



Perbandingan Algoritma C4.5 dengan C4.5+Particle Swarm Optimization untuk Klasifikasi Angkatan Kerja

Devy Safira¹, Mustakim²

^{1,2}Program Studi Sistem Informasi, Fakultas Sains dan Teknologi UIN Sultan Syarif Kasim Riau

^{1,2}Puzzle Research Data Tecnology, Fakultas Sains dan Teknologi UIN Sultan Syarif Kasim Riau

Jl. HR Soebrantas KM.18 Panam Pekanbaru - Riau

Email: ¹11753202034@students.uin-suska.ac.id, ²mustakim@uin-suska.ac.id

[1] Abstrak

Dalam suatu dataset yang besar, data mining merupakan sebuah proses penyelesaian yang menghasilkan beberapa pola baru menjadi pengetahuan yang berguna. Algoritma yang sering dipakai dalam machine learning salah satunya adalah C4.5. Algoritma ini terkenal sangat kuat dalam melakukan klasifikasi, namun masih memiliki beberapa kelemahan seperti sering terjadinya overlapping dan overfitting data yang membuat data menjadi tidak relevan sehingga dapat mengurangi tingkat akurasi dari algoritma. Untuk menangani ini dibutuhkannya seleksi atribut yang dapat mengidentifikasi atribut yang relevan, tanpa mengurangi akurasi dari algoritma itu sendiri. Algoritma optimasi Particle Swarm Optimization (PSO) merupakan salah satu algoritma yang mampu digunakan sebagai seleksi atribut. Keuntungan dari PSO ini mudah diterapkan, efisien dalam perhitungan dan memiliki konsep yang sederhana dibandingkan dengan teknik optimasi lainnya. Jumlah dataset yang akan digunakan pada penelitian ini sebanyak 2.518 dataset dimana sebelumnya dilakukan pembagian dataset dengan menggunakan algoritma K-Means dan K-Medoid. Atribut yang digunakan sebanyak 4 atribut yaitu, jenis kelamin, umur, tingkat produktivitas dan pendidikan terakhir. Dari penelitian ini, diperoleh hasil bahwa akurasi yang diberikan oleh C4.5 yang telah dioptimasi menggunakan algoritma Particle Swarm Optimization (PSO) terbukti lebih tinggi dibandingkan menggunakan algoritma C4.5 saja. Dimana algoritma C4.5+PSO memiliki akurasi sebesar 66,80% sedangkan algoritma C4.5 memiliki akurasi sebesar 76,32%.

Kata kunci: C4.5, K-Means, K-Medoid, Particle Swarm Optimization (PSO)

[2] Abstract

In a large dataset, data mining is a solution to arrange new models into useful information. The algorithm is often used in machine learning is C4.5. C4.5 is known to be very strong in classifying, but has several weaknesses such as overlapping and overfitting data which makes the data irrelevant that can reduce the accuracy of the algorithm. To handle this, it is necessary to select an attribute that can identify the relevant attribute without reducing the accuracy of the algorithm itself. The Particle Swarm Optimization (PSO) is an optimization algorithm which one can be used as an attribute selection. The PSO benefit is easy to use, efficient and has a simple concept when to compared of the other optimization techniques. Datasets will be used in this case is 2,518 datasets where previously, dataset was divided using K-Means and K-Medoid algorithms. The attributes used are 4 attributes, namely, gender, age, productivity level and the last education. In

this study, the precision of C4.5 which is optimized by Particle Swarm Optimization (PSO) algorithm is proven to be higher than using the C4.5 algorithm alone. Where the algorithm C4.5+PSO has an precision of 66.80% while the algorithm of C4.5 has an precision of 76.32%.

Keywords: C4.5, K-Means, K-Medoid, Particle Swarm Optimization (PSO)

1. Pendahuluan

Dalam suatu dataset yang besar, data mining merupakan sebuah bentuk proses penyelesaian yang menghasilkan beberapa pola baru menjadi suatu informasi yang berguna [1]. Beberapa teknik penyelesaian yang biasa digunakan dalam data mining, yaitu estimasi, asosiasi, klastering, dan klasifikasi [2]. Klastering biasanya digunakan dalam menemukan sejumlah kelompok data [3], teknik klasifikasi digunakan untuk menganalisis data menggunakan model kelas data untuk memprediksi label [4]. Salah satu dari sekian banyak teknik klasifikasi yang sering digunakan adalah C4.5[5].

