BAB 3

METODE PENELITIAN

3.1 Data Penelitian

Penelitian ini menggunakan *dataset* yang digunakan pada penelitian lain dan bersifat data kategori. Perolehan data dilakukan secara *online* sehingga bisa dilakukan dimana saja, data diambil pada bulan Januari 2017 selama satu minggu yang diakses dari https://archive.ics.uci.edu/ml/datasets.html. Alasan digunakannya *dataset* tersebut adalah karena *dataset* tersebut merupakan *dataset* yang sudah teruji kebenarannya dan sering digunakan pada beberapa penelitian di wilayah *data mining*. *Dataset* yang disediakan pada UCI ini merupakan *dataset* yang diambil berdasarkan kasus pada dunia nyata. Selain itu *dataset* tersebut juga digunakan pada penelitian-penelitian sebelumnya yang berkaitan dengan *clustering*. Penelitian ini menggunakan 4 *dataset* yang berbeda dalam hal jumlah atribut untuk membandingkan kinerja algoritma pada *dataset* yang memiliki data kategori. Tabel 3.1 merupakan karakteritik dari *dataset* tersebut.

Tabel 3.1 Karakteristik *Dataset* Bagian 1

Dataset	**Jumlah Data**	**Jumlah atribut**		**Kelas**	**Keterangan**
		Sebelum *preprocessing*	**Setelah *Preprocessing***		
Soybean Small	47	35	21	4	Pengelompokkan jenis kedelai berdasarkan daun, batang, dll.
Zoo	101	16	16	7	Pengelompokan hewan di kebun binatang berdasarkan ciri-ciri di tubuhnya seperti ekor, kaki, bertelur/tidak, dll
Balance Scale	625	4	4	3	Digunakan untuk memodelkan hasil eksperimen psikologis. Setiap contoh memiliki skala keseimbangan ke kanan, ujung ke kiri, atau seimbang.

Tabel 3.2 Karakteristik *Dataset* Bagian 2

Dataset	**Jumlah Data**	**Jumlah atribut**		**Kelas**	**Keterangan**
		Sebelum *preprocessing*	**Setelah *Preprocessing***		
Car Evaluation	1728	6	6	4	Berisi Informasi tingkat pembelian, perawatan, jumlah pintu, dan jumlah orang, keamanan pada mobil. Bermanfaat untuk metode penemuan struktur mobil

Data yang digunakan adalah data yang sebelumnya telah memiliki label kelas secara manual. Label kelas yang sudah ada sebelumnya digunakan untuk proses evaluasi hasil *clustering*, yaitu saat menghitung nilai akurasi cluster yang dihasilkan dari implementasi algoritma WDM-EA K-Modes. Untuk proses pengujian digunakan data kategori yang telah mengalami proses *preprocessing* yaitu berupa penyeleksian terhadap atribut yang bernilai numerik, karena algoritma ini hanya mampu menangani data yang bertipe kategori. Selain itu dilakukan penghapusan atribut yang hanya memiliki satu kategori, tujuannya adalah agar *cluster* yang dihasilkan lebih baik. Jika dalam satu atribut hanya memiliki satu kategori, menyebabkan kemungkinan data tidak beragam, hal ini kemungkinan data hanya mengarah pada satu *cluster* saja. Proses *preprocessing* dilakukan agar struktur data menjadi lebih baik sehingga hasil *clustering* lebih maksimal.

Pada data *Soybean Small* dilakukan penghapusan sebanyak 13 atribut karena memiliki 1 *value* saja. Sedangkan dataset *Zoo*, *Balance Scale*, dan *Car Evaluation* tidak dilakukan penyeleksian terhadap atribut (*preprocessing*) karena semua atributnya telah bernilai kategori dan masing-masing atributnya memiliki nilai kategori yang beragam. Contoh data setiap *dataset* dapat dilihat di Tabel 3.3 sampai Tabel 3.6.

