

**KLASIFIKASI STADIUM KANKER PAYUDARA BERBASIS
GEN-MIRNA DENGAN SVM DAN RIDGE REGRESSION
SERTA SELEKSI FITUR ANOVA-RFE**



SKRIPSI

**AGUSTINUS
00000053639**

**PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNIK DAN INFORMATIKA
UNIVERSITAS MULTIMEDIA NUSANTARA
TANGERANG
2025**

**KLASIFIKASI STADIUM KANKER PAYUDARA BERBASIS
GEN-MIRNA DENGAN SVM DAN RIDGE REGRESSION
SERTA SELEKSI FITUR ANOVA-RFE**



SKRIPSI

Diajukan sebagai salah satu syarat untuk memperoleh
Gelar Sarjana Komputer (S.Kom.)

**AGUSTINUS
00000053639**

**PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNIK DAN INFORMATIKA
UNIVERSITAS MULTIMEDIA NUSANTARA
TANGERANG
2025**

HALAMAN PERNYATAAN TIDAK PLAGIAT

Dengan ini saya,

Nama : Agustinus
Nomor Induk Mahasiswa : 00000053639
Program Studi : Informatika

Skripsi dengan judul:

Klasifikasi Stadium Kanker Payudara Berbasis Gen-miRNA dengan SVM dan Ridge Regression serta Seleksi Fitur ANOVA-RFE

merupakan hasil karya saya sendiri bukan plagiat dari laporan karya tulis ilmiah yang ditulis oleh orang lain, dan semua sumber, baik yang dikutip maupun dirujuk, telah saya nyatakan dengan benar serta dicantumkan di Daftar Pustaka.

Jika di kemudian hari terbukti ditemukan kecurangan/penyimpangan, baik dalam pelaksanaan maupun dalam penulisan laporan karya tulis ilmiah, saya bersedia menerima konsekuensi dinyatakan TIDAK LULUS untuk mata kuliah yang telah saya tempuh.

Tangerang, 01 Juli 2025



(Agustinus)

HALAMAN PENGESAHAN

Skripsi dengan judul

KLASIFIKASI STADIUM KANKER PAYUDARA BERBASIS GEN-MIRNA DENGAN SVM DAN RIDGE REGRESSION SERTA SELEKSI FITUR ANOVA-RFE

oleh

Nama : Agustinus
NIM : 00000053639
Program Studi : Informatika
Fakultas : Fakultas Teknik dan Informatika

Telah diujikan pada hari Senin, 14 Juli 2025

Pukul 08.00 s/d 10.00 dan dinyatakan

LULUS

Dengan susunan penguji sebagai berikut

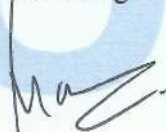
Ketua Sidang



(Dr. Ir. Winarno, M.Kom.)

NIDN: 0330106002

Pembimbing I



(Marinda Vasty Overbeek, S.Kom.,
M.Kom.)

NIDN: 0818038501

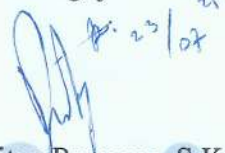
Ketua Program Studi Informatika,



(Arya Wicaksana, S.Kom., M.Eng.Sc., OCA)

NIDN: 0315109103

Penguji



(Angga Aditya Permana, S.Kom.,
M.Kom.)

NIDN: 0407128901

Pembimbing II



(David Agustriawan, S.Kom., M.Sc.,
Ph.D.)

NIDN: 0525088601

**HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK
KEPENTINGAN AKADEMIS**

Yang bertanda tangan di bawah ini:

Nama : Agustinus
NIM : 00000053639
Program Studi : Informatika
Jenjang : S1
Judul Karya Ilmiah : Klasifikasi Stadium Kanker Payudara
Berbasis Gen-miRNA dengan SVM
dan Ridge Regression serta Seleksi
Fitur ANOVA-RFE

Menyatakan dengan sesungguhnya bahwa saya bersedia:

- ☒ Saya bersedia memberikan izin sepenuhnya kepada Universitas Multimedia Nusantara untuk mempublikasikan hasil karya ilmiah saya ke dalam repositori Knowledge Center sehingga dapat diakses oleh Sivitas Akademika UMN/Publik. Saya menyatakan bahwa karya ilmiah yang saya buat tidak mengandung data yang bersifat konfidensial.
- ☐ Saya tidak bersedia mempublikasikan hasil karya ilmiah ini ke dalam repositori Knowledge Center, dikarenakan: dalam proses pengajuan publikasi ke jurnal/konferensi nasional/internasional (dibuktikan dengan *letter of acceptance*) **.

