PENDEKATAN KLASIFIKASI DISKRIMINAN LINIER FISHER UNTUK KABUPATEN/KOTA TERTINGGAL DI PULAU SUMATERA DENGAN *K-FOLD CROSS VALIDATION*

(Skripsi)

Oleh
NADHIRA DEWIANTARI

JURUSAN MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2023

ABSTRACT

FISHER'S LINEAR DISCRIMINANT CLASSIFICATION APPROACH FOR UNDERDEVELOPING DISTRICT/CITIES IN SUMATERA ISLAND USING *K-FOLD CROSS VALIDATION*

By

Nadhira Dewiantari

Sumatera Island is one of many island in Indonesia which has 7 lagging regencies/cities from 154 regencies/cities. Discriminant analysis is a technique related to the separation of objects into different predefined groups. One of the classification approaches is the fisher approach. Fisher's discriminant analysis assumes the covariance matrices of the two groups are the same. This study uses k-fold cross validation as validation. From this study, a model was obtained to classify regencies/cities on the Sumatera Island with an average prediction error probability of 5.8%, and has accurate and consistent results.

Key words: fisher's discriminant analysis, k-fold cross validation, lagging district

ABSTRAK

PENDEKATAN KLASIFIKASI DISKRIMINAN LINIER FISHER UNTUK KABUPATEN/KOTA TERTINGGAL DI PULAU SUMATERA DENGAN *K-FOLD CROSS VALIDATION*

Oleh

Nadhira Dewiantari

Pulau Sumatera merupakan salah satu dari banyaknya pulau di Indonesia yang masih memiliki 7 kabupaten/kota tertinggal dari 154 kabupaten/kota. Analisis diskriminan adalah teknik yang berkaitan dengan pemisahan objek ke dalam kelompok yang berbeda yang telah ditetapkan sebelumnya. Salah satu pendekatan klasifikasinya yaitu dengan pendekatan fisher. Analisis diskriminan linier fisher mengasumsikan matriks kovarians dari dua kelompok sama. Penelitian ini menggunakan *k-fold cross validation* sebagai validasi. Dari penelitian ini didapatkan model untuk mengklasifikasikan kabupaten/kota di Pulau Sumatera dengan rata-rata peluang kesalahan prediksi sebesar 5.8%, serta memiliki hasil yang akurat dan konsisten.

Kata kunci: analisis diskriminan fisher, *k-fold cross validation*, daerah tertinggal

PENDEKATAN KLASIFIKASI DISKRIMINAN LINIER FISHER UNTUK KABUPATEN/KOTA TERTINGGAL DI PULAU SUMATERA DENGAN *K-FOLD CROSS VALIDATION*

Oleh

NADHIRA DEWIANTARI

Skripsi

Sebagai Salah Satu Syarat untuk Mencari Gelar
SARJANA MATEMATIKA

Pada

Jurusan Matematika
Fakultas Matematika dan Ilmu Pengetahuan Alam

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2023

Judul	:	**Pendekatan Klasifikasi Diskriminan Linier Fisher Untuk Kabupaten/Kota Tertinggal Di Pulau Sumatera Dengan *K-Fold Cross Validation***
Nama Mahasiswa	:	**Nadhira Dewiantari**
Nomor Pokok Mahasiswa	:	**1757031011**
Jurusan	:	**Matematika**
Fakultas	:	**Matematika dan Ilmu Pengetahuan Alam**

MENYETUJUI

1. Komisi Pembimbing

Prof. Drs. Mustofa Usman, M.A., Ph.D.	**Agus Sutrisno, S.Si., M.Si**
NIP. 19570101 198404 1 001	NIP. 19700831 199903 1 002

2. Ketua Jurusan Matematika

Dr. Aang Nuryaman, S.Si., M.Si
NIP. 19740316 200501 1 001

MENGESAHKAN

1. Tim Penguji

   Ketua : **Prof. Drs. Mustofa, M.A., Ph.D.** .....

   Sekretaris: **: Agus Sutrisno, S.Si., M.Si.** .....

   Penguji : **Drs. Nusyirwan, M.Si.** .....

2. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Dr. Eng. Suripto Dwi Yuwono, S.Si., M. T.
NIP 19740705 200003 1 001

Tanggal Lulus Ujian Skrpsi: **24 Januari 2023**

PERNYATAAN SKRIPSI MAHASISWA

Yang bertanda tangan di bawah ini:

Nama Mahasiswa	: Nadhira Dewiantari
Nomor Pokok Mahasiswa	:1757031011
Jurusan	:Matematika
Judul Skripsi	: Pendekatan Klasifikasi Diskriminan Linier Fisher untuk Kabupaten/Kota Tertinggal Di Pulau Sumatera *Dengan K-Fold Cross Validation*

Dengan ini menyatakan bahwa penelitian ini adalah hasil pekerjan saya sendiridan semua tulisan yang tertuang dalam skripsi ini telah mengikuti kaidah penulisan karya ilmiah Universitas Lampung

Bandar Lampung, 24 Januari 2023

Penulis

Nadhira Dewiantari

RIWAYAT HIDUP

Penulis bernama lengkap Nadhira Dewiantari, anak pertama dari dua bersaudara yang dilahirkan di Bandar Lampung pada tanggal 4 September oleh pasangan Bapak Dwi Supriantoro dan Ibu Siluh Putu Sudewi. Penulis memiliki adik perempuuan bernama Elnaya Pricilia.

Penulis menyelesaikan pendidikan di TK Yapindo pada tahun 2005. Pendidikan sekolah dasar di SD Yapindo pada tahun 2011. Pendidikan sekolah menengah pertama di SMP Gula Putih Mataram pada tahun 2014. Pendidikan sekolah menengah atas di SMAS Sugar Group pada tahun 2017. Kemudian penulis diterima sebaga mahasiswa S1 Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan alam (FMIPA) Universitas Lampung (Unila) pada tahun 2017. Pada periode 2017/2018 penulis terdaftar sebagai anggota GEMATIKA Himpunan Mahasiswa FMIPA Unila. Penulis pernah menjadi anggota biro Eksternal Himpunan Mahasiswa Matematika tahun 2018 dan 2019.

Sebagai bentuk penerapan ilmu perkuliahan, penulis telah melakukan Kerja Praktik (KP) selama 40 hari di Telkom Witel Lampung pada tahun 2020. Pada tahun yang sama, sebagai bentuk pengabdian kepada masyarakat, penulis telah melaksanakan Kuliah Kerja Nyata (KKN) selama 40 hari di dusun Argopeni, Kecamatan Sumberejo, Kabupaten Tanggamus.

PERSEMBAHAN

Alhamdulillahirobbil'alamin,
Puji dan syukur tiada hentinya kepada Allah Subhanahu Wata'ala atas segala nikmat dan karunia-Nya, Shalawat dan salam selalu tercurah kepada Nabi Muhammad Shallahu'Alaihi Wasallam yang menjadi contoh dan panutan kepada umat manusia.

Kupersembahkan karya sederhana ini untuk:

Ayahanda Dwi Supriantoro dan Ibunda Siluh Putu Sudewi
Terimakasih atas limpahan kasih sayang, pengorbanan, doa dan seluruh motivasi di setiap langkah penulis. Karena atas doa dan ridho kalian, Allah memudahkan setiap perjalanan hidup ini.

Adik Elnaya Pricilia
Terimakasih telah mendukung dan mendoakan setiap waktu untuk keberhasilan penulis

Guru dan Dosen
Ilmu dan pengetahuan serta begitu banyak pembelajaran hidup yang telah diberikan adalah hal terbaik yang tak kalah ku syukur. Terimakasih atas segala jasa-jasamu selama ini.

Almamater Tercinta Universitas Lampung

SANWACANA

Alhamdulillahirabbil'alaamiin, puji dan syukur penulis kepada Allah SWT atas izin serta ridho-Nya dalam menyelesaikan skripsi yang berjudul "**Pendekatan Klasifikasi Diskriminan Linier Fisher untuk Kabupaten/Kota Tertinggal Di Pulau Sumatera Dengan *K-Fold Cross Validation***". Shalawat serta salam kepada Nabi Muhammad SAW yang telah menjadi suri tauladan yang baik sepanjang masa.

