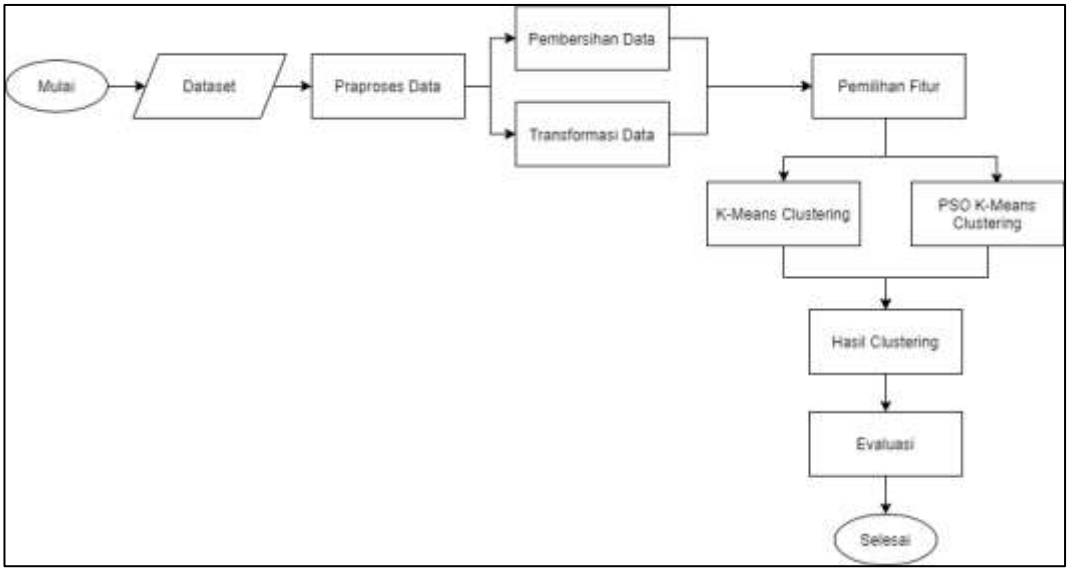
BAB III
METODOLOGI PENELITIAN

3.1 Diagram Alir Penelitian

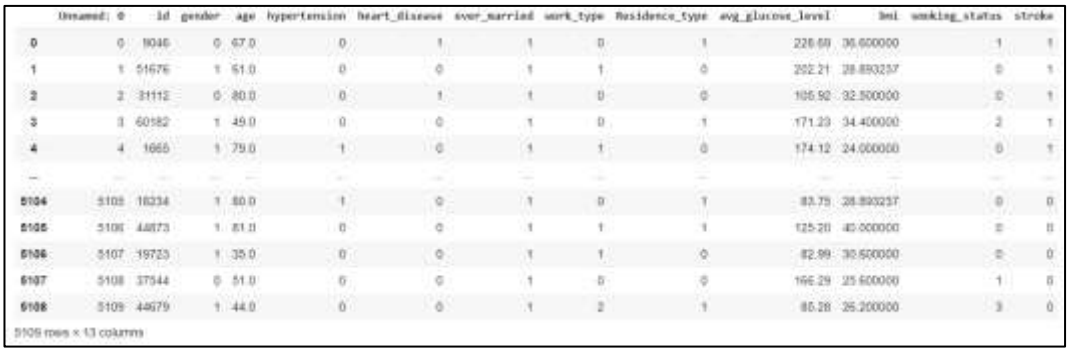
Tahapan-tahapan yang perlu dilakukan yaitu praproses data yang terdiri dari pembersihan data dan transformasi data, pemilihan fitur, tahap *clustering* dengan algoritma Kmeans dan PSO Kmeans, hasil *clustering*, dan evaluasi hasil *clustering*. Diagram alir penelitian dapat dilihat pada Gambar III.1.



Gambar III-1 Diagram Alir Penelitian

3.1.1 Data

Pada penelitian data yang digunakan adalah data penyakit stroke yang diperoleh dari Kaggle. Kaggle merupakan sebuah web atau situs yang menyediakan ribuan dataset tentang data science dan machine learning yang dapat diunduh dalam format csv. Data penyakit stroke (Gambar III) akan diolah dengan menggunakan algoritma Kmeans *clustering* dan PSO Kmeans *clustering*.



