METODE FITUR SELEKSI UNTUK PENINGKATAN KINERJA KLASIFIKASI PADA PROSES DETEKSI WEB PHISING

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana

Program Studi Teknik Komputer

diajukan oleh

DWIKY ALFIAN TAMA

18.83.0297

Kepada

PROGRAM SARJANA

PROGRAM STUDI TEKNIK KOMPUTER

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2022

METODE FITUR SELEKSI UNTUK PENINGKATAN KINERJA KLASIFIKASI PADA PROSES DETEKSI WEB PHISING

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana

Program Studi Teknik Komputer

diajukan oleh

DWIKY ALFIAN TAMA

18.83.0297

Kepada

PROGRAM SARJANA

PROGRAM STUDI TEKNIK KOMPUTER

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2022

HALAMAN PERSETUJUAN

SKRIPSI

METODE FITUR SELEKSI UNTUK PENINGKATAN KINERJA KLASIFIKASI PADA PROSES DETEKSI WEB PHISING

yang disusun dan diajukan oleh

Dwiky Alfian Tama

18.83.0297

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 24 Maret 2022

Dosen Pembimbing,
ii

Anggit Ferdita Nugraha,S.T.,M.Eng
NIK. 190302480

HALAMAN PENGESAHAN

SKRIPSI

METODE FITUR SELEKSI UNTUK PENINGKATAN KINERJA KLASIFIKASI PADA PROSES DETEKSI WEB PHISING

yang disusun dan diajukan oleh

Dwiky Alfian Tama

18.83.0297

Telah dipertahankan di depan Dewan Penguji
pada tanggal 21 Juni 2022

Susunan Dewan Penguji

Nama Penguji	Tanda Tangan
Jeki Kuswanto, M.Kom **NIK. 190302456**	_____
Supriatin, M.Kom **NIK. 190302239**	_____
Anggit Ferdita Nugraha, S.T., M.Eng **NIK. 190302480**	_____

Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 21 Juni 2022

DEKAN FAKULTAS ILMU KOMPUTER

Hanif Al Fatta,S.Kom., M.Kom.
NIK. 190302096

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Dwiky Alfian Tama
NIM : 18.83.0297

Menyatakan bahwa Skripsi dengan judul berikut:

Metode Fitur Seleksi Untuk Peningkatan Kinerja Klasifikasi Pada Proses Deteksi Web Phising

Dosen Pembimbing : Anggit Ferdita Nugraha,S.T.,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 Juni 2022

Yang Menyatakan,

Dwiky Alfian Tama

HALAMAN PERSEMBAHAN

Dengan rasa syukur yang mendalam, dengan telah diselesaikannya skripsi ini, penulis mempersembahkannya kepada:

1. Orangtua penulis yang selalu mendoakan penulis sehingga dapat menyelesaikan skripsi tanpa adanya gangguan.
2. Dosen pembimbing saya, Anggit Ferdita Nugraha,S.T,.M.Eng yang selalu memacu saya untuk menyelesaikan skripsi yang terkadang mentertawakan penulis karena kisah perjalanan hidup penulis.
3. Teman-teman penulis yang selalu ada untuk penulis dikala suka ataupun duka.
4. Serta semua pihak yang tidak bisa penulis sebutkan satu persatu.

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 "Metode Fitur Seleksi Untuk Peningkatan Kinerja Klasifikasi Pada Proses Deteksi Web Phising". Skripsi ini diajukan 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 di sekeliling saya yang membantu dan mendukung.Terimakasih saya sampaikan kepada:

1. Prof.Dr.M.Suyanto, M.M selaku Rektor Universitas AMIKOM Yogyakarta.
2. Anggit Ferdita Nugraha,S.T.,M.Eng selaku dosen pembimbing yang telah memberikan bimbingan kepada penulis.
3. Segenap Dosen Fakultas Ilmu Komputer yang telah memberikan ilmu yang bermanfaat selama menjalani studi.
4. Semua pihak yang telah membantu dan tidak dapat disebutkan satu persatu

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

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

Yogyakarta, 11 Juli 2022

Penulis

DAFTAR ISI

HALAMAN JUDUL i
HALAMAN PERSETUJUAN i
HALAMAN PENGESAHAN i
HALAMAN PERNYATAAN KEASLIAN SKRIPSI ii
HALAMAN PERSEMBAHAN iii
KATA PENGANTAR iv
DAFTAR ISI v
DAFTAR TABEL vii
DAFTAR GAMBAR viii
DAFTAR LAMBANG DAN SINGKATAN ix
DAFTAR ISTILAH x
INTISARI xi
ABSTRACT xii
BAB I PENDAHULUAN 1
1.1 Latar Belakang 1
1.2 Perumusan masalah 3
1.3 Tujuan Penelitian 3
1.4 Batasan Masalah 3
1.5 Manfaat Penelitian 3
BAB II TINJAUAN PUSTAKA 4
2.1 Literatur Review 4
2.2 Landasan Teori 8
2.2.1 Web Phising 8
2.2.2 Artificial Intelligence 8
2.2.3 Machine Learning 9
2.2.3.1 Supervised Learning 9
2.2.3.2 Unsupervised Learning 10
2.2.3.3 Semi Supervised Learning 10
2.2.3.4 Reinforcement Learning 10
2.2.4 Klasifikasi 11
2.2.5 Decision Tree 12

2.2.6 Fitur Seleksi.....13
2.2.6.1 Metode Filter.....13
2.2.6.2 Metode Wrapper.....14
2.2.7 Pearson Correlation.....14
2.2.8 Information Gain.....15
2.2.9 Gain Ratio.....17
2.2.10 Chi-Square.....17
2.2.11 Recursive Feature Elimination.....18
2.2.12 Evaluasi.....18
2.2.4.1 Akurasi.....19
2.2.4.2 Presisi.....19
2.2.4.3 Recall.....20
2.2.4.4 F1 Score.....20
BAB III METODOLOGI PENELITIAN.....21
3.1 Kebutuhan Alat dan Bahan.....21
3.2 Langkah Penelitian.....21
3.2.1 Data Acquisition.....22
3.2.2 Feature Engineering.....28
3.2.3 Klasifikasi.....29
3.2.4 Evaluasi.....29
BAB IV HASIL DAN PEMBAHASAN.....31
4.1 Implementasi.....31
4.2 Pengujian.....37
BAB V KESIMPULAN DAN SARAN.....45
5.1 Kesimpulan.....45
5.2 Saran.....45
DAFTAR PUSTAKA.....46

DAFTAR TABEL

Tabel 2.1. Referensi Penelitian 4
Tabel 3.1. Fitur Web Phising 22
Tabel 4.1. Nilai Korelasi Pearson Correlation 33
Tabel 4.2. Nilai Information Gain 34
Tabel 4.3. Nilai Gain Ratio 35
Tabel 4.4. Nilai Chi-Square 35
Tabel 4.5. Nilai RFE 37
Tabel 4.6. Akurasi 44

DAFTAR GAMBAR

Gambar 2.1. Bagian Artificial Intelligence 9
Gambar 2.2. Kategori Machine Learning 9
Gambar 2.3. Teknik Klasifikasi 11
Gambar 2.4. Binary Classification 11
Gambar 2.5. Multiclass Classification 12
Gambar 2.6. Struktur Decision Tree 12
Gambar 2.7. Alur Metode Filter 13
Gambar 2.8. Alur Metode Wrapper 14
Gambar 2.9. Confusion Metrik 19
Gambar 3.1. Alur Penelitian 21
Gambar 4.1. Phising Dataset 31
Gambar 4.2. Indeks Pertama Dataset 32
Gambar 4.3. Nilai Korelasi Pearson Correlation 34
Gambar 4.4. Nilai Hasil RFE 36
Gambar 4.5. Ranking RFE 36
Gambar 4.6. Fitur Bernilai True 36
Gambar 4.7. Confusion Matrix Klasifikasi Tunggal Decision Tree 38
Gambar 4.8. Akurasi Klasifikasi Tunggal 38
Gambar 4.9. Confusion Matrix Pearson Correlation 39
Gambar 4.10. Akurasi Pearson Correlation 39
Gambar 4.11. Confusion Matrix Information Gain 40
Gambar 4.12. Akurasi Information Gain 40
Gambar 4.13. Confusion Matrix Gain Ratio 41
Gambar 4.14. Akurasi Gain Ratio 41
Gambar 4.15. Confusion Matrix Chi-Square 42
Gambar 4.16. Akurasi Chi-Square 42
Gambar 4.17. Confusion Matrix RFE 43
Gambar 4.18. Akurasi RFE 44

DAFTAR LAMBANG DAN SINGKATAN

URL	Uniform Resource Locator
AI	Artificial Intelligence
RFE	Recursive Feature Elimination
SVM	Support Vector Machines

DAFTAR ISTILAH

Web-phising	website phising
Markov	tahapan pembelajaran pada reinforcement learning
Entropy(S)	hasil entropi sebelum pemisahan
Entropy(S, A)	hasil entropi setelah pemisahan
InfoGain(S, A)	hasil information gain
GainRatio(S, A)	hasil gain ratio
SplitInformation	nilai information yang dipisah

INTISARI

Web-phising menjadi salah satu kejahatan siber yang saat ini terus mengintai pengguna website di dunia maya. Korban dari serangan web-phising tidak menyadari website yang dikunjunginya adalah web-phising, sehingga membuat informasi penting yang dimilikinya bisa dimanfaatkan oleh pencuri data. Web-phising adalah tiruan dari website asli dengan tujuan untuk mengambil informasi dari korbannya. Deteksi web-phising terus dikembangkan untuk meminimalisir adanya kehilangan informasi akibat serangan web-phising. Berbagai penelitian terkait indentifikasi serangan web-phising memanfaatkan machine learning sampai saat ini masih terus dilakukan oleh para peneliti. Namun, sebagian besar dari penelitian yang dilakukan masih menerapkan klasifikasi tunggal dalam mendeteksi adanya serangan web-phishing.

Disisi lain, terdapat mekanisme fitur seleksi guna memungkinkan proses identifikasi web-phising dengan fitur yang dominan terhadap deteksi web-phising. Oleh karena itu, identifikasi fitur yang dominan dengan memanfaatkan mekanisme fitur seleksi menjadi fokus utama dalam penelitian ini. Teknik fitur seleksi populer seperti pearson correlation, information gain, gain ratio, chi-square dan recursive feature elimination menjadi metode yang diuji kinerjanya terhadap proses deteksi web-phising.

Dataset yang diujikan diperoleh dari UCI Dataset Repository, diketahui penggunaan pearson correlation menghasilkan kinerja akurasi sebesar 94.54%, penggunaan information gain sebesar 93.36%, penggunaan gain ratio sebesar 94.30%, penggunaan chi-square sebesar 93.97% dan penggunaan recursive feature elimination sebesar 97.88%. Dari hasil kinerja, diketahui sejumlah 5 fitur yang paling dominan dalam proses identifikasi web-phising yakni: ssl final state, url of anchor, prefix suffix, web traffic dan having sub domain yang bisa digunakan sebagai referensi untuk mengindikasi adanya web-phising.

Kata kunci: fitur seleksi, web phising, machine learning, decision tree

ABSTRACT

Web phishing is one of the cyber-crimes that currently continue to stalk website users in cyberspace. Victims of web-phishing attacks do not realize that the website they are visiting is a web-phishing one, thus making their important information can be used by data thieves. Web-phishing is a clone of the original website to extract information from its victims. Web-phishing detection continues to be developed to minimize information loss due to web-phishing attacks. Various studies related to the identification of web-phishing attacks using machine learning are still being carried out by researchers. However, most of the research conducted still applies a single classification in detecting web-phishing attacks.

On the other hand, there is a feature selection mechanism to enable the identification process of web-phishing with features that are dominant to web-phishing detection. Therefore, the identification of dominant features by utilizing the feature selection mechanism is the main focus of this study. Popular feature selection techniques such as Pearson correlation, information gain, gain ratio, chi-square and recursive feature elimination are the methods that are tested for their performance against the web-phishing detection process.

The dataset tested was obtained from the UCI Dataset Repository, it is known that the use of Pearson correlation produces an accuracy performance of 94.54%, the use of information gain is 93.36%, the use of a gain ratio is 94.30%, the use of chi-square is 93.97% and the use of recursive feature elimination is 97.88%. From the performance results, it is known that there are 5 most dominant features in the web-phishing identification process, namely: SSL final state, URL of anchor, prefix suffix, web traffic and having subdomains that can be used as references to indicate the existence of web-phishing.

Keyword: feature selection, web phishing, machine learning, decision tree