

**DETEKSI WEBSITE PHISHING BERDASARKAN
ANALISIS STRUKTUR URL MENGGUNAKAN
ALGORITMA MACHINE LEARNING**

SKRIPSI

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



disusun oleh

MUH. ILHAM NADZIRULLAH
22.11.4818

Kepada

PROGRAM SARJANA
PROGRAM STUDI S1-INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2026

**DETEKSI WEBSITE PHISHING BERDASARKAN
ANALISIS STRUKTUR URL MENGGUNAKAN
ALGORITMA MACHINE LEARNING**

SKRIPSI

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



disusun oleh

MUH. ILHAM NADZIRULLAH

22.11.4818

Kepada

**PROGRAM SARJANA
PROGRAM STUDI S1-INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA**

2026

HALAMAN PERSETUJUAN

SKRIPSI

**DETEKSI WEBSITE PHISHING BERDASARKAN ANALISIS
STRUKTUR URL MENGGUNAKAN ALGORITMA MACHINE
LEARNING**

yang disusun dan diajukan oleh

Muh. Ilham Nadzirullah

22.11.4818

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 21 April 2026

Dosen Pembimbing,



Afrig Aminuddin, S.Kom., M.Eng., Ph.D

NIK. 190302351

HALAMAN PENGESAHAN

SKRIPSI

**DETEKSI WEBSITE PHISHING BERDASARKAN ANALISIS
STRUKTUR URL MENGGUNAKAN ALGORITMA MACHINE
LEARNING**

yang disusun dan diajukan oleh

Muh. Ilham Nadzirullah

22.11.4818

Telah dipertahankan di depan Dewan Penguji
pada tanggal 21 April 2026

Susunan Dewan Penguji

Nama Penguji

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

Andika Agus Slameto, S.Kom., M.Kom.
NIK. 190302109

Afrig Aminuddin, S.Kom., M.Eng., Ph.D.
NIK. 190302351

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. Kusriani, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Muh. Ilham Nadzirullah
NIM : 22.11.4818

Menyatakan bahwa Skripsi dengan judul berikut:

Deteksi Website Phishing berdasarkan Analisis Struktur URL menggunakan Algoritma Machine Learning

Dosen Pembimbing : Afrig Aminuddin, S.Kom., M.Eng., Ph.D.

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 Menyatakan,



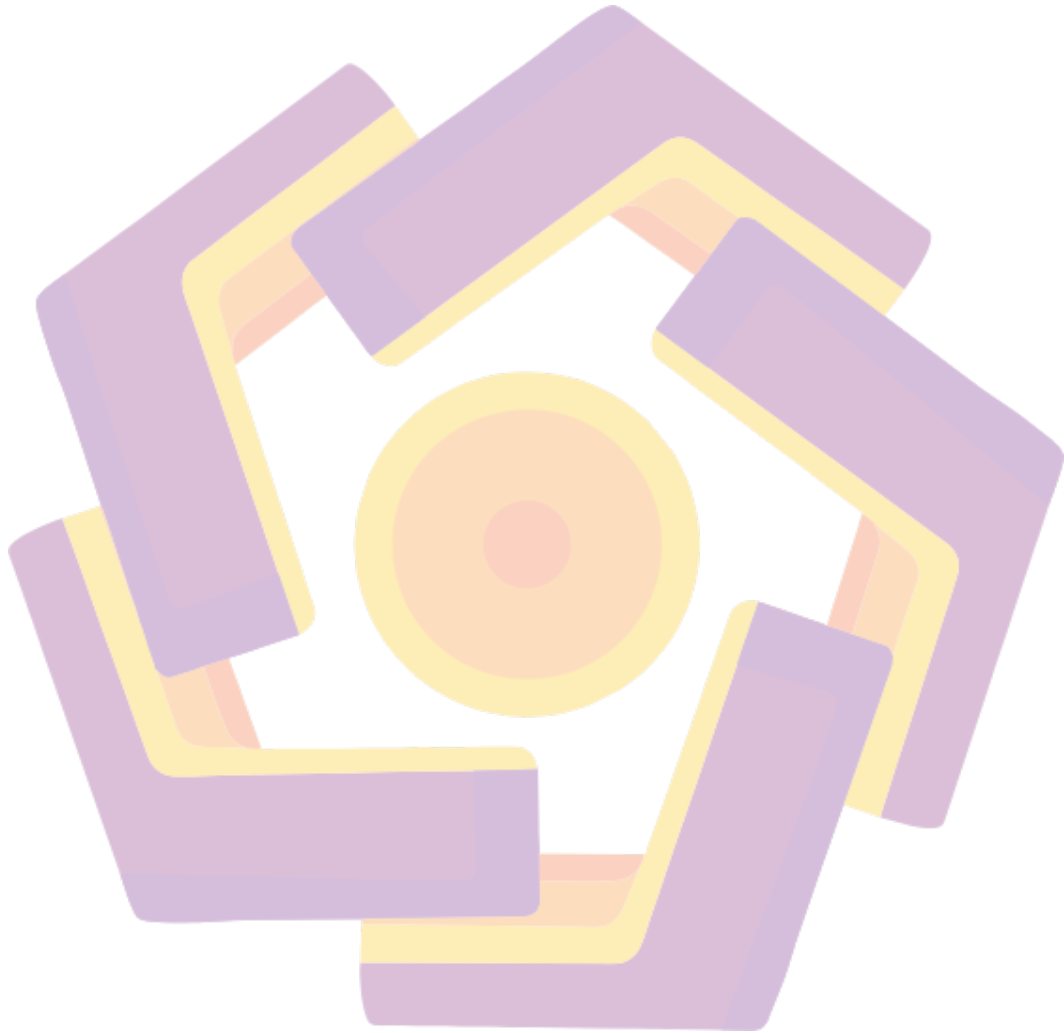
Muh. Ilham Nadzirullah

HALAMAN PERSEMBAHAN

Dengan penuh rasa syukur kehadiran Allah SWT atas segala rahmat, nikmat, dan karunia-Nya, sehingga penulis dapat menyelesaikan skripsi ini dengan baik. Karya sederhana ini penulis persembahkan dengan tulus kepada:

1. Bapak Wahyu Purnomo dan Ibu Nur Fitriawati, kedua orang tua tercinta yang telah menjadi alasan terbesar penulis untuk terus berjuang dan tidak menyerah. Terima kasih atas setiap doa yang tak pernah putus, kasih sayang yang tak pernah habis, serta pengorbanan yang tak ternilai dalam mendukung setiap langkah penulis hingga berada pada titik ini. Tidak ada kata yang cukup untuk menggambarkan rasa terima kasih penulis atas segala cinta, kesabaran, dan perjuangan yang telah diberikan.
2. Kakak saya Marina Febri Purnamawati dan Adik Annisa Nur Rahma, yang selalu memberikan semangat, kebahagiaan, dan warna dalam kehidupan penulis. Terima kasih telah menjadi bagian dari perjalanan hidup yang selalu menghadirkan dukungan dan motivasi.
3. Bapak Afrig Aminuddin, S.Kom., M.Eng., Ph.D., selaku dosen pembimbing yang telah dengan sabar membimbing, memberikan arahan, ilmu, serta motivasi kepada penulis selama proses penyusunan skripsi ini.
4. Niken Tri Mawati, yang selalu hadir menemani, mendukung, dan memberikan semangat dalam setiap proses perjuangan penulis. Terima kasih telah menjadi tempat berbagi cerita, keluh kesah, dan penyemangat di saat penulis merasa lelah selama proses penyusunan skripsi ini.
5. Serta para sahabat Iqbal Faris Ihsani, Rizky Fajar Putra Satria, Rizky Eka Haryadi dan yang tidak dapat disebutkan satu persatu telah menjadi bagian dari perjalanan hidup penulis selama masa perkuliahan. Terima kasih atas kebersamaan, dukungan, tawa, bantuan, serta kenangan indah yang telah diberikan selama ini.

6. Dan kepada semua pihak yang tidak dapat disebutkan satu per satu, yang telah memberikan bantuan, dukungan, serta doa kepada penulis dalam menyelesaikan karya ini. Semoga karya ini menjadi langkah awal bagi penulis untuk meraih cita-cita, membanggakan orang-orang terkasih, serta memberikan manfaat bagi banyak orang.



KATA PENGANTAR

Puji syukur penulis panjatkan kehadiran Allah SWT atas segala rahmat, hidayah, dan karunia-Nya sehingga penulis dapat menyelesaikan skripsi yang berjudul “DETEKSI *WEBSITE PHISHING* BERDASARKAN ANALISIS STRUKTUR URL MENGGUNAKAN ALGORITMA *MACHINE LEARNING*” dengan baik sebagai salah satu syarat untuk memperoleh gelar sarjana pada program studi yang penulis tempuh.

