ISSN 1978-3787 (Cetak)
ISSN 2615-3505 (Online)

PERBANDINGAN METODE ANALISIS DISKRIMINAN, NEURAL NETWORK, DISKRIMINAN KERNEL, REGRESI LOGISTIC, MARS UNTUK DATA BANGKITAN (KOMBINASI VARIANS, OVERLAP DAN KORELASI)

Rinda Nariswari[1)] & Elok Fitriani Rafikasari [2)]
[1]Bina Nusantara University
[2]IAIN Tulungagung
Email : [1]rinda.nariswari@binus.ac.id & [2]elokfitriani@ymail.com

Abstrak

Metode untuk pengklasifikasian data diantaranya menggunakan analisis diskriminan, analisis diskriminan kernel, analisis regresi logistik, neural network, dan MARS. Secara keseluruhan masing-masing metode jika diterapkan pada data mempunyai kelebihan maupun kekurangan. Pada pengelompokan data iris virginica dan vercicolor, metode MARS dan NN FeedForward paling baik digunakan. Sedangkan pada pengelompokan data iris setosa dan vercicolor, metode Analisis Diskriminan, NN RBF dan NN FeedForward adalah metode yang paling baik digunakan dalam pengelompokan. Namun berbeda dengan hasil analisis data simulasi yang dibangkitkan melalui Minitab, metode MARS adalah satu-satunya metode yang paling baik digunakan untuk data simulasi karena mempunyai rata-rata ketepatan klasifikasi yang paling besar diantara metode lainnya.

Kata kunci : Analisis Diskriminan, Analisis Diskriminan Kernel, Analisis Regresi Logistik, Neural Network, dan MARS.

PENDAHUALUAN

Ada beberapa macam metode untuk pengklasifikasian data diantaranya menggunakan analisis diskriminan, analisis diskriminan kernel, analisis regresi logistik, neural network, dan MARS. Dalam paper kali ini akan di bahas mengenai ketepatan klasifikasi dari masing-masing analisis, dimana data yang digunakan adalah data iris dengan criteria tertentu dan data bangkitan atau simulasi dengan criteria tertentu pula.

Pembagaian data *Training* dan *Testing* dengan perbandingan yang pertama digunakan 60:40 dan yang kedua digunakan pembagian sebanyak 70:30 yang dilakukan pada data bangkitan dan data iris. Dalam data bangkitan digunakan dua variabel bebas dan satu variabel respon dengan dua jenis respon, sedangkan pada data iris digunakan empat variabel bebas dan satu variabel respon dengan dua jenis respon.

TINJAUAN PUSTAKA

A. Analisis Diskriminan

Asumsi yang harus dipenuhi dalam melakukan analisis diskriminan adalah asumsi distribusi multivariat normal dan Homogenitas Matrik Varians-Kovarians.

Uji Asumsi Distribusi Multivariat Normal

Dalam melakukan analisis data multivariat, asumsi distribusi multivariat normal harus terpenuhi. Pengujian dilakukan dengan hipotesis sebagai berikut

H_0: data berdistribusi multivariat normal

H_1 : data tidak berdistribusi multivariat normal.

Adapun untuk melakukan pengujian multivariat normal dengan membuat plot tersebut adalah sebagai berikut (Morrison, 2005).

1. Menghitung d_j^2 yaitu jarak yang dikuadratkan dengan perhitungan sebagai berikut.

$$d_j^2 = \left(X_i - \overline{X}\right)' S^{-1} \left(X_i - \overline{X}\right)$$

dimana S^{-1} adalah invers matrik varian kovarian yang berukuran pxp

ISSN No. 1978-3787 (Cetak)
ISSN 2615-3505 (Online)

$$S_{ij} = \frac{\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)(X_{kj} - \bar{X}_k)}{n-1}$$

dimana :

p = banyak variabel

$j = 1, 2, \ldots, n$

n = banyak pengamatan

2. Mengurutkan nilai d_j^2 dari nilai yang terkecil sampai yang terbesar atau $d_{(1)}^2 \le d_{(2)}^2 \le \ldots \le d_{(n)}^2$
3. Langkah selanjutnya yaitu membuat plot dengan titik koordinat $\left(d_j^2 ; \chi^2_{\left(p, \frac{j-0,5}{n}\right)} \right)$ dimana nilai $\chi^2_{\left(p, \frac{j-0,5}{n}\right)}$ didapatkan dari tabel χ^2

Data berdistribusi multivariat normal jika plot ini membentuk garis lurus (linier) dan jika terdapat kelengkungan menunjukkan penyimpangan dari normalitas. Tolak H_0 atau data tidak berdistribusi multivariat normal jika terdapat kurang dari 50% jarak $d_j^2 \le \chi^2_{(p;0,5)}$

Uji Asumsi Homogenitas Matrik Varians-Kovarians

Asumsi yang harus dipenuhi dalam analisis diskriminan adalah bahwa matrik varians-kovariansnya adalah homogen. Untuk menguji homogenitas matrik varians-kovarians antar variabel independen digunakan statistik uji Box's M. Perumusan hipotesis dan statistik uji Box's M adalah sebagai berikut (Rencher, 2002).

$H_0 : \Sigma_1 = \Sigma_2 = \ldots = \Sigma_k$

H_1 : minimal ada satu kelompok yang berbeda, $\Sigma_i \neq \Sigma_j$ untuk $i \neq j$

Statistik uji Box's M :

$$\chi^2_{hitung} = -2(1-c_1)\left[\frac{1}{2}\sum_{i=1}^{k} v_i \ln|\mathbf{S}_i| - \frac{1}{2}\ln\left|\mathbf{S}_{pool}\right|\left|\sum_{ii=1}^{k} v_i\right|\right]$$

dimana

$$\mathbf{S}_{pool} = \frac{\sum_{i=1}^{k} v_i \mathbf{S}_i}{\sum_{i=1}^{k} v_i} \qquad c_1 = \left[\sum_{i=1}^{k}\frac{1}{v_i} - \frac{1}{\sum_{i=1}^{k} v_i}\right]\left[\frac{2p^2+3p-1}{6(p+1)(k-1)}\right] \qquad v_i = n_i - 1 \quad (4)$$

