Buletin Ilmiah Mat. Stat. dan Terapannya (Bimaster)
Volume 11, No. 1 (2022), hal 97-102

ANALISIS DISKRIMINAN DENGAN *K FOLD CROSS VALIDATION* UNTUK KLASIFIKASI KUALITAS AIR DI KOTA PONTIANAK

Linda Mardiana, Dadan Kusnandar, Neva Satyahadewi

INTISARI

Bagi masyarakat Kota Pontianak kebutuhan air untuk higiene sanitasi sebagian besar bersumber dari sungai. Maka penting bagi masyarakat untuk mengetahui sejauh mana tingkat kualitas air yang saat ini mereka manfaatkan untuk kebutuhan sehari-hari. Analisis diskriminan adalah metode untuk mengklasifikasi suatu objek atau sampel ke dalam salah satu dari dua kelompok atau lebih. Tujuan penelitian ini adalah mengklasifikasikan tingkat pencemaran air berdasarkan indeks pencemaran. Metode k fold cross validation digunakan untuk memperoleh hasil akurasi yang maksimal dari fungsi analisis diskriminan. Penelitian ini menggunakan data primer 42 sampel air di kota Pontianak. Variabel dependen yang digunakan merupakan hasil dari perhitungan indeks pencemaran terdiri dari kategori tercemar ringan dan tercemar sedang. Variabel independen terdiri dari fluorida, kesadahan, nitrat, dan DO. Langkah-langkah penelitian yaitu menentukan indeks pencemaran yang digunakan sebagai variabel dependen, menguji asumsi normal multivariat dan kesamaan matriks varian-kovarian, membagi data dengan metode k fold cross validation, dan melakukan proses klasifikasi analisis diskriminan. Model diskriminan terbaik diperoleh pada eksperimen kedua adalah $D_2 = -3{,}500 + 0{,}087x_2$ *dengan nilai Apparent Error Rate (APER) terendah yaitu sebesar 0,21.*

Kata kunci: *analisis diskriminan, k fold cross validation, indeks pencemaran.*

PENDAHULUAN

Air adalah sumber daya alam yang merupakan sumber kehidupan dari semua makhluk hidup. Sedangkan di Kota Pontianak sumber air yang digunakan sebagai sumber air baku berasal dari Sungai Kapuas. Status kualitas air sungai di Kalimantan Barat, khususnya Sungai Kapuas berdasarkan Peraturan Pemerintah No 82/2001 Kelas II, sudah mencapai kisaran tercemar ringan-tercemar sedang. Pengendalian pencemaran air sungai perlu dilakukan dengan menetapkan standar mutu air. Standar baku mutu kesehatan lingkungan dan persyaratan kesehatan air untuk keperluan higiene sanitasi diatur dalam Peraturan Menteri Kesehatan Republik Indonesia Nomor 32 Tahun 2017 [1].

Penelitian ini dilakukan untuk mengklasifikasi kualitas air di Kota Pontianak kedalam kriteria tercemar ringan atau tercemar sedang yang diperoleh dari perhitungan indeks pencemaran. Analisis diskriminan digunakan untuk mengklasifikasikan tingkat pencemaran air dengan memisahkan dan mengalokasikan objek pengamatan ke dalam kelompok sehingga setiap objek menjadi anggota dari salah satu kelompok dan tidak ada objek yang menjadi anggota lebih dari satu kelompok. Sedangkan validasi keakuratan model fungsi diskriminan dengan metode *k fold cross validation*. Dimana metode *k fold cross validation* digunakan dalam memperkirakan kesalahan prediksi untuk evaluasi kinerja model. Metode ini memiliki prinsip dasar membagi keseluruhan data menjadi data training dan data testing [2].

Proses penelitian diawali dengan menghitung nilai indeks pencemaran dari data yang sudah didapatkan. Kemudian menentukan vaiabel dependen berdasarkan nilai indeks pencemaran. Setelah itu melakukan uji asumsi analisis diskriminan yaitu data berdistribusi normal multivariat dan matriks varian-kovarian seragam. Data dibagi sebanyak k bagian dengan metode *k fold cross validation* lalu dilakukan analisis diskriman dengan mencari model diskiminan. Tahap akhir pada proses ini adalah melihat peluang kesalahan klasifikasi dengan nilai APER.

