JURNAL TEKNOINFO
Volume 19, Nomor 2, Juli 2025, Page 105-117
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index

ANALISIS KLASTERISASI DAFTAR PEMILIH KABUPATEN MANOKWARI MENGGUNAKAN METODE K-MEANS CLUSTERING BERBASIS ELBOW METHOD

Marselinda Rante Uma[1)], Christian Dwi Suhendra[2)], Josua Josen A. Limbong[3)]

[123]Teknik Informatika, Fakultas Teknik, Universitas Papua
[123]Jalan Gunung Salju Amban Manokwari, Indonesia
Email: [1]marselindauma10@gmail.com, [2]c.suhendra@unipa.ac.id, [3]jja.limbong@unipa.ac.id

Abstract

General Elections (Pemilu) are a crucial pillar in Indonesia's democratic system, ensuring public representation in government. As voter data becomes increasingly complex due to population growth and community mobility, electoral data management requires more efficient analytical approaches to support accurate decision-making. Therefore, methods capable of accurately grouping voters based on specific characteristics are needed. This study aims to cluster voter registration data in Manokwari Regency based on age and neighborhood unit (RT) using the K-Means algorithm. A total of 16,871 entries obtained from the General Election Commission of Manokwari Regency were used, but two outliers due to input errors were removed, leaving 16,869 valid entries analyzed using Jupyter Notebook. The Elbow Method was applied to determine the optimal number of clusters by calculating the Sum of Squared Errors (SSE) from K = 2 to K = 9. The most significant drop in SSE occurred from K = 2 to K = 3 and K = 3 to K = 4, with gradual decreases afterward, indicating the elbow point lies between K = 3 and K = 4. Considering data density and segmentation, K = 4 was chosen with an SSE value of 347,575. The K-Means algorithm then clustered the data based on age and RT through random centroid initialization, Euclidean distance calculation, reassignment, and iterative centroid updates until convergence. The results showed four clusters: Cluster_0 with 6,332 young voters aged 17–29 years (RT 0–17), Cluster_1 with 3,478 productive-age voters aged 43–56 years (RT 0–14), Cluster_2 with 1,768 elderly voters aged 57–93 years (RT 0–14), and Cluster_3 with 5,291 voters aged 30–42 years (RT 0–15). The broad RT distribution across clusters indicates diverse voter age groups across the region. These findings can help the Manokwari General Election Commission (KPU) and related institutions in planning effective voter education, outreach, and logistics distribution strategies.

Keyword: *Voter Data, Data Mining, Elbow Method, Jupyter Notebook, K-Means, Clustering.*

Abstrak

Pemilihan Umum (Pemilu) merupakan pilar penting dalam sistem demokrasi Indonesia yang menjamin keterwakilan rakyat dalam pemerintahan. Dengan meningkatnya kompleksitas data pemilih akibat pertumbuhan populasi dan mobilitas masyarakat, pengelolaan data kepemiluan memerlukan pendekatan analisis yang lebih efisien guna mendukung pengambilan keputusan yang tepat dalam proses demokrasi. Untuk itu, dibutuhkan metode yang mampu mengelompokkan pemilih secara akurat berdasarkan karakteristik tertentu. Salah satu tantangan utama dalam hal ini adalah bagaimana memahami karakteristik pemilih secara lebih mendalam agar dapat mendukung perencanaan strategi sosialisasi, edukasi, dan distribusi logistik yang tepat sasaran. Penelitian ini bertujuan untuk mengelompokkan data daftar pemilih di Kabupaten Manokwari berdasarkan usia dan Rukun Tetangga (RT) menggunakan algoritma K-Means. Data yang digunakan berjumlah 16871 *entry* yang diperoleh dari Komisi Pemilihan Umum Kabupaten Manokwari. Namun dalam tahap *preprocessing* terdapat dua *outlier* yang merupakan kesalahan input dan data tersebut dihapus jadi data valid yang digunakan berjumlah 16869 *entry* data. Data tersebut di analisis menggunakan aplikasi *Jupyter Notebook*. Untuk menentukan jumlah *cluster* optimal, digunakan metode *elbow* dengan menghitung nilai *Sum of Squared Errors* (SSE) pada rentang K = 2 hingga K = 9. Hasil perhitungan menunjukkan adanya penurunan SSE yang cukup signifikan dari K = 2 ke K = 3, dan dari K = 3 ke K = 4. Namun, setelah K = 4 ke K = 5, penurunan SSE tidak lagi sebesar sebelumnya dan terus menurun secara gradual hingga K = 9. Hal ini mengindikasikan bahwa titik *elbow* berada di antara K = 3 dan K = 4. Melalui pertimbangan tambahan terhadap distribusi kepadatan data dan variasi segmentasi antar *cluster*, maka dipilih K = 4 sebagai jumlah cluster yang optimal dengan nilai *error* SSE sebesar 347575 yang menjadi dasar pemilihan jumlah *cluster* optimal. Setelah jumlah *cluster* ditentukan, algoritma K-Means digunakan untuk mengelompokkan data berdasarkan dua atribut, yaitu usia dan RT. Proses ini dimulai dengan inisialisasi centroid secara acak, kemudian dilanjutkan dengan perhitungan jarak *Euclidean* dari setiap titik data ke centroid. Selanjutnya, data dikelompokkan ke dalam *cluster* berdasarkan jarak terdekat, dan centroid dihitung ulang. Proses ini dilakukan secara iteratif hingga algoritma mencapai konvergensi. Hasil klasterisasi menunjukkan empat kelompok yaitu, *cluster*_0 terdapat 6332 data yang berisi usia muda dengan rentang 17-29 tahun dengan sebaran RT 0-17, *cluster*_1 terdapat 3478 data mencakup pemilih usia

JURNAL TEKNOINFO
Volume 19, Nomor 2, Juli 2025, Page 105-117
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index

produktif dengan rentang 43–56 tahun dalam rentang RT 0–14, *cluster*_2 terdapat 1768 data yang berisi kelompok pemilih usia lanjut dengan rentang 57–93 tahun dalam rentang RT 0–14, dan *cluster*_3 terdapat 5291 data berisi pemilih usia produktif dengan rentang 30–42 tahun, yang tersebar dalam rentang RT 0–15. Penyebaran RT yang luas dan merata dalam setiap *cluster* menunjukkan keberagaman usia pemilih di hampir seluruh wilayah. Hasil dari penelitian ini dapat digunakan oleh Komisi Pemilihan Umum (KPU) Kabupaten Manokwari dan instansi terkait lainnya sebagai dasar dalam menyusun strategi sosialisasi, edukasi pemilih, serta distribusi logistik pemilu secara lebih efektif dan efisien.

