BAB III
METODE PENELITIAN

3.1 Objek Penelitian

Pada penelitian ini objek data yang diambil adalah data dari *website* SIHA yang masih berupa data mentah dengan format excel *csv*. berjumlah 972 *dataset* dengan jumlah 27 kabupaten/kota. Dalam kelompok data akan terdapat selisih data yang bernilai terbesar dan data yang bernilai kecil disebut dengan nilai rentang (*range*). Nilai rentang dataset dibawah pada kabupaten Bogor dari tahun 2019 ke 2021 dengan nilai range 45, kabupaten Sukabumi memiliki nilai range dari tahun 2019 ke 2021 yaitu 5, kabupaten Cianjur memiliki nilai range 78, kabupaten Bandung memiliki nilai range 60, kabupaten Garut memiliki nilai range 167, kabupaten Tasikmalaya memiliki nilai range 18, kabupaten Ciamis memiliki nilai range 19, kabupaten Kuningan memiliki nilai range 22, kabupaten Cirebon memiliki nilai range 25, kabupaten Majalengka memiliki nilai range 35, kabupaten Banjar memiliki nilai range 45.

NO	KABUPATEN	UMUR	JUMLAH KASUS	TAHUN
1	BOGOR	4 - 50+	475	2019
2	SUKABUMI	4 - 50+	112	2019
3	CIANJUR	4 - 50+	111	2019
4	BANDUNG	4 - 50+	179	2019
5	GARUT	4 - 50+	172	2019
6	TASIKMALAYA	4 - 50+	77	2019
7	CIAMIS	4 - 50+	95	2019
8	KUNINGAN	4 - 50+	91	2019
9	CIREBON	4 - 50+	257	2019
10	MAJALENGKA	4 - 50+	97	2019
.	.	-	-	-
.	.	-	-	-
.	.	-	-	-
972	BANJAR	4 - 50+	40	2021

Tabel 3.1 Objek penelitian *dataset* dari *website* SIHA

3.2 Lokasi dan Waktu Penelitian

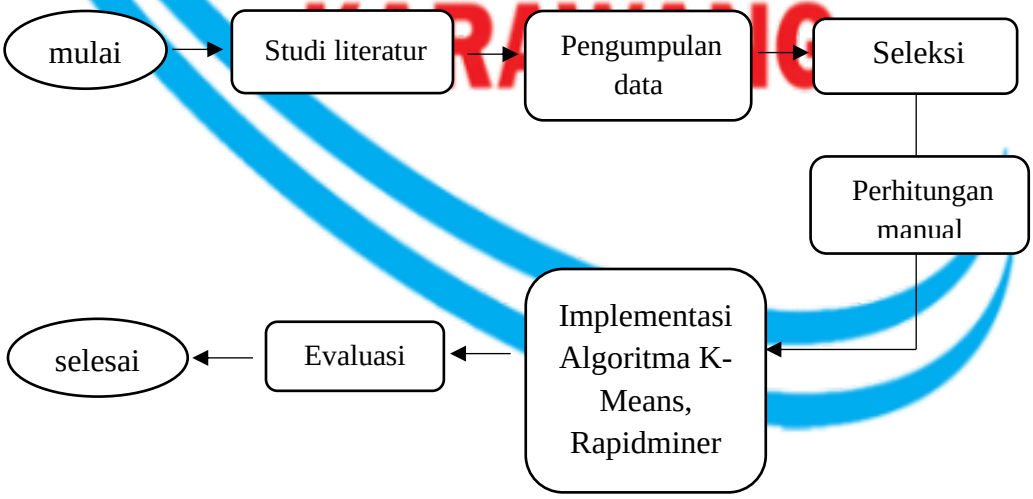
Penelitian dilakukan di Perpustakaan Universitas Buana Perjuangan Karawang dan berlangsung sejak Maret 2023 hingga Agustus 2023. Dari pengumpulan hingga tahap evaluasi.