Tangerang, 01 Juli 2025

Yang menyatakan



Agustinus

**Jika tidak bisa membuktikan LoA jurnal/HKI, saya bersedia mengizinkan penuh karya ilmiah saya untuk dipublikasikan ke KC UMN dan menjadi hak institusi UMN.

HALAMAN PERSEMBAHAN / MOTTO

"Turn your fear into adrenaline to fuel your passion, utilize that time to get it done and prove that you can"

Agustinus

UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA

KATA PENGANTAR

Puji syukur ke hadirat Tuhan Yang Maha Esa atas terselesaikannya penulisan skripsi ini sebagai bagian dari pemenuhan persyaratan akademik pada Program Studi Informatika, Universitas Multimedia Nusantara. Penyusunan skripsi ini melalui berbagai proses yang penuh tantangan, refleksi, dan pembelajaran yang berharga. Ucapan terima kasih disampaikan kepada pihak-pihak yang telah memberikan dukungan, bantuan, dan bimbingan, antara lain:

1. Bapak Dr. Ir. Andrey Andoko, M.Sc., selaku Rektor Universitas Multimedia Nusantara.
2. Bapak Dr. Eng. Niki Prastomo, S.T., M.Sc., selaku Dekan Fakultas Teknik dan Informatika Universitas Multimedia Nusantara.
3. Bapak Arya Wicaksana, S.Kom., M.Eng.Sc., OCA, selaku Ketua Program Studi Informatika Universitas Multimedia Nusantara.
4. Ibu Marlinda Vasty Overbeek, S.Kom., M.Kom., selaku Pembimbing Pertama yang dengan sabar membimbing dan meluangkan waktu serta pemikiran dalam setiap tahap penyusunan skripsi ini.
5. Bapak David Agustriawan, S.Kom., M.Sc., Ph.D., selaku Pembimbing Kedua yang telah memberikan masukan dan dukungan dalam penyusunan skripsi ini.
6. Keluarga yang senantiasa memberikan semangat, kepercayaan, serta dukungan moral dan material.

Skripsi ini diharapkan dapat berkontribusi nyata dalam pengembangan ilmu pengetahuan dan mendorong riset lanjutan di masa depan.

Tangerang, 01 Juli 2025



Agustinus

KLASIFIKASI STADIUM KANKER PAYUDARA BERBASIS GEN-MIRNA DENGAN SVM DAN RIDGE REGRESSION SERTA SELEKSI FITUR

ANOVA-RFE

Agustinus

ABSTRAK

Kanker payudara merupakan penyebab utama kematian pada wanita, di mana deteksi stadium dini krusial untuk meningkatkan prognosis pasien. Penelitian ini bertujuan mengembangkan model klasifikasi berbasis *machine learning* untuk membedakan stadium I dan III kanker payudara menggunakan data ekspresi gen dari *The Cancer Genome Atlas* (TCGA) pada wanita ras kulit putih. Pendekatan yang digunakan melibatkan seleksi fitur dengan *Analysis of Variance* (ANOVA) untuk memilih 1000 fitur awal, dilanjutkan *Recursive Feature Elimination* (RFE) hingga diperoleh 35 fitur optimal, dan klasifikasi menggunakan *Logistic Regression* dengan penalti L2. Hasil penelitian menunjukkan model terbaik mampu mencapai akurasi 97%, presisi 97%, *recall* 97%, dan F1-score 97% pada skenario stadium I vs. III, dengan durasi komputasi hanya 8 detik. Performa ini lebih unggul dibandingkan skenario stadium II vs. III (F1-score maksimum 88%) dan penggunaan data miRNA (F1-score maksimum 82%), menegaskan efektivitas data ekspresi gen dalam klasifikasi stadium. Sebanyak 35 gen kandidat biomarker berhasil diidentifikasi, dengan 10 gen didukung literatur terkait kanker payudara, sementara 25 lainnya berpotensi sebagai penanda baru. Analisis ROC AUC individual menunjukkan nilai maksimum 0,639, mengindikasikan kekuatan prediktif bergantung pada kombinasi multivariat. Penelitian ini membuktikan potensi pendekatan pembelajaran mesin dalam mendeteksi stadium kanker payudara secara akurat dan efisien.

