. Distribution of Performance Evaluation Results

The evaluation was conducted using two main metrics, namely accuracy and sensitivity, with 30 iterations in accordance with the hyperparameters for the best model obtained. The distribution of model performance evaluation values is shown in Fig. 5, indicating that LightGBM consistently achieves the best performance across both accuracy and sensitivity. The median accuracy is the highest, with relatively little variation, indicating stable performance across iterations. CatBoost is slightly below with median accuracy, and sensitivity is slightly lower than LightGBM, although the variation in results remains fairly stable. Meanwhile, Random Forest shows the lowest median accuracy and sensitivity with a wider spread, indicating performance instability in some experiments. Overall, LightGBM excels on both metrics, followed by CatBoost, while Random Forest performs the lowest among the three classification models.

**Table 3**. Average Performance Evaluation Results

| Model | Accuracy | Sensitivity | Specificity | Precision | AUC |
|---|---|---|---|---|---|
| LightGBM | 0.7458 ± 0.0012 | 0.7674 ± 0.0020 | 0.7670 ± 0.0030 | 0.7691 ± 0.0017 | 0.8272 ± 0.0008 |
| CatBoost | 0.7401 ± 0.0010 | 0.7533 ± 0.0021 | 0.7241 ± 0.0027 | 0.7687 ± 0.0015 | 0.8227 ± 0.0008 |
| Random Forest | 0.7364 ± 0.0012 | 0.7288 ± 0.0023 | 0.7457 ± 0.0024 | 0.7772 ± 0.0015 | 0.8167 ± 0.0011 |

Table 3 presents the average performance of LightGBM, CatBoost, and Random Forest in classifying public complaint data, reported as mean ± standard error across repeated experiments. Overall, LightGBM demonstrates the strongest performance, achieving the highest accuracy (0.7458), sensitivity (0.7674), and AUC (0.8272), indicating superior overall discrimination ability and better identification of positive cases. CatBoost shows competitive performance, particularly in precision (0.7687), but lags slightly behind LightGBM in accuracy, sensitivity, and AUC, suggesting that its advantages in handling categorical features do not fully translate into higher predictive power for this dataset. Random Forest attains the highest precision (0.7772), implying a lower false-positive rate when predicting positive cases, but it exhibits the lowest sensitivity and AUC, indicating reduced ability to capture true positive complaints. The relatively small standard errors across all metrics indicate stable model performance over repeated runs. These results suggest that while Random Forest may be preferable when minimizing false positives is critical, LightGBM offers the best overall balance between detection capability and discriminative performance, making it the most suitable model for automated classification of public complaints in the Ombudsman dataset.

Repeated-measures Analysis of variance was performed to test whether there were differences among the three classification models in terms of accuracy and sensitivity. The results of the analysis of variance showed a p-value of less than 0.05, indicating statistical significance. It can be concluded that at a 95% confidence level, there is sufficient evidence to state that at least one classification model has different evaluation values. Further testing was conducted using the Tukey HSD test. The results of the Tukey HSD post hoc test in Table 4 indicate significant differences in performance among the classification models. LightGBM has a significant difference in accuracy compared to CatBoost (mean difference = 0.0057; p-adj = 0.0018), indicating that LightGBM statistically provides higher accuracy. Meanwhile, no significant

difference was found between LightGBM and Random Forest (p-adj = 0.0685), so the null hypothesis for this pair was not rejected. Conversely, CatBoost and Random Forest showed a significant difference (mean difference = −0.0094; p-adj < 0.001), with CatBoost performing better.