Terima hipotesis nol yang berarti matriks varians-kovarians bersifat homogen jika $\chi^2_{hitung} \le \chi^2_{\frac{1}{2}(k-1)p(p+1)}$. (2)

Analisis Diskriminan

Analisis diskriminan merupakan suatu analisis dengan tujuan membentuk sejumlah fungsi diskriminan yang dapat digunakan sebagai cara terbaik untuk memisahkan kelompok-kelompok (Johnson dan Wichern, 2007). Fungsi diskriminan merupakan fungsi atau kombinasi linier peubah-peubah asal yang akan menghasilkan cara terbaik dalam pemisahan kelompok-kelompok tersebut. Dalam analisis diskriminan terdapat 2 metode berdasarkan jumlah kategori dari variabel dependennya. Apabila terdapat 2 kategori yang terlibat dalam pengklasifikasian, maka disebut dengan *two-group discriminant analysis.* Secara umum analisis diskriminan digunakan untuk mengetahui perbedaan yang jelas antar kelompok pada variabel dependen, jika terdapat perbedaan dapat diketahui pula variabel independen manakah pada fungsi diskriminan yang dapat membuat perbedaan. Apabila diketahui 2 populasi yang terdiri dari sampel $y_{11}, y_{12}, \ldots, y_{1n1}$ dan $y_{21}, y_{22}, \ldots, y_{2n2}$. Dengan vektor y_{ij} yang terdiri dari pengukuran sebanyak p variabel, maka didapatkan kombinasi linier $Z = a'y$ yang dibentuk dari setiap vektor skalar.

$Z_{ij} = a'y_{1i} = a_1 y_{1i1} + a_2 y_{1i2} + \ldots + a_p y_{1ip}$ j = 1, 2, …, n_1

$Z_{ij} = a'y_{2i} = a_1 y_{2i1} + a_2 y_{2i2} + \ldots + a_p y_{2ip}$ j = 1, 2, …, n_2

Maka vektor pengamatan untuk 2 sampel $n_1 + n_2$

$$\begin{matrix} y_{11} & y_{21} \\ y_{12} & y_{22} \\ \vdots & \vdots \\ y_{1n1} & y_{2n2} \end{matrix} \quad (3)$$

yang dalam bentuk skalar dapat dituliskan sebagai berikut.

$$\begin{matrix} z_{11} & z_{21} \\ z_{12} & z_{22} \\ \vdots & \vdots \\ z_{1n1} & z_{2n2} \end{matrix}$$

Dari bentuk skalar di atas didapatkan rata-rata $\bar{z} = \sum_{j=1}^{n_1} z_1 / n_1 = a' \bar{y}_1$ dan $\bar{z} = \sum_{j=1}^{n_2} z_2 / n_2 = a' \bar{y}_2$, selisih vektor rata-rata maksimum adalah $(\bar{z}_1 - \bar{z}_2)/s_z$. Karena $(\bar{z}_1 - \bar{z}_2)/s_z$ dapat bernilai negatif, maka digunakan jarak kuadrat $(\bar{z}_1 - \bar{z}_2)^2 / s_z^2$ dimana

$$(\bar{z}_1 - \bar{z}_2)^2 / s_z^2 = \frac{[a'(\bar{y}_1 - \bar{y}_2)]^2}{a' S_{pooled}\, a}$$

Nilai maksimum dari persamaan (7) dapat digunakan ketika

$$a = S_{pooled}^{-1}(\bar{y}_1 - \bar{y}_2)$$

Karena jarak kuadrat $(\bar{z}_1 - \bar{z}_2)^2 / s_z^2$ ekivalen terhadap jarak normal $\bar{y}_1$ dan $\bar{y}_2$. Maka diperoleh

$$(\bar{z}_1 - \bar{z}_2)^2 / s_z^2 = (\bar{y}_1 - \bar{y}_2)' S_{pooled}^{-1}(\bar{y}_1 - \bar{y}_2)$$

untuk $Z = a' y$ dengan $a = S_{pooled}^{-1}(\bar{y}_1 - \bar{y}_2)$.

B. Neural Networks

Cabang ilmu kecerdasan buatan cukup luas, dan erat kaitannya dengan disiplin ilmu yang lainnya. Hal ini bisa dilihat dari berbagai aplikasi yang merupakan hasil kombinasi dari berbagai ilmu. Seperti halnya yang ada pada peralatan medis yang berbentuk aplikasi. Sudah berkembang bahwa aplikasi yang dibuat merupakan hasil perpaduan dari ilmu kecerdasan buatan dan juga ilmu kedokteran atau lebih khusus lagi yaitu ilmu biologi.

Neural Network merupakan kategori ilmu Soft Computing. Neural Network sebenarnya mengadopsi dari kemampuan otak manusia yang mampu memberikan stimulasi/rangsangan, melakukan proses, dan memberikan output. Output diperoleh dari variasi stimulasi dan proses yang terjadi di dalam otak manusia. Kemampuan manusia dalam memproses informasi merupakan hasil kompleksitas proses di dalam otak. Misalnya, yang terjadi pada anak-anak, mereka mampu belajar untuk melakukan pengenalan meskipun mereka tidak mengetahui algoritma apa yang digunakan. Kekuatan komputasi yang luar biasa dari otak manusia ini merupakan sebuah keunggulan di dalam kajian ilmu pengetahuan.