Kata Kunci: Data Pemilih, Data Mining, Metode *Elbow*, *Jupyter Notebook*, *K-Means*, Klasterisasi

1. PENDAHULUAN

Indonesia adalah negara dengan masyarakat yang beragam dan memiliki populasi yang cukup besar, terdiri dari berbagai latar belakang. Untuk menjaga stabilitas dalam Negara Kesatuan Republik Indonesia (NKRI), diperlukan pemerintahan yang bijaksana serta mampu mewakili keberagaman masyarakat, baik dari aspek geografis maupun ideologis. Salah satu cara untuk memilih pemimpin di tingkat eksekutif maupun legislatif yang dapat merepresentasikan masyarakat adalah melalui Pemilihan Umum (Pemilu). Selain berfungsi sebagai sarana memilih pemerintahan yang representatif, pemilu juga menjadi alat untuk memastikan kedaulatan rakyat tetap terjaga, mencerminkan perkembangan serta kesehatan demokrasi di Indonesia pasca reformasi[1].

Pemilihan Umum (Pemilu) memiliki dasar hukum yang kuat dalam Pasal 22E ayat (1) Undang-Undang Dasar 1945, yang mengamanatkan penyelenggaraan pemilu secara berkualitas dengan melibatkan partisipasi rakyat seluas-luasnya. Pemilu harus dilaksanakan berdasarkan prinsip demokrasi, yaitu langsung, umum, bebas, rahasia, jujur, dan adil, sebagaimana telah diatur dalam perundang-undangan yang berlaku[2]. Pemilu merupakan ajang kompetisi untuk mengisi jabatan politik dalam pemerintahan melalui mekanisme pemilihan oleh warga negara yang memenuhi syarat. Secara umum, pemilu menjadi sarana bagi rakyat untuk menentukan pemimpin atau wakil mereka dalam pemerintahan serta merupakan hak politik masyarakat sebagai warga negara dalam memilih perwakilan yang akan menjalankan roda pemerintahan[3].

Seiring dengan bertambahnya jumlah pemilih, pengelolaan data kepemiluan khususnya Daftar Pemilih Tetap (DPT) menjadi semakin kompleks. Data pemilih yang tersedia umumnya berupa daftar mentah tanpa pengelompokan berdasarkan atribut penting seperti usia, jenis kelamin, atau wilayah RT/RW. Akibatnya, lembaga penyelenggara seperti KPU kesulitan dalam memetakan distribusi demografis pemilih, menyusun strategi sosialisasi yang efektif, serta mendistribusikan logistik secara efisien. Permasalahan ini juga terlihat dalam studi sebelumnya yang mengelompokkan partisipasi pemilih berdasarkan wilayah di Kabupaten Pasuruan, di mana segmentasi data menjadi kunci dalam menyusun strategi peningkatan partisipasi pemilih[4]. Demikian pula, penelitian yang dilakukan pada tingkat desa membuktikan bahwa pengelompokan berbasis usia dan alamat dapat membantu optimalisasi pengelolaan data pemilih secara lebih efisien[5]

Permasalahan spesifik yang dihadapi di Kabupaten Manokwari adalah belum tersedianya pemetaan pemilih berdasarkan usia dan wilayah RT, padahal hal ini sangat krusial dalam menentukan pendekatan sosialisasi dan kebutuhan logistik pada tiap wilayah. Selain itu, pengelompokan data secara manual membutuhkan waktu yang lama dan rawan kesalahan, terutama ketika jumlah entri data cukup besar. Untuk menjawab tantangan tersebut, pendekatan analitik seperti data mining menjadi penting. Salah satu metode yang banyak digunakan dalam klasterisasi data adalah algoritma K-Means. Metode ini memungkinkan pengelompokan data besar secara cepat dan akurat berdasarkan kesamaan karakteristik. Beberapa penelitian sebelumnya telah menerapkan algoritma K-Means dalam konteks pemilu. Misalnya, salah satu studi mengelompokkan opini masyarakat terhadap pemilu melalui media sosial Twitter menggunakan kombinasi algoritma K-Means dan Silhouette Coefficient, sehingga opini publik dapat dikategorikan menjadi sentimen positif, netral, dan negatif[6]. Penelitian lain di Kabupaten Pasuruan memanfaatkan K-Means untuk memetakan potensi partisipasi pemilih di berbagai kecamatan berdasarkan demografi wilayah[4]. Selain itu, K-Means juga digunakan dalam menganalisis DPT di tingkat desa dengan aplikasi RapidMiner, yang membuktikan efektivitasnya dalam segmentasi berdasarkan usia dan alamat[5]

Berbeda dengan penelitian sebelumnya yang masih memiliki keterbatasan, seperti cakupan wilayah yang sempit (hanya tingkat desa atau kecamatan) dan ukuran data yang relatif kecil. Penelitian ini mencoba menjembatani kesenjangan tersebut dengan menggunakan data DPT sebanyak 16871 entri dari Kabupaten Manokwari yang diklasifikasikan berdasarkan atribut usia, RT dan jenis kelamin. Pendekatan ini memungkinkan pemetaan yang lebih luas dan mendalam terhadap distribusi pemilih.

