BAB III
METODE PENELITIAN

3.1 Objek Penelitian

Objek yang digunakan dalam penelitian ini berupa data jumlah penerima layanan pencegahan *stunting* tahun 2023. Data tersebut masih berupa data mentah dengan format *excel*. Sumber data yang digunakan merupakan data sekunder yang diperoleh dari situs Satu Data Indonesia https://data.go.id/dataset/dataset/jumlah-penerima-layanan-pencegahan-stunting-tahun-2023 yang diakses pada 2 Desember 2024. Data bersumber dari Kementerian Desa, Pembangunan Daerah Tertinggal, dan Transmigrasi Republik Indonesia yang dipublikasikan pada 10 Oktober 2024.

Penelitian ini dilakukan di lab riset Universitas Buana Perjuangan Karawang yang dimulai pada bulan November 2024. Penelitian direncanakan dilakukan dalam jangka waktu 6 bulan. Dengan jangka waktu tersebut, diharapkan penelitian ini dapat menghasilkan model klasterisasi yang efektif untuk mencapai tujuan penelitian. Adapun gambaran waktu penelitian tersaji pada tabel 3.1 dibawah ini.

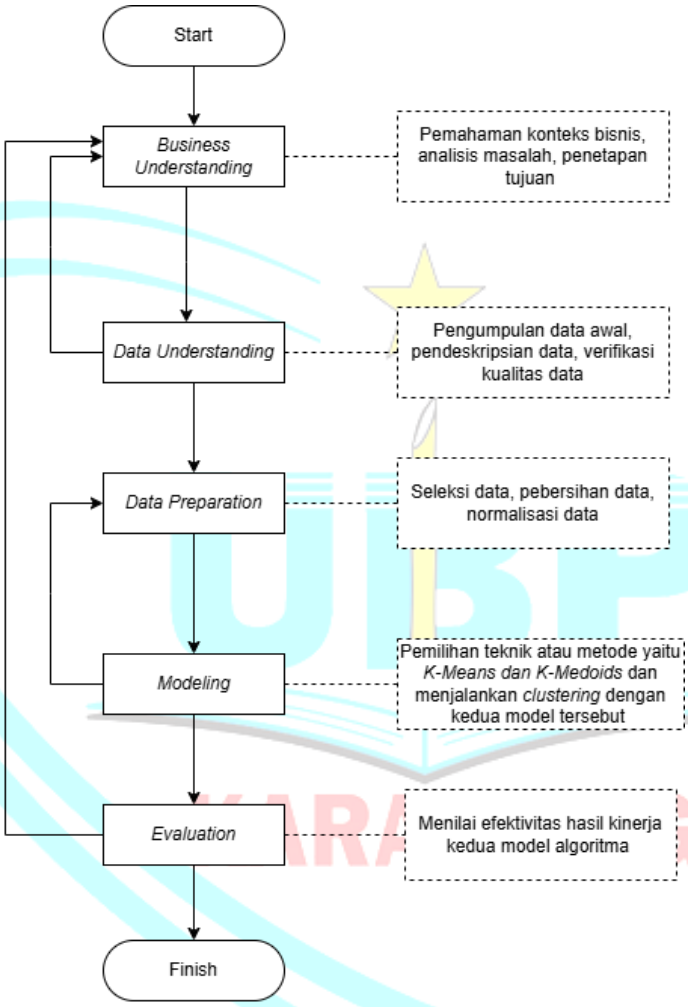
Tabel 3. 1 Waktu Penelitian

Kegiatan	Waktu Penelitian															
	Bulan 1				Bulan 2				Bulan 3				Bulan 4			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Business Understanding																
Data Understanding																
Data Preparation																
Modeling																

Kegiatan	Waktu Penelitian															
	Bulan 5				Bulan 6				Bulan 7				Bulan 8			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Evaluation																

3.2 Prosedur Penelitian

Cross Industry Standard Process of Data Mining (CRISP-DM) diterapkan dalam penelitian ini dengan rangkaian tahapan yang ditunjukan pada Gambar 3.1.



