Prediction of Public Satisfaction Levels with Public Services Using Random Forest and Decision Tree

Jheki Pranta Singarimbun, Muhammad Iqbal

Abstract
Evaluation of public satisfaction in government agencies generally stops at calculating an index, so there is no initial picture of groups of service users who tend to be less satisfied. This study builds a classification model of service user satisfaction levels at the Medan Center for Health Quarantine using Random Forest and Decision Tree, then compares the performance of the two. The data came from 1,045 respondents to the 2025 Service Quality Perception Survey and Anti-Corruption Perception Survey with 13 assessment items on a scale of 1 to 6. Labels were formed from respondent indices into four categories, and the data were divided into 80% for training and 20% for testing. On the test data, Random Forest achieved an accuracy of 93.30% and a weighted F1-score of 92.70%, while Decision Tree achieved 81.82% and 81.79%. Five-fold cross-validation with 3 repetitions showed the same pattern (F1-score 90.36% versus 82.72%, Wilcoxon test p = 0.0001). However, in the Poor class, which contained only 6 test respondents, Random Forest had a recall of 0.33 and Decision Tree 0.83. The infrastructure and online service system items were the most influential items in the three feature importance analyses. The labels were calculated from the same items used as features, so the results should be interpreted as the models' ability to learn the rules for grouping the index.

Keywords: Public Services, Public Satisfaction, Data Mining, Random Forest, Decision Tree, Classification.

Jheki Pranta Singarimbun
Master of Information Technology, University Of Pembangunan Panca Budi, Indonesia
e-mail: Jheky.Singarimbun28@gmail.com

Muhammad Iqbal
e-mail: muhammadiqbal@dosen.pancabudi.ac.id
2nd International Conference on Islamic Community Studies (ICICS)
Theme: History of Malay Civilisation and Islamic Human Capacity and Halal Hub in the Globalization Era

Pendahuluan

Public service is a form of government responsibility in meeting the needs of the community through the provision of effective, efficient, transparent, and user satisfaction-oriented services. The level of public satisfaction is one of the important indicators in assessing the quality of services provided by government agencies. The higher the level of public satisfaction, the better the quality of services provided. Therefore, evaluations of public satisfaction need to be conducted periodically as a basis for formulating policies to improve the quality of public services. The implementation of public satisfaction surveys has become one of the methods widely used to obtain an overview of service quality. The data obtained from survey results generally consists of various assessment indicators, such as service requirements, service procedures, completion time, service costs, staff competence, staff behavior, complaint handling, as well as the quality of facilities and infrastructure. The continuously increasing amount of data causes the manual analysis process to become less effective because it requires more time and has the potential to produce less consistent interpretations. This condition encourages the utilization of data analysis technology capable of processing large amounts of data quickly and objectively.The development of machine learning provides opportunities to utilize classification techniques to predict the level of public satisfaction based on survey data characteristics. The resulting predictions not only help classify levels of public satisfaction, but also provide initial information regarding the quality of services that need to be maintained or improved. Thus, government agencies can make more appropriate decisions based on the available data analysis results.

Decision Tree is one of the classification algorithms widely used because it has a simple and easy-to-understand model structure. This algorithm is capable of producing decision rules that can serve as a basis for the service evaluation process. Nevertheless, Decision Tree has the weakness of tending to produce models that fit the training data too closely (overfitting), so its performance may decrease when used on new data. As an alternative, Random Forest was developed by combining many decision trees built randomly, thereby improving model stability, reducing overfitting, and producing better accuracy.

Several studies from the Universitas Pembangunan Panca Budi environment have shown that tree-based algorithms have been used on various types of data. Iqbal and Efendi (2023) applied C4.5 to construct a credit risk decision tree using data from 30 banks. Sitompul and Iqbal (2025) used Random Forest and Naïve Bayes to classify telecommunications network customer experiences into satisfied and dissatisfied classes. Ernawati et al. (2025) compared Random Forest and K-Nearest Neighbors on 21,137 population data records and found a very small difference in accuracy, namely 0.6023 versus 0.5990. This latest finding serves as a reminder that the superiority of an algorithm cannot simply be assumed, but rather needs to be tested on the data actually being used. Various studies have applied Random Forest and Decision Tree to classification problems in the fields of healthcare, education, finance, and public services. However, most studies have focused only on the use of a single algorithm or have not provided an in-depth comparative analysis of the two methods in the case of predicting public satisfaction levels with public services. In addition, the characteristics of public satisfaction data, which consist of various assessment

indicators, allow each algorithm to produce different performance. Therefore, research is needed to compare the two algorithms in order to obtain the most suitable method for prediction. Based on these problems, this study aims to build a prediction model for public satisfaction levels with public services using the Random Forest and Decision Tree algorithms, and then compare the performance of the two algorithms based on accuracy, precision, recall, and F1-score values. The results of the study are expected to provide recommendations for the algorithm with the best performance in classifying public satisfaction levels. In addition to contributing to the development of machine learning applications in the field of public services, this study is also expected to serve as a reference for government agencies in developing faster, more objective, accurate, and data-driven service evaluation systems so that the process of improving service quality can be carried out continuously.

LITERATURE REVIEW

2.1 Public Satisfaction with Public Services

Public satisfaction refers to the level of public assessment of the quality of services provided by government agencies. Such satisfaction is influenced by various factors, such as service speed, timeliness, staff friendliness, ease of procedures, information transparency, and the quality of facilities and infrastructure. Measuring public satisfaction is an important indicator in evaluating public service performance and serves as a basis for developing strategies to improve service quality.

The data in this study were obtained from the Service Quality Perception Survey (SPKP) and the Anti-Corruption Perception Survey (SPAK), which were conducted as part of the development of the Integrity Zone. The questionnaires for both surveys refer to Regulation of the Minister of Administrative and Bureaucratic Reform Number 90 of 2021, and respondents rated each item on a scale of 1 to 6 (Balai Besar POM di Bogor, 2024). This scale differs from the Community Satisfaction Survey under Regulation of the Minister of Administrative and Bureaucratic Reform Number 14 of 2017, which uses four answer choices. This difference in scale needs to be taken into account when respondent scores are converted into satisfaction indices and categories.

2.2 Data Mining

Data mining is the process of extracting hidden information or patterns from a collection of data using statistical techniques, machine learning, and artificial intelligence. In this study, data mining is used to classify the level of public satisfaction based on public service survey data so that it can assist government agencies in making faster and data-driven decisions.

2.3 Decision Tree Algorithm

Decision Tree is a classification algorithm that forms a decision tree structure based on the best attribute as a data splitter. Each branch represents the decision result of an attribute, while each leaf shows the final class. This algorithm has the advantage of being easy to understand and producing simple decision rules.

One of the measures commonly used in constructing a Decision Tree is Gini Impurity, with the formula:

:$\text{Gini}(S) = 1 - \sum_{i=1}^{c} p_i^2$

Notes:

1. S = data set
2. pi = proportion of data in class i
3. c = number of classes

The smaller the Gini value, the better the attribute is for splitting the data. This measure is also the default splitting criterion in the scikit-learn library (Pedregosa et al., 2011).