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

1. Allah SWT, atas segala nikmat kesehatan, kekuatan, dan kelancaran yang diberikan kepada penulis selama proses penyusunan skripsi ini.
2. Prof. Dr. M. Suyanto, M.M., selaku Rektor yang telah memberikan kesempatan dan fasilitas sehingga proses pendidikan dapat berlangsung dengan baik.
3. Afrig Aminuddin, S.Kom., M.Eng., Ph.D., selaku dosen pembimbing yang telah meluangkan waktu, tenaga, dan pikiran dalam memberikan bimbingan, arahan, serta masukan selama proses penelitian dan penyusunan skripsi.
4. Tim dosen penguji yang telah memberikan saran dan evaluasi demi penyempurnaan skripsi ini.
5. Seluruh dosen dan civitas akademika, yang telah memberikan ilmu, pengalaman, dan bantuan selama penulis menempuh pendidikan.
6. Bapak Wahyu Purnomo dan Ibu Nur Fitriawati, selaku orang tua tercinta yang selalu memberikan doa, kasih sayang, dukungan moral maupun material, serta motivasi tanpa henti kepada penulis.
7. Kakak Marina Febri Purnamawati dan Adik Annisa Nur Rahma, yang senantiasa memberikan semangat dan dukungan selama proses perkuliahan hingga penyelesaian skripsi.

8. Niken Tri Mawati, yang selalu memberikan semangat, perhatian, dukungan, dan motivasi kepada penulis selama proses penyusunan skripsi.
9. Serta para sahabat terbaik Tegar Febrianto, Iqbal Faris Ihsani, Iman Saputra, Rizky Fajar Putra Satria, dan Rizky Eka Haryadi yang telah memberikan bantuan, semangat, serta kebersamaan selama masa perkuliahan dan penyusunan skripsi.

Penulis menyadari bahwa dalam penyusunan skripsi ini masih terdapat banyak kekurangan dan jauh dari kata sempurna. Oleh karena itu, penulis sangat mengharapkan kritik dan saran yang membangun demi perbaikan di masa yang akan datang.

Akhir kata, penulis berharap semoga skripsi ini dapat memberikan manfaat dan kontribusi bagi perkembangan ilmu pengetahuan, khususnya di bidang teknologi informasi dan keamanan siber.

Yogyakarta, 21 April 2026

Penulis

DAFTAR ISI

HALAMAN JUDUL	i
HALAMAN JUDUL	i
HALAMAN PERSETUJUAN.....	ii
HALAMAN PENGESAHAN	iii
HALAMAN PERNYATAAN KEASLIAN SKRIPSI	iv
HALAMAN PERSEMBAHAN	v
KATA PENGANTAR	vii
DAFTAR ISI.....	ix
DAFTAR TABEL.....	xiii
DAFTAR GAMBAR.....	xiv
DAFTAR LAMBANG DAN SINGKATAN	xvi
DAFTAR ISTILAH	xviii
INTISARI	xxii
ABSTRACT.....	xxiii
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	3
1.3 Batasan Masalah	3
1.4 Tujuan Penelitian	4
1.5 Manfaat Penelitian	5
1.6 Sistematika Penulisan	6
BAB II TINJAUAN PUSTAKA	8
2.1 Studi Literatur	8
2.2 Dasar Teori.....	20
2.2.1 Phishing Website	20
2.2.2 Struktur URL	21

2.2.3	Machine Learning	23
2.2.3.1	Supervised Learning.....	25
2.2.3.2	Unsupervised Learning	25
2.2.3.3	Reinforcement Learning.....	25
2.2.4	Tahapan (Pipeline) Pengembangan Model Machine Learning.....	26
2.2.4.1	Pengumpulan Data.....	26
2.2.4.2	Pra-pemrosesan Data	26
2.2.4.3	Ekstraksi Fitur.....	26
2.2.4.4	Pelatihan Model.....	26
2.2.4.5	Evaluasi dan Pengujian.....	26
2.2.5	Algoritma Machine Learning.....	27
2.2.5.1	Extreme Gradient Boosting (XGBoost).....	28
2.2.5.2	Light Gradient Boosting Machine (LightGBM).....	30
2.2.5.3	Random Forest.....	33
2.2.6	Persamaan Matematis Metrik Evaluasi.....	35
1	Akurasi (Accuracy).....	36
2	Presisi (<i>Precision</i>).....	36
3	Recall (Sensitivity)	37
4	F1-Score.....	37
2.2.7	K-Fold Cross Validation	37
BAB III METODE PENELITIAN		40
3.1	Objek Penelitian.....	40
3.2	Alur Penelitian	40
3.2.1	Mulai	42
3.2.2	Import Dataset.....	43