3.3 Prosedur Penelitian

Berdasarkan pada rincian jadwal kegiatan yang diatas, pada metode penelitian ini menjelaskan cara kerja algoritma yang digunakan dan proses yang akan dilakukan untuk melakukan pengelompokan *cluster*. Dalam metode penelitian ini akan membahas tentang metode pengelompokan *clustering k-means* yang akan digunakan dan tahapan-tahapan dalam penelitian ini. Tahapan tersebut terdiri dari :

1. Menentukan jumlah kluster "k" yang akan dibagi.
2. Menentukan data "k" sebagai titik pusat (*centroid*) awal tiap cluster.
3. Mengelompokan data kedalam "k" cluster sesuai dengan titik pusat yang telah ditentukan sebelumnya.
4. Memperbaharui pusat cluster dan mengulangi langkah ketiga sampai nilai dari titik pusat tidak berubah.

Tahapan prosedur penelitian dapat dilihat pada gambar flowchart 3.1



Gambar 3.1 Flowchart Tahapan Prosedur Penelitian

3.4 Studi Literatur

Tahapan studi literatur dilakukan upaya peneliti melakukan pengumpulan bahan landasan-landasan teori dan informasi yang relevan diperoleh dari

berbagai sumber seperti buku, internet, dan jurnal. Studi literatur yang digunakan disini meliputi pengolahan data kasus penyakit HIV/AIDS di Jawa Barat menggunakan algoritma k-means. Dengan melakukan studi literatur ini para peneliti memiliki pemahaman secara teoritis ataupun praktis yang lebih luas, memahami terhadap masalah yang akan diteliti, dan mendapat informasi tentang aspek-aspek mana dari suatu masalah yang sudah pernah diteliti untuk menghindari agar tidak meneliti hal yang sama.

3.5 Pengumpulan Data

Dalam pengumpulan data untuk penelitian ini digunakan metode pengumpulan studi pustaka yang mana pada metode ini adalah mempelajari, mencari dan mengumpulkan data yang berhubungan dengan penelitian ini. Data dalam penelitian ini berasal dari sumber data utama dan sumber data sekunder.

a. Sumber data utama ini diperoleh dari situs web Sistem Informasi HIV/AIDS dan IMS. Data pada penelitian ini berupa data penyebaran HIV yang ada diseluruh Kabupaten di Provinsi Jawa Barat. Data yang diperoleh berjumlah 972 data dari 27 kabupaten/kota. Atribut data meliputi sepuluh atribut yaitu id, kode provinsi, nama provinsi, kode kabupaten/kota, nama kabupaten/kota, jenis kelamin, kelompok umur, jumlah kasus, satuan dan tahun.

b. Sumber data sekunder diperoleh dari studi literatur yaitu dengan mencari materi yang berhubungan dengan permasalah, perancangan, metode *K-means Clustering*, penunjang keputusan dan mempermudah proses implementasi. Studi literatur diperoleh melalui buku, jurnal di internet dan pada penelitian sebelumnya. Data dapat dilihat pada gambar 3.3 dibawah.

Tabel 3.2 *Dataset*

NO	KABUPATEN	2019	2020	2021
1	BOGOR	475	417	430
2	SUKABUMI	112	110	117
3	CIANJUR	2	189	111
4	BANDUNG	179	239	225
5	GARUT	2	5	172

6	TASIKMALAYA	77	95	68
7	CIAMIS	95	2	76
8	KUNINGAN	91	48	113
9	CIREBON	257	251	232
10	MAJALENGKA	97	87	122
.	...	…	…	…
.	...	…	…	…
.	...	…	…	…
972	BANJAR	47	2	40

Pada gambar 3.2 diatas merupakan dataset kasus HIV/AIDS dengan jumlah 5 kolom yaitu no, kabupaten, umur, jumlah kasus dan tahun. Setelah dataset tersedia maka selanjutnya data tersebut di seleksi agar lebih sederhana dan spesifik.

