Terbit *online* pada laman : http://teknosi.fti.unand.ac.id/

**Jurnal Nasional Teknologi dan Sistem Informasi**

| ISSN (Print) 2460-3465 | ISSN (Online) 2476-8812 |

Research Article

# Classification Analysis of Market Service Levy Payment Compliance in South Tangerang Using Random Forest Algorithm

*Tiara Diba [a], Eri Rustamaji [a], Eva Khudzaeva [a,*], Nida'ul Hasanati [a], Nia Kumaladewi [a]*

[a] *Department of Information Systems, Faculty of Science and Technology, Syarif Hidayatullah Jakarta University, South Tangerang, Indonesia*

ARTICLE INFORMATION

*Sejarah Artikel:*
Diterima Redaksi: 03 Juni 2026
Revisi Akhir: 27 April 2026
Diterbitkan *Online*: 23 Agustus 2026

KEYWORDS

CRISP-DM,
Classification,
Machine Learning,
Random Forest

CORRESPONDENCE

E-mail: khudzaeva@gmail.com*

**A B S T R A C T**

Market service levies are among the primary sources of Local Own-Source Revenue (PAD) used to develop and revitalize markets and improve comfort, security, and local economic growth. Although the Electronic Transaction System for Local Government (ETPD) has been implemented in South Tangerang City to facilitate payments through various digital channels, the payment compliance rate remains low. In 2023, South Tangerang's market service revenue reached only 29.94% of the target. Of this amount, only 16.86% of levy obligors paid regularly and on time, while 83.14% failed to make timely payments. This condition may hinder optimal regional revenue generation. This study aims to develop a classification model to predict compliance with market service levy payments in South Tangerang City. The research utilizes the Random Forest algorithm and follows the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework to classify market service levy data from November 2021 to November 2024, obtained through the ETPD system. The developed model classifies payment compliance into two categories: timely and delinquent. Analysis results from K-Fold Cross Validation show that the Random Forest model achieves strong performance metrics: 94.94% accuracy, 99.49% precision, 90.34% recall, 94.69% F1-score, and 95% AUC, indicating the model's excellent classification capability. These findings can serve as recommendations for the South Tangerang City government to formulate strategies for improving levy payment compliance.

## 1. INTRODUCTION

Local Own-Source Revenue (PAD, *Pendapatan Asli Daerah* in Indonesian) constitutes one of the primary sources of regional income, with market service levies representing a significant component [1]. These levies are payments made to municipal governments for the provision of market facilities, including stalls, kiosks, and open trading areas. Revenue generated from market operations funds market development and revitalization, thereby enhancing user comfort and safety while stimulating local economic growth [2].

Historical data show that market service levies contributed only 1%-1.5% of the total PAD in South Tangerang City between 2016 and 2018 [3]. To improve this contribution, the municipal government implemented the Local Government Electronic Transaction System (ETPD) in November 2021. This digital platform aims to accelerate regional financial transactions while enhancing accountability, efficiency, and transparency. The ETPD facilitates levy payments through multiple digital channels, including virtual accounts, QRIS, mobile banking, ATMs, and e-commerce applications, while maintaining cash payment options via tellers or local government cashiers [4].

However, performance data indicates significant challenges. The Supreme Audit Agency (BPK) reported that South Tangerang's 2023 market service levy target of IDR 3.84 billion was only 29.94% achieved (IDR 1.15 billion). ETPD system data from the Communications and Information Office (Kominfo) show that of 34,055 transactions, 33,703 payments were made before the deadline, 315 after the deadline, and 37 remained unpaid. These figures demonstrate that despite digital payment accessibility through ETPD, compliance levels remain substantially below

https://doi.org/10.25077/TEKNOSI.v12i2.2026.190-202

targets, suggesting that payment convenience alone cannot significantly improve levy compliance and collection efficiency.

Local Own-Source Revenue (PAD, *Pendapatan Asli Daerah*) is a critical component of regional fiscal sustainability, with market service levies serving as a key revenue stream for infrastructure and public service improvements [1], [2]. In South Tangerang City, these levies face significant compliance challenges despite regulatory mandates under Regional Regulation No. 10 of 2023, which requires payment within 1 month of the levy assessment's issuance. Alarmingly, 2023 data reveal that only 16.86% of 1,061 obligors made regular payments, while 83.14% were delinquent—a pattern that undermines revenue targets and hampers market revitalization efforts. While the implementation of the Local Government Electronic Transaction System (ETPD) has diversified payment channels through digital platforms, the absence of enforcement mechanisms has limited its impact on compliance rates.

To address this gap, this study leverages machine learning classification, specifically the Random Forest algorithm, to analyze historical payment data and predict compliance behavior. Random Forest's ensemble approach—constructing multiple decorrelated decision trees to aggregate predictions [5]—offers distinct advantages in this context, including robustness to overfitting [6], handling of high-dimensional data [7], and interpretable feature-importance metrics [8]. Its efficacy is well documented in comparable classification tasks, such as disease diagnosis (96.6% accuracy) [9] and industrial failure prediction (99.5% accuracy) [10], and it outperforms alternatives such as SVM and Logistic Regression.

Adopting the CRISP-DM framework [11], this research systematically develops and evaluates a predictive model to identify payment patterns, with the dual aim of advancing methodological applications in public finance and informing targeted policy interventions. The findings are expected to equip South Tangerang's government with data-driven strategies to enhance levy compliance, optimize revenue collection, and ultimately improve public market services.

This study aims to develop a machine learning classification model to predict the timeliness of market service levy payments using the Random Forest algorithm. By analyzing historical payment data, the research seeks to identify patterns that distinguish compliance from non-compliant taxpayers, providing a data-driven tool for revenue optimization. Additionally, the study evaluates the performance of the Random Forest model in classifying payment behavior, assessing key metrics such as accuracy, precision, and recall determining its effectiveness in real-world fiscal management. The findings are expected to support local governments in designing targeted interventions to improve levy compliance and enhance revenue collection efficiency.

## 2. METHOD

The research methodology follows a systematic approach to analyze market service levy compliance, as illustrated in Figure 1. This comprehensive process begins with problem identification and proceeds through data collection and analysis, following the CRISP-DM framework [12].

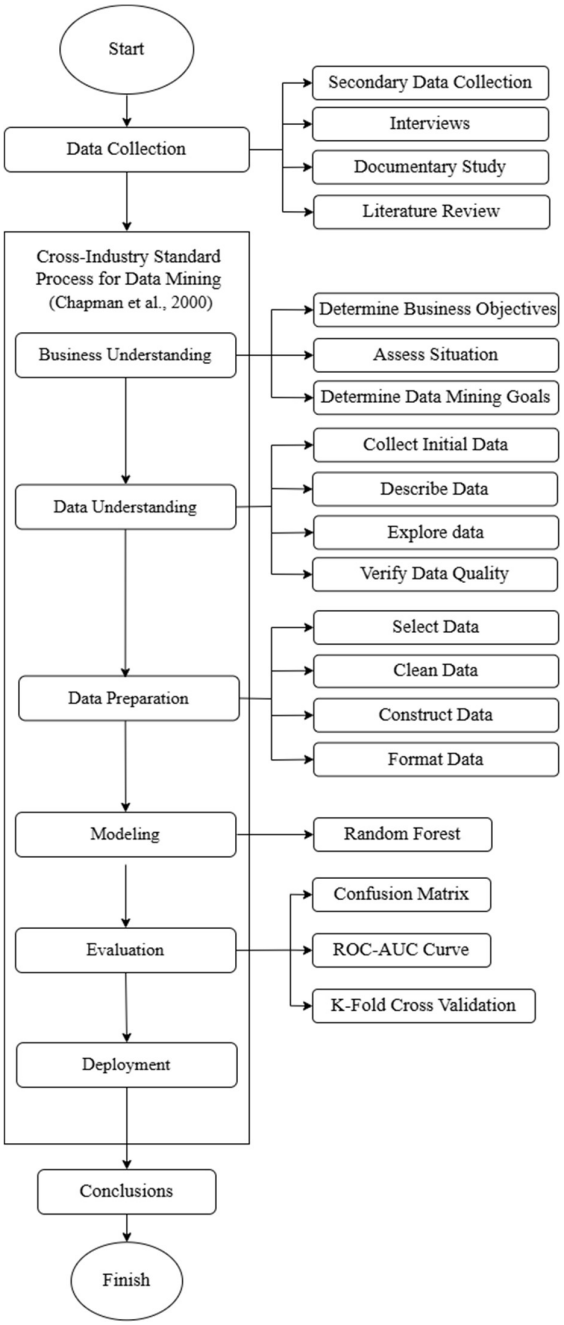


Figure 1. Research Method

As depicted in the workflow, the study combines qualitative methods with quantitative data mining techniques. The Random Forest algorithm's performance is validated through various metrics to ensure model robustness.

