Jurnal Algoritma
E-ISSN: 2302-7339 | P-ISSN: 1412-3622
INSTITUT TEKNOLOGI GARUT
https://jurnal.itg.ac.id/index.php/algoritma
DOI: 10.33364/algoritma/v.21-2.2155

Penerapan Algoritma *K-means Clustering* dan *Hierarchical Clustering* dalam Mengelompokkan Data Pengangguran di Karawang

Assyifa Alif Rahayu Mulyana[1*], Ayu Ratna Juwita[2], Amril Mutoi Siregar[3], Ahmad Fauzi[4]
Fakultas Ilmu Komputer, Program Studi Teknik Informatika
Universitas Buana Perjuangan Karawang

***email:** if20.assyifaalifrahayumulyana@mhs.ubpkarawang.ac.id

Info Artikel
Dikirim: 22 Oktober 2024
Diterima: 7 November 2024
Diterbitkan: 30 November 2024

Kata kunci:
Algoritma;
Cluster;
Karawang;
Pengangguran;
Pengelompokkan.

ABSTRAK

Adanya berbagai kawasan industri baru di Karawang dapat memicu penduduk luar daerah untuk bermigrasi. Bertambahnya jumlah penduduk dapat mempengaruhi tingkat pengangguran di suatu daerah. Dalam mengelompokkan sebuah data dapat menggunakan teknik *data mining*. Penelitian ini bertujuan memanfaatkan Algoritma *K-Means* dan *Hiererchical Clustering* dalam pengelompokkan untuk memperoleh informasi. Hasilnya Algoritma *K-Means* dan *Hierarchical Clustering* dapat mengelompokkan data berdasarkan karakteristik yang mirip dengan jumlah *cluster* yang sama tetapi memiliki perbedaan distribusi data dalam *cluster*. Metode evaluasi dengan *Silhouette Score* menunjukkan kedua algoritma tersebut memiliki kinerja yang sama dalam analisis di penelitian ini.

1. PENDAHULUAN

Berdasarkan perubahan struktur perekonomian dengan seiring berjalannya waktu, kota Karawang menjadi dikenal sebagai kota industri karena terdapat berbagai kawasan industri baru di Karawang. Hal tersebut dapat memicu penduduk luar daerah untuk melakukan migrasi yang akan berdampak pada jumlah penduduk sehingga dapat mempengaruhi tingkat pengangguran di suatu daerah [1]. Menurut Badan Pusat Statistik (BPS) kota Karawang, terdapat 8,95 persen pengangguran yang ada di Karawang berdasarkan bulan Agustus tahun 2023. Angka pengangguran diatas 5% dikategorikan sebagai tingkat pengangguran yang tinggi berdasarkan tolak ukur kondisi ekonomi pada daerah tersebut [2]. Suatu masalah dalam sebuah daerah jika terdapat tingkat penganggurannya yang tinggi yang akan berdampak di bidang sosial maupun ekonomi [3]. Salah satunya pada penelitian Amaliya dan Gunawan (2021) yang membahas lonjakan pengangguran akibat PHK massal di Karawang saat pandemi Covid-19. Hal tersebut memicu tindakan kriminal terutama faktor kesulitan di bidang ekonomi dan kondisi lingkungan yang mendukung kejahatan serta pengaruh dari kejahatan serupa di daerah lain [4]. Selain itu, pada penelitian Kusano dkk (2023) menemukan bahwa peningkatan angka pengangguran mempengaruhi tingkat bunuh diri di kalangan orang asia yang lebih muda dan orang kulit hitam [5].

Dalam memperoleh informasi yang bermanfaat, teknik *data mining* bisa digunakan untuk mengolah data salah satunya dengan *clustering*. Algoritma *clustering* menganalisis dengan mengelompokkan data berdasarkan kesamaan tertentu atau berdasarkan kelas yang memiliki karakteristik serupa [6]. Adapun Algoritma *K-Means Clustering* yang dapat menganalisis sampel dengan ukuran besar secara efisien tetapi memiliki kelemahan terhadap data outlier. Berikutnya, Algoritma *Hierarchical Clustering* yang dapat mempercepat pengolahan data sehingga menghemat waktu analisis tetapi memiliki kekurangan seperti ukuran jarak yang digunakan mempunyai perbedaan [7]. Berdasarkan kedua algoritma tersebut, Abdulhafedh (2021) menunjukkan bahwa pengelompokkan dengan menggunakan *K-Means* dan *Hierarchical Clustering* memiliki proses yang efektif dengan menerapkan reduksi dimensi seperti *Principal Component Analysis* (PCA) [8].

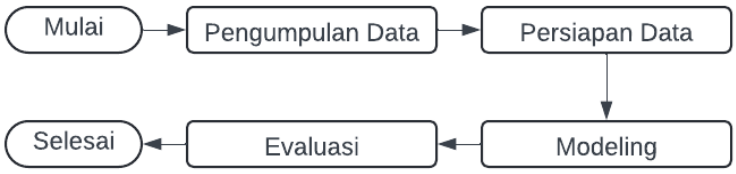
Penelitian terdahulu yang dilakukan Badri dan Habibi (2022) menerapkan Algoritma *K-Means* dalam pengelompokkan tingkat pengangguran pasca pandemi Covid-19 berdasarkan provinsi yang ada di Indonesia. Hasil analisis tersebut menunjukkan bahwa *clustering* terbagi menjadi dua kelompok yakni daerah dengan tingkat pengangguran tinggi sebanyak 10 data dan daerah dengan tingkat pengangguran rendah sebanyak 24 data [9]. Adapun penelitian Murniasih dkk (2020) yang mengaplikasikan Algoritma *K-Means* pada pengelompokkan tingkat buta aksara berdasarkan provinsi yang ada di Indonesia. Hasilnya terdapat 10 provinsi pengidap buta aksara dengan kelompok sedang dan 22 provinsi penederita buta aksara dengan kelompok rendah [10]. Penelitian yang dilakukan Junior dkk (2022) juga menunjukkan bahwa *K-Means* dapat digunakan dalam analisis sentimen pada Buletin APTIKOM. Hasil penelitian tersebut berupa nilai *Sum of Square Error* (SSE) yang mencapai 75% dari 2 *cluster* [11].