3.2.3	Eksploratory Data Analysis (EDA)	43
3.2.4	Data Preprocessing	43
3.2.5	Split Data 80:20 (Train & Test)	44
3.2.6	Training Model	44
3.2.7	Evaluasi Model	44
3.2.8	Perbandingan Model	45
3.3	Alat dan Bahan	45
3.3.1	Alat/Instrumen	46
BAB IV HASIL DAN PEMBAHASAN		48
4.1	Pengumpulan Data	48
4.2	Exploratory Data Analysis (EDA)	49
4.3	Pra-pemrosesan Data (Data Preprocessing)	57
4.4	Pembagian Data (Data Splitting)	58
4.5	Pelatihan Model (Model Training)	59
4.5.1	Pelatihan <i>XGBoost</i>	59
4.5.2	Pelatihan <i>LightGBM</i>	60
4.5.3	Pelatihan Random Forest	61
4.6	Evaluasi Model (<i>Model Evaluation</i>)	62
4.6.1	Evaluasi Kinerja <i>XGBoost</i>	63
4.6.2	Evaluasi Kinerja <i>LightGBM</i>	67
4.6.3	Evaluasi Kinerja <i>Random Forest</i>	73
4.7	Analisis Perbandingan Performa	79
4.7.1	Analisis Accuracy	79
4.7.2	Analisis <i>Precision</i>	80

4.7.3	Analisis <i>Recall</i>	80
4.7.4	Analisis <i>F1-Score</i>	80
4.7.5	Analisis <i>ROC-AUC</i>	81
4.8	Analisis Confusion Matriks	82
4.9	Perbandingan Dengan Penelitian Terdahulu.....	83
4.10	Analisis Feature Importance	86
4.11	Implementasi Model (Deployment).....	88
BAB V PENUTUP		90
5.1	Kesimpulan	90
5.2	Saran	91
REFERENSI		93
LAMPIRAN.....		96