Algoritma C4.5 disebut juga sebagai algoritma *decision tree* yang terkenal sangat kuat dalam melakukan klasifikasi, Algoritma ini biasa digunakan untuk mendeteksi hubungan tersembunyi antar variable [6] sehingga dapat membentuk sebuah pohon keputusan yang sederhana, mempunyai akurasi yang bagus, serta efektif dalam menangani atribut, baik diskrit maupun numerik. Namun, Algoritma ini memiliki beberapa kelemahan seperti sering terjadi *overlapping* dan *overfitting* data yang membuat data menjadi tidak relevan [2] sehingga dapat mengurangi tingkat akurasi dari suatu algoritma data mining [5].

Dalam data mining, permasalahan akurasi merupakan permasalahan mendasar dalam penelitian [7]. Biasanya, dataset yang digunakan dalam klasifikasi berisikan data yang noise, redundansi data, serta memiliki atribut yang tidak berguna [8]. Untuk menangani ini dibutuhkannya seleksi atribut yang mengidentifikasi atribut yang relevan tanpa mengurangi akurasi dari algoritma itu sendiri [5]. Beberapa teknik untuk menyeleksi atribut dapat menggunakan teknik *supervised learning*, Principle Component Analysis (PCA) atau beberapa algoritma optimasi seperti Particle Swarm Optimization (PSO) [9]. PSO adalah suatu algoritma optimasi yang dapat digunakan sebagai seleksi atribut. Algoritma ini dapat bekerja lebih baik dibandingkan algoritma genetika lainnya terutama dibidang optimasi [5]. Keuntungan dari PSO ini adalah algoritma ini mudah diterapkan, efisien dalam perhitungan dan memiliki konsep yang sederhana jika dibandingkan dengan algoritma data mining dan teknik optimasi lainnya [10].

Kesalahan dalam mendistribusikan dataset dapat juga dapat mempengaruhi hasil akurasi dari algoritma itu sendiri [9]. Algoritma K-Means dapat mendistribusikan data yang akan mewakili karakteristik kelompoknya [9]. Namun, karena algoritma ini dapat melakukan perubahan distribusi pada dataset, algoritma K-Means menjadi sangat rentan terhadap outlier data [11]. Algoritma K-Medoids merupakan algoritma klastering yang lebih kuat dari pada algoritma *K-Means* Algoritma ini hadir untuk menutupi kekurangan yang ada didalam algoritma K-Means yaitu sensitive terhadap *noise* dan *outlier* [12]. Selain itu, proses dari algoritma K-Medoids ini sendiri tidak bergantung pada urutan masuk dataset [13].

Data mining sangat berguna dan banyak diterapkan di berbagai bidang [14], salah satunya dapat berperan penting dalam keberhasilan suatu perusahaan, dimana dapat mengubah sejumlah data menjadi informasi yang penting dan berguna sehingga dapat membantu para pemangku keputusan untuk mengambil kebijakan yang lebih efisien [15]. Dari data BPS 2020, tingkat partisipasi angkatan kerja di Kota Sawahlunto dari tahun 2018 ke tahun 2019 banyak mengalami penurunan.

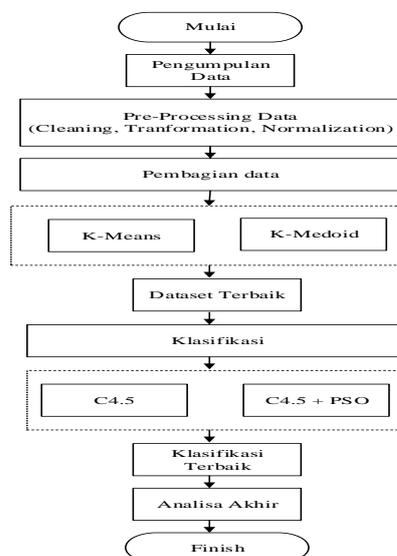
Oleh karena itu, dibutuhkan kebijakan baru dari pemerintah untuk mengatasi ketidakstabilan jumlah partisipasi angkatan kerja di Kota Sawahlunto [16].