### 2.1 *Data Collection*

#### *2.1.1. Secondary Data Collection*

The researchers obtained market service levy revenue data from the database of the Regional Revenue Electronic Transaction System (ETPD) managed by the Communication and Information Office (Kominfo) of South Tangerang City. The data was

obtained via direct request to the ETPD system database administrator, with researchers receiving it in CSV format. This study utilizes market service levy payment transaction data spanning from November 2021 to November 2024. The research population comprises all market levy transactions during this period, while the study sample includes both timely and delinquent payments.

#### *2.1.2. Interviews*

The researchers conducted interviews with key stakeholders to gain deeper insights into the ETPD system's operation and effectiveness. Specifically, they interviewed Mr. Bagus, Head of the Application and Information Systems Division at South Tangerang's Communication and Information Office (Kominfo), on January 31, 2024, at the South Tangerang Mayor's Office. The discussion focused on three main aspects: the procedures for collecting levy data, implementation challenges of the ETPD system, and the system's effectiveness in enhancing compliance with market levy payments. These interviews provided valuable qualitative data to complement the quantitative analysis of transaction records.

#### *2.1.3. Documentary Study*

Researchers analyzed local government policy documents to examine the regulatory framework and operational mechanisms governing the market levy system. This review focused on understanding the legal basis and implementation procedures relevant to the research.

#### *2.1.4. Literature Review*

A systematic examination of academic sources - including journal articles, books, and conference proceedings - was conducted to establish theoretical foundations. This scholarly review aimed to identify relevant concepts, methodologies, and findings from previous studies to inform the current research framework and analysis approach.

### 2.2 *Data Analysis*

The research methodology follows the structured phases of the Cross-Industry Standard Process for Data Mining (CRISP-DM), as outlined below:

#### *2.2.1. Business Understanding*

Researchers interviewed Kominfo officials in South Tangerang to analyze the implementation and effectiveness of the ETPD system in improving levy compliance. These insights helped define business objectives, assess current operations, and establish targeted data-mining goals aligned with organizational needs, ensuring that subsequent analysis remained relevant to both administrative and technical requirements.

#### *2.2.2. Data Understanding*

Researchers analyzed 84,984 market service levy transactions (November 2021-November 2024) with 29 attributes, identifying missing values, duplicates, and inconsistencies. This profiling informed essential preprocessing steps to ensure data reliability for subsequent analysis.

#### *2.2.3. Data Preparation*

The data preparation phase involved comprehensive preprocessing of market service levy payment records spanning November 2021 to November 2024, comprising an initial dataset of 84,984 observations with 29 attributes. Data cleaning addressed quality issues stemming from data entry errors and system inconsistencies [13] by implementing key methods, including duplicate removal (retaining unique entries), inconsistency correction (resolving format and value mismatches), and handling missing data by deleting sparse rows or irrelevant empty columns [14]. This process reduced the dataset to 83,437 records while ensuring validity, accuracy, and consistency.

Data transformation converted the cleaned data into analysis-ready formats through multiple techniques [15]. Categorical variables (revenue details, sub-district name, village name, and payment channel) underwent one-hot encoding to create binary vectors [16], avoiding artificial ordinal relationships inherent in label encoding approaches [17][18]. Continuous payment amounts were discretized into meaningful ranges, while the compliance target variable (timely/late payment) was label-encoded (1/0). Numerical features were standardized using StandardScaler to achieve $\mu = 0$ *and* $\sigma = 1$ distributions [19], enhancing machine learning algorithm performance [20].

To address severe class imbalance (66,304 compliant vs 445 non-compliant cases), the Synthetic Minority Over-sampling Technique (SMOTE) was implemented [21], [22]. This approach generated synthetic minority-class samples by linearly interpolating between existing observations, effectively balancing the class distribution while mitigating the risk of overfitting. The processed dataset was then divided following the Pareto principle's 80:20 ratio [23], yielding 66,749 training records and 16,688 testing records. While alternative splits (70:30, 60:40) exist depending on contextual factors [24], this conventional partitioning optimized the balance between training sufficiency and evaluation robustness.

#### *2.2.4. Modeling*

The modeling phase employed the Random Forest algorithm, an ensemble learning method that constructs multiple decision trees to form a "forest" [25]. Each decision tree independently classifies the input data via majority voting, thereby significantly improving prediction accuracy and stability compared to single decision trees [26]. This approach offers several advantages, including relatively low error rates, strong classification performance, and efficient handling of large training datasets [27]. The algorithm demonstrates robustness against outliers and high-dimensional features while maintaining computational efficiency [6], [7]. However, limitations include the complexity of model validation, which requires repeated testing on different datasets [28], and potential challenges with extreme class imbalance when minority classes represent less than 10% of the total data [29], though techniques like oversampling can mitigate this issue [30].

Hyperparameter optimization was conducted using Bayesian Optimization, which efficiently identifies optimal parameter combinations through probabilistic surrogate functions, primarily


https://doi.org/10.25077/TEKNOSI.v12i2.2026.190-202

Gaussian Processes [31]. These surrogates approximate the computationally expensive objective function (e.g., model accuracy), enabling iterative optimization without exhaustive evaluation [32]. The approach employs acquisition functions to balance exploration and exploitation during parameter-space search [33], outperforming traditional methods such as Grid Search (exhaustive combination evaluation) and Random Search (random sampling) in both convergence speed and computational efficiency [34], [35]. The optimization process iteratively approximates the computationally expensive objective function (e.g., model accuracy) without exhaustive evaluation [36], making it particularly suitable for complex model-tuning scenarios. Bayesian Optimization models the objective function as a Gaussian Process (GP):

$$f(x) \sim GP(\mu(x), k(x, x')) \quad (1)$$

Where $\mu(x)$ is the mean function and $k(x, x')$ is the covariance kernel (e.g., RBF kernel):

$$k(x, x') = \exp(-\frac{\|x - x'\|^2}{2\theta^2}) \quad (2)$$

Here, $\theta$ controls the length scale of correlation decay between parameter points. For a new candidate point $x^*$, the GP provides predictive mean $\mu(x^*)$ and variance $\sigma^2(x^*)$:

$$\mu(x^*) = k(x^*)^T . K^{-1} . y \quad (3)$$

$$\sigma^2(x^*) = k(x^*, x^*) - k(x^*)^T . K^{-1} . k(x^*) \quad (4)$$

Where $K$ is the kernel matrix of initial points and $y$ their observed performance. Acquisition Functions guide the search by balancing exploitation ($\mu$) and exploration ($\sigma$):

$$EI(x) = (\mu(x) - f^*)\Phi(z) + \sigma(x)\phi(z) \quad (5)$$

The tuned Random Forest model was subsequently evaluated on test data to assess its predictive performance for classifying levy payment compliance.

### 2.2.5. *Evaluation*

The model evaluation employed a comprehensive multi-metric approach to assess classification performance. The confusion matrix served as the foundational framework [37], This matrix organizes results into four key categories: True Positives (TP), representing class 1 instances correctly predicted as class 1; True Negatives (TN), denoting class 0 instances accurately classified as class 0; False Positives (FP), where class 0 instances are incorrectly predicted as class 1 (Type I error); and False Negatives (FN), indicating class 1 instances mistakenly classified as class 0 (Type II error). These metrics provide a comprehensive framework for assessing model accuracy and form the basis for deriving essential evaluation measures such as precision, recall, and F1-score.

Accuracy measures the overall correctness of predictions by calculating the ratio of correctly classified instances (both true positives and true negatives) to the total number of instances.

$$Accuracy = \frac{TP+T}{TP+TN+FP+FN} \quad (6)$$

Precision evaluates the model's exactness by determining the proportion of correctly predicted positive instances among all predicted positives.

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

Recall (also called sensitivity) assesses completeness by measuring the fraction of actual positives correctly identified.

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

The F1-Score provides a balanced measure between precision and recall through their harmonic mean.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (9)$$

These metrics collectively offer a comprehensive evaluation framework, where accuracy provides an overall performance indication, precision reflects prediction reliability, recall emphasizes positive class coverage, and F1-Score balances both concerns for robust model assessment.

