Application of K-Means and Decision Tree for Disease Prediction Using Data Mining Approach

[1]Riah Ukur Ginting, [2]Fernando H Sinaga, [3]Rianto Sitanggang, [4]Ivan Elisabeth Purba,[5]Aprima A Matondang
[123] Information Systems, Sari Mutiara Indonesia University, Medan
[4] Management, Sari Mutiara Indonesia University, Medan
[5] Electrical Engineering, Medan State Polytechnic, Medan
[1]riahukur@sari-mutiara.ac.id, [2]fernandosinaga2002@gmail.com, [3]rianto.sitanggang79@gmail.com, [4]ivanelisabeth.purba@sari-mutiara.ac.id,[5]aprimaamatondang@polmed.ac.id

Abtract - This study aims to analyze the distribution patterns of patient diseases using a data mining approach at UPTD Puskesmas Pakkat. The dataset consists of secondary data from 4,633 patients collected between January 2022 and December 2023, obtained from digital medical records, with variables including age, gender, and 22 disease diagnosis categories. The K-Means Clustering method was employed to identify disease grouping patterns based on patient characteristics. The optimal number of clusters was determined using the Silhouette Score, with the best value of 0.5556 at K=6. Cluster quality was further evaluated using the Davies-Bouldin Index (DBI) with a value of 0.6722, indicating good cluster separation. To support the classification process, the Decision Tree algorithm was applied to predict cluster membership for new patient data. Model evaluation was conducted using a train-test split scheme and k-fold cross-validation to enhance reliability and minimize the risk of overfitting. The results indicate distinct disease patterns across age groups, where infectious diseases such as acute respiratory infections (ARI) and diarrhea dominate in children, while non-communicable diseases such as hypertension and diabetes are more prevalent among adults and the elderly. This study contributes by integrating clustering and classification methods and provides data-driven epidemiological insights that can support decision-making in primary healthcare services.

Keywords — K-Means clustering, Decision Tree, data mining, disease pattern analysis, primary healthcare

I. Introduction

Data mining is a process of extracting valuable patterns and knowledge from large-scale datasets to support more effective and data-driven decision-making, particularly in the healthcare sector where comprehensive analysis of patient data is essential for improving service quality and outcomes [1][2][3]. The increasing availability of electronic medical records has enabled healthcare institutions to leverage advanced analytical techniques to identify disease patterns, predict health risks, and optimize resource allocation [4][5]. Community Health Centers (Puskesmas), as primary healthcare facilities, play a strategic role in promotive, preventive, curative, and rehabilitative services, and routinely collect patient visit data as part of health information systems [6]. However, in many cases, these data are primarily utilized for administrative reporting purposes, leading to the underutilization of their potential for deeper analytical insights and evidence-based decision-making [7][8]. Various studies have demonstrated the effectiveness of clustering techniques, particularly K-Means, in identifying hidden structures and grouping patterns within healthcare datasets due to its computational efficiency and scalability [9][10][11]. Similarly, supervised learning methods such as Decision Tree have been widely applied in healthcare data classification because of their interpretability and ability to generate decision rules that are easily understood by practitioners [12][13]. Despite these advantages, most existing studies employ clustering and classification techniques independently, without integrating both approaches into a unified analytical framework [14]. Moreover, limited attention has been given to epidemiological interpretation of the results, which is crucial for translating analytical findings into actionable healthcare policies and interventions [15]. In addition, prior research often lacks detailed descriptions of dataset characteristics, including data imbalance, feature representation, and preprocessing strategies, as well as the application of robust validation techniques such as k-fold cross-validation, which are essential to ensure model generalizability and reliability. These limitations highlight a research gap in the integration of unsupervised and supervised learning methods for primary healthcare data analysis, along with insufficient emphasis on comprehensive model evaluation and epidemiological interpretation [16]. To address these gaps, this study proposes an integrative approach by combining K-Means Clustering and Decision Tree algorithms within a unified analytical framework to identify disease distribution patterns based on patient demographic characteristics. This study utilizes a dataset of 4,633 patient records with 22 disease categories obtained from digital medical records at a Community Health Center, providing a contextual and representative dataset for primary healthcare analysis. Furthermore, model evaluation is conducted using train-test split and k-fold cross-validation techniques to minimize overfitting and enhance model robustness. The main contributions of this study include the integration of unsupervised and supervised learning methods, the identification of disease distribution patterns based on real-world primary healthcare data, and the provision of data-driven epidemiological insights that can support decision-making in healthcare planning and policy development. Therefore, this study is expected to contribute to the advancement of data-

driven health information systems and improve the effectiveness of primary healthcare services [17].