Penelitian ini bertujuan untuk menerapkan konsep klasifikasi data mining pada angkatan kerja di Kota Sawahlunto menggunakan algoritma C4.5 yang telah dioptimasi dengan algoritma Particle Swarm Optimazion (PSO). Hal ini bertujuan untuk memperoleh informasi mengenai variabel yang paling berpengaruh. Data yang akan digunakan pada penelitian ini merupakan data penduduk pada tahun 2020 yang berjumlah 2.518 dataset, yang mana sebelum dilakukannya pengklasifikasian akan dilakukan pembagian data terlebih dahulu dengan teknik klastering menggunakan algoritma K-Means dan K-Medoid. Atribut yang digunakan sebanyak 4 atribut yaitu jenis kelamin, umur, tingkat produktivitas dan pendidikan terakhir

Sebagai bahan acuan dalam penelitian, peneliti melakukan studi literatur terkait dengan metode-metode yang sebelumnya pernah menggunakan algoritma PSO diantaranya dilakukan oleh Sundaramurthy dkk [17], melakukan klasifikasi terhadap penyakit *Rheumatoid Arthritis* (RA) menggunakan algoritma C4.5 dengan optimasi PSO dan *Grey Wolf Optimization*, hasil penelitian ini menyimpulkan bahwa pengklasifikasian penyakit RA menggunakan algoritma C4.5 dengan algoritma optimasi, telah secara akurat membedakan penyakit RA dan non-RA. Pada penelitian lain yang dilakukan Pahlevi dkk tahun 2021, yang mana melakukan klasifikasi menggunakan C4.5 yang dioptimasi menggunakan PSO membuktikan bahwa algoritma ini dapat meningkatkan kinerja metode yang digunakan. Selanjutnya pada penelitian yang dilakukan Saputra dan Prasetyo [5] juga menyimpulkan bahwa PSO dapat digunakan sebagai seleksi atribut dan membuat kinerja dari algoritma C4.5 menjadi lebih akurat. Berdasarkan studi literatur ini dapat ditarik kesimpulan bahwa algoritma optimasi PSO dapat meningkatkan akurasi pada algoritma C4.5.

2. Metode Penelitian

Tahapan dari penelitian ini adalah mengumpulkan dataset sebanyak 2.518 *record* data, dimana data ini merupakan data penduduk angkatan kerja pada tahun 2020. Kegiatan pengumpulan data ini dilakukan dengan observasi dengan mendatangi langsung Dinas Penanaman Modal, Pelayanan Terpadu Satu Pintu dan Tenaga Kerja Kota Sawahlunto. Kemudian melakukan preprocessing yang meliputi pembersihan, transformasi dan normalisasi data. Pada tahap ini, dilakukan pembagian dataset menggunakan algoritma K-Means dan K-Medoid yang bertujuan agar dapat menghasilkan dataset terbaik dan seimbang. Penelitian ini akan membandingkan hasil akurasi algoritma C4.5 dengan C4.5 + PSO. Gambar berikut adalah langkah-langkah dari metode penelitian yang akan digunakan.



Gambar 1. Metodologi Penelitian

2.1. C4.5

Algoritma C4.5 merupakan algoritma tingkatan dari ID3 [18]. Algoritma C4.5 ini menggunakan nilai *gain* untuk menentukan atribut dalam setiap *node* yang selanjutnya menjadi penentu dalam pembuatan pohon keputusan [19]. Rumus information gain :

$$\text{Gain}(S,A) = \text{Entropy}(S) - \sum_{i=1}^n \left| \frac{S_i}{S} \right| \times \text{Entropy}(S_i) \quad (1)$$

$$\text{Entropy}(S) = \sum_{i=1}^n -p_i \times \log_2 p_i \quad (2)$$

Dimana S merupakan himpunan dari kasus, |S_i| jumlah pada partisi ke-i, A merupakan atribut, |S| jumlah dalam himpunan, n (jumlah atribut), serta p_i merupakan nilai seimbang dari S_i terhadap S