Fungsi dari Neural Network diantaranya adalah:

1. Pengklasifikasian pola
2. Memetakan pola yang didapat dari input ke dalam pola baru pada output
3. Penyimpan pola yang akan dipanggil kembali
4. Memetakan pola-pola yang sejenis
5. Pengoptimasi permasalahan
6. Prediksi

Metode neural network terdapat dua jenis yaitu Multilayer Nural Network dan Radial Basis Function. Multilayer Nural Network atau feedforward merupakan model jaringan syaraf tiruan dengan layar jamak, seperti halnya model jaringan saraf tiruan lainnya, jaringan untuk memberikan respon yang benar terhadap pola masukan yang serupa (tapi tidak sama) dengan pola yang dipakai selama pelatihan. Radial Basis Function merupakan salah satu bentuk multilayer perceptron yang unsupervised. Arsitektur dari RBF adalah fungsi basis sebagai fungsi aktifasi pada hidden layer dan linier pada output layer. Fungsi radial basis biasanya membutuhkan neuron lebih banyak jika dibandingkan dengan jaringan feedforward. Jaringan ini akan bekerja dengan baik apabila dat input yang diberikan cukup banyak. Input akan diolah oleh fungsi aktivasi bukan merupakan hasil penjumlahan terbobot dari data input, tapi berupa vector jarak antara vector bobot dan vector input yang dikalikan dengan bobot bias.

C. Diskriminan Kernel

Dalam klasifikasi dimana data tidak bisa dipisahkan secara linier, salah satu pendekatan yang dapat dipakai adalah dengan menggunakan metode kernel. Dalam hal ini, data dari input space dipetakan ke ruang baru yang disebut dengan kernel space. Dalam metoda kernel, suatu data *x* di input space dipetakan ke kernel space F

ISSN No. 1978-3787 (Cetak)
ISSN 2615-3505 (Online)

dengan dimensi yang lebih tinggi. Pemakaian fungsi kernel memungkinkan analisis diskriminan linier bekerja secara efisien dalam suatu kernel space berdimensi tinggi yang linier. Dengan pendekatan kernel ini, *Fisher Discriminant Analysis* bisa dikembangkan menjadi *Kernel Discriminant Analysis* (KFD) (Mika *et al*., 1999).

Dalam pendugaan fungsi kepadatan peluang metode kernel menggunakan Kernel Uniform, Normal, Epanechnikov, Biweight, atau Triweight. Namun fungsi kernel yang paling sering digunakan adalah kernel normal (Seber, 1984). Notasi berikut digunakan untuk menjelaskan metode pengklasifikasian :

x vektor berdimensi p berisi variabel kuantitatif dari suatu pengamatan

g sebuah subskrip untuk membedakan grup

n_g jumlah pengamatan dalam grup g

$f_g(\mathbf{x})$ pendugaan fungsi kepadatan peluang berasal dari grup g berdasarkan x

Jarak kuadrat antara dua pengamatan antara dua vektor **x** dan **y** dalam grup g diberikan sebagai berikut :

$$d_g^2(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})^{\mathbf{T}}\mathbf{V_g^{-1}}(\mathbf{x}-\mathbf{y})$$

dimana $\mathbf{V_g}$ mempunyai salah satu bentuk di bawah ini :

$\mathbf{V_g} = \mathbf{S_g}$ matrik kovarian dalam grup g

$\mathbf{V_g} = \mathbf{S_{pooled}}$ matrik kovarian gabungan

Pengklasifikasian dari sebuah vektor pengamatan **x** didasarkan pada estimasi kepadatan dari grup. Metode kernel menggunakan bandwith h dan fungsi kernel K_g untuk mengestimasi kepadatan kelompok g pada setiap vektor pengamatan **x.** Apabila **z** merupakan vektor p-dimensi pada kernel space, maka volume dari p-dimensi yang memenuhi $\mathbf{z^Tz}$=1 adalah (Ansys, 2004)

$$\upsilon_0 = \frac{\pi^{\frac{p}{2}}}{\Gamma\left(\frac{p}{2}+1\right)}$$

dimana Γ merupakan fungsi gamma. Jadi, volume p-dimensi dari grup g yang memenuhi $\left\{\mathbf{z} \mid \mathbf{z}^{\mathbf{T}}\mathbf{V}_g^{-1}\mathbf{z} = r^2\right\}$ adalah

$$\upsilon_h(g) = h^p\left|\mathbf{V}_g\right|^{\frac{1}{2}}\upsilon_0$$

Metode kernel menggunakan salah satu fungsi kepadatan kernel berikut pada grup g.

- Uniform Kernel

$$K_g(\mathbf{z}) = \begin{cases} \dfrac{1}{\upsilon_h(g)} & \text{jika } \mathbf{z}^{\mathbf{T}}\mathbf{V}_g^{-1}\mathbf{z} \le \mathrm{h}^2 \\ 0 & \text{lainnya} \end{cases}$$

- Normal Kernel (mean nol, varian $h^2\mathbf{V}_t$)

$$K_g(\mathbf{z}) = \frac{1}{c_0(g)}\exp\left(-\frac{1}{2h^2}\mathbf{z}^{\mathbf{T}}\mathbf{V}_g^{-1}\mathbf{z}\right)$$

dimana $c_0(g) = (2\pi)^{\frac{p}{2}}h^p\left|\mathbf{V}_g\right|^{\frac{1}{2}}$

- Epanechnikov Kernel

$$K_g(\mathbf{z}) = \begin{cases} c_1(g)\left(1-\dfrac{1}{h^2}\mathbf{z}^{\mathbf{T}}\mathbf{V}_g^{-1}\mathbf{z}\right) & \text{jika } \mathbf{z}^{\mathbf{T}}\mathbf{V}_g^{-1}\mathbf{z} \le h^2 \\ 0 & \text{lainnya} \end{cases}$$

dimana $c_1(g) = \frac{1}{\upsilon_h(g)}\left(1+\frac{p}{2}\right)$

- Biweight Kernel

$$K_g(\mathbf{z}) = \begin{cases} c_2(g)\left(1-\dfrac{1}{h^2}\mathbf{z}^{\mathbf{T}}\mathbf{V}_g^{-1}\mathbf{z}\right)^2 & \text{jika } \mathbf{z}^{\mathbf{T}}\mathbf{V}_g^{-1}\mathbf{z} \le h^2 \\ 0 & \text{lainnya} \end{cases}$$

dimana $c_2(g) = \left(1+\frac{p}{4}\right)c_1(g)$

- Triweight Kernel

$$K_g(\mathbf{z}) = \begin{cases} c_3(g)\left(1-\dfrac{1}{h^2}\mathbf{z}^{\mathbf{T}}\mathbf{V}_g^{-1}\mathbf{z}\right)^3 & \text{jika } \mathbf{z}^{\mathbf{T}}\mathbf{V}_g^{-1}\mathbf{z} \le h^2 \\ 0 & \text{lainnya} \end{cases}$$

dimana $c_3(g) = \left(1+\frac{p}{6}\right)c_2(g)$

Pengklasifikasian dari pengamatan vektor **x** berdasarkan fungsi kepadatan peluang suatu grup tertentu dimana $\mathbf{X_i} = (\mathbf{X_{i1}}, \mathbf{X_{i2}}, \ldots, \mathbf{X_{ip}})^{\mathbf{T}}$ sebagai berikut (Jones & Wand, 1995).

