BAB 4

HASIL PENGUJIAN DAN PEMBAHASAN

4.1 Hasil Pengujian

Pengujian *dataset* dilakukan dengan menggunakan *Purity Measure*. Pengujian dilakukan dengan mengubah-ubah jumlah *cluster* (K) terhadap nilai *purity* dari *cluster* yang terbentuk. Pemilihan jumlah *cluster* pada penelitian ini dimulai dari jumlah *cluster* sebenarnya sampai dengan penambahan 2 atau 3 cluster pada setiap *dataset*. 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.

Jika nilai *purity* semakin mendekati angka 1 maka menunjukkan bahwa *cluster* yang terbentuk merupakan hasil akurasi *clustering* yang semakin baik. Dan sebaliknya, semakin mendekati angka 0 maka hasil akurasi *clustering* semakin buruk. Untuk mengetahui kepresisian algoritma, maka untuk setiap *cluster* dilakukan 10 kali percobaan, lalu dihitung standar deviasi nya. Nilai standar deviasi 0 menunjukkan bahwa algoritma tersebut presisi.

4.1.1 Hasil pengujian pada dataset *soybean small*

Hasil pengujian *clustering Purity* untuk algoritma K-Modes, WDM K-Modes dan WDM-EA K-Modes pada *dataset soybean small* (47 data) sebanyak 10 kali percobaan disajikan di Tabel 4.1

Tabel 4.1 Hasil *Purity* percobaan *Dataset Soybean Small*

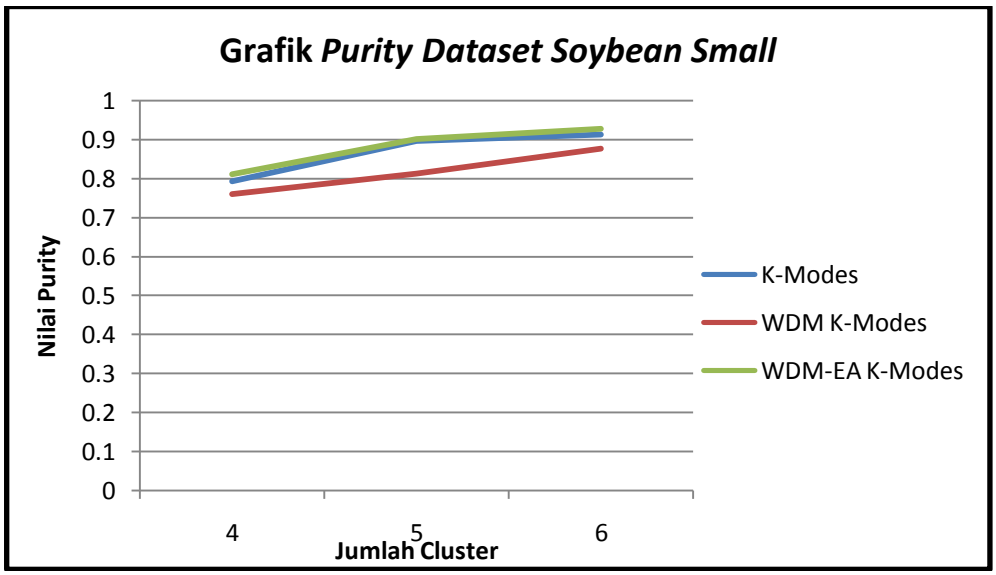
Percobaan Ke-	Jumlah *Cluster* (K = 4)			Jumlah *Cluster* (K = 5)			Jumlah *Cluster* (K = 6)		
	K-Modes	WDM	WDM-EA	K-Modes	WDM	WDM-EA	K-Modes	WDM	WDM-EA
1	0.787	0.766	0.830	1.000	0.766	0.957	0.979	0.979	0.979
2	1.00.	0.766	0.936	0.745	0.723	0.745	1.000	0.979	1.000
3	0.617	0.745	0.723	1.000	0.936	0.979	1.000	0.915	0.979
4	0.745	0.617	0.723	0.787	0.787	0.787	1.000	0.809	1.000
5	0.745	0.723	0.766	1.000	0.787	0.894	0.787	0.787	0.787
6	0.957	0.979	0.979	0.915	0.723	0.915	0.830	0.915	0.979
7	0.660	0.702	0.787	0.745	0.915	1.000	1.000	0.872	1.000
8	0.681	0.723	0.787	0.787	0.787	0.766	0.787	0.745	0.766
9	0.787	0.787	0.787	0.979	0.957	1.000	0.979	1.000	1.000
10	0.957	0.787	0.787	1.000	0.745	0.979	0.766	0.766	0.787
Rata-Rata	**0.794**	**0.760**	**0.811**	**0.896**	**0.813**	**0.902**	**0.913**	**0.877**	**0.928**
Std	**0.135**	**0.092**	**0.084**	**0.115**	**0.089**	**0.100**	**0.105**	**0.095**	**0.102**

Setelah dilakukan 10 kali percobaan, pada Tabel 4.1 dihitung rata-rata *purity* dan standar deviasi setiap jumlah *cluster*. Rata-rata *purity* tertinggi (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 4.2.