Selanjutnya penelitian Dewi dkk (2020) mengelompokkan sebaran tenaga kesehatan di provinsi Jawa Tengah untuk pemerataan dengan menerapkan Algoritma *K-Means*. Memperoleh hasil pengelompokkan dengan *cluster* tinggi terdapat 4 kota, kelompok dengan nilai sedang terdapat 25 kota serta kelompok dengan nilai rendah terdapat 6 kota [12]. *K-Means Clustering* juga digunakan oleh Adeliana dkk (2021) dalam mengelompokkan tingkat produksi daging sapi yang ada di Indonesia dan dibandingkan dengan Algoritma *K-Medoids*. Hasilnya menunjukkan bahwa dua algoritma yang dipakai cocok digunakan dalam pengelompokkan tersebut [13]. Berikutnya kedua algoritma tersebut juga digunakan dalam penelitian yang dilakukan Rofik dkk (2021) untuk mengelompokkan tingkat kepuasan siswa terhadap pelayanan sekolah. Jumlah *cluster* yang diperoleh dari kedua algoritma tersebut berbeda karena memiliki perhitungan masing-masing [14].

Berdasarkan uraian penelitian terkait, Algoritma *K-Means Clustering* mempunyai banyak keunggulan dalam pengelompokkan data. Berikutnya, angka pengangguran berdasarkan BPS kota Karawang per tahun 2023 masih menunjukkan angka yang masuk ke dalam kategori tinggi, sehingga hal tersebut membutuhkan informasi lebih lanjut mengenai pengangguran yang ada di kota Karawang. Penelitian Almaeira dkk (2021) memanfaatkan Algoritma *K-Means Clustering* untuk pengelompokkan tingkat pengangguran berdasarkan provinsi [15]. Pada penelitian ini Algoritma *K-Means Clustering* digunakan dengan Algoritma *Hierarchical Clustering* untuk mengelompokkan data pengangguran berdasarkan dua kecamatan di kota Karawang dengan tujuan memperoleh informasi mengenai pengangguran yang ada di Karawang.

2. METODE PENELITIAN

Penelitian ini menggunakan metode *data mining* untuk menganalisis atau memperoleh informasi mengenai data. Prosesnya meliputi pengumpulan data, persiapan data, pembuatan model (modeling) dan penilaian hasil (evaluasi). Tahapan-tahapan sesuai pada Gambar 1.



Gambar 1. Prosedur Penelitian

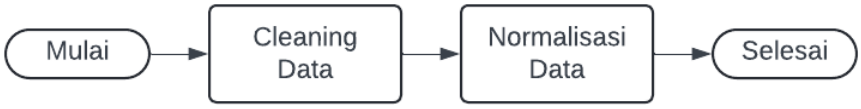
2.1 Pengumpulan Data

Pada penelitian ini menggunakan dataset yang diperoleh dari hasil menyebarkan kuisioner tahun 2024 di kecamatan yang ada di kota Karawang yakni Karawang Barat dan Telukjambe Timur dengan kriteria sedang tidak bekerja atau sedang mencari pekerjaan. Fokus kuisioner adalah mengidentifikasi faktor-faktor yang berkaitan dengan pengangguran. Variabel seperti tingkat pendidikan terakhir, pengalaman kerja, pelatihan yang diikuti, durasi menganggur dan lainnya. Kuisioner disebarkan dalam bentuk format digital yaitu Google Form. Dari hal tersebut memperoleh sebanyak 1500 responden, serta berformatkan CSV (*Comma Separated*

Value). Selanjutnya, memperoleh data jumlah penduduk dari Dinas Kependudukan dan Pencatatan Sipil (Disdukcapil) kota Karawang untuk dijadikan acuan dalam penelitian ini.

2.2 Persiapan Data

Pada langkah ini melakukan *pre-processing* seperti *cleaning data* dengan mengidentifikasi *missing value* pada data, menangani duplikat data dan melakukan normalisasi data. Hal tersebut dilakukan agar mendapatkan data yang berkualitas tinggi untuk dianalisis lebih lanjut [16]. Adapun tahapan *pre-processing* yang ditampilkan pada Gambar 2.



Gambar 2 *Pre-Processing*

2.3 Modeling

Pada tahap ini menerapkan teknik clustering untuk mengidentifikasi pola dalam analisis yang akan dilakukan. Model *machine learning* yang diterapkan yang pertama yaitu *K-Means Clustering* karena algoritma ini efisien dan dapat menangani dataset besar seperti dataset dalam penelitian ini serta salah satu kelebihan Algoritma *K-Means Clustering* ini mampu menangani data besar dengan cepat dan mudah diimplementasikan. Lalu yang kedua yaitu Algoritma *Hierarchical Clustering*, dipilih karena algoritma ini dapat membantu memahami hubungan hierarkis antar data sehingga memberikan informasi mengenai hubungan karakteristik antar responden yang lebih kompleks. Kedua Algoritma tersebut memberikan manfaat seperti kombinasi kecepatan dan presisi dimana *K-Means* digunakan untuk *clustering* cepat sedangkan *Hierarchical* membantu memberikan informasi lebih dalam struktur data. Adapun algoritma tersebut, yaitu:

1) *K-Means Clustering*
   Algoritma *K-Means* merupakan teknik pengelompokkan data yang membagi data ke dalam sejumlah *cluster*, objek-objek data yang memiliki kesamaan karakteristik akan dikelompokkan dalam satu *cluster* yang sama, proses Algoritma *K-Means* yaitu:
   a. Tentukan jumlah kelompok.
   b. Pilih titik tengah awal
   c. Kelompokkan data berdasarkan jarak terdekar, menggunakan rumus jarak *Euclidean* (*Euclidean Distance*) untuk melihat seberapa jauh suatu data dari suatu *centroid*. Persamaan *Euclidian Distance* yaitu:

      $$d(x_i, \mu_j) = \sqrt{\Sigma(x_i, \mu_j)^2} \quad (1)$$

      Keterangan:
      x_i : Data kriteria.
      μ_j : *Centroid* pada cluster ke-j
      d : Mengelompokkan masing-masing data berdasarkan jarak paling dekat dengan *centroid*.
   d. Menghitung ulang titik tengah dari setiap *cluster*. Rata-rata *cluster* menghasilkan nilai *centroid* baru berdasarkan rumus berikut:

      $$\mu j(t+1) = \frac{1}{Nsj} \sum j \in Sj\, \chi j \quad (2)$$

      Keterangan:
      $\mu j(t+1)$: *Centroid* baru pada iterasi.
      ke $(t+1)$ Nsj : Banyak data pada *cluster* Sj.
   e. Melakukan pengulangan sampai langkah d, hingga menemukan tidak ada yang berubah pada anggota dalam setiap *cluster* [7].
2) *Hierarchical Clustering*

