PENGKLASIFIKASIAN DATA MENGGUNAKAN ANALISIS DISKRIMINAN DAN ANALISIS GEROMBOL K-RATAAN

By Khoirin Nisa

WORD COUNT 2557 TIME SUBMITTED 17-SEP-2019 03:44PM
PAPER ID 50418460

PENGKLASIFIKASIAN DATA MENGGUNAKAN ANALISIS DISKRIMINAN DAN ANALISIS GEROMBOL K-RATAAN

Khoirin Nisa
Jurusan Matematika FMIPA Unila
Email : Nisa_stat@unila.ac.id

ABSTRACT

Discriminant analysis and K-means clustering analysis are widely used statistical techniques for data classification. Here we compare the two methods by data simulation experiment. We generated data of the type most commonly assumed when using K-means clustering. The misclassifications rates and the R- squares (R^2) of the two methods are discussed. The result shows that the discriminant analysis performs better than K-means clustering in misclassification rate, but worse in R^2.

Keywords : discriminant analysis, k-means clustering, misclassification rate, R-squares.

PENDAHULUAN

Berbagai penelitian tentang metode-metode klasifikasi dalam statistika telah banyak dilakukan oleh para statistikawan, dan diantaranya adalah membandingkan beberapa metode klasifikasi, misalnya membandingkan analisis diskriminan dengan analisis regresi logistik untuk data yang berasal dari dua populasi (Press & Wilson, 1978), membandingkan analisis gerombol k-rataan dengan analisis gerombol kelas laten (Magidson&Vermunt,2002a), dan lain sebagainya.

Makalah ini membahas tentang pembandingan analisis diskriminan dengan analisis gerombol k-rataan. Analisis diskriminan merupakan suatu metode peubah ganda (*multivariate*) yang digunakan untuk mengklasifikasikan objek berdasarkan suatu fungsi dari peubah-peubah penjelas X_i yang disebut sebagai "fungsi diskriminan". Sedangkan analisis gerombol K-Rataan merupakan metode peubah ganda yang digunakan untuk mengklasifikasikan objek berdasarkan kemiripannya, dimana kemiripan antar objek diukur berdasarkan rumus jarak. Pengklasifikasian data dengan menggunakan kedua metode ini membutuhkan pengetahuan awal peneliti untuk menentukan banyaknya jumlah gerombol (kelompok) dari mana data berasal, untuk selanjutnya tiap-tiap objek dapat ditentukan (diduga) kelompok asalnya. Untuk analisis diskriminan diperlukan beberapa data latihan (*training data*) dengan informasi tentang kelompok asal data tersebut, data ini digunakan untuk menghitung fungsi diskriminan, selanjutnya fungsi diskriminan yang diperoleh digunakan untuk mengklasifikasikan data lainnya. Sedangkan

analisis gerombol k-rataan tidak memerlukan data awal yang demikian. Perbedaan-perbedaan ini dapat memberikan hasil pengklasifikasian yang berbeda yang akan kami kaji dengan menggunakan beberapa contoh kasus dalam simulasi data.

Analisis Diskriminan

Aturan paling sederhana pada pengklasifikasian data dengan analisis diskriminan bisa dinyatakan dalam fungsi kuadrat jarak yaitu,

$$d_j(\mathbf{x}) = (\mathbf{x} - \bar{x}_j)'\mathbf{S}^{-1}(\mathbf{x} - \bar{x}_j) - 2\ ln(\pi_t) \qquad (1)$$

dengan S^{-1} merupakan matriks ragam koragam gabungan dan π_t adalah peluang posterior. Rumus di atas berlaku jika matriks ragam koragam populasi masing-masing kelompok sama. Suatu objek **x** diklasifikasikan kepada populasi yang terdekat dihitung dengan menggunakan rumus di atas. Atau, **x** akan diklasifikasikan berasal dari populasi ke-*t* jika :

$$d_t^2(\mathbf{x}) = \min_{j=1,\ldots,k}\left\{d_j^2(\mathbf{x})\right\}$$

Mengklasifikasikan objek pengamatan ke populasi yang terdekat setara dengan mengklasifikasikan objek ke populasi dengan peluang posterior yang paling besar. Pada kasus *k* buah populasi, peluang posterior diperoleh dari rumus berikut,

$$P(t|\mathbf{x}) = \frac{e^{-\frac{1}{2}d_t^2(x)}}{\sum_{j=1}^{k} e^{-\frac{1}{2}d_j^2(x)}} \qquad (2)$$

