BAB 2

LANDASAN TEORI

2.1. Regresi Logistik Biner

Regresi logistik biner merupakan salah satu pendekatan model matematis yang digunakan untuk menganalisis hubungan beberapa faktor dengan sebuah variabel yang bersifat dikotomus (biner). Pada regresi logistik jika variabel responnya terdiri dari dua kategori misalnya Y = 1 menyatakan hasil yang diperoleh "sukses" dan Y = 0 menyatakan hasil yang diperoleh "gagal" maka regresi logistik tersebut menggunakan regresi logistik biner. Menurut Agresti variabel *y* yang demikian lebih tepat dikatakan sebagai variabel indikator dan memenuhi distribusi Bernoulli. Fungsi distribusi peluang untuk *y* dengan parameter π_i adalah

$$f(y_i;\pi_i) = \begin{cases} \pi_i(1-\pi_i)^{1-y_i} \; untuk \; y_i = 0,1 \\ 0 \qquad\quad untuk \; y_i \; yang \; lain \end{cases}$$

dengan $\pi_i = P(Y_i = 1)$. Dari fungsi distribusi tersebut diperoleh rata-rata :

$$E(Y) = 1.P(Y = 1) + 0.P(Y = 0)$$
$$= P(Y = 1)$$

Misalkan probabilitas ini dinotasikan sebagai $\pi(x)$ yang bergantung dengan variabel penjelas $\boldsymbol{X} = (X_{1,\dots,}X_k)$ dengan $E(y) = \pi$ dan $0 \leq \pi \leq 1$, sehingga diperoleh

$$E(Y^2) = 1^2\pi(x) + 0^2[1-\pi(x)] = \pi(x)$$

Dan Varians dari *Y* adalah

$$V(Y) = E(Y^2) - [E(Y)]^2 = \pi(x)[1-\pi(x)]$$

Secara umum model probabilitas regresi logistik dengan melibatkan beberapa variabel prediktor (x) dapat diformulasikan sebagai berikut:

$$E(y|x) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}} \quad (1)$$

Dimana $E(y|x)$ merupakan penjumlahan dari $\pi(x)$. Fungsi $\pi(x)$ merupakan fungsi non linear sehingga perlu dilakukan transformasi logit untuk memperoleh fungsi yang linier agar dapat dilihat hubungan antara variabel respon (y)dengan variabel prediktornya (x). Bentuk logit dari $\pi(x)$ dinyatakan sebagai $g(x)$, yaitu:

$$g(x) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) \quad (2)$$

Persamaan (1) dan persamaan (2) disubtitusikan sehingga diperoleh:

$$ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \quad (3)$$

Untuk memperoleh estimasi dari parameter regresi logistik dapat dilakukan dengan dua cara yakni dengan cara *Maximum Likelihood Estimation (MLE)* dan iterasi *Newton Raphson*.

a. *Maximum Likelihood Estimation (MLE)*

Metode MLE digunakan untuk mengestimasi parameter-parameter dalam regresi logistik dan pada dasarnya metode maksimum likelihood memberikan nilai estimasi β dengan memaksimumkan fungsi likelihoodnya. (Hosmer dan Lemeshow, 1989). Secara matematis fungsi likelihood (x_i, y_i) dapat dinyatakan:

$$f(x_i) = \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i} \quad (4)$$

Karena setiap pengamatan diasumsikan independen maka fungsi likelihoodnya merupakan perkalian antara masing-masing fungsi likelihood yaitu:

$$l(\beta) = \prod_{i=1}^{n} f(x_i) \quad (5)$$

dan logaritma likelihoodnya dinyatakan sebagai:

$$L(\beta) = ln[l(\beta)]$$

$$= \sum_{i=1}^{n} \{y_i . ln(\pi(x_i)) + (1 - y_i) ln(1 - \pi(x_i))\} \quad (6)$$

