Journal of Information Technology and Computer Science
Volume 11, Number 2, August 2026, pp. 212-223
Journal Homepage: www.jitecs.ub.ac.id

Machine Learning Comparison for Housing Rehabilitation Beneficiary Selection in Surabaya

Cathrine Abigael Christy[1], Luluk Harida Budianti[2], Daniel Hary Prasetyo[3], Joko Siswantoro[4]

[1,2,3,4]Universitas Surabaya, Surabaya
[1]cathrine@staff.ubaya.ac.id, [2]s164223519@student.ubaya.ac.id,
[3]daniel@staff.ubaya.ac.id, [4]joko_siswantoro@staff.ubaya.ac.id
*Corresponding Author

Received 28 April 2026; accepted 30 July 2026

Abstract. The Surabaya City Government launched the 'Dandan Omah' Social Rehabilitation Program for Uninhabitable Houses (RTLH) in 2022. However, beneficiary selection during 2022–2024 period was conducted manually, leading to potentially subjective and less targeted decisions. This applied research established a data-driven support approach by executing structured procedure: constructing a preprocessing pipeline for a dataset of 1700 records with 11 features, implementing four supervised machine learning models (DT, RF, SVM, KNN), performing 10-fold cross-validation, and also hyperparameter tuning with 10-fold cross-validation across 2, 4, and 5 classification tasks. RF consistently achieved the best performance across all scenarios, while SVM demonstrated competitive baseline performance. After tuning, DT surpassed SVM on 5-class tasks and matched RF on 2-class tasks, making optimized DT a viable, interpretable alternative when decision accountability was essential. KNN consistently underperformed due to the curse of dimensionality. Overall performance declined as classification complexity increased, attributed to variability in field assessment standards. These findings suggested that RF and optimized DT could support more standardized and objective RTLH beneficiary selection. Future studies is recommended to identify influential input features to optimize model performance while maintaining explainability. Re-standardizing field assessment protocols is also recommended to improve data quality.

Keywords: decision support system; machine learning classification; random forest; uninhabitable houses (RTLH); beneficiary selection

1 Introduction

The Social Rehabilitation Program for Uninhabitable Houses (RTLH) is a poverty alleviation program whose implementation falls under the government's housing and settlement affairs and social affairs. In March 2022, the Surabaya City Government launched "Dandan Omah" as part of the RTLH Social Rehabilitation Program in Surabaya, as regulated in Surabaya Mayor Regulation Number 9 of 2022 concerning the RTLH Social Rehabilitation Program in Surabaya [1]. In 2025, the Surabaya City Housing, Settlement, and Land Agency targeted 2,069 housing units. However, at the beginning of the year the program had received 7,789 rehabilitation proposals, far exceeding the target and indicating a significant increase in public

demand [2].

To determine the recipients of the Uninhabitable House Rehabilitation (RTLH) program, each applicant's house is surveyed and assessed based on specific criterias. In accordance with Surabaya Mayor Regulation Number 7 of 2024, these factors include prior assistance history, as well as the conditions of the walls, ceiling, floor, and bathroom. Based on these criteria, the proposals are classified into five rehabilitation severity levels to determines their eligibility. Under factual conditions, however, both the criteria assessment and the final classification are handled manually on a case-by-case basis by the surveyors/officers. Because this evaluation relies entirely on manual judgment, it could introduces potential subjectivity, which may lead to poorly targeted aid distribution.

Given the massive volume of data, the complexity of the variables, the multiple severity classes, the current classification process faces significant challenges. Therefore, a systematic, data-driven approach is essential to ensure consistent, objective classification. Implementing such an approach would enable more objective recommendations and help reduce subjectivity in decision-making [3].

Table 1 highlights the positioning of this study relative to existing literature on RTLH beneficiary selection. In general, prior studies have been dominated by the Multi-Criteria Decision Making (MCDM) approach [4], [5], [6], [7]. Unlike MCDM methods which rely heavily on subjective weighting and ranking mechanisms and do not evaluate predictive performance across different classification scenarios, this study evaluates predictive performance across classification scenarios using Machine Learning (ML) methods that provide a standardized process to reduce subjectivity. Furthermore, this research expands the classification task from binary [8] to multiclass, providing deeper clarity and increasing the model's usability during field implementation. Finally, in contrast to previous studies that often utilized limited samples, this research leverages a larger-scale dataset, thereby enabling more robust model generalization.

Table 1. Comparison between the present study and previous research on RTLH assessment.

