

**KLASIFIKASI PENYAKIT PCOS (*POLYCYSTIC OVARY
SYNDROME*) MENGGUNAKAN SMOTE (*SYNTHETIC
MINORITY OVERSAMPLING TECHNIQUE*) DAN
ALGORITMA RANDOM FOREST**

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Informatika



disusun oleh

LUTHFIA RIDHO DAMAYANTI

21.11.4185

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA**

2026

**KLASIFIKASI PENYAKIT PCOS (*POLYCYSTIC OVARY
SYNDROME*) MENGGUNAKAN SMOTE (*SYNTHETIC
MINORITY OVERSAMPLING TECHNIQUE*) DAN
ALGORITMA RANDOM FOREST**

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Informatika



disusun oleh

LUTHFIA RIDHO DAMAYANTI

21.11.4185

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA**

2026

HALAMAN PERSETUJUAN

SKRIPSI

**KLASIFIKASI PENYAKIT PCOS (*POLYCYSTIC OVARY
SYNDROME*) MENGGUNAKAN SMOTE (*SYNTHETIC
MINORITY OVERSAMPLING TECHNIQUE*) DAN
ALGORITMA RANDOM FOREST**

yang disusun dan diajukan oleh

Luthfin Ridho Damayanti

21.11.4185

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 26 Februari 2026

Dosen Pembimbing,



Nurhikmah, M. Kom
NIK. 190302245

HALAMAN PENGESAHAN

SKRIPSI

KLASIFIKASI PENYAKIT PCOS (*POLYCYSTIC OVARY SYNDROME*) MENGGUNAKAN SMOTE (*SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE*) DAN ALGORITMA RANDOM FOREST

yang disusun dan diajukan oleh

Luthfia Ridho Damayanti

21.11.4185

Telah dipertahankan di depan Dewan Penguji
pada tanggal 26 Februari 2026

Susunan Dewan Penguji

Nama Penguji

Tanda Tangan

Ali Mustopa, S.Kom., M.Kom.
NIK. 190302192

Andriyan Dwi Putra, S.Kom., M.Kom.
NIK. 190302270

Norhikmah, S.Kom., M.Kom.
NIK. 190302245

Skrripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 26 Februari 2026

DEKAN FAKULTAS ILMU KOMPUTER



Prof. Dr. Kusriani, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Luthfia Ridho Damayanti

NIM : 21.11.4185

Menyatakan bahwa Skripsi dengan judul berikut:

Klasifikasi Penyakit PCOS (*Polycystic Ovary Syndrom*) Menggunakan SMOTE (*Synthetic Minority Oversampling Technique*) Dan Algoritma Random Forest

Dosen Pembimbing : Norhikmah, M.Kom

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 26 Februari 2025

Yang Menyatakan,

A rectangular stamp with a yellow background and a red border. It features the Garuda Pancasila emblem at the top center. Below the emblem, the text "KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, DAN KEMASYARAKATAN" is written in small letters. In the center, "KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, DAN KEMASYARAKATAN" is written in larger letters. At the bottom, the text "25040281832249" is printed.

Luthfia Ridho Damayanti

HALAMAN PERSEMBAHAN

Bismillahirrahmanirrahim

Dengan menyebut nama Allah Yang Maha Pengasih lagi Maha Penyayang, segala puji bagi Allah Swt atas limpahan rahmat, nikmat dan kesempatan yang tak henti diberikan hingga akhirnya skripsi ini dapat terselesaikan. Dengan penuh syukur dan ketulusan, karya sederhana ini penulis persembahkan kepada:

1. Kedua orang tua tercinta, atas doa, kasih sayang, dukungan, serta pengorbanan yang senantiasa mengiringi setiap langkah penulis.
2. Almamater tercinta, Universitas Amikom Yogyakarta sebagai tempat penulis menimba ilmu. Semoga karya ini dapat memberikan kontribusi positif bagi almamater.
3. Dosen Pembimbing, Ibu Norhikmah, M.Kom atas bimbingan, arahan, ilmu, dan waktu yang telah diberikan dengan penuh kesabaran. Bimbingan dan nasihat yang diberikan menjadi petunjuk bagi saya untuk menyelesaikan skripsi.
4. Teman seperjuangan saya Alivia Dyatama Arisusanto, yang telah memberikan bantuan, dukungan, serta kontribusi pemikiran melalui diskusi selama proses penyusunan skripsi.
5. Terakhir saya ucapkan terimakasih untuk Luthfia Ridho Damayanti, diri saya sendiri sebagai bentuk apresiasi atas komitmen dalam menyelesaikan tugas skripsi ini.