Jika matriks ragam koragam tidak sama, maka pengklasifikasian data ditentukan oleh fungsi berikut,

$$d_j^2(\mathbf{x}) = [\mathbf{x} - \bar{\mathbf{x}}_j]'\mathbf{S}^{-1}[\mathbf{x} - \bar{\mathbf{x}}_j] + ln\ |\mathbf{S}_j| - 2\ ln(\pi_t)$$

dengan peluang posterior π_t sama dengan rumus pada persamaan (2).

Analisis Gerombol K-Rataan

Pengklasifikasian data dengan dengan analisis gerombol didasarkan pada kemiripan antar objek yang diukur dengan menggunakan rumus jarak. Rumus jarak yang umum digunakan dalam analisis gerombol adalah jarak *euclid* yang dinyatakan dalam rumus berikut,

$$d(x,y) = \sqrt{(x-y)'(x-y)}$$
$$= \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$$

Analisis gerombol k-rataan memerlukan *k* buah centroid awal yang digunakan sebagai nilai tengah awal gerombol yang akan dibentuk. Suatu objek dimasukkan ke dalam suatu gerombol jika ia memiliki jarak terdekat dengan centroid gerombol tersebut dibandingkan dengan *k*-1 centroid lainnya. Setelah gerombol-gerombol terbentuk kemudian dilakukan perhitungan centroid yang baru untuk memperbaiki penggerombolan dengan menghitung jarak setiap objek dengan centroid yang baru. Langkah ini berlangsung terus sehingga centroid konvergen dan tidak ada lagi objek yang berpindah dari satu gerombol ke gerombol yang lain.

Salah satu statistik yang digunakan untuk mengevaluasi hasil penggerombolan adalah statistik R^2 (R-kuadrat), yaitu :

$$R^2 = \frac{JK_A}{JK_T} = \frac{JK_T - JK_D}{JK_T} = 1 - \frac{JK_A}{JK_T}^2$$

dengan JK_A merupakan jumlah kuadrat antar gerombol, JK_D adalah jumlah kuadrat dalam gerombol, sedangkan JK_T adalah jumlah kuadrat total. Rumus jumlah kuadrat secara umum adalah sebagai berikut,

$$JK = \sum (x_i - \bar{x})^2$$

Semakin tinggi nilai R^2 menunjukkan bahwa perbedaan antar gerombol semakin besar. Selain itu nilai R^2 yang besar juga menunjukkan bahwa nilai jumlah kuadrat dalam kelompok (JK_D) kecil, ini berarti bahwa objek-objek dalam gerombol yang terbentuk memiliki kemiripan yang tinggi.

METODOLOGI PENELITIAN

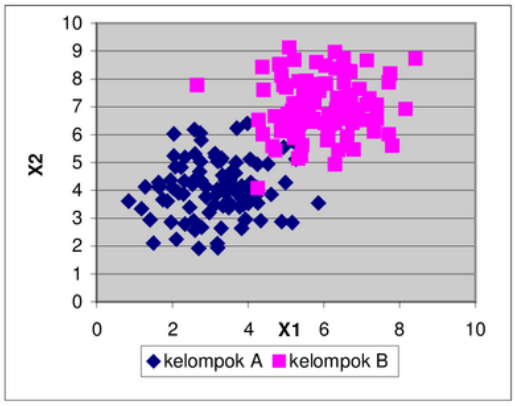
Metode penelitian yang dilakukan yaitu studi literatur dan simulasi data dengan menggunakan perangkat lunak Minitab. Langkah-langkah yang dilakukan adalah sebagai berikut :

1. Membangkitkan data yang terdiri dari dua kelompok. Data yang dibangkitkan sebanyak empat bentuk seperti yang digunakan oleh Magidson&Vermunt (2002b). Sampel masing-masing data yang dibangkitkan berukuran n=200, terdiri dari 2 peubah berdistribusi normal dan dibagi ke dalam dua kelompok dengan masing-masing kelompok terdiri dari 100 sampel. Gugus data pertama yang dibangkitkan adalah bentuk kasus dimana peubah saling bebas dengan ragam setiap kelompok sama, gugus data kedua adalah kasus di mana peubah saling bebas dengan ragam kelompok tidak sama, gugus data ketiga adalah kasus peubah saling bebas dengan ragam untuk peubah X_2 dan X_1 tidak sama pada masing-masing kelompok, sedangkan gugus data keempat adalah kasus untuk peubah saling berkorelasi. Sebaran data yang kami bangkitkan dapat dilihat pada Tabel 1 di lampiran.