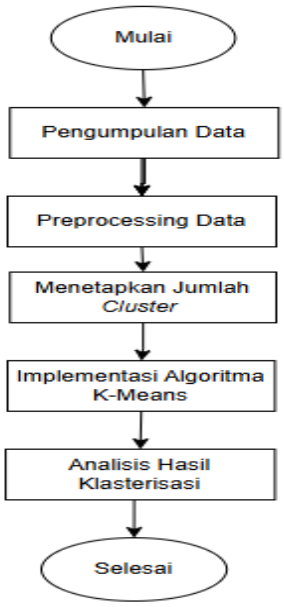
Selain itu, penelitian ini menghadapi beberapa tantangan, terutama pada tahap preprocessing data. Salah satu tantangan utama adalah keberadaan outlier yang dapat mempengaruhi posisi pusat klaster secara signifikan dan berdampak pada akurasi hasil pengelompokan. Oleh karena itu, perlu dilakukan proses identifikasi dan penanganan outlier[7], seperti mengevaluasi validitas data tersebut sebelum diputuskan untuk dipertahankan atau dihapus. Deteksi

JURNAL TEKNOINFO
Volume X, Nomor X, Bulan Tahun, Page x-x
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

outlier ini penting dalam memastikan hasil klasterisasi tetap representatif terhadap pola umum dalam data. Selain itu, penentuan jumlah klaster (K) yang optimal juga menjadi tantangan tersendiri. Penelitian ini menggunakan metode elbow, yang mengandalkan analisis nilai Sum of Squares Error (SSE) untuk mengidentifikasi titik tekuk sebagai indikasi jumlah klaster terbaik. Namun, metode ini memiliki kelemahan karena sifatnya yang subjektif, terutama ketika grafik tidak menunjukkan titik tekuk yang jelas. Penelitian ini dilakukan menggunakan Jupyter Notebook sebagai alat bantu eksplorasi dan visualisasi data. Tujuan utamanya adalah untuk melakukan segmentasi terhadap data pemilih tetap di Kabupaten Manokwari dengan memanfaatkan algoritma K-Means, sehingga pola distribusi pemilih berdasarkan usia dan RT dapat dipahami dengan lebih baik. Hasil klasterisasi ini diharapkan dapat menjadi dasar bagi KPU dan pihak terkait dalam menyusun strategi sosialisasi serta distribusi logistik secara lebih efektif dan tepat sasaran, khususnya di wilayah dengan komposisi usia dan penyebaran RT yang beragam.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif karena menggunakan data dari daftar pemilih yang telah diubah ke dalam bentuk data numerik di Kabupaten Manokwari. Pengolahan data dilakukan menggunakan *Jupyter Notebook* dengan menerapkan metode K-Means *Clustering*. Analisis data difokuskan pada pengelompokan pemilih berdasarkan beberapa atribut, yaitu usia, RT (Rukun Tetangga), dan jenis kelamin, untuk mengidentifikasi pola distribusi pemilih dan memahami karakteristik setiap kelompok. Berikut adalah langkah-langkah penelitian:



Gambar 1. Alur Penelitian

Alur penelitian pada Gambar 1 menunjukkan tahapan proses klasterisasi data menggunakan algoritma K-Means. Penelitian dimulai dengan tahap pengumpulan data sebagai fondasi utama, kemudian dilanjutkan dengan tahap preprocessing data yang mencakup proses pembersihan, seleksi atribut, transformasi data, serta penanganan outlier untuk memastikan kualitas dan validitas data. Setelah itu, dilakukan penentuan jumlah cluster optimal yang akan digunakan dalam proses klasterisasi, umumnya dengan menggunakan metode seperti Elbow Method. Tahap selanjutnya adalah implementasi algoritma K-Means untuk mengelompokkan data ke dalam cluster yang telah ditentukan. Setelah proses klasterisasi selesai, dilakukan analisis terhadap hasil cluster untuk memahami karakteristik masing-masing kelompok dan menarik kesimpulan yang relevan.

2.1 Pengumpulan Data

Pengumpulan data adalah proses mengidentifikasi variable yang diukur. Data dapat diperoleh melalui informasi data yang sudah tersedia sebelumya yang bersumber dari buku, jurnal, atau catatan. Selain itu, data juga dapat diperoleh dengan cara mencari data baru lewat survey oleh peneliti[8].

JURNAL TEKNOINFO
Volume X, Nomor X, Bulan Tahun, Page x-x
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

2.2 Pre-Processing Data

Preprocessing data adalah tahap awal dalam pengolahan data yang bertujuan untuk mempersiapkan data mentah agar siap digunakan dalam proses analisis atau pemodelan. Tahap ini penting untuk meningkatkan kualitas data dan memastikan bahwa data tersebut bersih, konsisten, dan relevan. *Preprocessing* data melibatkan beberapa tahapan, antara lain menghapus data yang duplikat, memeriksa adanya ketidaksesuaian dalam data, serta memperbaiki kesalahan yang terdapat pada data tersebut[9].

2.3 Menetapkan Jumlah Cluster

Dalam menentukan jumlah cluster yang optimal, salah satu metode yang umum digunakan adalah Metode *Elbow* dengan cara menghitung nilai SSE. *Sum of Square Error* (SSE) merupakan metrik evaluasi yang digunakan untuk mengukur total jarak kuadrat antara setiap data dengan pusat cluster-nya masing-masing. Semakin kecil nilai SSE, semakin tinggi keseragaman data dalam cluster, yang menandakan bahwa proses klasterisasi berjalan secara optimal[10] Metode *Elbow* adalah salah satu pendekatan yang digunakan untuk menentukan jumlah cluster optimal dengan menganalisis persentase hasil perbandingan antar cluster yang membentuk pola menyerupai siku pada grafik. Jika perbedaan antara nilai cluster pertama dan kedua menunjukkan sudut yang signifikan atau terjadi penurunan nilai yang paling besar, maka jumlah cluster pada titik tersebut dianggap sebagai jumlah yang optimal[11]. Metode *elbow* sering digunakan untuk menentukan jumlah *cluster* yang optimal dengan cara menghitung nilai SSE pada berbagai jumlah cluster. Hasil SSE tersebut kemudian diplot dalam sebuah grafik untuk mengidentifikasi titik "*elbow*", yaitu titik di mana penambahan jumlah cluster tidak lagi menghasilkan penurunan signifikan pada nilai SSE. Titik ini merepresentasikan jumlah *cluster* terbaik yang dapat digunakan untuk menghasilkan klasterisasi yang efisien dan bermakna[12].