Tabel 3.3 Contoh *Dataset Soybean Small*

No	*Date*	*Plant-stand*	*precip*	*temp*	*Hail*	*Leaves*	*Stem-cankers*	*Int-discolor*	…	*Roots*
1	*May*	*normal*	*Lt-norm*	*Lt-norm*	*Yes*	*Norm*	*Absent*	*none*	...	*Rotted*
2	*April*	*Normal*	*Lt-norm*	*Norm*	*No*	*Norm*	*Absent*	*Brown*	...	*Norm*
3	*April*	*Normal*	*Lt-norm*	*Norm*	*No*	*Abnorm*	*Absent*	*Brown*	...	*Galls-cysts*
4	*June*	*Normal*	*Lt-norm*	*Norm*	*No*	*Norm*	*Below-soil*	*Brown*	...	*Rotted*
5	*July*	*normal*	*Norm*	*Gt-norm*	*No*	*Abnorm*	*Absent*	*Brown*	...	*Norm*
6	*August*	*Normal*	*Norm*	*Gt-norm*	*No*	*Norm*	*Below-soil*	*Black*	...	*Galls-cysts*
7	*August*	*Normal*	*Norm*	*Gt-norm*	*No*	*Abnorm*	*Absent*	*black*	...	*Rotted*
8	*September*	*Normal*	*Gt-norm*	*Gt-norm*	*No*	*Norm*	*Absent*	*Brown*	...	*Rotted*
9	*September*	*Normal*	*Gt-norm*	*Norm*	*No*	*Abnorm*	*Absent*	*Brown*	...	*Rotted*
10	*October*	*Lt-normal*	*Gt-norm*	*Norm*	*No*	*Norm*	*Absent*	*Brown*	...	*Rotted*
11	*April*	*Lt-normal*	*Gt-norm*	*Norm*	*No*	*Abnorm*	*Above-soil*	*Black*	...	*Rotted*
12	*May*	*Lt-normal*	*Norm*	*Norm*	*No*	*Norm*	*Above-soil*	*black*	...	*Rotted*
13	*June*	*Normal*	*Norm*	*Norm*	*No*	*Norm*	*Absent*	*Brown*	...	*Norm*
14	*July*	*Normal*	*Norm*	*Norm*	*No*	*Abnorm*	*Absent*	*Brown*	...	*Galls-cysts*
15	*July*	*Normal*	*Norm*	*Norm*	*No*	*Norm*	*Absent*	*Brown*	...	*Rotted*
16	*July*	*Lt-normal*	*Norm*	*Gt-norm*	*Yes*	*Norm*	*Above-soil*	*Black*	...	*Rotted*
17	*July*	*Lt-normal*	*Norm*	*Norm*	*Yes*	*Abnorm*	*Above-soil*	*black*	...	*Norm*
18	*May*	*Lt-normal*	*Norm*	*Norm*	*Yes*	*Norm*	*Absent*	*Brown*	...	*Galls-cysts*
19	*May*	*Lt-normal*	*Gt-norm*	*Norm*	*Yes*	*Norm*	*Absent*	*Brown*	...	*Rotted*
20	*May*	*Lt-normal*	*Norm*	*Norm*	*Yes*	*Abnorm*	*Absent*	*Brown*	...	*Rotted*
21	*April*	*Lt-normal*	*Norm*	*Norm*	*Yes*	*Norm*	*Above-soil*	*Black*	...	*Norm*
22	*April*	*Lt-normal*	*Norm*	*Norm*	*No*	*Norm*	*Above-soil*	*black*	...	*Galls-cysts*
…	...	...	...	...	...	...	...	...	...	...
47	*April*	*Lt-normal*	*Norm*	*Norm*	*Yes*	*Norm*	*Absent*	*None*	...	*Galls-cysts*
48	*April*	*Lt-normal*	*Norm*	*Norm*	*yes*	*Norm*	*Absent*	*none*	...	*Rotted*