II. Research Method

A. *Method*

This study employs a data mining approach by integrating K-Means Clustering and Decision Tree algorithms to analyze disease distribution patterns based on patient characteristics. The research framework consists of several stages, including data collection, data preprocessing, clustering using K-Means, classification using Decision Tree, and model evaluation. The dataset used in this study consists of 4,633 patient records obtained from UPTD Puskesmas Pakkat, covering the period from January 2022 to December 2023. The data were collected from digital medical records in spreadsheet (CSV) format. The variables used in this study include patient age, gender, and 22 categories of disease diagnoses [18]. The dataset shows a certain level of imbalance, where some diseases appear more frequently than others, which is a common characteristic in healthcare data [19].

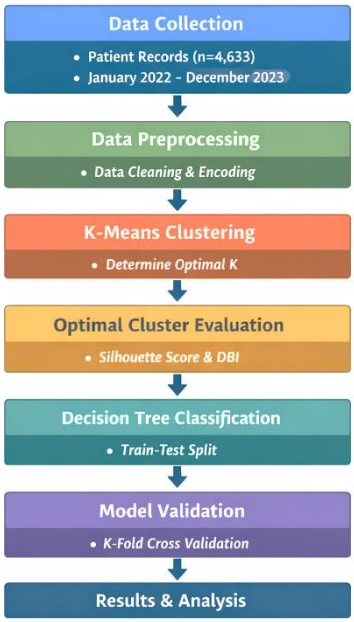

Figure 1. Steps In Research Methode

B. *Data Preprocesising*

Data preprocessing was conducted to ensure data quality and consistency. This stage includes data cleaning to remove inconsistencies and validate age ranges, as well as feature selection to retain relevant variables. Categorical variables such as gender and disease type were transformed into numerical values using label encoding. Although label encoding is simple and efficient, it may introduce ordinal bias; therefore, its use is justified due to the limited number of categorical features, while alternative encoding methods such as one-hot encoding are suggested for future research [20]. Furthermore, numerical normalization was applied using Min-Max Scaling to transform all variables into a uniform range between 0 and 1, thereby reducing the influence of different scales across variables [21].

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

C. *Implementation of K-Means Clustering*

The K-Means Clustering algorithm was applied to group patients based on similarity in demographic and clinical characteristics. The optimal number of clusters (K) was determined using the Silhouette Score by evaluating values of K ranging from 2 to 10. The highest Silhouette Score of 0.5556 was achieved at K=6, indicating the best cluster configuration. Additionally, cluster quality was evaluated using the Davies-Bouldin Index (DBI), resulting in a value of 0.6722, which indicates good cluster separation. The selection of K= 6 is based not only on the highest Silhouette Score but also on the interpretability of the resulting clusters in the context of disease distribution [23][24][25].

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (2)$$

Following the clustering process, the Decision Tree algorithm was employed to classify patient data into the generated clusters. The dataset was divided into training and testing sets using an 80:20 ratio [26][27]. The Decision Tree model was trained using the information gain criterion to determine the optimal splitting attributes. Model performance was evaluated using several metrics, including accuracy, precision, recall, and F1-score. The accuracy metric is calculated as (TP + TN) / Total, ensuring correct formulation of evaluation measures [29][30].