KATA PENGANTAR

Puji dan syukur penulis panjatkan ke hadirat Allah SWT atas rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan skripsi ini sebagai salah satu syarat untuk memperoleh gelar Sarjana pada Universitas Amikom Yogyakarta.

Penyusunan skripsi ini tidak terlepas dari bantuan, bimbingan, serta dukungan dari berbagai pihak. Oleh karena itu, pada kesempatan ini penulis menyampaikan ucapan terima kasih yang sebesar-besarnya kepada:

1. Bapak Prof. Dr. M. Suyanto, M.M., selaku Rektor Universitas Amikom Yogyakarta.
2. Ibu Prof. Dr. Kusriani, M.Kom., selaku Dekan Fakultas Ilmu Komputer Universitas Amikom Yogyakarta, atas dukungan akademik dan fasilitas yang diberikan selama penulis menempuh pendidikan.
3. Ibu Eli Pujastuti, M.Kom., selaku Ketua Program Studi Informatika, atas dukungan fasilitas yang disediakan selama masa studi hingga penyusunan skripsi ini.
4. Ibu Norhikmah, M.Kom., selaku Dosen Pembimbing, yang telah meluangkan waktu, tenaga, dan pikiran untuk memberikan bimbingan, arahan, serta masukan yang sangat berarti dalam penyusunan skripsi ini.
5. Tim Dosen Penguji, yang telah memberikan saran, kritik, dan masukan yang membangun guna penyempurnaan skripsi ini.
6. Kedua orang tua dan saudara penulis, atas doa, dukungan serta pengorbanan yang senantiasa diberikan sehingga penulis dapat menyelesaikan studi dan penyusunan skripsi ini dengan baik.
7. Seluruh pihak yang telah membantu penulis, baik secara langsung maupun tidak langsung, yang tidak dapat penulis sebutkan satu per satu, atas segala bantuan dan dukungan yang diberikan selama proses penyusunan skripsi ini.

Penulis menyadari bahwa skripsi ini masih memiliki keterbatasan dan kekurangan. Oleh karena itu, penulis mengharapkan kritik dan saran yang bersifat

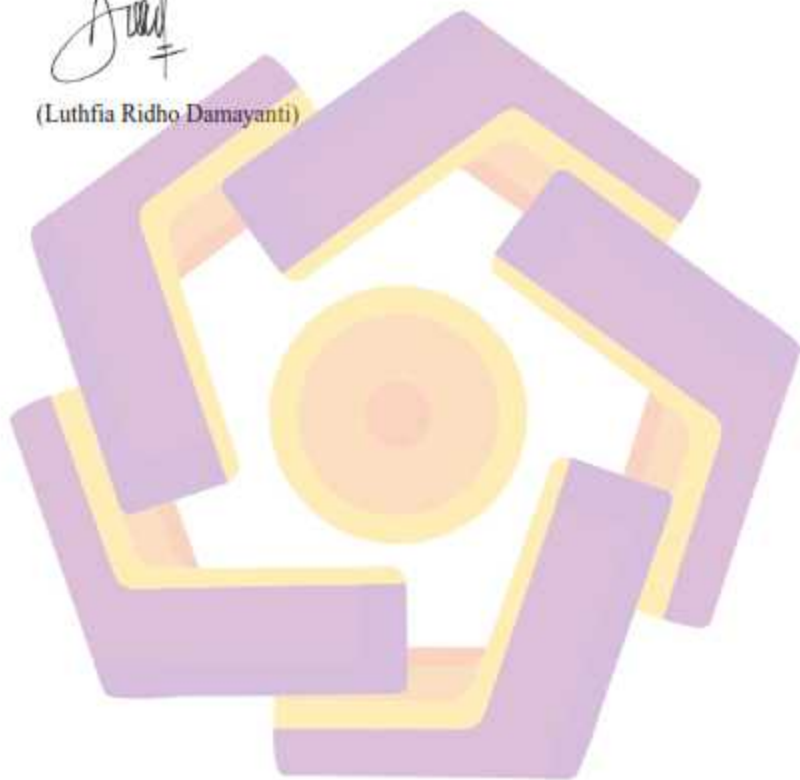
membangun demi penyempurnaan di masa yang akan datang. Semoga skripsi ini dapat memberikan manfaat bagi pengembangan ilmu pengetahuan.

