BAB 1

PENDAHULUAN

1.1. Latar Belakang

Pengklasifikasian merupakan salah satu metode statistika untuk mengelompok atau menglasifikasi suatu data yang disusun secara sistematis. Masalah klasifikasi sering dijumpai dalam kehidupan sehari-hari. Baik itu pengklasifikasian data pada bidang akademik, sosial, pemerintahan, maupun pada bidang lainnya. Masalah klasifikasi ini muncul ketika terdapat sejumlah ukuran yang terdiri dari satu atau beberapa kategori yang tidak dapat diidentifikasikan secara langsung tetapi harus menggunakan suatu ukuran.

Dalam banyak kasus pengklasifikasian dapat diasumsikan sebagai banyaknya kategori atau populasi dari suatu individu yang ada dan setiap populasi dikarakteristikkan dengan ukuran distribusi probabilitasnya (Anderson,1984). Sebagai contoh yakni: pengklasifikasian siswa yang akan mengikuti ujian masuk di suatu universitas. pengklasifikasian didasarkan atas jurusan yang akan dipilih. Dalam hal ini yang menjadi ukuran ialah jurusan yang dipilih. Pengklasifikasian yang ada terdiri dari tiga klasifikasi yakni IPA, IPS, dan IPC yang memilih IPA dan IPS. Bisa saja siswa tersebut masuk ke dalam kelompok IPA, IPS, ataupun IPC.

Dalam statistika ada beberapa metode klasifikasi yang digunakan untuk melakukan klasifikasi data seperti: analisis diskriminan, regresi logistik dan artificial neural network yang biasa disebut dengan jaringan saraf tiruan. Masing-masing metode tersebut memiliki kelebihan dan kelemahan. Artificial Neural Network (ANN) tidak lebih baik dibandingkan regresi logistik dan analisis diskriminan dalam hal efisiensi waktu pada proses analisisnya (Manel *et al*, 1999). Namun jika dibanding

dengan analisis diskriminan, regresi logistik merupakan metode klasifikasi yang cukup baik, setidaknya pada saat ada variabel independen berskala kuantitatif maupun kualitatif ataupun keduanya (Kurt *et al*, 2006).

Dalam penelitian ini yang akan dibahas ialah metode klasifikasi regresi logistik dan jaringan saraf tiruan (Artificial Neural Network). Regresi logistik adalah salah satu pendekatan model matematis yang digunakan untuk menganalisis hubungan antara satu atau beberapa variabel independen yang bersifat kontinu maupun biner dengan satu variabel dependen yang bersifat dikotomus (biner). Misalkan variabel dependen adalah *Y* dan variabel independen adalah *X*. Dalam hal ini 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.

Satu hal penting untuk menghasilkan prosedur klasifikasi ialah dengan menghitung tingkat error atau probabilitas misklasifikasi. Salah satu ukuran yang dapat digunakan adalah APER. Ukuran ini disebut dengan *apparent error rate* (APER) yang didefenisikan sebagai fraksi observasi dalam sampel yang salah diklasifikasikan pada fungsi klasifikasi (Johnson *et al*,2007). APER digunakan dalam perhitungan ini karena mudah untuk dihitung, dan untuk sampel yang lebih sedikit APER dapat dipergunakan. Disamping itu hasil pengukuran dengan APER tidak bergantung pada distribusi populasi dan dapat dihitung untuk setiap prosedur klasifikasi. Perhitungan APER dapat dilakukan dengan terlebih dahulu membuat tabel klasifikasi, n_1 observasi dari π_1 dan n_2 observasi dari π_2.

Tabel 1: Klasifikasi Actual dan Predicted Group

		Predicted group		
		π_1	π_2	
Actual Group	π_1	n_{1A}	$n_{1B} = n_1 - n_{1A}$	n_1
	π_2	$n_{2B} = n_2 - n_{2A}$	n_{2A}	n_2

Dari tabel diperoleh:

$$APER = \frac{n_{1B}+n_{2B}}{n_1+n_2}$$ (Johnson *et al*, 2007)

Dengan:

n_{1A} : Jumlah pengamatan dari $\pi_1$tepat diklasifikasikan sebagai π_1.

n_{1B} : Jumlah pengamatan dari π_1 salah diklasifikasikan sebagai π_2.

n_{2B} : Jumlah pengamatan dari π_2 salah diklasifikasikan sebagai π_1.

n_{2A} : Jumlah pengamatan dari π_2 tepat diklasifikasikan sebagai π_2.

Dalam evaluasi fungsi klasifikasi, khususnya pada regresi logistik ialah dengan terlebih dahulu membuat tabulasi antara *actual group* dan *predicted group* yang diperoleh dari fungsi logistik. Kemudian dihitung proporsi pengamatan yang salah diklasifikasikan. Diharapkan proporsi pengamatan misklasifikasi tersebut bisa minimum.