Untuk memperoleh nilai β maka dengan memaksimumkan nilai $L(\beta)$ dan mendiferensialkan $L(\beta)$ terhadap β dan menyamakannya dengan nol. Persamaan ini dapat ditulis dalam bentuk sebagai berikut:

$$\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0 \quad (7)$$

dan persamaan likelihood:

$$\sum_{i=1}^{n} x_i [y_i - \pi(x_i)] = 0 \quad (8)$$

b. Metode Newton Rhapson

Metode Newton Rhapson merupakan metode untuk menyelesaikan persamaan nonlinear seperti menyelesaikan persamaan likelihood dalam model regresi logistik (Agresti, A. 1990). Metode newton rhapson memerlukan taksiran awal untuk nilai fungsi maksimumnya, yang mana fungsi tersebut merupakan taksiran yang menggunakan pendekatan polinomial berderajat dua. Dalam hal ini untuk menentukan nilai $\hat{\beta}$ dari β yang merupakan fungsi maksimum dari $g(\beta)$. Andaikan: $q' = \left(\frac{\partial g}{\delta \beta_1}, \frac{\partial g}{\partial \beta_2}, \cdots\right)$, dan andaikan **H** dinotasikan sebagai matriks yang mempunyai anggota $h_{ab} = \frac{\partial g}{\delta \beta_1 . \partial \beta_2}$. Andaikan $q^{(t)}$dan $\boldsymbol{H}^{(t)}$ merupakan bentuk evaluasi dari $\beta^{(t)}$, taksiran ke *t* pada $\hat{\beta}$. Pada langkah *t* dalam proses iterasi (*t* = 0, 1, 2, ...), $g(\beta)$ ialah pendekatan $\beta^{(t)}$ yang merupakan bentuk orde kedua dari ekspansi deret Taylor,

$Q^{(t)}(\beta) = g\beta^{(t)} + q^{(t)'}\left(\beta - \beta^{(t)}\right) + \left(\frac{1}{2}\right)\left(\beta - \beta^{(t)}\right)' H^{(t)}\left(\beta - \beta^{(t)}\right)$.

Penyelesaian:

$$\frac{\delta Q^{(t)}}{\partial \beta} = q^{(t)} + H^{(t)}\left(\beta - \beta^{(t)}\right) = 0$$

$$\beta^{(t+1)} = \beta^{(t)} - \left(H^{(t)}\right)^{-1} q^{(t)}$$

dengan mengasumsikan $\boldsymbol{H}^{(t)}$ sebagai matriks nonsingular.

2.2. Fungsi Klasifikasi Regresi Logistik

Dalam regresi logistik penglasifikasian dilakukan dengan menghitung *"error rates"* atau probabilitas kesalahan klasifikasi (Johnson *et al*, 2007). Misalkan $f_1(x)$ dan $f_2(x)$ merupakan fungsi kepadatan peluang dengan *p x 1* variabel acak *X*. Dan misalkan Ω ialah ruang sampel yang merupakan semua observasi *x* yang mungkin. Andaikan R_1 merupakan nilai *x* sebagai objek klasifikasi π_1 dan $R_2 = \Omega - R_1$ sebagai objek klasifikasi π_2. Jika setiap objek disimbolkan dengan 1 atau hanya 1 dari 2 populasi maka himpunan R_1 dan R_2 merupakan *mutually exclusive dan exhaustive.* Sehingga probabbilitas kondisional $P(2|1)$ ialah

$$P(2|1) = P(x \epsilon R_2|\pi_1) = \int_{R_2=\Omega-R_1} f_1(x)\ dx$$

Sama halnya dengan $P(1|2)$

$$P(1|2) = P(x \epsilon R_1|\pi_2) = \int_{R_1} f_2(x)\ dx\ .$$

Andaikan P_1 merupakan probabilitas dari π_1 dan P_2 merupakan probabilitas dari π_2. Total probabilitas misklasifikasi (TPM) ialah:

$$TPM = p_1 \int_{R_1} f_1(x)\ dx + p_2 \int_{R_2} f_2(x)\ dx$$

Dalam hal ini untuk menentukan kesalahan klasifikasi dapat digunakan prosedur klasifikasi optimal yang disebut optimum error rate (OER) yaitu:

$$OER = p_1 \int_{R_1} f_1(x)\ dx + p_2 \int_{R_2} f_2(x)\ dx \qquad (9)$$

Dimana R_1: $\frac{f_1(x)}{f_2(x)} \geq \frac{p_2}{p_1}$ dan R_2: $\frac{f_1(x)}{f_2(x)} < \frac{p_2}{p_1}$,

Maka OER ialah error rate untuk aturan klasifikasi minimum TPM.