Yogyakarta, 26 Februari 2026

Penulis



(Luthfia Ridho Damayanti)



DAFTAR ISI

HALAMAN JUDUL	i
HALAMAN PERSETUJUAN.....	ii
HALAMAN PENGESAHAN	iii
HALAMAN PERNYATAAN KEASLIAN SKRIPSI	iv
HALAMAN PERSEMBAHAN	v
KATA PENGANTAR	vi
DAFTAR ISI.....	viii
DAFTAR TABEL.....	xi
DAFTAR GAMBAR.....	xii
DAFTAR LAMPIRAN.....	xiii
DAFTAR LAMBANG DAN SINGKATAN	xiv
DAFTAR ISTILAH	xv
INTISARI	xvi
ABSTRACT.....	xvii
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah.....	2
1.3 Batasan Masalah	2
1.4 Tujuan Penelitian	3
1.5 Manfaat Penelitian	4
1.6 Sistematika Penulisan	4
BAB II TINJAUAN PUSTAKA	6
2.1 Studi Literatur	6

2.2	Dasar Teori.....	19
2.2.1	<i>Polycystic Ovary Syndrome (PCOS)</i>	19
2.2.3	<i>Exploratory Data Analysis (EDA)</i>	20
2.2.4	<i>Preprocessing</i>	21
2.2.6	<i>Features Importance</i>	23
2.2.7	SMOTE.....	23
2.2.8	Random Forest.....	24
2.2.9	<i>Grid Search</i>	26
2.2.10	<i>Confusion Matrix</i>	27
2.2.11	ROC Curve dan AUC	28
2.2.12	Streamlit.....	29
BAB III METODE PENELITIAN		30
3.1	Objek Penelitian.....	30
3.2	Alur Penelitian	30
3.2.1	Data Understanding	31
3.2.2	Data Preparation.....	35
3.2.3	Modeling	40
3.2.4	Evaluasi Model	43
3.2.5	Simpan Model.....	43
3.2.6	Deployment.....	44
3.3	Alat dan Bahan	46
3.3.1	Data Penelitian	46
3.3.2	Alat/Instrumen	47
3.3.3	Bahan	47
BAB IV HASIL DAN PEMBAHASAN		48

4.1	Data Understanding	48
4.2	Data Preparation.....	57
4.2.1	Preprocessing Data.....	57
4.2.2	SMOTE	69
4.3	Modeling	70
4.3.1	Modeling Random Forest Tanpa SMOTE	70
4.3.2	Modeling Random Forest Dengan SMOTE	71
4.3.3	Studi Kasus Dengan Perhitungan Manual	73
4.3.4	Evaluasi Model Random Forest Tanpa SMOTE	79
4.3.5	Evaluasi Model Random Forest Dengan SMOTE.....	82
4.3.6	Ringkasan Perbandingan Kinerja Model	86
4.4	Deployment.....	88
4.4.1	Persiapan file untuk Deployment.....	88
4.4.2	Implementasi Model ke Web	89
4.4.3	Hasil Output	91
4.4.4	Testing Output	93
BAB V	PENUTUP	95
5.1	Kesimpulan	95
5.2	Saran	96
REFERENSI		98
LAMPIRAN		106

