Analisis Faktor Risiko Tingkat Kematian Pasien COVID-19 Menggunakan Algoritma K-Means Clustering

1st M. Rifadh Asjad
Sistem Informasi,
Telkom University
Bandung , Indonesia
rifadasjad@student.telkomuniveristy.ac.id1

2nd Rd.Rohmat Saedudin,S.T.,M.T.,Ph.D.
Sistem Informasi,
Telkom University
Bandung , Indonesia
rdrohmat@telkomuniversity.ac.idqw

3rd Taufik Nur Adi, S.Kom., M.T., Ph.D.
Sistem Informasi
Telkom University
Bandung , Indonesia
taufiknur@telkomuniversity.ac.id

Abstrak

Pandemi COVID-19 yang berlangsung sejak 2020 hingga 2023 telah menyebabkan jutaan kasus infeksi dan kematian di seluruh dunia, termasuk di Indonesia. Penyebaran yang tidak merata dan tingkat kematian yang bervariasi antar wilayah menunjukkan adanya perbedaan faktor risiko yang mempengaruhi kondisi pasien. Penelitian ini bertujuan untuk mengidentifikasi faktor-faktor tersebut—seperti usia, jenis kelamin, dan durasi sakit—guna mengelompokkan pasien berdasarkan risiko yang berdampak pada tingkat kematian dengan menggunakan algoritma *K-Means Clustering* . Data yang digunakan berasal dari laporan resmi pasien di Kota Depok, yang melalui tahapan preprocessing seperti pembersihan data, seleksi fitur, dan normalisasi. Hasil analisis menunjukkan terbentuknya empat kelompok utama, dengan Cluster 3 (rata-rata usia 56 tahun) memiliki status akhir dominan "meninggal", meskipun durasi sakit relatif singkat. Hal ini mengindikasikan adanya komorbid atau keterlambatan penanganan. Dengan demikian, pendekatan ini dapat digunakan sebagai dasar pengembangan sistem peringatan dini bagi pasien berisiko tinggi, serta membantu alokasi sumber daya medis secara lebih tepat sasaran

Kata kunci: COVID-19, *K-Means Clustering,* Faktor Risiko Kematian

I. PENDAHULUAN

Pandemi virus SARS-CoV-2 atau yang lebih dikenal sebagai COVID-19 telah memberikan dampak besar terhadap sistem kesehatan global sejak awal 2020. Di Indonesia, jumlah kasus dan kematian terus mengalami fluktuasi, dengan tingkat kematian yang tidak merata di setiap daerah. Beberapa faktor seperti usia, jenis kelamin, kondisi penyerta (komorbid), dan durasi perawatan menjadi variabel penting yang memengaruhi prognosis pasien

Penelitian ini mengusulkan penerapan algoritma K-Means Clustering untuk mengelompokkan pasien COVID-19 berdasarkan faktor klinis dan demografis. Metode ini dipilih karena kemampuannya dalam menemukan pola tersembunyi dalam data besar tanpa memerlukan label awal. Dengan memanfaatkan kerangka kerja CRISP-DM (Cross-Industry Standard Process for Data Mining) , proses analisis dilakukan secara sistematis, mulai dari pemahaman konteks kesehatan hingga evaluasi hasil

II. KAJIAN TEORI

A. COVID 19

COVID-19 merupakan sebuah penyakit menular yang mengakibatkan penderitanya mengalami infeksi pada sistem pernafasan. Penyakit ini diakibatkan oleh sebuah virus yang bernama Severe Acute respiratory Syndrome Coronavirus 2, atau biasa disingkat menjadi SARS-Cov 2. Virus ini dapat menyerang siapa saja baik itu bayi, anak-anak, dewasa, lansia, ibu hamil, dan ibu menyusui (Handayani, 2020). Hingga saat ini, penyebab virus corona belum diketahui, namun virus tersebut diketahui menular melalui hewan dan dapat ditularkan dari satu spesies ke spesies lainnya, termasuk manusia. Virus corona diketahui berasal dari kota Wuhan di Tiongkok pada bulan Desember 2019

