

**PENDEKATAN EXPLAINABLE MACHINE LEARNING
MENGUNAKAN CATBOOST DAN SHAP UNTUK
KLASIFIKASI SPESIES PENGUIN BERDASARKAN
MORFOLOGI**

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi S1 Informatika



disusun oleh
CORNELIA
22.11.4874

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2026**

**PENDEKATAN EXPLAINABLE MACHINE LEARNING
MENGUNAKAN CATBOOST DAN SHAP UNTUK
KLASIFIKASI SPESIES PENGUIN BERDASARKAN
MORFOLOGI**

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi S1 Informatika



disusun oleh
CORNELIA
22.11.4874

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2026**

HALAMAN PERSETUJUAN

SKRIPSI

**PENDEKATAN EXPLAINABLE MACHINE LEARNING
MENGUNAKAN CATBOOST DAN SHAP UNTUK
KLASIFIKASI SPESIES PENGUIN BERDASARKAN
MORFOLOGI**

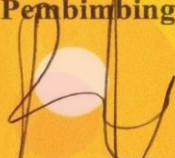
yang disusun dan diajukan oleh

CORNELIA

22.11.4874

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 09 April 2026

Dosen Pembimbing,



Theopilus Bayu Sasongko, S.Kom., M.Eng
NIK. 190302375

HALAMAN PENGESAHAN
SKRIPSI
PENDEKATAN EXPLAINABLE MACHINE LEARNING
MENGGUNAKAN CATBOOST DAN SHAP UNTUK
KLASIFIKASI SPESIES PENGUIN BERDASARKAN
MORFOLOGI

yang disusun dan diajukan oleh

CORNELIA

22.11.4874

Telah dipertahankan di depan Dewan Penguji
pada tanggal 21 April 2026

Susunan Dewan Penguji

Nama Penguji

Anna Baita, S.Kom., M.Kom
NIK. 190302290

Ferian Fauzi Abdulloh, S.Kom., M.Kom
NIK. 190302276

Theopilus Bayu Sasongko, S.Kom., M.Eng
NIK. 190302375

Tanda Tangan



Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 21 April 2026

DEKAN FAKULTAS ILMU KOMPUTER



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

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : CORNELIA
NIM : 22.11.4874

Menyatakan bahwa Skripsi dengan judul berikut:

PENDEKATAN EXPLAINABLE MACHINE LEARNING MENGGUNAKAN CATBOOST DAN SHAP UNTUK KLASIFIKASI SPESIES PENGUIN BERDASARKAN MORFOLOGI

Dosen Pembimbing : Theopilus Bayu Sasongko, S.Kom., M.Eng

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, 21 April 2026

Yang Menvatakan,



HALAMAN PERSEMBAHAN

Dengan rasa syukur yang mendalam, kepada Tuhan yang Maha Esa, penulis mempersembahkan skripsi ini kepada:

1. Allah Subhanahu wa Ta'ala atas segala rahmat, petunjuk, dan karunia-Nya sehingga penulis dapat menyelesaikan studi dan penyusunan skripsi ini dengan baik.
2. Kedua orang tua tercinta yang senantiasa memberikan doa, dukungan, dan kasih sayang tanpa henti selama penulis menempuh pendidikan hingga tahap akhir.
3. Dosen pembimbing dan seluruh dosen Program Studi Informatika Universitas Amikom Yogyakarta atas ilmu, bimbingan, serta arahan yang telah diberikan selama perkuliahan dan penyusunan skripsi.
4. Universitas Amikom Yogyakarta sebagai wujud kontribusi akademik dan hasil dari proses pembelajaran yang telah ditempuh, dengan harapan skripsi ini dapat memberikan manfaat bagi pengembangan ilmu pengetahuan di bidang Informatika.

KATA PENGANTAR

Puji dan syukur penulis panjatkan ke hadirat Allah Subhanahu wa Ta'ala atas segala rahmat dan karunia-Nya, sehingga penulis dapat menyelesaikan skripsi yang berjudul “Pendekatan Explainable Machine Learning Menggunakan CatBoost dan SHAP untuk klasifikasi Spesies Penguin Berdasarkan Morfologi”. Skripsi ini disusun sebagai salah satu syarat untuk memperoleh gelar Sarjana pada Program Studi S1 Informatika.

