



**KLASIFIKASI JENIS KANKER BERDASARKAN TINGKAT  
EKSPRESI GEN PASIEN PENDERITA KANKER  
MENGUNAKAN METODE VERTEX DISCRIMINANT  
ANALYSIS**

**SKRIPSI**

Sebagai salah satu syarat untuk memperoleh gelar Sarjana Statistika

Oleh:  
**GARNIS KOFSANOVA**  
**165090501111002**



**PROGRAM STUDI SARJANA STATISTIKA**  
**JURUSAN STATISTIKA**  
**FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM**  
**UNIVERSITAS BRAWIJAYA**  
**MALANG**  
**2020**



Halaman ini sengaja dikosongkan



**LEMBAR PENGESAHAN SKRIPSI**  
**KLASIFIKASI JENIS KANKER BERDASARKAN TINGKAT**  
**EKSPRESI GEN PASIEN PENDERITA KANKER**  
**MENGGUNAKAN METODE VERTEX DISCRIMINANT**  
**ANALYSIS**

Oleh:  
**GARNIS KOFSANOVA**  
**165090501111002**

Setelah dipertahankan di depan Majelis Penguji  
pada tanggal 11 Juni 2020  
dan dinyatakan memenuhi syarat untuk memperoleh gelar  
**Sarjana Statistika**

**Dosen Pembimbing**

**Dr. Suci Astutik, S.Si., M.Si.**  
**NIP. 197407221999032001**

**Mengetahui,**  
**Ketua Jurusan Statistika**  
**Fakultas MIPA Universitas Brawijaya**



**Rahma Fitriani, S.Si., M.Sc., Ph.D.**  
**NIP. 197603281999032001**



Repository Universitas Brawijaya

Repository Universitas Brawijaya



Halaman ini sengaja dikosongkan

Repository Universitas Brawijaya

Repository Universitas Brawijaya

Repository Universitas Brawijaya

Repository Universitas Brawijaya



## LEMBAR PERNYATAAN

**Saya yang bertanda tangan di bawah ini:**

**Nama** : Garnis Kofsanova

**NIM** : 165090501111002

**Program Studi** : Statistika

**Penulis Skripsi Berjudul** :

### **KLASIFIKASI JENIS KANKER BERDASARKAN TINGKAT EKSPRESI GEN PASIEN PENDERITA KANKER MENGUNAKAN METODE *VERTEX DISCRIMINANT* ANALYSIS**

**Dengan ini menyatakan bahwa:**

- 1. Isi dari skripsi yang saya buat adalah benar-benar karya sendiri dan tidak menjiplak karya orang lain, selain nama-nama yang termasuk di sini dan tertulis di daftar pustaka dalam skripsi ini.**
- 2. Apabila dikemudian hari ternyata skripsi yang saya tulis terbukti hasil jiplakan, maka saya akan bersedia menanggung resiko yang akan saya terima.**

**Demikian pernyataan ini dibuat dengan segala kesadaran.**

**Malang, 7 Juli 2020**

**Yang menyatakan,**

**Garnis Kofsanova**

**NIM. 165090501111002**



Halaman ini sengaja dikosongkan

# KLASIFIKASI JENIS KANKER BERDASARKAN TINGKAT EKSPRESI GEN PASIEN PENDERITA KANKER MENGUNAKAN METODE *VERTEX DISCRIMINANT*

## ANALYSIS

## ABSTRAK

Analisis diskriminan adalah analisis multivariat yang diterapkan untuk memodelkan hubungan antara peubah respon berupa data kualitatif dan peubah prediktor berupa data kuantitatif. Analisis diskriminan dengan jumlah peubah prediktor lebih banyak daripada jumlah amatan dapat ditangani dengan menggunakan metode *Vertex Discriminant Analysis* (VDA). VDA mengklasifikasikan objek dengan mengembangkan metode analisis diskriminan menggunakan simpul (*vertex*) dari simpleks pada ruang *Euclidean*. VDA dapat digunakan untuk menangani kasus data yang tidak memenuhi asumsi matriks *varians kovarians* yang homogen dan asumsi normal multivariat. Tujuan dari penelitian ini adalah untuk membentuk fungsi diskriminan untuk mengklasifikasikan pasien ke dalam lima kategori jenis kanker. Data yang digunakan pada penelitian ini berupa gen manusia yang menderita suatu jenis kanker sebanyak 685 dan amatan sebanyak 60 pasien yang diklasifikasikan ke dalam lima jenis kanker. Kelima jenis kanker pada penelitian ini, yaitu BRCA (*Breast Invasive Carcinoma*), COAD (*Colon Adenocarcinoma*), KIRC (*Kidney Renal Clear Cell Carcinoma*), LUAD (*Lung Adenocarcinoma*) dan PRAD (*Prostate Adenocarcinoma*). Hasil analisis menunjukkan bahwa metode VDA membentuk empat fungsi diskriminan untuk membedakan lima jenis kanker. VDA mampu mengklasifikasikan dengan tepat 48 dari 48 data *training*, sedangkan untuk data *testing* VDA mengklasifikasikan dengan tepat 6 dari 12 data dengan kesalahan klasifikasi (APER) sebesar 50%.

**Kata Kunci:** *Data Dimensi Tinggi, Gen, Vertex Discriminant Analysis* (VDA)



Halaman ini sengaja dikosongkan

# TYPE OF CANCER CLASSIFICATION BASED ON THE EXPRESSION LEVELS OF CANCER SUFFERERS' GENES USES THE VERTEX DISCRIMINANT ANALYSIS (VDA)

## METHOD

## ABSTRACT

Discriminant analysis is a multivariate analysis that is applied to modeling the relationship between qualitative data and predictor modification of quantitative data. Analysis of the discriminant with more than the number of predictor shifters can be completed by using the Vertex Discriminant Analysis (VDA) method. The VDA classifies the object by developing a method of discriminant analysis using the vertex of the simplex on the Euclidean room. VDA can be used to handle data cases that do not satisfy the assumption of homogeneous covariance matrix and multivariate normal assumptions. The purpose of this study is to establish discriminant functions to classify patients into five categories of cancer. The data used in this study consists of the human genes that suffer from a type of cancer of 685 and cancers of up to 60 patients classified into the five kinds. The study's fifth type of cancer, BRCA (Breast Invasive Carcinoma), COAD (Colon Adenocarcinoma), KIRC (Kidney Of Clear Cell Carcinoma), LUAD (Lung Adenocarcinoma), and PRAD (Prostate Adenocarcinoma). Results have shown that the VDA's method constitutes four discriminatory functions to distinguish the five types of cancer. VDA is able to classify exactly 48 from 48 training data, while for testing VDA data, it classifies exactly 6 of the 12 data by 50%.

**Keywords:** *High Dimensional Data, Gene, Vertex Discriminant Analysis (VDA)*



## DAFTAR ISI

Hal

LEMBAR PENGESAHAN SKRIPSI .....	iii
LEMBAR PERNYATAAN.....	v
ABSTRAK.....	vii
ABSTRACT.....	ix
KATA PENGANTAR .....	xi
DAFTAR ISI.....	xiii
DAFTAR GAMBAR.....	xvii
DAFTAR LAMPIRAN.....	xix
<b>BAB I PENDAHULUAN.....</b>	<b>1</b>
1.1. Latar Belakang.....	1
1.2. Rumusan Masalah .....	3
1.3. Tujuan Penelitian.....	3
1.4. Manfaat Penelitian.....	3
1.5. Batasan Masalah.....	3
<b>BAB II TINJAUAN PUSTAKA.....</b>	<b>5</b>
2.1. Analisis Diskriminan .....	5
2.2. Asumsi Analisis Diskriminan.....	5
2.2.1. Asumsi Normal Multivariat .....	5
2.2.2. Asumsi Homogenitas Matriks <i>Varians</i> <i>Kovarians</i> .....	6
2.2.3. Analisis Korelasi untuk Asumsi Multikolinieritas .....	7
2.3. Data Hilang ( <i>Missing Value</i> ).....	8
2.4. <i>Vertex Discriminant Analysis</i> (VDA).....	9
2.4.1. <i>Vertex</i> Pada Ruang $\mathbb{R}^{k-1}$ .....	9
2.4.2. Fungsi Kerugian ( <i>Loss Function</i> ).....	12
2.4.3. Fungsi Tujuan Pada VDA .....	13
2.4.4. Algoritma MM ( <i>Majorize Minimize</i> ) Pada VDA .....	14
2.4.5. Mayorisasi Fungsi Kerugian dan Fungsi Tujuan .....	15
2.5. Ukuran Kesalahan Klasifikasi dalam Analisis Diskriminan.....	17
2.6. Gen .....	18
	xiii



	Hal	
2.7. Ekspresi Gen.....	18	
2.8. <i>Microarray</i> .....	19	
2.9. Kanker.....	20	
<b>BAB III METODE PENELITIAN.....</b>	<b>25</b>	
3.1. Data.....	25	
3.2. Struktur Data.....	26	
3.3. Metode Analisis Data.....	26	
<b>BAB IV HASIL DAN PEMBAHASAN .....</b>	<b>31</b>	
4.1. Statistika Deskriptif .....	31	
4.2. Identifikasi Data Hilang ( <i>Missing Value</i> ) .....	31	
4.3. Asumsi Analisis Diskriminan .....	32	
4.3.1. Asumsi Normal Multivariat.....	32	
4.3.2. Asumsi Homogenitas Matriks <i>Varians</i> <i>Kovarians</i> .....	32	
4.3.3. Identifikasi Adanya Hubungan Antar Peubah Prediktor untuk Asumsi Multikolinieritas .....	32	
4.4. Vertex Pada Ruang $\mathbb{R}^{k-1}$ .....	33	
4.5. Pemodelan VDA.....	33	
4.6. Identifikasi Ketepatan Model Klasifikasi .....	37	
<b>BAB V PENUTUP.....</b>	<b>39</b>	
5.1. Kesimpulan.....	39	
5.2. Saran .....	39	
<b>DAFTAR PUSTAKA .....</b>	<b>41</b>	



# DAFTAR TABEL

Hal

Tabel 2.1. Contoh Nilai Titik Simpul ..... 10

Tabel 2.2. Klasifikasi Dua Kelompok ..... 17

Tabel 3.1. Peubah untuk Klasifikasi Jenis Kanker ..... 25

Tabel 3.2. Kategori Jenis Kanker ..... 26

Tabel 3.3. Struktur Data ..... 26

Tabel 4.1. *Vertex* Pada Ruang  $\mathbb{R}^{k-1}$  untuk 5 Kelompok ..... 33

Tabel 4.2. Ketepatan Model VDA Pada Data *Training* ..... 34

Tabel 4.3. Nilai *Vertex*, Jarak antara Objek dan Titik *Vertex* dan Dugaan Klasifikasi ..... 35

Tabel 4.4. Hasil Klasifikasi Model VDA Pada 12 Data *Testing* ..... 36

Tabel 4.5. Ketepatan Klasifikasi Data *Training* dan Data *Testing*... 37





Repository Universitas Brawijaya



Halaman ini sengaja dikosongkan



# DAFTAR GAMBAR

Hal

Gambar 2.1. Ilustrasi Sempul (*vertex*) untuk Tiga Kelompok ..... 12

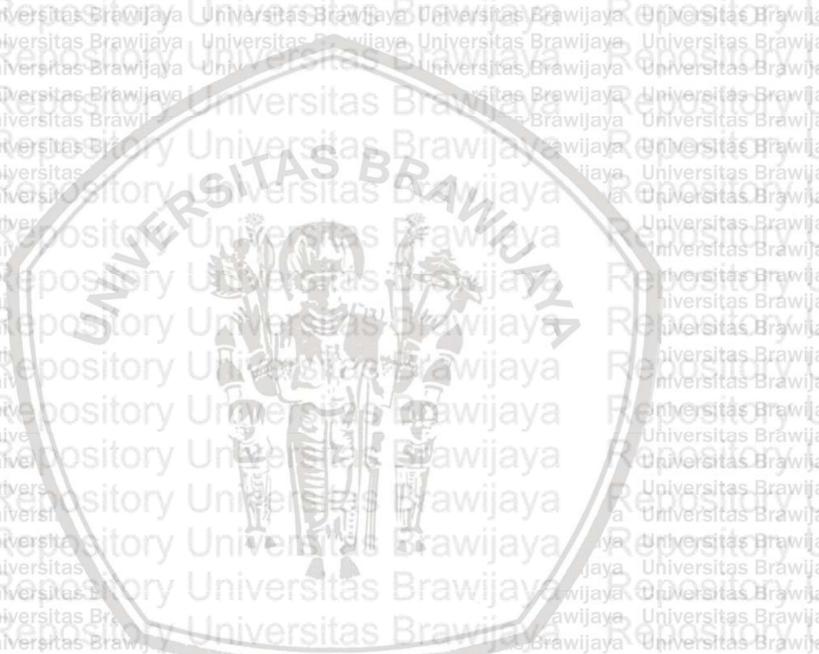
Gambar 2.2. Struktur Gen ..... 18

Gambar 2.3. Tahapan Ekspresi Gen ..... 19

Gambar 2.4. Bentuk *Microarray* ..... 20

Gambar 3.1. Diagram Alir Penelitian ..... 28

Gambar 4.1. Statistika Deskriptif ..... 31





Halaman ini sengaja dikosongkan



### DAFTAR LAMPIRAN

Hal

Lampiran 1. Data Tingkat Ekspresi Gen Pasien Penderita Kanker ..43

Lampiran 2. Banyaknya Data Hilang pada Setiap Peubah Prediktor .....47

Lampiran 3. *Output* Uji Asumsi Homogenitas Matriks *Varians Kovarians* Menggunakan *Software R* .....49

Lampiran 4. *Syntax R* dan *Output VDA* untuk data *training* .....51

Lampiran 5. Dugaan koefisien fungsi diskriminan pada metode VDA.....53

Lampiran 6. Jarak Amatan ke Titik Simpul, Kelompok Awal dan Dugaan Kelompok dengan VDA.....55





Halaman ini sengaja dikosongkan



## BAB I PENDAHULUAN

### 1.1. Latar Belakang

Metode klasifikasi adalah suatu metode pada data *mining* yang digunakan untuk memprediksi kelompok suatu data pada kelompok kelas yang sudah ada sebelumnya (Izenman, 2008). Metode klasifikasi banyak diterapkan dalam berbagai bidang ilmu seperti sosial, ekonomi, perbankan dan kedokteran. Menurut Johnson dan Wichern (2007), metode klasifikasi dalam statistika lebih menitikberatkan pada pemisahan objek-objek dan mengalokasikan objek baru pada kelompok yang telah terdefinisi.

Banyak metode statistika yang dapat digunakan untuk pengklasifikasian objek (observasi). Salah satu metode yang dapat digunakan untuk pengklasifikasian objek ke dalam suatu kelompok yang sudah ditentukan sebelumnya adalah analisis diskriminan. Menurut Mattjik dan Sumertajaya (2011), analisis diskriminan adalah salah satu teknik statistika yang digunakan pada kasus peubah respon berupa data kualitatif dan peubah prediktor berupa data kuantitatif.

