

**OPTIMALISASI *PREPROCESSING DATA* MENGGUNAKAN
PENDEKATAN *CRISP-DM* UNTUK MENINGKATKAN
KUALITAS KLASIFIKASI PENYAKIT JANTUNG**



SKRIPSI

Diajukan sebagai syarat meraih gelar Sarjana Komputer

Oleh:

WAHYU NURSAHID

NIM: 21195023

**PROGRAM STUDI TEKNIK INFORMATIKA
SEKOLAH TINGGI MANAJEMEN INFORMATIKA & KOMPUTER
(STMIK YMI TEGAL)
TEGAL
2025**

HALAMAN PERSETUJUAN

Pembimbing Skripsi memberikan rekomendasi kepada:

Nama : Wahyu Nursahid

NIM : 21195023

Program Studi : Teknik Informatika

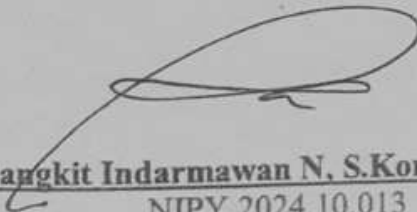
Judul Skripsi : OPTIMALISASI *PREPROCESSING* DATA MENGGUNAKAN
PENDEKATAN *CRISP-DM* UNTUK MENINGKATKAN
KUALITAS KUALIFIKASI PENYAKIT JANTUNG

Mahasiswa tersebut telah dinyatakan selesai melaksanakan bimbingan dan dapat mengikuti Ujian Skripsi pada tahun akademik 2024/2025.

Tegal, 17 Juli 2025

Pembimbing 1

Pembimbing 2



Bangkit Indarmawan N, S.Kom., M.Kom
NIPY.2024.10.013



Syefudin, S.Kom., M.Pd., M.Kom
NIPY.2024.10.010

HALAMAN PENGESAHAN

Nama : Wahyu Nursahid

NIM : 21195023

Program Studi : Teknik Informatika

Judul Skripsi : OPTIMALISASI *PREPROCESSING* DATA MENGGUNAKAN
PENDEKATAN *CRISP-DM* UNTUK MENINGKATKAN
KUALITAS KUALIFIKASI PENYAKIT JANTUNG

Dinyatakan LULUS setelah dipertahankan di hadapan Dewan Penguji Skripsi
Program Studi Teknik Informatika STMIK YMI Tegal.

Tegal, 24 Juli 2025

Dewan Penguji

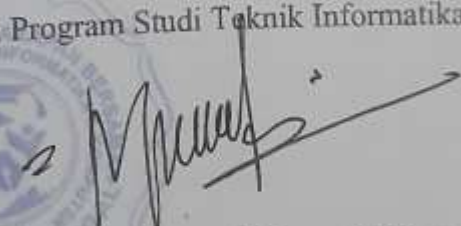
1. Ketua : Erni Unggul SU, S.E., M.Si
2. Anggota I : Sarif Surejo, S.E., M.Kom
3. Anggota II : Nugroho Adhi S, S.Kom., M.Kom

Tanda Tangan

1. 
2. 
3. 

Mengetahui

Ketua Program Studi Teknik Informatika



Aang Alim Murtopo, S.Kom., M.Kom
NIPY.2024.10.0002

LEMBAR PERNYATAAN ORISINALITAS

Saya yang bertanda tangan di bawah ini:

Nama : Wahyu Nursahid
NIM : 21195023
Tempat, Tanggal Lahir : Klaten, 10 Oktober 1990
Alamat : Jl. K. H. Ahmad Dahlan no.37 Rt. 002 / Rw. 009
Pasar batang, Kec. Brebes, Kab. Brebes

Dengan ini menyatakan bahwa Skripsi dengan judul **Optimalisasi Preprocessing Data Menggunakan Pendekatan CRISP-DM Untuk Meningkatkan Kualitas Klasifikasi Penyakit Jantung** adalah hasil penelitian saya bersama Dosen – Dosen Pembimbing saya dengan seluruh referensi, pernyataan atau materi dari sumber lain telah dikutip dengan cara penulisan referensi yang sesuai.

Pernyataan ini dibuat dengan sebenar-benarnya dan jika pernyataan ini tidak sesuai dengan kenyataan, maka saya bersedia menanggung sanksi yang akan dikenakan kepada saya termasuk pencabutan gelar Sarjana Komputer (S.Kom) yang telah saya dapatkan.

Tegal, 17 Juli 2025



Wahyu Nursahid

ABSTRAK

Peningkatan kualitas data merupakan langkah krusial dalam sistem klasifikasi berbasis *machine learning*, khususnya dalam kasus deteksi penyakit jantung. Penelitian ini bertujuan untuk menyusun dan menguji beberapa skema *preprocessing* data secara terstruktur dalam kerangka metodologi *CRISP-DM* (Cross Industry Standard Process for *Data mining*), serta mengevaluasi pengaruhnya terhadap performa klasifikasi menggunakan algoritma *K-Nearest Neighbors (KNN)*. Empat skema *preprocessing* dirancang secara bertahap dengan kombinasi tahapan *Cleaning*, *Transformation*, dan *Reduction*, termasuk penggunaan *Normalisasi*, *balancing* dengan *SMOTE*, dan seleksi *fitur* menggunakan *information Gain*. Evaluasi dilakukan menggunakan teknik *10-fold Cross-validation* dengan metrik akurasi, presisi, *Recall*, *F1-score*, dan *AUC*. Hasil penelitian menunjukkan bahwa variasi skema *preprocessing* memberikan dampak signifikan terhadap performa model. Skema paling lengkap, yang menggabungkan *Normalisasi*, *SMOTE*, dan seleksi fitur, menghasilkan performa terbaik dengan akurasi sebesar 81,26%, *Precision* 83,36%, *Recall* 83,25%, *F1-score* 83,00%, dan *AUC* 0,8460. Penelitian ini memberikan kontribusi teoritis dalam memperkuat pemahaman pentingnya kualitas data dalam siklus *data mining*, serta kontribusi praktis berupa panduan eksplisit dalam menyusun skema *preprocessing* yang dapat direplikasi oleh peneliti maupun praktisi. Temuan ini juga dapat menjadi acuan dalam pengembangan sistem klasifikasi penyakit yang lebih andal untuk mendukung diagnosis dini penyakit jantung dan pengambilan keputusan klinis secara lebih tepat.

Kata Kunci: *Preprocessing, CRISP-DM, KNN, Klasifikasi, Penyakit Jantung, Data Preparation*

ABSTRACT

Improving data quality is a crucial step in machine learning-based classification systems, especially in the case of heart disease detection. This study aims to design and evaluate several structured preprocessing schemes within the CRISP-DM (Cross Industry Standard Process for Data mining) framework, and to analyze their impact on classification performance using the K-Nearest Neighbors (KNN) algorithm. Four preprocessing schemes were developed incrementally, combining stages of Cleaning, Transformation, and Reduction, including normalization, class balancing with SMOTE, and Feature selection using information Gain. Evaluation was conducted using 10-fold Cross-validation with Accuracy, Precision, Recall, F1-score, and AUC as the performance metrics. The results show that the variation in preprocessing schemes significantly affects model performance. The most comprehensive scheme combining normalization, SMOTE, and Feature selection achieved the best performance with an Accuracy of 81.26%, Precision of 83.36%, Recall of 83.25%, F1-score of 83.00%, and an AUC of 0.8460. This study contributes theoretically by reinforcing the importance of data quality in the data mining process, and practically by providing an explicit and replicable preprocessing guideline for researchers and practitioners. The optimal scheme identified can serve as a reference for developing more reliable classification models to support early diagnosis of heart disease and more accurate clinical decision-making.

Keywords: Data Preprocessing, CRISP-DM, KNN, Classification, Heart Disease, Data Preparation

KATA PENGANTAR

Puji syukur ke hadirat Allah Yang Maha Esa atas segala rahmat dan hidayah-Nya, sehingga penulis dapat menyelesaikan tugas akhir dengan judul “Optimalisasi *Preprocessing Data* Menggunakan Pendekatan *CRISP-DM* Untuk Meningkatkan Kualitas Klasifikasi Penyakit Jantung” dengan lancar tanpa suatu halangan apa pun.

Pada kesempatan ini penulis juga mengucapkan terima kasih yang kepada:

1. Bapak Gunawan Adib Ahmadi, S.Pt., M.Pd selaku Ketua STMIK YMI Tegal yang telah memberikan fasilitas, dukungan serta suasana akademik yang kondusif sehingga penulis dapat menyelesaikan tugas akhir ini dengan baik.
2. Bapak Aang Alim Murtopo, M.Kom selaku Ketua Program Studi Informatika yang telah memberikan arahan, motivasi, serta kesempatan bagi penulis untuk terus berkembang dalam proses akademik.
3. Bapak Bangkit Indarmawan Nugroho, M.Kom selaku Dosen Pembimbing I dan Bapak Syefudin, S.Kom., M.Pd., M.Kom selaku Dosen Pembimbing II yang telah dengan sabar membimbing, memberikan masukan serta mendampingi penulis selama penyusunan tugas akhir ini.
4. Rizka Hidayatul Nurlaeli, istri tercinta atas segala doa, dukungan moral, serta semangat yang diberikan kepada penulis selama menjalani proses penyusunan tugas akhir ini.

Akhir kata, penulis berharap semoga Allah *Subhanahu wa Ta'la* senantiasa meridhoi setiap langkah kita dalam menuntut ilmu dan mengamalkannya.

Tegal, 17 Juli 2025

DAFTAR ISI

	Halaman
HALAMAN JUDUL	i
HALAMAN PERSETUJUAN.....	ii
HALAMAN PENGESAHAN	iii
LEMBAR PERNYATAAN ORISINALITAS	iv
ABSTRAK.....	v
ABSTRACT	vi
KATA PENGANTAR.....	vii
DAFTAR ISI.....	viii
DAFTAR GAMBAR	ix
DAFTAR TABEL.....	x
BAB 1 PENDAHULUAN.....	1
1.1. Latar Belakang Masalah.....	1
1.2. Rumusan Masalah	2
1.3. Batasan Penelitian	2
1.4. Tujuan Penelitian.....	3
1.5. Manfaat Penelitian.....	3
BAB 2 TINJAUAN PUSTAKA.....	4
2.1. Penelitian Terkait	4
2.2. Landasan Teori.....	5
BAB 3 METODOLOGI PENELITIAN	9
3.1. Metode Penelitian.....	9
3.2. Teknik Pengumpulan Data	10
3.3. Teknik Analisis Data	12
BAB 4 HASIL DAN PEMBAHASAN	21
4.1. Hasil Penelitian	21
4.2. Pembahasan Hasil Penelitian	35
BAB 5 KESIMPULAN DAN SARAN	38
5.1. Kesimpulan.....	38
5.2. Saran.....	38
DAFTAR PUSTAKA	39
LAMPIRAN.....	43

DAFTAR GAMBAR

	Halaman
Gambar 2.1. Metodologi <i>CRISP-DM</i>	6
Gambar 3.1. Alur Penelitian.....	9
Gambar 3.2. Alur <i>Data Preparation</i>	12
Gambar 3.3. Alur Evaluasi Model	18
Gambar 4.1. Deteksi Noise/ Outlier	23
Gambar 4.2. Hasil <i>Cleaning</i> data	23
Gambar 4.3. Tipe Data Tiap Fitur <i>Dataset</i>	24
Gambar 4.4. Tahap Transformasi Data - <i>Encoding</i>	25
Gambar 4.5. Tahap Transformasi Data - <i>Normalisasi</i>	26
Gambar 4.6. Tampilan Alur data Pada <i>RapidMiner</i>	30
Gambar 4.7. Tampilan dalam <i>Cross validation</i>	30

DAFTAR TABEL

	Halaman
Tabel 2.1. Penelitian Terdahulu	4
Tabel 3.1. Sampel <i>Dataset</i> Asli.....	11
Tabel 3.2. Skema Eksperimen.....	16
Tabel 4.1. Skema Eksperimen.....	21
Tabel 4.2. <i>One-Hot Encoding</i>	24
Tabel 4.3. <i>Label Encoding</i>	24
Tabel 4.4. Perbandingan IG dan IG+ <i>SMOTE</i>	29
Tabel 4.5. Hasil Perhitungan <i>Euclidean Distance</i>	32
Tabel 4.6. Hasil <i>Confusion matrix</i>	33
Tabel 4.7. Hasil Evaluasi Model Tiap Skema.....	34