DAFTAR TABEL

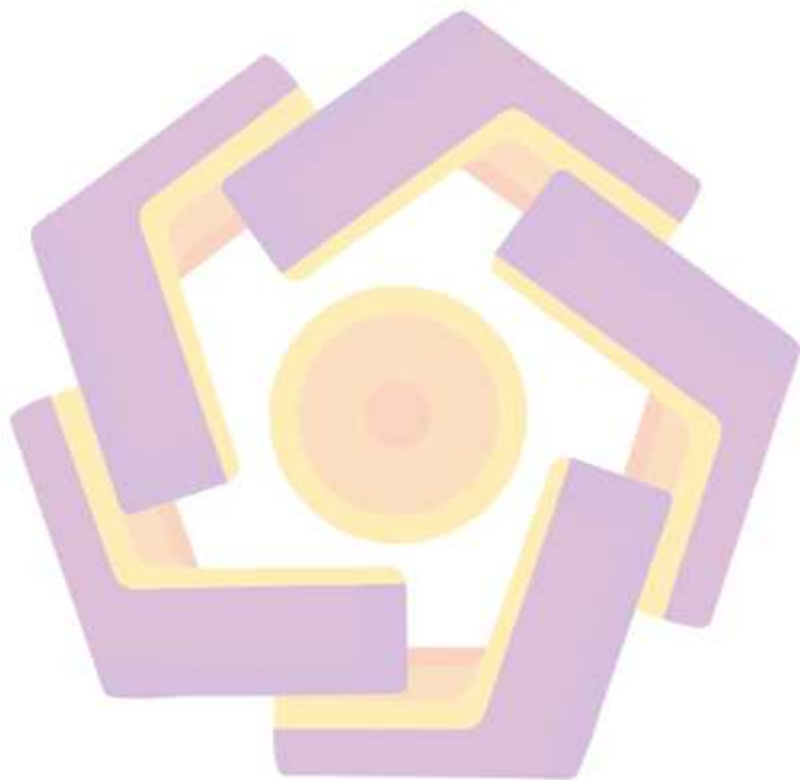
Tabel 2.1	Keaslian Penelitian	11
Tabel 3.1	Cuplikan Keterangan Dataset	33
Tabel 4.1	Jumlah Outlier.....	54
Tabel 4.2	Jumlah Nilai Hilang (<i>missing value</i>)	56
Tabel 4.3	Jumlah kolom sebelum dan sesudah dilakukan imputasi	58
Tabel 4.4	Perbandingan jumlah data sebelum dan sesudah drop outlier	59
Tabel 4.5	Jumlah outlier setelah drop data	60
Tabel 4.6	Perbandingan Jumlah Outlier pada Setiap Iterasi	63
Tabel 4.7	Perbandingan Jumlah Outlier Sebelum dan Sesudah Capping	65
Tabel 4.8	Perbandingan Jumlah Data Sebelum dan Sesudah Penanganan Outlier dengan Capping.....	65
Tabel 4.9	Nilai Tingkat Kepentingan Fitur yang Digunakan	69
Tabel 4.10	Hasil Parameter Terbaik dari Gird Search Tanpa SMOTE	71
Tabel 4.11	Hasil Parameter Terbaik dari Gird Search Menggunakan SMOTE.....	72
Tabel 4.12	Data Awal Fitur.....	73
Tabel 4.13	Hasil Term Frequency (TF)	74
Tabel 4.14	Document Frequency (DF)	74
Tabel 4.15	Inverse Document Frequency (IDF)	74
Tabel 4.16	Perhitungan TF-IDF.....	75
Tabel 4.17	Data yang diinput untuk perhitungan manual Random Forest	75
Tabel 4.18	Perhitungan Gini Root	75
Tabel 4.19	Hasil Perhitungan Gini Index Subset Kiri Fitur Wight Gain	76
Tabel 4.20	Hasil Perhitungan Gini Index Subset Kanan Fitur Wight Gain	77
Tabel 4.21	Hasil Perhitungan Gini Split pada Tree ke-1	77
Tabel 4.22	Parameter Random Forest yang Digunakan Untuk Perhitungan Manual	78
Tabel 4.23	Hasil Prediksi Perhitungan Manual Random Forest.....	79
Tabel 4.24	Hasil Classification Report Model Random Forest Tanpa SMOTE.....	80
Tabel 4.25	Hasil Classification Report Model Random Forest Dengan SMOTE	83
Tabel 4.26	Ringkasan perbandingan kinerja model Random Forest tanpa SMOTE dan dengan SMOTE.....	86
Tabel 4.27	Hasil Pengujian Output	94