Tabel 4.2 Hasil rata-rata *Purity Dataset Soybean Small*

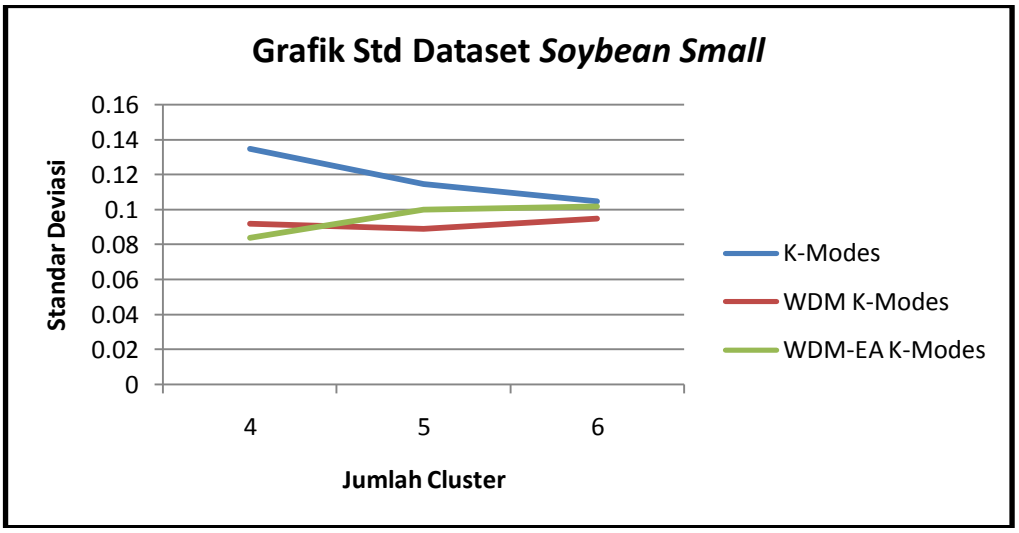
Jumlah *Cluster*	*Purity*			Standar Deviasi		
	K-Modes	WDM	WDM-EA	K-Modes	WDM	WDM-EA
4	0.794	0.760	0.811	0.135	0.092	0.084
5	0.896	0.813	0.902	0.115	0.089	0.100
6	**0.913**	**0.877**	**0.928**	**0.105**	**0.095**	**0.102**

Sedangkan grafik informasi yang akan dihasilkan dari hasil *purity* dan standar deviasi terbaik adalah seperti yang ditunjukkan Gambar 4.1 dan Gambar 4.2



Gambar 4.1 Grafik *Purity Dataset Soybean Small*

Pada Gambar 4.1, sumbu x menunjukkan jumlah *cluster* dan sumbu y menunjukkan nilai *invers Purity*. Sementara pada Gambar 4.2 merupakan grafik standar deviasi dimana sumbu x menunjukkan jumlah *cluster* dan sumbu y menunjukkan standar deviasi *Purity*.



Gambar 4.2 Grafik Standar Deviasi *Dataset Soybean Small*

Berdasarkan Tabel 4.2, Gambar 4.1 dan Gambar 4.2 terlihat bahwa nilai purity terbaik untuk *dataset soybean small* terdapat pada jumlah *cluster* 6, hal ini

dapat dilihat pada hasil standar deviasi pada jumlah *cluster* 6 juga nilainya dapat ditoleransi (mendekati 0). Nilai *purity* pada algoritma WDM-EA K-Modes lebih besar dibandingkan dengan *purity* pada algoritma WDM K-Modes ataupun K-Modes tradisional. Namun nilai *purity* pada algoritma WDM K-Modes lebih kecil dibandingkan *purity* pada algoritma K-Modes tradisional. Hal ini merupakan suatu kelemahan pada algoritma WDM K-Modes kurang optimal untuk data berjumlah kecil. Hal lain juga dapat disebabkan oleh pengambilan *centroid* awal yang kurang optimal dan jumlah data pada *dataset soybean small* sedikit.