Hierarchical Clustering mengelompokkan data berdasarkan bagan dalam bentuk hirarki, hal tersebut merupakan penggabungan dua kelompok yang lebih dekat di setiap iterasi atau *cluster* yang berisi pembagian dari seluruh set data. Adapun tahapan dari Algoritma *Hierarchical Clustering* sebagai berikut:

a. Mengidentifikasi item dengan jarak terdekat.
b. Menggabungkan ke dalam satu *cluster*
c. Menghitung jarak setiap *cluster*
d. Melakukan pengulangan dari awal hingga semua terhubung [7].

2.4 Evaluasi

Tahap ini merupakan pengujian untuk mengevaluasi efektivitas dari kinerja hasil perhitungan model *machine learning* yang digunakan agar menjadi informasi yang mudah dimengerti berdasarkan data yang sudah diolah [16]. Metode evaluasi *Silhouette Coefficient* dapat digunakan dalam mengukur kualitas model untuk tujuan penelitian yang ingin dicapai berdasarkan *cluster* yang sudah dihasilkan [17]. Metode ini merupakan serangkaian dari dua metode seperti *Cohession* dan *Separation* lalu penerapan kedua metode ini untuk memvalidasi hasil dari clustering. *Cohession* berfungsi untuk menghitung sebuah cluster berdasarkan keterkaitan antara objek data. Sedangkan *Separation* berguna dalam memperhitungkan jarak *cluster* terpisah dengan *cluster* yang lain [18]. Adapun persamaan dari *Silhouette Coefficient* sebagai berikut:

$$sil\ (c) = sil\ (k) \frac{1}{|k|} \sum_{i=1} k\ sil(ci) \quad (3)$$

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Data jumlah penduduk yang diperoleh dari disdukcapil berjumlah 53.281 yang merupakan jumlah dari penduduk kecamatan Karawang Barat dan Telukjambe Timur. Dari jumlah tersebut didapati 1.500 responden yang mengisi kuisioner yang dibagikan sehingga jumlah responden merupakan 2.81 persen dari jumlah penduduk Karawang Barat dan Telukjambe Timur. Berikutnya, data yang diperoleh dari menyebarkan kuisioner berisi biodata yang perlu diisi oleh responden dan atribut yang berbentuk pertanyaan-pertanyaan yang ditampilkan pada kuisioner. Adapun pertanyaan-pertanyaan yang terdapat dalam kuisioner berdasarkan Tabel 1.

Tabel 1 Kuisioner

Kode	Pertanyaan
P1	Apakah anda pernah bekerja sebelumnya?
P2	Jika Ya, berapa lama anda telah bekerja secara keseluruhan?
P3	Mengapa anda saat ini tidak bekerja?
P4	Berapa lama anda tidak bekerja?
P5	Apakah anda sedang aktif mencari pekerjaan?
P6	Jika ya, berapa banyak lamaran pekerjaan yang sudah anda kirimkan dalam 3 bulan terakhir?
	Apa kendala utama yang anda hadapi dalam mencari pekerjaan?
P7	Seberapa sering anda mencari pekerjaan melalui media berikut? (pilih yang sesuai)
P8	Apakah anda mengikuti pelatihan atau kursus untuk meningkatkan keterampilan?
P9	Seberapa sering anda mengikuti wawancara kerja dalam 3 bulan terakhir?
P10	Apakah anda merasa keterampilan anda sesuai dengan kebutuhan pasar kerja?
P11	Apakah anda pernah ditolak dalam wawancara kerja?
P12	Apakah anda sedang mengikuti studi setelah tingkat pendidikan terakhir?
P13	Jika ya, apa jenis studi yang sedang anda ikuti:
P14	Apakah anda pernah mengikuti program magang?
P15	Jika ya, sebutkan durasi magang anda:

P16	Apakah pekerjaan yang anda lamar saat ini sesuai dengan keterampilan anda?
P17	

Berikutnya adapun hasil jawaban responden dari kuisioner tersebut yang dapat dilihat pada Tabel 2.

Tabel 2 Hasil Kuisioner

No	P1	P2	…	P13	P14	P15	P16	P17
1	YA	1-3 tahun	…	TIDAK	Tidak studi	YA	3 bulan	YA
2	YA	1-3 tahun	…	TIDAK	-	TIDAK	-	YA
3	YA	1-3 tahun	…	TIDAK	Tidak ada	YA	3 bulan	YA
4	TIDAK	-	…	TIDAK	-	TIDAK	-	TIDAK
5	YA	Kurang dari 1 tahun	…	YA	S1	YA	-	YA
…	…	…	…	…	…	…	…	…
1500	YA	-		TIDAK	-	YA	-	-

3.2 Persiapan Data

Pada tahap ini mempersiapkan data untuk menyesuaikan dengan kebutuhan analisis, pada proses ini dilakukan penghapusan atribut yang tidak digunakan dan mengubah nama atribut sesuai tujuan analisis. Sehingga dapat dilihat pada Tabel 3

Tabel 3. Atribut untuk analisis

Nomor	Atribut	Deskripsi
1	Gender	Berisi nilai 1 : Perempuan dan 0 : Laki-Laki
2	Region	Berisi niai 1 : Karawang dan 0 : Luar Karawang
3	Exp_Work	Berisi nilai 1 : YA dan 0 : TIDAK
4	Unemp_Dur	Berisi beberapa kategori waktu menganggur
5	Training	Berisi nilai 1 : YA dan 0 : TIDAK
		Berisi nilai 1 : YA dan 0 : TIDAK
6	Studying	Berisi nilai 1 : YA dan 0 : TIDAK
		Berisi nilai 1 : YA dan 0 : TIDAK
7	Intern_Exp	Berisi nilai 1 : YA dan 0 : TIDAK
		Berisi nilai 1 : YA dan 0 : TIDAK
8	Diploma	Berisi nilai 1 : YA dan 0 : TIDAK
9	Sarjana	
10	SMA/SMK	
11	SMP	