Kata Kunci: Ekspresi Gen, Kanker Payudara, Klasifikasi Stadium, *Machine learning*, Seleksi Fitur

U N I V E R S I T A S
M U L T I M E D I A
N U S A N T A R A

**BREAST CANCER STAGE CLASSIFICATION BASED ON MIRNA GENES
USING SVM AND RIDGE REGRESSION WITH ANOVA-RFE FEATURE
SELECTION**

Agustinus

ABSTRACT

Breast cancer is a leading cause of death in women, where early stage detection is crucial to improve patient prognosis. This study aims to develop a machine learning-based classification model to distinguish between stages I and III of breast cancer using gene expression data from The Cancer Genome Atlas (TCGA) in white women. The approach used involved feature selection using Analysis of Variance (ANOVA) to select 1000 initial features, followed by Recursive Feature Elimination (RFE) to obtain 35 optimal features, and classification using Logistic Regression with L2 penalty. The results showed that the best model was able to achieve 97% accuracy, 97% precision, 97% recall, and F1-score. 97%, and F1-score 97% in the stage I vs. III scenario, with a computation duration of only 8 seconds. This performance is superior to the stage II vs. III scenario (maximum F1-score 88%) and the use of miRNA data (maximum F1-score 82%), confirming the effectiveness of gene expression data in stage classification. A total of 35 biomarker candidate genes were identified, with 10 supported by breast cancer-related literature, while the other 25 have potential as novel markers. ROC analysis of individual AUCs showed a maximum value of 0.639, indicating predictive power was dependent on multivariate combinations. This research proves the potential of machine learning approaches in detecting breast cancer stages accurately and efficiently.

Keywords: *Gene Expression, Breast Cancer, Stage Classification, Machine learning, Feature Selection*

U N I V E R S I T A S
M U L T I M E D I A
N U S A N T A R A

DAFTAR ISI

HALAMAN JUDUL	i
PERNYATAAN TIDAK MELAKUKAN PLAGIAT	ii
HALAMAN PENGESAHAN	iii
HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH	iv
HALAMAN PERSEMBAHAN/MOTO	v
KATA PENGANTAR	vi
ABSTRAK	vii
ABSTRACT	viii
DAFTAR ISI	ix
DAFTAR TABEL	xi
DAFTAR GAMBAR	xii
DAFTAR RUMUS	xiv
DAFTAR LAMPIRAN	xv
BAB 1 PENDAHULUAN	1
1.1 Latar Belakang Masalah	1
1.2 Rumusan Masalah	4
1.3 Batasan Permasalahan	4
1.4 Tujuan Penelitian	5
1.5 Manfaat Penelitian	5
1.6 Sistematika Penulisan	6
BAB 2 LANDASAN TEORI	7
2.1 Kanker Payudara	7
2.2 Ekspresi Gen (RNA-seq)	8
2.3 Ekspresi miRNA (stem-loop)	9
2.4 <i>Differentially Expressed Genes</i> (DEG)	9
2.5 <i>Feature Selection</i>	10
2.5.1 <i>Analysis of Variance</i> (ANOVA)	11
2.5.2 <i>Recursive Feature Elimination</i> (RFE)	12
2.6 <i>Logistic Regression</i> dengan L2 <i>Ridge Regularization</i>	13
2.7 <i>Support Vector Machine</i> (SVM)	15
2.8 <i>Confusion Matrix</i>	20
BAB 3 METODOLOGI PENELITIAN	24
3.1 Gambaran Umum Penelitian	24
3.2 Spesifikasi Perangkat	25
3.2.1 Hardware (Lokal)	25
3.2.2 Software (Lokal)	25
3.2.3 Spesifikasi Cloud (Kaggle - Versi Gratis)	25
3.3 Studi Literatur	26
3.4 Pengumpulan Data	26
3.5 Praproses Data	26
3.6 Seleksi Fitur	29
3.7 Pembangunan Model	30
3.8 Evaluasi Model	31
3.9 Skenario Eksperimen	32
3.10 Interpretasi Biomarker	32
BAB 4 HASIL DAN DISKUSI	34
4.1 Implementasi Alur Kerja Penelitian	34
4.1.1 Struktur dan Format Dataset	34

