Penerapan Metode Clustering Dengan Algoritma K-Means Untuk Analisa Persebaran Varian Covid-19 (Studi Kasus Kelurahan Antapani Kidul)

1st Mochamad Noverian Zhafar
Fakultas Teknik Elektro
Universitas Telkom
Bandung, Indonesia
noverianzhafar@student.telkomuniversity.ac.id

2nd Korediantо Usman
Fakultas Teknik Elektro
Universitas Telkom
Bandung, Indonesia
korediantousman@telkomuniversity.co.id

3rd Fityanul Akhyar
Fakultas Teknik Elektro
Universitas Telkom
Bandung, Indonesia
fityanul@telkomuniversity.ac.id

***Abstrak* — Pandemi COVID-19 merupakan peristiwa persebaran penyakit yang terjadi di seluruh dunia. Berbagai negara telah berupaya untuk memberhentikan pergerakan virus tersebut agar tidak terjadi gelombang akibat virus yang berevolusi dan melahirkan varian baru. Pada umumnya, data persebaran dari suatu wilayah sangat diperlukan oleh para praktisi Kesehatan untuk meneliti lajur dan kemungkinan terjadinya gelombang ataupun ditemukannya varian baru dari virus SARS-Cov-2. Penelitian ini bertujuan untuk menganalisa persebaran varian COVID-19 di kelurahan Antapani Kidul, kota Bandung dari segala aspek perbedaan dari setiap variannya.**

Metode dalam penelitian ini yaitu menggunakan teknik clustering dengan penggunaan alur data mining yang menerapkan algoritma K-Means. Algoritma K- Means menggunakan dataset yang digunakan untuk mengelompokan data berdasarkan kriteria pendukung berupa tingkat penularan, tingkat transmisi komunitas, dan juga sejumlah dampaknya pada imunitas pasien pengidap COVID- 19.

Dalam penelitian ini juga diukur akurasi dari performansi metode clustering menggunakan algoritma K-Means dengan membandingkannya dengan empat metode lain, yaitu DBSCAN, Gaussian Mixture, Agglomerative Clustering, dan Spectral Clustering dengan menggunakan tabel Performance Metrics dengan empat parameter pengukuran yang disebut main metrics, yang merupakan Silhouette Score, Calinski-Harbasz Index, Davies Bouldien Index, dan Rand Index.

***Kata kunci:* COVID-19, Varian, Algoritma K-Means, Clustering, Kriteria, Persebaran, Performance Metrics.**

I. PENDAHULUAN

Perkembangan teknologi informasi sangat berkembang pesat dalam memajukan beberapa sektor pekerjaan, termasuk dalam bidang kesehatan. Kecanggihan teknologi informasi juga membuat eksistensi sistem pengolahan data secara konvensional tergantikan oleh digital yang jauh lebih efisien. Sistem pengolahan data konvensional secara manual kini mulai ditinggalkan mengingat banyaknya data berlimpah yang memerlukan kecerdasan buatan untuk mengolahnya. Dengan majunya teknologi informasi, pengolahan data akan lebih mudah untuk dilakukan dalam menganalisa data di bidang kesehatan, termasuk untuk menganalisa data persebaran varian dari pandemi COVID-19.

Pandemi COVID-19 diprediksi akan terus berevolusi menjadi berbagai jenis varian setelah penelitian mengemukakan bahwa varian Omicron telah berkembang menjadi 30 mutasi dan menjadi landasan berbagai negara untuk mempercepat laju vaksinasi [2]. Analisis *cluster* adalah teknik untuk mengelompokkan objek-objek berdasarkan kesamaan karakteristik di antara objek-objek tersebut. Pemanfaatan analisis *cluster* seringkali dilakukan untuk melakukan klasifikasi pada produk, benda, bahkan manusia. Analisis *cluster* juga memiliki tujuan untuk mengelompokkan objek-objek yang bersifat mirip untuk dimasukkan ke dalam satu *cluster* yang sama. Dalam melakukan melakukan analisa *cluster* persebaran varian COVID-19 di kelurahan Antapani Kidul, varian akan dikelompokan berdasarkan frekuensi kejadian sehingga setiap varian yang paling banyak kesamaannya dengan varian lain akan berada dalam satu *cluster* yang sama menggunakan algoritma K-Means. Algoritma K-means merupakan algoritma *unsupervised learning* yang digunakan untuk menggelompokkan data berdasarkan *variable* atau *feature*. Algoritma K-Means diperlukan dalam penelitian ini karena memiliki tingkat efisiensi yang tinggi untuk mengolah objek dalam jumlah yang besar.

Metode *clustering* dapat digunakan untuk menganalisis persebaran virus COVID-19 dengan membagi wilayah atau populasi menjadi kelompok-kelompok berdasarkan karakteristik tertentu yang terkait dengan persebaran virus. beberapa cara dalam penerapan metode *clustering* untuk analisis persebaran virus COVID-19 adalah analisis jenis kelamin, usia, status vaksinasi, dsb. Dengan menggunakan metode *clustering*, pemerintah dan otoritas kesehatan dapat membuat strategi yang lebih tepat dan efektif untuk mengatasi persebaran virus COVID-19 dan membantu mencegah wabah yang lebih besar.

II. KAJIAN TEORI

A. Metode Cluster

Data *clustering* atau pengelompokan data adalah proses untuk mengidentifikasi kelompok atau cluster dalam data multidimensional berdasarkan beberapa ukuran kesamaan. Data yang digunakan adalah tabel dengan jumlah ratusan hingga ribuan kolom dan berisikan nilai yang memiliki poin yang unik atau berbeda di antara satu sama lainnya, kumpulan dari data ini biasa disebut dengan *dataset*.