Receiver Operating Characteristic (ROC) analysis further evaluated the model's discriminative capability by plotting the relationship between True Positive Rate (TPR = Recall) and False Positive Rate across classification thresholds [38]. False Positive Rate (FPR) is the ratio of the proportion of negative data that is incorrectly classified as positive. FPR can be obtained using the following calculation:

$$FPR = \frac{FP}{FP+TN} \quad (10)$$

A low FPR indicates that the model rarely misclassifies negative cases as positive. The ideal ROC curve will curve from the bottom left to the top right, indicating that the model can achieve high TPR with low FPR simultaneously. This indicates good classification performance [39]. Area Under the Curve (AUC) is used to calculate the area under the ROC curve. The AUC value is calculated by summing the areas of the trapezoids formed under the ROC curve [40]. The AUC ranges from 0 to 1; the closer it is to 1, the better the model is at distinguishing between positive and negative classes.

To ensure reliable performance estimation, 5-fold cross-validation was implemented [41], [42]. This technique partitioned data into five equal subsets, iteratively using each subset once for testing while training on the remaining four [43]. The repeated validation process mitigated the risk of overfitting while providing stable performance estimates across different data partitions, enhancing the generalizability of the results. Together, these evaluation methods provided a rigorous, multifaceted assessment of the model's effectiveness in predicting levy payment compliance.

### 2.2.6. *Deployment*

This study's deployment phase focuses on operationalizing the Random Forest model for compliance prediction and generating interpretable visual reports.

## 3. RESULT

### *3.1. Business Understanding*

The initial foundation in understanding the business context being studied. This understanding is the basis for determining business objectives and determining data mining processes that are relevant to needs.

### *3.1.1 Determine Business Objectives*

Market service levies are payments made by traders to the South Tangerang City government for the use of market facilities, such as stalls, kiosks, and trading areas. As a key local revenue source, these levies fund market operations and maintenance. The city has implemented the Local Government Electronic Transaction System (ETPD) to replace cash payments with digital transactions, enhancing efficiency and transparency through real-time payment recording.

The ETPD system serves three core functions: enhancing financial transparency and governance, accelerating digital

payment adoption, and advancing the city's sustainable development vision of "Excellent Tangsel: Towards a Sustainable, Connected, Effective, and Efficient City." By establishing integrated transaction networks, the system strengthens infrastructure connectivity (Mission 2) while its transparent processes build trader confidence in municipal governance (Mission 4). Furthermore, the digitization of levy payments significantly reduces manual administrative burdens (Mission 5), directly supporting the development of an efficient, technology-driven bureaucracy. This multifaceted approach simultaneously modernizes revenue collection while aligning with broader urban development objectives.

### *3.1.2 Assess Situation*

The Local Government Electronic Transaction System (ETPD) was introduced in South Tangerang City in November 2021 with the dual objectives of increasing transparency and optimizing the collection of local levy revenue. While the system has successfully digitized payment processes and improved the accuracy of transaction recording, it has yet to achieve its targeted revenue improvements. Current usage of ETPD data remains limited to basic payment monitoring, lacking comprehensive analysis of compliance patterns or predictive capabilities to identify non-compliant vendors.

Several key challenges hinder the system's effectiveness. First, many merchants demonstrate low compliance rates, primarily due to insufficient awareness about payment obligations. Second, the system underutilizes valuable transaction data that could reveal patterns based on location, business type, and payment channels. Third, a significant portion of traditionally cash-dependent vendors resist adopting the digital system, creating barriers to adoption.

The ETPD operates within a clear regulatory framework, including Mayoral Instruction No. 920/81/BPKAD and Presidential Decree No. 3/2021. The payment process involves multiple stages: initial issuance of levy assessment notices (SKRD), digital validation via electronic signatures, generation of payment references and QR codes, bank payment verification, and final issuance of receipts. Enhancing the system's analytical capabilities could transform ETPD from a transaction recorder to a strategic compliance tool. By implementing advanced data analysis techniques, the government could develop targeted intervention strategies, improve monitoring of payment compliance, and ultimately increase revenue collection efficiency. This would address current limitations while leveraging existing digital infrastructure.

### *3.1.3 Determine Data Mining Goals*

Classification analysis of historical transaction data can effectively identify compliance patterns. This analytical approach enables governments to develop targeted strategies that promote timely payments while optimizing payment methods. By distinguishing between compliant and non-compliant payment behaviors, the model provides actionable insights for improving revenue collection systems.

## 3.2. *Data Understanding*

The process includes initial analysis to examine data patterns, detect missing values, identify duplicate entries, and assess data inconsistencies. These preliminary investigations establish a foundational understanding of the dataset's condition, informing subsequent preprocessing steps and ensuring the reliability of analytical outcomes.

### *3.2.1 Collect Initial Data*

The study uses market service levy payment transaction data from November 2021 to November 2024, comprising 84,984 records and 29 attributes, as shown in Figure 2. This dataset was obtained from the South Tangerang Communication and Information Office in CSV format.

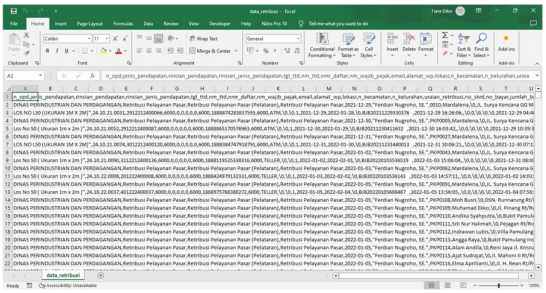

Figure 2. Market Service Levy Data

The transaction dataset for market service levies contains comprehensive payment records, including transaction dates, revenue categories, payment amounts, compliance status (timely/late payments), and geographic details such as sub-district and village locations.

### *3.2.2 Describe Data*

This section describes the attributes contained within the dataset. A detailed explanation of each data attribute is provided in Table 1.

Table 1. Dataset Attribute Explanation

| Attribute Name | Data Type | Explanation |
|---|---|---|
| 'n_opd' | *String* | Name of Regional Apparatus Organization |
| 'jenis_pendapatan' | *String* | Types of Income received |
| 'rincian_pendapatan' | *String* | More specific details regarding the type of income |
| 'tgl_ttd' | *String* | SKRD Signature Date |
| 'nm_ttd' | *String* | Name of the official who signed the document |
| 'nmr_daftar' | *String* | Levy transaction registration number |
| 'nm_wajib_pajak' | *String* | Taxpayer Name |
| 'email' | *String* | Taxpayer's email address |
| 'alamat_wp' | *String* | Taxpayer Address |
| 'lokasi' | *String* | Market location or levy payment area |
| 'n_kecamatan' | *String* | District Name |


https://doi.org/10.25077/TEKNOSI.v12i2.2026.190-202

| | | |
|---|---|---|
| 'n_kelurahan' | *String* | Village Name |
| 'uraian_retribusi | *String* | A brief description of the levy paid |
| 'no_skrd' | *String* | SKRD Number (Regional Levy Decree) |
| 'no_bayar' | *Integer* | Payment umber |
| 'jumlah_bayar' | *Integer* | Payment amount due |
| 'status_ttd' | *Integer* | Signature status (already/unsigned) |
| 'nomor_va_bjb' | *String* | BJB Account Vitual Number |
| 'chanel_bayar' | *String* | Payment channels used |
| 'invoice_id' | *String* | Unique Invoice ID or Identification for payment invoices |
| 'text_qris' | *String* | QRIS Text or QRIS Information for payment |
| 'status_bayar' | *Integer* | Payment status (successful or failed) |
| 'tgl_skrd_awal' | *String* | SKRD effective date |
| 'tgl_skrd_akhir' | *String* | SKRD's last deadline date |
| 'tgl_bayar' | *String* | Date the payment transaction was made |
| 'created_at' | *String* | Date transaction data is created in the system |
| 'updated_at' | *String* | The last date the data was updated |
| 'created_by' | *String* | The identity of the user who created the transaction data |
| 'updated_by' | *String* | The identity of the user who last updated the data |

### *3.2.3 Explore Data*

The data exploration stage aims to understand the dataset's characteristics. This process involves identifying patterns in the data, such as calculating the frequency of each category, analyzing data distribution, identifying dominant and minority categories, and exploring relationships between attributes.

To understand data variation and identify distribution patterns in each category, the number of unique values for every attribute is calculated. The dataset reveals complete uniformity in the 'n_opd' and 'jenis_pendapatan' attributes for all Department of Industry and Trade records related to Market Service Retribution. Analysis of revenue details shows a highly uneven distribution, with Pelataran comprising 60,029 entries (60.9%), Los accounting for 20,366 (20.7%), and Kios making up just 4,589 (4.7%). Geographically, the data is concentrated in Ciputat sub-district (48,541 entries, 49.3%) and Pamulang (36,031, 36.6%), while Ciputat Timur is significantly underrepresented with only 412 entries (0.4%). This imbalance intensifies at the village level, where Ciputat (48,438 entries) and Pamulang Barat (35,329) collectively account for 85.1% of all data, dwarfing the minimal representation of other villages. Such pronounced disparities across revenue types, sub-districts, and villages indicate a substantial data imbalance that could distort analytical findings and subsequent policy decisions derived from this dataset.