2.4 Random Forest Algorithm

Random Forest is an ensemble learning algorithm that combines multiple Decision Trees to improve classification accuracy. Each tree is built using different data samples (bootstrap sampling) and random attribute selection. The final prediction is obtained based on the majority vote (majority voting) of all the trees formed. This approach is able to reduce the risk of overfitting and produce a more stable model compared to a single Decision Tree. The Random Forest concept was introduced by Breiman (2001).

The simple formula for Random Forest prediction is:

$$\hat{Y} = \text{Mode}(T_1, T_2, \ldots, T_n)$$

Note:

1. $\hat{Y}$ = final prediction result
2. T_1, T_2, …, T_n = prediction results from each Decision Tree
3. Mode = class that receives the most votes (majority voting)

2.5 Model Evaluation

The performance of the Decision Tree and Random Forest algorithms is evaluated using a confusion matrix. Several metrics used include Accuracy, Precision, Recall, and F1-Score. The values of these metrics are used to compare the ability of the two algorithms to predict the level of public satisfaction with public services. Through this comparison, the algorithm with the best performance can be identified and used as a basis for supporting decision-making in the evaluation of public services more effectively and objectively.

The confusion matrix contains the number of correct and incorrect predictions for each class. From the table, the following four metrics are calculated for one class (Sokolova & Lapalme, 2009).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

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

TP, TN, FP, and FN respectively are the numbers of true positives, true negatives, false positives, and false negatives. In multiclass classification, the metric values are calculated for each class and then averaged. The weighted average assigns weights according to the number of data points in each class, while the macro average assigns equal weight to all classes.

$$F1_{\text{weighted}} = \sum_{k=1}^{K} \frac{n_k}{N} F1_k$$

$$F1_{\text{macro}} = \frac{1}{K} \sum_{k=1}^{K} F1_k$$

In the formula above, K is the number of classes, n with index k is the number of data points in the k-th class, and N is the total number of data points. Because the number of respondents in each satisfaction category may be imbalanced, macro F1 is also reported so that classes with fewer data points are not overshadowed by classes with larger numbers of data points.

2.6 Previous Related Studies

a. Many previous studies relevant to this research include:

a. Nelviony Parhusip, Muhammad Iqbal, and Sri Wahyuni (2024) conducted a study entitled Classification Analysis of the Assessment of the Community Satisfaction Index for Public Service Mall Services Based on Education Using Decision Tree. This study applied the Decision Tree algorithm to classify the Community Satisfaction Index (IKM) at Public Service Malls based on education level. The results showed that Decision Tree was able to produce an easily understandable classification model and was effective in supporting public service evaluation. The difference from this study is that the previous study only used Decision Tree, whereas this study compares Decision Tree and Random Forest.

b. Septhya, D., Rahayu, K., Rabbani, S., Fitria, V., Rahmaddeni, R., Irawan, Y., & Hayami, R. (2023). Implementation of Decision Tree and Support Vector Machine Algorithms for Lung Cancer Disease Classification. MALCOM: Indonesian Journal of Machine Learning and Computer Science, 3(1), 15–19.

c. Mardiani, E., Rahmansyah, N., Ningsih, S., Lantana, D. A., Wirawan, A. S. P., Wijaya, S. A., & Putri, D. N. (2023). Comparison of KNN, Naive Bayes, Decision Tree, Ensemble, and Linear Regression Methods for Analyzing High School Student Performance. INNOVATIVE: Journal Of Social Science Research, 3(2), 13880–13892.

d. Purbolaksono, M. D., Irvan Tantowi, M., Imam Hidayat, A., & Adiwijaya, A. (2021). Comparison of Support Vector Machine and Modified Balanced Random Forest for Diabetes Patient Detection. Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 5(2), 393–399.

e. Muhammad Abdurohman, Ranu Husna, Irfan Ali, Gifthera Dwilestari, and Nining Rahaningsih (2022) conducted a study entitled Application of a Classification Model for the Level of Public Service Satisfaction in Karyamulya Urban Village Using the Decision Tree Algorithm. The study achieved an accuracy of 90.66% and showed that Decision Tree was effective in classifying public satisfaction levels toward public services. This study differs because it compares the performance of Decision Tree and Random Forest.

f. Muzakki Hafizh Setiono (2022) conducted a study entitled Comparison of Decision Tree, Random Forest, SVM, and K-NN Algorithms for Airline Passenger Satisfaction Classification. The results showed that Random Forest achieved the highest accuracy of 96%, outperforming Decision Tree, SVM, and K-NN. This study provides a basis for the conclusion that Random Forest performs very well on satisfaction data.

g. Safira Amalia, Irene Deborah, and Intan Nurma Yulita (2022), through their study Comparative Analysis of Classification Algorithm: Random Forest, SPAARC, and MLP for Airlines Customer Satisfaction, concluded that Random Forest achieved an accuracy of 95.827% with a more efficient training time compared with the other algorithms.

h. Andi Dwi Riyanto, Arif Mu'amar Wahid, and Aniec Anafisah Pratiwi (2024) conducted a study entitled Analysis of Factors Determining Student Satisfaction Using Decision Tree, Random Forest, SVM, and Neural Networks. The study compared four classification algorithms for predicting student satisfaction and showed that tree-based algorithms had competitive performance in satisfaction analysis.

i. Nora Fitaria and Dedi Darwis (2025) conducted a study entitled Performance Analysis of Decision Tree and Naïve Bayes in Community Satisfaction Level Classification. The study showed that Decision Tree provided good performance in classifying Community Satisfaction Survey (SKM) data and was suitable for use as a public service prediction model.

j. Sigit Umar Anggono, Danny Manongga, and Ade Iriani (2021) conducted a study entitled Evaluation of One-Stop Integrated Services (PTSP) Using Decision Tree. The study utilized the ID3 algorithm to evaluate the quality of government services and generated decision rules that could be used as recommendations for service improvement.

k. Tyastuti Sri Lestari, Ismaniah, and Wowon Priatna (2024), through their study Particle Swarm Optimization for Optimizing Public Service Satisfaction Level Classification, compared several classification algorithms on public service satisfaction data. The results showed that Decision Tree achieved high accuracy, and optimization using Particle Swarm Optimization was able to improve the performance of several classification algorithms.

l. Agil Irman Fadri, Abid Zahfran, Taylan Irak, Naufal Helga Firjatullah, and Jelita Ekaraya Herianto (2024) conducted a study entitled Comparison of Supervised Learning Algorithms for Predicting Airline Passenger Satisfaction. The study compared Decision Tree, Random Forest, K-NN, Naïve Bayes, and SVM. Random Forest achieved the highest accuracy of 95.78%, while Decision Tree achieved 93.82%.

m. Hilmi Ammar and Ade Davy Wiranata (2025) conducted a study entitled Comparison of the Accuracy Levels of Naïve Bayes, Decision Tree, and Random Forest Algorithms in Sentiment Analysis of National Digital Samsat Application Users. The study showed that Random Forest achieved the best performance with an accuracy of 92.9%, outperforming Decision Tree and Naïve Bayes. These results further support the strong generalization capability of Random Forest in classification tasks.

n. Muhammad Iqbal and Syahril Efendi (2023) conducted a study entitled Data-Driven Approach for Credit Risk Analysis Using C4.5 Algorithm. This study used a sample of 30 banks with six factors, namely credit growth, net interest margin, bank type, capital ratio, company size, and bank compliance level. Credit risk was evaluated by

constructing a decision tree using the RapidMiner application. The study showed that decision trees can be applied to relatively small datasets. The difference is that this study works with public service survey data and compares Decision Tree with Random Forest.

o. Jelly Rolleys Sitompul and Muhammad Iqbal (2025), through their study An Intelligent Predictive Model for Customer Experience Evaluation in Telecom Networks, classified customer experience into satisfied and dissatisfied categories using Random Forest and Naïve Bayes. The data were tested using a 70:30 split and k-fold cross-validation. Random Forest was reported to perform better and also provided feature importance analysis. The topic is closely related to this study, but its target consisted of only two classes and its comparison algorithm was Naïve Bayes, whereas this study uses four satisfaction categories with Decision Tree as the comparison algorithm.

p. Robin Antoni and Muhammad Iqbal (2025) conducted a study entitled Classification of Study Program Competitiveness Levels in SNBT Using the Random Forest Method Based on Capacity and Regional Distribution Data. The Low, Medium, and High classes were automatically formed from quota and applicant data using a quantile approach. The Random Forest model achieved accuracy, precision, recall, and F1-score of 68.00%. Mean Decrease Impurity analysis showed that seat capacity was the main factor, with a contribution of 53.60%. The method of forming labels from the data itself is similar to that used in this study, so its limitations also need to be openly discussed.

q. Andi Ernawati, Khairul, Zulham Sitorus, Muhammad Iqbal, and Darmeli Nasution (2025) compared Random Forest and K-Nearest Neighbors for classifying poverty status using 21,137 population records from Labuhanbatu Regency. The research stages included preprocessing, training, hyperparameter optimization, and evaluation using five metrics. Random Forest achieved an accuracy of 0.6023 and KNN achieved 0.5990, resulting in only a small difference. This result provides a reason why this study does not rely only on a single data split, but also performs cross-validation and statistical testing.

r. Muhammad Syahputra Novelan, Syahril Efendi, Poltak Sihombing, and Herman Mawengkang (2023), through their study Vehicle Routing Problem Optimization with Machine Learning in Imbalanced Classification Vehicle Route Data, classified imbalanced vehicle route data using K-NN and tested K values of 11, 13, and 15. This study is relevant because it demonstrates that class imbalance needs to be considered from the beginning of the classification process.

s. Muhammad Syahputra Novelan and Solly Aryza (2025) conducted a study entitled Optimization CVRP with Machine Learning for Improved Classification of Imbalanced Data Food Distribution. This study used K-Means and K-NN on imbalanced demand data so that priority areas would not be overlooked. According to its abstract, K-NN with K = 11 achieved an accuracy of 95, precision of 97, and recall of 94. In public service satisfaction data, respondents who are less satisfied function as the minority class, similar to the priority areas in that study.

T. Based on the sources reviewed by the authors, Random Forest and Decision Tree have been widely used, but comparisons between the two on SPKP and SPAK data, in which labels are formed from respondent indices, have not been widely discussed. Some studies also report only a single data split without repeated testing. This study

addresses this research gap by incorporating cross-validation, the Wilcoxon test, hyperparameter tuning, and permutation importance.

2.7 Relationship Between Variables

In this study, the independent variable (X) is the public service quality indicator obtained from the SPKP and SPAK results, while the dependent variable (Y) is the level of public satisfaction.

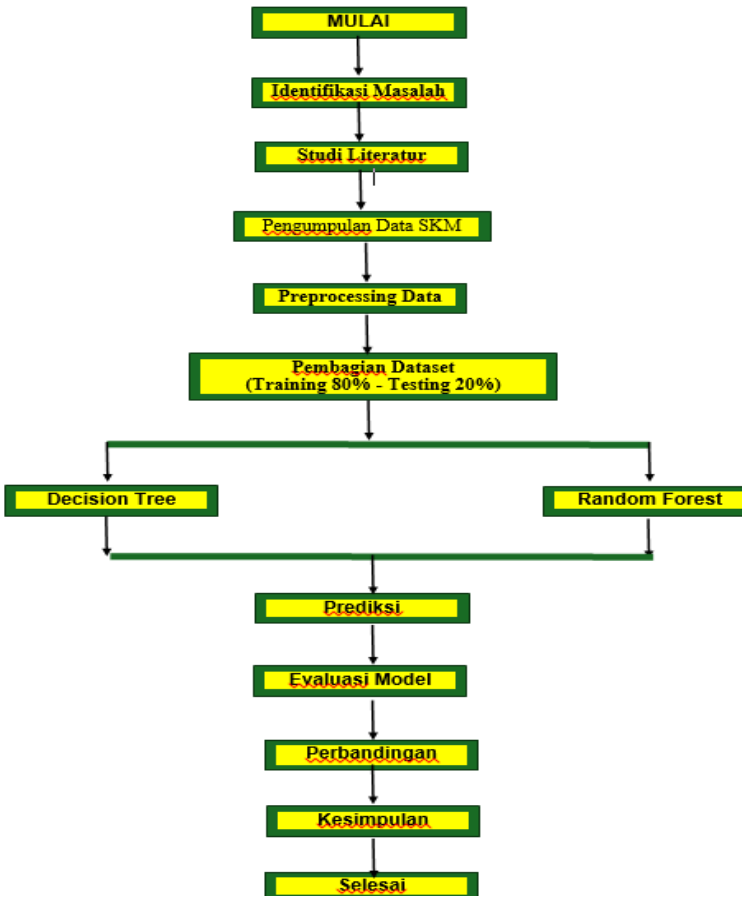
2.8 Variable X (Independent) : The independent variable (X) is a factor that influences the dependent variable. In this study, variable X consists of 13 assessment items, namely eight service quality items (information, requirements, procedures, completion time, fees, facilities and infrastructure, staff responsiveness, as well as consultation and complaint services) and five anti-corruption items (non-discriminatory, in accordance with procedures, no requests for compensation, no illegal charges, and no intermediaries). Respondent attributes such as gender, age, service work area, occupation, and highest level of education are also used as features.

2.9 Variable Y (Dependent) : The dependent variable (Y) is the variable influenced by the independent variable. In this study, Y consists of four respondent satisfaction categories, namely D (Poor), C (Fair), B (Good), and A (Very Good). These classes are calculated from each respondent's index and become the classification results predicted by Decision Tree and Random Forest.

The relationship between variables in this study shows that variable X (public service indicators) is a factor that influences variable Y (the level of public satisfaction). The value of each service indicator is used as input data to build a classification model using the Decision Tree and Random Forest algorithms. Both algorithms then learn the relationship patterns between variables X and Y to produce predictions of the level of public satisfaction. Furthermore, the performance of both algorithms is compared using the Accuracy, Precision, Recall, and F1-Score metrics to determine the algorithm that provides the most accurate prediction results for the level of public satisfaction with public services.

Research Methodology

This study uses a quantitative approach with a comparative method. The quantitative approach is used to process SPKP and SPAK data into information that can be analyzed statistically, while the comparative method is used to compare the performance of the Random Forest and Decision Tree algorithms in predicting the level of public satisfaction with public services.



Data Figure 1. Data Analysis

This research method is suitable for a proceedings article because it shows a clear research flow, starting from data collection through to the comparison of the performance of the Decision Tree and Random Forest algorithms.

3.1 Data Collection

The data were obtained from the 2025 service quality perception survey and anti-corruption perception survey files at the Medan Center for Health Quarantine service unit. The files contain 1.046 rows and 24 columns. The columns consist of respondent data (gender, age, city of residence, service work area, occupation, and highest level of education), eight service quality items, five anti-corruption items, as well as suggestion and follow-up columns. Each item was rated on a scale of 1 to 6. The data were read directly from the research storage folder into Google Colab.

3.2 Data Preprocessing

This stage aims to improve data quality before the classification process is carried out. The preprocessing stages include:

1. Checking for missing values
2. Removing duplicate data
3. Checking for data inconsistencies

4. Standardizing category entries, for example "SLTA atau Sederajat" and "SLTA atau sederajat", which were previously counted as two different values
5. Transforming categorical data into numerical data (if necessary)
6. Selecting the attributes used in the research
7. Forming class labels from the respondent index (explained below)

The class labels are formed from the survey question items. Each respondent has an average score from 13 items, and then the average is converted into an index on a scale of 0 to 100 using the following formula.

$$\text{Indeks} = \frac{\overline{x}}{6} \times 100$$

In the formula, x with a line above it is the average score of the 13 items belonging to one respondent and the number 6 is the highest score on the response scale. The index is then grouped into four categories following the interval boundaries in Ministerial Regulation of PANRB Number 14 of 2017, namely D (Not Good) for an index below 65, C (Less Good) for 65 to less than 76,61, B (Good) for 76,61 to less than 88,31, and A (Very Good) for 88,31 to 100.The labels and features come from the same items. This condition is known as data leakage because the model obtains information that is very close to the target (Kaufman et al., 2012). As a result, the high metric values most likely indicate the model's ability to guess the index grouping rules, rather than its ability to predict satisfaction from independent information. Therefore, the results of this study are interpreted with this caveat and supplemented with cross validation and permutation importance.

3.3 Dataset Splitting

The dataset is divided into two parts, namel

- Training Data = 80%
- Testing Data = 20%

The training data are used to build the model, while the testing data are used to evaluate model performance. The split uses class stratification and random_state 42 so that the class proportions in the training data and testing data are the same and the results can be reproduced.

3.4 Model Building

At this stage, two classification models are built using:

- Decision Tree algorithm
- Random Forest algorithm

Both algorithms are trained using the same training data so that their results can be compared objectively.

Both models are created using the scikit-learn library (Pedregosa et al., 2011) in a single pipeline. Missing values in numerical features are filled with the median and those in categorical features are filled with the mode, then categorical features are transformed using one-hot encoding. All of these steps are learned only from the training data so that testing data information does not leak into the model. The splitting criterion for both models is Gini. The initial Random Forest setting uses 100 trees, while the Decision Tree is allowed to grow without a depth limit.

3.5 Prediction

The trained models are used to predict the level of public satisfaction on the testing data.**6.**

Model Evaluation

The performance of both algorithms was evaluated using a Confusion Matrix with the following metrics:

- Accuracy
- Precision
- Recall
- F1-Score

The algorithm that obtained the best evaluation results was considered the most optimal model for predicting the level of public satisfaction.

In addition to a single test on the testing data, this study employed repeated stratified k-fold cross-validation with a maximum of five folds and three repetitions. The difference in F1 scores per fold between the two models was tested using the Wilcoxon signed-rank test (Wilcoxon, 1945). Because the folds across repetitions are not completely independent, the p-value was interpreted as an indication rather than absolute evidence. Hyperparameter tuning was performed using Grid Search only on the training data with the weighted F1 score. The Decision Tree search space included max_depth (3, 5, 7, 10, and unlimited), min_samples_split (2, 5, 10), and min_samples_leaf (1, 2, 4). The Random Forest search space included n_estimators (100, 200, 300), max_depth (unlimited, 5, 10), min_samples_leaf (1, 2, 4), and max_features (sqrt and log2). Finally, feature importance was calculated using impurity-based importance for Random Forest and permutation importance on the testing data for both models (Breiman, 2001).

3.6 Research Workflow

The research workflow consists of several stages carried out systematically as follows:

1. Problem identification.
2. Literature review on public services, data mining, Decision Tree, and Random Forest.
3. Collection of SPKP and SPAK data.
4. Data preprocessing.
5. Formation of class labels from respondent indices.
6. Division of the data into training data and testing data.
7. Model development using the Decision Tree algorithm.
8. Model development using the Random Forest algorithm.
9. Model testing using the testing data.
10. Evaluation using Accuracy, Precision, Recall, and F1-Score.
11. Cross-validation, Wilcoxon test, hyperparameter tuning, and featureimportance analysis.
12. Comparison of the results of both algorithms.
13. Drawing conclusions and preparing recommendations.

Results and Discussion

4.1 Data Overview and Class Distribution

The initial file contained 1,046 rows. During the cleaning stage, 1 duplicate row was removed, leaving 1,045 respondents. The email, sequence number, type of service (constant value), city of residence, suggestions, and follow-up columns were not used as features. No respondents had missing labels. The remaining features totaled 18, consisting of five respondent attributes and 13 assessment items. The class distribution resulting from label formation is shown in Table 1.

Table 1. Class Distribution of Satisfaction Categories

Kelas	Jumlah responden	Persentase (%)
D (Not Good)	29	2,78
C (Less Good)	82	7,85
B (Good)	305	29,19
A (Very Good)	629	60,19
Total	1.045	100

Class A accounts for around 60% of respondents, whereas class D accounts for only 2.78%. The ratio of the largest to the smallest class is 21.7 to 1. Under such conditions, accuracy can easily appear high even when the small class is not predicted well. This is consistent with the attention to imbalanced data in the studies by Novelan et al. (2023) and Novelan and Aryza (2025), so macro F1 is reported alongside weighted F1. The data split resulted in 836 training data points and 209 testing data points with the same class proportions in both subsets (classes D, C, B, and A in the testing data consisted of 6, 16, 61, and 126 respondents, respectively).