Pengumpulan data pasien COVID-19 selama durasi tahun 2020-2022 berasal dari dinas kesehatan kota Depok. Dengan memadukan teori dan metode yang telah dibahas, penelitian ini diharapkan dapat memberikan wawasan baru dalam pemetaan risiko kematian akibat COVID-19 di Kota Depok

Pengumpulan data pasien COVID-19 selama durasi tahun 2020-2022 berasal dari dinas kesehatan kota Depok. Dengan memadukan teori dan metode yang telah dibahas, penelitian ini diharapkan dapat memberikan wawasan baru dalam pemetaan risiko kematian akibat COVID-19 di Kota Depok

B. Data Mining

Data mining adalah teknik untuk menggali informasi penting dari dataset besar guna menemukan pola, tren, atau hubungan tersembunyi. Contohnya, data mining dapat digunakan untuk menganalisis hubungan antara usia pasien dan tingkat kematian akibat COVID-19

Menurut Yuli Asriningtias dan Rodhya Mardhiyah (2014), data mining diterapkan melalui pendekatan Knowledge Discovery in Databases (KDD) yang mencakup tahapan pemilihan data, pra-pemrosesan, transformasi, clustering, dan interpretasi hasil.

Contoh variabel penelitian yang didapatkan :

1. Variabel independen: Karakteristik data *COVID-19* (misalnya Tanggal Lapor, tanggal lahir, tanggal sembuh/meninggal, jenis kelamin, dan status akhir pasien).
2. Variabel dependen: Usia, durasi masa sakit, status akhir pasien dan faktor risiko terjadi peningkatan tingkat kematian pasien *COVID-19*.

Teknik data mining digunakan untuk menganalisis faktor risiko tingkat kematian pasien COVID-19 di Kota Depok berdasarkan variabel seperti usia, jenis kelamin, durasi sakit, dan status akhir pasien. Hasilnya diharapkan dapat menjadi dasar pengambilan keputusan dalam mitigasi pandemi

C. CRISP-***DM***

Cross-Industry Standard Process for Data Mining (CRISP-DM) adalah metodologi standar lintas industri untuk data mining, yang menawarkan pendekatan terstruktur dan sistematis untuk proyek data mining. CRISP-DM terdiri dari enam tahapan utama, yaitu:

1. *Business Understanding*
2. *Data Understanding*
3. *Data Preparation*
4. *Modelling*
5. *Evaluation*
6. *Deployment*

D. ***K***-Means ***Clustering***

K-Means Clustering adalah algoritma yang digunakan untuk mengelompokkan data berdasarkan kesamaan karakteristik. Algoritma ini termasuk dalam kategori unsupervised learning, yang berarti tidak memerlukan label atau target dalam proses pengelompokannya. K-Means bekerja dengan membagi dataset menjadi beberapa cluster (kelompok) tertentu, dimana setiap data dalam cluster memiliki kemiripan yang tinggi satu sama lain, tetapi berbeda secara signifikan dengan data dari cluster lain

Berikut adalah langkah-langkah utama dalam K-Means sebagai berikut :

1. Inisialisasi *Centroid* : menentukan jumlah *cluster* (k) dan memilih k *centroid* awal secara acak.
2. Alokasi Data ke *Cluster* : menghitung jarak setiap data ke semua centroid dan mengalokasikan data ke *cluster* dengan *centroid* terdekat (biasanya menggunakan jarak Euclidean).
3. Perbarui *Centroid* : menghitung ulang *centroid* untuk setiap *cluster* berdasarkan rata-rata data dalam *cluster* tersebut.
4. Iterasi : Mengulangi langkah 2 dan 3 hingga tidak ada perubahan signifikan pada centroid atau konvergensi tercapai.

Untuk menentukan jumlah cluster optimal (k), metode Elbow Method sering digunakan. Metode ini melibatkan:

1. Menghitung *Sum of Squared Errors* (SSE) untuk berbagai nilai k.
2. Membuat grafik SSE vs k dan memilih nilai k pada titik "siku" (*elbow* point), yaitu titik di mana penurunan SSE mulai melambat.