2. Melakukan analisis diskriminan dan analisis gerombol k-rataan.
3. Menentukan nilai salah klasifikasi hasil analisis kedua metode.
4. Menghitung R^2 berdasarkan pengklasifikasian kedua metode.
5. Membandingkan hasil yang diperoleh kedua metode berdasarkan nilai salah klasifikasi dan jumlah kudratnya.

HASIL DAN PEMBAHASAN

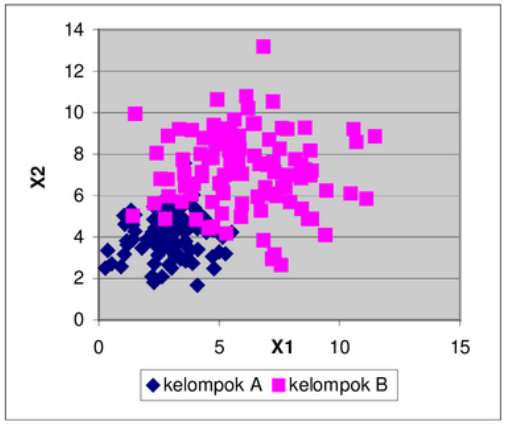
Plot sebaran gugus data pertama dapat dilihat pada Gambar 1. Untuk gugus data ini hasil analisis diskriminan memberikan hasil yang lebih baik dari analisis gerombol. Analisis dikriminan memberikan hasil sebagai berikut : ada 4 objek pada kelompok A yang diklasifikasikan ke dalam kelompok B dan ada 1 objek pada kelompok B yang diklasifikasikan ke dalam kelompok A. Jadi analisis diskriminan mengidentifikasi terdapat 97 objek dalam kelompok A dan ada 103 objek dalam kelompok B. Sehingga analisis diskriminan menghasilkan salah klasifikasi sebanyak 5 kasus, atau sama dengan 2,5% dari seluruh jumlah objek dalam kedua kelompok. Sedangkan analisis gerombol memberikan nilai salah klasifikasi yang lebih besar dari analisis diskriminan, yaitu ada 5 objek pada kelompok A yang diklasifikasikan ke dalam kelompok B dan 1 objek pada kelompok B diklasifikasikan ke dalam kelompok A. Secara keseluruhan ada 6 kasus (3%) salah klasifikasi yang dihasilkan oleh analisis gerombol.



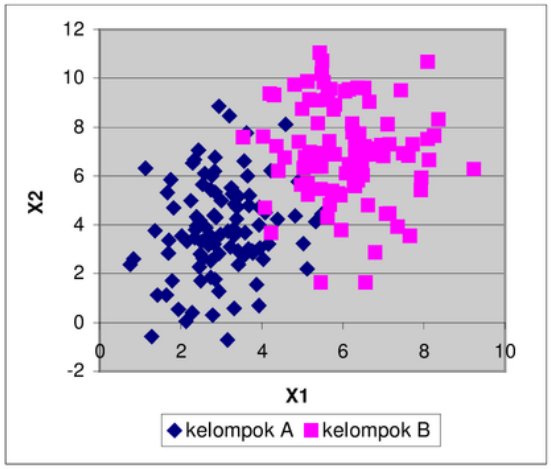
Gambar 1. Sebaran gugus data 1, kasus ragam sama