2. Model Performance on the Testing Data

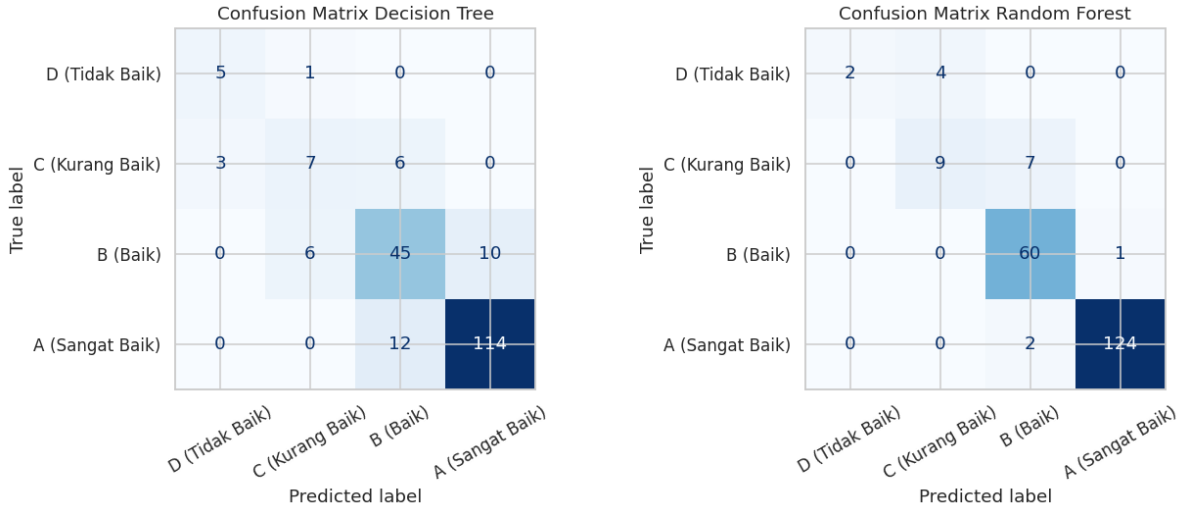
Table 2 presents the evaluation results of both models on the 209 testing data points. Figure 2 shows the confusion matrices of both models and Figure 3 compares the four main metrics.

Table 2. Evaluation Results on the Testing Data (%)

Model	Accuracy	Precision	Recall	F1 weighted	F1 macro	Akurasi training
Decision Tree	81,82	81,89	81,82	81,79	70,47	100,00
Random Forest	93,30	93,36	93,30	92,70	75,80	100,00

Random Forest outperforms on all metrics. The accuracy difference reaches 11.48 percentage points, the weighted F1 difference is 10.91 points, and the macro F1 difference is 5.33 points. Both models achieve 100% accuracy on the training data, but the difference from testing accuracy varies greatly, namely 18.18 points for Decision Tree and 6.70 points for Random

Forest. The larger difference in Decision Tree indicates a stronger tendency toward overfitting.



Gambar 2. Confusion Matrix Decision Tree dan Random Forest

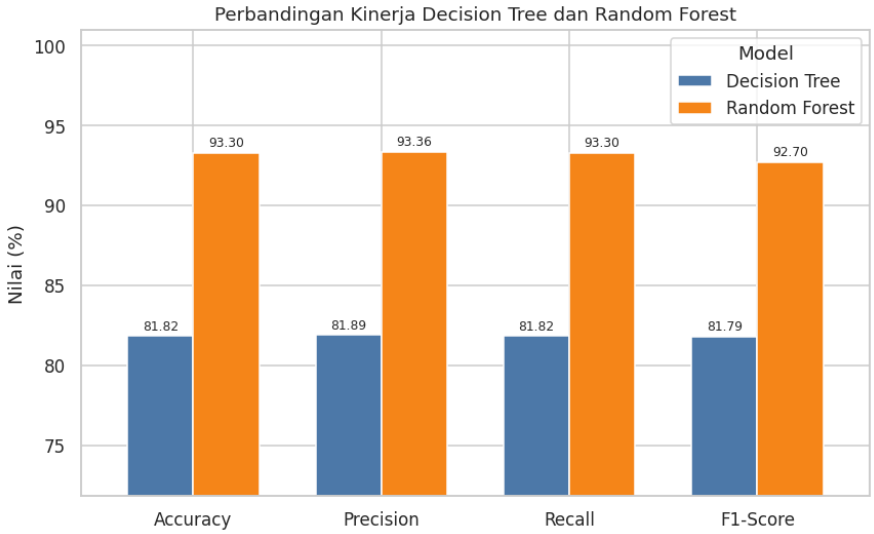


Figure 3. Comparison of Decision Tree and Random Forest Metrics

The per-class details in Table 3 reveal something that is not apparent from the average figures. For class D (Poor), Random Forest has a precision of 1.00 but a recall of only 0.33, meaning that only 2 of the 6 class D respondents were detected. Decision Tree detects 5 of those 6 respondents (recall 0.83), but its precision is 0.62. Because class D contains only 6 test respondents, a single prediction changes the recall by 16.7 percentage points, so the figures for this class should be interpreted as indicative.

Table 3. Per-Class Performance on Testing Data

Kelas	Jumlah uji	DT Precision	DT Recall	DT F1	RF Precision	RF Recall	RF F1
D (Not Good)	6	0,62	0,83	0,71	1,00	0,33	0,50
C (Less Good)	16	0,50	0,44	0,47	0,69	0,56	0,62
B (Good)	61	0,71	0,74	0,73	0,87	0,98	0,92
A (Very Good)	126	0,92	0,90	0,91	0,99	0,98	0,99

Note: DT is Decision Tree and RF is Random Forest.

4.2 Cross Validation Results and Wilcoxon Test

Because the testing data was split only once, both models were retested using repeated stratified k-fold with 5 folds and 3 repetitions (15 folds). Table 4 presents the mean and standard deviation of each metric.

Table 4. Cross Validation Results (mean ± standard deviation, %)

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	82,78 ± 2,73	82,95 ± 2,93	82,78 ± 2,73	82,72 ± 2,80
Random Forest	90,88 ± 1,56	90,75 ± 1,52	90,88 ± 1,56	90,36 ± 1,63

The pattern in cross-validation is the same as in the single test. Random Forest is higher on all metrics and has a smaller standard deviation (1.63 versus 2.80 on the F1-score), which means its performance is more stable across folds. The average difference in F1-score is 7.64 percentage points in favor of Random Forest. The Wilcoxon signed-rank test on the F1 scores per fold yielded a statistic of 0.000 with a p-value of 0.0001. The folds in the repetitions are not fully independent, so the p-value tends to be optimistic, but the direction of the difference is consistent across all folds.

4.3 Hyperparameter Tuning Results

Grid Search on the training data selected max_depth 10, min_samples_leaf 4, and min_samples_split 10 for the Decision Tree. For the Random Forest, the best parameters were n_estimators 300, max_depth unlimited, min_samples_leaf 1, and max_features sqrt. Table 5 compares the initial models with the models after tuning.

Table 5. Comparison of Initial Models and Tuned Models on Testing Data (%)

Model	Accuracy	F1 weighted	F1 macro	Akurasi training	Selisih training-testing
Decision Tree (awal)	81,82	81,79	70,47	100,00	18,18
Random Forest (awal)	93,30	92,70	75,80	100,00	6,70
Decision Tree (tuning)	80,38	80,43	70,01	92,11	11,72
Random Forest (tuning)	93,30	92,78	76,82	100,00	6,70

Tuning does not change the conclusion. For the Decision Tree, testing accuracy decreased by 1.44 points (81.82 to 80.38) even though the training-testing gap shrank from 18.18 to 11.72 because the tree depth was limited. For the Random Forest, accuracy remained at 93.30, weighted F1 increased slightly by 0.08 points, and macro F1 increased by 1.02 points. This means that the initial Random Forest settings were already close to the best combination for this dataset, and the Random Forest remained superior to the Decision Tree after both were tuned.

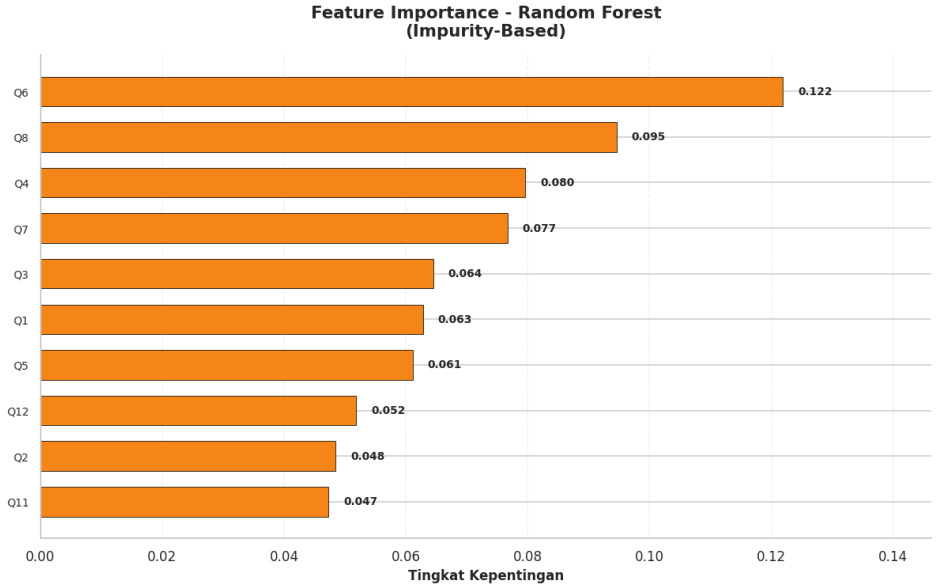
4.4 Feature Importance

Figures 4 through 6 show the top ten features according to impurity-based importance for the Random Forest, permutation importance for the Random Forest, and permutation importance for the Decision Tree. The Q codes in the three graphs refer to the 13 assessment items according to Table 6. The same feature importance approach was used by Antoni and Iqbal (2025) to explain the dominant factors in their Random Forest model.

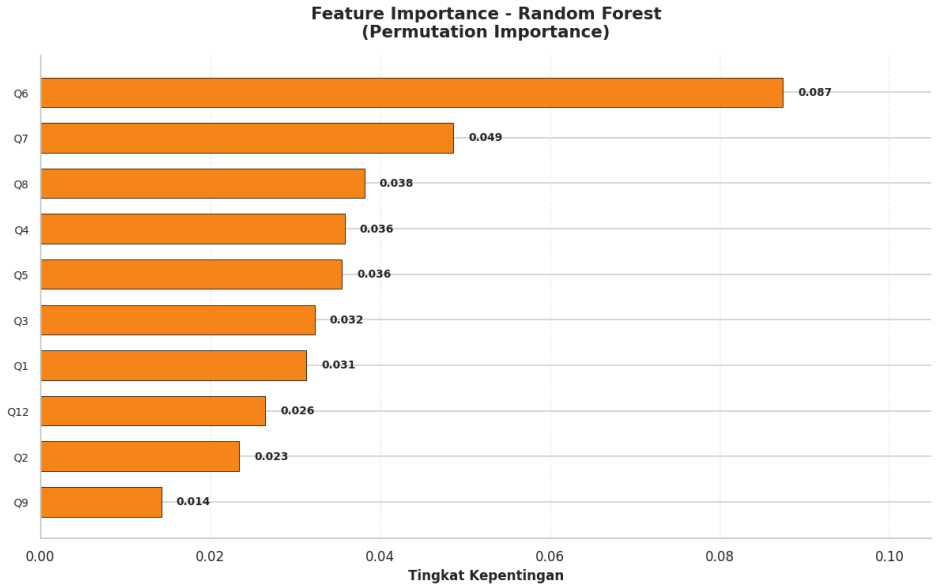
Table 6. Item Code Legend in Figures 4 through 6

Kode	Assessment items
Q1	A. Service information is available
Q2	B. Service requirements are appropriate
Q3	C. Service procedures or flow are easy to follow
Q4	D. Completion time is appropriate
Q5	E. Tariffs or fees are appropriate
Q6	F. Facilities and infrastructure and the online service system are comfortable and easy to use

Q7	G. Staff or the online system responds quickly to needs
Q8	H. Consultation and complaint services are easy to access
Q9	Anti-corruption A. Services without discrimination
Q10	Anti-corruption B. Services in accordance with procedures and without fraud
Q11	Anti-corruption C. Never asks for compensation
Q12	Anti-corruption D. No illegal charges
Q13	Anti-corruption E. No brokering



Gambar 4. Feature Importance Random Forest (Impurity-Based)



Gambar 5. Feature Importance Random Forest (Permutation Importance)

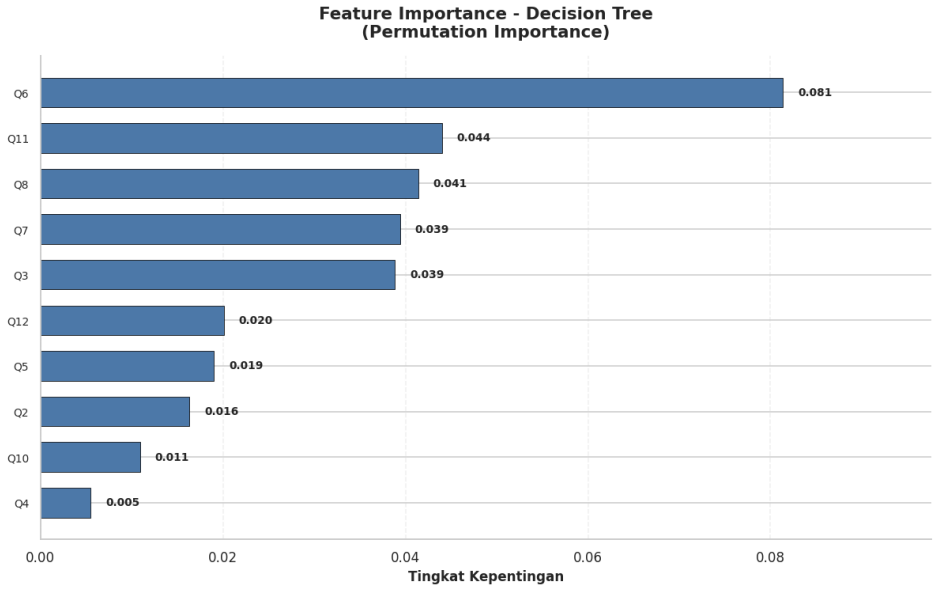
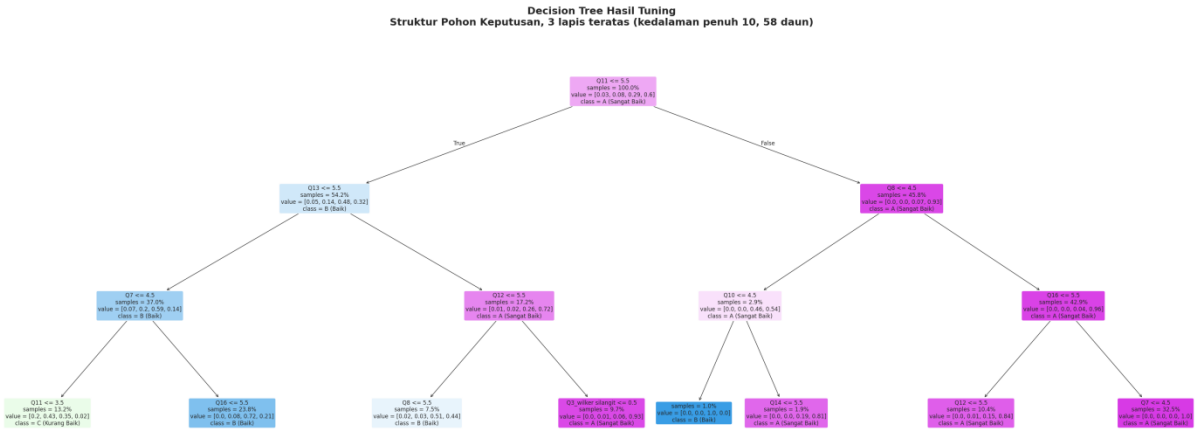