Table 2 presents the results of an analysis of the distribution of market retribution data by sub-district (kecamatan), village (kelurahan), and revenue type (Kios, Los, Pelataran). The analysis reveals a significant imbalance in transaction volumes across different regions.

Table 2. Revenue Detail Distribution

| District | Villages | Income Details | Amount |
|---|---|---|---|
| Ciputat | Cipayung | Kios | 2 |
| | | Los | 12 |
| | Ciputat | Kios | 4.196 |
| | | Los | 43.920 |
| | | Pelataran | 322 |
| | Jombang | Los | 5 |
| | Serua | Pelataran | 84 |
| Ciputat Timur | Rempoa | Kios | 391 |
| | | Pelataran | 21 |
| Pamulang | Bambuapus | Pelataran | 484 |
| | Pamulang Barat | Los | 16.092 |
| | | Pelataran | 19.237 |
| | Pamulang Timur | Pelataran | 218 |

The data reveals a heavy concentration of transactions in specific areas, while others show minimal or no activity, suggesting uneven adoption of the ETPD system across districts and villages. This imbalance likely reflects implementation challenges, such as inadequate outreach or public preference for traditional payment methods.

Table 3 presents the descriptive statistics of the 'jumlah_bayar' (payment amount) attribute for market service fees. The data shows an average payment of IDR 27,900 with a standard deviation of IDR 85,709, indicating significant transaction variability. Payments range from IDR 100 to IDR 2,340,000, with the maximum likely representing outlier transactions.

Table 3. Descriptive Statistics of Payment Amounts

| Informasi | Nilai |
|---|---|
| Count | 84.984 |
| Mean | 27.900 |
| Standard Deviation | 85.709 |
| Minimun | 100 |
| 25% (Q1) | 3.400 |
| 50% (Median) | 6.750 |
| 75% (Q2) | 12.750 |
| Maximum | 2.340.000 |

Payment data show a right-skewed distribution: 50% of transactions are ≤IDR 6,750, and 75% are ≤IDR 12,750, with high-value outliers inflating the mean. The scatter plot in Figure 3 shows the distribution of payments across revenue categories (*pelataran*, *los*, *kios*).

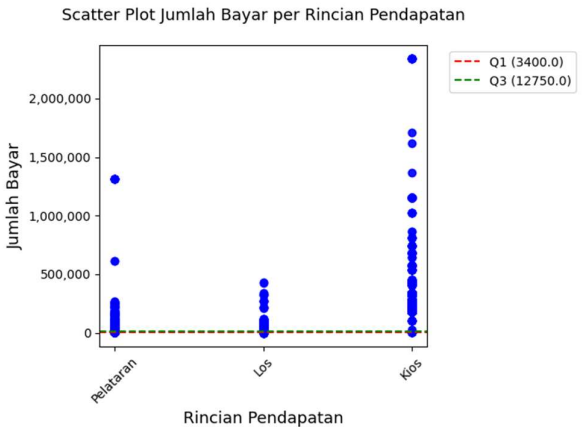


Figure 3. Payment Amount Scatter Plot by Revenue Category

Most payments fall below the third quartile (Q3) of IDR 12,750 (green line), while the first quartile (Q1) is IDR 3,400 (red line). The majority of transactions cluster between Q1 and Q3, but 9,886 high-value outliers exist. The plot reveals significant variation in payments across revenue categories, with a concentration of small payments and a few large transactions skewing the distribution.

### *3.2.4 Verify Data Quality*

The data quality verification phase ensures dataset integrity and reliability before further analysis. This stage involves comprehensive checks of the market service fee dataset to identify potential issues, including missing values, duplicates, and inconsistencies. The dataset contains 926 missing values in the 'invoice_id' and 'text_qris' attributes, likely due to cash transactions or non-digital payment methods that don't require these fields. No duplicate entries were found, confirming that each transaction is uniquely recorded and eliminating potential double-counting issues in subsequent analysis. The 'chanel_bayar' attribute exhibits multiple unique values (e.g., ATM, Teller, Virtual Account, QRIS-user), reflecting diverse payment methods. Standardization of formatting is required to improve data consistency. The string-type date attributes required validation for format consistency (length, time notation) and data integrity, with findings detailed in Table 4.

Table 4. Date Format Inconsistencies

| Atribut | Jumlah *String* | Jumlah Data | Contoh Data |
|---|---|---|---|
| tgl_bayar | 19 | 83.437 | 2021–12–29 16:26:06 |
| | 3 | 1.547 | NaT |
| tgl_ttd | 10 | 84.984 | 2023–05–16 |
| tgl_skrd_awal | 10 | 63.132 | 2023–05–23 |
| | 19 | 21.852 | 2024–07–29 00:00:00 |
| **tgl_skrd_akhir** | 10 | 63.132 | 2023–08–15 |
| | 19 | 21.852 | 2024–08–11 00:00:00 |

Date attributes contain invalid 'NaT' values and inconsistent formats (YYYY-MM-DD vs. YYYY-MM-DD HH:MM:SS), necessitating conversion to a standardized YYYY-MM-DD datetime format.

### *3.3. Data Preparation*

This stage involves cleaning, processing, and preparing data for machine learning analysis. The goal is to ensure the data meets quality standards, relevance, and consistency requirements for modeling.

### *3.3.1 Select Data*

Before data selection, a correlation matrix was used to identify relationships among variables in the dataset. Before data selection, a correlation matrix was used to examine relationships between variables (Figure 4). The matrix displays the strength and direction of linear correlations, with values ranging from -1 to 1.

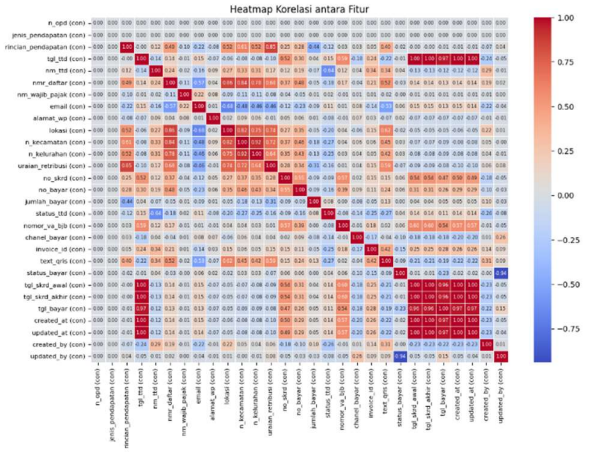


Figure 4. Correlation Matrix

Geographic factors like *n_kecamatan* and *n_kelurahan* show a strong correlation with each other (0.92) and with location (0.82 and 0.75), indicating that geographic location significantly influences payment patterns. This may be due to economic conditions, infrastructure, or local policies affecting payment accessibility and preferences. *Chanel_bayar* has a weak negative correlation with *jumlah_bayar* (-0.08), suggesting certain payment methods are used for smaller transactions. Its low correlations
with rincian_pendapatan (0.03), n_kecamatan (0.06),
and n_kelurahan (0.04) imply that payment methods are less influenced by fee types or geographic location. However, *chanel_bayar* may still impact compliance, though it is not a primary factor. Then, *rincian_pendapatan*, *n_kecamatan*, and *n_kelurahan* are key interrelated features, while *status_bayar* is the main compliance indicator. Although *jumlah_bayar* and *chanel_bayar* show weak correlations with compliance, they remain relevant in understanding payment behavior.