Dalam hal lain OER dapat dihitung jika fungsi densitas populasi diketahui. Namun , dalam kasus lain populasi parameter harus di estimasikan terlebih dahulu sehingga evaluasi error ratenya menjadi tidak seimbang. Untuk itu sampel fungsi klasifikasinya dapat dihitung dengan menghitung *actual error rate (AER).*

$$AER = p_1 \int_{\widehat{R_1}} f_1(x)\ dx + p_2 \int_{\widehat{R_2}} f_2(x)\ dx$$

AER akan mengindikasikan bagaimana fungsi klasifikasi yang akan diperlihatkan pada sampel berikutnya seperti OER namun tidak dapat menghitung secara umum karena tergantung pada fungsi densitas yang tidak diketahui yaitu $f_1(x)$ dan $f_2(x)$. Sehingga untuk mempermudah perhitungan dalam proses klasifikasi dan tidak bergantung pada distribusi populasi dengan menghitung error rate atau probabilitas kesalahan klasifikasi pada APER (*apperent error rate)* yang merupakan fraksi observasi dalam sampel yang salah diklasifikasikan atau misclassified pada fungsi klasifikasi (Johson *et al*, 2007). Perhitungan APER terlebih dahulu dibuat matriks konfusinya yang diperlihatkan dalam tabel 1 sebelumnya. Sehingga diperoleh:

$$APER = \frac{n_{1B} + n_{2B}}{n_1 + n_2}$$

2.3. Model Logit

Pada umumnya variabel respon data kategorik hanya mempunyai 2 kategorik yaitu sukses dan gagal, ya atau tidak, hidup atau mati dan sebagainya. Hasil observasi untuk setiap objek diklasifikasikan sebagai sukses dan gagal. Untuk sukses dinyatakan dengan 1, gagal dinyatakan dengan 0. Seperti halnya distribusi Bernaulli/Binomial untuk variabel random dengan probabilitas sukses $P(Y = 1) = \pi$ dan gagal $P(Y = 0) = 1 - \pi$ dengan π ialah $E(y)$ dimana y_i berdistribusi binomial dalam parameter π_i dan fungsi padat peluangnya ialah

$$\begin{aligned} f(y_i, \pi_i) &= \pi_i^{y_i}(1 - \pi_i)^{1-y_i} \\ &= (1 - \pi_i)\left[\frac{\pi_i}{1 - \pi_i}\right]^{y_i}, y_i = 0,1 \\ &= 1 - \pi_i exp\left[y_i log\left(\frac{\pi_i}{1 - \pi_i}\right)\right], \pi_i = P(Y_i = 1) \end{aligned}$$

Distribusi ini termasuk dalam exponensial sejati dengan parameter sejatinya ialah

$$Q(\pi) = log\left(\frac{\pi}{1-\pi}\right)$$

dan

$$\frac{\pi}{1-\pi} = odds$$

Dalam regresi logistik untuk variable biner model natural odds rasio disebut logit(π) sehingga

$$logit(\pi) = ln(odds) = ln\left(\frac{\pi}{1-\pi}\right)$$

Fungsi logit merupakan fungsi probabilitas π, jika diasumsikan ke dalam variabel predictor variable *Z* maka

$$logit(\pi) = ln(odds) = ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_{1z}$$

