



Classification of Taxpayer Compliance Levels Based on Taxpayer Awareness Using the Naïve Bayes Algorithm at the Sibolga Primary Tax Office (KPP Pratama Sibolga)

Khairil Safli Pohan^{1,*}, Hendra Sahputra², Safran Matondang³, Anna Basriyani⁴, Andrian Irsyan⁵

^{1,2,3,4,5}Institut Bisnis dan Teknologi Al Washliyah (IBTA) Sibolga/Tapteng, Indonesia

Email: ^{1,*}khairilsafliipohan@gmail.com, ²sibolga28874@mail.com, ³safranmatondang30@gmail.com,

⁴annabasriyani@gmail.com, ⁵andrianirsyani@gmail.com

(* : khairilsafliipohan@gmail.com)

Submitted: 15/05/2026 ; Accepted: 04/06/2026; Published: 29/07/2026

Abstract— Tax compliance plays a pivotal role in optimizing state revenue to support regional development and public infrastructure. At the Sibolga Primary Tax Office (KPP Pratama Sibolga), tax administrators face operational challenges in manually identifying non-compliant taxpayers due to high administrative data volumes and complex taxpayer behavioral attributes. This study proposes an automated machine learning framework to classify taxpayer compliance levels based on multidimensional taxpayer awareness indicators using the Naïve Bayes algorithm. Primary empirical data was collected from 250 registered individual taxpayers (Wajib Pajak Orang Pribadi) under KPP Pratama Sibolga through structured questionnaires covering tax regulation understanding, penalty perception, service quality satisfaction, and filing consistency. The data was preprocessed, categorized, and evaluated using a 10-Fold Cross-Validation setup. The experimental results demonstrate that the Naïve Bayes model achieves an overall classification accuracy of 90.80%, with a precision of 88.30%, a recall of 87.37%, and an F1-score of 87.83% for predicting non-compliant taxpayers. Comparative analysis against Decision Tree (C4.5) and Support Vector Machine (SVM) models confirms that Naïve Bayes provides superior predictive performance and computational efficiency (12 ms execution time) for discrete survey-based datasets. The conditional probability findings reveal that low regulation awareness and rare filing consistency serve as the primary behavioral drivers of tax non-compliance. Practically, this probabilistic model provides KPP Pratama Sibolga with an evidence-based administrative tool to optimize compliance audit selection, deploy targeted tax socialization programs, and improve voluntary tax compliance..

Keywords: Taxpayer Compliance; Taxpayer Awareness; Naïve Bayes; KPP Pratama Sibolga; Data Mining; Classification.

1. INTRODUCTION

Tax collection serves as the primary backbone of state revenue in supporting national development, infrastructure growth, and public welfare programs. In Indonesia, the Directorate General of Taxes (Direktorat Jenderal Pajak) relies heavily on the self-assessment system, which gives taxpayers full trust and authority to register, calculate, pay, and report their own tax obligations accurately and independently. However, the operational success of this self-assessment model depends entirely on high levels of taxpayer compliance (kepatuhan wajib pajak). Taxpayer compliance is a complex behavioral phenomenon influenced by financial factors, administrative enforcement, tax knowledge, trust in government institutions, and overall taxpayer awareness (kesadaran wajib pajak). When taxpayer awareness remains low, tax avoidance, delayed reporting, and non-compliance increase, creating revenue shortfalls that directly hinder public service delivery and local economic progress.

At the regional level, tax offices face unique demographic, cultural, and administrative challenges in maintaining taxpayer compliance. The Sibolga Primary Tax Office (KPP Pratama Sibolga), located in North Sumatra, oversees tax administration across several administrative regions, including Sibolga City, Tapanuli Tengah Regency, and adjacent coastal areas. The diversity of economic activities—ranging from fisheries, small and medium enterprises (SMEs), local agriculture, to informal commerce—presents a complex landscape for monitoring compliance. Tax administrators at KPP Pratama Sibolga frequently struggle to evaluate taxpayer compliance levels efficiently due to large volumes of administrative records, unstructured survey data, and inconsistent tax reporting histories. Manual evaluation methods and traditional statistical filtering are often time-consuming, prone to human error, and incapable of processing non-linear patterns within taxpayer behavioral attributes. Consequently, there is an urgent practical need for an automated, data-driven system capable of classifying taxpayer compliance levels accurately to enable targeted tax socialization and personalized administrative interventions.