4.1.2	Kategorisasi dan Labeling Kelas	34
4.1.3	Pengendalian Bias Rasial	35
4.1.4	Penggabungan Label ke Data Ekspresi	36
4.1.5	Ringkasan Dataset Pasca-Praproses	37
4.2	Hasil Tahapan Seleksi Fitur	38
4.2.1	DEG Limma	38
4.2.2	Tahap Penyaringan Awal	39
4.2.3	Seleksi Fitur Akhir dengan RFE	41
4.2.4	Pembangunan Model dan Evaluasi	42
4.2.5	Kriteria Pemilihan Model Optimal	42
4.3	Hasil dan Evaluasi Skenario Pendekatan Seleksi Fitur Limma + RFE	43
4.3.1	Model Klasifikasi SVM	43
4.3.2	Model Klasifikasi <i>Logistic Regression</i>	47
4.4	Hasil dan Evaluasi Skenario Pendekatan Seleksi Fitur ANOVA + RFE	52
4.4.1	Model Klasifikasi SVM	52
4.4.2	Model Klasifikasi <i>Logistic Regression</i>	57
4.5	Model Optimal	63
4.5.1	Kontribusi Fitur dalam Model Optimal	63
4.5.2	Evaluasi Perbandingan Strategi Seleksi Fitur	66
4.6	Analisis dan Interpretasi Kandidat Biomarker	68
BAB 5	SIMPULAN DAN SARAN	73
5.1	Simpulan	73
5.2	Saran	74
DAFTAR PUSTAKA		75



DAFTAR TABEL

Tabel 2.1	Pengelompokan stadium kanker payudara menurut AJCC edisi ke-8	8
Tabel 2.2	Struktur <i>Confusion Matrix</i> untuk Klasifikasi Biner	20
Tabel 2.3	Interpretasi Nilai AUC dalam Studi Klinis	23
Tabel 3.1	Informasi dataset yang digunakan	26
Tabel 3.2	Tuning <i>hyperparameter</i> untuk <i>SVM</i> dan <i>Logistic Regression</i>	31
Tabel 3.3	Skenario eksperimen	32
Tabel 4.1	Dimensi data sebelum dan sesudah transposisi	34
Tabel 4.2	Distribusi data awal berdasarkan <code>ajcc_pathologic_stage.diagnoses</code>	35
Tabel 4.3	Distribusi ras berdasarkan <code>race.demographic</code>	36
Tabel 4.4	Statistik 5 sampel pertama sebelum dan sesudah normalisasi	37
Tabel 4.5	Hasil praproses data berdasarkan skenario klasifikasi dan jenis data	37
Tabel 4.6	Lima hasil analisis diferensial ekspresi gen lima	39
Tabel 4.7	Jumlah fitur setelah penyaringan awal	40
Tabel 4.8	Model SVM optimal stadium I vs. III (Limma+RFE)	45
Tabel 4.9	Model SVM optimal stadium II vs. III (Limma+RFE)	47
Tabel 4.10	Model LogReg optimal stadium I vs. III (Limma+RFE)	49
Tabel 4.11	Model LogReg optimal stadium II vs. III (Limma+RFE)	52
Tabel 4.12	Model SVM optimal stadium I vs. III (ANOVA+RFE)	54
Tabel 4.13	Model SVM optimal stadium II vs. III (ANOVA+RFE)	57
Tabel 4.14	Model LogReg optimal stadium I vs. III (ANOVA+RFE)	60
Tabel 4.15	Model LogReg optimal stadium II vs. III (ANOVA+RFE)	62
Tabel 4.16	Perbandingan performa model terbaik berdasarkan strategi seleksi fitur	67
Tabel 4.17	Daftar gen kandidat biomarker berdasarkan model terbaik	69