2.4 Implementasi Algoritma K-Means

Pada tahap ini, dilakukan penerapan algoritma *K-Means Clustering*. *K-Means Clustering* adalah teknik pengelompokan data non-hirarki yang memisahkan data ke dalam cluster, mengelompokkan data dengan fitur yang sama bersama-sama dan mengelompokkan data dengan karakteristik yang berbeda ke dalam kelompok yang berbeda[13]. *Clustering K-Means* merupakan salah satu metode pembentukan *cluster* berbasis data, dimana objek secara acak dalam *cluster* pertama yang terbentuk dijadikan sebagai titik tengah/titik pusat (*centroid*). Nilai akhir dari metode *K-Means* sendiri adalah memaksimalkan kemiripan data di dalam satu *cluster* dan meminimalkan kemiripan data antar *cluster*. Ukuran kemiripan dalam *cluster* dengan fungsi jarak yang berarti jarak yang terdekat merupakan data yang mirip dengan titik pusat *cluster*[10]. Berikut adalah langkah-langkah algoritma *K-Means*:

a. Pemilihan titik awal *cluster*
b. Pengukuran jarak setiap data ke dalam masing-masing *cendroid*. Menggunakan rumus *Euclidean distance* :

$$d(x,y) = \sqrt{\sum(x_i - y_i)^2} \quad (1)$$

Yang dimana :
$d(x,y)$ = Jarak anatara 2 titik data x dan data y
x_i = Nilai x dari anggota ke-i dalam data
y_i = Nilai y dari anggota ke-i dalam data

c. Mengelompokkan data *cluster* dengan *centroid* terdekat
d. Menyesuaikan letak *cendroid*

$$T_j = \frac{1}{|n_j| \sum x_i, x_i} \in n_j \quad (2)$$

Yang dimana:
T_j = Titik pusat *cluster* ke-*j*
n_j = Kumpulan data dalam *cluster* ke-*j*

e. Pengulangan langkah b-d hingga hasil stabil

2.5 Analisis

Analisis adalah kegiatan berpikir untuk menguraikan suatu pokok menjadi bagian-bagian atau komponen sehingga dapat diketahui ciri atau tanda tiap bagian kemudian hubungan satu sama lain serta fungsi masingmasing bagian dari keseluruhan[14].

JURNAL TEKNOINFO
Volume X, Nomor X, Bulan Tahun, Page x-x
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Pengumpulan data yang digunakan dalam penelitian ini adalah data sekunder. Data sekunder adalah data yang didapatkan secara tidak lansung melalui sebuah perantara[8]. Data ini diperoleh dari Kantor Komisi Pemilihan Umum (KPU) Kabupaten Manokwari. Data tersebut berisi 16871 *entry* data yang mencakup enam atribut, yaitu nama, jenis kelamin, usia, desa/kelurahan, RT, dan RW. Data ini digunakan sebagai dasar dalam analisis klasterisasi daftar pemilih menggunakan algoritma K-Means.

3.2 Pre-processing

Data *preprocessing* merupakan salah satu tahapan penting dalam proses data mining. Sebelum data diproses lebih lanjut, data mentah perlu diolah terlebih dahulu agar lebih bersih dan siap digunakan. Tahapan ini biasanya melibatkan eliminasi data yang tidak relevan atau tidak sesuai, serta transformasi data ke dalam format yang lebih mudah dipahami oleh sistem. Dengan melakukan data *preprocessing*, kualitas data dapat ditingkatkan sehingga analisis yang dilakukan menjadi lebih akurat dan efektif. Jumlah data yang digunakan dalam penelitian ini sebanyak 16871 dataset. Karena jumlahnya yang cukup besar, maka saya hanya akan menampilkan 5 data teratas yang dimulai dari indeks 0 dan 5 data terbawah yang dimulai dari indeks 16866. Pada tahap ini, dilakukan pembersihan data Daftar Pemilih di Kabupaten Manokwari, yaitu dengan menghilangkan atribut yang tidak relevan, menangani data yang tidak lengkap atau tidak valid, serta melakukan transformasi data agar sesuai dengan kebutuhan analisis lebih lanjut.

a. Data Selection

Data *selection* adalah proses memilih data atau sampel yang akan digunakan dalam pengolahan data mining agar sesuai dengan tujuan atau informasi yang ingin diperoleh. Data yang dipilih berasal dari sumber yang sudah ada sebelumnya dan telah melewati proses tertentu, sehingga dapat digunakan secara efektif dalam analisis lebih lanjut[15]. Dalam data selection ini ada 3 atribut dataset yang digunakan yaitu Jenis Kelamin, Usia, dan RT. Hasil dari seleksi data tersebut ditampilkan pada tabel 1 berikut.

Tabel 1. Hasil data selection

No	Jenis kelamin	Usia	RT
0	L	41	1
1	L	22	1
2	L	77	1
3	L	18	1
4	L	48	1
...	...	...	...
16866	P	17	5
16867	L	45	2
16868	P	38	6
16869	P	41	3
16870	P	28	4

b. Transformasi

Pada tahap transformasi data, atribut Jenis Kelamin dikonversi kedalam bentuk numerik menggunakan one-hot encoding untuk memastikan data dapat diolah dalam algoritma K-Means clustering. Hasil transformasi ini menghasilkan dua kolom baru, yaitu "L" (Laki-laki) dan "P" (Perempuan). Untuk menghindari redundansi, salah satu kategori dihapus, sehingga hanya kolom P yang dipertahankan, dengan nilai 1 untuk Perempuan dan 0 untuk Laki-laki. Selain itu, atribut yang tidak relevan, seperti atribut Jenis Kelamin dalam bentuk aslinya, dihapus agar dataset lebih ringkas. Transformasi ini dilakukan untuk memastikan semua data berbentuk numerik agar dapat digunakan secara optimal dalam proses klasterisasi.

JURNAL TEKNOINFO
Volume X, Nomor X, Bulan Tahun, Page x-x
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

Tabel 2. Hasil one-hot encoding jenis kelamin

No	L	P
0	1	0
1	1	0
2	1	0
3	1	0
4	1	0
...	...	...
16866	0	1
16867	1	0
16868	0	1
16869	0	1
16870	0	1

Tabel 2 dapat dilihat, data tersebut menunjukkan hasil transformasi atribut Jenis Kelamin menjadi dua kolom baru, yaitu "L" untuk laki-laki dan "P" untuk perempuan dan hasil konversi, yaitu "P" yang direpresentasikan dengan nilai 1, sementara "L" direpresentasikan dengan nilai 0.

Tabel 3. Dataset setelah menghapus atribut jenis kelamin

No	Usia	RT
0	41	1
1	22	1
2	77	1
3	18	1
4	48	1
...	...	...
16866	17	5
16867	45	2
16868	38	6
16869	41	3
16870	28	4

Pada tabel 3 dapat dilihat dataset yang ditampilkan setelah atribut asli Jenis Kelamin dihapus karena telah dikonversi ke dalam bentuk numerik dalam kolom baru, sehingga dataset yang tersisa hanya berisi atribut lain seperti, Usia dan RT.