E. Hubungan *CRISP -DM dengan K-Means Clustering*

Penerapan CRISP-DM dalam penelitian ini memberikan kerangka kerja yang jelas untuk analisis data menggunakan K-Means Clustering . Setiap tahapan CRISP-DM saling terintegrasi dengan proses clustering:

a. Data *Understanding* dan *Preparation* : memastikan bahwa dataset siap untuk dimodelkan.
b. *Modeling* : K-Means *Clustering* digunakan sebagai alat utama untuk mengelompokkan data.
c. Evaluation : Hasil *clustering* dievaluasi untuk memastikan kualitas model.
d. *Deployment* : Hasil clustering diimplementasikan dalam bentuk rekomendasi praktis.ahuan, bisnis,

Artikel ditulis dalam ukuran kertas A4, maksimal 5000 kata dan ditulis menggunakan spasi 1

III. METODE

Penelitian ini menggunakan pendekatan kuantitatif dengan metode data mining , mengikuti kerangka kerja CRISP-DM yang terdiri dari enam tahap: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, dan Deployment.

1. Business Understanding

konteks ini, tujuan utama adalah membantu instansi kesehatan mengidentifikasi kelompok pasien berisiko tinggi agar dapat dilakukan tindakan preventif secara cepat dan tepat.

2. Data *Understanding*

Data primer diperoleh dari laporan resmi pasien COVID-19 di Kota Depok, dengan jumlah total 190.579 baris dan 33 kolom. Beberapa variabel utama yang digunakan meliputi: usia, jenis kelamin, hasil laboratorium, durasi sakit, dan status akhir pasien (sembuh/meninggal).

3. Data Preparation

Tahap ini mencakup tiga sub-proses utama:

1. Pembersihan Data :
   a. Menghapus bari duplikat.
   b. Mengisi nilai yang hilang (missing values) menggunakan imputasi media (untuk numerik) dan modus (untuk kategorik).
   c. Memperbaiki format tanggal dan kesalahan input (typo).

2. Seleksi Fitur :
   Variabel yang dipilih berdasarkan relevansi terhadap tujuan penelitian :
   a. Usia (dihitung dari tanggal lahir dan tanggal lapor)
   b. Jenis Kelamin
   c. Durasi Sakit (tanggal sembuh/meninggal – tanggal lapor)
   d. Hasil lab akhir
   e. Jenis pemeriksaan
   f. Status akhir pasien

Kolom seperti Tanggal Hasil Lab 5 dihapus karena hanya memiliki 2.178 data non-null, sehingga dianggap tidak representatif.

3. Normalisasi Data
   Untuk menghindari bias akibat perbedaan skala, data dinormalisasi menggunakan Standard Scaler , yang mengubah distribusi menjadi mean = 0 dan standar deviasi = 1. Variabel kategorikal seperti jenis kelamin diubah menjadi numerik menggunakan One-Hot Encoding .

4. Modeling

Algoritma K-Means diterapkan untuk mengelompokkan data menjadi beberapa cluster. Jumlah cluster optimal ditentukan menggunakan Elbow Method dan Silhouette Score . Proses iteratif dilakukan untuk nilai k = 2 hingga 6, dan dipilih k = 4 karena memberikan elbow point yang jelas dan nilai Silhouette tertinggi (0,48).

5. Evaluation

Setelah model clustering selesai diimplementasikan, hasilnya dievaluasi untuk memastikan kualitas cluster yang dihasilkan. Evaluasi dilakukan melalui dua pendekatan utama:

- Silhouette Score :
  Metrik ini digunakan untuk mengevaluasi seberapa baik data dikelompokkan dalam cluster. Nilai Silhouette Score berkisar antara -1 hingga 1, di mana nilai mendekati 1 menunjukkan pengelompokan yang baik
- Analisis Deskriptif :
  Setiap cluster dianalisis secara deskriptif untuk memahami karakteristik utamanya. Hasil evaluasi menunjukkan:

a. Cluster 0 : Usia luas, durasi sakit Panjang dan status akhir sembuh.
b. Cluster 1 : Usia muda, durasi sakit singkat dan status akhir sembuh.
c. Cluster 2 : Usia lanjut, durasi sakit singkat dan status akhir sembuh.
d. Cluster 3 : Usia luas, durasi sakit bervariasi dan status akhir, meninggal

Hasil evaluasi ini memberikan wawasan yang berharga tentang faktor-faktor yang memengaruhi tingkat kematian pasien COVID-19.

