BAB III
METODE PENELITIAN

3.1 Objek Penelitian

Penelitian ini berfokus pada tingkat perceraian di provinsi Jawa Barat dengan cara pengelompokan data perceraian berdasarkan administratif Kabupaten/Kota di Provinsi tersebut. Sumber data yang digunakan berasal dari situs resmi *opendata.jabarprov.go.id* dan mencakup data kasus perceraian yang terjadi sepanjang tahun 2023.

Wilayah penelitian meliputi seluruh Kabupaten/Kota di Provinsi Jawa Barat, yang dikenal sebagai provinsi dengan jumlah penduduk terbesar di Indonesia serta mencatat tren perceraian yang sangat signifikan. Dengan tren tersebut, Jawa Barat secara konsisten masuk dalam tiga besar provinsi dengan jumlah perceraian tertinggi di Indonesia. Penelitian ini bertujuan untuk menganalisis bagaimana algoritma *K-Means* dan *K-Medoids* bekerja dalam mengelompokkan tingkat perceraian di Jawa Barat, dengan mempertimbangkan berbagai faktor yang mempengaruhi perceraian di setiap Kabupaten/Kota dalam Provinsi tersebut.

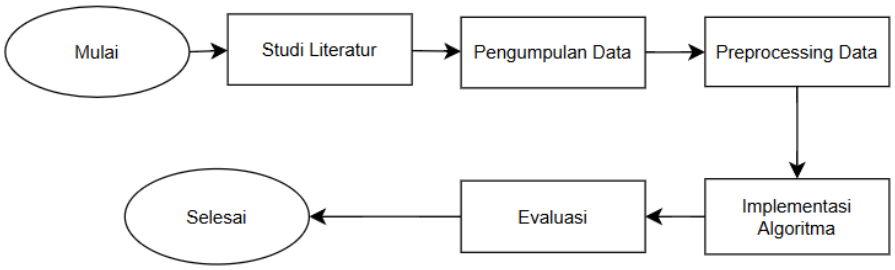
3.2 Waktu Penelitian

Penelitian ini dilaksanakan selama enam bulan dan terdiri dari lima tahapan utama. Tahap pertama, yaitu studi literatur, dilakukan pada akhir bulan pertama hingga pertengahan bulan kedua. Pengumpulan data berlangsung dari pertengahan bulan kedua hingga awal bulan ketiga. Selanjutnya, tahap *preprocessing* data dilaksanakan mulai pertengahan bulan ketiga hingga pertengahan bulan keempat. Implementasi algoritma dilakukan dari akhir bulan keempat hingga pertengahan bulan kelima. Tahap akhir, yaitu evaluasi, berlangsung dari akhir bulan kelima hingga akhir bulan keenam. Seluruh kegiatan ini disusun agar penelitian berjalan sistematis dan efisien sesuai dengan target waktu yang telah ditetapkan.

3.3 Prosedur Penelitian

Proses perancangan dalam penelitian ini dilakukan secara sistematis melalui beberapa tahapan, yaitu studi literatur, pengumpulan data, *preprocessing* data,

Implementasi algoritma dan evaluasi data. Rangkaian tahapan penelitian tersebut dapat dilihat pada Gambar 3.1.



Gambar 3.1 Tahapan Penelitian

3.3.1. Studi Literatur

Studi literatur dilakukan dengan mengeksplorasi berbagai sumber referensi yang relevan, seperi jurnal ilmiah, skripsi, atau dokumen pendukung lainnya yang membahas penerapan *data mining*, khususnya metode *clustering* dengan algoritma *K-Means* dan *K-Medoids*. Sumber informasi dalam studi literatur ini diperoleh melalui penelusuran secara daring (internet) maupun secara luring, seperti perpustakaan.

3.3.2. Pengumpulan Data

Data yang digunakan dalam penelitian ini diperoleh dari situs resmi *opendata.jabarprov.go.id*. Dataset tersebut memuat informasi mengenai berbagai faktor penyebab perceraian di Provinsi Jawa Barat pada tahun 2023.

```
Informasi dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 338 entries, 0 to 337
Data columns (total 8 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   id                       338 non-null    int64
 1   kode_provinsi            338 non-null    int64
 2   nama_provinsi            338 non-null    object
 3   wilayah_pengadilan_agama 338 non-null    object
 4   faktor_penyebab          338 non-null    object
 5   jumlah_perceraian        338 non-null    int64
 6   satuan                   338 non-null    object
 7   tahun                    338 non-null    int64
dtypes: int64(4), object(4)
memory usage: 21.3+ KB
```

Gambar 3.2 df.info()

Dataset ini berisi 338 entri dengan 8 kolom yang mencakup informasi terkait perceraian dan pengadilan agama di Jawa Barat. Kolom-kolom dalam dataset mencakup identifikasi unik (id), kode dan nama provinsi, wilayah pengadilan agama, faktor penyebab perceraian, jumlah perceraian, satuan pengukuran, serta tahun pencatatan data. Data ini mencakup tipe data numerik (int64) untuk kode provinsi, jumlah perceraian, dan tahun, serta teks (object) untuk nama provinsi, wilayah pengadilan agama, faktor penyebab, dan satuan. Semua kolom dalam dataset ini memiliki nilai lengkap (*non-null*).

