Jurnal Barekeng
Vol. 7 No. 2 Hal 47 – 51 (2013)

Barekeng
jurnal ilmu matematika dan terapan

PENGGUNAAN METODE ANALISIS DISKRIMINAN, REGRESI LOGISTIK, NEURAL NETWORK, DAN MARS UNTUK ANALISIS PERMASALAHAN KLASIFIKASI DATA HBAT DAN DATA IRIS

The Applications of Discriminant Analysis, Laogistic Regression, Neural Network, and MARS to analyze data classification of HBAT and IRIS

FERRY KONDO LEMBANG[1], DORTEUS LODEWYIK RAHAKBAUW[2]
[1,2] *Staf Jurusan Matematika Fakultas MIPA Universitas Pattimura*
Jl. Ir. M. Putuhena, Kampus Unpatti, Poka-Ambon
e-mail: ferrykondolembang@gmail.com[1], lodewyik@gmail.com[2]

ABSTRAK

Masalah mendasar dalam permasalahan klasifikasi umumnya berkaitan dengan ketidakstabilan performansi atau kebaikan model mencakup aspek interpretasi model yang didapat dikaitkan dengan hubungan antara prediktor (*input)* dan respon (*output*), serta aspek ketepatan klasifikasi terutama pada objek baru yang tidak dimasukkan dalam pembentukan model (data *testing*). Analisis Diskriminan dan Regresi Logistik merupakan dua model klasik dari beberapa metode statistika yang digunakan untuk mengatasi masalah tersebut. Prinsip dasar kedua model klasik dalam permasalahan klasifikasi adalah adanya asumsi yang harus dipenuhi berkaitan dengan skala pengukuran prediktor, keterkaitan antara prediktor, dan distribusi bersama dari prediktor. Agar asumsi dari model klasik ini tidak menjadi syarat utama dalam masalah klasifikasi maka dikembangkan metode klasifikasi modern yaitu *Neural Network* (NN) dan MARS. Data HBAT dan data IRIS akan digunakan dalam penelitian ini untuk menilai kekonsistenan dan performansi model klasifikasi klasik dan model klasifikasi modern. Hasil empirik menunjukkan bahwa kekonsistenan performansi model klasifikasi klasik lebih baik daripada model klasifikasi modern.

Keywords: Analisis Diskriminan, Regresi Logistik, MARS, *Neural Network* Performansi, ketepatan klasifikasi.

PENDAHULUAN

Secara umum permasalahan utama dalam penelitian metode statistika dengan pendekatan regresi adalah mendapatkan model terbaik. Permasalahan klasifikasi dalam statistika juga memiliki permasalahan yang sama. Namun dalam masalah klasifikasi bukan hanya bagaimana mendapatkan model terbaik, tetapi juga menilai performansi atau kebaikan model mencakup interpretasi model yang didapat dikaitkan dengan hubungan antara variabel prediktor (*input*) dan respon (*output*), serta aspek ketepatan klasifikasi terutama pada objek baru yang tidak dimasukkan dalam pembentukan model (data *testing*). Ada beberapa metode statistika yang dapat digunakan untuk menyelesaikan permasalahan klasifikasi, antara lain Analisis Diskriminan dan Regresi Logistik. Pada Analisis Diskriminan akan diperoleh suatu fungsi linear atau kuadratik yang dikenal dengan fungsi diskriminan yang dapat digunakan untuk mengelompokkan obyek. Sedangkan pada Regresi Logistik akan diperoleh suatu model logistik yang digunakan untuk menjelaskan hubungan antara variabel prediktor dan variabel respon yang bersifat dikotomus, serta untuk mengelompokkan obyek ke dalam salah satu dari dua kategori respon. Kedua metode statistika ini dalam beberapa *literature* klasifikasi sering disebut sebagai model klasik. Metode-metode ini mempunyai beberapa asumsi yang harus dipenuhi berkaitan dengan skala pengukuran prediktor, keterkaitan antara prediktor, dan distribusi bersama dari prediktor.

