

**PREDIKSI STADIUM KANKER PROSTAT BERDASARKAN
EKSPRESI GEN-MIRNA MENGGUNAKAN SVM DAN LR
DENGAN SELEKSI FITUR DESEQ2-RFE**



SKRIPSI

**GREGORIUS IVAN HALIM
00000054295**

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

**PREDIKSI STADIUM KANKER PROSTAT BERDASARKAN
EKSPRESI GEN-MIRNA MENGGUNAKAN SVM DAN LR
DENGAN SELEKSI FITUR DESEQ2-RFE**



SKRIPSI

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

**GREGORIUS IVAN HALIM
00000054295**

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

HALAMAN PERNYATAAN TIDAK PLAGIAT

Dengan ini saya,

Nama : Gregorius Ivan Halim

Nomor Induk Mahasiswa : 00000054295

Program Studi : Informatika

Skripsi dengan judul:

Prediksi Stadium Kanker Prostat Berdasarkan Ekspresi Gen-miRNA menggunakan SVM dan LR dengan Seleksi Fitur DESeq2-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



(Gregorius Ivan Halim)

UNIVERSITAS
MULTIMEDIA
NUSANTARA

HALAMAN PENGESAHAN

Skripsi dengan judul

**PREDIKSI STADIUM KANKER PROSTAT BERDASARKAN EKSPRESI
GEN-MIRNA MENGGUNAKAN SVM DAN LR DENGAN SELEKSI
FITUR DESEQ2-RFE**

oleh

Nama : Gregorius Ivan Halim
NIM : 00000054295
Program Studi : Informatika
Fakultas : Fakultas Teknik dan Informatika

Telah diujikan pada hari Jumat, 18 Juli 2025

Pukul 08.00 s/d 10.00 dan dinyatakan

LULUS

Dengan susunan penguji sebagai berikut

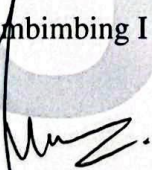
Ketua Sidang

Penguji


(Moeljono Widjaja, B.Sc., M.Sc.,
Ph.D.)

NIDN: 0311106903

Pembimbing I


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

NIDN: 0818038501


(Dr. Maria Irmira Prasetiyowati, S.Kom.,
M.T.)

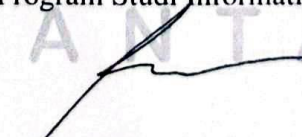
NIDN: 0725057201

Pembimbing II


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

NIDN: 0525088601

Ketua Program Studi Informatika,


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

NIDN: 0315109103

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

Yang bertanda tangan di bawah ini:

Nama : Gregorius Ivan Halim
NIM : 00000054295
Program Studi : Informatika
Jenjang : S1
Judul Karya Ilmiah : Prediksi Stadium Kanker Prostat
Berdasarkan Ekspresi Gen-miRNA
menggunakan SVM dan LR dengan
Seleksi Fitur DESeq2-RFE

Menyatakan dengan sesungguhnya bahwa saya bersedia (**pilih salah satu**):

- ☒ 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*) **.
- ☐ Lainnya, pilih salah satu:
- Hanya dapat diakses secara internal Universitas Multimedia Nusantara
 - Embargo publikasi karya ilmiah dalam kurun waktu tiga tahun.

Tangerang, 01 Juli 2025

Yang menyatakan



Gregorius Ivan Halim

HALAMAN PERSEMBAHAN / MOTTO

"It's not the hours you put in your work that counts, it's the work you put in the hours."

Sam Ewing



KATA PENGANTAR

Puji syukur penulis panjatkan ke hadirat Tuhan Yang Maha Esa atas rahmat-Nya sehingga Tugas Akhir ini dapat diselesaikan. Tugas ini disusun sebagai syarat kelulusan Program Studi Informatika, Universitas Multimedia Nusantara. Penelitian ini mengembangkan model prediksi kanker prostat berbasis data gen dan miRNA dengan pendekatan komputasional. Terima kasih kepada semua pihak yang telah mendukung penyusunan tugas ini diantaranya:

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 telah memberikan bimbingan, arahan, dan motivasi.
5. Bapak David Agustiawan, S.Kom., M.Sc., Ph.D, selaku Pembimbing kedua, yang turut memberikan bimbingan dan masukan.
6. Keluarga saya yang selalu memberikan bantuan, dukungan moral maupun material.
7. Serta semua pihak yang tidak dapat disebutkan satu per satu namun telah membantu dalam proses penyusunan tugas akhir ini.

Penulis berharap karya ilmiah ini dapat memberikan manfaat bagi pengembangan ilmu pengetahuan, khususnya dalam penelitian terkait deteksi kanker prostat, serta menjadi referensi bagi studi lanjutan di masa mendatang.

Tangerang, 01 Juli 2025



Gregorius Ivan Halim

PREDIKSI STADIUM KANKER PROSTAT BERDASARKAN EKSPRESI GEN-MIRNA MENGGUNAKAN SVM DAN LR DENGAN SELEKSI FITUR DESEQ2-RFE

Gregorius Ivan Halim

ABSTRAK

Klasifikasi stadium kanker prostat yang akurat sangat penting untuk meningkatkan hasil pengobatan dan mendukung pengambilan keputusan klinis. Penelitian ini bertujuan mengembangkan model prediktif untuk klasifikasi stadium kanker prostat berdasarkan data ekspresi gen dan miRNA, dengan memanfaatkan kombinasi teknik seleksi fitur dan algoritma pembelajaran mesin. Fokus utama penelitian ini adalah membandingkan kemampuan prediktif dari kedua jenis data dalam membedakan antara kanker prostat stadium II dan III pada pasien pria non-Hispanik kulit putih. Proses seleksi fitur dilakukan secara bertahap, dimulai dengan metode filter seperti DESeq2, Limma, ANOVA, dan MRMR, kemudian dilanjutkan dengan teknik wrapper Recursive Feature Elimination (RFE). Model dibangun menggunakan algoritma Support Vector Machine (SVM) dan Logistic Regression (LR) serta dilatih pada data RNA-Seq yang telah distandarisasi. Evaluasi performa dilakukan dengan menggunakan metrik akurasi, presisi, recall, dan skor F1 makro. Hasil menunjukkan bahwa model berbasis ekspresi gen secara konsisten mengungguli model berbasis miRNA pada semua metrik evaluasi. Model terbaik diperoleh dari kombinasi DESeq2–RFE–SVM, dengan akurasi 96,1% dan skor F1 makro 96%, hanya menggunakan 18 fitur gen. Meskipun sebagian besar gen memiliki skor AUC rendah dan korelasi lemah hingga sedang terhadap stadium kanker ketika dievaluasi secara individu, integrasi multigen memungkinkan klasifikasi stadium yang sangat efektif. Temuan ini menegaskan bahwa fitur ekspresi gen lebih informatif dibandingkan miRNA dalam klasifikasi stadium kanker prostat. Ke depan, pengembangan klasifikasi multikelas, penerapan metode deep learning, penggunaan RFE secara langsung, dan pengolahan fitur pada sistem berperforma tinggi disarankan untuk meningkatkan akurasi dan relevansi klinis model.

Kata kunci: Biomarker, Ekspresi gen, Kanker prostat, *Machine learning*, miRNA, Seleksi fitur

**PREDICTION OF PROSTATE CANCER STAGE BASED ON GENE-MIRNA
EXPRESSION USING SVM AND LR WITH DESEQ2-RFE FEATURE
SELECTION**

Gregorius Ivan Halim

ABSTRACT

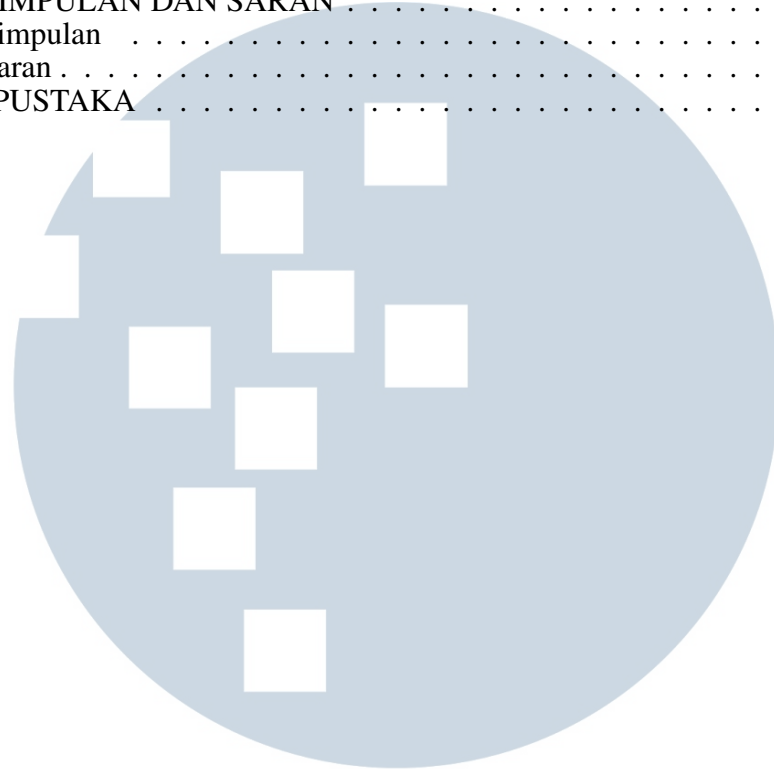
Accurate staging classification of prostate cancer is crucial for improving treatment outcomes and supporting clinical decision-making. This study aims to develop predictive models for prostate cancer staging using gene and miRNA expression data, employing a combination of feature selection techniques and machine learning algorithms. The primary objective is to compare the predictive capabilities of both data types in distinguishing between stage II and stage III prostate cancer in non-Hispanic white male patients. Feature selection was performed in a multi-step process, beginning with filter methods such as DESeq2, Limma, ANOVA, and MRMR, followed by the wrapper-based Recursive Feature Elimination (RFE) technique. Predictive models were built using Support Vector Machine (SVM) and Logistic Regression (LR) algorithms, trained on standardized RNA-Seq data. Model performance was evaluated using accuracy, precision, recall, and macro F1-score metrics. The results demonstrated that gene expression-based models consistently outperformed miRNA-based models across all evaluation metrics. The best performance was achieved by the DESeq2-RFE-SVM model, which reached 96.1% accuracy and a 96% macro F1-score using only 18 gene features. Although most individual genes showed low AUC scores and weak to moderate correlations with cancer stage, their collective integration enabled highly effective stage classification. These findings confirm that gene expression features are more informative than miRNA in the context of prostate cancer stage classification. Future work should consider exploring multiclass classification, incorporating deep learning approaches, applying RFE directly, and executing feature selection on higher-performance systems to further enhance model accuracy and clinical relevance.

Keywords: *Feature selection, Gene expression, Machine learning, miRNA, Prostate cancer.*

DAFTAR ISI

HALAMAN JUDUL	i
PERNYATAAN TIDAK MELAKUKAN PLAGIAT	ii
HALAMAN PENGESAHAN	iii
HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH	iv
HALAMAN PERSEMBAHAN/MOTO	vi
KATA PENGANTAR	vii
ABSTRAK	viii
ABSTRACT	ix
DAFTAR ISI	x
DAFTAR TABEL	xii
DAFTAR GAMBAR	xiii
DAFTAR RUMUS	xv
DAFTAR LAMPIRAN	xvi
BAB 1 PENDAHULUAN	1
1.1 Latar Belakang Masalah	1
1.2 Rumusan Masalah	3
1.3 Batasan Permasalahan	4
1.4 Tujuan Penelitian	4
1.5 Manfaat Penelitian	5
1.6 Sistematika Penulisan	5
BAB 2 LANDASAN TEORI	7
2.1 Kanker Prostat	7
2.2 Ekspresi Gen RNA Sequence	8
2.3 MikroRNA	8
2.4 Analisis DEG	9
2.5 Seleksi Fitur	9
2.5.1 Analysis of Variance (ANOVA)	10
2.5.2 Minimum Redundancy Maximum Relevance (MRMR)	11
2.5.3 Recursive Feature Elimination (RFE)	12
2.5.4 Support Vector Machine (SVM)	13
2.5.5 L2-penalized Logistic Regression	18
2.5.6 Model Evaluation	20
BAB 3 METODOLOGI PENELITIAN	23
3.1 Pengambilan & Praproses Data	23
3.2 Seleksi Fitur Gen & miRNA	24
3.3 Pembuatan Model	25
3.4 Analisis Biomarker	27
3.5 Spesifikasi Perangkat	28
BAB 4 HASIL DAN DISKUSI	29
4.1 Praproses Data	29
4.2 Seleksi Fitur	34
4.2.1 Metode DEG	34
4.2.2 Metode Statistik	38
4.2.3 RFE	39
4.3 Pembangunan Model	41
4.4 Hasil dan Evaluasi Model	43
4.4.1 Hasil Skenario Model Ekspresi Gen	44
4.4.2 Hasil Skenario Model miRNA	47

4.4.3	Evaluasi Model	51
4.4.4	Analisis Biomarker	52
BAB 5	SIMPULAN DAN SARAN	59
5.1	Simpulan	59
5.2	Saran	60
DAFTAR PUSTAKA		61



UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA

DAFTAR TABEL

Tabel 2.1	Pengelompokan stadium prostat AJCC dan UICC edisi 2010	7
Tabel 2.2	<i>Confusion matrix</i> untuk klasifikasi biner	20
Tabel 3.1	Kriteria seleksi fitur gen menggunakan DEG	25
Tabel 3.2	Selisih negatif rata-rata F1-score makro dan akurasi antara data pelatihan dan pengujian	25
Tabel 3.3	Penyesuaian <i>hyperparameter</i> untuk SVM	26
Tabel 3.4	Penyesuaian <i>hyperparameter</i> untuk LR-L2	27
Tabel 3.5	Kombinasi seleksi fitur dan model	27
Tabel 4.1	Jumlah sampel dan fitur pada masing-masing jenis data	29
Tabel 4.2	Contoh data fenotipe dengan tiga fitur utama (5 baris teratas)	30
Tabel 4.3	Contoh data fenotipe setelah pemrosesan label stadium kanker (5 baris teratas)	31
Tabel 4.4	Normalisasi contoh fitur ekspresi gen ENSG000000000003.15	34
Tabel 4.5	Normalisasi fitur ekspresi miRNA hsa-let-7a-1	34
Tabel 4.6	Lima baris pertama hasil analisis ekspresi gen menggunakan <i>DESeq2</i>	35
Tabel 4.7	Lima baris pertama hasil analisis ekspresi miRNA menggunakan <i>DESeq2</i>	36
Tabel 4.8	Lima baris pertama hasil analisis ekspresi gen menggunakan <i>limma</i>	37
Tabel 4.9	Lima baris pertama hasil analisis ekspresi miRNA menggunakan <i>Limma</i>	37
Tabel 4.10	Hasil dari seleksi fitur menggunakan DEG pada data ekspresi gen dan miRNA	38
Tabel 4.11	Hasil dari seleksi fitur menggunakan statistik pada data ekspresi gen dan miRNA	39
Tabel 4.12	Nilai koefisien 5 fitur gen terpilih menggunakan model SVM	40
Tabel 4.13	Nilai koefisien 5 fitur gen terpilih menggunakan model LR-L2	40
Tabel 4.14	Nilai koefisien 5 fitur miRNA terpilih menggunakan model SVM	41
Tabel 4.15	Nilai koefisien 5 fitur miRNA terpilih menggunakan model LR-L2	41
Tabel 4.16	Hasil evaluasi model terbaik pada data ekspresi gen (dalam persen)	52
Tabel 4.17	Daftar gen terpilih dari seleksi fitur	53
Tabel 4.18	Daftar gen dengan hubungan pada kanker prostat	54
Tabel 4.19	Daftar gen dengan hubungan pada jenis kanker lain	55

DAFTAR GAMBAR

Gambar 2.1	Ilustrasi pemisahan kelas oleh hyperplane pada SVM . . .	14
Gambar 2.2	Ilustrasi pencarian margin pada SVM	15
Gambar 2.3	Ilustrasi <i>kernel trick</i> untuk transformasi ke ruang fitur berdimensi lebih tinggi	17
Gambar 2.4	kurva ROC-AUC pada berbagai nilai ambang keputusan . .	22
Gambar 3.1	Diagram alur jalan dari penelitian	23
Gambar 4.1	Distribusi stadium kanker pada data ekspresi gen akhir . .	32
Gambar 4.2	Distribusi stadium kanker pada data miRNA akhir	33
Gambar 4.3	Hasil akurasi skenario seleksi fitur dan klasifikasi terbaik pada data ekspresi gen	44
Gambar 4.4	Hasil <i>F1-Score</i> makro skenario seleksi fitur dan klasifikasi terbaik pada data ekspresi gen	45
Gambar 4.5	Hasil ROC-AUC skenario seleksi fitur dan klasifikasi terbaik pada data ekspresi gen	47
Gambar 4.6	Hasil akurasi skenario seleksi fitur dan klasifikasi terbaik pada data miRNA	48
Gambar 4.7	Hasil <i>F1-Score</i> makro skenario seleksi fitur dan klasifikasi terbaik pada data miRNA	48
Gambar 4.8	Hasil ROC-AUC skenario seleksi fitur dan klasifikasi terbaik pada data miRNA	50
Gambar 4.9	Kurva ROC dari masing-masing gen untuk menilai kemampuan diskriminatifnya dalam membedakan stadium awal dan lanjut kanker prostat.	56
Gambar 4.10	<i>Heatmap</i> korelasi antara ekspresi gen dengan stadium kanker prostat	57



DAFTAR KODE

Kode 4.1	Membaca data ekspresi gen, miRNA, dan fenotipe	29
Kode 4.2	Transposisi data ekspresi gen dan miRNA	29
Kode 4.3	Penyaringan fitur penting dari data fenotipe	30
Kode 4.4	Klasifikasi stadium kanker dan pelabelan biner	30
Kode 4.5	Menyaring data fenotipe untuk kelompok etnis kulit putih	31
Kode 4.6	Penggabungan data ekspresi gen dan miRNA dengan data fenotipe serta penghapusan fitur nol	31
Kode 4.7	Penghapusan fitur dengan seluruh nilai nol dari data ekspresi gen dan miRNA	32
Kode 4.8	Fungsi untuk praproses dan normalisasi data	33
Kode 4.9	Pra-pemrosesan data untuk analisis DEG	35
Kode 4.10	Analisis DEG dengan DESeq2 untuk data miRNA dan ekspresi gen	35
Kode 4.11	Analisis DEG menggunakan Limma untuk data miRNA dan ekspresi gen	36
Kode 4.12	Filter fitur miRNA berdasarkan hasil DESeq2 dan Limma	37
Kode 4.13	Fungsi seleksi fitur menggunakan ANOVA	38
Kode 4.14	Fungsi seleksi fitur menggunakan mRMR	38
Kode 4.15	Fungsi seleksi fitur menggunakan RFE	39
Kode 4.16	Fungsi pembuatan model dengan <i>stratified train-test split</i> dan <i>grid search</i> pada berbagai <i>seed</i>	41
Kode 4.17	Fungsi lanjutan pembuatan model dengan <i>stratified train-test split</i> dan <i>grid search</i> pada berbagai <i>seed</i>	42