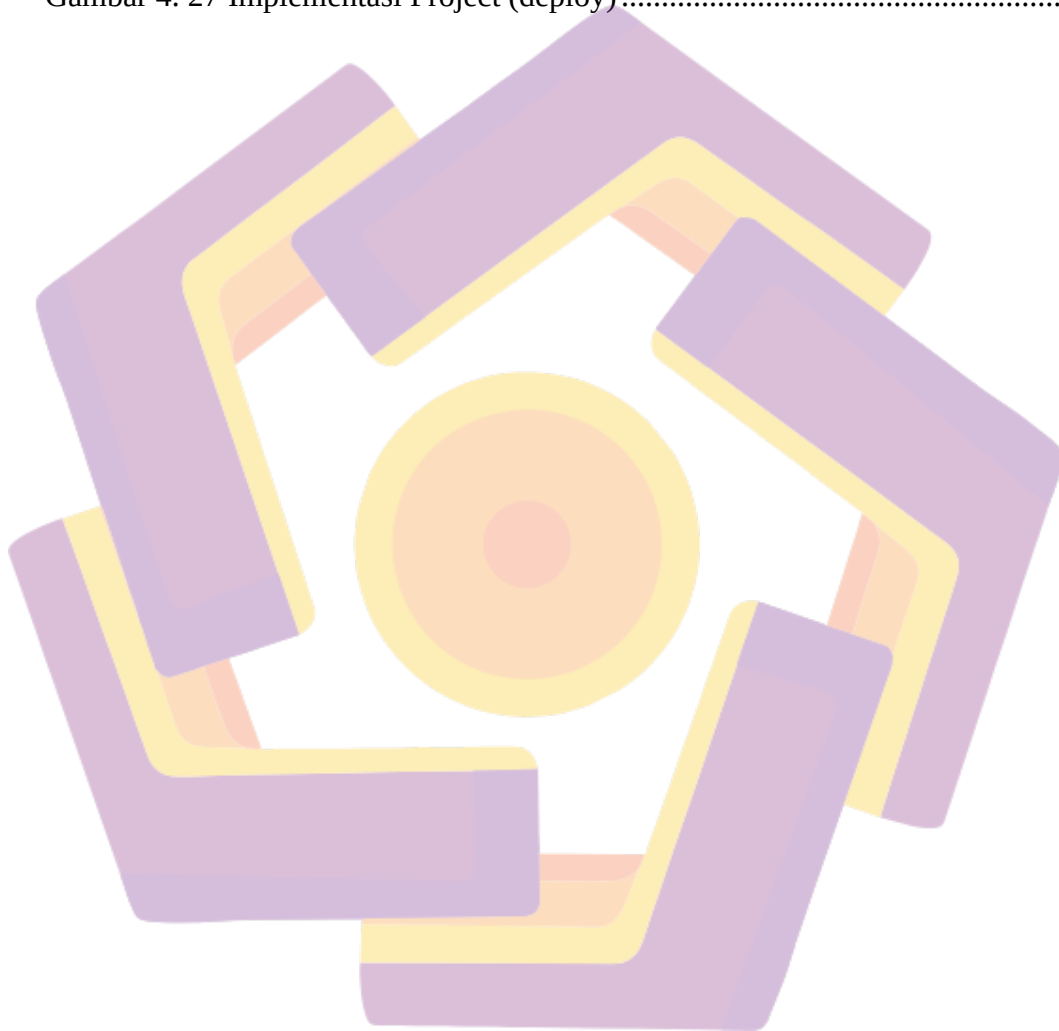
DAFTAR TABEL

Tabel 2. 1 Keaslian Penelitian	11
Tabel 3. 1 Spesifikasi Perangkat Keras.....	46
Tabel 3. 2 Spesifikasi Perangkat Lunak.....	46
Tabel 4. 1 Sampel Lima Baris Pertama Dataset Mentah Phishing Websites.....	49
Tabel 4. 2 Statistik Deskriptif	50
Tabel 4. 3 Hasil Evaluasi 5-Fold Cross-Validation XGBoost	66
Tabel 4. 4 Rekapitulasi Metrik Evaluasi Akhir XGBoost	67
Tabel 4. 5 Hasil Evaluasi 5-Fold Cross-Validation LightGBM	72
Tabel 4. 6 Rekapitulasi Metrik Evaluasi Akhir LightGBM.....	72
Tabel 4. 7 Hasil Evaluasi 5-Fold Cross-Validation Random Forest.....	78
Tabel 4. 8 Rekapitulasi Metrik Evaluasi Akhir Random Forest.....	78
Tabel 4. 9 Perbandingan Performa Model	79
Tabel 4. 10 Perbandingan Confusion Matrix Model Klasifikasi	82
Tabel 4. 11 Perbandingan Penelitian	84

DAFTAR GAMBAR

Gambar 2. 1 Struktur URL.....	23
Gambar 2. 2 Tiga Paradigma Utama Machine Learning	24
Gambar 2. 3 Skema Arsitektur Ensemble Learning pada Pembelajaran Mesin	29
Gambar 2. 4 Mekanisme Leaf-wise Tree Growth pada Algoritma LightGBM.....	32
Gambar 2. 5 Arsitektur Klasifikasi Ansambel pada Algoritma Random Forest ...	33
Gambar 2. 6 Arsitektur Confusion Matrix dan Persamaan Metrik Evaluasi	35
Gambar 2. 7 Arsitektur Pembagian Data dan Iterasi pada 5-Fold Cross Validation	39
Gambar 3. 1 Diagram Alir Tahapan Penelitian	42
Gambar 4. 1 Distribusi Kelas Label Dataset Phishing.....	57
Gambar 4. 2 Proses Penghapusan Atribut index dan Pemisahan Variabel.....	58
Gambar 4. 3 Tahapan Pelatihan Model XGBoost	60
Gambar 4. 4 Pelatihan Model LightGBM.....	61
Gambar 4. 5 Pelatihan Model Random Forest.....	62
Gambar 4. 6 Hasil Evaluasi Data Split	63
Gambar 4. 7 Confusion Matrix Model XGBoost.....	64
Gambar 4. 8 Classification Report XGBoost.....	64
Gambar 4. 9 Grafik Learning Curve Accuracy XGBoost.....	65
Gambar 4. 10 Grafik Learning Curve Loss XGBoost	66
Gambar 4. 11 Evaluasi Data Split LightGBM.....	68
Gambar 4. 12 Confusion Matrix model LightGBM	68
Gambar 4. 13 Classification Report LightGBM	69
Gambar 4. 14 Learning Curve Accuracy	70
Gambar 4. 15 Learning Curve Loss LightGBM	71
Gambar 4. 16 Evaluasi Data Split Random Forest	73
Gambar 4. 17 Confusion Matrix Model Random Forest	74
Gambar 4. 18 Classification Report Random Forest	75
Gambar 4. 19 Grafik Learning Curve Accuracy Random Forest	76
Gambar 4. 20 Grafik Learning Curve Loss Random Forest.....	77
Gambar 4. 21 Kurva ROC XGBoost	81

Gambar 4. 22 Kurva ROC LightGBM.....	81
Gambar 4. 23 Kurva ROC Random Forest.....	82
Gambar 4. 24 Bar Chart Feature Importance XGBoost.....	86
Gambar 4. 25 Bar Chart Feature Importance LightGBM.....	86
Gambar 4. 26 Bar Chart Feature Importance Random Forest	87
Gambar 4. 27 Implementasi Project (deploy)	88



DAFTAR LAMBANG DAN SINGKATAN

X	Matriks fitur / variabel independen
Y	Vektor target / label klasifikasi
TP	True Positive
TN	True Negative
FP	False Positif
FN	False Negatif
%	Persentase
@	Simbol karakter khusus pada URL
Σ	Operasi penjumlahan matematis
=	Simbol kesetaraan matematis
>	Lebih besar dari
<	Lebih kecil dari
URL	Uniform Resource Locator
ML	Machine Learning
KNN	K-Nearest Neighbors
SVM	Support Vector Machine
GBM	Gradient Boosting Machine
XGBoost	Extreme Gradient Boosting
LightGBM	Light Gradient Boosting Machine

AUC

Area Under Curve

ROC

Receiver Operating Characteristic

EDA

Exploratory Data Analysis

CSV

Comma Separated Values

HTTP

Hypertext Transfer Protocol

HTTPS

Hypertext Transfer Protocol Secure

SSL

Secure Sockets Layer

TLS

Transport Layer Security

DNS

Domain Name System

SLD

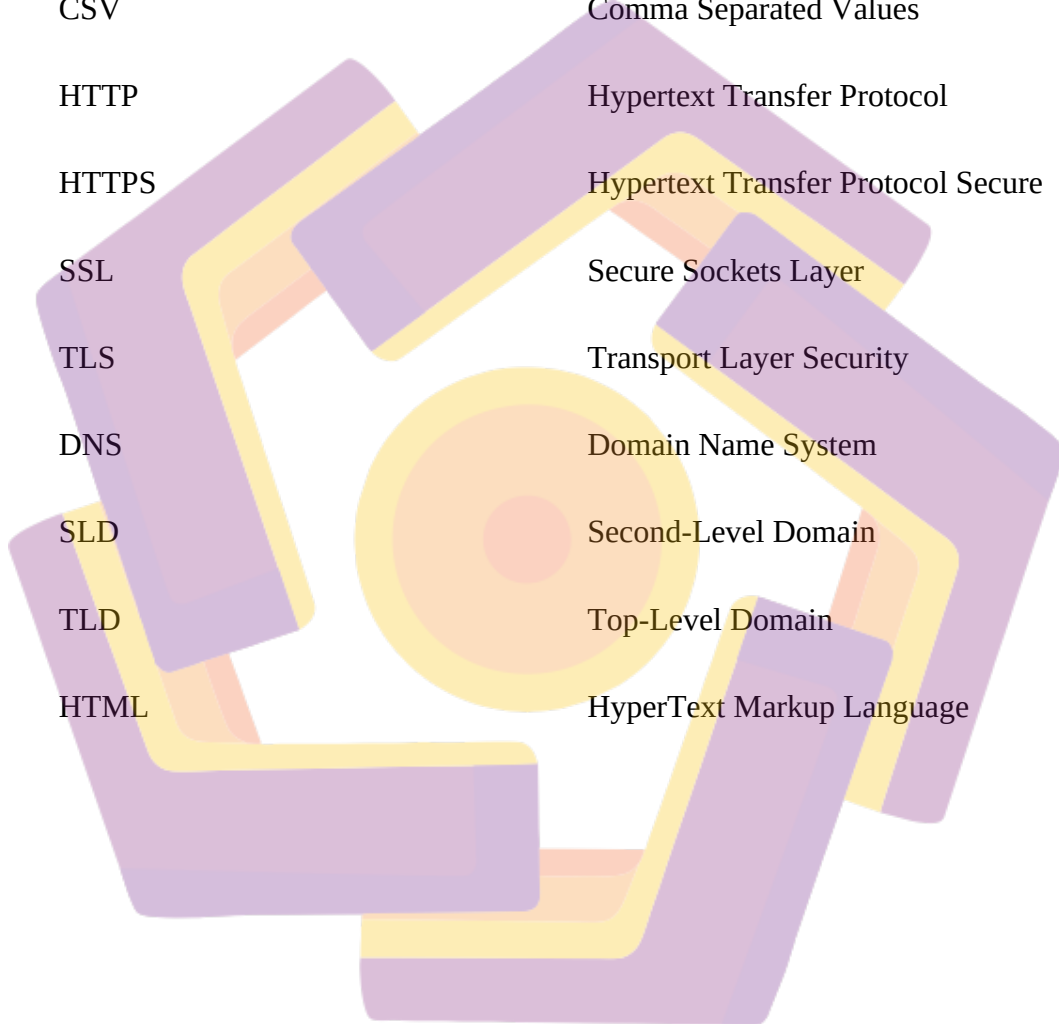
Second-Level Domain

TLD

Top-Level Domain

HTML

HyperText Markup Language



DAFTAR ISTILAH



Phishing	Bentuk kejahatan siber yang bertujuan memperoleh informasi sensitif dengan menyamar sebagai pihak terpercaya.
Website Phishing	Situs web palsu yang dibuat menyerupai website resmi untuk menipu pengguna agar memberikan data pribadi.
URL (Uniform Resource Locator)	Alamat standar yang digunakan untuk mengakses sumber daya di internet.
Struktur URL	Susunan atau komponen pembentuk alamat website yang digunakan sebagai objek analisis dalam penelitian.
