JIMT
Vol. 12 No. 2 Desember 2016 (Hal 115 - 124)

Jurnal Ilmiah Matematika dan Terapan

ISSN : 2450 – 766X

PERBANDINGAN ANALISIS DISKRIMINAN DAN *K-NEAREST NEIGHBOR* (KNN) UNTUK MENGKLASIFIKASIKAN PENDERITA PENYAKIT TUBERKULOSIS (TB)

Nurfajri[1], Rais[2], dan I.TriUtami [3]

[1,2,3] Program Studi Matematika Jurusan Matematika
Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Tadulako
Jalan Sukarno-Hatta Km. 9 Palu 94118, Indonesia
nurfajry90@gmail.com, rais76_untad@yahoo.co.id, triutami.iut@gmail.com

ABSTRACT

Classification is one of statistical methods that arranged data systematically. There are two classify the object are parametric and non parametric. The parametric used in this research is discriminant analysis and *K-Nearest Neighbor* method as non parametric method. The research revealed that discriminant analysis resulted three significant variable, which are a history of disease, means of sanitation, and healthy home. Moreover, discriminant analysis succeeded in classifying Tuberculosis patients into severe and not severe about 93% while the *K-Nearest Neighbor* is about 96%. It is based on classification table.

Keywords : Discriminant Analysis, K-Nearest Neighbor, Classification, Tuberculosis.

ABSTRAK

Pengklasifikasian merupakan salah satu metode statistik dalam pengelompokkan suatu data yang disusun secara sistematis. Pengklasifikasian suatu objek dapat dilakukan dengan dua pendekatan yaitu parametrik dan non parametrik. Parametrik yang digunakan dalam penelitian ini adalah metode Analisis Diskriminan dan K-*Nearest Neighbor* sebagai non parametrik. Hasil penelitian ini diperoleh bahwa faktor-faktor yang mempengaruhi penyakit Tuberkulosis dengan metode Analisis diskriminan menghasilkan 3 variabel yang signifikan yaitu riwayat penyakit, sarana sanitasi dan rumah sehat. Dengan menggunakan tabel ketepatan klasifikasi diketahui bahwa metode analisis diskriminan mengklasifikasikan status pasien penderita Tuberkulosis kedalam kategori parah dan tidak parah dengan tepat sebesar 93%, sedangkan pada K-*Nearest Neighbor* dengan persentase sebesar 96%.

Kata Kunci : Analisis Diskriminan, K-Nearest Neighbor, Klasifikasi, Tuberkulosis.

I. PENDAHULUAN

1.1. Latar Belakang

Klasifikasi merupakan pengelompokkan obyek kedalam satu atau beberapa kelompok berdasarkan variabel yang diamati. Metode statistik yang sering digunakan untuk menyelesaikan masalah pengklasifikasian suatu obyek dapat dilakukan dengan dua pendekatan yaitu parametrik dan non parametrik. Parametrik yang akan digunakan dalam penelitian ini adalah metode analisis diskriminan dan non parametrik yaitu metode *K-Nearest Neighbor.*

Analisis diskriminan merupakan salah satu teknik statistik yang digunakan untuk memisahkan beberapa kelompok obyek yang sudah terkelompokkan sebelumnya dengan cara membentuk fungsi diskriminan. Analisis ini digunakan untuk memeriksa ketepatan suatu pengklasifikasian dan untuk mengetahui besarnya kesalahan pengklasifikasian (misklasifikasi) saat awal pengklasifikasian. *K-Nearest Neighbor* (KNN) adalah suatu metode yang menggunakan algoritma pembelajaran dimana hasil dari *testing sample* yang baru diklasifikasikan berdasarkan mayoritas dari kategori pada KNN. Tujuan dari metode ini adalah mengklasifikasikan objek baru berdasarkan atribut dan *training sample*. Pengklasifikasian tidak menggunakan model apapun untuk dicocokkan dan hanya berdasarkan pada memori.