$$f_g(\mathbf{x}) = \frac{1}{n_g}\sum_{i=1}^{n_g}K_g(\mathbf{x}-\mathbf{X}_{ig})$$

Berdasarkan pendugaan fungsi kepadatan peluang ini, maka probabilitas posterior dari grup **x** dapat dihitung. Penghitungan probabilitas posterior $p(\pi_g \mid \mathbf{x})$ menggunakan aturan Bayes berdasarkan probabilitas posterior terbesar (Johnson dan Wichern, 2007).

$$p(\pi_1 \mid \mathbf{x}) = \frac{p_1 f_1(\mathbf{x})}{p_1 f_1(\mathbf{x}) + p_2 f_2(\mathbf{x})}$$

$$p(\pi_2 \mid \mathbf{x}) = 1 - p(\pi_1 \mid \mathbf{x}) = \frac{p_2 f_2(\mathbf{x})}{p_1 f_1(\mathbf{x}) + p_2 f_2(\mathbf{x})}$$

Jika $p(\pi_1 \mid \mathbf{x}) > p(\pi_2 \mid \mathbf{x})$ maka pengamatan **x** diklasifikasikan ke π_1, demikian pula sebaliknya. Dimana p_1 dan p_2 merupakan probabilitas prior dari grup 1 dan grup 2 yang diperoleh dari :

$p_1 = {n_1}/{n_1 + n_2}$ dan $p_2 = {n_2}/{n_1 + n_2}$

Pada diskriminan kernel, hal yang perlu diperhatikan adalah bagaimana menentukan nilai bandwith h terbaik. Nilai h yang kecil mengakibatkan estimasi kepadatan yang dihasilkan tidak mulus, dan nilai h yang besar membuat estimasi kepadatan semakin mulus. Salah satu cara untuk menentukan bandwith h yang tepat adalah dengan cara meminimalkan *Approximate Mean Integrated Square Error* (AMISE) dari estimasi kepadatan (Rosenblatt, 1956).

$$AMISE = \frac{1}{4} h^g \left(\int_g g^2 K(g) dg \right)^2 \int_x \left(f''(x) \right)^2 dy + \frac{1}{nh} \int_g K(g)^2 dg$$

AMISE merupakan pengembangan dari *Mean Integrated Square Error* (MISE)

$$MISE = \int_x \left\{ E\left(\hat{f}_h(x) \right) - f(x) \right\}^2 dx + \int_x Var\left(\hat{f}_h(x) \right) dx$$

yang merupakan integral dari kuadrat bias dan integral varians. Kriteria untuk meminimalkan MISE dari fungsi kepadatan terduga. Nilai optimal h yang dihasilkan tergantung pada fungsi kepadatan dan kernel. Pemilihan bandwith h dilakukan dengan mengoptimalkan kriteria dengan mengasumsikan bahwa grup g berdistribusi normal dengan matrik varian kovarian $\mathbf{V}_g$. Sehingga nilai bandwith h optimal yang dihasilkan pada grup g yaitu (Ansys, 2004):

$$\left(\frac{A(K_g)}{n_g} \right)^{1/p+4}$$

dimana konstanta optimal $A(K_g)$ tergantung pada kernel K_g. Konstanta $A(K_g)$ dapat diperoleh dari :

$$A(K_g) = \frac{2^{p+1}(p+2)\Gamma(p/2)}{p}$$

Dengan Kernel Uniform

$$A(K_g) = \frac{4}{2p+1}$$

Dengan Kernel Normal

$$A(K_g) = \frac{2^{p+1} p^2 (p+2)(p+4)\Gamma(p/2)}{2p+1}$$

Dengan Kernel Epanechnikov

Untuk memudahkan dalam memahami metode diskriminan kernel maka diberikan ilustrasi tahapan diskriminan kernel untuk 2 grup dengan jumlah n_1=3 dan n_2=4 sebagai berikut.

1. Apabila diketahui suatu pengamatan terdiri dari p = 3 variabel, maka dapat dituliskan dalam bentuk vektor $\mathbf{x} = [x_1, x_2, x_3]^T$. Ingin diketahui pengamatan tersebut berdasarkan metode diskriminan kernel akan diklasifikasikan ke grup 1 atau grup 2.
2. Membagi keseluruhan data berdasarkan respon grup 1 dan grup 2 dimana $\mathbf{X_i} = (\mathbf{X_{i1}}, \mathbf{X_{i2}}, \mathbf{X_{i3}})^T$, i=1,2, ... ,n maka apabila data dituliskan dalam bentuk matrik menjadi :

$$\underset{(pxn)}{\mathbf{X_{i1}}} = \begin{bmatrix} x_{11} & x_{12} & x_{31} \\ x_{12} & x_{22} & x_{32} \\ x_{13} & x_{23} & x_{33} \end{bmatrix}$$

$$\underset{(pxn)}{\mathbf{X_{i2}}} = \begin{bmatrix} x_{11} & x_{21} & x_{31} & x_{41} \\ x_{12} & x_{22} & x_{32} & x_{42} \\ x_{13} & x_{23} & x_{33} & x_{43} \end{bmatrix}$$

3. Menghitung nilai $f_g(\mathbf{x}) = \frac{1}{n_g} \sum_{i=1}^{n_g} K_g(\mathbf{x} - \mathbf{X}_{ig})$

$$f_1(\mathbf{x}) = \frac{1}{3}\left[K_1(\mathbf{x} - \mathbf{X}_{11}) + K_1(\mathbf{x} - \mathbf{X}_{21}) + K_1(\mathbf{x} - \mathbf{X}_{31}) \right]$$

ISSN No. 1978-3787 (Cetak)
ISSN 2615-3505 (Online)

$$f_2(\mathbf{x}) = \frac{1}{4}\left[K_2(\mathbf{x}-\mathbf{X}_{12}) + K_2(\mathbf{x}-\mathbf{X}_{22}) + K_2(\mathbf{x}-\mathbf{X}_{32}) + K_2(\mathbf{x}-\mathbf{X}_{42}) \right]$$

Kemudian dilanjutkan dengan :

a. menghitung nilai masing-masing $\mathbf{z} = \mathbf{x} - \mathbf{X}_{ig}$ yang kemudian disubstitusi pada persamaan fungsi kepadatan kernel. Namun sebelumnya harus ditentukan dulu jenis metode kernel serta nilai bandwith optimal yang digunakan.
b. Apabila menggunakan bandwith sama untuk semua grup maka matrik varian kovarian $\mathbf{V_g}$ merupakan matrik kovarians gabungan ($\mathbf{S_{pooled}}$). Sedangkan apabila ingin setiap grup memiliki bandwith berbeda maka digunakan $\mathbf{V_g = S_g}$.
c. Misal digunakan metode kernel normal maka **z** disubstitusi pada persamaan :

$$K_g(\mathbf{z}) = \frac{1}{c_0(g)} \exp\left(-\frac{1}{2h^2} \mathbf{z}^{\mathrm{T}} \mathbf{V}_g^{-1} \mathbf{z} \right)$$

$$c_0(g) = (2\pi)^{\frac{p}{2}} h^p \left| \mathbf{V}_g \right|^{\frac{1}{2}}$$

4. Nilai $f_1(\mathbf{x})$ dan $f_2(\mathbf{x})$ yang dihasilkan kemudian digunakan untuk menghitung probabilitas posterior dari masing-masing grup dengan persamaan (2.12). Jika $p(\pi_1 \mid \mathbf{x}) > p(\pi_2 \mid \mathbf{x})$ maka pengamatan **x** diklasifikasikan ke grup 1, demikian pula sebaliknya.

D. Regresi Logistik

Regresi Logistik adalah suatu analisis regresi yang digunakan untuk menggambarkan hubungan antara variabel respon dengan sekumpulan variabel prediktor dimana variabel respon bersifat biner atau dikotomus. Regresi logistik Biner digunakan saat variabel dependen merupakan variabel dikotomus (kategorik dengan 2 macam kagegori),. Regresi Logistik tidak memodelkan secara langsung variabel dependen (Y) dengan variabel independen (X), melainkan melalui transformasi variabel dependen ke variabel logit yang merupakan natural log dari odds rasio (Fractal, 2003).

Dalam penerapannya, regresi logistik tidak memerlukan asumsi multivariat normal atau kesamaan matrik varian kovarian seperti halnya analisis diskriminan (Hosmer dan Lemeshow, 1989). Oleh karena itu metode ini cukup tahan (*robust*) untuk dapat diterapkan dalam berbagai skala/keadaan data (Tatham et. al, 1998). Model regresi logistik multivariate dengan k variabel prediktor adalah :

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}$$

Apabila model persamaan (3) ditransformasi dengan transformasi logit, akan diperoleh bentuk logit :

$$g(x) = \beta_0 + \beta_1 x_1 + \ldots\ldots + \beta_k x_k$$

dengan

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)} \right)$$

Penelitian selama ini menunjukkan apabila tidak ada data kategori, regresi logistik dapat digunakan apabila ada pelanggaran asumsi, dan apabila tidak ada pelanggaran asumsi maka analisis diskriminan sebaiknya digunakan, karena analisis diskriminan memberikan perhitungan yang lebih efisien (Sharma (1996), dan Efron (1975)). Selain itu perbandingan kedua metode ini juga telah dilakukan oleh Press dan Wilson (1978) dan Krzanowski (1975). Kedua penelitian tersebut menunjukkan bahwa untuk data dengan skala campuran (*mixed*), antara diskret dan kontinu, regresi logistik memberikan ketepatan klasifikasi yang lebih baik dari pada analisis diskriminan.

E. MARS

Pada tahun 1991, Jerome H. Friedman, seorang ahli statistik terkenal di dunia dan merupakan salah satu *co-developer* dari CART memperkenalkan suatu metode yang relatif baru, inovatif dan fleksibel yang mengotomatiskan pembangunan model-model prediktif akurat untuk variabel-variabel dependen yang kontinu dan biner, metode ini dikenal dengan nama MARS. Model MARS ini berguna untuk mengatasi permasalahan data yang berdimensi tinggi serta untuk menghasilkan prediksi variabel respon yang akurat. Selain itu MARS juga

merupakan pengembangan dari pendekatan *Recursive Partitioning Regression* (RPR) yang masih memiliki kelemahan dimana model yang dihasilkan tidak kontinu pada titik knotnya.

MARS ini digunakan untuk mengurangi efek dari outlier pada model yang terakhir. Beberapa hal yang perlu diperhatikan dalam menggunakan model MARS adalah (1) *Knot*, yaitu apabila suatu garis regresi tidak bisa menjelaskan keseluruhan data maka beberapa garis regresi digunakan untuk menjelaskan seluruh data yang ada dari variabel yang independen. Tempat perubahan pola itulah yang dinamakan knot. Knot ini merupakan akhir dari sebuah garis regresi (*region*) dan awal dari sebuah garis regresi (*region*) yang lain. Di setiap titik knot, diharapkan adanya kontinuitas dari fungsi basis antar satu *region* dengan *region* lainnya. (2) *Basis Function*, yaitu suatu fungsi yang digunakan untuk menjelaskan hubungan antara variabel dependen dan variabel independen. Fungsi basis ini merupakan fungsi parametrik yang didefinisikan pada tiap *region.* Pada umumnya fungsi basis yang dipilih adalah berbentuk polinomial dengan turunan yang kontinu pada setiap titik knot. Friedman menyarankan jumlah maksimum fungsi basis (BF) adalah 2 sampai dengan 4 kali jumlah variabel prediktornya. Sedangkan untuk jumlah maksimum interaksi (MI) adalah 1, 2 dan 3 dengan pertimbangan jika lebih dari 3 akan menghasilkan bentuk model yang semakin kompleks. Minimum jarak antara knots atau minimum observasi antara knots sebesar 0, 10, 25, 50 atau 100. Banyaknya variabel prediktor yang diduga berpengaruh terhadap curah hujan sebanyak tiga variabel, maka banyaknya fungsi basis (BF) yang digunakan dalam pembentukan variabel adalah 9. Banyaknya fungsi basis ini diperoleh dari tiga kali dari variabel prediktor dengan maksimum interaksi (MI) sebesar 2. Untuk observasi minimum di bawah *knot* (MO) digunakan 0, 1 dan 2, karena di atas nilai itu GVC semakin besar

METODE PENELITIAN

Analisis yang dilakukan terhadap data bangkitan dan data iris. Data bangkitan meliputi kombinasi data yang varian sama, varian tidak sama, overlap, tidak terjadi overlap, korelasi dan tidak terjadi korealasi dengan jumlah data sebanyak 100 data yang terediri dua variabel bebas dan satu variabel respon dengan katagori variabel respon 1 dan 2. Dalam pengolahan data digunakan data *Testing* dan data *Training* yang terbagi menjadi 2 bagian, bagian pertama digunakan sebagai data *Training* dan satu bagian digunakan sebagai *Testing*. Pembagaian data *Training* dan *Testing* dengan perbandingan yang pertama digunakan 60:40 dan yang kedua digunakan pembagian sebanyak 70:30. Kombinasi untuk data bangkitan dengan keterangan sebagai berikut:

Data1= var sama,overlap dan tidak ada korelasi
Data2= var sama,tidak overlap dan korelasi
Data3= var sama,tidak overlap dan tidak ada korelasi
Data4= var tidak sama,overlap dan tidak ada korelasi
Data5= var tidak sama, tidak ada overlap dan korelasi
Data6= var tidak sama, tidak ada overlap dan tidak ada korelasi

Data selanjutnya dianalisis menggunakan Analisis Diskriminan, Neural Network, Diskriminan Kernel, Regresi Logistic dan MARS. Dari hasil pengolahan semua metode akan dilihat metode mana yang paling cocok digunakan berdasarkan ketepatan prediksi yang telah diperoleh.

HASIL DAN PEMBAHASAN

Analisis yang digunakan dalam penelitian ini yaitu dengan pendekatan metode Analisis Diskriminan, Neural Network, Diskriminan Kernel, Regresi Logistic dan MARS. Hasil pengolahan untuk masing-masing metode adalah sebagai berikut:

Analisis Diskriminan

Asumsi data berdistribusi Multivariat Normal dan asumi Homogenitas Matrik Varians-

ISSN No. 1978-3787 (Cetak)
ISSN 2615-3505 (Online)

Kovarians telah terpenuhi sehingga tahapan selanjutnya dapat dilakukan proses analisis diskriminan. Hasil dari pengolahan analisis diskriminan dengan perbandingan 60(*Training*) – 40 (*Testing)* pada data bangkitan adalah sebagai berikut:

Tabel 1 Hasil analisis diskriminan 60:40

Data	Ketepatan Klasifikasi	
	Data *Training*	Data *Testing*
Data1	85%	92.50%
Data2	100%	100%
Data3	100%	100%
Data4	91.67%	100%
Data5	100%	100%
Data6	100%	55%
Rata-rata	96%	91.25%

Dari tabel hasil pengolahan di atas didapatkan bahwa ketepatan klasifikasi yang paling tinggi pada data *Training* daripada data *Testing* yaitu sebesar 96%.

Hasil dari pengolahan analisis diskriminan dengan perbandingan 70(*Training*) – 30 (*Testing)* pada data bangkitan adalah sebagai berikut:

Tabel 2. Hasil analisis diskriminan 70-30

Data	Ketepatan Klasifikasi	
	Data *Training*	Data *Testing*
Data1	87,14%	96,67%
Data2	100%	100%
Data3	98,57%	100%
Data4	92,86%	86,67%
Data5	100%	100%
Data6	100%	100%
Rata-rata	**96,43%**	**97,22%**

Dari tabel hasil pengolahan di atas didapatkan bahwa ketepatan klasifikasi yang paling tinggi pada data *Testing* ketimbang data *Training* sebesar 97,22%.

Neural Network

Data bangkitan pada analisis neural network ini keterangan data nya sama hal nya dengan analisis pada diskriminan. Hasil pengolahan pada neural network adalah sebagai berikut:

NN Radial basis function

Tabel 3. Hasil analisis neural network Radial basis function 60-40

Data	Ketepatan Klasifikasi	
	Data *Training*	Data *Testing*
Data1	90.3%	86.8%
Data2	100%	100%
Data3	100%	100%
Data4	89.2%	100%
Data5	100%	100%
Data6	100%	100%
Rata-rata	96.58%	97.80%

Pada table 5 di atas tampak bahwa ketepatan klasifikasi yang paling tinggi pada data *Testing* yaitu sebesar 97.80%

Tabel 4. Hasil analisis neural network Radial basis function 70-30

Data	Ketepatan Klasifikasi	
	Data *Training*	Data *Testing*
Data1	93.2%	88.9%
Data2	100%	100%
Data3	98.6%	100%
Data4	91.2%	87.5%
Data5	100%	100%
Data6	100%	100%
Rata-rata	97.17%	96.07%

Pada table 4 di atas tampak bahwa rata-rata ketepatan klasifikasi yang paling tinggi pada data *Training* yaitu sebesar 97.17%.