5. Deployment

Tahap Deployment bertujuan untuk menerapkan hasil penelitian ke dalam praktik nyata. Implementasi hasil mencakup beberapa aspek berikut:

1. Rekomendasi Mitigasi Pandemi Berbasis Risiko :
   Hasil clustering digunakan untuk merancang strategi mitigasi pandemi yang lebih tepat sasaran. Misalnya, pasien dengan risiko tinggi dapat diprioritaskan untuk mendapatkan penanganan medis lebih intensif.
2. Alokasi Sumber Daya Medis :
   Informasi dari hasil clustering membantu pihak rumah sakit atau fasilitas kesehatan dalam mengalokasikan sumber daya medis secara lebih efektif kepada kelompok pasien dengan risiko tinggi.

3. Manfaat Penelitian :
   a. Bagi Universitas Telkom: Penelitian ini memberikan tambahan referensi dan pengetahuan dalam bidang analisis data kesehatan.
   b. Bagi Peneliti Lain: Hasil penelitian dapat digunakan sebagai dasar untuk penelitian lebih lanjut terkait teknik clustering pada data pasien.
   c. Bagi Masyarakat: Informasi yang dihasilkan membantu meningkatkan pemahaman masyarakat tentang bahaya COVID-19 dan pentingnya mitigasi berbasis risiko.

Dengan demikian, penelitian ini tidak hanya memberikan kontribusi teoritis tetapi juga praktis dalam upaya menghadapi tantangan pandemi.

Teknik pengumpulan data dalam penelitian ini dilakukan secara sistematis dan terstruktur untuk memastikan keakuratan serta validitas informasi yang diperoleh. Langkah-langkah yang dilakukan meliputi: pengumpulan data dilakukan melalui langkah-langkah berikut :

1. Surat Permohonan :
   Penelitian dimulai dengan mengirimkan surat permohonan resmi kepada Dinas Kesehatan Kota Depok . Surat ini bertujuan untuk meminta izin penggunaan data pasien COVID-19 yang relevan dengan penelitian, termasuk variabel seperti tanggal laporan, jenis kelamin, usia, status akhir pasien (sembuh/meninggal), durasi sakit, dan hasil laboratorium.
2. Rekomendasi Penelitian :
   Setelah surat permohonan dikirimkan, penelitian mendapatkan rekomendasi dari Badan Kesatuan Bangsa dan Politik (KESBANGPOL) Kota Depok . Rekomendasi ini menjadi salah satu syarat administratif yang menjamin legalitas dan kredibilitas penelitian
3. Surat Pernyataan :
   Untuk menjaga kerahasiaan data, peneliti menandatangani surat pernyataan kerahasiaan data . Hal ini memastikan bahwa data yang diberikan oleh Dinas Kesehatan Kota Depok hanya digunakan untuk keperluan penelitian dan tidak disebarluaskan kepada pihak lain tanpa izin.
4. Pembimbingan Teknis :
   Selama proses pengumpulan data, peneliti mendapatkan pembimbingan teknis dari pembimbing lapangan yang ditetapkan oleh Dinas Kesehatan Kota Depok. Pembimbingan ini bertujuan untuk memastikan bahwa data yang dikumpulkan sesuai dengan kebutuhan penelitian dan memiliki tingkat validitas yang tinggi.

4. Metode Pengolahan Data

Metode pengolahan data dalam penelitian ini mencakup tahapan sistematis untuk menyiapkan dataset agar siap dianalisis. Proses dimulai dengan pembersihan data untuk mengidentifikasi, memperbaiki, dan menghapus data yang tidak akurat, tidak lengkap, atau inkonsisten, termasuk penghapusan duplikasi dan pelengkapan data yang hilang.

Selanjutnya, seleksi fitur dilakukan untuk memilih variabel relevan, mengurangi dimensi data, meningkatkan akurasi model, serta mengeliminasi noise atau outlier. Kemudian, normalisasi data menggunakan teknik One-Hot Encoding mengubah data kategorik menjadi numerik agar lebih mudah diproses oleh algoritma clustering.