Pengklasifikasian data penderita penyakit TB dapat dilakukan dengan kedua metode dengan mendapatkan hasil misklasifikasi antara diagnosa dari dokter yang mendiagnosa penyakit seseorang parah atau tidak parah dan kedua metode tersebut. Pengklasifikasian tersebut dilihat berdasarkan faktor-faktor yang mempengaruhi penyakit TB diantaranya jenis kelamin, umur, riwayat hidup, lingkungan rumah, sarana sanitasi dan rumah sehat dengan menggunakan tabel ketepatan klasifikasi.

1.2. Rumusan Masalah

1. Bagaimana menerapkan metode Analisis Diskriminan dan metode *K-Nearest Neighbor* pada data penderita penyakit TB?
2. Bagaimana membandingkan nilai ketepatan klasifikasi menggunakan metode Analisis Diskriminan dan metode *K-Nearest Neighbor* pada data penderita Penyakit TB?

1.3. Tujuan

1. Menerapkan metode Analisis Diskriminan dan *K-Nearest Neighbor* dalam mengklasifikasikan penderita TB.
2. Membandingkan nilai ketepatan klasifikasi antara metode Analisis Diskriminan dan *K-Nearest Neighbor* dalam mengklasifikasikan penderita TB.

1.4. Manfaat Penelitian

1. Menambah wawasan dan pengetahuan tentang analisis diskriminan serta *K-Nearest Neighbor* sebagai metode pengklasifikasian.
2. Memberikan informasi tentang faktor-faktor yang mempengaruhi penyakit TB. Dari informasi tersebut diharapkan dapat meminimalisir kondisi tersebut.

1.5. Ruang Lingkup Penelitian

Ruang lingkup dalam penelitian ini adalah :

1. Sampel yang menjadi penelitian adalah pada penderita penyakit TB di Kota Palu.
2. Faktor-faktor yang mempengaruhi penyakit TB yaitu jenis kelamin, umur, riwayat penyakit, lingkungan rumah, sarana sanitasi dan rumah sehat yang akan dklasifikasikan dengan menggunakan metode Analisis Diskriminan dan *K-Nearest Neighbor.*

II. METODE PENELITIAN

Langkah-langkah yang dilakukan dalam penelitian ini yaitu

1. Pengumpulan Data
2. Uji Validitas dan reliabilitas
3. Jika ada data yang tidak valid dan reliabel maka pertanyaan dihilangkan dalam kuesioner.
4. Jika sudah valid dan reliabel maka dapat dilanjutkan kedalam metode analisis diskriminan dan k-nearest neighbor.

- Analisis Diskriminan
    1. Uji Normalitas
    2. Uji Homoskedastisitas
    3. Menentukan variabel bebas dengan menggunakan metode Stepwise
    4. Pembentukan model analisis diskriminan
    5. Mengukur ketepatan pengklasifikasian analisis diskriminan.
- *K-Nearest Neighbor*
    1. Menentukan parameter K (user)
    2. Membagi data menjadi data training dan data testing
    3. Menghitung jarak data ke data training

4. Kemudian mengurutkan objek-objek tersebut kedalam kelompok yang mempunyai jarak *euclid* terkecil.
5. Mengumpulkan kategori Y (klasifikasi *nearest neighbor*).

III. Hasil dan Pembahasan

1. Uji Validitas

Pengujian validitas masing-masing pertanyaan menggunakan korelasi *product moment*, dengan persamaan keputusan pada sebuah butir pertanyaan dianggap valid, jika $r > 0.174$. Berdasarkan hasil output dari program SPSS 21 yang diperoleh nilai r sebagai berikut :

Tabel 1 : Nilai Product Moment (r)

Nomor	Nilai *Product Moment*	Keterangan
1	0,319	Valid
2	0,354	Valid
3	0,391	Valid
4	0,821	Valid
5	0,451	Valid
6	0,552	Valid
7	0,643	Valid
8	0,685	Valid
9	0,731	Valid
10	0,672	Valid
11	0,615	Valid
12	0,621	Valid
13	0,626	Valid

Sebuah pertanyaan dianggap valid, jika nilai signifikan ≤ α. Dengan menggunakan α = 5%, terlihat dari hasil output SPSS bahwa keseluruhan pertanyaan menunjukkan nilai *sign* = 0.000. Karena 0.000 < 0.005, maka seluruh pertanyaan dalam penelitian ini semuanya valid.