4.1.2 Hasil pengujian pada data *zoo*

Hasil pengujian *clustering Purity* untuk algoritma K-Modes, WDM K-Modes dan WDM-EA K-Modes pada *dataset zoo* (101 data) sebanyak 10 kali percobaan disajikan di Tabel 4.3

Tabel 4.3 Hasil *Purity* percobaan *Dataset Zoo*

Percobaan Ke-	Jumlah *Cluster* (K = 7)			Jumlah *Cluster* (K = 8)			Jumlah *Cluster* (K = 9)		
	K-Modes	WDM	WDM-EA	K-Modes	WDM	WDM-EA	K-Modes	WDM	WDM-EA
1	0.802	0.723	0.812	0.733	0.733	0.802	0.871	0.842	0.842
2	0.911	0.842	0.891	0.891	0.881	0.950	0.851	0.861	0.861
3	0.871	0.812	0.861	0.881	0.891	0.871	0.812	0.822	0.832
4	0.851	0.812	0.851	0.921	0.723	0.723	0.901	0.901	0.911
5	0.792	0.812	0.812	0.812	0.812	0.871	0.832	0.802	0.802
6	0.713	0.683	0.713	0.733	0.792	0.832	0.812	0.832	0.832
7	0.812	0.802	0.832	0.832	0.802	0.802	0.861	0.871	0.871
8	0.812	0.792	0.832	0.842	0.832	0.832	0.832	0.842	0.861
9	0.832	0.871	0.861	0.812	0.822	0.832	0.802	0.832	0.871
10	0.792	0.832	0.851	0.861	0.812	0.871	0.950	0.901	0.901
Rata-Rata	**0.819**	**0.798**	**0.832**	**0.832**	**0.810**	**0.839**	**0.852**	**0.850**	**0.858**
Std	**0.053**	**0.056**	**0.048**	**0.063**	**0.054**	**0.060**	**0.046**	**0.033**	**0.033**

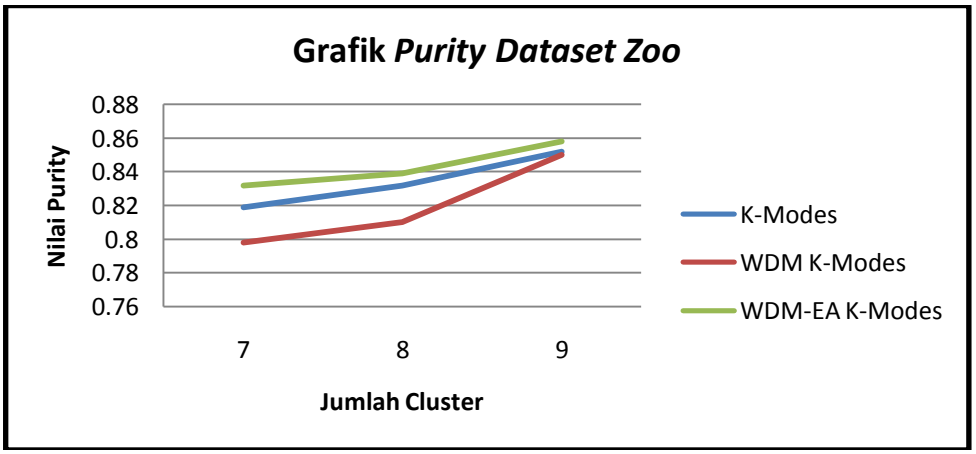
Setelah dilakukan 10 kali percobaan, pada Tabel 4.3 dihitung rata-rata *purity* dan standar deviasi setiap jumlah *cluster*. Rata-rata *purity* tertinggi