3.3.3. Preprocessing Data

Tahapan berikutnya dalam penelitian ini adalah *preprocessing* data. Proses *preprocessing* mencakup beberapa langkah penting, yaitu seleksi data, pembersihan data, transformasi data, serta normalisasi data.

a. Seleksi Data

Proses seleksi data bertujuan untuk memilih atribut-atribut yang relevan untuk analisis. Sebelum seleksi, dataset memiliki 8 atribut, yaitu id, kode_provinsi, nama_provinsi, wilayah_pengadilan_agama,faktor_penyebab, jumlah_perceraian, satuan, dan tahun. Dalam proses seleksi, fungsi .drop() digunakan untuk menghapus kolom yang tidak relevan. Setelah seleksi, data yang tersisa hanya terdiri dari 2 atribut, yaitu wilayah pengadilan agama dan faktor penyebab.

b. Pembersihan Data

Pembersihan data bertujuan untuk menghapus informasi yang tidak relevan atau tidak memberikan kontribusi pada analisis, seperti data kosong *(missing values)*. Data yang hilang dapat ditangani dengan beberapa cara, seperti menghapus baris yang memiliki nilai kosong, mengganti nilai kosong dengan angka 0, atau mengisinya menggunakan nilai rata-rata (mean), nilai tengah (median), atau niali yang sering muncul (modus). Nilai rata-rata dihitung dari keseluruhan data yang tersedia, median diisi dengan nilai tengah dari data yang telah diurutkan, dan modus diisi dengan nilai yang paling sering muncul. Selain itu, data duplikat juga perlu diidentifikasi dan dihapus agar analisis tidak

terganggu. data yang terduplikasi dapat dihapus jika tidak memiliki informasi tambahan, atau digabungkan jika terdapat nilai berbeda yang relevan.

c. Transformasi Data

Transformasi data merupakan proses mengubah format, struktur kolom, atau skala data menjadi bentuk yang lebih sesuai untuk analisis. Salah satu bentuk transformasi dilakukan dengan menyederhanakan data kategori melalui proses pelabelan, seperti mengubah "CACAT BADAN" menjadi "CD" atau "EKONOMI" menjadi "EK" dan seterusnya, selanjutnya mengonversi kolom wilayah menjadi nilai numerik, seperti "BANDUNG" menjadi 0, "BEKASI" menjadi 1, "BOGOR" menjadi 2, dan seterusnya.

d. Seleksi Fitur

Pada tahap seleksi fitur, dipilih faktor-faktor penyebab perceraian dengan jumlah kasus lebih dari 1.000 secara nasional, guna memfokuskan analisis pada variabel variabel yang paling dominan serta mengurangi gangguan dari fitur dengan frekuensi yang terlalu rendah.

e. Normalisasi Data

Normalisasi data merupakan proses untuk mengubah nilai-nilai atribut ke dalam skala tertentu agar semua atribut memiliki bobot seimbang dalam analisis. Salah satu tektik normalisasi yang digunakan adalah *Min-Max Scaling*, yaitu metode yang mengkonversi nilai data ke dalam rentang 0 hingga 1 dengan mengurangkan nilai minimum dari setiap data dan membaginya dengan rentang (selisih antara nilai maksimum dan minimum) dari data tersebut.

3.3.4. Implementasi Algoritma

Setelah melakukan tahapan *processing* data, langkah selanjutnya adalah implementasi *elbow method* untuk menentukan jumlah *cluster*. Kemudian, proses klasterisasi dilanjutkan dengan menggunakan algoritma *K-Means* dan *K-Medoids* dengan pemrograman *Python*.

Algoritma 1 menentukan jumlah *cluster*

Menentukan jumlah *cluster* dengan *elbow method*
Input : Membaca dataset hasil *preprocessing*

Output : Jumlah *cluster* optimal, plot *Elbow Curve*, dan nilai *SSE*

1. Input dataset hasil *preprocessing*
2. Menentukan rentang jumlah *cluster* yang akan diuji
3. Melakukan *clustering* untuk setiap jumlah *cluster* lakukan proses *clustering* menggunakan algoritma *k-means* dan *k-medoids* untuk setiap jumlah *cluster* yang telah ditentukan
4. Menghitung inertia (*Sum Square Error-SSE*) berdasarkan persamaan (1)
5. Membuat plot jumlah *cluster* pada sumbu-x dan nilai inertia pada sumbu-y untuk membentuk *Elbow Curve*
6. Mengidentifikasi titik yang membentuk sudut siku-siku (*elbow*) pada kurva
7. Menampilkan hasil *cluster* optimal, plot *Elbow Curve*, dan nilai *SSE*

Algoritma 2 *clustering* tingkat perceraian berdasarkan Kabupaten/Kota di Provinsi Jawa Barat

Algoritma : *Clustering* tingkat perceraian berdasarkan Kabupaten/Kota di Provinsi Jawa Barat menggunakan *K-Means*

Input : Membaca dataset hasil *preprocessing*

Output : Centroid *cluster* dan label *cluster* untuk setiap data.