2. Uji Reliabilitas

Pengujian reliabilitas dihitung menggunakan rumus "*Spearman Brown"* yaitu :

$$r_i = \frac{2r_b}{1+r_b} \quad (1)$$

Keterangan : r_i = Angka reliabilitas keseluruhan instrumen

r_b = Korelasi "*Product Moment*" dari keseluruhan pertanyaan yang terbagi atas belahan pertama dan belahan kedua.

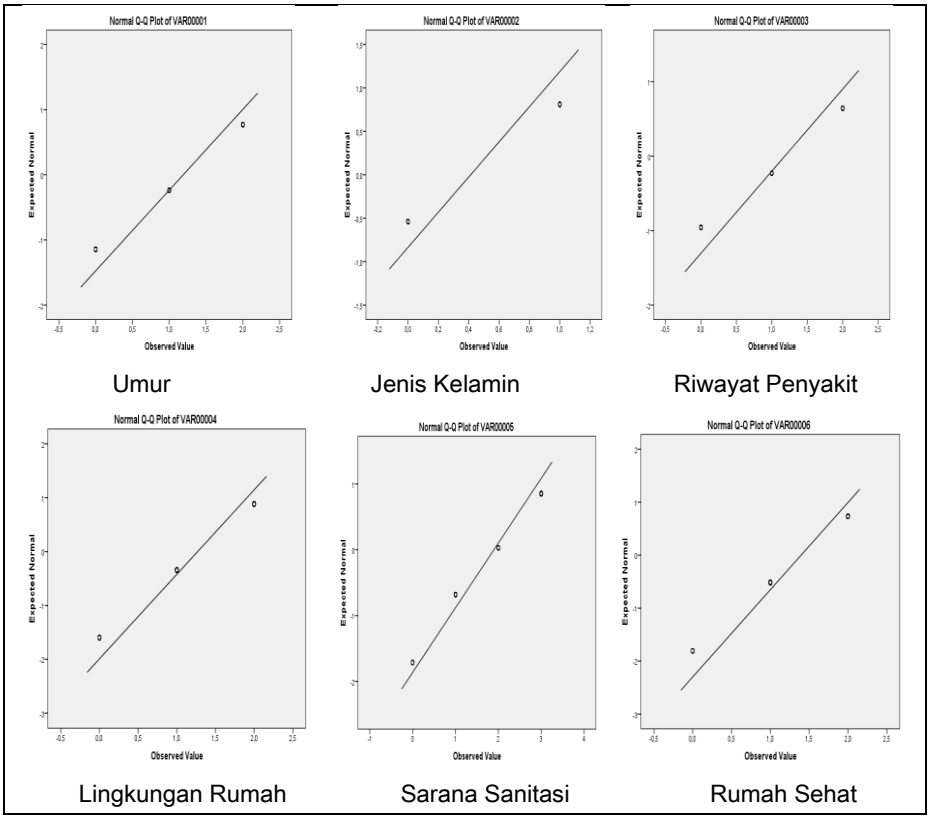
(Sugiyono.,2004)

Dengan mengggunakan rumus diatas maka diperoleh nilai relabilitasnya adalah 0,93. Sehingga untuk mengetahui apakah pertanyaan-pertanyaan tersebut reliabel atau tidak, maka r_i diatas dibandingkan dengan nilai r_{tabel}. Dengan ukuran sampel (n) sebanyak 126 responden dan taraf kesalahan 5%, diperoleh $r_{tabel} = 0.174$. Oleh karena itu nilai $r_{hitung} = 0.93$ > nilai $r_{tabel} = 0.174$, maka kesuluruhan pertanyaan dalam penelitian ini reliabel.