Salah satu metode klasifikasi yang berkembang dari kelompok *machine learning* adalah *Neural Network* (NN), Model ini tidak mensyaratkan skala pengukuran dan distribusi tertentu dari prediktor atau input dalam terminologi NN. Secara umum, ada dua kelompok besar dalam NN dikaitkan dengan ada tidaknya respon, yaitu *supervised* dan *unsupervised* NN. Dalam kasus analisis klasifikasi ini, NN yang digunakan termasuk dalam

kelompok *supervised* NN, karena proses pembelajarannya (optimasi fungsi) terawasi oleh suatu respon (output klasifikasi). Metode klasifikasi lain yang dikembangkan dari pendekatan nonparametrik, khususnya spline adalah MARS (*Multivariate Adaptive Regression Spline*). Dalam beberapa *literature* klasifikasi, kedua model ini seringkali disebut sebagai bagian dari model klasifikasi modern. Pada penelitian Mulyono (2004) model ANN dan MARS mampu menghilangkan misklasifikasi dan meminimalkan kesalahan pengklasifikasian dan kemudian membandingkan Analisis Diskriminan dengan metode ANN dan MARS.

Tertarik dengan penelitian diatas maka tujuan dari paper ini adalah menerapkan dan membandingkan Analisis Diskriminan, Regresi Logistik, NN dan MARS pada dua data, yaitu data HBAT dan data IRIS dengan melihat performansi keempat metode untuk pengklasifikasian. Keempat metode tersebut akan diaplikasikan dengan menggunakan paket statistika yang menyediakan fasilitas untuk analisis data dengan metode-metode tersebut, khususnya R dan SPSS.

Beberapa metode statistika yang akan dipakai dalam penelitian ini akan dijelaskan secara singkat dalam bagian ini.

1.1 Analisis Diskriminan

Analisis Diskriminan atau *discriminant analysis* adalah bagian dari analisis multivariat yang bertujuan untuk memisahkan beberapa kelompok data yang sudah terkelompokkan dengan cara membentuk fungsi diskriminan. Analisis Diskriminan merupakan suatu alat dalam analisis data, jika variabel tak bebas merupakan kategori (nominal atau ordinal, bersifat kualitatif) sedangkan variabel bebas sebagai prediktor merupakan metrik (interval atau rasio, bersifat kuantitatif) (Supranto, 2004). Fungsi ini selanjutnya dapat juga digunakan untuk memprediksi group dari suatu objek baru yang diamati (Hair, 1996).

1.2 Regresi Logistik

Regresi logistik merupakan metode yang dipergunakan untuk menganalisis hubungan antara variabel respon dan variabel prediktor. Variabel prediktor yang dipergunakan berupa data kategori atau kontinu dan variabel responnya berupa data data dengan skala nominal atau ordinal (Agresti, 1990). Regresi logistik ini dapat dipergunakan untuk pengklasifikasian sejumlah obyek ke dalam beberapa kelompok Variabel respon Y yang bersifat random dan dikotomus, yakni bernilai 1 dengan probabilitas π dan bernilai 0 dengan probabilitas $1-\pi$, disebut sebagai *point-binomial* (Lee, 1998).

Model regresi logistik yang dinyatakan sebagai fungsi x adalah (Hosmer and Lemeshow, 1989) :

$$\pi(x)=\frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}$$

Model regresi logistik dengan lebih dari satu variabel prediktor disebut juga model multivariat (Hosmer and Lemeshow, 1989).