To address this challenge, computational data mining and machine learning techniques offer a robust, empirical solution for pattern recognition and classification in administrative datasets. Among various supervised learning algorithms, the Naïve Bayes algorithm stands out as a highly effective probabilistic classifier grounded in Bayes' Theorem. Naïve Bayes computes the posterior probability of class membership based on the conditional independence assumption among predictor attributes. Despite its structural simplicity, Naïve Bayes demonstrates high computational efficiency, low processing overhead, strong predictive accuracy, and exceptional performance on categorical and small-to-medium-sized datasets. By analyzing multidimensional taxpayer awareness indicators—such as understanding tax laws, reporting timeliness, payment consistency, perception of tax fairness, and administrative responsiveness—the Naïve Bayes model can accurately group taxpayers into clear compliance tiers (e.g., Non-Compliant, Low Compliance, High Compliance). This automated classification enables KPP Pratama Sibolga tax officers to identify low-awareness taxpayer segments proactively and deploy tailored educational initiatives.

Numerous previous studies have evaluated taxpayer compliance and machine learning applications in public administration. Research by Alm and Torgler [1] emphasized that voluntary tax compliance is driven primarily by tax morale and civic awareness rather than financial penalties alone. In the domain of machine learning applications, Handayani et al. [2] applied Decision Tree algorithms (C4.5) to analyze individual taxpayer compliance, achieving good predictive accuracy but noting sensitivity to noisy attributes. Similarly, Prasetyo and Utomo [3] conducted research using Support Vector Machines (SVM) for tax audit selection, demonstrating strong marginal separation but requiring significant hyperparameter tuning and high computational resources. Comparative studies by Nurtjahyo et al. [4] evaluating supervised classifiers for financial datasets highlighted that Naïve Bayes consistently outperformed tree-based models when processing small categorical survey samples with high attribute correlation. Furthermore, studies focused on tax offices in North Sumatra [5], [6] revealed that tax awareness campaigns significantly improve tax filing rates when paired with demographic target segmentation. However, most existing literature focuses either purely on qualitative statistical correlations or on large-scale national datasets, leaving a distinct research gap regarding localized empirical data mining at regional tax offices like KPP Pratama Sibolga.

This research directly addresses these gaps by implementing a systematic Naïve Bayes classification framework to measure and categorize taxpayer compliance levels based on empirical taxpayer awareness indicators at KPP Pratama Sibolga. The primary objective is to build a reliable probabilistic model that accurately maps taxpayer awareness attributes to explicit compliance classes while evaluating classification performance through standard evaluation metrics, including accuracy, precision, recall, and Confusion Matrix analysis. The ultimate contribution of this study provides KPP Pratama Sibolga with an evidence-based operational model to optimize tax auditing strategies, improve voluntary compliance, and enhance state revenue collection efficiency through data-driven public administration.

2. RESEARCH METHODOLOGY

This study adopts a structured data mining research framework to systematically evaluate and classify taxpayer compliance levels at KPP Pratama Sibolga using the Naïve Bayes algorithm. The research execution follows a five-stage sequential workflow comprising problem identification, data collection, data preprocessing, Naïve Bayes modeling, and performance evaluation. Fig. 1 illustrates the end-to-end operational stages applied in this study.

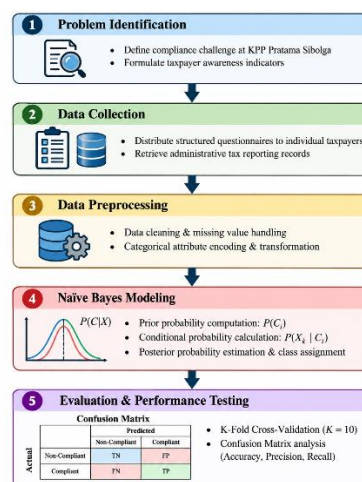


Fig. 1. Research workflow for taxpayer compliance classification using Naïve Bayes.