2.2. Particle Swarm Optimization (PSO)

Algoritma PSO merupakan algoritma optimasi yang berawal dari gagasan perilaku gerombolan ikan dan gerombolan burung yang terbang secara berkelompok dalam mencapai tempat tujuan mereka [17]. Langkah dasar dari algoritma ini yaitu setiap partikel dalam Particle Swarm Optimization memiliki kecenderungan untuk bergerak menuju area pencarian yang lebih baik [10]. Rumus Particle Swarm Optimization (PSO):

$$V_i(t) = V_i(t-1) + c_1 r_1 [Xp_{best\ i} - X_i(t)] + c_2 r_2 [Xg_{best} - X_i(t)] \quad (3)$$

$$X_i(t) = X_i(t-1) + V_i(t) \quad (4)$$

Dimana V_i(t) adalah kecepatan partikel, X_i(t) adalah posisi partikel, c₁ dan c₂ adalah *learning rate*, r₁ dan r₂ adalah bilangan acak, Xp_{best i} adalah posisi terbaik dari partikel i, dan Xg_{best} adalah posisi terbaik global.

2.3. Algoritma K-Means

Algoritma K-Means dapat memisahkan data ke cluster terdekat yang mana data yang memiliki kemiripan dapat membentuk suatu cluster [20]. Langkah awal dari algoritma ini adalah pemilihan nilai k pada masing-masing kelompok, yang selanjutnya dilakukan perbandingan nilai n pada setiap data. Perbandingan nilai n digunakan menggunakan jarak *euclidean* yang diikuti dengan jarak nilai terdekat [21]. Persamaan Euclidean:

$$d(x,y) = ||x-y||^2 \quad (5)$$

Dimana, d adalah jarak data ke titik pusat, x adalah letak data pada atribut, dan y adalah letak titik pusat pada atribut.

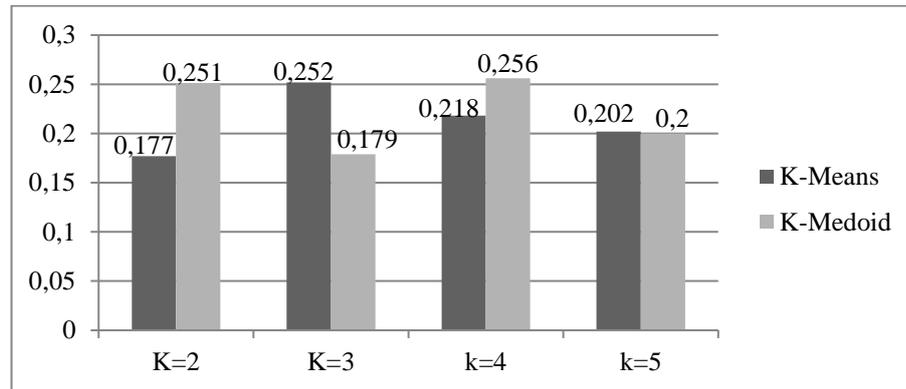
2.4. Algoritma K-Medoid

Algoritma K-Medoid juga dikenal sebagai sebutan algoritma PAM (*Partitioning Around Medoid*) pertama kali dikenalkan oleh Leonard Kaufman serta Peter J. Rousseeuw. K-Medoid ini hadir untuk menangani kekurangan dari K-Means yang *sensitive* terhadap *noise* dan *outlier* [13].

3. Hasil dan Pembahasan

3.1. Pembagian Data

Pembahasan dan hasil diperoleh dari pengklasifikasian data angkatan kerja yang telah dilakukan. Dalam penelitian ini, dilakukan pembagian data terlebih dahulu menggunakan algoritma klastering yaitu K-Means dan K-Medoid. Gambar 2 menjelaskan nilai validitas *cluster* K-Means dan K-Medoid.



Gambar 2. Nilai validitas cluster K-Means dan K-Medoid.

Dari Gambar 2 dapat dilihat bahwa pembagian kluster K-Means terbaik terdapat pada kluster =2 dengan nilai validitas klaster yang terendah, sedangkan pembagian kluster menggunakan algoritma K-Medoid terbaik terdapat pada kluster 3, semakin rendah validitas kluster, membuat hasil kluster yang dihasilkan menjadi baik dapat disimpulkan bahwa pembagian kluster menggunakan algoritma K-Means [22] mempunyai validitas kluster Davies Bouldin Index (DBI) lebih kecil daripada pembagian data menggunakan algoritma K-Medoid.

3.2. Klasifikasi

Klasifikasi dilakukan dengan menggunakan bahasa pemrograman Python. Berdasarkan eksperimen yang telah dilakukan pada data ini, maka diperoleh pembobotan atribut yang mana akan digunakan untuk data training C4.5+PSO dapat dilihat pada tabel 1.

