Seleksi Variabel Menggunakan Algoritma Genetika untuk Klasifikasi Data *Indian Liver Patient Dataset* (ILDP)

Rizky Kusumawardani*, Irhamah dan Heri Kuswanto

Abstrak

Analisis diskriminan merupakan salah satu metode statistika yang dapat digunakan untuk klasifikasi, metode ini sesuai digunakan untuk klasifikasi ketika variabel dependen yang digunakan bertipe kategorikal dan variabel independennya bertipe kontinu. Data penelitian yang tersedia untuk analisis diskriminan rata-rata memuat banyak variabel independen. Oleh karena itu, analisis diskriminan membutuhkan tahapan seleksi variabel untuk memilih variabel independen yang berkontribusi besar terhadap fungsi diskriminannya. Metode statistika yang sering digunakan untuk seleksi variabel adalah stepwise method. Penggunaan metode ini belum memberikan hasil yang optimal. Metode stepwise terlalu sensitif terhadap perubahan tingkat signifikasi, rentan terhadap kasus multikolinieritas, dan tidak memberikan jaminan kebaikan hasil. Oleh karena itu, diperlukan metode baru yang dapat memperbaiki kekurangan tersebut. Metode yang dapat digunakan adalah algoritma genetika, yang merupakan metode iteratif untuk mendapatkan hasil global optimum. Pada penelitian ini digunakan metode algoritma genetika untuk seleksi variabel pada klasifikasi data Indian Liver Patient Dataset (ILDP). Hasil yang diperoleh adalah metode seleksi variabel untuk analisis diskriminan yang memberikan nilai kesalahan klasifikasi terkecil adalah algoritma genetika.

Kata-kata kunci: Analisis Diskriminan, Algoritma Genetika, Indian Liver Patient Dataset (ILDP).

Pendahuluan

Tahapan seleksi variabel pada analisis diskriminan sangat penting untuk mengetahui variabel independen yang memiliki kontribusi besar pada fungsi diskriminan. Metode yang biasanya digunakan untuk seleksi variabel adalah metode *stepwise*. [1] menyatakan bahwa metode ini rentan digunakan ketika ada kasus multikolinieritas. Penelitian [2] yang berkaitan dengan kebaikan metode klasik (*forward selection, backward elimination,* dan *stepwise method*) juga memberikan kesimpulan bahwa metode seleksi variabel klasik pada pemodelan regresi akan memberikan informasi yang kurang lengkap karena adanya kasus multikolinieritas. Penelitian terkait metode seleksi variabel pada analisis diskriminan juga telah banyak dilakukan. [3] membandingkan metode seleksi variabel *stepwise* untuk analisis diskriminan dan regresi logistik, serta algoritma genetika untuk *neural network*, hasilnya adalah algoritma genetika baik digunakan untuk data 1 tahun dan 3 tahun sebelum kegagalan, sedangkan metode *stepwise* baik digunakan untuk data 2 tahun sebelum kegagalan, selain itu [4] juga melakukan penelitian dengan menggunakan algoritma genetika yang dikombinasikan dengan analisis diskriminan untuk mengidentifikasi variabel, penelitian ini memberikan hasil bahwa algoritma genetika dapat mengidentifikasi variabel yang menjadi penyebab masalah.

Berdasarakan pemaparan tersebut dapat ditarik sebuah permasalahan bahwa metode seleksi variabel klasik khususnya metode *stepwise* untuk seleksi variabel pada analisis diskriminan rentan akan masalah multikolinieritas, sehingga diperlukan metode lain untuk seleksi variabel salah satunya adalah dengan menggunakan algoritma genetika. Penggunaan algoritma genetika untuk seleksi variabel sudah terbukti baik berdasarkan pemaparan tersebut. Oleh karena itu, pada penelitian ini akan digunakan metode algoritma genetika untuk seleksi variabel yang akan dibandingkan kebaikannya dengan metode *stepwise* pada data (ILDP) yang diambil dari [5].

Teori

Analisis diskriminan merupakan salah satu metode satatistika yang dapat digunakan untuk klasifikasi ketika variabel dependenya bertipe kategori dan variabel independennya bertipe kontinu. Persamaan linier dari fungsi diskriminan dapat dilihat pada Persamaan (1)

$$Z_{ij} \equiv \alpha + w_1 X_{1j} + w_2 X_{2j} + ... + w_p X_{pj} \,. \quad (1)$$