Selain regresi logistik, klasifikasi juga dapat dilakukan menggunakan jaringan saraf tiruan (artificial neural network). Jaringan saraf tiruan merupakan suatu sistem pemrosesan data yang memiliki karakteristik mirip dengan jaringan biologis, baik cara kerjanya maupun susunannya juga meniru jaringan saraf biologis.

Jaringan saraf tiruan (artificial neural network) bisa dibayangkan seperti otak buatan di dalam cerita-cerita fiksi. Otak buatan ini dapat berpikir seperti manusia dan juga sepandai manusia dalam menyimpulkan sesuatu dari potongan-potongan informasi yang diterima dan jaringan saraf tiruan ini juga dikatakan mengambil ide dari cara kerja jaringan saraf biologis. Salah satu contoh pengambilan ide dari jaringan saraf tiruan ini ialah adanya elemen-elemen pemrosesan pada jaringan saraf tiruan yang saling terhubung dan beroperasi secara paralel. Ini meniru jaringan saraf biologis yang tersusun dari sel-sel saraf neuron. Namun hal yang perlu diperhatikan adalah bahwa jaringan saraf tiruan cara kerjanya tidak diprogram untuk menghasilkan suatu keluaran tertentu tetapi semua keluaran ataupun kesimpulan yang ditarik oleh jaringan didasarkan pengalamannya selama mengikuti proses pembelajaran. Pada proses pembelajaran, ke dalam jaringan saraf tiruan dimasukkan pola-pola input dan output lalu jaringan akan diajari untuk memberikan jawaban yang bisa diterima (Diyah, 2006).

Masalah yang umum dijumpai pada sistem klasifikasi pola adalah cara menemukan data ciri yang tepat bisa membedakan satu objek dengan objek lainnya. Cara membedakan atau dikenal dengan metode klasifikasi pola juga harus dipertimbangkan. Algoritma artificial neural network banyak digunakan untuk mengatasi masalah-masalah penyimpangan dan pemanggilan data, klasifikasi dan identifikasi pola, pemetaan pola input dan output, pengelompokan pola, hingga pada pencarian nilai-nilai optimasi.

Dalam hal melakukan perbandingan metode yang dinyatakan terbaik ialah metode yang memiliki tingkat error lebih kecil. Error dapat diketahui dari hasil akhir perhitungan masing-masing metode klasifikasi.

Untuk lebih jelasnya bagaimana regresi logistik dan jaringan saraf tiruan bekerja dan metode klasifikasi yang manakah lebih baik dalam proses pengklasifikasian maka penulis mengambil contoh pada data demografi. Data demografi yang diambil merupakan data tentang kependudukan Indonesia pada 30 provinsi yang diambil secara garis besarnya saja atau secara umum.

1.2. Perumusan Masalah

Pada bagian sebelumnya telah diuraikan secara singkat mengenai perbandingan metode klasifikasi regresi logistik dan Jaringan Saraf Tiruan. Namun, yang akan dibahas ialah bagaimana melakukan pengklasifikasian pada data demografi jumlah penduduk 30 provinsi di Indonesia secara umum dengan menggunakan metode klasifikasi regresi logistik dan jaringan saraf tiruan (artificial neural network) dan kemudian membandingkan kedua metode tersebut.

1.3. Pembatasan Masalah

Ruang lingkup penelitian ini dibatasi pada data demografi jumlah penduduk 30 provinsi di Indonesia secara umum yang diklasifikasikan dengan menggunakan metode klasifikasi regresi logistik dan jaringan saraf tiruan (artificial neural network).

1.4. Tujuan Penelitian

Tujuan penelitian ini ialah untuk membandingkan hasil metode klasifikasi regresi logistik dengan jaringan saraf tiruan (artificial neural network) pada data demografi kemudian memilih metode klasifikasi yang manakah lebih baik dalam pengklasifikasian.

1.5. Metodologi Penelitian

Dalam penelitian ini penulis melakukan studi literatur dan mencari bahan dari internet yang membahas mengenai regresi logistik dan artificial neural network (jaringan saraf tiruan). Kemudian mengambil sampel data demografi di 30 provinsi di Indonesia dari internet. Adapun langkah-langkahnya adalah sebagai berikut:

a. Menguraikan penyelesaian pengklasifikasian dengan menggunakan metode klasifikasi regresi logistik dan artificial neural network (jaringan saraf tiruan).
b. Melakukan pengklasifikasian data demografi dengan bantuan komputer spss17
c. Membandingkan kedua metode klasifikasi regresi logistik dan artificial neural network (jaringan saraf tiruan).
d. Menyimpulkan hasil pengklasifikasian, metode klasifikasi yang dianggap terbaik adalah metode klasifikasi yang memberikan kesalahan terkecil.

1.6. Tinjauan Pustaka

Berikut diberikan tinjauan pustaka tentang pengklasifikasian dengan metode regresi logistik dan neural artificial network (jaringan saraf tiruan). Untuk itu dalam hal pengklasifikasian di ambil sampel data demografi jumlah penduduk Indonesia dari 30 provinsi secara umum.

1.6.1. Regresi Logistik

Regresi logistik merupakan salah satu metode klasifikasi yang sering digunakan. Regresi logistik biner digunakan saat variabel dependen merupakan variabel dikotomus. Regresi logistik multinomial digunakan pada saat variabel dependen adalah variabel kategorik dengan lebih dari 2 kategori. Secara umum model regresi logistik multivariate adalah:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}} \qquad \text{(Hosmer } et\ al\text{, 1989)}$$

dimana $\pi(x)$ merupakan nilai probabilitas sehingga $0 \leq \pi(x) \leq 1$, yang berarti bahwa regresi logistik menggambarkan suatu probabilitas. Dengan mentransformasikan $\pi(x)$ pada persamaan di atas dengan transformasi logit $g(x)$, dimana:

$$g(x) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) \qquad \text{(Hosmer } et\ al\text{, 1989)}$$

maka diperoleh bentuk logit:

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \qquad \text{(Hosmer } et\ al\text{, 1989)}$$

Untuk memperoleh estimasi dari parameter regresi logistik dapat dilakukan dengan 2 cara yaitu dengan cara *Maximum Likelihood Estimation* dan metode *iterasi newton rhapson.* Menurut Hosmer dan Lemeshow (1989), metode estimasi maximum likelihood digunakan untuk mengestimasi parameter-parameter dalam regresi logistik.

Pada dasarnya metode maximum likelihood memberikan nilai estimasi β dengan memaximumkan fungsi likelihoodnya.

Dalam melakukan evaluasi fungsi klasifikasi dilakukan dengan membagi data menjadi 2 bagian. Bagian pertama akan dipergunakan sebagai training set, yang diperlukan untuk membentuk model klasifikasi regresi logistik. Berikutnya, bagian kedua akan dipergunakan sebagai validasi set, yang berfungsi sebagai cross-validasi fungsi klasifikasi regresi logistik.

Dalam melakukan pengklasifikasian diharapkan untuk meminimalkan kesalahan klasifikasi atau meminimalkan rata-rata efek buruk dari kesalahan klasifikasi.

1.6.2. Jaringan Saraf Tiruan

Jaringan saraf tiruan merupakan model tiruan bagaimana makhluk hidup memahami dan mengenali informasi disekitarnya (eka *et al*, 2006). Secara sederhana, jaringan saraf tiruan adalah sebuah alat pemodelan data statistik non-linier. Jaringan saraf tiruan dapat digunakan untuk memodelkan hubungan yang kompleks antara input dan output untuk menemukan pola-pola pada data (Wikipedia, 2010).

Jaringan saraf tiruan ini disusun oleh elemen-elemen pemroses yang berada pada lapisan-lapisan yang berhubungan dan diberi bobot. Dengan serangkaian inputan diluar sistem yang diberikan kepadanya jaringan ini dapat memodifikasi bobot yang akan dihasilkannya, sehingga akan menghasilkan output yang konsisten sesuai dengan input yang diberikan kepadanya.

Ada pun langkah-langkah klasifikasi data pada jaringan saraf tiruan diantaranya ialah :

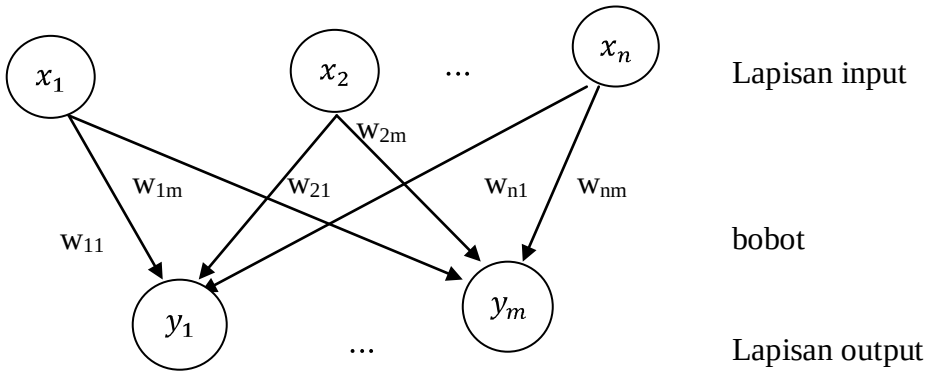
a. Pembangunan model dari set data pelatihan atau pembelajaran. Dalam hal ini dilakukan pembentukan jaringan dan penghitungan nilai-nilai parameter jaringan (bobot, bias, dll).

b. Penggunaan model untuk mengklasifikasikan data baru. Dalam hal ini, sebuah record “diumpankan” ke model, dan model akan memberikan jawaban “kelas” hasil perhitungannya.

Jaringan Saraf Tiruan dibagi ke dalam 3 macam arsitektur, yaitu:

a. Jaringan Lapis Tunggal

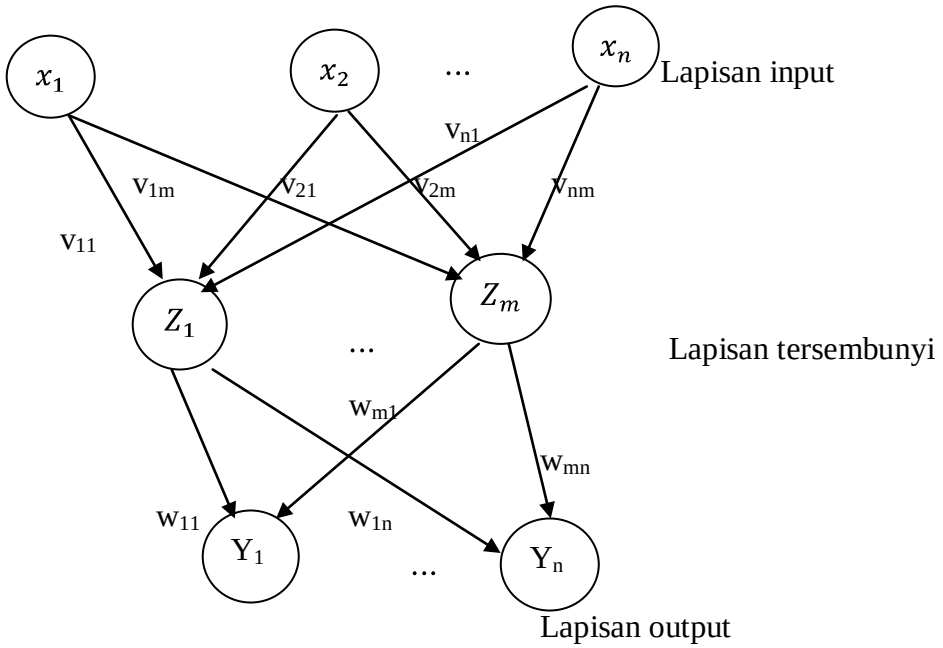
Jaringan yang memiliki arsitektur jenis ini hanya memiliki satu buah lapisan bobot koneksi. Jaringan lapisan tunggal terdiri dari unit-unit input yang menerima sinyal dari dunia luar, dan unit-unit output dimana kita bisa membaca respons dari jaringan saraf tiruan tersebut.



Gambar 1.1: Arsitektur Jaringan Saraf Tiruan Dengan Satu Jaringan Lapis Tunggal

b. Jaringan Multilapis

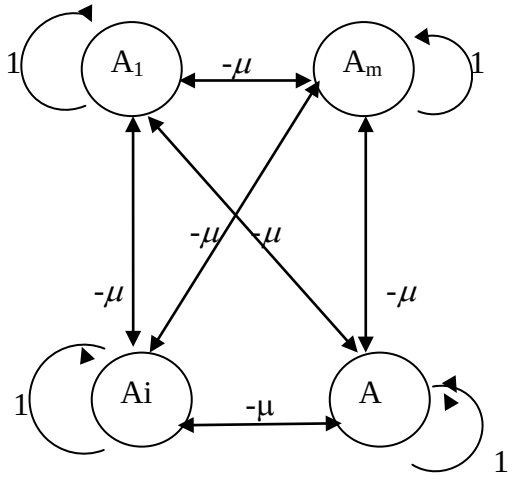
Merupakan jaringan dengan satu atau lebih lapisan tersembunyi. Multilayer net ini memiliki kemampuan lebih dalam memecahkan masalah bila dibandingkan dengan single layer net, namun pelatihannya mungkin lebih rumit.



Gambar 1.2: Arsitektur Jaringan Saraf Tiruan Dengan Jaringan Multilapis

c. Jaringan Kompetitif

Pada jaringan ini sekumpulan neuron bersaing untuk mendapatkan hak menjadi aktif.



Gambar 1.3: Arsitektur Jaringan Saraf Tiruan Dengan Jaringan Kompetitif

Secara umum, tiap unit pada lapisan (layer) yang sama atau dapat juga disebut neuron mempunyai tingkah laku yang sama untuk pemrosesan sinyal data. Hanya hal terpenting yang perlu diperhatikan adalah penentuan penggunaan jenis fungsi aktifasi pada masing-masing unit pada lapisan tersebut dan pola koneksi antar lapisan.