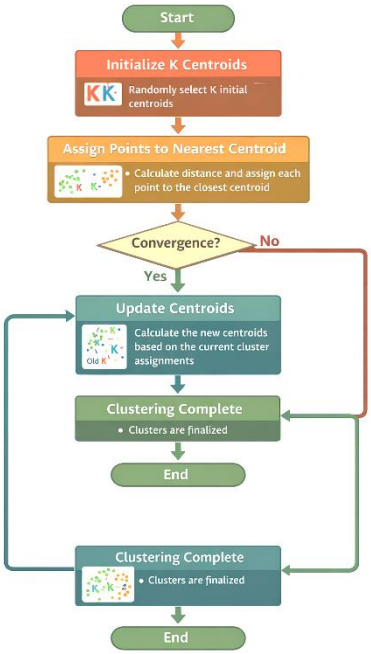

Figure 2. K – Means Clustering flowchart

D. Implementasi Decision Tree

The Decision Tree algorithm was implemented as a supervised learning method to classify patient data into clusters generated from the K-Means Clustering process. The input features used in this model include patient age, gender, and disease category, while the cluster labels obtained from the clustering stage were used as the target variable. Prior to model training, the dataset was divided into training and testing sets using an 80:20 ratio [31]. The training set was used to build the classification model, while the testing set was used to evaluate its performance. The Decision Tree model was constructed using the information gain criterion to select the optimal splitting attributes at each node, aiming to maximize the reduction of entropy [32]. To control model complexity and prevent overfitting, several parameters were defined, including a maximum tree depth of 5 and a maximum number of leaf nodes of 8. These parameters were selected to balance model interpretability and performance. Model evaluation was conducted using multiple performance metrics, including accuracy, precision, recall, and F1-score [33]. The accuracy metric is calculated using the formula (TP + TN) / Total, while precision, recall, and F1-score were used to provide a more comprehensive evaluation of classification performance. To improve the reliability and generalization capability of the model, k-fold cross-validation (k=5) was applied. This validation technique allows the dataset to be partitioned into multiple subsets, where each subset is used alternately as training and validation data, thereby reducing bias and improving robustness. It is important to note that the classification labels used in this study are derived from clustering results. Therefore, the evaluation results should be interpreted cautiously, as high accuracy values may indicate the model's ability to replicate clustering patterns rather than true predictive performance.

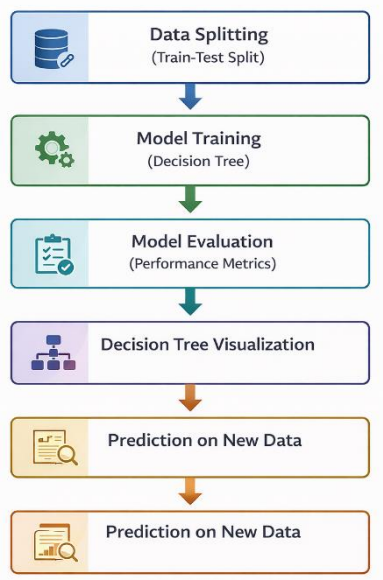

Figure 3. Decision Tree Implementation Flowchart

III. Results and Discussion

1. Results

This section presents the results of clustering and classification processes, followed by an in-depth discussion that integrates technical evaluation and epidemiological interpretation. The K-Means clustering algorithm was applied to group patient data based on age, gender, and disease category. The optimal number of clusters was determined using the Silhouette Score, with the highest value of 0.5556 obtained at K=6. This value indicates a moderate clustering structure with acceptable cohesion and separation. In addition, cluster quality was evaluated using the Davies-Bouldin Index (DBI), resulting in a value of 0.6722, which confirms that the clusters are well-separated and relatively compact. The selection of K=6 is not only based on the highest Silhouette Score but also on the interpretability of the resulting clusters in representing meaningful disease groupings [34].

Table 1. K-Means Clustering Evaluation Results

Parameter	Value	Interpretation
Optimal Number of Clusters	K = 6	Best clustering configuration
Silhouette Score	0.5556	Moderate cluster separation
Davies-Bouldin Index (DBI)	0.6722	Good cluster compactness and separation

The distribution of patients across the six clusters reveals distinct patterns associated with demographic characteristics. Clusters dominated by younger age groups are primarily associated with infectious diseases such as acute respiratory infections (ARI), fever, and diarrhea. In contrast, clusters representing older age groups are characterized by non-communicable diseases such as hypertension, diabetes, and rheumatism. This finding is consistent with epidemiological evidence, where infectious diseases are more prevalent among children, while chronic diseases increase with age. From a healthcare perspective, these results highlight the importance of targeted interventions, such as strengthening immunization and nutrition programs for children, and improving screening and management of non-communicable diseases among adults and the elderly.