Terakhir, clustering dengan K-Means diterapkan untuk mengelompokkan data berdasarkan kesamaan karakteristik. Jumlah cluster optimal ditentukan menggunakan Elbow Method , sehingga hasil clustering memiliki kualitas dan interpretasi yang baik. Melalui metode ini, dataset disiapkan secara optimal untuk menghasilkan analisis yang akurat, valid, dan relevan.

5. Pembersihan Data

Dalam penelitian ini, proses pembersihan data dilakukan berdasarkan dataset yang diperoleh dari Dinas Kesehatan Kota Depok, mencakup informasi terkait pasien COVID-19 selama periode 2020–2022 dan terlihat pada table dibawah ini. Dataset awal terdiri dari 33 kolom dengan total 190.579 baris .

Tabel 4. *3* Hasil Data Clustering

Data #	Columns	Non-Null Count	Dtype
0	No.	190579 non-null	int64
1.	NIK/No Passport	190579 non-null	Object
2.	Tanggal Lapor	190579 non-null	Object
3.	Provinsi	190579 non-null	Object
4.	Kabupaten/Kota	190579 non-null	Object
5.	Nama	190579 non-null	Object
6.	Jenis Kelamin	190579 non-null	Object
7.	Tanggal Lahir	189620 non-null	Object
8.	Usia Tahun	189928 non-null	Object
9.	Warganegara	190518 non-null	Object
10.	Alamat Desa/Kelurahan	189757 non-null	Object
11.	Alamat Kecamatan	189769 non-null	Object
12.	Alamat Kabupaten/Kota	190576 non-null	Object
13.	Alamat Provinsi	190578 non-null	Object

14.	Status Epidemiologi	190579 non-null	Object
15.	Tanggal Ambil Spesimen	190039 non-null	Object
16.	Tanggal Hasil Lab 1	190109 non-null	Object
17.	Hasil Lab. 1	190579 non-null	Object
18.	Tanggal Hasil Lab 2	62909 non-null	Object
19.	Hasil Lab. 2	62835 non-null	Object
20	Tanggal Hasil Lab 3	19890 non-null	Object
21	Hasil Lab. 3	19877 non-null	Object
22.	Tanggal Hasil Lab 4	6249 non-null	Object
23.	Hasil Lab. 4	6246 non-null	Object
24.	Tanggal Hasil Lab 5	2180 non-null	Object
25.	Hasil Lab. 5	2178 non-null	Object
26.	Hasil Lab Akhir	190579 non-null	Object
27.	Asal Faskes	189738 non-null	Object
28.	Lab Pemeriksa	166649 non-null	Object
29.	Jenis Pemeriksaan	190579 non-null	Object
30.	Status Akhir Pasien	190561 non-null	Object
31.	Tanggal Sembuh	188276 non-null	Object
32.	Tanggal Meninggal	2285 non-null	Object

6. Seleksi Fitur Data

Seleksi fitur merupakan proses pemilihan variabel atau atribut yang paling relevan dan bermanfaat untuk analisis. Dalam penelitian ini, fitur yang dipilih meliputi **usia, jenis kelamin, durasi sakit, hasil lab, jenis pemeriksaan, dan status akhir pasien**. Fitur-fitur ini dianggap memiliki kontribusi signifikan terhadap tujuan analisis, yaitu mengidentifikasi faktor risiko tingkat kematian pasien COVID-19

7. Normalisasi Data

Dalam penelitian ini, normalisasi data dilakukan melalui serangkaian langkah-langkah sistematis, yaitu :

1. Membersihkan data : Menghapus kesalahan, duplikat, dan data yang tidak relevan.
2. Mengubah data menjadi format yang seragam dan konsisten : Memastikan bahwa semua data memiliki format yang sesuai untuk analisis.
3. Mengubah skala data : Menyelaraskan skala data agar seragam dan konsisten
4. Menghapus data yang tidak relevan : Mengeliminasi data yang tidak mendukung tujuan analisis atau modeling.

