Information Technology Journal
Vol. 4 No. 2 December 2022
Homepage: https://jurnal.amikom.ac.id/index.php/intechno
INTECHNO Journal
INFORMATION TECHNOLOGY JOURNAL
e-ISSN:2655-1438
p-ISSN: 2655-1632

.

Implementasi Metode Random Forest Klasifikasi untuk Phishing Link Detection

Adi Kresna Kencana [1)], Fadhilah Dwi Ananda [2)], Anggit Dwi Hartanto [3)], Hartatik [4)]
[1,2,3,4]Universitas AMIKOM Yogyakarta, Ring Road Utara, Sleman, 55283, Indonesia

Info Artikel

Kata Kunci:
Phishing Website
Random Forest
Klasifikasi
Decision Tree

Keywords:
Phishing Website
Random Forest
Classification
Decision Tree

ABSTRAK (9 PT)

Internet sangat dibutuhkan saat ini. Masalah yang muncul dari perkembangan internet dan teknologi saat ini adalah keamanan dan privasi, dimana data privasi sangat rentan untuk dicuri oleh seseorang melalui internet. Contohnya adalah situs web phishing yang telah tersebar luas di internet yang dapat mencuri data seperti, data pribadi, data kartu kredit, perbankan online, dan data email tanpa diketahui oleh pengguna internet. Bisa dibilang sulit membedakan situs web asli atau palsu. Karenanya diperlukan klasifikasi untuk membedakan situs web asli atau palsu. Penelitian ini menggunakan algoritma Random Forest untuk memilih situs web phishing dari pohon keputusan. Berdasarkan penerapan algoritma Random Forest untuk mendeteksi phishing situs web, hasil akurasi adalah 94,36% dan hasil validasi adalah 94,77% menggunakan 2.457 dataset yang diperoleh dari situs web www.kaggle.com. Dari penelitian ini terbukti bahwa algoritma ini memiliki akurasi tinggi untuk memprediksi situs web phishing dan hasil yang diperoleh diimplementasikan dalam bentuk ekstensi dari browser secara realtime yang nantintya akan memberikan popup peringatan jika situs website yang dibuka adalah phishing website.

ABSTRACT (9PT)

The internet is very much needed nowadays. The problem that arises from the development of the internet and technology today is security and privacy, where privacy data is very vulnerable to being stolen by someone via the internet. An example is a phishing website that is widespread on the internet that can steal data such as personal data, credit card data, online banking, and email data without internet users knowing. You could say it's difficult to distinguish genuine or fake web sites. Therefore classification is needed to distinguish genuine or fake websites. This study uses the Random Forest algorithm to select phishing websites from the decision tree. Based on the application of the Random Forest algorithm to detect website phishing, the accuracy result is 94.36% and the validation result is 94.77% using 2,457 datasets obtained from the website www.kaggle.com. From this research it is proven that this algorithm has a high accuracy for predicting phishing websites and the results obtained are implemented in the form of extensions from the browser in realtime which will then provide a warning popup if the website being opened is phishing website.

Corresponding Author:

Adi Kresna Kencana
Email: Adi.kencana@students.amikom.ac.id

1. PENDAHULUAN

Seiring meningkatnya penggunaan internet dari tahun ke tahun. Berbagai aktifitas ataupun transaksi yang awalnya dilakukan secara offline sekarang sudah dapat diakses dan dilakukan di mana saja secara online. Disamping dari kemudahan tersebut menimbulkan celah keamanan bagi para pengguna internet yang masih awam dalam melakukan transaksi di dunia maya yang dimanfaatkan oleh penjahat internet untuk mencuri data ataupun informasi seperti data pribadi, data kartu kredit, online banking dan data e-mail tanpa diketahui oleh pengguna internet. Website palsu atau phishing website adalah website yang di gunakan untuk melakukan kejahatan di dunia maya yang popularitasnya meningkat seiring meningkatnya pengguna internet di dunia. Hingga saat ini tercatat pengguna internet pada tahun 2014 mencapai 3,078 Triliyun atau 4,41% penduduk (internetworldstats.com). Berdasarkan data dari (insectpro.com) kejahatan internet berasal dari phishing website dengan 21% dari total kejahatan internet, menempati kedudukan di atas social engineering dan webbase attack. Phishing website yang digunakan untuk kejahatan didesain sedemikian rupa agar menyerupai website otentuk (konten,URL,Domain, tampilan dan lainnya) yang bertujuan mengelabui korban seolah-olah sedang mengakses suatu halaman website dari sumber yang sah [1].