Table 2. Distribution of Data in Each Cluster

Cluster	Number of Data	Percentage	Dominant Characteristics
1	906	19.6%	Children – ISPA, fever
2	707	15.3%	Elderly – hypertension
3	1391	30.0%	Adults – hypertension, dyspepsia
4	506	10.9%	Elderly – diabetes, gout
5	589	12.7%	Children – diarrhea, allergy
6	534	11.5%	Children – ISPA, tonsillitis

The Decision Tree algorithm was implemented to classify patient data into clusters generated from the K-Means process. The model achieved a high accuracy of 99.89%, along with high precision, recall, and F1-score values. However, this performance should be interpreted critically. Since the classification labels are derived from clustering results, the model essentially learns to reproduce the clustering structure, which may lead to overfitting. Therefore, the reported accuracy does not fully reflect the model's predictive capability for unseen real-world data.

Table 3. Decision Tree Model Evaluation

Matric	Value
Accuracy	99.89%
Precision	99.87%
Recall	99.88%
F1-Score	99.88%

Table 4. Example of Decision Tree Rules

Rule	Condition	Predicted Cluster	Interpretation
1	Age ≤ 11 AND Disease = ISPA	Cluster 1	Pediatric infectious diseases
2	Age ≤ 11 AND Disease = Diarrhea	Cluster 5	Child digestive disorders
3	Age > 45 AND Disease = Hypertension	Cluster 2	Chronic disease in elderly
4	Age > 45 AND Disease = Diabetes	Cluster 4	Metabolic disorders
5	Age 26–45 AND Disease = Dyspepsia	Cluster 3	Adult digestive conditions

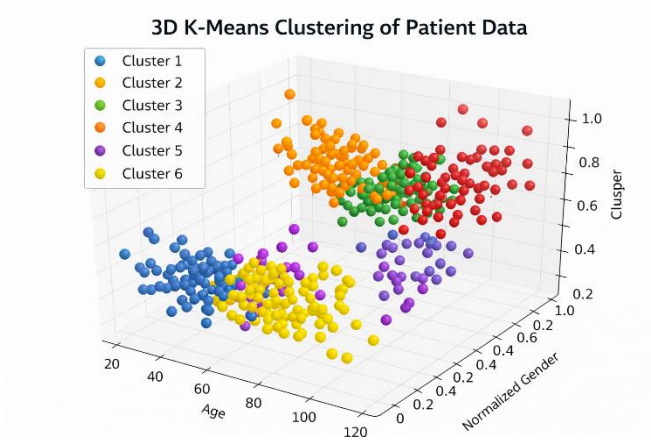

Figure 4. 3D Visualization of K-Means Clustering

2. Discussion

To address this issue, k-fold cross-validation (k=5) was applied to evaluate the model's stability and generalization performance. The results indicate relatively consistent performance across folds, suggesting that the model is stable within the dataset used. Nevertheless, further validation using independent datasets is necessary to confirm its generalizability. In addition to performance metrics, the Decision Tree model provides interpretable classification rules. These rules indicate that age is a dominant factor influencing cluster membership, followed by disease type. For example, younger patients diagnosed with ARI or diarrhea are consistently classified into clusters associated with pediatric conditions, while older patients with hypertension or diabetes are grouped into clusters representing chronic disease patterns. This interpretability is a key advantage of the Decision Tree algorithm, as it allows healthcare practitioners to understand and utilize the model in practical decision-making contexts.

Compared to previous studies that applied clustering or classification independently, this study demonstrates the added value of integrating both approaches. The K-Means algorithm effectively identifies hidden patterns in the data, while the Decision Tree provides a mechanism for classifying new patient data based on those patterns. This integrated approach enhances both analytical capability and practical applicability in healthcare settings. However, this study has several limitations. The use of limited variables, namely age, gender, and disease category, may restrict the model's ability to capture more complex clinical relationships. Additionally, the use of label encoding may introduce bias in distance-based methods such as K-Means. Furthermore, the absence of comparison with other classification algorithms limits the ability to evaluate relative performance. Future studies are recommended to incorporate additional clinical variables, apply alternative encoding techniques, and compare multiple machine learning algorithms such as Random Forest and Naïve Bayes to obtain more robust results [35].