Dengan kata lain log odds merupakan variabel prediktor linear. Jika dimasukkan bentuk logit atau log odds ke dalam probabilitas π diperoleh:

$$ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_{1z}$$

Sehingga dapat ditentukan:

$$\theta(z) = \frac{\pi(z)}{1-\pi(z)} = exp(\beta_0 + \beta_{1z})$$

2.4. Distribusi Binomial

Distribusi Binomial merupakan suatu distribusi probabilitas yang dapat digunakan bilamana suatu proses sampling dapat diasumsikan sesuai dengan proses Bernoulli. Misalnya, dalam perlemparan sekeping uang logam sebanyak 5 kali, hasil setiap ulangan mungkin muncul sisi gambar atau sisi angka. Begitu pula, bila kartu diambil berturut-turut, kita dapat memberi label “berhasil” bila kartu yang terambil adalah kartu merah atau “gagal” bila yang terambil adalah kartu hitam. Ulangan-ulangan

tersebut bersifat bebas dan peluang keberhasilan setiap ulangan tetap sama,yaitu sebasar 0,5 (Cyber-learn, 2011).

Secara umum bentuk distribusi binomial yaitu

$$b(x,n,p) = \binom{n}{x} p^x q^{n-x}$$

Dengan probabilitas sukses p (atau probabilitas gagal q=1-p).

2.5. Deret Taylor

Deret taylor dapat memberikan nilai hampiran bagi suatu fungsi pada suatu titik, berdasarkan nilai fungsi dan turunannya pada titik yang lain.(Kholijah, S. 2008).
Andaikan suatu fungsi $f(x)$ dan turunannya, yaitu $f'(x), f"(x), f'''(x), \ldots, f(n)(x)$ kontinu dalam selang $[a,b]$, dan $x_0 \in [a,b]$, maka untuk nilai *x* disekitar x_0, $f(x)$ dapat diekspansikan (diperluas) ke dalam deret Taylor sebagai:

$$f(x) = f(x_0) + \frac{(x-x_0)}{1!} f'(x_0) + \frac{(x-x_0)^2}{2!} f''(x_0) + \cdots + \frac{(x-x_0)^m}{m!} f^m(x_0) + \cdots$$

Aproksimasi orde nol pada deret taylor merupakan suku pertama dari deret taylor tersebut. Bila dalam deret taylor terdapat penambahan suku maka akan berkembang menjadi aproksimasi orde 2 dan seterusnya. Misalkan $r_n(x)$ merupakan suku tambahan dalam deret taylor setelah bentuk ke *n* dalam deret dan $x_0 = a$, maka diperoleh deret taylor secara umum:

$$f(x) = f(a) + \frac{(x-a)}{1!} f'(a) + \cdots + \frac{(x-a)^n}{n!} f^n(a) + r_n(x)$$

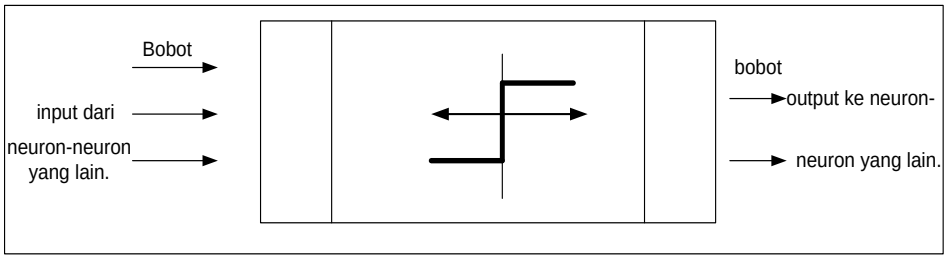
Dengan

$$r_n(x) = \frac{(x-a)^{n-1}}{(n-1)!} f^{n-1}(a)$$

2.6. Konsep Dasar Jaringan Saraf Tiruan

Jaringan saraf tiruan merupakan salah satu bagian dari metode dalam bidang Artificial Intelligence yang dikenal sebagai machine learning (Negnevitsky dan Michael (dalam Pujiati, S. A), 2002).

