

Classification of Pension Program Service Workload Levels Across Payment Cities at PT ASABRI (Persero) Bandung Branch Office Using Machine Learning Models

Angela Nisrina Zhafir^{1*}, Nur Alamsyah²

Sistem Informasi^{1,2}
Universitas Informatika dan Bisnis Indonesia, Bandung, Indonesia^{1,2}
<https://unibi.ac.id/>^{1,2}
angela.nz22@student.unibi.ac.id¹, nuralamsyah@unibi.ac.id²

Abstract. This study classifies service workload levels across 13 cities served by the PT ASABRI (Persero) Bandung Branch Office to mitigate evaluation bias by implementing the CRISP-DM methodology on transaction data from 2023 to 2024. The target variable was constructed by combining multiple operational variables into a composite score index, using percentile cutoffs to define High, Medium, and Low workload categories. Under an Out-of-Time Validation (OOTV) framework, data from 2023 served as training data and data from 2024 as testing data, followed by the application of the Synthetic Minority Oversampling Technique (SMOTE) exclusively to the training data to neutralize class imbalance. Comparative evaluation identified a Random Forest model optimized via RandomizedSearchCV as the top-performing model, achieving an accuracy of 83.33%, a macro F1-score of 81.17%, and minimal performance discrepancy between training and test sets. Architecture stability was demonstrated through 5-fold cross-validation, yielding an average accuracy of 72.42% with a minimal standard deviation of 5.69%. Global Feature Importance analysis identified the total number of family dependents (TOTAL_JIWA_DILAYANI) as the dominant operational factor, with an importance score of 0.4145. This comprehensive modeling provides a measurable Early Warning System that enables management to prioritize evaluation areas more systematically while minimizing subjective assumptions.

Key words: ASABRI; service workload; percentile composite score; Random Forest; global feature importance

1. INTRODUCTION

Public services at PT ASABRI (Persero) in the digital transformation era generate historical operational data for each payment city. In practice, the utilization of this historical data remains limited to the company's administrative and operational reporting needs [1], [2]. Consequently, operational data has not been sustainably leveraged for future decision-making evaluations or for identifying service workload characteristics across locations. Transforming this approach from passive administrative reporting to an objective, data-driven evaluation tool is now critical [3], [4]. This transformation is particularly relevant for pension program administrators at the Bandung Branch Office of PT ASABRI (Persero), which oversees operations across 13 payment cities with diverse activity characteristics.

Mapping efforts to date have been hampered by the lack of standardized quantitative indicators to systematically

evaluate service workload. Historical pension program payment data lacks target variables that directly represent service workload levels. This absence of labeled data hinders objective identification of workload characteristics, leaving evaluations vulnerable to management bias. Previous research has indeed implemented machine learning as a classification solution in the public sector, [5], [6], [7]. However, research specific to the PT ASABRI (Persero) Bandung Branch Office remains very limited. Furthermore, the majority of existing studies rely on fully available target variables [8], [9]. A methodological gap exists in developing target variable engineering with multiple operational variables prior to classification, as well as addressing potential class imbalance in data distribution [10]. To date, much predictive research focuses solely on accuracy metrics without providing diagnostic capabilities through Explainable AI (XAI) for stakeholders to identify key underlying factors [11], [12].



To address these issues and methodological gaps, this study employs a machine learning classification approach using the CRISP-DM methodology, which generates a foundation for organizational evaluation by transforming raw data into objective and measurable insights [13], beginning with the formulation of target variables for service workload levels. This study integrates multiple operational variables using composite scores and percentile cutoff segmentation to map service workload levels into three categories: High, Medium, and Low. Subsequently, model selection among several classification algorithms is performed to identify the optimal model [12]. followed by Global Feature Importance analysis to diagnose dominant factors affecting service workload [14]. Comprehensively, this study aims to construct target variables to map service workload across 13 payment cities of the PT ASABRI (Persero) Bandung Branch Office into three categories using the top-performing algorithm, as well as to identify primary factors influencing service workload. The novelty of this work lies in the integration of multi-variable target engineering techniques using composite scores and percentile cutoffs prior to classification. The contribution of this research is transforming historical data into an evaluation support tool for objective managerial decision-making to identify essential service workload characteristics needed in the future.

2. RELATED WORK

Several studies have explored the application of machine learning across various domain sectors and organizations. Driven by the shift toward data-driven decision-making and organizational big data evaluation, machine learning offers promising solutions to complex operational challenges.

D. Setiawan et al. [15] applied the CRISP-DM framework and a Random Forest classifier to predict bank customer churn based on behavioral and financial indicators, achieving an accuracy of 86% with a 70:30 holdout split. Their findings demonstrate high reliability in identifying customer churn risk to support proactive retention strategies. A similar study conducted by F. Dong [16] applied human resource (HR) data to predict employee turnover, categorizing metrics into two domains performance evaluation and effectiveness and validating model performance using AUC and F1-score metrics. Furthermore, Random Forest was also applied by A. Zermane et al. [17] using a regression approach oriented toward public service operations. Similarly, I. Aruleba et al. [18] utilized this algorithm alongside feature importance analysis to support in-depth evaluation. Public service organizations were highlighted by P. Pampouktsi et al. [19], who compared four classification algorithms to identify the optimal model with high relevance while mapping the outcomes into four categories. Similar model comparison approaches were also employed by Nofriani [20], R. Das [21], and E.