IV. Conslusion

1. This study successfully integrates K-Means Clustering and Decision Tree to analyze disease distribution patterns in primary healthcare data.
2. The optimal clustering result is achieved at K=6, with a Silhouette Score of 0.5556 and DBI of 0.6722, indicating a reasonably good clustering structure.
3. The findings show that infectious diseases dominate younger age groups, while non-communicable diseases are more prevalent in older populations.
4. The Decision Tree model achieves high accuracy, but this should be interpreted cautiously due to potential overfitting caused by the use of cluster labels.
5. The integration of clustering and classification methods provides meaningful insights for data-driven healthcare decision-making.

V. Reference

[1] M. B. Fajri and S. D. Purnamasari, "Klasterisasi Pola Penyebaran Penyakit Pasien Berdasarkan Usia Pasien Menggunakan K-Means Clustering," Journal of Information Technology Ampera, vol. 3, no. 3, pp. 317–334, 2022.

[2] A. A. N. Mostafa and H. E. A. Mahmoud, "Review of Data Mining Concept and its Techniques," International Journal of Academic Research in Business and Social Sciences, vol. 12, no. 6, pp. 611–619, 2022, doi: 10.6007/ijarbss/v12-i6/13135.

[3] S. Maulia, B. S. Ginting, and A. Sihombing, "Implementasi Data Mining Pengelompokan Jenis Penyakit Pasien Menggunakan Metode Clustering (Studi Kasus : Puskesmas Sambirejo)," Jurnal Informatika Kaputama (JIK), vol. 5, no. 1, pp. 71–80, 2021, doi: 10.59697/jik.v5i1.304.

[4] A. Lingga, "Penerapan Algoritma K-Means Clustering Untuk Klasterisasi Penyakit Pasien," vol. 1, no. 3, pp. 93–102, 2024.

[5] I. Kanedi and E. Suryana, "Penerapan Metode K-Means Clustering Dalam Pengelompokan Data Pasien Rawat Inap Peserta BPJS Di Rumah Sakit Umum Daerah Kabupaten Kaur," vol. 20, no. 2, pp. 493–500, 2024.

[6] N. Cholifatul Izza and A. I. Rizmayanti, "Analisis Rekam Medis dengan Metode Data Mining untuk Memprediksi Faktor Risiko Stunting dalam Kesehatan Masyarakat," Jurnal Manajemen Informasi dan Administrasi Kesehatan (JMIAK), vol. 7, no. 1, pp. 1–9, 2024, doi: 10.32585/jmiak.v7i1.5192.

[7] R. Bayu Prasetyo, Y. Agus Pranoto, and R. Primaswara Prasetya, "Implementasi Data Mining Menggunakan Algoritma K-Means Clustering Penyakit Pasien Rawat Jalan Pada Klinik Dr. Atirah Desa Sioyong, Sulteng," JATI (Jurnal Mahasiswa Teknik Informatika), vol. 7, no. 4, pp. 2144–2151, 2023, doi: 10.36040/jati.v7i4.7419.

[8] Syaiful Hasan Abdullah and Zaehol Fatah, "Analisis Produksi Cabai Rawit Indonesia Menggunakan Algoritma K-Means Clustering," Jurnal Ilmiah Sains Teknologi Dan Informasi, vol. 3, no. 1, pp. 66–74, 2024, doi: 10.59024/jiti.v3i1.1024.

[9] J. Jumadi, Y. Yupianti, and D. Sartika, "Pengolahan Citra Digital Untuk Identifikasi Objek Menggunakan Metode Hierarchical Agglomerative Clustering," JST (Jurnal Sains dan Teknologi), vol. 10, no. 2, pp. 148–156, 2021, doi: 10.23887/jstundiksha.v10i2.33636.