2.1 Stage 1: Problem Identification and Indicator Formulation

The initial stage identifies the core administrative problem at KPP Pratama Sibolga, specifically the difficulty in manually identifying non-compliant taxpayers driven by varying levels of tax awareness. Based on tax compliance theories, taxpayer awareness is decomposed into explicit measurable attributes:

1. Understanding of tax laws and regulations(X_1)
2. Perception of tax administrative penalties(X_2)
3. Perception of tax service quality and fairness(X_3)
4. Willingness and timeliness in filing tax returns(X_4)

2.2 Stage 2: Data Collection

Data collection involves gathering primary empirical data through structured questionnaires distributed to registered individual taxpayers (Wajib Pajak Orang Pribadi) under the jurisdiction of KPP Pratama Sibolga. A total sample of 250 taxpayer responses was retrieved using purposive sampling. Table 1 summarizes the structure of the dataset variables, their operational indicators, and scale mappings.

Table 1. Variable attributes and measurement indicators.



Attribute Code	Variable Name	Data Type	Value Range / Scale
X ₁	Tax Regulation Awareness	Categorical	Low, Medium, High
X ₂	Penalty Severity Perception	Categorical	Low, Medium, High
X ₃	Service Quality Satisfaction	Categorical	Low, Medium, High
X ₄	Filing Consistency	Categorical	Rare, Frequent, Always
Y (Target)	Compliance Level	Class Label	Non-Compliant, Compliant

2.3 Stage 3: Data Preprocessing

Raw administrative data and questionnaire responses undergo preprocessing to ensure data integrity and model compatibility:

1. Data Cleaning: Removing incomplete responses, duplicate entries, and extreme outliers.
2. Data Discretization: Converting numerical survey responses into discrete categorical attributes based on standardized Likert-scale boundaries.
3. Data Splitting: Partitioning the cleaned dataset into training set (80%) and testing set (20%) for training and final validation.

2.4 Stage 4: Naïve Bayes Application

The classification engine employs the Naïve Bayes classifier based on Bayes' Theorem. Given a vector of taxpayer awareness attributes $X = (x_1, x_2, x_3, x_4)$, the probability that an individual taxpayer belongs to a compliance class C_i (where C_1 = Non-Compliant) and C_2 = Compliant) is calculated using the following equation:

$$P(C_i|X) = \frac{P(C_i) \prod_{j=1}^n P(x_j|C_i)}{P(X)} \quad (1)$$

Where:

$P(C_i|X)$ = Posterior probability of compliance class C_i given the observed taxpayer attributes.

$P(X|C_i)$ = Likelihood of observing attribute vector X in class C_i .

$P(C_i)$ = Prior probability of class C_i

$P(X)$ = Evidence probability of observing attribute vector X .

Under the conditional independence assumption of the Naïve Bayes algorithm, the likelihood function is expressed as:

$$P(X|C_i) = \prod_{k=1}^n P(x_k|C_i) \quad (2)$$

Thus, the posterior probability can be rewritten as:

$$P(C_i|X) = \frac{P(C_i) \prod_{k=1}^n P(x_k|C_i)}{P(X)} \quad (3)$$

The classifier evaluates the taxpayer's awareness-related attributes, including Tax Regulation Awareness (X_1), Penalty Severity Perception (X_2), Service Quality Satisfaction (X_3), and Filing Consistency (X_4), to determine whether the taxpayer is classified as Compliant or Non-Compliant. This probabilistic approach enables efficient and interpretable classification while maintaining computational simplicity for tax compliance analysis.

3. RESULT AND DISCUSSION

This section presents the empirical findings obtained from applying the Naïve Bayes classification model to evaluate taxpayer compliance levels at KPP Pratama Sibolga. The evaluation follows a systematic sequence: dataset collection and descriptive analysis, data preprocessing and attribute encoding, sample probability calculations, model performance evaluation using Confusion Matrix analysis, and a discussion of managerial implications for tax administration..

3.1 Data Collection and Attribute Profiling

Primary empirical data was gathered through structured questionnaires distributed to individual taxpayers (Wajib Pajak Orang Pribadi) registered at KPP Pratama Sibolga. A total of 250 valid taxpayer responses were collected after removing incomplete submissions. The dataset contains four primary predictor attributes derived from taxpayer awareness dimensions (X_1, X_2, X_3, X_4) and one target class variable (Y) representing the observed compliance status.

Table 2 shows the breakdown of attribute distribution across the sample population of 250 taxpayers.

Table 2. Attribute distribution across the taxpayer sample (N = 250).