Model regresi logistik dengan k variabel prediktor adalah (Le, 1998) :

$$\pi(x)=\frac{\exp(\beta_0+\beta_1 x_1+\cdots+\beta_k x_k)}{1+\exp(\beta_0+\beta_1 x_1+\cdots+\beta_k x_k)}$$

1.3 MARS

MARS merupakan sebuah pendekatan untuk memodelkan regresi nonparametrik multivariat yang dicetuskan pertama kali oleh Friedman (1991). Model MARS digunakan untuk mengatasi kelemahan RPR yaitu menghasilkan model yang kontinu pada knots. Penentuan knots secara otomatis pada MARS menggunakan algoritma *forward stepwise* dan *backward stepwise* yang didasarkan pada nilai GCV minimum. Model MARS dapat ditulis seperti pada persamaan (1) berikut (Friedman, 1991).

$$\hat{f}(x)=a_0+\sum_{m=1}^{M}a_m\prod_{k=1}^{K_m}\left[s_{km}\left(x_{v,(k,m)}-t_{km}\right)\right]$$

dimana

a_0 = parameter fungsi basis induk

a_m = parameter dari fungsi basis ke-m

M = jumlah fungsi basis

K_m = derajat interaksi

s_{km} = nilainya ± 1

$x_{v(k,m)}$ = variabel prediktor

t_{km} = nilai knots dari variabel prediktor $x_{v(k,m)}$

1.4 Neural Network (NN)

Jaringan syaraf tiruan (*Neural Network*) adalah suatu sistem komputasi dengan arsitektur dan operasinya diilhami oleh pengetahuan tentang sel syaraf biologis di dalam otak. Arsitektur ini merupakan salah satu representasi buatan dari otak manusia yang selalu mencoba menstimulasi proses pembelajaran pada otak manusia tersebut. *Neural Network* dapat digambarkan sebagai model matematika dan komputasi untuk fungsi aproksimasi *non-linier*, klasifikasi data *cluster* dan regresi nonparametrik atau sebuah simulasi dari koleksi model Syaraf biologis (Hermawan, 2006).

1.5 Stepwise pada pengklasifikasian

Metode stepwise merupakan salah satu solusi menyelesaikan masalah pengklasifikasian untuk variabel prediktor yang saling berkorelasi. Tidak semua variabel prediktor yang diduga memiliki pengaruh terhadap variabel respon. Salah satu variabel kadang berkorelasi atau berhubungan dengan variabel prediktor yang lain. Oleh karena itu, cukup memasukkan salah satu variabel ke dalam model karena variabel tersebut dianggap sudah mewakili variabel lainnya. Pemilihan variabel yang akan yang akan dimasukkan ke dalam model tentu berdasarkan kriteria-kriteria tertentu, misalnya variabel prediktor yang memiliki korelasi parsial paling besar dengan variabel responnya (Iriawan, 2006). Prosedur stepwise dibuat agar menghasilkan model terbaik. Dalam metode stepwise,

variabel dibuang dan ditambahkan ke model untuk membuat model terbaik

METODOLOGI

Di dalam penelitian ini digunakan dua data sekunder antara lain, Data HBAT dan Data IRIS. Pada data HBAT dipilih variabel prediktor persepsi pelanggan HBAT pada kualitas produk (X_6), aktifitas E-Commerce (X_7), *technical support* (X_8), *Complaint resolution* (X_9), *advertising* (X_{10}), *product line* (X_{11}), *salesforce image* (X_{12}), kompetisi harga (X_{13}), Garansi dan klaim (X_{14}), produk baru (X_{15}), pemesanan dan penagihan (X_{16}), fleksibilitas harga (X_{17}) serta kecepatan pengiriman (X_{18}) dan variabel responnya yaitu *customer intention* yang menjelaskan kemauan pelanggan HBAT meneruskan kerjasama pada waktu yang akan datang (X_{23}). Sedangkan untuk data IRIS terdiri dari empat variabel prediktor, antara lain panjang dan lebar dari PETAL dan SEPAL bunga IRIS, serta kelompok bunga SETOSA, VERSICOLOR, dan VIRGINICA sebagai variabel responnya.