Data berdimensi tinggi merupakan data yang memiliki jumlah peubah prediktor lebih banyak daripada jumlah amatan. Semakin tinggi nilai dimensi, maka akan semakin sulit memperoleh informasi. Hal ini akan menyebabkan prediksi yang didapat kurang tepat. Salah satu cara yang dapat dilakukan untuk menangani kasus berdimensi tinggi adalah dengan metode *Vertex Discriminant Analysis* (VDA).

*Vertex Discriminant Analysis* (VDA) adalah salah satu metode statistika multivariat yang digunakan untuk pengklasifikasian objek dengan mengembangkan metode analisis diskriminan menggunakan simpul (*vertex*) dari simpleks pada ruang *Euclidean* (Lange dan Wu, 2008). Metode VDA pertama kali dikembangkan oleh Lange dan Wu pada kasus multikategori linier (Lang dan Wu, 2010). Klasifikasi pada VDA dilakukan dengan cara meminimumkan fungsi tujuan yang melibatkan  $\varepsilon$ -insensitive loss dan penalti kuadrat. Penambahan penalti kuadrat digunakan dalam seleksi peubah dengan cara memperkecil koefisien parameter yang berkorelasi, yakni menuju nilai nol (Hastie dkk., 2008). Analisis diskriminan klasik mengasumsikan bahwa data dari masing-masing kelompok memiliki matriks *varians kovarians*



yang homogen dan data berdistribusi normal multivariat. Apabila data tidak memenuhi asumsi, maka dapat menggunakan metode VDA.

Penelitian menggunakan data berdimensi tinggi telah dilakukan sebelumnya oleh Zharfani (2019), tentang model klasifikasi analisis diskriminan dengan metode *Partial Least Squares Discriminant Analysis* (PLSDA) pada data berdimensi tinggi. Penelitian ini menghasilkan nilai akurasi sebesar 100%. Penelitian lain juga telah dilakukan oleh Nurmaleni (2015), tentang perbandingan metode klasifikasi antara *Vertex Discriminant Analysis* (VDA) dan *Fisher Discriminant Analysis* (FDA). Dalam penerapannya, FDA tidak dapat digunakan untuk klasifikasi ketika matriks  $X$  tidak berpangkat penuh, yaitu pada saat banyaknya peubah prediktor jauh lebih besar dari banyak observasi (data berdimensi tinggi). Hal ini akan menyebabkan matriks *varians kovarians* singular, sehingga tidak memiliki matriks kebalikan (*inverse matrix*). VDA mampu mengklasifikasikan objek ketika matriks  $X$  berpangkat penuh dan tidak berpangkat penuh. Oleh karena itu, pada penelitian ini, peneliti menggunakan metode VDA untuk memodelkan klasifikasi penyakit kanker berdasarkan tingkat ekspresi gen pasien penderita kanker pada kasus data berdimensi tinggi.

Kanker merupakan salah satu penyakit yang menjadi penyebab utama kematian di dunia. Kanker muncul akibat pertumbuhan sel abnormal dalam tubuh manusia yang berkembang dengan cepat dan tak terkendali. Kanker terjadi karena perubahan atau mutasi pada gen dalam sebuah sel yang menyebabkan sel tumbuh di luar kendali dari bagian tubuh lainnya, sehingga sel kanker dapat menyebar ke bagian tubuh yang jauh dari tempat sel tersebut berasal (Chang dkk., 2010).

Salah satu tahapan penting dalam penanganan kanker adalah klasifikasi kanker. Hal ini sangat diperlukan sebagai pendukung diagnosa, sehingga mempermudah ahli medis untuk memberikan penanganan yang tepat sesuai dengan kondisi sebenarnya. Terdapat beberapa jenis kanker dengan ciri fisik yang sama tetapi jenisnya berbeda. Apabila terdapat kesalahan dalam diagnosa, maka akan memperburuk kondisi pasien. Tingkat ekspresi gen dapat digunakan untuk menentukan jenis sel kanker dari seorang penderita. Karena terdapat banyak gen yang menyusun penyakit kanker, maka dibutuhkan metode untuk mengklasifikasikan jenis kanker yang

diderita oleh pasien dimana peubah prediktor cukup banyak, sehingga dapat memberikan prediksi data baru dengan tepat.

## 1.2. Rumusan Masalah

Rumusan masalah pada penelitian ini adalah:

1. Bagaimana fungsi diskriminan yang terbentuk untuk mengklasifikasikan pasien ke dalam lima kategori jenis kanker, yaitu BRCA (*Breast Invasive Carcinoma*), COAD (*Colon Adenocarcinoma*), KIRC (*Kidney Renal Clear Cell Carcinoma*), LUAD (*Lung Adenocarcinoma*) dan PRAD (*Prostate Adenocarcinoma*)?
2. Berapa besar ketepatan hasil klasifikasi pada fungsi diskriminan yang terbentuk?

## 1.3. Tujuan Penelitian

Tujuan dari penelitian ini adalah:

1. Membentuk fungsi diskriminan untuk mengklasifikasikan pasien ke dalam lima kategori jenis kanker (BRCA, COAD, KIRC, LUAD dan PRAD).
2. Memperoleh tingkat ketepatan hasil klasifikasi dari fungsi diskriminan yang terbentuk.

## 1.4. Manfaat Penelitian

Manfaat dari penelitian ini adalah:

1. Model diharapkan mampu mengklasifikasikan objek ketika jumlah peubah prediktor lebih banyak dibandingkan dengan jumlah amatan, sehingga diperoleh ketepatan prediksi.
2. Bagi instansi kesehatan atau ahli medis, diharapkan mampu mempermudah dalam menentukan klasifikasi seseorang yang terkena kanker, sehingga dapat memberikan penanganan yang tepat sesuai dengan kondisi sebenarnya.

## 1.5. Batasan Masalah

Batasan masalah pada penelitian ini adalah:

1. Kanker yang digunakan hanya lima jenis, yaitu (BRCA, COAD, KIRC, LUAD dan PRAD).
2. Jumlah gen yang digunakan hanya sebanyak 685 gen pertama dengan amatan sebanyak 60 pasien yang telah dipublikasikan oleh UCI *Machine Learning Repository*.



Halaman ini sengaja dikosongkan



## BAB II TINJAUAN PUSTAKA

### 2.1. Analisis Diskriminan

Analisis multivariat merupakan metode dalam melakukan penelitian terhadap lebih dari dua peubah secara bersamaan. Berdasarkan hubungan antar peubah, analisis multivariat dibedakan menjadi teknik dependensi dan teknik interdependensi. Analisis diskriminan termasuk ke dalam teknik dependensi, yaitu peubahnya saling bergantung satu dengan yang lain.

Analisis diskriminan adalah analisis multivariat yang digunakan pada kasus yang memiliki peubah respon berupa data nonmetrik (nominal atau ordinal) dan peubah prediktor berupa data metrik (interval atau rasio) (Johnson dan Wichern, 2002). Tujuan dari analisis diskriminan adalah untuk menentukan peubah prediktor mana yang memberikan pengaruh terbesar terhadap terjadinya perbedaan kelompok dan untuk mengklasifikasikan peubah prediktor ke dalam suatu kelompok berdasarkan nilai peubah respon (Hair dkk., 2010).

Menurut Hair dkk. (2010), fungsi diskriminan ditunjukkan pada persamaan (2.1).

$$D_k = a + W_1X_{1k} + W_2X_{2k} + W_3X_{3k} + \dots + b_pX_{pk} \quad (2.1)$$

Keterangan:

$D$  : skor diskriminan

$a$  : intersep

$W$  : koefisien diskriminan atau bobot

$X$  : peubah prediktor

$p$  : banyaknya peubah prediktor

$k$  : banyaknya kelompok

### 2.2. Asumsi Analisis Diskriminan

#### 2.2.1. Asumsi Normal Multivariat

Menurut Johnson dan Winchern (1992), terpenuhinya asumsi normal multivariat pada peubah prediktor dapat dilakukan dengan melihat Q-Q plot antara jarak Mahalanobis untuk pengamatan ke- $i$  ( $D_i^2$ ) dengan nilai quantil dari distribusi *Chi-Square*  $\left(\chi_p^2 \left(\frac{i-0,5}{n}\right)\right)$ . Jarak Mahalanobis untuk pengamatan ke- $i$  ( $D_i^2$ ) dapat dihitung dengan menggunakan persamaan (2.2).

$$D_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) \quad (2.2)$$

Keterangan:

$i : 1, 2, 3, \dots, n$

$\mathbf{x}_i$  : vektor pengamatan ke- $i$

$\mathbf{S}^{-1}$  : inverse matriks varians kovarians

Masing-masing  $D_i^2$  akan mengikuti distribusi *Chi-Square*. Plot *Chi-Square* akan memeriksa apakah  $D_i^2$  mengikuti distribusi *Chi-Square* dengan mengurutkan  $D_i^2$  dari terkecil ke terbesar ( $D_1^2 \leq D_2^2 \leq \dots \leq D_n^2$ ), kemudian memplotkan  $D_i^2$  dengan  $\left(\chi_p^2 \left(\frac{i-0,5}{n}\right)\right)$ . Apabila plot yang terbentuk cenderung membentuk garis lurus dan terdapat lebih dari 50% dari seluruh jumlah amatan yang memiliki nilai  $D_i^2 < \chi_{p, (0,5)}^2$ , maka data mengikuti distribusi normal multivariat (Johnson dan Winchern, 1992).

Pengujian asumsi normal multivariat juga dapat dilakukan dengan menggunakan statistik uji seperti pada persamaan (2.3).

Hipotesis:

$H_0$ : Data berdistribusi normal multivariat

$H_1$ : Data tidak berdistribusi normal multivariat

Statistik Uji:

$$r_q = \frac{\sum_{i=1}^n (x_i - \bar{x})(q_i - \bar{q})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (q_i - \bar{q})^2}} \quad (2.3)$$

Apabila  $r_q < r_{n,\alpha}$ , artinya terima  $H_0$ , maka data mengikuti distribusi normal multivariat.

### 2.2.2. Asumsi Homogenitas Matriks Varians Kovarians

Pengujian asumsi homogenitas matriks *variens kovarians* dapat dilakukan dengan menggunakan statistik uji Box's M. Hipotesis uji Box's M, yaitu:

$H_0$ :  $\Sigma_1 = \Sigma_2 = \dots = \Sigma_k$

$H_1$ : setidaknya ada sepasang kelompok yang memiliki matriks *variens kovarians* berbeda antar kelompok

Prosedur pada uji Box's M menggunakan pendekatan distribusi *Chi-Square* dengan derajat bebas  $\frac{1}{2}(k-1)p(p+1)$ .

Statistik uji Box's M ditunjukkan pada persamaan (2.4).

$$M = \sum_{i=1}^k (n_i - 1) \ln|S| - \sum_{i=1}^k (n_i - 1) \ln|S_i| \quad (2.4)$$

dengan:

$$S = \frac{\sum_{i=1}^k (n_i - 1) S_i}{\sum_{i=1}^k n_i - k}$$

$$S_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)^T$$

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

Keterangan:

$S$  : matriks kombinasi *varians kovarians* setiap kelompok

$S_i$  : matriks *varians kovarians* kelompok ke- $i$

$\bar{x}_i$  : vektor rata-rata kelompok ke- $i$

Apabila  $M < \chi_{\alpha/2, (k-1)p(p+1)}^2$ , artinya terima  $H_0$ , maka matriks *varians kovarians* homogen.

### 2.2.3. Analisis Korelasi untuk Asumsi Multikolinieritas

Menurut Walpole (1993), analisis korelasi merupakan suatu analisis yang digunakan untuk mengukur kekuatan hubungan antara dua peubah melalui sebuah bilangan yang disebut koefisien korelasi. Koefisien korelasi memiliki rentang nilai antara -1 sampai 1. Apabila nilai koefisien korelasi bernilai 0, maka dapat dikatakan bahwa kedua peubah tidak memiliki hubungan. Apabila nilai koefisien korelasi bernilai -1 atau 1, maka dapat dikatakan bahwa kedua peubah memiliki hubungan yang erat.

Jenis-jenis analisis korelasi yang sering digunakan, yaitu korelasi *pearson*, korelasi *rank spearman*, korelasi *kendall tau* dan koefisien kontingensi. Analisis korelasi dapat dilakukan dengan beberapa pengujian tergantung pada jenis data. Uji korelasi Pearson merupakan pengujian korelasi ketika jenis data bersifat kuantitatif (data berskala interval atau rasio). Dalam pengujian akan menghasilkan koefisien korelasi yang digunakan untuk menyatakan besar hubungan linier antara dua peubah.

Menurut Walpole (1993), rumus untuk menghitung koefisien korelasi *Pearson* ditunjukkan pada persamaan (2.5).



$$r = \frac{\sum X_1 X_2 - \frac{\sum X_1 \sum X_2}{n}}{\sqrt{\sum X_1^2 - \frac{(\sum X_1)^2}{n}} \sqrt{\sum X_2^2 - \frac{(\sum X_2)^2}{n}}} \quad (2.5)$$

Keterangan:

- $r$  : koefisien korelasi
- $n$  : banyaknya amatan

Statistik uji yang digunakan adalah statistik uji t, sesuai dengan persamaan (2.6).

$$t_{hitung} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad (2.6)$$

Hipotesis untuk pengujian korelasi adalah:

- $H_0 : \rho = 0$  vs
- $H_1 : \rho \neq 0$

Keputusan yang diambil, yaitu tolak  $H_0$  apabila  $|t_{hitung}| > t_{(\frac{\alpha}{2}, n-2)}$  atau dengan melihat  $p$ -value  $< \alpha$ , artinya terdapat hubungan yang signifikan antara dua peubah. Sedangkan terima  $H_0$  apabila  $|t_{hitung}| \leq t_{(\frac{\alpha}{2}, n-2)}$  atau dengan melihat  $p$ -value  $\geq \alpha$ , artinya tidak terdapat hubungan yang signifikan antara dua peubah.

### 2.3. Data Hilang (Missing Value)

Salah satu permasalahan yang sering dijumpai dalam suatu penelitian adalah data hilang (*missing value*). Data hilang adalah suatu kondisi hilangnya sebagian informasi atau fitur pada sebuah data set, yang terjadi karena informasi tentang objek tidak diberikan, sulit dicari, atau memang informasi tersebut tidak ada. Data hilang juga dapat disebabkan oleh kesalahan sistem, dapat pula disebabkan oleh kesalahan manusia (*human error*), seperti kelalaian dalam pengumpulan data maupun ketidaklengkapan ketika proses *input* data.

Banyaknya data hilang pada suatu data set pada dasarnya tidak akan bermasalah apabila hanya terdapat sekitar 1% dari keseluruhan data, sehingga data hilang tersebut dapat diabaikan. Namun, apabila persentase data hilang cukup besar maka perlu dilakukan penanganan.

Menurut Davey dan Savla (2010), beberapa cara yang dapat dilakukan untuk menangani data hilang adalah *listwise deletion*, *pairwise deletion* dan imputasi. Salah satu cara imputasi, yaitu dengan

metode *mean* (menghitung rata-rata dari semua nilai yang diketahui pada suatu peubah).

