PENDUGAAN *MISCLASSIFICATION ERROR* DENGAN METODE RESUBSTITUSI DAN *JACKKNIFE* PADA ANALISIS DISKRIMINAN FISHER

(Skripsi)

Oleh

Vera Yuni Arsa

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2022

ABSTRAK

PENDUGAAN *MISCLASSIFICATION ERROR* DENGAN METODE RESUBSTITUSI DAN *JACKKNIFE* PADA ANALISIS DISKRIMINAN FISHER

Oleh

VERA YUNI ARSA

Analisis diskriminan merupakan suatu teknik analisis data multivariat yang digunakan untuk mengklasifikasikan objek ke dalam populasi-populasi yang berbeda, yaitu dengan menggunakan suatu aturan klasifikasi berdasarkan sampel latihan yang telah diketahui asal populasinya. Namun, setiap aturan klasifikasi tidak selalu memberikan ketepatan dalam mengklasifikasikan objek. Sehingga dalam setiap aturan klasifikasi tersebut akan selalu terdapat kesalahan klasifikasi. Tingkat kesalahan klasifikasi ini dapat diduga dengan menggunakan metode pendugaan, dua diantaranya adalah metode Resubstitusi dan *Jackknife*.

Penelitian ini bertujuan untuk menduga tingkat kesalahan klasifikasi menggunakan metode Resubstitusi dan *Jackknife* dalam analisis diskriminan Fisher. Data yang digunakan adalah data simulasi yang dihasilkan menggunakan *software* SAS 9.4 yang berisi dua populasi dengan matriks kovarians yang sama yaitu $\boldsymbol{\Sigma_1} = \boldsymbol{\Sigma_2}$. Dua kasus yang diperhatikan yaitu populasi dengan 1) matriks kovarian diagonal dan 2) matriks kovarian nondiagonal. Perbandingan kedua metode tersebut dianalisis berdasarkan kuadrat tengah galat (KTG) dari menduga kesalahan klasifikasi. Hasil penelitian ini menunjukkan bahwa penduga *Jackknife* baik dalam menduga *misclassification error* pada kasus pertama hanya untuk ukuran sampel kecil. Sementara itu, penduga Resubstitusi cenderung lebih baik pada ukuran sampel yang besar. Namun, penduga Resubstitusi selalu lebih baik daripada *Jackknife* untuk kasus kedua. Dapat juga disimpulkan bahwa semakin banyak peubah dan ukuran sampel yang digunakan dalam simulasi semakin kecil nilai MSE dari menduga kesalahan klasifikasi yang diperoleh dari penduga Resubstitusi dan *Jackknife*.

Kata Kunci: *Analisis Diskriminan Fisher, Resubstitusi, Jackknife, Misclassification error*

ABSTRACT

ESTIMATION OF MISCLASSIFICATION ERROR USING RESUBSTITUTION AND JACKKNIFE METHODS IN FISHER DISCRIMINANT ANALYSIS

By

VERA YUNI ARSA

Discriminant analysis is a multivariate data analysis technique that is used to classify objects into different populations, by using a classification rule based on an exercise sample whose origin is known. However, each classification rule does not always provide accuracy in classifying objects. So that in each classification rule there will always be a classification error. The level of misclassification can be estimated by using estimation methods, two of them are Resubstitution and Jackknife methods.

This study aimed to predict misclassification rate using the Resubstitution and Jackknife methods in Fisher discriminant analysis. The data used were simulated data generated using SAS 9.4 software containing two populations with equal covariance matrices i.e $\boldsymbol{\Sigma_1 = \Sigma_2}$. Two cases were concerned namely populations with 1) diagonal covariance matrix and 2) nondiagonal covariance matrices. The comparison of the two methods was analyzed based on the mean squared error (MSE) of their misclassification estimates. The results of this study indicate that the Jackknife estimator was satisfactory in the misclassification error estimation of the first case only for small sample sizes. Meanwhile, the Resubstitution estimator tends to be better at large sample sizes. However, the Resubstitution estimator is always better than the Jackknife for the second case. It also can be concluded that the more variables and the sample sizes used in the simulation the smaller the value of the MSE of misclassification estimates obtained from Resubstitution and Jackknife estimators.

Keywords: *Fisher Discriminant Analysis, Resubstitution, Jackknife, Misclassification error*

PENDUGAAN *MISCLASSIFICATION ERROR* DENGAN METODE RESUBSTITUSI DAN *JACKKNIFE* PADA ANALISIS DISKRIMINAN FISHER

Oleh

Vera Yuni Arsa

Skripsi

Sebagai salah satu syarat untuk mencapai gelar
SARJANA MATEMATIKA

Pada

Jurusan Matematika
Fakultas Matematika dan Ilmu Pengetahuan Alam

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2022

Judul Skripsi : **PENDUGAAN *MISCLASSIFICATION ERROR* DENGAN METODE RESUBSTITUSI DAN *JACKKNIFE* PADA ANALISIS DISKRIMINAN FISHER**

Nama Mahasiswa : *Vera Yuni Arsa*

Nomor Pokok Mahasiswa : 1517031150

Program Studi : Matematika

Fakultas : Matematika dan Ilmu Pengetahuan Alam

MENYETUJUI

1. Komisi Pembimbing

Dr. Khoirin Nisa, S.Si., M.Si.	**Amanto, S.Si., M.Si.**
NIP 19740726 200003 2 001	NIP 19730314 200012 1 002

2. Ketua Jurusan Matematika

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

MENGESAHKAN

1. Tim Penguji

Ketua : **Dr. Khoirin Nisa, S.Si., M.Si.** ....

Sekretaris : **Amanto, S.Si., M.Si.** ....

Penguji
Bukan Pembimbing : **Drs. Eri Setiawan, 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 Skripsi : **28 Januari 2022**

PERNYATAAN SKRIPSI MAHASISWA

Yang bertanda tangan di bawah ini :

Nama	: **Vera Yuni Arsa**
Nomor Pokok Mahasiswa	: **1517031150**
Jurusan	: **Matematika**
Judul Skripsi	: **Pendugaan *Misclassification Error* dengan Metode Resubstitusi dan *Jackknife* pada Analisis Diskriminan Fisher**

Dengan ini menyatakan bahwa penelitian ini adalah hasil pekerjaan saya sendiri. Dan apabila kemudian hari terbukti bahwa skripsi ini merupakan hasil salinan atau dibuat oleh orang lain, maka saya bersedia menerima sanksi sesuai dengan ketentuan akademik yang berlaku.

Bandar Lampung, Juni 2022

Yang Menyatakan,

Vera Yuni Arsa

RIWAYAT HIDUP

Penulis bernama lengkap Vera Yuni Arsa, dilahirkan pada tanggal 01 Februari 1997 di Desa Riang Bandung, Sumatera Selatan. Penulis merupakan putri sulung dari Bapak Muslim dan Ibu Yusroh, dan kakak dari Tiara Septiana dan Hafid Zuddin Arsyad.

