



KLASIFIKASI TINGKAT KEPARAHAN PENYAKIT PARU KRONIS BERDASARKAN FAKTOR KLINIS DENGAN REGRESI LOGISTIK MULTINOMIAL DAN ANALISIS KOMPONEN UTAMA

NUR NABILA



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 “Klasifikasi Tingkat Keparahan Penyakit Paru Kronis Berdasarkan Faktor Klinis dengan Regresi Logistik Multinomial dan Analisis Komponen Utama” 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

Nur Nabila
G5401221097

ABSTRAK

NUR NABILA, Klasifikasi Tingkat Keparahan Penyakit Paru Kronis Berdasarkan Faktor Klinis Dengan Regresi Logistik Multinomial dan Analisis Komponen Utama. Dibimbing oleh NGAKAN KOMANG KUTHA ARDANA dan FENDY SEPTYANTO.

Penyakit Paru Obstruktif Kronis (PPOK) merupakan penyakit paru kronis yang tingkat keparahannya dipengaruhi oleh berbagai faktor klinis, seperti usia, riwayat merokok, kapasitas aktivitas, gejala, kualitas hidup, dan komorbiditas. Hubungan antar variabel klinis tersebut menyebabkan analisis tingkat keparahan PPOK memerlukan pendekatan statistik multivariat. Data yang digunakan merupakan data sekunder dari dataset PPOK *Gold Stages* yang diperoleh dari Kaggle. Model tanpa AKU menggunakan 13 variabel prediktor klinis asli, sedangkan model dengan AKU menggunakan tiga komponen utama yang dibentuk dari tujuh prediktor kontinu serta enam variabel biner. Hasil pengujian kelayakan AKU menunjukkan nilai KMO sebesar 0.7042 dan *Bartlett's Test* yang signifikan ($p\text{-value} < 0.001$), sehingga data layak dilakukan reduksi dimensi. Tiga komponen utama yang terbentuk mampu menjelaskan 75.30% keragaman data asli. Hasil penelitian menunjukkan bahwa model regresi logistik multinomial tanpa AKU menghasilkan performa terbaik pada data uji dengan akurasi sebesar 0.7, *Macro F1-score* sebesar 0.668, dan AUC sebesar 0.853. Sementara itu, model dengan AKU menghasilkan akurasi sebesar 0.6, *Macro F1-score* sebesar 0.605, dan AUC sebesar 0.770. Meskipun AKU berhasil mengurangi jumlah prediktor dan menghasilkan model yang lebih sederhana serta lebih stabil selama proses validasi silang, penggunaan AKU belum mampu meningkatkan performa klasifikasi dibandingkan penggunaan variabel klinis asli secara langsung. Hasil penelitian ini menunjukkan bahwa informasi diskriminatif yang penting untuk klasifikasi tingkat keparahan PPOK pada data yang digunakan masih lebih baik dipertahankan dalam bentuk variabel klinis asli.

Kata kunci: Analisis Komponen Utama, PPOK, klasifikasi, reduksi dimensi, regresi logistik multinomial.

ABSTRACT

NUR NABILA, Classification of Chronic Lung Disease Severity Based on Clinical Factors Using Multinomial Logistic Regression and Principal Component Analysis. Supervised by NGAKAN KOMANG KUTHA ARDANA and FENDY SEPTYANTO.

Chronic Obstructive Pulmonary Disease (COPD) is a chronic respiratory disease whose severity is influenced by various clinical factors, including age, smoking

history, exercise capacity, symptoms, quality of life, and comorbidities. The relationships among these clinical variables require a multivariate statistical approach for analyzing COPD severity. This study used secondary data from the COPD *Gold Stages* dataset obtained from Kaggle. The model without Principal Component Analysis (PCA) employed 13 original clinical predictor variables, whereas the PCA-based model utilized three principal components derived from seven continuous predictors along with six binary variables. The PCA suitability assessment yielded a Kaiser-Meyer-Olkin (KMO) value of 0.7042 and a significant Bartlett's *Test* result ($p\text{-value} < 0.001$), indicating that the data were appropriate for dimensionality reduction. The three retained principal components explained 75.30% of the total variance in the original dataset. The results showed that the multinomial logistic regression model without PCA achieved the best performance on the *test* data, with an accuracy of 0.7, a *Macro F1-score* of 0.668, and an AUC of 0.853. Meanwhile, the PCA-based model achieved an accuracy of 0.6, a *Macro F1-score* of 0.605, and an AUC of 0.770. Although PCA successfully reduced the number of predictors and produced a simpler and more stable model during cross-validation, it did not improve classification performance compared with using the original clinical variables directly. These findings indicate that the discriminative information required for COPD severity classification in the studied dataset is better preserved in the original clinical variables than in their reduced-dimensional representation.