**Table 4**. Tukey HSD Test Results

| Group 1 | Group 2 | Meandiff | p-adj | Lower | Upper | $H_0$ |
|---|---|---|---|---|---|---|
| LightGBM | CatBoost | 0.0057 | 0.0018 | 0.0019 | 0.0096 | Reject |
| LightGBM | Random Forest | -0.0036 | 0.0685 | -0.0075 | 0.0002 | Accept |
| CatBoost | Random Forest | -0.0094 | 0.0000 | -0.0132 | -0.0055 | Reject |

After calculating the average performance differences between models, a Tukey HSD post hoc test was conducted to determine whether the differences were statistically significant. The test results showed that the comparison between LightGBM and CatBoost produced a p-value smaller than 0.05, indicating that the difference in performance between the two was significant, even though the metric difference was within a narrow range. Conversely, the comparison between LightGBM and Random Forest yielded a p-value above 0.05, indicating no significant difference in performance between the two based on the test average.

These findings indicate that CatBoost exhibits a statistically different model response compared to LightGBM for complex categorical data, while Random Forest maintains a similar performance to LightGBM. In practice, the Tukey HSD results confirm that CatBoost is more suitable for datasets with many categorical variables or non-linear interactions; LightGBM is in the middle, balancing efficiency and accuracy; and Random Forest is a competitive alternative, especially when the data structure is simpler. However, in simple scenarios, Random Forest's performance can match the two boosting-based models.

These results confirm that LightGBM is the most superior algorithm in terms of accuracy and sensitivity, followed by CatBoost, while Random Forest consistently performs worst across evaluation metrics. LightGBM benefits from a leaf-wise tree growth strategy that captures complex interactions more efficiently than level-wise splitting, enabling higher predictive accuracy on heterogeneous feature spaces [3]. The results also indicate that high-cardinality categorical features strongly influence model performance, where CatBoost remains competitive through ordered boosting and target-based encoding that mitigate target leakage and maintain stability in the presence of skewed category distributions [4]. These findings are consistent with a study conducted by Zhu [18], which states that the Random Forest algorithm struggles with datasets containing high-cardinality categorical variables and unevenly distributed category levels. This occurs because the number of possible candidate splits grows exponentially with the number of categories, making split enumeration computationally expensive and less reliable, leading Random Forest to miss the optimal split and ultimately reducing model accuracy and efficiency [19]. In addition, Random Forest exhibits greater performance variability because it relies on independent trees without iterative error correction, making it more prone to variance under complex data conditions [20]. Overall, the results confirm that boosting-based methods, particularly LightGBM and CatBoost, are more adaptive and robust than bagging when handling categorical data with many levels and uneven distributions.

The modelling results indicate that the LightGBM and CatBoost classification models are indeed more capable of handling categorical variables. Both models also demonstrate good adaptability to more complex data structures. Therefore, CatBoost and LightGBM can be considered as reliable and robust models for application on empirical data with similar characteristics. In general, LightGBM and CatBoost have proven to be robust classification models for complex categorical data structures, such as social and demographic data. Although this study was limited to three algorithms, the findings confirm that selecting a classification algorithm requires careful consideration of the data characteristics. This study also shows that boosting models are more robust in modelling more complex categorical data structures.

## 4. CONCLUSION

This study compares the performance of LightGBM, CatBoost, and Random Forest algorithms in modeling public complaints data related to maladministration, which contains categorical variables. The results show that LightGBM achieved the highest accuracy and sensitivity after 30 iterations, followed by

CatBoost with slightly lower but still competitive performance, while Random Forest showed much weaker predictive performance across all evaluation metrics. Statistical tests confirm that the differences between LightGBM, CatBoost, and Random Forest are significant, supporting the recommendation to use LightGBM as the primary model, given its accuracy and reliability in identifying cases of maladministration, as well as its potential to support data-driven public service improvement strategies. However, this study has several limitations. First, it only evaluates three machine learning models, namely LightGBM, CatBoost, and Random Forest, ignoring other approaches based on boosting, neural networks, or probabilistic methods. Second, the encoding strategy is not explicitly evaluated, which may affect applicability in datasets with different categorical characteristics. Third, the hyperparameter search space was deliberately limited, meaning that better performance may be achievable with more in-depth tuning. Finally, the experiments did not include basic statistical models, limiting the interpretability gap between traditional and modern modeling approaches. Future research should expand the comparison to include a wider range of algorithms—particularly deep learning and hybrid boosting architectures—and explicitly analyze the impact of categorical encoding schemes. Broader hyperparameter optimization, the inclusion of basic classification, and validation using external datasets or cross-agency public complaint repositories would strengthen the generalization and operational relevance of these findings.