Dalam penelitian ini, metode normalisasi yang digunakan adalah *One-Hot Encoding* . Teknik ini dipilih karena mampu mengubah data kategorikal, seperti jenis kelamin, hasil lab akhir, jenis pemeriksaan, dan status akhir pasien, menjadi format numerik yang dapat diproses oleh *algoritma K-Means Clustering*. Berikut adalah contoh hasil normalisasi data menggunakan One-Hot Encoding

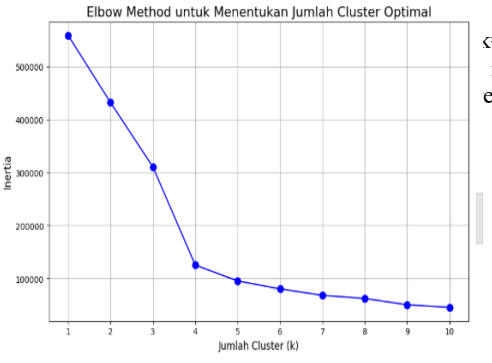
Variabel kategorikal seperti Jenis Kelamin , Hasil Lab Akhir , Jenis Pemeriksaan , dan Status Akhir Pasien telah dikonversi menjadi nilai numerik menggunakan teknik One-Hot Encoding . Contohnya, label "SEMBUH" pada kolom Status Akhir Pasien diubah menjadi nilai numerik 1 . Proses ini memastikan bahwa data dapat diproses dalam algoritma K-Means Clustering secara konsisten tanpa bias atau kesalahan interpretasi

Melalui normalisasi data menggunakan One-Hot Encoding, dataset menjadi lebih seragam dan konsisten, sehingga hasil clustering yang dihasilkan lebih akurat dan relevan. Langkah ini juga membantu mempermudah interpretasi hasil analisis, terutama dalam mengidentifikasi faktor-faktor risiko tingkat kematian pasien COVID-19 di Kota Depok.

8. Clustering dengan K-Means

Untuk menentukan jumlah cluster (k) yang optimal, digunakan metode ***Elbow Method***. Berikut adalah langkah-langkah yang dilakukan dalam proses clustering dengan algoritma K-Means :

1. **Menetapkan Nilai Awal untuk k :**
   Jumlah cluster (k) ditentukan sebagai input awal. Nilai ini biasanya didasarkan pada eksplorasi awal terhadap data atau menggunakan metode seperti *Elbow Method*.
2. **Menjalankan Algoritma *K-Means* :**
   Algoritma dijalankan menggunakan aplikasi seperti *Google Collabs*, yang akan membagi data ke dalam *k-cluster* dan menghitung centroid untuk setiap cluster berdasarkan nilai k yang telah ditentukan.
3. **Menghitung *Sum of Square Errors* (SSE) :**
   Untuk setiap nilai k, dihitung nilai SSE , yaitu jumlah kuadrat jarak antara setiap data dengan centroid cluster-nya. SSE digunakan untuk mengevaluasi seberapa baik data dikelompokkan.
4. **Membuat grafik yang menunjukkan hubungan antara SSE dan nilai :**
   Plot hubungan antara SSE dan nilai k dibuat untuk membantu menentukan jumlah cluster yang optimal. Pada plot ini, nilai k yang optimal biasanya ditunjukkan oleh titik "siku" (*elbow point*), yaitu titik di mana penambahan jumlah cluster tidak lagi memberikan pengurangan SSE yang signifikan
5. **Memilih nilai k yang paling optimal :**
   Berdasarkan hasil plot Elbow Method, nilai k yang paling optimal dipilih sebagai jumlah cluster yang akan digunakan dalam analisis.
6. **Melakukan validasi untuk hasil yang didapat.**
   Setelah menentukan jumlah *cluster*, hasil *clustering* divalidasi untuk memastikan kualitas pengelompokan. Metrik seperti *Silhouette Score* dapat digunakan untuk mengevaluasi seberapa baik data dikelompokkan dalam cluster-cluster yang dihasilkan.
7. **Hasilnya dapat terlihat sesuai gambar berikut :**