Terselesaikannya skripsi ini tidak lepas dari bantuan, kerjasama dan dukungan berbagai pihak. Untuk itu, penulis ingin mengucapkan terimakasih kepada:

1. Bapak Prof. Drs. Mustofa, M.A., Ph.D. Selaku dosen pembimbing 1, yang senantiasa membimbing dan memberikan arahan, ide, kritik dan saran kepada penulis selama pembuatan skripsi ini.
2. Bapak Agus Sutrisno, S.Si., M.Si. selaku dosen pembimbing II, yang telah membimbing, memberi masukan, dan mengarahkan penulis selama proses penyusunan skripsi ini.
3. Bapak Drs. Nusyirwan, M.Si., selaku dosen penguji yang telah memberikan kritik dan saran yang membangun kepada penulis selama proses penyelesaian skripsi ini.
4. Ibunda Siluh Putu Sudewi, ayahanda Dwi Supriantoro, adik Elnaya Pricilia yang tak pernah berhenti memberi semangat, doa, dorongan, kasih sayang dan nasihat untuk selalu berjuang setiap harinya.
5. Sahabat-sahabat penulis, Renny Andrelia Antika, Syifa Nailul Fu'ikah, Della Egidia, Vina Nurmadani, Maria Ulfa, Dindha Agustina, dan Shintia Anjarwati yang senantiasa memberikan bantuan, dukungan, serta menemani suka duka penulis.
6. Teman-teman yang telah membantu penulis yang tidak dapat disebutkan satu persatu atas peran dan dukungannya dalam menyusun skripsi ini.

DAFTAR ISI

Halaman

DAFTAR TABEL

Tabel Halaman

DAFTAR GAMBAR

Gambar Halaman

I. PENDAHULUAN

1.1 Latar Belakang

Analisis diskriminan merupakan salah satu metode yang digunakan dalam analisis multivariat dengan metode dependen (hubungan antar variabel yang sudah bisa dibedakan antara peubah respon dan peubah penjelas). Pada dasarnya analisis diskriminan dapat digunakan untuk mengetahui variabel karakteristik yang membedakan kelompok populasi yang ada. Analisis diskriminan linier Fisher adalah teknik multivariat konvesional untuk pengurangan dimensi dan klasifikasi. Analisis diskriminan Fisher adalah kombinasi linier dari variabel yang diamati atau diukur yang paling baik menggambarkan pemisahan antara kelompok pengamatan. Ide dasarnya adalah untuk mengklasifikasikan atau memprediksi masalah.

Dalam analisis diskriminan akan dihasilkan suatu fungsi yang dapat menjelaskan perbedaan atau memisahkan kelompok-kelompok yang disebut fungsi diskriminan. Fungsi diskriminan merupakan kombinasi linear dari variabel-variabel bebas yang digunakan untuk menduga nilai suatu variabel tak bebas. Menurut Rencher (2002), fungsi diskriminan adalah kombinasi linier dari variabel yang paling baik memisahkan kelompok.

Pada analisis diskriminan bisa dilakukan suatu validasi keakuratan model fungsi diskriminan. Validasi keakuratan model bisa menggunakan metode *cross validation*. Seperti penelitian Arisona (2015) yang mengklasifikasikan nasabah yang menunggak dan tidak menunggak dengan metode *cross validation*. Prinsip

dasar metode *cross validation* adalah membagi keseluruhan data menjadi data *training* dan data *testing* (Davidson & Hinkley, 1997). Salah satu metode dalam *cross-validation* yaitu *k-fold cross-validation*. Menurut Berrar (2018), dalam teknik ini data set dibagi menjadi sejumlah *k*-buah partisi. Kemudian dilakukan sejumlah *k*-kali pengujian, dimana masing-masing pengujian menggunakan data partisi ke-*k* sebagai data *testing* dan memanfaatkan sisa partisi lainnya sebagai data *training*.

Berdasarkan Kementerian Negara Pembangunan Daerah Tertinggal (KNPDT) 2010-2014 yang ditetapkan dengan Perpres No. 5 tahun 2010, pengertian kabupaten/kota tertinggal adalah kabupaten/kota yang masyarakat serta wilayahnya relatif kurang berkembang dibandingkan daerah lain dalam skala nasional. Suatu daerah dikategorikan sebagai daerah tertinggal, disebabkan oleh beberapa faktor, yaitu perekonomian masyarakat, sumber daya manusia, prasarana (infrastruktur), kemampuan keuangan lokal, aksesbilitas, dan karakteristik daerah. Pulau Sumatera memiliki 10 Provinsi yang didalamnya terdapat 154 kabupaten/kota. Terdiri dari 120 kabupaten dan 34 kota, 7 diantaranya masuk sebagai kabupaten tertinggal yaitu Nias, Nias Selatan, Nias Utara, Nias Barat, Kep. Mentawai, Musi Rawas Utara, dan Pesisir Barat. Berdasarkan latar belakang tersebut penulis akan melakukan pendekatan klasifikasi diskriminan linier fisher untuk kabupaten/kota di Pulau Sumatera ke dalam kelompok kabupaten/kota tertinggal atau tak tertinggal dengan menggunakan *k-fold cross validation*.

1.2 Tujuan Penelitian

Adapun tujuan yang diharapkan dari hasil penelitian ini adalah

- Mengestimasi persamaan fungsi diskriminan linear *fisher* dengan metode *k-fold cross validation*.
- Menghitung rata-rata peluang kesalahan klasifikasi prediksi menggunakan metode *k-fold cross validation*.

- Menghitung kontribusi setiap variabel terhadap perbedaan kabupaten/kota tertinggal atau tak tertinggal di Pulau Sumatera.

1.3 Manfaat Penelitian

Adapun manfaat yang diharapkan dari hasil penelitian ini adalah

- Mengetahui faktor yang mempengaruhi pengklasifikasian kabupaten/kota di Pulau Sumatera.
- Memperdalam pengetahuan mengenai metode analisis diskriminan linear *fisher* dengan *k-fold cross validation*.

II. TINJAUAN PUSTAKA

2.1 Data Multivariat dan Vektor Acak

Menurut Johnson dan Wichern (2007), data multivariat adalah data yang diperoleh dari hasil pengukuran terhadap n observasi-observasi berdasarkan variabel-variabel random p. Secara umum data multivariat disajikan dalam bentuk matriks **X** berukuran $n \times p$:

$$\boldsymbol{X} = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{np} \end{bmatrix} \quad (1)$$

Matriks **X** memuat data yang terdiri dari seluruh data pengujian terhadap seluruh peubah penjelasnya.

Vektor acak adalah vektor yang elemennya merupakan variabel acak. Pengukuran pada baris ke-i yaitu $x_{i1}, x_{i2}, \dots, x_{ip}$ merupakan pengukuran pada individu yang sama, jika disusun sebagai vektor kolom x_i diperoleh:

$$x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix} \quad (2)$$

Maka x_i disebut sebagai pengujian vektor acak.

2.2 Matriks Varians-kovarians

Menurut Raykov dan Marcoulides (2008), matriks varians kovarians merupakan

suatu matriks simetris yang berisi varians pada diagonal utamanya dan kovarians pada elemen lainnya. Koefisien varians menggambarkan sebuah indeks tidak baku dari hubungan linear antara dua peubah penjelas.

Menurut Everitt (2005), varians populasi dari dua peubah, x_i dan x_j didefinisikan oleh:

$$Cov(x_i, x_j) = E\left[(x_i - \mu_i)(x_j - \mu_j)\right] \quad (3)$$

Kovarians dari x_i dan x_j biasanya dinotasikan oleh σ_{ij}. Jadi, varians dari peubah x_i sering dinotasikan oleh σ_{ii} dari pada σ_i^2.