Study	Approach/Method	Dataset Size	Features	Output
[4]	WP (MCDM)	33	9	Ranking
[7]	SAW (MCDM)	4	3	Ranking
[5]	TOPSIS (MCDM)	30	8	Ranking
[6]	AHP-VIKOR (MCDM)	30	5	Ranking
[8]	KNN (ML)	1289	13	Binary class
Current study	**DT, RF, SVM, KNN (ML)**	**1700**	**11**	**Multiclass**

Considering the strengths and limitations of prior studies, this study employs DT, RF, SVM, and KNN. The selection of these ML methods is based on their widespread recognition as state-of-the-art approaches across many classification tasks. From a theoretical perspective, RF as an ensemble model is often considered stronger. In some previous studies, RF outperformed other models such as SVM, KNN, and DT [9], [10]. However, machine learning model performance is influenced by data characteristics [11]. For example, RF has a slightly different performance from DT [10], while SVM and KNN actually show the best performance in the other cases [12], [13]. This is in line with the "No Free Lunch" Theorem, which states that

no single machine learning algorithm consistently performs best across all problem domains or scenarios [14].

Specifically, DT and RF were chosen for their ability to handle both categorical and numerical data, identify important features, manage complex classification structures, handle missing values flexibly and achieve excellent performance on large datasets, thereby addressing limitations observed in prior study methods [15], [16]. SVM was selected because it is widely adopted in industrial and research applications and is reliable in maximizing the decision margin, resulting in models that are more stable and robust to noise [17]. Lastly, KNN was included to enable direct benchmark against prior research in this case study that employed the same method [8].

A comparative evaluation of the four ML methods was conducted under three classification scenarios: 2-class eligibility, 4-class rehabilitation severity, and 5-class rehabilitation severity. Model performance was assessed using accuracy and macro F1-score across 10-fold stratified cross-validation. Based on this experimental design, this study proposes the following research questions: "How do DT, RF, SVM, and KNN compare in classifying RTLH beneficiary proposals under the 2, 4, and 5 rehabilitation severity settings in terms of accuracy and macro F1?"

These research questions are proposed to support the study objective of identifying an ML-based approach that enables more standardized and objective RTLH beneficiary decisions. The main contribution of this study is a comparative evaluation of multiple machine learning algorithms for RTLH classification under varying levels of classification complexity.

2 Research Methods

This research is applied research that was conducted in three sequential steps: preprocessing pipeline construction, supervised machine learning models implementation, 10-fold cross-validation, and hyperparameter tuning. The dataset used in this study comprises the 2023 RTLH rehabilitation proposals obtained from the Department of Public Housing and Settlement Areas, Land, and Settlements of Surabaya City (DPRKPP).

The data collection process followed a structured administrative workflow [18]. Initially, DPRKPP issued official requests to the village administration head (lurah) to submit verified proposals. These local administrations then performed administrative cross-checking and verification on the applicant data before submitting the prospective beneficiary list back to DPRKPP, which was subsequently used to determine the beneficiary quota. Following this administrative phase, DPRKPP surveyors conducted direct field surveys to assess the proposals. This process yielded a dataset with 1700 rows data with 14 input variables (*id*, *name*, *building_area*, *num_occupants*, *household_income*, *area_per_capita*, *prior_assistance*, *poverty_status*, *wall_condition*, *floor_condition*, *roof_condition*, *ceiling_condition*, *bathroom_condition*,) and a target label (*rehab_severity*).

This workflow suggests a direct dependency between the input variables and the target labels, as both were derived from the same manual field assessment process. Consequently, the machine learning models may be inherently learning the surveyor's subjective decision-making logic rather than purely objective patterns. This in turn could lead to an artificially inflated classification performance. While this may appear as a dependency in a traditional predictive context, the objective of this study is not to develop a predictive model but rather to evaluate the consistency of the existing decision-making process. By using machine learning to approximate this decision-

making process, this study aim to provide a standardized and objective RTLH beneficiary selection.

In the *prior_assistance* feature, there are 4 categories: 'received no intervention', 'self-renovation', 'received BSPS assistance', 'received CSR assistance', and 'prior work by RTLH'. The feature was simplified into two groups: 'never renovated' and 'already renovated,' producing a derived feature named *prior_assistance_enc*. Before further processing, the *id* and *name* columns were dropped because they do not contribute to the classification, leaving 11 input features and 1 target column.

Of the 11 input features, 3 are numerical: *building_area* (total floor area, range −1 to 84,344 m²), *num_occupants* (household size, range 1–21), and *household_income* (monthly income of head of household, range 0–6,000,000 IDR). The remaining 8 features are categorical: *area_per_capita* (building area per occupant, reflecting occupancy density), *prior_assistance_enc* (whether the dwelling has received prior renovation support), *poverty_status* (official poverty classification), and five physical-condition assessments: *wall_condition*, *floor_condition*, *roof_condition*, *ceiling_condition*, and *bathroom_condition*.

The *rehab_severity* (later renamed to *rehab_severity_5*) target variable represents 5 classes of rehabilitation severity determined through field surveys and assessment: 'Heavy Rehab', 'Moderate Rehab' (1 million/m²), 'Light Rehab' (500 thousand/m²), 'Partial Rehab' (300 thousand/m²), and 'No repair needed' (0 thousand/m²). The target labels are based on 11 housing-condition indicators used as input features. To evaluate model robustness and practical applicability, two additional target formulations were derived: *rehab_severity_4* (four-class configuration) and *rehab_eligibility* (binary configuration. The reformulation into four classes (rehab_severity_4) was conducted by merging 'Partial Rehab' and 'Light Rehab' because these two classes share a narrow cost treshold (300 thousand/m² and 500 thousand/m², respectively), which potentially led to feature overlap and labeling ambiguity among surveyors. Merging them helps reduce classification noise. Meanwhile, the binary formulation enables comparison with prior studies adopting a yes/no categorization. Table 2 reports the class distribution across all three target formulations.

Table 2. Class distribution across target formulation

Target	Class	N	Target	Class	N
rehab_eligibility	Yes	1588	rehab_severity_5	Heavy Rehab	283
	No	112		Moderate Rehab	517
rehab_severity_4	Heavy Rehab	283		Light Rehab	606
	Moderate Rehab	517		Partial Rehab	182
	Light Rehab	788		No repair needed	112
	No repair needed	112			

2.1 Preprocessing pipeline

Preprocessing was implemented as a single pipeline in which all transformers (imputers, domain-specific operators, binary encoders, one-hot encoders, and scalers) were fit exclusively on the training partition and subsequently applied to both the training and test sets. A fixed random_state=42 was applied uniformly across the train–test split, all classifiers, the oversampler, and all cross-validation procedures to ensure full reproducibility.

Train-Test Split

The goal is to avoid data leakage, since the next steps are based solely on

statistical training. In this study, the dataset was split into training (80%) and test (20%) sets using stratified sampling to preserve class proportions to ensure robustness and generalization and to prevent overfitting [19].

Handling Missing Values

Missing values were identified in *building_area* (55 entries), *household_income* (206 entries), and *num_occupants* (55 entries). Median imputation was applied to all three numerical features; mode imputation was applied to categorical features with missing entries.

Domain-Based Numeric Preprocessing

For the feature *building_area*, CAP winsorization was applied to limit extreme values to the range between 10 m^2 and 100 m^2. Values beyond these limits may indicate an input error, as evidenced by the detection of several extreme values in the descriptive statistics. For the *household_income* feature, a log transform was applied to address the highly right-skewed distribution, even though the dataset showed no extreme invalid values. For feature *num_occupants*, the threshold for outliers (IQR) was calculated using the upper fence rule by [20]. Values below zero were removed, while values above the upper threshold were flagged only as extreme cases, since large numbers of residents from a large family house can be valid.

Categorical Encoding

Categorical features were encoded into numerical values using one hot encoding to enable machine learning processing. One hot encoding method was selected to ensure that each category is treated independently without imposing an arbitrary numerical structure [21]. This approach is also suitable for models sensitive to margin and distance, such as SVM and KNN.

Feature Scaling

Scaling was applied to models that are sensitive to feature magnitude. In margin-based models such as SVM, features with larger values may dominate margin optimization, leading to biased decision boundaries. Similarly, KNN relies on Euclidean distance, where unscaled features with larger ranges can dominate distance calculations. In contrast, tree-based models such as Decision Tree (DT) and Random Forest (RF) are rule-based and therefore insensitive to feature scale. In this study, Standard Scaler was used to standardize features to a mean of 0 and a standard deviation of 1, ensuring consistent transformation between training and testing data [22].

Handling Class Imbalance

To address potential class imbalance, RandomOversampler was applied to prevent synthetic minority-class samples from inflating cross-validation performance estimates, which applied resampling strictly to the training fold of each cross-validation split and never to the validation or test partitions

2.2 Classification Models

This study employs four ML methods, which are DT, RF, SVM, and KNN. DT is an ML algorithm that produces predictive models [23]. In many cases, this model can handle uneven distributions and provide consistent performance across different test data [15]. It works by recursively partitioning the feature space using binary splits selected to maximize the reduction in node impurity [24]. In this study, Gini impurity

was used as the splitting criterion. Equation (1) defines the Gini impurity at node t, where K is the number of classes and p_k is the proportion of samples in node t that belong to class k .

$$Gini(t) = 1 - \sum_{k=1}^{K} P_k^2 \quad (1)$$

RF is an ensemble classifier due to its ability to handle data with high-dimensional features, as well as its stability in providing accurate predictions [25]. It aggregates predictions from multiple trees trained on different subsets of the data while randomly selecting a subset of features at each split. Instead of relying on a single tree, RF combines the outputs of multiple trees and predicts the final class by majority voting across trees [26]. Compared to DT, this technique reduces variance and improves predictive stability [27].

SVM is a supervised learning machine learning method commonly used for classification and regression analysis [28]. In this study, SVM was employed with feature standardization using Standard Scaler. Using the Radial Basis Function kernel, non-linear separation is achieved by implicitly mapping the input space into a higher-dimensional feature space. The kernel mentioned before is defined in Equation (2) where x and x' denote feature vectors and γ controls the kernel width.

$$K(x, x') = exp\left(-\gamma \left| |x - x'| \right|^2\right) \quad (2)$$

KNN is a classification method that classifies a new data point by computing the distance to its nearest neighbors in the training set [29]. The main principle of KNN is to classify data points based on their nearest neighbors in feature space, assuming that closer points are more likely to share the same class [30]. In the baseline evaluation phase (Phase 1), five separate KNN (k=1,3,5,7,9) were evaluated individually, adopting the configuration in prior classification of uninhabitable houses in Jepara [8]. In the tuning phase, KNN was treated as a single model and k is selected automatically by GridSearchCV.

Because the method is distance-driven, features were standardized using Standard Scaler to prevent variables with larger numeric ranges from dominating the distance computation. KNN does not assume any parametric data distribution; however, its performance may degrade in very high-dimensional spaces, hence, preprocessing and feature scaling are important. Distances between a query point Q and a training point P were computed using the Euclidean metric, as defined in Equation (3). In Equation (3), m denotes the number of features, and P_j and Q_j are the values of the j-th feature for points P and Q, respectively.

$$d(P, Q) = \sqrt{\sum_{j=1}^{m} \left(P_j - Q_j\right)^2} \quad (3)$$

2.3 Cross-Validation and Hyperparameter Tuning

In the first phase, all models were evaluated using cross_validate with StratifiedKFold (n_splits=10) applied within the training set, following prior study configuration [8]. No hyperparameter tuning was performed; classifiers used their default scikit-learn configurations. Metrics recorded include accuracy and macro F1-score (mean and standard deviation).

In the second phase, hyperparameter tuning was conducted to determine the optimal set of machine learning model parameters to improve predictive performance [31]. This study uses a grid search to evaluate all possible hyperparameter

combinations by forming a grid across each parameter value. The grid search was nested entirely within the training set, so the test set is never seen during this procedure. After grid search, the best-configured model was refit on the full training set and evaluated exactly once on the held-out test set, with the classification report and confusion matrix reported. Results are ranked by the CV macro F1-score and holdout balanced accuracy.

3 Results and Discussion

To answer the research question, Table 3 compares the accuracy and macro F1-score scores across models in each scenario with the format (mean ± std). In the baseline comparison, RF dominated almost all classification targets, both in the accuracy matrix and in macro F1-score. In comparison, SVM performance is close to RF. On the other hand, KNN's performance consistently ranks lowest, with a downward trend that is directly proportional to the increase in k-value. Meanwhile, DT sits in the middle of the standings, indicating moderate performance.

Table 3. Model evaluation results for various target categories

Target	Metrics	RF	SVM	DT	KNN (k=1)	KNN (k=3)	KNN (k=5)	KNN (k=7)	KNN (k=9)
rehab_ eligibility	Accuracy	**0.993 ± 0.016**	**0.993 ± 0.016**	0.988 ± 0.022	0.992 ± 0.016	0.992 ± 0.016	0.990 ± 0.016	0.987 ± 0.015	0.985 ± 0.015
	Macro F1	**0.988 ± 0.014**	**0.988 ± 0.014**	0.988 ± 0.015	0.980 ± 0.018	0.977 ± 0.017	0.965 ± 0.026	0.945 ± 0.033	0.938 ± 0.034
rehab _severity_4	Accuracy	0.861 ± 0.030	**0.863 ± 0.020**	0.811 ± 0.030	0.796 ± 0.048	0.807 ± 0.033	0.812 ± 0.027	0.817 ± 0.031	0.830 ± 0.030
	Macro F1	**0.852 ± 0.029**	0.847 ± 0.018	0.808 ± 0.031	0.784 ± 0.046	0.776 ± 0.040	0.772 ± 0.034	0.773 ± 0.040	0.782 ± 0.043
rehab _severity_5	Accuracy	**0.714 ± 0.037**	0.698 ± 0.035	0.650 ± 0.035	0.655 ± 0.026	0.661 ± 0.030	0.669 ± 0.039	0.680 ± 0.039	0.677 ± 0.037
	Macro F1	**0.711 ± 0.038**	0.685 ± 0.033	0.647 ± 0.029	0.647 ± 0.030	0.635 ± 0.026	0.629 ± 0.036	0.636 ± 0.036	0.633 ± 0.033

For multiclass targets, RF outperforms DT because ensemble aggregation mitigates single-tree variance, capturing complex building damage patterns far better than an individual decision tree. Similarly, in high-dimensional feature spaces expanded by one-hot encoding, SVM outperforms KNN; while the curse of dimensionality degrades KNN's distance-based similarity metrics, SVM effectively leverages maximum-margin hyperplanes to establish clear decision boundaries.

It was observed that performance decreased across all models as the number of target classes increased. This indicates that the model faces greater difficulty maintaining balanced performance across all classes when the classification problem becomes more complex. Because target determination is done manually through field

surveys, labeling results are vulnerable to human subjectivity, leading to class overlap across damage levels. Besides that, the boundaries between the repair categories are ambiguous, and the model has difficulty precisely separating the classes.

In addition to the baseline comparison, the inclusion of the KNN model in this study was also motivated by its use in previous research [8] on a similar problem domain. Although the datasets used differ in size and feature composition, the results in this study show that the KNN model achieves competitive performance, with an accuracy of 0.992 (k=1 and 3) in the binary classification scenario. This result is higher than the performance reported in the previous study, which is 0.979 (k=5). However, since the datasets and feature sets differ, the comparison should be interpreted cautiously and mainly indicates that KNN can still perform well under the dataset characteristics used in this study.

Following the baseline comparison, grid search was performed to optimize all models. Table 4 presents the result of tuned models across all scenarios, while Table 5 shows the best parameters of all tuned models.

Table 4. Tuned model evaluation results for various target categories

Target	Metrics	RF	SVM	DT	KNN
rehab_ eligibility	Accuracy	**0.997**	**0.997**	0.995	0.995
	CV Macro F1	**0.992**	0.988	**0.992**	0.981
	Holdout Macro F1	**0.977**	**0.977**	0.966	0.966
rehab_ severity_4	Accuracy	**0.886**	0.838	0.848	0.808
	CV Macro F1	**0.858**	0.847	0.833	0.800
	Holdout Macro F1	0.**872**	0.825	0.841	0.770
rehab_ severity_5	Accuracy	0.712	**0.723**	0.717	0.652
	CV Macro F1	**0.732**	0.689	0.717	0.654
	Holdout Macro F1	0.699	0.704	0.703	0.625

Table 5. Best parameters of all tuned models

Model	Best parameters		
	rehab_eligibility	**rehab_severity_4d**	**rehab_severity_5**
DT	{'criterion': 'entropy', 'max_depth': 5, 'min_samples_leaf': 1, 'min_samples_split': 2}	{'criterion': 'gini', 'max_depth': 5, 'min_samples_leaf': 1, 'min_samples_split': 10}	{'criterion': 'gini', 'max_depth': 5, 'min_samples_leaf': 1, 'min_samples_split': 2}
RF	{'max_depth': 10, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 200}	{'max_depth': 10, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 300}	{'max_depth': 10, 'max_features': 'log2', 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 100}
SVM	{'C': 1, 'gamma': 'scale', 'kernel': 'rbf'}	{'C': 1, 'gamma': 'scale', 'kernel': 'rbf'}	{'C': 1, 'gamma': 'scale', 'kernel': 'linear'}
KNN	{'metric': 'manhattan', 'n_neighbors': 1, 'weights': 'distance'}	{'metric': 'manhattan', 'n_neighbors': 9, 'weights': 'distance'}	{'metric': 'manhattan', 'n_neighbors': 9, 'weights': 'distance'}

Consistent with previous evaluation, RF maintained its performance across all classification scenarios. However, hyperparameter tuning enables DT to match RF performance on 2-class tasks and outperform SVM on 5-class tasks. Conversely, KNN remains in the lowest position, confirming its reliance on spatial proximity

rather than learning complex feature patterns.

To further validate the significance of the performance shifts from the baseline to the tuned configuration, a Wilcoxon signed-rank test was conducted. A p-value below 0.05 indicates a statistically significant improvement resulting from hyperparameter optimization. As shown in Table 6, grid search optimization yielded valid enhancements for DT and RF in the 4-class task, and for RF, DT, and KNN in the 5-class task. In contrast, the SVM model exhibited no statistically significant improvement from the grid search tuning.

Table 6. Wilcoxon Test Results for Baseline vs. Tuned Models

Target	p-value (baseline vs tuned configuration)			
	RF	**SVM**	**DT**	**KNN**
rehab_eligibility	1	1	0.5	0.0625
rehab_severity_4	0.23242	1	**0.01953**	**0.00977**
rehab_severity_5	**0.01953**	0.43164	**0.00195**	**0.01367**

Furthermore, the Wilcoxon signed-rank test was also conducted to evaluate the superiority among models. Given its consistent outperformance across both baseline and tuned setups, RF was selected as the reference model. To mitigate Type I errors, a Bonferroni correction was applied by dividing the significance level by the number of comparisons, establishing a stricter threshold of $\alpha = 0.0167$. A p-value below this adjusted threshold demonstrates that RF is statistically superior to the competing model.

Based on the results in Table 7, RF performs comparably to all other models in the 2-class task. However, in the 4-class scenario, RF proves statistically superior to KNN, and in the 5-class scenario, it significantly outperforms both SVM and KNN. Notably, across all three classification tasks, RF shares a statistically comparable performance with DT.

Table 7. Pairwise Model Superiority Test with Bonferroni Correction ($\alpha = 0.0167$)

Target	p-value (RF vs other model)		
	SVM	**DT**	**KNN**
rehab_eligibility	1	1	0.25
rehab_severity_4	0.03711	0.1953	**0.00195**
rehab_severity_5	**0.00195**	0.13086	**0.00195**

The lack of significant difference between RF and DR demonstrates that through grid search tuning, DT as a transparent model can achieve performance nearly equivalent to RF, a much more complex opaque model, without sacrificing interpretability. This finding challenges the trade-off highlighted by [32], who found that opaque models consistently outperformed transparent ones in terms of accuracy, while transparent models were only advantaged in response time and interpretability. This is particularly valuable for real-world implementation in public policy, such as social assistance distribution, where policymakers require clear justification and decision accountability.

Beyond accountability, the transparency of DT supports algorithmic fairness by allowing officers to inspect decision boundary rules, ensuring that eligibility thresholds do not introduce bias or unfair exclusion of eligible recipients. Meanwhile,

the opaque nature of RF can be mitigated through feature importance analysis to identify influential variables and maintain model explainability.

The feature importance analysis reveals that RF relies heavily on three main variables: *bathroom_condition*, *wall_condition*, and *ceiling_condition*, with importance scores of 0.116, 0.089, and 0.087, respectively. Further investigation of the target *rehab_severity_5* revealed that the low performance in the 'Partial Rehab' class was due to the stagnation of information on these dominant features. The high importance score failed to provide a clear separation between the 'Light Rehab' and 'Partial Rehab' because the physical attributes recorded in the dataset, specifically the condition of the bathrooms, walls, and ceilings, were statistically identical. This issue largely arises due to the ambiguity of assessment standards during manual field surveys, which introduces surveyors' subjectivity and creates labeling noise. As a result, the model lacks sufficient sensitivity to accurately distinguish between the two categories. Consequently, from a deployment perspective, integrating human validation as a final decision step is essential when deploying either DT or RF into established decision-making systems.

In addition, the confusion matrix in Figure 2 shows that 22 of the 37 'Partial Rehab' samples were incorrectly classified as 'Light Rehab'. This confirms a severe feature overlap. Logically, this is supported by the small price difference between 'Partial Rehab' (300 thousand/m²) and 'Light Rehab' (500 thousand/m²), which creates a very narrow threshold in real-world conditions compared to the wider gap in other classes. Therefore, merging these classes into a 4-class scenario is the right methodological decision to improve the model's predictive stability and reliability.

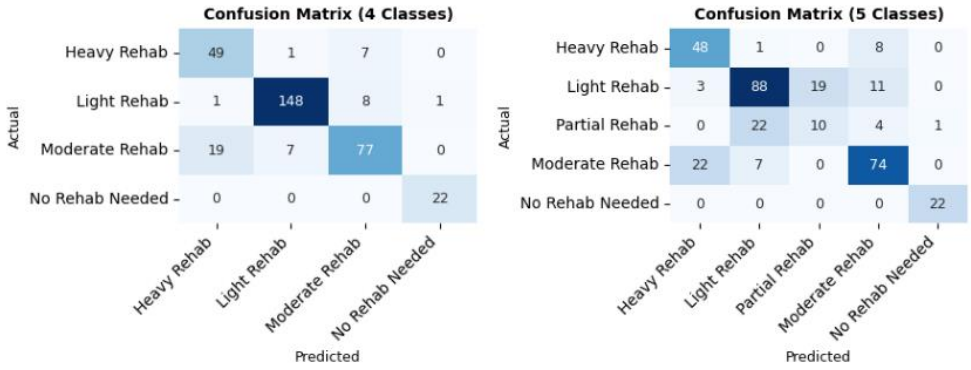

Figure 2. Confusion matrices for *rehab_severity_4* and *rehab_severity_5* (tuned RF)

4 Conclusion

The study demonstrates that machine learning models, particularly RF and optimized DT offer a viable and data-driven solution to standardize the RTLH beneficiary selection process. Across all scenarios (2, 4, and 5 rehabilitation classes), RF consistently achieved the best performance while optimized DT emerged as an alternative when interpretability and decision accountability is critical for public policy transparency. The results also indicate that classification difficulty increases with more granular rehabilitation classes, primarily due to feature overlap and labeling inconsistencies caused by surveyor subjectivity. Ultimately, while machine learning can significantly reduce human bias, re-standardization of input criteria and

target variables in the future is recommended as enhanching data quality will yield greater gains than algorithmic optimization alone.

References

1. M. Chabibi, "Labour-Intensive Program: A Transformation of Public Welfare Policy in Supporting Sustainable Development," *Entita: Jurnal Pendidikan Ilmu Pengetahuan Sosial dan Ilmu-Ilmu Sosial*, pp. 129–146, May 2025, doi: 10.19105/EJPIS.V1I.19140.
2. G. Arnanda and A. Widiyarta, "Implementation of the Dandan Omah Program in Realizing Livable Homes in Kedung Cowek Urban Village Surabaya City," *Jurnal Kebijakan Publik |*, vol. 16, no. 2, 2025, Accessed: Apr. 28, 2026. [Online]. Available: http://jkp.ejournal.unri.ac.id
3. R. Ali *et al.*, "Intelligent Decision Support Systems—An Analysis of Machine Learning and Multicriteria Decision-Making Methods," *Applied Sciences 2023, Vol. 13,* vol. 13, no. 22, Nov. 2023, doi: 10.3390/APP132212426.
4. P. Ponidi, R. Renaldo, S. Mukodimah, and S. Abadi, "Application of the Weighted Product Method to Determine House Renovation Assistance in Pringsewu Regency," *Tech-E*, vol. 5, no. 1, pp. 42–56, Sep. 2021, doi: 10.31253/TE.V5I1.664.
5. C. Acuña-Soto, V. Liern, and B. Pérez-Gladish, "Normalization in TOPSIS-based approaches with data of different nature: application to the ranking of mathematical videos," *Ann. Oper. Res.*, vol. 296, no. 1, pp. 541–569, Jan. 2021, doi: 10.1007/S10479-018-2945-5.
6. S. Armonitha Lusinia and N. Rahmansyah, "Implementation of the VIKOR Method and Analytical Hierarchy Process (AHP) in Prioritizing the Recipients of Uninhabitable House Assistance Funds (RTLH)"," *The Indonesian Journal of Computer Science*, vol. 13, no. 1, pp. 2024–132, Jan. 2024, doi: 10.33022/IJCS.V13I1.3665.
7. B. S. Manopo, A. A. Kenap, and K. Santa, "Citizen-Centric Decision Support Systems for Housing Assistance: A Low-Barrier E-Government Approach," *Edumatic: Jurnal Pendidikan Informatika*, vol. 10, no. 1, pp. 120–129, Apr. 2026, doi: 10.29408/edumatic.v10i1.33945.
8. A. Naas, S. Na'iema, H. Mulyo, and A. Widiastuti, "Klasifikasi penerima bantuan program rehabilitasi rumah tidak layak huni menggunakan algoritme K-Nearest Neighbor," *Jurnal Teknologi dan Sistem Komputer*, vol. 10, no. 1, pp. 32–37, Jan. 2022, doi: 10.14710/JTSISKOM.2021.14110.
9. S. Eva and M. Purba, "A Comparative Study of Drug Prediction Models using KNN, SVM, and Random Forest," *Journal of Information Systems and Informatics*, vol. 7, no. 1, pp. 378–392, Mar. 2025, doi: 10.51519/JOURNALISI.V7I1.1013.
10. M. Y. Iskandar and H. W. Nugroho, "Comparative Evaluation of Decision Tree and Random Forest for Lung Cancer Prediction Based on Computational Efficiency and Predictive Accuracy," *Jurnal Teknik Informatika (Jutif)*, vol. 6, no. 5, pp. 3392–3404, Oct. 2025, doi: 10.52436/1.JUTIF.2025.6.5.4877.
11. S. Uddin and H. Lu, "Dataset meta-level and statistical features affect machine learning performance," *Sci. Rep.*, vol. 14, no. 1, p. 1670, Dec. 2024, doi: 10.1038/S41598-024-51825-X.
12. L. Yang *et al.*, "Comparative Analysis of the Optimized KNN, SVM, and Ensemble DT Models Using Bayesian Optimization for Predicting Pedestrian Fatalities: An Advance towards Realizing the Sustainable Safety of Pedestrians," *Sustainability 2022, Vol. 14, Page 10467*, vol. 14, no. 17, p. 10467, Aug. 2022, doi: 10.3390/SU141710467.
13. Y. Rimal, N. Sharma, S. Paudel, A. Alsadoon, M. Parsad Koirala, and S. Gill, "Comparative analysis of heart disease prediction using logistic regression, SVM, KNN, and random forest with cross-validation for improved accuracy", doi: 10.1038/s41598-025-93675-1.
14. S. P. Adam, S. A. N. Alexandropoulos, P. M. Pardalos, and M. N. Vrahatis, "No Free Lunch Theorem: A Review," *Springer Optimization and Its Applications*, vol. 145, pp. 57–82, 2019, doi: 10.1007/978-3-030-12767-1_5.

15. B. T. Jijo and A. M. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, Mar. 2021, doi: 10.38094/jastt20165.
16. D. Kurniawan, *Pengenalan Machine Learning dengan Python*. Jakarta: PT Elex Media Komputindo, 2020. Accessed: Dec. 22, 2025. [Online]. Available: https://books.google.co.id/books?id=ZutsEAAAQBAJ&printsec=frontcover&hl=id#v=onepage&q&f=false
17. R. Guido *et al.*, "An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review," *Information 2024, Vol. 15,* vol. 15, no. 4, Apr. 2024, doi: 10.3390/INFO15040235.
18. *Peraturan Walikota Surabaya Nomor 7 Tahun 2024 tentang Perubahan Kedua atas Peraturan Walikota Surabaya Nomor 9 Tahun 2022 tentang Rehabilitasi Sosial Rumah Tidak Layak Huni Kota Surabaya*. Indonesia: JDIH Kota Surabaya, 2024.
19. J. V. Ferro, C. E. de Farias Silva, and B. M. V. da Gama, "Machine learning-based prediction of carbohydrate productivity in continuous cultivation of Chlorella vulgaris," *Bioresour. Technol.*, vol. 445, p. 134064, Apr. 2026, doi: 10.1016/j.biortech.2026.134064.
20. J. Tukey, *Exploratory data analysis*, vol. 2. Reading, MA: Addison-Wesley, 1977. Accessed: Feb. 24, 2026. [Online]. Available: https://scholar.google.com/scholar_lookup?&title=Exploratory%20data%20analysis&publication_year=1977&author=Tukey%2CJW
21. J. Lee, "HeartSense: Leveraging Machine Learning to Predict Cardiovascular Risk," *Journal of Advanced Artificial Intelligence, Engineering and Technology*, vol. 1, Jun. 2025, doi: 10.56147/1.4.44.
22. M. F. Naufal, A. F. Susanto, C. N. Kansil, and S. Huda, "Analisis Perbandingan Algoritma Machine Learning untuk Prediksi Potensi Hilangnya Nasabah Bank," *Techno.Com*, vol. 22, no. 1, pp. 1–11, Feb. 2023, doi: 10.33633/TC.V22I1.7302.
23. P. H. Putra, M. S. Novelan, and M. Rizki, "Analysis K-Nearest Neighbor Method in Classification of Vegetable Quality Based on Color," *Journal of Applied Engineering and Technological Science (JAETS)*, vol. 3, no. 2, pp. 126–132, Jun. 2022, doi: 10.37385/jaets.v3i2.763.
24. G. Zhang and A. Gionis, "Regularized impurity reduction: Accurate decision trees with complexity guarantees," 2022.
25. I. D. Mienye and Y. Sun, "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects," *IEEE Access*, vol. 10, pp. 99129–99149, Jan. 2022, doi: 10.1109/ACCESS.2022.3207287.
26. Z. Sun, G. Wang, P. Li, H. Wang, M. Zhang, and X. Liang, "An improved random forest based on the classification accuracy and correlation measurement of decision trees," *Expert Syst. Appl.*, vol. 237, no. 18, p. 121549, Mar. 2024, doi: 10.1016/j.eswa.2023.121549.
27. G. Louppe, "Understanding Random Forests: From Theory to Practice," Jun. 2014, Accessed: Feb. 26, 2026. [Online]. Available: http://arxiv.org/abs/1407.7502
28. P. Sarang, "Support Vector Machines: A Supervised Learning Algorithm for Classification and Regression," pp. 153–165, 2023, doi: 10.1007/978-3-031-02363-7_8.
29. Z. Xu *et al.*, "FinTruthQA: A Benchmark Dataset for Evaluating the Quality of Financial Information Disclosure," Feb. 2025, Accessed: Feb. 27, 2026. [Online]. Available: http://arxiv.org/abs/2406.12009
30. N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *American Statistician*, vol. 46, no. 3, pp. 175–185, 1992, doi: 10.1080/00031305.1992.10475879.
31. B. Bischl *et al.*, "Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 13, no. 2, p. e1484, Mar. 2023, doi: 10.1002/widm.1484.
32. A. Assis, J. Dantas, and E. Andrade, "The performance-interpretability trade-off: a comparative study of machine learning models," *Journal of Reliable Intelligent Environments 2024 11:1*, vol. 11, no. 1, pp. 1-, Dec. 2024, doi: 10.1007/S40860-024-00240-0.