[10] M. R. Nugroho, I. E. Hendrawan, and P. P. Purwantoro, "Penerapan Algoritma K-Means Untuk Klasterisasi Data Obat Pada Rumah Sakit ASRI," Nuansa Informatika, vol. 16, no. 1, pp. 125–133, 2022, doi: 10.25134/nuansa.v16i1.5294.

[11] M. F. Al Halik and L. Septiana, "Analisa Data Untuk Prediksi Daerah Rawan Bencana Alam Di Jawa Barat Menggunakan Algoritma K-Means Clustering," Journal of Information System, Applied, Management, Accounting and Research, vol. 6, no. 4, pp. 856–870, 2022, doi: 10.52362/jisamar.v6i4.939.

[12] L. Pantanowitz et al., "Nongenerative Artificial Intelligence in Medicine: Advancements and Applications in Supervised and Unsupervised Machine Learning," Modern Pathology, vol. 38, no. 3, p. 100680, 2025, doi: 10.1016/j.modpat.2024.100680.

[13] E. Helmud, E. Helmud, F. Fitriyani, and P. Romadiana, "Classification Comparison Performance of Supervised Machine Learning Random Forest and Decision Tree Algorithms Using Confusion Matrix," Jurnal Sisfokom (Sistem Informasi dan Komputer), vol. 13, no. 1, pp. 92–97, 2024, doi: 10.32736/sisfokom.v13i1.1985.

[14] S. Pokhrel, "Klasterisasi Penyakit Menular Di Indonesia Menggunakan Metode K-Means Clustering," Αγαη, vol. 15, no. 1, pp. 37–48, 2024.

[15] R. Wang, J. Xu, and M. He, "Blood leukocyte-based clusters in patients with traumatic brain injury," Front. Immunol., vol. 15, no. January, pp. 1–8, 2024, doi: 10.3389/fimmu.2024.1504668.

[16] Dwita Elisa Sinaga, Agus Perdana Windarto, and Rizki Alfadillah Nasution, "Analisis Data Mining Algoritma Decision Tree Pada Prediksi Persediaan Obat (Studi Kasus : Apotek Franch Farma)," KLIK: Kajian Ilmiah Informatika dan Komputer, vol. 2, no. 4, pp. 123–131, 2022, doi: 10.30865/klik.v2i4.328.

[17] O. P. Moerdyanto and I. K. D. Nuryana, "Prediksi Kelulusan Tepat Waktu Menggunakan Pendekatan Pohon Keputusan Algoritma Decision Tree," Journal of Informatics and Computer Science, vol. 05, no. 1, pp. 90–96, 2023.

[18] F. Saptawan et al., "Prediksi epidemiologi penyakit tidak menular menggunakan algoritma random forest pada puskesmas," vol. XIII, no. 2, pp. 192–201, 2024.

[19] E. A. Herdiaman, A. Sudiarjo, M. Hikmatyar, T. Informatika, U. Perjuangan, and J. Barat, "Rsud Menggunakan Metode K-Means," vol. 12, no. 3, 2024.

[20] M. Sholeh, D. Andayati, and Rr. Y. Rachmawati, "Data Mining Model Klasifikasi Menggunakan Algoritma K-Nearest Neighbor Dengan Normalisasi Untuk Prediksi Penyakit Diabetes," TeIKa, vol. 12, no. 02, pp. 77–87, 2022, doi: 10.36342/teika.v12i02.2911.

[21] W. A. Prastyabudi, A. N. Alifah, and A. Nurdin, "Segmenting the Higher Education Market: An Analysis of Admissions Data Using K-Means Clustering," Procedia Comput. Sci., vol. 234, no. 2023, pp. 96–105, 2024, doi: 10.1016/j.procs.2024.02.156.

[22] M. Orisa, "Optimasi Cluster pada Algoritma K-Means," Prosiding SENIATI, vol. 6, no. 2, pp. 430–437, 2022, doi: 10.36040/seniati.v6i2.5034.