J=1,2,....,*q*

Z_{ij} merupakan nilai diskriminannya, α merupakan konstanta, dan w_i merupakan pembobot. Tahapan awal sebelum melakukan pemodelan analisis diskriminan adalah melakukan beberapa pengujian yang diperlukan, yaitu pengujian multikolinieritas, distribusi data multivariat normal, kesamaan matriks varians-kovarians, dan pengujian perbedaan rata-rata

antar kategori. Pengujian multikolinieritas menggunakan nilai korelasinya, menurut [6] apabila korelasi antar variabel penelitian lebih dari 0,8 maka terdapat kasus multikolinieritas. Pengujian data berdistribusi normal multivariat menggunakan pengujian *Henze-Zirkler's Multivariate Normality Test*, apabila p_{value} statistik uji kurang dari 0,05 maka data tidak berdistribusi normal multivariat. Pengujian kesamaan matriks varians-kovarians dapat menggunakan *Box's M*, apabila p_{value} statistik uji kurang dari 0,05 maka matriks varians-kovarians data berbeda. Pengujian perbedaan rata-rata antar kategori dapat menggunakan *wilk's lamda,* apabila p_{value} statistik uji kurang dari 0,05 maka rata-rata antar kategori berbeda

Apabila data penelitian telah memenuhi asumsi, maka tahapan selanjutnya adalah memilih variabel yang memiliki kontribusi besar terhadap fungsi diskriminan. Metode yang dapat digunakan adalah *stepwise method.* Kriteria yang digunakan untuk seleksi variabel adalah *wilk's lamda.* Selain itu juga digunakan metode yang lain sebagai pembanding, yaitu algoritma genetika. Algoritma genetika merupakan metode yang berbasis perulangan untuk mendapatkan nilai global optimum. Terdapat 6 tahapan utama dalam algoritma genetika, yaitu.

1. Membentuk populasi awal yang terdiri dari 100 kromosom, dimana kromosom yang digunakan bertipe *bitstring* untuk proses seleksi variabel.
2. Mengevaluasi masing-masing kromosom berdasarkan nilai fitnesnya. Nilai fitnes yang digunakan untuk seleksi variabel adalah kesalahan klasifikasi model.
3. Memilih orang tua untuk digandakan dengan menggunakan metode *roulette whelel selection* (RWS).
4. Melakukan proses pindah silang.
5. Melakukan proses mutasi
6. Mengulangi proses tahapan 2-5 apabila belum didapatkan kromosom yang optimum yang menjadi solusi dari permasalahan.

Data penelitian yang digunakan merupakan data sekunder, yaitu data orang terkena penyakit liver dan tidak terkena penyakit liver. Terdapat dua jenis variabel penelitian yang digunakan yaitu variabel dependen dan variabel independen. Variabel dependen bertipe kategori yang terdiri dari pasien tidak liver (1) dan pasien liver (2). Variabel independennya bertipe kontinu yang terdiri dari usia, *total bilirubin* (TB), *direct bilirubin* (DB), *alkphos alkaline phosphotase* (AAP), *Sgpt alamine aminotransferase* (ALT), *Sgot aspartate aminotransferase* (AST), *total protiens* (TP), *Albumin* (ALB), dan *albumin and globumin ratio* (A/G).

Hasil dan diskusi

Terdapat 583 data pengamatan yang dibagi menjadi data training 525 pengamatan dan data testing 58 pengamatan. Sebelum data dibagi menjadi data training dan testing, data dianalisis terlebih dahulu menggunakan analisis deskriptif. Tabel 1 berikut ini memberikan informasi terkait analisis deskriptif data pengamatan.

Tabel 1. Analisis Deskriptif Data Pengamatan Tiap Kategori.

Variabel	Rata-Rata	Standar Deviasi	Modus
Usia	44,75	16,19	60,00
TB	3,29	6,21	0,80
DB	1,49	2,81	0,20
AAP	290,60	242,90	198,00 215,00 298,00
ALT	80,71	182,62	25,00
AST	109,90	288,90	23,00
TP	6,48	1,09	7,00
ALB	3,14	0,79	3,00
A/G	0,95	0,32	1,00

Informasi yang dapat diperoleh berdasarkan Tabel 1 adalah pasien paling banyak berusia 60 tahun dengan rata-rata usia sebesar 44,75 tahun dan keragaman data sebesar 16,19. Informasi untuk variabel yang lain dapat dilihat pada Tabel 1. Keragaman data paling besar terdapat pada variabel AST dan yang paling kecil terdapat pada variabel A/G.

Tahapan selanjutnya setelah analisis deskriptif adalah pengujian asumsi data. Berdasarkan perhitungan nilai korelasi antar variabelnya didapatkan bahwa variabel yang memiliki korelasi paling besar ada antara variabel TB dan DB dengan nilai korelasi sebesar 0,863. Tabel 2 berikut ini menunjukkan rangkuman hasil pengujian asumsi analisis diskriminan, dimana (1) menunjukkan pengujian data berdistribusi normal multivariat, (2) menunjukkan pengujian kesamaan matriks varians-kovarians, dan (3) menunjukkan pengujian perbedaan rata-rata.