## Author Contributions

Oktaviyani Daswati: Conceptualization, Data Curation, Formal Analysis, Methodology, Visualization, Writing – Original Draft. Hari Wijayanto: Conceptualization, Supervision, Validation, Writing – Review and Editing. Farit Mochamad Afendi: Formal Analysis, Supervision, Validation, Writing – Review and Editing. All authors discussed the results and contributed to the final manuscript.

## Funding Statement

This research was funded by the Education Fund Management Institution (LPDP) of the Republic of Indonesia. The publication of this work was supported by the School of Data Science, Mathematics, and Informatics, IPB University.

## Aknowledgement

The authors would like to express their deepest gratitude and appreciation to the Ombudsman of the Republic of Indonesia for the valuable contribution in providing public complaint data, which served as the basis for this study.

## Declarations

The authors declare no conflicts of interest to report.

## Declaration of Generative AI and AI-assisted technologies

Generative AI tools (e.g., ChatGPT) were used solely for language refinement (grammar, spelling, and clarity). The scientific content, analysis, interpretation, and conclusions were developed entirely by the authors. The authors reviewed and approved all final text.

## REFERENCES

[1] S. C. and S. R. Balasundaram, "DATA ANALYSIS IN CONTEXT-BASED STATISTICAL MODELING IN PREDICTIVE ANALYTICS," pp. 96–114, 2021, doi: https://doi.org/10.4018/978-1-7998-3053-5.ch006

[2] X. Wang, X. Y. Lou, S. Y. Hu, and S. C. He, "EVALUATION OF SAFE DRIVING BEHAVIOR OF TRANSPORT VEHICLES BASED ON K-SVM-XGBOOST," in *2020 3rd International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)*, IEEE, pp. 84–92, Apr. 2020, doi: https://doi.org/10.1109/AEMCSE50948.2020.00026

[3] G. Ke *et al.*, "LIGHTGBM: A HIGHLY EFFICIENT GRADIENT BOOSTING DECISION TREE," *Adv Neural Inf Process Syst*, vol. 30, pp. 3146–3154, 2017.

[4] L. Prokhorenkova, G. Gusev, A. Vorobev, A. Dorogush, and A. Gulin, "CATBOOST: UNBIASED BOOSTING WITH CATEGORICAL FEATURES," *Adv Neural Inf Process Syst*, vol. 31, pp. 6638–6648, 2018.

[5] S. M. Intani, B. I. Nasution, M. E. Aminanto, Y. Nugraha, N. Muchtar, and J. I. Kanggrawan, "AUTOMATING PUBLIC COMPLAINT CLASSIFICATION THROUGH JAKLAPOR CHANNEL: A CASE STUDY OF JAKARTA, INDONESIA," in *2022 IEEE International Smart Cities Conference (ISC2)*, IEEE, pp. 1–6, Sep. 2022, doi: https://doi.org/10.1109/ISC255366.2022.9922346

[6] E. D. Madyatmadja, C. P. M. Sianipar, C. Wijaya, and D. J. M. Sembiring, "CLASSIFYING CROWDSOURCED CITIZEN COMPLAINTS THROUGH DATA MINING: ACCURACY TESTING OF K-NEAREST NEIGHBORS, RANDOM FOREST, SUPPORT VECTOR MACHINE, AND ADABOOST," *Informatics*, vol. 10, no. 4, p. 84, Nov. 2023, doi: https://doi.org/10.3390/informatics10040084

[7] 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: https://doi.org/10.3390/math8050765

[8] J. T. Hancock and T. M. Khoshgoftaar, "CATBOOST FOR BIG DATA: AN INTERDISCIPLINARY REVIEW," *J Big Data*, vol. 7, no. 1, p. 94, Dec. 2020, doi: https://doi.org/10.1186/s40537-020-00369-8