	A	B	C	D	E	F	G	H	I	J	K	L
1	area_nam	area_code	date	dose	age_band	age_highe	age_lowe	cum_dose	new_dose	populatio	new_prop	cum_prop
2	Barking ar	E09000002	########	1st dose	12 - 15 yea	15	12	0	0	13415	0	0
3	Barking ar	E09000002	########	1st dose	16 - 17 yea	17	16	0	0	5541	0	0
4	Barking ar	E09000002	########	1st dose	18 - 24 yea	24	18	0	0	17719	0	0
5	Barking ar	E09000002	########	1st dose	25 - 29 yea	29	25	0	0	15880	0	0
6	Barking ar	E09000002	########	1st dose	30 - 34 yea	34	30	1	1	17683	5.66E-05	5.66E-05
7	Barking ar	E09000002	########	1st dose	35 - 39 yea	39	35	0	0	17445	0	0
8	Barking ar	E09000002	########	1st dose	40 - 44 yea	44	40	0	0	15380	0	0
9	Barking ar	E09000002	########	1st dose	45 - 49 yea	49	45	0	0	14293	0	0
10	Barking ar	E09000002	########	1st dose	50 - 54 yea	54	50	0	0	12735	0	0
11	Barking ar	E09000002	########	1st dose	55 - 59 yea	59	55	0	0	10845	0	0
12	Barking ar	E09000002	########	1st dose	60 - 64 yea	64	60	0	0	8228	0	0
13	Barking ar	E09000002	########	1st dose	65 - 69 yea	69	65	0	0	6062	0	0
14	Barking ar	E09000002	########	1st dose	70 - 74 yea	74	70	0	0	4823	0	0
15	Barking ar	E09000002	########	1st dose	75 - 79 yea	79	75	0	0	3399	0	0
16	Barking ar	E09000002	########	1st dose	80+ years	90	80	0	0	5523	0	0
17	Barking ar	E09000002	########	2nd dose	12 - 15 yea	15	12	0	0	13415	0	0
18	Barking ar	E09000002	########	2nd dose	16 - 17 yea	17	16	0	0	5541	0	0
19	Barking ar	E09000002	########	2nd dose	18 - 24 yea	24	18	0	0	17719	0	0
20	Barking ar	E09000002	########	2nd dose	25 - 29 yea	29	25	0	0	15880	0	0
21	Barking ar	E09000002	########	2nd dose	30 - 34 yea	34	30	0	0	17683	0	0
22	Barking ar	E09000002	########	2nd dose	35 - 39 yea	39	35	0	0	17445	0	0

GAMBAR 2.1
Contoh Tampilan Dataset

Data *clustering* memiliki algoritma yang terdiri dari : Algoritma K-Means, Algoritma DBSCAN, Algoritma Gaussian Mixture Model, Algoritma Agglomerative Hierartical Clustering.

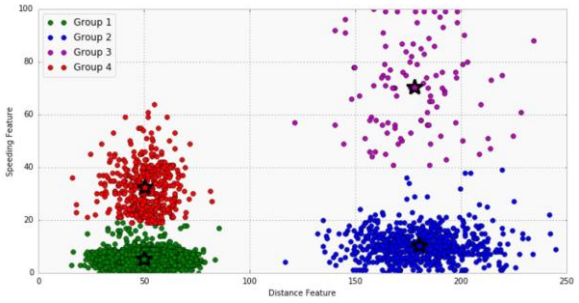
B. K-Means

Algoritma K-Means merupakan salah satu algoritma yang bersifat *unsupervised learning*. K-Means memiliki fungsi untuk mengelompokkan data kedalam data *cluster*.[7] Metode *clustering* pada algoritma K-Means menggunakan metode *non-hierarchy*, yaitu metode yang bertujuan untuk mengelompokkan n obyek kedalam kelompok. Metode tersebut seringkali digunakan sebagai alternatif metode *cluster* untuk data dengan ukuran yang besar karena memiliki kecepatan yang lebih tinggi dan lebih efisien dibandingkan metode *hierarchy*. Algoritma K-Means juga menerapkan metode *cluster sampling*, yaitu metode dimana sampel akan diambil dari unit-unit populasi yang dipilih secara acak dari kelompok atau *cluster*. *Cluster sampling* adalah landasan utama Algoritma K-Means sebagai salah satu algoritma *unsupervised learning*. Algoritma K-means bertujuan untuk memilih centroid yang meminimalkan inersia, atau kriteria jumlah kuadrat dalam sebuah cluster dengan rumus:

$$\sum_{i=0}^{n} \min_{\mu_j \in C}(||x_i - \mu_j||^2)$$

Algoritma K-Means sering diterapkan dalam menentukan parameter jumlah data yang berukuran besar, salah satunya adalah mengenai analisis yang penulis lakukan, yaitu persebaran varian COVID-19 di wilayah kelurahan Antapani Kidul. Beberapa kekurangan yang dimiliki oleh algoritma K-Means diantaranya adalah sulit dalam memilih jumlah *cluster* yang tepat, *overlapping*, dan kegagalan dalam melakukan *converge*.

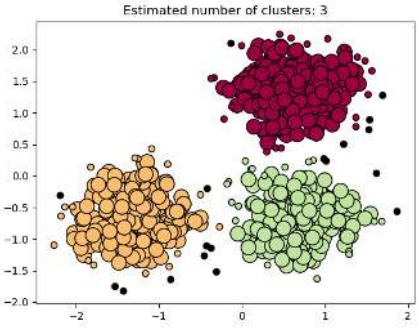
Algoritma k-Means juga merupakan algoritma partisional yang dapat meminimalisir kesalahan dalam melakukan *clustering*. Hasil dari pengelompokan menggunakan algoritma K-Means ditampilkan pada **Gambar 2.2** dimana bobot fitur *centroid* dapat digunakan untuk memahami jenis grup yang diwakili oleh setiap *cluster* secara kualitatif.[3]



GAMBAR 2.2
Hasil Pengelompokan Algoritma K-Means

C. DBSCAN

Algoritma DBSCAN *(Density-Based Spatial Clustering of Applications with Noise)* merupakan salah satu contoh pelopor perkembangan teknik pengelompokan berdasarkan kepadatan atau yang biasa dikenal dengan sebutan *density based clustering*. [13] DBSCAN dapat berbentuk apa saja, berbeda dengan K-Means yang mengasumsikan bahwa *cluster* akan selalu berbentuk cembung. Komponen utama DBSCAN adalah konsep sampel inti, yaitu sampel yang berada di area dengan kepadatan tinggi.