Informasi yang dapat diperoleh dari Tabel 2 adalah data ILDP tidak memenuhi asumsi data berdistribusi normal multivariat dan asumsi matriks varians-kovarians antar kategori sama, karena p_{value} yang dihasilkan dari kedua pengujian tersebut kurang dari 0,05. Namun data ILDP telah memenuhi asumsi bahwa rata-rata antar kategorinya telah berbeda. Berdasarkan

pengujian asumsi diketahui bahwa terdapat dua asumsi yang tidak terpenuhi, pada penelitian ini kedua asumsi tersebut diasumsikan telah terpenuhi, karena menurut [3] terlanggarnya asumsi tersebut tidak menyebabkan kemampuan klasifikasinya menjadi buruk.

Tabel 2. Rangkumann Pengujian Asumsi Analisis Diskriminan.

Pengujian	Statistik Uji	Nilai Statistik Uji	P_{value}
(1)	HZ	20,71	0,00
(2)	*Box's M*	2450,09	0,00
(3)	*Wilk's Lamda*	69,51	0,00

Berdasarkan ulasan tersebut secara multivariat telah diketahui bahwa rata-rata antar kategorinya berbeda, Tabel 3 ini menyajikan hasil pengujian perbedaan rata-rata antar kategori untuk setiap variabel independen.

Tabel 3. Pengujian Perbedaan Rata-Rata antar Kategori untuk setiap Variabel Independen.

Variabel	*Wilk's Lamda*	F_{hitung}	P_{value}
Usia	0,982	9,439	0,002
TB	0,950	27,801	0,000
DB	0,937	35,308	0,000
AAP	0,957	23,554	0,000
ALT	0,972	15,311	0,000
AST	0,977	12,221	0,001
TP	0,998	1,166	0,281
ALB	0,968	17,464	0,000
A/G	0,969	16,670	0,000

Tabel 3 memebrikan informasi bahwa dari 9 variabel independen terdapat 8 variabel independen yang memiliki perbedaan rata-rata antar kategorinya atau dengan kata lain berpengaruh terhadap variabel dependennya. Variabel independen yang tidak berpengaruh terhadap variabel dependen tersebut adalah variabel TP. Variabel yang memiliki pengaruh paling besar adalah variabel ALB sedangkan yang paling kecil pengaruhnya adalah variabel Usia.

Tahapan selanjutnya adalah melakukan seleksi variabel untuk memilih variabel yang memiliki kontribusi besar terhadap fungsi diskriminan. Tabel 4 dan tabel 5 menyajikkan tahapan seleksi variabel menggunakan *stepwise method.* Tabel 4 merupakan tahapan pemilihan variabel independen yang akan dimasukkan ke dalam fungsi diskriminan.

Tabel 4. Tahapan Pemilihan Variabel di luar Model.

Tahapan	Variabel	*Wilk's Lamda*	F_{hitung}	Variabel masuk dalam model
0	Usia	0,982	9,439	DB
	TB	0,950	27,801	
	DB	0,937	35,308	
	AAP	0,957	23,554	
	ALT	0,972	15,311	
	AST	0,977	12,221	
	TP	0,998	1,166	
	ALB	0,968	17,464	
	A/G	0,969	16,670	
1	Usia	0,920	9,380	AAP
	TB	0,937	0,075	
	AAP	0,914	12,852	
	ALT	0,924	6,938	
	AST	0,929	4,565	
	TP	0,934	1,300	
	ALB	0,921	8,814	
	A/G	0,920	9,282	
2	Usia	0,901	7,791	Usia
	TB	0,914	0,075	
	ALT	0,905	5,497	
	AST	0,909	2.986	
	TP	0,912	1,335	
	ALB	0,902	7,207	
	A/G	0,904	6,061	
3	TB	0,901	0,060	ALT
	ALT	0,889	6,886	
	AST	0,895	3,287	
	TP	0,900	0,428	
	ALB	0,894	4,032	
	A/G	0,894	3,686	
4	TB	0,889	0,033	A/G
	AST	0,889	0,228	
	TP	0,889	0,216	
	ALB	0,882	3,956	
	A/G	0,882	4,010	
5	TB	0,882	0,002	
	AST	0,882	0,345	
	TP	0,882	0,002	
	ALB	0,882	0,832	

Informasi yang diperoleh dari Tabel 4 adalah terdapat 5 variabel independen yang akan dimasukkan dalam model yaitu variabel DB, AAP,

ISBN: 978-602-19655-8-0

Usia, ALT, dan A/G. Tabel 5 berikut ini menyajikan tahapan evaluasi variabel yang ada di dalam fungsi diskriminan harus dikeluarkan atau tidak.