Tabel 4. Dataset final setelah penggabungan

No	Usia	RT	P
0	41	1	0
1	22	1	0
2	77	1	0
3	18	1	0
4	48	1	0
...	...	...	...
16866	17	5	1
16867	45	2	0
16868	38	6	1
16869	41	3	1

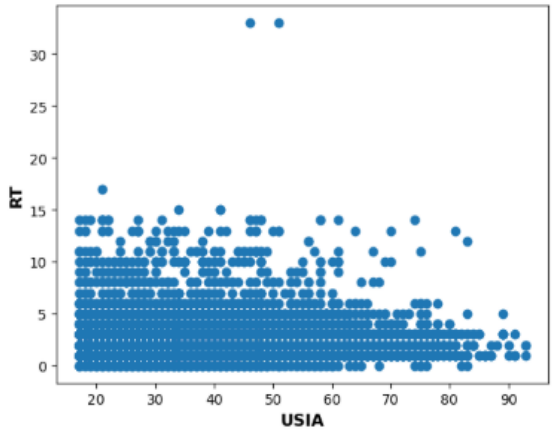
JURNAL TEKNOINFO
Volume X, Nomor X, Bulan Tahun, Page x-x
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

No	Usia	RT	P
16870	28	4	1

Pada tabel 4 menunjukkan hasil akhir setelah atribut baru P ditambahkan kedalam dataset sebagai representasi dari atribut Jenis Kelamin dalam bentuk numerik, di mana hanya satu kategori yang dipertahankan untuk menghindari redundansi, yaitu P = 1 untuk Perempuan dan P = 0 untuk Laki-laki. Meskipun atribut aslinya dihapus, informasi tentang Jenis Kelamin tetap ada dalam dataset.

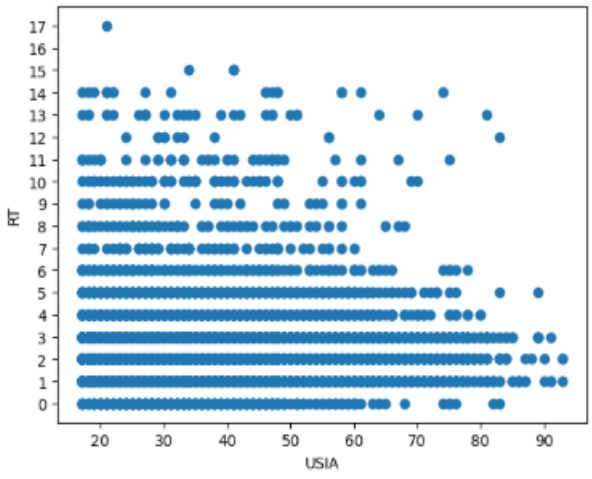
c. Deteksi dan Penanganan *Outlier*

Outlier adalah observasi yang menyimpang sangat jauh dari observasi lainnya sehingga menimbulkan kecurigaan bahwa data tersebut berasal dari mekanisme yang berbeda dibandingkan dengan sebagian besar data lainnya [16]. Dalam penelitian ini, *outlier* dideteksi menggunakan scatter plot pada atribut Usia dan RT. Berikut adalah scatter plot sebaran datanya.



Gambar 2. Scatter plot sebaran data

Gambar 2 menunjukkan scatter plot sebaran data. Pada gambar tersebut dapat dilihat ada dua *outlier* yang memiliki nilai RT lebih dari 30 yaitu RT 33, sedangkan mayoritas data berada di bawah angka tersebut. Untuk memastikan validitas data, dilakukan konfirmasi langsung ke Kantor Komisi Pemilihan Umum (KPU) Kabupaten Manokwari. Berdasarkan klarifikasi dari pihak KPU, diketahui bahwa data tersebut tidak valid dan diduga merupakan hasil dari kesalahan input. Maka data tersebut dihapus dari dataset karena datanya tidak valid. Berikut scatter plot setelah perbaikan data.



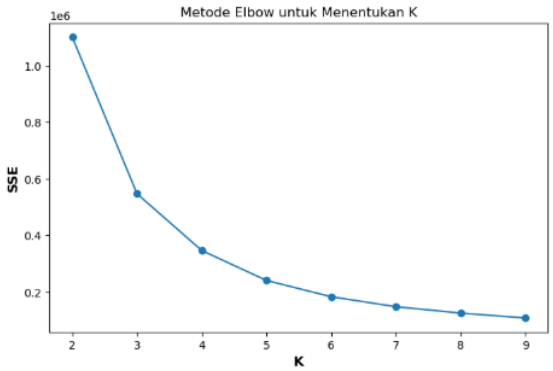
Gambar 3. Scatter plot data valid

JURNAL TEKNOINFO
Volume X, Nomor X, Bulan Tahun, Page x-x
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

Gambar 3 di atas menunjukkan scatter plot data valid. Pada gambar tersebut dapat dilihat tidak ada lagi data yang menyimpang jauh(*outlier*). Jumlah data berkurang dari data awal yaitu, 16871 *entry* data menjadi 16869 *entry* data setelah penghapusan. Kini data telah melalui tahap *preprocessing*, data telah dikonversi sepenuhnya kedalam bentuk numerik dan data yang ekstrem telah di hapus maka dataset kini siap untuk dianalisis menggunakan teknik data mining dengan algoritma *K-Means clustering* guna mengidentifikasi pola dan pengelompokkan lebih akurat.

3.3 Menentukan Jumlah Cluster

Menentukan jumlah *cluster* dalam algoritma *K-Means* dapat dilakukan menggunakan Metode *Elbow*, yang melibatkan perhitungan *Sum of Squared Errors* (SSE) untuk berbagai jumlah *cluster* (k). SSE mengukur jumlah kuadrat jarak antara setiap titik data dan pusat klasternya. Penurunan signifikan pada nilai SSE hingga mencapai titik siku pada grafik menunjukkan jumlah *cluster* optimal.



Gambar 4. Grafik metode elbow

Gambar 4 di atas merupakan grafik metode elbow. Berdasarkan hasil analisis menggunakan metode *elbow*, diperoleh bahwa titik siku (*elbow*) berada di antara K=3 dan K=4. Hal ini menunjukkan bahwa baik K=3 maupun K=4 merupakan jumlah klaster yang layak dipertimbangkan. Namun, untuk menentukan jumlah *cluster* yang paling representatif, tidak hanya dilihat dari nilai *Sum of Squared Errors (SSE)* saja, melainkan juga dari sebaran *density* (kepadatan) dan kualitas segmentasi *cluster*. Namun dari hasil perbandingan *density* K=4 memiliki kepadatan yang lebih merata di banding K=3. Oleh karena itu K=4 ditetapkan menjadi K optimal.