3.1 Analisis Diskriminan

3.1.1 Uji Kenormalan Variabel Bebas

Pengujian kenormalan variabel bebas dilakukan dengan melihat pada normal Q_Q plot dalam paket program SPSS 21 hasil pengujiannya adalah sebagai berikut :



Gambar 1 : Uji Distribusi Normal

Dari Gambar 1 menunjukkan bahwa hasil plot untuk variabel umur, jenis kelamin, riwayat penyakit, lingkungan rumah, sanitasi rumah dan rumah sehat pada normal *Q_Q* plot mendekati garis diagonal, maka dapat disimpulkan bahwa semua variabel berdistribusi normal.

3.1.2 Uji Homoskedastisitas

Uji Homoskedastisitas menggunakan uji statistik Box's M dengan hipotesis yaitu:

H_0 : Kedua kelompok yang ada mempunyai matriks kovarians yang sama.

H_1 : Ada dua kelompok yang berbeda

Berdasarkan Hasil yang diperoleh dengan menggunnakan paket program SPSS 21 uji Homoskedastisitas diperoleh bahwa nilai Sig. = 0.464 > 0.05 atau H_0 diterima sehingga dapat disimpulkan bahwa kedua kelompok yang ada mempunyai matriks kovarians yang sama atau memenuhi asumsi homoskedastisitas.

3.1.3 Penentuan Variabel Bebas dengan Metode Stepwise

Penentuan variabel bebas dengan metode Stepwise dilakukan untuk mengetahui variabel bebas yang paling berpengaruh yaitu variabel yang dapat diikutsertakan dalam pembentukan fungsi diskriminan. Hasil yang diperoleh dengan menggunakan paket program SPSS 21 dari enam variabel ada tiga variabel yang dimasukkan kedalam fungsi diskriminan yaitu tahap pertama variabel yang dimasukkan dalam analisis diskriminan adalah variabel riwayat penyakit karena variabel ini memiliki nilai F_{hitung} (statistik) yang tertinggi yaitu sebesar 139.067, pada tahap kedua variabel yang dimasukkan dalam analisis adalah variabel sarana sanitasi dengan nilai statistik sebesar 122.881, dan tahap ketiga variabel yang dimasukkan dalam analisis yaitu rumah sehat dengan nilai statistik sebesar 99.335. Dengan nilai *Wilk's Lamda* berturut-turut adalah untuk masing-masing variabel yaitu sebesar 0.471, 0.334 dan 0.290

3.1.4 Pembentukan Model Fungsi Diskriminan

Fungsi diskriminan merupakan fungsi kombinasi linear variabel-variabel yang akan menghasilkan cara terbaik dalam pemisahan kelompok-kelompok. Hasil yang diperoleh dengan menggunakan paket program SPSS 21 yaitu :

Tabel 2 : Fungsi Diskriminan

	function
	1
VAR0003	1,074

VAR0005	,910
VAR0006	,795
(constant)	-4,107

Berdasarkan Tabel 2 di atas terlihat bahwa fungsi diskriminan yang terbentuk adalah sebagai berikut :

$\square Z = -4,107 + 1,074\, x_3 + 0,910\, x_5 + 0,795\, x_6$

3.1.5 Pengukuran Ketepatan Pengklasifikasian

Setelah pembentukan dan klasifikasi dilakukan, maka selanjutnya akan dilihat seberapa besar hasil klasifikasi yang tepat dengan kata lain berapa persen kesalahan klasifikasi pada proses klasifikasi tersebut.