Jaringan saraf ini diimplementasikan dengan menggunakan program komputer yang mampu menyelesaikan sejumlah proses perhitungan selama proses pembelajaran. Ada beberapa tipe jaringan saraf yang sebagian besar memiliki komponen-komponen yang sama. Sama halnya otak manusia, jaringan saraf juga terdiri dari beberapa neuron dan memiliki hubungan antara neuron-neuron tersebut. Neuron-neuron tersebut akan menstranformasikan informasi yang diterima melalui sambungan keluarnya menuju ke neuron yang lain. Pada jaringan saraf, hubungan ini dikenal dengan nama bobot. Informasi tersebut disimpan pada suatu nilai tertentu pada bobot tersebut. Gambar 2.1 berikut menunjukkan struktur neuron pada jaringan saraf tiruan:



Gambar 2.1: Struktur Neuron Jaringan Saraf.
(Kusumadewi, 2004)

Pada gambar dapat dilihat bahwa neuron buatan ini mirip dengan sel neuron biologis dan cara kerjanya juga sama dengan neuron-neuron biologis. Imformasi akan dikirim ke neuron dengan bobot kedatangan yang akan menjumlahkan nilai-nilai semua bobot yang datang. Hasil penjumlahan ini kemudian akan dibandingkan dengan suatu nilai ambang tertentu melalui fungsi aktivasi setiap neuron. Fungsi aktivasi merupakan fungsi yang menggambarkan hubungan antara tingkat aktivasi internal y Apabila input tersebut melewati suatu nilai ambang tertentu, maka neuron tersebut akan mengirimkan output melalui bobot-bobot outputnya ke semua neuron ang mungkin berbentuk linear atau nonlinear yang berhubungan dengannya (Diyah, 2006). Pada jaringan saraf, neuron-neuron akan dikumpulkan dalam lapisan-lapisan (layer) yang disebut dengan lapisan neuron.

Adapun lapisan-lapisan penyusun jaringan saraf tiruan dapat dibagi menjadi tiga, yaitu:

1. Lapisan Input

   Node-node di dalam lapisan input disebut unit-unit input. Unit-unit input menerima input dari dunia luar. Input yang dimasukkan merupakan penggambaran dari suatu masalah.

2. Lapisan Tersembunyi

   Node-node di dalam lapisan tersembunyi disebut unit-unit tersembunyi. Output dari lapisan ini tidak secara langsung dapat diamati.

3. Lapisan Output

   Node-node pada lapisan output disebut unit-unit output. Keluaran atau output dari lapisan ini merupakan output jaringan saraf tiruan terhadap suatu permasalahan.

2.7. Aturan Pembelajaran Jaringan Saraf Tiruan

Aturan kerja atau aturan pembelajaran jaringan saraf tiruan secara umum terdiri dari 4 tipe dasar (Diyah, 2006), yaitu:

1. Aturan Pengoreksian Error (*Error Correcting)*

   Prinsip dasar dari aturan pembelajaran pengoreksian error ialah memodifikasi bobot-bobot koneksi dengan menggunakan sinyal kesalahan (output target–output aktual) untuk mengurangi besarnya kesalahan secara bertahap.

2. Aturan Pembelajaran Boltzmann

   Aturan Boltzmann dapat juga dikatakan sebagai kasus lain dari aturan pembelajaran pengoreksian error, yang membedakan ialah kesalahan (error) diukur bukan sebagai perbedaan langsung antara output actual dengan output yang diinginkan, melainkan sebagai perbedaan antara output aktual dengan output yang diinginkan, melainkan sebagai perbedaan antara korelasi output-output dari 2 buah neuron dalam kondisi operasi *clamped* dan *free-running*. Pada *clamped*, neuron-neuron visible maupun *hidden* dapat beroperasi dengan bebas. Neuron-neuron yang berinteraksi dengan lingkungan disebut

neuron yang visible, sedangkan neuron-neuron yang tidak berinteraksi dengan lingkungan disebut neuron tersembunyi (*hidden neurons*).

3. Aturan Hebbian

   Pada aturan hebbian kekuatan koneksi antara 2 buah neuron akan meningkat jika kedua neuron memilik tingkah laku yang sama (keduanya memiliki aktivasi positif atau keduanya memiliki aktivasi negatif).