3.6 Seleksi

Seleksi Data merupakan proses pemilihan data yang akan digunakan dalam proses *data mining*. Setelah data penelitian terkumpul, peneliti harus memilih atau menyeleksi data yang ada. Dalam proses pemilihan data, peneliti harus berhati-hati dan selalu berpedoman pada tujuan penelitian. Dengan menyeleksi data sesuai kebutuhan penelitian, maka peneliti bisa mendapatkan data-data yang lebih sederhana dan spesifik. Dalam penelitian ini peneliti hanya menggunakan data tahun 2021. Data asli yang didapat oleh peneliti terdiri dari tahun 2019 sampai tahun 2021 dengan sepuluh atribut yaitu id, kode provinsi, nama provinsi, kode kabupaten/kota, nama kabupaten/kota, jenis kelamin, kelompok umur, jumlah kasus, satuan dan tahun. Setelah itu tentukan penamaan *cluster* tersebut yaitu *Cluster* 0 = Rendah, *Cluster* 1 = Sedang*Cluster* 2 = Tinggi. Setelah menentukan cluster, lakukan penyeleksian data maka atribut data meliputi nama kota, tahun diantaranya 2019, 2020, 2021 seperti pada tabel 3.2 dibawah.

Tabel 3.2 Seleksi Data

NO	KABUPATEN	2019	2020	2021
1	BOGOR	475	417	430
2	SUKABUMI	112	110	117
3	CIANJUR	2	189	111

4	BANDUNG	179	239	225
5	GARUT	2	5	172
6	TASIKMALAYA	77	95	68
7	CIAMIS	95	2	76
8	KUNINGAN	91	48	113
9	CIREBON	257	251	232
10	MAJALENGKA	97	87	122
11	SUMEDANG	104	2	120
12	INDRAMAYU	151	275	113
13	SUBANG	339	247	222
14	PURWAKARTA	197	234	130
15	KARAWANG	255	315	244
16	BEKASI	221	134	239
17	BANDUNG BARAT	36	67	67
18	PANGANDARAN	26	16	4
19	KOTA BOGOR	443	364	333
20	KOTA SUKABUMI	57	41	43
21	KOTA BANDUNG	357	82	43
22	KOTA CIREBON	189	342	254
23	KOTA BEKASI	335	322	390
24	KOTA DEPOK	247	220	199
25	KOTA CIMAHI	44	361	513
26	KOTA TASIKMALAYA	105	2	99
27	KOTA BANJAR	47	2	40

3.7 Perhitungan K-Means Manual (Excel)

Pada tahapan ini *clustering* terhadap *dataset* dihitung secara manual menggunakan *Microsoft excel* agar mengetahui berapa nilai jarak antar klaster secara terperinci pada K-Means dengan rumus *Euclidiance Distance*. Adapun tahapannya sebagai berikut :

1. Menentukan jumlah *cluster*

   Data penyebaran hiv di provinsi Jawa Barat yang telah lolos tahap seleksi kemudian ditentukan jumlah clusternya yaitu pada penelitian ini akan dibentuk 3 *cluster*. Setelah itu ditentukan penamaan *cluster* tersebut yaitu *Cluster* 0 = Rendah, *Cluster* 1 = Sedang, *Cluster* 2 = Tinggi.

2. Tentukan pusat *centroid*

   Setelah ditentukan banyaknya *cluster* kemudian ditentukan nilai *centroid* awalnya, untuk pusat *centroid* terdiri dari Kabupaten Cirebon (C0),

Kabupaten Bekasi (C1), Kabupaten Bogor (C2). Adapun tabel centroid awalnya seperti dibawah ini.

Tabel 3.4 Centroid Awal

Centroid	Cirebon	Bekasi	Bogor
C0	20	35	53
C1	66	81	134
C2	887	1179	1852

Tabel diatas berisikan kolom *centroid* dan umur, lalu barisnya ada C0, C1, C2 dan jumlah kasus pada kabupaten Cirebon, Bekasi, dan Bogor.

3. Menghitung jarak objek

   Setiap *dataset* hiv dari tahun 2019-2021 berdasarkan Kabupaten di Provinsi Jawa Barat yang telah diseleksi, dihitung menggunakan rumus persamaan jarak *Euclidean Distance* sebagai berikut :

   Tabel 3.5 Rumus *Eulclidean Distance*

$$D(x,y) = \sqrt{\sum_{i=1}^{n} (xi - yi)^2}$$

   Keterangan :

   D(*x*,*y*) = jarak data *x* ke pusat cluster *y*

   x_i dan y_i = atribut ke-k dari objek data x dan y (pusat centroid)

   n = jumlah atribut

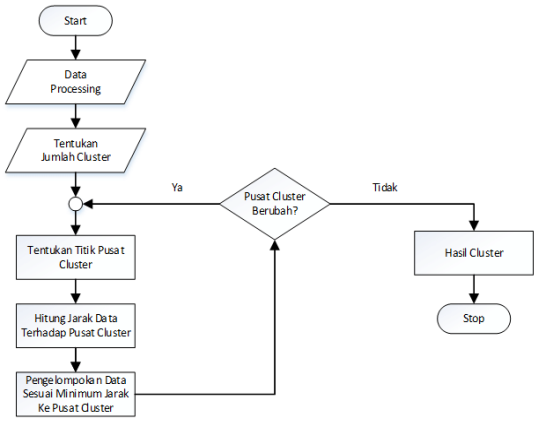
4. Mengelompokan objek berdasarkan jarak minimum

   Setiap data atribut yang telah dihitung menggunakan rumus *Euclidean Distance* akan didapatkan hasil atau nilai jarak clusternya.

3.8 Implementasi Algoritma K-Means dan RapidMiner

Perhitungan dengan program *python* dimaksudkan untuk memperkuat perhitungan sebelumnya yakni perhitungan manual dan menggunakan *rapidminer*. Dapat dilihat pada gambar 3.2 dibawah merupakan flowchart dari metode *k-means* yang digunakan dalam pengelompokan penyakit HIV/AIDS

di Jawa Barat, pada umumnya kinerja metode *k-means* secara berurutan adalah sebagai berikut :



Gambar 3.2 Flowchart Implementasi Algoritma *K-Means*

Pada gambar 3.2 diatas menjelaskan bahwa sebagai berikut :

1. Langkah pertama dimulai (***start***) dengan menggunkan simbol "**Terminator**" sebagai tanda awal atau akhir flowchart.
2. Selanjutanya masukan (***input***) data *processing* untuk menghasilkan informasi atau menghasilkan pengetahuan dari data mentah dengan menggunakan simbol "***Input/Output***" sebagai tanda memasukan dan keluaran suatu data.
3. Tentukan jumlah cluster yang ingin dikelompokan dengan menggunakan simbol "***Input/Output***" sebagai tanda memasukan dan keluaran suatu data.
4. Tentukan titik pusat *cluster* (*centroid*) yang akan dihitung kembali sampai semua komponen data digolongkan kedalam tiap-tiap cluster dan terakhir akan terbentuk *cluster* baru dengan menggunakan simbol "***Process***" sebagai tanda untuk menyatakan suatu proses yang dilakukan komputer.
5. Kemudian hitung jarak data terhadap pusat *cluster* dengan menggunakan simbol "***Process***".
6. Kelompokan data sesuai minimum jarak ke pusat *cluster* dengan menggunakan symbol "***Process***".
7. Jika "**Ya**" pusat *cluster* berubah, maka dihitung kembali titik pusat clusternya sampai terbentuk *cluster* baru dengan menggunakan simbol "***Decission***" sebagai tanda kondisi tertentu yang akan menghasilkan dua kemungkinan yaitu "Ya/Tidak" dan lakukan kembali langkah 5 dan 6.

8. Jika "**Tidak**" tidak berubah pusat clusternya, maka perhitungan hasil cluster sudah selesai dengan menggunakan simbol "**Terminator**".

3.8.1 Implementasi Pengolahan Data pada *RapidMiner*

Berikut adalah cara implementasi pengolahan data menggunakan algoritma k-means dengan *tools RapidMiner* :

1. Data *Selection*

   Pada aplikasi RapidMiner, fungsi "*Excel Reading*" berfungsi sebagai pembaca file Excel. Operator ini digunakan untuk mengimpor atau memasukkan data Excel di komputer pengguna ke dalam proses speedminer. *Read Excel* adalah operator dasar yang digunakan sebelum memulai suatu proses. Operator ini dapat digunakan untuk memuat data dari spreadsheet Microsoft Excel. Basis data yang digunakan adalah basis data kasus HIV/AIDS di Jawa Barat.

2. *Preprocessing*

   Dikarenakan pada *dataset* kasus HIV/AIDS tidak ditemukan *missing value* atau data yang tidak memiliki nilai maka *preprocessing* tidak dilakukan.

3. Data *Transformation*

   Nominal to Polynominal, digunakan untuk mengubah tipe atribut yang bersifat non-numerik menjadi tipe polynominal. Dalam dataset kasus HIV/AIDS atribut tahun menjadi polynominal.

4. Data *Mining*

   Pada tahapan ini dilakukan dengan metode algoritma *kmeans clustering*.

5. *Evaluation*

   Setelah melakukan perbandingan SSE dengan metode *K-Means* dari k-3 sampai k-15 yang terdapat pada tabel diatas, dapat dilihat bahwa cluster yang mendekati 0 yaitu k-3, dengan nilai SSE sebesar 0,511 Karena nilai k-3 merupakan nilai terkecil dibandingkan k lainnya, maka dapat disimpulkan bahwa k-3 dengan nilai 0,511 yang paling mendekati 0 merupakan hasil *cluster* terbaik.

3.9 Evaluasi

Pada penelitian ini, tujuan evaluasi dilakukan untuk mengetahui nilai validitas dari pengelompokan atau *clustering* yang telah dihasilkan dari perhitungan k-means dengan menggunakan metode *Sum of Square Error* (SSE). Untuk menghitung SSE menggunakan rumus :

Tabel 3.6 Rumus Menghitung SSE

$$SSE = \sum_{K=1}^{K} \sum_{xisSk} ||Xi - Ck||2$$

Keterangan :

K = jumlah cluster

xi = data ke – i

Ck = centroid cluster

Sum of squared errors atau SSE adalah perhitungan statistik asli yang digunakan untuk menghitung nilai lain. Jika memiliki kumpulan data, maka dapat mencari hubungan antara angka-angka dalam data tersebut dan mengatur data ke dalam tabel dan kemudian melakukan perhitungan sederhana. Setelah SSE dari kumpulan data diperoleh, maka dapat melanjutkan untuk menghitung varians dan standar deviasi.

1. Buat tabel dengan tiga kolom itu dengan Angka, Deviasi, dan Deviasi^2. Masukan data pada kolom pertama angka hasil pengukuran perhitungan persentase bisa diperoleh dari hasil eksperimen, penelitian statistik, atau data yang sudah diambil dari *website* tertentu.
2. Hitung rata-rata
3. Kurangkan rata-rata jumlah value dengan nilai value pertama dan seterusnya, lalu dari setiap hasil pengukuran dimasukan ke kolom kedua dengan hasilnya.
4. Untuk kolom ketiga, cari nilai kuadrat dari setiap nilai pada kolom tengah. Angka ini adalah kuadrat selisih dari setiap pengukuran dengan rata-rata.
5. Untuk data ini, SSE diperoleh dengan menjumlahkan seluruh nilai pada kolom ketiga.

3.10 Peralatan Penelitian

Peralatan yang digunakan pada penelitian ini meliputi perangkat keras dan perangkat lunak dengan rincian sebagai berikut :

1. Perangkat Keras

   Model Laptop HP Convertible x360 11-ab0XX dengan spefisikasi sebagai berikut :

   a. Processor Intel ® Celeron CPU N3060
   b. RAM 4GB
   c. X64-based PC

2. Perangkat Lunak

   a. Microsoft Office Excel 2013
   b. Google Colab
   c. *RapidMiner Studio Version* 10.2