Tabel 3.4 Contoh *Dataset Zoo*

No	Hair	Feathers	Eggs	Legs	…	Tail
1	Yes	Yes	No	0	…	No
2	No	No	Yes	4	…	Yes
3	Yes	Yes	No	6	…	No
4	Yes	No	Yes	5	…	Yes
5	Yes	No	Yes	2	…	Yes
6	Yes	No	Yes	0	…	Yes
7	No	No	Yes	0	…	Yes
8	Yes	No	No	0	…	No
9	No	No	Yes	0	…	Yes
10	No	No	No	2	…	No
11	No	No	No	2	…	No
12	No	No	Yes	2	…	Yes
13	No	Yes	No	4	…	No
14	No	No	Yes	4	…	Yes
15	No	Yes	Yes	4	…	Yes
16	No	Yes	Yes	4	…	Yes
17	No	Yes	Yes	0	…	No
18	Yes	Yes	No	0	…	No
19	Yes	No	Yes	0	…	No
20	No	Yes	No	0	…	No
21	Yes	No	No	2	…	No
22	Yes	No	Yes	2	…	No
23	Yes	Yes	No	2	…	Yes
…	…	…	…	…	…	…
1728	Yes	No	No	2	…	Yes

Tabel 3.5 Contoh *Dataset Balance Scale*

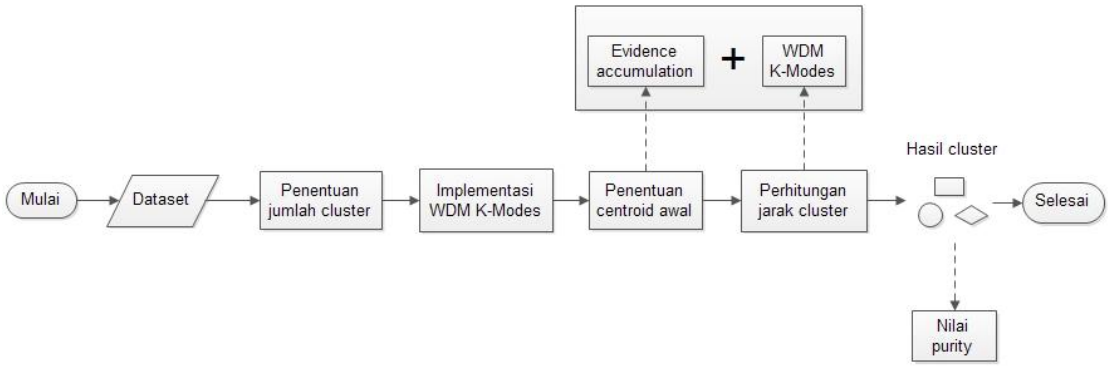
No	***Left-weight***	***Left-distance***	***Right-weight***	***Right-distance***
1	1	1	2	5
2	1	1	2	5
3	1	1	2	5
4	1	1	2	5
5	1	1	1	5
6	1	1	1	5
7	2	1	1	5
8	2	1	1	4
9	2	1	1	4
10	2	1	3	4
11	2	2	3	4
12	2	2	3	3
13	2	2	3	3
14	2	2	5	3
15	3	2	5	3
16	3	2	5	2
17	3	2	5	2
18	3	3	4	2
19	4	3	4	2
20	4	3	4	1
21	4	4	4	1
22	4	4	4	1
23	4	5	4	1
…	…	…	…	…
625	5	5	4	1

Tabel 3.6 Contoh *Dataset Car Evaluation*

No	*buying*	*maintenance*	*Doors*	*Person*	*Lug-boot*	*Safety*
1	*vhigh*	*vhigh*	*Two*	*two*	*small*	*High*
2	*vhigh*	*vhigh*	*Three*	*four*	*big*	*Low*
3	*med*	*Low*	*Three*	*more*	*med*	*Med*
4	*low*	*Low*	*Four*	*four*	*med*	*High*
5	*low*	*high*	*Fivemore*	*two*	*big*	*Med*
6	*med*	*med*	*Fivemore*	*more*	*small*	*Low*
7	*vhigh*	*high*	*Fivemore*	*two*	*small*	*High*
8	*med*	*Low*	*Three*	*more*	*med*	*Med*
9	*low*	*Low*	*Four*	*four*	*med*	*High*
10	*low*	*high*	*Fivemore*	*two*	*big*	*Med*
11	*vhigh*	*vhigh*	*Three*	*four*	*big*	*Low*
12	*med*	*Low*	*Three*	*more*	*med*	*Med*
13	*low*	*Low*	*Four*	*four*	*med*	*High*
14	*low*	*high*	*Fivemore*	*two*	*big*	*Med*
15	*med*	*med*	*Fivemore*	*more*	*small*	*Low*
16	*low*	*Low*	*Four*	*four*	*med*	*High*
17	*low*	*high*	*Fivemore*	*two*	*big*	*Med*
18	*vhigh*	*vhigh*	*Three*	*four*	*big*	*Low*
19	*med*	*Low*	*Three*	*more*	*med*	*Med*
20	*low*	*Low*	*Four*	*four*	*med*	*High*
21	*med*	*Low*	*Three*	*more*	*med*	*Med*
22	*vhigh*	*vhigh*	*Four*	*more*	*big*	*Low*
23	*vhigh*	*vhigh*	*Four*	*four*	*big*	*Med*
…	...	...	...	...	...	...
1728	*vhigh*	*vhigh*	*Four*	*more*	*med*	*High*

