BAB III

METODOLOGI PENELITIAN

3.1. Objek Penelitian

Objek penelitian pada penelitian ini adalah penyebaran Covid-19 di DKI Jakarta. Dimana peneliti memakai informasi mengenai kasus Covid-19 di tahun 2020 – 2022 dengan perincian yaitu bulan Maret – Desember 2020, Januari – Desember 2021, Januari – November 2022, selain itu penelitian yang akan dilakukan memiliki tujuan yaitu mengelompokan penyebaran Covid-19 di DKI Jakarta. Hasil penelitian ini diharapkan akan bisa digunakan sebagai sebuah masukan untuk menentukan kebijakan terhadap masing-masing daerah di DKI Jakarta guna menekan angka penyebaran Covid-19 dengan data terbaru yaitu tahun 2020 sampai 2022. Data tersebut dikategorikan dengan melihat mean terdekat, sehingga hasil penelitian adalah kluster dari masing-masing daerah yang terkena dampak Covid-19.

3.2. Metode Penelitian

Metode *k-means* klastering dan *fuzzy c-means* digunakan untuk membentuk klasterisasi penyebaran Covid-19 di DKI Jakarta. Metode *k-means* dipilih karena mudah diterapkan dan dijalankan, relatif cepat, mudah diadaptasi serta banyak dipraktikkan dalam tugas data mining [27]. Selain itu, nilai akhir yang diperoleh dari metode *k-means* dapat memaksimalkan kemiripan data pada satu klaster dan meminimalkan kemiripan data antar klaster [14]. Sedangkan, metode *fuzzy c-means* dipilih karena memiliki kelebihan yang terletak pada penempatan pusat klaster yang lebih tepat dibandingkan dengan metode klaster lain [28].

3.2.1. K-Means Klastering

Termasuk bagian dari metode klastering *non-hirarki* yang dipakai untuk pengklasteran data ke satu atau beberapa kelompok. Penerapan metode ini yaitu ketika data dengan ciri yang serupa akan diklasterkan dalam sebuah kelompok,

sedangkan data yang karakteristiknya berbeda dikelompokkan dengan kelompok yang berbeda.

3.2.2. Fuzzy C-Means

Algoritma *Fuzzy C-Means* merupakan teknik pengelompokan yang terawasi, karena pada algoritma ini jumlah klaster yang akan dibentuk perlu diketahui terlebih dahulu. Konsep dasar algoritma *fuzzy c-means* adalah menentukan pusat kelompok yang akan menandai lokasi rata-rata untuk setiap tiap-tiap klaster. Tujuan dari algoritma *fuzzy c-means* untuk mendapatkan pusat klasteryang nantinya akan digunakan untuk mengetahui data yang masuk ke dalam sebuah klaster [29].

3.2.3. Perbandingan Metode

Metode yang digunakan dalam penelitian ini adalah *K-means Klastering* dan *fuzzy c-means*. Kedua metode ini dipakai untuk melakukan pengklasteran. Kedua metode tersebut memiliki karakteristik yang berbeda dalam penerapannya.

Tabel 3.1 Perbandigan Metode

Parameter	***K-Means Klastering***	***Fuzzy C-Means***
Fase	Tahapan K-Means [30]: 1. Tentukan k sebagai jumlah klaster yang dibentuk. 2. Tentukan k centroid (titik pusat klaster) awal secara random. 3. Hitung jarak setiap objek ke masing-maisng centroid dari masing-masing klaster. 4. Alokasi masing-masing objek ke dalam centoid yang paling dekat. 5. Lakukan iterasi, kemudian tentukan posisi centroid baru. 6. Ulangi langkah 3 jika posisi centroid baru tidak sama.	Tahapan *Fuzzy C-Means* [31]: 1. Inisialisasi K (Jumlah klaster), f(x) (fungsi objektif) awal, centoid awal (K data sebagai centroid awal), dan tetapkan ambang batas (threshold), pembobot (w) dan maksimum iterasi. 2. Hitung nilai centroid tiap-tiap klaster. 3. Hitung nilai derajat keanggotaan setiap data ke setiap Klaster. 4. Hitung f(x) dengan metode perhitungan jarak Euclidean. 5. Perbaiki derajat keanggotaan setiap data pada setiap klaster. 6. Ulangi langkah 2 dan 5 hingga konvergen tercapai, yaitu apabila derajat keanggotaan ≤ ambang batas, atau perubahan f(x) ≤ ambang batas, atau apabila

		perubahan centroid ≤ ambang batas, atau telah mencapai iterasi maksimum yang ditentukan di awal.
Kelebihan	Kelebihan K-*Means* [30]: 1. Metode yang sangat simple dan fleksibel. Artinya perhitungan komputasinya tidak telalu rumit dan dapat diimplementasikan pada segala bidang. 2. Metode ini sangat mudah dipahami, teruama dalam implementasi data yang sangat besar serta dapat mengurangi kompleksitas data yang dimiliki.	Kelebihan *Fuzzy C-Means* [32]*:* *1.* Metode ini bersifat *unsupervized.* 2. Dapat mencapai pusat *klaster* yang *konvergen.* 3. Dalam kondisi tertentu FCM merupakan model klastering yang mempunyai ketangguhan jika dilihat dari nilai fungsi obyektifnya, jumlah iterasinya dan waktu yang diselesaikan.
Kekurangan	Kekurangan K-*Means* [30]: 1. Memerlukan angka yang tepat dalam menentukan jumlah klaster sebanyak k. Karena terkadang pusat klaster awal dapat berubah sehingga kejadian ini bisa mengakibatkan pengelompokkan data menjadi tidak stabil. 2. Metode ini tidak bisa maksimal dalam menentukan atau menginisialisasi nilai centroid awalnya, karena pada pengelompokkan data dengan metode K-Means sangat bergantung pada nilai centroid-nya. 3. Ditemukannya beberapa model klaster yang berbeda. 4. Harus melakukan pemilihan jumlah klaster yang tepat. 5. Kegagalan untuk converge 6. Input dari K-Means tergantung pada nilai pusat yang dipilih pada *Klastering*.	Kekurangan *Fuzzy C-Means* [32]: *1.* Mudah terjebak dalam *local optima.* 2. Sensitif terhadap pusat klaster awal.
Hasil Penerapan	Output dari *K-Means* [33]: 1. Klaster: vector yang berisikan lokasi klaster tiap objek. 2. *Centers*: matriks yang berisikan centroid/rata-ratanilai tiap klaster. 3. *Withinss*: vektor yang berisikan simpangan tiap klaster yang terbentuk. 4. Jumlah objek pada tiap *klaster*.	Berbentuk deret dari pusat klaster dan keanggotaan terhadap setiap titik data. Hasil akan dipakai dalam membangun *fuzzy interface system* [34].

Perbandingan metode penelitian yang umum digunakan pada proses data mining, yaitu metode KDD, CRISP-DM dan SEMMA.

Tabel 3.2 Perbandingan Metode *Data Mining* [35]

	KDD	**CRISP-DM**	**SEMMA**
Tahapan	1. *Data selection* 2. *Pre-processing* atau *cleaning* 3. *Transformation* 4. *Data mining* 5. *Interpretation* atau *Evaluation*	1. *Business understanding* 2. *Data understanding* 3. *Data preparation* 4. *Modeling* 5. *Evaluation* 6. *Deployment*	1. *Sample* 2. *Explore* 3. *Modify* 4. *Model* 5. *Assessment*

Dari tabel 3.2 mengenai perbandingan metode penelitian *data mining* tersebut pada tabel diatas, peneliti menentapkan penggunaan metode CRISP-DM (*Cross-Industry Standard Process for Data Mining*) pada penelitian ini dikarenakan CRISP-DM sebagai salah satu standard untuk menghasilkan *data driven decision making* yang berkualitas berdasarkan hasil *pooling* datascience-pm, yang dimana CRISP-DM digunakan 2 sampai dengan 3 kali lebih banyak dari 5 teratas standard yang paling banyak digunakan. Metode CRISP-DM dipilih karena model proses standar terbuka yang menggambarkan pendeketan umum yang digunakan untuk *data mining*. Selain itu, CRISP-DM bertujuan untuk menyediakan proses *data mining* yang andal dan berulang. Hal tersebut mempunyai kerangka kerja yang seragam dengan pedoman dan fleksibel untuk beradaptasi dengan masalah bisnis dan kumpulan data yang berbeda [36].

3.3. Teknik Pengumpulan Data

Penelitian ini memakai pendekatan kuantitatif. Tahapan yang digunakan dalam pengumpulan data adalah observasi untuk memperoleh data yang diperlukan berupa data melalui website Corona Jakarta dan kemudian dilakukan *scrapping*.

3.4. Teknik Pengambilan Sampel

Teknik *Purposive Sampling* digunakan dalam penelitian ini untuk pengambilan sampel. Tenik *purposive sampling* merupakan metode jenis *sampling*

non-random yang dilakukan dengan mempertimbangkan hal tertentu dalam pengambilan sampel untuk tujuan tertentu [37]. Sampel yang digunakan adalah data Covid-19 di DKI Jakarta pada tahun 2020-2022 yang diambil melalui laman corona.jakarta.go.id.

3.5. Variabel Penelitian

Variabel penelitian menggunakan data tahun 2020 – 2022 yang diperoleh dari data jumlah yang terdeteksi positif dan kasus aktif yang berasal dari website Corona Jakarta. Kasus positif merupakan pasien yang yang terkonfirmasi covid-19 dan kasus aktif merupakan pasien yang masih dalam perawatan yang dilakukan pada fasilitas kesehatan maupun secara mandiri. Pada penelitian ini menggunakan dua buah variabel data yang meliputi independen dan dependen.

1. Variabel Independen

   Variabel independen merupakan variabel data-data yang digunakan untuk membuat klasterisasi yang direncanakan. Variabel ini akan memberikan pengaruh terhadap perubahan pada variabel dependen. Variabelindependent dalam penelitian ini yang terdiri dari: Positif, Dirawat, Sembuh, Meninggal dan isolasi dirumah atau isoman.

2. Variabel Dependen

   Variabel dependen merupakan luaran klasterisasi yang dihasilkan dari model data mining yang dibuat, varaibel ini dipengaruhi oleh varaibel independent. Variabel dependen yang digunakan dalam penelitian ini yang terdiri dari: klaster merah untuk mengindikasi daerah yang rawan dengan jumlah pasien Covid-19 terbanyak. Klaster kuning untuk mengindikasi daerah yang cukup rawan dengan jumlah pasien Covid-19 hampir banyak. Klaster hijau untuk mengindikasi daerah yang aman dengan jumlah pasien Covid-19 tidak terlalu banyak.

3.6. Kerangka Kerja

Metode CRISP-DM (*Cross-Industry Standard Process for Data Mining*) merupakan model proses standar terbuka yang menggambarkan pendeketan umum yang digunakan untuk penambangan data. Tujuan dari CRISP-DM untuk menyediakan proses penambangan data yang andal dan berulang. Hal tersebut mempunyai kerangka kerja yang seragam dengan pedoman dan fleksibel untuk beradaptasi dengan masalah bisnis dan kumpulan data yang berbeda [36].

Dalam penelitian ini akan melakukan klasterisasi wilayah penyebaran Covid-19 di DKI Jakarta dengan penerapan data mining menggunakan metode *K-Means* dan *Fuzzy C-Means,* oleh karena itu untuk menerapkan tenik data mining dalam penelitian ini terdpaat metode penelitian yang digunakan yaitu metode CRISP-DM. Berikut terdapat tahapan dalam CRISP-DM (Cross-Industry Standard Process for Data Mining) yaitu:

1. *Business Understanding*

   Pada penelitian ini tujuan bisnisnya adalah memetakan klasterisasi wilayah penyebaran Covid-19 di DKI Jakarta yang diperlukan sebagai referensi menetapkan prioritasi dan kebijakan penanganan pada penyebaran Covid-19 agar pemerintah setempat dapat menentukan kebijakan yang tepat dan sesuai dengan kasus disetiap daerah provinsi.

2. *Data understanding*

   Pada tahap data understanding merupakan tahap persiapan yang dilakukan untuk melakukan penetapan dan validasi data yang dibutuhkan untuk pembentukan klasterisasi wilayah penyebaran Covid-19 di DKI Jakarta.

3. Data *Preparation*

   Pada tahap data preparation merupakan tahap mempersiapkan data dengan menyesuaikan dataset sesuai dengan kebutuhan pada tahap pemodelan. Pada data *preparation* dilakukan seleksi terhadap data yang digunakan, data *preparation* yang bertujuan untuk mempersiapkan data mentah menjadi data yang siap untuk tahap pemodelan dan melakukan transformasi data.

4. *Modeling*

   Pada tahap *modeling* merupakan tahapan pembuatan model menggunakan teknik data mining klastering, menggunakan metode K-*means Klastering* dan *fuzzy* C-*means Klastering* terhadap dataset penyebaran Covid-19 di DKI Jakarta. Pada tahap ini diimplementasikan dalam bentuk visualiasasi menggunkan bahasa pemrograman Python dengan *tools Jupyter Notebook* dan *Microosft Excel* untuk melakukan analisis hasil dari metode K-*means Klastering* dan *fuzzy* C-*means Klastering* terhadap klasterisasi penyebaran Covid-19 di DKI Jakarta.

5. *Evaluation*

   Tahap *evaluation* merupakan tahap pengukuran terhadap model klasterisasi yang dihasilkan dari tahap *modeling*, dalam hal ini mengukur performasi dan akurasi dari masing-masing algoritma yang digunakan pada model tersebut yaitu *K-Means* dan *Fuzzy C Means*.

6. *Deployment*

   Pada tahap *deployment* merupakan tahap implementasi hasil klasterisasi dari metode *K-Means* dan *fuzzy C-Means* pada kasus Covid-19 di DKI Jakarta yang digunakan untuk website corona Jakarta.

3.7. Teknik Analisis Data

Analisis data merupakan upaya mencari dan menata secara sistematis catatan hasil observasi, wawancara, dan lainnya untuk meningkatkan pemahaman peneliti tentang kasus yang diteliti dan menyajikannya sebagai temuan bagi orang lain. Teknik analisis data terdapat beberapa tahapan yaitu reduksi data (*Data Reduction)*, penyajian data *(Data Display)* dan penarikan simpulan/verifikasi *(Conclusion drawing/verification)* [38].

1. Reduksi Data *(Data Reduction)*

   Pada tahap reduksi data dilakukan dengan melakukan analisis yang menjelaskan, menggolongkan, menyederhanakan dan mengorganisasi data sehingga bisa ditarik kesimpulan dan diverifikasi. Pada tahap ini berfokus pada

tujuan yang ingin dicapai. Jika terdapat data yang dianggap tidak penting dari data kasus Covid-19 di DKI Jakarta akan dilakukan reduksi data untuk menelusuri pola kasus Covid-19 di DKI Jakarta.

2. Penyajian Data *(Data Display)*

   Pada tahap penyajian data merupakan kegiatan ketika sekumpulan informasi disusun sehingga memberikan kemungkinan akan adanya penarikan kesimpulan dan pengambilan tindakan. Bentuk penyajikan data dalam penelitian ini berupa visualisasi hasil klasterisasi kasus Covid-19 di DKI Jakarta yang menggunakan alat bantu bahasa pemrograman *Python*. *Python* merupakan bahasa pemrograman simpel yang digunakan untuk komputasi statistik, data mining dan grafis [39]. Dalam penelitian ini editor yang digunakan untuk menulis kode program yaitu *Jupyter Notebook*. Karena *Jupyter notebook* merupakan sebuah aplikasi untuk menulis kode bahasa pemrograman Python dalam bentuk website yang berada pada localhost komputer, setiap perintah dapat dijelaskan dalam satu halaman dan terdapat fitur yang menampilkan hasil visualisasi berupa grafik yang berada pada satu *cell* yang sama [39].

3. Penarikan simpulan/verifikasi *(Conclusion/verivication)*

   Pada tahap ini yaitu penarikan kesimpulan yang didukung dengan hasil penelitian berupa klasterisasi kasus Covid-19 di DKI Jakarta dalam bentuk visualisasi.

Dari Tabel 3.3 terdapat definisi, kelebihan dan kekurangan dari *tools* yang digunakan dalam penelitian ini. Dengan menggunakan *Jupyter Notebook* dapat memvisualisasi data dalam bentuk grafik dan dapat menjelaskan setiap perintah dari setiap bagian program pemodelan yang dibuat [39].

Tabel 3.3 Perbandingan *Tools Data Mining* Yang Digunakan

Tools	**Definisi**	**Kelebihan**	**Kekurangan**
Jupyter Notebook	Alat *open source*	● Visualisasi data,	● *Jupyter Notebook*

	berbasis Python yang berinteraksi langsung dengan data para ilmuwan dikarenakan reatif mudah dipelajarinya.	dapat melakukan visualiasasi data dalam bentuk grafik, tableau. ● Berbagi kode, dapat melakukan berbagi kode secara *cloud* atau melalui jaringan internet untuk melihat kode, menjalankan kode dan menampilkannya pada browser, seperti GitHub ● Interaksi langsung dengan kode, kode dalam *jupyter* tidak statis sehingga dapat diedit dan dijalankan kembali dan menghasilkan umpan balik pada browser. ● Mendokumentasikan kode, dalam *jupyter* dapat menjelaskan secara bertahap pada potongan kode yang ingin dijelaskan.	tidak mandiri yaitu membutuhkan *runtime jupyter*. ● Status sesi tidak dapat disimpan dengan mudah yaitu kode yang dijalankan tidak dapat dipertahankan dan dikembalikan sehingga setaip membuat notebook harus dijalankan kembali kodenya untuk memulihkan kondisinya. ● Tidak ada *debuging* interaktif atau IDE lain
Google Colaboratory	Perangkat lunak yang dapat digunakan untuk mengeksekusi kode python melalui *web browser* [40].	● Dapat digunakan untuk *machine learning*, analisis data dan pendidikan. ● Layanan mirip *jupyter notebook* yang telah dihosting dan disediakan oleh	● *Resource* yang disediakan tidak terjamin dan terbatas yang memiliki batas penggunaan *resource* tersebut [40]. ● Tidak dapat menjamin jenis

		google yang memiliki akses bebas dan *resource* gratis [40].	*hardware* yang diperoleh atau kapasitas memori yang diperlukan [41].

Dari tabel 3.3 merupakan perbandingan *tools* pada data mining, dimana setiap *tools* memiliki kelebihan dan kekurangannya. Dalam penelitian yang dilakukan, peneliti menggunakan *tools Jupyter Notebook* karena pada *Jupyter* terdapat berbagai teknik dalam data mining yaitu teknik *supervised* maupun *unsupervised learning.* Dalam *tools* tersebut juga menyediakan *library* yang dapat digunakan dalam pembuatan sistem, *library* merupakan kumpulan modul yang berisi kumpulan kode yang dapat digunakan berulang kali dalam program yang berbeda [18]. Dalam *library* pada Python yang tersedia diantaranya yaitu numPy, Pandas, matplotlib, sklearn, serta menyediakan berbagai algoritma dan menghasilkan *output* berupa visualisasi dalam bentuk grafis [18]. Selain itu, *tools* ini sangat berguna di bidang *data science* dan *machine learning* mengingat penelitian yang dilakukan menggunakan metode klasterisasi yang akan diimplementasikan kedalam bahasa pemrograman Python, karena *tools* ini terintegrasi dengan Python dan *library* yang ada tidak mempunyai batasan waktu. Hal ini berbeda dengan *google colaboratory* yang memiliki *resource* yang tidak terjamin dan terbatas serta tidak dapat menjamin kapasitas memori yang diperlukan.