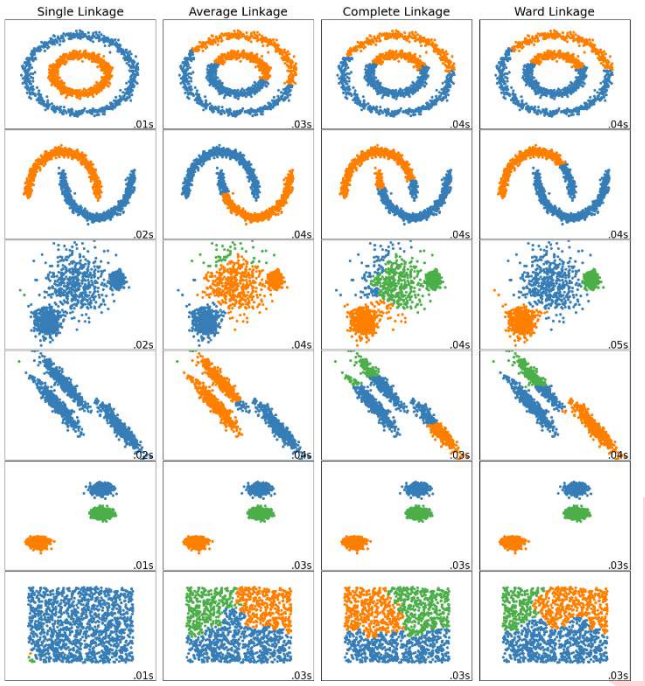
GAMBAR 2.3
Hasil Pengelompokan Metode DBSCAN

D. Gaussian Mixture Model

Algoritma Gaussian Mixture Model adalah sebuah metode kepadatan model yang terdiri dari komponen fungsi-fungsi pada gaussian. Gaussian Mixture Model juga a mempresentasikan gabungan dari beberapa hal linier pada Gaussian. Metode ini sering digunakan pada *speaker recognition, speech recognition, voice recognition* dan lainnya.

E. Agglomerative Clustering

Algoritma Agglomerative Hierarchical Clustering merupakan metode analisis cluster yang bertujuan untuk mengelompokkan objek-objek berdasarkan karakteristik yang dimilikinya, yang dimulai dengan objek-objek individual sampai objek-objek tersebut bergabung menjadi satu cluster tunggal. Metode Agglomerative Hierarchical Clustering terbagi menjadi beberapa algoritma, di antaranya metode *single linkage, complete linkage, average linkage, dan ward*. [18]



GAMBAR 2.4
Hasil Pengelompokan Metode Agglomerative Clustering

F. Data Mining

Data mining adalah proses pengumpulan dan pengolahan data yang bertujuan untuk mengekstrak informasi penting pada data. Proses pengumpulan dan ekstraksi informasi tersebut dapat dilakukan menggunakan perangkat lunak dengan bantuan statistika, matematika, ataupun kecerdasa buatan. Dalam jurnal yang dilakukan oleh Saefudin,M.Kom dan Septian DN dengan judul "*Penerapan Data Mining dengan Metode Algoritma Apriori untuk Menentukan Pola Pembelian Ikan", Jurnal Sistem Informasi*, vol. 6, no. 2, pp. 110-114, Sept. 2019. Data mining juga dapat diartikan sebagai proses pencarian secara otomatis yang berguna dalam tempat penyimpanan data berukuran besar. [4] Tahapan dari data mining adalah:

1. Data *Cleaning* (untuk menghilangkan noise data yang tidak konsisten).
2. Data *Integration* (dimana sumber data yang terpecah dapat disatukan).
3. Data *Selection* (di mana data yang relevan dengan tugas analisis dikembalikan ke dalam database).
4. Data *Transformation* (di mana data berubah atau bersatu menjadi bentuk yang tepat untuk menambang dengan ringkasan performa atau operasi).
5. Data *Mining* (proses esensial di mana metode yang intelijen digunakan untuk mengekstrak pola data).
6. *Pattern Evolution* (untuk mengidentifikasi pola yang benar-benar menarik yang mewakili pengetahuan berdasarkan atas beberapa Tindakan yang menarik).
7. *Knowledge Presentation* (di mana gambaran Teknik visualisasi dan pengetahuan digunakan untuk memberikan pengetahuan yang telah ditambang kepada user).

G. Bahasa Python

Python adalah salah satu bahasa pemrograman dan perangkat lunak serbaguna yang bisa dijalankan pada hampir semua arsitektur sistem. Python biasanya digunakan dalam berbagai pengaplikasian di banyak bidang, seperti pengembangan *website, game,* analisa data hingga *machine learning*. Python diciptakan oleh Guido van Rossum di Centrum Wiskunde & Informatica (CWI) di Belanda sebagai penerus bahasa ABC dan pertama kali dirilis pada tahun 1991. Python kini dikembangkan oleh Python Software Foundation, perusahaan nonprofit yang memegang hak kekayaan intelektual atas Python. R menjadi standar di antara *developer* untuk mengembangkan perangkat lunak dan juga digunakan secara luas untuk pengembangan analisis data. Kini Python memiliki lebih dari 100.000 library. [11]

Python merupakan bagian dari proyek GNU. GNU adalah suatu sistem operasi computer yang sepenuhnya terdiri dari perangkat-perangkat lunak bebas. Sumber kode Python tersedia secara bebas di bawah Lisensi Publik Umum GNU, dan versi biner perkompilasinya tersedia untuk berbagai sistem operasi.[8] Python menyediakan berbagai teknik statistika seperti pemodelan linier dan nonlinier, analisis deret waktu, klasifikasi, klasterisasi, dan sebagainya.