Dengan p peubah $x_1, x_2, \dots, x_p$ ada p varians dan $\frac{p(p-1)}{2}$ kovarians. Secara umum, perhitungan ini dihasilkan dari suatu $p \times p$ matriks simetris **Σ**, yaitu:

$$\boldsymbol{\Sigma} = E[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})']$$

$$= E\left(\begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \\ \vdots \\ X_p - \mu_p \end{bmatrix} [X_1 - \mu_1 \quad X_2 - \mu_2 \quad \dots \quad X_p - \mu_p]\right)$$

$$= \begin{bmatrix} E(X_1 - \mu_1)^2 & E(X_1 - \mu_1)(X_2 - \mu_2) & \dots & E(X_1 - \mu_1)(X_p - \mu_p) \\ E(X_2 - \mu_2)(X_1 - \mu_1) & E(X_2 - \mu_2)^2 & \dots & E(X_2 - \mu_2)(X_p - \mu_p) \\ \vdots & \vdots & \ddots & \vdots \\ E(X_p - \mu_p)(X_1 - \mu_1) & E(X_p - \mu_p)(X_2 - \mu_2) & \dots & E(X_p - \mu_p)^2 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \mathrm{Cov}(\mathbf{X}) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{bmatrix} \quad (4)$$

dengan $\sigma_{ij} = \sigma_{ji}$. Matriks ini biasanya disebut matriks varians kovarians atau matriks kovarians. Matriks **Σ** diduga oleh matriks **S**.

S adalah penduga matriks varians kovarians kelompok ke-i yang didefinisikan oleh:

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})' \quad (5)$$

dengan $x_i' = [x_{i1}, x_{i2}, \dots, x_{ip}]$ adalah vektor pengujian untuk i pengujian.

Diagonal utama dari matriks **S** berisi varians dari peubah lainnya.

2.3 Kombinasi Linear

Kombinasi linier dari variabel $y_1, y_2, \dots, y_p$. Misal $a_1, a_2, \dots, a_p$ adalah kontanta dan bentuk kombinasi liner dari elemen vektor $\boldsymbol{y}$ sebagai berikut:

$$z = a_1 y_1 + a_2 y_2 + \cdots + a_p y_p = \boldsymbol{a'y} \tag{6}$$

dimana $\boldsymbol{a'} = (a_1, a_2, \dots, a_p)$. Jika koefisien vektor $\boldsymbol{a}$ yang sama diterapkan pada setiap $\boldsymbol{y}_i$ pada sampel, kita dapatkan:

$$z_i = a_1 y_{i1} + a_2 y_{i2} + \cdots + a_p y_{ip} = \boldsymbol{a'y}_i\,, \quad i = 1, 2, \dots, n \tag{7}$$

Rata-rata sampel dari z yaitu:

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i = \boldsymbol{a'\bar{y}} \tag{8}$$

Varians sampel dari $z_i = \boldsymbol{a'y_i}\,, \quad i = 1, 2, \dots, n$ dapat ditemukan seperti varians sampel dari $z_1, z_2, \dots, z_n$ atau langsung dari $\boldsymbol{a}$ dan $\boldsymbol{S}$, dimana $\boldsymbol{S}$ adalah matriks kovarians dari $\boldsymbol{y_1}, \boldsymbol{y_2}, \dots, \boldsymbol{y_n}$:

$$s_z^2 = \frac{\sum_{i=1}^{n}(z_i - \bar{z})^2}{n-1} = \boldsymbol{a'Sa} \tag{9}$$

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})' \tag{10}$$

Jika kita definisikan kombinasi linier lain $w = \boldsymbol{b'y} = b_1 y_1 + b_2 y_2 + \cdots + b_p y_p$, dimana $\boldsymbol{b'} = (b_1, b_2, \dots, b_p)$ adalah vektor konstanta yang berbeda dari $\boldsymbol{a'}$, maka kovarians sampel dari z dan w diberikan oleh:

$$s_{zw} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(w_i - \bar{w})}{n-1} = \boldsymbol{a'Sb}$$

Korelasi sampel antara z dan w diperoleh sebagai berikut:

$$r_{zw} = \frac{s_{zw}}{\sqrt{s_z^2 s_w^2}} = \frac{\boldsymbol{a'Sb}}{\sqrt{(\boldsymbol{a'Sb})(\boldsymbol{b'Sb})}} \tag{11}$$

(Rencher, 2002)

2.4 Analisis Diskriminan

Menurut Johnson & Wichern (2007), analisis diskriminan merupakan suatu teknik peubah ganda yang digunakan untuk memisahkan pengujian atau objek ke dalam kelompok atau himpunan yang berbeda dan untuk mengklasifikasikan objek baru ke dalam salah satu kelompok yeng telah ditentukan sebelumnya

Terdapat dua tujuan utama pemisahan kelompok dalam analisis diskriminan, yaitu:

1. Aspek deskriptif yaitu menggambarkan pemisahan kelompok, dimana fungsi linier variabel (fungsi diskriminan) digunakan untuk mendeskripsikan atau menjelaskan perbedaan antara dua atau beberapa kelompok. Tujuan dari analisis diskriminan meliputi identifikasi kontribusi p variabel untuk memisahkan kelompok dan mendapatkan hasil yang optimal dimana titik-titik tersebut dapat menjelaskan gambaran terbaik dari masing-masing kelompok.
2. Aspek prediksi yaitu mengelompokkan pengujian ke dalam kelompok, dimana fungsi linier atau kuadratik dari beberapa variabel digunakan untuk menentukan satu sampel individu atau objek ke dalam salah satu dari beberapa kelompok. Nilai-nilai yang diukur dalam vektor observasi dari individu atau objek dievaluasi oleh fungsi pengelompokkan untuk mencari kelompok dimana individu atau objek tersebut berada didalamnya.

Model dasar analisis diskriminan mirip regresi berganda. Perbedaannya adalah jika variabel dependen regresi berganda dilambangkan dengan Y, maka dalam analisis diskriminan dilambangkan dengan Z. Fungsi diskriminan dapat dituliskan sebagai berikut:

$$Z = a_1Y_1 + a_2Y_2 + \cdots + a_pY_p \tag{12}$$

dimana

Z = skor diskriminan

a_p = bobot diskriminan untuk variabel ke-p

Y_p = prediktor atau variabel ke-p

2.5 Pendekatan Fisher untuk Klasifikasi

2.5.1 Pembentukan Fungsi Diskriminan

Fisher menggunakan gagasan yang berbeda untuk membentuk fungsi diskriminan. Gagasan Fisher adalah mengubah pengamatan multivariat **y** menjadi pengamatan univariat z sehingga z diperoleh dari populasi π_1 dan π_2 yang terpisah sejauh mungkin. Fisher menyarankan untuk mengambil kombinasi linear dari **y** untuk membuat z, karena kombinasi linear dari **y** merupakan fungsi dari **y** yang cukup sederhana untuk ditangani. Pendekatan Fisher ini tidak mengasumsikan bahwa populasi-populasi harus berdistribusi normal, namun secara implisit mengasumsikan bahwa matriks-matriks kovarians populasi adalah sama, karena menggunakan penduga gabungan dari matriks-matriks kovarians.

Andaikan terdapat n_1 observasi terhadap variabel random multivariat $\boldsymbol{Y}' = [Y_1, Y_2, \dots, Y_P]$ dari populasi π_1 dan n_2 dari populasi π_2 dengan $n_1 + n_2 - 2 \geq p$, maka matriks-matriks data ditulis sebagai berikut:

$$\underset{(n_1 \times p)}{Y_1} = \begin{bmatrix} y'_{11} \\ y'_{12} \\ \vdots \\ y'_{1n_1} \end{bmatrix} \qquad \underset{(n_2 \times p)}{Y_2} = \begin{bmatrix} y'_{21} \\ y'_{22} \\ \vdots \\ y'_{2n_2} \end{bmatrix} \tag{13}$$

Vektor rata-rata sampel dan matriks kovarians dapat dihitung dengan:

$$\underset{(p \times 1)}{\bar{y}_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} y_{1i} \quad ; \quad \underset{(p \times p)}{S_1} = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(y_{1i} - \bar{y}_1)(y_{1i} - \bar{y}_1)'$$

$$\underset{(p \times 1)}{\bar{y}_2} = \frac{1}{n_2}\sum_{i=1}^{n_2} y_{2i} \quad ; \quad \underset{(p \times p)}{S_2} = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(y_{2i} - \bar{y}_2)(y_{2i} - \bar{y}_2)' \tag{14}$$

Karena diasumsikan bahwa kedua populasi mempunyai matriks kovarians Σ yang sama, maka matriks-matriks kovarians sampel S_1 dan S_2 digabung untuk mendapatkan sebuah penduga tak bias dari Σ.

$$S_{pl} = \left[\frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)}\right] S_1 + \left[\frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)}\right] S_2 \tag{15}$$

Fungsi diskriminan yang terbentuk mempunyai bentuk umum berupa *Fisher's Sample Linear Discriminant Function* (persamaan linier) yaitu:

$$z = (\bar{y}_1 - \bar{y}_2)' S_{pl}^{-1} y = \boldsymbol{a'y} \tag{16}$$

(Johnson & Wichern, 2007)

Sebuah kombinasi linear tertentu dari nilai-nilai $\boldsymbol{y}$ yang bernilai $z_{11}, z_{12}, \dots, z_{1n_1}$ untuk observasi-observasi dari populasi pertama, dan nilai-nilai $z_{21}, z_{22}, \dots, z_{2n_2}$ dari populasi kedua. Pemisahan dari dua himpunan univariat ini dijelaskan dengan selisih antara $\bar{z}_1$ dan $\bar{z}_2$ yang dinyatakan dalam satuan standar deviasi sebagai berikut:

$$pemisahan = \frac{|\bar{z}_1 - \bar{z}_2|}{s_z} \tag{17}$$

dimana $s_z^2 = \frac{\sum_{i=1}^{n_1}(z_{1i} - \bar{z}_1)^2 + \sum_{i=1}^{n_2}(z_{2i} - \bar{z}_2)^2}{n_1 + n_2 - 2}$ dengan $z_{1i} = \boldsymbol{a'y_{1i}}$, $z_{2i} = \boldsymbol{a'y_{2i}}$, $\boldsymbol{a} = \boldsymbol{S_{pl}^{-1}(\bar{y}_1 - \bar{y}_2)}$, adalah penduga gabungan dari varians. Tujuannya adalah untuk memilih kombinasi linear dari **y** yang memaksimumkan pemisahan rata-rata sampel $\bar{z}_1$ dan $\bar{z}_2$.

Karena $\frac{|\bar{z}_1 - \bar{z}_2|}{s_z}$ bisa negatif, kita menggunakan jarak kuadrat. Kombinasi linear $z = \boldsymbol{(\bar{y}_1 - \bar{y}_2)' S_{pl}^{-1} y = a'y}$ akan memaksimumkan rasio

$$\frac{(jarak\ kuadrat\ antara\ rata - rata\ sampel\ z)}{(variansi\ sampel\ z)} = \frac{(\bar{z}_1 - \bar{z}_2)^2}{s_z^2}$$

$$= \frac{\boldsymbol{(a'\bar{y}_1 - a'\bar{y}_2)^2}}{\boldsymbol{a' S_{pl} a}} = \frac{\boldsymbol{(a'(\bar{y}_1 - \bar{y}_2))^2}}{\boldsymbol{a' S_{pl} a}}$$

$$= \frac{\boldsymbol{(a'd)}^2}{\boldsymbol{a' S_{pl} a}} \tag{18}$$

Untuk semua vektor-vektor koefisien $\boldsymbol{a}$ yang mungkin, dimana $\boldsymbol{d} = (\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2)$. Dengan mensubtitusikan $\boldsymbol{a} = \boldsymbol{S_{pl}^{-1}(\bar{y}_1 - \bar{y}_2)}$ kedalam persamaan (18) maka didapatkan maksimum dari rasio pada persamaan tersebut sebagai berikut:

$$\frac{(\bar{z}_1-\bar{z}_2)^2}{s_z^2} = (\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2)'\boldsymbol{S_{pl}^{-1}}(\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2) \quad (19)$$

dimana $z = \boldsymbol{a'y}$ dengan $\boldsymbol{a} = \boldsymbol{S_{pl}^{-1}}(\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2)$. Karena $\boldsymbol{a'} = (\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2)'\boldsymbol{S_{pl}^{-1}}$, kita bisa menulis persamaan diatas menjadi:

$$\frac{(\bar{z}_1-\bar{z}_2)^2}{s_z^2} = \boldsymbol{a'}(\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2) \quad (20)$$

dimana

$\bar{y}_1$ = matriks rata-rata kelompok 1

$\bar{y}_2$ = matriks rata-rata kelompok 2

$\bar{z}_1$ = matriks rata-rata nilai diskriminan kelompok 1

$\bar{z}_2$ = matriks rata-rata nilai diskriminan kelompok 2

S_{pl}^{-1} = matriks invers kovarians gabungan

a = koefisien diskriminan

s_z = standar deviasi nilai diskriminan

(Rencher, 2002).

2.5.2 Klasifikasi Diskriminan Fisher

Terdapat $\boldsymbol{y}$ yaitu vektor pengukuran pada unit sampling baru yang ingin kita klasifikasikan kedalam salah satu dari dua kelompok (populasi). Untuk melihat apakah $\boldsymbol{y}$ lebih dekat ke $\bar{\boldsymbol{y}}_1$ atau $\bar{\boldsymbol{y}}_2$, kita periksa apakah z pada persamaan (16) lebih dekat dengan rata-rata transformasi rata-rata $\bar{z}_1$ atau $\bar{z}_2$. Kita evaluasi persamaan (16) untuk setiap pengamatan $\boldsymbol{y}_{1i}$ dari sampel pertama dan diperoleh $z_{11}, z_{12}, \ldots, z_{1n_1}$, dengan

$$\bar{z}_1 = \sum_{i=1}^{n_1} \frac{z_{1i}}{n_1} = \boldsymbol{a'}\bar{\boldsymbol{y}}_1 = (\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2)'\boldsymbol{S_{pl}^{-1}}\bar{\boldsymbol{y}}_1 \quad (21)$$

$$\bar{z}_2 = \sum_{i=1}^{n_2} \frac{z_{2i}}{n_2} = \boldsymbol{a'}\bar{\boldsymbol{y}}_2 = (\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2)'\boldsymbol{S_{pl}^{-1}}\bar{\boldsymbol{y}}_2 \quad (22)$$

Nyatakan dua kelompok dengan π_1 dan π_2. Prosedur klasifikasi linier Fisher mengklasifikasikan **y** ke π_1 jika $z = \boldsymbol{a}'\boldsymbol{y}$ lebih dekat dengan $\bar{z}_1$ daripada $\bar{z}_2$ dan mengklasifikasikan **y** ke π_2 jika z lebih dekat dengan $\bar{z}_2$. Dengan itu kita lihat bahwa z ke $\bar{z}_1$ jika:

$$z > \frac{1}{2}(\bar{z}_1 + \bar{z}_2) \tag{23}$$

Secara umum karena $\bar{z}_1$ selalu lebih besar dari $\bar{z}_2$, yang mana ditunjukkan sebagai berikut:

$$\bar{z}_1 - \bar{z}_2 = \boldsymbol{a}'(\bar{\boldsymbol{y}}_{\boldsymbol{1}} - \bar{\boldsymbol{y}}_{\boldsymbol{2}}) = (\bar{\boldsymbol{y}}_{\boldsymbol{1}} - \bar{\boldsymbol{y}}_{\boldsymbol{2}})'\boldsymbol{S}_{\boldsymbol{pl}}^{-\boldsymbol{1}}(\bar{\boldsymbol{y}}_{\boldsymbol{1}} - \bar{\boldsymbol{y}}_{\boldsymbol{2}}) > 0 \tag{24}$$

Karena $\frac{1}{2}(\bar{z}_1 + \bar{z}_2)$ adalah titik tengah, $z > \frac{1}{2}(\bar{z}_1 + \bar{z}_2)$ mengimplikasikan bahwa z lebih dekat ke $\bar{z}_1$. Dengan persamaan diatas jarak dari $\bar{z}_1$ ke $\bar{z}_2$ sama dengan jarak $\bar{\boldsymbol{y}}_1$ ke $\bar{\boldsymbol{y}}_2$.