Metode yang akan digunakan untuk mendeteksi phishing website yaitu Random Forest. Random Forest adalah algoritma yang di kembangkan dari Decision Tree [7] digunakan untuk mengklasifikasi data dalam jumlah yang besar[8]. Klasifikasi yang dilakukan di random forest yaitu dengan penggabungkan pohon (tree) lalu dilakukan training pada data yang ada. Jika pohon (tree) semakin bertambah banyak maka tingkat akurasi yang di dapatkan akan semakin tinggi. Hasil klasifikasi random forest di ambil dari setiap pohon (tree) yang di bentuk. Pengutamaan dari tree dibentuk berdasakan vote terbanyak [2] .

Random forest di bantu Decision Tree untuk menseleksi suatu data. Pohon (tree) yang di gunakan di bagi secara sendiri-sendiri artinya pemanggil dirinya-sendiri secara langsung ataupun tidak langsung yang di sebut dengan rekursif dari data yang sama. Pemecahan digunakan untuk membagi-bagi data berdasarkan type data ataupun jenis atributnya. Klasifikasi berjalan jika tree telah di bentuk. Proses klasifikasi di awali dari pemecahan data yang ada ke dalam decision tree secara acak [3]. Setelah tree ada maka voting dilakukan pada setiap kelas dari dataset. Lalu kombinasi vote dari setiap kelas di ambil yang paling banyak menggunakan random forest, maka hasil vote yang paling baik tercipta [2].

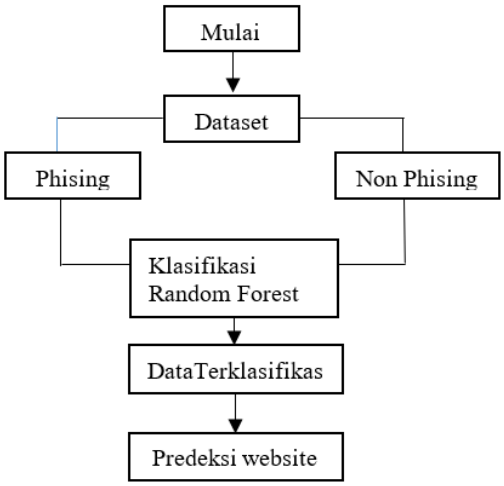
Penelitian ini bertujuan untuk membuktikan bahwa dengan menggunakan algoritma Random Forest bisa untuk mendeteksi sebuah phishing website, yang akan di aplikasikan dalam extensi di browser nantinya akan memberikan persentase dari website yang kita buka phishing atau non-phishing.

2. METODE

Penelitian tentang klasifikasi phishing website maupun algoritma random forest sudah sangat banyak, salah satunya tentang email spam dan non-spam yang dilakukan oleh Aswan Supriyadi (Sekolah Tinggi Teknologi Pelita Bangsa) yang dipublikasi oleh Jurnal Komputer dan Informatika Akademi Bina Saran Informatika.

Penelitian tersebut menggunakan database phishing website dari hasil komputasi digital pada UCI Neda Abdelhamid Auckland Institute of Studies. Dimana data yang didapat terdiri dari Variabel rendah (0), sedang(-1) dan tinggi (1). Parameter yang digunakan ada 9 yaitu SFH, popUpWindow, SSLfinal_State, Request_URL, URL_of_Anchor, web_traffic, having_IP_Address, URL_Length, Age_of_domain. Data yang digunakan berjumlah 1353 data. Yang mendapatkan hasil akurasi dari algoritma C4.5 untuk akurasinya sebesar 83.81% dan untuk hasil dari algoritma genetika mendapatkan akurasi sebesar 86.40%. penelitian ini menunjukkan bahwa dari perbandingan algoritma C4.5 dan algoritma genetika yang lebih baik akurasinya ada pada algoritma genetika [4].

Di dalam melakukan penelitian ini, peneliti mencari dataset dan kemudiaan dari dataset tersebut dilakukan proses Data Mining, dalam hal ini peneliti melakukan percobaan validasi dari hasil data training yang menjadi sebuah data testing dari masing-masing sampel menghasilkan value dari 20% - 100%. Proses validasi tersebut dilakukan dengan metode Random Forest, metode ini dapat meningkatkan hasil akurasi, karena di dalam membangkitkan simpul anak untuk node-node dilakukan dengan cara acak.selanjutnya peneliti menguji dengan menggunakan operasi Test dengan menggunakan algoritma dari Random Forest yang hasilnya dapat menentukan sebuah website palsu atau tidak secara akurat [5]. Kerangka alur dari penelitian dapat dilihat pada table 1.



Gambar 1. Alur Penelitian

A. Dataset

Dataset yang digunakan dalam penelitian phishing website diambil dari situs https://www.kaggle.com/akashkr/phising-website#Training%20Dataset.arfff sebanyak 2457 data phising website, yang terdiri dari 30 variable. Pada setiap variabel tersebut memiliki kelasnya tersendiri, pada kelas tersebut memiliki batasan tersendiri. [6]. Struktur data dapat dilihat pada table 2.

Tabel 2. Struktur dataset

No	NamaVariabel	TipeData	Keterangan
1	having_IP_Address	Integer	Attribut
2	URL_Length	Integer	Attribut
3	Shortining_Service	Integer	Attribut
4	having_At_Symbol	Integer	Attribut
5	double_slash_ redirecting	Integer	Attribut
6	Prefix_Suffix	Integer	Attribut
7	having_Sub_Domain	Integer	Attribut
8	SSLfinal_State	Integer	Attribut
9	Domain_registeration_ length	Integer	Attribut
10	Favicon	Integer	Attribut
11	Port	Integer	Attribut
12	HTTPS_token	Integer	Attribut
13	Request_URL	Integer	Attribut
14	URL_of_Anchor	Integer	Attribut
15	Links_in_tags	Integer	Attribut
16	SFH	Integer	Attribut
17	Submitting_to_email	Integer	Attribut
18	Abnormal_URL	Integer	Attribut
19	Redirect	Integer	Attribut
20	on_mouseover	Integer	Attribut
21	RightClick	Integer	Attribut
22	popUpWidnow	Integer	Attribut
23	Iframe	Integer	Attribut
24	age_of_domain	Integer	Attribut
25	DNSRecord	Integer	Attribut
26	web_traffic	Integer	Attribut
27	Page_Rank	Integer	Attribut
28	Google_Index	Integer	Attribut
29	Links_pointing_to _page	Integer	Attribut
30	Statistical_report	Integer	Attribut
31	Result	Integer	Attribut

B. Random Forest

Random Forst merupakan metode berdasarkan pohon yang acak kemudian di kombinasikan ke satu model. Random Forest menggunakan seleksi input yang acak. Setiap training data set di ambil dari training set yang

asli. Lalu tree(pohon) ada di dalam sebuah data training menggunakan seleksi secara acak ini yang disebut dengan Bagging. Random Forest memiliki kombinasi input yang linear Misalnya data input C, K mengambil fraksi di K yang akan meningkatkan kekuatan korelasi.Pendekatan lainnya lebih banyak fitur dengan acak dari variable inputnya. Variabel P adalah jumlah yang di kombinasikan. Variabel P diseleksi secara acak(random) di tambahkan dengan nomor yang random [-1,0,1] maka kombinasi P di hasilkan prosedur ini di sebut Forest-RC. Pohon keputusan di mulai dengan cara menghitung suatu nilai entropy yang digunakan sebagai penentu tingkat ketidakmurnian dari atribut.

3. HASIL DAN DISKUSI

Entropy yaitu parameter untuk mengukur tingkat keberagaman(heterogenitas) dari banyak data[9]. Jika value dari entropy semakin besar maka, tingkat keberagaman suatu banyak data semakin besar rumus untuk menghitung entropy seperti di bawah ini:

$$Entropy(S) = -\sum_{i=1}^{m} log2\ (pi) \quad (1)$$

di mana *m* adalah jumlah kelas klasifikasi dan *Pi* adalah jumlah proporsi sampel (peluang) untuk kelas i

Untuk rumus entropy pada masing-masing variable yaitu:

$$Entropy_A\ (S) = \sum_{v} \frac{|S_v|}{|S|}\ entropy\ (S_v) \quad (2)$$

di mana A adalah variabel, v adalah nilai yang mungkin untuk variable A, $|S_v|$ adalah jumlah sampel untuk nilai v, $|S|$ adalah jumlah sampel untuk seluruh sampel data, $entropy\ (S_v)$ adalah entropy untuk sampel yang memiliki nilai v.

Pada penelitian ini di lakukan training data yang melibatkan dua kelas yaitu kelas phishing website dannon-phishing website dengan menggunakan 30 variabel dan jumlah data 2.457. proses training data ini menggunakan algoritma Random Forest yang menghasilkan nilai akurasi sebesar 94.36% dan nilai validasinya sebesar 94.77%. Untuk hasil dari training dataset dapat dilihat pada table 3.

Table.3 Nilai pengujian training algoritma

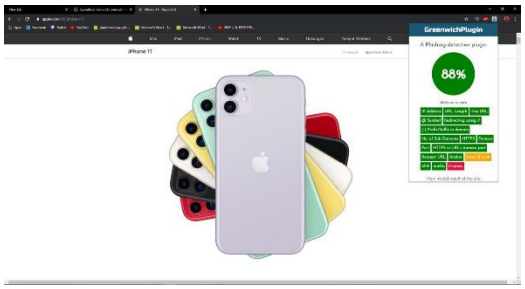
Metode	Validation	Akurasi	Jumlah Data
Random Forest	94,77%	94,36%	2.457

Untuk hasil dari pengujian kita membutuhkan sampel yang di ambil dari website, dimana untuk pengaplikasiannya menggunakan extensi yang di pasang di browser yang nantinya extensi tersebut melacak secara realtime apakah website yang dibuka tersebut phishing atau non-phishing. Contoh website yang telah di uji statusnya ada pada table.4.

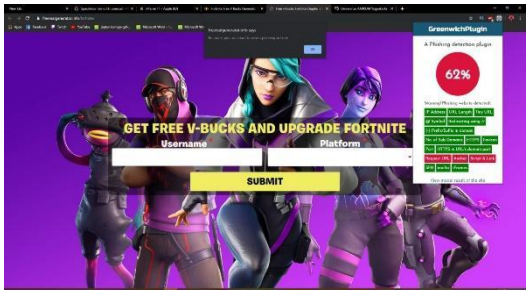
Tabel 4. Website uji status

No	URL	Status
1	https://www.kompas.com/	Non-Phising
2	https://freerealgenerator.info/fortnite/	Phising
3	https://free.bitcoin.com/	Phising
4	https://www.youtube.com/	Non-Phising
5	https://www.reddit.com/	Non-Phising

Untuk tampilan dari extensi yang di install di browser seperti gambar 2 dan 3. Jika terdeteksi sebuah website adalah phising maka akan ada popup notifikasi yang muncul di atas untuk membuat user agar lebih waspada contoh ada pada gambar 3.



Gambar 2. Hasil dari percobaan untuk website resmi Apple.com yang berstatus non-Phishing



Gambar 3. Hasil dari percobaan untuk website freerealgenerator.info/fortnite/ yang berstatus Phishing

4. KESIMPULAN

Dalam penelitian ini dilakukan klasifikasi phishing website dan non-phishing website dengan menggunakan algoritma Random Forest, didapatkan hasil akurasi sebesar 94,36% dan yang menggunakan 2.457 dataset.hasil tersebut di aplikasikan menggunakan exstensi pada browser, yang akan menampilkan persentase dari setiap website yang di buka, jika teridentifikasi phishing web maka akan ada popup peringatan. Terdeteksi phishing website rata-rata karena tidak menggunakan SSL (Secure Socket Layer), terlalu banyak script yang ada di dalam website tersebut dan Panjang URL juga berpengaruh dalam pendeteksian phishing website.

Hasil penelitian yang kami lakukan dengan metode Random Forest untuk akurasi yang kami dapatkan cukup besar tetapi penelitian ini belum di uji coba dengan membandingkan algoritma lain yang memungkinkan algoritma lain mendapatkan nilai akurasi di atas algoritma yang kami gunakan.

REFERENSI

[1] Y. C. G. Tomy Salim, "DATA MINING IDENTIFIKASI WEBSITE PHISING MENGGUNAKAN ALGORITMA C4.5," *Jurnal TAM (Technology Acceptance Model)* , vol. 8, pp. 130-135, 2017.

[2] M. Z. Z. R. W. S. Saifullah, "ANALISA TERHADAP PERBANDINGAN ALGORITMA DECISION TREE DENGAN ALGORITMA RANDOM TREE UNTUK PRE-PROCESSING DATA," *Jurnal Sains Komputer & Informatika (J-SAKTI)* , vol. 1, pp. 180-185, 2017.

[3] L. Y. S. X. a. H. P. Mingming Zhu, "Random Forests for Object Detection," pp. 267-274, 2015.

[4] A. S. Sunge, "Optimasi Algoritma C4.5 Dalam Prediksi Web Phishing Menggunakan Seleksi Fitur Genetic Algoritma," *Paradigma*, vol. XX, pp. 27-32, 2018.

[5] Sunaryono, "PENELITIANKOMPARASIALGORITMAKLASIFIKASI DALAMMENENTUKANWEBSITEPALSU," *Teknikom*, vol. 1, pp. 1-12, 2017.

[6] A. Kumar, "Data phising Website," 20 01 2018. [Online]. Available: https://www.kaggle.com/akashkr/phising-website#Training%20Dataset.arff.[

[7] Su dan Zhang, "A Fast Decision Tree Learning Algorithm", American Association for Artificial Intelligence (www.aaai.org), 2006.

[8] Belgiu dan Dragut, "Random forest in remote sensing: A review of applications and future directions", ISPRS Journal of Photogrammetry and Remote Sensing, 2016.

[9] Chairunnisa dkk, "Perbandingan CART dan Random Forest untuk Deteksi Kanker berbasis Klasifikasi Data Microarray", Jurnal RESTI, 2021