INDEKS PENCEMARAN

Indeks pencemaran (PI) merupakan metode yang digunakan untuk menentukan status mutu air dengan cara membandingkan kondisi mutu air sumber dan baku mutu air yang telah ditetapkan. Standar mutu air ini kemudian dapat digunakan untuk seluruh bagian badan air atau sebagian dari suatu sungai [3].

$$PI_j = \sqrt{\frac{\left(\frac{C_{ij}}{L_{ij}}\right)_M^2 + \left(\frac{C_{ij}}{L_{ij}}\right)_R^2}{2}} \quad (1)$$

dimana PI_j adalah indeks pencemaran pada lokasi ke-j , C_{ij} adalah nilai hasil penelitian pada variabel ke-i dan lokasi ke-j, sedangkan L_{ij} adalah nilai baku mutu pada variabel ke-i lokasi ke-j , R merupakan nilai rata-rata rasio C_{ij} dibagi L_{ij} , dan M adalah nilai maksimum rasio C_{ij} dibagi L_{ij} .

ANALISIS DISKRIMINAN

Analisis diskriminan merupakan metode statistik yang bisa digunakan pada hubungan antarvariabel dimana sudah bisa dibedakan mana variabel respon dan mana variabel penjelas. Variabel dependen pada analisis diskriminan berupa data kualitatif dan variabel independennya berupa data kuantitatif [4].

Asumsi utama yang harus dipenuhi pada analisis diskriminan adalah sejumlah p variabel penjelas harus berdistribusi normal dan matriks varian-kovarian variabel penjelas berukuran $p \times p$ pada kedua kelompok harus sama. Klasifikasi dalam analisis diskriminan bertujuan untuk memasukkan observasi baru ke dalam kelompok yang telah mempunyai label kelompok [5].
Bentuk umum dari model diskriminan yaitu:

$$D_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip} \quad (2)$$

dengan D_i merupakan nilai diskriminan ke- i, $b_p$merupakan nilai koefisien fungsi diskriminan ke- p dan x_{ip} adalah variabel pada baris ke-i kolom ke- p.

Uji Normal Ganda

Pemeriksaan distribusi normal ganda dilakukan menggunakan plot nilai jarak Mahalanobis dan nilai chi-kuadrat yang ditulis dengan $\left(d_i{}^2, \chi^2 p\left(\left(i-\frac{1}{2}\right)/n\right)\right)$. nilai jarak mahalanobis dihitung dengan persamaan sebagai berikut:

$$d_i{}^2 = (x_i - \bar{x})' S^{-1} (x_i - \bar{x});\ i = 1,2, \ldots,n.$$

dimana x_i merupakan sampel pengamatan ke i, $\bar{x}$ merupakan vektor rata-rata, dan S^{-1}= invers matriks ragam-peragam. Nilai $d_i{}^2$ kemudian diturunkan dari nilai terkecil sampai nilai terbesar dan dibuat plot $d_i{}^2$ dan nilai $\chi^2 p\left(\left(i-\frac{1}{2}\right)/n\right)$. Sehingga jika plot cenderung membentuk garis lurus dan lebih dari 50% nilai$d_i{}^2 \leq \chi^2 p_{(0,05)}$ artinya data berdistribusi normal ganda [6].

Uji M Box

Hipotesis Uji M Box dilakukan untuk menguji asumsi kehomogenan ragam. Hipotesis yang digunakan adalah sebagai berikut:

H_0: memliki kelompok kovarian yang sama. Dengan $\sum_1 = \sum_2 = \cdots \sum_K$,

H_1: minimal ada dua kelompok yang berbeda. Dengan $\sum_i \neq \sum_j untuk\ i \neq j$ dengan *i* dan *j* = 1, 2, ..., k

Statistik Uji M Box yaitu:

$$-2 \ln \lambda = (n - k) \ln \left|\frac{W}{(n-k)}\right| - \sum_{i=1}^{k}(n_i - 1) ln|S_i|, \text{ dengan } \lambda = \frac{\prod_{i=1}^{k}|S_i|^{(n_i-1)/2}}{\left|\frac{W}{(n-k)}\right|^{(n-k)/2}}$$

Jika diperoleh $\frac{-2ln\lambda}{b} \leq F_{V_1,V_2,\alpha}$ dan *p-value* > α maka H_0 diterima. Hal ini diartikan bahwa semua kelompok mempunyai matriks varian-kovarian yang homogen [4].