Dalam proses penyusunan skripsi ini, penulis menyadari bahwa keberhasilan penyelesaian karya ilmiah ini tidak terlepas dari bantuan, bimbingan, dukungan, serta doa dari berbagai pihak. Oleh karena itu, pada kesempatan ini penulis ingin menyampaikan rasa terima kasih yang sebesar-besarnya kepada:

1. Allah Subhanahu wa Ta'ala, Tuhan yang Maha Esa, atas limpahan rahmat, kesehatan, dan kemudahan yang diberikan kepada penulis sehingga skripsi ini dapat diselesaikan dengan baik.
2. Ayah dan Ibu yang senantiasa memberikan dukungan moral maupun doa tanpa henti sejak awal perkuliahan hingga tahap penyelesaian skripsi ini.
3. Bapak Theopilus Bayu Sasongko, S.Kom., M.Eng., selaku dosen pembimbing, yang dengan penuh kesabaran telah memberikan bimbingan, arahan, serta masukan yang sangat berarti selama proses penyusunan skripsi ini.
4. Vania Putri yang telah memberikan dukungan, motivasi, dan semangat kepada penulis, terutama pada saat penulis mengalami penurunan motivasi dalam menyelesaikan skripsi ini.

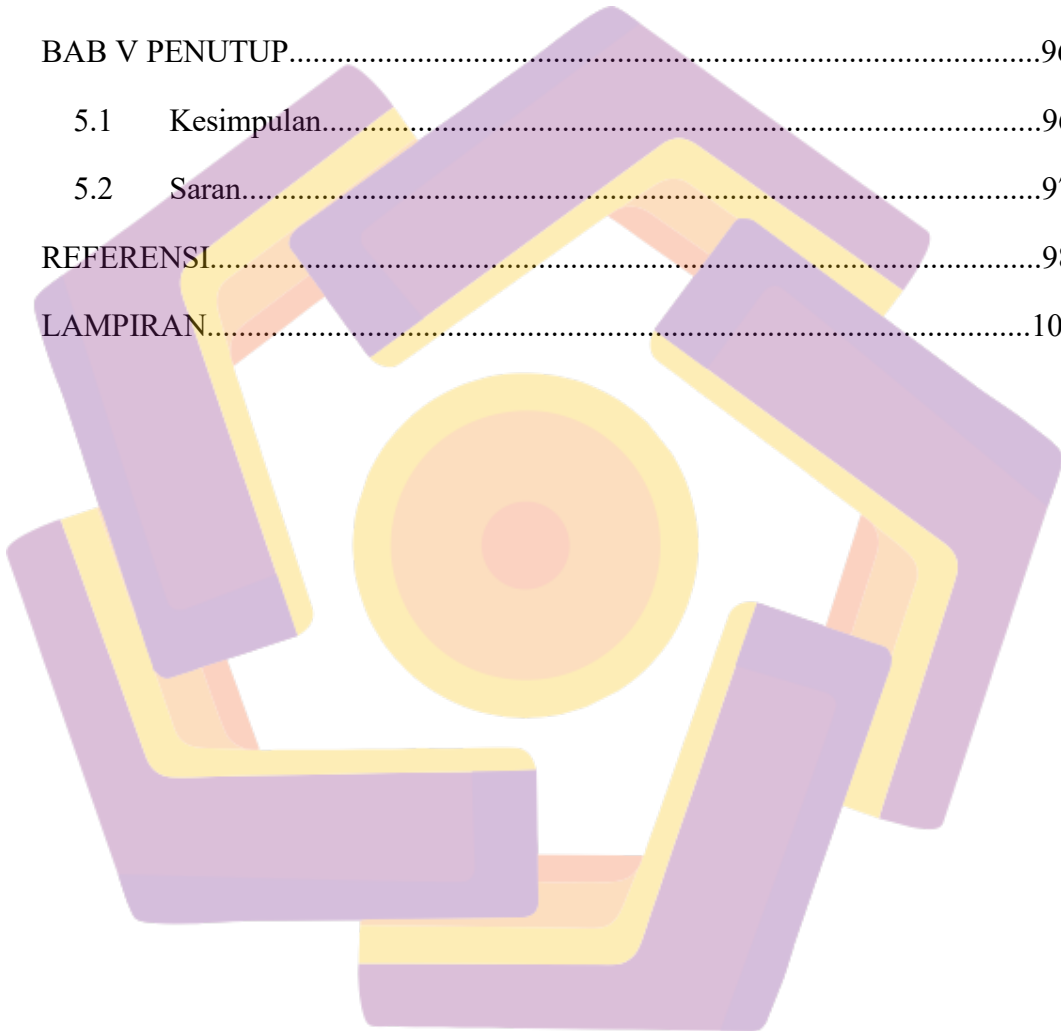
Yogyakarta, 21 April 2026

Penulis

DAFTAR ISI

HALAMAN JUDUL	i
HALAMAN PERSETUJUAN	ii
HALAMAN PENGESAHAN	iii
HALAMAN PERNYATAAN KEASLIAN SKRIPSI	iv
HALAMAN PERSEMBAHAN	v
KATA PENGANTAR.....	vi
DAFTAR ISI	vii
DAFTAR TABEL	ix
DAFTAR GAMBAR	xi
DAFTAR LAMPIRAN	xiii
DAFTAR LAMBANG DAN SINGKATAN	xiv
DAFTAR ISTILAH	xv
INTISARI.....	xvii
<i>ABSTRACT</i>	xviii
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah.....	3
1.3 Batasan Masalah.....	3
1.4 Tujuan Penelitian.....	3
1.5 Manfaat Penelitian.....	4
1.6 Sistematika Penulisan.....	5
BAB II TINJAUAN PUSTAKA.....	6
2.1 Studi Literatur.....	6