3.4 Penerapan Algoritma K-means

Pada tahap ini, dilakukan penerapan algoritma K-Means *Clustering* untuk mengelompokkan data daftar pemilih berdasarkan Usia, Jenis Kelamin dan RT. Berikut ini merupakan penerapan algoritma K-Means yang dilakukan yaitu:

a. Menentukan jumlah *cluster*
   Berdasarkan grafik pada Gambar 1, jumlah *cluster* optimal yang diperoleh adalah K = 4, yang ditentukan melalui metode *Elbow* dengan melihat titik siku pada grafik perubahan nilai SSE.
b. Melakukan *clustering* dengan algoritma k-means
   Setelah menetapkan jumlah *cluster* optimal sebanyak 4, proses *clustering* dilakukan menggunakan algoritma *K-Means*, yang dimulai dari *cluster* 0 hingga 3. Hasil *clustering* tersebut ditampilkan pada tabel 5 berikut:

Tabel 5. Hasil *clustering*

No	Usia	RT	P	Cluster
0	41	1	0	3
1	22	1	0	0
2	77	1	0	2
3	18	1	0	0
4	48	1	0	1
...	...	...	...	...

JURNAL TEKNOINFO
Volume X, Nomor X, Bulan Tahun, Page x-x
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

16864	17	5	1	0
16865	45	2	0	1
16866	38	6	1	3
16867	41	3	1	3
16868	28	4	1	0

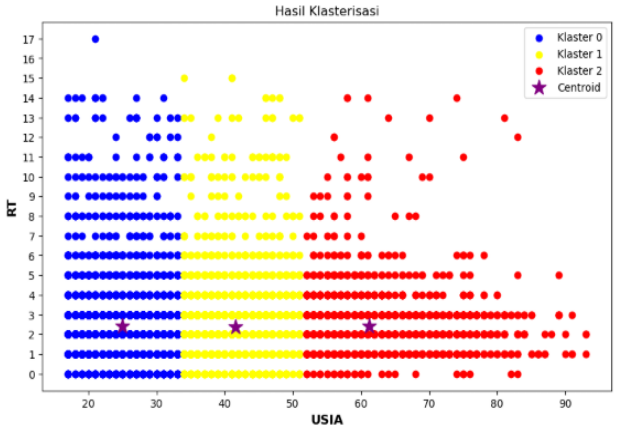
c. Evaluasi dengan SSE
Pada tahap ini, nilai error ditampilkan berdasarkan delapan kali percobaan dimulai dari K = 2 hingga K = 9. Hasil SSE dapat dilihat pada tabel 6 berikut.

Tabel 6. Hasil SSE

Percobaan	Nilai SSE
K = 2	1100028
K = 3	547000
K = 4	347575
K = 5	240621
K = 6	182858
K = 7	147427
K = 8	124121
K = 9	107361

Berdasarkan evaluasi nilai SSE dari tabel, terjadi penurunan besar dari K = 2 ke K = 3 dan dari K = 3 ke K = 4. Namun, setelah K = 4 ke K = 5, penurunannya tidak lagi sebesar sebelumnya dan terus berlanjut hingga K = 9, yang mengindikasikan bahwa titik elbow berada di antara K=3 dan K=4. Namun dengan pertimbangan hasil dari sebaran data setelah klasterisasi, berikut ini akan dibandingkan antara K=4 dan K=3 untuk menetapkan K optimal.

d. Visualisasi Hasil Clustering K=3

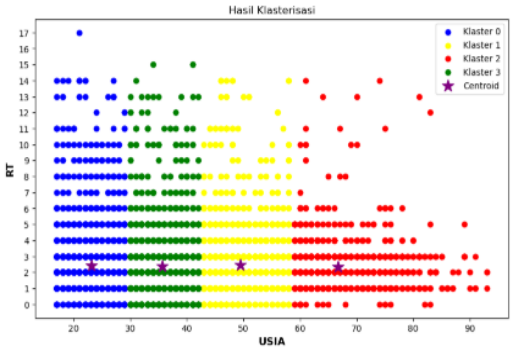


Gambar 5. Scatter plot hasil clustering untuk K=3

Gambar 5 menunjukkan visualisasi hasil clustering untuk K=3 menggunakan scatter plot berdasarkan atribut Usia dan RT. Terlihat bahwa dua cluster (biru dan merah) mendominasi dari sisi kepadatan, sedangkan cluster kuning berada di tengah-tengah dengan sebaran yang lebih luas dan *density* lebih rendah. Hal ini mencerminkan bahwa segmentasi pada K=3 masih terlalu umum, terutama dalam membedakan kelompok usia menengah, yang menyimpan banyak variasi tetapi gabungkan dalam satu *cluster*

JURNAL TEKNOINFO
Volume X, Nomor X, Bulan Tahun, Page x-x
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

e. Visualisasi hasil *clustering* K=4



Gambar 6. Scatter plot hasil clustering untuk K=4

Gambar 6 menunjukkan visualisasi hasil clustering untuk K=4 menggunakan scatter plot berdasarkan atribut Usia dan RT, pembagian *cluster* menjadi lebih seimbang dan informatif, karena tidak ada *cluster* yang terlalu dominan dalam jumlah titik. *Density* tiap *cluster* lebih merata, dan variasi pada rentang usia menengah berhasil dipisahkan menjadi dua *cluster* berbeda (hijau dan kuning), sehingga informasi tidak terkompresi dan dapat dianalisis dengan lebih mendalam.

Melalui hasil perbandingan antara K=3 dan K=4, maka K=4 ditetapkan sebagai jumlah *cluster* yang optimal karena mampu memberikan segmentasi yang lebih detail dan representatif sehingga lebih tepat digunakan dalam konteks analisis ini.