Analisis yang akan dilakukan pada data penelitian ini adalah dilakukan pemodelan dengan menggunakan teknik analisis data antara lain, analisis diskriminan, regresi logistik, MARS dan NN. Dua cara pemodelan yang dilakukan adalah pemodelan dengan semua prediktor (*input*) dan pemodelan dengan prediktor terbaik menggunakan teknik-teknik analisis pada setiap metode. Performansi atau kebaikan mencakup aspek interpretasi model yang didapat dikaitkan dengan hubungan antara prediktor (*input)* dan respon (*output*), serta aspek ketepatan klasifikasi terutama pada objek baru yang tidak dimasukkan dalam pembentukan model (data *testing*). Sebagai langkah awal sebelum dilakukan pemodelan, data dikelompokkan menjadi dua bagian antara lain data untuk pemodelan (*training*) dan evaluasi (*testing*) dengan deskripsi perbandingan data *training* dan *testing* adalah 50:50, 60:40, 70:30, 80:20, dan 40:60.

HASIL DAN PEMBAHASAN

Analisis data pada paper ini dilakukan dengan metode analisis diskriminan, regresi logistik, MARS dan NN menggunakan software SPSS, Minitab, Mars 2,0, dan OSR.

1.6 Analisis data HBAT

Tabel 1. Hasil ketepatan klasifikasi perbandingan training dan testing data HBAT

Data (80:20)								
Konsep yang dianalisis	AD		Regresi Logistik		MARS		NN	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
Ketepatan	78,8%	85%	80%	80%	75%	80%	80%	80%
Data (70:30)								
Ketepatan	85,7%	70 %	85,7%	66,67%	75%	70%	87,14%	66,67%
Data (60:40)								
Ketepatan	86,7%	62,50%	85%	67,50%	86,67%	70%	88,33%	67,5%
Data (50:50)								
Ketepatan	88%	62%	90%	66%	88%	72%	88%	66%
Data (40:60)								
Ketepatan	85%	68,33%	87,5%	68,33%	82,50%	71,67%	87,5%	71,67%

Secara umum dalam analisis permasalahan klasifikasi biasanya sebagai langkah awal adalah data sekunder yang dipakai perlu dirandom dan kemudian dibagi dalam dua bagian yaitu untuk data training dan testing dengan deskripsi perbandingan yang telah ditentukan. Analisis permasalahan klasifikasi untuk data HBAT yang telah di-random menggunakan software SPSS, Minitab, Mars 2,0. dan OSR diperoleh perbandingan training dan testing untuk Analisis Diskriminan (AD), Regresi Logistik, MARS, dan *Neural Network* (NN) pada tabel 1.

Interpretasi Hasil :

Dari tabel diatas dapat dianalisis data HBAT untuk masalah klasifikasi perbandingan training dan testing setelah melalui tahap Kriteria pemilihan model terbaik *stepwise* bahwa konsistensi perbandingan data training lebih besar daripada data testing digunakan pendekatan regresi logistik dan NN artinya untuk pembentukan model lebih bagus menggunakan pendekatan regresi logistik dan NN , namun untuk validasi model kedua metode ini kurang baik. Untuk metode Analisis Diskriminan dan MARS tidak memiliki konsistensi, namun terkhusus untuk perbandingan data terbesar pada perbandingan data (80:20) validasi model dengan mengggunakan kedua metode ini sangat baik.

Hasil Ketepatan klasifikasi perbandingan data training dan testing diatas didapat dari model terbaik dengan menggunakan kriteria pemilihan model terbaik *stepwise*. Dari metode stepwise untuk data random didapat variabel prediktor yang mempengaruhi adalah *Complaint resolution* (X_9) dan *salesforce image* (X_{12}). Data random yang biasanya dipilih sebagai data acuan untuk perbandingan data yang lain adalah data yang memiliki perbandingan data training terbesar daripada data testing, dalam hal ini adalah perbandingan data (80:20).