K FOLD CROSS VALIDATION

K fold cross validation digunakan untuk mengestimasi kesalahan prediksi dalam mengevaluasi kinerja model. Data dibagi menjadi himpunan bagian k berjumlah hampir sama. Model dalam klasifikasi dilatih dan diuji sebanyak k. Disetiap pengulangan, salah satu himpunan bagian akan digunakan sebagai data *training* dan data *testing*. [7]. Langkah-langkah dari *k fold cross validation* yaitu:

1. Total data dibagi menjadi *k* bagian.
2. Fold ke-1 adalah ketika bagian ke-1 menjadi data uji (*testing data*) dan sisanya menjadi data latih (*training data*). kemudian, hitung akurasi atau kesamaan atau kedekatan suatu hasil pengukuran dengan angka atau data yang sebenarnya berdasarkan porsi data tersebut. Perhitungan akurasi tersebut menggunakan persamaan sebagai berikut.

$$akurasi = \frac{\sum data\ uji\ benar\ klasifikasi}{\sum total\ data\ uji} \times 100$$

3. Fold ke-2 adalah ketika bagian ke-2 menjadi data uji (*testing data*) dan sisanya menjadi data latih (*training data*). kemudian hitung akurasi berdasarkan porsi data tersebut.
4. Demikian seterusnya hingga mencapai fold ke-*k*. Hitung rata-rata akurasi dari *k* buah akurasi diatas. Rata-rata akurasi ini menjadi akurasi final.

UJI APER

Apparent Error Rates (APER) atau peluang kesalahan klasifikasi menyatakan bahwa secara ekuivalen kinerja sebuah model dapat dinyatakan dalam bentuk *error rate-nya*, yang dinyatakan oleh persamaan berikut [8]:

$$APER = \frac{\sum n_{ij}}{N}; i \neq j \quad (3)$$

Dengan $\sum n_{ij}$ merupakan jumlah dari nilai pada kelompok aktual ke-i terhadap nilai pada kelompok prediksi ke-j dan *N* merupakan jumlah seluruh sampel yang diteliti.

HASIL DAN PEMBAHASAN

Data penelitian menggunakan 42 titik lokasi di Kota Pontianak. Indeks pencemaran (PI) diperoleh berdasarkan indikator kekeruhan, warna, besi, BOD, dan COD [8]. Variabel dependen yang digunakan merupakan hasil dari perhitungan indeks pencemaran terdiri dari kategori tercemar ringan dan tercemar sedang. Variabel independen terdiri dari fluorida, kesadahan, nitrat, dan DO. Setelah melakukan perhitungan indeks pencemaran menggunakan persamaan (1) pada tiap-tiap sampel, didapat nilai indeks pencemaran untuk sampel pertama adalah 5,888839. Sehingga sampel pertama tergolong dalam kategori

tercemar sedang dan perhitungan indeks dilanjutkan hingga sampel data terakhir. Sehingga didapat variabel dependen terdiri dari dua kategori, yaitu tercemar ringan dan tercemar sedang.

Tabel 1 Kriteria Indeks Pencemaran

Kelas Indeks Pencemaran	Keterangan
$0 \leq PI \leq 1{,}0$	Memenuhi baku mutu
$1{,}0 < PI \leq 5{,}0$	Tercemar ringan
$5{,}0 < PI \leq 10$	Tercemar sedang
$PI > 10$	Tercemar berat

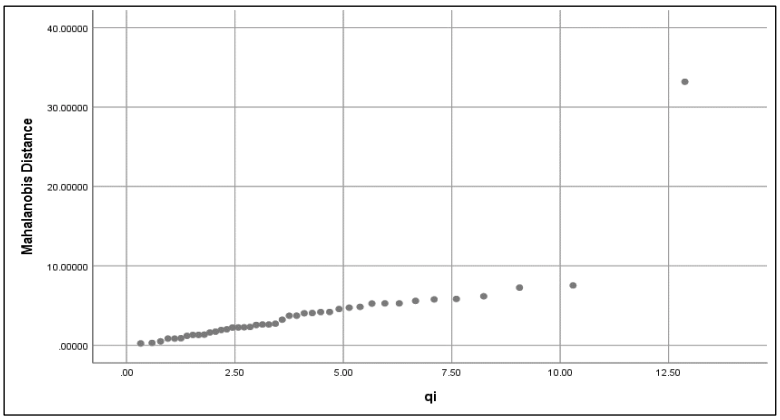
Sumber: Keputusan Menteri Negara Lingkungan Hidup Nomor 115 Tahun 2003

Berdasarkan Tabel 1 dapat ditentukan kelas indeks pencemaran $1,0 < PI \leq 5,0$ yang termasuk kategori tercemar ringan yaitu terdiri dari 16 titik lokasi. Kelas indeks pencemaran $5{,}0 < PI \leq 10$ yang termasuk kategori tercemar sedang yaitu terdiri dari 26 titik lokasi.