Visualisasi hasil *clustering* berdasarkan gambar 6 menunjukkan bahwa data terbagi kedalam empat *cluster* berdasarkan atribut USIA dan RT, dengan distribusi data yang bervariasi di setiap *cluster*. *cluster* dengan jumlah data terbanyak terdapat pada *cluster*_0(biru), yang mencakup pemilih dengan rentang usia 17-29 tahun, diikuti oleh *cluster*_3(hijau) dengan pemilih berusia 30-42 tahun, selanjutnya *cluster*_1(kuning) yang terdiri dari pemilih berusia 43-56 tahun dan cluster dengan jumlah data paling sedikit adalah *cluster*_2(merah), yang mayoritas berisi pemilih berusia 57-93 tahun. Titik *centroid*, yang ditandai dengan tanda bintang berwarna ungu, mewakili rata-rata usia dan RT dalam setiap *cluster*. *Clustering* usia berdasarkan RT ini bertujuan untuk memahami distribusi usia pemilih di setiap wilayah, sehingga KPU dapat menyusun strategi sosialisasi dan distribusi logistik pemilu yang lebih efektif serta menyesuaikan metode kampanye sesuai dengan karakteristik usia pemilih di masing-masing *cluster*.

f. Menampilkan jumlah data dalam setiap cluster.
Tabel 7 berikut ini menunjukkan jumlah data pada masing-masing cluster yang dimulai dari cluster 0 hingga cluster 3.

Tabel 7. Jumlah data dalam setiap cluster

Cluster	Jumlah Data
Cluster_0	6332
Cluster_1	3478
Cluster_2	1768
Cluster_3	5291

3.5 Analisis Hasil Klasterisasi

Tabel 8 menyajikan distribusi usia dan RT yang terdapat dalam masing-masing cluster, sebagai hasil dari proses klasterisasi menggunakan algoritma K-Means. Penjelasan dalam tabel ini bertujuan untuk menunjukkan bagaimana karakteristik usia dan RT tersebar di setiap cluster yang terbentuk.

JURNAL TEKNOINFO
Volume X, Nomor X, Bulan Tahun, Page x-x
ISSN: 1693-0010(Print), ISSN: 2615-224X(Online)
Available online at https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

Tabel 8. Distribusi usia dan RT pada setiap cluster

Cluster	Usia	Daftar RT
Cluster_0	17-29	1, 9, 2, 3, 0, 4, 5, 10, 8, 11, 7, 6, 14, 13, 17, 12
Cluster_1	43-56	1, 2, 3, 0, 4, 9, 11, 10, 5, 6, 14, 13, 7, 8, 12
Cluster_2	57-93	1, 2, 3, 0, 4, 5, 11, 9, 10, 6, 12, 14, 8, 13, 7
Cluster_3	30-42	1, 3, 0, 2, 4, 5, 8, 10, 9, 6, 7, 14, 13, 11, 15, 12

Berdasarkan hasil klasterisasi menggunakan algoritma K-Means, data pemilih tetap di Kabupaten Manokwari berhasil dikelompokkan ke dalam empat *cluster* berdasarkan atribut usia dan RT. *Cluster*_0 terdiri dari pemilih berusia 17–29 tahun, dengan sebaran RT dalam rentang RT 0–17, *cluster*_1 mencakup pemilih usia 43–56 tahun, tersebar dalam rentang RT 0–14, *cluster*_2 merepresentasikan kelompok pemilih berusia 57–93 tahun, dengan rentang RT 0–14. Sementara itu, *cluster*_3 berisi pemilih usia 30–42 tahun, yang tersebar dalam rentang RT 0–15. Penyebaran RT yang luas dan merata dalam setiap *cluster* menunjukkan keberagaman usia pemilih di hampir seluruh wilayah.

Analisis ini menunjukkan bahwa setiap cluster tidak hanya dibentuk berdasarkan usia dan jenis kelamin, tetapi juga mempertimbangkan keterkaitan spasial melalui keberadaan RT. Dengan demikian, informasi ini dapat membantu dalam memahami sebaran pemilih per kelompok serta dapat menjadi dasar strategis bagi KPU dalam menyusun pendekatan sosialisasi, edukasi, dan distribusi logistik pemilu yang lebih efektif dan tepat sasaran sesuai karakteristik usia di masing-masing *cluster*.

4. KESIMPULAN

Berdasarkan hasil penelitian ini, dapat disimpulkan bahwa proses klasterisasi terhadap data daftar pemilih di Kabupaten Manokwari menggunakan algoritma K-Means berhasil memberikan segmentasi yang jelas dan representatif terhadap karakteristik pemilih berdasarkan atribut usia, jenis kelamin, dan RT. Data awal sebanyak 16.871 entry yang diperoleh dari KPU Kabupaten Manokwari telah melalui proses preprocessing secara menyeluruh, meliputi seleksi atribut relevan (Jenis Kelamin, Usia, dan RT), transformasi data dengan one-hot encoding pada atribut Jenis Kelamin, serta deteksi dan penghapusan outlier untuk menjamin validitas dan kualitas data. Setelah proses pembersihan, data berkurang menjadi 16.869 entry dan dikonversi seluruhnya menjadi bentuk numerik agar sesuai dengan kebutuhan analisis klasterisasi. Dalam penentuan jumlah cluster optimal, metode Elbow digunakan dengan mempertimbangkan nilai Sum of Squared Errors (SSE), di mana penurunan SSE paling signifikan terjadi antara K=3 dan K=4. Berdasarkan pertimbangan tambahan terhadap distribusi kepadatan dan variasi segmentasi antar cluster, K=4 dipilih sebagai jumlah cluster yang optimal dengan nilai error 347575. Visualisasi hasil klasterisasi menghasilkan empat *cluster* yang mencerminkan kelompok pemilih berdasarkan rentang usia dan distribusi RT, yaitu: *Cluster*_0 (usia 17–29 tahun, RT 0–17), *cluster*_1 (usia 43–56 tahun, RT 0–14), *cluster*_2 (usia 57–93 tahun, RT 0–14), dan *cluster*_3 (usia 30–42 tahun, RT 0–15). Sebaran RT yang merata dalam setiap *cluster* menunjukkan bahwa kelompok usia pemilih tersebar luas di berbagai wilayah, sehingga hasil klasterisasi ini memberikan gambaran menyeluruh mengenai karakteristik pemilih berdasarkan usia dan lokasi, yang dapat dimanfaatkan dalam menyusun strategi sosialisasi, edukasi pemilih, dan distribusi logistik pemilu secara lebih efisien dan tepat sasaran, dengan mempertimbangkan konsentrasi kelompok usia di setiap wilayah RT. Dengan adanya pemetaan ini, KPU dapat meningkatkan efisiensi dan efektivitas dalam pelaksanaan tahapan pemilu di daerah tersebut. Secara keseluruhan, penerapan algoritma K-Means dalam penelitian ini terbukti mampu mengidentifikasi pola dan segmentasi data secara optimal, serta memberikan wawasan strategis yang bermanfaat dalam pengambilan keputusan berbasis data. Penelitian selanjutnya disarankan untuk menambahkan atribut lain yang relevan seperti alamat, tingkat pendidikan atau pekerjaan guna menghasilkan *cluster* yang mencerminkan profil pemilih yang lebih komprehensif. Selain itu, penggunaan algoritma *cluster*isasi lain seperti DBSCAN atau *Hierarchical Clustering* juga dapat dijadikan alternatif pembanding untuk memperoleh hasil analisis yang lebih variatif dan mendalam.