Penulis menempuh pendidikan taman kanak-kanak di TK Yayasan Pendidikan Islam Al-Lathifiyah Jakarta Timur pada tahun 2002 sampai tahun 2003. Kemudian melanjutkan ke sekolah dasar di SD Negeri Jatinegara 02 Pagi Jakarta Timur pada tahun 2003 sampai 2009. Kemudian melanjutkan ke sekolah menengah pertama di SMP Negeri 194 Jakarta Timur pada tahun 2009 sampai 2012. Kemudian belajar pada jenjang sekolah menengah atas di SMA Negeri 107 Jakarta Timur pada tahun 2012- sampai 2015.

Pada tahun 2015, melalui jalur SBMPTN, penulis diterima dan terdaftar sebagai mahasiswa S1 Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.

Di awal tahun 2018 tepatnya pada tanggal 18 Januari sampai 28 Februari 2018 penulis melakukan Kerja Praktik (KP) di Kantor AJB Bumiputera 1912 cabang Tanjung Karang jalan Raden Intan no.95, Bandar Lampung. Pada pertengahan tahun 2018, sebagai bentuk aplikasi bidang ilmu kepada masyarakat, penulis telah melaksanakan Kuliah Kerja Nyata (KKN) selama 40 hari di Pekon Srimengaten, Kecamatan Pulaupanggung, Kabupaten Tanggamus.

KATA INSPIRASI

MOTTO

"Kebahagiaan orang tua adalah kunci keberhasilanku"
(Penulis)

"Sesungguhnya bersama kesulitan ada kemudahan"
(Q.S. Al-Insyirah : 6)

"Di atas langit masih ada langit, di bawah tanah masih ada tanah"
(Anonim)

"Apapun yang orang-orang katakan. Aku hanya hidup sebagaimana yang aku inginkan, dipandu oleh keyakinanku sendiri"
(Intro: Never Mind-BTS)

PERSEMBAHAN

Alhamdulillah Wa Syukurillah.
Kuucapkan syukur dan terimakasihku kepada Allah SWT.
yang selalu memberikan petunjuk dan kemudahan dalam
menyelesaikan skripsi ini.

Kepada dua orang yang paling berharga dalam hidupku, Bapak dan Mama. Ketika dunia menutup pintunya padaku, kalian membuka lengan dengan sangat lebar untukku. Terima kasih karena selalu menjagaku dalam doa-doa serta selalu membiarkanku mengejar impianku apapun itu. **Kupersembahkan sebuah karya kecil ini yang menjadi bukti kudapatkan gelar, Vera Yuni Arsa, S.Mat.**

Semua usaha dan kerja keras yang kulakukan untuk membanggakan Bapak dan Mama, semoga dengan karya kecilku ini bisa menambah kebahagiaan kalian.

Terima kasih untuk kedua Adikku dan sahabat-sahabatku karena selalu mendukungku dan bersamaku.

SANWACANA

Puji dan syukur kepada Allah SWT, atas segala rahmat dan hidayah-Nya sehingga penulis dapat menyelesaikan skripsi ini. Shalawat dan salam semoga selalu tercurah kepada Nabi Muhammad SAW.

Dalam penulisan skripsi ini, penulis banyak mendapatkan bantuan, bimbingan, saran, dan motivasi dari berbagai pihak. Untuk itu pada kesempatan ini penulis ingin menyampaikan terima kasih kepada:

1. Ibu Dr. Khoirin Nisa, S.Si., M.Si., sebagai Dosen Pembimbing I yang telah banyak meluangkan waktu dan memberikan bimbingan, saran, serta dukungan selama penyusunan skripsi.
2. Bapak Amanto, S.Si., M.Si., sebagai Dosen Pembimbing II yang telah memberikan saran serta arahan kepada penulis.
3. Bapak Drs. Eri Setiawan, M.Si., sebagai Dosen Penguji yang telah memberikan saran serta evaluasi kepada penulis.
4. Bapak Ir. Warsono, M.S., Ph.D., dan Bapak Drs. Tiryono Ruby, M.Sc., Ph.D., sebagai Dosen Pembimbing Akademik yang telah memberikan pengarahan selama masa perkuliahan.
5. Bapak Dr. Aang Nuryaman, S.Si., M.Si., sebagai Ketua Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam.
6. Bapak Dr. Eng. Suripto Dwi Yuwono, S.Si., M.T., sebagai Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam.
7. Seluruh dosen, staf, dan karyawan Jurusan Matematika FMIPA Universitas Lampung yang telah memberikan ilmu serta bantuan kepada penulis.

8. Bapak, Mama, dan Adik-adikku serta seluruh keluarga yang selalu memberikan doa dan kasih sayang, mendukung dan memotivasi penulis agar dapat menjadi kebanggaan keluarga dan meraih kesuksesan.
9. Sahabat-sahabat Kamar 13: Fitri, Hana, Ditha, Suci, dan Josa yang mendukung, memotivasi, memberikan bantuan, dan mewarnai masa perkuliahan penulis.
10. Teman- teman satu bimbingan skripsi yang banyak membantu dan memberikan saran kepada penulis.
11. Teman-teman Jurusan Matematika Angkatan 2015 yang telah memberikan semangat dan ide kepada penulis.
12. Sahabat-sahabat terbaik penulis dari zaman SMP, FEIGOVA yang selalu mendukung dan menghibur penulis di saat sedih dan susah.
13. Semua pihak yang terlibat dalam penyelesaian skripsi ini yang tidak dapat penulis tuliskan satu per satu.

Penulis berharap skripsi ini dapat bermanfaat bagi penulis khususnya dan bagi pembaca pada umumnya. Penulis menyadari bahwa masih banyak kekurangan dalam skripsi ini. Untuk itu, penulis mengharapkan kritik dan saran yang membangun untuk skripsi ini.

Bandarlampung, Juni 2022
Penulis

Vera Yuni Arsa

DAFTAR ISI

Halaman

DAFTAR TABEL

Tabel **Halaman**

DAFTAR GAMBAR

Gambar **Halaman**

I. PENDAHULUAN

1.1 Latar Belakang dan Masalah

Analisis diskriminan merupakan suatu teknik analisis data multivariat yang digunakan untuk mengklasifikasikan objek ke dalam populasi-populasi yang berbeda, yaitu dengan menggunakan suatu aturan klasifikasi yang dibangun berdasarkan sampel latihan (*training sample*) yang telah diketahui asal populasinya. Aturan klasifikasi tersebut selanjutnya digunakan untuk mengklasifikasikan objek yang baru ke dalam populasi-populasi yang ada. Aturan klasifikasi disebut juga sebagai fungsi diskriminan.