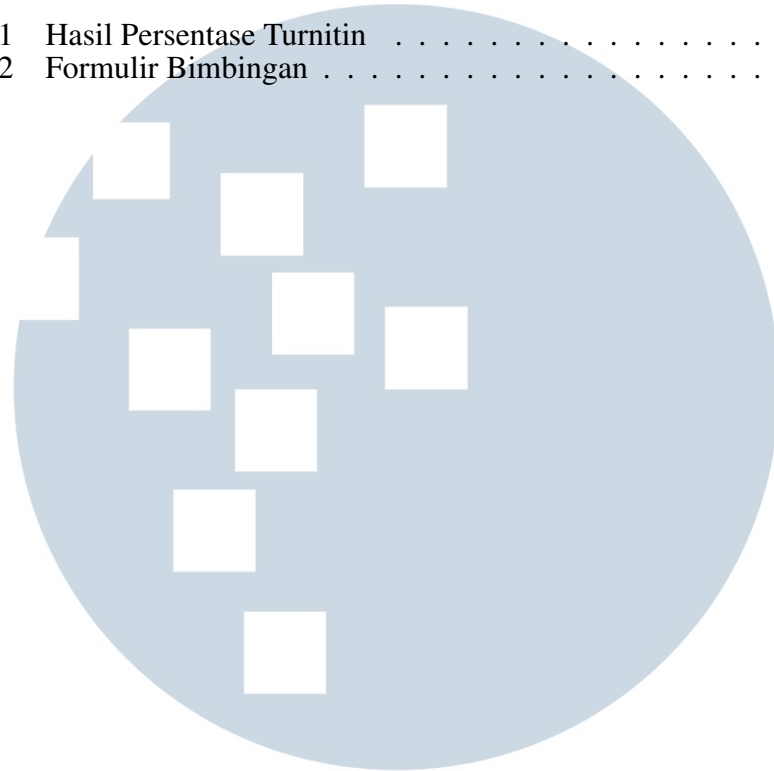


DAFTAR RUMUS

Rumus 2.1	<i>Analysis of Variance (ANOVA)</i>	10
Rumus 2.2	<i>Mean Square Between (MSB)</i>	10
Rumus 2.3	<i>Sum of Squares Between groups (SSB)</i>	10
Rumus 2.4	<i>Degrees of Freedom Between (DFB)</i>	10
Rumus 2.5	<i>Mean Square Within (MSW)</i>	11
Rumus 2.6	<i>Sum of Squares Within (SSW)</i>	11
Rumus 2.7	<i>Degrees of Freedom Within (DFW)</i>	11
Rumus 2.8	<i>Mutual Information</i>	11
Rumus 2.9	Relevansi Maksimum	11
Rumus 2.10	Redundansi Minimum	12
Rumus 2.11	Minimum Redundancy Maximum Relevance (MRMR)	12
Rumus 2.12	<i>Hyperplane SVM</i>	13
Rumus 2.13	kondisi pemisahan <i>hyperplane</i>	13
Rumus 2.14	<i>Soft margin SVM</i>	13
Rumus 2.15	margin SVM	14
Rumus 2.16	permasalahan optimisasi SVM	15
Rumus 2.17	bentuk dual SVM dengan metode Lagrange	16
Rumus 2.18	fungsi keputusan SVM metode dual	16
Rumus 2.19	pembaruan nilai bias metode dual	16
Rumus 2.20	fungsi kernel dalam SVM	16
Rumus 2.21	Fungsi keputusan SVM dengan kernel	17
Rumus 2.22	Kernel linear SVM	17
Rumus 2.23	Kernel Polinomial SVM	17
Rumus 2.24	Kernel RBF SVM	18
Rumus 2.25	Fungsi linier regresi logistik	18
Rumus 2.26	Fungsi sigmoid	18
Rumus 2.27	Probabilitas regresi logistik	19
Rumus 2.28	Fungsi <i>binary cross-entropy</i>	19
Rumus 2.29	Regularisasi L2	19
Rumus 2.31	Fungsi <i>gradient descent</i> regresi logistik	20
Rumus 2.33	kerugian <i>gradient descent</i>	20
Rumus 2.34	Metrik akurasi	21
Rumus 2.35	Metrik presisi	21
Rumus 2.36	Metrik <i>recall</i>	21
Rumus 2.37	Metrik F1-score	21
Rumus 2.38	Metrik AUC-ROC	22
Rumus 2.39	Metrik AUC	22

DAFTAR LAMPIRAN

Lampiran 1	Hasil Persentase Turnitin	68
Lampiran 2	Formulir Bimbingan	77



UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA