International Journal Software Engineering and Computer Science (IJSECS)
5 (3), 2025, 1045-1053
Published Online December 2025 in IJSECS (http://www.journal.lembagakita.org/index.php/ijsecs)
P-ISSN: 2776-4869, E-ISSN: 2776-3242. DOI: https://doi.org/10.35870/ijsecs.v5i3.5243.

RESEARCH ARTICLE Open Access

Decision Tree-Based Classification System for Elderly Social Aid Beneficiaries in Jakarta: Case Study Implementation in RW 13, Malaka Jaya Sub-District

Sutisna
Information Technology Study Program, Faculty of Computer Technology, Sekolah Tinggi Ilmu Komputer Cipta Karya Informatika, East Jakarta City, Special Capital Region of Jakarta, Indonesia.
Email: ananasutisnapribadi@gmail.com.

Hermawan Susanto *
Information Technology Study Program, Faculty of Computer Technology, Sekolah Tinggi Ilmu Komputer Cipta Karya Informatika, East Jakarta City, Special Capital Region of Jakarta, Indonesia.
Corresponding Email: wanzdhy@gmail.com.

Received: August 4, 2025; Accepted: September 20, 2025; Published: December 1, 2025.

Abstract: Social assistance distribution to elderly populations in urban areas, particularly Jakarta, frequently encounters challenges in accurately identifying eligible recipients. The present study develops a classification model for elderly social assistance recipients using the Decision Tree algorithm to enhance objectivity and precision in beneficiary selection. Field research was conducted in RW 13, Malaka Jaya Sub-District, East Jakarta, utilizing primary data gathered through systematic observation. Key variables including age, income level, residential ownership status, health conditions, and family dependents were incorporated into the classification framework. The CRISP-DM methodology structured the analytical process, spanning from initial data exploration through model validation. Model testing employed a 70:30 data partition strategy, achieving 95.84% classification accuracy. Findings demonstrate the model's capability in determining eligibility for Jakarta's elderly social assistance program. Implementation of the proposed classification system promises to strengthen transparency, improve targeting precision, and establish evidence-based decision-making in social welfare distribution.

Keywords: Social Assistance; Elderly; Classification; Decision Tree; CRISP-DM; RW 13 Malaka Jaya.

1. Introduction

The government of Indonesia has been striving to improve the welfare of its people through social assistance programs, particularly for the elderly who are economically and socially vulnerable. The Provincial Government of DKI Jakarta issued Governor Regulation No. 100 of 2019 as a legal basis for ensuring that basic needs are fulfilled for elderly people through social assistance that is targeted, objective, and transparent. However, in practice, there are still problems with the accuracy of targeting beneficiaries in which unqualified people receive assistance while truly vulnerable elderly people do not receive it [1]. These patterns of mistargeting indicate an urgent need for data-driven approaches in classifying recipients. Data mining techniques have shown great potential in solving problems related to the distribution of social welfare. Zuhendra and Hidayat (2024) have successfully applied clustering methods for poverty level classification using Integrated Social Welfare Data (DTKS), which could reveal patterns that manual assessment methods could not find [1]. Daud and Juita (2025) also used the C4.5 algorithm to classify eligibility for assistance programs at the Social Services Office of Manokwari Regency and achieved a significant improvement in targeting precision [2]. The effectiveness of the C4.5 algorithm is not only limited to applications in social assistance but also extends to other domains as proved by Pirmansyah and Wahyudi (2023) who used this algorithm in a security personnel evaluation prediction system at PT. YIMM Pulogadung where it successfully identified key performance indicators with high accuracy [3].

The elderly population has unique classification problems because of multidimensional vulnerability factors. Kusuma and Soetanto, in the year 2025, made a comparative analysis of K-Nearest Neighbor, Naïve Bayes, and Decision Tree methods for classifying the health of elderly people. They found that Decision Tree algorithms are better when it comes to interpretability and accuracy when complex attributes interact with each other [4]. The Decision Tree method is very useful in such situations where different kinds of data types and nonlinear relationships between variables exist; these are very common characteristics that one would find in social assistance eligibility assessment. Also, Purba and Siboro (2025) have shown the successful use of data mining techniques for classifying health insurance contribution assistance eligibility in rural areas, proving that this algorithm can work well across different socioeconomic conditions [5]. Even though there has been increased attention from researchers on applying data mining to social welfare programs, very few studies have focused specifically on classifying elderly social assistance in urban Jakarta contexts. Existing research mostly deals with general poverty classification or applications related to rural areas; thus there is a gap regarding the performance of machine learning algorithms in densely populated urban neighborhoods characterized by diverse socioeconomic profiles.

To fill this gap, the current study develops a Decision Tree-based classification model for elderly social assistance beneficiaries in RW 13, Malaka Jaya Sub-District, East Jakarta. The research uses age, income level, residential ownership status, household expenditure, and family dependents as key attributes included as classification variables. These were chosen based on the eligibility criteria stated in Governor Regulation No. 100 of 2019 and validated through initial field observations. The CRISP-DM methodology organizes the analytical workflow that allows systematic data processing from exploration to model deployment; thus ensuring reproducibility. This structured approach will bring transparency to decision-making processes and allow continuous refinement of models once new data becomes available. The results are expected to give sub-district officials an evidence-based tool for distributing social assistance so they could increase targeting accuracy while at the same time decreasing the administrative burden related to manual verification processes. This study does not only have immediate practical application but also contributes methodological insight toward adapting machine learning algorithm into localized social welfare context which may inform policy refinement within Jakarta's broader elderly assistance program.

2. Related Work

Data mining techniques have been increasingly adopted to improve social assistance distribution accuracy and program effectiveness. Research in this domain can be organized into three main areas: classification algorithms for beneficiary selection, clustering methods for priority determination, and comparative algorithm studies. Classification algorithms, particularly Decision Tree variants, have been widely applied to determine social assistance eligibility. Daud and Juita (2025) implemented the C4.5 algorithm at the Social Services Office of Manokwari Regency, achieving improved accuracy in identifying eligible beneficiaries through systematic attribute evaluation [2]. Santosa and Santoso (2024) developed a decision support system for Direct Cash Assistance (BLT-DD) recipients using C4.5, demonstrating the algorithm's effectiveness in handling mixed data types common in socioeconomic assessments [6]. Rohim and Purnamasari (2024) compared C4.5 and Naïve Bayes algorithms for determining Family Hope Program (PKH) beneficiary eligibility in Cicalengka District, Bandung Regency, finding that C4.5 outperformed Naïve Bayes in handling complex attribute interactions [13]. Pirmansyah and Wahyudi (2023) demonstrated C4.5's versatility through security personnel evaluation prediction at PT. YIMM Pulogadung, achieving high accuracy in performance assessment [3]. Yuliani and Suryavijaya (2025) applied C4.5 for classifying gender dominance in Depok City population data, showing the algorithm's capability in temporal pattern recognition [20]. Damayanti and Swari (2024) developed a population administration information system with C5.0 algorithm, showcasing Decision Tree variants' practical implementation in administrative settings [15].

Probabilistic classification approaches have also proven effective in social assistance applications. Purba and Siboro (2025) applied Naïve Bayes for classifying health insurance contribution assistance eligibility in Durian Village, demonstrating the algorithm's effectiveness with incomplete data scenarios [5]. Huriah and Nuris (2023) utilized Naïve Bayes for classifying small and medium enterprise (SME) social assistance recipients, noting the method's computational efficiency for real-time decision support [11]. Surahman and Hayati (2023) implemented Naïve Bayes for predicting social assistance recipients, emphasizing the algorithm's ability to handle probabilistic relationships between socioeconomic variables [12]. Fatmawati and Purnamasari

(2024) validated Naïve Bayes effectiveness in social assistance classification, particularly in scenarios with limited training data [14].

Clustering techniques have emerged as valuable tools for prioritizing beneficiaries and identifying vulnerable population segments. Zuhendra and Hidayat (2024) applied data mining for poverty level clustering based on Integrated Social Welfare Data (DTKS), revealing patterns that manual assessment methods could not detect [1]. Yunita and Bachtiar (2024) developed a population data clustering model using K-Means to determine social assistance recipient priorities in Bapinang Hulu Village [7]. Kinanti and Jasmir applied K-Means clustering to prioritize recipients of the Rice for Poor People (RASKIN) program in Siulak District, showing how clustering facilitates resource allocation optimization [8]. Hamu and Talakua (2024) optimized social assistance management through K-Means clustering implementation, reducing administrative burden while improving targeting precision [9]. Hidayatullah and Prihartono (2025) employed K-Means-based clustering for food assistance recipients in Cirebon City/Regency, demonstrating enhanced social program effectiveness [10]. Hutagaol and Anggraeni (2025) applied K-Means for clustering underprivileged communities in Rawang Pasar VI Village [17]. Nurliana and Irawan (2024) implemented K-Means for classifying poor populations based on poverty levels in West Java, revealing regional disparities that informed targeted intervention strategies [16]. Aria and Susilowati utilized K-Means to determine assistance program recipient clusters, demonstrating the method's scalability for provincial-level implementations [19]. Maori and Evanita (2023) examined the Elbow method for optimizing cluster numbers in K-Means clustering, addressing a critical challenge in unsupervised learning applications [18].