3.2 *Weighted Dissimilarity Measure – Evidence Accumulation K-Modes*

Algoritma WDM-EA K-Modes dilakukan untuk melengkapi kekurangan masing-masing algoritma terdahulu. Dengan diterapkannya WDM-EA K-Modes sebagai perbaikan perhitungan jarak *cluster* dan *centroid* awal, maka dapat memperbaiki hasil *purity* terbaik dari suatu *dataset*. Tahap-tahap penggabungan algoritma WDM-EA K-Modes dapat dilihat pada Gambar 3.1



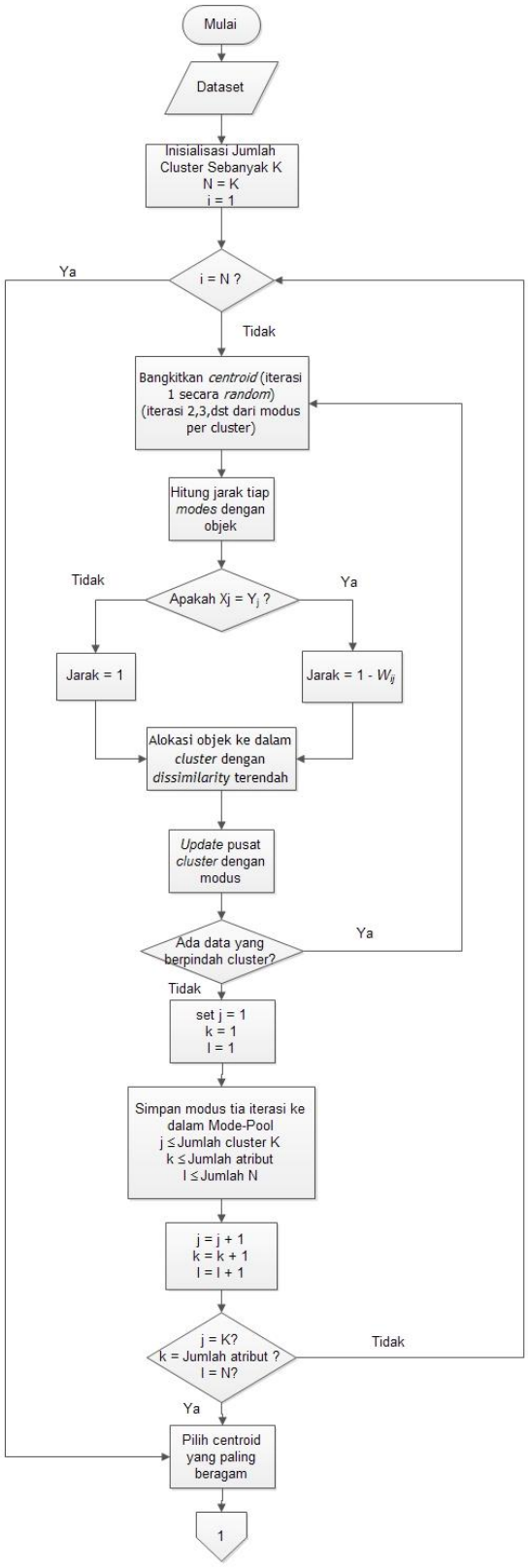
Gambar 3.1 Penggabungan Algoritma WDM K-Modes dan *Evidence Accumulation* K-Modes