IV. HASIL DAN PEMBAHASAN

4.1. Hasil Implememtasi K-Means

Setelah proses preprocessing, data yang digunakan untuk clustering berjumlah 188.276 pasien. Dari hasil Elbow Method, dipilih k = 4 sebagai jumlah cluster optimal. Berikut adalah deskripsi masing-masing cluster:

Cluster 0
34 tahun
45 hari
Sembuh
Pasien muda, gejala ringan, pemulihan lambat

Cluster 1
28 tahun
12 hari
Sembuh
Pasien sangat muda, cepat sembuh

Cluster 2
64 tahun
21 hari
Sembuh
Lansia, tetapi sembuh

Cluster 3
56 tahun
8 hari
Meninggal
Lansia, durasi singkat, komorbid tinggi

4.2. Interpretasi Cluster

Cluster 3 menjadi fokus utama karena meskipun usia tidak tertua, durasi sakit sangat singkat namun berakhir dengan kematian. Hal ini mengindikasikan bahwa pasien mungkin memiliki komorbid seperti diabetes, hipertensi, atau gangguan paru-paru yang menyebabkan perkembangan penyakit sangat cepat.

Cluster 0 dan 1 menunjukkan bahwa pasien usia muda cenderung memiliki prognosis baik, meskipun durasi sakit bisa panjang.

Cluster 2 menunjukkan bahwa lansia bisa sembuh jika penanganan dilakukan tepat waktu.

4.3. Validasi Hasil

Nilai Silhouette Score sebesar 0,48 menunjukkan bahwa pengelompokan cukup baik, meskipun tidak sempurna. Hal ini wajar karena data kesehatan bersifat heterogen dan banyak dipengaruhi faktor eksternal yang tidak terukur dalam dataset.. Misalnya, pada awal peneliti memaparkan narasi temuannya, kemudian didukung dengan sajian data dalam bentuk tabulasi, tabel atau grafik. Peneliti juga menyajikan data-data hasil penelitian, kemudian didukung grafik dilanjutkan deskripsi naratif [10 pts]. Berikan kemungkinan pengembangan atau penelitian ke depan terkait penelitian ini

V. KESIMPULAN DAN SARAN

Dari penelitian ini dapat disimpulkan :

1. Terdapat beberapa metode atau algoritma yang sering digunakan dalam proses *data mining*, dan dalam Penelitian ini menggunakan Metode *k-Means Clustering* yaitu sebuah algoritma yang digunakan untuk mengelompokkan data ke dalam beberapa kategori atau *cluster* yang berbeda berdasarkan fitur atau variabel yang ada. Di dalam Tugas Akhir ini, penggunaan metode *K-means clustering* dianggap sebagai metode yang cepat dan efisien, terutama jika data yang digunakan cukup besar, output yang dihasilkan mudah dipahami, output yang dihasilkan yang bersifat numerik, dan dapat digunakan pada data yang tidak terstruktur atau tidak berlabel.
   Setelah dilakukan tahapan Persiapan Dataset, Pengolahan Data, Pre-processing Data dan penetapan nilai k-, Algoritma K-Means Clustering berhasil mengelompokkan faktor risiko yaitu **Rentan Usia, Durasi Sakit dan Status Akhir Pasien** yang berdampak pada tingkat kematian Pasien COVID-19.
2. Algoritma K-Means Clustering berhasil mengelompokkan data pasien berdasarkan faktor-faktor risiko, dimana dari data-data tersebut, kemudian ditetapkan beberapa kelompok yang memiliki faktor risiko tingkat kematian. Dengan menggunakan Aplikasi Google Collabs, didapatkan nilai k=4, maka didapatkan Cluster 3 (risiko sangat tinggi), Cluster 2 (risiko sedang), Cluster 0 (risiko rendah), Cluster 1 (risiko sangat rendah)
3. Dari data-data yang dimiliki dan berdasarkan penjelasan diatas, penetapan kelompok / Cluster dengan menggunakan Algoritma K-Means Clustering telah mengurutkan urutan risiko kematian dari tinggi ke rendah adalah:Disimpulkan bahwa Cluster 3 memiliki karakteristik yang sangat jelas, antara lain : *Rentang Usia yang Luas : 0–91 tahun, Durasi Sakit yang Panjang : 0–475 hari, Status Akhir Pasien : Meninggal dunia*.