H. Jupyter Notebook

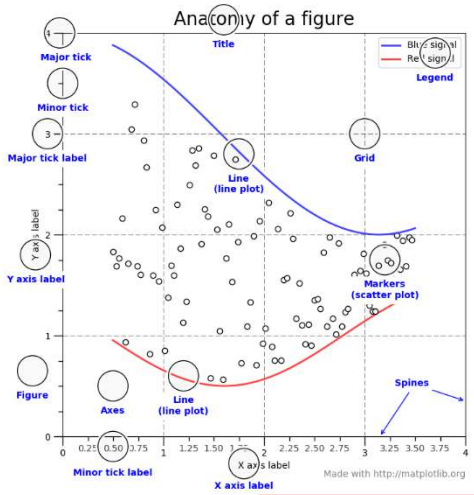
Jupyter Notebook adalah *Integrated Development Environment* (IDE) untuk bahasa pemrograman Python yang banyak digunakan hingga saat ini untuk melakukan komputasi statistik dan grafik. Jupyter Notebook memiliki berbagai cakupan untuk membantu pengguna dalam melakukan komputasi, diantaranya adalah konsol, editor sintaks yang mendukung eksekusi kode, serta fitur lainnya untuk mengelola ruang kerja. Jupyter Notebook tersedia di berbagai sistem operasi, diantaranya Windows, Mac, Linux, ataupun untuk browser yang terhubung ke Jupyter Notebook seperti Debian, Ubuntu, RedHat, CentOS, dan SUSE Linux.[9] Berikut ini merupakan Library dari Jupyter Notebook:

1. Matpolib

Library Matpotlib merupakan salah satu library penting di Jupyter Notebook yang digunakan untuk menciptakan dan memuat berbagi bentuk grafik yang dihasilkan oleh data yang kompleks. Library Matpotlib merupakan library yang digunakan dalam melakukan visualisasi berdasarkan sumbu x dan y, sehingga untuk membuat dasar grafik menggunakan bahasa Python, hanya dapat dilakukan hanya lewat library tersebut. [5]

Dalam melakukan visualisasi data melalui Matpotlib, pengguna harus memastikan mana sampel yang harus dipilih dari data. Sampel tersebut kemudian akan ditampilkan dalam bentuk grafik menggunakan sintaks yang sepenuhnya akan dikomputasi oleh library Matpotlib. Dalam membuat suatu grafik, obyek geometri yang harus diterapkan pada sintaks adalah pivot_table, xlabel, ylabel, dst. Tujuan dari proses visualisasi menggunakan library Maptoptilb adalah untuk melakukan visualisasi yang lebih kompleks dan simpel sehingga data yang menjadi sampel

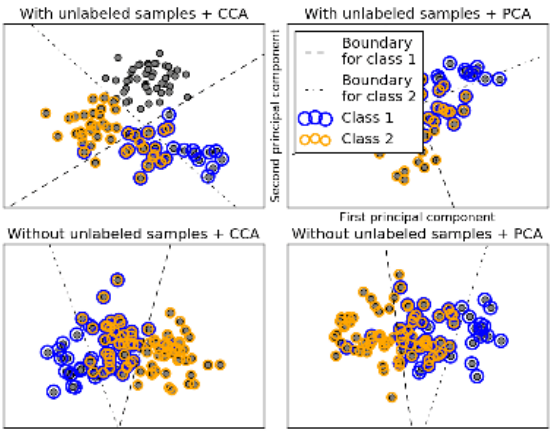
tidak perlu dalam format data.frame.



GAMBAR 2.5
Anatomi Kinerja Library Matpotlib

2. Scikit-Learn

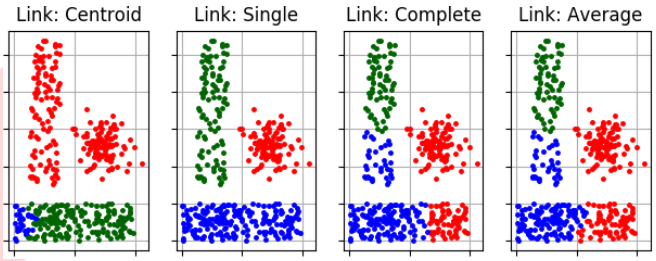
Library Scikit-learn atau Sklearn adalah salah satu library pada bahasa Python yang dibangun berdasarkan NumPy, SciPy, dan Matpotlib. Library Scikit-learn bekerja dengan melakukan ekstraksi menggunakan hasil analisa yang berasal dari multivariasi data, lalu Scikit-learn akan membantu melakukan processing data ataupun melakukan training data untuk kebutuhan *machine learning* [6]. Library Scikit-Learn dapat menganalisa berbagai fungsi atau model analisa dalam bahasa pemrograman Python, seperti model klasifikasi, *clustering*, regresi berbasis model *machine learning,* dan proses-proses yang dapat dimanfaatkan pada tahap *feature engineering* seperti reduksi menggunakan PCA. Library Scikit-learn akan melakukan simpelisasi pada beberapa langkah analisa *clustering* dan menyediakannya pada library Matpotlib untuk menyempurnakan visualisasi data yang lebih baik.



GAMBAR 2.6
Ekstraksi pada Library Scikit-Learn

3. PyClystering

Library PyClustering atau Clustering merupakan salah satu library pada bahasa Python yang digunakan dalam melakukan analisa data berbentuk klaster, melakukan variasi pada grafik, mengaplikasikan algoritma pada proses kluster, mengosilasi *neural networks model,* memvisualisasikan hasil analisa, dsb. Library PyClustering merupakan library yang terdiri atas layanan *centroid-based clustering algorithm* yang menyediakan analisa algoritma k-Means. [12] Selain *centroid-based*, library PyClustering menyediakan layanan *distribution-based clustering algorithm* yang menyediakan fungsi *Gaussian Mixture Models* (GMM) yang merepresentasikan distribusi subpopulasi secara normal tanpa melibatkan populasi secara keseluruhan. Tujuan dari digunakannya library PyClustering adalah untuk melakukan validasi pada output, memplot hasil baik secara siluet maupun plot 2 dimensional, melakukan prediksi dalam observasi yang akan dilakukan selanjutnya, dan mengestimasi jumlah optimal dari *cluster* untuk setiap algoritma secara terpisah.



GAMBAR 2.7
Metode Clustering Menggunakan Library PyClustering

4. Pandas

Library Pandas adalah salah satu library dalam bahasa pemrograman Python yang berlisensi BSD dan open source yang menyediakan struktur data dan analisis data yang mudah digunakan. Pandas digunakan untuk membuat tabel, mengubah dimensi data, dan mengecek data. Struktur data pada pandas dinamakan sebagai DataFrame, dan DataFrame ini adalah struktur yang memudahkan pengguna bahasa Python dalam membaca sebuah file dengan banyak jenis format seperti file .txt, .csv, dan .tsv. Fitur ini akan menjadikannya sebagai sebuah tabel dan juga dapat mengelola suatu data dengan menggunakan seperti join, distinct, group by, agregasi, dan berbagai perintah lain dalam SQL.