The selected features, as in Figure 5, include economic (*jumlah_bayar*), administrative (*rincian_pendapatan*, *n_kecamatan*, *n_kelurahan*), and behavioral (*chanel_bayar*, *status_bayar*) factors. *Status_bayar* serves as the target variable for tax compliance, with *tgl_skrd_akhir* as the due date reference and *tgl_bayar* determining timeliness.


https://doi.org/10.25077/TEKNOSI.v12i2.2026.190-202

| | rincian_pendapatan | n_kecamatan | n_kelurahan | jumlah_bayar | chanel_bayar | status_bayar | tgl_skrd_akhir | tgl_bayar |
|---|---|---|---|---|---|---|---|---|
| 0 | Retribusi Pelayanan Pasar (Pelataran) | Pamulang | Pamulang Barat | 6000 | ATM | 1 | 2022-01-28 | 2021-12-29 16:26:06 |
| 1 | Retribusi Pelayanan Pasar (Pelataran) | Pamulang | Pamulang Barat | 6000 | ATM | 1 | 2022-01-29 | 2021-12-30 16:03:42 |
| 2 | Retribusi Pelayanan Pasar (Pelataran) | Pamulang | Pamulang Barat | 6000 | ATM | 1 | 2022-01-30 | 2021-12-31 10:06:21 |
| 3 | Retribusi Pelayanan Pasar (Pelataran) | Pamulang | Pamulang Barat | 6000 | TELLER | 1 | 2022-02-01 | 2022-01-03 15:06:04 |
| 4 | Retribusi Pelayanan Pasar (Pelataran) | Pamulang | Pamulang Barat | 6000 | TELLER | 1 | 2022-02-02 | 2022-01-03 14:57:11 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 84979 | Retribusi Pelayanan Pasar (Kios) | Ciputat | Ciputat | 432000 | TELLER | 1 | 2024-12-30 00:00:00 | 2024-11-30 10:41:10 |
| 84980 | Retribusi Pelayanan Pasar (Kios) | Ciputat | Ciputat | 432000 | TELLER | 1 | 2024-12-30 00:00:00 | 2024-11-30 10:41:18 |
| 84981 | Retribusi Pelayanan Pasar (Kios) | Ciputat | Ciputat | 405000 | TELLER | 1 | 2024-12-30 00:00:00 | 2024-11-30 10:41:26 |
| 84982 | Retribusi Pelayanan Pasar (Kios) | Ciputat | Ciputat | 576000 | TELLER | 1 | 2024-12-30 00:00:00 | 2024-11-30 10:41:33 |
| 84983 | Retribusi Pelayanan Pasar (Kios) | Ciputat | Ciputat | 576000 | TELLER | 1 | 2024-12-30 00:00:00 | 2024-11-30 10:41:41 |

84984 rows × 8 columns

Figure 5. Selected Dataset

These attributes will be used for classification analysis. Table 5 lists the selected variables used to build the classification model that predicts compliance (on-time/late) based on these features.

Table 5. Attributes for Classification Analysis

| No | Classification Variables | Targeted Variables |
|---|---|---|
| 1 | Income Details | Compliance (On-Time/Off-Time) |
| 2 | District Name | |
| 3 | Village Name | |
| 4 | Payment Amount | |
| 5 | Chanel Pay | |

Compliance is derived by calculating the difference between *tgl_skrd_akhir* (due date) and *tgl_bayar* (payment date). This difference determines whether the payment was made before or after the deadline.

### 3.3.2 Clean Data

Categorical attributes (*rincian_pendapatan* and *chanel_bayar*) were standardized to eliminate ambiguity in Figure 6 for standardized values.

```
Rincian Pendapatan: ['Pelataran' 'Los' 'Kios']
Chanel Bayar: ['ATM' 'Teller' '\\0' 'Virtual Account' 'QRIS']
```

Figure 6. Consistency of Income Details and Pay Channels

Date-time attributes (*tgl_skrd_akhir* and *tgl_bayar*) were split to isolate date values in Figure 4.7, then converted to a standardized datetime format ("YYYY-MM-DD") for consistent time-based analysis.

| | tgl_skrd_akhir | tgl_bayar | waktu_bayar | waktu_bayar_akhir |
|---|---|---|---|---|
| **84979** | 2024-12-30 | 2024-11-30 | 10:41:10 | 00:00:00 |
| **84980** | 2024-12-30 | 2024-11-30 | 10:41:18 | 00:00:00 |
| **84981** | 2024-12-30 | 2024-11-30 | 10:41:26 | 00:00:00 |
| **84982** | 2024-12-30 | 2024-11-30 | 10:41:33 | 00:00:00 |
| **84983** | 2024-12-30 | 2024-11-30 | 10:41:41 | 00:00:00 |

Figure 7. Date and Time Separation

Invalid entries (e.g., '\0' in *chanel_bayar*, indicating missing payment method) were removed (1,547 rows deleted). Post-cleaning, the dataset was reduced from 84,984 to 83,437 rows, improving the quality of the analysis.

### 3.3.3 Construct Data

The *construct data* phase creates new features. Figure 8 displays the newly generated compliance feature, calculated as the time difference between payment date (*tgl_bayar*) and due date (*tgl_skrd_akhir*).

| | tgl_skrd_akhir | tgl_bayar | selisih_hari |
|---|---|---|---|
| **84981** | 2024-12-30 | 2024-11-30 | 30 |
| **84982** | 2024-12-30 | 2024-11-30 | 30 |
| **84983** | 2024-12-30 | 2024-11-30 | 30 |

Figure 8. Feature Creation

The calculated time difference indicates the timeliness of payment. If the time difference is positive or close to zero, the data is labeled as an on-time payment. Conversely, if the time difference is negative, the data is labeled as a late payment. The calculation results show 82,878 records labeled as on-time payments and 559 as late payments, indicating a significant disparity between timely and delayed payments in the dataset.

### 3.3.4 Format Data

The data formatting stage involves several steps, including data transformation and processing, to ensure the data is ready for analysis. This process aims to convert raw data into a format suitable for the intended algorithms. The transformation begins with using label encoding to convert the compliance attribute from categorical values into numerical ones, where "on-time" payments are encoded as 1 and "late" payments as 0.

Subsequently, a discretization process is applied to the payment amount (jumlah_bayar) attribute. This process groups continuous data into specific range categories, facilitating easier analysis of payment patterns through more structured values. The applied value ranges for payment amounts are presented in Table 6.

Table 6. Value of the Range of Payment Amounts

| Kategori Jumlah Bayar |
|---|
| '0–1.000', '1.001–10.000', '10.001–50.000', '50.001–100.000', '100.001–500.000', '500.001–1.000.000', '1.000.001–1.500.000', '1.500.001–2.000.000', '> 2.000.000' |

Next, one-hot encoding is applied to categorical attributes, converting them into binary columns (1 for present, 0 for absent). Figure 9 shows the transformed dataset, with categorical values now represented numerically for machine-learning compatibility.

| | kepatuhan | rincian_pendapatan_Kios | rincian_pendapatan_Los | rincian_pendapatan_Pelataran | n_kecamatan_Ciputat | n_kecamatan_Ciputat Timur | n_kecamatan_Pamulang | n_kelurahan_Bambuapus |
|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 1.00 | 0.00 |
| 1 | 1 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 1.00 | 0.00 |
| 2 | 1 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 1.00 | 0.00 |
| 3 | 1 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 1.00 | 0.00 |
| 4 | 1 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 1.00 | 0.00 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 84979 | 1 | 1.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 |
| 84980 | 1 | 1.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 |
| 84981 | 1 | 1.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 |
| 84982 | 1 | 1.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 |
| 84983 | 1 | 1.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 |

83437 rows × 28 columns

Figure 9. One Hot Encoded Dataset

The dataset is split into 80% training data (66,749 samples) and 20% test data (16,688 samples), for a total of 83,437 records. Standard Scaler is then applied to normalize all features. The training data shows significant class imbalance, with 66,314 "On-Time" (Class 1) samples compared to only 435 "Late" (Class 0) samples. To address this bias, SMOTE (Synthetic Minority Oversampling Technique) is employed to generate synthetic minority class samples. After SMOTE is applied, both classes are

balanced at 66,314 samples each, ensuring fair learning and improved predictive performance for both classes. This balancing is crucial for developing an unbiased machine learning model.

### *3.4. Modeling*

The modeling phase involves applying machine learning algorithms to build a classification model. This study employs the Random Forest algorithm with hyperparameter optimization using Bayesian Optimization. Figure 10 shows the training data used to train the Random Forest model while searching for optimal parameters through Bayesian Optimization.

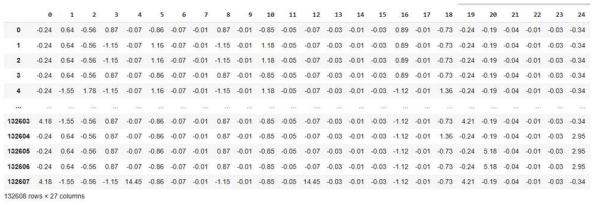

Figure 10. Train Data

Subsequently, Figure 11 presents the test data used to evaluate the model's performance after training with the optimized parameters. Both the training and test datasets are used in the hyperparameter tuning process to improve model accuracy.