NN Feedforward

Tabel 5. Hasil analisis neural network Feedforward 60-40

Data	Ketepatan Klasifikasi	
	Data *Training*	Data *Testing*
Data1	89.7%	95.2%
Data2	100%	100%
Data3	97.9%	98.1%
Data4	91.2%	93.8%
Data5	100%	100%
Data6	100%	100%
Rata-rata	96.47%	97.85%

Pada table 5 di atas tampak bahwa rata-rata ketepatan klasifikasi yang paling tinggi pada data *Testing* yaitu sebesar 97.85%

Tabel 6. Hasil analisis neural network Feedforward 70-30

Data	Ketepatan Klasifikasi	
	Data *Training*	Data *Testing*
Data1	95.2%	86.8%
Data2	100%	100%
Data3	98.7%	96%
Data4	95.7%	93.5%
Data5	100%	100%
Data6	100%	100%
Rata-rata	98.27%	96.05%

Pada table 6 di atas tampak bahwa rata-rata ketepatan klasifikasi yang paling tinggi pada data *Training* yaitu sebesar 98.27%

Diskriminan Kernel

Metode diskriminan kernel merupakan metode yang bergantung pada dua hal, yaitu pemilihan fungsi kernel (K) dan penentuan besarnya bandwith (h). Fungsi kernel terdiri dari tujuh macam yaitu *Uniform, Triangle, Epannechnikov, Quartic/Biweight, Triweight, Gaussian/* Normal, *dan Cosinus* (Jones & Wand, 1995). Dari beberapa macam fungsi kernel tersebut yang paling sering digunakan adalah fungsi kernel *Gaussian*/Normal (Seber, 1984). Oleh karena itu, pada analisis diskriminan kernel ini digunakan fungsi kernel normal.

Sedangkan dalam pemilihan bandwith, akan dipilih berdasarkan penghitungan bandwith optimal dimana konstanta optimal $A(K_t)$ tergantung pada Kernel (K_t). Dengan menggunakan kernel normal ditentukan nilai $A(K_t)$ dengan perhitungan sebagai berikut (Ansys, 2004) :

$$A(K_t) = \frac{4}{2p+1}$$

$$= \frac{4}{2.4+1} = 0{,}447$$

Sehingga didapatkan untuk data *Training* 70

$$h = \left(A(K_t) \Big/ n_t \right)^{1/(p+4)}$$

$$h = \left(0{,}447 \Big/ 70 \right)^{1/(4+4)} = 0{,}5313$$

Sedangkan untuk data *Training* 60

$$h = \left(A(K_t) \Big/ n_t \right)^{1/(p+4)}$$

$$h = \left(0{,}447 \Big/ 60 \right)^{1/(4+4)} = 0{,}5416$$

Dimana p merupakan dimensi data yaitu 4 dan n_t merupakan banyaknya data *Training* yaitu sebanyak 70.

Berikut adalah metode diskriminan kernel yang diterapkan pada data bangkitan. Dengan menggunakan kernel normal ditentukan nilai $A(K_t)$ dengan perhitungan sebagai berikut (Ansys, 2004):

$$A(K_t) = \frac{4}{2p+1} = \frac{4}{2.2+1} = 0{,}8$$

Sehingga didapatkan untuk data *Training* 70

$$h = \left(A(K_t) \Big/ n_t \right)^{1/(p+4)}$$

$$h = \left(0{,}8 \Big/ 70 \right)^{1/(2+4)} = 0{,}4765 = 0{,}5$$

Sedangkan untuk data *Training* 60

$$h = \left(A(K_t) \Big/ n_t \right)^{1/(p+4)}$$

$$h = \left(0{,}8 \Big/ 60 \right)^{1/(2+4)} = 0{,}4869 = 0{,}5$$

Tabel 7. Hasil analisis Diskriminan Kernel Data Bangkitan 60-40

Data	Ketepatan Klasifikasi	
	Data *Training*	Data *Testing*
Data1	86,67%	97,5%
Data2	93,3%	97,5%
Data3	90%	92,5%
Data4	95,7%	97,5%
Data5	93,3%	92,5%
Data6	97,1%	100%
Rata-rata	92.6%	96.2%

ISSN No. 1978-3787 (Cetak)
ISSN 2615-3505 (Online)

Tabel 8. Hasil analisis Diskriminan Kernel Data Bangkitan 70-30

Data	Ketepatan Klasifikasi	
	Data *Training*	Data *Testing*
Data1	95,7%	90%
Data2	92,8%	93,3%
Data3	90%	93,3%
Data4	97,1%	98,3%
Data5	95,7%	93,3%
Data6	98,3%	100%
Rata-rata	94.9%	94.7%

Berdasarkan tabel di atas, data simulasi yang menghasilkan nilai ketepatan klasifikasi tertinggi pada data *Training* maupun *Testing* adalah data dengan kombinasi varians tidak sama, tidak ada overlap, dan tidak ada korelasi, yaitu sebesar 98,3% pada data *Training* dengan proporsi 70% sedangkan pada data *Testing*, untuk kedua proprorsi data yaitu 70:30 dan 60:40 sama – sama menghasilkan nilai ketepatan klasifikasi 100%.

Regresi Logistik

Berikut adalah metode regresi logistik yang diterapkan pada data bangkitan dengan perbandingan 60(*Training*) – 40 (*Testing)* dan 70(*Training*) – 30 (*Testing).*

Tabel 9. Hasil analisis Regresi Logistik Data Bangkitan 60-40

Data	Ketepatan Klasifikasi	
	Data *Training*	Data *Testing*
Data1	86,7%	85%
Data2	90%	95%
Data3	*	*
Data4	96,7%	100%
Data5	91,7%	87,5%
Data6	88,3%	80%
Rata-rata	90.68%	89.5%

Tabel 10. Hasil analisis Regresi Logistik Data Bangkitan 70-30

Data	Ketepatan Klasifikasi	
	Data *Training*	Data *Testing*
Data1	87,1%	86,7%
Data2	90%	90%
Data3	*	*
Data4	97,1%	100%
Data5	91,4%	90%
Data6	85,7%	90%
Rata-rata	90.26	91.34%

* Tidak dapat dicari ketepatan klasifikasi karena dalam analisis regresi logistik tidak ada variabel prediktor yang signifikan sehingga tidak dapat dibentuk model.

