

**KLASIFIKASI *MALICIOUS WEB CONTENT* DENGAN
ALGORITME RANDOM FOREST DAN
K-NEAREST NEIGHBORS**

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Teknik Komputer



disusun oleh

RONI SEPTIAN

18.83.0330

Kepada

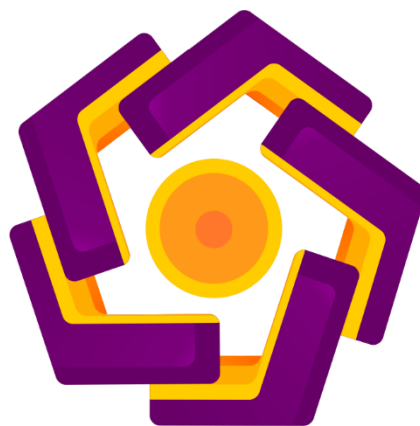
**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA**

2022

**KLASIFIKASI MALICIOUS WEB CONTENT DENGAN
ALGORITME RANDOM FOREST DAN
K-NEAREST NEIGHBORS**

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Teknik Komputer



disusun oleh

RONI SEPTIAN

18.83.0330

Kepada:

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA**

2022

HALAMAN PERSETUJUAN
SKRIPSI

KLASIFIKASI MALICIOUS WEB CONTENT DENGAN ALGORITME
RANDOM FOREST DAN K-NEAREST NEIGHBORS

SKRIPSI

yang disusun dan diajukan oleh
Roni Septian
18.83.0330

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 21 Maret 2023

Dosen Pembimbing.



Wahid Miftahul Ashari, S.Kom., M.T.
NIK. 190302452



CS Dipindai dengan CamScanner

HALAMAN PENGESAHAN
SKRIPSI

**KLASIFIKASI MALICIOUS WEB CONTENT DENGAN ALGORITME RANDOM
FOREST DAN K-NEAREST NEIGHBORS**

yang disusun dan diajukan oleh

Roni Septian

18.83.0330

Telah dipertahankan di depan Dewan Penguji
pada tanggal 21 Maret 2023

Susunan Dewan Penguji

Nama Penguji

Melwin Syafrizal, S.Kom., M.Eng.

NIK. 190302105

Banu Santoso, S.T., M.Eng

NIK. 190302327

Wahid Miftahul Ashari, S.Kom., M.T

NIK. 190302452

Tanda Tangan



Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 21 maret 2023

DEKAN FAKULTAS ILMU KOMPUTER



Hanif Al Fatta, S.Kom., M.Kom.

NIK 190307096

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Roni Septian
NIM : 18.83.0330

Menyatakan bahwa Skripsi dengan judul berikut:

KLASIFIKASI MALICIOUS WEB CONTENT DENGAN ALGORITME RANDOM FOREST DAN K-NEAREST NEIGHBORS

Dosen Pembimbing : Wahid Miftahul Ashari, S.Kom., M.T.

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 Maret 2023
Yang Menyatakan,



10000
REPUBLIK INDONESIA
METERAI
TEMPEL
637C/A/X554335257

Roni Septian

KATA PENGANTAR

Alhamdulillah, puji dan syukur selalu kita panjatkan kepada Allah SWT karena dengan ridhanya penulis dapat menyelesaikan penyusunan skripsi ini. Adapun judul skripsi yang diajukan adalah “*KLASIFIKASI MALICIOUS WEB CONTENT DENGAN ALGORITME RANDOM FOREST DAN K-NEAREST NEIGHBOR*”. Skripsi ini ditunjukkan untuk memenuhi syarat kelulusan mata kuliah skripsi di Fakultas Ilmu Komputer Universitas AMIKOM Yogyakarta.

Dibutuhkan usaha yang keras dalam penyelesaian skripsi ini. Skripsi ini tidak akan selesai tanpa orang-orang disekeliling saya yang membantu dan mendukung. Terimakasih saya sampaikan kepada:

1. Prof.Dr.M.Suyanto, M.M selaku Rektor Universitas AMIKOM Yogyakarta.
2. Wahid Miftahul Ashari, S.Kom., M.T selaku pembimbing yang telah membantu penulis dalam membimbing untuk menyelesaikan skripsi ini.
3. Segenap Dosen Fakultas Ilmu Komputer yang telah memberikan ilmu yang bermanfaat selama menjalani masa studi.
4. Serta semua pihak yang tidak bisa disebutkan satu persatu.