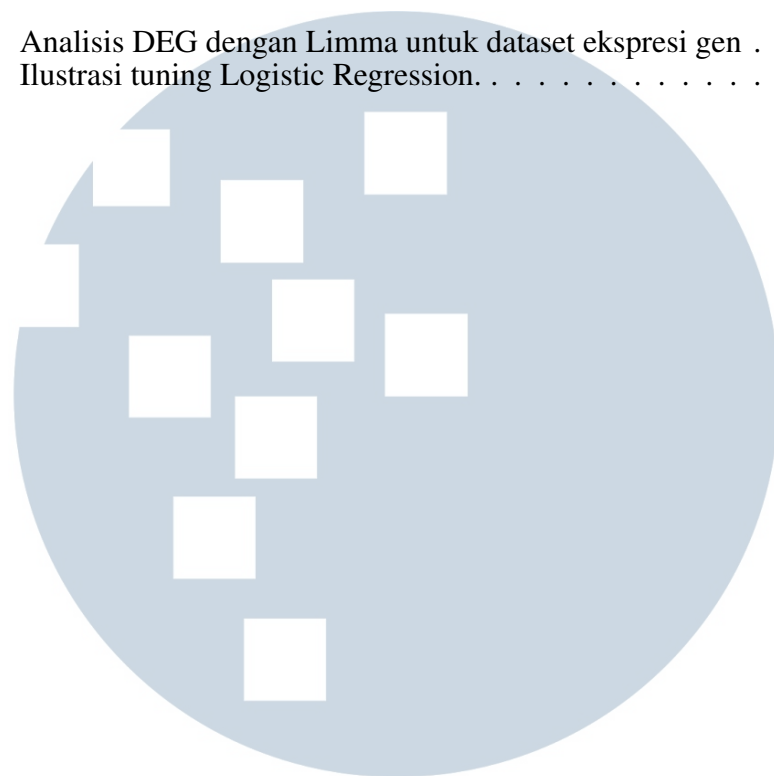


DAFTAR GAMBAR

Gambar 2.1	Ilustrasi optimal <i>hyperplane</i> dengan margin maksimum . .	16
Gambar 3.1	Gambaran umum alur kerja penelitian	24
Gambar 3.2	Alur kerja praproses data	27
Gambar 3.3	Alur kerja seleksi fitur untuk pendekatan DEG <i>analysis</i> dan statistika	29
Gambar 4.1	Akurasi model SVM dengan Limma+RFE (stage I vs. III) .	44
Gambar 4.2	F1-Score model SVM dengan Limma+RFE (stage I vs. III)	44
Gambar 4.3	Akurasi model SVM dengan Limma+RFE (stage II vs. III)	46
Gambar 4.4	F1-Score model SVM dengan Limma+RFE (stage II vs. III)	46
Gambar 4.5	Akurasi model <i>logistic regression</i> dengan Limma+RFE (stage I vs. III)	48
Gambar 4.6	F1-Score model <i>logistic regression</i> dengan Limma+RFE (stage I vs. III)	49
Gambar 4.7	Akurasi model <i>logistic regression</i> dengan Limma+RFE (stage II vs. III)	50
Gambar 4.8	F1-Score model <i>logistic regression</i> dengan Limma+RFE (stage II vs. III)	51
Gambar 4.9	Akurasi model SVM dengan ANOVA+RFE (stadium I vs. III)	53
Gambar 4.10	F1-Score model SVM dengan ANOVA+RFE (stadium I vs. III)	54
Gambar 4.11	Akurasi model SVM dengan ANOVA+RFE (stadium II vs. III)	55
Gambar 4.12	F1-Score model SVM dengan ANOVA+RFE (stadium II vs. III)	56
Gambar 4.13	Akurasi model <i>logistic regression</i> dengan ANOVA+RFE (stadium I vs. III)	58
Gambar 4.14	F1-Score model <i>logistic regression</i> dengan ANOVA+RFE (stadium I vs. III)	59
Gambar 4.15	Akurasi model <i>logistic regression</i> dengan ANOVA+RFE (stadium II vs. III)	61
Gambar 4.16	F1-Score model <i>logistic regression</i> dengan ANOVA+RFE (stadium II vs. III)	62
Gambar 4.17	Tingkat kepentingan pada kombinasi 10 fitur	64
Gambar 4.18	Tingkat kepentingan pada kombinasi 35 fitur (model optimal)	65
Gambar 4.19	Kurva ROC untuk 35 fitur gen kandidat secara individual. .	71

DAFTAR KODE

Kode 4.1	Analisis DEG dengan Limma untuk dataset ekspresi gen	39
Kode 4.2	Ilustrasi tuning Logistic Regression.	42



UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA

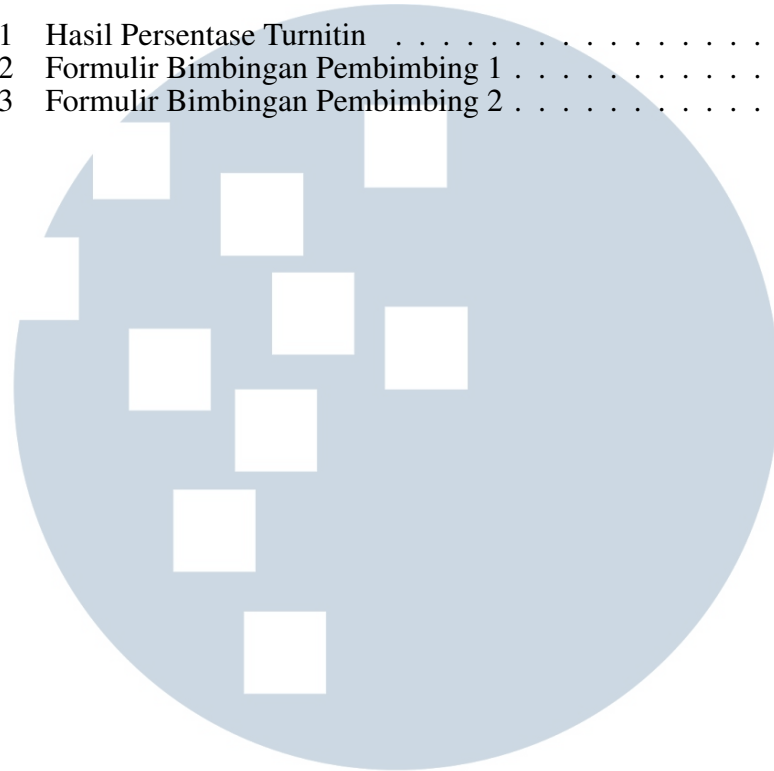
DAFTAR RUMUS

Rumus 2.1	ANOVA <i>Sum of Squares Between</i>	11
Rumus 2.2	ANOVA <i>Mean Square Between</i>	11
Rumus 2.3	ANOVA <i>Sum of Squares Within</i>	11
Rumus 2.4	ANOVA <i>Mean Square Within</i>	12
Rumus 2.5	ANOVA F-statistik	12
Rumus 2.6	<i>Logistic Regression</i> Fungsi linear	13
Rumus 2.7	<i>Logistic Regression</i> Aktivasi sigmoid	14
Rumus 2.8	<i>Logistic Regression</i> Cross-entropy loss	14
Rumus 2.10	<i>Logistic Regression</i> Gradient descent $J(w,b)$	14
Rumus 2.12	<i>Logistic Regression</i> Pembaruan parameter	15
Rumus 2.13	<i>Support Vector Machine</i> Hyperplane	16
Rumus 2.15	<i>Support Vector Machine</i> Syarat Klasifikasi Benar	16
Rumus 2.16	<i>Support Vector Machine</i> Hyperplane optimal	17
Rumus 2.17	<i>Support Vector Machine</i> Margin	17
Rumus 2.19	<i>Support Vector Machine</i> Linear	17
Rumus 2.20	<i>Support Vector Machine</i> Lagrangian	18
Rumus 2.22	<i>Support Vector Machine</i> Lagrangian w dan b	18
Rumus 2.24	<i>Support Vector Machine</i> Optimasi Lagrangian	18
Rumus 2.26	<i>Support Vector Machine</i> Slack variable	19
Rumus 2.27	<i>Support Vector Machine</i> Hard Margin	19
Rumus 2.28	<i>Support Vector Machine</i> kernel	19
Rumus 2.29	<i>Support Vector Machine</i> Linear Kernel	20
Rumus 2.30	<i>Support Vector Machine</i> RBF Kernel	20
Rumus 2.31	<i>Confusion Matrix</i> Accuracy	21
Rumus 2.32	<i>Confusion Matrix</i> Precision	21
Rumus 2.33	<i>Confusion Matrix</i> Recall	22
Rumus 2.34	<i>Confusion Matrix</i> Specificity	22
Rumus 2.35	<i>Confusion Matrix</i> F1-Score	22
Rumus 2.36	ROC Curve TPR FPR	22
Rumus 4.1	Kompleksitas komputasi RFE	41
Rumus 4.2	Kompleksitas iterasi RFE SVM linear	41
Rumus 4.3	Kompleksitas RFE SVM	42
Rumus 4.4	Kompleksitas pelatihan RFE <i>Logistic Regression</i>	42
Rumus 4.5	Kompleksitas RFE <i>Logistic Regression</i>	42

UNIVERSITAS
MULTIMEDIA
NUSANTARA

DAFTAR LAMPIRAN

Lampiran 1	Hasil Persentase Turnitin	82
Lampiran 2	Formulir Bimbingan Pembimbing 1	91
Lampiran 3	Formulir Bimbingan Pembimbing 2	94



UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA