



L A P O R A N S K R I P S I

KOMPARASI ALGORITMA *NAIVE BAYES*, *SIMPLE LOGISTIC*, *RANDOM FOREST*, DAN *DECISION TREE J48* DALAM KLASIFIKASI SITUS WEB *PHISHING*

Nabilla Artamevia
NIM. 202151044

DOSEN PEMBIMBING

Alif Catur Murti, S.Kom, M.Kom
Rizkysari Meimaharani, S.Kom, M.Kom

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNIK
UNIVERSITAS MURIA KUDUS
2025



L A P O R A N S K R I P S I

**KOMPARASI ALGORITMA *NAIVE BAYES*, *SIMPLE LOGISTIC*, *RANDOM FOREST*, DAN *DECISION TREE J48*
DALAM KLASIFIKASI SITUS WEB *PHISHING***

**Nabilla Artamevia
NIM. 202151044**

DOSEN PEMBIMBING

**Alif Catur Murti, S.Kom, M.Kom
Rizkysari Meimaharani, S.Kom, M.Kom**

**PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNIK
UNIVERSITAS MURIA KUDUS
2025**

HALAMAN PERSETUJUAN

KOMPARASI ALGORITMA *NAIVE BAYES*, *SIMPLE LOGISTIC*, *RANDOM FOREST*, DAN *DECISION TREE J48* DALAM KLASIFIKASI SITUS WEB *PHISHING*

NABILLA ARTAMEVIA

NIM. 202151044

Kudus, 19 Desember 2024

Menyetujui,

Pembimbing Utama,

Pembimbing Pendamping,

Alif Catur Murti, S.Kom., M.Kom
NIDN. 0629077402

Rizkysari Meimaharani, S.Kom., M.Kom
NIDN. 0620058501

Koordinator Skripsi,

Alvin Rainaldy Hakim, S.Kom., M.Kom
NIDN. 9990586218

HALAMAN PENGESAHAN

KOMPARASI ALGORITMA *NAIVE BAYES*, *SIMPLE LOGISTIC*, *RANDOM FOREST*, DAN *DECISION TREE J48* DALAM KLASIFIKASI SITUS WEB *PHISHING*

NABILLA ARTAMEVIA

NIM. 202151044

Kudus, 20 Januari 2025

Menyetujui,

Ketua Penguji,

Anggota Penguji I,

Anggota Penguji II,

Dr. Ahmad Abdul Chamid., M.Kom.
NIDN. 0616109101

Endang Supriyati., M.Kom
NIDN. 0629077402

Alif Catur Murti., M.Kom
NIDN. 0610129001

Mengetahui

Dekan Fakultas Teknik

Ketua Program Studi Teknik Informatika

Dr. Eko Darmanto, S.Kom., M.Cs
NIY. 0610701000001171

Ir. Muhammad Imam Ghozali., M.Kom
NIY. 0610701000001289

PERNYATAAN KEASLIAN

Saya yang bertanda tangan dibawah ini:

Nama : Nabilla Artamevia
NIM : 202151044
Tempat & Tanggal Lahir : Kudus, 19 Desember 2003
Judul Skripsi : Komparasi Algoritma Naive Bayes, Simple Logistic, Random Forest, Dan Decision Tree J48 Dalam Klasifikasi Situs Web Phishing

Dengan ini, saya menyatakan dengan sungguh-sungguh bahwa seluruh isi skripsi ini, mulai dari perumusan masalah, analisis data, hingga penulisan laporan akhir, merupakan hasil karya intelektual saya sendiri. Saya telah melakukan penelitian secara mandiri dan bertanggung jawab atas setiap informasi yang tertuang dalam skripsi ini. Segala ide, pendapat, atau temuan yang berasal dari sumber lain telah saya kutip dan rujuk sesuai dengan kaidah penulisan ilmiah yang berlaku.

Saya menyatakan dengan sungguh-sungguh bahwa seluruh isi skripsi ini merupakan hasil karya orisinal saya. Saya telah menjunjung tinggi prinsip-prinsip integritas akademik dalam proses penulisannya. Segala sumber yang digunakan telah dikutip dan dicantumkan dengan benar. Apabila di kemudian hari ditemukan adanya penyimpangan atau ketidaksesuaian dalam pernyataan ini, saya bersedia menerima segala bentuk sanksi akademik yang berlaku di Universitas Muria Kudus

Demikian pernyataan ini saya buat dalam keadaan sadar tanpa paksaan dari pihak manapun.

Kudus, 19 Desember 2024

Yang memberi pernyataan,

Nabilla Artamevia
NIM. 202151044

KATA PENGANTAR

Syukur Alhamdulillah, segala puji dan syukur penulis panjatkan ke hadirat Allah SWT atas rahmat, hidayah, dan karunia-Nya, sehingga penulis dapat menyelesaikan skripsi yang berjudul "Komparasi Algoritma Naive Bayes, Simple Logistic, Random Forest, dan Decision Tree J48 dalam Klasifikasi Situs Web Phishing". Skripsi ini disusun sebagai salah satu syarat untuk memperoleh gelar Sarjana Teknik (S1) pada Program Studi Teknik Informatika.

Proses penyusunan skripsi ini tidak terlepas dari bantuan, bimbingan, dan dukungan dari berbagai pihak. Oleh karena itu, dengan segala kerendahan hati, penulis ingin menyampaikan ucapan terima kasih yang sebesar-besarnya kepada:

1. Bapak Prof. Dr. Ir. Darsono, M.Si., selaku Rektor Universitas Muria Kudus, yang telah memberikan fasilitas dalam proses studi hingga penyelesaian skripsi ini.
2. Bapak Ir. Muhammad Imam Ghozali, M.Kom, selaku Ketua Program Studi Teknik Informatika, yang telah memberikan arahan dan dukungan.
3. Bapak Alif Catur Murti, S.Kom., M.Kom, selaku pembimbing utama, yang dengan penuh kesabaran dan kebijaksanaan memberikan bimbingan, saran, dan kritik dalam penyelesaian skripsi ini.
4. Ibu Rizkysari Mei Maharani, S.Kom., M.Kom, atas masukan dan dukungannya selama proses penelitian dan penulisan skripsi ini.
5. Ibu Indah Fajarwati yang selalu memberikan doa, dukungan, serta motivasi yang tiada henti kepada penulis.

Penulis menyadari bahwa skripsi ini masih jauh dari sempurna. Oleh karena itu, kritik, saran, dan masukan yang membangun dari para pembaca sangat penulis harapkan demi penyempurnaan karya di masa mendatang.

Akhir kata, penulis berharap skripsi ini dapat memberikan kontribusi kecil bagi pengembangan ilmu pengetahuan, khususnya dalam bidang klasifikasi situs web phishing.

Kudus, 19 Desember 2024

Penulis

JUDUL SKRIPSI
KOMPARASI ALGORITMA NAIVE BAYES, SIMPLE LOGISTIC,
RANDOM FOREST, DAN DECISION TREE J48 DALAM
KLASIFIKASI SITUS WEB PHISHING

Nama mahasiswa : Nabilla Artamevia

NIM : 202151044

Pembimbing :

1. Alif Catur Murti, S.Kom, M.Kom
2. Rizkysari Meimaharani, S.Kom, M.Kom

RINGKASAN

Phishing adalah salah satu ancaman keamanan siber terbesar yang bertujuan mencuri informasi pribadi pengguna melalui situs web palsu. Penelitian ini bertujuan untuk membandingkan kinerja empat algoritma machine learning, yaitu Naive Bayes, Simple Logistic, Random Forest, dan Decision Tree J48, dalam mendeteksi situs web *phishing*. Selain itu, penelitian ini juga mengidentifikasi fitur paling signifikan yang berkontribusi dalam klasifikasi situs *phishing*, dengan harapan dapat meningkatkan efektivitas deteksi *phishing*.

Metode penelitian melibatkan penggunaan dataset berisi 11.055 data situs web yang diolah dengan teknik SMOTE (*Synthetic Minority Over-sampling Technique*) untuk mengatasi ketidakseimbangan kelas. Evaluasi algoritma dilakukan menggunakan 10-Fold Cross Validation dan metrik evaluasi seperti akurasi, presisi, recall, F1 Score, MCC, dan ROC AUC. Setiap algoritma diuji untuk menentukan algoritma terbaik dalam mendeteksi situs *phishing* berdasarkan performa metrik tersebut.

Hasil penelitian menunjukkan bahwa algoritma Random Forest memberikan performa terbaik, dengan akurasi 97,60%, presisi 97,14%, recall 98,08%, dan F1 Score 97,61%. Fitur *SSLfinal_State* ditemukan sebagai atribut paling signifikan dalam mendeteksi situs *phishing*. Penelitian ini diharapkan dapat menjadi referensi bagi pengembangan sistem deteksi *phishing* yang lebih andal dan akurat.

Kata kunci: Phishing, Naive Bayes, Simple Logistic, Random Forest, Decision Tree J48.

COMPARISON OF NAIVE BAYES, SIMPLE LOGISTIC, RANDOM FOREST, AND DECISION TREE J48 ALGORITHMS IN CLASSIFYING PHISHING WEBSITES

Student Name : Nabilla Artamevia

Student Identity Number : 202151044

Supervisor :

1. Alif Catur Murti, S.Kom, M.Kom
2. Rizkysari Meimaharani, S.Kom, M.Kom

ABSTRACT

Phishing is one of the most significant cybersecurity threats that aims to steal users' private information through fraudulent websites. This research aims to compare the performance of four machine learning algorithms, namely Naive Bayes, Simple Logistic, Random Forest, and Decision Tree J48, in detecting phishing websites. Additionally, this study identifies the most significant features contributing to phishing classification to enhance detection effectiveness.

The research methodology involved using a dataset containing 11,055 website entries processed with the SMOTE (Synthetic Minority Over-sampling Technique) method to address class imbalance. Algorithm evaluation was conducted using 10-Fold Cross Validation and performance metrics such as accuracy, precision, recall, F1 Score, MCC, and ROC AUC. Each algorithm was tested to determine the best algorithm for phishing detection based on these performance metrics.

The results indicate that the Random Forest algorithm achieves the best performance, with 97.60% accuracy, 97.14% precision, 98.08% recall, and 97.61% F1 Score. The SSLfinal_State feature was identified as the most significant attribute in detecting phishing websites. This research is expected to serve as a reference for developing more reliable and accurate phishing detection systems.

Keywords: *Phishing, Naive Bayes, Simple Logistic, Random Forest, Decision Tree J48.*

DAFTAR ISI

HALAMAN PERSETUJUAN.....	iii
HALAMAN PENGESAHAN.....	iv
PERNYATAAN KEASLIAN.....	v
KATA PENGANTAR	vi
RINGKASAN	vii
ABSTRACT.....	viii
DAFTAR ISI.....	ix
DAFTAR GAMBAR	xii
DAFTAR TABEL.....	xiii
DAFTAR SIMBOL.....	xiv
DAFTAR LAMPIRAN.....	xv
DAFTAR ISTILAH DAN SINGKATAN	xvi
BAB I PENDAHULUAN	1
1.1. Latar Belakang.....	1
1.2. Perumusan Masalah.....	2
1.3. Batasan Masalah.....	3
1.4. Tujuan.....	4
1.5. Sistematika Penulisan.....	4
BAB II TINJAUAN PUSTAKA.....	7
2.1. Penelitian Terkait	7
2.2. GAP Penelitian.....	9
2.3. Landasan Teori.....	10
2.3.1. Definisi Phishing.....	10
2.3.2. Algoritma Machine Learning.....	10
2.3.2.1. Naïve Bayes	11
2.3.2.2. Simple Logistic Regresion	12
2.3.2.3. Random Forest	13
2.3.2.4. Decision Tree J48.....	15
BAB III METODOLOGI.....	19
3.1. Sumber Dataset	19
3.2. Deskripsi Dataset	19
3.2.1. Fitur Website Phishing Berbasis Address Bar	21
3.2.1.1. Menggunakan Alamat IP (<i>having_IP_Address</i>)	21

3.2.1.2. URL Panjang (<i>URL_Length</i>).....	21
3.2.1.3. Menggunakan Layanan Pemendekan URL (<i>Shortening_Service</i>)	22
3.2.1.4. URL memiliki Simbol “@” (<i>having_At_Symbol</i>).....	22
3.2.1.5. Mengarahkan ulang menggunakan “//” (<i>double_slash_redirecting</i>)	23
3.2.1.6. Terdapat Simbol “-” Pada URL (<i>Prefix_Suffix</i>)	23
3.2.1.7. Sub Domain dan Multi Sub Domain (<i>having_Sub_Domain</i>)	23
3.2.1.8. HTTPS dan Sertifikat HTTPS (<i>SSLfinal_State</i>)	24
3.2.1.9. Durasi Pendaftaran Domain (<i>Domain_registration_length</i>)	24
3.2.1.10. Favicon (<i>Favicon</i>)	24
3.2.1.11. Menggunakan Port Non-Standar (<i>port</i>)	25
3.2.1.12. Terdapat Token HTTPS Pada Bagian Domain (<i>HTTPS_token</i>)	26
3.2.2. Fitur Website Phishing Berbasis Abnormal	26
3.2.2.1. Jumlah Request URL (<i>Request_URL</i>)	26
3.2.2.2. Anchor Yang Mengarah Ke Domain Lain (<i>URL_of_Anchor</i>)	26
3.2.2.3. Link dalam tag <Meta>, <Script> dan <Link> (<i>Links_in_tags</i>)..	27
3.2.2.4. Status <i>Server Form Handler</i> (<i>SFH</i>)	27
3.2.2.5. Mengirimkan Informasi ke Email (<i>Submitting_to_email</i>)	27
3.2.2.6. Ketidaknormalan URL (<i>Abnormal_URL</i>)	28
3.2.3. Fitur Website Phishing Berbasis HTML dan JavaScript	28
3.2.3.1. Jumlah Beberapa Kali Website Forwarding / Redirecting (<i>Redirect</i>)	28
3.2.3.2. Kustomisasi Status Bar (<i>on_mouseover</i>)	28
3.2.3.3. Menonaktifkan Klik Kanan (<i>RightClick</i>).....	29
3.2.3.4. Menggunakan Pop-up Window (<i>popUpWidnow</i>).....	29
3.2.3.5. Pengalihan IFrame (<i>Iframe</i>)	29
3.2.4. Fitur Website Phishing Berbasis Domain	30
3.2.4.1. Usia Domain (<i>age_of_domain</i>).....	30
3.2.4.2. DNS Record (<i>DNSRecord</i>)	30
3.2.4.3. Website Traffic (<i>web_traffic</i>).....	30
3.2.4.4. PageRank (<i>Page_Rank</i>)	31
3.2.4.5. Indeks Google (<i>Google_Index</i>).....	31
3.2.4.6. Jumlah Link yang Menunjuk ke Halaman (<i>Links_pointing_to_page</i>)	31
3.2.4.7. Laporan Statistik PhishTank dan StopBadware (<i>Statistical_report</i>)	32
3.3. Implementasi Algoritma.....	32

3.4. Preprocessing Data.....	34
3.5. K-Fold Cross Validation	34
3.6. Metrik Evaluasi	35
3.7. Identifikasi Atribut.....	36
3.7.1. Naive Bayes	36
3.7.2. Simple Logistic	37
3.7.3. Random Forest	37
3.7.4. Decision Tree J48.....	37
BAB IV HASIL DAN PEMBAHASAN	39
4.1. Analisis Dataset.....	39
4.2. Preprocessing	41
4.3. Evaluasi Kinerja Algoritma.....	45
4.3.1. Hasil Evaluasi Naive Bayes	45
4.3.2. Hasil Evaluasi Simple Logistic	46
4.3.3. Hasil Evaluasi Random Forest	47
4.3.4. Hasil Evaluasi Decision Tree J48	48
4.4. Perbandingan Hasil Evaluasi Algoritma	49
4.5. Identifikasi Atribut Paling Relevan pada Model Terbaik	50
4.6. Pembuatan Sistem Prediksi	52
4.6.1. Desain Antarmuka Pengguna (UI) dengan Tkinter.....	52
4.6.2. Implementasi Sistem Prediksi	53
4.7. Uji Coba Sistem dan Analisis Hasil	54
4.8. Pembahasan Hasil Penelitian	57
BAB V PENUTUP.....	59
5.1. Kesimpulan.....	59
5.2. Saran	59
DAFTAR PUSTAKA	61
LAMPIRAN 1	63
LAMPIRAN 2	67
LAMPIRAN 3	68
LAMPIRAN 4	71
LAMPIRAN 5.....	Error! Bookmark not defined.
BIODATA PENULIS	72

DAFTAR GAMBAR

Gambar 3.1. Persebaran Data Fitur Dataset (1)	20
Gambar 3.2. Persebaran Data Fitur Dataset (2)	21
Gambar 3.3. Implementasi Algoritma.....	32
Gambar 4.1. Heatmap Korelasi Antar Fitur	40
Gambar 4.2. Distribusi Dataset Sebelum SMOTE.....	42
Gambar 4.3. Distribusi Dataset Sesudah SMOTE	42
Gambar 4.4. Persebaran Data Sebelum SMOTE	43
Gambar 4.5. Persebaran Data Setelah SMOTE	44
Gambar 4.6. Grafik ROC Model Naive Bayes	46
Gambar 4.7. Grafik ROC Model Simple Logistic	47
Gambar 4.8. Grafik ROC Model Random Forest	48
Gambar 4.9. Grafik ROC Model Decision Tree	49
Gambar 4.10. WireFrame Desain Antarmuka.....	52
Gambar 4.11. Flowchart Sistem Prediksi.....	53
Gambar 4.12. Hasil Implementasi Sistem.....	54
Gambar 4.13 Hasil Prediksi https://sunan.umk.ac.id/	57

DAFTAR TABEL

Tabel 3. 1. Port dan Status Rekomendasi.....	25
Tabel 3.2. Tabel Proses 10-Fold Cross Validation	35
Tabel 3.3. Model Confusion Matrix.....	35
Tabel 4.1. Analisis Fitur Dataset.....	39
Tabel 4.2. Confusion Matrix Model Naive Bayes	45
Tabel 4.3. Nilai Metrik Evaluasi Model Naive Bayes	45
Tabel 4.4. Confusion Matrix Model Simple Logistic	46
Tabel 4.5. Nilai Metrik Evaluasi Model.....	46
Tabel 4.6. Confusion Matrix Model Random Forest	47
Tabel 4.7. Nilai Metrik Evaluasi Model.....	47
Tabel 4.8. Confusion Matrix Model Decision Tree	48
Tabel 4.9. Nilai Metrik Evaluasi Model Decision Tree	48
Tabel 4.10. Perbandingan Hasil Metrik Evaluasi.....	49
Tabel 4.11. Hasil Perhitungan <i>Feature Importance</i>	50
Tabel 4.12 Uji Coba Sistem	54
Tabel 4.13 Hasil Analisis https://sunan.umk.ac.id/	55

DAFTAR SIMBOL

Simbol	Keterangan
P	Probabilitas atau kemungkinan yang terjadi.
Σ	Operator penjumlahan, digunakan untuk menghitung total nilai dalam suatu kumpulan data.
$\in$	Digunakan untuk menunjukkan bahwa suatu elemen termasuk dalam suatu himpunan.
argmax	Operator yang memaksimalkan fungsi tertentu.
$\prod$	Operasi perkalian berulang dari sekumpulan elemen dalam rentang tertentu.
$ $	Simbol untuk probabilitas bersyarat, berarti "dengan asumsi".
$=$	Simbol kesetaraan, menunjukkan bahwa kedua sisi memiliki nilai yang sama.
β_0	Intersep atau konstanta dalam model regresi logistik.
β_1	Koefisien regresi yang menunjukkan pengaruh variabel X terhadap Y.
$-$	Operasi pengurangan.
$+$	Operasi penjumlahan.
θ	Parameter yang digunakan untuk mengestimasi nilai dalam model machine learning.
$\neq$	Tidak sama dengan, digunakan untuk menyatakan ketidaksamaan antara dua nilai.
Θ	Parameter dalam model statistik atau machine learning.
$\leq$	Lebih kecil atau sama dengan, digunakan untuk membandingkan dua nilai.
ρ	Koefisien korelasi yang mengukur hubungan antara dua variabel.
$\log$	Logaritma, operasi matematika untuk menemukan eksponen basis tertentu yang menghasilkan nilai tertentu.

DAFTAR LAMPIRAN

Lampiran 1	Lampiran Bimbingan Dosen Pendamping 1 (1)	63
	Lampiran Bimbingan Dosen Pendamping 1 (2)	64
	Lampiran Bimbingan Dosen Pendamping 2 (1)	65
	Lampiran Bimbingan Dosen Pendamping 2 (2)	66
Lampiran 2	Hasil Turnitin	67
Lampiran 3	Lampiran Revisi Sidang Penguji 1	68
	Lampiran Revisi Sidang Penguji 2	69
	Lampiran Revisi Sidang Penguji 3	70
Lampiran 4	Poster	71

DAFTAR ISTILAH DAN SINGKATAN

SMOTE	: Synthetic Minority Over-sampling Technique. Teknik untuk menangani ketidakseimbangan kelas dalam dataset dengan membuat data sintetis pada kelas minoritas.
ROC AUC	: Receiver Operating Characteristic Area Under the Curve. Metrik evaluasi untuk mengukur kemampuan model membedakan antara kelas positif dan negatif.
MCC	: Matthews Correlation Coefficient. Koefisien yang mengukur hubungan antara prediksi model dan kelas sebenarnya.
K-Fold Cross Validation	: Teknik validasi model dengan membagi data ke dalam beberapa subset untuk pelatihan dan pengujian bergantian.
Phishing	: Aktivitas kriminal yang bertujuan mencuri informasi pribadi pengguna melalui situs web palsu.
SSL	: Secure Sockets Layer. Teknologi keamanan untuk mengenkripsi data yang dikirimkan antara server dan browser.
DNS	: Domain Name System. Sistem yang menghubungkan nama domain dengan alamat IP.
URL	: Uniform Resource Locator. Alamat yang digunakan untuk mengakses sumber daya di internet.
Redirect	: Mekanisme untuk mengarahkan pengguna dari satu URL ke URL lainnya.
HTTPS	: Hypertext Transfer Protocol Secure. Protokol keamanan untuk transfer data melalui internet.
F1 Score	: Rata-rata harmonis antara presisi dan recall, digunakan sebagai metrik kinerja model klasifikasi.
HTML	: Hypertext Markup Language. Bahasa markup untuk membuat struktur halaman web.
IP	: Internet Protocol. Protokol komunikasi untuk mengidentifikasi perangkat di jaringan.
WHOIS	: Layanan pencarian yang memberikan informasi mengenai kepemilikan nama domain.
PageRank	: PageRank
Anchor Tag	: Elemen HTML (<a>) yang digunakan untuk membuat hyperlink pada halaman web.

Iframe	: Elemen HTML untuk menyisipkan dokumen atau konten dari halaman lain ke dalam halaman web.
Favicon	: Ikon grafis kecil yang muncul di tab browser, biasanya merepresentasikan identitas situs web.
SMB	: Server Message Block. Protokol jaringan yang memungkinkan berbagi file, printer, dan sumber daya lainnya.
PCA	: Principal Component Analysis. Teknik pengurangan dimensi dalam dataset untuk meningkatkan efisiensi algoritma.
JavaScript	: Bahasa pemrograman yang digunakan untuk membuat halaman web interaktif.