1.7 Analisis data IRIS

Dengan analisis yang sama dengan data HBAT, permasalahan klasifikasi untuk data IRIS menggunakan Analisis Diskriminan, Regresi Logistik, MARS, dan *Neural Network* (NN) dengan perbandingan data training dan data testing yang telah di-random. Adapun variabel prediktor yang mempengaruhi untuk perbandingan data training dan testing dianalisis menggunakan kriteria pemilihan model terbaik *stepwise* sama halnya dengan data HBAT. Dari hasil analisis menggunakan software SPSS, MINITAB, MARS 2.0, dan OSR didapat hasil ketepatan klasifikasi data training dan data testing yang ditujukkan pada tabel 2 dibawah ini :

Tabel 2. Hasil ketepatan klasifikasi perbandingan training dan testing data IRIS

Data (80:20) n1 = 120, n2 = 30								
Konsep yang dianalisis	AD		Regresi Logistik		MARS		NN	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
Ketepatan	99,2%	96,67%	100%	error	95,83%	90%	99,17%	96,67%
Data (70:30) n1 = 105, n2 = 45								
Ketepatan	99%	95,5 %	100%	91,11%	100%	91,1%	98,09%	91,11%
Data (60:40) n1 = 90, n2 = 60								
Ketepatan	98,9%	98,3%	100%	93,33%	96,67%	95%	97,78%	91,67%
Data (50:50) n1 = 75, n2 = 75								
Ketepatan	98,7%	98,67%	100%	94,67%	100%	93,33%	97,33%	97,33%
Data (40:60) n1 = 60, n2 = 90								
Ketepatan	96,7%	95,56%	100%	95,55%	96,67%	95,56%	96,67%	93,33%

Interpretasi Hasil :

Dari tabel diatas dapat dianalisis data IRIS untuk masalah klasifikasi perbandingan training dan testing setelah melalui tahap Kriteria pemilihan model terbaik *stepwise* bahwa konsistensi perbandingan data training lebih besar daripada data testing dapat disimpulkan baik Analisis Diskriminan, Regresi Logistik, MARS, dan NN memenuhi syarat tersebut. Berarti bisa disimpulkan bahwa model yang didapat dapat dikategorikan sebagai model yang paling baik, diperkuat juga dengan rata-rata nilai training dari masing-masing metode yang didapat lebih besar dari 80 %. Hal ini disebabkan oleh karena dari hasil pemilihan model terbaik yang dilakukan dengan menggunakan metode *stepwise* didapat bahwa semua variabel prediktor dari data IRIS merupakan variabel yang signifikan atau mempengaruhi. Khusus untuk metode regresi logistik nilai ketepatan untuk data training semuanya mencapai 100%. Sehingga apabila dibandingkan keempat metode tersebut berarti metode yang sangat baik dalam menghasilkan model terbaik adalah dengan pendekatan regresi logistik.

Untuk data testing atau validasi model dari keempat metode yang dianalisis bisa disimpulkan juga sangat baik, dengan mengesampingkan data testing untuk metode regresi logistik dimana dari analisis dengan software Minitab diperoleh error. Hal ini disebabkan pada model logit yang didapat baik itu intersep maupun koefisien variabel prediktor menghasilkan nilai yang sangat besar. Perbandingan metode yang paling baik dalam menghasilkan validasi model, dapat terlihat bahwa pendekatan Analisis Diskriminan menjadi pilihan terbaik untuk data IRIS. Dapat diinterpretasikan juga bahwa dari hasil ketepatan klasifikasi data testing pada regresi logistik diperoleh konsistensi linier antara jumlah sampel (nilai n) dengan nilai testing tanpa menyertakan hasil ketepatan data testing pada data (80:20). Artinya semakin besar jumlah sampel yang ada, maka semakin besar pula juga keakuratan validasi model yang didapat. Untuk metode yang lain tidak ditemukan konsistensi linier antara jumlah sampel dengan nilai validasi model yang didapat.

Setelah melakukan analisis dengan menggunakan software SPSS, Minitab, Mars 2,0, dan OSR untuk metode Analisis Diskriminan, Regresi Logistik, MARS, dan *Neural Network* (NN), dari sudut pandang konsistensi terhadap ketepatan klasifikasi untuk data training dan data testing maka terlihat bahwa teknik analisis data dengan menggunakan metode regresi logistik merupakan pendekatan yang paling baik dibandingkan ketiga metode yang lain.