Berdasarkan Gambar 3.1, pembentukan hasil *cluster* pada gabungan algoritma WDM K-Modes dan *Evidence Accumulation* K-Modes dapat dijelaskan sebagai berikut:

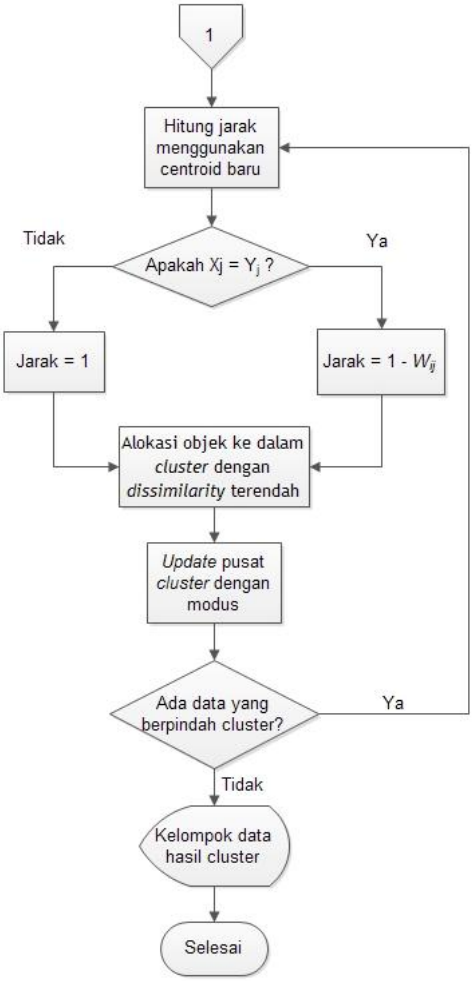
1) *Dataset*, langkah pertama yaitu mengolah data dari *dataset*, berupa matriks data yang dibutuhkan untuk penelitian ini.
2) Penentuan jumlah *cluster*, mengalokasikan *modes* sejumlah *k* (secara *random*)
3) Implementasi WDM K-Modes untuk memperoleh modus dari tiap iterasi.

4) *Evidence Accumulation* bekerja menyimpan modus di setiap iterasi K-Modes ke dalam *Mode-pool* dan memilih modus yang paling beragam yang dapat dimanfaatkan sebagai *centroid* awal untuk Algoritma K-Modes sehingga algoritma *Evidence Accumulation* mampu menentukan *centroid* awal
5) Perhitungan jarak *cluster*, menghitung jarak masing-masing data terhadap semua titik pusat *cluster* menggunakan ukuran jarak *dissimilarity* WDM
6) Hasil *cluster*, dari perhitungan jarak *cluster* akan menghasilkan kelompok-kelompok data yang memiliki kesamaan ciri-ciri atau karakteristik *intra cluster.*
7) Nilai *purity*, *cluster* yang dikatakan *pure* (murni) yaitu semua objek dengan *class* yang sama berada pada *cluster* yang sama. Akurasi terhadap hasil *clustering* menggunakan metode *Purity Measure*.

Algoritma WDM-EA K-Modes diperlihatkan pada Gambar 3.2 dan Gambar 3.3



Gambar 3.2 *Flowchart* WDM-EA K-Modes Bagian 1



Gambar 3.3 *Flowchart* WDM-EA K-Modes Bagian 2

3.3 Teknik Pengujian dan Analisis Data

3.3.1 Teknik Pengujian

Pengujian dilakukan dengan cara mengimplementasikan dan membandingkan kinerja algoritma *clustering* yang diusulkan yaitu gabungan algoritma *Weighted Dissimilarity Measure* dengan *Evidence Accumulation* (WDM-EA). Pengujian data ini dilakukan pada dataset yang telah ditentukan sesuai dengan kebutuhan algoritma.