REFERENSI

Achmad Solichin, K. K. (2020). Klasterisasi Persebaran Virus Corona (Covid-19) Di DKI Jakarta. *Fountain of Informatics Journal*, 8.

Andi Akram Nur Risal (2021). Penerapan data mining dalam mengklasifikasikan tingkat kasus Covid -19 Di Sulawesi Selatan menggunakan Naïve Bayes.

Dunham, Margareth H., (2003). Data Mining Introductory and Advanced Topics. New Jersey:Prentice Hall.

Etikasari, B.dkk, (2020). Sistem Informasi Deteksi Dini Covid-19.

Fauziah Nur, P. Z. (2017). Penerapan Algoritma K-Means pada siswa baru Sekolah Menengah Kejuruan. Jurnal Nasional Informatika dan Teknologi Jaringan , 6.

Fayyed et al (1996). The KDD Process Model.

Ghozali, Imam. (2011). Aplikasi Analisis Multivariate dengan Program SPSS. Semarang

Handayani. (2020). Metode Penelitian Kualitatif & Kuantitatif. CV. Pustaka Ilmu.

Hairunisa and Amalia (2020). Jurnal Biomedika dan Kesehatan (JBK) review : penyakit virus Corona baru 2019 (COVID-19).

Heri Susanto dan Sudiyanto (2019). Data mining untuk memprediksi prestasi siswa berdasarkan sosial ekonomi, motivasi, kedisiplinan dan prestasi masa lalu Jurnal Pendidikan Vokasi, Vol 4, Nomor 2.

Imam Cholissodin (2021). Klasifikasi tingkat laju data covid-19 untuk mitigasi penyebaran

menggunakan metode Modified K-Nearest Neighbour (MKNN)

Joko Suntoro, M. (2019). Data Mining: Algoritma dan Implementasi dengan pemograman PHP. Semarang: Joko Suntoro, M.Kom.

Jogiyanto, H.M., (2005). Analisa dan Desain Sistem Informasi: Pendekatan terstruktur teori dan praktik Aplikasi Bisnis, ANDI, Yogyakarta

Mahmudan, A. (2020). Pengelompokan kabupaten atau kota di Jawa Tengah berdasarkan kasus COVID-19 menggunakan K-Means Clustering. Jurnal Matematika Statistika dan komputansi, 13.

Mulyanti, S. (2015). Penerapan data mining dengan metode clustering untuk pengelompokkan data pengiriman burung. Prosiding Seminar Ilmah Nasional Teknologi Komputer, 30-35.

Prastyadi Wibawa Rahayu, d. (2024). Buku Ajar Data Mining. Jambi: PT.Sonpedia Publishing Indonesia.

Sari, D. (2022). Impelementasi Algoritma K-Means Dalam Menentukan Tingkat Penyebaran Pandemi. Computer Based Information System Journal, 9.

S.P Tamba, F. t. (2019). Penerapan data mining untuk menentukan penjualan sparepart Toyota dengan menggunakan K-Means clustering. Jurnal sistem informasi komputer prima, 67-72.

S.Sindi, W. N. (2020). Analisis Algoritma K-Medoids Clustering dalam pengelompokan penyebaran COVID-19 di Indonesia. Jurnal Teknol informasi , 166-173.

Susanto, H. (2014). Data Mining Untuk Memprediksi Prestasi Siswa. Jurnal Pendidikan Vokasi, 222-231.

Tikaridha Hardiani (2022). Analisis Clustering Kasus Covid 19 Di Indonesia Menggunakan Algoritma K-Means

Yoga Puja Aritama Juli (2022). Penerapan Metode K-Means Clustering Untuk Mengelompokkan Data Kasus Covid-19 Di Indonesia

Yuli Asriningtias, Rodhya Mardhiyah (2014). Aplikasi Data Mining Untuk Menampilkan Informasi Tingkat Kelulusan Mahasiswa. Jurnal Informatika, 837-848.

Yulia Darmi, A. S. (2016). Penerapan Metode Clustering K-Means dalam pengelompokkan penjualan produk. Jurnal Media infotama, 10