4. Aturan Pembelajaran Kompetitif (*competitive Learning*)

   Unit-unit output pada aturan pembelajaran kompetitif ini harus saling bersaing untuk beraktivasi. Jadi hanya satu unit output yang aktif pada satu waktu. Bobot-bobotnya diatur setelah satu node pemenang terpilih.

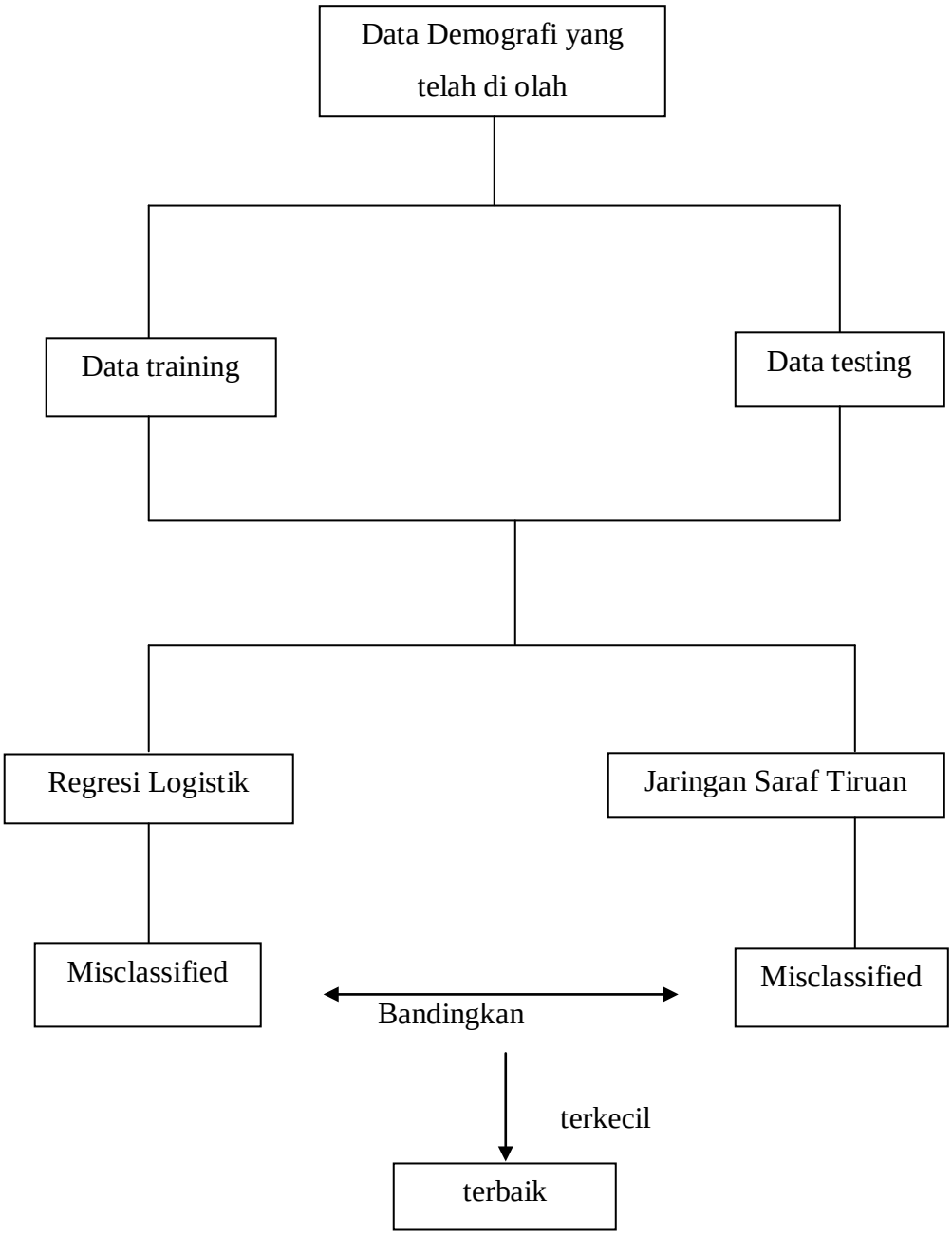
Adapun hal yang ingin dicapai dalam pembelajaran jaringan saraf tiruan ialah untuk mencapai keseimbangan antara kemampuan *memorisasi* dan *generalisasi.* Yang dimaksud dengan kemampuan memorisasi ialah kemampuan jaringan saraf tiruan untuk memanggil kembali secara sempurna sebuah pola yang telah dipelajari. Kemampuan generalisasi ialah kemampuan jaringan saraf tiruan untuk menghasilkan respons yang bisa diterima terhadap pola-pola input yang serupa (tidak identik) dengan pola-pola yang sebelumnya telah dipelajari. Hal ini sangat bermanfaat bila pada suatu saat ke dalam jaringan saraf tiruan itu diinputkan informasi baru yang belum pernah dipelajari, maka jaringan saraf tiruan itu masih akan tetap dapat memberikan tanggapan yang baik, memberikan keluaran yang paling mendekati.