Akhir kata penulis menyadari bahwa tidak ada yang sempurna, penulis masih melakukan kesalahan dalam proses penyusunan skripsi ini. Oleh karena itu, penulis meminta maaf atas kersalahan yang dilakukan penulis.

Peneliti berharap semoga skripsi ini dapat bermanfaat bagi pembaca dan dijadikan referensi demi pengembangan yang lebih baik. Semoga Allah SWT, senantiasa melimpahkan Rahmat dan rida-Nya kepada kita semua

Yogyakarta, 19 Maret 2023

Penulis

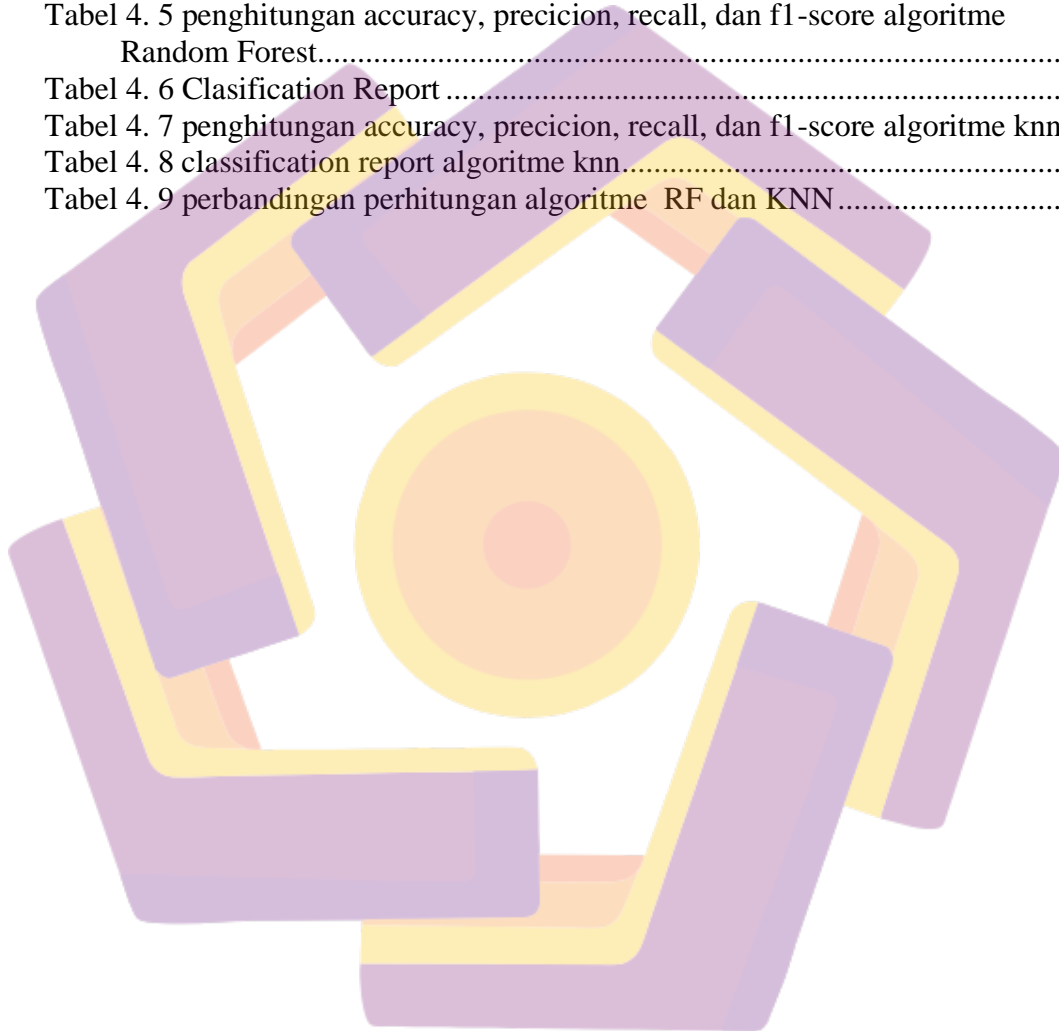
DAFTAR ISI

HALAMAN PERSETUJUAN.....	ii
HALAMAN PENGESAHAN	iii
.....	iv
HALAMAN PERNYATAAN KEASLIAN SKRIPSI	iv
KATA PENGANTAR	v
DAFTAR ISI.....	vi
DAFTAR TABEL.....	viii
DAFTAR GAMBAR	ix
DAFTAR LAMBANG DAN SINGKATAN	x
DAFTAR ISTILAH	xi
INTISARI	xiii
ABSTRACT.....	xiv
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah.....	2
1.3 Batasan Masalah.....	3
1.4 Tujuan Penelitian	3
1.5 Manfaat Penelitian	4
BAB II TINJAUAN PUSTAKA	5
2.1 Studi Literatur.....	5
2.2 Rangkuman Keaslian Penelitian	12
2.3 Perbedaan Penelitian.....	12
2.4 Kemajuan.....	12
2.5 Dasar Teori	12
2.6 Algoritma <i>Random Forest</i>	12
2.7 Algoritma <i>K-Nearest Neighbors (KNN)</i>	14
2.8 <i>Malicious Web Content</i>	15
2.9 <i>Phising</i>	15
2.10 <i>Confusion Matrix</i>	15
2.11 <i>ROC AUC (Receiver Operating Characteristic Area Under the Curve)</i>	17
2.12 <i>Classification Report</i>	17
2.13 <i>Dataset</i>	18
BAB III METODE PENELITIAN	19
3.1 Alur Penelitian	19
3.2 Kerangka Penelitian.....	19
3.2.1 Pemilihan Topik.....	20
3.2.2 Rumusan Masalah.....	20