(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 4.4

Tabel 4.4 Hasil rata-rata *Purity Dataset Zoo*

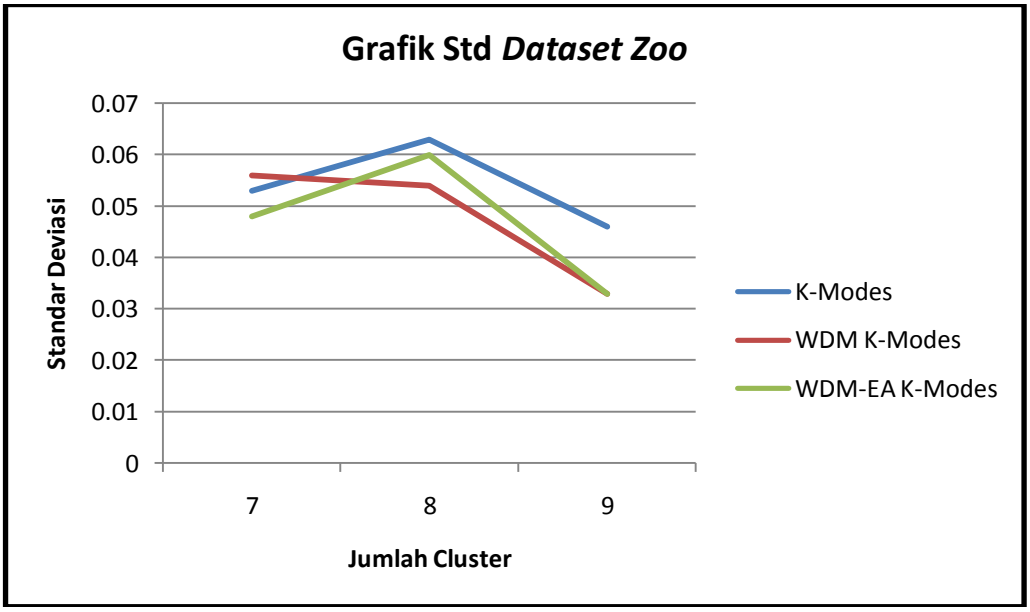
Jumlah *Cluster*	*Purity*			Standar Deviasi		
	K-Modes	WDM K-Modes	WDM-EA K-Modes	K-Modes	WDM K-Modes	WDM-EA K-Modes
7	0.819	0.798	0.832	0.053	0.056	0.048
8	0.832	0.810	0.839	0.063	0.054	0.060
9	**0.852**	**0.850**	**0.858**	**0.046**	**0.033**	**0.033**

Sedangkan grafik informasi yang akan dihasilkan dari hasil purity dan standar deviasi terbaik adalah seperti yang ditunjukkan Gambar 4.3 dan Gambar 4.4



Gambar 4.3 Grafik *Purity Dataset Zoo*

Pada Gambar 4.3, sumbu x menunjukkan jumlah *cluster* dan sumbu y menunjukkan nilai *invers Purity*. Sementara pada Gambar 4.4 merupakan grafik standar deviasi dimana sumbu x menunjukkan jumlah *cluster* dan sumbu y menunjukkan standar deviasi *Purity*.



Gambar 4.4 Grafik Standar Deviasi *Dataset Zoo*

Berdasarkan Tabel 4.4, Gambar 4.3 dan Gambar 4.4 terlihat bahwa nilai *purity* terbaik untuk *dataset zoo* terdapat pada jumlah cluster 9, hal ini dapat dilihat pada hasil standar deviasi pada jumlah *cluster* 9 juga nilainya dapat ditoleransi (mendekati 0). Nilai *purity* pada algoritma WDM-EA K-Modes lebih besar dibandingkan dengan *purity* pada algoritma WDM K-Modes ataupun K-Modes tradisional. Kemudian nilai *purity* pada algoritma WDM K-Modes juga lebih besar dibandingkan *purity* pada algoritma K-Modes tradisional. Pada *dataset zoo*, nilai *purity* sedikit meningkat dibandingkan *purity* pada *soybean small*.

4.1.3 Hasil pengujian pada data *balance scale*

Hasil pengujian *clustering Purity* untuk algoritma K-Modes, WDM K-Modes dan WDM-EA K-Modes pada dataset *balance scale* (625 data) sebanyak 10 kali percobaan disajikan di Tabel 4.5