2.8. Jaringan Saraf Back Propagation

Jaringan saraf back propagation merupakan jaringan saraf tiruan dengan topologi multi lapis (multilayer) atau biasa disebut juga dengan Multilayer Perceptron yang menggunakan pembelajaran terawasi, dengan satu lapis masukan (lapis X), satu atau lebih lapis hidden atau tersembunyi (lapis Z) dan satu lapis keluaran (lapis Y). setiap lapis memiliki neuron-neuron (unit-unit). Di antara neuron pada satu lapis dengan neuron pada lapis berikutnya dihubungkan dengan model koneksi yang memiliki bobot-bobot (*weights*), *w* dan *v*. Lapis tersembunyi dapat memiliki bias, yang memiliki bobot sama dengan satu (Daneswara *et al*, 2004).

2.9. Prosedur Klasifikasi

Untuk dapat menyatakan metode klasifikasi yang terbaik dari perbandingan metode klasifikasi regresi logistik dan jaringan saraf tiruan yaitu dengan menghitung *misclassified* pada kedua metode tersebut. Atau dengan kata lain *misclassified* digunakan sebagai indikator dalam melakukan proses klasifikasi. Prosesdur pengklasifikasian dapat dilihat pada Gambar 2.2 berikut:



Gambar 2.2: Diagram Proses Pengklasifikasian

Dalam melakukan klasifikasi, digunakan data training yang berfungsi untuk membentuk model, baik untuk regesi logistik maupun pada jariangan saraf tiruan. Sedangkan data testing digunakan untuk menguji ketepatan klasifikasi dari model yang telah terbentuk. Misclassified kedua model nantinya akan dibandingkan. Pemilihan data training dan data testing di pilih secara acak.

2.9. Demografi

Berdasarkan Multilingual Demographic Dictionary (IUSSP, 1982) defenisi demografi ialah: *Demography is the scientific study of human populations in primarily with the respect to their structure (composition) and their development (change)*.
Menurut Donald J. Bogue mengatakan bahwa demografi ialah ilmu yang mempelajari secara statistik dan matematik tentang besar, komposisi, dan distribusi penduduk dan perubahan-perubahannya sepanjang masa melalui bekerjanya lima komponen demografi yaitu kelahiran (fertilitas), kematian (mortalitas), perkawinan, migrasi, dan mobilitas sosial.

Dari kedua defenisi di atas dapat disimpulkan bahwa dmografi mempelajari tentang suatu penduduk di suatu wilayah. Dapat juga dikatakan bahwa demografi tidaklah mempelajari penduduk sebagai individu, tetapi penduduk sebagai suatu kumpulan (*agregas* atau *collection*) (Mantra, 2003). Selain itu demografi bersifat analitis matematis, yang berarti ananlisis demografi didasarkan atas analisis kuantitatif, dan karena sifatnya yang demikian maka demografi sering juga disebut dengan statistik penduduk. Demografi formal dengan teknik-teknik analisis kuantitatif dapat dibuat pekiraan variabel-variabel demografi berdasarkan data kependudukan yang didapat dari sensus penduduk dan dapat jugadibuat proyeksi penduduk untuk masa-masa mendatang dan masa-masa yang lalu.