Dataset	Sekumpulan data yang digunakan dalam proses pelatihan dan pengujian model machine learning.
Fitur	Karakteristik atau atribut data yang digunakan model sebagai dasar proses klasifikasi.
Ekstraksi fitur	Proses mengambil informasi penting dari data mentah menjadi atribut terstruktur untuk model.
Machine Learning	Metode kecerdasan buatan yang memungkinkan sistem belajar dari data untuk membuat prediksi otomatis.



Random Forest

Algoritma ensemble berbasis banyak decision tree untuk meningkatkan akurasi klasifikasi.

Confusion Matrix

Matriks evaluasi untuk membandingkan hasil prediksi model dengan label sebenarnya.

Accuracy

Persentase prediksi benar dari total keseluruhan data.

Precision

Tingkat ketepatan model dalam memprediksi kelas positif.

Recall

Kemampuan model mendeteksi seluruh data positif aktual.

F1-Score

Nilai harmonik antara precision dan recall untuk mengukur keseimbangan performa model.

AUC-ROC

Metrik evaluasi berdasarkan luas area di bawah kurva ROC untuk mengukur kualitas klasifikasi.

K-Fold Cross Validation

Metode validasi model dengan membagi data ke dalam beberapa subset/lipatan pengujian.

Ensemble Learning

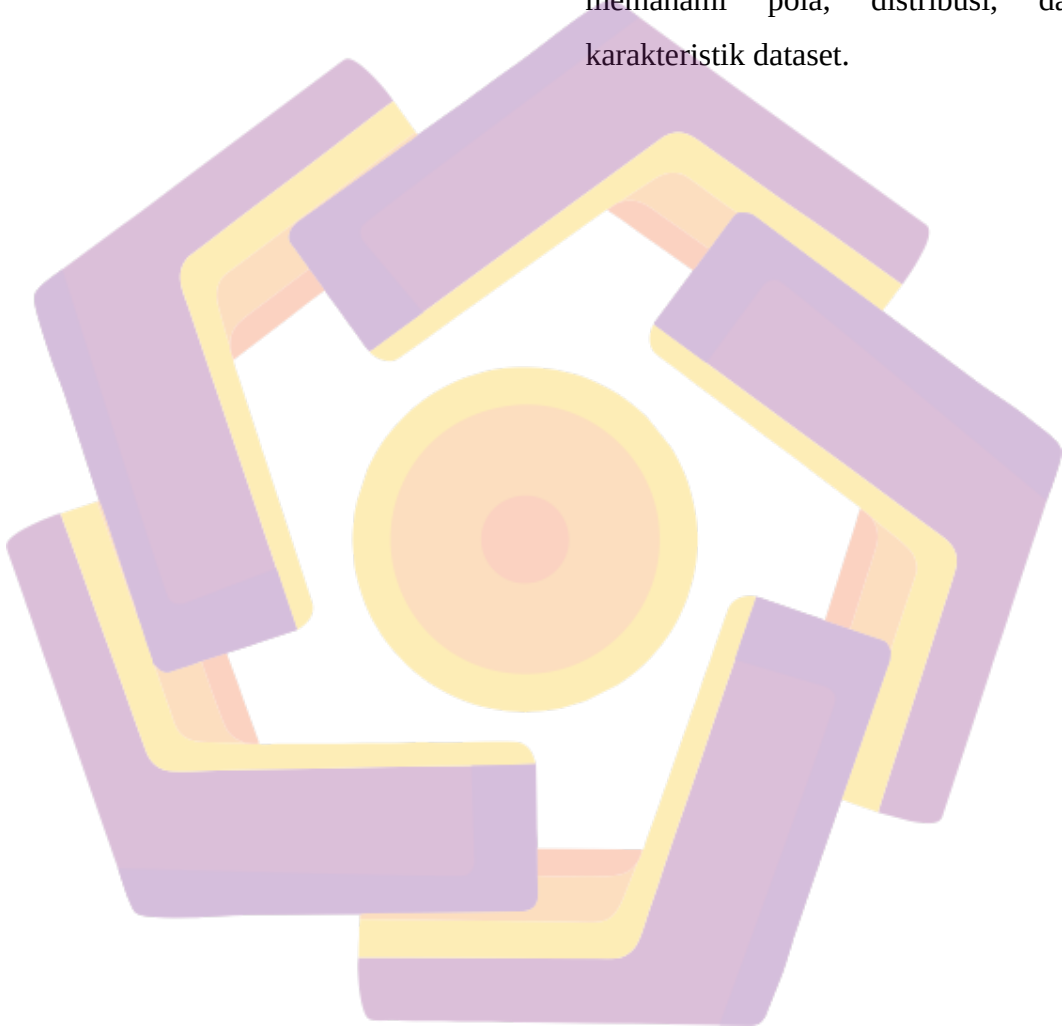
Teknik pembelajaran dengan menggabungkan beberapa model untuk meningkatkan performa prediksi.