Tabel 4.5 Hasil *Purity* percobaan *Dataset balance scale*

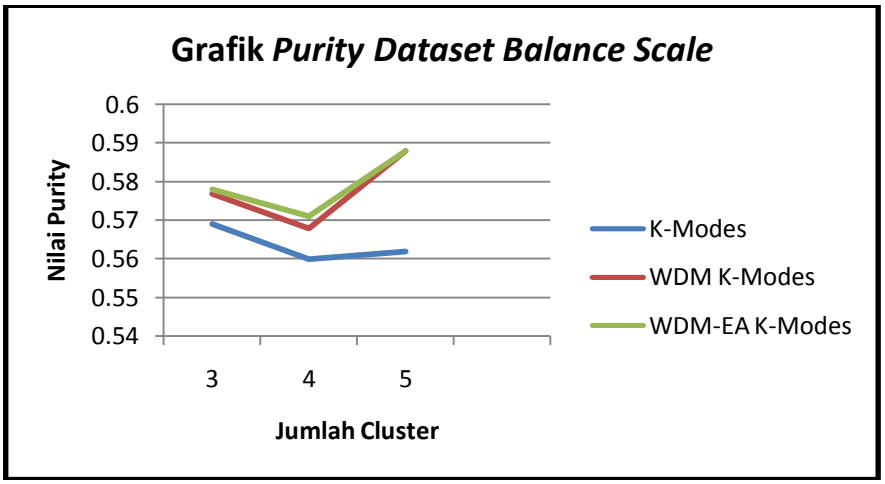
Percobaan Ke-	Jumlah *Cluster* (K = 4)			Jumlah *Cluster* (K = 5)			Jumlah *Cluster* (K = 6)		
	K-Modes	WDM	WDM-EA	K-Modes	WDM	WDM-EA	K-Modes	WDM	WDM-EA
1	0.578	0.509	0.509	0.648	0.635	0.635	0.616	0.610	0.610
2	0.558	0.554	0.554	0.510	0.531	0.531	0.571	0.648	0.648
3	0.579	0.592	0.592	0.506	0.530	0.530	0.571	0.544	0.544
4	0.613	0.624	0.624	0.517	0.536	0.536	0.509	0.518	0.518
5	0.474	0.501	0.501	0.618	0.570	0.570	0.554	0.637	0.637
6	0.558	0.552	0.562	0.578	0.600	0.600	0.491	0.534	0.534
7	0.594	0.635	0.635	0.614	0.571	0.571	0.587	0.646	0.646
8	0.515	0.525	0.525	0.576	0.576	0.576	0.570	0.595	0.595
9	0.587	0.621	0.621	0.504	0.552	0.552	0.552	0.558	0.558
10	0.635	0.661	0.661	0.534	0.584	0.610	0.600	0.594	0.594
Rata-Rata	**0.569**	**0.577**	**0.578**	**0.560**	**0.568**	**0.571**	**0.562**	**0.588**	**0.588**
Std	**0.047**	**0.057**	**0.056**	**0.053**	**0.033**	**0.036**	**0.038**	**0.048**	**0.048**

Setelah dilakukan 10 kali percobaan, pada Tabel 4.5 dihitung rata-rata *purity* dan standar deviasi setiap jumlah *cluster*. Rata-rata *purity* tertinggi (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 4.6

Tabel 4.6 Hasil rata-rata *Purity Dataset Balance Scale*

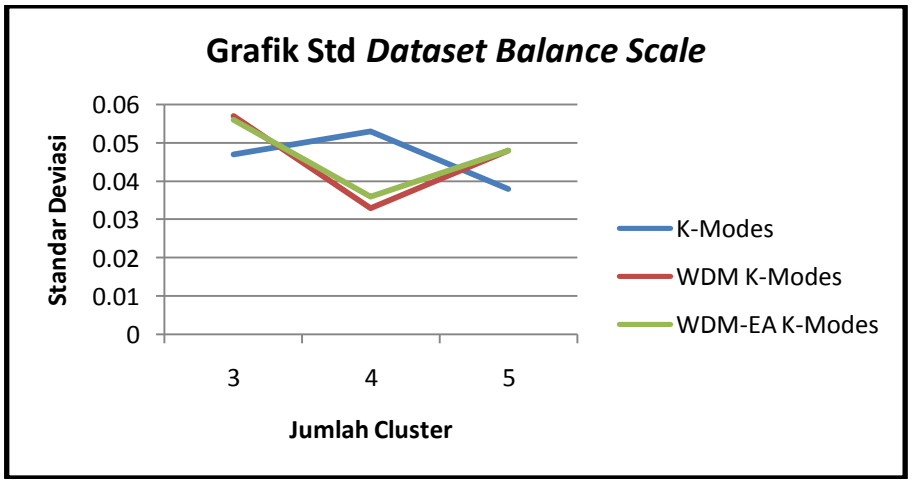
Jumlah *Cluster*	*Purity*			Standar Deviasi		
	K-Modes	WDM K-Modes	WDM-EA K-Modes	K-Modes	WDM K-Modes	WDM-EA K-Modes
3	0.569	0.577	0.578	0.047	0.057	0.056
4	0.560	0.568	0.571	0.053	0.033	0.036
5	**0.562**	**0.588**	**0.588**	**0.038**	**0.048**	**0.048**

Sedangkan grafik informasi yang akan dihasilkan dari hasil purity dan standar deviasi terbaik adalah seperti yang ditunjukkan Gambar 4.5 dan Gambar 4.6



Gambar 4.5 Grafik *Purity Dataset balance scale*

Pada Gambar 4.5, sumbu x menunjukkan jumlah *cluster* dan sumbu y menunjukkan nilai *invers Purity*. Sementara pada Gambar 4.6 merupakan grafik standar deviasi dimana sumbu x menunjukkan jumlah *cluster* dan sumbu y menunjukkan standar deviasi *Purity*.