Tabel 5. Tahapan Evaluasi Variabel di dalam Model.

Tahapan	Variabel	*Wilk's Lamda*	F_{hitung}	Variabel dikeluarkan dari model
1	DB	0,937	35,308	-
2	DB	0,957	24,354	-
	AAP	0,937	12,852	
3	DB	0,944	24,775	-
	AAP	0,920	11,246	
	Usia	0,914	7,791	
4	DB	0,921	18,646	-
	AAP	0,905	9,605	
	Usia	0,905	9,181	
	ALT	0,901	6,886	
5	DB	0,909	15,624	-
	AAP	0,895	7,444	
	Usia	0,893	6,520	
	ALT	0,894	7,206	
	A/G	0,889	4,010	

Berdasarkan Tabel 5 dapat diperoleh informasi bahwa tidak ada variabel yang akan dikeluarkan dari dalam model, karena kelima variabelnya berpengaruh signifikan. Pengaruh multikolinieritas pada *stepwise method* tampak dari variabel TB yang tidak dimasukkan dalam model karena berkorelasi dengan variabel DB. Seleksi variabel menggunakan algoritma genetika menghasilkan 5 variabel yaitu usia, AAP, ALT, PT, dan A/G. Rangkuman hasil kesalahan klasifikasi dari perhitungan model dan prediksi dari keempat metode dapat dilihat pada Tabel 6.

Tabel 6. Rangkumann Nilai Kesalahan Klasifikasi Model dan Prediksi.

Metode Seleksi Variabel	Kesalahan Klasifikasi Model	Kesalahan Klasifikasi Prediksi
Stepwise method	0,3638	0,3448
Algoritma Genetika	0,3562	0,2414

Informasi yang dapat diperoleh berdasarkan Tabel 6 adalah untuk pembentukan model fungsi diskriminan, kesalahan klasifikasi terkecil diperoleh pada saat menggunakan metode algoritma genetika untuk seleksi variabel. Apabila ditinjau dari hasil kesalahan klasifikasi prediksi, nilai kesalahan klasifikasi terkecil didapatkan ketika menggunakan metode algoritma genetika. Berdasarkan tahapan seleksi variabel maka variabel yang akan digunakan untuk estimasi parameter adalah hasil dari algoritma genetika. Model yang didapatkan adalah sebagai berikut ini

$$Z = 0{,}8312 - 0{,}0225 * usia - 0{,}0022 * AAP - 0{,}0030 * ALT + 0{,}0165 * PT + 0{,}4483 * A/G.$$

Kesimpulan

Berdasarkan hasil dan diskusi dapat disimpulkan bahwa metode seleksi variabel untuk analisis diskriminan yang memberikan nilai kesalahan klasifikasi terkecil baik untuk model maupun prediksi adalah metode algoritma genetika.

Referensi

[1] Hair J. F., Black W. C., Babin B. J., and Anderson R. E, *Multivariate Data Analysis* (7th Edition ed.), Pearson Prentice Hall, New Jersey, 2010

[2] Sartono B., Pengenalan Algoritma Genetika untuk Pemilihan Peubah Penjelas dalam Model Regresi Menggunakan SAS/IML. Forum Statistika dan Komputasi, 2010, pp. 10-15

[3] Back B., Laitinen T., Sere K., and Wezel M. V, "Choosing Bankruptcy Predictors Using Discriminan Analysis, Logit Analysis, and Genetic Algorithms", Turku Centre for Computer Science, (1996)

[4] Chiang L. H., and Pell R. J, "Genetic Algorithms Combined with Discriminant Analysis for Key Variable Identification", *Journal of Process Control 14*, 143-155 (2004)

[5] Lichman M., "UCI Machine Learning Repository", http://arc-hive.ics.uci.edu/ml], [accessed 25 May 2015]

[6] Ayinla A. S., and Adekunle B. K, "An Overview and Application of Discriminant Analysis in Data", *Journal of Mathematics (IOSR-JM)*, 12-15 (2015)

Rizky Kusumawardani*
Faculty of Mathematics and Natural Sciences
Institut Teknologi Sepuluh Nopember
rizky.kusumawardani@gmail.com

Irhamah
Faculty of Mathematics and Natural Sciences
Institut Teknologi Sepuluh Nopember
irhamahn@yahoo.com

Heri Kuswanto
Faculty of Mathematics and Natural Sciences
Institut Teknologi Sepuluh Nopember
Kuswanto.its@gmail.com

*Corresponding author