Setelah tahap mempersiapkan data untuk analisis lebih lanjut dilakukan *pre-processing* sesuai tahapan yang sudah dijelaskan sebelumnya pada Bab 2, adapun proses tersebut seperti:

1) Cleaning Data
   Menggunakan fungsi '*isnull*()' untuk melihat jumlah *missing value* pada setiap kolom dalam dataset. Setelah menjalankan fungsi tersebut ditemukan *missing value* pada atribut Training, Studying dan Intern_Exp berdasarkan Gambar 3.

```
Gender          0
Pendidikan      0
Region          1
Exp_Work        0
Unemp_Dur       0
Training        8
Studying       10
Intern_Exp      6
dtype: int64
```

Gambar 3 Identifikasi *Missing Value*

Berikutnya mengidentifikasi duplikasi data dengan menggunakan fungsi '*duplicated*()', setelah menggunakan fungsi tersebut terdapat duplikasi data berdasarkan Gambar 4.

```
Jumlah Duplikasi Data: 1171
```

Gambar 4 Jumlah Duplikasi Data

Setelah *missing value* dan duplikasi data teridentifikasi langkah selanjutnya adalah menangani hal tersebut dengan menghapusnya menggunakan fungsi '*dropna*()' untuk menghapus *missing value* dan menghapus duplikasi data dengan fungsi '*drop_duplicates()'*. Hal ini diperlukan terutama dalam melakukan *clustering* [19].

2) Normalisasi Data

Tahap ini melakukan normalisasi data pada satu atribut yaitu Unemployment_Duration, karena atribut yang lain sudah memiliki nilai yang sama yaitu 0 dan 1, hal ini dilakukan agar data memiliki skala yang sama. Adapun hasil normalisasi tersebut sebagaimana pada Tabel 4.

Tabel 4. Hasil normalisasi

No	Unemp_Dur
0	0.000000
1	0.571429
2	0.000000
3	0.571429
4	0.571429

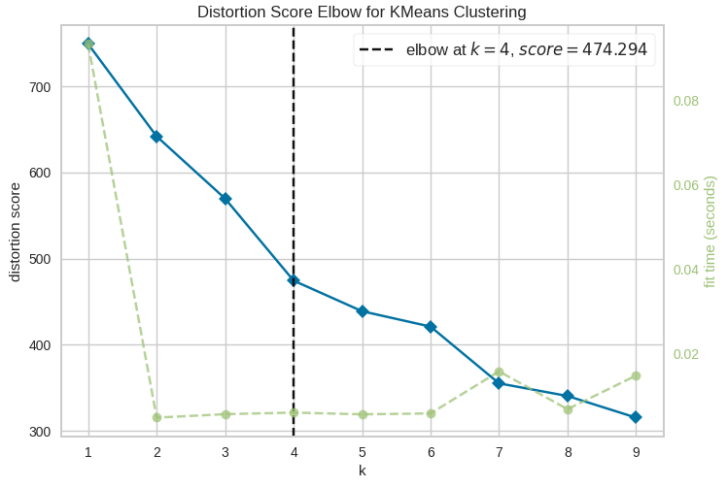
3.3 Modeling

1) Algoritma K-Means Clustering

Algoritma *K-Means Clustering* diterapkan untuk menemukan *cluster* dengan mengelompokkan objek berdasarkan data yang akan di cluster yang sudah diidentifikasi di awal. Data yang akan di *cluster* berisi atribut yang sudah dinormalisasi dan sudah melalui tahap konversi data. Pada langkah pertama adalah menetapkan titik pusat cluster secara bebas untuk pertama kali melakukan iterasi. Hal tersebut mengikuti sesuai langkah-langkah Algoritma *K-Means Clustering* yang ada pada metode penelitian.

a. Menentukan jumlah K

Pada tahap ini metode *elbow* akan digunakan untuk mengoptimalkan jumlah *cluster*. Metode yang digunakan dapat menampilkan grafik *elbow* berupa titik sudut siku. Dapat dilihat pada Gambar 5.



Gambar 5 Metode *Elbow*

Pada grafik *elbow* diatas menunjukkan bahwa cluster yang optimal dimulai dari 4, untuk penelitian ini memilih 3 *cluster* untuk melakukan pengelompokkan sesuai kebutuhan analisis.

b. Inisialisasi pusat cluster

Hal ini dilakukan untuk mendapatkan nilai acak sebagai pusat awal *cluster* sebanyak K. Hasilnya dapat dilihat pada Gambar 6.

```
[0 2 1 2 0 1 0 2 2 0 0 2 1 1 0 0 2 0 2 2 0 0 1 0 0 0 0 0 2 1 2 1 2 0 1 1 1
 0 2 0 1 1 1 1 2 1 0 0 1 0 2 0 1 1 1 2 1 2 2 1 2 0 1 1 2 2 0 0 1 2 0 2 2 1
 2 2 2 1 0 1 2 2 2 1 2 2 0 1 2 2 2 2 0 2 0 0 0 2 0 2 0 0 0 2 0 0 0 0 0 0 1
 0 1 0 0 2 0 0 0 1 1 1 0 2 0 2 2 0 0 0 0 0 0 1 0 0 2 2 0 2 2 1 2 0 0 2 0 2
 0 0 0 0 0 2 0 2 2 2 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 2 0 0 2 0 0 0 0 0 2 0 2 0 0 0 0 0 0 0 0 0 2 0 2 0 2 0 0 0 0 0 0 0 0 0 2 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0
 0 0 2 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 2 0
 0 0 0 2 0 0 0 0 0 0 0 0 0 2 2 2 2 2 1 1 2 2 2 2 1 1 2 2 2 1 1 1 1 1 1 1 1
 1 1 1 2 2 1 2 2 2 2 1 2 2 1 2 1 2 2 2 2 2 2 1 2 2 2 1 2 1 2 2 2 2 2 2 1 2 2
 2 1 2 1 2 2 2 2 2 2 2 1 1 1 1 2 1 2 2 2 2 2 2 2 2 2 2 1 2 1 2 1 2 1 2 2 1
 1 2 1 2 2 2 1 2 2 1 1 2 2 2 2 2 1 1 2 2 2 2 2 1 1 2 2 1 1 2 1 2 1 2 1 1 1 2 2
 1 1 1 2 1 2 1 1 1 2 2 2 1 1 2 1 1 1 1 1 1 1 1 2 1 1]
```

Gambar 6 Inisialisasi Pusat Cluster

Pada Gambar 6 menunjukkan label *cluster* untuk setiap titik data dalam dataset penelitian ini, label tersebut berdasarkan *cluster* optimal yang sudah ditentukan sebelumnya untuk analisis ini.

c. Menghitung jarak setiap data ke setiap centroid

Pada langkah ini dapat melihat hasil *clustering* dan memahami setiap titik data terdistribudi di antara cluster-cluster. Sebagaimana pada Tabel 5.

Tabel 5 Jarak Data ke *Centroid*

No	Gender	Region	…	Centroid_1	Centroid_2	Centroid_3	Cluster
0	1.0	1.0	…	1.203	1.787	1.330	0
1	1.0	1.0	…	1.336	1.557	0.996	2
2	1.0	1.0	…	1.457	1.424	1.895	1
3	1.0	1.0	…	1.440	1.275	0.364	2
4	0.0	1.0	…	1.298	2.394	2.068	0

Berdasarkan Tabel 3, kolom *centroid* merupakan jarak *Euclidean* dari titik data ke *centroid* atau menunjukkan seberapa dekat setiap titik data dengan masing-masing *centroid cluster*. *K-Means Clustering* menggunakan nilai-nilai ini untuk menentukan ke *cluster* mana titik data akan ditetapkan.

d. Memperbarui nilai centroid

Tahap ini Algoritma *K-Means* menghitung dan memperbarui posisi *centroid* hingga proses iterasi maksimum tercapai. Dapat dilihat pada Gambar 7.

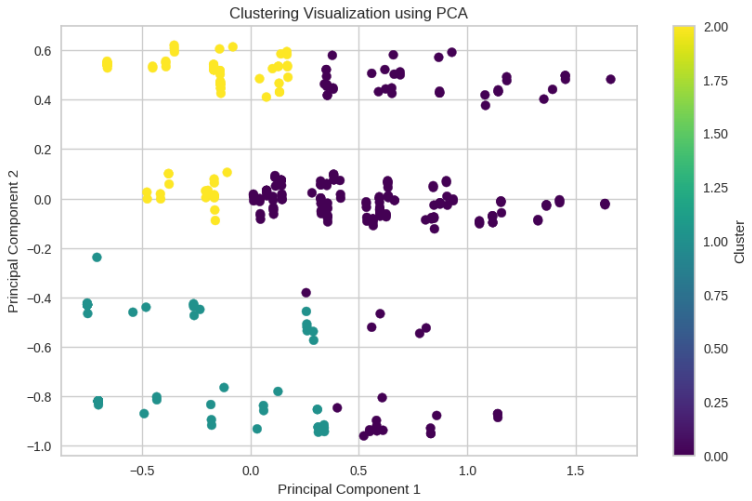
```
Centroid Akhir:
[[ 3.34975369e-01  5.76354680e-01  6.45320197e-01  2.71111893e-01
   3.94088670e-01  3.30049261e-01  6.60098522e-01  1.38186936e-01
   2.60507965e-01  5.99033971e-01  2.27112789e-03]
 [ 8.63636364e-01  9.81818182e-01  1.00000000e-01  3.05194805e-01
   5.45454545e-02  2.72727273e-02  1.27272727e-01  4.09090909e-01
   5.81818182e-01 -6.66133815e-16  9.09090909e-03]
 [ 9.23566879e-01  8.91719745e-01  7.00636943e-02  2.90263876e-01
   9.55414013e-02  7.00636943e-02  1.14649682e-01  2.06799570e-02
   3.30879312e-02  9.45818513e-01  4.13599140e-04]]
Labels:
[0 2 1 2 0 1 0 2 2 0 0 2 1 1 0 0 2 0 2 2 0 0 1 0 0 0 0 0 2 1 2 1 2 0 1 1 1
 0 2 0 1 1 1 1 2 1 0 0 1 0 2 0 1 1 1 2 1 2 2 1 2 0 1 1 2 2 0 0 1 2 0 2 2 1
 2 2 2 1 0 1 2 2 2 1 2 2 0 1 2 2 2 2 0 2 0 0 0 2 0 2 0 0 0 2 0 0 0 0 0 0 1
 0 1 0 0 2 0 0 0 1 1 1 0 2 0 2 2 0 0 0 0 0 0 1 0 0 2 2 0 2 2 1 2 0 0 2 0 2
 0 0 0 0 0 2 0 2 2 2 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 2 0 0 2 0 0 0 0 0 2 0 2 0 0 0 0 0 0 0 0 0 0 2 0 2 0 2 0 0 0 0 0 0 0 0 2 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0
 0 0 2 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 2 0
 0 0 0 2 0 0 0 0 0 0 0 0 0 2 2 2 2 1 1 2 2 2 1 1 2 2 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 2 2 1 2 2 2 2 2 1 2 2 1 2 1 2 2 2 2 2 1 2 2 2 1 2 1 1 2 2 2 2 2 2 1
 2 1 2 1 2 2 2 2 2 2 2 1 1 1 1 2 1 2 2 2 2 2 2 2 2 2 2 1 2 1 2 1 2 1 2 2 1
 1 2 1 2 2 2 1 2 2 1 1 2 2 2 2 1 1 2 2 2 2 1 1 2 2 1 1 2 1 2 1 2 1 1 1 2 2
 1 1 1 2 1 2 1 1 1 2 2 2 1 1 2 1 1 1 1 1 1 1 1 2 1 1]
Jarak ke centroid:
[[1.20355427 1.78700638 1.33001573]
 [1.33619923 1.55745748 0.99651476]
 [1.45794201 1.42469486 1.89589108]
 ...
 [1.40937021 1.24813455 0.23608609]
 [1.63198541 0.62785627 1.37156765]
 [1.63198541 0.62785627 1.37156765]]
```

Gambar 7 Perbarui *Centroid*

Centroid akhir merupakan koordinat titik pusat dari setiap *cluster* setelah iterasi Algoritma *K-Means* selesai. *Centroid* tersebut merupakan titik rata-rata dari semua titik data dalam *cluster* tersebut.

e. Hasil clustering menggunakan K-Means

Berdasarkan hasil *clustering* dari Algoritma *K-Means* menunjukkan jumlah 203 data pada cluster 0 dengan memiliki karakteristik berisi pendidikan terakhir SMA/SMK, memiliki pengalaman bekerja dan durasi pengangguran yang pendek, 110 data pada cluster 1 berisi rata-rata pendidikan terakhir Sarjana (S1), tidak memiliki pengalaman kerja dan durasi pengangguran yang panjang serta 157 data pada cluster 2 yang berisi penganggur dengan pendidikan terakhir SMA/SMK tanpa pengalaman kerja. Adapun plot dari Algoritma *K-Means* berdasarkan visualisasi pada Gambar 8.