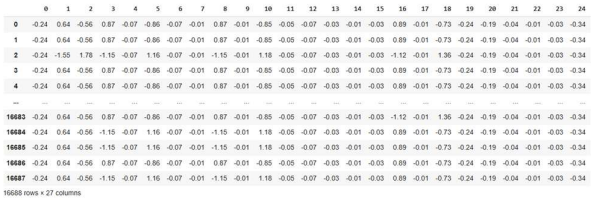

Figure 11. Test Data

#### *3.4.1 Hyperparameter Tuning*

The optimal parameter tuning for the Random Forest algorithm was achieved using Bayesian Optimization, a probabilistic approach that identifies hyperparameter combinations yielding the highest accuracy. The optimized parameters achieved 90% test accuracy, demonstrating the model's high reliability in predicting payment compliance. Table 7 presents the hyperparameter tuning results from the Bayesian Optimization process.

Table 7. Best Parameters

| Parameters | Nilai Terbaik |
|---|---|
| ***n_estimators*** | 16 |
| ***max_depth*** | 10 |
| ***max_features*** | None |
| ***min_samples_split*** | 2 |
| ***min_samples_leaf*** | 1 |

These refined parameters enable the model to learn effectively from training data and generate accurate predictions.

#### *3.4.2 Train Model*

The model is trained on the training data using the Random Forest algorithm, with optimized parameters to learn data patterns effectively. Once trained, the model's predictive performance is evaluated using test data to verify its accuracy on new, unseen data. This final evaluation validates that the hyperparameter optimization successfully produced a robust model without overfitting to the training data.

### *3.5. Evaluation*

The evaluation phase assesses the performance of the built and optimized model. This critical step verifies the model's ability to deliver accurate and reliable predictions for both training data and new, unseen data.

#### *3.5.1 Confusion Matrix*

The Random Forest model demonstrates excellent classification performance, as indicated by the confusion matrix in Figure 12. The results show: 110 True Negatives (TN), 4 False Positives (FP), 1,651 False Negatives (FN), and 14,923 True Positives (TP). While the model correctly identifies most positive samples, it shows some limitations in detecting all positive cases.

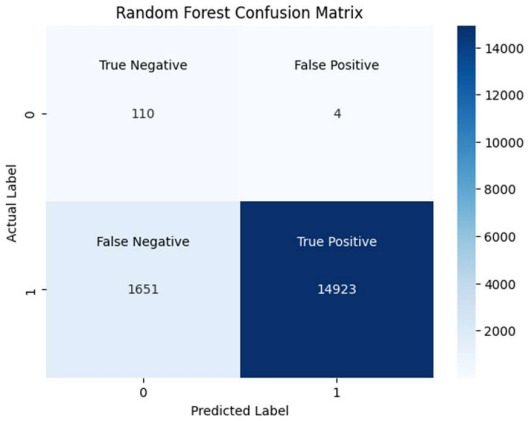


Figure 12. Confusion Matrix

The model achieves 90% accuracy, indicating strong overall predictive capability. It maintains an exceptional 99% precision, demonstrating nearly perfect positive prediction accuracy with minimal false positives. However, the 90% recall reveals room for improvement in detecting all positive samples (1,589 false negatives). The balanced 94% F1-score reflects good harmony between precision and recall, confirming the model's effectiveness for classification tasks.

#### *3.5.2 ROC-AUC*

Figure 13 displays the Random Forest model's ROC-AUC curve, plotting True Positive Rate (TPR) against False Positive Rate (FPR). The gray diagonal line represents random chance performance – the farther the curve deviates from this line, the better the model. With an AUC of 0.95 (nearly perfect 1.0), the model demonstrates excellent positive-class detection (high TPR) while maintaining low false positives (low FPR). The curve's proximity to the top-left corner confirms robust classification performance, accurately identifying positive samples with minimal misclassification of negatives.


https://doi.org/10.25077/TEKNOSI.v12i2.2026.190-202

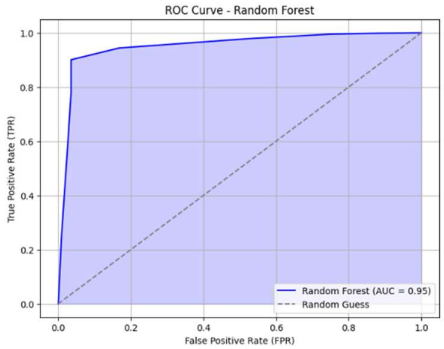


Figure 13. ROC-AUV Curve

### 3.5.3 *K-Fold Cross Validation*

The model's stability was assessed using 5-fold cross-validation, where the dataset was divided into 5 folds. In each iteration, the model was trained on 4 folds and tested on the remaining fold. This process was repeated 5 times, with each fold serving as the test set once. The evaluation results across all iterations are shown in Table 8.

Table 8. 5-Fold Cross Validation

| Metric | Iteration | | | | | *Mean* |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | |
| *Accuracy* | 95.14 | 94.99 | 94.85 | 94.76 | 94.94 | 94.94 |
| *Precision* | 99.43 | 99.61 | 99.45 | 99.43 | 99.53 | 99.49 |
| *Recall* | 90.79 | 90.33 | 90.23 | 90.02 | 90.30 | 90.34 |
| *F1-Scores* | 94.92 | 94.75 | 94.60 | 94.50 | 94.69 | 94.69 |

The 5-fold cross-validation results demonstrate that the Random Forest model delivers stable and reliable performance across all folds. The model achieved an average accuracy of 94.94%, while maintaining an exceptional precision of 99.49% in correctly identifying positive cases. Although the recall score of 90.34% suggests that some positive samples were missed, the well-balanced F1 Score of 94.69% indicates good overall performance. These consistent results across all five validation folds confirm the model's robustness in making accurate predictions while effectively minimizing false positives.

### 3.6. *Deployment*

The deployment phase focuses on implementing the optimized Random Forest classification model (using Bayesian Optimization) for South Tangerang's Communication and Information Office. This model predicts market service fee compliance ("On-Time" or "Late") to support data-driven decision-making in segmentation, billing schedules, and policy development.

The trained model is packaged as rf_model.pkl using Python's joblib/pickle for easy reuse. A robust data processing pipeline ensures consistency, with separate components (scaler.pkl and encoder.pkl) maintaining standardized preprocessing. New data from the ETPD system undergo the same transformation before being fed into the model for compliance prediction. For enhanced accessibility, the model is deployed via a Flask/FastAPI API, enabling real-time predictions through the/predict endpoint in a cloud-based environment.

The implementation includes targeted operational approaches: early notifications for predicted non-compliant payers and incentive programs (discounts/rewards) for consistently compliant ones. These strategies promote timely payments while maintaining positive engagement with service users.

## 4. DISCUSSION

The visualization in Figure 14 shows our Random Forest model's 16 decision trees (max depth=10 to prevent overfitting). Each tree makes independent predictions (blue="On-Time", orange="Late"), with their combined results creating more accurate compliance forecasts than any single tree could achieve alone. The varying tree structures demonstrate how the model learns different payment patterns while maintaining reliable performance.

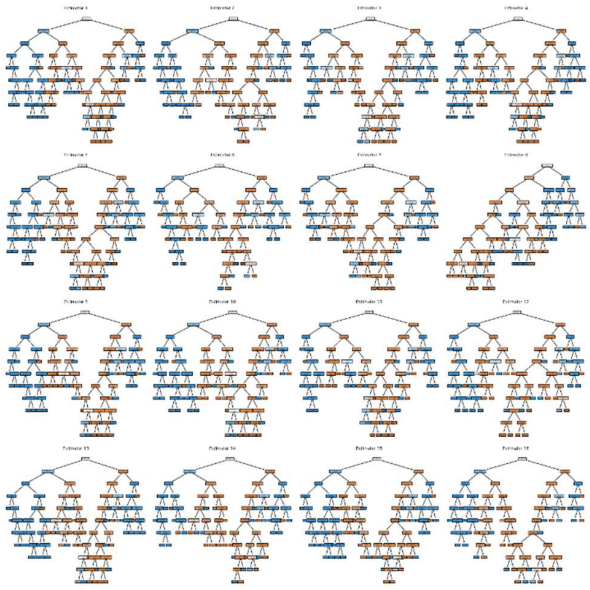

Figure 14. Random Forest Estimators

Figure 15 displays the Random Forest's predictions across 16,688 test samples. Each of the 16 decision trees (estimators) makes independent predictions, represented by colored dots (one color per tree), while black "x" marks show the final ensemble prediction through majority voting. The x-axis indexes samples, and the y-axis shows the predicted probability of "Compliant" classification (Class 1).