Pengujian *dataset* dilakukan dengan cara mengimplementasikan algoritma *clustering* kemudian melakukan evaluasi hasil *clustering* dengan *Purity Measure*. Hal ini dilakukan untuk mengetahui bagaimana kinerja algoritma kemudian membandingkan K-Modes tradisional, WDM K-Modes dan WDM-EA K-Modes, *dataset* mana yang akan menghasilkan nilai *purity* yang lebih tinggi. Tahapan pengujian *dataset* adalah sebagai berikut:

A. Melakukan implementasi algoritma *clustering* dan evaluasi hasil *clustering*

Implementasi algoritma dan evaluasi hasil *clustering* dilakukan berdasarkan jumlah *cluster* pada setiap *dataset*. Evaluasi hasil *clustering* dilakukan dengan cara menghitung nilai *purity* dari setiap hasil *clustering* pada dataset untuk setiap jumlah *cluster*. Untuk menentukan jumlah *cluster*, dimulai dari jumlah kelas yang sebenarnya pada *dataset*, dan seterusnya meningkat sampai jumlah *cluster* bertambah 3 dari jumlah kelas sebenarnya. Ini digunakan untuk membandingkan nilai *purity* yang diperoleh jika jumlah *cluster*nya berbeda karena tidak ada metode khusus pada penelitian ini untuk menentukan jumlah *cluster* yang terbaik.

Sedangkan untuk mengetahui kepresisian algoritma, maka untuk setiap *cluster* dilakukan 10 kali percobaan. Sehingga informasi *purity* yang akan dihasilkan dari setiap percobaaan adalah seperti yang ditunjukkan pada Tabel 3.7.

Tabel 3.7 Hasil *Purity* Percobaan *Dataset*

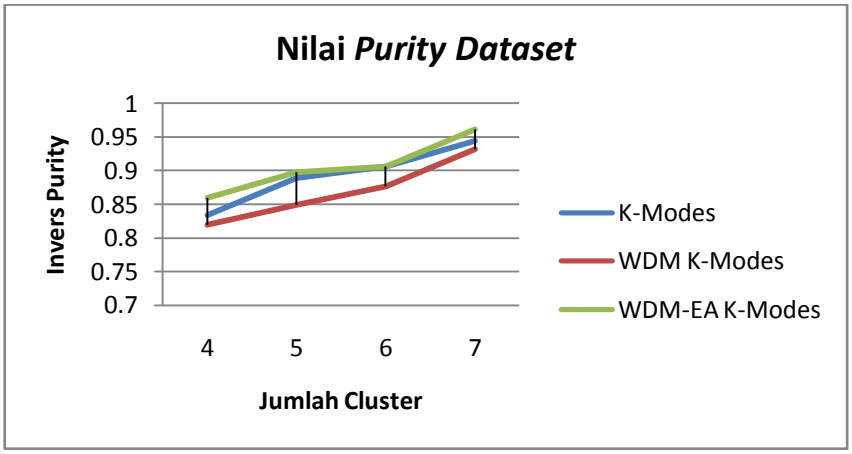
Percobaan Ke-	Jumlah *Cluster* (K = 4)			Jumlah *Cluster* (K = 5)			Jumlah *Cluster* (K = 6)		
	K-Modes	WDM	WDM-EA	K-Modes	WDM	WDM-EA	K-Modes	WDM	WDM-EA
1	0.834	0.819	0.859	0.141	0.0745	0.091	0.141	0.0745	0.091
2	…	…	…	…	…	…	…	…	…
3	…	…	…	…	…	…	…	…	…
4	…	…	…	…	…	…	…	…	…
5	…	…	…	…	…	…	…	…	…
6	…	…	…	…	…	…	…	…	…
7	…	…	…	…	…	…	…	…	…
8	…	…	…	…	…	…	…	…	…
9	…	…	…	…	…	…	…	…	…
10	…	…	…	…	…	…	…	…	…
Rata-Rata	…	…	…	…	…	…	…	…	…
Std	…	…	…	…	…	…	…	…	…

Setelah dilakukan 10 kali percobaan, pada Tabel 3.7 dihitung rata-rata *purity* dan standar deviasi setiap jumlah *cluster*. Rata-rata *purity* (mendekati angka 1) dan standar deviasi presisi (mendekati angka 0) menyatakan bahwa algoritma tersebut baik. Sehingga informasi yang akan dihasilkan dari hasil purity dan standar deviasi terbaik adalah seperti yang ditunjukkan pada Tabel 3.8.