Gambar 3. 1 Tahapan Metode Penelitian

Seluruh langkah pada prosedur penelitan dilakukan secara berkesinambungan Setiap tahapan saling berkaitan satu sama lain. Sehingga jika terdapat tahapan yang belum selesai dilakukan, maka tidak dapat dilanjutkan ke tahapan berikutnya.

3.2.1 Business Understanding

Pada tahap *business understanding*, dilakukan pemahaman terhadap tujuan bisnis, penilaian situasi dan penentuan tujuan yang ingin dicapai dengan *data mining*. Tahapan ini menjadi fondasi awal dalam proyek analisis karena akan menentukan arah dan ruang lingkup penelitian. Didalamnya dilakukan identifikasi secara rinci mengenai tujuan utama bisnis atau penelitian. Dalam konteks penelitian

ini akan mengelompokan jumlah keluarga beresiko stunting seluruh desa di Jawa Barat.

Selain itu, dilakukan pula penilaian situasi termasuk kendala yang ada, seperti tingginya angka *stunting* di Jawa Barat dan belum optimalnya pemetaan wilayah prioritas. Proses ini juga mencakup perumusan tujuan *data mining* yang spesifik, yaitu dengan mengimplementasikan algoritma *K-Means* dan *K-Medoids* untuk pengelompokan jumlah keluarga beresiko stunting seluruh desa di Jawa Barat.

3.2.2 Data Understanding

Pada langkah ini, data akan dikumpulkan berdasarkan kebutuhan yang ditetapkan pada tahap sebelumnya. Selanjutnya, data akan dieksplorasi untuk memahami karakteristik data. Evaluasi terhadap data juga dilakukan untuk memastikan kualitas data.

Penelitian ini akan menggunakan data jumlah penerima layanan pencegahan *stunting* tahun 2023. Data tersebut diperoleh dari situs Satu Data Indonesia Indonesia https://data.go.id/dataset/dataset/jumlah-penerima-layanan-pencegahan-stunting-tahun-2023 yang berupa data mentah dengan format *excel (xlsx)*.

Tabel 3. 1 Jumlah penerima layanan pencegahan *stunting* tahun 2023

Kode_ Prov	Nama_ Provinsi	Kode_ Kab	Nama_ Kabupaten	…	…	Kendala_yang_ dihadapi_ (Jelaskan)
11	Aceh	1101	Aceh Selatan	…	…	0
11	Aceh	1101	Aceh Selatan	…	…	0
11	Aceh	1101	Aceh Selatan	…	…	Tidak Ada
11	Aceh	1101	Aceh Selatan	…	…	Kurangnya sumber daya manusia
…	…	…	…	…	…	…
…	…	…	…	…	…	…
91	Papua	9110	Sarmi	…	…	Kurangnya Anggaran

Data jumlah penerima layanan pencegahan *stunting* tahun 2023 berisi 75.265 *record* dan 67 atribut yang mencakup seluruh wilayah di Indonesia. Namun, penelitian ini akan fokus terhadap data yang terkait

wilayah Provinsi Jawa Barat. Hal ini dilakukan dengan menyimpan data dimana Nama_Provinsi berisi Jawa Barat. Data tersebut berjumlah 5.311 *record.* Berikut adalah tabel 3. 3 yang berisi data jumlah penerima layanan pencegahan *stunting* tahun 2023 Provinsi Jawa Barat

Tabel 3. 2 Jumlah penerima layanan pencegahan *stunting* tahun 2023 Provinsi Jawa Barat

Kode_ Prov	Nama_ Provinsi	Kode_ Kab	Nama_ Kabupaten	…	…	Kendala_yang_ dihadapi_ (Jelaskan)
32	Jawa Barat	3201	Bogor	…	…	0
32	Jawa Barat	3201	Bogor	…	…	0
32	Jawa Barat	3201	Bogor	…	…	0
32	Jawa Barat	3201	Bogor	…	…	0
…	…	…	…	…	…	…
…	…	…	…	…	…	…
32	Jawa Barat	3279	Kota Banjar	…	…	0

Data yang akan digunakan perlu dipastikan kualitasnya. Evaluasi terhadap data jumlah penerima layanan pencegahan *stunting* tahun 2023 diperlukan untuk memastikan data memiliki kualitas baik. Salah satu kriterianya yaitu tidak terdapat nilai yang hilang. Dataset yang baik diharapkan dapat mencapai tujuan penelitian.