Gambar 4.6 Grafik Standar Deviasi *Dataset Balance Scale*

Berdasarkan Tabel 4.6, Gambar 4.5 dan Gambar 4.6 terlihat bahwa nilai *purity* terbaik untuk *dataset balance scale* terdapat pada jumlah cluster 5, hal ini

dapat dilihat pada hasil standar deviasi pada jumlah *cluster* 5 juga nilainya dapat ditoleransi (mendekati 0). Nilai *purity* pada algoritma WDM-EA K-Modes lebih besar dibandingkan dengan *purity* pada algoritma WDM K-Modes ataupun K-Modes tradisional. Kemudian nilai *purity* pada algoritma WDM K-Modes juga lebih besar dibandingkan *purity* pada algoritma K-Modes tradisional.

4.1.4 Hasil pengujian pada data *car evaluation*

Hasil pengujian *clustering Purity* untuk algoritma K-Modes, WDM K-Modes dan WDM-EA K-Modes pada dataset *car evaluation* (1728 data) sebanyak 10 kali percobaan disajikan di Tabel 4.7

Tabel 4.7 Hasil *Purity* percobaan *Dataset car evaluation*

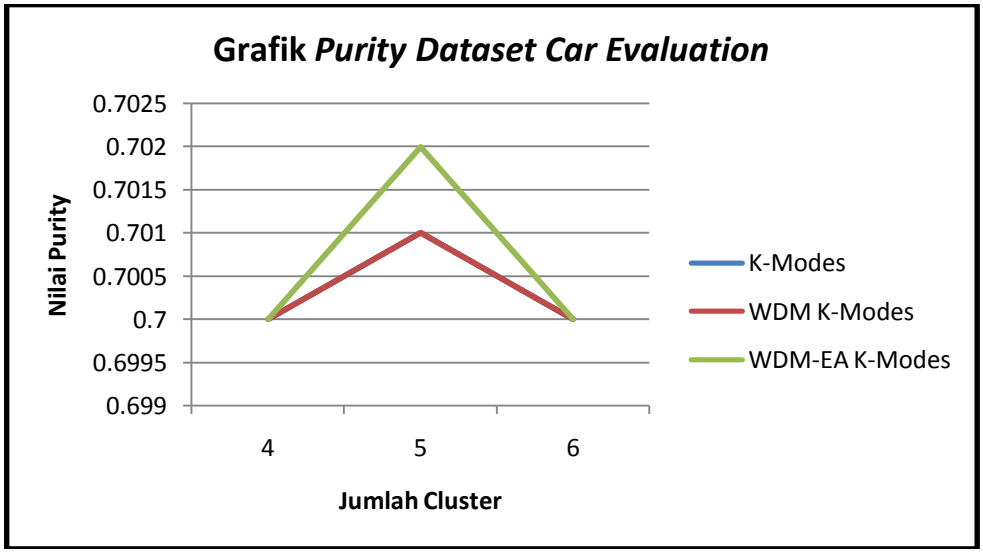
Percobaan Ke-	**Jumlah *Cluster* (K = 4)**			**Jumlah *Cluster* (K = 5)**			**Jumlah *Cluster* (K = 6)**		
	K-Modes	**WDM**	**WDM-EA**	**K-Modes**	**WDM**	**WDM-EA**	**K-Modes**	**WDM**	**WDM-EA**
1	0.70	0.70	0.70	0.700	0.700	0.700	0.70	0.70	0.70
2	0.70	0.70	0.70	0.700	0.704	0.704	0.70	0.70	0.70
3	0.70	0.70	0.70	0.700	0.700	0.700	0.70	0.70	0.70
4	0.70	0.70	0.70	0.700	0.704	0.704	0.70	0.70	0.70
5	0.70	0.70	0.70	0.700	0.700	0.700	0.70	0.70	0.70
6	0.70	0.70	0.70	0.700	0.700	0.700	0.70	0.70	0.70
7	0.70	0.70	0.70	0.700	0.704	0.704	0.70	0.70	0.70
8	0.70	0.70	0.70	0.700	0.702	0.704	0.70	0.70	0.70
9	0.70	0.70	0.70	0.705	0.700	0.700	0.70	0.70	0.70
10	0.70	0.70	0.70	0.707	0.700	0.700	0.70	0.70	0.70
Rata-Rata	**0.700**	**0.700**	**0.700**	**0.701**	**0.701**	**0.702**	**0.70**	**0.70**	**0.70**
Std	**0**	**0**	**0**	**0.003**	**0.002**	**0.002**	**0.00**	**0.00**	**0.00**

Setelah dilakukan 10 kali percobaan, pada Tabel 4.7 dihitung rata-rata *purity* dan standar deviasi setiap jumlah *cluster*. Rata-rata *purity* tertinggi (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 4.8.