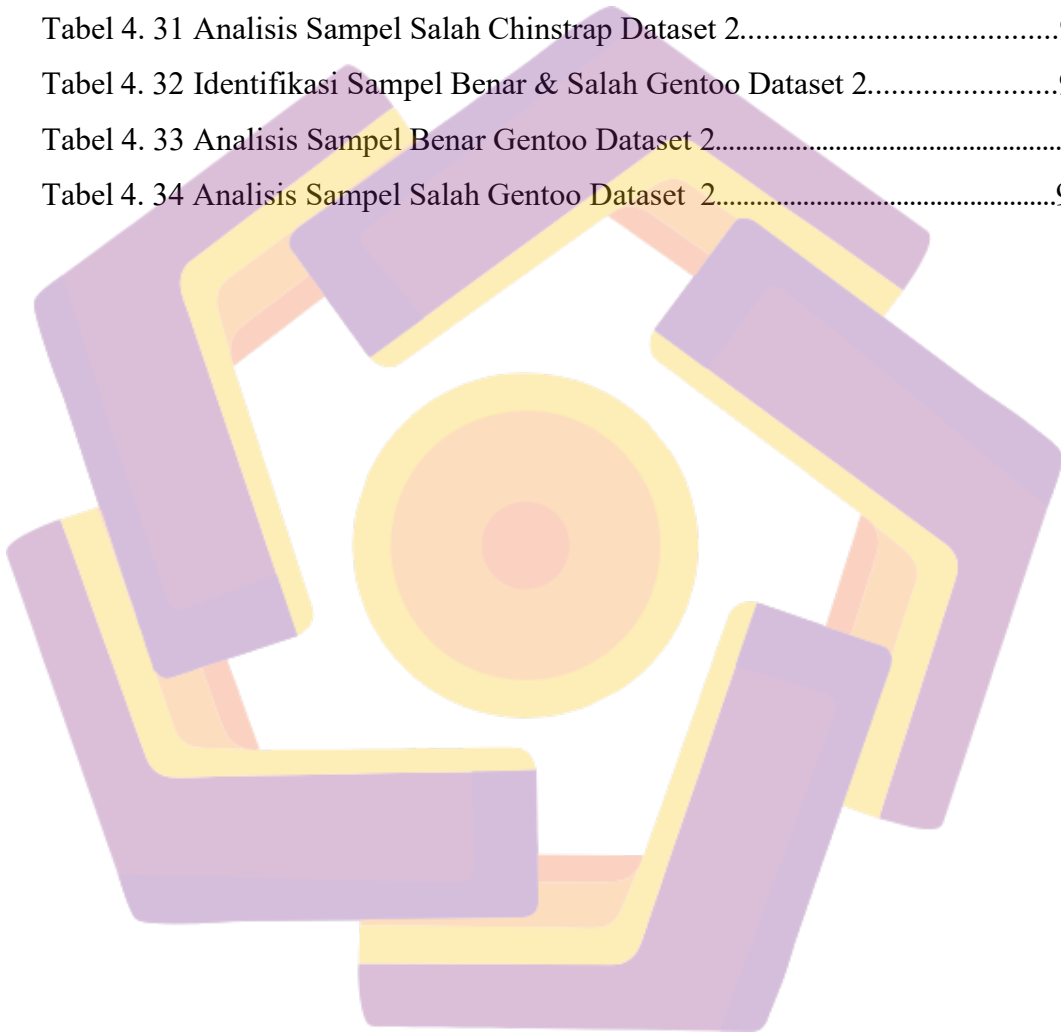
2.2	Dasar Teori.....	16
BAB III METODE PENELITIAN.....		28
3.1	Objek Penelitian.....	28
3.2	Alur Penelitian.....	28
3.3	Alat dan Bahan.....	35
BAB IV HASIL DAN PEMBAHASAN.....		37
BAB V PENUTUP.....		96
5.1	Kesimpulan.....	96
5.2	Saran.....	97
REFERENSI.....		98
LAMPIRAN.....		100



DAFTAR TABEL

Tabel 2. 1 Keaslian Penelitian.....	15
Tabel 3. 1 Alat dan Bahan Penelitian.....	36
Tabel 4. 1 Distribusi Data.....	37
Tabel 4. 2 Distribusi Kelas Target.....	38
Tabel 4. 3 Encoding Data.....	42
Tabel 4. 4 Statistik Scaling.....	43
Tabel 4. 5 Balancing Data.....	44
Tabel 4. 6 Holdout Validation Dataset 1.....	46
Tabel 4. 7 Holdout Validation Dataset 2.....	48
Tabel 4. 8 Cross Validation Dataset 1.....	54
Tabel 4. 9 Cross Validation Dataset 2.....	55
Tabel 4. 10 Search Space Hyperparameter.....	59
Tabel 4. 11 Trial-based Optimization.....	61
Tabel 4. 12 Hasil Optimasi Hyperparameter.....	63
Tabel 4. 13 Classification Report Tuned.....	64
Tabel 4. 14 Evaluasi Model Dataset 1.....	67
Tabel 4. 15 Evaluasi Model Dataset 2.....	68
Tabel 4. 16 Classification Report Data Eksternal.....	71
Tabel 4. 17 Identifikasi Sampel Benar & Salah Adelie Dataset 1.....	76
Tabel 4. 18 Analisis Sampel Benar Adelie Dataset 1.....	77
Tabel 4. 19 Analisis Sampel Salah Adelie Dataset 1.....	78
Tabel 4. 20 Identifikasi Sampel Benar & Salah Chinstrap Dataset 1.....	79
Tabel 4. 21 Analisis Sampel Benar Chinstrap Dataset 1.....	80
Tabel 4. 22 Analisis Sampel Salah Chinstrap Dataset 1.....	81

Tabel 4. 23 Identifikasi Sampel Benar & Salah Gentoo Dataset 1.....	82
Tabel 4. 24 Analisis Sampel Benar Gentoo Dataset 1.....	83
Tabel 4. 26 Identifikasi Sampel Benar & Salah Adelie Dataset 2.....	85
Tabel 4. 27 Analisis Sampel Benar Adelie Dataset 2.....	86
Tabel 4. 28 Analisis Sampel Salah Adelie Dataset 2.....	87
Tabel 4. 29 Identifikasi Sampel Benar & Salah Chinstrap Dataset 2.....	88
Tabel 4. 30 Analisis Sampel Benar Chinstrap Dataset 2.....	89
Tabel 4. 31 Analisis Sampel Salah Chinstrap Dataset 2.....	90
Tabel 4. 32 Identifikasi Sampel Benar & Salah Gentoo Dataset 2.....	91
Tabel 4. 33 Analisis Sampel Benar Gentoo Dataset 2.....	92
Tabel 4. 34 Analisis Sampel Salah Gentoo Dataset 2.....	93