5. NumPy

Library NumPy atau Numerical Python adalah library Python yang menyediakan fungsi untuk memproses komputasi numerik. NumPy memiliki kemampuan untuk membuat objek N-dimensi *array*. *Array* merupakan sekumpulan variabel yang memiliki tipe data yang sama. Kelebihan dari NumPy adalah dapat memudahkan operasi komputasi pada data, cocok untuk melakukan akses secara acak, dan elemen *array* merupakan sebuah nilai yang independen sehinga penyimpanannya dianggap sangat efisien.

I. Data Mining

Performance metrics adalah ukuran yang digunakan untuk menilai kualitas hasil *clustering*. Metrik ini bertujuan untuk menentukan apakah metode *clustering* berhasil membuat asumsi yang tepat mengenai data dan membagi data menjadi kelompok yang memiliki homogenitas di dalamnya. Ada beberapa metrik populer yang digunakan dalam evaluasi performa *clustering*, seperti *Silhouette Score, Rand Index, Calinski-Harabasz Index, dan Davies-Bouldin Index.*

1. *Silhouette Score*

Silhouette Score adalah metrik yang digunakan untuk menilai kualitas dari *clustering* data. Metrik ini mengukur jarak antara *data point* dengan *cluster*-nya sendiri dan jarak antara *data point* dengan *cluster* terdekat lainnya. Skor *Silhouette* berkisar antara -1 dan 1, di mana nilai yang lebih tinggi menunjukkan bahwa *data point* lebih baik diklasifikasikan dalam *cluster* mereka.

2. *Calinski-Harabasz Index*

Calinski-Harabasz Index adalah metrik yang digunakan untuk menilai kualitas *clustering* data. Ini mengukur rasio antara varians internal dalam setiap *cluster* dan varian antar *cluster*. Secara umum, nilai *Calinski-Harabasz* yang lebih tinggi menunjukkan bahwa *clustering* yang lebih baik, karena ini menunjukkan bahwa cluster memiliki varians internal yang lebih rendah dan varians antar *cluster* yang lebih tinggi.

3. *Davies-Bouldin Index*

Davies-Bouldin Index adalah metrik yang digunakan untuk menilai kualitas dari *clustering* data. Ini mengukur rata-rata jarak antara setiap *data point* pada suatu *cluster* dengan *centroid* (titik rata-rata) dari *cluster* terdekat lainnya. Nilai *Davies-Bouldin* berkisar antara 0 dan infiniti, di mana nilai yang lebih rendah menunjukkan bahwa *clustering* yang lebih baik, karena ini menunjukkan bahwa jarak antara *cluster* terdekat lebih kecil.

4. *Rand Index*

Rand Index adalah metrik yang digunakan untuk menilai kualitas dari *clustering* data. Ini mengukur persentase pasangan *data points* yang diklasifikasikan dengan benar dalam suatu *clustering*. Nilai *Rand Index* berkisar antara 0 dan 1, di mana nilai yang lebih tinggi menunjukkan bahwa *clustering* yang lebih baik, karena ini menunjukkan bahwa lebih banyak pasangan *data points* yang diklasifikasikan dengan benar.

III. METODE

A. Desain Sistem

Sebelum melakukan penelitian mengenai persebaran varian COVID-19 di kelurahan Antapani Kidul, penulis mengumpulkan data acak berformat .xlsx yang sebelumnya telah diciptakan dan dikumpulkan oleh Dinas Kesehatan Kota Bandung dan UPT Puskesmas Jajaway Antapani. Data tersebut berisikan mengenai laporan setiap kasus yang diterima oleh UPT Puskesmas Jajaway di wilayah kelurahan Antapani Kidul, Kota Bandung. Lalu akan dilakukan impor data yang dilakukan oleh sintaks pemrograman. Sintaks pemrograman yang dirancang oleh penulis diketik pada halaman konsol IDE untuk mengkompilasi bahasa Python, yaitu Jupyter Notebook.

Pertama data persebaran varian COVID-19 di kelurahan Antapani Kidul akan diimpor langsung dari file penyimpanan perangkat penulis untuk dimuat dan dilakukan pengecekan kelengkapan data pada Jupyter Notebook.

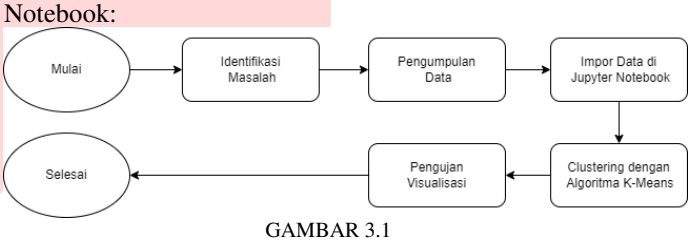
Lalu ketika data sudah sepenuhnya berhasil dimuat, sintaks yang telah dipersiapkan akan menjadi perintah bagi tiap library (Matpotlib, Scikit-Learn, PyClustering) untuk melakukan analisa, pengambilan dan pemilahan, dan memvisualisasi data. Data akan dianalisa oleh library PyClustering yang akan mengambil sampel dari kelompok acak untuk dikelompokkan pada unit yang mirip satu sama lain.

Lalu sampel yang telah diambil akan diekstrak oleh library Scikit-Learn untuk membantu library selanjutnya dalam memudahkan langkah-langkah analisa demi mendapatkan hasil visualisasi yang tepat.

Terakhir, library Matpotlib akan merancang sebuah visualisasi berbentuk grafik kompleks dari sampel yang telah dicocokkan satu sama lain. Library Matpotlib melakukan finalisasi dalam tahap *clustering* data persebaran varian COVID-19 di kelurahan Antapani Kidul dalam bentuk grafik statis yang memuat informasi data jumlah persebaran virus COVID-19 pada setiap variannya.