Pada kasus dua kelompok, analisis diskriminan dibedakan menjadi dua jenis, yaitu analisis diskriminan linear dan analisis diskriminan kuadratik. Analisis diskriminan kuadratik digunakan jika matriks kovarian kedua populasi diasumsikan berbeda, sedangkan analisis diskriminan linear digunakan jika matriks kovarian kedua populasi diasumsikan sama. Fungsi diskriminan linear yang paling awal dan paling terkenal adalah fungsi diskriminan linear Fisher (Nisa, 2009).

Karena tujuan utama dari analisis diskriminan adalah mengklasifikasikan data, maka merupakan suatu hal yang sangat penting untuk mengetahui peluang kesalahan klasifikasi atau tingkat kesalahan klasifikasi. Menurut Mangku (2002), ada tiga jenis tingkat kesalahan, yaitu tingkat kesalahan optimal, tingkat kesalahan bersyarat, dan tingkat kesalahan harapan. Tingkat kesalahan optimal dan tingkat kesalahan harapan sulit untuk diduga, sehingga banyak penelitian yang terfokus pada tingkat kesalahan bersyarat. Yaitu tingkat kesalahan yang diduga berdasarkan statistik yang dihitung dari sampel percobaan.

Dalam menduga tingkat kesalahan bersyarat ada dua macam penduga yang dapat digunakan, yaitu penduga parametrik dan penduga empirik. Penduga parametrik terdiri dari penduga Okamoto, penduga Mclachan, dan penduga diperhalus. Sedangkan penduga empirik terdiri dari penduga Resubstitusi, Penduga U, penduga $\overline{U}$, dan penduga *Jackknife*. Berbagai penelitian tentang penaksiran tingkat kesalahan telah dilakukan. Diantaranya Lachenbruch & Mickey (1968), Snapinn & Knoke (1984), Nisa (2009), dan Osuji, dkk (2013).

Berdasarkan uraian di atas, penulis tertarik melakukan analisa dalam menduga *misclassification error* dalam fungsi diskriminan linear Fisher, dengan penduga Resubstitusi dan *Jackknife*.

1.2 Tujuan Penelitian

Tujuan penelitian ini adalah untuk mengetahui penduga kesalahan klasifikasi (*Misclassification Error*) dengan metode Resubstitusi dan metode *Jackknife* pada analisis diskriminan Fisher secara empiris, yaitu dengan simulasi data.

1.3 Manfaat Penelitian

Adapun manfaat dari penelitian ini adalah:

1. Memperdalam pengetahuan mengenai pendugaan tingkat kesalahan klasifikasi (*Misclassification Error*), khususnya mengenai metode *Jackknife* dan Resubstitusi.
2. Sebagai sumbangan pemikiran mengenai penduga *Misclassification Error* dengan metode Resubstitusi dan *Jackknife* pada analisis Diskriminan Fisher

II. TINJAUAN PUSTAKA

2.1 Konsep-Konsep Aljabar Matriks

2.1.1 Definisi Matriks

Matriks adalah susunan empat persegi panjang atau bujur sangkar dari bilangan-bilangan yang diatur dalam baris dan kolom ditulis antara dua tanda kurung, yaitu ( ) atau [ ] (Gazali, 2005).

Matriks **A** berukuran m x n, dapat ditulis sebagai berikut:

$$\boldsymbol{A_{mxn}} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

2.1.2 Definisi Transpos Matriks

Jika **A** adalah sebarang matriks m x n, maka transpos **A** dinyatakan oleh $\mathbf{A}^{\mathbf{T}}$ dan didefinisikan dengan matriks n x m yang kolom pertamanya adalah baris pertama dari **A**, kolom keduanya adalah baris kedua dari **A**, dan seterusnya (Anton, dkk, 2019).

2.1.3 Definisi Matriks Simetrik

Matriks bujur sangkar **A** dikatakan simetrik, jika $\mathbf{A}^{\mathbf{T}}$ = **A** (Anton, dkk, 2019).

2.1.4 Definisi Invers Matriks

Jika **A** adalah matriks bujur sangkar, dan jika terdapat matriks **B** yang ukurannya sama sedemikian rupa sehingga **AB** = **BA** = **I**, maka **A** disebut dapat dibalik (*invertible*) dan **B** dinamakan invers (*inverse*) dari **A** (Anton, dkk, 2019).

2.1.5 Definisi Nilai Tengah dan Matriks Kovarian

Menurut Johnson & Wichern (2007), nilai tengah dan kovarian dari vektor peubah acak **X** berukuran p x l dapat ditampilkan dalam bentuk matriks. Nilai harapan dari setiap elemen (peubah acak) terdapat dalam vektor nilai tengah $\boldsymbol{\mu}$ = E(**X**). Varians σ_{ii} yang berjumlah p dan kovarian berbeda σ_{ik} $(i < k)$ berjumlah p(p – 1)/2, terdapat dalam matriks simetrik varians kovarians, yaitu

$$\boldsymbol{\Sigma} = E(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^{t} \tag{2.1}$$

Secara spesifik dituliskan sebagai berikut:

$$E(\mathbf{X}) = \begin{bmatrix} E(\boldsymbol{X_1}) \\ E(\boldsymbol{X_2}) \\ \vdots \\ E(\boldsymbol{X_p}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \\ \vdots \\ \boldsymbol{\mu_p} \end{bmatrix} = \boldsymbol{\mu} \tag{2.2}$$

dan

$\boldsymbol{\Sigma}$ = E($\mathbf{X}$ – $\boldsymbol{\mu}$)($\mathbf{X}$ – $\boldsymbol{\mu}$)^t

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

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

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

atau

$$\boldsymbol{\Sigma} = \mathrm{Cov}(\mathbf{X}) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix} \qquad (2.3)$$

2.2 Analisis Peubah Ganda

Menurut Sartono, dkk (2003), analisis peubah ganda dilakukan pada data yang memiliki n buah pengamatan dan disetiap pengamatan dilakukan pengukuran p buah karakteristik atau peubah, misalnya $\mathbf{X_1}$, $\mathbf{X_2}$, …, $\mathbf{X_p}$. Data tersebut dapat digambarkan sebagai matriks n x p, yaitu : $\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$

Pengukuran pada baris ke-i, yaitu $x_{i1}, x_{i2}, \dots, x_{ip}$, merupakan pengukuran pada objek yang sama. Jika pengamatan tersebut disusun sebagai vektor kolom x_i maka diperoleh:

$$\boldsymbol{x_i} = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix} \tag{2.4}$$

$\boldsymbol{x_i}$ disebut pengamatan peubah ganda.

2.3 Analisis Diskriminan

Analisis diskriminan adalah teknik statistika yang dipergunakan untuk mengklasifikasikan suatu individu atau observasi ke dalam suatu kelas atau kelompok berdasarkan sekumpulan peubah-peubah (Johnson & Wichern, 2007). Analisis diskriminan adalah suatu metode yang dapat menghasilkan klasifikasi yang terbaik antara berbagai macam populasi, klasifikasi dilakukan dengan fungsi diskriminan yang merupakan kombinasi linear dari peubah-peubah bebasnya. Analisis diskriminan mempunyai asumsi bahwa data berasal dari distribusi multivariat normal dan matriks kovarian semua kelompok sama.

$$Z_{jk} = a + W_1X_{1k} + W_2X_{2k} + \cdots + W_nX_{nk} \tag{2.5}$$

dengan

Z_{jk} = Nilai diskriminan Z dari fungsi diskriminan j untuk objek k

a = Intersep

W_i = Bobot diskriminan untuk variabel bebas i

X_{ik} = Variabel bebas i untuk objek k

(Hair, dkk, 2019)

2.3.1 Analisis Diskriminan Fisher

R. A Fisher bekerja untuk mendapatkan fungsi diskriminan pada kasus yang terdiri atas hanya dua populasi, melalui suatu pencarian fungsi linear sedemikian rupa sehingga kedua populasi terpisah sejauh mungkin. Metode Fisher tidak mengharuskan kenormalan multivariate pada peubah-peubah penjelasnya namun mengasumsikan kesamaan matriks varian kovarian (Johnson & Wichern, 2007). Misalkan Π_t adalah populasi yang mengikuti sebaran $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$, dengan t = 1 dan 2, dan **x** adalah pengamatan berdimensi p yang diklasifikasikan ke dalam satu dari dua populasi.

Ditentukan :

$$\boldsymbol{\mu}_1 = \mathrm{E}\ (\boldsymbol{X}|\Pi_1) \text{ adalah nilai harapan pengamatan dari } \Pi_1$$

$$\boldsymbol{\mu}_2 = \mathrm{E}\ (\boldsymbol{X}|\Pi_2) \text{ adalah nilai harapan pengamatan dari } \Pi_2 \qquad (2.6)$$

Sehingga misalkan terdapat kombinasi linear :

$$Y = \mathbf{a}^t\mathbf{X}$$

Dengan menggunakan persamaan (2.2) diperoleh nilai tengah dari Y adalah :

$$\boldsymbol{\mu}_{1Y} = \mathrm{E}\ (Y|\Pi_1) = E(\boldsymbol{a^t X}|\Pi_1) = \boldsymbol{a^t}E(\boldsymbol{X}|\Pi_1) = \boldsymbol{a^t\mu_1}$$

$$\boldsymbol{\mu}_{2Y} = \mathrm{E}\ (Y|\Pi_2) = E(\boldsymbol{a^t X}|\Pi_2) = \boldsymbol{a^t}E(\boldsymbol{X}|\Pi_2) = \boldsymbol{a^t\mu_2} \qquad (2.7)$$

Jika di asumsikan matriks peragam $\boldsymbol{\Sigma}$ sama untuk setiap populasi, maka ragam dari Y adalah :

$$\sigma_Y^2 = \mathrm{Var}\ (\mathrm{Y}) = \mathrm{Var}\ (\boldsymbol{a^t X}) = \boldsymbol{a^t}\ \mathrm{Cov}\ (\mathbf{X})\ \mathbf{a} = \boldsymbol{a^t}\boldsymbol{\Sigma}\mathbf{a} \qquad (2.8)$$

Misal Y = $\mathbf{a}^t\,\mathbf{x}$

$$\begin{aligned}
\text{Var }(\mathbf{X}) &= \boldsymbol{\Sigma}_{\mathbf{pxp}} = E(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^t \\
\text{Var }(Y) &= E(Y - \boldsymbol{\mu}_Y)(\mathbf{Y} - \boldsymbol{\mu}_Y)^t \\
&= E(Y - E(Y))(Y - E(Y))^t \\
&= E(\mathbf{a}^t\,\mathbf{x} - E(\mathbf{a}^t\,\mathbf{x}))(\,\mathbf{a}^t\,\mathbf{x} - E(\mathbf{a}^t\,\mathbf{x}))^t \\
&= E(\mathbf{a}^t\,\mathbf{x} - \mathbf{a}^t\,E(\mathbf{x}))(\,\mathbf{a}^t\,\mathbf{x} - \mathbf{a}^t\,E(\mathbf{x}))^t \\
&= E(\mathbf{a}^t\,(\mathbf{x} - E(\mathbf{x})))(\,\mathbf{a}^t\,(\mathbf{x} - E(\mathbf{x})))^t \\
&= E[\mathbf{a}^t(\mathbf{x} - E(\mathbf{x}))((\mathbf{x} - E(\mathbf{x}))^t\,\mathbf{a})] \\
&= E[\mathbf{a}^t(\mathbf{x} - E(\mathbf{x}))(\mathbf{x} - E(\mathbf{x}))^t\,\mathbf{a}] \\
&= \mathbf{a}^t\,\underbrace{E(\mathbf{x} - E(\mathbf{x}))(\mathbf{x} - E(\mathbf{x}))^t}_{\boldsymbol{\Sigma}}\,\mathbf{a} \\
&= \mathbf{a}^t\,\boldsymbol{\Sigma}\,\mathbf{a} \\
\text{Var }(Y) &= \text{Var }(\mathbf{a}^t\,\mathbf{x}) = \mathbf{a}^t\,\boldsymbol{\Sigma}\,\mathbf{a}
\end{aligned}$$

Jadi, jika diasumsikan matriks peragam $\boldsymbol{\Sigma}$ sama untuk setiap populasi, maka ragam dari Y adalah :

$$\sigma_Y^2 = \text{Var (Y)} = \text{Var }(\boldsymbol{a^t X}) = \boldsymbol{a^t}\text{ Cov }(\mathbf{X})\,\mathbf{a} = \boldsymbol{a^t\Sigma}\mathbf{a}$$

Kombinasi linear terbaik diperoleh dari perbandingan

$$\begin{aligned}
\left(\frac{jarak\ kuadrat\ antara\ nilai\ tengah\ Y}{ragam\ Y}\right) &= \frac{(\boldsymbol{\mu_{1Y}} - \boldsymbol{\mu_{2Y}})^2}{\sigma_Y^2} \\
&= \frac{(\boldsymbol{a^t\mu_1} - \boldsymbol{a^t\mu_2})^2}{\boldsymbol{a^t\Sigma a}} \\
&= \frac{(\boldsymbol{a^t}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2}))^2}{\boldsymbol{a^t\Sigma a}} \\
&= \frac{(\boldsymbol{a^t\delta})^2}{\boldsymbol{a^t\Sigma a}} \qquad (2.9)
\end{aligned}$$

dengan $\boldsymbol{\delta} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ adalah perbedaan vektor nilai tengah.

Misalkan $\boldsymbol{\delta} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ dan Y = $\boldsymbol{a^t X}$, maka :

$$\left(\frac{jarak\ kuadrat\ antara\ nilai\ tengah\ Y}{ragam\ Y}\right) = \frac{(\boldsymbol{a}^t\boldsymbol{\delta})^2}{\boldsymbol{a}^t\boldsymbol{\Sigma}\boldsymbol{a}}$$

akan maksimum dengan memilih **a** = c $\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} = c\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ untuk c ≠ 0. Jika dipilih c = 1, maka akan menghasilkan kombinasi linear

$$Y \quad = \boldsymbol{a}^t\boldsymbol{X} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X} \tag{2.10}$$

yang dikenal sebagai fungsi diskriminan linear Fisher. Maksimum dari persamaan tersebut adalah :

$$\max_a \frac{(\boldsymbol{a}^t\boldsymbol{\delta})^2}{\boldsymbol{a}^t\boldsymbol{\Sigma}\boldsymbol{a}} = \boldsymbol{\delta}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} \tag{2.11}$$

Bukti :

$\boldsymbol{\Sigma}$ adalah matriks definit positif, sehingga $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^t$ dan $\boldsymbol{a}^t\boldsymbol{\Sigma}\boldsymbol{a} > 0$ untuk setiap **a**.

Dengan menggunakan persamaan *Cauchy-Schwarz* :

$$(\boldsymbol{a}^t\boldsymbol{\delta})^2 \leq (\boldsymbol{a}^t\boldsymbol{\Sigma}\boldsymbol{a})(\boldsymbol{\delta}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta})$$

karena **a** ≠ 0 dan $\boldsymbol{\Sigma}$ adalah definit positif, maka $\boldsymbol{a}^t\boldsymbol{\Sigma}\boldsymbol{a} > 0$. Dengan membagi kedua ruas pertidaksamaan dengan $\boldsymbol{a}^t\boldsymbol{\Sigma}\boldsymbol{a}$, maka menghasilkan batas atas :

$$\frac{(\boldsymbol{a}^t\boldsymbol{\delta})^2}{\boldsymbol{a}^t\boldsymbol{\Sigma}\boldsymbol{a}} \leq \boldsymbol{\delta}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$$

sehingga untuk **a** = c $\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$, batas maksimum untuk persamaan (2.9) menjadi :

$$\max_a \frac{(\boldsymbol{a}^t\boldsymbol{\delta})^2}{\boldsymbol{a}^t\boldsymbol{\Sigma}\boldsymbol{a}} = \boldsymbol{\delta}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$$

Misalkan $\overline{\boldsymbol{x}}_1, \overline{\boldsymbol{x}}_2, \boldsymbol{S}_1$, dan $\boldsymbol{S}_2$ masing-masing merupakan penduga tak bias dari $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1$, dan $\boldsymbol{\Sigma}_2$. Misalkan juga kita mempunyai n_1 dan n_2 pengamatan dari peubah acak **X** = $[X_1, X_2, \ldots, X_p]$ yang masing-masing berasal dari Π_1 dan Π_2.

$$\boldsymbol{X}_{1\,(pxn_1)} = [\boldsymbol{x}_{11}, \boldsymbol{x}_{12}, \ldots, \boldsymbol{x}_{1n_1}]$$

$$\boldsymbol{X}_{2\,(pxn_2)} = [\boldsymbol{x}_{21}, \boldsymbol{x}_{22}, \ldots, \boldsymbol{x}_{2n_2}] \tag{2.12}$$

Vektor nilai tengah sampel dan matriks peragamnya ditentukan dengan :

$$\bar{\boldsymbol{x}}_{1\,(px1)} = \frac{1}{n_1}\sum_{j=1}^{n_1}\boldsymbol{x}_{1j}\ ;\ \boldsymbol{S}_{1\,(pxp)} = \frac{1}{n_1-1}\sum_{j=1}^{n_1}(\boldsymbol{x}_{1j}-\bar{\boldsymbol{x}}_1)(\boldsymbol{x}_{1j}-\bar{\boldsymbol{x}}_1)^t$$

$$\bar{\boldsymbol{x}}_{2\,(px1)} = \frac{1}{n_2}\sum_{j=1}^{n_2}\boldsymbol{x}_{2j}\ ;\ \boldsymbol{S}_{2\,(pxp)} = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(\boldsymbol{x}_{2j}-\bar{\boldsymbol{x}}_2)(\boldsymbol{x}_{2j}-\bar{\boldsymbol{x}}_2)^t \qquad (2.13)$$

Diasumsikan bahwa kedua matriks **Σ** sama, sehingga penggabungan dari matriks $\boldsymbol{S}_1$ dan $\boldsymbol{S}_2$ akan memperoleh dugaan tak bias tunggal untuk **Σ**. Sehingga rata-rata terboboti :

$$\begin{aligned}\boldsymbol{S}_p &= \left[\frac{n_1}{(n_1-1)+(n_2-1)}\right]\boldsymbol{S}_1 + \left[\frac{n_2}{(n_1-1)+(n_2-1)}\right]\boldsymbol{S}_2 \\ &= \frac{(n_1-1)\boldsymbol{S}_1+(n_2-1)\boldsymbol{S}_2}{(n_1+n_2-2)} \qquad (2.14)\end{aligned}$$

Persamaan (2.14) merupakan penduga tak bias untuk **Σ** jika matriks data $\boldsymbol{X}_1$ dan $\boldsymbol{X}_2$ masing-masing berisi sampel acak dari Π_1 dan Π_2. Besaran sampel $\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2$, dan $\boldsymbol{S}_p$ disubstitusikan untuk $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$, dan **Σ** dalam persamaan (2.10), sehingga penduga yang sesuai bagi **a** misal dinotasikan sebagai **b** adalah :

$$\mathbf{b} = \mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

maka menghasilkan fungsi diskriminan linear sampel Fisher :

$$\mathrm{y} = \boldsymbol{b}^t\boldsymbol{x} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^t\boldsymbol{S}_p\boldsymbol{x} \qquad (2.15)$$

kombinasi linear y = $\boldsymbol{b}^t\boldsymbol{x} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^t\boldsymbol{S}_p\boldsymbol{x}$ memaksimumkan perbandingan :

$$\begin{aligned}\left(\frac{jarak\ kuadrat\ antara\ nilai\ tengah\ sampel\ y}{ragam\ sampel\ y}\right) &= \frac{(\hat{y}_1-\hat{y}_2)^2}{S_y^2} \\ &= \frac{\left(\boldsymbol{b}^t\bar{\mathbf{x}}_1-\boldsymbol{b}^t\bar{\mathbf{x}}_2\right)^2}{\boldsymbol{b}^t\boldsymbol{S}_p\boldsymbol{b}} \\ &= \frac{\left(\boldsymbol{b}^t(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)\right)^2}{\boldsymbol{b}^t\boldsymbol{S}_p\boldsymbol{b}} \\ &= \frac{(\boldsymbol{b}^t\boldsymbol{d})^2}{\boldsymbol{b}^t\boldsymbol{S}_p\boldsymbol{b}} \qquad (2.16)\end{aligned}$$

dengan $\mathbf{d} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$

Misalkan $\mathbf{S}_p$ merupakan penduga tak bias bagi $\boldsymbol{\Sigma}$. Dengan menggunakan persamaan *Cauchy-Schwarz* :

$$(\boldsymbol{b}^t\boldsymbol{d})^2 \leq (\boldsymbol{b}^t\boldsymbol{S}_p\boldsymbol{b})(\boldsymbol{d}^t\boldsymbol{S}_p^{-1}\boldsymbol{d})$$

karena $\mathbf{b} \neq 0$ dan $\boldsymbol{S}_p$ adalah definit positif, maka $\boldsymbol{b}^t\boldsymbol{S}_p\boldsymbol{b} > 0$. Dengan membagi kedua ruas pertidaksamaan dengan $\boldsymbol{b}^t\boldsymbol{S}_p\boldsymbol{b}$, maka menghasilkan batas atas :

$$\frac{(\boldsymbol{b}^t\boldsymbol{d})^2}{\boldsymbol{b}^t\boldsymbol{S}_p\boldsymbol{b}} \leq \boldsymbol{d}^t\boldsymbol{S}_p^{-1}\boldsymbol{d}$$

sehingga maksimum dari perbandingan sampel dalam (2.16) diberikan dengan memisalkan $\mathbf{b} = \boldsymbol{S}_p^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$, sehingga :

$$\begin{aligned}\max_b \frac{(\boldsymbol{b}^t\boldsymbol{d})^2}{\boldsymbol{b}^t\boldsymbol{S}_p\boldsymbol{b}} &= \boldsymbol{d}^t\boldsymbol{S}_p^{-1}\boldsymbol{d} \\ &= (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^t\boldsymbol{S}_p^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \\ &= D^2\end{aligned}$$

di mana D^2 merupakan jarak kuadrat sampel.

Fungsi linier diskriminan Fisher tersebut kemudian dipakai oleh Andeson untuk memperoleh statistik klasifikasi W(x) yaitu;

$$W(x) = \left(\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)\right)^t \boldsymbol{S}_p^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \qquad (2.17)$$

Uraiannya adalah sebagai berikut:

Misalkan $\mathbf{x}$ akan dikelompokkan ke Π_1 jika $\mathbf{b}^t\,\mathbf{x} \geq h$, dan jika sebaliknya maka $\mathbf{x}$ akan dikelompokkan ke dalam Π_2, dengan $h = \frac{\boldsymbol{b}^t(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)}{2}$. Dengan kata lain $\mathbf{x}$ akan dimasukkan ke dalam populasi yang paling dekat dengannya.

$$h = \frac{1}{2}\mathbf{b}^t(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$$

$$= \frac{1}{2}\left[\boldsymbol{S}_p^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\right]^{\boldsymbol{t}}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$$

$$= \tfrac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^{\mathbf{t}}\ \boldsymbol{S}_p^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$$

$$L(\mathbf{x}) = \mathbf{b}^{\mathbf{t}}\mathbf{x}$$

$$\mathbf{b} = \boldsymbol{S}_p^{-1}\,(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

$$W(x) = L(\mathbf{x}) - h$$

$$= \mathbf{b}^{\mathbf{t}}\mathbf{x} - \tfrac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^{\mathbf{t}}\ \boldsymbol{S}_p^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$$

$$= \left(\boldsymbol{S}_p^{-1}\,(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\right)^{\mathbf{t}}\mathbf{x} - \tfrac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^{\mathbf{t}}\ \boldsymbol{S}_p^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$$

$$= (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^{\mathbf{t}}\,\boldsymbol{S}_p^{-1}\mathbf{x} - \tfrac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^{\mathbf{t}}\ \boldsymbol{S}_p^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$$

$$W(x) = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^{\mathbf{t}}\,\boldsymbol{S}_p^{-1}\left[\mathbf{x} - \tfrac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)\right]$$

Dimana

$$\mathbf{x} = \mathbf{x}_{px1} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$

$$\bar{\mathbf{x}}_{\mathbf{1}(\mathrm{px1})} = \begin{bmatrix} \bar{x}_{11} \\ \bar{x}_{21} \\ \vdots \\ \bar{x}_{p1} \end{bmatrix}, \qquad \bar{\mathbf{x}}_{\mathbf{2}(\mathrm{px1})} = \begin{bmatrix} \bar{x}_{12} \\ \bar{x}_{22} \\ \vdots \\ \bar{x}_{p2} \end{bmatrix}$$

$$\mathbf{S}_p = \boldsymbol{S}_{(p\mathrm{x}p)}\begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \cdots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \cdots & S_{pp} \end{bmatrix}$$

Maka:

$$W(x) = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^{t}_{(p\mathrm{x}1)}\boldsymbol{S}_p^{-1}{}_{(\mathrm{pxp})}\left[\mathbf{x}_{(\mathrm{px1})} - \tfrac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)_{(\mathrm{px1})}\right]$$

Sehingga diperoleh matriks W(x) berukuran 1x1 (simetris), dan berlaku

$W(x) = W(x)^t$

atau

$$W(x) = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^{\mathbf{t}} \boldsymbol{S}_p^{-1} \left[\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)\right]$$

$$= \left((\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^{\mathbf{t}} \boldsymbol{S}_p^{-1} \left[\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)\right]\right)^{t}$$

$$= \left(\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)\right)^{\mathbf{t}} \boldsymbol{S}_p^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

Menggunakan aturan ini, objek yang baru yaitu **x** akan dimasukkan atau diklasifikasikan ke dalam Π_1 jika $W(\underline{x}) \geq 0$, dan akan dimasukkan ke dalam Π_2 jika $W(\underline{x}) < 0$ (Mangku, 2004). Aturan klasifikasi W(x) ini dapat juga dinotasikan dengan W(x, t) untuk menegaskan bahwa aturan klasifikasi ini dibangun menggunakan training sampel t, untuk mengklasifikasikan pengamatan/observasi baru yaitu **x**.

2.4 Peluang Kesalahan Klasifikasi

Sesuai dengan aturan klasifikasi, terdapat peluang kesalahan klasifikasi jika digunakan aturan klasifikasi tersebut untuk mengklasifikasikan objek baru ke dalam salah satu dari dua kelompok. Peluang kesalahan klasifikasi disebut juga tingkat kesalahan.