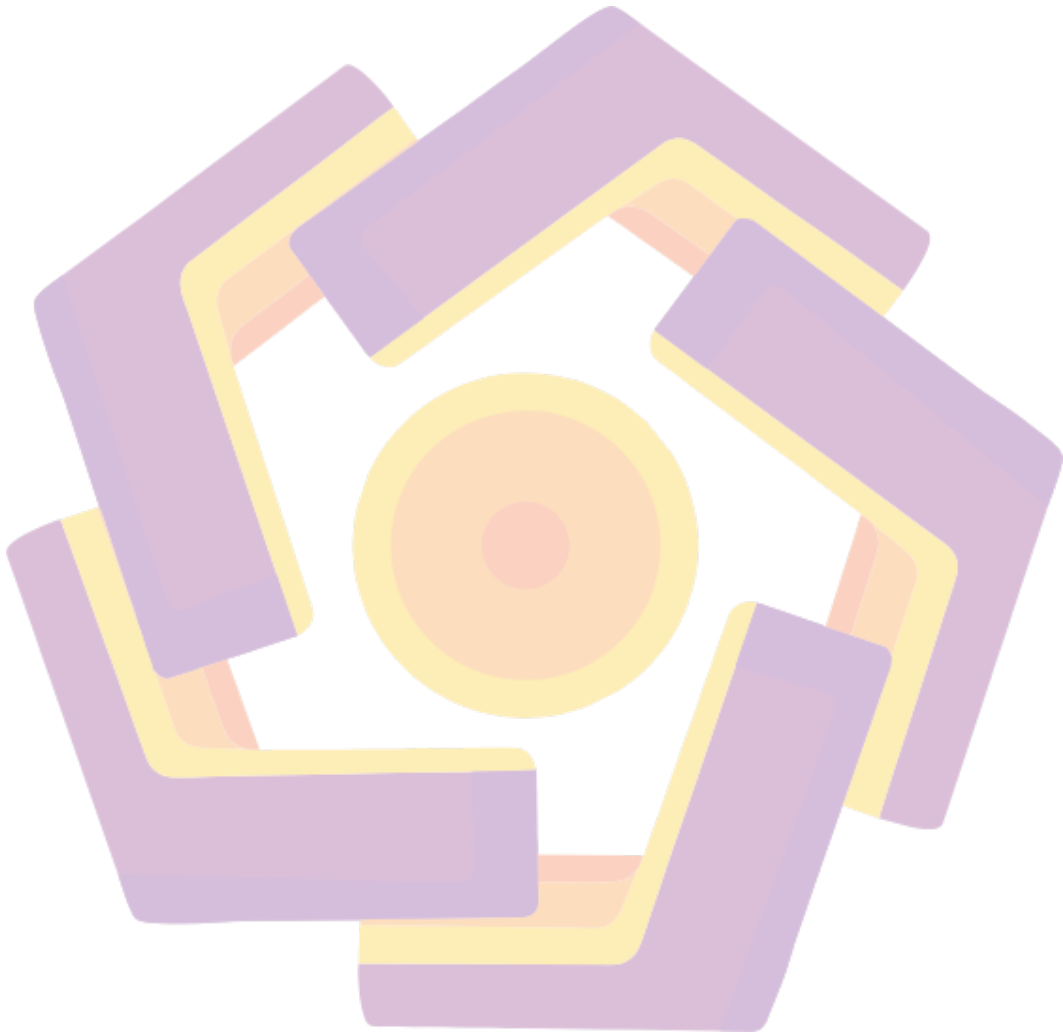
DAFTAR GAMBAR

Gambar 2. 1 Penguin Adelie.....	17
Gambar 2. 2 Penguin Chinstrap.....	17
Gambar 2. 3 Penguin Gentoo.....	17
Gambar 3. 1 Alur Penelitian.....	29
Gambar 4. 1 Presentase Kelas Target.....	39
Gambar 4. 2 Kode Program Penanganan Missing Value.....	40
Gambar 4. 3 Kode Program Encoding Data.....	41
Gambar 4. 4 Kode Program Standar Scaling.....	42
Gambar 4. 5 Kode Program Balancing Data.....	43
Gambar 4. 6 Kode Program Holdout Validation.....	46
Gambar 4. 7 Confusion Matrix Rasio 60:40 Dataset 1.....	51
Gambar 4. 8 Confusion Matrix Rasio 80:20 Dataset 2.....	51
Gambar 4. 9 Kode Program Cross Validation.....	53
Gambar 4. 10 Kode Program Best KFold Dataset 1.....	56
Gambar 4. 11 Kode Program Best KFold Dataset 2.....	56
Gambar 4. 12 Confusion Matrix K-Fold 5 Dataset 1.....	57
Gambar 4. 13 Confusion Matrix K-Fold 10 Dataset 2.....	57
Gambar 4. 14 Kode Program Search Space Hyperparameter.....	59
Gambar 4. 15 Kode Program Trial-based Optimization Dataset 1.....	61
Gambar 4. 16 Kode Program Trial-based Optimization Dataset 2.....	61
Gambar 4. 17 Kode Program Classification Report Tuned.....	64
Gambar 4. 18 Confusion Matrix Tuned Dataset 1.....	66
Gambar 4. 19 Confusion Matrix Tuned Dataset 2.....	66
Gambar 4. 20 Kode Program Inferensi.....	70
Gambar 4. 21 Confusion Matrix Data Eksternal Dataset 1.....	73
Gambar 4. 22 Confusion Matrix Data Eksternal Dataset 2.....	73
Gambar 4. 23 Morfologi Global Dataset 1.....	75
Gambar 4. 24 SHAP Summary Plot Adelie Dataset 1.....	76
Gambar 4. 25 Sampel Benar Adelie Dataset 1.....	77
Gambar 4. 26 Sampel Salah Adelie Dataset 2.....	78

Gambar 4. 27 SHAP Summary Plot Chinstrap Dataset 1.....	79
Gambar 4. 28 Sampel Benar Chinstrap Dataset 1.....	80
Gambar 4. 29 Sampel Salah Chinstrap Dataset 1.....	81
Gambar 4. 30 SHAP Summary Plot Gentoo Dataset 1.....	82
Gambar 4. 31 Sampel Benar Gentoo Dataset 1.....	83
Gambar 4. 32 Sampel Salah Gentoo Dataset 1.....	84
Gambar 4. 33 Morfologi Global Dataset 2.....	84
Gambar 4. 34 SHAP Summary Plot Adelie Dataset 2.....	85
Gambar 4. 35 Sampel Benar Adelie Dataset 2.....	86
Gambar 4. 36 Sampel Salah Adelie Dataset 2.....	87
Gambar 4. 37 SHAP Summary Plot Chinstrap Dataset 2.....	88
Gambar 4. 38 Sampel Benar Chinstrap Dataset 2.....	89
Gambar 4. 39 Sampel Salah Chinstrap Dataset 2.....	90
Gambar 4. 40 SHAP Summary Plot Gentoo Dataset 2.....	91
Gambar 4. 41 Sampel Benar Gentoo Dataset 2.....	92
Gambar 4. 42 Sampel Salah Gentoo Dataset 2.....	93
Gambar 4. 43 Halaman Utama Web.....	94
Gambar 4. 44 Menghasilkan Prediksi Penguin Adelie.....	95
Gambar 4. 45 Menghasilkan Prediksi Penguin Chinstrap.....	95
Gambar 4. 46 Menghasilkan Prediksi Penguin Gentoo.....	95

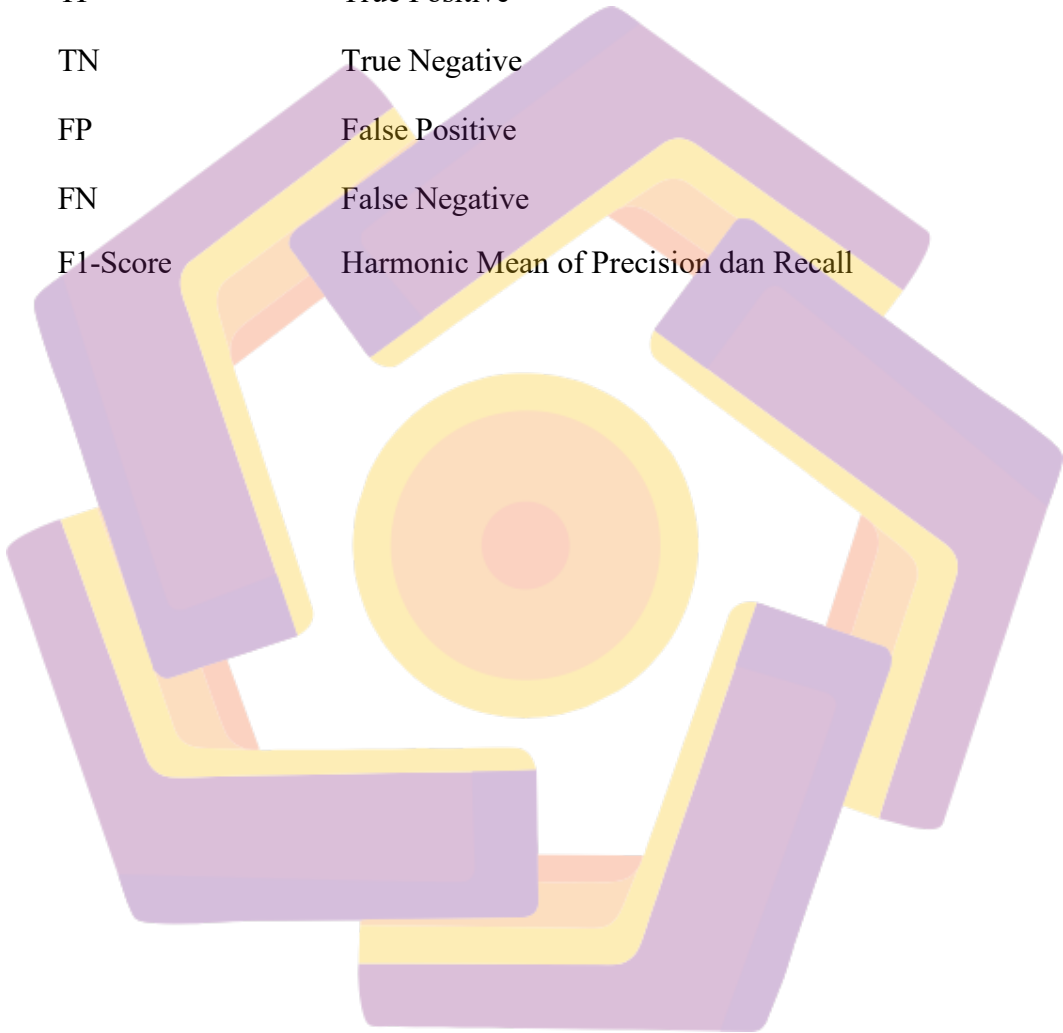
DAFTAR LAMPIRAN

Lampiran 1. 1 File CSV Inferensi Dataset 1.....	100
Lampiran 1. 2 File CSV Inferensi Dataset 2.....	101



DAFTAR LAMBANG DAN SINGKATAN

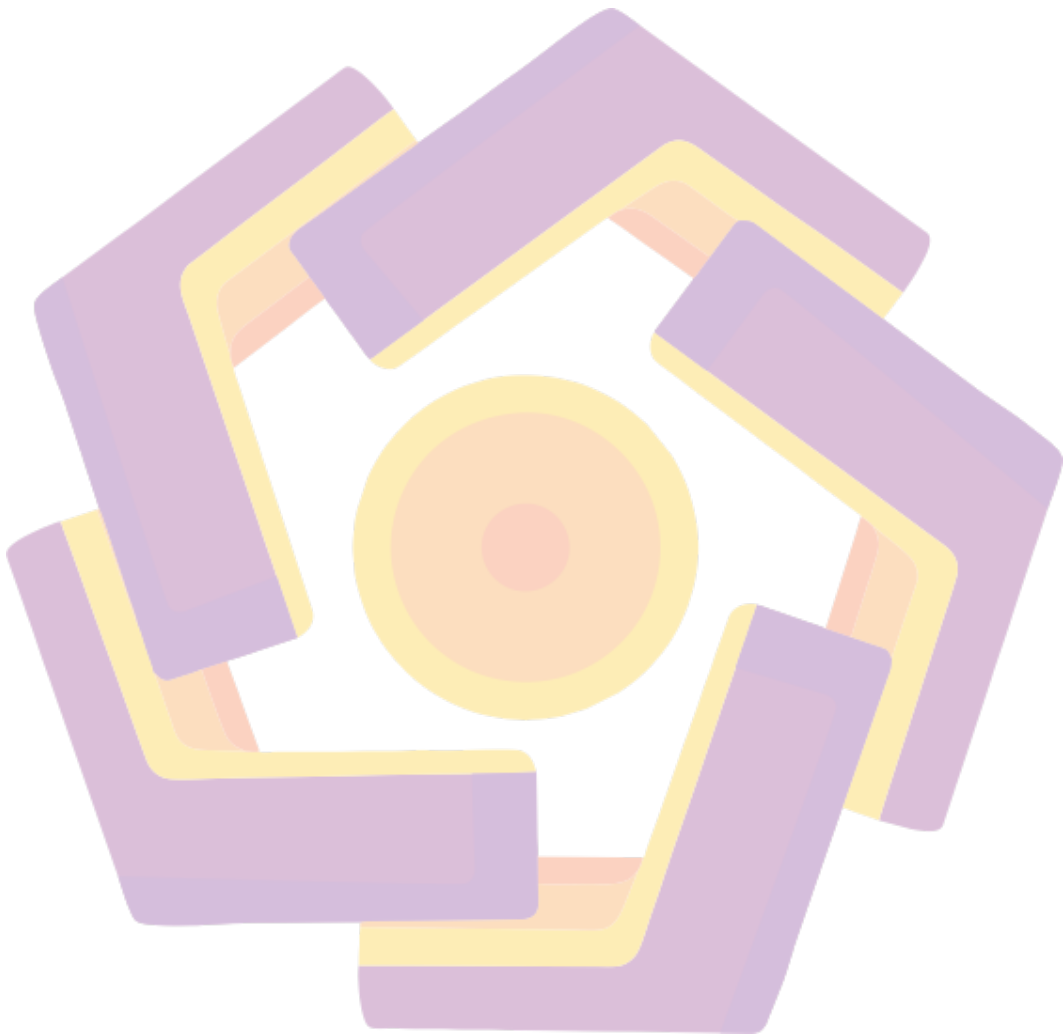
XAI	Explainable Artificial Intelligence
SHAP	Shapley Additive Explanations
ADASYN	Adaptive Synthetic Sampling
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
F1-Score	Harmonic Mean of Precision dan Recall



DAFTAR ISTILAH

Dataset	Kumpulan data yang digunakan dalam proses pelatihan, validasi, dan pengujian model
Klasifikasi	Proses pengelompokan data ke dalam kelas atau kategori berdasarkan pola yang dipelajari model
CatBoost	Algoritma berbasis gradient boosting yang mampu menangani fitur kategorikal secara efisien
CatBoostClassifier	Varian CatBoost yang digunakan untuk tugas klasifikasi.
Hyperparameter	Proses optimasi parameter awal model untuk meningkatkan kinerja
XAI	Konsep AI yang bertujuan membuat keputusan model dapat dipahami manusia
SHAP	Metode interpretasi yang mengukur kontribusi setiap fitur terhadap prediksi
ADASYN	Metode oversampling adaptif untuk menyeimbangkan distribusi kelas
Holdout Validation	Metode evaluasi dengan satu kali pembagian data latih dan data uji
Cross Validation	Metode evaluasi dengan pembagian data ke beberapa lipatan
Cross-Dataset Evaluation	Metode evaluasi model dengan menguji model dari satu dataset pada dataset lain untuk menilai kemampuan generalisasi dan ketahannya terhadap perbedaan distribusi data
Morfologi	Bentuk dan struktur fisik organisme yang digunakan sebagai dasar pengamatan atau analisis
Confusion Matrix	Matriks evaluasi yang menunjukkan perbandingan antara label prediksi dan label sebenarnya
Akurasi	Persentase prediksi yang benar dibandingkan dengan seluruh data pengujian
Presisi	Perbandingan jumlah prediksi positif yang benar terhadap seluruh prediksi positif

Recall	Perbandingan jumlah prediksi positif yang benar terhadap seluruh data positif sebenarnya
F1-Score	Rata-rata harmonis antara presisi dan recall yang menggambarkan keseimbangan keduanya



INTISARI

Data penguin di Kepulauan Palmer, Antartika, merupakan salah satu contoh data yang banyak digunakan untuk penelitian ilmiah karena mencerminkan variasi biologis yang kompleks antarspesies. Namun, sebagian besar penelitian terdahulu hanya berfokus pada peningkatan akurasi model tanpa menjelaskan faktor-faktor yang memengaruhi hasil klasifikasi. Hal ini menyebabkan keterbatasan dalam memahami secara mendalam hubungan antara fitur morfologi dengan prediksi spesies. Penelitian ini bertujuan untuk menerapkan algoritma CatBoostClassifier dalam dua konfigurasi, yaitu parameter default dan hasil hyperparameter tuning, pada Dataset 1 dan Dataset 2 yang telah diselaraskan. Tahapan penelitian meliputi eksplorasi dan preprocessing data, pembagian dataset, pemodelan, evaluasi kinerja menggunakan metrik klasifikasi, serta interpretasi model dengan pendekatan Explainable Machine Learning berbasis SHAP. Hasil penelitian menunjukkan bahwa model CatBoostClassifier hasil tuning mampu menghasilkan kinerja yang optimal dan stabil dengan tingkat akurasi sebesar 99% pada Dataset 1 dan 85% pada Dataset 2, yang selanjutnya divalidasi melalui pengujian menggunakan data eksternal. Analisis SHAP mengungkap bahwa fitur morfologi seperti panjang paruh (Culmen Length (mm) / Bill Length (mm)), kedalaman paruh (Culmen Depth (mm) / Bill Depth (mm)), panjang sirip (Flipper Length (mm)), dan massa tubuh (Body Mass (g)) memiliki kontribusi yang berbeda dalam membedakan setiap spesies penguin. Dengan demikian, penelitian ini menghasilkan model klasifikasi yang tidak hanya akurat tetapi juga interpretatif, sehingga dapat dimanfaatkan sebagai alat bantu analisis berbasis data dalam bidang biologi, khususnya dalam identifikasi spesies, serta mendukung kajian keanekaragaman hayati dan konservasi.

Kata kunci: CatBoost, SHAP, Klasifikasi Penguin, Morfologi, Explainable Machine Learning

ABSTRACT

The penguin dataset from the Palmer Archipelago, Antarctica, is one of the data widely used in scientific research due to its representation of complex biological variations among species. However, most previous studies have primarily focused on improving model accuracy without explaining the factors that influence classification outcomes. This limitation hinders a deeper understanding of the relationship between morphological features and species prediction. This study aims to implement the CatBoostClassifier algorithm in two configurations, namely default parameters and hyperparameter tuning results, on Dataset 1 and Dataset 2 that have been aligned. The research stages include data exploration and preprocessing, dataset splitting, modeling, performance evaluation using classification metrics, and model interpretation using the Explainable Machine Learning approach based on SHAP. The results show that the tuned CatBoostClassifier model is capable of achieving optimal and stable performance, with an accuracy of 99% on Dataset 1 and 85% on Dataset 2, which is further validated through testing using external data. SHAP analysis reveals that morphological features such as bill length (Culmen Length (mm) / Bill Length (mm)), bill depth (Culmen Depth (mm) / Bill Depth (mm)), flipper length (Flipper Length (mm)), and body mass (Body Mass (g)) contribute differently to distinguishing each penguin species. Therefore, this study produces a classification model that is not only accurate but also interpretable, making it useful as a data-driven analytical tool in the field of biology, particularly for species identification, as well as supporting biodiversity and conservation studies.

Keyword: Catboost, SHAP, Penguin Classification, Morphology, Explainable Machine Learning