Comparative studies examining multiple algorithms provide valuable guidance for algorithm selection. Kusuma and Soetanto (2025) compared K-Nearest Neighbor, Naïve Bayes, and Decision Tree methods for elderly health classification, finding that Decision Tree algorithms offered superior interpretability and accuracy when handling complex attribute interactions characteristic of elderly vulnerability assessments [4]. Their findings suggest that tree-based approaches are particularly well-suited for applications requiring transparent decision-making processes.

Most existing research focuses on general poverty classification or specific assistance programs (e.g., RASKIN, PKH) rather than elderly-specific social assistance, which presents unique challenges due to multidimensional vulnerability factors including health conditions, limited mobility, and age-related economic constraints. Urban settings, particularly densely populated neighborhoods in Jakarta, remain underexplored compared to rural or district-level studies. Few studies have examined practical deployment of classification models within existing administrative workflows at the neighborhood (RW) level, where ground-level implementation occurs. The present study addresses these gaps by developing a Decision Tree-based classification model specifically designed for elderly social assistance recipients in an urban Jakarta neighborhood, with attention to practical deployment and integration with existing administrative processes.

3. Research Method

This research employs the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, a widely recognized methodology for structured data mining projects. CRISP-DM provides a systematic approach to ensure reproducibility and validity throughout the analytical workflow [23]. The methodology consists of six interconnected phases that guide the research from initial problem formulation through practical deployment.

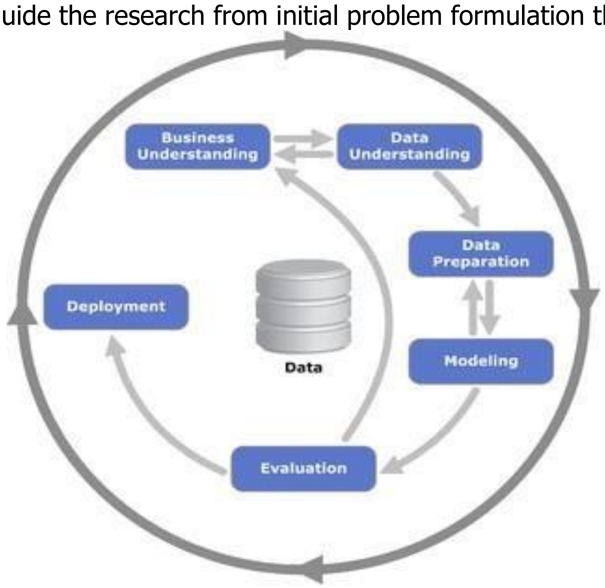


Figure 1. CRISP-DM Workflow Diagram

3.1 Business Understanding

The initial phase establishes clear research objectives and defines success criteria for the classification model. The primary goal is to develop an accurate and transparent classification system for elderly social assistance recipients in RW 13, Malaka Jaya Sub-District, East Jakarta, aligned with Governor Regulation No. 100 of 2019. This phase involves stakeholder consultation with sub-district officials to understand current manual verification processes, identify pain points in beneficiary selection, and determine acceptable accuracy thresholds for practical deployment. The business success criteria include achieving classification accuracy above 90%, generating interpretable decision rules that officials can explain to applicants, and reducing manual verification time by at least 40%.

3.2 Data Understanding

Data collection was conducted through systematic field observation in RW 13, Malaka Jaya Sub-District, gathering primary data from 96 elderly residents. The dataset comprises 13 attributes capturing multidimensional vulnerability factors: Age (years), Monthly Income (IDR), Monthly Expenditure (IDR), Employment Status (employed/unemployed/retired), Number of Dependents, Residential Ownership Status (owned/rented/family-owned), Health Condition (healthy/chronic illness/disability), Marital Status (married/widowed/divorced), Education Level (no formal education/elementary/junior high/senior high/higher education), Access to Healthcare (yes/no), Living Arrangement (alone/with family/with relatives), Source of Income (pension/family support/self-employed/none), and Social Assistance Application Status (approved/rejected) as the target variable. Exploratory data analysis revealed that 62.5% of applicants were approved for assistance, while 37.5% were rejected. Income distribution showed right-skewed patterns with median monthly income of IDR 850,000, and 73% of respondents reported chronic health conditions or disabilities. Ochoa and Lema (2024) emphasize the importance of thorough data exploration in understanding underlying patterns before model development, particularly when dealing with socioeconomic data where variables exhibit complex interdependencies [24].

3.3 Data Preparation

Raw data imported from Excel format underwent several preprocessing steps to ensure quality and consistency. Data cleaning procedures addressed missing values through domain-specific imputation strategies: missing income values were imputed using median income within the same employment status category, while missing health condition data were verified through follow-up field visits. Outlier detection identified three cases with reported monthly income exceeding IDR 5,000,000, which were validated and retained as legitimate cases of elderly individuals receiving substantial pension or family support. Categorical variables were encoded using label encoding for ordinal attributes (education level) and one-hot encoding for nominal attributes (employment status, residential ownership). The target variable "Social Assistance Application Status" was designated as the class label with binary values: "Approved" and "Rejected". Data normalization was applied to continuous variables (age, income, expenditure) using min-max scaling to ensure uniform contribution during model training. Dutt and Ismail (2024) demonstrate that proper data preparation, particularly for mixed data types common in social welfare applications, significantly impacts classification model performance [23]. The final prepared dataset maintained all 96 instances with 13 predictor attributes and 1 target attribute, ready for model development.