Gugus data 2 memiliki sebaran data yang berbeda pada kedua kelompok, yaitu ragam peubah pada kelompok B digandakan. Plot sebaran data untuk gugus data ini dapat dilihat pada Gambar 2. Untuk gugus data ke dua analisis diskriminan memberikan hasil yang lebih baik dari analisis gerombol. Analisis dikriminan memberikan hasil sebagai berikut : ada satu objek pada kelompok A yang diklasifikasikan ke dalam kelompok B, dan ada 15 objek pada

kelompok B yang diklasifikasikan ke dalam kelompok A. Jadi analisis diskriminan mengidentifikasi terdapat 114 objek dalam kelompok A dan ada 86 objek dalam kelompok B. Sehingga analsisi diskriminan menghasilkan salah klasifikasi sebanyak 16 kasus, atau sama dengan 8% dari seluruh jumlah objek dalam kedua kelompok. Sedangkan analisis gerombol memberikan nilai salah klasifikasi yang lebih besar dari analisis diskriminan, yaitu ada 1 objek pada kelompok A diklasifikasikan ke dalam kelompok B, dan ada 23 objek pada kelompok B yang diklasifikasikan ke dalam kelompok A. Jadi ada 24 kasus (12%) salah klasifikasi yang dihasilkan oleh analisis gerombol.



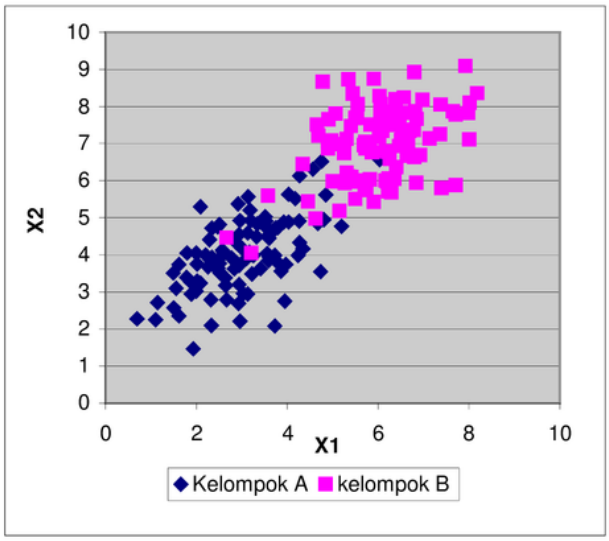
Gambar 2. Sebaran data gugus data 2, ragam peubah kelompok B digandakan



Gambar 3. Sebaran gugus data 3, ragam X_2 digandakan

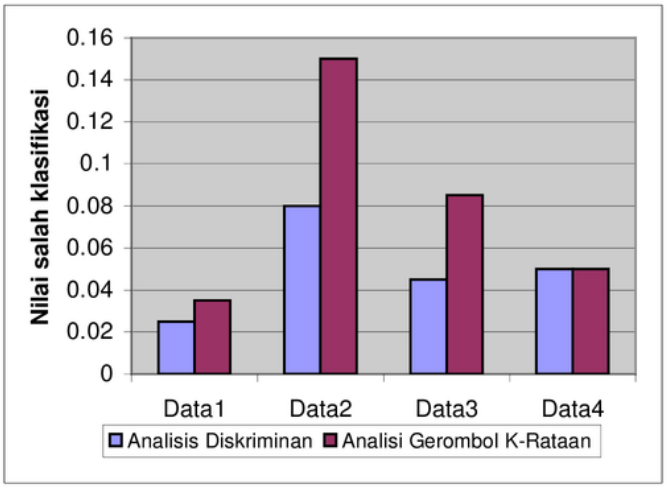
Gugus data 3 serupa dengan gugus data 1 hanya saja ragam peubah kedua, yaitu $X_{2,}$ digandakan ($\sigma_2=2\sigma_1$) untuk tiap kelompok. Hasil klasifikasi untuk data ini memberikan hasil yang berbeda daru kedua metode. Untuk gugus data 3 analisis diskriminan juga memberikan hasil yang lebih baik dari analisis gerombol. Analisis dikriminan memberikan hasil sebagai berikut : ada 5 objek pada kelompok A yang dikasisifikasikan ke dalam kelompok B, dan ada 4 objek pada kelompok B yang diklasifikasikan ke dalam kelompok A. Sehingga secara keseluruhan analisis diskriminan menghasilkan salah klasifikasi sebanyak 9 asus, atau sama dengan 4.5% (0,045) dari seluruh jumlah objek dalam kedua kelompok. Sedangkan analisis gerombol memberikan hasil sebagai berikut: ada 10 objek pada kelompok A dikasifikasikan ke dalam kelompok B, dan ada 7 objek pada kelompok B yang diklasifikasikan ke dalam kelompok A. Berarti ada 17 kasus salah klasifikasi sehingga nilai salah klasifikasinya adalah 8.5% (0,085).