Tabel 4.8 Hasil rata-rata *Purity Dataset car evaluation*

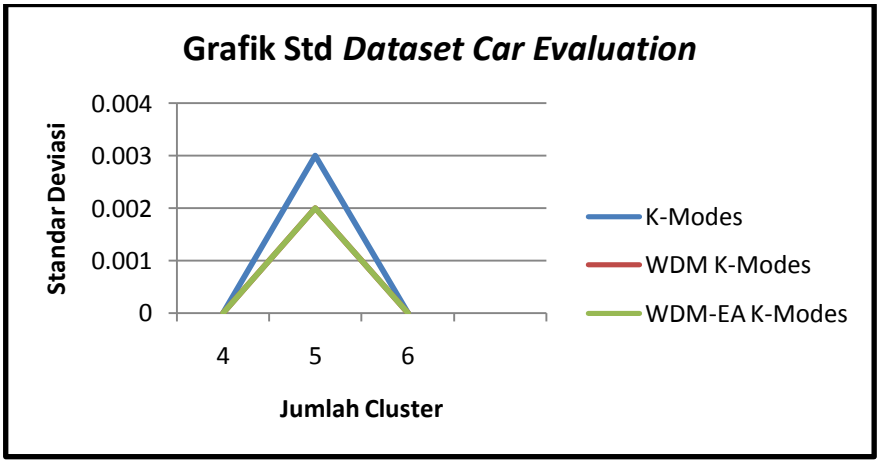
Jumlah *Cluster*	*Purity*			Standar Deviasi		
	K-Modes	WDM K-Modes	WDM-EA K-Modes	K-Modes	WDM K-Modes	WDM-EA K-Modes
4	0.700	0.700	0.700	0.000	0.000	0.000
5	**0.701**	**0.701**	**0.702**	**0.003**	**0.002**	**0.002**
6	0.700	0.700	0.700	0.000	0.000	0.000

Sedangkan grafik informasi yang akan dihasilkan dari hasil purity dan standar deviasi terbaik adalah seperti yang ditunjukkan Gambar 4.7 dan Gambar 4.8



Gambar 4.7 Grafik *Purity Dataset car evaluation*

Pada Gambar 4.7, sumbu x menunjukkan jumlah *cluster* dan sumbu y menunjukkan nilai *invers Purity*. Sementara pada Gambar 4.8 merupakan grafik standar deviasi dimana sumbu x menunjukkan jumlah *cluster* dan sumbu y menunjukkan standar deviasi *Purity*.



Gambar 4.8 Grafik Standar Deviasi *Dataset Car Evaluation*

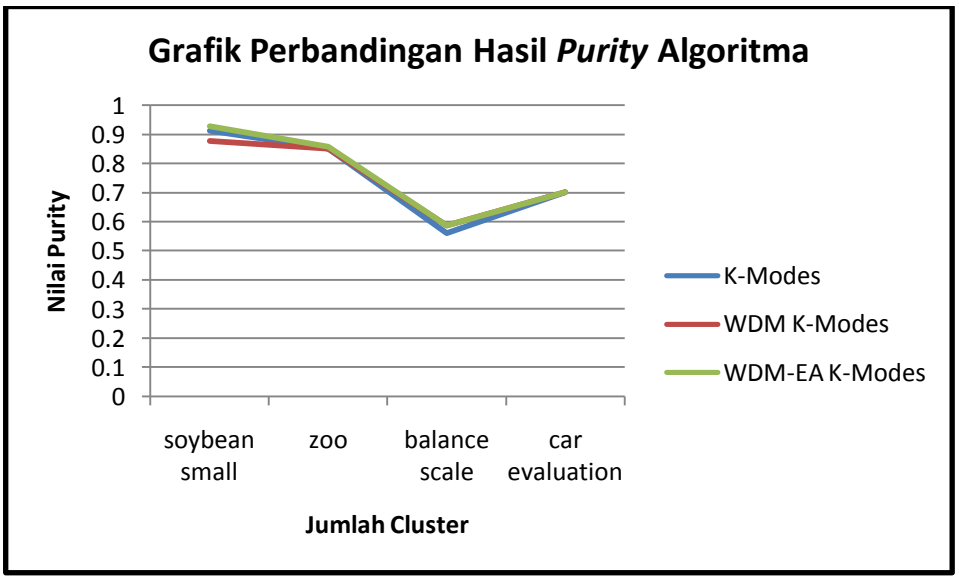
Berdasarkan Tabel 4.8, Gambar 4.7 dan Gambar 4.8 terlihat bahwa nilai purity terbaik untuk dataset *car evaluation* terdapat pada jumlah *cluster* 5, namun pada hasil standar deviasi pada jumlah cluster 7 nilai lebih besar dibandingkan pada jumlah *cluster* 4 dan 6. Nilai *purity* pada algoritma WDM K-Modes dan K-Modes tradisional adalah sama. Namun *purity* pada algoritma WDM-EA K-Modes lebih unggul sedikit dari kedua algoritma terdahulu. Hal ini menunjukkan semakin banyak data, maka nilai *purity* yang diperoleh setiap algoritma akan mendekati atau sama nilainya.