Tabel 3. Pengelompokkan Kekonsistensian Metode untuk data HBAT dan data IRIS

Teknik Analisis Data	**Data HBAT** **Kekonsistenan**	**Data IRIS** **Kekonsistenan**
Analisis Diskriminan	Tidak konsisten	Konsisten
Regresi Logistik	Konsisten	Konsisten
MARS	Tidak Konsisten	Konsisten
Neural Network	Konsisten	Konsisten

Pengelompokkan ketepatan klasifikasi untuk dua kategori konsisten dan tidak konsisten pada data HBAT dan data IRIS menggunakan teknik analisis data antara lain Analisis Diskriminan, Regresi Logistik, MARS, dan *Neural Network* (NN) dapat disimpulkan pada Tabel 3.

Interpretasi hasil:

Dari hasil yang didapat dan dikelompokkan kedalam table 3 diatas terlihat bahwa regresi logistik dan *Neural Network* (NN) memiliki konsistensi performansi dalam mendapatkan ketepatan klasifikasi training untuk data HBAT dan data IRIS. Artinya model yang dihasilkan dikatakan model terbaik. Ketepatan klasifikasi untuk kedua metode ini dikatakan baik karena nilai yang didapat melebihi kriteria yang ditentukan yaitu diatas 80 %. Bahkan berdasarkan analisa software SPSS, metode regresi logistik pada data IRIS mencapai 100% untuk keseluruhan perbandingan training dan testing yang ditentukan. Model Analisis Diskriminan dan MARS tidak memiliki kestabilan performansi pada data HBAT, namun untuk data IRIS kestabilan performansi model disimpulkan stabil karena ketepatan klasifikasi training pada seluruh perbandingan yang telah ditentukan lebih besar daripada testing. Berarti dikatakan bahwa model yang didapat baik.

KESIMPULAN

Berdasarkan hasil analisis dan pembahasan dapat disimpulkan bahwa untuk permasalahan ketepatan klasifikasi untuk pembentukan model (*training*) dan evaluasi atau validasi (*testing*) menggunakan metode Analisis Diskriminan, Regresi Logistik, MARS, dan *Neural Network* (NN) pada data HBAT dan data IRIS disimpulkan bahwa dengan pendekatan regresi logistik dan *Neural Network* (NN) memiliki kekonsistenan dalam pembentukan model dan evaluasi model dimana pada data IRIS nilai training regresi logistic mencapai 100%. Disamping itu metode regresi logistik memiliki konsistensi linier antara jumlah sampel (n buah) dengan nilai testing yang diperoleh, dimana semakin banyak n pada data testing maka semakin besar pula nilai ketepatan klasifikasi testing atau evaluasi model. Untuk ketiga metode yang lain, masing-masing memiliki variasi kekonsistenan ketepatan model pada data sekunder HBAT dan IRIS.

DAFTAR PUSTAKA

Agresti, A., 1990, *Categorical Data Analysis,* John Wiley and Sons, Inc, New York, USA

Friedman, J.H., 1991, *Multivariate Adaptive Regression Splines*, *Tech Report 102 Rev,* Department of Statistics Stanford University Stanford, California.

Hair, J.F., Anderson, R.E, Black, W.C., Babin, B.J., and Tatham,R.L, 2006, *Multivariate Data Analysis*, Sixth edition, Prentice Hall International, UK.

Hermawan, A., 2006, *Jaringan Syaraf Tiruan Teori dan Aplikasi*. ANDI Yogyakarta.

Le, C. T., 1998**,** *Applied Categorical Data Analysis*, John Wiley and Sons, Inc, New York. USA

Mulyono, A., 2004, A*nalisis Diskriminan dengan Metode Fisher, Metode Artificial Neural Networks (ANN) dan Metode Multivariate Adaptive Regression Spline (MARS)*, http://adln.lib.unair.ac.id/go.php?id=jiptunair-gdl-s2-2004-mulyonoagu-1279, Rabu, 28 Oktober 2009, pukul 20.43 WIB.

Supranto, J., 2004, *Analisis Multivariat Arti dan Interpretasi*, Rineka Cipta, Jakarta.