3.2.3 Data Preparation

Tahapan ini dilakukan dengan tujuan untuk membangun dataset yang layak digunakan dalam pemodelan. Tahapan ini melibatkan beberapa langkah, diantaranya yaitu *select data, cleaning data,* dan *construct data*. *Select data* atau seleksi data dilakukan untuk menyeleksi fitur-fitur yang diperlukan sesuai dengan kebutuhan penelitian. *Select data* akan dilakukan dengan metode *pearson correlation coeficient* untuk mencari fitur-fitur dengan nilai korelasi yang kuat dengan fitur yang dijadikan target.Berikut merupakan persamaan untuk menghitung *pearson correlation coeficient.*

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\ \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (3.1)$$

Dimana :

r : nilai *pearson correlation coefficient*

x_i, y_i : nilai ke-i dari masing-masing variabel

$\bar{x}$,$\bar{y}$: rata-rata dari masing-masing variabel x, dan y

Selanjutnya dilakukan *cleaning data* atau pembersihan data. Pada saat proses *cleaning dat*a, akan dilakukan pembersihan pada data untuk memastikan data bersih dan layak digunakan. Beberapa kriterianya yaitu bersih dari *missing value* dan *duplicated data.*

Tahap selanjutnya yaitu *construct data* yang bertujuan membuat dan memastikan struktur data sesuai untuk proses *modeling.* Salah satu proses yang dapat dilakukan dalam tahap *construct data* yaitu normalisasi data. Normalisasi atau standarisasi diperlukan untuk memastikan data pada skala yang sama. Berikut merupakan rumus untuk mencari nilai *z-score* seperti pada persamaan 2..

$$z = \frac{x - \mu}{\sigma} \tag{3.2}$$

Dimana:

z = Nilai *z-score*

x = Nilai yang ingin diubah menjadi *z-score*

μ = Rata-rata populasi

σ = Standar deviasi populasi

Untuk mencari nilai rata-rata perlu menghitung jumlah nilai-nilai dalam populasi dibagi dengan jumlah banyaknya data dalam populasi. Berikut adalah rumusnya.

$$\mu = \frac{\sum_{i=1}^{n} X_i}{n} \tag{3.3}$$

Dimana:

μ = Nilai rata-rata populasi

X_i = Nilai data ke- i

n = Jumlah total data (banyaknya nilai dalam dataset)

Adapun berikut merupakan formula yang digunakan untuk menghitung standar deviasi.

$$\sigma = \sqrt{\frac{1}{n}\sum_{i-1}^{n}(X_i - \mu)^2} \qquad (3.4)$$

3.2.4 Modeling

Tahap *modeling* diawali dengan menentukan model algoritma yang akan digunakan. Penelitian ini akan menggunakan algoritma *K-Means* dan *K-Medoids* dalam mengelompokkan desa di Jawa Barat berdasarkan kelompok keluarga beresiko *stunting*. Kedua model algoritma tersebut akan diuji dan dibandingkan untuk mengetahui kinerja model terbaik diantara keduanya. Berikut merupakan penjelasan klasterisasi dengan model algoritma klasterisasi *K-Means* dan *K-Medoids.*

Algoritma 1 Menentukan jumlah klaster dengan metode elbow

Algoritma menentukan jumlah klaster
Input : Dataset Ouput : Jumlah klaster 1. Membaca dataset 2. Menentukan jumlah klaster dengan metode *elbow* dengan persamaan 2.2 3. Hasil jumlah klaster optimal metode *elbow*