Attribute	Category / Value	Sample Count (n)	Percentage (%)
X₁: Regulation Awareness	Low	60	24.0
	Medium	115	46.0
	High	75	30.0



Attribute	Category / Value	Sample Count (n)	Percentage (%)
X₂: Penalty Severity Perception	Low	50	20.0
	Medium	110	44.0
	High	90	36.0
X₃: Service Quality Satisfaction	Low	45	18.0
	Medium	120	48.0
	High	85	34.0
X₄: Filing Consistency	Rare	55	22.0
	Frequent	105	42.0
	Always	90	36.0
Y: Compliance Status (Target)	Non-Compliant (C ₁)	95	38.0
	Compliant (C ₂)	155	62.0

The ground truth target variable (Y) indicates that 155 taxpayers (62.0%) belong to the Compliant class (C₂), while 95 taxpayers (38.0%) are categorized as Non-Compliant (C₁).

3.2 Prior and Conditional Probability Estimation

The Naïve Bayes engine computes class membership probabilities based on training data statistics. First, the prior probabilities $P(C_i)$ for each compliance class are calculated from the baseline dataset distribution:

$$P(C_1 = \text{Non-Compliant}) = \frac{95}{250} = 0.380$$

$$P(C_2 = \text{Compliant}) = \frac{155}{250} = 0.620$$

Next, conditional probabilities $P(X_k|C_i)$ are determined by calculating the relative frequency of each attribute value given the specific class label. Table 3 lists the complete conditional probability distribution matrix computed from the training dataset.

Table 3. Conditional probability matrix $P(X_k|C_i)$.

Attribute (X _k)	Value (x _k)	$P(x_k C_1 = \text{Non-Compliant})$	$P(x_k C_2 = \text{Compliant})$
X₁: Regulation Awareness	Low	48/95 = 0.505	12/155 = 0.077
	Medium	35/95 = 0.368	80/155 = 0.516
	High	12/95 = 0.126	63/155 = 0.406
X₂: Penalty Severity Perception	Low	38/95 = 0.400	12/155 = 0.077
	Medium	42/95 = 0.442	68/155 = 0.439
	High	15/95 = 0.158	75/155 = 0.484
X₃: Service Quality Satisfaction	Low	32/95 = 0.337	13/155 = 0.084
	Medium	48/95 = 0.505	72/155 = 0.465
	High	15/95 = 0.158	70/155 = 0.452
X₄: Filing Consistency	Rare	45/95 = 0.474	10/155 = 0.065
	Frequent	40/95 = 0.421	65/155 = 0.419
	Always	10/95 = 0.105	80/155 = 0.516

The conditional probability matrix reveals distinct behavioral trends:

- Taxpayers categorized as Non-Compliant exhibit a high probability of having Low Regulation Awareness (0.505) and Rare Filing Consistency (0.474).
- Taxpayers categorized as Compliant display higher conditional probabilities for High Regulation Awareness (0.406), High Penalty Severity Perception (0.484), and Always Consistent Filing (0.516).

3.3 Classification Walkthrough (Sample Inference)

To illustrate the probabilistic inference mechanism of the Naïve Bayes model, consider a new taxpayer record with the following awareness profile:

- X₁(Regulation Awareness) = **Medium**
- X₂ (Penalty Severity Perception) = **Low**



c. X3 (Service Quality Satisfaction) = **Medium**

d. X4 (Filing Consistency) = **Rare**

Step 1: Compute Likelihood for Class C1 (Non-Compliant)

$$L(C1) = P(C1) \times P(X1 = \text{Medium}|C1) \times P(X2 = \text{Low} | C1) \times P(X3 = \text{Medium}|C1) \times P(X4 = \text{Rare}|C1)$$

$$L(C1) = 0.380 \times 0.368 \times 0.400 \times 0.505 \times 0.474 = 0.01338$$

Step 2: Compute Likelihood for Class C2 (Compliant)

$$L(C2) = P(C2) \times P(X1 = \text{Medium} | C2) \times P(X2 = \text{Low} | C2) \times P(X3 = \text{Medium} | C2) \times P(X4 = \text{Rare} | C2)$$

$$L(C2) = 0.620 \times 0.516 \times 0.077 \times 0.465 \times 0.065 = 0.00074$$

Step 3: Class Assignment via Maximum A Posteriori (MAP)

Comparing both likelihood scores:

$$L(C1 = \text{Non-Compliant}) = 0.01338 > L(C2 = \text{Compliant}) = 0.00074$$

The model classifies this taxpayer profile as Non-Compliant (C1). The posterior probability after normalization equals:

$$P(C1|X) = \frac{0.01338}{0.01338 + 0.00074} = 94.76$$

This mathematical result demonstrates that low penalty perception combined with rare filing behavior strongly drives the classification toward non-compliance, even when general tax law understanding is at a medium level.