Gambar 8 Plot *K-Means*

Pada visualisasi plot *K-Means Clustering* menunjukkan bahwa data berhasil dibagi menjadi beberapa *cluster* yang berbeda berdasarkan kesamaan karakterisriknya, setiap titik mewakili satu data point dan warna yang berbeda mewakili cluster yang berbeda.

2) Algoritma Hierarchical Clustering

Pada langkah 1 dan 2 Algoritma *Hierarchical Clustering*, dilakukan identifikasi item dengan jarak terdekat berdasarkan *Euclidean Distance*. Selanjutnya menerapkan *ward's method* untuk membentuk *cluster* berdasarkan jarak yang sudah dihitung.

a. Jarak setiap cluster

Pada langkah ini mengukur jarak terdekat antara setiap pasangan cluster dalam dataset yang sudah ditetapkan nilai clusternya. Hal ini dapat memahami seberapa terpisah atau dekat pada setiap cluster yang ada. Selain itu, pada tahap ini juga untuk memahami distribusi data setelah melakukan *clusterimg*. Hasilnya dapat dilihat pada Gambar 9.

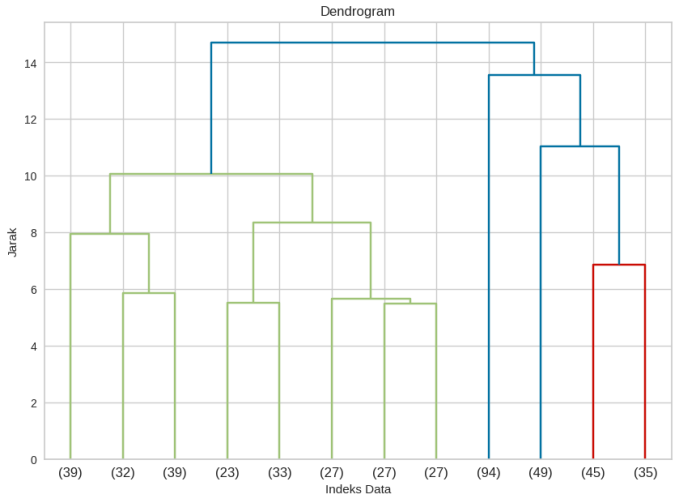
```
Jarak antar Cluster 0 - Cluster 1: 0.9511047628205896
Jarak antar Cluster 0 - Cluster 2: 0.5241470595070435
Jarak antar Cluster 1 - Cluster 2: 0.9511047628205896
```

Gambar 9 Jarak Terdekat *Hierarchical*

Gambar tersebut menunjukkan bahwa *cluster* 0 ke 1, dan 1 ke 2 memiliki karakterisitik yang mirip satu sama lain. Sedangkan *cluster* 0 ke 2 memiliki jarak yang lebih jauh diantara *cluster* lainnya, hal tersebut menunjukkan jarak *cluster* 0 ke 2 memiliki karakteristik yang berbeda dibandingkan *cluster* lainnya.

b. Visualisasi dendogram

Pada visualisasi ini, dendogram akan menunjukkan bagaimana data dikelompokkan secara hierarkis dengan bertahap. Adapun visualisasi tersebut seperti pada Gambar 10.

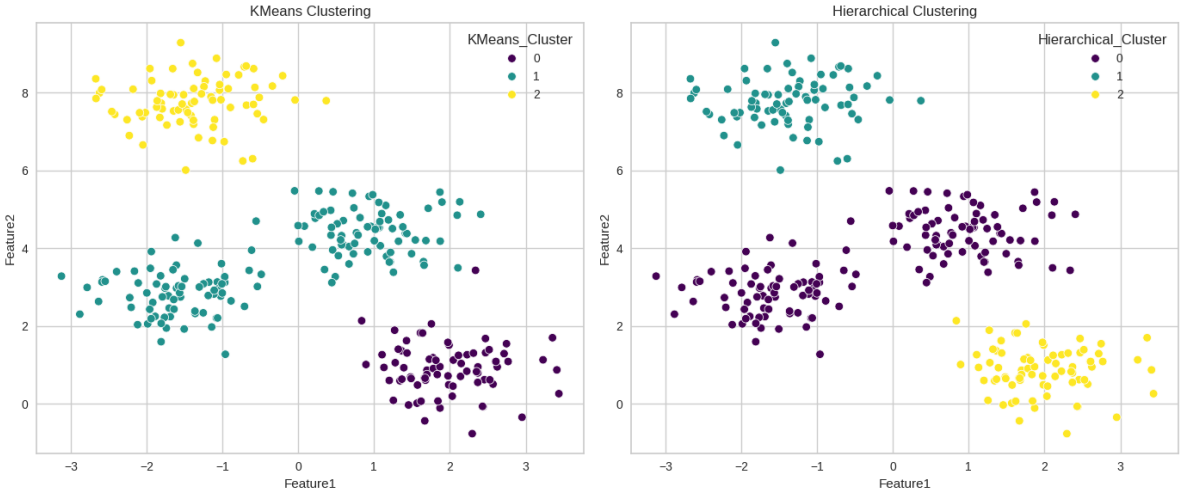


Gambar 10 Visualisasi Dendogram

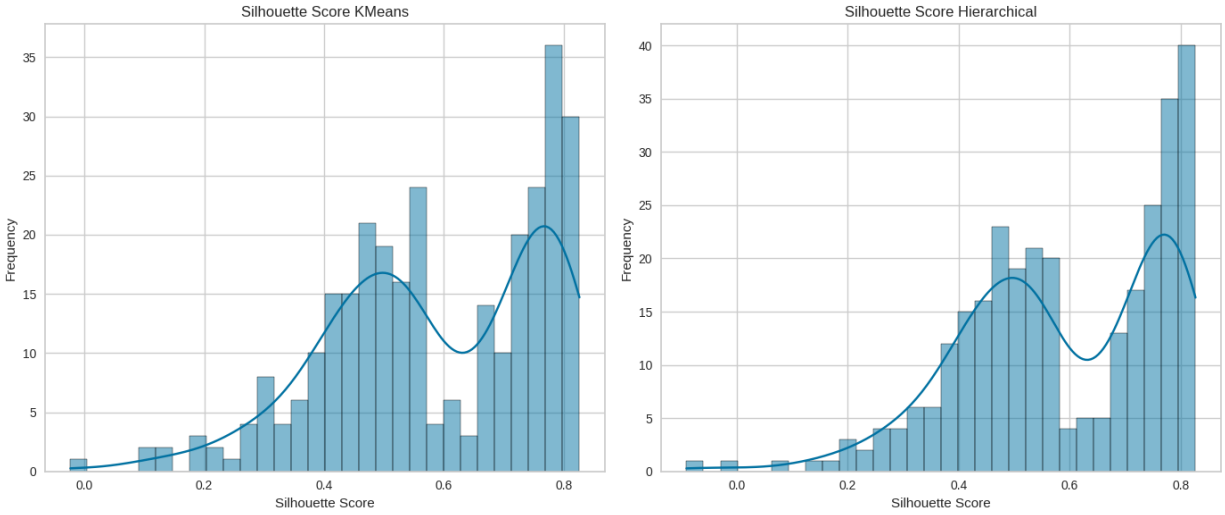
Pada pengelompokkan *Hierarchical Clustering* memiliki label baru berdasarkan label yang sudah ada sebelumnya yaitu hasil pengelompokkan *K-Means Clustering*. Label baru tersebut dibuat kolom baru dengan nama Hierarchical_Cluster dan menghasilkan 108 data pada *cluster* 0 yang memiliki karakteristik yang sama pada *cluster* 1 di *K-Means Clustering*, 205 data pada *cluster* 1 berisi karakteristik yang sama dengan *cluster* 0 pada *K-Means Clustering* dan 157 data pada *cluster* 2 yang memiliki karakteristik yang hampir sama dengan *cluster* 2 pada *K-Means Clustering* tetapi memiliki perbedaan pada durasi pengangguran.

3.4 Evaluasi

Silhouette Score digunakan untuk evaluasi model dengan menghitung seberapa baik *cluster* terpisah satu sama lain. Selain itu, evaluasi ini digunakan untuk menilai kualitas *clustering* antara Algoritma *K-Means Clustering* dan *Hierarchical Clustering*. Hal tersebut dapat dilihat pada Gambar 11 dan Gambar 12.



Gambar 11. Visualisasi *Clustering*



Gambar 12. *Silhouette* Plot

Pada Gambar 11 menunjukkan bahwa kedua algortima yakni *K-Means Clustering* dan *Hierarchical Clustering* berhasil mengidentifikasi 3 *cluster* dalam data. Terdapat sedikit perbedaan dari hasil *clustering* tersebut dikarenakan kedua algoritma tersebut memiliki pendekatan yang berbeda. *K-Means Clustering* menghasilkan *cluster* yang lebih bulat dan terpisah sedangkan *Hierarchical Clustering* menghasilkan *cluster* dengan bentuk yang lebih kompleks. Berikutnya Gambar 12 menunjukkan kedua algoritma tersebut menghasilkan hasil yang cukup baik dan memiliki nilai *silhouette score* yang serupa yaitu 58 persen.

Selain digunakan untuk mengelompokkan tingkat pengangguran berdasarkan hasil penelitian Almeira dkk (2021) [15], Algoritma *K-Means* juga dapat menghasilkan cluster yang optimum dalam mengelompokkan pengangguran berdasarkan karakteristiknya pada penelitian ini serta memiliki hasil yang sama dalam membentuk cluster dengan Algoritma *Hierarchical Clustsering*.

4. KESIMPULAN

Berdasarkan penelitian yang dilakukan, sekiranya terdapat 2,81 persen pengangguran pada dua kecamatan yang ada di Karawang yaitu kecamatan Karawang Barat dan Teluk Jambe Timur. Hal tersebut dapat digunakan

sebagai acuan bagi pemerintah untuk mengimplementasikan kebijakan yang terarah demi menurunkan angka pengangguran yang ada di Karawang. Selanjutnya, Algoritma *K-Means Clustering* dan *Hierarchical Clustering* mampu mengelompokkan data pengangguran dengan 3 *cluster*. Meskipun memiliki jumlah *cluster* yang sama, distribusi data dalam *cluster* memiliki perbedaan antara kedua algoritma. Pada Algoritma *K-Means Clustering* menghasilkan *cluster* yang seimbang dikarenakan melakukan pendekatan *centroid* tetapi membutuhkan penentuan jumlah *cluster* di awal. Sedangkan Algoritma *Hierarchical Clustering* tidak memerlukan penentuan jumlah *cluster* di awal sehingga *cluster* yang terbentuk merupakan pengelompokan alami dari data. *K-Means* menghasilkan perbedaan yang jelas pada faktor pengalaman kerja dan durasi pengangguran sedangkan *Hierarchical Clustering* dapat lebih baik menangkap hubungan dalam data seperti peran pendidikan dan *internship*. Selanjutnya pada visualisasi *Silhouette Score* yang menunjukkan nlai yang sama yaitu 0.58 dapat disimpulkan bahwa Algoritma *K-Means Clustering* maupun *Hierarchical Clustering* menghasilkan kualitas *clustering* yang serupa dengan 3 *cluster*. Hal tersebut berarti kedua algoritma yang digunakan dapat mengelompokkan data dengan cara yang sangat mirip berdasarkan evaluasi tersebut. Disarankan untuk penelitian selanjutnya dapat menganalisis lebih detail mengenai faktor-faktor apa saja yang menyebabkan seseorang menganggur.

REFERENSI

[1] T. Rahmawati and N. Nurwati, "Pengaruh Pertumbuhan Industri terhadap Pengangguran Terbuka di Kabupaten Karawang," *J. Polit. Indones.*, vol. 6, no. 1, pp. 51–61, 2021, doi: 10.35706/jpi.v6i1.5165.

[2] StudySmarter, "Unemployment Rate," StudySmarter.

[3] M. D. Kahfi, F. R. Umbara, and H. Ashaury, "Prediksi Pengagguran Menggunakan Decision Tree Dengan Algoritma C5.0 Pada Data Penduduk Kecamatan Caringin Kabupaten Bogor," *Informatics Digit. Expert*, vol. 4, no. 2, pp. 75–80, 2023, doi: 10.36423/index.v4i2.913.

[4] A. Mayssara A. Abo Hassanin Supervised, "Dampak Penurunan Ekonomi Karena Pandemi Covid-19 Terhadap Jumlah Kriminalitas Di Kelurahan Nagasari Kabupaten Karawang Dalam Perspektif Kriminologi," *Pap. Knowl. . Towar. a Media Hist. Doc.*, pp. 1147–1159, 2014.

[5] K. Kusano, A. K. Uskul, and M. Kemmelmeier, "Suicide during the COVID-19 pandemic: Uncovering demographic and regional variation in the United States and associations with unemployment and depression," *Curr. Res. Ecol. Soc. Psychol.*, vol. 5, no. August, p. 100144, 2023, doi: 10.1016/j.cresp.2023.100144.

[6] S. Nofita, H. H. Hanny, and M. S. Amril, "Implementasi Clustering Data Kasus Covid 19 Di Indonesia Menggunakan Algoritma K-Means," *Bianglala Inform.*, vol. 11, no. 1, pp. 7–12, 2023.

[7] I. Rahma, P. P. Arhandi, and A. T. Firdausi, "Penerapa Metode Hierarchical Clustering Dan K-Means Clustering Untuk Mengelompokkan Potensi Lokasi Penjualan Linkaja," *J. Inform. Polinema*, vol. 6, no. 1, pp. 15–22, 2020, doi: 10.33795/jip.v6i1.287.

[8] A. Abdulhafedh, "Incorporating K-means, Hierarchical Clustering and PCA in Customer Segmentation," *J. City Dev.*, vol. 3, no. 1, pp. 12–30, 2021, doi: 10.12691/jcd-3-1-3.

[9] F. Badri and A. Habibi, "Implementasi Metode K-Means Clustering Analysis pada Pengelompokan Pengangguran di Indonesia sebagai Dampak dari Pandemi Covid-19," *Ilk. J. Comput. Sci. Appl. Informatics*, vol. 4, no. 2, pp. 171–179, 2022, doi: 10.28926/ilkomnika.v4i2.471.

[10] A. M. Siregar and D. Wahiddin, "Penerapan Algoritma K-Means Dalam Mengurangi Tingkat Buta Aksara Di Indonesia Sebagai Penunjang Keputusan," *Sci. Student J. ...*, vol. 1, pp. 55–60, 2020.

[11] A. R. Junior, H. H. Handayani, A. Fitri, and N. Masruriyah, "Analisis Sentimen Menggunakan Algoritma K-Means untuk Mengetahui Kalimat Positif maupun Negatif pada Buletin APTIKOM," *Sci. Student J. Information, Technol. Sci.*, vol. III, no. 1, p. 113, 2022.

[12] S. C. Dewi, A. M. Siregar, and D. S. Kusumaningrum, "Pengelompokan Jumlah Sumber Daya Manusia Kesehatan Puskesmas untuk Menunjang Pemerataan pada Provinsi Jawa Tengah Menggunakan Algoritma K-Means," *Sci. Student J. Information, Technol. Sci.*, vol. 1, no. 2, pp. 86–94, 2020.

[13] L. Adeliana, A. M. Siregar, and D. S. Kusumaningrum, "Pengelompokan Kabupaten dan Kota di Indonesia Berdasarkan Hasil Produksi Daging Sapi Menggunakan Algoritma K-Means dan K-Medoids," *Sci. Student J. Information, Technol. Sci.*, vol. II, no. 1, pp. 15–21, 2021.

[14] M. A. Rofik, A. M. Siregar, and ..., "Perbandingan Tingkat Kepuasan Siswa Terhadap Pelayanan

Sekolah Menggunakan Algoritma K-Means Dan K-Medoids," ... *Student J.* ..., vol. II, pp. 21–30, 2021.

[15] D. Almeira and G. Graciella Juanda, "Analisis Multidimensional Scaling dan k-Means Clustering untuk Pengelompokan Provinsi Berdasarkan Tingkat Pengangguran," *E-Prosiding Nas. | Dep. Stat. FMIPA Univ. Padjadjaran*, vol. 10, p. 08, 2021.

[16] A. Pramudya, I. Maulana, and R. Mayasari, "Pengelompokan Hasil Belajar Siswa Sdn Tunas Jaya Dengan Algoritma K-Means," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 6, pp. 3960–3967, 2024, doi: 10.36040/jati.v7i6.7970.

[17] D. Amelia, T. N. Padilah, and A. Jamaludin, "Optimasi Algoritma K-Means Menggunakan Metode Elbow dalam Pengelompokan Penyakit Demam Berdarah Dengue (DBD) di Jawa Barat," *J. Ilm. Wahana Pendidik.*, vol. 8, no. 11, pp. 207–215, 2022.

[18] I. Alfian, "Penerapan Metode K-Means Dalam Melakukan Pengelompokan Bencana Alam di Indonesia Dilakukan dengan Memanfaatkan Teknik Text Mining," *J. Algoritm.*, vol. 20, no. 1, pp. 139–147, 2023, doi: 10.33364/algoritma/v.20-1.1275.

[19] N. Pooja, M. Saputra, S. Aisyah, and P. Juanta, "Implementasi Data Mining Clustering Data Valuasi Ekspor Kertas Indonesia Menggunakan Algoritma K-Means," *J. Sist. Inf. dan Ilmu Komput. Prima(JUSIKOM PRIMA)*, vol. 5, no. 2, pp. 86–90, 2022, doi: 10.34012/jurnalsisteminformasidanilmukomputer.v5i2.2372.