DAFTAR PUSTAKA

[1] M. I. Rantau, "Penguatan Sistem Presidensial Di Indonesia: Analisis Terhadap Undang Undang No 7 Tahun 2017 Tentang Pemilihan Umum," *J. Penelit. Dan Karya Ilm.*, vol. 19, no. 2, pp. 181–193, 2019, doi: 10.33592/pelita.vol19.iss2.120.

[2] L. Nasution, "Pemilu dan Kedaulatan Rakyat," *'Adalah*, vol. 1, no. 9, pp. 83–84, 2017, doi: 10.15408/adalah.v1i9.11323.

[3] D. Anggara, "Kajian Umum Pilkada," *Africa's potential Ecol. Intensif. Agric.*, vol. 53, no. 9, pp. 1689–1699, 2019.
[4] D. M. Chulloh, A. S. Fitrani, I. R. Indra Astutik, and A. Eviyanti, "Uji Akurasi K-Means dalam Prediksi Partisipasi Pemilu pada Demografi Wilayah Kabupaten Pasuruan," *Jutisi J. Ilm. Tek. Inform. dan Sist. Inf.*, vol. 13, no. 1, p. 201, 2024, doi: 10.35889/jutisi.v13i1.1753.
[5] S. D. Hilda, A. Voutama, and Y. Umaidah, "Analisis Daftar Pemilih Tetap Pemilihan Gubernur dan Wakil Gubernur menggunakan Algoritma K-Means," *JATISI (Jurnal Tek. …*, vol. 10, no. 3, pp. 398–408, 2023, [Online]. Available: https://jurnal.mdp.ac.id/index.php/jatisi/article/view/4921%0Ahttps://jurnal.mdp.ac.id/index.php/jatisi/article/download/4921/1600
[6] Y. Andrianus, W. Wasino, and T. Sutrisno, "Implementasi Algoritma K-Means Terhadap Opini Masyarakat Mengenai Perkiraan Pemilu 2024 Pada Twitter," *Simtek J. Sist. Inf. dan Tek. Komput.*, vol. 8, no. 2, pp. 305–308, 2023, doi: 10.51876/simtek.v8i2.271.
[7] M. F. Anggarda, I. Kustiawan, D. R. Nurjanah, and N. F. A. Hakim, "Pengembangan Sistem Prediksi Waktu Penyiraman Optimal pada Perkebunan: Pendekatan Machine Learning untuk Peningkatan Produktivitas Pertanian," *J. Budid. Pertan.*, vol. 19, no. 2, pp. 124–136, 2023, doi: 10.30598/jbdp.2023.19.2.124.
[8] R. Arviyanda, E. Fernandito, and P. Landung, "Analisis Perbedaan Bahasa dalam Komunikasi Antarmahasiswa," *J. Harmon. Nusa Bangsa*, vol. 1, no. 1, p. 67, 2023, doi: 10.47256/jhnb.v1i1.338.
[9] A. Winarta and W. J. Kurniawan, "Optimasi Cluster K-Means Menggunakan Metode Elbow Pada Data Pengguna Narkoba Dengan Pemrograman Python," *JTIK (Jurnal Tek. Inform. Kaputama)*, vol. 5, no. 1, pp. 113–119, 2021, doi: 10.59697/jtik.v5i1.593.
[10] L. P. Refialy, H. Maitimu, and M. S. Pesulima, "Perbaikan Kinerja Clustering K-Means pada Data Ekonomi Nelayan dengan Perhitungan Sum of Square Error (SSE) dan Optimasi nilai K cluster," *Techno.Com*, vol. 20, no. 2, pp. 321–329, 2021, doi: 10.33633/tc.v20i2.4572.
[11] D. A. I. C. Dewi and D. A. K. Pramita, "Analisis Perbandingan Metode Elbow dan Silhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali," *Matrix J. Manaj. Teknol. dan Inform.*, vol. 9, no. 3, pp. 102–109, 2019, doi: 10.31940/matrix.v9i3.1662.
[12] R. Nainggolan, R. Perangin-Angin, E. Simarmata, and A. F. Tarigan, "Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method," *J. Phys. Conf. Ser.*, vol. 1361, no. 1, 2019, doi: 10.1088/1742-6596/1361/1/012015.
[13] L. Azzahra and Amru Yasir, "Metode K-Means Clustering Dalam Pengelompokan Penjualan Produk Frozen Food," *J. Ilmu Komput. dan Sist. Inf.*, vol. 3, no. 1, pp. 1–10, 2024, doi: 10.70340/jirsi.v3i1.88.
[14] M. S. Ummah, "ANALISIS KINERJA KEUANGAN DENGAN METODE ALTMAN Z – SCORE PADA PT. MATAHARI DEPARTMENT STORE TBK," *Sustain.*, vol. 11, no. 1, pp. 1–14, 2019.
[15] A. Srirahayu and L. S. Pribadie, "Review Paper Data Mining Klasifikasi Data Mining," *J. Ilm. Inform. Glob.*, vol. 14, no. 1, 2023, doi: 10.36982/jiig.v14i1.2981.
[16] R. Yuliani, B. Pusat, S. Provinsi, and K. Utara, "Identifikasi Nilai Esensial Dari Outlier Non-Extreme Menggunakan Metode Minimum Volume Ellipsoid," *AJurnal TEKINKOM*, vol. 6, no. 1, pp. 236–244, 2023, doi: 10.37600/tekinkom.v6i1.572.