## 2.4. Vertex Discriminant Analysis (VDA)

*Vertex Discriminant Analysis* (VDA) adalah salah satu metode statistika multivariat yang digunakan untuk pengklasifikasian objek dengan mengembangkan metode analisis diskriminan menggunakan simpul (*vertex*) dari simpleks pada ruang *Euclidean* (Lange dan Wu, 2008). VDA dapat digunakan pada kasus data berdimensi tinggi (*high dimensional*), yaitu ketika banyaknya peubah prediktor lebih besar dari banyaknya amatan.

Masalah yang akan muncul akibat dimensi tinggi adalah *overfitting*. *Overfitting* merupakan suatu keadaan ketika model sudah sesuai dengan data contoh (data *training*), namun kurang baik ketika digunakan untuk memprediksi data yang bukan bagian dari data penyusun model (data *testing*). Hal ini dapat ditangani dengan cara regularisasi, yaitu mengatur estimasi koefisien regresi dengan menambahkan aturan penalti yang mampu mengecilkan estimasi menuju nilai sebenarnya. Regularisasi yang dilakukan menggunakan kalibrasi menggunakan *cross validation* (Lange dan Wu, 2008).

### 2.4.1. Vertex Pada Ruang $\mathbb{R}^{k-1}$

Simpul (*vertex*) adalah suatu titik yang memiliki jarak sama pada ruang simpleks dalam *Euclidean* yang digunakan untuk pemberian kode indikator kelas dari masing-masing kelompok. Masing-masing simpul menunjukkan kelompok yang berbeda. VDA menggunakan simpul berjarak sama di ruang *Euclidean*  $\mathbb{R}^{k-1}$  untuk pemberian kode indikator kelas yang berbeda pada setiap kelompok, agar setiap simpul mempunyai jarak yang sama.

Menurut Lange dan Wu (2008), penentuan titik simpul pada simpleks reguler dapat diperoleh menggunakan persamaan (2.7).

$$v_j = \begin{cases} (k-1)^{-1/2} \mathbf{1}, & \text{untuk } j = 1 \\ c \mathbf{1} + d e_{j-1}, & \text{untuk } 2 \leq j \leq k \end{cases} \quad (2.7)$$

dengan:

$$c = -\frac{1+\sqrt{k}}{(k-1)^{3/2}}$$

$$d = \sqrt{\frac{k}{k-1}}$$



Keterangan:

$k$  : banyaknya kelompok

$\mathbf{1}$  : matriks berukuran  $(1 \times n_j)$

$\mathbf{e}_j$  : vektor koordinat ke- $j$  dalam  $\mathbb{R}^{k-1}$  (bernilai 1 pada saat ke- $j$  dan 0 untuk lainnya).

Beberapa contoh nilai titik simpul untuk beberapa kategori di ruang  $\mathbb{R}^{k-1}$  disajikan pada Tabel 2.1.

Tabel 2.1. Contoh Nilai Titik Simpul

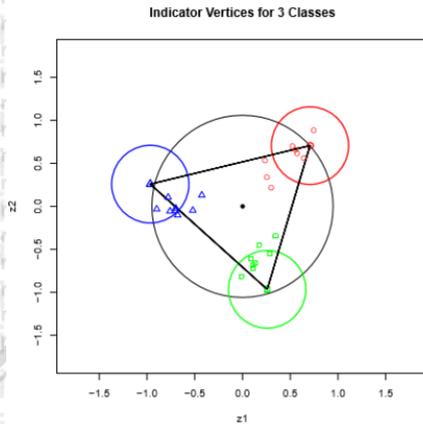
Banyak kelompok (k)	$\mathbf{v}_j$	Keterangan	Bentuk
2	$\mathbf{v}_1 = 1,000$	$c = -\frac{1 + \sqrt{2}}{\sqrt{1^3}}$ ,	Garis (line)
	$\mathbf{v}_2 = -1,000$	$d = \sqrt{\frac{2}{1}}$	
3	$\mathbf{v}_1 = \begin{bmatrix} 0,707 \\ 0,707 \end{bmatrix}$	$c = -\frac{1 + \sqrt{3}}{\sqrt{2^3}}$ ,	Tiga simpul segitiga sama sisi yang dibatasi oleh lingkaran (bidang)
	$\mathbf{v}_2 = \begin{bmatrix} 0,258 \\ -0,965 \end{bmatrix}$	$d = \sqrt{\frac{3}{2}}$ ,	
	$\mathbf{v}_3 = \begin{bmatrix} -0,965 \\ 0,258 \end{bmatrix}$	$\mathbf{e}_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ , $\mathbf{e}_3 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$	

Tabel 2.1. Contoh Nilai Titik Simpul (Lanjutan)

Banyak kelompok (k)	$v_j$	Keterangan	Bentuk
4	$v_1 = \begin{bmatrix} 0,577 \\ 0,577 \\ 0,577 \end{bmatrix}$	$c = -\frac{1 + \sqrt{4}}{\sqrt{3^3}}$	
	$v_2 = \begin{bmatrix} -0,422 \\ -0,577 \\ -0,577 \end{bmatrix}$	$d = \sqrt{\frac{4}{3}}$	
	$v_3 = \begin{bmatrix} -0,577 \\ -0,422 \\ -0,577 \end{bmatrix}$	$e_2 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$	
	$v_4 = \begin{bmatrix} -0,577 \\ -0,577 \\ -0,422 \end{bmatrix}$	$e_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$	
		$e_4 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$	

Keterangan:  
 $v_j$  : titik simpul kelompok ke-j

Ilustrasi mengenai simpul untuk tiga kelompok pada  $\mathbb{R}^2$  disajikan pada Gambar 2.1.



Gambar 2.1. Ilustrasi Simpul untuk Tiga Kelompok

$Z_1$  merupakan titik koordinat pertama atau absis dan  $Z_2$  merupakan titik koordinat kedua atau ordinat dalam ruang *Euclidean*. Untuk  $k$ -kelompok, simpul  $v_1, \dots, v_k$  dari simpleks regular berada dalam  $\mathbb{R}^{k-1}$ .

#### 2.4.2. Fungsi Kerugian (*Loss Function*)

Fungsi kerugian (*loss function*) adalah fungsi yang menunjukkan hubungan antara *error* dengan bagaimana *error* tersebut dikenai sebuah penalti. Fungsi kerugian dinotasikan dengan  $L(y, x)$ . Fungsi kerugian yang digunakan pada VDA adalah  $\varepsilon$ -insensitive loss.  $\varepsilon$ -insensitive loss pada regresi dijelaskan pada persamaan (2.8).

$$L(y, x) = |y - a^t x - b|_\varepsilon \quad (2.8)$$

dengan  $|u|_\varepsilon = \max\{|u| - \varepsilon, 0\}$  (Vapnik 1995; Hastie dkk., 2001).

Kerugian  $\varepsilon$ -insensitif lebih resisten terhadap pencilan dibandingkan dengan *squared error loss* (Liu dkk., 2005). Pengklasifikasian dilakukan dengan memilih jarak terdekat antara penduga linier dengan indikator kelas yang mungkin. Kerugian  $\varepsilon$ -insensitif baik digunakan pada dimensi tinggi.

Agar nilai dugaan dapat mendekati nilai populasi, maka fungsi kerugian diminimumkan. Misalkan  $Y$  menunjukkan indikator kelas



dan  $X$  menunjukkan vektor peubah pencari dari observasi acak. Vektor  $Y$  bertepatan dengan salah satu simpul tersebut. Fungsi tujuan merupakan nilai harapan dari fungsi kerugian. Meminimumkan fungsi tujuan sama dengan meminimumkan nilai harapan kerugian. Diketahui fungsi kerugian  $L(y, x)$ , analisis diskriminan berusaha untuk meminimumkan nilai harapan kerugian seperti pada persamaan (2.9).

$$E[L(Y, X)] = E\{[L(Y, X)|X]\} \quad (2.9)$$

Untuk meminimumkan kerugian dapat dilakukan diferensiasi/turunan, tetapi fungsi kerugian  $\epsilon$ -insensitif tidak dapat diturunkan. Hal ini sulit dilakukan, sehingga untuk pendugaan parameter dilakukan dengan cara meminimumkan rata-rata kerugian bersyarat  $n^{-1} \times \sum_{i=1}^n L(y_i, x_i)$  dengan menambahkan batas penalti.

### 2.4.3. Fungsi Tujuan Pada VDA

VDA linier mengasumsikan model regresi linier  $f(x) = Ax + b$ , dengan  $A = (a_{jm})$  adalah matriks koefisien *slope* berukuran  $(k - 1) \times p$ , dan  $b = (b_j)$  adalah vektor kolom dari intersep yang berukuran  $(k - 1) \times 1$ . Fungsi diskriminan VDA ditunjukkan pada persamaan (2.10).

$$f(x) = Ax + b \quad (2.10)$$

Setelah dilakukan identifikasi terhadap kelas indikator dengan *vertex*, fungsi tujuan (fungsi kerugian yang dinormalisasi) terdiri dari fungsi kerugian dan penalti, didefinisikan sesuai dengan persamaan (2.11).

$$R(\theta) = \frac{1}{n} \sum_{j=1}^k \sum_{i=1}^{n_j} \|v_{j(i)} - Ax_{j(i)} - b\|_{\epsilon} + \lambda \sum_{j=1}^{k-1} \|a_j\|^2 \quad (2.11)$$

dengan:

$$\|v_{j(i)} - Ax_{j(i)} - b\|_{\epsilon} = \max \begin{cases} 0, & \text{jika } \|u\| < \epsilon \\ \|u\| - \epsilon, & \text{jika } \|u\| \geq \epsilon \end{cases} \quad (2.12)$$

$$\epsilon = \frac{1}{2} \sqrt{\frac{2k + 2}{k}}$$

Keterangan:

$\theta(A, b)$  : himpunan semua parameter yang tidak diketahui

$v_{j(i)}$  : penempatan titik *vertex* ke- $j$  untuk observasi ke- $i$

$a_j$  : baris ke- $j$  dari matriks koefisien regresi  $A$  yang berukuran  $(k - 1) \times p$

$b$  : vektor kolom intersep berukuran  $(k - 1) \times 1$

Fungsi tujuan pada VDA dapat diminimumkan dengan menggunakan algoritma *Majorize-Minimize* (MM) (Lange 2004; Hunter dan Lange 2004). Algoritma MM digunakan untuk mendekati fungsi kerugian dengan fungsi kuadrat dan diminimumkan dengan menggunakan *reweighted least squares* (Lange, 2004).

#### 2.4.4. Algoritma MM (*Majorize Minimize*) Pada VDA

Algoritma MM bukanlah suatu algoritma, melainkan suatu deskripsi untuk membangun algoritma optimasi. Algoritma MM (*Majorize Minimize*) mampu memecahkan masalah optimasi dan estimasi dimensi tinggi. Menurut Hunter dan Lange (2003); Lange dan Wu (2010), algoritma MM dapat digunakan untuk menyederhanakan masalah optimasi yang sulit dengan cara iterasi. Dalam meminimumkan, MM berarti mayorisasi minimum (*Majorize Minimize*), sedangkan dalam memaksimumkan, MM berarti minorisasi maksimum (*Minorize Maximize*) (Lange dan Wu, 2008). Algoritma MM pada VDA berarti mayorisasi minimum (*Majorize Minimize*) karena algoritma ini dibutuhkan untuk meminimumkan fungsi tujuan.

Algoritma MM dapat digunakan untuk meminimisasi fungsi kerugian dengan optimasi iterasi. Persamaan (2.11) dapat diminimalisasi dengan cara menggabungkan algoritma MM dan estimasi *ridge* terkecil dari algoritma MM. Mayorisasi dilakukan dengan memperoleh fungsi pengganti  $g(\theta|\theta^m)$  untuk fungsi tujuan  $f(\theta)$  pada titik tertentu di iterasi  $\theta^m$ . Fungsi  $g(\theta|\theta^m)$  dapat dikatakan sebagai mayorisasi (*majorize*) fungsi  $f(\theta)$  pada  $\theta^m$  apabila  $f(\theta^m) = g(\theta^m|\theta^m)$

$f(\theta) \leq g(\theta|\theta^m)$ , untuk  $\theta \neq \theta^m$  (2.13)

Untuk mencari fungsi pengganti  $g(\theta|\theta^m)$  yang optimum dapat diperoleh melalui tahapan mayorisasi dan meminimumkan pada tahap minimisasi, sehingga menghasilkan nilai dugaan  $\theta^{m+1}$ . Ketika fungsi pengganti dioptimalkan, fungsi tujuan akan didorong agar meningkat atau menurun sampai optimum lokal. Jika  $\theta^{m+1}$  meminimumkan  $g(\theta|\theta^m)$  maka tahapan minimisasi memaksa sifat turun  $f(\theta^{m+1}) \leq f(\theta^m)$ . Fungsi ini mengikuti bentuk petidaksamaan (2.14).

$$\begin{aligned}
 f(\theta^{m+1}) &= g(\theta^{m+1}|\theta^m) + f(\theta^{m+1}) - g(\theta^{m+1}|\theta^m) \\
 &\leq g(\theta^m|\theta^m) + f(\theta^m) - g(\theta^m|\theta^m) \\
 &= f(\theta^m)
 \end{aligned}
 \tag{2.14}$$

sehingga mengakibatkan  $g(\theta^{m+1}|\theta^m) \leq g(\theta^m|\theta^m)$ . Sifat turun ini akan menghasilkan algoritma MM yang sangat stabil.

Algoritma MM ideal digunakan untuk tujuan komputasi fungsi *ridge*. Menurut Lange dan Wu (2010), algoritma MM bertujuan untuk mencapai konvergen minimum global untuk fungsi tujuan yang sangat cembung (*non-trivial*), sehingga solusi yang dihasilkan adalah solusi yang unik.

#### 2.4.5. Mayorisasi Fungsi Kerugian dan Fungsi Tujuan

Mayorisasi fungsi kerugian  $f(x) = \|x\|_\epsilon$  yang didefinisikan seperti pada persamaan (2.12) diperoleh dari pertidaksamaan *Cauchy-Schwarz*. Menurut Lange dan Wu (2008), pertidaksamaan *Cauchy-Schwarz*  $g(x|x^m)$  untuk mayorisasi fungsi  $f(x) = \|x\|_\epsilon$  dijelaskan pada persamaan (2.15).

$$g(x|x^m) = \begin{cases} \frac{1}{2\|x^m\|} \|x\|^2 + \frac{1}{2} \|x^m\| - \epsilon, & \text{untuk } \|x^m\| \geq 2\epsilon \\ \frac{1}{4(\epsilon - \|x^m\|)} \|x - x^m\|^2, & \text{untuk } \|x^m\| < \epsilon \\ \frac{1}{4(\epsilon - \|z\|)} \|x - z\|^2, & \text{untuk } \epsilon < \|x^m\| < 2\epsilon \end{cases}
 \tag{2.15}$$

Mayorisasi dari fungsi tujuan dengan meminimalisasi fungsi pengganti dengan cara mereduksi menjadi pendugaan kuadrat terkecil terboboti (Lange dan Wu, 2008). Berdasarkan persamaan (2.15), mayorisasi fungsi tujuan dapat ditulis seperti persamaan (2.16).

$$\begin{aligned}
 R(\mathbf{A}, \mathbf{b}) &\leq \frac{1}{n} \sum_{i=1}^n w_i \|r_i - s_i\|^2 + \lambda \sum_{j=1}^k \|a_j\|^2 + d \\
 &= \sum_{j=1}^k \left[ \frac{1}{n} \sum_{i=1}^n w_i (r_{ij} - s_{ij})^2 + \lambda \|a_j\|^2 \right] + d
 \end{aligned}
 \tag{2.16}$$

untuk  $r_i = y_i - Ax_i - b$ , dengan pembobot:

$$w_i = \begin{cases} \frac{1}{2 \|r_i^m\|} & \text{untuk } \|r_i^m\| \geq 2\varepsilon \\ \frac{1}{4(\varepsilon - \|r_i^m\|)} & \text{untuk } \|r_i^m\| < \varepsilon \\ \frac{1}{4(\|r_i^m\| - \varepsilon)} & \text{untuk } \varepsilon < \|r_i^m\| < 2\varepsilon \end{cases} \quad (2.17)$$

dan argument shift:

$$s_i = \begin{cases} 0, & \text{untuk } \|r_i^m\| \geq 2\varepsilon \\ r_i^m, & \text{untuk } \|r_i^m\| < \varepsilon \\ \left(\frac{2\varepsilon}{\|r_i^m\|} - 1\right) r_i^m, & \text{untuk } \varepsilon < \|r_i^m\| < 2\varepsilon \end{cases} \quad (2.18)$$

konstanta  $d$  bergantung pada sisaan  $r_i$  pada iterasi ke- $m$ .

Menurut Golub dan Loan (1996), meminimalisasi fungsi pengganti akan memperkecil dugaan pembobot kuadrat terkecil. Meminimumkan fungsi pengganti pada algoritma MM dilakukan untuk mendapatkan penduga  $A^{(m+1)}$  dan  $b^{(m+1)}$  pada iterasi  $(m + 1)$ . Hal tersebut dilakukan dengan cara memecahkan  $k$  pasang persamaan linier menggunakan *Cholesky Decomposition* tunggal pada setiap iterasi. Digunakan *Cholesky Decomposition* tunggal karena matriks  $x^t W x$  untuk semua  $k$  sama, dengan  $x$  adalah baris ke- $i$  ( $x_i^t, 1$ ) dan  $W$  adalah matriks diagonal dengan nilai diagonalnya adalah  $w_i$ . Prediksi  $k$  pasang persamaan linier  $y_i = Ax_i + b$  pada observasi ke- $i$  ditunjukkan pada persamaan (2.19).

$$v_{j(i)} = \begin{pmatrix} v_{i,1} \\ \vdots \\ v_{i,k} \end{pmatrix} = \begin{pmatrix} a_{11} & \dots & a_{1p} \\ \vdots & \ddots & \vdots \\ a_{k,1} & \dots & a_{k,p} \end{pmatrix} \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix} + \begin{pmatrix} b_1 \\ \vdots \\ b_k \end{pmatrix} \quad (2.19)$$

Sedangkan untuk  $n$  observasi ditunjukkan pada persamaan (2.20).

$$\begin{pmatrix} v_{i,1} \\ \vdots \\ v_{i,k} \end{pmatrix} = \begin{pmatrix} a_{11} & \dots & a_{1p} \\ \vdots & \ddots & \vdots \\ a_{k,1} & \dots & a_{k,p} \end{pmatrix} \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix} + \begin{pmatrix} b_1 \\ \vdots \\ b_k \end{pmatrix} \quad (2.20)$$

Bentuk persamaan (2.20) jika diperhatikan sama dengan model regresi multivariat untuk  $n$  observasi. Klasifikasi respon untuk kasus baru pada metode VDA diprediksi dengan persamaan (2.21).

$$\hat{y} = \operatorname{argmin}_{j=1, \dots, k+1} \| v_j - \hat{A}x_i - \hat{b} \| \quad (2.21)$$

## 2.5. Ukuran Kesalahan Klasifikasi dalam Analisis Diskriminan

*Apparent Error Rate* (APER) adalah salah satu metode yang digunakan untuk mengevaluasi analisis diskriminan dalam pengklasifikasian. Nilai APER menurut Johnson dan Wichern (2002) adalah banyaknya persentase yang salah dalam pengelompokannya oleh fungsi klasifikasi. Perhitungan nilai APER dilakukan dengan menggunakan matriks konfusi (*confusion matrix*). Contoh perhitungan persentase kesalahan klasifikasi untuk analisis diskriminan dua kelompok ( $k=2$ ) ditunjukkan pada Tabel 2.2. Perhitungan tersebut dilakukan dengan terlebih dahulu membentuk tabulasi antara nilai aktual dan nilai prediksi dari setiap kelompok. Kelompok 1 dinotasikan dengan  $\pi_1$ , kelompok 2 dinotasikan dengan  $\pi_2$ , dan  $n$  merupakan banyaknya observasi.

Tabel 2.2 Klasifikasi Dua Kelompok

Aktual	Prediksi		Total
	$\pi_1$	$\pi_2$	
$\pi_1$	$n_{11}$	$n_{12}$	$n_{11} + n_{12}$
$\pi_2$	$n_{21}$	$n_{22}$	$n_{21} + n_{22}$
Total	$n_{11} + n_{21}$	$n_{12} + n_{22}$	$N = n_{11} + n_{12} + n_{21} + n_{22}$

Sehingga nilai APER dapat didefinisikan sesuai dengan persamaan (2.22).

$$APER = \frac{n_{12} + n_{21}}{N} \quad (2.22)$$

Keterangan:

$n_{11}$  : banyaknya objek dari  $\pi_1$  yang tepat diklasifikasikan sebagai  $\pi_1$

$n_{12}$  : banyaknya objek dari  $\pi_1$  yang salah diklasifikasikan sebagai  $\pi_2$

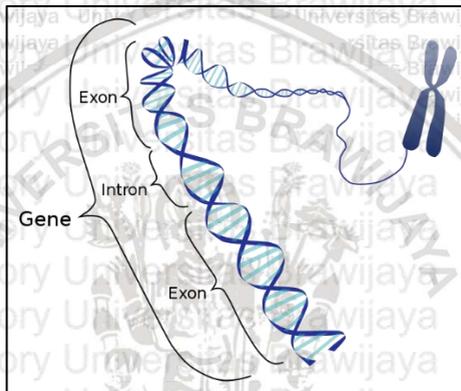
$n_{21}$  : banyaknya objek dari  $\pi_2$  yang salah diklasifikasikan sebagai  $\pi_1$

$n_{22}$  : banyaknya objek dari  $\pi_2$  yang tepat diklasifikasikan sebagai  $\pi_2$

Semakin banyak objek yang terdeteksi benar maka nilai kesalahan klasifikasinya semakin kecil. Nilai kesalahan klasifikasi berada pada rentang nilai 0 sampai 1, dimana 0 merupakan nilai dengan kesalahan klasifikasi paling rendah.

## 2.6. Gen

Gen merupakan unit terkecil materi genetik. Gen adalah substansi genetik terkecil yang tersusun atas DNA (*Deoxyribonucleic Acid*) yang menentukan sifat individu melalui pembentukan polipeptida. Sekumpulan DNA ini akan membentuk struktur padat yang disebut kromosom. Bentuk fisik gen adalah urutan DNA yang berada di suatu protein, polipeptida, atau seuntai RNA yang berfungsi bagi organisme yang memilikinya. Ilustrasi mengenai struktur gen dapat dilihat pada Gambar 2.2.



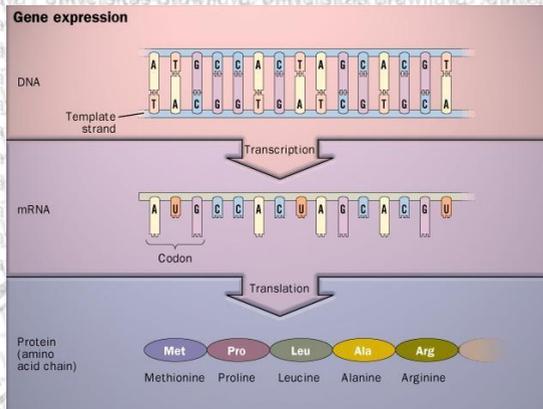
Gambar 2.2. Struktur Gen

Gen berfungsi untuk perkembangan dan aktivitas dari organ-organ dan segala hal yang berhubungan dengan tubuh manusia. Untuk melaksanakan fungsi dari gen yang berada dalam tubuh manusia, dibutuhkan suatu proses yang disebut dengan ekspresi gen.

## 2.7. Ekspresi Gen

Ekspresi gen adalah suatu rangkaian proses penerjemahan informasi genetik dalam bentuk urutan basa pada DNA atau RNA menjadi produk sintesis berupa protein atau polipeptida hingga fenotip, yaitu sifat yang dapat diamati. Ekspresi gen merupakan suatu proses yang dimulai dari inisialisasi untuk proses transkripsi DNA, kemudian berinteraksi dengan sekuens yang sesuai (menghasilkan *primary RNA*) dan proses *post transcription modification* yang menghasilkan mRNA. Proses selanjutnya, yaitu proses translasi mRNA yang menghasilkan polipeptida, kemudian membentuk asam

amino yang akan menjadi protein yang akan digunakan untuk fungsi seluler. Tahapan ekspresi gen dapat dilihat pada Gambar 2.3.



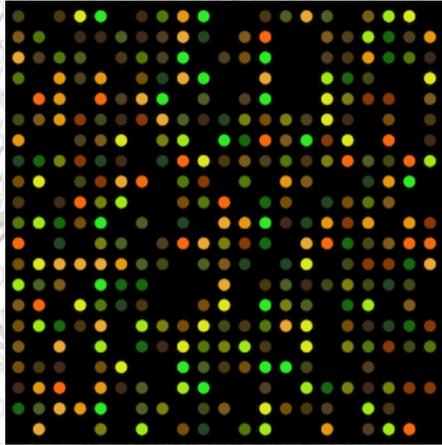
Gambar 2.3. Tahapan Ekspresi Gen

Menurut Stefanka dan MacEwan (2017), terdapat RNA *microarrays* dan RNA *sequence* (untaian RNA) yang membantu mempelajari transkriptom. Transkriptom merupakan seperangkat transkrip RNA lengkap yang diproduksi oleh genom dalam keadaan tertentu dalam sel atau jaringan tertentu. Perbandingan transkriptom memungkinkan identifikasi gen yang diekspresikan secara berbeda dalam populasi sel yang berbeda. Hal ini akan membantu menjelaskan mekanisme penyakit, terutama penyakit kanker.

## 2.8. Microarray

Seiring dengan perkembangan ilmu biologi molekuler yang semakin pesat, terdapat suatu teknologi *microarray* yang dapat menyajikan data tingkat ekspresi gen. *Microarray* merupakan suatu *chip* yang berisikan serangkaian sampel berupa DNA, RNA, protein dan jaringan (Pasanen, 2003). *Chip gen microarray* terbuat dari silikon atau kaca dan memiliki struktur seperti grid sebagai tempat bahan genetik. Satu sampel pada *microarray* memiliki ribuan atau bahkan sampai ratusan ribu gen. Sampel ini berupa banyaknya orang atau organisme yang diteliti pada suatu jaringan tertentu atau berupa data seorang atau satu organisme saja yang diambil pada waktu yang berbeda.

Data *microarray* berupa sebuah gambar, kemudian gambar tersebut dianalisis untuk mendapatkan tingkat ekspresi. Data ekspresi gen yang didapat berupa matriks berukuran besar dengan entri matriks yang menunjukkan level atau tingkatan ekspresi gen pada suatu sampel. Bentuk *microarray* dapat dilihat pada Gambar 2.4.



Gambar 2.4. Bentuk *Microarray*

Informasi yang diperoleh dari hasil *microarray* dapat digunakan untuk diagnosis penyakit, penemuan obat-obatan, pengelompokan ekspresi gen yang terlibat dalam organogenesis, cekaman dan interaksi dengan mikroorganisme.

## 2.9. Kanker

Kanker adalah penyakit genetik yang disebabkan oleh perubahan tertentu pada gen yang mengendalikan fungsi sel, terutama cara sel untuk tumbuh dan membelah. Gen membawa perintah untuk membuat protein, yang melakukan banyak pekerjaan di sel. Perubahan gen tertentu dapat menyebabkan sel menghindari kontrol pertumbuhan normal dan menjadi kanker. Sebagai contoh, beberapa perubahan gen yang menyebabkan kanker meningkatkan produksi protein yang membuat sel tumbuh, sedangkan perubahan gen yang lain menghasilkan produksi cacat. Karena tidak berfungsi, bentuk protein biasanya memperbaiki kerusakan sel tersebut.

Menurut WHO (2019), kanker adalah istilah umum untuk satu kelompok besar penyakit yang ditandai dengan pertumbuhan sel-sel

baru secara abnormal yang tumbuh melampaui batas normal yang kemudian dapat menyerang bagian sebelah tubuh atau menyebar ke organ lain. Istilah umum lainnya yang digunakan adalah tumor ganas dan neoplasma. Kanker dapat mempengaruhi hampir seluruh bagian tubuh dan memiliki banyak subtype anatomi dan molekuler yang masing-masing memerlukan strategi pengelolaan yang spesifik.

Menurut *National Cancer Institute* (2019), kanker adalah suatu istilah untuk penyakit dimana sel-sel membelah secara abnormal tanpa kontrol dan dapat menyerang jaringan di sekitarnya. Sel-sel kanker juga dapat menyebar ke bagian tubuh yang lain melalui sistem darah dan getah bening. Kanker merupakan penyebab kematian nomor dua di dunia yakni terhitung sebanyak 9,6 juta kematian terjadi pada tahun 2018. Kanker yang paling umum menyerang laki-laki adalah kanker paru-paru, prostat, usus, lambung, dan hati, sementara kanker yang paling umum menyerang perempuan adalah kanker payudara, usus, paru-paru, leher rahim, dan perut (WHO, 2019).

Beberapa jenis kanker menurut *National Cancer Institute* (NCI), yaitu:

#### 1. BRCA (*Breast Invasive Carcinoma*)

Kanker payudara merupakan kanker yang berkembang pada jaringan sekitar payudara yang menyerang sel-sel tertentu dengan frekuensi kejadian yang tinggi. Kanker payudara menjadi penyebab kematian kedua akibat kanker pada wanita di dunia (Jemal dkk., 2002; Chan dan Moris, 2006). Gejala paling umum yang dialami penderita kanker payudara adalah benjolan yang terasa jelas berbeda dari jaringan payudara di sekitarnya. Penyebab kanker payudara belum diketahui secara pasti, namun diduga terjadi dikarenakan kerusakan pada gen pengatur pertumbuhan dan diferensiasi sel.

BRCA (*Breast Invasive Carcinoma*) adalah jenis dari kanker payudara invasive, lebih tepatnya, yaitu *ductal carcinoma* dan *lobular carcinoma*. *Invasive ductal carcinoma* adalah jenis kanker payudara yang paing umum dan berkembang dari saluran payudara. Sedangkan *invasive lobular breast carcinoma* merupakan kanker yang berkembang di lobulus atau kelenjar penghasil ASI.

#### 2. COAD (*Colon Adenocarcinoma*)

Kanker usus besar merupakan kanker yang terbentuk di jaringan usus besar (bagian terpanjang dari usus besar). COAD adalah bagian

dari kanker usus besar yang dimulai dari dalam sel yang membuat dan melepaskan lendir dan cairan lain.

### 3. KIRC (*Kidney Renal Clear Cell Carcinoma*)

KIRC (*Kidney Renal Clear Cell Carcinoma*) merupakan jenis kanker ginjal yang paling umum. Kanker ini berkembang dari sel-sel yang melapisi tubulus kecil pada ginjal yang menyaring limbah dari darah dan membentuk urin. Kanker ini lebih sering terjadi pada pria. Sebagian besar orang yang mengidap kanker ginjal biasanya berusia 55 tahun. Apabila kanker sudah terdeteksi sejak dini, maka kanker ginjal dapat diobati secara efektif. Namun, apabila kanker telah menyebar ke bagian lain dari tubuh, maka tingkat kelangsungan hidup penderita rendah.

Merokok, penyalahgunaan obat penghilang rasa sakit tertentu, memiliki tekanan darah tinggi, memiliki riwayat keluarga yang mempunyai kanker ginjal, dapat menjadi resiko terkena kanker ginjal. Tanda-tanda apabila seseorang mengidap kanker ginjal, yaitu terdapat benjolan di bagian perut, darah dalam urin, kehilangan selera makan, penurunan berat badan tanpa alasan yang diketahui, anemia. Tanda-tanda tersebut bisa saja muncul pada tahap awal dan bisa juga muncul pada saat tumor telah tumbuh.

### 4. LUAD (*Lung Adenocarcinoma*)

Terdapat dua jenis utama kanker paru-paru, yaitu *Non-Small Cell Lung Cancer* (NSCLC) dan *Small Cell Lung Cancer* (SCLC). Sekitar 80% hingga 85% kanker paru-paru adalah NSCLC. LUAD (*Lung Adenocarcinoma*) adalah salah satu subtype utama dari NSCLC yang merupakan jenis kanker paru-paru yang paling umum. Kanker tersebut bermula pada sel yang melapisi alveoli dan membuat zat seperti lendir.

Penyebab utama kanker paru-paru adalah merokok. Semakin dini, semakin sering dan semakin lama seseorang merokok, maka akan semakin besar resiko terkena kanker paru-paru. Tidak hanya perokok aktif, namun kanker paru-paru juga dapat menyerang perokok pasif, yaitu seseorang yang terkena asap rokok. Faktor lain seseorang terkena kanker paru-paru adalah polusi udara, paparan radon di rumah atau tempat kerja, memiliki riwayat keluarga terkena kanker paru-paru (*genetic*), terinfeksi HIV dan lingkungan.

Beberapa gejala apabila seseorang terkena kanker paru-paru adalah rasa nyeri di dada, batuk yang semakin memburuk hingga keluar darah dalam dahak, kesulitan bernafas, suara serak, kehilangan

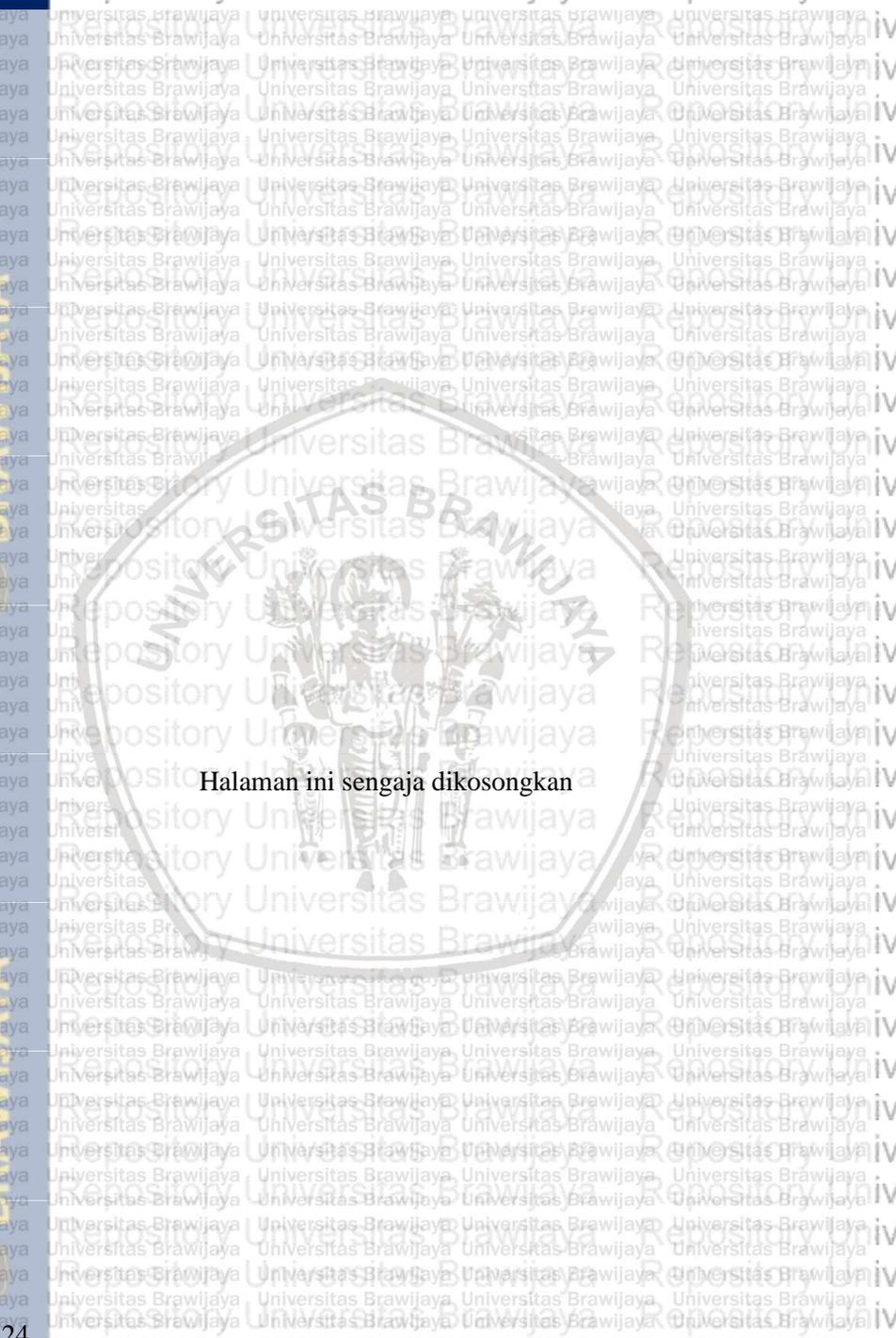
selera makan, penurunan berat badan tanpa alasan yang diketahui, pembengkakan di wajah dan/atau pembuluh darah di leher.

### 5. PRAD (*Prostate Adenocarcinoma*)

Kanker prostat adalah kanker paling umum dan merupakan penyebab utama kedua kematian akibat kanker pada kalangan pria di Amerika Serikat. Kanker prostat adalah sebuah kelenjar yang terdapat dalam sistem reproduksi pria. Adenokarsinoma adalah jenis kanker yang paling umum terjadi di prostat karena sebagian besar sel dalam kelenjar prostat adalah jenis kelenjar.

Kanker prostat biasanya tumbuh sangat lambat. Kanker prostat dapat didiagnosis menggunakan skor *Gleason*. Diagnosis tersebut didasarkan pada sel-sel yang terlihat pada mikroskop dan memiliki kisaran antara dua hingga sepuluh. Apabila skor *Gleason* tersebut rendah, maka jaringan kanker memiliki kemiripan dengan sel-sel normal dan tidak memiliki kemungkinan untuk menyebar. Sedangkan apabila skor *Gleason* tinggi, maka sel-sel kanker tersebut sangat berbeda dari sel-sel normal dan memiliki kecenderungan untuk menyebar. Pasien yang kankernya telah menyebar biasanya memiliki waktu bertahan hidup satu hingga tiga tahun.





Halaman ini sengaja dikosongkan

### BAB III METODE PENELITIAN

#### 3.1. Data

Penelitian ini menggunakan data sekunder yang didapatkan dari UCI *Machine Learning Repository* pada [link https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RN](https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RN) *A-Seq*. Data yang digunakan merupakan data pada tahun 2016 berupa panjang untai RNA pada tingkat ekspresi gen pasien penderita kanker yang dilambangkan dengan *dummy* (gen\_XX). Terdapat lima jenis kanker, yaitu BRCA (*Breast Invasive Carcinoma*), COAD (*Colon Adenocarcinoma*), KIRC (*Kidney Renal Clear Cell Carcinoma*), LUAD (*Lung Adenocarcinoma*) dan PRAD (*Prostate Adenocarcinoma*). Data yang digunakan pada penelitian ini dapat dilihat pada Lampiran 1.

Peubah prediktor dalam penelitian ini adalah tingkat ekspresi gen pasien sebanyak 685 gen dengan sampel sebanyak 60 amatan, sedangkan peubah responnya adalah klasifikasi jenis kanker. Peubah yang digunakan pada penelitian ini disajikan pada Tabel 3.1.

Tabel 3.1. Peubah untuk Klasifikasi Jenis Kanker

Peubah	Keterangan
Y	BRCA ( <i>Breast Invasive Carcinoma</i> )
	COAD ( <i>Colon Adenocarcinoma</i> )
	KIRC ( <i>Kidney Renal Clear Cell Carcinoma</i> )
	LUAD ( <i>Lung Adenocarcinoma</i> )
	PRAD ( <i>Prostate Adenocarcinoma</i> )
$X_1$	Panjang untai RNA pada tingkat ekspresi gen pasien penderita kanker pada ID gen_1 (kbps)
$X_2$	Panjang untai RNA pada tingkat ekspresi gen pasien penderita kanker pada ID gen_2 (kbps)
$X_3$	Panjang untai RNA pada tingkat ekspresi gen pasien penderita kanker pada ID gen_3 (kbps)
$X_4$	Panjang untai RNA pada tingkat ekspresi gen pasien penderita kanker pada ID gen_4 (kbps)
$X_5$	Panjang untai RNA pada tingkat ekspresi gen pasien penderita kanker pada ID gen_5 (kbps)
$\vdots$	$\vdots$
$X_{685}$	Panjang untai RNA pada tingkat ekspresi gen pasien penderita kanker pada ID gen_685 (kbps)

Jenis kanker dilambangkan dengan angka seperti pada Tabel 3.2.

Tabel 3.2. Kategori Jenis Kanker

Jenis Kanker	Kategori
BRCA ( <i>Breast Invasive Carcinoma</i> )	1
COAD ( <i>Colon Adenocarcinoma</i> )	2
KIRC ( <i>Kidney Renal Clear Cell Carcinoma</i> )	3
LUAD ( <i>Lung Adenocarcinoma</i> )	4
PRAD ( <i>Prostate Adenocarcinoma</i> )	5

### 3.2. Struktur Data

Struktur data dalam penelitian ini ditunjukkan pada Tabel 3.3.

Tabel 3.3. Struktur Data Penelitian

Peubah Respon ( $Y_j$ )	Pengamatan ( $n$ )	Peubah Prediktor (X)				
		$X_1$	$X_2$	$X_3$	...	$X_{685}$
1	1	$X_{1,1,1}$	$X_{2,1,1}$	$X_{3,1,1}$	...	$X_{685,1,1}$
	2	$X_{1,1,2}$	$X_{2,1,2}$	$X_{3,1,2}$	...	$X_{685,1,2}$
	⋮	⋮	⋮	⋮	⋮	⋮
	60	$X_{1,1,60}$	$X_{2,1,60}$	$X_{3,1,60}$	...	$X_{685,1,60}$
⋮	⋮	⋮	⋮	⋮	⋮	
5	1	$X_{1,5,1}$	$X_{2,5,1}$	$X_{3,5,1}$	...	$X_{685,5,1}$
	2	$X_{1,5,2}$	$X_{2,5,2}$	$X_{3,5,2}$	...	$X_{685,5,2}$
	⋮	⋮	⋮	⋮	⋮	⋮
	60	$X_{1,5,60}$	$X_{2,5,60}$	$X_{3,5,60}$	...	$X_{685,5,60}$

Keterangan:

$X$  : Panjang untaian RNA pada tingkat ekspresi gen pasien penderita kanker pada ID gen\_1 sampai ID gen\_685

$Y$  : Jenis kanker dengan kategori 1 untuk BRCA, 2 untuk COAD, 3 untuk KIRC, 4 untuk LUAD dan 5 untuk PRAD

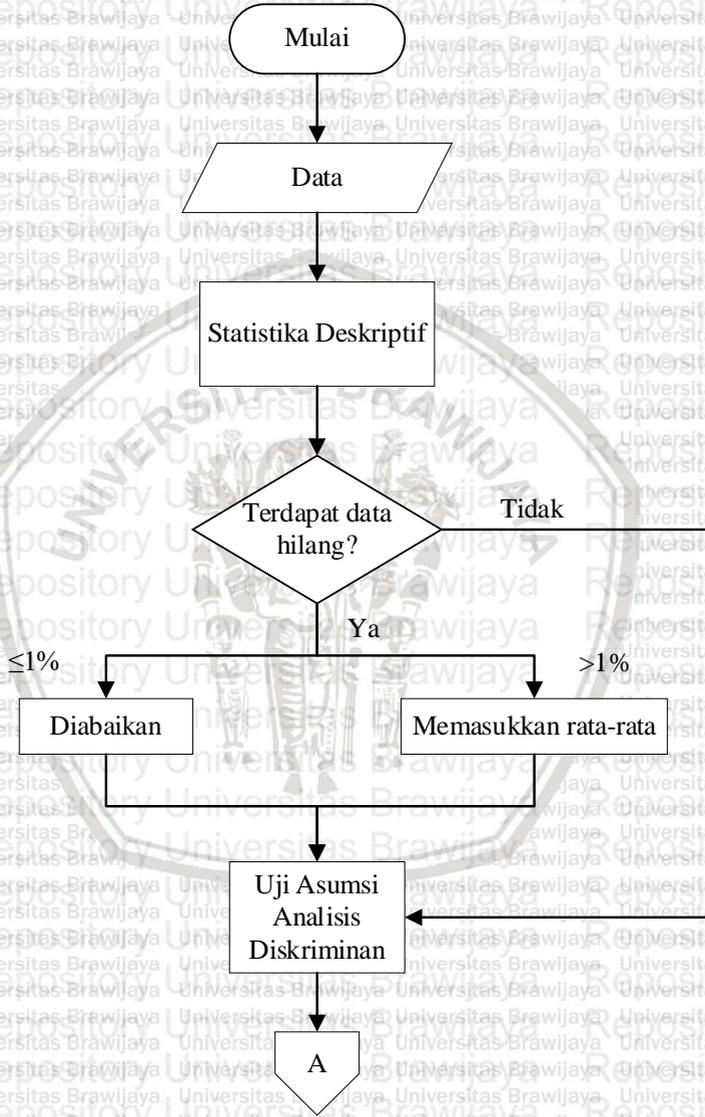
### 3.3. Metode Analisis Data

Metode analisis data yang digunakan dalam penelitian ini adalah analisis diskriminan dengan metode *Vertex Discriminant Analysis* (VDA) menggunakan bantuan dua *software*, yaitu Minitab dan R dengan *packages* VDA. Langkah-langkah analisis data yang dilakukan, yaitu:

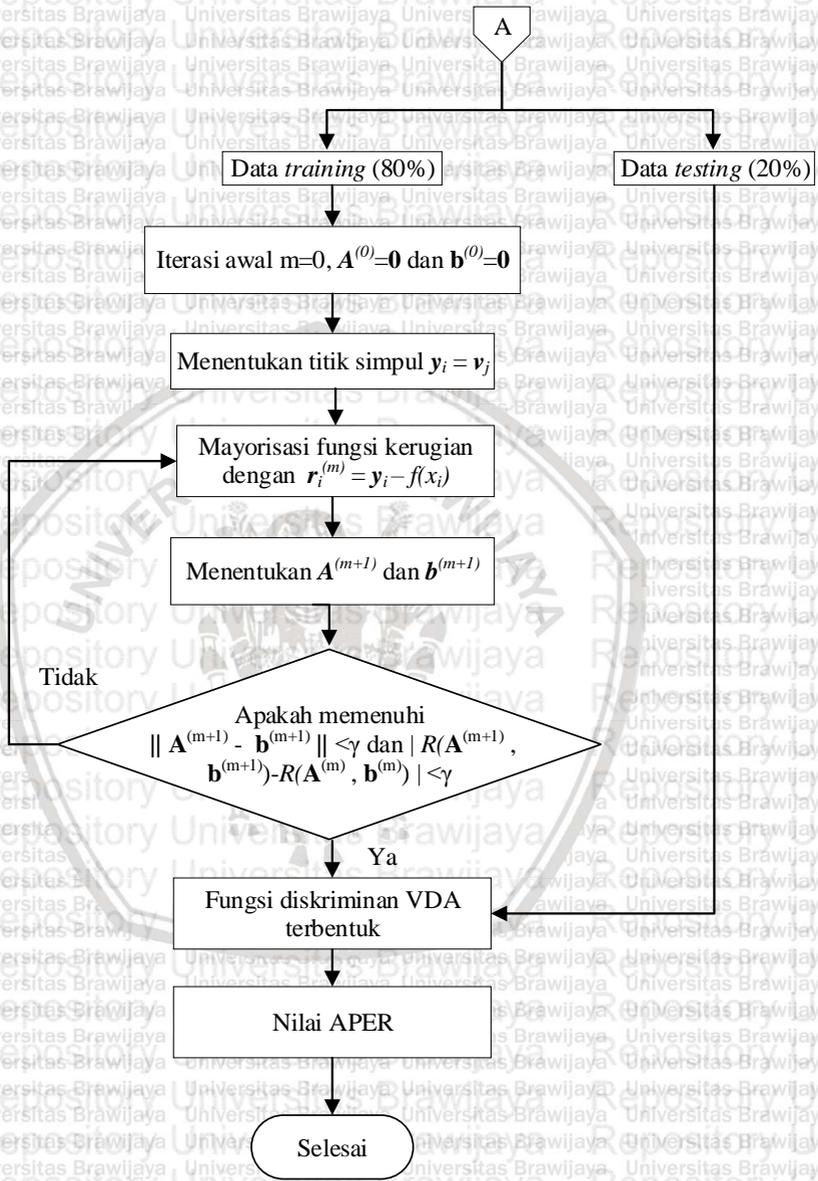
1. Analisis statistika deskriptif.

2. Mengidentifikasi adanya data hilang (*missing value*) sesuai dengan subbab 2.3.
3. Melakukan uji asumsi analisis diskriminan.
4. Membagi data menjadi data *training* dan data *testing* dengan persentase masing-masing 80% dan 20%. Data *training* digunakan untuk membentuk fungsi diskriminan. Fungsi diskriminan yang terbentuk dapat digunakan untuk memprediksi kelas jika data pada setiap peubah dimasukkan, sehingga perbandingan antara kelas prediksi dan kelas sebenarnya menghasilkan ketepatan klasifikasi. Sedangkan data *testing* digunakan untuk menguji ketepatan klasifikasi.
5. Membentuk fungsi diskriminan dengan menggunakan metode *Vertex Discriminant Analysis* (VDA). Menurut Lange dan Wu (2008), tahapan pembentukan fungsi VDA adalah sebagai berikut:
  - a. Menentukan nilai iterasi awal  $m = 0$  dan inialisasi  $\mathbf{A}^{(0)} = \mathbf{0}$  dan  $\mathbf{b}^{(0)} = \mathbf{0}$ .
  - b. Menentukan nilai titik simpul dari masing-masing kelompok dengan persamaan (2.7) dan mendefinisikan  $\mathbf{y}_i = \mathbf{v}_j$ .
  - c. Mayorisasi fungsi kerugian yang diregularisasi pada persamaan (2.11) dengan residual ke- $i$   $\mathbf{r}_i^{(m)} = \mathbf{y}_i - \mathbf{A}^{(m)}\mathbf{x}_i - \mathbf{b}^{(m)}$ .
  - d. Meminimumkan fungsi pengganti dengan menentukan  $\mathbf{A}^{(m+1)}$  dan  $\mathbf{b}^{(m+1)}$  dengan cara menyelesaikan  $k$  pasang persamaan linier.
  - e. Jika  $\|\mathbf{A}^{(m+1)} - \mathbf{A}^{(m)}\| < \gamma$  dan  $|R(\mathbf{A}^{(m+1)}, \mathbf{b}^{(m+1)}) - R(\mathbf{A}^{(m)}, \mathbf{b}^{(m)})| < \gamma$  dengan  $\gamma = 10^{-4}$  maka berhenti. Jika tidak, maka ulangi langkah c sampai e.
  - f. Mengklasifikasikan objek ke dalam sejumlah kelas dengan memilih jarak terdekat antara penduga dengan indikator kelas yang mungkin (titik verteks ke- $j$ ) seperti pada persamaan (2.21).
6. Menghitung kesalahan klasifikasi berdasarkan nilai APER menggunakan persamaan (2.22).

Gambar 3.1. Langkah-langkah penelitian secara lengkap disajikan dalam



Gambar 3.1. Diagram Alir Penelitian



Gambar 3.1. Diagram Alir Penelitian (lanjutan)

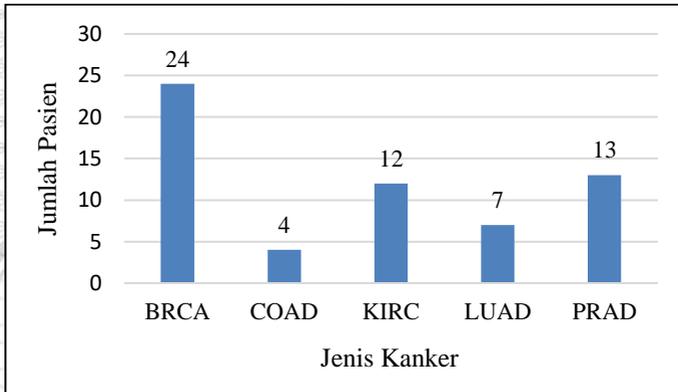


Halaman ini sengaja dikosongkan

## BAB IV HASIL DAN PEMBAHASAN

### 4.1. Statistika Deskriptif

Statistika deskriptif digunakan untuk mengetahui gambaran data. Diagram data gen pasien yang menderita kanker disajikan pada Gambar 4.1.



Gambar 4.1. Statistika Deskriptif

Gambar 4.1. menunjukkan bahwa dari 60 pasien penderita kanker, BRCA merupakan jenis kanker dengan persentase tertinggi, yaitu sebesar 40% (24 pasien). Sedangkan COAD merupakan jenis kanker dengan persentase terendah, yaitu sebesar 7% (4 pasien). KIRC, LUAD dan PRAD masing-masing memiliki persentase 20% (12 pasien), 11% (7 pasien) dan 22% (13 pasien).

### 4.2. Identifikasi Data Hilang (*Missing Value*)

Data gen pasien penderita kanker merupakan data berdimensi tinggi, yaitu data dengan jumlah yang cukup banyak. Oleh karena itu, perlu dilakukan identifikasi adanya data hilang. Banyaknya data hilang pada suatu data set pada dasarnya tidak akan bermasalah apabila hanya terdapat sekitar 1% dari keseluruhan data, sehingga data hilang tersebut dapat diabaikan. Namun, apabila persentase data hilang cukup besar maka perlu dilakukan penanganan.

Identifikasi data hilang dapat dilakukan dengan analisis statistika deskriptif dengan bantuan *software* Minitab. Berdasarkan analisis statistika deskriptif, terdapat 472 data hilang dari 41.100 data gen

pasien penderita kanker. Hal ini berarti bahwa hanya terdapat 0,01% data hilang pada data gen pasien penderita kanker. Persentase data hilang tersebut cukup kecil sehingga tidak perlu dilakukan penanganan pada data hilang. Banyaknya data hilang pada masing-masing peubah prediktor dapat dilihat pada Lampiran 2.

### **4.3. Asumsi Analisis Diskriminan**

#### **4.3.1. Asumsi Normal Multivariat**

Data berdimensi tinggi, yaitu ketika banyaknya peubah prediktor jauh lebih besar dari banyak observasi akan menyebabkan matriks *varians kovarians* singular, sehingga tidak memiliki matriks kebalikan (*inverse matrix*). Pengujian asumsi distribusi normal multivariat tidak dapat dilakukan, plot yang dihasilkan akan membentuk garis lurus. VDA mampu mengklasifikasikan objek pada saat asumsi normal multivariat tidak terpenuhi.

#### **4.3.2. Asumsi Homogenitas Matriks *Varians Kovarians***

Berdasarkan hasil uji Box's M diperoleh nilai-p mendekati 0,000. Dapat disimpulkan bahwa matriks *varians kovarians* antar kelompok berbeda. Hal tersebut dapat ditangani dengan menggunakan analisis diskriminan kuadratik. Menurut Duda dkk. Tao Li dkk. (2006) analisis diskriminan seringkali menghasilkan hasil klasifikasi yang lebih baik meskipun asumsi normalitas multivariat dan homogenitas matriks ragam peragam terlanggar. Selain itu, VDA mampu mengklasifikasikan objek pada saat asumsi homogenitas matriks *varians kovarians* tidak terpenuhi. Pada penelitian ini pelanggaran pada asumsi yang terlanggar tidak ditangani.

#### **4.3.3. Identifikasi Adanya Hubungan Antar Peubah Prediktor untuk Asumsi Multikolinieritas**

Peubah prediktor sebanyak 685 dilakukan uji korelasi antara dua peubah untuk mengetahui hubungan antar peubah prediktor. Analisis korelasi yang digunakan adalah uji korelasi Pearson. Terdapat 234.270 pengujian korelasi yang dilakukan. Sebanyak 196.400 pengujian menghasilkan keputusan terima  $H_0$ , artinya sebanyak 196.400 pengujian tidak terdapat hubungan yang signifikan antara dua peubah prediktor. Sisanya, sebanyak 37.870 pengujian menghasilkan keputusan tolak  $H_0$ , artinya sebanyak 37.870 pengujian terdapat hubungan yang signifikan antara dua peubah prediktor.

Pengujian korelasi yang dilakukan memberikan kesimpulan bahwa jumlah pengujian yang menghasilkan hubungan antar dua peubah prediktor lebih sedikit daripada tidak adanya hubungan antara dua peubah prediktor.

#### 4.4. Vertex Pada Ruang $\mathbb{R}^{k-1}$

Berdasarkan perhitungan menggunakan persamaan (2.7), *vertex* pada ruang  $\mathbb{R}^{k-1}$  untuk 5 kelompok dapat dilihat pada Tabel 4.1.

Tabel 4.1. *Vertex* Pada Ruang  $\mathbb{R}^{k-1}$  untuk 5 Kelompok

$v_1$	$v_2$	$v_3$	$v_4$	$v_5$
0,5	0,7135	-0,4045	-0,4045	-0,4045
0,5	-0,4045	0,7135	-0,4045	-0,4045
0,5	-0,4045	-0,4045	0,7135	-0,4045
0,5	-0,4045	-0,4045	-0,4045	0,7135

#### 4.5. Pemodelan VDA

Setelah menentukan *vertex* pada ruang  $\mathbb{R}^{k-1}$ , langkah selanjutnya adalah mengacak data kemudian membagi data pasien penderita kanker menjadi dua, yaitu data *training* sebanyak 48 pasien dan data *testing* sebanyak 12 pasien. Data *training* digunakan untuk membentuk fungsi dikriminan sedangkan data *testing* untuk mengevaluasi model diskriminan. Acak data yang dilakukan menggunakan metode *simple random sampling*. Setelah dihitung banyaknya sampel yang dibutuhkan ( $n=60$ ), kemudian membangkitkan angka *random* antara 1-N sebanyak  $n$ . Misal, angka *random* yang keluar pertama adalah 23, maka sampel pertama yang digunakan untuk data *testing* adalah data ke 23. Dalam melakukan analisis diskriminan menggunakan metode VDA dibutuhkan bantuan *software R* dengan *packages VDA*.

Metode VDA membentuk 4 fungsi diskriminan untuk membedakan 5 jenis kanker berdasarkan banyaknya amatan 48 pasien penderita kanker. Lambda yang digunakan sebesar 0,05. Dugaan koefisien yang dibentuk oleh VDA untuk membentuk 4 fungsi diskriminan disajikan pada Lampiran 5.

Berdasarkan dugaan koefisien pada Lampiran 5, diperoleh fungsi diskriminan pertama ( $D_1$ ), fungsi diskriminan kedua ( $D_2$ ), fungsi diskriminan ketiga ( $D_3$ ) dan fungsi diskriminan keempat ( $D_4$ ) yang ditunjukkan pada persamaan (4.1) sampai dengan persamaan (4.4).



$$D_1 = 0,0113 - 7,0914(10^{-4})X_1 - \dots - 0,0012X_{685} \quad (4.1)$$

$$D_2 = 0,0670 - 1,2326(10^{-3})X_1 - \dots - 0,0015X_{685} \quad (4.2)$$

$$D_3 = 0,0119 - 6,6328(10^{-4})X_1 - \dots + 0,0006X_{685} \quad (4.3)$$

$$D_4 = 0,0887 - 6,5951(10^{-5})X_1 + \dots + 0,0011X_{685} \quad (4.4)$$

Model yang diperoleh merupakan model penuh. VDA dengan penambahan penalti digunakan pada saat seleksi peubah dengan cara memperkecil koefisien parameter yang berkorelasi mendekati nilai nol untuk peubah-peubah yang berkorelasi, sehingga tidak dilakukan reduksi parameter.