	Unnamed: 0	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	0	9046	0	67.0	0	1	1	0	1	228.69	36.600000	1	1
1	1	51676	1	61.0	0	0	1	1	0	202.21	28.893237	0	1
2	2	31112	0	80.0	0	1	1	0	0	105.92	32.500000	0	1
3	3	60182	1	49.0	0	0	1	0	1	171.23	34.400000	2	1
4	4	1665	1	79.0	1	0	1	1	0	174.12	24.000000	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...
5104	5103	18234	1	80.0	1	0	1	0	1	83.75	28.893237	0	0
5105	5106	44873	1	81.0	0	0	1	1	1	125.20	40.000000	0	0
5106	5107	19723	1	35.0	0	0	1	1	0	82.99	30.600000	0	0
5107	5108	37544	0	51.0	0	0	1	0	0	166.29	25.600000	1	0
5108	5109	44679	1	44.0	0	0	1	2	1	85.28	26.200000	3	0

5109 rows × 13 columns

Gambar III-2 Data Stroke

3.1.2 Seleksi Fitur

Seleksi fitur atau *fitur selection* dilakukan dengan menggunakan metode *Extra Trees Clasifier* (ETC). *Extra Trees Clasifier* (ETC) akan melakukan pemilihan fitur dengan cara membagi node dan membangun setiap *tree* dengan menggunakan data stroke. Seleksi fitur bertujuan untuk memilih atribut dari factor penyebab penyakit stroke yang relevan terhadap penyakit stroke. Berikut ini merupakan langkah-langkah dalam seleksi fitur:

1. Data yang akan digunakan dalam seleksi fitur

Gender	**Work_type**	**Resident_type**	**Smoking_status**	**stroke**
Male	Private	Urban	Formerly smoked	1
Female	Self-employed	Rural	Never smoked	1
Male	Private	Rural	Never smoked	0
Female	Private	Urban	Smokes	1
…	…	…	…	…
Male	Private	Rural	Never smokes	1

2. Mengitung nilai entropy dan gain

		Jumlah kasus	**(1)**	**(0)**	**Entropy**	**Gain**
Total		9	5	4	0,891	

		Jumlah kasus	(1)	(0)	Entropy	Gain
Gender						0,007
	Male	4	2	2	1	
	Female	5	3	2	0,969	
Work_type						0,091
	Private	6	3	3	1	
	Self-employed	2	1	1	1	
	Govt_job	1	1	0	0	
Resident_type						0,016
	Urban	3	2	1	0,825	
	Rural	6	3	3	1	
Smoking_status						0,207
	formerly smoked	2	1	1	1	
	never smoked	5	2	3	0,969	
	smokes	2	2	0	0	

a. Perhitungan nilai entropy untuk “gender”

$$Entropy\ (S) = \sum_{i=1}^{n} -pi * log_2 pi$$

$$Entropy\ (S) = -\frac{5}{9} * log_2 \frac{5}{9} + -\frac{4}{9} * log_2 \frac{4}{9}$$

$$Entropy\ (S) = (-0{,}5) * (-0{,}848) + (-0{,}4) * (-1{,}169)$$

$$Entropy\ (S) = 0{,}891$$

b. Perhitungan nilai gain untuk "gender"

$$Gain\ (S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i)$$

$$Gain\ (S, A) = 0{,}891 - \left(\frac{4}{9} * 1\right) + \left(\frac{5}{9} * 0{,}969\right)$$

$$Gain\ (S, A) = 0{,}891 - (0{,}4 + 0{,}484)$$

$$Gain\ (S, A) = 0{,}007$$

3. Mengurutkan skor untuk setiap atribut

Fitur	Skor
Gender	0,007
Work_type	0,091
Resident_type	0,016
Smoking_status	0,207

3.1.3 Praproses Data

Pada tahap praproses data atau *Preprocessing* data akan dilakukan proses normalisasi data seperti penyeleksian atribut, pembersihan data (*data cleaning*) dan transformasi data (*data transformation*).

1. Pembersihan Data (*Data Cleaning*)

Pada tahap ini data-data yang tidak sesuai akan dihapus atau dibersihkan, contohnya data noise atau missing value (Tabel III-1). Pembersihan data pada penelitian ini dilakukan dengan menggunakan nilai Mean.

Tabel III-1 Data Null atau Kosong

Nama	Jenis Kelamin	Usia	Alamat
Ana	Wanita	21	Kalianda
Bima	Pria	**NaN**	Palas
Angga	Pria	25	Sukaraja
Reni	Wanita	**NaN**	Pringsewu
Dini	Wanita	22	Pesawaran

2. **Transformasi Data**

   Transformasi data merupakan tahap dilakukannya perubahan data atau konversi data, seperti perubahan dari tipe nominal ke tipe kategori ataupun sebaliknya dengan tujuan agar terhindar dari data yang rusak atau tidak valid. Pada penelitian ini teknik transformasi dilakukan dengan menggunakan label encoding yaitu teknik mengubah setiap nilai yang ada dalam kolom menjadi angka yang berurutan (Gambar 3.1-7).