Overfitting

Kondisi ketika model terlalu menyesuaikan data latih sehingga kurang baik pada data baru.

Exploratory Data Analysis (EDA)

Proses eksplorasi awal data untuk memahami pola, distribusi, dan karakteristik dataset.



INTISARI

Perkembangan teknologi informasi telah meningkatkan penggunaan internet secara luas, namun juga diikuti dengan meningkatnya ancaman keamanan siber, salah satunya website phishing. Website phishing merupakan situs palsu yang dirancang menyerupai situs resmi untuk memperoleh informasi sensitif pengguna, seperti username, password, dan data finansial. Metode deteksi konvensional seperti blacklist dan rule-based detection dinilai kurang efektif dalam mendeteksi URL phishing baru, sehingga diperlukan pendekatan yang lebih adaptif melalui penerapan machine learning.

Penelitian ini bertujuan untuk mengembangkan model deteksi website phishing berbasis analisis struktur URL menggunakan algoritma machine learning. Algoritma yang digunakan dalam penelitian ini adalah XGBoost, LightGBM, dan Random Forest. Metode penelitian meliputi pengumpulan dataset, exploratory data analysis (EDA), data preprocessing, pembagian dataset dengan rasio 80:20, pelatihan model, evaluasi performa, serta analisis perbandingan antar algoritma. Evaluasi model dilakukan menggunakan metrik Accuracy, Precision, Recall, F1-Score, dan AUC-ROC serta divalidasi dengan metode 5-Fold Cross Validation.

Hasil penelitian menunjukkan bahwa seluruh algoritma memiliki performa yang sangat baik dalam mendeteksi website phishing. Algoritma LightGBM memperoleh hasil terbaik dengan accuracy sebesar 97,01%, precision 96,82%, recall 97,86%, F1-score 97,33%, dan AUC-ROC 99,69%. Berdasarkan hasil tersebut, LightGBM dinilai sebagai model paling optimal dalam mendeteksi website phishing berbasis analisis struktur URL.

Kata Kunci: *Website Phishing, Struktur URL, Machine Learning, LightGBM, XGBoost, Random Forest.*

ABSTRACT

The rapid development of information technology has increased internet usage significantly, but it has also led to growing cybersecurity threats, one of which is phishing websites. Phishing websites are fraudulent websites designed to imitate legitimate websites in order to obtain sensitive user information such as usernames, passwords, and financial data. Conventional detection methods such as blacklist and rule-based detection are considered less effective in identifying new phishing URLs, therefore a more adaptive approach is needed through machine learning implementation.

This study aims to develop a phishing website detection model based on URL structure analysis using machine learning algorithms. The algorithms implemented in this research are XGBoost, LightGBM, and Random Forest. The research methodology includes dataset collection, exploratory data analysis (EDA), data preprocessing, dataset splitting using an 80:20 ratio, model training, performance evaluation, and comparative analysis among algorithms. Model evaluation is conducted using Accuracy, Precision, Recall, F1-Score, and AUC-ROC metrics and validated using 5-Fold Cross Validation.

The results show that all algorithms achieved excellent performance in phishing website detection. LightGBM obtained the best performance with an accuracy of 97.01%, precision of 96.82%, recall of 97.86%, F1-score of 97.33%, and AUC-ROC of 99.69%. Based on these results, LightGBM is considered the most optimal model for phishing website detection based on URL structure analysis.

Keywords: *Phishing Website, URL Structure, Machine Learning, LightGBM, XGBoost, Random Forest.*