Untuk menyatakan aturan klasifikasi dalam bentuk y, kita tulis $\frac{1}{2}(\bar{z}_1 + \bar{z}_2)$ kedalam bentuk:

$$\text{Titik tengah} = \frac{1}{2}(\bar{z}_1 + \bar{z}_2) = \frac{1}{2}(\bar{\boldsymbol{y}}_{\boldsymbol{1}} - \bar{\boldsymbol{y}}_{\boldsymbol{2}})'\boldsymbol{S}_{\boldsymbol{pl}}^{-\boldsymbol{1}}(\bar{\boldsymbol{y}}_{\boldsymbol{1}} + \bar{\boldsymbol{y}}_{\boldsymbol{2}}) \tag{25}$$

Maka aturan klasifikasinya menjadi:
Klasifikasikan $\boldsymbol{y}$ ke π_1 jika:

$$\boldsymbol{z} \geq \text{titik tengah} \tag{26}$$

Klasifikasikan $\boldsymbol{y}$ ke π_2 jika:

$$\boldsymbol{z} < \text{titik tengah} \tag{27}$$

(Rencher, 2002)

2.6 Uji Asumsi Kehomogenan Matriks Kovarians

Pendekatan Fisher tidak mengasumsikan bahwa populasi-populasi harus berdistribusi normal, namun mengasumsikan matriks kovarians populasi sama, karena menggunakan gabungan dari matriks-matriks kovarians. Salah satu asumsi saat membandingkan dua atau vektor rata-rata adalah matriks kovarians dari populasi yang berbeda adalah sama. Sebelum menggabungkan variasi antar sampel untuk membentuk matriks kovarians gabungan saat membandingkan vektor matriks, lebih baik menguji kesamaan matriks kovarians populasi. Salah satu uji yang umum digunakan untuk matriks kovarians yang sama adalah uji Box's M.

Untuk menguji kesamaan matriks kovarians ($\boldsymbol{\Sigma}$) antar g populasi multivariat hipotesis yang digunakan:

$H_0: \Sigma_1 = \Sigma_2 = \cdots = \Sigma_g$ (matrik kovarians sama)

$H_1: \Sigma_i \neq \Sigma_j$(sedikitnya ada dua kelompok yang berbeda) $i \neq j$=1,2,…g

Statistik uji yang digunakan adalah statistic Box's M yaitu:

$$u = \left[\sum_i \frac{1}{(n_i-1)} - \frac{1}{\sum_i (n_i-1)}\right]\left[\frac{2p^2+3p-1}{6(p+1)(g-1)}\right] \quad (28)$$

Maka

$$C = (1-u)M = (1-u)\left\{[\textstyle\sum_i (n_i-1)] ln\left|S_{pl}\right| - \sum_i [(n_i-1) ln|S_i|]\right\} \quad (29)$$

dengan

$$M = [\textstyle\sum_i (n_i-1)] \ln\left|S_{pl}\right| - \sum_i [(n_i-1) \ln|S_i|] \quad (30)$$

$$S_{pl} = \frac{1}{\sum_i (n_i-1)}\left\{(n_i-1)S_1 + (n_2-1)S_2 + \cdots + \left(n_g-1\right)S_g\right\} \quad (31)$$

memiliki pendekatan sebaran χ^2 dengan derajat kebebasan

$$v = g\frac{1}{2}p(p+1) - \frac{1}{2}p(p+1) = \frac{1}{2}p(p+1)(g-1) \quad (32)$$

dimana

p = jumlah variabel bebas

g = jumlah kelompok

S_{pl} = matriks varians kovarians dalam kelompok gabungan

S_i = matriks varians kovarians kelompok ke-i (i = 1,2,…,g)

n_i = ukuran sampel untuk kelompok ke- i (i =1,2,...g)

Pada tingkat signifikan α, tolak H_0 jika $C > \chi^2_{\frac{p(p+1)(g-1)}{2}}(\alpha)$ atau $p - value < \alpha$. Maka matriks kovarians antar populasi berbeda (Johnson and Wichern, 2007).

2.7 Uji Kesamaan Vektor Rata-rata

Pemisahan rata-rata yang ditransformasi, $(\bar{z}_1 - \bar{z}_2)^2/s_z^2$, yang dicapai dengan fungsi diskriminan pada $z = \boldsymbol{a}'\boldsymbol{y}$ setara dengan jarak antara vektor rata-rata $\bar{\boldsymbol{y}}_{\boldsymbol{1}}$ dan $\bar{\boldsymbol{y}}_{\boldsymbol{2}}$. Jarak standar ini sebanding dengan dua kelompok T^2 berikut. Pengujian terhadap perbedaan vektor rata-rata dapat dipandang sebagai pengujian terhadap "signifikansi" dari pemisahan yang akan dicapai. Andaikan kita memiliki populasi π_1 dan π_2 dengan matriks kovarians gabungan $\boldsymbol{\Sigma}$.

Vektor rata-rata sampel yaitu $\bar{\boldsymbol{y}}_{\boldsymbol{1}} = \frac{1}{n_1}\sum_{i=1}^{n_1} \boldsymbol{y}_{\boldsymbol{1i}}$ dan $\bar{\boldsymbol{y}}_{\boldsymbol{2}} = \frac{1}{n_2}\sum_{i=1}^{n_2} \boldsymbol{y}_{\boldsymbol{2i}}$. Tentukan $\boldsymbol{W}_1$ dan $\boldsymbol{W}_2$ untuk menjadi matriks jumlah kuadrat dan hasil kali dua sampel:

$$\boldsymbol{W}_1 = \sum_{i=1}^{n_1}(\boldsymbol{y}_{\boldsymbol{1i}} - \bar{\boldsymbol{y}}_{\boldsymbol{1}})(\boldsymbol{y}_{\boldsymbol{1i}} - \bar{\boldsymbol{y}}_{\boldsymbol{1}})' = (n_1 - 1)\boldsymbol{S}_{\boldsymbol{1}} \quad (33)$$

$$\boldsymbol{W}_2 = \sum_{i=1}^{n_2}(\boldsymbol{y}_{\boldsymbol{2i}} - \bar{\boldsymbol{y}}_{\boldsymbol{2}})(\boldsymbol{y}_{\boldsymbol{2i}} - \bar{\boldsymbol{y}}_{\boldsymbol{2}})' = (n_2 - 1)\boldsymbol{S}_{\boldsymbol{2}} \quad (34)$$

Karena $(n_1 - 1)\boldsymbol{S}_{\boldsymbol{1}}$ adalah penduga tak bias dari $(n_1 - 1)\boldsymbol{\Sigma}$ dan $(n_2 - 1)\boldsymbol{S}_{\boldsymbol{2}}$ adalah penduga tak bias dari $(n_2 - 1)\boldsymbol{\Sigma}$ kita bisa menggabungkan mereka dan memperoleh penduga tak bias dari matriks kovarians populasi gabungan $\boldsymbol{\Sigma}$:

$$\boldsymbol{S}_{pl} = \frac{1}{n_1+n_2-2}(\boldsymbol{W}_{\boldsymbol{1}} + \boldsymbol{W}_{\boldsymbol{2}}) \quad (35)$$

$$= \frac{1}{n_1 + n_2 - 2}[(n_1 - 1)\boldsymbol{S}_1 + (n_2 - 1)\boldsymbol{S}_2]$$

Kuadrat dari statistik t univariat dapat dinyatakan sebagai berikut:

$$t^2 = \frac{n_1 n_2}{n_1+n_2}(\bar{y}_1 - \bar{y}_2)\left(s_{pl}^2\right)^{-1}(\bar{y}_1 - \bar{y}_2) \quad (37)$$

Ini bisa digeneralisasikan ke p variabel dengan mensubtitusi $(\bar{\boldsymbol{y}}_{\boldsymbol{1}} - \bar{\boldsymbol{y}}_{\boldsymbol{2}})$ untuk $(\bar{y}_1 - \bar{y}_2)$ dan $\boldsymbol{S}_{pl}$ untuk s_{pl}^2 untuk mendapatkan:

$$T^2 = \frac{n_1 n_2}{n_1+n_2}(\bar{\boldsymbol{y}}_{\boldsymbol{1}} - \bar{\boldsymbol{y}}_{\boldsymbol{2}})'\boldsymbol{S}_{\boldsymbol{pl}}^{-\boldsymbol{1}}(\bar{\boldsymbol{y}}_{\boldsymbol{1}} - \bar{\boldsymbol{y}}_{\boldsymbol{2}}) \quad (38)$$

Yang berdistribusi T^2_{p,n_1+n_2-2} . Untuk melakukan pengujian dua sampel, hitung T^2, dan tolak H_0 jika $T^2 \geq T^2_{p,n_1+n_2-2}$.

Statistic T^2 dapat ditransformasikan ke statistic F menjadi:

$$F = \frac{n_1+n_2-p-1}{(n_1+n_2-2)p}T^2 \quad (39)$$

Yang berdistribusi F dengan derajat bebas $v_1 = p$ dan $v_2 = n_1 + n_2 - p - 1$.
dimana

$\boldsymbol{\Sigma}$ = matriks kovarians gabungan

$\boldsymbol{W}_1$ = matriks jumlah kuadrat dan hasil kali kelompok 1

$\boldsymbol{W}_2$ = matriks jumlah kuadrat dan hasil kali kelompok 2

Untuk menguji perbedaan rata-rata dua populasi, hipotesis yang digunakan:

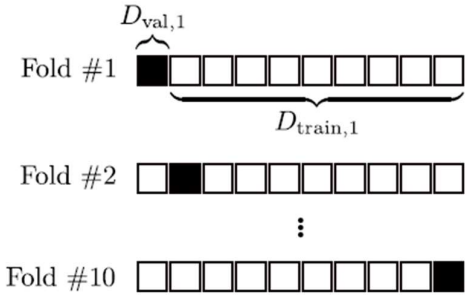
$H_0 : \mu_1 = \mu_2$ (rata-rata populasi sama)

$H_1 : \mu_1 \neq \mu_2$ (rata-rata populasi berbeda)

Tolak H_0 jika $F_{hitung} \geq F_{\alpha;p,n_1+n_2-p-1}$. Jika H_0 ditolak, maka dapat disimpulkan bahwa rata-rata populasi berbeda sehingga pemisahan antara populasi π_1 dan π_2 adalah signifikan (Rencher, 2002).

2.8 Metode *K-Fold Cross-Validation*

Cross validation atau validasi silang adalah *resampling* data untuk menilai kemampuan generalisasi model prediktif dan untuk mencegah *overfitting*. Dalam validasi silang *k-fold*, data dipartisi menjadi *k* himpunan bagian yang terputus-putus dengan ukuran yang kira-kira sama. Disini "*fold*" mengacu pada jumlah himpunan bagian yang dihasilkan. Model dibentuk menggunakan himpunan bagian *k-1*, yang merepresentasikan data *training*/latih. Kemudian model diterapkan pada himpunan bagian yang tersisa, yang mana merupakan data *testing*/uji, dan kinerjanya diukur. Prosedur ini diulang sampai masing-masing dari *k* himpunan bagian telah berfungsi sebagai himpunan *testing*. Rata-rata pengukuran kinerja *k* pada data *testing k* adalah kinerja yang divalidasi silang.



Gambar 2.1 10-*fold cross-validation*

Gambar 2.1 mengilustrasikan proses untuk *k*=10, yaitu 10-*fold cross-validation*. Dalam *fold* pertama, himpunan bagian pertama adalah data *testing*/uji dan himpunan bagian yang tersisa adalah data *training*/latih, dan seterusnya. Akurasi cross-validation, misalnya, adalah rata-rata dari sepuluh akurasi yang dicapai pada data *testing*/uji. Validasi silang seringkali melibatkan pengambilan sampel acak berlapis, yang berarti bahwa pengambilan sampel dilakukan sedemikian rupa hingga proporsi kelas dalam himpunan bagian individu mencerminkan proporsi data awal (Berrar, 2018).

2.9 Interpretasi Fungsi Diskriminan

Terdapat korespondensi yang erat antara menafsirkan fungsi diskriminan dan menentukan kontribusi setiap variabel. Dalam interpretasi, tanda koefisien diperhitungkan; dalam memastikan kontribusi, tanda diabaikan, dan koefisien diberi rank (peringkat) dalam nilai absolut. Terdapat pendekatan umum untuk menilai kontribusi masing-masing variabel (dengan adanya variabel lain) untuk memisahkan kelompok. Metode tersebut yaitu menentukan koefisien fungsi diskriminan standar, menghitung uji-F parsial untuk setiap variabel, dan menghitung korelasi setiap variabel dengan fungsi diskriminan (korelasi struktur) (Rencher, 2002).

2.9.1 Koefisien Standar

Untuk mengimbangi skala yang berbeda di antara variabel, koefisien fungsi diskriminan dapat distandarisasi. Untuk vektor observasi ke-i $\boldsymbol{y}_{1i}$ atau $\boldsymbol{y}_{2i}$ dalam kelompok 1 atau 2, dapat dinyatakan fungsi diskriminan dalam variabel standar sebagai berikut:

$$z_{1i} = a_1^* \frac{y_{1i1} - \bar{y}_{11}}{s_1} + a_2^* \frac{y_{1i2} - \bar{y}_{12}}{s_2} + \cdots + a_p^* \frac{y_{1i} \ - \bar{y}_{1p}}{s_p}, \quad i = 1,2, \dots, n_1$$

$$z_{2i} = a_1^* \frac{y_{2i1} - \bar{y}_{21}}{s_1} + a_2^* \frac{y_{2i2} - \bar{y}_{22}}{s_2} + \cdots + a_p^* \frac{y_{2ip} - \bar{y}_{2p}}{s_p}, \quad i = 1,2, \dots, n_2 \qquad (40)$$

dimana $\bar{\boldsymbol{y}}_1' = (\bar{y}_{11}, \bar{y}_{12}, \dots, \bar{y}_{1p})$ dan $\bar{y}_2' = (\bar{y}_{21}, \bar{y}_{22}, \dots, \bar{y}_{2p})$ merupakan vektor rata-rata untuk dua kelompok, dan s_r adalah standar deviasi dalam sampel dari variabel ke-r, diperoleh sebagai akar kuadrat dari elemen diagonal ke-r dari S_{pl}. Bentuk koefisien standar adalah sebagai berikut:

$$a_r^* = s_r a_r, \quad r = 1,2, \dots, p \qquad (41)$$

Dalam bentuk vektor menjadi:

$$\boldsymbol{a}^* = (diag \boldsymbol{S}_{pl})^{1/2} \boldsymbol{a} \qquad (42)$$

Koefisien standar a_r^* mencerminkan kontribusi bersama dari variabel ke fungsi diskriminan Z karena memisahkan kelompok secara maksimal. Nilai absolut dari koefisien dapat digunakan untuk mengurutkan variabel kontribusi mereka dalam memisahkan kelompok. Jika ingin menafsirkan fungsi diskriminan, tanda-tandanya dapat diperhitungkan. Tanda hanya menunjukkan bahwa variabel memberikan kontribusi positif atau negatif. Namun koefisien standar fungsi diskriminan memiliki batasan yaitu koefisien suatu variabel dapat berubah terutama jika variabel ditambahkan atau dihapus (Rencher, 2002).

2.9.2 Nilai F Parsial

Untuk variabel y_r, kita dapat menghitung uji-F parsial yang menunjukkan signifikansi y_r setelah variabel-variabel lain disesuaikan, yaitu pemisahan yang diberikan oleh y_r sebagai tambahan karena variabel lainnya. Setelah menghitung F parsial untuk masing-masing variabel, variabel kemudian dapat diberi peringkat.

Dalam kasus dua kelompok, F parsial dituliskan sebagai

$$F = (v - p + 1)\frac{T_p^2 - T_{p-1}^2}{v + T_{p-1}^2} \tag{43}$$

$$T_p^2 = \frac{n_1 n_2}{n_1 + n_2}(\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2)'\boldsymbol{S_{pl}^{-1}}(\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2) \tag{44}$$

dimana

T_p^2 = Hotelling T^2 dua sampel dengan semua variabel p

T_{p-1}^2 = statistik T^2 dengan semua variabel kecuali y_r

v = $n_1 + n_2 - 2$

p = jumlah variabel bebas

Yang berdistribusi F dengan derajat bebas 1 dan $v - p + 1$. Jika $F_{hitung} \geq F_{1,v-p+1}$ atau p-value < α maka dapat disimpulkan bahwa variabel memberikan pemisahan yang signifikan.

Nilai F parsial tidak terkait satu dimensi pemisahan kelompok, seperti halnya koefisien standar. Misalnya y_2 akan memiliki kontribusi yang berbeda di setiap

fungsi diskriminan, tetapi F parsial untuk y_2 merupakan indeks keseluruhan dari kontribusi y_2 untuk pemisahan kelompok dengan mempertimbangkan semua dimensi. Namun, nilai F parsial akan sering memeringkat variabel dalam urutan yang sama dengan koefisien standar fungsi diskriminan (Rencher, 2002).

2.9.3 Korelasi Struktur

Muatan diskriminan atau sering disebut sebagai korelasi struktur. Mengukur korelasi linier sederhana antara setiap variabel independent dan fungsi diskriminan. Muatan diskriminan mencerminkan varians yang variabel independent bagi dengan fungsi diskriminan. Salah satu karakterisktik unik dari pemuatan adalah pemuatan dapat dihitung untuk semua variabel, baik yang digunakan dalam estimasi fungsi diskriminan atau tidak.

Koefisien standar diskriminan dapat mengalami ketidakstabilan. Korelasi struktur dianggap relaif lebih valid daripada koefisien standar sebagai alat untuk menginterpretasikan kekuatan pembeda variabel independen karena sifat korelasionalnya (Hair *et al*., 2006).

Menurut Rencher (2002), untuk menghitung hubungan linier kita bisa membakukan kovarians dengan membaginya dengan standar deviasi kedua variabel. Kovariansi standar ini disebut korelasi. Korelasi sampel dari dua variabel acak y dan z adalah sebagai berikut:

$$r_{yz} = \frac{s_{yz}}{s_y s_z} = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(z_i-\bar{z})}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2 \sum_{i=1}^{p}(z_i-\bar{z})^2}} \quad (45)$$

dimana

y_i = data amatan ke-i

$\bar{y}$ = rata-rata data amatan

z_i = nilai fungsi diskriminan ke-i

$\bar{z}$ = rata-rata nilai fungsi diskriminan

2.10 Korelasi Kanonik

Ukuran hubungan antara variabel independen $y_1, y_2, \dots, y_p$ dan variabel pengelompokkan dependent i terkait dengan $\boldsymbol{\mu}_i, i = 1,2, \dots, k$. Ukuran ini mencoba menjawab pertanyaan, seberapa baik variabel memisahkan kelompok. Telah diketahui bahwa statistik Roy θ berfungsi sebagai ukuran hubungan seperti R^2, karena merupakan rasio antara jumlah total kuadrat untuk fungsi diskriminan pertama, $z_1 = \boldsymbol{a}_1'\boldsymbol{y}$:

$$\eta_\theta^2 = \theta = \frac{\lambda_1}{1+\lambda_1} = \frac{SSH_{(z_1)}}{SSE_{(z_1)}+SSH_{(z_1)}} \tag{46}$$

$$\lambda_1 = \frac{SSH_{(z_1)}}{SSE_{(z_1)}} \tag{46}$$

$$SSH = n\sum_{i=1}^{k}(\bar{z}_i - \bar{z})^2 \tag{47}$$

$$SSE = \sum_{ij}(z_{ij} - \bar{z}_i)^2 \tag{48}$$

dimana

z_{ij} = nilai diskriminan dari fungsi diskriminan ke-i untuk objek ke-j

$\bar{z}_i$ = rata-rata nilai diskriminan ke-i

$\bar{z}$ = rata-rata nilai diskriminan keseluruhan

SSH = jumlah kuadrat antar kelompok

SSE =,jumlah kuadrat dalam kelompok

Interpretasi lain dari η_θ^2 adalah korelasi kuadrat maksimum antara fungsi diskriminan dengan kombinasi linier terbaik dari variabel keanggotaan kelompok. Korelasi maksimal disebut korelasi kanonik. Korelasi kanonik kuadrat dapat dihitung untuk setiap fungsi diskriminan:

$$r_i^2 = \frac{\lambda_i}{1+\lambda_i}, \qquad i = 1,2, \dots, s \tag{49}$$

dimana λ adalah nilai eigen (Rencher, 2002).

2.11 APER (*Apparent Error Rate*)

Satu hal yang penting untuk menilai performa dari semua prosedur klasifikasi adalah menghitung "tingkat kesalahan", atau peluang kesalahan klasifikasi.
Ukuran yang dapat digunakan adalah *Apparent Error Rate* (APER). APER (Apparent Error Rate) merupakan bagian pengujian yang mengalami kesalahan klasifikasi menurut fungsi klasifikasi. Nilai APER menyatakan fraksi atau proporsi sampel yang salah diklasifikasikan oleh fungsi klasifikasi.Tingkat kesalahan dapat dihitung dari confusion matrix yang menunjukkan keanggotaan kelompok aktual dan prediksi. Untuk n_1 dari grup 1 dan n_2 dari grup 2, bentuk tabel klasifikasi sebagai berikut.

Tabel 2.1 Klasifikasi Prediksi

Kelompok Aktual	Kelompok Prediksi		Jumlah Observasi
	π_1	π_2	
π_1	n_{1C}	$n_{1M} = n_1 - n_{1C}$	n_1
π_2	$n_{2M} = n_2 - n_{2C}$	n_{2C}	n_2

dimana

n_{1C} = jumlah obyek dari π_1 tepat diklasifikasikan sebagai π_1

n_{1M} = jumlah obyek dari π_1 salah diklasifikasikan sebagai π_2

n_{2C} = jumlah obyek dari π_2 tepat diklasifikasikan sebagai π_2

n_{2M} = jumlah obyek dari π_2 salah diklasifikasikan sebagai π_1

$$APER = \frac{n_{1M}+n_{2M}}{n_1+n_2} \tag{50}$$

yang disebut sebagai proporsi objek dalam kumpulan percobaan yang salah diklasifikasikan (Johnson dan Wichern, 2007).

2.12 Uji Keakuratan

Persentase keseluruhan yang diklasifikasikan dengan benar disebut hit ratio. Akurasi prediksi dari fungsi diskriminan diukur dengan hit ratio. Ukuran tersebut membandingkan banyaknya sampel yang diklasifikasikan benar dengan total sampel yaitu:

$$Hit\ Ratio = (n_{benar}/N) \times 100\% \quad (51)$$

Perbandingan standar untuk hit ratio dengan ukuran kelompok yang tidak sama yaitu dengan kriteria kemungkinan proporsional (*proportional chance criterion*). Supaya prediksi akurat, akurasi (hit ratio) paling sedikit 1.25% lebih besar dari kriteria kemungkinan proporsional. Rumus kriteria kemungkinan proporsional sebagai berikut:

$$C_{pro} = p^2 + (1-p)^2 \quad (52)$$

dimana

p = proporsi jumlah sampel di kelompok 1

$(1-p)$ = proporsi jumlah sampel di kelompok 2

N = jumlah sampel total

n_{benar} = jumlah sampel yang diklasifikasikan benar

Jika kriteria kemungkinan proporsional lebih kecil dari hit ratio, maka akurasi prediksi dapat diterima (Hair *et al*., 2006).

2.13 Press's Q

Press's Q adalah uji statistik untuk kekuatan diskriminatif dari matriks klasifikasi. Ukuran sederhana ini membandingkan jumlah klasifikasi yang benar dengan ukuran sampel total dan jumlah kelompok. Nilai yang dihitung kemudian dibandingkan dengan nilai χ^2. jika melebihi nilai kritis ini, maka matriks

klasifikasi dapat dianggap baik secara statistic. Statistik Q dihitung dengan rumus berikut :

$$Press's\ Q = \frac{[N-(nK)]^2}{N(K-1)} \quad (53)$$

dimana

N = ukuran sampel total

n = jumlah sampel yang diklasifikasikan benar

K = jumlah kelompok

Jika $Press's\ Q > \chi^2_{1,\alpha}$ maka analisis diskriminan stabil (Hair *et al.*, 2006).

2.14 Scatter Plot

Untuk memplot dua fungsi diskriminan pertama untuk vektor pengamatan individual $\boldsymbol{y}_{ij}$, cukup hitung $z_{1i} = \boldsymbol{a}_1'\boldsymbol{y}_{ij}$ dan $z_{2ij} = \boldsymbol{a}_2'\boldsymbol{y}_{ij}$ for $i = 1,2,\dots,k; j = 1,2,..,n_i$, dan memplot scatter plot dari $N = \sum_i n_i$ N= nilai dari

$$\boldsymbol{z}_{ij} = \begin{pmatrix} z_{1ij} \\ z_{2ij} \end{pmatrix} = \begin{pmatrix} \boldsymbol{a}_1'\boldsymbol{y}_{ij} \\ \boldsymbol{a}_2'\boldsymbol{y}_{ij} \end{pmatrix} = \begin{pmatrix} \boldsymbol{a}_1' \\ \boldsymbol{a}_2' \end{pmatrix}\boldsymbol{y}_{ij} = \boldsymbol{A}\boldsymbol{y}_{ij}$$

Vektor rata-rata yang ditransformasi,

$$\bar{\boldsymbol{z}}_i = \begin{pmatrix} \boldsymbol{a}_1' \\ \boldsymbol{a}_2' \end{pmatrix}\bar{\boldsymbol{y}}_i = \boldsymbol{A}\bar{\boldsymbol{y}}_i, \quad i = 1,2,\dots,k$$

Harus diplot bersama dengan nilai-nilai individu $\boldsymbol{z}_{ij}$. Dalam beberapa kasus, sebuah plot hanya akan menampilkan vektor rata-rata $\bar{\boldsymbol{z}}_i, \bar{\boldsymbol{z}}_2, \dots, \bar{\boldsymbol{z}}_k$ yang telah ditransformasikan (Rencher, 2002).

III. METODOLOGI PENELITIAN

3.1 Waktu dan Tempat Penelitian

Penelitian ini dilakukan pada semester ganjil tahun akademik 2022/2023, bertempat di Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.

3.2 Data Penelitian

Data yang digunakan dalam laporan ini adalah faktor penyebab daerah tertinggal untuk wilayah kabupaten/kota di Sumatera tahun 2020. Klasifikasi daerah tempat tinggal dibagi dua, yaitu kelompok 1 daerah tidak tertinggal dan kelompok 2 daerah tertinggal. Terdapat 5 peubah penjelas, yaitu $Y_1 =$ angka harapan hidup (%), $Y_2 =$ jumlah SMP, $Y_3 =$ persentase rumah tangga pengguna listrik (%), $Y_4 =$ persentase penduduk miskin (%), dan $Y_5 =$ jumlah klinik KB. Data tersebut merupakan data yang diperoleh dari Badan Pusat Statistik setiap Provinsi di Sumatera. Pada penelitian ini penulis akan membuat fungsi diskriminan untuk mengetahui perbedaan daerah tertinggal dan tidak tertinggal di Sumatera dengan *k-fold cross validation*.

3.3 Metode Penelitian

Metode yang digunakan dalam penulisan skripsi ini adalah studi pustaka, yaitu dengan mempelajari buku-buku teks penunjang yang berhubungan dengan tugas akhir ini. Kemudian digunakan *software* SAS dan R dalam pengujian asumsi dan analisis data.

Dalam penelitian ini, langkah-langkah yang dilakukan adalah sebagai berikut.

1. Melakukan analisis statistika deskriptif terhadap data.
2. Menguji asumsi kehomogenan matriks varians kovarians dengan uji Box M.
3. Menguji kesamaan vektor rata-rata.
4. Melakukan *k-fold cross validation* pada data dengan $k = 7$.
5. Menentukan fungsi diskriminan.
6. Menginterpretasi hasil analisis diskriminan.
7. Menghitung rata-rata persentase kesalahan prediksi dalam pengklasifikasian dengan APER.
8. Menguji rata-rata keakuratan pengklasifikasian menggunakan C_{pro} dan *Hit ratio*.
9. Menguji kestabilan pengklasifikasian menggunakan $Press's\ Q$.

V. KESIMPULAN

Dari hasil analisis dan pembahasan, maka dapat diambil kesimpulan sebagai berikut:

1. Model pengklasifikasian kabupaten/kota di Pulau Sumatera dengan teknik analisis diskriminan fisher menggunakan 7-*fold cross validation* diperoleh yaitu:

   $Z = -0.20929\, Y_1 + 0.00459\, Y_2 + 0.48341\, Y_3 - 0.44576\, Y_4 + 0.00578\, Y_5$

   Dengan titik tengah yaitu 19.20777

2. Model diskriminan fisher tersebut mempunyai rata-rata peluang kesalahan klasifikasi prediksi yaitu sebesar 5.8% dan mempunyai hasil yang akurat dan konsisten.

3. Variabel yang memberikan kontribusi terhadap perbedaan daerah tertinggal dan tak tertinggal yaitu Y_3 (persentase rumah tangga pengguna listrik) sebesar 85.6%, variabel Y_4 (persentase penduduk miskin) sebesar 50.1%, Y_5 (jumlah klinik KB) memberikan perbedaan sebesar 9.4%, dan Y_2 (jumlah SMP) memberikan perbedaan sebesar 6.2%, dan Y_1 (angka harapan hidup) sebesar 3%.

DAFTAR PUSTAKA

Anton, H. & Rorres, C. 2003. *Elementary Linear Algebra*. Ninth Edition. John Wiley & Sons, Inc., New York.

Badan Pusat Statistik. 2021. *Provinsi Sumatera Utara Dalam Angka 2021.* Badan Pusat Statistik, Jakarta.

Berrar, D. 2018. *Cross-validation*. Tokyo Institute of Technology, Japan.

Davidson, A. & Hinkley, D. 1997. *Bootstrap Methods and Their Application*. Cambridge University Press, New York.

Everitt, B.,S. 2005. *An Rand-S-PLUS Companion to Multivariate Analysis.* Springer, London.

Hair, *et al*. 2006. *Multivariate Data Analysis.* Prentice Hall, New Jersey.

Johnson, R.A. & Wichern, D.W. 2007. *Applied Multivariate Statistical Analysis* Sixth Edition. Prentice Hall, Inc., New York.

Raycov, T. & Marcoulides, G. 2008. *An Introduction to Applied Multivariate Analysis*. Sixth Edition. Prentice Hall, Inc., New York.

Rencher, A.C. 2002. *Methods of Multivariate Analysis*. Second Edition. John Wiley and Sons, Inc. Canada.