Algoritma 2 Klasterisasi dengan *K-Means*

Algoritma klasterisasi *K-Means*
Input : • Dataset • Jumlah klaster Ouput : Data memberntuk klaster 1. Membaca dataset 2. Membaca jumlah klaster hasil metode *elbow* 3. Memilih *centroid* awal 4. Menghitung jarak setiap data ke titik *centroid* dengan persamaan 2.1

5. Menghitung *centroid* baru dengan nilai rata-rata dari *cluster* yang terbentuk
6. Melakukan iterasi sampai posisi *centroid* stabil (tidak berubah)
7. Menampilkan data yang telah membentuk klaster

Algoritma 3 Klastersasi dengan *K-Medoids*

Algoritma klasterisasi *K-Medoids*
Input : • Dataset • Jumlah klaster
Ouput : Data membentuk klaster
1. Membaca dataset yang sudah disiapkan 2. Membaca jumlah klaster optimal hasil metode *elbow* 3. Memilih *medoid* awal 4. Menghitung jarak setiap data ke *medoid* untuk membentuk *cluster* berdasarkan jarak terdekat dan hitung total jarak terdekat dengan persamaan 2.1. 5. Membentuk *medoid* baru dan jalankan proses ke 4. 6. Hitung deviasi total (S) dengan cara membandingkan nilai total jarak terdekat baru dengan total jarak terdekat sebelumnya. Jika S>0, maka *clustering* selesai. 7. Menampilkan data yang telah membentuk klaster

3.2.5 *Evaluation*

Pada tahap ini, evaluasi akan dilakukan dengan penerapan metode *Davies-Bouldin Index (DBI)* dan *silhouette score.* Evaluasi dilakukan untuk mengetahui efektifitas dan kualitas model yang dihasilkan. Berdasarkan perbandingan hasil evaluasi dari model *K-Means* dan *K-Medoids,* akan diketahui model dengan kinerja terbaik diantara keduanya.

Dalam evaluasi menggunakan metode DBI, nilai DBI yang semakin rendah (non-negatif, ≥ 0) menunjukan kualitas *clustering* semakin optimal.

Rumus yang digunakan dalam menghitung nilai kohesi menggunakan *Sum of square within cluster (SSW)* pada *cluster-i* adalah sebagai berikut.

$$SSW_i = \frac{1}{m_i} \sum_{j=1}^{m_i} d\,(x_j, c_i) \qquad (3.5)$$

Dalam persamaan SSW, m_i merupakan jumlah data dalam *cluster-i* sementara c_i merupakan *centroid* dari *cluster* tersebut. Sedangkan, nilai separasi dihitung menggunakan *Sum of square between cluster (SSB)* dengan rumus sebagai berikut.

$$SSB_{i,j} = d(c_i, c_j) \tag{3.6}$$

Setelah nilai kohesi dan separasi didapatkan, tahap berikutnya yaitu menghitung rasio pengukuran ($R_{i,j}$) dengan rumus sebagai berikut.

$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{i,j}} \tag{3.7}$$

Kemudian, nilai rasio tersebut digunakan untuk menghitung nilai DBI dengan rumus sebagai berikut.

$$DBI = \frac{1}{k}\sum_{i=1}^{k} max_{j \neq j}\,(R_{i,j}) \tag{3.8}$$

Dalam rumus DBI, nilai k menunjukan jumlah *cluster* yang digunakan. Nilai DBI yang minimal menunjukan bahwa pengelompokkan antar *cluster* tersebut optimal. Adapun dalam menghitung evaluasi dengan *silhouette score* dapat menggunakan rumus sebagai berikut.

$$sil(c) = sil(k)\frac{1}{|k|}\sum_{i=1}^{k} sil(c_i) \tag{3.9}$$

Dimana:

$sil(k)$ = nilai *silhouette* semua *cluster*

|k| = jumlah *cluster* k

$sil(c)$ = rata-rata nilai *silhouette*