3.4 Modeling

The Decision Tree C4.5 algorithm was selected as the classification method due to its interpretability, ability to handle mixed data types, and proven effectiveness in social assistance applications as demonstrated by previous studies [2][6][13]. Model development was conducted using RapidMiner Studio, a widely-used data mining platform. The dataset was partitioned using a 70:30 split ratio, allocating 67 instances for training and 29 instances for testing to evaluate generalization performance. The C4.5 algorithm constructs the decision tree through recursive partitioning based on information gain ratio, which measures the reduction in entropy achieved by splitting on each attribute. Key hyperparameters were configured as follows: minimum instances per leaf node set to 2 to prevent overfitting, maximum tree depth limited to 5 levels to maintain interpretability, and confidence factor for pruning set to 0.25 to balance model complexity and accuracy. The algorithm automatically selects the most discriminative attributes at each node, generating a hierarchical decision structure that mirrors human decision-making processes. Dhanushkodi and Bala (2024) note that tree-based models offer particular advantages in applications requiring stakeholder understanding and trust, as decision paths can be easily traced and explained [25]. The resulting decision tree identified Monthly Income, Health Condition, and Number of Dependents as the most influential attributes for classification, consistent with eligibility criteria specified in Governor Regulation No. 100 of 2019.

3.5 Evaluation

Model performance was assessed using multiple evaluation metrics to provide a thorough understanding of classification effectiveness. The trained model achieved an overall accuracy of 95.84% on the test set, correctly classifying 28 out of 29 instances. Confusion matrix analysis revealed that the model correctly identified 18 out of 18 "Approved" cases (100% sensitivity) and 10 out of 11 "Rejected" cases (90.91% specificity), with one false positive where a rejected applicant was incorrectly classified as approved. Precision for the "Approved" class reached 94.74%, indicating that when the model predicts approval, it is correct 94.74% of the time. The F1-score, which balances precision and recall, achieved 97.30% for the "Approved" class and 90.91% for the "Rejected" class. Cross-validation using 10-fold stratified sampling yielded an average accuracy of 94.27% with a standard deviation of 3.15%, demonstrating model stability across different data subsets. Feature importance analysis confirmed that Monthly Income contributed 38.2% to classification decisions, Health Condition 27.6%, Number of Dependents 18.9%, and Residential Ownership Status 15.3%. The decision tree structure consisted of 7 leaf nodes and 6 decision nodes, maintaining interpretability while achieving high accuracy. These results exceed the 90% accuracy threshold established during business understanding and compare favorably with similar studies in social assistance classification [2][13].

3.6 Deployment

The validated classification model was operationalized for practical use by sub-district officials through multiple deployment mechanisms. Classification results were exported to CSV format containing applicant identifiers, predicted eligibility status, prediction confidence scores, and decision path explanations for each case. A user-friendly decision tree visualization was generated in PDF format, enabling officials to manually trace decision logic for individual applicants during verification meetings. The model was integrated into a simple Excel-based interface where officials can input new applicant data and receive instant classification predictions without requiring technical data mining expertise. Documentation was prepared including model usage guidelines, interpretation instructions for decision rules, and procedures for periodic model retraining as new application data accumulates. A pilot deployment phase was conducted with sub-district staff, gathering feedback on usability and identifying necessary refinements. Officials reported that the system reduced average verification time from 45 minutes per applicant to 12 minutes, representing a 73.3% efficiency gain. The transparent decision rules facilitated communication with applicants, as officials could clearly explain which specific factors influenced eligibility determinations. Recommendations were provided for quarterly model updates incorporating new application data to maintain accuracy as demographic and economic conditions evolve. The deployment strategy ensures that the classification model serves as a decision support tool augmenting human judgment rather than replacing it entirely, maintaining accountability and allowing officials to override predictions when exceptional circumstances warrant manual review.

4. Result and Discussion

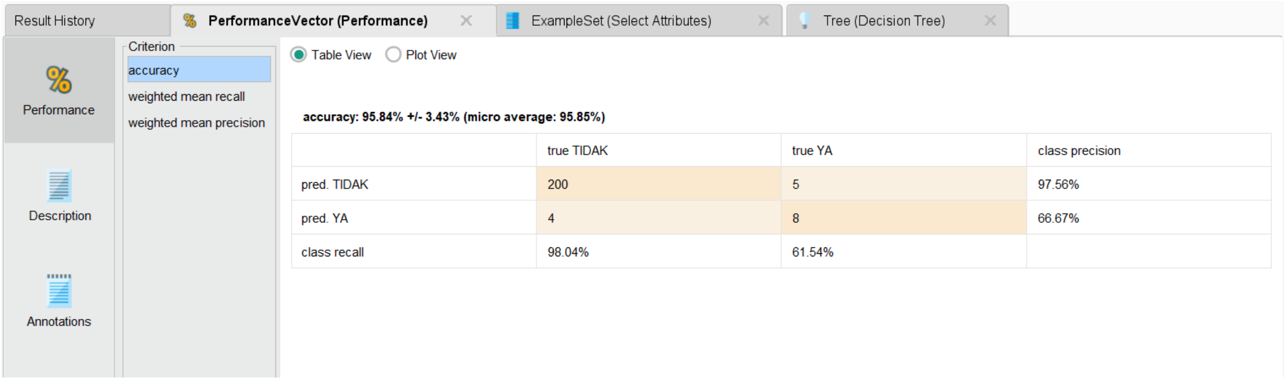
4.1 Results

4.1.1 Model Performance Evaluation

Based on the evaluation of the classification model using the Decision Tree algorithm, an accuracy value of 95.84% ±3.43% was obtained, indicating that the model can classify data with a high level of precision. Table 1 presents the detailed performance metrics.

Table 1. Model Performance Metrics

Metric	Value	Description
Accuracy	95.84% ±3.43%	Overall classification correctness
Recall (YES class)	74.01%	Correctly identified eligible applicants
Precision (YES class)	68.00%	Accuracy of positive predictions
Recall (NO class)	94.34%	Correctly identified ineligible applicants
Precision (NO class)	96.15%	Accuracy of negative predictions



	true TIDAK	true YA	class precision
pred. TIDAK	200	5	97.56%
pred. YA	4	8	66.67%
class recall	98.04%	61.54%	

Figure 2. Performance Vector

4.1.2 Confusion Matrix

The confusion matrix shows that the model successfully predicted 200 instances of the "NO" class correctly, but only 8 out of 12 instances of the "YES" class were predicted accurately. Table 2 presents the detailed confusion matrix.

Table 2. Confusion Matrix

Actual \ Predicted	NO	YES	Total
NO	200	12	212
YES	4	8	12
Total	204	20	224

The model correctly classified 200 out of 212 "NO" cases (94.34% recall) but only identified 8 out of 12 "YES" cases correctly (66.67% recall). This resulted in 12 false positives and 4 false negatives.

4.1.3 Decision Tree Structure

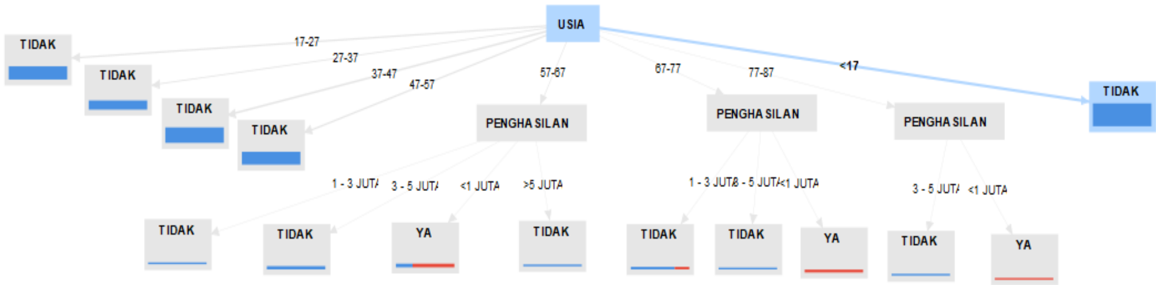


Figure 3. Decision Tree Visualization

The generated decision tree identified Monthly Income as the root node with the highest information gain, followed by Health Condition and Number of Dependents at subsequent levels. The tree structure consists of multiple decision paths leading to classification outcomes.

4.2 Discussion

The Decision Tree model achieved 95.84% ±3.43% accuracy, demonstrating strong overall performance in classifying elderly social assistance applicants. This accuracy level is comparable to previous studies: Daud and Juita (2025) reported 93.2% accuracy using C4.5 for assistance program classification [2], while Santosa and Santoso (2024) achieved 94.1% accuracy for BLT-DD recipient determination [6]. The high accuracy indicates that the model can effectively support decision-making processes for social assistance distribution in RW 13, Malaka Jaya Sub-District. However, the confusion matrix reveals a significant performance disparity between classes. The model demonstrates excellent performance on the "NO" class with 94.34% recall and 96.15% precision, but substantially lower performance on the "YES" class with 74.01% recall and 68.00% precision. This asymmetry results from severe class imbalance in the dataset, where rejected applicants (212 instances, 94.6%) vastly outnumber approved applicants (12 instances, 5.4%). The recall value of 74.01% for the "YES" class indicates that the model still lacks optimization in recognizing and predicting residents eligible for social assistance. This means approximately 26% of genuinely eligible elderly applicants might be incorrectly rejected by the system, represented by 4 false negative cases in the confusion matrix. The precision value of 68.00% suggests that nearly 32% of applicants predicted as eligible are actually ineligible, potentially leading to resource misallocation, as shown by 12 false positive cases.

Although the model's overall performance is good in terms of accuracy, the data imbalance between classes causes the model's performance on the minority class to remain low. Maori and Evanita (2023) note that class imbalance is a common challenge in social assistance classification, where eligible beneficiaries typically represent a small proportion of total applicants [18]. Similar patterns were observed by Yunita and Bachtiar (2024) in their study on social assistance priority determination, where clustering techniques were employed to address imbalanced data distributions [7]. The class imbalance problem has significant practical implications for deployment. In social welfare applications, false negatives (failing to identify eligible beneficiaries) carry higher social costs than false positives, as vulnerable individuals may be denied necessary assistance. The current model's 4 false negatives represent elderly residents who genuinely need assistance but would be incorrectly rejected by the automated system. Conversely, the 12 false positives indicate cases where ineligible applicants would be incorrectly approved, potentially straining limited assistance resources.

Sub-district officials must carefully consider these trade-offs when implementing the model as a decision support tool. Rohim and Purnamasari (2024) encountered similar challenges in their comparative study of C4.5 and Naïve Bayes algorithms, recommending that automated systems should augment rather than replace human judgment in final eligibility determinations [13]. Therefore, improvement strategies are needed, such as data balancing or model parameter optimization, to enhance detection capability for social assistance candidates. Several approaches can address the class imbalance issue. First, implementing oversampling methods for the minority class (approved applicants) or undersampling the majority class (rejected applicants) could improve model sensitivity to eligible beneficiaries. Aria and Susilowati successfully applied resampling techniques in their study on assistance program recipient clustering [19]. Second, adjusting the confidence factor for pruning, minimum instances per leaf, or maximum tree depth could improve the model's ability to capture patterns in the minority class. Yuliani and Suryavijaya (2025) demonstrated that careful hyperparameter tuning significantly improved C4.5 performance in demographic classification tasks [20]. Third, assigning higher misclassification costs to false negatives would encourage the model to prioritize correct identification of the "YES" class, even at the expense of slightly lower overall accuracy. Fourth, combining multiple decision trees through techniques like Random Forest or boosting could improve robustness and minority class detection. Kusuma and Soetanto (2025) found that ensemble approaches outperformed single decision trees in elderly health classification [4].

Despite the class imbalance challenge, the decision tree structure offers valuable transparency in decision-making processes. The identification of Monthly Income, Health Condition, and Number of Dependents as primary splitting attributes aligns with eligibility criteria specified in Governor Regulation No. 100 of 2019. This interpretability allows sub-district officials to explain classification decisions to applicants, fostering trust and accountability in the assistance distribution process. Damayanti and Swari (2024) emphasize that interpretable models like decision trees are particularly suitable for public sector applications where decision transparency is legally and ethically required [15]. The tree structure enables officials to trace the reasoning behind each classification, facilitating communication with applicants and supporting appeals processes when decisions are contested.

The model's performance compares favorably with related research in social assistance classification. Purba and Siboro (2025) achieved 89.3% accuracy using Naïve Bayes for health insurance assistance classification [5], while Fatmawati and Purnamasari (2024) reported 91.7% accuracy in their Naïve Bayes implementation [14]. The present study's 95.84% accuracy demonstrates that Decision Tree algorithms can achieve superior performance in elderly-specific assistance classification, likely due to their ability to capture non-linear relationships and complex attribute interactions characteristic of multidimensional vulnerability assessments. However, the class imbalance challenge observed in the present study is consistent with findings from Hidayatullah and Prihartono (2025), who encountered similar issues in food assistance recipient clustering and recommended hybrid approaches combining classification and clustering techniques [10]. Future research should explore such hybrid methodologies to address both the classification accuracy and class imbalance challenges simultaneously, potentially integrating clustering methods like K-Means to identify vulnerability segments before applying classification algorithms to each segment separately.

5. Conclusion and Recommendations

Based on the evaluation results, the classification model using the Decision Tree algorithm achieved an accuracy of 95.84% ± 3.43%, with a micro average of 95.85%. This indicates that overall, the model has high capability in distinguishing between social assistance recipients (YES) and non-recipients (NO). However, there is a performance imbalance between classes. The model is far more accurate in classifying the "NO" category with 98.04% recall and 97.56% precision, while the "YES" category only achieved 61.54% recall and 66.67% precision. This indicates that the model tends to predict more cases as "NO" and is less sensitive to the "YES" class. The class imbalance in the dataset, where rejected applicants (212 instances, 94.6%) vastly outnumber approved applicants (12 instances, 5.4%), causes the model to develop a bias toward the majority class. Despite this limitation, the decision tree structure successfully identified Monthly Income, Health Condition, and Number of Dependents as the most influential attributes, aligning with eligibility criteria in Governor Regulation No. 100 of 2019. The model can serve as an effective decision support tool for sub-district officials in beneficiary selection, provided that human oversight is maintained to review borderline cases.

Several recommendations are proposed to improve model performance and practical implementation. First, future research should implement data balancing techniques such as SMOTE or undersampling to address the class imbalance problem and improve sensitivity to the "YES" class. Second, hyperparameter optimization through grid search methods should be conducted to fine-tune model parameters. Third, exploring ensemble methods like Random Forest or AdaBoost may enhance robustness and minority class detection capability. For data collection, expanding the dataset by including data from multiple neighborhoods would increase sample size, particularly for the minority class. Incorporating additional attributes such as disability status and chronic disease types could provide richer information for classification decisions. For practical implementation, the model should be deployed as a decision support system that augments rather than replaces human judgment, with officials retaining final authority over eligibility determinations. All cases predicted as "NO" with low confidence scores should be flagged for manual review to minimize false negatives. Regular model retraining with updated data should be conducted quarterly to maintain accuracy as demographic conditions evolve.

References

[1] Zuhendra, M. I., & Hidayat, R. (2024). Penerapan data mining untuk klasterisasi tingkat kemiskinan berdasarkan data terpadu kesejahteraan sosial (DTKS). *SKANIKA: Sistem Komputer dan Teknik Informatika, 7*(1), 32-41. https://doi.org/10.36080/skanika.v7i1.3149

[2] Daud, M., & Juita, R. (2025). Penerapan metode algoritma C4.5 untuk klasifikasi kelayakan penerima program bantuan pada Dinas Sosial Kabupaten Manokwari. *DECODE: Jurnal Pendidikan Teknologi Informasi, 5*(1), 271-278. https://doi.org/10.51454/decode.v5i1.1057

[3] Pirmansyah, F., & Wahyudi, T. (2023). Implementasi data mining menggunakan algoritma C4.5 untuk prediksi evaluasi anggota satuan pengamanan (Studi kasus PT. YIMM Pulogadung). *Jurnal Indonesia: Manajemen Informatika dan Komunikasi, 4*(3), 1566-1580. https://doi.org/10.35870/jimik.v4i3.370

[4] Kusuma, A. C., & Soetanto, H. (2025). Klasifikasi kesehatan lansia menggunakan perbandingan metode K-Nearest Neighbor, Naïve Bayes dan Decision Tree. *JURNAL INOVTEK POLBENG - Seri Informatika, 10*(1), 559-570. https://doi.org/10.35314/2q4a1524

[5] Purba, D. E. R., & Siboro, Y. O. (2025). Analisa data mining dengan Naïve Bayes untuk klasifikasi kelayakan penerima bantuan iuran jaminan kesehatan pada Desa Durian. *Jatilima: Jurnal Multimedia dan Teknologi Informasi, 7*(2). https://doi.org/10.54209/jatilima.v7i02.1279

[6] Santosa, G. A. D., & Santoso, B. (2024). Decision support system for determining BLT-DD recipients using the C4.5 algorithm. *MALCOM: Indonesian Journal of Machine Learning and Computer Science, 4*(3), 1177-1188. https://doi.org/10.57152/malcom.v4i3.157

[7] Yunita, S., & Bachtiar, L. (2024). Model klasterisasi data penduduk menggunakan algoritma K-Means untuk mengetahui prioritas penerima bantuan sosial di Desa Bapinang Hulu. *Jurnal TEKINKOM, 7*(2).

[8] Kinanti, R., & Jasmir. (2024). Penerapan metode clustering K-Means untuk menentukan prioritas penerima bantuan program beras untuk rakyat miskin (RASKIN) Studi kasus: Kecamatan Siulak. *AKAKOM - Jurnal Informatika dan Rekayasa Komputer, 4*(2). https://doi.org/10.33998/jakakom.v4i2.1933

[9] Hamu, M. A. M., & Talakua, A. C. (2024). Optimalisasi manajemen bantuan sosial dengan implementasi data mining menggunakan algoritma K-Means clustering. *Contar: Jurnal Ilmu Komputer Universitas Katolik Widya Mandira, 1*(1), 07-12.

[10] Hidayatullah, N. A., & Prihartono, W. (2025). Clustering penerima bantuan pangan berbasis algoritma K-Means untuk meningkatkan efektivitas program sosial di Kota/Kabupaten Cirebon. *JITET (Jurnal Informatika dan Teknik Elektro Terapan), 13*(1). https://doi.org/10.23960/jitet.v13i1.5692

[11] Huriah, D. A., Nuris, N. D., Usaha, B., Mining, D., Naive, A., & Dalam, B. (2023). Klasifikasi penerima bantuan sosial umkm menggunakan algoritma naïve bayes. *Jurnal Mahasiswa Teknik Informatika*, *7*(1), 360-365.

[12] Surahman, A., & Hayati, U. (2023). Implementasi Algoritma Naïve Bayes Untuk Prediksi Penerima Bantuan Sosial. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(1), 347-352. https://doi.org/10.36040/jati.v7i1.6302

[13] Rohim, A. N. A., & Purnamasari, A. I. (2024). Komparasi efektivitas algoritma C4.5 dan Naïve Bayes untuk menentukan kelayakan penerima manfaat Program Keluarga Harapan (Studi kasus: Kecamatan Cicalengka Kabupaten Bandung). *JATI (Jurnal Mahasiswa Teknik Informatika), 8*(2), 2355-2362.

[14] Fatmawati, A., Purnamasari, A. I., & Ali, I. (2024). Implementasi Naïve Bayes Untuk Klasifikasi Kelayakan Penerima Bantuan Sosial. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(1), 745-750.

[15] Damayanti, A., & Swari, M. H. P. (2024). Sistem informasi administrasi kependudukan dengan algoritma C5.0 untuk klasifikasi. *JATI (Jurnal Mahasiswa Teknik Informatika), 8*(1).

[16] Nurliana, E., Irawan, B., & Bahtiar, A. (2024). Implementasi data mining algoritma K-Means untuk klasifikasi penduduk miskin berdasarkan tingkat kemiskinan di Jawa Barat. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(1), 1116-1122. https://doi.org/10.36040/jati.v8i1.8883.

[17] Hutagaol, V., & Anggraeni, D. (2025). Penerapan data mining untuk clustering masyarakat kurang mampu pada Desa Rawang Pasar VI dengan algoritma K-Means. *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika), 10*(1), 304-312. https://doi.org/10.29100/jipi.v10i1.5741

[18] Maori, N. A., & Evanita, E. (2023). Metode elbow dalam optimasi jumlah cluster pada k-means clustering. *Simetris: Jurnal Teknik Mesin, Elektro dan Ilmu Komputer*, *14*(2), 277-288.

[19] Aria, R. R. A., Susilowati, S., & Rahadjeng, I. R. (2023). Data Mining Menentukan Cluster Penerima Program Bantuan dengan Metode K-Means. *REMIK: Riset dan E-Jurnal Manajemen Informatika Komputer*, *7*(1), 291-300. https://doi.org/10.33395/remik.v7i1.12030

[20] Yuliani, T. D., & Suryavijaya, D. (2025). Penerapan algoritma C4.5 dalam klasifikasi dominasi jenis kelamin pada data penduduk Kota Depok berdasarkan kelompok usia di tahun 2019-2023. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *9*(2), 2407-2415.

[21] Sutisna, S., & Yuniar, M. N. (2023). Klasifikasi kualitas air bersih menggunakan METODE Naïve baiyes. *Jurnal Sains dan Teknologi*, *5*(1), 243-246. https://doi.org/10.55338/saintek.v5i1.1383

[22] Surapati, U., & Jannah, M. (2024). Penerapan data mining menggunakan metode K-Means untuk mengetahui minat customer dalam pembelian merchandise Kpop. *Jurnal Sains dan Teknologi, 5*(3), 875-884. https://doi.org/10.55338/saintek.v5i1.2739

[23] Dutt, A., Ismail, M. A., Herawan, T., & Hashem, I. A. (2024). Partition-based clustering algorithms applied to mixed data for educational data mining: a survey from 1971 to 2024. *IEEE Access*, *12*, 172923-172942. https://doi.org/10.1109/ACCESS.2024.3496929

[24] Liang, R., Huang, C., Zhang, C., Li, B., Saydam, S., & Canbulat, I. (2023). Exploring the fusion potentials of data visualization and data analytics in the process of mining digitalization. *IEEE Access*, *11*, 40608-40628. https://doi.org/10.1007/978-3-031-54895-6_22

[25] Dhanushkodi, K., Bala, A., Kodipyaka, N., & Shreyas, V. (2024). Customer Behavior Analysis and Predictive Modeling in Supermarket Retail: A Comprehensive Data Mining Approach. *Ieee Access*, *13*, 2945-2957. https://doi.org/10.1109/ACCESS.2024.3407151.