

ANALISIS PENGARUH PENGURANGAN FITUR BERKORELASI TINGGI SERTA RFE TERHADAP KLASIFIKASI KANKER PAYUDARA DENGAN SVM DAN RANDOM FOREST

ADINDA RACHMALIA



**PROGRAM STUDI MATEMATIKA
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2026**



PERNYATAAN MENGENAI SKRIPSI DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa skripsi dengan judul “Analisis Pengaruh Pengurangan Fitur Berkorelasi Tinggi serta RFE terhadap Klasifikasi Kanker Payudara dengan SVM dan *Random Forest*” adalah karya saya dengan arahan dari dosen pembimbing dan belum diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir skripsi ini.

Dengan ini saya melimpahkan hak cipta dari karya tulis saya kepada Institut Pertanian Bogor.

Bogor, Juli 2026

Adinda Rachmalia
G5401221005

- Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
 2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



ABSTRAK

ADINDA RACHMALIA. Analisis Pengaruh Pengurangan Fitur Berkorelasi Tinggi serta RFE terhadap Klasifikasi Kanker Payudara dengan SVM dan *Random Forest*. Dibimbing oleh MOCHAMAD TITO JULIANTO dan ELIS KHATIZAH.

Kanker payudara merupakan salah satu jenis kanker dengan tingkat kejadian yang tinggi pada perempuan sehingga diperlukan metode klasifikasi yang akurat untuk mendukung diagnosis dini. Penelitian ini bertujuan menganalisis pengaruh penghapusan fitur berkorelasi tinggi berdasarkan nilai *Variance Inflation Factor* (VIF) serta penggunaan *Recursive Feature Elimination* (RFE) terhadap klasifikasi kanker payudara menggunakan algoritma *Support Vector Machine* (SVM) dan *random forest*. *Dataset* yang digunakan adalah *Wisconsin Breast Cancer Diagnostic* dari *UCI Machine Learning Repository*. Penelitian dilakukan menggunakan empat skenario, yaitu *baseline*, penghapusan fitur berkorelasi tinggi berdasarkan nilai VIF, RFE dengan 15 fitur terpilih, serta kombinasi VIF dan RFE. Hasil penelitian menunjukkan bahwa kombinasi VIF dan RFE menghasilkan performa terbaik pada model SVM menggunakan 15 fitur terpilih, sedangkan model *random forest* menghasilkan performa terbaik pada skenario RFE tanpa VIF. Secara umum, penghapusan dan seleksi fitur mampu mengurangi jumlah fitur dari 30 menjadi 15 tanpa menurunkan performa klasifikasi secara signifikan.

Kata kunci: klasifikasi kanker payudara, *random forest*, *recursive feature elimination*, *support vector machine*, *variance inflation factor*

ABSTRACT

ADINDA RACHMALIA. Analysis of the Effect of Highly Correlated Feature Reduction and RFE on Breast Cancer Classification Using SVM and Random Forest. Supervised by MOCHAMAD TITO JULIANTO and ELIS KHATIZAH.

Breast cancer is one of the most common cancers among women, making accurate classification methods essential for supporting early diagnosis. This study aims to analyze the effect of removal of highly correlated features based on *Variance Inflation Factor* (VIF) values and the application of *Recursive Feature Elimination* (RFE) on breast cancer classification using *Support Vector Machine* (SVM) and *random forest* algorithms. The *dataset* used in this study was the *Wisconsin Breast Cancer Diagnostic* obtained from the *UCI Machine Learning Repository*. Four experimental scenarios were evaluated, namely the *baseline* scenario, removal of highly correlated features based on VIF values, RFE with 15 selected features, and a combination of VIF and RFE. The results showed that the combination of VIF and RFE produced the best performance for the SVM model using 15 selected features, whereas the *random forest* model achieved its best performance under the RFE-only scenario. Overall, feature elimination and feature selection reduced the number of features from 30 to 15 without significantly decreasing classification performance.

Keywords: breast cancer classification, random forest, recursive feature elimination, support vector machine, variance inflation factor.



Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

© Hak Cipta milik IPB, tahun 2026
Hak Cipta dilindungi Undang-Undang

Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan atau menyebutkan sumbernya. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik, atau tinjauan suatu masalah, dan pengutipan tersebut tidak merugikan kepentingan IPB.

Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apa pun tanpa izin IPB.

ANALISIS PENGARUH PENGURANGAN FITUR BERKORELASI TINGGI SERTA RFE TERHADAP KLASIFIKASI KANKER PAYUDARA DENGAN SVM DAN RANDOM FOREST

ADINDA RACHMALIA

Skripsi
sebagai salah satu syarat untuk memperoleh gelar
Sarjana pada
Program Studi Matematika

**PROGRAM STUDI MATEMATIKA
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2026**



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



Judul Skripsi : Analisis Pengaruh Pengurangan Fitur Berkorelasi Tinggi serta RFE terhadap Klasifikasi Kanker Payudara dengan SVM dan *Random Forest*

Nama : Adinda Rachmalia

NIM : G5401221005

Disetujui oleh

Pembimbing 1:

Mochamad Tito Julianto, S.Si., M.Kom.

Pembimbing 2:

Elis Khatizah, S.Si., M.Si., Ph.D.

Diketahui oleh

Ketua Program Studi:

Dr. Donny Citra Lesmana, S.Si., M.Fin.Math.

NIP 197902272005011001

Tanggal Ujian:

24 Juni 2026

Tanggal Lulus:



PRAKATA

Puji dan syukur penulis panjatkan kepada Allah SWT atas segala karunia-Nya sehingga karya ilmiah ini berhasil diselesaikan. Tema yang dipilih dalam penelitian yang dilaksanakan sejak bulan Oktober 2025 sampai bulan Juni 2026 ini ialah klasifikasi keganasan kanker payudara menggunakan metode *machine learning*, dengan judul “Analisis Pengaruh Pengurangan Fitur Berkorelasi Tinggi serta RFE terhadap Klasifikasi Kanker Payudara dengan SVM dan *Random Forest*”. Ucapan terima kasih penulis ucapkan kepada berbagai pihak, antara lain:

1. Kedua orang tua penulis Bapak Demi Abdullah dan Ibu Ipah Latifah, serta kakak penulis Aulia Defanti yang senantiasa memberikan doa, dukungan, kasih sayang, semangat dan motivasi kepada penulis selama proses perkuliahan hingga penyusunan skripsi ini.

2. Bapak Mochamad Tito Julianto, S.Si., M.Kom. dan Ibu Elis Khatizah, S.Si., M.Si., Ph.D. selaku dosen pembimbing penulis, serta Prof. Dr. Ir. Sri Nurdianti, M.Sc. selaku dosen penguji yang telah meluangkan waktu, memberikan arahan, masukan, serta bimbingan kepada penulis sehingga skripsi ini dapat diselesaikan dengan baik.

3. Bapak Prof. Dr. Ir. Bib Paruhum Silalahi, M.Kom. selaku dosen penggerak akademik penulis yang telah memberikan bantuan, arahan, dan dukungan kepada penulis selama menempuh pendidikan di Program Studi Matematika IPB.

4. Seluruh dosen dan tenaga kependidikan Program Studi Matematika IPB yang telah memberikan ilmu, pengalaman, bantuan, serta pelayanan kepada penulis selama masa perkuliahan.

5. Ghonniyu Hiban Saputra, Nida Hanaa, Haifa Shabira, Sarah Firjani, Tiara Markhairina, Fadhila Hasna, Syarifah Nafisah serta Jodi At-Takbir yang selalu memberikan dukungan, semangat, doa, serta menemani penulis dalam berbagai proses hingga skripsi ini dapat diselesaikan.

6. Ika Nurjannah, Hana Nadifa, Risma Rustanti, Najwaa Alifya, Nadhira Maulida, Arnida Riani, Alfadhiyya Cheryl, Mirlan Sujana, Alief Nur Widiyanto serta teman-teman Matematika 59 yang telah memberikan bantuan, kebersamaan, dukungan, dan pengalaman yang sangat baik selama menempuh pendidikan di IPB University.

7. Keluarga besar KSR PMI IPB, Medbrenk, dan GiziQu yang telah menjadi tempat bertumbuh, belajar, berbagi pengalaman, serta memberikan dukungan dan kebersamaan yang berharga bagi penulis selama masa perkuliahan.

Semoga karya ilmiah ini bermanfaat bagi pihak yang membutuhkan dan bagi kemajuan ilmu pengetahuan.