Tabel 1. Hasil Seleksi Atribut Menggunakan PSO

Atribut	Weight
Umur	0,623
Tingkat Produktivitas	0,0
Pendidikan	0,863
Jenis Kelamin	0,565

Berdasarkan tabel 1 dapat disimpulkan bahwa atribut tingkat produktivitas memiliki nilai bobot nol, yang mana atribut yang memiliki nilai bobot sama dengan 0 (nol) dinyatakan terseleksi [23]

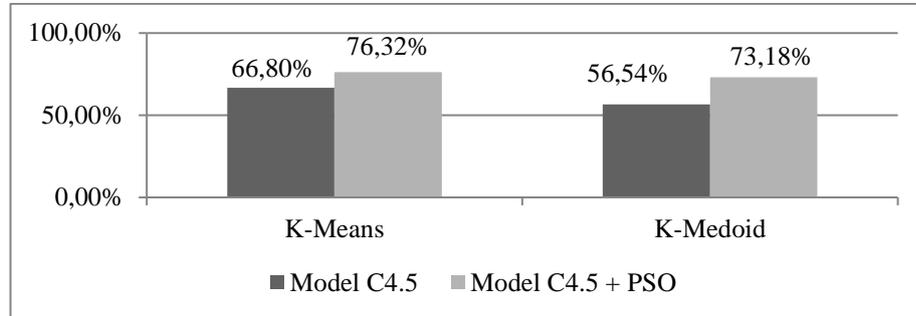
Pengukuran model dilakukan dengan mengujinya menggunakan 2 dataset metode pembagian data yaitu K-Means dan K-Medoid.

Tabel 2. Rekap Pengukuran Akurasi Model C4.5

Dataset	C4.5	C4.5 + PSO
K-Means	66,80%	76,32%
K-Medoid	56,54%	73,18%

Dari tabel 2 dapat disimpulkan bahwa hasil proses pembagian data menggunakan K-Means lebih tinggi daripada menggunakan K-Medoid. Untuk mengetahui tingkat perbandingan akurasi antara

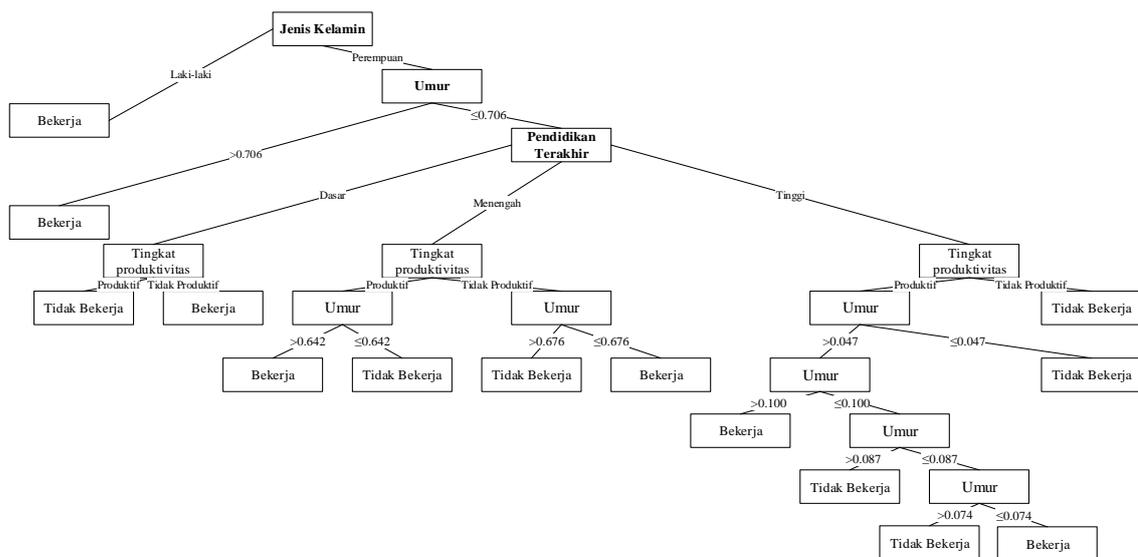
K-Means dan K-Medoid terhadap algoritma C4.5 dengan algoritma C4.5+PSO dapat dilihat pada gambar 3.



Gambar 3. Diagram Perbandingan Akurasi

Hasil pengukuran model C4.5 untuk pengukuran akurasi menunjukkan bahwa klasifikasi menggunakan algoritma C4.5+ PSO meningkat sebesar 9.52% dibandingkan dengan klasifikasi menggunakan algoritma C4.5 saja

3.3. Analisa Akhir



Gambar 4. Hasil Pohon Keputusan C4.5+PSO

Dari gambar 4 dapat dianalisa bahwa jenis kelamin, umur, dan pendidikan berada paling puncak pada pohon keputusan yang telah dihasilkan, yang mana dapat disimpulkan bahwa faktor yang paling mempengaruhi angkatan kerja dikota sawahlunto adalah jenis kelamin

4. Kesimpulan

Pada penelitian yang telah dilakukan, pembagian data menggunakan K-Means memiliki akurasi yang lebih baik dibanding algoritma K-Medoid dengan nilai DBI K-Means sebesar 0.179. Selanjutnya akurasi yang diberikan oleh C4.5 yang telah dioptimasi menggunakan algoritma PSO terbukti lebih tinggi dibandingkan menggunakan algoritma C4.5 saja. Dimana hasil pada penelitian menggunakan data ini adalah algoritma C4.5+PSO memiliki tingkat akurasi sebesar 76.32% sedangkan algoritma C4.5 memiliki akurasi sebesar 66.80%. Hasil dari klasifikasi ini yaitu faktor yang paling mempengaruhi angkatan kerja dikota sawahlunto adalah jenis kelamin, umur dan pendidikan terakhir.

Daftar Pustaka

- [1] R. S. Kodeeshwari and K. T. Ilakkiya, "Different Types of Data Mining Techniques Used in Agriculture - A Survey," *Int. J. Adv. Eng. Res. Sci.*, vol. 4, no. 6, pp. 17–23, 2017, doi: 10.22161/ijaers.4.6.3.
- [2] A. Waluyo, H. Jatnika, M. R. S. Permatasari, T. Tuslaela, I. Purnamasari, and A. P. Windarto, "Data Mining Optimization uses C4.5 Classification and Particle Swarm Optimization (PSO) in the location selection of Student Boardinghouses," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 874, no. 1, pp. 1–9, 2020, doi: 10.1088/1757-899X/874/1/012024.
- [3] *Clustering algorithms 3.1*. 2020.
- [4] Z. E. Rasjid and R. Setiawan, "Performance Comparison and Optimization of Text Document Classification using k-NN and Naïve Bayes Classification Techniques," *Procedia Comput. Sci.*, vol. 116, pp. 107–112, 2017, doi: 10.1016/j.procs.2017.10.017.
- [5] R. H. Saputra and B. Prasetyo, "Improve the Accuracy of C4.5 Algorithm Using Particle Swarm Optimization (PSO) Feature Selection and Bagging Technique in Breast Cancer Diagnosis," *J Soft Comp. Exp*, vol. 1, no. 1, pp. 47–55, 2020.
- [6] O. Pahlevi, "JITE (Journal of Informatics and Telecommunication Engineering) Data Mining Optimization Based on Particle Swarm Optimization," vol. 5, no. July, pp. 152–159, 2021.
- [7] T. Eftimov and P. Korošec, "A novel statistical approach for comparing meta-heuristic stochastic optimization algorithms according to the distribution of solutions in the search space," *Inf. Sci. (Ny)*, vol. 489, pp. 255–273, 2019, doi: 10.1016/j.ins.2019.03.049.
- [8] A. Adamu, M. Abdullahi, S. B. Junaidu, and I. H. Hassan, "An hybrid particle swarm optimization with crow search algorithm for feature selection," *Mach. Learn. with Appl.*, vol. 6, no. April, p. 100108, 2021, doi: 10.1016/j.mlwa.2021.100108.
- [9] Mustakim, "Effectiveness of K-means clustering to distribute training data and testing data on K-nearest neighbor classification," *J. Theor. Appl. Inf. Technol.*, vol. 95, no. 21, pp. 5693–5700, 2017.
- [10] I. Romli, T. Pardamean, S. Butsianto, T. N. Wiyatno, and E. Bin Mohamad, "Naive Bayes Algorithm Implementation Based on Particle Swarm Optimization in Analyzing the Defect Product," *J. Phys. Conf. Ser.*, vol. 1845, no. 1, 2021, doi: 10.1088/1742-6596/1845/1/012020.
- [11] P. Kumar and D. Sirohi, "Comparative analysis of FCM and HCM algorithm on Iris data set," *Int. J. Comput. Appl.*, vol. 5, no. 2, pp. 33–37, 2017, doi: 10.5120/888-1261.