1. Membaca dataset hasil *preprocessing*
2. Menentukan banyaknya *cluster* yang dibentuk gunakan hasil dari *elbow* untuk menentukan jumlah *cluster* optimal
3. Memilih secara acak sejumlah data sebagai titik pusat *cluster* (*centroid*) sesuai denngan jumlah *cluster* yang telah ditentukan.
4. Menghitung jarak setiap data menggunakan rumus persamaan (2)
5. Mengelompokkan data ke dalam *cluster* berdasarkan jarak terdekat terhadap *centeroid*.
6. Menghitung ulang pusat *cluster* (*centeroid*) berdasarkan rata-rata dari anggota dalam setiap *cluster*, sesuai dengan rumus pada persamaan (3)
7. Melakukan perulangilah langkah ke-3 sampai ke-5 hingga keanggotaan setiap *cluster* tetap
8. Menampilkan Centroid *cluster* dan label *cluster* untuk setiap data.

Algoritma 3 *clustering* tingkat perceraian berdasarkan Kabupaten/Kota di Provinsi Jawa Barat

Algoritma : *Clustering* tingkat perceraian berdasarkan Kabupaten/Kota di Provinsi Jawa Barat *K-Medoids*

Input : Membaca dataset hasil *preprocessing*

Output : Centroid *cluster* dan label *cluster* untuk setiap data.

1. Membaca dataset hasil *preprocessing*
2. Tentukan pusat awal *cluster* sebanyak, sejumlah dengan jumlah *cluster* yang diinginkan.
3. tempatkan setiap data atau objek ke dalam *cluster* terdekat menggunakan ukuran jarak *Euclidean Distance* dengan persamaan (4)
4. Tentukan secara acak objek pada masing-masing *cluster* sebagai kandidat medoid baru.
5. Hitung jarak setiap objek yang berada pada setiap masing-masing *cluster* dengan menempuh medoids baru.
6. Hitung selisih total jarak (S) dengan mengurangkan total jarak baru dengan total jarak sebelumnya. Jika Skudang dari 0, maka objek tersebut dapat diganti dengan data dari *cluster* untuk memperoleh himpunan baru sebanyak k objek sebagai medoids.
7. Lakukan pengulangan dari langkah ke 3 hingga ke 5 sampai tidak terjadi perubahan pada medoid, sehingga diperoleh *cluster* beserta anggota-anggotanya. Selanjutnya pemilihan jumlah k yang optimal dalam *clustering* masing-masing. Kemudian untuk mendapatkan nilai k di sebuah data yang ada *K-Medoid* dapat ditentukan berdasarkan nilai *DBI (Davies Bouldin Index)* yang paling rendah.
8. Menampilkan Centroid *cluster* dan label *cluster* untuk setiap data.

3.3.5. Evaluasi

Pada penelitian ini, proses evaluasi dilakukan untuk menilai seberapa baik kualitas pengelompokan objek dalam setiap *cluster* yang dihasilkan oleh algoritma *K-Means* dan *K-Medoids*. Evaluasi tersebut menggunakan dua metode, yaitu *Davies Bouldin Index (DBI)* dan *Silhouette Coefficient*. Untuk menilai kualitas hasil

clustering, dapat dihitung nilai *silhouette* untuk setiap *cluster (*Kurniawan *et al., 2023).*

a. *Davies Bouldin Index (DBI)*

Rumus untuk menghitung nilai *Davies-Bouldin Index (DBI)* adalah dengan menjumlahkan nilai maksimum dari rasio antara *cluster-cluster*, kemudian membaginya dengan jumlah *cluster* atau K. Berikut adalah persamaan untuk menghitung nilai *DBI* :

$DBI = \frac{1}{k}\Sigma_{i=1}^{k} max_{i \neq j}$R(i,j) (5)

Keterangan :

K:Jumlahcluster

$max_{i \neq j}$R(i,j) : nilai maksimum dari rasio antara *cluster* i dan *cluster* j.

b. *Silhouette Coefficient*

Mengevaluasi seberapa baik suatu objek ditempatkan dalam *cluster* dengan membandingkan jarak rata-rata objek ke anggota *cluster* lain dengan jarak rata-rata ke *cluster* lain. Dengan persamaan :

$sil(c) = sil(k)\frac{1}{|k|}\Sigma_{i=1}^{k}\ sil(c_i)$ (6)

Dimana :

sil(k) : nilai *silhouette* semua *cluster*

|k| : banyaknya *cluster* k

sil (ci) : rata-rata nilai *silhouette*