Bogor, Juli 2026

Adinda Rachmalia



Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

DAFTAR ISI

DAFTAR TABEL	viii
DAFTAR GAMBAR	viii
DAFTAR LAMPIRAN	ix
I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	2
1.3 Tujuan	2
II TINJAUAN PUSTAKA	3
2.1 Penelitian Terdahulu	3
2.2 Kanker Payudara	3
2.3 <i>Machine Learning</i>	4
2.4 Multikolinearitas	4
2.5 <i>Outlier</i>	6
2.6 <i>Standard Scaler</i>	6
2.7 <i>Support Vector Machine (SVM)</i>	7
2.8 <i>Random Forest</i>	10
2.9 <i>Recursive Feature Elimination (RFE)</i>	12
2.10 Evaluasi Model	13
III METODE	16
3.1 Bahan dan Data	16
3.2 Tahapan Penelitian	17
IV HASIL DAN PEMBAHASAN	21
4.1 Pra-pemrosesan Data	21
4.2 <i>Exploratory Data Analysis</i>	23
4.3 Penghapusan Fitur Berkorelasi Tinggi	26
4.4 Pembagian Data	27
4.5 Transformasi Data	27
4.6 Pembangunan Model	28
4.7 Perbandingan Model	36
V SIMPULAN DAN SARAN	38
5.1 Simpulan	38
5.2 Saran	38
DAFTAR PUSTAKA	39
LAMPIRAN	42
RIWAYAT HIDUP	59



DAFTAR TABEL

1	Deskripsi Fitur pada <i>Dataset</i> WBCD	16
2	Ringkasan Skenario Pengujian	18
3	Pasangan Fitur dengan Korelasi Tinggi	24
4	Daftar Fitur Setelah Eliminasi Fitur dengan $VIF > 10$	26
5	Hasil Pembagian Data Latih dan Data Uji	27
6	Hasil Evaluasi Skenario 1	29
7	Hasil Evaluasi Skenario 2	30
8	Hasil Evaluasi Skenario 3	32
9	Fitur Terpilih Skenario 3	33
10	Hasil Evaluasi Skenario 4	34
11	Fitur Terpilih Skenario 4	35
12	Perbandingan Performa Model Klasifikasi pada Seluruh Skenario	36

DAFTAR GAMBAR

1	Ilustrasi <i>Support Vector Machine</i> (SVM)	8
2	Ilustrasi <i>Random Forest</i>	11
3	Ilustrasi <i>Recursive Feature Elimination</i> (RFE)	13
4	<i>Confusion Matrix</i> Klasifikasi	14
5	Diagram Alir Penelitian	17
6	<i>Boxplot</i> Fitur <i>Mean</i>	21
7	<i>Boxplot</i> Fitur <i>se (standard error)</i>	22
8	<i>Boxplot</i> Fitur <i>Worst</i>	22
9	Distribusi Kelas Jinak vs Ganas	23
10	<i>Heatmap</i> Korelasi Antarfitur	24
11	Visualisasi Nilai <i>VIF</i> pada Setiap Fitur	25
12	Visualisasi nilai <i>VIF</i> setelah Eliminasi Fitur Berkorelasi Tinggi	26
13	Distribusi Kelas pada (A) Data Latih dan (B) Data Uji	27
14	Visualisasi Distribusi Data pada Setiap Fitur	28
15	Visualisasi Distribusi Data setelah Penskalaan dengan <i>Standard Scaler</i>	28
16	<i>Confusion Matrix</i> Skenario 1 dengan Model (A) <i>Support Vector Machine</i> dan (B) <i>Random Forest</i>	29
17	Kurva ROC Skenario 1	30
18	<i>Confusion Matrix</i> Skenario 2 dengan Model (A) <i>Support Vector Machine</i> dan (B) <i>Random Forest</i>	31
19	Kurva ROC Skenario 2	31
20	<i>Confusion Matrix</i> Skenario 3 dengan Model (A) <i>Support Vector Machine</i> dan (B) <i>Random Forest</i>	32
21	Kurva ROC Skenario 3	33
22	<i>Confusion Matrix</i> Skenario 4 dengan Model (A) <i>Support Vector Machine</i> dan (B) <i>Random Forest</i>	34
23	Kurva ROC Skenario 4	35

DAFTAR LAMPIRAN

1	<i>Dataset WBCD</i>	43
2	<i>Syntax Import Library</i>	43
3	<i>Syntax Import Dataset</i>	44
4	<i>Syntax Deteksi Missing Value</i>	44
5	<i>Syntax Label Encoding</i>	44
6	<i>Syntax Analisis Korelasi Pearson</i>	44
7	<i>Syntax Deteksi Multikolinearitas</i>	44
8	<i>Syntax Penghapusan fitur VIF > 10</i>	45
9	<i>Hasil Iterasi Eliminasi Fitur Menggunakan VIF</i>	45
10	<i>Syntax Fungsi Bantu Evaluasi Model dan Visualisasi</i>	52
11	<i>Syntax Split Data dan Standardisasi</i>	54
12	<i>Syntax Skenario 1</i>	55
13	<i>Syntax Skenario 2</i>	56
14	<i>Syntax Skenario 3</i>	56
15	<i>Syntax Skenario 4</i>	57