Keywords: classification, COPD, dimensionality reduction, multinomial logistic regression, principal component analysis.



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.



KLASIFIKASI TINGKAT KEPARAHAN PENYAKIT PARU KRONIS BERDASARKAN FAKTOR KLINIS DENGAN REGRESI LOGISTIK MULTINOMIAL DAN ANALISIS KOMPONEN UTAMA

NUR NABILA

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

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



@Hak cipta milik IPB University

IPB University



IPB University
— Bogor Indonesia —

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.

Tim Penguji pada Ujian Skripsi:

1. Bidayatul Masulah, S.Si., M.Sc.

Perpustakaan IPB University

Judul Skripsi : Klasifikasi Tingkat Keparahan Penyakit Paru Kronis Berdasarkan Faktor Klinis dengan Regresi Logistik Multinomial dan Analisis Komponen Utama

Nama : Nur Nabila
NIM : G5401221097

@Hak cipta milik IPB University

Disetujui oleh

Pembimbing 1:

Ir. Ngakan Komang Kutha Ardana, M.Sc

Pembimbing 2:

Fendy Septyanto, B.Sc., M.Si.

Diketahui oleh

Ketua Program Studi:

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

NIP. 197902272005011001

Tanggal Ujian: 11 Juni 2026

Tanggal Lulus:



PRAKATA

Puji dan syukur penulis panjatkan kepada Allah *subhanahu wa ta'ala* atas segala rahmat, karunia, dan pertolongan-Nya sehingga penulis dapat menyelesaikan karya ilmiah dengan judul “Klasifikasi Tingkat Keparahan Penyakit Paru Kronis Berdasarkan Faktor Klinis dengan Regresi Logistik Multinomial dan Analisis Komponen Utama”.

Proses penyusunan skripsi ini tidak terlepas dari bantuan, arahan, dukungan, dan doa dari berbagai pihak. Oleh karena itu, penulis menyampaikan terima kasih kepada:

1. Kedua orang tua saya yaitu Bapak Sudiyono dan Ibu Nurul Laili yang senantiasa memberikan doa, kasih sayang, dukungan, semangat, dan motivasi kepada penulis selama menjalani perkuliahan hingga penyusunan skripsi ini.
2. Bapak Ir. Ngakan Komang Kutha Ardana, M.Sc. selaku dosen pembimbing pertama dan Bapak Fendy Septyanto, B.Sc., M.Si. selaku dosen pembimbing kedua yang telah memberikan bimbingan, arahan, masukan, dan motivasi selama proses penyusunan skripsi ini. Ibu Bidayatul Masulah, S.Si., M.Sc. selaku dosen penguji yang telah memberikan masukan, arahan, dan saran dalam penyempurnaan skripsi ini, serta atas bimbingan dan dukungan yang telah diberikan kepada penulis selama proses penyusunan skripsi
3. Seluruh dosen dan staf Program Studi Matematika, Sekolah Sains Data, Matematika, dan Informatika IPB University yang telah memberikan ilmu, bantuan, dan dukungan selama penulis menempuh pendidikan.
4. Namira, Salsa, Talenta, Rere, dan Hariri selaku teman-teman terdekat penulis yang telah memberikan bantuan, dukungan, semangat, serta menjadi tempat berbagi selama proses penyusunan skripsi ini.
5. Ahmad Fakhrudin selaku orang terkasih yang membersamai penulis dengan penuh sabar dan perhatian.
6. Seluruh teman-teman IPB University yang telah menjadi bagian dari perjalanan akademik dan non-akademik penulis selama menempuh pendidikan di IPB University.

Penulis menyadari bahwa skripsi ini masih memiliki keterbatasan. Oleh karena itu, kritik dan saran yang membangun sangat diharapkan untuk penyempurnaan karya ilmiah ini. Semoga skripsi ini dapat memberikan manfaat bagi pembaca dan menjadi kontribusi bagi pengembangan ilmu pengetahuan, khususnya dalam penerapan pemodelan matematika dan statistika pada bidang kesehatan.