Untuk gugus data 4 kami tambahkan korelasi antar peubah X1 dan X2, dengan demikian maka asumsi saling bebas (*independent*) dalam analisisis diskriminan tidak lagi terpenuhi. Untuk data ini hasil pengklasifian dengan analisis diskriminan dan analisis gerombol k-rataan memberikan hasil yang sama. Yaitu terdapat 10 kasus salah klasifikasi atau 5% dari seluruh pengamatan dengan perincian sebagai berikut : ada 5 objek dalam kelompok A diklasifikasikan ke dalam kelompok B, dan 5 objek dalam kelompok B diklasifikasikan ke dalam kelompok A. Sehingga kedua metode mengidentifikasi masing-masing kelompok terdiri dari 100 objek.



Gambar 4. Sebaran gugus data 4, peubah X_1 dan X_2 saling berkorelasi

Ringkasan hasil pengklasifikasian dan nilai salah klasifikasi untuk kedua metode dapat dilihat pada Tabel 2 dan Tabel 3 di lampiran. Nilai salah klasifikasi pada Tabel 3 disajikan bentuk diagram pada Gambar 5 berikut ini.

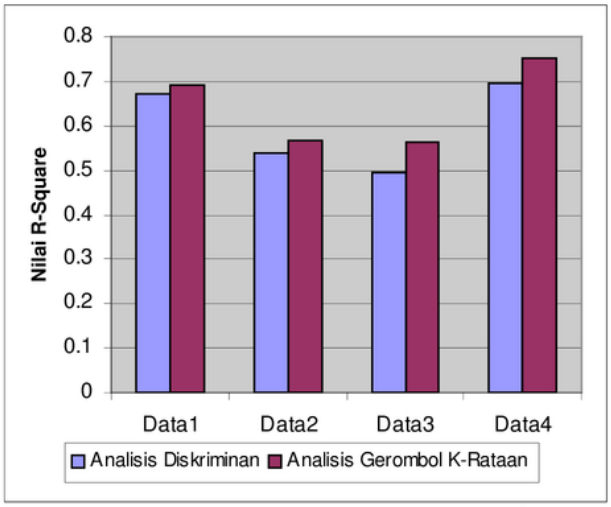


Gambar 5. Diagram batang nilai salah klasifikasi

Berdasarkan nilai salah klasifikasinya maka dapat disimpulkan bahwa analisis diskriminan selalu memberikan hasil yang lebih baik dibandingkan dengan analisis gerombol k-rataan dalam mengklasifikasikan objek dengan tepat untuk tiap kasus. Hal ini terlihat dari nilai salah klasifikasi analisis diskriminan yang lebih kecil untuk 3 data pertama. Meskipun nilai salah klasifikasi yang dihasilkan oleh analisis gerombol masih cukup rendah (kurang dari 20%), namun untuk data dengan sebaran tertentu nilai ini dapat meningkat seiring dengan semakin kompleksnya sebaran data antar kelompok. Berarti akan semakin banyak pula objek yang diklasifikasikan bukan kepada kelompok sebenarnya.

Untuk mengevaluasi hasil pengelompokan maka dapat dilakukan dengan menggunakan statistik R^2. Nilai R^2 hasil analisis diskriminan dan analisis gerombol k-rataan untuk empat gugus data di atas dapat dilihat pada Tabel 4 dan disajikan dalam bentuk diagram pada Gambar 7.

Untuk setiap bentuk data, nilai R^2 kedua metode memiliki nilai yang berbeda. Nilai R^2 untuk analisis gerombol lebih besar dari pada analisis diskriminan. Ini berarti bahwa pengklasifikasian objek dengan analisis gerombol menghasilkan gerombol-gerombol yang lebih homogen dibandingkan dengan analisis diskriminan. Semakin tinggi R^2 juga juga berarti keheterogenan antar gerombol semakin tinggi.



