



**KOMPARASI TEKNIK IMPUTASI UNTUK NILAI NOL DAN  
PARADIGMA SELEKSI FITUR (*GA* DAN *RFE-SHAP*) UNTUK  
KLASIFIKASI PENYAKIT DIABETES MENGGUNAKAN *RANDOM  
FOREST***

**Skripsi**

**Untuk Memenuhi Persyaratan  
Dalam Menyelesaikan Strata-1 Ilmu Komputer**

**Oleh  
MUHAMMAD HENRY ADITYA  
NIM 2211016210016**

**PROGRAM STUDI S-1 ILMU KOMPUTER  
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM  
UNIVERSITAS LAMBUNG MANGKURAT  
BANJARBARU  
FEBRUARI 2026**



**KOMPARASI TEKNIK IMPUTASI UNTUK NILAI NOL DAN  
PARADIGMA SELEKSI FITUR (*GA* DAN *RFE-SHAP*) UNTUK  
KLASIFIKASI PENYAKIT DIABETES MENGGUNAKAN *RANDOM  
FOREST***

**Skripsi**

**Untuk Memenuhi Persyaratan  
Dalam Menyelesaikan Strata-1 Ilmu Komputer**

**Oleh**

**MUHAMMAD HENRY ADITYA**

**NIM 2211016210016**

**PROGRAM STUDI S-1 ILMU KOMPUTER  
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM  
UNIVERSITAS LAMBUNG MANGKURAT  
BANJARBARU  
FEBRUARI 2026**

# SKRIPSI

## KOMPARASI TEKNIK IMPUTASI UNTUK NILAI NOL DAN PARADIGMA SELEKSI FITUR (*GA* DAN *RFE-SHAP*) UNTUK KLASIFIKASI PENYAKIT DIABETES MENGGUNAKAN *RANDOM FOREST*

Oleh:

**MUHAMMAD HENRY ADITYA**

**NIM 2211016210016**

Telah dipertahankan di depan Dosen Penguji pada tanggal 25 Februari 2026.

Susunan Dosen Penguji:

**Pembimbing I**



Muhammad Itqan Mazdadi, S.Kom., M.Kom.  
NIP. 199006122019031013

**Dosen Penguji I**



Triando Hamonangan Saragih, S.Kom., M.Kom.  
NIP. 199308242019031012

**Pembimbing II**



Muliadi, S.Kom., M.Cs.  
NIP. 197804222010121002

**Dosen Penguji II**



Fatma Indriani, S.T., M.I.T., Ph.D.  
NIP. 198404202008122004

Banjarbaru, 26 Februari 2026

**Koordinator Program Studi Ilmu Komputer**



M. Kartini, S.Kom., M.Kom.  
NIP. 1987042120121220003

## PERNYATAAN

Dengan ini saya menyatakan bahwa dalam skripsi ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Banjarbaru, 25 Februari 2026

Yang Menyatakan,



Muhammad Henry Aditya

NIM. 2211016210016

## ABSTRAK

### KOMPARASI TEKNIK IMPUTASI UNTUK NILAI NOL DAN PARADIGMA SELEKSI FITUR (GA DAN RFE-SHAP) UNTUK KLASIFIKASI PENYAKIT DIABETES MENGGUNAKAN RANDOM FOREST

(Oleh: Muhammad Henry Aditya; Pembimbing: Muhammad Itqan Mazdadi, S.Kom., M.Kom. dan Muliadi, S.Kom., M.Cs.; 2026; 78 halaman)

Diabetes salah satu penyakit kronis yang menjadi tantangan kesehatan global dan memerlukan deteksi dini yang akurat serta andal. Pemanfaatan *machine learning*, khususnya menggunakan *dataset Pima Indians Diabetes*, telah banyak dilakukan untuk mendukung klasifikasi penyakit ini. Namun, *dataset* tersebut memiliki permasalahan kualitas data berupa nilai nol yang tidak valid secara medis. Selain itu, perbedaan teknik imputasi dan paradigma seleksi fitur berpotensi menghasilkan variasi performa sekaligus memengaruhi keterbacaan model. Penelitian ini bertujuan mengevaluasi secara komparatif pengaruh kombinasi teknik imputasi (*mean*, KNN, dan MICE) dan seleksi fitur dengan paradigma berbeda, yaitu *Genetic Algorithm* (GA) dan RFE-SHAP, terhadap performa dan keterbacaan model *Random Forest*. Hasil penelitian menunjukkan bahwa kombinasi MICE–GA menghasilkan performa paling tinggi dan efisien pada akurasi (0,7727), *precision* (0,6610), dan *F1-score* (0,6903), *mean*–RFE-SHAP unggul pada *recall* (0,7407) dengan jumlah fitur lebih sedikit, sedangkan *mean*–GA memperoleh AUC tertinggi (0,8406), di mana seluruh kombinasi mampu menghasilkan  $AUC > 0,8$  dan RFE-SHAP mengidentifikasi *Glucose* sebagai fitur paling dominan diikuti BMI, *Age*, dan Insulin berdasarkan nilai kepentingan fitur (*SHAP importance*) berada pada rentang 0,0740 hingga 0,1733 serta arah kontribusi (*directional contribution*) pada rentang  $-0,0028$  hingga  $+0,0031$ .

**Kata Kunci:** Diabetes, Imputasi Data, Seleksi Fitur, Random Forest, RFE-SHAP

## **ABSTRACT**

### **COMPARISON OF IMPUTATION TECHNIQUES FOR ZERO VALUES AND FEATURE SELECTION PARADIGM (GA AND RFE-SHAP) FOR DIABETES DISEASE CLASSIFICATION USING RANDOM FOREST**

*(By: Muhammad Henry Aditya; Supervisor: Muhammad Itqan Mazdaadi, S.Kom., M.Kom. & Muliadi, S.Kom., M.Cs.; 2026; 78 pages)*

*Diabetes is a chronic disease that poses a global health challenge and requires accurate and reliable early detection. The use of machine learning, particularly using the Pima Indians Diabetes dataset, has been widely used to support the classification of this disease. However, this dataset has data quality issues in the form of medically invalid null values. Furthermore, differences in imputation techniques and feature selection paradigms have the potential to produce performance variations and affect model readability. This study aims to comparatively evaluate the effect of a combination of imputation techniques (mean, KNN, and MICE) and feature selection with different paradigms, namely the Genetic Algorithm (GA) and RFE-SHAP, on the performance and readability of the Random Forest model. The results showed that the combination of MICE–GA produced the highest and most efficient performance in accuracy (0,7727), precision (0,6610), and F1-score (0,6903), mean–RFE-SHAP excelled in recall (0,7407) with fewer features, while mean–GA obtained the highest AUC (0,8406), where all combinations were able to produce  $AUC > 0.8$  and RFE-SHAP identified Glucose as the most dominant feature followed by BMI, Age, and Insulin based on feature importance values (SHAP importance) in the range of 0,0740 to 0,1733 and directional contribution in the range of  $-0,0028$  to  $+0,0031$ .*

**Keywords:** *Diabetes, Data Imputation, Feature Selection, Random Forest, RFE-SHAP*

## PRAKATA

Puji syukur penulis haturkan kepada Tuhan Yang Maha Esa, Allah SWT, karena atas berkat, rahmat, dan karunia-Nya penulis dapat menyelesaikan skripsi yang berjudul “Komparasi Teknik Imputasi untuk Nilai Nol dan Paradigma Seleksi Fitur (*GA* dan *RFE-SHAP*) untuk Klasifikasi Penyakit Diabetes Menggunakan *Random Forest*”.

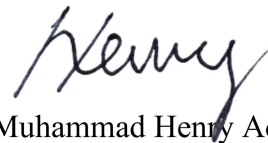
Melalui lembar ini, penulis ingin menyampaikan ucapan terima kasih sebesar-besarnya kepada berbagai pihak yang sangat mendukung penulis dalam pembuatan dan penyusunan skripsi sehingga pada akhirnya skripsi ini dapat diselesaikan, adapun pihak-pihak yang dimaksud adalah sebagai berikut.

1. Keluarga besar, terutama Abah, Mama, Wina, Kai, Nini, Tante Ati, Bapak H. M. Isra Ismail, S.H., M.H., Ibu Hj. Hairiyati, S.H. dan sekeluarga, Ibu Ir. Hj. Ernawati dan sekeluarga, Ibu Hairiyah dan sekeluarga, Bapak Dr.Eng Ir. Irfan Prasetya, S.T., M.T. dan sekeluarga, yang selalu memberikan doa, semangat, dukungan, hingga kepercayaan, sehingga penulis termotivasi untuk mengejar pendidikan setinggi-tingginya serta menjadi bahan bakar semangat dalam menyelesaikan skripsi ini sebaik mungkin.
2. Bapak Muhammad Itqan Mazdadi, S.Kom., M.Kom. selaku dosen pembimbing utama serta Bapak Muliadi, S.Kom., M.Cs. selaku dosen pembimbing pendamping dan dosen pembimbing akademik yang turut serta membantu dan meluangkan waktu demi kelancaran dalam penyelesaian skripsi ini.
3. Ibu Dwi Kartini, S.Kom., M.Kom. selaku Koordinator Program Studi Ilmu Komputer FMIPA ULM, atas bantuan dan izin beliau skripsi ini dapat diselesaikan.
4. Seluruh dosen dan staf Program Studi Ilmu Komputer FMIPA ULM atas ilmu dan bantuan yang diberikan selama ini yang sangat bermanfaat.
5. Teman-teman seperjuangan, keluarga besar NVIDIA'22, yang telah menjadi bagian atas perjuangan kuliah selama ini.

6. Semua pihak yang tidak dapat disebutkan satu persatu yang telah turut membantu dalam penyelesaian skripsi ini.

Akhir kata penulis menyadari sepenuhnya bahwa penulisan ini jauh dari sempurna dan penulis mengharapkan bantuan berupa kritik dan saran yang membangun dari semua pihak demi kesempurnaan dan mutu penulisan skripsi ini. Semoga tulisan ini dapat bermanfaat bagi ilmu pengetahuan dan pembaca khususnya serta mendapat keridhaan dari Tuhan Yang Maha Esa, Allah SWT.

Banjarbaru, 25 Februari 2026



Muhammad Henry Aditya

## DAFTAR ISI

HALAMAN JUDUL.....	i
HALAMAN PENGESAHAN.....	ii
PERNYATAAN.....	iii
ABSTRAK.....	iv
ABSTRACT.....	v
PRAKATA.....	vi
DAFTAR ISI.....	viii
DAFTAR TABEL.....	xi
DAFTAR GAMBAR.....	xiii
DAFTAR LAMPIRAN.....	xvi
<b>BAB I PENDAHULUAN</b>	
1.1. Latar Belakang.....	1
1.2. Rumusan Masalah.....	3
1.3. Tujuan Penelitian.....	4
1.4. Manfaat Penelitian.....	4
1.5. Batasan Masalah.....	5
<b>BAB II TINJAUAN PUSTAKA</b>	
2.1. Kajian Terdahulu.....	6
2.2. Landasan Teori.....	13
2.2.1 Diabetes Melitus.....	13
2.2.2 <i>Exploratory Data Analysis (EDA)</i> .....	13
2.2.3 Pra-pemrosesan Data.....	14
2.2.3.1 Pembagian Data.....	14

2.2.3.2	Imputasi Data .....	15
2.2.3.2.1	Imputasi Mean.....	15
2.2.3.2.2	Imputasi K-Nearest Neighbors (KNN).....	15
2.2.3.2.3	Imputasi Multivariate Imputation by Chained Equations (MICE) .....	17
2.2.3.3	Seleksi Fitur.....	18
2.2.3.3.1	Algoritma Genetika .....	18
2.2.3.3.2	RFE-SHAP .....	20
2.2.3.4	Penanganan <i>Imbalanced Data</i> dengan SMOTE.....	22
2.2.4	<i>Random Forest</i> .....	23
2.2.5	Evaluasi Model.....	25
2.2.6	Evaluasi Keterbacaan (Interpretabilitas) .....	27

### **BAB III METODE PENELITIAN**

3.1.	Alat Penelitian.....	29
3.2.	Bahan Penelitian .....	29
3.3.	Variabel Penelitian.....	29
3.4.	Prosedur Penelitian .....	30

### **BAB IV HASIL DAN PEMBAHASAN**

4.1.	Hasil.....	34
4.1.1	Pengumpulan <i>Dataset</i> .....	34
4.1.2	<i>Exploratory Data Analysis (EDA)</i> .....	36
4.1.3	<i>Preprocessing Data</i> .....	39
4.1.3.1	Konversi Nilai Nol ke NaN ( <i>Not a Number</i> ).....	39
4.1.4	Pembagian Data.....	40
4.1.5	Imputasi Data.....	40

4.1.6	Seleksi Fitur.....	42
4.1.7	Klasifikasi dengan <i>Random Forest</i> .....	43
4.2.	Pembahasan .....	51

## **BAB V PENUTUP**

5.1.	Kesimpulan.....	72
5.2.	Saran .....	72

## DAFTAR PUSTAKA

## LAMPIRAN

## DAFTAR TABEL

<b>Tabel</b>		<b>Halaman</b>
Tabel 1	Keaslian Penelitian.....	9
Tabel 2	Perancangan Penelitian .....	12
Tabel 3	Kategorisasi Nilai AUC .....	26
Tabel 4	Nama-nama fitur <i>dataset</i> dan deskripsinya.....	34
Tabel 5	Contoh <i>Pima Indians Diabetes Dataset</i> .....	35
Tabel 6	Jumlah distribusi kelas pada <i>dataset</i> .....	35
Tabel 7	Distribusi frekuensi nilai nol pada setiap fitur .....	39
Tabel 8	Parameter Imputasi KNN.....	40
Tabel 9	Parameter Imputasi MICE .....	40
Tabel 10	Parameter <i>genetic algorithm</i> .....	42
Tabel 11	Parameter RFE-SHAP.....	43
Tabel 12	Parameter <i>Random Forest</i> .....	44
Tabel 13	Evaluasi performa model pada data <i>testing</i> untuk kombinasi imputasi <i>mean</i> dan seleksi fitur <i>Genetic Algorithm (GA)</i> .....	46
Tabel 14	Evaluasi performa model pada data <i>testing</i> untuk kombinasi imputasi <i>K-Nearest Neighbors (KNN)</i> dan seleksi fitur <i>Genetic Algorithm (GA)</i> .....	46
Tabel 15	Evaluasi performa model pada data <i>testing</i> untuk kombinasi imputasi <i>Multivariate Imputation by Chained Equations (MICE)</i> dan seleksi fitur <i>Genetic Algorithm (GA)</i> .....	46
Tabel 16	Evaluasi performa model pada data <i>testing</i> untuk kombinasi imputasi <i>mean</i> dan seleksi fitur <i>Recursive Feature Elimination with SHAP (RFE-SHAP)</i> 47	
Tabel 17	Evaluasi performa model pada data <i>testing</i> untuk kombinasi imputasi <i>K-Nearest Neighbors (KNN)</i> dan seleksi fitur <i>Recursive Feature Elimination with SHAP (RFE-SHAP)</i> .....	47
Tabel 18	Evaluasi performa model pada data <i>testing</i> untuk kombinasi imputasi <i>Multivariate Imputation by Chained Equations (MICE)</i> dan seleksi fitur <i>Recursive Feature Elimination with SHAP (RFE-SHAP)</i> .....	48

Tabel 19	Hasil keterbacaan fitur dengan seleksi fitur RFE-SHAP dengan kombinasi imputasi <i>mean</i> .....	48
Tabel 20	Hasil keterbacaan fitur dengan seleksi fitur RFE-SHAP dengan kombinasi imputasi <i>K-Nearest Neighbors</i> (KNN).....	49
Tabel 21	Hasil keterbacaan fitur dengan seleksi fitur RFE-SHAP dengan kombinasi imputasi <i>Multivariate Imputation by Chained Equations</i> (MICE).....	50
Tabel 22	Set fitur yang terpilih dengan seleksi fitur GA dan RFE-SHAP ....	67

## DAFTAR GAMBAR

<b>Gambar</b>	<b>Halaman</b>
Gambar 1	Ilustrasi tahapan <i>crossover</i> (Huang & Wang, 2006)..... 19
Gambar 2	Ilustrasi tahapan <i>mutation</i> (Huang & Wang, 2006) ..... 19
Gambar 3	Siklus algoritma genetika (Huang & Wang, 2006)..... 20
Gambar 4	<i>Confusion matrix</i> dengan dua kelas (Widya Astuti <i>et al.</i> , 2020) ..... 25
Gambar 5	Diagram Alur Penelitian..... 31
Gambar 6	Perbandingan persentase penderita diabetes dan bukan penderita diabetes..... 36
Gambar 7	Distribusi data untuk fitur <i>Pregnancies</i> ..... 37
Gambar 8	Distribusi data untuk fitur <i>Glucose</i> ..... 37
Gambar 9	Distribusi data untuk fitur <i>BloodPressure</i> ..... 37
Gambar 10	Distribusi data untuk fitur <i>SkinThickness</i> ..... 38
Gambar 11	Distribusi data untuk fitur <i>Insulin</i> ..... 38
Gambar 12	Distribusi data untuk fitur BMI..... 38
Gambar 13	Distribusi data untuk fitur <i>DiabetesPedigreeFunction</i> ..... 39
Gambar 14	Distribusi data untuk fitur Age..... 39
Gambar 15	Perbandingan jumlah kelas pada data latih untuk lipatan pertama ... 44
Gambar 16	Grafik hasil performa model <i>Random Forest</i> berdasarkan <i>accuracy</i> untuk metode imputasi <i>mean</i> dan kedua jenis teknik seleksi fitur..... 54
Gambar 17	Grafik hasil performa model <i>Random Forest</i> berdasarkan <i>precision</i> untuk metode imputasi <i>mean</i> dan kedua jenis teknik seleksi fitur..... 54
Gambar 18	Grafik hasil performa model <i>Random Forest</i> berdasarkan <i>recall</i> untuk metode imputasi <i>mean</i> dan kedua jenis teknik seleksi fitur..... 55
Gambar 19	Grafik hasil performa model <i>Random Forest</i> berdasarkan <i>F1-score</i> untuk metode imputasi <i>mean</i> dan kedua jenis teknik seleksi fitur..... 55
Gambar 20	Grafik hasil performa model <i>Random Forest</i> berdasarkan AUC untuk metode imputasi <i>mean</i> dan kedua jenis teknik seleksi fitur..... 56

Gambar 21	Grafik hasil performa model <i>Random Forest</i> berdasarkan <i>accuracy</i> untuk metode imputasi KNN dan kedua jenis teknik seleksi fitur.....	57
Gambar 22	Grafik hasil performa model <i>Random Forest</i> berdasarkan <i>precision</i> untuk metode imputasi KNN dan kedua jenis teknik seleksi fitur.....	58
Gambar 23	Grafik hasil performa model <i>Random Forest</i> berdasarkan <i>recall</i> untuk metode imputasi KNN dan kedua jenis teknik seleksi fitur.....	58
Gambar 24	Grafik hasil performa model <i>Random Forest</i> berdasarkan <i>F1-score</i> untuk metode imputasi KNN dan kedua jenis teknik seleksi fitur.....	58
Gambar 25	Grafik hasil performa model <i>Random Forest</i> berdasarkan AUC untuk metode imputasi KNN dan kedua jenis teknik seleksi fitur.....	59
Gambar 26	Grafik hasil performa model <i>Random Forest</i> berdasarkan <i>accuracy</i> untuk metode imputasi MICE dan kedua jenis teknik seleksi fitur .....	60
Gambar 27	Grafik hasil performa model <i>Random Forest</i> berdasarkan <i>precision</i> untuk metode imputasi MICE dan kedua jenis teknik seleksi fitur .....	61
Gambar 28	Grafik hasil performa model <i>Random Forest</i> berdasarkan <i>recall</i> untuk metode imputasi MICE dan kedua jenis teknik seleksi fitur.....	61
Gambar 29	Grafik hasil performa model <i>Random Forest</i> berdasarkan <i>F1-score</i> untuk metode imputasi MICE dan kedua jenis teknik seleksi fitur .....	62
Gambar 30	Grafik hasil performa model <i>Random Forest</i> berdasarkan AUC untuk metode imputasi MICE dan kedua jenis teknik seleksi fitur.....	62
Gambar 31	Grafik perbandingan hasil performa model <i>Random Forest</i> berdasarkan <i>accuracy</i> .....	64
Gambar 32	Grafik perbandingan hasil performa model <i>Random Forest</i> berdasarkan <i>precision</i> .....	64
Gambar 33	Grafik perbandingan hasil performa model <i>Random Forest</i> berdasarkan <i>recall</i> .....	65
Gambar 34	Grafik perbandingan hasil performa model <i>Random Forest</i> berdasarkan <i>F1-score</i> .....	65
Gambar 35	Grafik perbandingan hasil performa model <i>Random Forest</i> berdasarkan AUC .....	66

Gambar 36	Hasil <i>beeswarm plot</i> untuk keseluruhan fitur dengan imputasi <i>mean</i> .....	68
Gambar 37	Hasil <i>beeswarm plot</i> untuk keseluruhan fitur dengan imputasi KNN .....	69
Gambar 38	Hasil <i>beeswarm plot</i> untuk keseluruhan fitur dengan imputasi MICE.....	70

## DAFTAR LAMPIRAN

### Lampiran

Lampiran 1. Sumber kode untuk impor *library*

Lampiran 2. Sumber kode untuk memuat data, konversi nol, dan *data split*

Lampiran 3. Sumber kode untuk visualisasi data sebelum proses imputasi

Lampiran 4. Sumber kode untuk imputasi data

Lampiran 5. Sumber kode untuk visualisasi pengaruh SMOTE

Lampiran 6. Sumber kode untuk evaluasi pipeline dan seleksi fitur

Lampiran 7. Sumber kode untuk data *loading*

Lampiran 8. Sumber kode untuk *main program*