Qumsiyeh et al. [22] as a benchmark to ensure that machine learning classification achieves optimal performance.

Furthermore, C. Satria et al. [23] applied the Decision Tree algorithm to enhance classification models for regencies and cities across Indonesia, successfully categorizing regional socio-economic strata to support targeted government programs. Similarly, D. Qin et al. [24] applied spatial machine learning to estimate county-level rural digital economy and infrastructure readiness.

Based on the literature review, the majority of research on machine learning classification focuses on comparing model performance using datasets with predefined variables. A research gap exists regarding how to construct target variables from large-scale raw operational data. Therefore, this study offers novelty through the integration of target variable engineering (composite scores and percentiles) prior to modeling, alongside Global Feature Importance. This approach explicitly transforms historical data into an objective service workload evaluation tool for the PT ASABRI (Persero) Bandung Branch Office.

3. METHODS

The research methodology is based on the CRISP-DM framework, with stages adapted to align with the operational needs of the PT ASABRI (Persero) Bandung Branch Office. Each stage is designed to ensure that the process yields valid and practically relevant outcomes for implementation.

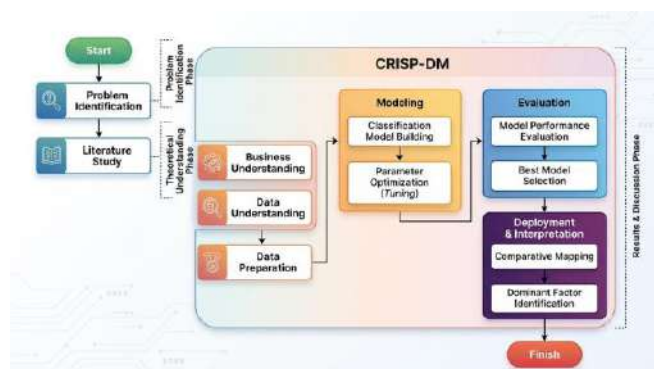


Fig. 1 Research Workflow

3.1. Problem Identification

Observations were conducted to perform a gap analysis on service workload monitoring across 13 payment cities, where current evaluations remain subjective. The absence of quantitative evaluation standards potentially triggers inconsistent regional workload mapping. This review aims to formulate model requirements and generate accurate outputs.



3.2. Literature Study

Theoretical foundations were collected to establish the methodological background. This stage aims to ensure that the selected method aligns with the problem statement, specifically classifying service workload levels accurately and identifying dominant factors influencing the company's operational workload based on a literature review

3.3. Business Understanding

This stage aligns managerial requirements with technical objectives. Execution is carried out through the design of a classification model with three service workload levels (High, Medium, Low) across 13 payment cities as a measurable, objective, and reusable evaluation instrument. The process begins with raw data from the Pension Program Historical Transaction Reports for the 2023–2024 period, exceeding 1,000,000 record rows across 13 columns. The study utilizes a cloud-based system via Google Colaboratory and the Python programming language. The computational power of cloud servers ensures that large-scale data processing does not strain the organization's local hardware.

3.4. Data Understanding

The study focused on secondary data and involved discussions with operational staff to align business logic, variable definitions, and labeling parameters. The process began with data profiling and attribute selection to protect agency and retiree confidentiality by filtering relevant variables aligned with research objectives. This process yielded 11 candidate variables, comprising four Y (target) variables (NOPENS, NRP, JML_NETTO, JNSBYR) and seven X (feature) variables (UNOR, PANGKAT, JENIS_PENSIUN, STATUS_PESERTA_PENSIUN, KODE_JIWA, NAMA_MITRA, and MATA_ANGGARAN_KELUARAN/MAK). This was followed by data quality verification, addressing missing values, duplicate data, outliers, data and formatting inconsistencies, noise, and structural errors. Exploratory Data Analysis (EDA) was also dynamically integrated throughout this phase.

3.5. Data Preparation

Data transformation into a structured dataset is performed prior to the modeling phase. Data cleansing is integrated into the data verification process to ensure that data quality does not compromise algorithm performance stability. This process includes attribute harmonization, city filtering to match the 13 payment cities under the Bandung Branch Office, and handling missing values and duplicate data using the row deletion method, resulting in final observation counts of 580,258 for 2023 and 592,426 for 2024. Candidate target variables between NOPENS and NRP were evaluated by comparing unique population counts, total observation ratios, and technical feasibility tests, which revealed NOPENS accounted for 8.71%–8.83% of total

observations; thus, the selected target variables were set to NOPENS, IS_LATE, and JML_NETTO. Feature transformation adjusts variables for machine learning execution through various techniques based on variable characteristics, as detailed below:

TABLE 1. Feature transformation

Initial Attributes	Processing Method	Transformed Attributes
JNBYR, UNOR, JENIS_PENSIUN, STATUS_PESERTA_PENSIUN, MAK	Label Encoding	IS_LATE, UNOR_LE, JENIS_PENSIUN_LE, STATUS_LE, MAK_LE
PANGKAT, NAMA MITRA	Grouping & Label Encoding	PANGKAT_LE, MITRA_LE
KODE_JIWA	Feature Construction	TOTAL_JIWA_DILAYANI

The cleaned dataset is subsequently aggregated to align the analysis unit across three dimensions: Payment City, Month, and Year, with the operational mapping structured as follows:

TABLE 2. Aggregation Method

Attribute	Method
NOPENS	Nunique
JML_NETTO, TOTAL_JIWA_DILAYANI	Sum
IS_LATE, UNOR_LE, JENIS_PENSIUN_LE, STATUS_LE, MAK_LE, PANGKAT_LE, MITRA_LE	Mean

Outlier processing was applied to the JML_NETTO variable using logarithmic transformation. Target variable construction involves normalizing operational indicators and integrating them into a multivariate target variable Y using a composite score. An equal weighting of 33.33% was assigned to each indicator (NOPENS, IS_LATE, and JML_NETTO) based primarily on operational considerations, as field conditions show that service volume, processing delay, and financial scale contribute equally to the overall workload. This balanced weighting prevents any single indicator from dominating the composite score, defined as follows:

$$Y = \sum_{i=1}^n w_i x_i \quad (1)$$

where Y represents the final composite workload score, w_i denotes the assigned weight or relative importance of the i -th variable, and x_i is the normalized value of the i -th input variable. This weighted formulation ensures that each operational criterion contributes proportionally to the final service workload evaluation.

Mapping service workload categories (High, Medium, Low) employs a percentile-based threshold approach to promote objective class segmentation aligned with data distribution, defined as follows:



$$\begin{aligned} \text{idx}_{\{25\}} &= 0,25(n - 1) + 1 \\ \text{idx}_{\{75\}} &= 0,75(n - 1) + 1 \\ Q &= y_i + f(y_{\{i+1\}} - y_i) \end{aligned} \quad (2)$$

where $\text{idx}_{\{25\}}$ and $\text{idx}_{\{75\}}$ define the rank index positions for percentiles $P_{\{25\}}$ and $P_{\{75\}}$, f represents the fractional rank component, and y_i is the actual observation value at position i . These percentile thresholds explicitly categorize the composite workload score (Y) into Low ($Y \leq P_{25}$), Medium ($P_{25} < Y < P_{75}$), and High ($Y \geq P_{75}$) classes. Cutoff thresholds are established using the $P_{\{25\}}$ and $P_{\{75\}}$ percentiles, facilitating the identification of observations with extreme characteristics. This segmentation categorizes the dataset into 158 data points in the Medium category, 78 in the Low category, and 76 in the High category. Subsequently, feature selection is performed using a correlation matrix heatmap, as illustrated in Fig. 2.

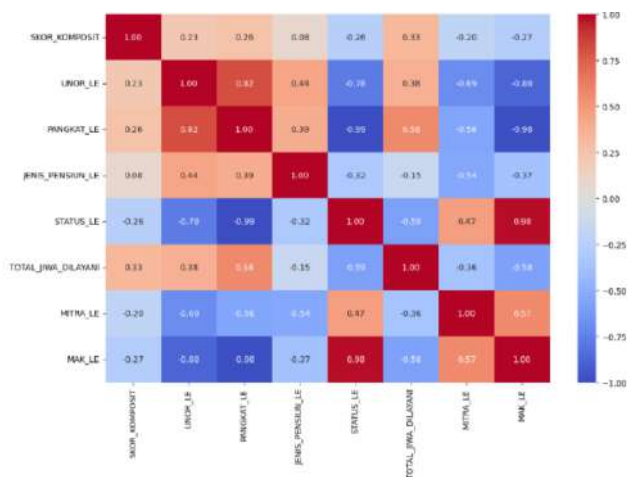


Fig. 2 Inter-Feature Correlation Matrix

A strong correlation occurred within the group of membership identity attributes, where redundancy served as the primary basis for simplification into a single representative variable, namely PANGKAT_LE, based on the core operational framework of retiree entitlements. Consequently, PANGKAT_LE, TOTAL_JIWA_DILAYANI, MITRA_LE, and JENIS_PENSIUN_LE were selected as the primary feature set (X variables).

Data splitting was performed using Out-of-Time Validation (OOTV), utilizing 2023 data as the training set (156 samples) and 2024 data as the independent test set (156 samples), aligning with the annual operational cycle of PT ASABRI (Persero) Bandung Branch Office. To effectively mitigate data leakage, all preprocessing steps including Min-Max feature scaling and class rebalancing via Synthetic Minority Over-sampling Technique (SMOTE) were fitted exclusively on the 2023 training data. The 2023 training set initially comprised 78 Medium, 39 Low, and 39 High samples, which were balanced via SMOTE to 78 samples per category (234 total resampled training samples). Conversely, the independent 2024 test set (80 Medium, 39 Low, and 37 High samples) was maintained in its original

distribution without any SMOTE oversampling to evaluate real-world performance.

3.6. Modeling

The model-building process to identify the optimal algorithm is executed in two main steps:

- Classification Model Building:** Algorithm training evaluates four distinct mathematical logic representation approaches, including Logistic Regression as a linear probability model, K-Nearest Neighbors (KNN) as a distance-based classification method, Decision Tree as a single hierarchical decision-rule model, and Random Forest as an ensemble model combining multiple decision trees.
- Parameter Optimization (Tuning):** To configure an appropriate architecture, improve generalization capability, and mitigate overfitting risks, RandomizedSearchCV integrated with 5-Fold Cross-Validation was conducted exclusively within the resampled 2023 training dataset. Hyperparameter tuning incorporates $n_estimators$ (10, 30, 70, 100, 130, 180, 200, 300), max_depth (5, 8, 12, 15, 25, None), $min_samples_split$ (2, 4, 6, 8, 12), and $min_samples_leaf$ (1, 2, 3, 5). Once optimal hyperparameters were identified during training cross-validation, the champion model was evaluated against the unseen 2024 test dataset.

3.7. Evaluation

The developed models are validated through two main components:

- Model Performance Evaluation:** Model capability is measured using Accuracy, Precision, Recall, and F1-Score metrics. Confusion Matrix analysis complements the testing to evaluate the model's ability to distinguish among the three service workload categories while minimizing prediction bias. Stratified K-Fold cross-validation with $shuffle=True$ and 5 folds is employed to measure stability across data splits. The standard deviation serves as a variance metric to validate model consistency and resilience against overfitting.
- Best Model Selection:** The optimal algorithm is selected by comparing performance results across evaluation metrics. The Champion Model selection identifies a single final model that demonstrate stable, balanced, and operationally relevant performance.

3.8. Deployment and Interpretation

Model implementation and variable analysis results serve as the primary research deliverables, yielding two key components:

- Comparative Mapping:** Two matrices are presented side by side: the Actual Field Matrix, representing operational reality, and the Model Classification Matrix, serving as the benchmark for organic capacity. Management can objectively detect anomalies,



inefficiencies, and system disruptions by evaluating color discrepancies (deltas) between the two matrices across 13 payment cities in the service region.

- b. Identification of Dominant Factors: Global feature importance serves as a metric and interpretation tool for identifying significant variables influencing service workload, supporting objective policy decisions by the organizational team during post-mortem evaluations.

4. RESULTS AND DISCUSSIONS

4.1. Target variable mapping

Target variable construction was specifically mapped using the training data to prevent data leakage, yielding standardized quantitative indicators for the pension service workload at the PT ASABRI (Persero) Bandung Branch Office. The composite score statistics are detailed as follows:

TABLE 3. 2023 Composite Score Statistics

Statistical Metric	Score
Count	156,0000
Mean	0,2864
Standard Deviation	0,1465
Minimum	0,0000
P ₂₅	0,2092
P ₅₀ (Median)	0,2598
P ₇₅	0,3423
Maximum	0,8032

The concentration of observations occurs within the low-to-medium score range. The extreme difference between the quartile boundaries forms a right-skewed distribution pattern, reflecting organic variations in field service workload dynamics that peak sharply across a small portion of the operational area, as illustrated in Fig. 3.

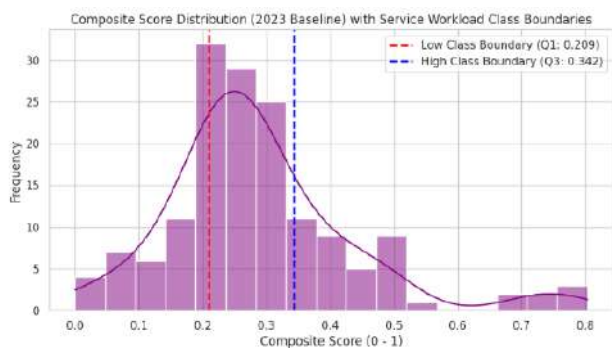


Fig. 3 Composite Score Distribution (Service Workload) 2023

Threshold values were established at 0.2093 for P₂₅ and 0.3423 for P₇₅. These obtained values were established as baseline reference thresholds, allowing data from subsequent periods to be evaluated consistently against standard service workload criteria. Data labeling converts the numerical composite scores into discrete categories. The division across two annual periods categorizes the

population into 158 Medium, 78 Low, and 76 High data points.

4.2. Modeling Performance Evaluation

This study evaluated the performance of four baseline classification algorithms (Logistic Regression, K-Nearest Neighbors, Decision Tree, and Random Forest) trained on the un-resampled 2023 training data and evaluated against the unseen 2024 test dataset, with the initial baseline results presented in Table 4.

TABLE 4. Baseline Classification Model Performance Evaluation Results

Algorithm	Accuracy	Precision	Recall	F1-Score
Random Forest	83,97%	88,77%	78,03%	81,41%
K-Nearest Neighbors	78,21%	87,14%	70,34%	74,35%
Logistic Regression	76,28%	85,93%	67,78%	71,44%
Decision Tree	50,64%	51,98%	54,43%	51,76%

The high initial evaluation scores indicated an "accuracy paradox" caused by class imbalance within the training dataset, as illustrated in Fig. 4.