Figure 6. Feature Importance Decision Tree (Permutation Importance)

The three graphs point to the same item in first place, namely Q6 (infrastructure and online service system) with a value of 0.122 for impurity-based importance, 0.087 for Random Forest permutation importance, and 0.081 for Decision Tree permutation importance. The value of the second item is far below it. In Random Forest permutation importance, Q7 (staff responsiveness) has a value of 0.049 and Q8 (consultation and complaints) 0.038. In the Decision Tree, Q11 (does not request compensation) has a value of 0.044 and Q8 0.041. No respondent attributes such as gender, age, work area, occupation, or education are among the top ten in the three graphs. Impurity-based importance can be biased toward features that have many values, so the order in Figure 4 (Q6, Q8, Q4, Q7) is slightly different from permutation importance, but the top item remains the same. Because the label is formed from the average of the same items, this order indicates the items that most help the model reproduce the index in this data and cannot be interpreted as a causal effect on satisfaction.

4.5 Decision Tree Rules

The tuned Decision Tree has a full depth of 10 and 58 leaves. Figure 7 displays the top three layers. Because the full tree is too large to print, the rules at the upper layers are summarized in Table 7. Item scores are integer values from 1 to 6, so a threshold of 5.5 in the tree means a score of 6 for the larger branch and a score of 5 or below for the smaller branch. The Q codes in Figure 7 differ from Figures 4 to 6 because they use the numbering of all 18 attributes, and their meanings are explained in Table 8.



Gambar 7. Struktur Pohon Keputusan (3 lapis teratas dari kedalaman penuh 10 dengan 58 daun)

Tabel 7. Summary of the Main Nodes of the Decision Tree

Condition (respondent score)	Data share (%)	Dominant class	Composition D / C / B / A (%)
Facilities and infrastructure = 6	45,8	A (Very Good)	0 / 0 / 7 / 93
Facilities and infrastructure = 6 and procedures ≥ 5	42,9	A (Very Good)	0 / 0 / 4 / 96
Facilities and infrastructure ≤ 5 and consultation ≤ 5	37,0	B (Good)	7 / 20 / 59 / 14
Facilities ≤ 5, consultation ≤ 5, requirements ≤ 4	13,2	C (Less Good)	20 / 43 / 35 / 2
Facilities ≤ 5, consultation ≤ 5, requirements ≥ 5	23,8	B (Good)	0 / 8 / 72 / 21
Facilities and infrastructure ≤ 5 and consultation = 6	17,2	A (Very Good)	1 / 2 / 26 / 72

The root of the tree is the infrastructure item. Respondents who gave a score of 6 on this item (45.8% of the data) are almost entirely in class A (93%). The group of greatest interest to the agency is respondents with low infrastructure, consultation services, and requirements scores, because this group (13.2% of the data) contains 63% of class D and C respondents combined (20% and 43%). This finding is in line with the feature importance results in the previous

sectionTable 8. Code Legend in Figure 7

Kode	Attribute
Q3	Service work area (Silangit Wilker branch)
Q7	B. Service requirements are appropriate
Q8	C. Service procedures or flow are easy to follow
Q10	E. Tariffs or fees are appropriate
Q11	F. Facilities and infrastructure and online service system
Q12	G. Staff or online system responsiveness
Q13	H. Consultation and complaint services
Q14	Anti-corruption A. Service without discrimination
Q16	Anti-corruption C. Never asks for compensation

Discussion

The three sources of evidence point to the same model. On the testing data, Random Forest was higher with an accuracy of 93,30% compared with 81,82%. In cross validation, the Random Forest F1-score was 90,36% compared with 82,72% and the Wilcoxon test gave a p-value of 0,0001. After tuning, this advantage persisted. Thus, Random Forest was more accurate and more stable than Decision Tree for the SPKP and SPAK data in this study.

This pattern is consistent with other studies. Fadri et al. (2024) reported Random Forest at 95,78% and Decision Tree at 93,82% for airline passenger satisfaction. Ammar and Wiranata (2025) obtained Random Forest at 92,9% and Decision Tree at 86,8% in sentiment analysis of the National Digital Samsat application. Amalia et al. (2022) also found Random Forest to be superior on airline customer satisfaction data with an accuracy of 95,827%. The difference in this study (11,48 points) was larger than the difference in Fadri et al. (1,96 points). One possible cause is that the Decision Tree without a depth limit overfit the training data of around 836 rows, but this hypothesis was not examined further. Sitompul and Iqbal (2025) also reported that Random Forest was superior on telecommunications customer

experience data, while Ernawati et al. (2025) found a very small difference between Random Forest and KNN, indicating that the magnitude of the advantage indeed depends on the data.

Random Forest was not superior in all respects. For class D (Poor), this model detected only 2 of 6 respondents, whereas Decision Tree detected 5 of 6. For service agencies, this is the most important class to identify. The high overall accuracy of Random Forest was largely supported by classes A and B, which had large numbers. Therefore, the macro F1 (75,80 for Random Forest and 70,47 for Decision Tree) more honestly describes performance on minority classes, and its value was far below the accuracy for both models.

Decision Tree was superior in interpretability. Rules such as those in Table 7 can be read directly by service unit staff, whereas Random Forest only provides a ranking of feature importance. Iqbal and Efendi (2023) also presented their classification results in the form of a decision tree. The choice of model therefore depends on the objective. For the most accurate and stable predictions, Random Forest is more suitable. For explaining to management which items need to be improved, Decision Tree is easier to use.