Dari tabel di atas, terlihat bahwa data terbaik yang menghasilkan nilai ketepatan klasifikasi tertinggi pada data *Testing* adalah data dengan kombinasi varians tidak sama, tidak overlap, dan tidak ada korelasi. Ketepatan klasifikasi pada kombinasi data ini menghasilkan nilai 100% baik untuk proporsi data *Training* dan *Testing* 70:30 maupun 60:40.

Selain itu, ada beberapa kombinasi data yang menghasilkan nilai ketepatan klasifikasi pada data *Training* nilainya lebih tinggi dibandingkan nilai ketepatan klasifikasi pada data *Testing*. Namun hal ini belum bisa dikatakan sebagai kondisi *overfitting* karena selisih antara nilai ketepatan klasifikasi pada data *Training* tidak terlalu jeuh berbeda dengan nilai ketepatan klasifikasi pada data *Testing*. *Overfitting* terjadi apabila nilai ketepatan klasifikasi di data *Training* tinggi, namun pada data *Testing* sangat rendah.

MARS

Berikut adalah analisis data bangkitan atau simulasi dengan menggunakan analisis MARS:

Tabel 11. Hasil analisis MARS Data Bangkitan 60-40

Data	Ketepatan Klasifikasi	
	Data *Training*	Data *Testing*
Data1	90.6%	93.5%
Data2	100%	100%
Data3	100%	100%
Data4	100%	100%
Data5	100%	100%
Data6	96.6%	97.5%
Rata-rata	97.8%	98.5%

Tabel 12. Hasil analisis MARS Data Bangkitan 70-30

Data	Ketepatan Klasifikasi	
	Data *Training*	Data *Testing*
Data1	100%	100%
Data2	100%	100%
Data3	100%	100%
Data4	100%	100%
Data5	98.6%	100%
Data6	98.7%	96%
Rata-rata	99.55%	99,33%

Berdasarkan tabel ketepatan klasifikasi di atas,, untuk data *Training-Testing* 60-40 data 2,3,4,5 memperoleh ketepatan klasifikasi sempurna, sedangkan pada data *Training-Testing* 70-30 data 1,2,3,4 memperoleh ketepatan klasifikasi sempurna.

PENUTUP

Kesimpulan

Berdasarkan hasil analisis dan pembahasan didapatkan beberapa kesimpulan sebagai berikut:

Tabel 13. Hasil analisis Data Bangkitan 60-40

Data	Ketepatan Klasifikasi	
	Rata^2 Data *Training*	Rata^2 Data *Testing*
A. Diskriminan	96%	91.25%
NN RBF	96.58%	97.80%
NN Feedforward	96.47%	97.85%
Kernel. Diskriminan	92.6%	96.2%
Regresi Logistik	90.68%	89.5%
MARS	97.8%	98.5%

Dari hasil analisis data bangkitan dengan *Training-Testing* 60-40 dapat disimpulkan metode MARS adalah metode yang paling bagus untuk pengklasifikasian data simulasi ini karena rata-rata ketepatan klasifikasinya paling tinggi diantara yang lain baik untuk data *Training* maupun data *Testing*.

Tabel 14. Hasil analisis Data Bangkitan 70-30

Data	Ketepatan Klasifikasi	
	Rata^2 Data *Training*	Rata^2 Data *Testing*
A. Diskriminan	96,43%	97,22%
NN RBF	97.17%	96.07%
NN Feedforward	98.27%	96.05%
Kernel. Diskriminan	94.9%	94.7%
Regresi Logistik	90.68%	89.5%
MARS	99.55%	99,33%

Dari hasil analisis data bangkitan dengan *Training-Testing* 70-40 dapat disimpulkan metode MARS adalah metode yang paling bagus untuk pengklasifikasian data simulasi ini karena rata-rata ketepatan klasifikasinya paling tinggi diantara yang lain baik untuk data *Training* maupun data *Testing*.

Saran

Pada penelitian selanjutnya dapat digunakan data yang lebih besar agar dapat dibandingkan pula ketepatan klasifikasinya meningkat ataukah sama jika menggunakan data dengan ukuran 100 dengan menggunakan metode-metode analisis dalam penelitian ini.

DAFTAR PUSTAKA

[1] ANSYS , (2004), "ANSYS Modeling and Meshing Guide: ANSYS Release 9,0", ANSYS, Inc.

[2] Efron, B., (1975), "The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis". Journal of the American Statisitical Association, 70;892-898.

[3] Fractal , (2003), Comparative Analysis of Classification Techniques, A Fractal White Paper.

[4] Hair, J. F. Jr. ; Rolph E.A; Tatham R. L.,(1998), Multivariate Data Analysis. Fifth Edition. New York. Macmillan College Pub. Co.

ISSN No. 1978-3787 (Cetak)
ISSN 2615-3505 (Online)

[5] Hosmer, D.W., dan Lemeshow, S., (1989), Applied Logistic Regression. New York: John Wiley & Sons.
[6] Johnson, R,. dan Wichern, D., (2007), Applied Multivariate Statistical Analysis, 6nd edition, Prentice-Hall.
[7] Krzanowski, W..J., (1975), "Discrimination and Classification using Both Binary and Continuous Variable", Journal of the American Statisitical Association, 70;782-352.
[8] Morrison, A. M., 2002, "Hospitality and Travel Marketing", 3rd edition, USA: Delmar.
[9] Muller,K.R., dan Mika,S., et al, 2003. "An Introduction to Kernel-Based Learning Algorithms. IEEE Trans. On Neural Networks", vol. 12, No. 2.
[10] Rencher, A.C., (2002), "Methods of Multivariate Analysis"2 rd edition, John Wiley & Sons Ltd., Chichester, England.
[11] Rosenblatt, M., (1956),"Remarks on Some Nonparametric Estimates of a Density Function". Annals of Mathematical Statistics. 27, 832 -837.
[12] Sharma, S., (1996), Applied Multivariate Techniques, John Wiley & Sons, Inc, New York.
[13] Seber,G.A.F, (1984), Multivariate Observation. John Wiley & Sons Ltd., New York.
[14] Wurm, S.A. and Wilson,B. (1978), English Finderlist of Reconstructions in Austronesian Languages (Post-Brandstetter). Canberra, Australia: The Australian National University.