Tabel III-2 Teknik Label Encoding

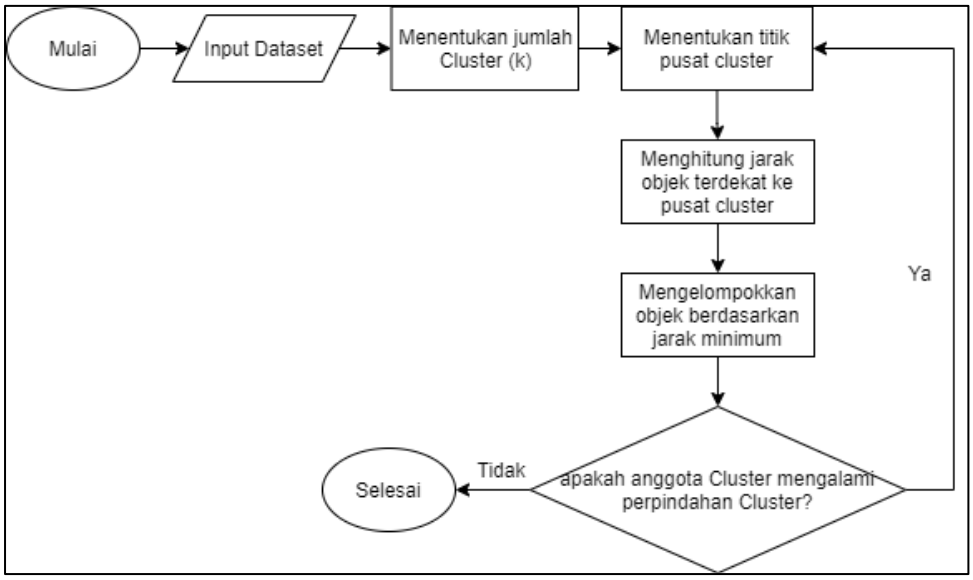
Nama	**Jenis Kelamin**		**Nama**	**Jenis Kelamin**
Ana	Wanita		Ana	1
Bima	Pria		Bima	0
Angga	Pria		Angga	0
Reni	Wanita		Reni	1
Dini	Wanita		Dini	1

3.1.4 *Clustering* atau Pengelompokkan Data

Pengelompokan data dilakukan setelah data melewati tahap praproses data dan pemilihan fitur. Pengelompokkan data akan dilakukan dengan menggunakan Kmeans *clustering* dan PSO Kmeans *clustering*.

1. **Algoritma Kmeans *Clustering***

   Alur proses atau flowchart pengelompokkan data pada Kmeans *Clustering* terdapat pada Gambar III-11.