Koefisien pada fungsi diskriminan yang terbentuk menunjukkan tingkat ekspresi gen pasien penderita kanker. Semakin besar nilai koefisien, maka gen semakin ekspresif. Begitu pula sebaliknya, semakin kecil nilai koefisien, maka gen semakin kurang ekspresif. Pada fungsi diskriminan  $D_1$ ,  $X_1$ , yaitu ID gen 1 mengekspresikan gen pasien penderita kanker sebesar -7,0915, artinya ID gen 1 tersebut kurang ekspresif dalam mengekspresikan gen pasien penderita kanker.

Fungsi diskriminan yang telah terbentuk selanjutnya dapat digunakan untuk membandingkan antara kelompok awal dengan kategori kelompok prediksi. Kategori kelompok prediksi didapatkan dari nilai jarak terkecil antara objek dengan *vertex*.

Nilai  $D$  dalam fungsi diskriminan yang terbentuk merupakan koordinat dalam *Euclidean* yang akan menentukan klasifikasi pengamatan ke- $i$ . Pada penelitian ini, VDA menggunakan lima *vertex* sebagai indikator titik pusat kelompok. Kelompok 1 (0,5; 0,5; 0,5; 0,5) hingga kelompok 5 (-0,4045; -0,4045; -0,4045; 0,7135) seperti yang telah dijelaskan pada Tabel 4.1.

Ketepatan klasifikasi fungsi diskriminan pada data *training* menggunakan metode VDA pada setiap kelompok dapat dilihat pada Tabel 4.2.

Tabel 4.2. Ketepatan Model VDA pada Data *Training*

	Klasifikasi Model					Total	Benar
	Kelompok	1	2	3	4		
Klasifikasi Sebenarnya	1	18	0	0	0	0	18
	2	0	4	0	0	0	4
	3	0	0	10	0	0	10
	4	0	0	0	4	0	4

Tabel 4.2. Ketepatan Model VDA pada Data *Training* (lanjutan)

	Klasifikasi Model					Total	Benar
	Kelompok	1	2	3	4		
Klasifikasi Sebenarnya	5	0	0	0	0	12	12
Total		18	4	10	4	12	48
Misklasifikasi						0	

Tabel 4.2 menunjukkan bahwa keempat fungsi diskriminan mengklasifikasikan secara tepat sebanyak 48 dari 48 data *training* dengan tidak terdapat misklasifikasi. Klasifikasi model pada kelompok 1 (BRCA), kelompok 2 (COAD), kelompok 3 (KIRC), kelompok 4 (LUAD) dan kelompok 5 (PRAD) semua sesuai dengan klasifikasi sebenarnya. Berdasarkan Lampiran 4 terlihat bahwa kesalahan klasifikasi (nilai APER) sebesar 0. Hal ini berarti bahwa metode VDA memiliki kemampuan dalam mengklasifikasikan amatan dengan tepat pada data *training* sebesar 100%.

Setelah fungsi diskriminan terbentuk, langkah selanjutnya adalah melakukan evaluasi pada 12 data *testing*. Misalkan klasifikasi dilakukan pada amatan pertama data *testing* (pasien ke-23) dengan klasifikasi awal adalah kelompok 4 (LUAD).

Nilai masing-masing peubah pasien ke-23 disubstitusi ke dalam empat fungsi diskriminan yang telah terbentuk, sehingga diperoleh  $\hat{D}_1 = -0,5545$ ,  $\hat{D}_2 = -0,4330$ ,  $\hat{D}_3 = 0,0015$ ,  $\hat{D}_4 = -0,3898$ . Nilai *vertex* untuk 5 kelompok, jarak antara objek dengan *vertex* dan dugaan klasifikasi pasien ke-23 disajikan pada Tabel 4.3.

Tabel 4.3. Nilai *vertex*, jarak antara objek dan titik *vertex* dan dugaan klasifikasi

Kelompok ( $Y_i$ )	Nilai <i>Vertex</i> ( $v_j$ )	$\ v_j - \hat{A}x_i - \hat{b}\ $ $= \ v_j - \hat{D}_k\ $
1	$v_1$	0,5
		0,5
		0,5
		0,5
2	$v_2$	0,7135
		-0,4045
		-0,4045
		-0,4045

Tabel 4.3. Nilai *vertex*, jarak antara objek dan titik *vertex* dan dugaan klasifikasi (lanjutan)

Kelompok ( $Y_i$ )	Nilai <i>Vertex</i> ( $v_j$ )	$\ v_j - \hat{A}x_i - \hat{b}\ $ $= \ v_j - \hat{D}_k\ $
3	$v_3$	-0,4045
		0,7135
		-0,4045
		-0,4045
4	$v_4$	-0,4045
		-0,4045
		0,7135
		-0,4045
5	$v_5$	-0,4045
		-0,4045
		-0,4045
		0,7135

Berdasarkan nilai jarak terkecil antara objek dan titik simpul pada Tabel 4.3 dari kelima kelompok, data *testing* pasien ke-23 diklasifikasikan ke dalam kelompok 4. Hal ini sesuai dengan klasifikasi sebenarnya, artinya hasil klasifikasi metode VDA ini mengklasifikasikan dengan tepat data *testing* pasien ke-23. Hasil analisis klasifikasi untuk 12 data *testing* secara lengkap dapat dilihat pada Lampiran 6.

Hasil klasifikasi fungsi diskriminan untuk semua data *testing* ditunjukkan pada Tabel 4.4.

Tabel 4.4. Hasil Klasifikasi Model VDA pada 12 Data *Testing*

	Klasifikasi Model						Total	Benar
		1	2	3	4	5		
Klasifikasi Sebenarnya	1	0	0	0	5	1	6	0
	2	0	0	0	0	0	0	0
	3	0	0	2	0	0	2	2
	4	0	0	0	3	0	3	3
	5	0	0	0	0	1	1	1
Total		6	0	2	3	1	12	6
Misklasifikasi							6	

Berdasarkan Tabel 4.4 dapat dilihat bahwa keempat fungsi diskriminan mengklasifikasikan secara tepat sebanyak 6 dari 12 data *testing*. Fungsi diskriminan yang terbentuk mampu memprediksi data baru sebesar 50% dengan nilai APER sebesar 0,500.

#### 4.6. Identifikasi Ketepatan Model Klasifikasi

Ketepatan klasifikasi pada data *training* dan data *testing* secara keseluruhan disajikan pada Tabel 4.5.

Tabel 4.5. Ketepatan Klasifikasi Data *Training* dan Data *Testing*

Ketepatan Klasifikasi Data <i>Training</i>	100%
Ketepatan Klasifikasi Data <i>Testing</i>	50%

Berdasarkan Tabel 4.5 dapat dilihat bahwa metode VDA memiliki kemampuan dalam mengklasifikasikan amatan dengan tepat pada data *training* sebesar 100%, sedangkan untuk data *testing*, fungsi diskriminan yang terbentuk mampu memprediksi data baru sebesar 50%.





Halaman ini sengaja dikosongkan

## BAB V PENUTUP

### 5.1. Kesimpulan

Kesimpulan dari penelitian ini adalah:

- 1) Fungsi diskriminan yang terbentuk untuk mengklasifikasikan pasien ke dalam lima kategori jenis kanker (BRCA, COAD, KIRC, LUAD dan PRAD) dijelaskan pada persamaan sebagai berikut:

$$Y_1 = 0,0113 - 7,0914(10^{-4})X_1 - \dots - 0,0012X_{685}$$

$$Y_2 = 0,0670 - 1,2326(10^{-3})X_1 - \dots - 0,0015X_{685}$$

$$Y_3 = 0,0119 - 6,6328(10^{-4})X_1 - \dots + 0,0006X_{685}$$

$$Y_4 = 0,0887 - 6,5951(10^{-5})X_1 + \dots + 0,0011X_{685}$$

Koefisien pada fungsi diskriminan yang terbentuk menunjukkan tingkat ekspresi gen pasien penderita kanker. Semakin besar nilai koefisien, maka gen semakin ekspresif. Fungsi diskriminan yang telah terbentuk selanjutnya dapat digunakan untuk membandingkan antara Y awal dengan kategori Y prediksi. Kategori Y prediksi didapatkan dari nilai jarak terkecil antara objek dengan *vertex*.

- 2) Fungsi diskriminan yang terbentuk menggunakan metode VDA mampu mengklasifikasikan dengan tepat 48 dari 48 data *training*, sedangkan untuk data *testing* VDA mengklasifikasikan dengan tepat 6 dari 12 data dengan kesalahan klasifikasi (APER) sebesar 50%.

### 5.2. Saran

Saran yang dapat diberikan untuk penelitian selanjutnya adalah dapat dilakukan *preprocessing/cleaning* pada data awal penelitian. Data berdimensi tinggi kemungkinan masih terdapat data yang *incomplete*, *noisy*, dan *inconsistent*, sehingga akan menyebabkan hasil data yang kurang berkualitas yang akan mempengaruhi hasil penelitian.



Halaman ini sengaja dikosongkan

**DAFTAR PUSTAKA**

- Davey, A. dan Savla, J. 2010. *Statistical Power Analysis with Missing Data*. New York: Taylor and Francis Group.
- Golub, G.H. dan Loan, C. F. 1996. *Matrix Computations*. Edisi 3. Baltimore (MD): Johns Hopkins University.
- Hair, J. F., Black, W. C., Babin, B. J. dan Anderson, R. E. 1998. *Multivariate Data Analysis*. Edisi 5. New Jersey: Prentice Hall International Inc.
- Hair, J. F., Black, W. C., Babin, B. J. dan Anderson, R. E. 2010. *Multivariate Data Analysis*. Edisi 7. New Jersey: Prentice Hall International Inc.
- Hastie, T., Tibshirani, R. dan Friedman, J. 2008. *The Elements Of Statistical Learning, Data Mining, Inference, and Prediction*. Edisi 2. Stanford California (US): Springer.
- Hunter, D. R. dan Lange K. 2004. *A Tutorial on MM algorithms. The American statistician*. 58:30-37.
- Izenman, A. J. 2008. *Modern Mulivariate Statistical Techniques*. New York (US): Springer.
- Jemal, A., Thomas, A., Murray, T. dan Thun, M. 2002. Cancer Statistics, *CA Cancer Journal Clinicians*. Volume 52, pp. 23-47.
- Johnson dan Wichern. 2002. *Applied Multivariate Statistical Analysis*. Edisi 3. London (US): Pearson Education.
- Johnson dan Wichern. 2007. *Applied Multivariate Statistical Analysis*. Edisi 6. London (US): Pearson Education.
- Lange, K. 2004. *Optimization*. New York (US): Springer-Verlag.
- Lange, K. dan Wu, T. T., 2008. An MM Algorithm for Multicategory Vertex Discriminant Analysis, *Journal of Computational Graphical Statistics*. Volume 17, No. 3, pp. 527-544.
- Lange, K. dan Wu, T. T., 2010, *Multicategory Vertex Discriminant Analysis for High Dimensional Data. The Annals of Applied Statistics*. Volume 4, No. 4, pp. 1698-1721.
- Liu, Y., Shen, X. dan Doss, H. 2005. Multicategory  $\Psi$ -learning and Support Vector Machine: Computational Tools. *Journal Comput Graph Stat*. Volume 14, pp. 219-236.
- Mattjik, A. A. dan Sumertajaya, I. M. 2011. *Sidik Peubah Ganda*. Bogor (ID): IPB.

National Cancer Institute. 2019. *The Cancer Genome Atlas*. <https://www.cancer.gov/aboutnci/organization/ccg/research/structural-genomics/tcga/studied-cancers>. Diakses pada 20 September 2019.

Nurmaleni. 2015. *Perbandingan Metode Multikategori Vertex Discriminant Analysis dan Analisis Diskriminan Fisher* [tesis]. Bogor (ID): Institut Pertanian Bogor.

Pasanen, Tomi. 2003. *DNA Microarray Data Analysis*. *Curr Top Med Chem*. 2004; 4(13):1357-70.

Stefanska dan MacEwan. 2017. *Epigenetics and Gene Expression in Cancer, Inflammatory and Immune Disease*. New York: Springer Science+Business Media LLC.

Vapnik, V. 1995. *The Nature of Statistical Learning Theory*. New York (US): Springer.

Walpole, R. E. 1993. *Pengantar Statistika*. Edisi Ketiga. Jakarta: PT Gramedia Pustaka Utama.

Zharfani. 2019. *Model Klasifikasi Analisis Diskriminan dengan Metode Partial Least Squares Discriminant Analysis (PLSDA) pada Data Berdimensi Tinggi* [skripsi]. Malang (ID): Universitas Brawijaya.

Lampiran 1. Data Tingkat Ekspresi Gen Pasien Penderita Kanker

No.	$X_1$	$X_2$	...	$X_{685}$	Y
1	2,017209	3,265527	...	8,09189	5
2	0,592732	1,588421	...	5,709045	4
3	3,511759	4,327199	...	9,044247	5
4	3,663618	4,507649	...	7,511942	5
5	2,655741	2,821547	...	6,571285	1
6	3,467853	3,581918	...	9,18152	5
7	1,224966	1,691177	...	4,48542	3
8	2,854853	1,750478	...	8,484702	5
9	3,992125	2,77273	...	8,399398	1
10	3,642494	4,423558	...	8,102443	5
11	3,492071	3,553373	...	5,274243	1
12	2,941181	2,663276	...	4,865409	3
13	3,970348	2,364292	...	7,917575	5
14	1,551048	3,529846	...	7,14582	1
15	1,964842	2,18301	...	5,603448	1
16	2,901379	3,685368	...	7,307647	1
17	3,460913	3,618474	...	9,305647	4
18	3,004519	3,007178	...	5,069165	3
19	1,541465	2,54154	...	7,844982	3
20	4,167583	3,841389	...	7,42177	5
21	2,066916	2,619953	...	5,246913	1
22	3,529783	2,976712	...	7,354047	3
23	1,131853	2,351515	...	9,852687	4
24	3,121844	2,473943	...	4,464198	1

## Lampiran 1. (Lanjutan)

No.	$X_1$	$X_2$	...	$X_{685}$	Y
25	4,801852	2,648465	...	3,430205	3
26	4,317702	3,642678	...	8,669565	4
27	2,325242	3,247092	...	5,967355	2
28	0,657091	1,026304	...	3,41715	1
29	3,670081	3,382792	...	9,469176	1
30	5,688295	4,899949	...	10,97036	1
31	2,662479	2,742869	...	6,046229	1
32	2,090786	3,769116	...	6,702464	1
33	0	2,499578	...	7,209717	3
34	0	1,633152	...	7,087993	1
35	3,19022	5,690191	...	9,48823	5
36	2,464721	3,25408	...	6,08762	1
37	3,384243	2,673624	...	4,392956	3
38	4,416259	4,188978	...	9,096388	4
39	0	2,72275	...	3,567107	1
40	0	2,536525	...	4,593145	1
41	3,729379	2,293871	...	4,236577	3
42	3,026375	4,364103	...	6,834054	5
43	4,618861	5,648986	...	8,662658	5
44	3,757322	3,231048	...	5,474595	3
45	3,47241	3,007465	...	5,312462	3
46	3,092055	2,210327	...	7,419817	1
47	1,149259	2,575385	...	8,292069	5
48	3,44619	3,620962	...	3,392963	2
49	3,768449	2,736172	...	6,246425	1
50	4,577822	3,800237	...	10,45224	4

### Lampiran 1. (Lanjutan)

No.	$X_1$	$X_2$	...	$X_{685}$	Y
51	3,818472	3,215772	...	6,083198	1
52	3,858916	4,324523	...	6,952019	4
53	2,396708	2,399062	...	8,583534	1
54	3,639278	3,402176	...	7,403097	5
55	2,554196	1,296134	...	3,771695	2
56	2,350384	3,249369	...	8,389322	1
57	1,252355	2,40806	...	5,443172	1
58	3,667699	3,083179	...	6,923506	2
59	1,985792	3,640031	...	5,796351	3
60	1,870306	3,768777	...	6,189321	1

#### Keterangan:

Y : Jenis kanker dengan kategori 1 untuk BRCA, 2 untuk COAD, 3 untuk KIRC, 4 untuk LUAD dan 5 untuk PRAD.