Menurut Mangku (2002), ada tiga macam tingkat kesalahan, yaitu:

1. Tingkat kesalahan optimal (*Optimal Error Rate*), yang menjelaskan hasil dari aturan klasifikasi berdasarkan parameter yang diketahui.
2. Tingkat kesalahan bersyarat (*Conditional Error Rate*), yang menjelaskan hasil dari aturan klasifikasi berdasarkan parameter yang diduga dengan statistik yang dihitung dari sampel percobaan.

3. Tingkat kesalahan harapan (*Expected Error Rate*), yang menjelaskan hasil yang diharapkan dari aturan klasifikasi berdasarkan parameter yang diduga dengan sampel percobaan yang dipilih secara acak.

Dalam prakteknya tingkat kesalahan optimal dan tingkat kesalahan harapan sulit diduga, karena parameter jarang diketahui dan tingkat kesalahan harapan bergantung pada distribusi fungsi diskriminannya, dimana hal ini sangat kompleks. Sehingga penelitian lebih banyak terfokus pada tingkat kesalahan bersyarat dalam kajiannya.

Peluang kesalahan klasifikasi untuk tingkat kesalahan bersyarat oleh:

P_1 = P (W(**x,t**) < 0 ketika x berasal dari Π_1 $|\boldsymbol{t}$ tetap)

P_2 = P (W(**x,t**) ≥ 0 ketika x berasal dari Π_2 $|\boldsymbol{t}$ tetap)

t adalah sampel acak yang disebut juga sampel percobaan (*training sampel*). Sampel acak adalah suatu himpunan bagian dari populasi yang dipilih sedemikian rupa sehingga setiap himpunan bagian tersebut mempunyai peluang terpilih yang sama. Peluang kesalahan totalnya didefinisikan sebagai :

$$P = \frac{n_1P_1+n_2P_2}{n_1+n_2} \qquad (2.18)$$

Dalam menduga tingkat kesalahan bersyarat ini terdapat dua macam penduga, yaitu yang pertama penduga parametrik yang terdiri dari penduga Okamoto, penduga Mclachan, dan penduga yang diperhalus, dan yang kedua penduga empirik yang terdiri dari penduga resubstitusi, penduga U, penduga $\overline{U}$, dan penduga *Jackknife* (Mangku, 2004).

2.4.1 Penduga Resubstitusi

Penduga ini diperkenalkan oleh Smith pada tahun 1947 (Mangku, 2004). Ide dasarnya adalah merealokasi setiap individu dalam sampel percobaan t dengan menggunakan aturan W (x, t). misalkan kriteria perhitungannya adalah Q (i, j) dengan ketentuan sebagai berikut:

$$Q(i, j) = \begin{cases} 0, jika\ i = j \\ 1, jika\ i \neq j \end{cases} \quad \text{untuk setiap i dan j} \tag{2.19}$$

maka penduga Resubstitusi dapat didefinisikan sebagai:

$$\hat{P}_1(R) = \frac{1}{n_1} \sum_{j=1}^{n_1} Q\left[1, W\left(x_{1j}, t\right)\right] \tag{2.20}$$

dan

$$\hat{P}_2(R) = \frac{1}{n_2} \sum_{j=1}^{n_2} Q\left[2, W\left(x_{2j}, t\right)\right] \tag{2.21}$$

dengan W (x_{ij}, t) merupakan hasil pengklasifikasian suatu unit dalam Π_1, sedangkan W (x_{2j}, t) merupakan hasil pengklasifikasian suatu unit dalam Π_2.

2.4.2 Penduga Jackknife

Jackknife adalah metode resampling yang diperkenalkan oleh Quenouille sebagai metode non parametrik untuk penduga bias (Efron, 1982). Menurut Mangku (2004), pada prosedur *Jackknife*, yang pertama kali dilakukan adalah menghitung tingkat kesalahan Resubstitusi setiap kali pengamatan dihilangkan dari sampel percobaan t. Kemudian teknik *Jackknife* standar digunakan untuk memperbaiki bias keseluruhan dari kesalahan Resubstitusi.

Penduga *Jackknife* untuk penduga tingkat kesalahan diberikan oleh rumus berikut:

$$\hat{P}_1(JK) = \hat{P}_1(R) + (n-1)(R_1^+ - R_{1(.)}) \quad (2.22)$$

$$\hat{P}_2(JK) = \hat{P}_2(R) + (n-1)(R_2^+ - R_{2(.)}) \quad (2.23)$$

dengan $\hat{P}_1(R)$ dan $\hat{P}_2(R)$ merupakan penduga Resubstitusi pada persamaan (2.20) dan (2.21), R_1^+ dan R_2^+ masing-masing diberikan oleh rumus sebagai berikut:

$$R_1^+ = \left(\frac{1}{n_1}\right)^2 \sum_{j=1}^{n_1} \sum_{k=1}^{n_1} Q\left[1, W\left(x_{1k}, T_{[1j]}\right)\right] \quad (2.24)$$

$$R_2^+ = \left(\frac{1}{n_2}\right)^2 \sum_{j=1}^{n_2} \sum_{k=1}^{n_2} Q\left[2, W\left(x_{2k}, T_{[2j]}\right)\right] \quad (2.25)$$

sedangkan, $R_{1(.)}$ dan $R_{2(.)}$ masing-masing diberikan rumus sebagai berikut:

$$R_{1(.)} = \frac{1}{n_1} \sum_{j=1}^{n_1} \left(\frac{1}{n_1 - 1} \sum_{k \neq j}^{n_1} Q\left[1, W\left(x_{1k}, T_{[1j]}\right)\right]\right) \quad (2.26)$$

$$R_{2(.)} = \frac{1}{n_2} \sum_{j=1}^{n_2} \left(\frac{1}{n_2 - 1} \sum_{k \neq j}^{n_2} Q\left[2, W\left(x_{2k}, T_{[2j]}\right)\right]\right) \quad (2.27)$$

III. METODOLOGI PENELITIAN

3.1 Tempat dan Waktu Penelitian

Penelitian ini dilakukan di jurusan Matematika FMIPA Universitas Lampung pada semester ganjil 2021-2022.

3.2 Data Penelitian

Data yang akan digunakan dalam penelitian ini merupakan data simulasi yaitu membangkitkan data dengan menggunakan *software* SAS 9.4. Data yang dibangkitkan adalah dua populasi dengan sebaran normal multivariate, masing-masing berukuran 1000 dan mempunyai 5 peubah. Kemudian dari setiap populasi tersebut diambil sampel, ukuran sampel untuk setiap kelompok ditetapkan sama yaitu $n_1 = n_2$, dalam penelitian ini akan dicobakan n = 20, 50, 80, 120, 200, dan 300 dengan $n = n_1 + n_2$. Pengambilan sampel dilakukan masing-masing sebanyak 1000 kali.

Data yang dibangkitkan terdiri dari:

1. Dua populasi dengan matriks kovarian diagonal, $\mathbf{\Sigma_1} = \mathbf{\Sigma_2}$
2. Dua populasi dengan matriks kovarian nondiagonal, $\mathbf{\Sigma_1} = \mathbf{\Sigma_2}$

3.3 Metode Penelitian

Metode yang digunakan dalam penelitian ini adalah studi pustaka, yaitu dengan mempelajari buku-buku teks penunjang dan karya ilmiah yang disajikan dalam bentuk jurnal serta melakukan pengolahan data dengan menggunakan *software* SAS 9.4. Pengolahan data dilakukan untuk mendapatkan perbandingan tingkat kesalahan klasifikasi dengan menggunakan penduga Resubstitusi dan penduga *Jackknife*.

Adapun langkah-langkah yaitu sebagai berikut:

1. Membangkitkan data populasi Φ sebanyak N = 2000 yang dibagi ke dalam dua kelompok Π_1 dan Π_2 masing-masing berukuran $N_1 = N_2 = 1000$ dengan dimensi p = 5 dari sebaran normal dengan vektor nilai tengah dan matriks kovarian sama, yaitu $\Pi_1 \sim N_p(\mu_1,\ \Sigma_1)$ dan $\Pi_2 \sim N_p(\mu_2,\ \Sigma_2)$.
2. Mengambil sampel percobaan t yaitu $X_1 = \{x_{11}, x_{12}, \dots, x_{1n_1}\}$ dari kelompok pertama Π_1 dan $X_2 = \{x_{21}, x_{22}, \dots, x_{2n_2}\}$ dari kelompok kedua Π_2, dengan $t = X_1 + X_2 = \{x_{11}, x_{12}, \dots, x_{1n_1}, x_{21}, x_{22}, \dots, x_{2n_2}\}$ dan $n = n_1 + n_2 = 20$.
3. Menghitung nilai tengah kelompok satu $\bar{x}_1$, nilai tengah kelompok dua $\bar{x}_2$, dan matriks kovarian gabungan dari sampel percobaan.
4. Menduga fungsi diskriminan Fisher menurut persamaan (2.17) berdasarkan nilai tengah dan matriks kovarian yang diperoleh pada langkah 3.
5. Mengklasifikasikan data menggunakan fungsi diskriminan yang diperoleh pada langkah 4, dan menghitung banyaknya objek pada kelompok 1 dan kelompok 2 yang salah diklasifikasi (*misclassified*).

6. Menghitung nilai tingkat kesalahan klasifikasi dari fungsi diskriminan yang diperoleh dengan menggunakan penduga Resubstitusi, dan penduga *Jackknife*.
7. Ulangi langkah 2 sampai 6 sebanyak 1000 kali.
8. Menghitung rata-rata dari π yaitu:

$$\bar{\pi} = \frac{1}{1000} \sum_{i=1}^{1000} \pi_i$$

9. Menghitung nilai Kuadrat Tengah Galat (KTG) dari penduga tingkat kesalahan kedua metode dengan rumus sebagai berikut:

$$\text{KTG}\left(\hat{P}_i\right) = \frac{1}{1000} \sum_{i=1}^{1000} \left(\hat{P}_i - \bar{\pi}\right)^2$$

10. Ulangi langkah 2 sampai 9 untuk ukuran sampel n=50, 80, 120, 200, dan 300.

V. KESIMPULAN

Dari hasil penelitian yang dilakukan dapat disimpulkan bahwa :

1. Penduga *Jackknife* cukup baik pada data dua populasi dengan matriks kovarian diagonal untuk ukuran sampel kecil. Sedangkan, penduga Resubstitusi cenderung lebih baik pada ukuran sampel besar.
2. Penduga Resubstitusi lebih baik dari pada penduga *Jackknife* dalam menduga kesalahan klasifikasi pada data dua populasi dengan matriks kovarian nondiagonal pada setiap ukuran sampel dan jumlah peubah.
3. Banyaknya peubah (p = 3, 4, 5) dan ukuran sampel (n = 20, 50, 80, 120, 200, 300) dalam penelitian ini yang digunakan untuk membentuk fungsi diskriminan Fisher, memberikan pengaruh terhadap nilai penduga Resubstitusi dan *Jackknife* yang diperoleh, yaitu Semakin banyak jumlah peubah maupun ukuran sampel maka nilai KTG dugaan *Misclassification Error* yang diperoleh dari kedua penduga semakin lebih kecil.

DAFTAR PUSTAKA

Anton, H., Rorres, C., & Kaul, A. 2019. *Elementary Linear Algebra Applications Version*. Ed. ke-12. John Wiley & Sons Inc., United States.

Efron, B. 1982. *The Jackknife, The Bootstrap and Other Resampling Plans*. SIAM-CBMS Monograph 38. Philadelphia, S.I.A.M.

Gazali, W. 2005. *Matriks dan Transformasi Linear*. Graha Ilmu, Yogyakarta.

Hair Jr, J.F., Black, W.C., Babin, B.J., & Anderson R.E. 2019. *Multivariate Data Analysis*. Ed. Ke-8. Cengage Learning, United Kingdom.

Johnson, R.A. & Wichern, D.W. 2007. *Applied Multivariate Statistical Analysis*. Ed. ke-6. Prentice Hall Internasional, New Jersey.

Lachenbruch, P.A., & Mickey, M.R. 1968. Estimation of Error Rates in Discriminant Analysis. *Technometrics*. **10**(1): 1-11.

Mangku, I.W. 2002. Discriminant Functions and Their Misclassification Errors. *Jurnal Matematika dan Aplikasinya*. **1**(2): 37-48.

Mangku, I.W. 2004. Estimating The Probability of Misclassifications in Two Groups Discriminant Analysis. *Jurnal Matematika dan Aplikasinya*. **3**(1): 1-10.

Nisa, K. 2009. Pembandingan Beberapa Penduga Tingkat Kesalahan Klasifikasi Pada Analisis Diskriminan Kuadratik. *Seminar Hasil Penelitian & Pengabdian Kepada Masyarakat*. Unila, Bandar Lampung. A183-A192.

Osuji, G.A., Onyeagu, S.I., & Ekezie, D.D. 2013. Comparison of Jackknife and Resubstitution Methods in the Estimation of Error Rates in Discriminant Analysis. *International Journal of Mathematics and Statistics Studies*. **1**(3): 29-37.

Sartono, B., Affendi, F.M., Syafitri, U.D., Sumertajaya, I.M., & Anggraeni, Y. 2003. *Analisis Peubah Ganda*. Institut Pertanian Bogor, Bogor.

Snapinn, S.M., & Knoke, J.D. 1984. Classification Error Rate Estimators Evaluated by Unconditional Mean Square Error. *Technometrics*. **26**(4): 371-378.