Hasil pengklasifikasian yang diperoleh dengan menggunakan paket program SPSS 21 jumlah pengklasifikasian yang benar untuk kelompok parah adalah 53, dan untuk kelompok tidak parah adalah 64. Ketepatan pengklasifikasian di atas kemudian dihitung dengan menggunakan rumus *hit ratio* sebagai berikut :

$$Hit\ Ratio = \frac{n_{11c}+n_{22c}}{\sum_{i=1}^{k} n_i} \qquad (2)$$

Keterangan : n_{1c} = Banyaknya observasi dari kelompok satu yang tepat dikelompokkan pada kelompok satu.

n_{2c} = Banyaknya observasi dari kelompok dua yang tepat dikelompokkan pada kelompok dua.

n_i = Banyaknya seluruh observasi, dengan i=1,2,...,k.

(Supranto,J.,2004).

Jadi, perhitungan $Hit\ Ratio$ dari hasil pengklasifikasian tersebut, data yang terklasifikasikan dengan benar sebesar 93%.

3.2 Metode K-*Nearest Neighbor* (K-NN)

3.2.1 Pembagian Data Training dan Data Testing

Sebelum data diproses dalam MATLAB, terlebih dahulu data harus dibagi menjadi dua yaitu untuk data training dan data testing. Data *training* digunakan untuk membentuk model, sedangkan data *testing* digunakan untuk menguji ketepatan klasifikasi dari model yang telah terbentuk. Dalam kasus ini terdapat 126 data yang terdiri dari 55 data parah dan 71 data tidak parah yang akan dibagi menjadi data training dan data testing, yaitu untuk data training dibagi menjadi 80% dan untuk data testing dibagi menjadi 20%. Dari hasil data

Pembagian data training dan data testing untuk 55 data parah langsung dibagi menjadi 2 bagian yaitu 11 untuk data testing dan 44 untuk data training. Untuk 71 data tidak parah yaitu 14 data untuk data testing dan 57 untuk data training. Maka data training sebanyak 101 data, sedangkan data testing sebanyak 25 data.

3.2.2 Proses Training Data

Pada proses training data, KNN hanya melakukan penyimpanan vektor-vektor fitur dan klasifikasi dari data training sampel tersebut. *Classifier* tidak menggunakan model apapun untuk dicocokkan dan hanya berdasarkan pada memori.

3.2.3 Proses Testing Data

Untuk nilai akurasi sendiri tergantung pada nilai k. Nilai k yang terbaik untuk metode ini tergantung pada data secara umum, nilai k yang tinggi terkadang akan membuat batasan antara setiap klasifikasi menjadi lebih kabur dan sebaliknya. Dari hasil yang didapatkan, nilai k yang memiliki akurasi tertinggi adalah k=1. Hasil testing data untuk k=1 dapat dilihat pada Tabel 3 yang dihasilkan dari program KNN pada perangkat lunak MATLAB R2008b, output yang dihasilkan berupa kelas hasil prediksi dan nilai akurasi.

Tabel 3 : Data Testing Hasil Prediksi untuk k =1

No	U	JK	RP	L.R	S.S	R.S	Hasil Nyata	Hasil Pengenalan	Ket.
1	1	0	0	0	2	0	Parah	Parah	BENAR
2	1	0	0	1	1	2	Parah	Parah	BENAR
3	1	0	0	1	0	1	Parah	Parah	BENAR
4	1	0	0	2	1	2	Parah	Parah	BENAR
5	1	0	0	1	1	2	Parah	Parah	BENAR
6	1	1	2	1	1	1	Parah	Parah	BENAR
7	1	1	1	2	0	1	Parah	Parah	BENAR
8	0	0	0	1	1	0	Parah	Parah	BENAR
9	0	0	0	1	1	1	Parah	Parah	BENAR
10	1	0	0	1	1	1	Parah	Parah	BENAR
11	1	0	0	1	1	1	Parah	Parah	BENAR
12	0	1	2	1	3	1	Tidakparah	Tidak Parah	BENAR
13	0	1	2	2	2	1	Tidakparah	Tidak Parah	BENAR
14	0	1	2	2	2	2	Tidakparah	Tidak Parah	BENAR
15	0	1	2	2	2	2	Tidakparah	Tidak Parah	BENAR
16	0	0	2	1	3	2	Tidakparah	Tidak Parah	BENAR
17	0	0	2	1	3	2	Tidakparah	Tidak Parah	BENAR
18	2	0	2	2	2	2	Tidakparah	Tidak Parah	BENAR
19	1	1	2	0	3	1	Tidakparah	Tidak Parah	BENAR
20	1	1	2	2	3	1	Tidakparah	Tidak Parah	BENAR
21	2	1	2	2	1	2	Tidakparah	Tidak Parah	BENAR
22	2	1	2	1	3	2	Tidakparah	Tidak Parah	BENAR
23	2	0	0	1	1	2	Tidakparah	Parah	SALAH
24	1	0	2	2	3	2	Tidakparah	Tidak Parah	BENAR
25	1	0	2	2	3	1	Tidakparah	Tidak Parah	BENAR

Dari Tabel 3 di atas, hasil proses testing pada data untuk k=1 menunjukkan bahwa terdapat 24 data yang benar, dalam hal ini data hasil pengenalan program sesuai dengan data nyata, dan 1 data yang salah atau kata lain data hasil pengenalan program tidak sesuai dengan data nyata. Maka, untuk metode *K-Nearest Neighbor* data yang diperoleh nilai akurasinya sebesar 96% dari data awal.

IV. Kesimpulan

Berdasarkan hasil dari penelitian mengenai faktor-faktor yang mempengaruhi terjadinya penyakit TB yang telah dibahas serta perhitungan statistik yang dilakukan, maka penulis menarik kesimpulan sebagai berikut :

1. Dari enam faktor-faktor yang mempengaruhi penyakit TB, pada analisis diskriminan hanya tiga faktor yang bisa dimasukkan kedalam fungsi diskriminan yaitu riwayat penyakit, sarana sanitasi dan rumah sehat. Sedangkan untuk metode K-*Nearest Neighbor* menghasilkan nilai tetangga terdekat yaitu k=1.

2. Bedasarkan tabel klasifikasi dari kedua metode yang digunakan untuk metode analisis diskriminan hasil klasifikasinya sebesar 93% dan untuk metode *K-Nearest Neighbor* hasil klasifikasi yang didapatkan sebesar 96%. Sehingga metode yang tepat dan akurat klasifikasinya untuk penentuan solusi penyakit TB yang digunakan yaitu metode *K-Nearest Neighbor.* Metode non parametrik seperti *K-Nearest Neighbor* mempunyai kelebihan karena tidak menggunakan asumsi apapun dalam penggunaannya.

DAFTAR PUSTAKA

[1]. Boedy, Cged., 2012. *Pengertian, Kelebihan, dan Kekurangan K-nearest Neighbor (K-NN). http://cgeduntuksemua.blogspot.com/2012/03/pengertian kelebihan- dan-kekurangan-k.html. Diakses pada tanggal 16 April 2014.*

[2]. Rusmasari, A., 2004, *Profil dan Analisis Keterkaitan Berbagai Karakteristik Kusir dan sarana Angkutan Tradisional Andong Dengan Pendapatan Harian (Studi Kasus Andong di Kabupaten/Kota Magelang tahun 2004)*, Sekolah Tinggi Ilmu Statistik, Jakarta.

[3] Rencher, A. C., 1996, *Methods of Multivariate Analysis,* Inc, New York.

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

[5] Santosa, P. B., Ashari., 2005, *Analisis Statistik Dengan Microsoft Excel dan SPSS*, Andi Yogyakarta, Yogyakarta.

[6] Suliyanto., 2005, *Analisis Data dalam Aplikasi Pemasaran,* Ghalia Indonesia, Bogor.

[7] Sugiyono., 2004, *Statistika Untuk Penelitian*, Cetakan III, CV Alfabeta, Bandung.