$X_1$  : Panjang untaian RNA pada tingkat ekspresi gen pasien penderita kanker pada ID gen\_1 (kbps).

$X_2$  : Panjang untaian RNA pada tingkat ekspresi gen pasien penderita kanker pada ID gen\_2 (kbps).

$X_3$  : Panjang untaian RNA pada tingkat ekspresi gen pasien penderita kanker pada ID gen\_3 (kbps).

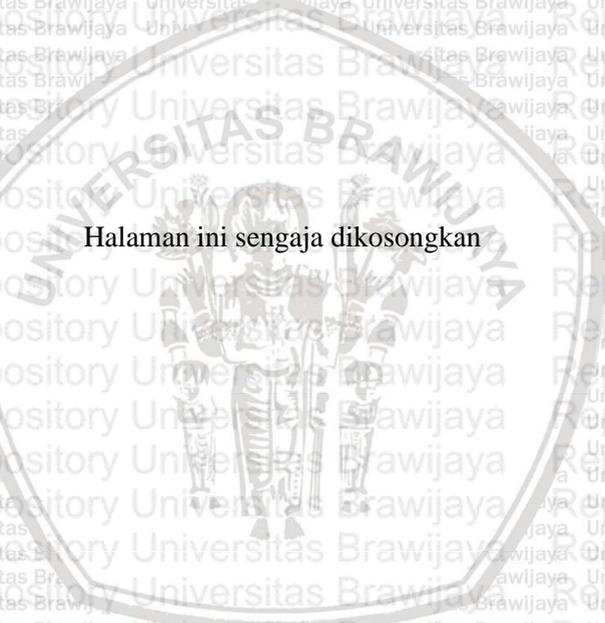
$X_4$  : Panjang untaian RNA pada tingkat ekspresi gen pasien penderita kanker pada ID gen\_4 (kbps).

$X_5$  : Panjang untaian RNA pada tingkat ekspresi gen pasien penderita kanker pada ID gen\_5 (kbps).

$X_{685}$  : Panjang untaian RNA pada tingkat ekspresi gen pasien penderita kanker pada ID gen\_685 (kbps).



Halaman ini sengaja dikosongkan



## Lampiran 2. Banyaknya Data Hilang pada Setiap Peubah Prediktor

<i>Variable</i>	<i>Total Count</i>	<i>N</i>	<i>N Missing</i>
gen_1	60	56	4
gen_2	60	60	0
gen_3	60	60	0
gen_4	60	60	0
gen_5	60	60	0
gen_6	60	60	0
gen_7	60	60	0
gen_8	60	58	2
gen_9	60	60	0
gen_10	60	48	12
gen_11	60	60	0
gen_12	60	60	0
gen_13	60	60	0
gen_14	60	60	0
gen_15	60	60	0
gen_16	60	60	0
gen_17	60	60	0
gen_18	60	60	0
gen_19	60	60	0
gen_20	60	60	0
gen_21	60	60	0
gen_22	60	60	0
gen_23	60	60	0
gen_24	60	60	0
gen_25	60	60	0
gen_26	60	60	0
⋮	⋮	⋮	⋮
gen_685	60	60	0
Total	41.100	40.628	472

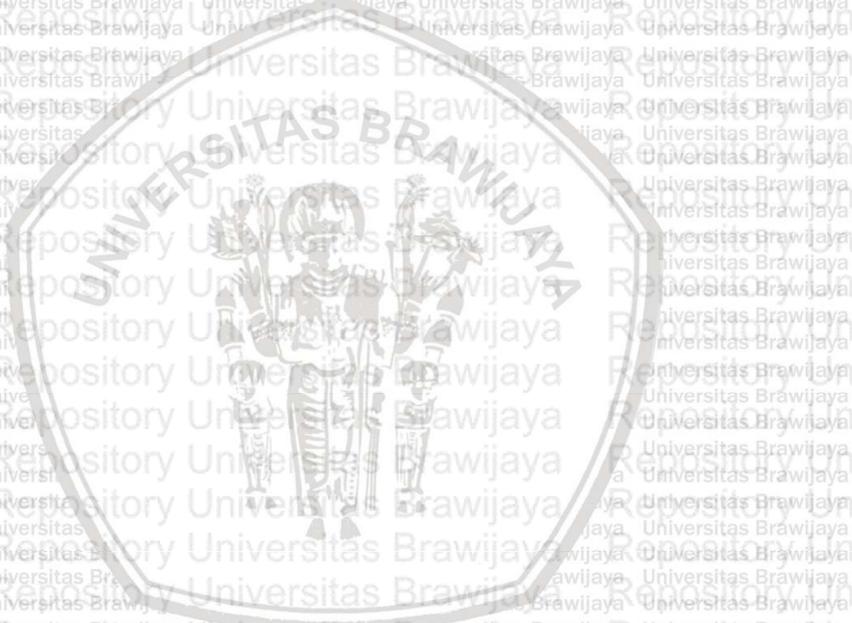


Halaman ini sengaja dikosongkan



Lampiran 3. *Output Uji Asumsi Homogenitas Matriks Varians Kovarians Menggunakan Software R*

Box's M-test for Homogeneity of Covariance Matrices  
data: datagen [, 1:685]  
Chi-Sq (approx.) = NaN, df = 939820, p-value = NA





Halaman ini sengaja dikosongkan

#### Lampiran 4. *Syntax R dan Output VDA untuk data training*

```
> library(VDA)
> library(csv)
> data <- read.csv(file.choose(), header=TRUE)
> data
> x <- data[,1:685]
> y <- data[,686]
> lam.vec <- (1:5/100)
> cv <- cv.vda.r(x, y,5,lam.vec)
> plot(cv)
> out <- vda.r(x,y,cv$lam.opt)
> print(out)
Call:
vda.r.default(x, y, cv$lam.opt)
# of cases= 48
# of classes= 5
# of features= 685
Lambda used: 0.05
Estimated Coefficients:
      [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 0,01126175 -7,091402e-04 -0,0008487133 0,0009963253 0,0012275764
[2,] 0,06703040 -1,232634e-03 -0,0011638405 -0,0011503506 0,0005257080
[3,] 0,001193860 -6,632845e-04 -0,0005102462 -0,0010526064 0,0003039559
[4,] 0,08865187 -6,595054e-05 0,0006191625 -0,0007781621 0,0012309263
      [,6]      [,7]      [,8]      [,9]     [,10]
[1,] 0,0002119148 0,0009701657 -0,0004743466 -0,0007429364 -6,925109e-04
[2,] -0,0008312396 0,0000298621 0,0001791168 -0,0008468143 -2,953169e-04
[3,] 0,0005084174 -0,0005936033 0,0005990754 0,0007938516 7,195301e-04
[4,] 0,0002578227 0,0001473572 -0,0009500985 -0,0004441117 2,424291e-05
      [,11]      [,12]      [,13]      [,14]      [,15]
[1,] -0,0008449580 -1,449396e-04 -9,266203e-04 0,0010247089 -0,0003764927
[2,] -0,0010841658 -4,357295e-04 -4,467436e-05 0,0012167285 -0,0003679754
[3,] 0,0013579079 -1,055775e-03 4,530960e-04 0,0014961523 -0,0014119688
[4,] 0,0009291525 8,408950e-06 1,431935e-03 0,0009318442 -0,0005058074
      :
      [,682]      [,683]      [,684]      [,685]      [,686]
[1,] 7,946132e-04 -0,0012457614 5,235051e-04 -0,0008735150 -0,0012008493
[2,] -8,456153e-05 -0,0004620933 -9,119395e-04 -0,0008552187 -0,0015203673
[3,] 8,240575e-04 0,0008281376 -9,818422e-05 0,0002288349 0,0006370532
[4,] 6,658743e-04 0,0002397233 -1,459708e-03 -0,0002074668 0,0010504072
```

## Lampiran 4. *Syntax R dan Output VDA untuk data training* (lanjutan)

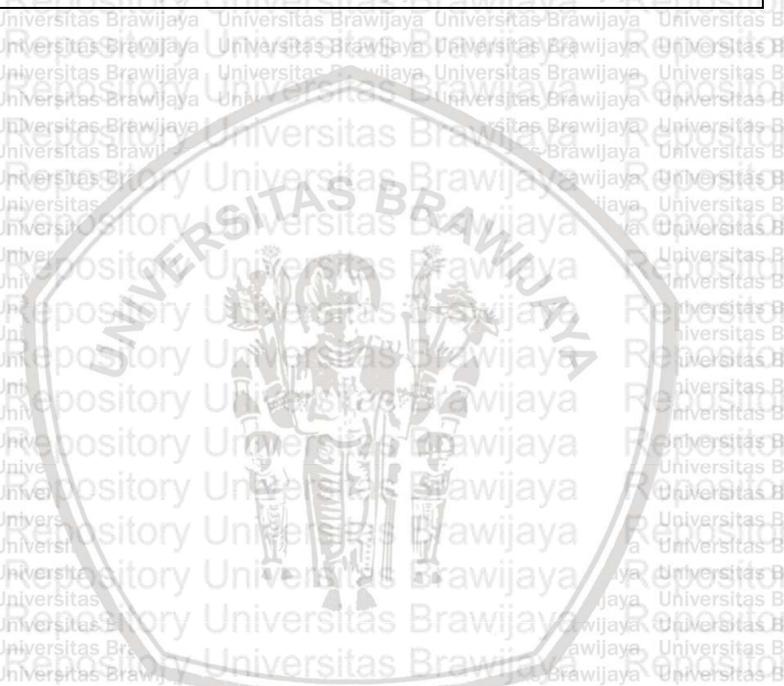
Predicted classification:

```
[1] 3 3 1 2 1 1 3 5 3 2 1 1 4 5 1 1 5 5 1 1 2 3 1 3 3 5 1 4 5 1 3 3 3 5 1 1 5 5
```

```
[39] 4 5 1 4 1 2 5 1 1 5
```

Training Error:

```
[1] 0
```



## Lampiran 5. Dugaan koefisien fungsi diskriminan pada metode VDA

Peubah	Koefisien fungsi diskriminan 1	Koefisien fungsi diskriminan 2	Koefisien fungsi diskriminan 3	Koefisien fungsi diskriminan 4
Konstanta	0,01126	0,06703	0,01194	0,08865
$X_1$	-0,00071	-0,00123	-0,00066	-0,00007
$X_2$	-0,00085	-0,00116	-0,00051	0,00062
$X_3$	0,00100	-0,00115	-0,00105	-0,00078
$X_4$	0,00123	0,00053	0,00030	0,00123
$X_5$	0,00021	-0,00083	0,00051	0,00026
$X_6$	0,00097	0,00003	-0,00059	0,00015
$X_7$	-0,00047	0,00018	0,00060	-0,00095
$X_8$	-0,00074	-0,00085	0,00079	-0,00044
$X_9$	-0,00069	-0,00030	0,00072	0,00002
$X_{10}$	-0,00084	-0,00108	0,00136	0,00093
$X_{11}$	-0,00014	-0,00044	-0,00106	0,00001
$X_{12}$	-0,00093	-0,00004	0,00045	0,00143
$X_{13}$	0,00102	0,00122	0,00150	0,00093
$X_{14}$	-0,00038	-0,00037	-0,00141	-0,00051
$X_{15}$	-0,00123	0,00023	0,00048	-0,00069
$X_{16}$	-0,00082	0,00110	-0,00046	0,00011
$X_{17}$	0,00050	-0,00084	-0,00027	0,00092
$X_{18}$	0,00028	-0,00101	0,00032	0,00103
$X_{19}$	-0,00107	-0,00212	-0,00109	0,00010
$X_{20}$	0,00100	0,00003	0,00083	0,00142
$X_{21}$	-0,00072	-0,00072	-0,00108	0,00046
$X_{22}$	-0,00028	-0,00028	-0,00075	0,00025
$X_{23}$	-0,00099	-0,00126	-0,00058	-0,00028
$X_{24}$	0,00019	0,00016	-0,00059	0,00055
$X_{25}$	-0,00076	-0,00027	-0,00024	0,00007
$X_{26}$	-0,00031	0,00070	-0,00010	-0,00049
$X_{27}$	-0,00062	-0,00061	0,00035	0,00031
$X_{28}$	-0,00194	-0,00123	-0,00056	-0,00026
$X_{29}$	-0,00007	0,00029	0,00002	-0,00076
$X_{30}$	-0,00070	-0,00077	-0,00145	-0,00102
$X_{31}$	0,00040	-0,00084	-0,00066	0,00082
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$X_{685}$	-0,0012	-0,00152	0,000637	0,00105



Halaman ini sengaja dikosongkan



Lampiran 6. Jarak Amatan ke Titik Simpul, Kelompok Awal dan Dugaan Kelompok dengan VDA

No. Pasien	Jarak Amatan					Y Awal	Y Prediksi
	Kel. 1	Kel. 2	Kel. 3	Kel. 4	Kel. 5		
23	1,7380	1,3309	1,2254	0,7281	1,1854	4	4
30	1,3019	1,1363	1,0593	0,8995	0,8416	1	5
52	1,7881	1,2208	1,2181	0,8869	1,1549	4	4
7	1,7283	1,3793	0,9099	1,0531	1,0917	3	3
9	1,4002	1,2140	1,1003	0,7130	1,0046	1	4
50	1,6319	1,2248	1,1386	0,7666	1,1128	4	4
19	1,9085	1,4504	0,8409	1,2182	1,3233	3	3
8	1,9113	1,5399	1,3938	1,2146	0,7858	5	5
46	1,1697	1,1571	1,0626	0,8169	0,9301	1	4
49	1,2016	1,1774	1,0710	0,705	0,9358	1	4
39	1,2662	1,1564	1,1652	0,7177	0,9614	1	4
24	1,0507	1,1069	1,1599	0,8867	1,8915	1	4