3.2.3 Penentuan Jenis Data	20
3.2.4 Pengumpulan Data	21
3.2.5 Pembersihan Data	22
3.2.6 Pemilihan Algoritma Klasifikasi.....	22

3.2.7	Pembuatan Model	22
3.2.8	Evaluasi Model.	23
3.2.9	Penyajian Hasil,	23
3.3	Pengembangan model klasifikasi dengan RF dan KNN.....	24
3.3.1	Load dataset	26
3.3.2	Exploratory Data Analysis (EDA)	26
3.3.4	Preprocessing	27
3.3.5	Feature extraction.....	27
3.3.6	Split data	28
3.3.7	Data train dan Data test.....	28
3.3.8	Algoritma Random Forest dan <i>K-Nearest Neighbors</i> (KNN)	29
3.3.9	Prediction Model.....	29
3.3.10	Evaluasi.....	30
3.3.11	Classification Report.....	31
3.4	Alat dan Bahan.....	31
3.4.1	Data Penelitian,	31
3.4.2	Alat/instrument	31
BAB IV	HASIL DAN PEMBAHASAN	33
4.	Implementasi	33
4.1	Proses klasifikasi dengan algoritme Random Forest.....	33
4.2	Proses klasifikasi dengan algoritme K-Nearest Neighbors (KNN).....	47
4.3	Analisa Data Klasifikasi	60
4.4	Kesimpulan Dari Hasil Penelitian.....	66
BAB V	PENUTUP	69
5.1	Kesimpulan.....	69
5.2	Saran	70
REFERENSI	71

DAFTAR TABEL

Tabel 4. 1 library yang digunakan	33
Tabel 4. 2 Menjelaskan isi dari dataset.....	35
Tabel 4. 3 Library Klasifikasi	47
Tabel 4. 4 Menampilkan data pada dataset.....	49
Tabel 4. 5 penghitungan accuracy, precision, recall, dan f1-score algoritme Random Forest.....	62
Tabel 4. 6 Clasification Report	62
Tabel 4. 7 penghitungan accuracy, precision, recall, dan f1-score algoritme knn.....	65
Tabel 4. 8 classification report algoritme knn.....	65
Tabel 4. 9 perbandingan perhitungan algoritme RF dan KNN.....	67



DAFTAR GAMBAR

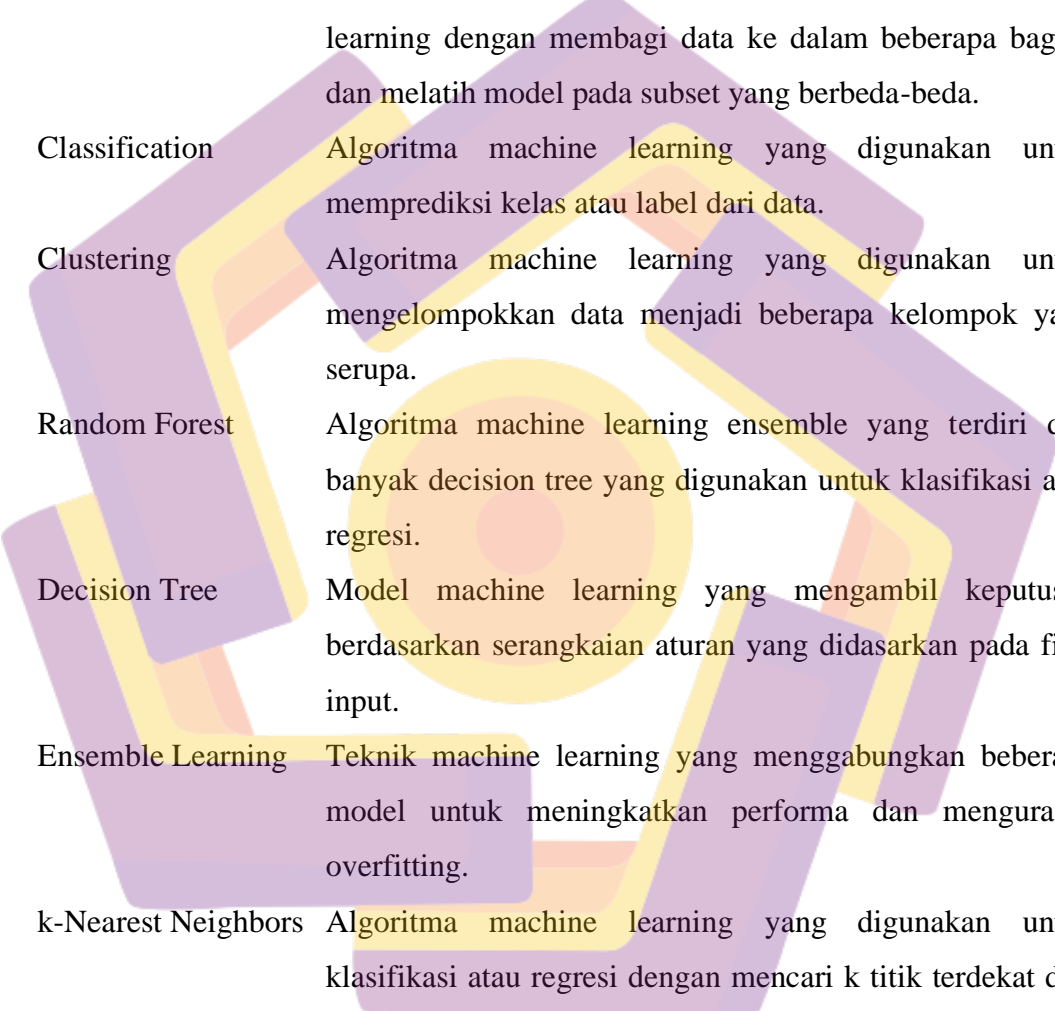
Gambar 2. 1 Struktur Random forestm forest.....	13
Gambar 2. 2 struktur KNN.....	14
Gambar 3. 1 kerangka penelitian	19
Gambar 3. 2 Pengembangan model klasifikasi dengan RF dan KNN.....	24
Gambar 4. 1 import library	33
Gambar 4. 2 Load Dataset.....	34
Gambar 4. 3 menampilkan data yang sudah di muat	35
Gambar 4. 4 Memeriksa Informasi dasar dataframe.....	37
Gambar 4. 5 menampilkan distribusi data dengan pie chart	38
Gambar 4. 6 menampilkan matriks korelasi antar kolom pada DataFrame.....	38
Gambar 4. 7 proses profiling report	39
Gambar 4. 8 overview data	40
Gambar 4. 9 heatmap pada dataframe.....	40
Gambar 4. 10 onehot encod	41
Gambar 4. 11 Preprocessing data.....	42
Gambar 4. 12 mengevaluasi kinerja model yang dilatih.....	43
Gambar 4. 13 pemrosesan imput.....	44
Gambar 4. 14 menghitung jumlah kemunculan nilai unik pada variable	45
Gambar 4. 15 melatih model klasifikasi	45
Gambar 4. 16 confsusion matrix random forest.....	46
Gambar 4. 17 clasification report random forest	46
Gambar 4. 18 import library	47
Gambar 4. 19 load dataset.....	48
Gambar 4. 20 menampilkan dataset.....	48
Gambar 4. 21 melihat informasi pada dataframe	50
Gambar 4. 22 distribusi data dengan pie chart.....	51
Gambar 4. 23 korelasi antar variabel	52
Gambar 4. 24 profiling report	53
Gambar 4. 25 overview data	53
Gambar 4. 26 heatmap data.....	54
Gambar 4. 27 onehot encode.....	54
Gambar 4. 28preprocessing data KNN	55
Gambar 4. 29 evaluasi model.....	56
Gambar 4. 30 membagi data latih dan data test	57
Gambar 4. 31 menghitung fungsi frekuensi nilai pada dataframe	58
Gambar 4. 32 inisiasi onjek	58
Gambar 4. 33 confusionsion matirx algoritme knn.....	59
Gambar 4. 34 classification report algoritme knn.....	59
Gambar 4. 35 confusion matrix algoritme random forest.....	61
Gambar 4. 36 ROC-AUC algoritme random forest	63
Gambar 4. 37 confusion matrix algoritme knn	64
Gambar 4. 38 ROC-AUC algoritme knn	66
Gambar 4. 39 ROC-AUC algoritme RF dan KNN	67

DAFTAR LAMBANG DAN SINGKATAN

K-NN	<i>K-Nearest Neighbors</i>
RF	<i>Random Forest</i>
URL	<i>Uniform Resource Locator</i>
FPR	<i>False Positive Rate</i>
FNR	<i>False Negative Rate</i>
TP	<i>True Positive</i>
TN	<i>True Negative</i>
FP	<i>False Positive</i>
FN	<i>False Negative</i>
ROC	<i>Receiver Operating Characteristic</i>
AUC	<i>Area Under the Curve</i>
SVM	<i>Support Vector Machine</i>
PRECISION	<i>Presisi</i>
RECALL	<i>Recall</i>
F1-SCORE	<i>F1 Score</i>
ACCURACY	<i>Akurasi</i>
OVERFITTING	<i>Overfitting</i>
AI	<i>Artificial Intelligence</i>
CV	<i>Cross Validation</i>
Ensemble Learning	<i>Pembelajaran Gabungan</i>

DAFTAR ISTILAH

Malware	Perangkat lunak jahat
Deep Learning	Jenis machine learning yang menggunakan jaringan saraf tiruan untuk mempelajari pola yang kompleks pada data.
Phising	penipuan online yang dilakukan dengan maksud untuk memancing informasi sensitif dari korban, seperti username, password, atau informasi keuangan.
Scam	jenis penipuan yang biasanya dilakukan dengan menggunakan metode online atau internet.
Konten ilegal	Konten yang melanggar hukum, seperti konten yang berisi pornografi anak, obat-obatan terlarang, atau konten yang melanggar hak cipta
Machine Learning	Disiplin ilmu yang memungkinkan komputer untuk belajar dari data dan melakukan prediksi atau tindakan tertentu tanpa di-program secara eksplisit.
Data Set	Kumpulan data yang digunakan untuk melatih model machine learning.
Label	Nilai target yang harus diprediksi oleh model machine learning.
Feature	Variabel input yang digunakan dalam model machine learning.
Training Set	Bagian dari data set yang digunakan untuk melatih model machine learning.
Validation Set	Bagian dari data set yang digunakan untuk mengevaluasi performa model machine learning selama pelatihan.
Test Set	Bagian dari data set yang digunakan untuk menguji performa model machine learning setelah pelatihan.
Model	Representasi matematika dari hubungan antara fitur dan label dalam data.



Algorithm	Metode atau teknik yang digunakan untuk membangun model machine learning.
Overfitting	Kondisi di mana model machine learning terlalu kompleks sehingga dapat mempelajari karakteristik yang spesifik pada data training, tetapi tidak dapat diterapkan pada data baru.
Cross-Validation	Teknik untuk mengevaluasi performa model machine learning dengan membagi data ke dalam beberapa bagian dan melatih model pada subset yang berbeda-beda.
Classification	Algoritma machine learning yang digunakan untuk memprediksi kelas atau label dari data.
Clustering	Algoritma machine learning yang digunakan untuk mengelompokkan data menjadi beberapa kelompok yang serupa.
Random Forest	Algoritma machine learning ensemble yang terdiri dari banyak decision tree yang digunakan untuk klasifikasi atau regresi.
Decision Tree	Model machine learning yang mengambil keputusan berdasarkan serangkaian aturan yang didasarkan pada fitur input.
Ensemble Learning	Teknik machine learning yang menggabungkan beberapa model untuk meningkatkan performa dan mengurangi overfitting.
k-Nearest Neighbors	Algoritma machine learning yang digunakan untuk klasifikasi atau regresi dengan mencari k titik terdekat dari data yang sedang diprediksi.
Nearest Neighbors	Poin data terdekat dari data yang sedang diprediksi.
Data Preprocessing	Proses mengubah data mentah menjadi bentuk yang dapat digunakan untuk melatih dan menguji model k-NN.

INTISARI

Tujuan dari penelitian ini adalah untuk mengklasifikasi dataset phising menggunakan algoritma Random Forest dan KNN. Dataset yang digunakan berasal dari Kaggle dan telah diproses serta dibagi menjadi data latih dan data uji. Penelitian ini dilakukan karena semakin meningkatnya kerentanan pada aplikasi web dan internet yang dapat dimanfaatkan oleh pihak yang tidak bertanggung jawab dengan menggunakan metode phising, yang dapat berdampak pada kerugian pada pengguna internet dan website.

Pada proses klasifikasi menggunakan Random Forest, dibuat sebuah model dengan data latih dan kemudian digunakan untuk memprediksi label kelas pada data uji. Selanjutnya, akurasi model dihitung dengan membandingkan label kelas prediksi dengan label kelas sebenarnya pada data uji. Hal yang sama dilakukan pada proses klasifikasi menggunakan KNN.

Hasil penelitian menunjukkan bahwa model Random Forest memiliki performa yang lebih baik dalam hal akurasi dan presisi, serta dalam membedakan kelas positif dan negatif pada data pengujian. Hal ini dibuktikan dengan nilai ROC-AUC yang lebih tinggi pada model Random Forest. Selain itu, perhitungan selisih tiap metrik menunjukkan bahwa model Random Forest memiliki skor yang lebih tinggi dalam setiap metrik evaluasi yang digunakan, meskipun perbedaannya bervariasi dari 0,22% hingga 4,28%. Rata-rata selisih tiap metrik kemudian dihitung dan dibagi menjadi 4, sehingga didapatkan hasil 2,47%. Berdasarkan keseluruhan metrik evaluasi yang sudah dilakukan, dapat disimpulkan bahwa kinerja model Random Forest lebih baik dibandingkan dengan K-nearest neighbour dalam klasifikasi data.

Kata kunci: *Random Forest, K-Nearest Neighbors (KNN), Confusion Matrix, ROC-AUC, Phising,*

ABSTRACT

The purpose of this study is to classify the phishing dataset using Random Forest and KNN algorithms. The dataset used in this study was obtained from Kaggle and has been processed and divided into training and testing data. This research was conducted due to the increasing vulnerability of web applications and the internet, which can be exploited by irresponsible parties using phishing methods, resulting in losses for internet users and websites.

In the classification process using Random Forest, a model was created using the training data and then used to predict the class label on the testing data. Next, the accuracy of the model was calculated by comparing the predicted class label with the actual class label on the testing data. The same process was done in the classification process using KNN.

The results of the study show that the Random Forest model performs better in terms of accuracy and precision, as well as in distinguishing positive and negative classes on the testing data. This is evidenced by the higher ROC-AUC value on the Random Forest model. In addition, the calculation of each metric difference shows that the Random Forest model has a higher score in each evaluation metric used, although the difference varies from 0.22% to 4.28%. The average difference of each metric is then calculated and divided into 4, resulting in a value of 2.47%. Based on the overall evaluation metrics performed, it can be concluded that the performance of the Random Forest model is better than K-nearest neighbor in classifying data.

Keywords: *Random Forest, K-Nearest Neighbors (KNN), Confusion Matrix, ROC-AUC, Phishing.*