STUDI PENGARUH SELEKSI FITUR DAN HYPERPARAMETER TUNING PADA RANDOM FOREST UNTUK KLASIFIKASI WEB PHISING

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana

Program Studi Informatika

disusun oleh

LUTHFA SOBRIAN PRAMASTA

21.11.3943

Kepada

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

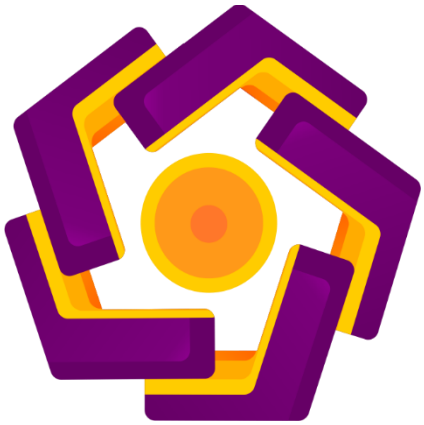
2024

STUDI PENGARUH SELEKSI FITUR DAN HYPERPARAMETER TUNING PADA RANDOM FOREST UNTUK KLASIFIKASI WEB PHISING

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana

Program Studi Informatika



disusun oleh

LUTHFA SOBRIAN PRAMASTA

21.11.3943

Kepada

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2024

HALAMAN PERSETUJUAN

SKRIPSI

STUDI PENGARUH SELEKSI FITUR DAN HYPERPARAMETER TUNING PADA RANDOM FOREST UNTUK KLASIFIKASI WEB PHISING

yang disusun dan diajukan oleh

Luthfa Sobrian Pramasta

21.11.3943

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 17 Desember 2024

Dosen Pembimbing,

Anna Baita, S.Kom., M.Kom.
NIK. 190302290

HALAMAN PENGESAHAN

SKRIPSI

STUDI PENGARUH SELEKSI FITUR DAN HYPERPARAMETER TUNING PADA RANDOM FOREST UNTUK KLASIFIKASI WEB PHISING

yang disusun dan diajukan oleh

Luthfa Sobrian Pramasta

21.11.3943

Telah dipertahankan di depan Dewan Penguji
pada tanggal 17 Desember 2024

Susunan Dewan Penguji

Nama Penguji	Tanda Tangan
Arifiyanto hadinegoro, S.Kom., M. T. **NIK. 190302289**	
Donni Prabowo, S.Kom., M.Kom. **NIK. 190302253**	
Anna Baita, S.Kom., M.Kom. **NIK. 190302290**	

Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 17 Desember 2024

DEKAN FAKULTAS ILMU KOMPUTER

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

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa :Luthfa Sobrian Pramasta
NIM : 21.11.3943

Menyatakan bahwa Skripsi dengan judul berikut:

Studi Pengaruh Seleksi Fitur dan Hyperparameter Tunning Pada Random Forest Untuk Klasifikasi Web Phising

Dosen Pembimbing : Anna Baita, S.Kom., M.Kom.

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, 24 Desember 2024

Yang Menyatakan,

Luthfa Sobrian Pramasta

HALAMAN PERSEMBAHAN

Segala puji dan syukur saya persembahkan kepada Allah subḥānahu wa ta'āla yang selalu memberikan rahmat dan karunia-Nya sehingga saya dapat menyelesaikan naskah skripsi ini dengan lancar dan barokah. Naskah skripsi ini saya persembahkan kepada:

1. Kedua orang tua saya yang selalu memberikan dukungan berupa doa dan semangat.
2. Universitas Amikom Yogyakarta sebagai tempat menimba ilmu melanjutkan studi saya.
3. Bapak dosen pembimbing yang telah membimbing saya dalam penyelesaian naskah skripsi ini.
4. Teman – teman saya yang telah membantu saya dalam pemberian arahan ketika saya membutuhkan bantuan.
5. Seluruh pihak yang telah memberikan kontribusi sekecil apapun dalam proses penyelesaian naskah skripsi ini yang tidak dapat saya sebutkan satu per satu.

KATA PENGANTAR

Dengan puji syukur saya persembahkan kepada Allah subḥānahu wa ta'āla, Tuhan Yang Maha Esa, atas karunia-Nya dan rido-Nya sehingga penulis dapat menyelesaikan skripsi yang berjudul "Studi Pengaruh Seleksi Fitur dan Hyperparameter Tunning Pada Random Forest Untuk Klasifikasi Web Phising" yang mana naskah skripsi ini saya ajukan sebagai salah satu syarat kelulusan pada Program Studi S1 Informatika.

Selanjutnya penulis juga mengucapkan terima kasih atas dukungan dan bantuannya sehinga dapat sampai pada tahap ini kepada :

1. Muhammad Bunyamin dan Tutik Marchamah selaku kedua orang tua yang selalu memberi doa dan dukungan kepada penulis.
2. Prof, Dr. M. Suyanto, MM., selaku Rektor Universitas Amikom Yogyakarta.
3. Anna Baita, S.Kom., M.Kom., selaku Dosen pembimbing yang telah membantu dan memberikan arahan dan bimbingan selama proses pengerjaan skripsi.
4. Segenap Dosen Teknik Komputer yang telah memberikan wawasan kepada penulis selama proses menimba ilmu di Universitas Amikom Yogyakarta.
5. Semua pihak yang telah berkontribusi membantu penulis yang tidak dapat disebutkan satu per satu.

Penulis menyadari bahwa dengan keterbatasan ilmu yang dimiliki saat ini maka skripsi yang dibuat masih jauh dari kata sempurna. Oleh karena itu, penulis mengharapkan kritik dan saran yang sifatnya membangun dari semua pihak demi memperbaiki laporan penelitian ini.

Yogyakarta, 17 Desember 2024

Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

DAFTAR LAMBANG DAN SINGKATAN

ID	User Identification
NCA	Neighborhood Component Analysis
APWG	Anti-Phishing Working Group
RFE	Recursive Feature Elimination
GIs	Genomic Islands
SVM	Support Vector Machines
BO	Bayesian Optimization
GP	Gaussian Process
TPE	Tree-structutrd Parzen Estimator
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
X^2	Statistik chi-square
O_i	Frekuensi yang diamati dalam kategori i
E_i	Frekuensi yang diharapkan dalam kategori i
p-value	Nilai probabilitas untuk menentukan signifikansi statistic hasil uji chi-square
α	Nilai alpha
F	Kumpulan fitur yang dievaluasi
arg	Parameter untuk konfigurasi RFE
w_i	Bobot atau tingkat kepentingan fitur I yang dihasilkan oleh model
f_i	Fitur ke-i
n	Total fitur dalam model
y	Nilai loss
y^*	Threshold untuk memisahkan baik l(x) dan buruk g(x)

$l(x)$	Probabilitas dari nilai hyperparameter x ketika nilai loss y lebih kecil dari y^*
$g(x)$	Probabilitas dari nilai hyperparameter x ketika nilai loss y lebih besar atau sama dengan y*

DAFTAR ISTILAH

Phishing	upaya mencuri informasi penting melalui situs palsu
Random Forest	algoritma ensemble berbasis pohon keputusan
Machine learning	pembelajaran mesin untuk membuat model
Chi-square	uji statistik untuk mengukur independensi variabel
RFE	seleksi fitur iteratif untuk menemukan fitur terbaik
BO	pendekatan optimasi berbasis probabilitas
Hyperparameter tunning	Penyetelan parameter untuk performa optimal
Cross-validation	teknik evaluasi model dengan data berbeda
Overfitting	model terlalu cocok dengan data latih
Decision Tree	algoritma berbasis struktur pohon keputusan
Naïve Bayes	algoritma probabilistik sederhana
Pearson Correlation	korelasi linier antar variabel
Spyware	perangkat lunak mencuri data pengguna
Random Search	metode pencarian hyperparameter secara acak
Grid Search	pencarian sistematis hyperparameter
Feature selection	proses memilih fitur relevan
SVM	algoritma klasifikasi berbasis margin
Metode stacking	kombinasi model untuk hasil lebih baik
E-commerce	perdagangan elektronik
Trade-off	pengorbanan satu aspek untuk yang lain
Septuple	kelompok terdiri dari tujuh elemen
Ansamble kanonik	kombinasi model dengan pendekatan kanonik
Supervised learning	pembelajaran dengan label data
Unsupervised learning	pembelajaran tanpa label data
Reinforcement learning	pembelajaran berbasis penghargaan
Filter method	seleksi fitur berbasis statistic
Wrapper method	seleksi fitur dengan model prediksi
Embedded method	seleksi fitur dalam proses pelatihan model
Dataset	kumpulan data untuk analisis/model

Lasso Regression	regresi dengan penalti untuk seleksi fitur
Observed	nilai aktual yang diamati
Expected	nilai yang diharapkan berdasarkan model
Null hypothesis	hipotesis awal untuk diuji
Interpretabilitas	kemampuan memahami model prediksi
Gini impurity	ukuran ketidakmurnian dalam Decision Tree
Ensemble	kombinasi model untuk akurasi lebih baik
Underfitting	model tidak cukup belajar dari data
Sqrt	akar kuadrat
Auto	pilihan otomatis oleh sistem
Log2	logaritma basis 2
Framework	kerangka kerja pengembangan perangkat lunak
Confusion matrix	matriks evaluasi performa model
Legitimate	website sah atau valid
Suspicious	website mencurigakan
Alpha	parameter tingkat signifikansi statistik
Scikit-learn	library Python untuk machine learning
Default	pengaturan awal bawaan
Search space	ruang pencarian parameter
Oversampling	menambah data dari kelas minoritas
Undersampling	mengurangi data dari kelas mayoritas
Feature importance	nilai pentingnya fitur untuk model
Folds	pembagian data untuk cross-validation
Goodfit	model yang sesuai dengan data

INTISARI

Phishing merupakan upaya ilegal untuk mencuri informasi penting seperti user ID dan kata sandi melalui situs palsu [1]. Laporan APWG mencatat 165.772 situs phishing terdeteksi pada kuartal pertama 2020, dengan rekor 245.771 serangan di Januari 2021, yang mayoritas menargetkan lembaga keuangan [2], [3]. Untuk mencegah ancaman ini, pengembangan model *machine learning*, seperti *Random Forest*, digunakan karena akurasinya tinggi dan kemampuannya menangani dataset kompleks. Namun, untuk mengurangi kompleksitas komputasi, seleksi fitur seperti *Chi-Square* dan *RFE* diterapkan, serta *hyperparameter tuning* dengan *Bayesian Optimization* dilakukan untuk meningkatkan performa model. Penelitian ini menggunakan dataset *Phishing Websites* dari *UCI Machine Learning Repository* dengan 31 fitur kategorikal. Langkah-langkah penelitian meliputi prapemrosesan data, seleksi fitur dengan *Chi-Square* dan *RFE*, serta *tuning hyperparameter* menggunakan *Bayesian Optimization*. Evaluasi dilakukan dengan *cross-validation* 10 *fold* untuk mengukur akurasi dan *recall*. Hasil penelitian menunjukkan kombinasi *Chi-Square* dan *Bayesian Optimization* menghasilkan akurasi tertinggi sebesar 97,02% dengan *cross-validation* 97,29%, lebih baik dibandingkan lainnya. Hal ini membuktikan bahwa *Chi-Square* efektif untuk seleksi fitur pada data nominal, sedangkan *Bayesian Optimization* membantu meningkatkan performa model secara efisien. Penelitian ini dapat dimanfaatkan oleh institusi keamanan siber dan lembaga pemerintah untuk mendeteksi situs web phishing dengan lebih andal.

Kata kunci: Bayesian Optimization, Chi-Square, phishing, Random Forest, RFE.

ABSTRACT

Phishing is an illegal attempt to steal important information such as user IDs and passwords through fake sites [1]. The APWG report noted 165,772 phishing sites detected in the first quarter of 2020, with a record 245,771 attacks in January 2021, the majority of which targeted financial institutions [2], [3]. To prevent these threats, the development of machine learning models, such as Random Forest, is used due to its high accuracy and ability to handle complex datasets. However, to reduce computational complexity, feature selection such as Chi-Square and RFE are applied, and hyperparameter tuning with Bayesian Optimization is performed to improve model performance. This study uses the Phishing Websites dataset from the UCI Machine Learning Repository with 31 categorical features. The research steps include data preprocessing, feature selection with Chi-Square and RFE, and hyperparameter tuning using Bayesian Optimization. Evaluation is done with 10 fold cross-validation to measure accuracy and recall. The results showed that the combination of Chi-Square and Bayesian Optimization produced the highest accuracy of 97.02% with a cross-validation of 97.29%, better than other combinations of methods. This proves that Chi-Square is effective for feature selection on nominal data, while Bayesian Optimization helps improve model performance efficiently. This research can be utilized by cybersecurity institutions and government agencies to detect phishing websites more reliably.

Keyword**:** Bayesian Optimization, Chi-Square, phishing, Random Forest, RFE.