Pengujian Asumsi Analisis Diskriminan

a. Distribusi Normal Ganda

Data berdistribusi normal ganda diketahui dengan melihat plot *chi-square*. Hasil pengujian normal ganda diinterpretasikan ke dalam plot *chi-square* yang disajikan pada Gambar 2:



Gambar 2 Plot Jarak Mahalanobis

Berdasarkan Gambar 2 uji normalitas terpenuhi dengan titik-titik menyebar mendekati garis lurus. Uji hipotesis untuk mengetahui apakah data telah berdistribusi normal ganda atau tidak yaitu:
H_0: Data berdistribusi normal ganda dan H_1: Data tidak berdistribusi normal ganda. Statistik uji yaitu sebesar 79,6% nilai dari $d_i^2 < \chi^2_{(0,05;db)}$. Sehingga kesimpulan dengan menggunakan $\alpha = 0{,}05$ lebih dari 50% nilai dari jarak mahalanobis $d_i^2 < \chi^2_{(0,05;db)}$ maka terima H_0 [9].

b. Kesamaan Matriks Varian-Kovarian

Hipotesis uji yang digunakan adalah adalah:
H_0: kedua kategori variabel terikat mempunyai matrik varian-kovarian yang sama
H_1: kedua kategori variabel terikat mempunyai matrik varian-kovarian yang berbeda. Setelah data diolah, diperoleh nilai signifikan dengan $\alpha = 0,05$ adalah 0,544. Karena 0,544 > 0,05 maka H_0 diterima, sehingga dapat disimpulkan bahwa matriks varian-kovarian bersifat homogen [2].

Analisis Diskriminan dengan metode *K Fold Cross Validation*

Data dipartisi menjadi 3 bagian $(k = 3)$sehingga dari 42 data menjadi masing-masing partisi berjumlah 14 data secara acak. Pada tiap eksperimen sebanyak 28 data sebagai data *training* yang akan digunakan untuk membentuk model dan 14 data sebagai data *testing* digunakan untuk memeriksa model. Kelompok kategori yang digunakan adalah kelompok 1 untuk kategori tercemar ringan dan kelompok 2 untuk kategori tercemar sedang. Dalam penelitian ini akan dilakukan 3 kali eksperimen.

***Cross Validation* 1**

Pada eksperimen pertama digunakan data *training* yaitu data partisi ke-1 dan ke-2. Sehingga terbentuklah model diskriminan menggunakan metode *Stepwise* $D_1 = -3{,}401 + 0{,}086x_2$. Sehingga diperoleh hasil pada eksperimen pertama menggunakan data *training* terdapat 11 titik sampel yang salah klasifikasi pada kelompok 1. Sedangkan pada kelompok 2 tidak terdapat titik sampel yang salah klasifikasi. Jadi, total kesalahan klasifikasi ada 11 titik sampel.

Selanjutnya Validasi dilakukan pada data *testing* yaitu data partisi ke-3, hasil klasifikasi terdapat 5 titik sampel pada kelompok 1 yang salah klasifikasi. Sedangkan pada kelompok 2 tidak terdapat titik sampel yang salah klasifikasi. Jadi, total kesalahan klasifikasi ada 5 titik sampel.

***Cross Validation* 2**

Proses pada eksperimen kedua data *training* adalah data partisi ke-1 dan ke-3. model diskriminan yang diperoleh: $D_2 = -3{,}500 + 0{,}087x_2$. Sehingga diperoleh hasil 6 titik sampel pada kelompok 1 yang salah klasifikasi. Sedangkan pada kelompok 2 terdapat 3 titik sampel yang salah klasifikasi. Jadi, total kesalahan klasifikasi ada 9 titik sampel.

Selanjutnya Validasi dilakukan pada data *testing* yaitu data partisi ke-2. Sehingga hasil klasifikasi terdapat 3 titik sampel pada kelompok 2 yang salah klasifikasi. Sedangkan pada kelompok 1 tidak terdapat titik sampel yang salah klasifikasi. Jadi, total kesalahan klasifikasi ada 3 titik sampel.

***Cross Validation* 3**

Proses pada eksperimen ketiga data yang digunakan pada data *training* adalah data partisi ke-2 dan ke-3. Dan diperoleh model $D_3 = -3{,}176 + 0{,}086x_2$. Sehingga diperoleh hasil pada eksperimen ketiga terdapat 8 titik sampel pada kelompok 1 yang salah klasifikasi. Sedangkan pada kelompok 2 tidak terdapat titik sampel yang salah klasifikasi. Jadi, total kesalahan klasifikasi ada 8 titik sampel.

Selanjutnya Validasi dilakukan pada data *testing* yaitu data partisi ke-1. Sehingga didapatkan hasil klasifikasi terdapat 6 titik sampel pada kelompok 1 yang salah klasifikasi. Sedangkan pada kelompok 2 tidak terdapat titik sampel yang salah klasifikasi. Jadi, total kesalahan klasifikasi ada 6 titik sampel.