DAFTAR GAMBAR

Gambar 2.1	Ilustrasi Algoritma Random Forest	25
Gambar 2.2	Confusion Matrix	27
Gambar 2.3	Logo Streamlit	29
Gambar 3.1	Alur Penelitian	31
Gambar 4.1	Cuplikan dataset	49
Gambar 4.2	Informasi dataset	50
Gambar 4.3	Cuplikan informasi statistik dataset	50
Gambar 4.4	Visualisasi distribusi kelas target	51
Gambar 4.5	Hasil Korelasi antar Fitur	52
Gambar 4.6	Visualisasi boxplot deteksi outlier	55
Gambar 4.7	Grafik feature importance terhadap target PCOS	68
Gambar 4.8	Distribusi Kelas Sebelum dan Sesudah SMOTE	70
Gambar 4.9	Confusion Matrix Model Random Forest Tanpa SMOTE	81
Gambar 4.10	Kurva ROC Model Random Forest Tanpa SMOTE	82
Gambar 4.11	Confusion Matrix Model Random Forest Dengan SMOTE	84
Gambar 4.12	Kurva ROC Model Random Forest Dengan SMOTE	85
Gambar 4.13	Perbandingan kinerja Random Forest tanpa SMOTE dan dengan SMOTE	87
Gambar 4.14	Cuplikan struktur file pada repository GitHub proyek klasifikasi PCOS	89
Gambar 4.15	Tampilan halaman utama web	90
Gambar 4.16	Tampilan hasil prediksi "terindikasi PCOS" pada web	91
Gambar 4.17	Tampilan hasil prediksi "tidak terindikasi PCOS" pada web	92
Gambar 4.18	Tampilan riwayat prediksi	93

DAFTAR LAMPIRAN

Lampiran 1. Link Repository Github	106
Lampiran 2. Link Web Klasifikasi PCOS	106



DAFTAR LAMBANG DAN SINGKATAN



<i>PCOS</i>	<i>Polycystic Ovary Syndrome</i>
<i>EDA</i>	<i>Exploratory Data Analysis</i>
<i>IQR</i>	<i>Interquartile Range</i>
<i>SMOTE</i>	<i>Synthetic Minority Oversampling Technique</i>
<i>TP</i>	<i>True Positive</i>
<i>TN</i>	<i>True Negative</i>
<i>FP</i>	<i>False Positive</i>
<i>FN</i>	<i>False Negative</i>
<i>API</i>	<i>Application Programming Interface</i>
<i>Q3</i>	<i>Kuartil ketiga</i>
<i>Q1</i>	<i>Kuartil pertama</i>
<i>AUC</i>	<i>Area Under Curve</i>
<i>ROC</i>	<i>Receiver Operating Characteristic</i>
<i>TF-IDF</i>	<i>Term Frequency - Inverse Document Frequency</i>

DAFTAR ISTILAH

Dataset	Kumpulan data untuk analisis pada pelatihan
Machine Learning	Cabang keilmuan yang membuat mesin bisa belajar dari data tanpa harus diprogram secara eksplisit
Confusion Matrix	Tabel yang menunjukkan performa model klasifikasi dengan membandingkan prediksi dan label sebenarnya
Overfitting	Model belajar terlalu banyak detail dari data pelatihan, sehingga tidak bekerja baik pada data baru
Preprocessing	Pemrosesan data bertujuan untuk menyiapkan data sebelum diberikan kepada model
Outlier	Data yang memiliki nilai ekstrem dan menyimpang jauh dari mayoritas data lainnya
Class Imbalance	Kondisi ketika jumlah data pada satu kelas jauh lebih banyak dibandingkan kelas lainnya
Hyperparameter	Parameter yang ditentukan sebelum proses pelatihan model dan memengaruhi kinerja algoritma
True Positive	Jumlah data kelas positif yang diprediksi benar sebagai positif oleh model.
True Negative	Jumlah data kelas negatif yang diprediksi benar sebagai negatif oleh model.
False Positive	Jumlah data kelas negatif yang keliru diprediksi sebagai positif oleh model.
False Negative	Jumlah data kelas positif yang keliru diprediksi sebagai negatif oleh model.
Deployment	Tahap penerapan model machine learning ke dalam sistem atau aplikasi yang dapat digunakan oleh pengguna
Streamlit	Framework Python yang digunakan untuk membangun aplikasi web interaktif berbasis data dan machine learning

INTISARI

Polycystic Ovary Syndrome (PCOS) merupakan salah satu gangguan endokrin yang umum terjadi pada wanita usia reproduksi dengan karakteristik gejala yang beragam, sehingga menyulitkan proses analisis data secara manual. Kondisi ini dapat berdampak pada gangguan siklus menstruasi, infertilitas, obesitas, diabetes tipe 2, serta peningkatan risiko penyakit kardiovaskular. Ketidaktepatan dalam mengidentifikasi karakteristik data terkait PCOS dapat menghambat proses analisis informasi kesehatan. Oleh karena itu, diperlukan klasifikasi berbasis data untuk mendukung proses identifikasi PCOS secara lebih akurat. Penelitian ini bertujuan untuk melakukan klasifikasi PCOS menggunakan pendekatan machine learning guna meningkatkan kinerja identifikasi kasus PCOS.

Penelitian ini menggunakan algoritma Random Forest dengan penerapan Synthetic Minority Oversampling Technique (SMOTE) untuk menangani permasalahan ketidakseimbangan kelas pada dataset yang bersumber dari Kaggle. Tahapan penelitian meliputi data understanding, exploratory data analysis, preprocessing data, penanganan outlier menggunakan metode capping, pemilihan fitur, penyeimbangan data latih menggunakan SMOTE, pemodelan Random Forest, serta evaluasi menggunakan confusion matrix dan ROC-AUC. Hasil penelitian menunjukkan bahwa model Random Forest tanpa penerapan SMOTE menghasilkan akurasi sebesar 87% dengan nilai AUC sebesar 0,948. Sementara itu, penerapan SMOTE menghasilkan akurasi sebesar 89% dengan nilai AUC sebesar 0,950, serta mampu meningkatkan recall kelas PCOS dari 0,75 menjadi 0,78. Peningkatan nilai recall tersebut menunjukkan bahwa model menjadi lebih sensitif dalam memprediksi dan klasifikasi kasus PCOS, meskipun diikuti dengan penurunan nilai precision dan F1-score yang dipengaruhi oleh peningkatan jumlah false positive akibat proses penyeimbangan data. Model terbaik kemudian diimplementasikan dalam bentuk aplikasi web prototipe sebagai pemenuhan kebutuhan akademik dan demonstrasi hasil penelitian.

Kata kunci: PCOS, Random Forest, SMOTE, klasifikasi, machine learning.

ABSTRACT

Polycystic Ovary Syndrome (PCOS) is one of the most common endocrine disorders affecting women of reproductive age, characterized by diverse clinical symptoms that make manual data analysis challenging. This condition may lead to menstrual cycle disorders, infertility, obesity, type 2 diabetes mellitus, and an increased risk of cardiovascular diseases. Inaccurate identification of data characteristics related to PCOS can hinder health information analysis; therefore, a data-driven classification approach is required to support more accurate PCOS identification. This study aims to classify PCOS using a machine learning approach to improve the performance of PCOS case identification.

This study employs the Random Forest algorithm with the application of the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance in a dataset obtained from Kaggle. The research stages include data understanding, exploratory data analysis, data preprocessing, outlier handling using the capping method, feature selection, balancing the training data using SMOTE, Random Forest modeling, and performance evaluation using a confusion matrix and ROC-AUC. The results show that the Random Forest model without SMOTE achieves an accuracy of 87% with an AUC value of 0.948. Meanwhile, the application of SMOTE results in an accuracy of 89% with an AUC value of 0.950, and improves the recall of the PCOS class from 0.75 to 0.78. The increase in recall indicates that the model becomes more sensitive in detecting PCOS cases, although it is accompanied by a decrease in precision and F1-score due to an increase in false positive predictions caused by data balancing. The best-performing model is then implemented as a web-based prototype application for academic purposes and research demonstration.

Keyword: PCOS, Random Forest, SMOTE, classification, machine learning.