This study has several limitations. First, the labels were formed from the same items as the features, creating a risk of data leakage. The accuracy did not reach 100%, possibly because the tree estimates the average of 13 items through splitting one item at a time, but the high values still more strongly reflect the model's ability to learn the index formula than its ability to independently predict satisfaction. Second, the classes were imbalanced and class D contained only 6 test respondents, making the metrics for this class less stable. Third, the data came from one institution and one period. Fourth, the category thresholds were adopted from a 4-point scale, so their application to a 6-point scale needs to be justified. Fifth, several categorical columns still had variations in spelling, for example in the highest education field there were values that differed only in punctuation. This model is more appropriately used as an initial support tool to flag groups of respondents with a tendency toward low satisfaction and does not replace the calculation of the official index.

Conclusion

This study developed and compared Decision Tree and Random Forest models to predict user satisfaction categories based on SPKP and SPAK data from 1.045 respondents. On the testing data, Random Forest achieved an accuracy of 93,30% and a weighted F1-score of 92,70%, while Decision Tree achieved 81,82% and 81,79%. Five-fold cross validation with 3 repetitions showed the same pattern (F1-score 90,36% compared with 82,72%) and the Wilcoxon test gave a p-value of 0,0001, so Random Forest was considered more accurate and more stable. This advantage persisted after hyperparameter tuning.

Random Forest was weak on the Poor class with a recall of 0,33, while Decision Tree reached 0,83 on the same class although its precision was lower. Service infrastructure and online service system items were the most influential items in all three feature importance analyses and became the root of the decision tree. Decision Tree remains useful because its rules are easy to read, but the results of this study need to be interpreted with the caveat that the labels were calculated from the same items as the features.

For future research, it is recommended to use satisfaction labels that are separate from the feature items, address class imbalance with class weight or oversampling, add data from several periods or institutions, and try other algorithms such as Gradient Boosting.

References

[1] Amalia, S., Deborah, I., & Yulita, I. N. (2022). Comparative analysis of classification algorithm: Random Forest, SPAARC, and MLP for airlines customer satisfaction. SINERGI, 26(2), 213-222. https://doi.org/10.22441/sinergi.2022.2.010

[2] Ammar, H., & Wiranata, A. D. (2025). Perbandingan tingkat akurasi algoritma Naïve Bayes, Decision Tree dan Random Forest dalam analisis sentimen pengguna aplikasi Samsat Digital Nasional. IJAI (Indonesian Journal of Applied Informatics), 10(1), 59-73. https://doi.org/10.20961/ijai.v10i1.111010

[3] Antoni, R., & Iqbal, M. (2025). Classification of study program competitiveness levels in SNBT using the random forest method based on capacity and regional distribution data. Proceedings of International Conference on Islamic Community Studies, 7622-7632. https://proceeding.pancabudi.ac.id/index.php/ICIE/article/view/1452

[4] Balai Besar POM di Bogor. (2024). Laporan pelaksanaan survei persepsi kualitas pelayanan dan survei persepsi anti korupsi Mei 2024. https://bogor.pom.go.id/storage/informasipublik/Laporan%20SPKP%20SPAK%20Mei%202024%20Bogor.pdf

[5] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

[6] Ernawati, A., Khairul, Sitorus, Z., Iqbal, M., & Nasution, D. (2025). Penerapan data mining untuk klasifikasi penduduk miskin di Kabupaten Labuhanbatu menggunakan Random Forest dan K-Nearest Neighbors. Bulletin of Information Technology (BIT), 6(2), 23-35. https://doi.org/10.47065/bit.v6i1.1783

[7] Fadri, A. I., Zahfran, A., Irak, T., Firjatullah, N. H., & Herianto, J. E. (2024). Comparison of supervised learning algorithms for predicting airline passenger satisfaction. https://journal.irpi.or.id/index.php/ijatis/article/view/1868

[8] Fitaria, N., & Darwis, D. (2025). Analisis kinerja Decision Tree dan Naïve Bayes pada klasifikasi tingkat kepuasan masyarakat. https://ejurnal.seminar-id.com/index.php/bits/article/view/9325

[9] Iqbal, M., & Efendi, S. (2023). Data-driven approach for credit risk analysis using C4.5 algorithm. ComTech: Computer, Mathematics and Engineering Applications, 14(1), 11-20. https://journal.binus.ac.id/index.php/comtech/article/view/8243

[10] Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery from Data, 6(4), 15. https://doi.org/10.1145/2382577.2382579

[11] Kementerian PANRB. (2017). Peraturan Menteri PANRB Nomor 14 Tahun 2017 tentang Pedoman Penyusunan Survei Kepuasan Masyarakat Unit Penyelenggara Pelayanan Publik.

[12] Kementerian PANRB. (2021). Peraturan Menteri PANRB Nomor 90 Tahun 2021 tentang Pembangunan dan Evaluasi Zona Integritas Menuju Wilayah Bebas dari Korupsi dan Wilayah Birokrasi Bersih dan Melayani di Instansi Pemerintah.

[13] Novelan, M. S., Efendi, S., Sihombing, P., & Mawengkang, H. (2023). Vehicle routing problem optimization with machine learning in imbalanced classification vehicle route data. Eastern-European Journal of Enterprise Technologies, 5(3(125)), 49-56. https://doi.org/10.15587/1729-4061.2023.288280

[14] Novelan, M. S., & Aryza, S. (2025). Optimization CVRP with machine learning for improved classification of imbalanced data food distribution. JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer), 10(4), 917-925. https://doi.org/10.33480/jitk.v10i4.6467

[15] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830. https://jmlr.org/papers/v12/pedregosa11a.html

[16] Purbolaksono, M. D., Tantowi, M. I., Hidayat, A. I., & Adiwijaya, A. (2021). Perbandingan Support Vector Machine dan Modified Balanced Random Forest dalam deteksi pasien penyakit diabetes. Jurnal RESTI, 5(2), 393-399. https://doi.org/10.29207/resti.v5i2.3008

[17] Septhya, D., Rahayu, K., Rabbani, S., Fitria, V., Rahmaddeni, R., Irawan, Y., & Hayami, R. (2023). Implementasi algoritma Decision Tree dan Support Vector Machine untuk klasifikasi penyakit kanker paru. MALCOM: Indonesian Journal of Machine Learning and Computer Science, 3(1). https://doi.org/10.57152/malcom.v3i1.591

[18] Setiono, M. H. (2022). Komparasi algoritma Decision Tree, Random Forest, SVM dan K-NN dalam klasifikasi kepuasan penumpang maskapai penerbangan. INTI Nusa Mandiri. https://ejournal.nusamandiri.ac.id/index.php/inti/article/download/3420/961

[19] Sitompul, J. R., & Iqbal, M. (2025). An intelligent predictive model for customer experience evaluation in telecom networks. Proceedings of International Conference on Islamic Community Studies, 6644-6646. https://proceeding.pancabudi.ac.id/index.php/ICIE/article/view/1280

[20] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427-437.https://doi.org/10.1016/j.ipm.2009.03.002Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics Bulletin, 1(6), 80-83. https://doi.org/10.2307/3001968