B. Diagram Blok

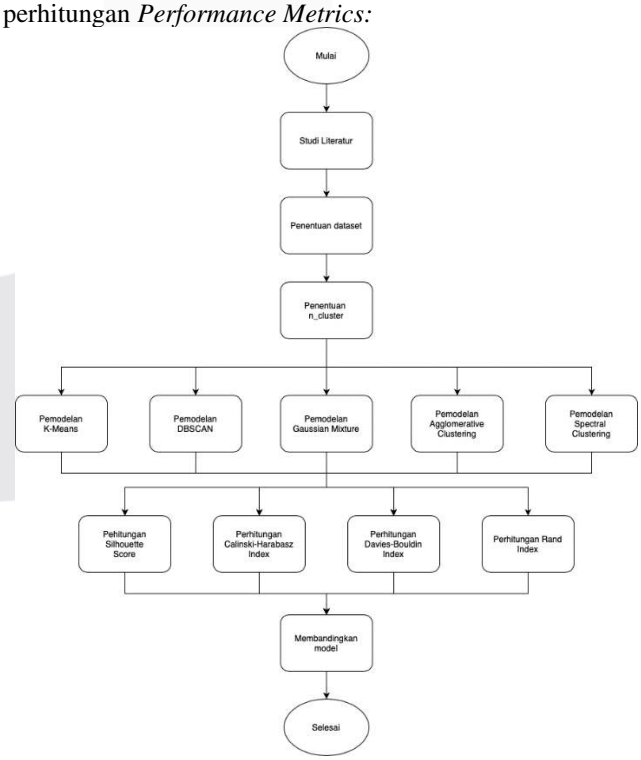
Berikut ini merupakan diagram blok dari implementasi algoritma K-Means menggunakan *software* atau IDE Jupyter Notebook:



GAMBAR 3.1
Diagram Blok Clustering dan Visualisasi Grafik

C. Diagram Perhitungan Performance Metrics

Pembuatan tabel perbandingan parameter dengan format tabel performansi atau *Performance Metrics,* dengan membandingkan algoritma K-Means dengan 4 algoritma *clustering* lainnya. Berikut merupakan langkah-langkah perhitungan *Performance Metrics:*



GAMBAR 3.2
Diagram Perhitungan Performance Metrics

D. Fungsi dan Fitur

Komponen Pendukung Perangkat Keras:

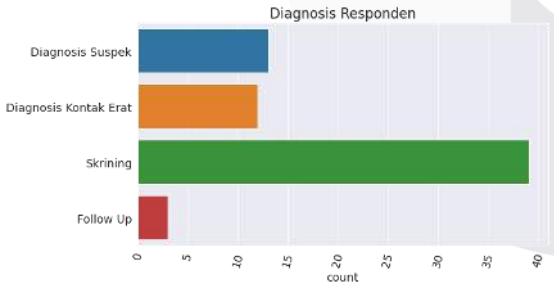
TABEL 3.1
Komponen Pendukung Perangkat Keras

Spesifikasi	Nilai
OS	Windows 2000/XP/Vista/7/8/8.1/10/11, Windows Server 2003/2008/2012/2016/2019/2022, macOS 10.15/11/12/13, Ubuntu Linux 18/20/22, RHEL/CentOS Linux 7, RHEL 8, Debian Linux 9/10/11, SUSE Linux 15 SP3/openSUSE 15.3, Amazon Linux 2
Memori	7200 RPM STATA dengan 20GB available space
Prosesor	Intel Core i3 2.5 GHz – Intel Core i5
RAM	4 GB
CPU	4+ CPU
Browser	Microsoft Edge, Safari, Chrome, Firefox

Program dirancang menggunakan bahasa pemrograman Python dengan Jupyter Notebook sebagai IDE atau *Integrated Development Environment* sebagai perangkat untuk menyusun sintaks perintah yang dapat menjalankan metode *clustering* menggunakan algoritma K-Means. Program juga menggunakan 3 library untuk memuat visualisasi dari data persebaran varian COVID-19, yaitu library Matpotlib, Scikit-Learn, dan PyClustering. Ketiga library tersebut memiliki fungsi masing-masing. Pertama data persebaran varian COVID-19 dimuat diimpor dari penyimpanan komputer penulis, lalu data yang berbentuk tabel atau yang dapat disebut sebagai *dataset* akan ditransformasi menjadi file .py atau file Python dan akan ditampilkan di halaman utama Jupyter Notebook.

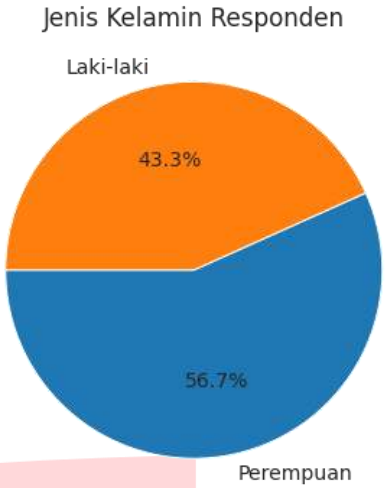
IV. HASIL DAN PEMBAHASAN

A. Hasil Grafik Diagnosis Responden



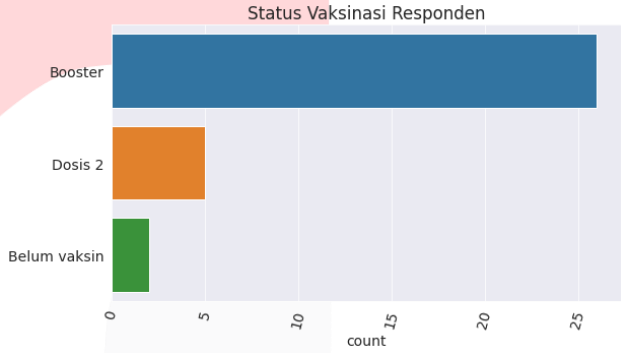
GAMBAR 4.1
Grafik Diagnosis Responden

B. Hasil Grafik Jenis Kelamin



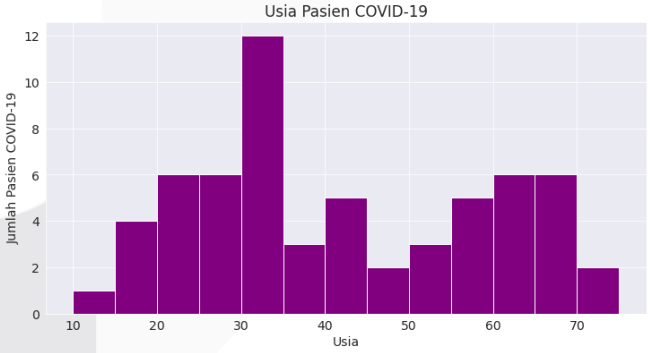
GAMBAR 4.2
Grafik Jenis Kelamin Pasien

C. Hasil Grafik Status Vaksinasi Pasien



GAMBAR 4.3
Grafik Status Vaksinasi Pasien dengan Bar Chart

D. Hasil Grafik Usia Pasien



GAMBAR 4.4
Grafik Status Usia Pasien dengan Bar Chart

E. Perhitungan dan Analisis Hasil

Berikut ini merupakan uraian seraca rinci analisis dan perhitungan data terhadap keakuratan masing-masing algoritma menggunakan empat *main metrics,* yaitu *Silhouette Score, Calinski-Harabasz Index, Davies-Bouldin Index,* dan *Rand Index.*

Pengujian pada perhitungan data terhadap keakuratang masing-masing algoritma ini dicantumkan pada sebuah tabel *Performace Metrics* yang berindekskan keempat *main metrics*, tanpa adanya parameter dan label yang digunakan seperti untuk melakukan perbandingan terhadap metode klasifikasi. Untuk melakukan *Performance Metrics* pada metode *clustering*, penulis menggunakan dataset sebagai sampel perbandingan. Dataset akan menjadi acuan dan tumpuan kelima algoritma untuk diklasifikasikan (diprediksikan) secara acak sesuai dengan nilai n_cluster yang juga menjadi tumpuan dari kelima algoritma untuk melihat nilai performansi mereka.

TABEL 4.1
Tabel *Performance Metrics*

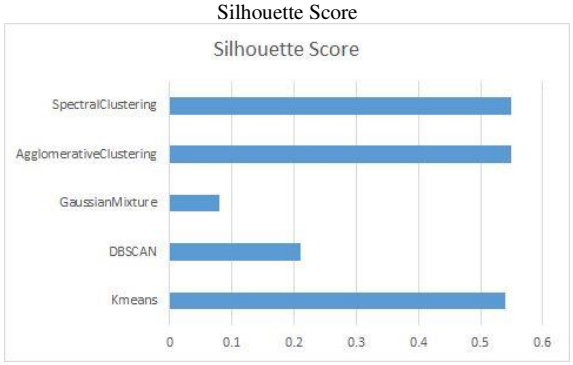
	K-Means	*DBSCAN*	*Gaussian Mixture*	*Agglomerative Clustering*	*Spectral Clustering*
Silhouette Score	0.54	0.21	0.08	0.55	0.55
Calinski Harabasz Index	266.21	18.25	30.75	252.73	147.73
Davies Bouldien Index	0.53	3.76	1.14	0.54	0.49
Rand Index	1.00	1.00	1.00	1.00	1.00

F. Analisis Perbandingan Metode Clustering Menggunakan Performance Metrics

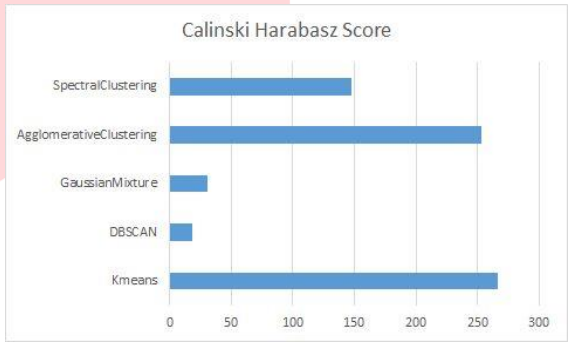
Hasil visualisasi grafik perbandingan keempat metode *clustering* dengan menggunakan pengujian *performance metrics* menggunakan empat parameter pengujian, yaitu *Silhouette Score, Calinski Harabasz Index, Davies Boudin Index,* dan *Rand Index.* Hasil visualisasi grafik ini ditujukan untuk membantu Analisa perbandingan keakuratan dari performansi kelima metode *clustering* berdasarkan nilai dari setiap keakuratan metode dan ditampilkan dalam bentuk grafik sederhana berjenis *bar plot.*

Keunikan yang dimilikki setiap metode *clustering* dalam melakukan proses *data mining* dan *clustering* menjadikan setiap metode memiliki keunggulan dan kekurangannya masing-masing dan juga berdasarkan parameter keempat *main metrics*, Rand Score memiliki hasil yang selalu sama karena tidak memiliki nilai minimal bersifat desimal ketika menguji sebuah parameter atau label dari n_cluster yang digunakan oleh setiap metodenya.

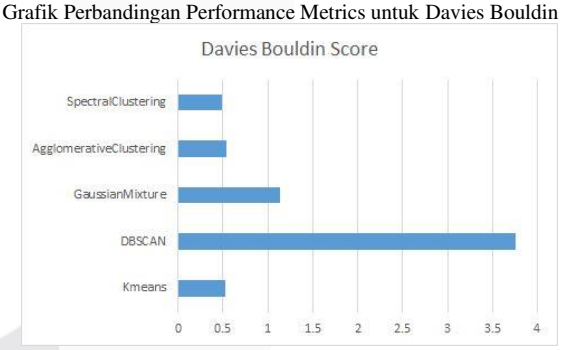
TABEL 4.2 Grafik Perbandingan Performance Metrics untuk Silhouette Score



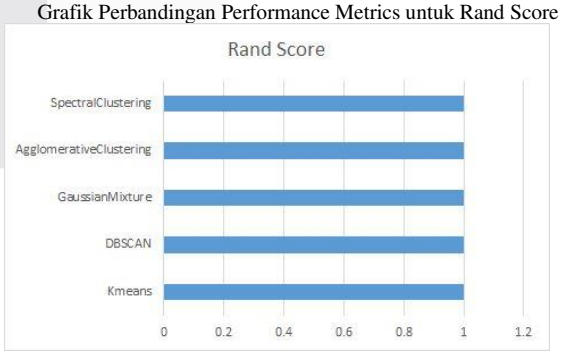
TABEL 4.3
Grafik Perbandingan Performance Metrics untuk Calinski Harabasz Score



TABEL 4.4
Grafik Perbandingan Performance Metrics untuk Davies Bouldin Score



TABEL 4.5
Grafik Perbandingan Performance Metrics untuk Rand Score



V. KESIMPULAN

A. Kesimpulan

Berdasarkan hasil pengujian yang telah dipaparkan di bab sebelumnya, maka dapat ditarik kesimpulan sebagai berikut:

1. Jika dibandingkan hasil n_cluster pada simulasi *Performance Metrics* dari setiap metode *clustering* yang telah

dicoba pada dataset persebaran varian COVID-19, Agglomerative Clustering merupakan algoritma clustering paling baik karena nilai rata-rata main metricsnya lebih unggul dari metode lain setelah dicoba dengan perubahan nilai n_cluster sebanyak 4 kali.
2. Nilai *Performance Metrics* pada metode K-Means yang diuji didapatkan hasil bahwa algoritma/metode tersebut bukan metode yang paling akurat untuk melakukan *clustering,* namun Algoritma K-Means sangat cocok untuk mengatasi data numerik dan dapat memberikan hasil yang baik untuk data dengan distribusi normal.
3. Dataset COVID-19 yang dijadikan parameter penelitian ini hanya bisa divisualisasikan setelah dilakukan *clustering* oleh metode algoritma K-Means.

B. Saran

Hal-hal yang dapat ditingkatkan berdasarkan penelitian yang sudah dilakukan adalah:
1. Penelitian ini dilakukan dengan menggunakan metode algoritma k-Means yang juga sering digunakan penulis lain dalam melakukan Analisa dan penelitian, namun metode Agglomerative Clustering lebih direkomendasikan karena teruji lebih akurat dari metode K-Means.
2. Sangat disarankan untuk membuat perhitungan *Performance Metrics* pada perhitungan keakuratan berbagai metode *clustering* menggunakan 4 *main metrics*, yaitu Silhouette Score, Calinski Harbasz index, David Bouldin Index, dan Rand Score karena merupakan parameter yang kredibel dan tepat dalam mengukur keakuratan suatu metode *clustering*.
3. Seringkali IDE Jupyter Notebook mengalami disfungsi sehingga perlu menyiapkan IDE opsional atau cadangan seperti Google Collabs yang dapat dengan mudah diakses dari internet.
Grafik visualisasi persebaran COVID-19 di Kelurahan Antapani Kidul dengan mengambil periode setelah gelombang varian Omicron dan sebelum gelombang varian XBB yang dihasilkan sangat sederhana sehingga perlu dilakukan Analisa menggunakan dataset yang lebih Panjang periodenya dan mencakup area yang lebih besar seperti Kota Bandung atau Provinsi Jawa Barat

REFERENSI

[1] A. R. Jannah, D. Arifianto, “Penerapan Metode Clustering dengan Algoritma K-means untuk Prediksi Kelulusan Mahasiswa Jurusan Teknik Informatika di Universitas Muhammadiyah Jember,” 2015.
[2] Burhan E, Rachmadi R. A, “Omicron Surge and The Future of COVID-19 Vaccinations,” Medical Journal Indonesia, vol 31 No 1, April 2022 [Internet]. Available: https://mji.ui.ac.id/journal/index.php/mji/article/view/6066
[3] T. S. M. Duarte, “Learning Analytics of K-means Clustering: A Pilot Study,” in Magnetism, 2018. Available: https://repositorio-aberto.up.pt/handle/10216/116970
[4] Saefudin, Septian D.N, “Penerapan Data Mining dengan Algoritma Apriori untuk Menentukan Pola Pembelian Ikan,” Jurnal Sistem Informasi, vol. 6, no. 2, pp. 110-114, Sept. 2019.
[5] H. Wickham, “ggplot2: Elegant Graphics for Data Analysis,” Journal of Statistical Software, vol. 35, pp. 216, 2009. Available: http://www.springer.com/978-0-387-98140-6
[6] A. Kassambra, F. Mundt, “factoextra: Extract and Visualize The Results of Multivariate Data Analyses”, 2020. [Online]. Available: [6] https://cran.r-project.org/web/packages/factoextra
[7] B. Orleans, E.P Putra, “Clustering Algoritma (K-Means), Jan. 2022. [Online]. Available: https://sis.binus.ac.id/2022/01/31/clustering-algoritma-k-means/
[8] A. Robbins, “What’s GNU” Linux Journal, vol. 1994, pp. 7-es, May 1994.
[9] W. Hasbi, “Pengklasifikasian Breast Cancer dengan Metode Naïve Bayes”, 2018.
[10] A. Salsabilla, “Penerapan Algoritma K-Means dan C4.5 untuk Clustering Jurusan Siswa Baru pada Sekolah Menengah Kejuruan (Studi Kasus: SMK Negeri 1 Paron)”, 2021.
[11] A. Bogdanchikov, M. Zhaparov, R Suliyev, “Python to Learn Programming,” Journal of Phyriscs: Conference Series, vol. 423, Jan 2013.
[12] A.V. Novikov, “PyClustering: Data Mining Library,” Journal Of Open Source Software, vol. 4, Apr 2019.
[13] Mumtaz K, “An Analysis on Density Based Clustering of Multi Dimensional Spatial Data,” Indian Journal of Computer Science and Engineering (IJCSE), vol. 1, no. 1, pp. 8–12, 2010.
[14] E.P. Prameswari, I. Zaid, Suroso, “Penerapan Metode Gaussian Mixture Model pada Monitoring Pulsa Listrik dengan Masukkan Suara,” Seminar Nasional Inovasi dan Aplikasi Teknologi di Industri 2019, Feb 2019.
[15] F. R. Bach and M. I. Jordan, “Learning spectral clustering,” Adv. Neural Inf. Process. Syst., 2004.
[16] M. S. Hatta, F. Azmi, C. Setianingsih, “Clustering pada Sentimen Penggunaan Transportasi Online Menggunakan Algoritma Spectral Clustering,” e-Proceeding of Engineering, vol. 8, no. 6, pp 11945-11951, Des 2021.
[17] Frisca, A. Bustaman, “Implementasi Spectral Clustering pada Data Microarray Gen Karsinoma Menggunakan Algoritma K-Means,” pp 37-39, 2017.
[18] A. Fadliana, F. Rozi, “Penerapan Metode Agglomerative Hierarchical Clustering untuk Klasifikasi Kabupaten/Kota di Provinsi Jawa Timur berdasarkan Kualitas Pelayanan Keluarga Berencana,” vol. 4, no. 1, pp. 35-40, Nov 2015.

..