- [12] Mustakim, M. Z. Fauzi, Mustafa, A. Abdullah, and Rohayati, "Clustering of Public Opinion on Natural Disasters in Indonesia Using DBSCAN and K-Medoids Algorithms," *J. Phys. Conf. Ser.*, vol. 1783, no. 1, 2021, doi: 10.1088/1742-6596/1783/1/012016.
- [13] D. F. Pramesti, Lahan, M. Tanzil Furqon, and C. Dewi, "Implementasi Metode K-Medoids Clustering Untuk Pengelompokan Data," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 9, pp. 723–732, 2017, doi: 10.1109/EUMC.2008.4751704.
- [14] L. D. Yulianto, A. Triayudi, and I. D. Sholihati, "Implementation Educational Data Mining For Analysis of Student Performance Prediction with Comparison of K-Nearest Neighbor Data Mining Method and Decision Tree C4.5," *J. Mantik*, vol. 4, no. 1, pp. 441–451, 2020.
- [15] M. R. Khalilpour Darzi, S. T. A. Niaki, and M. Khedmati, "Binary classification of imbalanced datasets: The case of CoIL challenge 2000," *Expert Syst. Appl.*, vol. 128, pp. 169–186, 2019, doi: 10.1016/j.eswa.2019.03.024.
- [16] L. Rahmi, "Analisis Proyeksi Pertumbuhan Penduduk Terhadap Kondisi Ketenagakerjaan Di Kota Sawahlunto Sumatera Barat," *Georafflesia*, vol. 2, no. 1, pp. 95–106, 2017.
- [17] S. Sundaramurthy and P. Jayavel, "A hybrid Grey Wolf Optimization and Particle Swarm Optimization with C4.5 approach for prediction of Rheumatoid Arthritis," *Appl. Soft Comput. J.*, vol. 94, p. 106500, 2020, doi: 10.1016/j.asoc.2020.106500.
- [18] H. Bin Wang and Y. J. Gao, "Research on C4.5 algorithm improvement strategy based on MapReduce," *Procedia Comput. Sci.*, vol. 183, pp. 160–165, 2021, doi: 10.1016/j.procs.2021.02.045.
- [19] X. Meng, P. Zhang, Y. Xu, and H. Xie, "Construction of decision tree based on C4.5 algorithm for online voltage stability assessment," *Int. J. Electr. Power Energy Syst.*, vol. 118, no. October 2019, p. 105793, 2020, doi: 10.1016/j.ijepes.2019.105793.
- [20] W. Utomo, "The comparison of k-means and k-medoids algorithms for clustering the spread of the covid-19 outbreak in Indonesia," *Ilk. J. Ilm.*, vol. 13, no. 1, pp. 31–35, 2021, doi: 10.33096/ilkom.v13i1.763.31-35.
- [21] R. M. Adnan, P. Khosravinia, B. Karimi, and O. Kisi, "Prediction of hydraulics performance in drain envelopes using Kmeans based multivariate adaptive regression spline," *Appl. Soft Comput.*, vol. 100, p. 107008, 2021, doi: 10.1016/j.asoc.2020.107008.
- [22] A. K. Singh, S. Mittal, P. Malhotra, and Y. V. Srivastava, "Clustering Evaluation by Davies-Bouldin Index(DBI) in Cereal data using K-Means," *Proc. 4th Int. Conf. Comput. Methodol. Commun. ICCMC 2020*, no. Iccmc, pp. 306–310, 2020, doi: 10.1109/ICCMC48092.2020.ICCMC-00057.
- [23] I. Yulianti, R. A. Saputra, M. S. Mardiyanto, and A. Rahmawati, "Optimasi Akurasi Algoritma C4.5 Berbasis Particle Swarm Optimization dengan Teknik Bagging pada Prediksi Penyakit Ginjal Kronis," *Techno.Com*, vol. 19, no. 4, pp. 411–421, 2020, doi: 10.33633/tc.v19i4.3579.