Tightly clustered dots indicate high prediction certainty, while dispersed points reflect challenging samples where trees disagree. This variation demonstrates how the ensemble combines diverse perspectives from different data subsets to produce stable, accurate results - a core strength of Random Forest methodology.

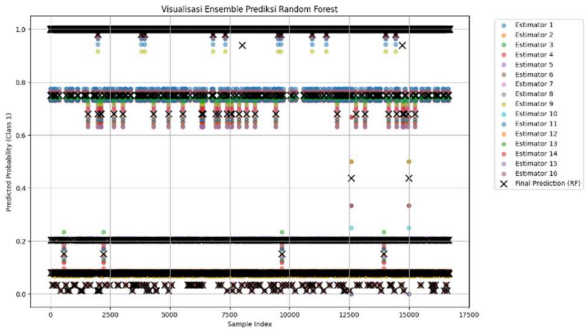


Figure 15. Predict Random Forest Model By Each Estimator

This study developed a Random Forest classification model to predict compliance with market service fee payments. The dataset consisted of 83,437 records, split into 80% for training (66,749 records) and 20% for testing (16,688 records). Key preprocessing steps included label encoding for compliance status (1 = On-Time, 0 = Late), one-hot encoding for categorical features, and addressing class imbalance—with 66,314 On-Time and only 435 Late payments—by applying SMOTE to synthesize samples from the minority class.

Model optimization was conducted using Bayesian Optimization, resulting in the following optimal parameters: *n_estimators* = 16, *max_depth* = 10, *max_features* = None, *min_samples_split* = 2, and *min_samples_leaf* = 1. The optimized model demonstrated consistent performance, achieving 94.94% accuracy based on 5-fold cross-validation. Additionally, it achieved a precision of 99.49%, a recall of 90.34%, an F1 Score of 94.69%, and an AUC of 95%.

Analysis revealed that payment compliance patterns were influenced by several key predictors, including revenue type (retribution category), payment amount, payment method, and geographic location (subdistrict and village). The model provides actionable insights to identify compliance trends and optimize revenue collection strategies.

## 5. CONCLUSION

This study successfully developed a classification model using the Random Forest algorithm to predict compliance with market service fee payments. The model used 16 decision trees (*n_estimators*=16) with a maximum depth of 10 (*max_depth*=10) to prevent overfitting and utilized all available features (*max_features*=None) to capture data patterns effectively. With this configuration, the model achieved 90% accuracy, demonstrating strong classification performance. These findings can support local governments in improving revenue collection, for example, by implementing early warning systems for payers predicted to be late, allowing for preventive measures.

Evaluation results show that the Random Forest model performs well, achieving 90% accuracy, 99% precision, 90% recall, and a 94% F1-score. An AUC of 95% indicates a strong ability to distinguish between on-time and late payments. To ensure reliability, 5-Fold Cross Validation was conducted, yielding consistent results with an average accuracy of 94.94%, precision of 99.49%, recall of 90.34%, and F1-score of 94.69%. These consistently high scores confirm the model's accuracy and strong generalization to new data.

It is recommended to explore other algorithms and apply hyperparameter tuning methods, such as Grid Search or Random Search, to improve model accuracy and generalization further. Temporal analysis is also suggested to identify payment patterns over specific time intervals, such as daily or weekly trends, to understand better how timing affects payment compliance.

## BIBLIOGRAPHY

[1] V. P. Tappi, "Analisis Pendapatan Asli Daerah (PAD) Kabupaten Jayapura," *Jurnal Ekonomu & Bisnis*, vol. 12, no. 1, pp. 16–24, 2021, Accessed: Mar. 07, 2024. [Online]. Available: https://doi.org/10.55049/jeb.v12i1.66

[2] Sumiati, Sulkarnain, Sitti Jamilah Amin, and Damirah, "Analisis Potensi Pasar Tradisional dalam Meningkatkan Perekonomian Daerah," *BALANCA : Jurnal Ekonomi dan Bisnis Islam*, vol. 4, no. 2, pp. 8–15, Nov. 2023, doi: 10.35905/balanca.v4i2.4823.

[3] S. Ruddin and A. Y. Nasution, "Analisis Revitalisasi Pasar Tradisional untuk Meningkatkan Pendapatan Daerah Kota Tangerang Selatan, Provinsi Banten," *Jurnal Mandiri : Ilmu Pengetahuan, Seni, dan Teknologi*, vol. 3, no. 2, pp. 294–306, Dec. 2019, doi: 10.33753/mandiri.v3i2.91.

[4] O. B. Saputri, "Analisis SWOT Transformasi Digital Transaksi Keuangan Pemerintah Daerah dalam Mendukung Inklusi Keuangan," *Jurnal Ekonomi Keuangan dan Manajemen*, vol. 17, no. 3, pp. 482–494, 2021, Accessed: Mar. 18, 2024. [Online]. Available: https://doi.org/10.30872/jinv.v17i3.9339

[5] L. Breiman, "Random Forests," *Mach Learn*, vol. 45, no. 1, pp. 5–32, 2001, Accessed: Sep. 17, 2024. [Online]. Available: https://doi.org/10.1023/A:1010933404324

[6] G. Biau and E. Scornet, "A Random Forest Guided Tour," *Test*, vol. 25, no. 2, pp. 197–227, Jun. 2016, doi: 10.1007/s11749-016-0481-7.

[7] Y. Ao, H. Li, L. Zhu, S. Ali, and Z. Yang, "The Linear Random Forest Algorithm and Its Advantages in Machine Learning Assisted Logging Regression Modeling," *J Pet Sci Eng*, vol. 174, pp. 776–789, Mar. 2019, doi: 10.1016/j.petrol.2018.11.067.

[8] C. Janiesch, P. Zschech, and K. Heinrich, "Machine Learning and Deep Learning," *Electronic Markets*, vol. 31, no. 3, pp. 685–695, Sep. 2021, doi: 10.1007/s12525-021-00475-2.

[9] Narayanan and Jayashree, "Implementation of Efficient Machine Learning Techniques for Prediction of Cardiac Disease using SMOTE," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 558–569. doi: 10.1016/j.procs.2024.03.245.

[10] Y. Kanoun, A. M. Aghbash, T. Belem, B. Zouari, and H. Mrad, "Failure Prediction in The Refinery Piping System Using Machine Learning Algorithms: Classification and Comparison," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 1663–1672. doi: 10.1016/j.procs.2024.01.164.


https://doi.org/10.25077/TEKNOSI.v12i2.2026.190-202

[11] C. Schröer, F. Kruse, and J. M. Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 526–534. doi: 10.1016/j.procs.2021.01.199.

[12] P. Chapman *et al.*, *CRISP-DM 1.0 Step-By-Step Data Mining Guide*. Amsterdam: SPSS Inc, 2000.

[13] I. N. Hidayati, T. F. Kusumasari, and F. Hamami, "Comparison of Apache SparkSQL and Oracle Performance: Case Study of Data Cleansing Process," *International Journal on Informatics Visualization*, vol. 6, no. 2, pp. 208–213, 2022, Accessed: Jul. 24, 2024. [Online]. Available: http://dx.doi.org/10.30630/joiv.6.1-2.928

[14] F. Ridzuan, W. M. Nazmee, and W. Zainon, "A Review on Data Cleansing Methods for Big Data," in *Procedia Computer Science*, Elsevier B.V., 2019, pp. 731–738. doi: 10.1016/j.procs.2019.11.177.

[15] U. Suriani, "Penerapan Data Mining untuk Memprediksi Tingkat Kelulusan Mahasiswa Menggunakan Algoritma Decision Tree C4.5," *Journal of Computer and Information Systems Ampera*, vol. 3, no. 2, pp. 55–66, 2023, doi: 10.51519/journalcisa.v4i2.393.

[16] K. Potdar, T. S. Pardawala, and C. D. Pai, "A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers," *Int J Comput Appl*, vol. 175, no. 4, pp. 7–9, 2017, Accessed: Jul. 24, 2024. [Online]. Available: https://doi.org/10.5120/ijca2017915495

[17] R. Larose and B. Coyle, "Robust data encodings for quantum classifiers," *Phys Rev A (Coll Park)*, vol. 102, no. 3, pp. 1–24, Sep. 2020, doi: 10.1103/PhysRevA.102.032420.

[18] M. S. Irwanto, F. A. Bachtiar, and N. Yudistira, "Klasifikasi Aktivitas Manusia Menggunakan Algoritme Computed Input Weight Extreme Learning Machine dengan Reduksi Dimensi Principal Component Analysis," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 9, no. 6, pp. 1195–1202, 2022, doi: 10.25126/jtiik.202295504.

[19] F. Rachmawati, J. Jaenudin, N. B. Ginting, and P. Laksono, "Machine Learning for the Model Prediction of Final Semester Assessment (FSA) using the Multiple Linear Regression Method," *Jurnal Teknik Informatika*, vol. 17, no. 1, pp. 1–9, May 2024, doi: 10.15408/jti.v17i1.28652.

[20] F. Aldi, F. Hadi, N. A. Rahmi, and S. Defit, "Standardscaler's Potential in Enhancing Breast Cancer Accuracy Using Machine Learning," *Journal of Applied Engineering and Technological Science*, vol. 5, no. 1, pp. 401–413, 2023, Accessed: Jul. 25, 2024. [Online]. Available: https://doi.org/10.37385/jaets.v5i1.3080

[21] Ridwan, E. H. Hermaliani, and M. Ernawati, "Penerapan Metode SMOTE Untuk Mengatasi Imbalanced Data pada Klasifikasi Ujaran Kebencian," *Computer Science* , vol. 4, no. 1, pp. 80–88, 2024, Accessed: Jul. 25, 2024. [Online]. Available: https://doi.org/10.31294/coscience.v4i1.2990

[22] E. Sutoyo and M. A. Fadlurrahman, "Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Television Advertisement Performance Rating Menggunakan Artificial Neural Network," *Jurnal Edukasi dan Penelitian Informatika*, vol. 6, no. 3, pp. 379–385, 2020, Accessed: Sep. 13, 2024. [Online]. Available: http://dx.doi.org/10.26418/jp.v6i3.42896

[23] V. R. Joseph, "Optimal Ratio for Data Splitting," *Stat Anal Data Min*, vol. 15, no. 4, pp. 531–538, Aug. 2022, doi: 10.1002/sam.11583.

[24] N. M. Kebonye, "Exploring The Novel Support Points-Based Split Method On A Soil Dataset," *Measurement (Lond)*, vol. 186, no. 58, pp. 1–4, Dec. 2021, doi: 10.1016/j.measurement.2021.110131.

[25] C. Zucco, "Multiple learners combination: Bagging," *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, vol. 1, no. 3, pp. 525–530, Jan. 2018, doi: 10.1016/B978-0-12-809633-8.20345-2.

[26] D. Nishfi, I. Huda, C. Prianto, and R. M. Awangga, "Analisis Sentimen Perbandingan Layanan Jasa Pengiriman Kurir pada Ulasan Play Store Menggunakan Metode Random Forest dan Descision Tree," *Jurnal Ilmiah Informatika (JIF)*, vol. 11, no. 2, pp. 150–158, 2023, Accessed: Aug. 07, 2024. [Online]. Available: https://doi.org/10.33884/jif.v11i02.7952

[27] F. Y. Pamuji and V. P. Ramadhan, "Jurnal Teknologi dan Manajemen Informatika Komparasi Algoritma Random Forest Dan Decision Tree untuk Memprediksi Keberhasilan Immunotheraphy," *Jurnal Teknologi dan Manajemen Informatika*, vol. 7, no. 1, pp. 46–50, 2021, Accessed: Aug. 07, 2024. [Online]. Available: https://doi.org/10.26905/jtmi.v7i1.5982

[28] L. Langsetmo *et al.*, "Advantages and Disadvantages of Random Forest Models for Prediction of Hip Fracture Risk Versus Mortality Risk in the Oldest Old," *Journal of Bone and Mineral Research Plus*, vol. 7, no. 8, pp. 1–11, Aug. 2023, doi: 10.1002/jbm4.10757.

[29] T. M. Khoshgoftaar, M. Golawala, and J. Van Hulse, "An empirical study of learning from imbalanced data using random forest," in *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, 2007, pp. 310–317. doi: 10.1109/ICTAI.2007.46.

[30] E. Renata and M. Ayub, "Penerapan Metode Random Forest untuk Analisis Risiko pada Dataset Peer to Peer Lending," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 6, no. 3, pp. 462–474, Dec. 2020, doi: 10.28932/jutisi.v6i3.2890.

[31] V. Nguyen, "Bayesian optimization for accelerating hyper-parameter tuning," in *Proceedings - IEEE 2nd International Conference on Artificial Intelligence and Knowledge Engineering, AIKE 2019*, Institute of Electrical and Electronics Engineers Inc., Jun. 2019, pp. 302–305. doi: 10.1109/AIKE.2019.00060.

[32] Q. Yang *et al.*, "Research on Surrogate Models and Optimization Algorithms of Compressor Characteristic Based on Digital Twins," *Journal of Engineering Research (Kuwait)*, vol. 5, no. 4, pp. 1–13, 2024, doi: 10.1016/j.jer.2024.01.025.

[33] E. Elgeldawi, A. Sayed, A. R. Galal, and A. M. Zaki, "Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis," *Informatics*, vol. 8, no. 4, pp. 1–21, Dec. 2021, doi: 10.3390/informatics8040079.

[34] D. M. Belete and M. D. Huchaiah, "Grid Search in Hyperparameter Optimization of Machine Learning Models for Prediction of HIV/AIDS Test Results," *International Journal of Computers and Applications*, vol. 44, no. 9, pp. 875–886, 2022, doi: 10.1080/1206212X.2021.1974663.

[35] S. Andradóttir, "A Review of Random Search Methods," *International Series in Operations Research Management & Science*, vol. 7, no. 4, pp. 277–292, 2014, doi: 10.1007/978-1-4939-1384-8_10.

[36] S. Greenhill, S. Rana, S. Gupta, P. Vellanki, and S. Venkatesh, "Bayesian Optimization for Adaptive Experimental Design: A Review," *IEEE Access*, vol. 8, no. 5, pp. 13937–13948, 2020, doi: 10.1109/ACCESS.2020.2966228.

[37] D. Normawati and S. A. Prayogi, "Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter," *Jurnal Sains Komputer & Informatika (J-SAKTI*, vol. 5, no. 2, pp. 697–711, 2021, Accessed: Jul. 31, 2024. [Online]. Available: http://dx.doi.org/10.30645/j-sakti.v5i2.369

[38] Z. A. Dwiyanti and C. Prianto, "Prediksi Cuaca Kota Jakarta Menggunakan Metode Random Forest," *Jurnal Tekno Insentif*, vol. 17, no. 2, pp. 127–137, Oct. 2023, doi: 10.36787/jti.v17i2.1136.

[39] A. H. Bik, F. T. Anggraeny, and E. Y. Puspaningrum, "Klasifikasi Penyakit Ginjal Menggunakan Algoritma Hibrida CNN-ELM," *Jurnal Mahasiswa Teknik Informatika*, vol. 8, no. 3, pp. 3836–3844, 2024, Accessed: Aug. 02, 2024. [Online]. Available: https://doi.org/10.36040/jati.v8i3.9807

[40] D. T. Wilujeng, M. Fatekurohman, and I. M. Tirta, "Analisis Risiko Kredit Perbankan Menggunakan Algoritma K-Nearest Neighbor dan Nearest Weighted K-Nearest Neighbor," *Indonesian Journal of Applied Statistics*, vol. 5, no. 2, pp. 142–148, Oct. 2023, doi: 10.13057/ijas.v5i2.58426.

[41] R. N. Irawan, K. M. Hindrayani, and M. Idhom, "Penerapan Cross Validation sebagai Analisis Sentimen Pelayanan Publik Kereta Api Lokal Daop 8 Menggunakan Metode Multinomial Naïve Bayes," *G-Tech: Jurnal Teknologi Terapan*, vol. 8, no. 2, pp. 954–963, Apr. 2024, doi: 10.33379/gtech.v8i2.4117.

[42] H. Azis, Purnawansyah, F. Fattah, and I. P. Putri, "Performa Klasifikasi K-NN dan Cross Validation Pada Data Pasien Pengidap Penyakit Jantung," *ILKOM Jurnal Ilmiah*, vol. 12, no. 2, pp. 81–86, Aug. 2020, doi: 10.33096/ilkom.v12i2.507.81-86.

[43] H. Hafid, "Penerapan K-Fold Cross Validation untuk Menganalisis Kinerja Algoritma K-Nearest Neighbor pada Data Kasus Covid-19 di Indonesia," *Journal of Mathematics, Computations, and Statistics*, vol. 6, no. 2, pp. 161–168, 2023, Accessed: Jul. 29, 2024. [Online]. Available: https://doi.org/10.35580/jmathcos.v6i2.53043