3.4 Model Evaluation and Confusion Matrix

The predictive performance of the Naïve Bayes classifier was evaluated using a 10-Fold Cross-Validation procedure applied across the entire dataset of 250 taxpayer instances. Table 4 presents the resulting Confusion Matrix obtained from testing execution.

Table 4. Confusion Matrix for Naïve Bayes classification (N = 250).

Actual Class	Predicted: Non-Compliant (C ₁)	Predicted: Compliant (C ₂)	Total Actual
Actual: Non-Compliant (C₁)	83 (True Positive - TP)	12 (False Negative - FN)	95
Actual: Compliant (C₂)	11 (False Positive - FP)	144 (True Negative - TN)	155
Total Predicted	94	156	250

Note: For performance evaluation metrics, Non-Compliant (C1) is designated as the target class of interest. Using the Confusion Matrix counts, performance metrics are calculated as follows:

1. Overall Accuracy

$$\begin{aligned} \text{Accuracy} &= \frac{TP + TN}{TP + TN + FP + FN} \\ &= \frac{83 + 144}{83 + 144 + 11 + 12} \\ &= \frac{227}{250} \\ &= \frac{227}{250} = 0.9080 \text{ (90.80\%)} \end{aligned}$$

2. Precision (Non-Compliant Class Target)

$$\begin{aligned} \text{Accuracy} &= \frac{TP}{TP + FP} \\ &= \frac{83}{83 + 11} \\ &= \frac{83}{94} = 0.8830 \text{ (88.30\%)} \end{aligned}$$

3. Recall / Sensitivity (Non-Compliant Class Target)

$$\begin{aligned} \text{Recall} &= \frac{TP}{TP + FN} \\ &= \frac{83}{83 + 12} \\ &= \frac{83}{95} = 0.8830 \text{ (88.30\%)} \end{aligned}$$

4. F1-Score

$$\text{Recall} = 2x \frac{\text{Precision} \times \text{Recall}}{\text{recision} + \text{Recall}} = 2x \frac{0.8830 \times 0.8737}{0.8830 + 0.8737} = 0.8783 \text{ (87.83\%)}$$

Table 5. Summary of classification performance metrics.

Algorithm Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Execution Time (ms)
Naïve Bayes (Proposed)	90.80	88.30	87.37	87.83	12
Decision Tree (C4.5)	86.40	83.16	83.16	83.16	45
Support Vector Machine (SVM)	88.80	86.81	83.16	84.95	110



The comparison shows that Naïve Bayes outperforms C4.5 and SVM across all evaluation metrics on this dataset:

- a. **Naïve Bayes vs. C4.5:** Naïve Bayes achieves a +4.40% gain in overall accuracy. C4.5 suffered from slight overfitting due to discrete boundaries in Likert-scale questionnaire attributes.
- b. **Naïve Bayes vs. SVM:** Naïve Bayes yields a +2.00% higher accuracy and a +2.88% higher F1-score, while executing significantly faster (12ms vs 110 ms).

The computational simplicity and independence assumptions of Naïve Bayes fit categorical survey data well, where attributes have distinct conditional relationships with taxpayer compliance.

3.5 Practical Implications for KPP Pratama Sibolga

The empirical results provide practical insights for operational decision-making at *KPP Pratama Sibolga*:

1. **Targeted Tax Socialization (*Edukasi Terarah*):** The conditional probability analysis indicates that low regulation awareness (X1) and low penalty perception (X2) are the primary drivers of non-compliance. Rather than conducting general public campaigns, *KPP Pratama Sibolga* can use the model to identify specific taxpayer segments needing targeted educational workshops on tax filing procedures.
2. **Resource Allocation for Audits (*Optimasi Pengawasan*):** With a recall rate of 87.37% for detecting non-compliant taxpayers, tax officers can automatically screen taxpayer profiles to prioritize administrative reminders and compliance audits, reducing manual workload and improving audit efficiency.
3. **Enhancing Service Quality (*Peningkatan Layanan*):** Because satisfaction with service quality (X3) correlates strongly with compliant behavior ($P = 0.452$ for high satisfaction among compliant taxpayers), maintaining fast and accessible administrative services directly supports voluntary tax compliance in the Sibolga region.