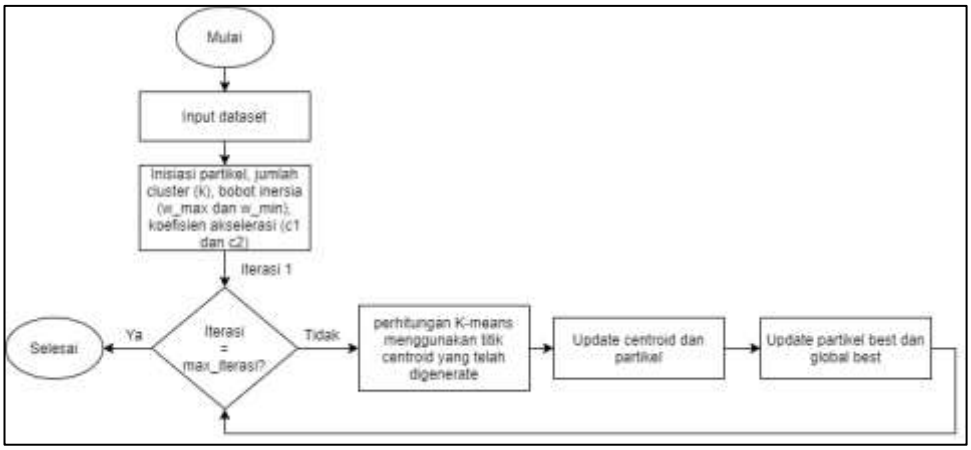


Gambar III-3 Flowchart Kmeans *Clustering*

Proses pengelompokkan data diawali dengan menginput dataset yang akan dikelompokkan, kemudian menentukan jumlah *cluster*. Pada penelitian ini jumlah *cluster* yang akan digunakan adalah k=3. Kemudian menentukan titik pusat *cluster* pada setiap *cluster*, menghitung jarak setiap data ke masing-masing pusat *cluster* menggunakan persamaan *euclidean distance*, mengelompokkan data ke dalam *cluster* berdasarkan jarak terdekat atau jarak minimum, selanjutnya yaitu menghitung rata-rata dari setiap pusat *cluster* untuk mendapatkan pusat *cluster* yang baru. Setelah itu membandingkan pusat *cluster* yang baru dengan pusat *cluster* yang lama, jika posisi anggota tiap *cluster* berubah maka ulangi menghitung jarak data ke masing-masing pusat *cluster* sampai posisi anggota setiap *cluster* tidak berubah.

2. Particle Swarm Optimization (PSO) dan Kmeans

Alur proses atau flowchart pengelompokkan data pada PSO Kmeans *Clustering* dapat terdapat pada Gambar III-12.



Gambar III-4 Flowchart PSO Kmeans *Clustering*

Pengelompokkan data pada PSO Kmeans *clustering* diawali dengan menginput dataset yang akan dikelompokkan, kemudian menginisialisasi partikel dan kecepatan partikel. Selanjutnya menentukan *partikel best* (Pbest) dan *global best* (Gbest) awal. Selanjutnya melakukan perhitungan Kmeans *clustering* menggunakan titik centroid yang telah degenerate, menentukan *cluster* setiap objek dengan Kmeans, mengupdate nilai titik *centroid* menggunakan persamaan update kecepatan pada persamaan (2.6) dan update partikel dengan persamaan (2.5). Update *particle best* (Pbest) dan *global best* (Gbest) dengan persamaan (2.7). Langkah terakhir yaitu mengulang iterasi sampai mencapai *stopping condition* atau iterasi maksimum.

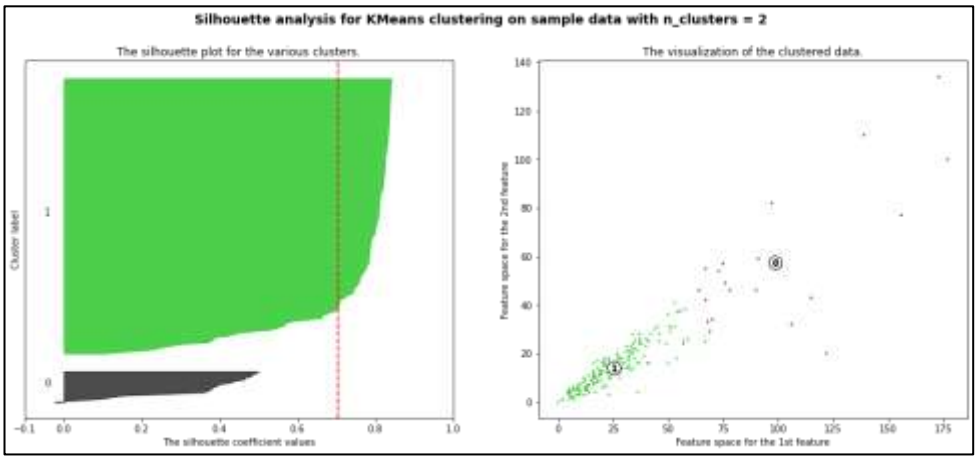
3.1.5 Evaluasi

Tahapan ini merupakan tahap untuk melakukan pengujian terhadap hasil dari tahapan sebelumnya.

1. **Pengujian Jumlah *Cluster***

   Pada tahap ini akan dilakukan pengujian jumlah *cluster* untuk mengetahui jumlah *cluster* yang memiliki tingkat validasi tertinggi. Pengujian jumlah *cluster* dilakukan pada beberapa dataset. Kualitas hasil *clustering* dievaluasi menggunakan *Silhouette Coefficient*. *Silhouette coefficient* juga digunakan untuk membandingkan efektivitas dari algoritma Kmeans dan PSO Kmeans [30]. Skor *Sillhouette* akan semakin baik saat mendekati nilai +1, dan akan

semakin buruk saat skor mendekati nilai -1 [27]. Jumlah *cluster* yang digunakan dalam penelitian adalah k=3.



Gambar III-5 *Shillhouette Plot*