Gambar 7. Diagram batang nilai R^2

Berdasarkan nilai R^2 yang diperoleh maka dapat disimpulkan bahwa kemiripan antar objek dalam gerombol-gerombol yang dihasilkan oleh analisis gerombol k-rataan lebih tinggi dibandingkan dengan kemiripan antar objek dalam gerombol-gerombol yang dihasilkan oleh analisis diskriminan. Dengan demikian maka jika seorang peneliti menginginkan kelompok yang "solid" maka analisis gerombol k-rataan memberikan hasil yang lebih baik. Namun jika pengklasifikasian data lebih ditujukan untuk menduga kelompok asal yang sebenarnya (*the true groups*), maka analisis diskriminan akan memberikan hasil yang lebih baik.

KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian ini maka dapat disimpulkan bahwa berdasarkan nilai salah klasifikasinya maka analisis diskriminan lebih baik dalam mengklasifikasikan data dari pada analisis gerombol k-rataan, namun jika berdasarkan nilai R^2 maka analisis gerombol k-rataan lebih baik dari pada analisis diskriminan

Disarankan bagi para pembaca jika ingin mengklasifikasikan data berdasarkan sifat kelompok asalnya (meminimumkan nilai salah klasifikasi) maka gunakanlah analisis diskriminan, yaitu dengan mengumpulkan beberapa data awal sebagai "training data" untuk menghitung fungsi diskriminan dan selanjutnya fungsi ini digunakan sebagai aturan dalam pengklasifikasian. Sedangkan jika ingin mengklasifikasikan data untuk mendapatkan kelompok-kelompok yang bersifat sehomogen mungkin (memaksimumkan nilai R^2) maka analisis gerombol k-rataan lebih tepat.

DAFTAR PUSTAKA

Anderson, H. & Black, T. 1998. *Multivariate Data Analysis*. Prentice Hall Inc. New Jersey.

Dudoit, S. & Gentleman . R. 2002. Cluster Analysis in DNA Microarray Experiments. Bioconductor Short Course Paper.

Feighner, J.P., Sverdlov, L. 2001. The Use of Discriminant Analysis to Separate a Study Population by Treatment Subgroups in a Clinical Trial with a New Pentapeptide Antidepressant. *Journal of Applied Research in Clinical and Experimental Therapeutics.*

Grubesic, T. H. 2002 *Detecting hot spots using cluster analysis and GIS*. http://www.ojp.usdoj.gov/nij/maps/Conferences/ 01conf/ Grubesic.doc

Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R. & Wu. A. Y. 2002. An Efficient k-Means Clustering Algorithm: Analysis and Implementation. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, Vol. 24, No. 7.

Magidson J. & Vermunt, J.K. 2002a. Latent class models for clustering: A comparison with K-means. *Canadian Journal of Marketing Research*, 20, 37-44.

Magidson J. & Vermunt, J.K. 2002b. Latent class modeling as a probabilistic extension of K-means clustering. *Quirkís Marketing Research Review*, March.

Ng, A., Jordan, M., & Weiss, Y. 2001. On spectral clustering: Analysis and an algorithm. *Proc. Natural Info.Processing Systems* (NIPS 2001).

Press, S. J. & Wilson, S. 1978. Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, Vol. 73: 699-705

Sharma S. 1996. *Applied Multivariate Techiques In Statistics.* Academic Press, San Diego.

Tang, B., Shepherd, M., Milios, E. & Heywood, M.I. 2005. Comparing and Combining Dimension Reduction Techniques for Efficient Text Clustering. *SIAM Journal.*

Yeung, K.Y., Fraley, C., Murua, A., Raftery, A.E., Ruzzo, W.L. 2001. *Model-based Clustering and Data Transformations for Gene Expression Data*. University of Washington, Dept. of Statistics, Technical Report no 396.