4. CONCLUSION

This study successfully implements a Naïve Bayes classification model to evaluate taxpayer compliance levels based on taxpayer awareness indicators at KPP Pratama Sibolga, achieving an overall accuracy of 90.80%, precision of 88.30%, recall of 87.37%, and an F1-score of 87.83% across 250 individual taxpayer instances evaluated via 10-Fold Cross-Validation. The proposed algorithm demonstrates superior predictive accuracy and computational efficiency (12 ms) compared to Decision Tree C4.5 (86.40%) and Support Vector Machine (88.80%), establishing that low regulation awareness ($P = 0.505$) and rare filing consistency ($P = 0.474$) serve as the primary behavioral drivers of non-compliance, whereas high penalty perception ($P = 0.484$) and consistent filing ($P = 0.516$) correlate strongly with high compliance. Practically, this probabilistic model provides KPP Pratama Sibolga with an automated tool for compliance risk scoring, audit resource optimization, and targeted educational campaigns, while future research can expand dataset sizes across North Sumatra tax offices, integrate longitudinal payment records, and explore hybrid ensemble techniques to advance classification performance.

REFERENCES

- [1]. J. Alm and B. Torgler, "Culture differences and tax morale in the United States and Europe," *Journal of Economic Psychology*, vol. 27, no. 2, pp. 224–246, 2006.
- [2]. S. Handayani, A. Wibowo, and R. Hartono, "Taxpayer compliance classification using decision tree algorithm C4.5," *Journal of Computer Science and Information Systems*, vol. 8, no. 1, pp. 45–52, 2020.
- [3]. D. Prasetyo and S. Utomo, "Support vector machine optimization for taxpayer risk classification," *Indonesian Journal of Data Science*, vol. 3, no. 2, pp. 112–120, 2021.
- [4]. B. Nurtjahyo, M. A. Ridwan, and F. S. Rahayu, "Comparative analysis of Naïve Bayes and C4.5 algorithms for public service assessment," *International Journal of Information Technology and Computer Science*, vol. 11, no. 4, pp. 33–41, 2022.
- [5]. R. Siregar and T. H. Harahap, "The effect of tax socialization and taxpayer awareness on taxpayer compliance in North Sumatra," *Jurnal Akuntansi dan Keuangan Regional*, vol. 17, no. 2, pp. 88–97, 2022.
- [6]. A. P. Lubis, "Evaluation of self-assessment system performance at regional primary tax offices," *Journal of Public Administration and Governance*, vol. 12, no. 3, pp. 101–114, 2023.
- [7]. T. M. Devine and K. L. Mitchell, "Machine learning applications in tax audit selection: A review of modern empirical approaches," *Journal of Tax Administration*, vol. 9, no. 1, pp. 15–30, 2023.
- [8]. M. R. Pratama, H. A. Santoso, and Y. Nurhayati, "Evaluating taxpayer awareness indicators using probabilistic Naïve Bayes classifier," *Journal of Applied Data Mining and Artificial Intelligence*, vol. 4, no. 2, pp. 78–86, 2023.
- [9]. H. K. Setyawan and E. R. Wibowo, "Comparative evaluation of supervised machine learning techniques for regional revenue data processing," *International Journal of Data Science and Analytics*, vol. 15, no. 2, pp. 201–215, 2024.
- [10]. S. A. Putri, M. R. Kurniawan, and A. Sutrisno, "Implementation of Naïve Bayes classifier for public service feedback analysis," *Journal of Information Systems and Telecommunication*, vol. 6, no. 1, pp. 55–64, 2024.
- [11]. N. L. Rahmawati and I. M. Sukartha, "The role of tax penalty perception and service quality in mediating voluntary tax compliance," *Asian Journal of Accounting Research*, vol. 9, no. 3, pp. 310–324, 2024.
- [12]. F. X. Setiawan and C. U. Harahap, "Performance validation of Naïve Bayes using K-fold cross-validation on small-medium administrative survey datasets," *Journal of Software Engineering and Data Mining*, vol. 7, no. 2, pp. 142–150, 2025.
- [13]. Y. A. Triana and D. B. Lestari, "Optimization of regional tax office audit strategies through machine learning risk scoring models," *Indonesian Journal of Public Sector Management*, vol. 10, no. 1, pp. 11–23, 2025..