Tabel 3.8 Hasil rata-rata *Purity Dataset*

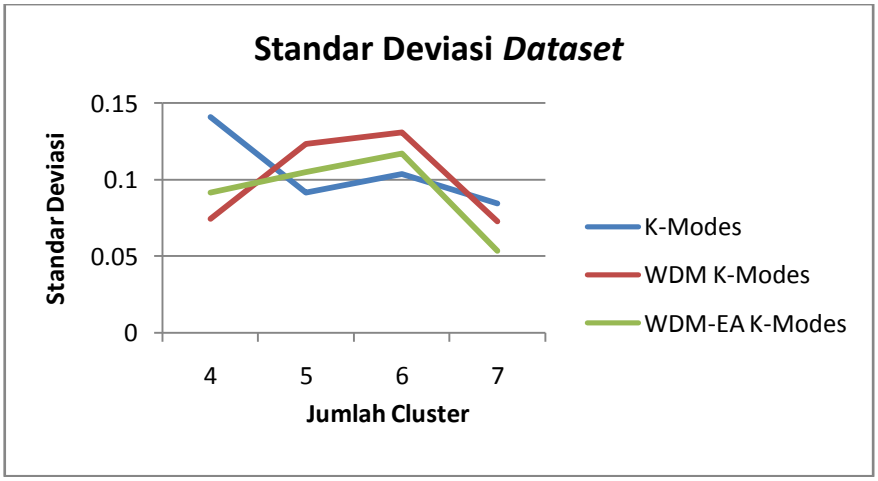
Jumlah *Cluster*	*Purity*			Standar Deviasi		
	K-Modes	WDM	WDM-EA	K-Modes	WDM	WDM-EA
4	0.834	0.819	0.859	0.141	0.074	0.091
5	…	…	…	…	…	…
6	…	…	…	…	…	…

Sedangkan grafik kinerja terbagi menjadi 2 yaitu grafik kinerja algoritma dan grafik standar deviasi kinerja. Pada grafik kinerja algoritma dan grafik standar deviasi, data yang digunakan adalah pada hasil evaluasi setelah dilakukan 10 kali percobaan pada setiap *dataset*. Contoh grafik kinerja algoritma dapat dilihat pada Gambar 3.4.



Gambar 3.4 Grafik Nilai *Purity Dataset*

Pada Gambar 3.4, sumbu x menunjukkan jumlah *cluster* dan sumbu y menunjukkan nilai *invers Purity*. Sementara grafik standar deviasi dapat dilihat pada Gambar 3.5 dimana sumbu x menunjukkan jumlah *cluster* dan sumbu y menunjukkan standar deviasi *Purity*.



Gambar 3.5 Grafik standar deviasi

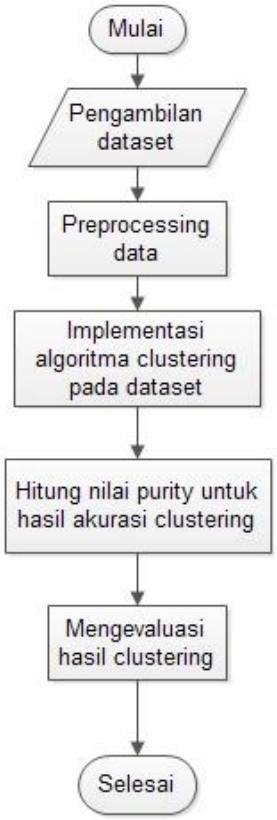
B. Melakukan analisis dengan cara membandingkan nilai kinerja

Setelah dilakukan percobaan pada semua *dataset*, selanjutnya akan membandingkan hasil *purity* yang telah diperoleh kemudian menyimpulkan

informasi nilai terbaik setiap dataset ke dalam sebuah tabel dan grafik. Semakin besar nilai *Purity* menunjukkan semakin baik hasil *clustering* tersebut, sebaliknya semakin kecil nilai *Purity* menunjukkan semakin buruk hasil *clustering* yang dihasilkan. Standar deviasi menunjukkan kepresisian algoritma. Nilai standar deviasi 0 menunjukkan bahwa algoritma tersebut presisi.

3.3.2 Analisis Data

Teknik analisis data penelitian dilakukan dengan langkah-langkah seperti pada Gambar 3.6.



Gambar 3.6 Teknik Analisis Data

1) Pengambilan *Dataset*

   Langkah awal dalam teknik analisis data adalah melakukan pengambilan data. Data yang diperoleh merupakan data yang bersifat kategori sesuai dengan kebutuhan karakteristik algoritma yang digunakan. Proses pengambilan data dilakukan secara *online* dengan mengakses situs https://archive.ics.uci.edu/ml/datasets.html.

2) *Preprocessing Data*

   Melakukan *preprocessing data* terhadap data mentah dari *Soybean Small, Zoo, Balance Scale* dan *Car Evaluation* dengan melakukan *cleaning data* yaitu membuang data yang tidak diperlukan saat proses pencarian pola dan mengubah data ke dalam format yang sesuai dengan kebutuhan.

3) Implementasi algoritma *clustering* pada *dataset*

   Dalam tahap ini akan dilakukan mengimplementasi algoritma menggunakan K-Modes Tradisional, WDM K-Modes dan WDM-EA K-Modes pada masing-masing *dataset* yaitu *Soybean Small, Zoo, Balance Scale* dan *Car Evaluation*.

4) Hitung nilai *Purity* untuk hasil akurasi *clustering*

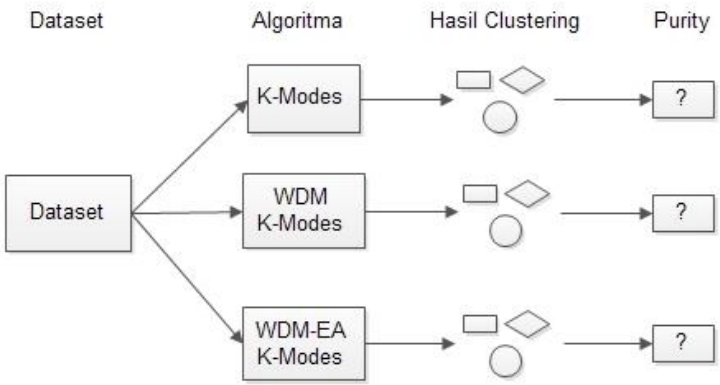
   Setelah terbentuk hasil *clustering* untuk masing-masing algoritma pada *dataset*, selanjutnya menghitung akurasi terhadap hasil *clustering* menggunakan metode *Purity Measure.*

5) Mengevaluasi hasil *clustering*

   Setelah didapatkan nilai *purity* hasil *clustering* untuk masing-masing algoritma, selanjutnya dilakukan perbandingan hasil untuk mengetahui

apakah nilai *purity* hasil *clustering* untuk algoritma WDM-EA K-Modes lebih tinggi dibandingkan dengan algoritma K-Modes tradisional dan WDM K-Modes. Jika nilai *purity* untuk algoritma WDM-EA K-Modes lebih besar (mendekati angka 1) maka menunjukkan bahwa *cluster* yang terbentuk merupakan hasil akurasi *clustering* yang semakin baik dibandingkan dengan algoritma terdahulu.

Sedangkan alur penelitian dilakukan dengan langkah-langkah seperti pada Gambar 3.7.



Gambar 3.7 Alur Penelitian

1) *Dataset*

   Data yang diperoleh merupakan data yang bersifat kategori sesuai dengan kebutuhan karakteristik algoritma yang digunakan. Proses pengambilan data dilakukan secara *online* dengan mengakses situs https://archive.ics.uci.edu/ml/datasets.html.

2) Implementasi algoritma *clustering* pada *dataset*

   Dataset yang telah di *preprocessing* akan dilakukan implementasi algoritma menggunakan K-Modes Tradisional, WDM K-Modes dan WDM-EA K-

Modes pada masing-masing *dataset* yaitu *Soybean Small, Zoo, Balance Scale* dan *Car Evaluation*.

3) Hitung nilai *Purity* untuk hasil akurasi *clustering*

Setelah terbentuk hasil *clustering* untuk masing-masing algoritma pada *dataset*, selanjutnya menghitung akurasi terhadap hasil *clustering* menggunakan metode *Purity Measure.*