4.2 Analisis Perbandingan Algoritma

Berdasarkan hasil pengujian pada Sub Bab 4.1, dapat disimpulkan nilai purity terbaik nya di Tabel 4.9 dan Grafiknya dapat dilihat di Gambar 4.9

Tabel 4.9 Perbandingan hasil *purity* setiap algoritma

Dataset	**Jumlah *Data***	**Jumlah Atribut**	**Jumlah *Cluster***	***Purity***		
				K-Modes	**WDM K-Modes**	**WDM-EA K-Modes**
soybean small	47	21	6	0.913	0.877	0.928
zoo	101	16	9	0.852	0.850	0.858
balance scale	625	4	5	0.562	0.588	0.588
car evaluation	1728	6	5	0.701	0.701	0.702



Gambar 4.9 Grafik Perbandingan hasil *purity* setiap algoritma

Berdasarkan Tabel 4.9 dan Gambar 4.9 terlihat nilai *purity* pada algoritma WDM-EA K-Modes sama atau lebih besar dari *purity* pada algoritma WDM K-Modes. Pada *dataset* yang berjumlah sedikit, algoritma WDM K-Modes cenderung lebih kecil dibandingkan algoritma K-Modes. Namun semakin banyak jumlah datanya, nilai *purity* semakin meningkat atau sama seperti nilai *purity* WDM-EA K-Modes. Begitu juga dengan algoritma K-Modes, pada data sedikit, nilai *purity*-nya lebih besar namun semakin banyak jumlah *dataset* nya, nilai *purity* pada algoritma K-Modes justru semakin menurun.

Nilai *purity* yang diperoleh setiap *dataset* berbeda-beda, hal ini dipengaruhi beberapa faktor, diantaranya oleh sebaran data, banyaknya data, pemilihan *centroid* awal dan jumlah atributnya. Pada dataset *soybean small* yang memiliki jumlah data sedikit, cenderung menghasilkan nilai *purity* yang tinggi, namun pada setiap percobaan terjadi ketidakstabilan hasil *purity*. Sedangkan pada dataset *balance scale* memiliki *purity* yang rendah dibanding *purity* pada *dataset*

lain, hal ini disebabkan kurang beragamnya data yang ada pada dataset *balance scale*, sehingga *purity* nya selalu menghasilkan sekitar 50% dari *purity* keseluruhan.

Perubahan jumlah *cluster* (K) dapat berpengaruh terhadap nilai akurasi hasil pengelompokkan. Hal ini dibuktikan untuk *dataset soybean small, zoo* dan *balance scale,* sejumlah K yaitu K=3 sampai K=6, menghasilkan nilai *purity* yang semakin besar. Jumlah kemunculan kategori pada setiap atribut juga mempengaruhi dalam pengelompokan menggunakan WDM-EA K-modes. Jika jumlah kemunculan kategori dalam setiap atribut bernilai sama maka memungkinkan data tidak berpindah *cluster* dan menyebabkan iterasi berhenti sebelum pengelompokkan data maksimal. Selain itu hasil akurasi *clustering* juga dipengaruhi oleh data yang digunakan. Semakin besar jumlah *record* data, maka akurasi pengelompokan akan semakin baik. Akan tetapi, hal ini berlaku jika data tersebut memiliki struktur yang baik seperti variasi nilai dalam setiap atribut.