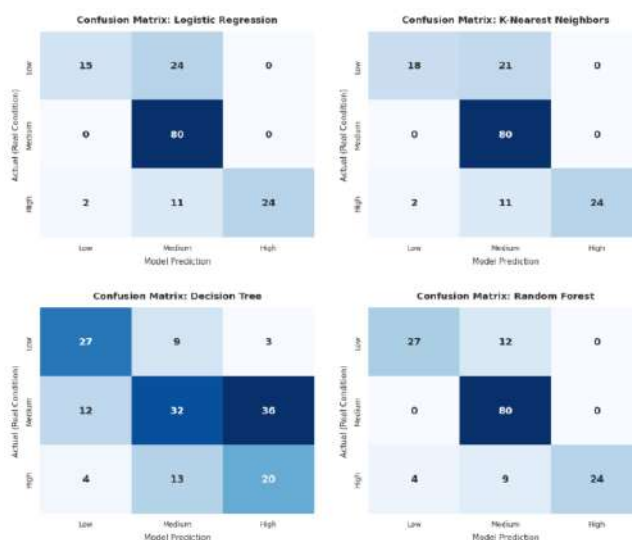


Fig. 4 Confusion Matrix (Before SMOTE)

The confusion matrix revealed that this high performance was artificially inflated due to prediction bias toward the majority medium class. This necessitated applying SMOTE solely to the training data to rebalance class representations prior to model fitting, ensuring objective performance evaluation when evaluated against the unseen 2024 test set, as shown in Table 5.

TABLE 5. Optimal Classification Model Performance Evaluation Results

Algorithm	Accuracy (Before)	Accuracy (After)	F1-Score (Before)	F1-Score (After)
Logistic Regression	76,28%	69,87%	71,44%	66,96%
K-Nearest Neighbors	78,21%	75,64%	74,35%	73,40%
Decision Tree	50,64%	60,90%	51,76%	59,09%



Algorithm	Accuracy (Before)	Accuracy (After)	F1-Score (Before)	F1-Score (After)
Random Forest	83,97%	84,62%	81,41%	82,25%

The flexibility of the rule-based architecture proved adaptive in accommodating minority observations without compromising the precision of the majority class, as evidenced by shifts in accuracy and F1-score dominance across algorithms. Consequently, Random Forest with SMOTE which yielded the highest and most stable evaluation matrix results was selected as the primary algorithm for subsequent model development. This performance stability is further illustrated by the confusion matrix in Fig. 5.

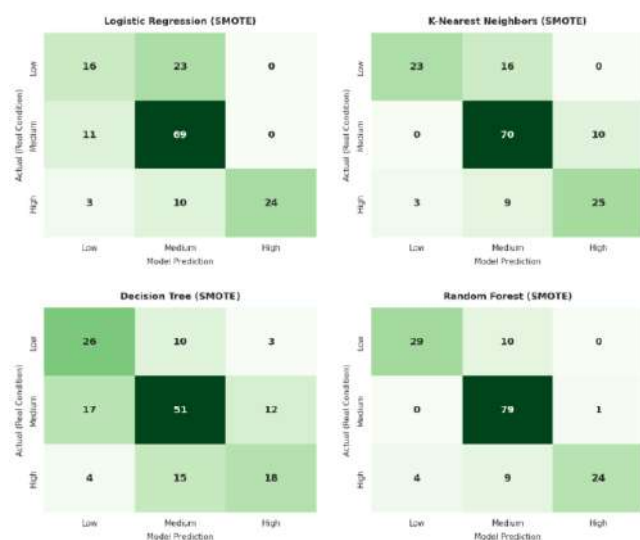


Fig. 5 Confusion Matrix (After SMOTE)

Optimal parameter combinations were identified using RandomizedSearchCV with the following settings: max_depth = 5, min_samples_leaf = 2, min_samples_split = 8, and n_estimators = 100. The resulting evaluation matrix is detailed in Table 6.

TABLE 6. RandomizedSearchCV Evaluation Results

	precision	recall	f1-score	support
Low	0.81	0.74	0.77	39
Medium	0.80	0.96	0.88	80
High	1.00	0.65	0.79	37
accuracy			0.83	156
macro avg	0.87	0.78	0.81	156
weighted avg	0.85	0.83	0.83	156

The overfitting gap analysis is presented under two distinct conditions:

TABLE 7. Overfitting Gap Analysis

Statistical Metric	Standard Random Forest Accuracy	RandomizedSearchCV Random Forest Accuracy
Training Data Accuracy	100,00%	83,33%
Test Data Accuracy	84,62%	83,33%
Gap	15,38%	0,00%

An accuracy of 83.33% across both training and testing datasets indicates good generalization with no apparent evidence of substantial overfitting. The equivalent accuracy observed in the RandomizedSearchCV model reflects balanced performance across evaluation sets, supported by the symmetrical dataset of 156 observations in both periods following annual data aggregation, resulting in a comparable number of correct predictions. To demonstrate the independence of the evaluation sets and address potential concerns regarding data leakage, the target class distribution and descriptive statistics of the training (2023) and test (2024) datasets are presented in Table 8, Table 9, and Table 10.

TABLE 8. Target Class Distribution (y_train vs. y_test)

Service Workload Category	Training Data (2023)	Test Data (2024)
Medium	78 Observations	80 Observations
High	39 Observations	37 Observations
Low	39 Observations	39 Observations
Total	156 Observations	156 Observations

TABLE 9. Descriptive Statistics of the Training Dataset (2023)

Statistical Metric	PANGKAT (Label Encoded)	TOTAL_JIWA_DILAYANI	MITRA (Label Encoded)	JENIS_PENSIUN (Label Encoded)
Data Count	156.0000	156.0000	156.0000	156.0000
Mean	0.8683	6332.3141	1.3041	1.4042
Standard Deviation	0.2220	9613.4564	0.2107	0.0237
Minimum	0.4000	288.0000	0.9990	1.3267
P ₂₅	0.7339	2129.5000	1.1996	1.4008
P ₅₀ (Median)	0.8628	3489.0000	1.2603	1.4093
P ₇₅	1.0010	4758.5000	1.3228	1.4196
Maximum	1.2172	38511.0000	1.9097	1.4372

TABLE 10. Descriptive Statistics of the Test Dataset (2024)

Statistical Metric	PANGKAT (Label Encoded)	TOTAL_JIWA_DILAYANI	MITRA (Label Encoded)	JENIS_PENSIUN (Label Encoded)
Data Count	156.0000	156.0000	156.0000	156.0000
Mean	0.8643	6419.5962	1.2982	1.4100



Statistical Metric	PANGKAT (Label Encoded)	TOTAL_JIWA_DILAYANI	MITRA (Label Encoded)	JENIS_PENSIUN (Label Encoded)
Standard Deviation	0.2134	9722.0255	0.1933	0.0209
Minimum	0.4611	315.0000	0.9882	1.3542
P_{25}	0.7254	2234.0000	1.2343	1.4008
P_{50} (Median)	0.8521	3523.0000	1.2641	1.4130
P_{75}	0.9935	4800.5000	1.3370	1.4270
Maximum	1.2078	39029.0000	1.8444	1.4384

Based on the aforementioned analysis and empirical evaluation, the tuned RandomizedSearchCV model demonstrates the most balanced performance among the evaluated architectures, with its confusion matrix illustrated in Fig. 6.

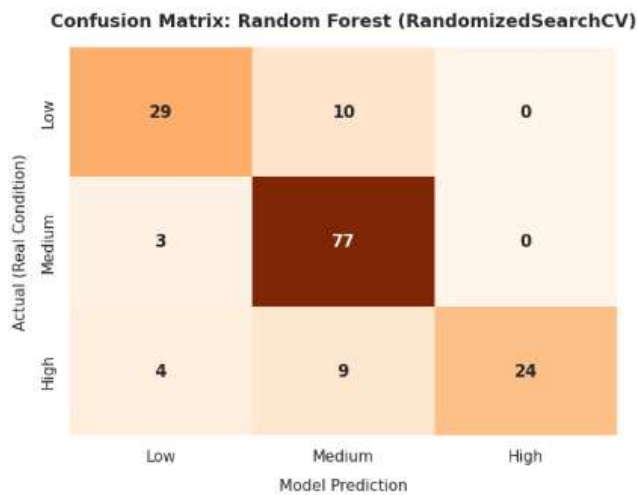


Fig. 6 Confusion Matrix (Random Tuning)

Cross-validation evaluation was mapped under two conditions for comparison to evaluate generalization capability, as shown in Fig. 7.

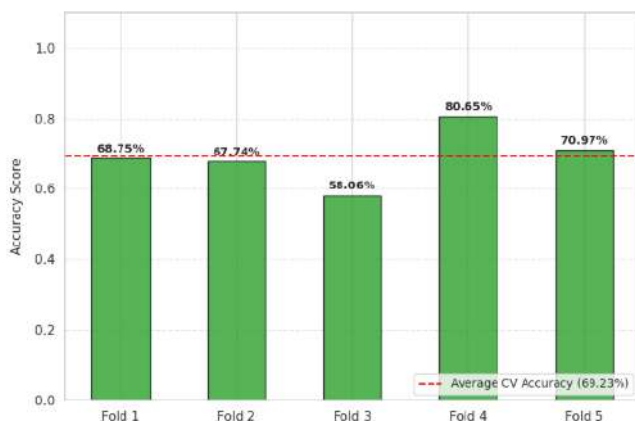


Fig. 7 Cross-Validation (Without Tuning)

The baseline Random Forest model exhibits a wide deviation range with a relatively lower mean score, whereas the performance distribution of the tuned Random Forest using RandomizedSearchCV is shown in Fig. 8.

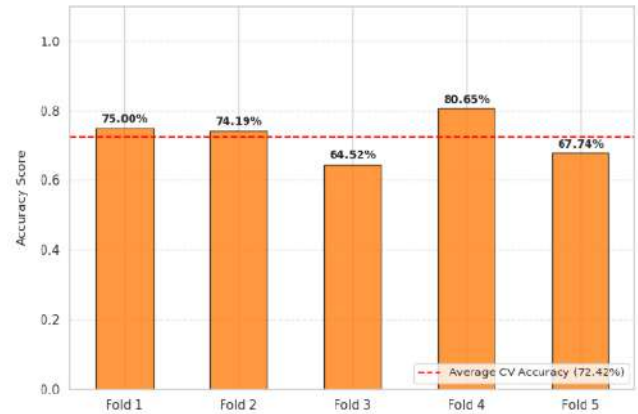


Fig. 8 Cross Validation RandomizedSearchCV Tuning

The tight and consistent accuracy distribution across folds, along with a minimal standard deviation, demonstrates the efficacy of the dynamic random search method in reducing algorithm sensitivity to noise. Consequently, the Champion Model was established as the Random Forest model with SMOTE and RandomizedSearchCV hyperparameter tuning. Model results are mapped through a dual visualization framework comparing field reality with standard regional organic capacity, as shown in Fig. 9.



Actual vs Predicted Workload Comparison Matrix (Tuned Randomized Search Model)
PT ASABRI KC Bandung - 2024



Fig. 9 Visualization of Classification Results

The left matrix represents actual field conditions, while the right matrix displays machine learning predictions for standard regional organic capacity. Standard organic capacity defines the reasonable workload baseline calculated purely from retiree demographic profiles. The visualization establishes a consistent baseline, reflecting minimal drastic demographic shifts over short timeframes. Color analysis indicates that matrix alignment signifies normal operational continuity; an actual level below the prediction indicates successful performance optimization exceeding baseline standards; whereas an actual status exceeding the prediction reveals service workload inefficiencies beyond reasonable demographic limits.

4.3. Indicators of Service Workload Influence

Global feature importance provides an empirical indicator of the mathematical significance of service workload factors. The evaluation results demonstrate the contribution magnitude of the four primary variables shaping the organic service workload of retirees, as illustrated in Fig. 10.

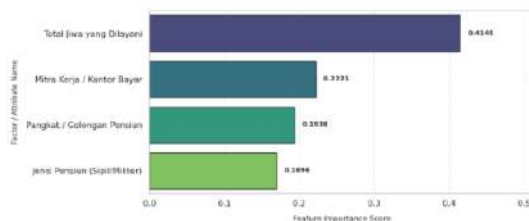


Fig. 10 Global Feature Importance Visualization management realities. The mathematical identification of dominant factors establishes the strategic foundation for management

evaluation priorities at the PT ASABRI (Persero) Bandung Branch Office across the Low, Medium, and High service workload categories.

4.4. Discussions

The implementation of the CRISP-DM framework in this study processed large-scale unlabeled historical data. Methodological novelty was achieved through target engineering, where multivariate complexity was consolidated into a single composite score index and categorized by percentiles ($P_{25} = 0.2093$ and $P_{75} = 0.3423$). This approach addresses the limitations of subjective management evaluations [1], [2], while distinguishing this study from prior research relying on instant targets [8], [9]. The interaction between SMOTE and RandomizedSearchCV optimization in the Random Forest algorithm yielded stable performance compared to single baseline algorithms [19], [20]. The model achieved an accuracy of 83.33%, an F1-score of 81.17%, and a performance gap of 0.00% (optimal fit), aligned with previous findings [15], [16] regarding the advantages of ensemble learning in mitigating overfitting risks. Using global feature importance, the variable TOTAL_JIWA_DILAYANI was identified as the dominant factor with a contribution of 41.45%. This analysis complements previous research that was limited to evaluating accuracy metrics without the diagnostic capabilities of XAI [11], [12]. Empirical findings demonstrate that service complexity across the 13 payment cities under the PT ASABRI (Persero) Bandung Branch Office is predominantly determined by the verified family dependents population size for each retiree, rather than other membership characteristics.



Comparative mapping of actual conditions against predicted organic capacity serves as an early warning system. In line with organizational evaluation classification concepts [13], these results serve as a decision-support tool for measurable managerial policy. High-priority regions are designated for additional human resource allocation and server capacity enhancement, whereas Low-priority regions are evaluated for resource efficiency. Future research is recommended to expand the operational scope to other branch offices of PT ASABRI (Persero).

5. CONCLUSIONS

This study classified the service workload levels of the pension program across 13 payment cities at the PT ASABRI (Persero) Bandung Branch Office into High, Medium, and Low categories based on historical data, while identifying dominant factors for continuous performance monitoring. The implementation of the CRISP-DM framework and multivariate target construction based on composite scores and percentiles successfully transformed transaction data that previously lacked objective labels. The Random Forest machine learning model, integrated with SMOTE and RandomizedSearchCV, demonstrated superior performance among the evaluated algorithms, achieving an optimal fit and high classification accuracy. Based on global feature importance analysis, the verified family dependents population (TOTAL_JIWA_DILAYANI) was identified as the most influential factor driving service workload. In practical implementation, combining the classification model with comparative matrix visualization functions directly as a reusable early warning system. This tool provides measurable guidance for operational management, establishes priority areas for evaluation, optimizes the allocation of human resources and infrastructure, and informs strategies for targeted service interventions in subsequent operational periods.

REFERENCES

- [1] N. I. Taufik, I. Sulistianti, and H. A. Pradesa, "Penilaian risiko pada layanan pembayaran pensiun PT Asabri Bandung : Sebuah praktek terbaik untuk penguatan tata kelola perusahaan," vol. 5, no. 2, pp. 857–867, 2022.
- [2] N. Aulia, S. Wanti, M. F. Yani, R. Anggraini, N. Hidayati, and M. Huda, "Analisis Pentingnya Kompensasi Pensiun Dan Tabungan Hari Tua Sebagai Hak Peserta PT Taspen (Persero)," vol. 1, no. June, pp. 279–286, 2024.
- [3] H. Raza, A. Singh, T. Erdenetsogt, M. M. Kabeer, M. S. Aslam, and M. Farooq, "Machine Learning Driven Decision Making in the Modern Data Era," PERFECT J. Smart Algorithms, vol. 3, no. 1, pp. 11–22, 2026.
- [4] L. Theodorakopoulos, A. Theodoropoulou, and C. Halkiopoulos, "Cognitive bias mitigation in executive decision-making: A data-driven approach integrating big data analytics, AI, and explainable systems," Electronics, vol. 14, no. 19, p. 3930, 2025.
- [5] D. Ruete et al., "Improving Public Sector Efficiency Through Rapid Machine Learning-Based Product Classification," in International Conference on Advanced Research in Technologies, Information, Innovation and Sustainability, Springer, 2025, pp. 3–15.
- [6] I. A. Liza et al., "Heart disease risk prediction using machine learning: A data-driven approach for early diagnosis and prevention," Br. J. Nurs. Stud., vol. 5, no. 1, pp. 38–54, 2025.
- [7] E. Elabd, H. M. Hamouda, M. A. M. Ali, A. S. Hamid, and Y. Fouad, "Air quality index AQI classification based on hybrid particle swarm and grey wolf optimization with ensemble machine learning model," Sci. Rep., vol. 16, no. 1, p. 4152, 2026.
- [8] N. A. Tohfa et al., "Predicting Heart Disease Using Machine Learning and Ensemble Models: A Comparative Study," in 23 RD INTERNATIONAL CONFERENCE ON COMPUTER APPLICATIONS, 2026.
- [9] S. Sinharoy, H. Reese, T. Clasen, and S. S. Sinharoy, "Applying machine learning to predict stunting in children under 5 years old based on water, sanitation and hygiene behaviors and infrastructure," PLoS One, vol. 21, no. 3, p. e0343796, 2026.
- [10] M. Carvalho, A. J. Pinho, and S. Brás, "Resampling approaches to handle class imbalance: a review from a data perspective," J. Big Data, vol. 12, no. 1, p. 71, 2025.
- [11] A. Wang, S. Kapoor, S. Barocas, and A. Narayanan, "Against predictive optimization: On the legitimacy of decision-making algorithms that optimize predictive accuracy," ACM J. Responsible Comput., vol. 1, no. 1, pp. 1–45, 2024.
- [12] M. S. M. AL-Inizi, "Enhancing governmental decision-making through predictive analytics with machine learning-based data-driven framework," Babylonian J. Mach. Learn., vol. 2025, pp. 86–96, 2025.
- [13] M. A. Souames, L. A. Mohammedi, I. Zouaghi, A. Gunasekaran, S. Beldjoudi, and A. Laghouag, "Estimating import lead times using business intelligence and machine learning within the CRISP-DM framework: A case study in oil and gas services industry," IEEE Access, vol. 13, pp. 119215–119234, 2025.
- [14] I. Ahmad et al., "An efficient feature selection and explainable classification method for EEG-based epileptic seizure detection," J. Inf. Secur. Appl., vol. 80, p. 103654, 2024.
- [15] D. Setiawan, M. Christina, K. Stefani, and R. Selamat, "Predicting Bank Customer Churn Based on Activity and Financial Factors with Random Forest Classifier Using the CRISP-DM Framework," vol. 05, 2025.
- [16] F. Dong, "Random Forest Algorithm for HR Data Classification and Performance Analysis in Cloud Environments," vol. 15, no. 11, pp. 501–508, 2024.
- [17] A. Zermane, M. Z. M. Tohir, H. Zermane, M. R. Baharudin, and H. M. Yusoff, "Predicting fatal fall from heights accidents using random forest classification machine learning model," Saf. Sci., vol. 159, p. 106023, 2023.
- [18] I. Aruleba and Y. Sun, "Effective Credit Risk Prediction Using Ensemble Classifiers With Model Explanation," no. June, 2024.
- [19] P. Pampoukts, S. Avdimiotis, M. Maragoudakis, and M. Avlonitis, "Journal of Information System Techniques of Applied Machine Learning Being Utilized for the Purpose of Selecting and Placing Human Resources within the Public Sector," vol. 1, no. 1, pp. 1–16, 2023.
- [20] Nofriani, "Machine Learning Application for Classification Prediction of Household ' s Welfare Status," vol. 02, pp. 72–82, 2020.
- [21] R. C. Das, "Machine Learning-Based Financial Distress Prediction : Evidence from Bangladesh ' s Banking Sector," pp. 1–28.
- [22] E. Qumsiyeh, "Comparative Analysis of Credit Card Approval Prediction Using Machine Learning in Python and PySpark," 2025.
- [23] C. Satria, P. W. Sugijanto, A. Anggrawan, I. N. Y. Sumadewa, and A. D. Dayani, "Multi-Algorithm Approach to Enhancing Social Assistance Efficiency Through Accurate Poverty Classification," vol. 24, no. 1, pp. 167–178, 2024, doi: 10.30812/matrik.v24i1.4275.
- [24] D. Qin and N. Puttanapong, "Estimating county-level digital economy development in china using satellite data : insights from explainable spatial machine learning," vol. 6049, 2025, doi: 10.1080/10106049.2025.2596971.

