# 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**

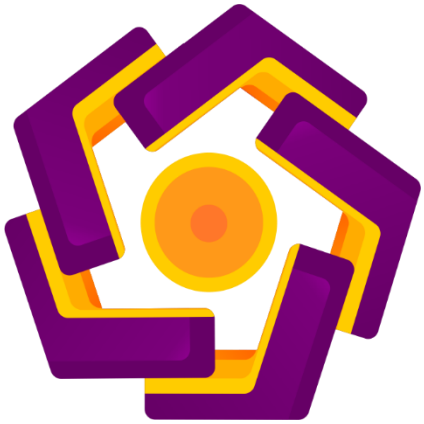
**2022**

# 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**

**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 | Tanda Tangan |
|---|---|
| **Melwin Syafrizal, S.Kom., M.Eng.**<br>**NIK. 190302105** | |
| **Banu Santoso, S.T., M.Eng**<br>**NIK. 190302327** | |
| **Wahid Miftahul Ashari, S.Kom., M.T**<br>**NIK. 190302452** | |

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. 190302096**

## 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,


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 ditunjukan 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

# DAFTAR TABEL

# DAFTAR GAMBAR

# DAFTAR LAMBANG DAN SINGKATAN

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

# 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 illegal | 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 metriks 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 metriks kemudian dihitung dan dibagi menjadi 4, sehingga didapatkan hasil 2,47%. Berdasarkan keseluruhan metriks 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.*