Tabel 1. Sebaran Data Simulasi

Gugus	Kelompok A	Kelompok B
1	X_1~N(3,1) X_2~N(4,1)	X_1~N(6,1) X_2~N(7,1)
2	X_1~N(3,1) X_2~N(4,1)	X_1~N(6,2) X_2~N(7,2)
3	X_1~N(3,1) X_2~N(4,2)	X_1~N(6,1) X_2~N(7,2)
4	X_1~N(3,1) X_2~N(4,1) Corr (X_1,X_2) = 0,6	X_1~N(6;1) X_2~N(7;1) Corr (X_1,X_2)=0,45

Tabel 2. Hasil pengklasifikasian data

Gugus Data	Kelompok Sebenarnya	Analisis Diskriminan		Analisis Gerombol K-rataan	
		A	B	A	B
Data1	A	96	4	95	5
	B	1	99	1	99
	Jumlah	97	103	96	104
Data2	A	99	1	99	1
	B	15	85	23	77
	Jumlah	114	86	122	78
Data3	A	95	5	90	10
	B	4	96	7	93
	Jumlah	99	101	97	103
Data4	A	95	5	95	5
	B	5	95	5	95
	Jumlah	100	100	100	100

Tabel 3. Nilai salah klasifikasi kedua metode

Gugus Data	Nilai salah klasifikasi	
	Analisis Diskriminan	Analisis gerombol K-Rataan
Data1	0,025	0,030
Data2	0,080	0,120
Data3	0,045	0,085
Data4	0,050	0,050

Tabel 4. Nilai R^2 untuk kedua metode

Gugus Data	Nilai R^2 (R-kuadrat)	
	Analisis Diskriminan	Analisis gerombol K-Rataan
Data1	0,671784	0,690193
Data2	0,536814	0,566370
Data3	0,492887	0,562369
Data4	0,694269	0,753522

PENGKLASIFIKASIAN DATA MENGGUNAKAN ANALISIS DISKRIMINAN DAN ANALISIS GEROMBOL K-RATAAN

ORIGINALITY REPORT

13%

SIMILARITY INDEX

PRIMARY SOURCES

1	www.statisticalinnovations.com Internet	58 words — 2%
2	Young Yong Kim. "A NETwork COding based Multicasting (NETCOM) over IEEE 802.11 Multi-hop", 2007 IEEE 65th Vehicular Technology Conference - VTC2007-Spring, 04/2007 Crossref	32 words — 1%
3	Xiaofeng He. "Linearized cluster assignment via spectral ordering", Twenty-first international conference on Machine learning - ICML 04 ICML 04, 2004 Crossref	22 words — 1%
4	Laith Mohammad Qasim Abualigah. "Chapter 1 Introduction", Springer Nature, 2019 Crossref	21 words — 1%
5	acikarsiv.ankara.edu.tr Internet	18 words — 1%
6	www.iaca.net Internet	17 words — 1%
7	file.scirp.org Internet	15 words — 1%
8	etd.lib.metu.edu.tr Internet	14 words — 1%
9	www.airitilibrary.com Internet	13 words — < 1%

10	Daniel Stahl, Andrew Pickles, Mayada Elsabbagh, Mark H. Johnson, The BASIS Team. "Novel Machine Learning Methods for ERP Analysis: A Validation From Research on Infants at Risk for Autism", Developmental Neuropsychology, 2012 Crossref	12 words — < 1%
11	greendom-afc.blogspot.com Internet	12 words — < 1%
12	docplayer.info Internet	12 words — < 1%
13	eprints.uny.ac.id Internet	12 words — < 1%
14	soddis.blogspot.com Internet	10 words — < 1%
15	wapresri.go.id Internet	10 words — < 1%
16	digilib.unila.ac.id Internet	10 words — < 1%
17	fahmiqbo.blogspot.com Internet	10 words — < 1%
18	Mingyang Li, Lijun Zhang, Yunxin Wu, Hao Chi. "Comprehensive assessment of optical network in power grid", 2016 15th International Conference on Optical Communications and Networks (ICOCN), 2016 Crossref	10 words — < 1%
19	isamveri.org Internet	9 words — < 1%
20	www.umpalangkaraya.ac.id Internet	8 words — < 1%
	anzdoc.com	

21	Internet	8 words — < 1%
22	Urszula Ledzewicz. "A High-Order Generalized Local Maximum Principle", SIAM Journal on Control and Optimization, 2000 Crossref	7 words — < 1%

EXCLUDE QUOTES	ON	EXCLUDE MATCHES	OFF
EXCLUDE BIBLIOGRAPHY	ON		