[23] N. T. M. Sagala and A. A. S. Gunawan, "Discovering the Optimal Number of Crime Cluster Using Elbow, Silhouette, Gap Statistics, and NbClust Methods," ComTech: Computer, Mathematics and Engineering Applications, vol. 13, no. 1, pp. 1–10, 2022, doi:

10.21512/comtech.v13i1.7270.
[24] R. Ishak, “Optimasi K-Means pada Clustering Penyakit Ibu Hamil Menggunakan Random Forest Optimization of K-Means in Disease Clustering of Pregnant Women Using Random Forest,” vol. 7, pp. 41–47, 2024.
[25] S. E. A. Buananta, M. A. Ahmad, J. Mahmood, and P. Paradise, “Identification of Evaluation Results in E-Banking Services Transaction for Product Recommendation using the BIRCH and Davies Bouldin Index Method,” Jurnal Infotel, vol. 16, no. 2, pp. 427–440, 2024, doi: 10.20895/infotel.v16i2.1116.
[26] I. T. Umagapi, B. Umaternate, S. Komputer, P. Pasca Sarjana Universitas Handayani, B. Kepegawaian Daerah Kabupaten Pulau Morotai, and B. Riset dan Inovasi, “Uji Kinerja K-Means Clustering Menggunakan Davies-Bouldin Index Pada Pengelompokan Data Prestasi Siswa,” Seminar Nasional Sistem Informasi dan Teknologi (SISFOTEK), vol. 7, no. 1, pp. 303–308, 2023.
[27] H. Sulastri, H. Mubarok, and S. S. Iasha, “Implementasi Algoritma Machine Learning Untuk Penentuan Cluster Status Gizi Balita,” Jurnal Rekayasa Teknologi Informasi (JURTI), vol. 5, no. 2, p. 184, 2021, doi: 10.30872/jurti.v5i2.6779.
[28] M. A. Vidi et al., “Proceeding National Conference of Research and Community Service Sisi Indonesia,” 2025.
[29] T. Novianti, S. A. Mandati, and E. K. Andana, “Peningkatan Evaluasi Risiko Kredit Menggunakan Decision Tree C 4.5,” Journal of Manufacturing in Industrial Engineering & Technology, vol. 2, no. 2, pp. 1–9, 2023, doi: 10.30651/mine-tech.v2i2.21749.
[30] A. W. Safitri, S. N. Anisa, A. Al-habsyi, T. Elektro, and P. N. Jakarta, “Implementasi Algoritma Decision Tree dalam rangka Peningkatan Efisiensi Energi Penggunaan Beban Listrik dalam Ruangan Abstrak SNIV : SEMINAR NASIONAL INOVASI VOKASI,” vol. 3, no. 1, pp. 506–515, 2022.
[31] M. Ferdyandi, N. Y. Setiawan, and F. Abdurrachman Bachtiar, “Prediksi Potensi Penjualan Makanan Beku Berdasarkan Ulasan Pengguna Shopee Menggunakan Metode Decision Tree Algoritma C4.5 Dan Random Forest (Studi Kasus Dapur Lilis),” Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, vol. 6, no. 2, pp. 588–596, 2022.
[32] S. I. R. Adi, B. Bakkara, K. A. Zega, F. N. Vielita, and N. A. Rakhmawati, “Analisis Sentimen Masyarakat Terhadap Progress Ikn Menggunakan Model Decision Tree,” JIKA (Jurnal Informatika), vol. 8, no. 1, p. 57, 2024, doi: 10.31000/jika.v8i1.9803.
[33] P. Studi, T. Informasi, S. T. M. Ik, and P. Nusantara, “Pemodelan Classification and Regression Tree (CART) Pada Klasifikasi Gaya Hidup Sehat Menggunakan Pendekatan User-Based Classification,” vol. 4, pp. 1028–1036, 2025.
[34] P. Di and D. Pasawahan, “Penerapan Algoritma K-Means Dalam Analisis Data Kependudukan Untuk Optimalisasi,” Vol. 13, No. 1, Pp. 439–445, 2025.
[35] I. Hamzah, Muhammad Iqbal, and Rian Farta Wijaya, “Klasifikasi Jenis Penyakit dengan Algoritma Decision Tree Menggunakan Rapid Miner,” Jurnal Nasional Teknologi Komputer, vol. 4, no. 1, pp. 34–39, 2024, doi: 10.61306/jnastek.v4i1.127.