Peluang Kesalahan Klasifikasi (APER)

Menghitung nilai APER digunakan pada data *testing* dengan rumus pada persamaan (3) sehingga diperoleh nilai APER sebagai berikut**:**

Untuk eksperimen ke– 1: $APER = \frac{0+5}{5+9} = 0{,}35$ artinya peluang kesalahan klasifikasi untuk data *testing* sebesar 35%.

Untuk eksperimen ke– 2: $APER = \frac{0+3}{3+11} = 0{,}21$ artinya peluang kesalahan klasifikasi untuk data *testing* sebesar 21%.

Untuk eksperimen ke– 3: $APER = \frac{0+8}{6+8} = 0{,}57$ artinya peluang kesalahan klasifikasi untuk data *testing* sebesar 57%.

PENUTUP

Berdasarkan hasil pembahasan pada penelitian ini model diskriminan yang diperoleh pada data kualitas air di 42 titik di Kota Pontianak adalah: $D_1 = -3{,}401 + 0{,}086x_2$, $D_2 = -3{,}500 + 0{,}087x_2$, dan $D_3 = -3{,}176 + 0{,}086x_2$. Model terbaik ditentukan dengan nilai APER terkecil yaitu pada eksperimen pertama didapat nilai APER sebesar 0,35 eksperimen kedua sebesar 0,21 dan eksperimen ketiga sebesar 0,57. Sehingga model terbaik adalah pada eksperimen kedua. Berdasarkan hasil analisis yang didapatkan pada model terbaik, diketahui bahwa terdapat 30 data yang tepat diklasifikasikan atau terdapat 12 data tidak tepat diklasifikasikan.

DAFTAR PUSTAKA

[1]. Republik Indonesia, Peraturan Menteri Kesehatan Republik Indonesia Nomor 32 Tahun 2017 Tentang Standar Baku Mutu Kesehatan Lingkungan dan Persyaratan Kesehatan Air Untuk Keperluan Higiene Sanitasi, Kolam Renang, Solus per Aqua dan Pemandian Umum. Sekretariat Negara. Jakarta; 2017.

[2]. Davidson, A., & Hinkley, D., *Bootstrap Methods and their Application*, Cambridge University Press; 1997.

[3]. Republik Indonesia,Peraturan Menteri Lingkungan Hidup Nomor 115 Tahun 2003 Tentang Pedoman Penentuan Status Mutu Air. Jakarta; 2003.

[4]. Mattjik, A.S, dan Sumertajaya I.M., *Sidik Peubah Ganda dengan Menggunakan SAS*, IPB PRESS, Bogor; 2011.

[5]. Johnson, R. A., and Wichern, W. D., *Applied Multivariate Statistical Analysis 6th ed*, 2007.

[6]. Supartini, I. A. M., Sukarsa, I. K. G., dan Srinadi, I. G. A. M,Analisis Diskriminan Pada Klasifikasi Desa di Kabupaten Tabanan Menggunakan Metode K-Fold Cross Validation. *E-Jurnal Matematika*. 6(2), 106-115; 2017.

[7]. Nurhayati., Iwan K, Hadihardaja., Indratmo Soekarno., & M. Cahyono,*A Study of Hold-Out and K-Fold Cross Validation for Accuracy of Groundwater Modeling in Tidal Lowland Reclamation Using Extreme Learning Machine. 2nd International Conference on Technology, Informatics, Management, Engineering & Environment*. pp: 228 – 233; 2014.

[8]. Fikri, M., Debataraja, N. N., dan Kusnandar, D., Analisis Deskriptif Kualitas Air di Kawasan Pemukiman di Kota Pontianak. *Media Statistika*. 8(2), 345–348; 2019.

[9]. Sonya, G., Khairinda N.A., dan Salam, N., *Pengujian Distribusi Multivariat Normal Dan Vektor Rataan Pada Jumlah Penduduk Menurut Jenis Kelamin, Tingkat Pendidikan Dan Kemiskinan Di Provinsi Kalimantan Selatan Dan Kalimantan Barat;* 2019.

LINDA MARDIANA **:** Jurusan Statistika FMIPA UNTAN, Pontianak
lindastatuntan14@student.untan.ac.id
DADAN KUSNANDAR : Jurusan Matematika FMIPA UNTAN, Pontianak
dkusnandar@math.untan.ac.id
NEVA SATYAHADEWI : Jurusan Matematika FMIPA UNTAN, Pontianak
neva.satya@math.untan.ac.id