[9] D. Setiawan, H. Wijayanto, and L. O. A. Rahman, "BAGGING AND RANDOM FOREST CLASSIFICATION METHODS FOR UNBALANCED DATA SCHOOL DROPOUT CASES IN LAMPUNG PROVINCE," p. 020026, 2022, doi: https://doi.org/10.1063/5.0109130

[10] A. Pratiwi, K. A. Notodiputro, and H. Wijayanto, "PEMODELAN LOYALITAS KONSUMEN SUSU PERTUMBUHAN DALAM MENGIKUTI PROGRAM REWARDS MENGGUNAKAN METODE RANDOM FOREST DAN NEURAL NETWORK," *Xplore: Journal of Statistics*, vol. 2, no. 2, pp. 41–48, Aug. 2018, doi: https://doi.org/10.29244/xplore.v2i2.104

[11] F. Izzati, M. Masjkur, and F. M. Afendi, "COMPARISON OF CHI-SQUARE AUTOMATIC INTERACTION DETECTOR (CHAID) AND RANDOM FOREST METHODS IN THE CLASSIFICATION OF HOUSEHOLD POVERTY STATUS IN CENTRAL JAVA," *Indonesian Journal of Statistics and Its Applications*, vol. 8, no. 1, pp. 1–13, Jun. 2024. https://doi.org/10.29244/ijsa.v8i1p1-13

[12] T.-H. Lee, A. Ullah, and R. Wang, "BOOTSTRAP AGGREGATING AND RANDOM FOREST," 2020, pp. 389–429. https://doi.org/10.1007/978-3-030-31150-6_13

[13] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A COMPARATIVE ANALYSIS OF GRADIENT BOOSTING ALGORITHMS," *Artif Intell Rev*, vol. 54, no. 3, pp. 1937–1967, Mar. 2021, doi: https://doi.org/10.1007/s10462-020-09896-5

[14] D. Zhang and Y. Gong, "THE COMPARISON OF LIGHTGBM AND XGBOOST COUPLING FACTOR ANALYSIS AND PREDIAGNOSIS OF ACUTE LIVER FAILURE," *IEEE Access*, vol. 8, pp. 220990–221003, 2020, doi: https://doi.org/10.1109/ACCESS.2020.3042848

[15] A. V. Dorogush, V. Ershov, and A. Gulin, "CATBOOST: GRADIENT BOOSTING WITH CATEGORICAL FEATURES SUPPORT," *ArXiv*, vol. abs/1810.11363, 2018.

[16] H. A. Salman, A. Kalakech, and A. Steiti, "RANDOM FOREST ALGORITHM OVERVIEW," *Babylonian Journal of Machine Learning*, vol. 2024, pp. 69–79, Jun. 2024, doi: https://doi.org/10.58496/BJML/2024/007

[17] M. Heydarian, T. E. Doyle, and R. Samavi, "MLCM: MULTI-LABEL CONFUSION MATRIX," *IEEE Access*, vol. 10, pp. 19083–19095, 2022, doi: https://doi.org/10.1109/ACCESS.2022.3151048

[18] T. Zhu, "ANALYSIS ON THE APPLICABILITY OF THE RANDOM FOREST," *J Phys Conf Ser*, vol. 1607, no. 1, p. 012123, Aug. 2020, doi: https://doi.org/10.1088/1742-6596/1607/1/012123

[19] M. N. Wright and I. R. König, "SPLITTING ON CATEGORICAL PREDICTORS IN RANDOM FORESTS," *PeerJ*, vol. 7, p. e6339, Feb. 2019, doi: https://doi.org/10.7717/peerj.6339

[20] G. Biau, B. Cadre, and L. Rouvière, "ACCELERATED GRADIENT BOOSTING," *Mach Learn*, vol. 108, no. 6, pp. 971–992, Jun. 2019, doi: https://doi.org/10.1007/s10994-019-05787-1