Bogor, Juli 2026

Nur Nabila



DAFTAR ISI

DAFTAR ISI	viii
DAFTAR TABEL	ix
DAFTAR GAMBAR	ix
DAFTAR LAMPIRAN	ix
PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	2
1.3 Tujuan	2
TINJAUAN PUSTAKA	3
2.1 Penyakit Paru Obstruktif Kronis (PPOK)	3
2.2 Data Tidak Seimbang	4
2.3 Analisis Statistik Multivariat	5
2.4 Analisis Komponen Utama (AKU)	6
2.5 Regresi Logistik Multinomial	11
2.6 <i>5-fold cross-validation</i>	13
2.7 <i>Imbalance ratio</i>	13
2.8 Evaluasi Model	14
III METODE PENELITIAN	17
3.1 Jenis Penelitian	17
3.2 Data dan Sumber Data	17
3.3 Variabel Penelitian	17
3.4 Tahapan Penelitian	18
IV HASIL DAN PEMBAHASAN	22
4.1 Analisis Deskriptif	22
4.2 Pemeriksaan Multikolinearitas Variabel Prediktor	23
4.3 Analisis Komponen Utama (AKU)	26
4.4 Regresi Logistik Multinomial	29
4.5 Evaluasi Model	32
4.6 Perbandingan Model	33
V KESIMPULAN DAN SARAN	36
5.1 Kesimpulan	36
5.2 Saran	36
DAFTAR PUSTAKA	38
LAMPIRAN	40
RIWAYAT HIDUP	62

DAFTAR TABEL

1	<i>Confusion matrix</i> kasus multinomial berkelas 3 untuk kasus A1	14
2	Kelas tingkat keparahan PPOK	17
3	Kelas tingkat keparahan PPOK hasil penggabungan	17
4	Variabel prediktor penelitian	18
5	Statistik deskriptif variabel prediktor kontinu penelitian	23
6	Nilai VIF variabel prediktor	25
7	Nilai KMO per variabel	26
8	Hasil Bartlett's <i>Test of Sphericity</i>	27
9	Nilai eigen, variansi, dan variansi kumulatif dari setiap komponen utama	27
10	Nilai <i>loading</i> dari variabel kontinu pada setiap komponen	28
11	Perbandingan performa <i>cross-validation</i> model regresi logistik multinomial tanpa AKU dan dengan AKU	30
12	Variabel prediktor signifikan pada model regresi logistik multinomial tanpa AKU dan dengan AKU berdasarkan uji <i>bootstrap</i>	31
13	Hasil evaluasi metrik pada model regresi logistik tanpa AKU	32
14	Hasil evaluasi metrik pada model regresi logistik dengan AKU	33
15	Perbandingan kinerja model regresi logistik multinomial tanpa AKU dan dengan AKU pada data uji	33

DAFTAR GAMBAR

1	Diagram alir tahapan penelitian	21
2	Distribusi jumlah pasien dan proporsi kelas tingkat keparahan PPOK	22
3	Heatmap korelasi variabel prediktor	24
4	Plot sebaran data latih pada KU1 dan KU2	29
5	Hasil <i>5-fold cross-validation 5-fold</i> pada data latih	30
6	<i>Confusion matrix</i> pada model regresi logistik tanpa AKU	32
7	<i>Confusion matrix</i> pada model regresi logistik dengan AKU	33
8	Perbandingan metrik evaluasi model regresi logistik multinomial pada data uji	34

DAFTAR LAMPIRAN

1	Hasil estimasi parameter lengkap model regresi logistik multinomial tanpa AKU	40
2	Hasil estimasi parameter lengkap model regresi logistik multinomial dengan AKU	41
3	Distribusi kelas pada data pelatihan dan data pengujian	42
4	Distribusi frekuensi variabel prediktor biner penelitian	42
5	Hasil <i>cross-validation 5-fold</i> pada data latih	42
6	Kode program rekode target: 4 kelas menjadi 3 kelas	42
7	Kode program analisis model regresi logistik multinomial tanpa AKU	43
8	Kode program analisis model regresi logistik multinomial dengan AKU	54