P-ISSN : 2549-7480
E-ISSN : 2549-6301

JURNAL ILMIAH INFORMATIKA
E-ISSN : 2549-6301
ISSN : 2549-7480
Faculty of Sains and Technology, Ibrahimy University
S4-Accredited : 85/M/KPT/2020
Journal page is available to : https://journal.ibrahimy.ac.id/index.php/JIMI

# IDENTIFIKASI *WEBSITE PHISHING* MENGGUNAKAN ALGORITMA *CLASSIFICATION AND REGRESSION TREES* (CART)

**Pungkas Subarkah[1)], Ali Nur Ikhsan[2)]**

[1,2] *Program Studi Informatika, Universitas Amikom Purwokerto*

*email: [1] subarkah@amikompurwokerto.ac.id, [2] alinurikhsan@amikompurwokerto.ac.id*

**ARTICLE INFO**

**Article History:**
Recieved : 08 August 2021
Revised : -
Accepted : 30 December 2021
Published : 31 December 2021

**Keywords:**
Phishing
Website
Identification
CART

**IEEE style in citing this article:**
P. Subarkah and A. N. Ikhsan, "Identifikasi Website Phishing Menggunakan Algoritma Classification And Regression Trees (CART)", *Jurnal.ilmiah.informatika*, vol. 6, no. 2, pp. 127-136, Dec. 2021.

**Corresponding Author:**
Pungkas Subarkah
Universitas Amikom Purwokerto

**ABSTRACT**

With the increase in internet users and the development of technology, the threats to its security are increasingly diverse. One of them is phishing which is the most important issue in cyberspace. Phishing is a threatening and trapping activity someone by luring the target to indirectly provide information to the trapper. The number of phishing crimes, this has the potential to cause several losses, one of which is namely about the loss of privacy of a person or company. This study aims to identify phishing websites. The Classification And Regression Trees (CART) algorithm is one of the classification algorithms, and the dataset in this research taken from the UCI Repository Learning obtained from the University of Huddersfield. The method used in this research is problem identification, data collection, pre-processing stage, use of the CART algorithm, validation and evaluation and withdrawal conclusion. Based on the test results obtained the value of accuracy of 95.28%. Thus the value of the accuracy obtained using the CART algorithm of 95.28% categorized very good classification.



## 1. PENDAHULUAN

Dengan adanya kemajuan teknologi, hal ini membantu bagi masyarakat dalam menunjang aktivitas, khususnya dalam pemanfaatan internet, masyarakat menjadi lebih mudah dan efektif dalam berkomunikasi maupun mencari informasisistem informasi [1]. Khususnya masyarakat Indonesia, 202,6 juta jiwa dari total penduduk Indonesia 274,9 juta jiwa atau sekitar 73,7% penduduk Indonesia telah aktif menggunakan internet [2].

Dengan meningkatnya pengguna internet dan berkembangnya teknologi,


DOI : https://doi.org/10.35316/jimi.v6i2.1342

P-ISSN : 2549-7480
P. Subarkah dkk/ JIMI 6 (2) pp. 127-136 E-ISSN : 2549-6301

ancaman terhadap keamanannya semakin kian beragam. Salah satunya adalah *phishing*.

*Phishing* merupakan kegiatan yang bersifat mengancam dan menjebak seseorang dengan cara memancing target untuk secara tidak langsung memberikan informasi kepada penjebak [3]. Selain itu *phishing* bertujuan untuk mengirimkan tautan berbahaya, biasanya menyamar sebagai yang legal, melalui spam atau jejaring sosial untuk mendorong pengguna untuk mengunjungi dan memperoleh informasi pribadi mereka [4]. Hal ini sejalan dengan pendapat [5] bahwa *phishing* ialah skema cyber berdasarkan kegiata kriminal yang menarik perhatian.

Banyaknya kejahatan *phishing*, hal ini berpotensi menimbulkan beberapa kerugian, salah satunya yaitu tentang kerugian *privacy* seseorang atau perusahaan. Dikemukukan oleh APWG (*Anti-Phishing Working Gruop*) dalam [6] bahwa dari tahun ke tahun, masyarakat Indonesia semakin sadar tentang adanya *website phishing*. Maka salah satu upaya yang dapat dilakukan adalah dengan cara identifikasi untuk mendeteksi website yang terindikasi *phishing*, oleh karena itu dibutuhkanna klasifikasi dalam data mining untuk mengetahui data maupun parameter yang dijadikan acuan dalam pendeteksian *phishing* [7].

Penelitian terdahulu [6] dengan judul model klasifikasi untuk deteksi situs *phishing* di Indonesia. Hasil yang didapatkan dalam penelitian ini dengan Algoritma *Multilayer Perceptron* memiliki tingkat akurasi terbaik sebesar 91,8%. Selanjutnya penelitian sejenis yang melakukan prediksi *website phishing* [8]. Penelitian ini bertujuan untuk memberikan gambaran metode yang paling efisie dalam memprediksi *website phishing*. Hasilnya Algoritma SVM memperoleh nilai akurasi sebesar 92.34%.

Penelitian yang dilakukan oleh [9] cukup mendalam menggunakan optimasi Algoritma C4.5 dengan seleksi fitur Genetic Algoritma dalam memprediksi *web phishing*. Berdasarkan penerapan Algoritma C4.5 dihasilkan akurasi sebesar 83.82% dan dengan menerapkan seleksi fitur menggunakan Algoritma genetika meningkat sebesar 3.22% menjadi 86.47%.

Penelitian yang dilakukan [10] dengan menerapkan Algoritma CART dan Algoritma *Naive Bayes*. Hasil yang didapatkan dalam penelitian yaitu dalam mengklasifikasi nasabah bank telemarketing ialah nilai akurasi CART sebesar 89.51% lebih besar dari akurasi Algoritma *Naive Bayes* sebesar 86.88%.

Peneliti ini akan mengujii keberhasilan *dataset website phishing* dengan menerapkan *machine learning* sehingga luarannya mendapatkan sebuah informasi dari pola yang terbentuk dan mengekstraksi sebuah informasi. Algoritma klasifikasi yang digunakan ialah Algoritma *Classification And Regression Trees* (CART) untuk mengindentifikasi *website phishing*. Data yang digunakan dalam penelitian ini diperoleh dari data *public* UCI yaitu *Phishing Website Data Set* [11].

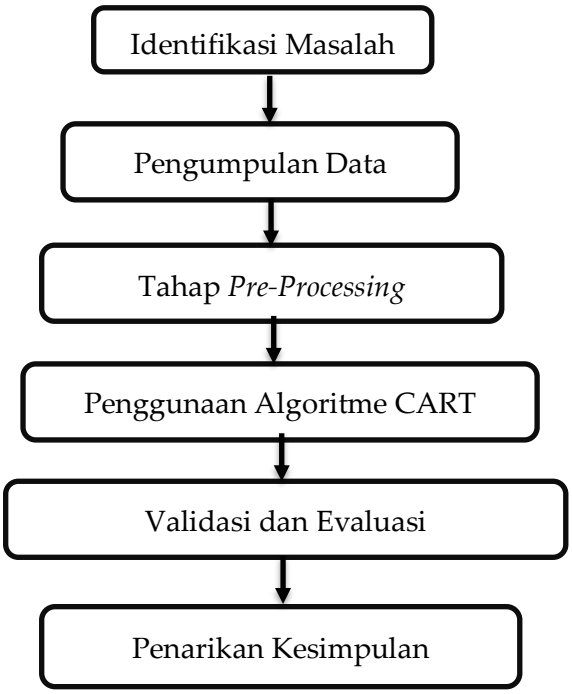
## 2. METODE PENELITIAN

Metode yang digunakan pada peneliian ini ialah Algoritma *Classification and Regression Trees* (CART). CART adalah suatu metodologi yang berjalan dengan baik dan cocok sebagai metode untuk prediksi dan klasifikasi[12]. Algoritma CART mempunyai kelebihan, salah satunya yaitu bersifat nonparametrik / cocok digunakan untuk data yang berjenis *numeric*. Pada Algoritma CART, baris / *record* akan diklasifikasikan pada varibel tujuan berdasarkan nilai-nilai variabel prediktornya [13].

Alur penelitian digunakan untuk


DOI : https://doi.org/10.35316/jimi.v6i2.1342

mempermudah dalam melakukan kegiaan penelitian. Berikut langkah-langkah penelitian dapat dilihat pada bagan di bawah ini :




Gambar 1. Alur Penelitian

Berikut penjelasan tiap-tiap alur penelitian sebagai berikut :

A. Identifikasi Masalah

Pada tahapan ini dilakukan untuk menentukan masalah dan menganalisis penelitian-penelitian sebelumnya, serta teknik yang digunakan dapat mengindentifikasi *website phishing.*

B. Pengumpulan Data

Pengumpulan data yaitu tahapan mengumpulkan data, pada penelitian ini menggunakan data sekunder, diperoleh dari *database* UCI *Repository*. Pada *dataset phishing website* terdiri dari 2456 *record* dengan atribut 31 (30 atribut dan 1 target atribut). Target atibut mempunyai dua kategori yaitu bukan *phishing* dan *phishing*[11]. Berikut jumlah *dataset* berdasarkan *class*, dapa dilihat pada tabel 1.

Tabel 1. *Dataset* Berdasarkan *Class*

| No | Klasifikasi | Jumlah Record *Dataset* |
|---|---|---|
| 1 | Bukan *Phishing* | 1362 |
| 2 | *Phishing* | 1094 |
| | Jumlah | 2356 |

C. Tahap *Pre-Processing*

*Pre-Processing* merupakan tahapan yang bertujuan untuk menyeleksi data dari *missing values* sehingga mendapatkan data yang bersih dan siap digunakan.

D. Penggunaan Algoritma CART

Algoritma CART yang akan digunakan dalam penelitian ini, dari hasil *confusion matrix* dapat dihitung dari nilai *precision, recall* dan *F-Measure*. Rincian pehitungan *confusion matrix* dapat dilihat dari Tabel 2., sebagai berikut [14] :


DOI : https://doi.org/10.35316/jimi.v6i2.1342

P-ISSN : 2549-7480
P. Subarkah dkk/ JIMI 6 (2) pp. 127-136 E-ISSN : 2549-6301

Tabel 2. *Confusion Matrix*

| *Correct Classiication* | *Classifiacation As* | |
|---|---|---|
| | *Yes* | *No* |
| *Yes* | TP | FN |
| *No* | FP | TN |

Dari tabel 2., diatas, adapun rumus untuk *confusion matrix* berasal dari nilai *precision, recall* dan *F-Measure* dari sebuah nilai *accuracy* [15] *:*

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$F\text{-}Measure = \frac{2\times Precision \times Recall}{Precision + Recall} \quad (3)$$

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (4)$$

E. Validasi dan Evaluasi

Tahapan validasi dan evaluasi bertujuan untuk mengukur keakuratan hasil yang dicapai oleh model menggunakan teknik yang digunakan yaitu *confusion matrix* dan *10-fold cross validation*

F. Penarikan Kesimpulan

Tahapan penarikan kesimpulan berfungsi untuk menyimpulkan hasil yang didapatkan dari penelitian dengan menggunakan Algoritma CART yang memberikan hasil akurasi terbaik untuk mengindentifikasi dan klasifikasi *website phishing* berdasarkan nilai *precision, recall* dan *F-Meeasure*. Tingkatan klasifikasi pada data mining untuk Algoritma sebagai berikut [16] :

1. Klasifikasi sangat baik = 0.90 – 1.00
2. Klasifikasi baik = 0.80 – 0.90
3. Klasifikasi cukup = 0.70-0.80
4. Klasifikasi rendah = 0.60-0.70

## 3. HASIL DAN PEMBAHASAN

3.1 Identifikasi Masalah

Tahapan identifikasi masalah dilakukan untuk menentukan masalah dan teknik yang baik sehingga dapat digunakan dengan tujuan untuk mengindentifikasi *website phishing*.

3.2 Pengumpulan Data

Tahapan ini meliputi mencari informasi yang dipakai pada penelitian ini yaitu menggunakan data sekunder. Data tersebut didapatkan dari data *public* UCI *Repository Learning*. Pada *dataset phishing website* terdiri dari 2456 *record* dengan atribut 31 (30 atribut dan 1 target atribut). Target atibut mempunyai dua kategori yaitu bukan *phishing* dan *phishing*[11]. Kedua kelas target akan dijadikan sebagai *output* dari *dataset website phishing*. Berikut penjelasan dari tiap atribut pada *dataset website phishing* :

1) Atribut *Having_IP_Address*
   Adanya IP address sebagai domain pada url (biner) -1 (tidak), 1 (iya).
2) Atribut URL_*Length*
   Panjang url (polinominal) -1(kurang dari 54 karakter), 0 (antara 54 sampai 75 karakter), 1 (lebih dari 75 karakter).
3) Atribut *Shortining_Service*
   Penggunaan layanan penyingkatan URL (biner) 0(tidak), 1 (iya)
4) Atribut *Having_At_Symbol*
   Penggunaan simbol "@" pada url (biner) 0 (tidak), 1 (iya)
5) Atribut *Double_slash_redirecting*
   Penggunaan simbol "//" pada url untuk mengalihkan website (biner) 0 (tidak), 1 (iya).
6) Atribut *Prefix_Suffix*
   Pengunaan simbol "-" pada domain dalam url (biner) -1 (tidak), 1 (iya)
7) Atribut *Having_Sub_Domain*
   Penggunaan subdomain polinominal) -1 (tidak punya), (1 subdomain), 1 (lebih dari 1 subdomain).


DOI : https://doi.org/10.35316/jimi.v6i2.1342

P-ISSN: 2549-7480
P. Subarkah dkk/ JIMI 6 (2) pp. 127-136 E-ISSN: 2549-6301

8) Atribut *SSLfinal_State*
Memiliki sertifikat SSL dimana sertifikat yang dipercaya berasal dari penyedia ternama seperti "GeoTrust, GoDaddy, Network Solutions, Thawte, Comodo, Doster dan VeriSign" polinominal) -1 (punya sertifikat yang dipercaya), 0 (punya sertifikat yang belum dipercaya), 1 (tidak memiliki sertifikat)
9) Atribut *Domainregistrationlength*
Batas berlakunya domain (biner) 0 (lebih dari 1 tahun), 1 (kurang dari 1 tahun).
10) Atribut *Favicon*
Memiliki favicon dari link ekternal (biner) 0(tidak), 1 (iya)
11) Atribut *Port*
Penggunaan port seperti 21, 22, 23, 445 dan lainnya (biner) 0 (tidak), 1 (iya)
12) Atribut HTTPS_*token*
Penggunaan https ke dalam bagian domain pada url (biner) 0 (tidak), 1 (iya)
13) Atribut *Request*_URL
Persentase permintaan url ekternal dari keseluruhan (polinominal) -1 (kurang dari 22%), 0 (antara 22% sampai 61%), 1 (lebih dari 61%)
14) Atribut URL_*of_Anchor*
Persentase penggunaan tag <a> yang mengarah selain ke domain yang sama dari keseluruhan (polinominal) -1 (kurang dari 31%), 0 (antara 31% sampai 67%), 1 (lebih dari 67%)
15) Atribut *Links_in_tags*
Persentase penggunaan tag <*link*>, <*meta*>, dan <*script*> yang mengarah selain ke domain yang sama dari keseluruhan (polinominal) -1 (kurang dari 17%), 0 (antara 17% sampai 81%), 1 (lebih dari 81%)
16) Atribut SFH
Domain pemrosesan Server Form Handler (polinominal) -1 (pada domain yang sama), 0 (pada domain yang berbeda), 1 (kosong)
17) Atribut *Submitting_to_email*
Penggunaan fungsi "mail() atau mailto" dalam php untuk mengirim informasi *user* (biner) 0 (tidak), 1 (iya).
18) Atribut *Abnormal*_URL
Kecocokan website dengan catatannya yang ditunjukan pada basis data WHOIS (biner) -1 (cocok), 1 (tidak cocok)
19) Atribut *Redirect*
Jumlah pengalihan website yang dilakukan (polinominal) -1 (kurang dari 2 kali), 0 ( 2,3, atau 4 kali), 1 (lebih dari 4 kali)
20) Atribut *on_mouseover*
Perubahan status bar ketika *event onMouseOver* aktif (biner) 0 (tidak), 1 (iya)
21) Atribut *RightClick*
Keadaan klik kanan pada website (biner) 0 (diaktifkan), 1 (dinonaktifkan)
22) Atribut *popUpWindow*
Penggunaan popUpWindow untuk meminta user mengisi data mereka (biner) 0 (tidak) ,1 (iya)
23) Atribut *Iframe*
Penggunaan fungsi iframe (biner) 0 (tidak), 1 (iya)
24) Atribut *age_of_domain*
Umur domain (biner) -1 (lebih dari atau sama dengan 6 bulan), 1 (kurang dari 6 bulan)
25) Atribut *DNSRecord*
Adanya catatan DNS pada domain (biner) -1 (ada), 1 (tidak ada).
26) Atribut *web_traffic*
Rank lalu lintas website dalam basis data Alexa (polinominal) -1 (diatas 100,000), 0 (dibawah 100,000) ,1 (tidak terdaftar)
27) Atribut *Page_Rank*
Nilai PageRank website (biner) -1 (lebih dari atau sama dengan 0.2) 1 (kurang dari 0.2)
28) Atribut *Google_Index*


DOI : https://doi.org/10.35316/jimi.v6i2.1342

P-ISSN : 2549-7480
P. Subarkah dkk/ JIMI 6 (2) pp. 127-136 E-ISSN : 2549-6301

Adanya website dalam indeks pencarian Google (biner) -1 (iya), 1 (tidak)

29) Atribut *Links_pointing_to_page*
Jumlah link eksternal yang menunjuk ke website (polinominal) -1 (lebih dari 2) , 0 (1 atau 2), 1 (tidak ada)
30) Atribut *Statistical_report*
*Host* berasal dari *Top Phishing* IPs atau *Top Phishing Domains* yang dibuat oleh beberapa pihak seperti *StopBadware* dan *PhishTank* (biner) -1 (tidak), 1 (iya)
31) Atribut *Result* (Label)
Hasil identifikasi website (biner) -1 (bukan *phishing*), 1 (*phishing*)

3.3 Tahap *Pre-Processing*

Tahap *pre-processing* dilakukan untuk memperoleh data yang benar-benar bersih atau tidak ada *noise* pada atribut dan *missing values* pada atribut. Hasil yang didapatkan pada *dataset website phishing* yang telah disesuaikan atribut untuk digunakan pada aplikasi Weka, sebagai berikut pada tabel 3.

Tabel 3. Data *Pre-processing*

| Data Asli | Data hasil *pre-processing* | Keterangan |
|---|---|---|
| 1 | 1 | IP address sebagai domain pada url (biner) -1 (tidak), 1 (iya |
| 1 | 1 | Panjang url (polinominal) -1(kurang dari 54 karakter), 0 (antara 54 sampai 75 karakter), 1 (lebih dari 75 karakter). |
| 0 | 0 | Penggunaan layanan penyingkatan URL (biner) 0 (tidak), 1 (iya) |
| 0 | 0 | Penggunaan simbol "@" pada url (biner) 0 (tidak), 1 (iya) |
| 1 | 1 | Penggunaan simbol "//" pada url untuk mengalihkan website (biner) 0 (tidak), 1 (iya). |
| -1 | -1 | Pengunaan simbol "-" pada domain dalam url (biner) -1 (tidak), 1 (iya) |
| -1 | -1 | Penggunaan subdomain polinominal) -1 (tidak punya), (1 subdomain), 1 (lebih dari 1 subdomain). |
| -1 | -1 | Memiliki sertifikat SSL , -1 (punya sertifikat yang dipercaya), 0 (punya sertifikat yang belum dipercaya), 1 (tidak memiliki sertifikat |
| 0 | 0 | Batas berlakunya domain (biner) 0 (lebih dari 1 tahun), 1 (kurang dari 1 tahun |
| 0 | 0 | Memiliki *favicon* dari link ekternal (biner) 0(tidak), 1 (iya |
| 0 | 0 | Penggunaan *port* seperti 21, 22, 23, 445 dan lainnya (biner) 0 (tidak), 1 (iya) |
| 1 | 1 | Penggunaan https ke dalam bagian domain pada url (biner) 0 (tidak), 1 (iya) |
| 1 | 1 | Persentase permintaan url ekternal dari keseluruhan (polinominal) -1 (kurang dari 22%), 0 (antara 22% sampai 61%), 1 (lebih dari 61%) |
| -1 | -1 | Persentase penggunaan tag <a> yang mengarah selain ke domain yang sama dari keseluruhan (polinominal) -1 (kurang dari 31%), 0 (antara 31% sampai 67%), 1 (lebih dari 67%) |
| 1 | 1 | Persentase penggunaan tag <*link*>, <*meta*>, dan <*script*> yang mengarah selain ke domain yang sama dari keseluruhan |


DOI : https://doi.org/10.35316/jimi.v6i2.1342

P-ISSN: 2549-7480
P. Subarkah dkk/ JIMI 6 (2) pp. 127-136 E-ISSN: 2549-6301

| Data Asli | Data hasil *pre-processing* | Keterangan |
|---|---|---|
| | | (polinominal) -1 (kurang dari 17%), 0 (antara 17% sampai 81%), 1 (lebih dari 81%) |
| -1 | -1 | Domain pemrosesan *Server Form Handler* (polinominal) -1 (pada domain yang sama), 0 (pada domain yang berbeda), 1 (kosong) |
| 1 | 1 | Penggunaan fungsi "mail() atau mailto" dalam php untuk mengirim informasi user (biner) 0 (tidak), 1 (iya). |
| 1 | 1 | Kecocokan website dengan catatannya yang ditunjukan pada basis data WHOIS (biner) -1 (cocok), 1 (tidak cocok) |
| 0 | 0 | Jumlah pengalihan website yang dilakukan (polinominal) -1 (kurang dari 2 kali), 0 ( 2,3, atau 4 kali), 1 (lebih dari 4 kali) |
| 0 | 0 | Perubahan status bar ketika event onMouseOver aktif (biner) 0 (tidak), 1 (iya) |
| 0 | 0 | Keadaan klik kanan pada website (biner) 0 (diaktifkan), 1 (dinonaktifkan) |
| 0 | 0 | Penggunaan popUpWindow untuk meminta user mengisi data mereka (biner) 0 (tidak) ,1 (iya) |
| 0 | 0 | Penggunaan fungsi *iframe* (biner) 0 (tidak), 1 (iya) |
| -1 | -1 | Umur domain (biner) -1 (lebih dari atau sama dengan 6 bulan), 1 (kurang dari 6 bulan) |
| 1 | 1 | Adanya catatan DNS pada domain (biner) -1 (ada), 1 (tidak ada). |
| -1 | -1 | *Rank* lalu lintas website dalam basis data Alexa (polinominal) -1 (diatas 100,000), 0 (dibawah 100,000) ,1 (tidak terdaftar) |
| -1 | -1 | Nilai PageRank website (biner) -1 (lebih dari atau sama dengan 0.2) 1 (kurang dari 0.2) |
| 0 | 0 | Adanya website dalam indeks pencarian Google (biner) -1 (iya), 1 (tidak) |
| 1 | 1 | Jumlah link eksternal yang menunjuk ke website (polinominal) -1 (lebih dari 2) , 0 (1 atau 2), 1 (tidak ada) |
| 1 | 1 | *Host* berasal dari Top *Phishing* IPs atau Top *Phishing* Domains yang dibuat oleh beberapa pihak seperti StopBadware dan PhishTank (biner) -1 (tidak), 1 (iya) |
| 1 | 1 | Hasil identifikasi website (biner) -1 (bukan *phishing*), 1 (*phishing*) |

Hasil yang dijabarkan pada tabel 3., diatas merupakan hasil data *pre-processing dataset website phishing* ialah semua atribut tidak ada yang *missing values* atau nilainya lengkap dan siap digunakan pada tahap berikutnya.

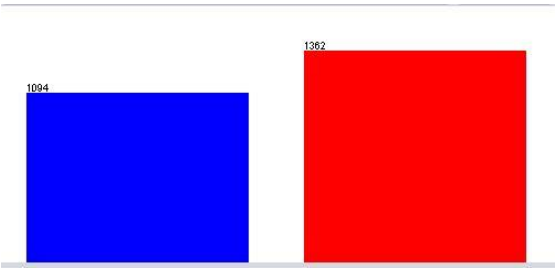
3.4 Penggunaan Algoritma CART

Setelah data melewati tahapan *pre-process*simg dan hasilnya data bersih dari *missing values* atau atributnya tidak lengkap. Selanjuntya *dataset* diolah dengan menggunakan aplikasi Weka.

Penggunaan Algoritma CART dengan menggunakan metove evaluasi 10-*fold cross-validation* untuk menghasilkan *confusion matrix*. Pada gambar 2., merupakan visualisasi dataset secara keseluruhan.


DOI : https://doi.org/10.35316/jimi.v6i2.1342

P-ISSN : 2549-7480
P. Subarkah dkk/ JIMI 6 (2) pp. 127-136 E-ISSN : 2549-6301




Gambar 2. Visualisasi *Dataset*

Dari gambar 2 dapat diketahui bahwa grafik yang berwarna biru dengan data 1094 berlabel *phishing*, sedangkan grafik warna merah dengan data 1392 berlabel bukan *phishing*.

Langkah selanjutnya ialah menghitung nilai akurasi yang didapatkan dari nilai *precision*, *recall* dan *F-Measure*. Dibawah ini merupakan proses perhitungan nilai *confusion matrix*.

Tabel 3. *Confusion* Kelas "1"

| | |
|---|---|
| 1046 (TP) | 48 (FN) |
| 68 (FP) | 1294 (TN) |

Tabel 3., merupakan penjabaran dari nilai *confusion matrix* yang, Sedangkan nilai *confusion matrix* pada kelas "-1", bisa dilihat pada tabel 4.

Tabel 4. *Confusion matrix* Kelas "-1"

| | |
|---|---|
| 1294 (TP) | 68 (FN) |
| 48 (FP) | 1046 (TN) |

Dari tabel 3 dan tabel 4., diperoleh nilai *precision*, *recall*, dan *F-Measure* dan nilai akurasi berdasarkan *confusion matrix*, sebagai berikut :

Tabel 5. Nilai Akurasi *dataset website phishing*

| ***Class*** | ***Precision*** | ***Recall*** | ***F-Measure*** |
|---|---|---|---|
| 1 | 0.939 | 0.956 | 0.947 |
| -1 | 0.964 | 0.950 | 0.957 |
| *Weighted Avg* | 0.953 | 0.953 | 0.953 |

Tabel 5., merupakan rincian dari nilai *precision*, *recall* dan *F-Measure* dari *dataset website phishing* dengan Algoritma CART menggunakan *software* WEKA 3.9 membutuhkan waktu 1.76 *second*

3.5 Evaluasi dan Validasi

Tahapan evaluasi dan validasi digunakan untuk mengukut tingkat akurasi dari Algoritma klasifikasi yang disajikan pada tabel 6. Pada tabel 6., yaitu tabel yang dihasilkan oleh *confusion matrix* dari pengujian *dataset* menggunakan Algoritma CART dengan metode 10-*fold cross validation*.


DOI : https://doi.org/10.35316/jimi.v6i2.1342

Tabel 6. *Confusion Matrix Dataset Website Phishing*

| | *Website Phishing* | *Bukan Website Phishing* |
|---|---|---|
| *Website Phishing* | 1046 | 48 |
| *Bukan Website Phishing* | 68 | 1294 |
| 2456 | 1114 | 1342 |

Tabel 6, dapat dijabarkan sebagai berikut bahwa jumlah data hasil bentukan *rule* yang berkategori *website phishing* yang sama dengan data testing yang juga *website phishing* sebanyak 1046. Selanjutnya, jumlah data hasil bentukan rule yang bukan *website Phishing* dengan data testing yang *Website Phishing* sebanyak 48. Kemudian jumlah data hasil bentukan rule yang *website phishing* dan data testing yang bukan *website phishing* sebanyak 68. Terakhir, jumlah data hasil bentukan *rule* yang bukan *website phishing* yang sama dengan data testing yang juga bukan *website phishing* sebanyak 1294.

3.6 Penarikan Kesimpulan

Berdasarkan hasil perhitungan yang sudah dilakukan menggunakan Algoritma CART pada *dataset website phishing* diperoleh nilai akurasi sebesar 95.28%, dengan rincian nilai *precision* sebesar 0.953%, nilai *recall* sebesar 0.953% dan nilai *F-Measure* sebesar 0.953%. Hal ini sejalan dari pendapat referensi [17] bahwa Algoritma CART cocok digunakan untuk data *numeric* dan data yang berjumlah besar.

## 4. UCAPAN TERIMA KASIH

Penulis ucapkan terima kasih kepada LPPM Universitas Amikom Purwokerto atas dukungannya sehingga penelitian terlaksana.

## 5. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan mengenai identifikasi *website phishing* menggunakan Algoritma *Classification And Regression Trees* (CART) dapat disimpulkan bahwa dari nilai *confusion matrix* diperoleh hasil akurasi sebesar 95.28%, dengan rincian nilai *precision* sebesar 0.953%, nilai *recall* sebesar 0.953% dan nilai *F-Measure* sebesar 0.953%. Dari hasil nilai akurasi dikategorikan klasifikasi sangat baik. Saran-saran yang dapat dianjurkan melalui penelitian ini untuk penelitian selanjutnya antara lain : 1) Menambahkan Algoritma klasifikasi yang bertujuan untuk perbandingan atau komparasi sehingga mendapatkan nilai akurasi yang lebih baik pada identifikasi *website phishing*., 2) Menambahkan proses seleksi fitur dalam identifikasi *website phishing,* dan 3) Melakukan proses optimasi pada Algoritma *data mining* guna meningkatkan nilai akurasi yang lebih baik.

## 6. REFERENSI

[1] A. S. Gulo, S. Lasmadi, and K. Nawawi, "Cyber Crime dalam Bentuk Phising Berdasarkan Undang-Undang Informasi dan Transaksi Elektronik," *PAMPAS J. Crim. Law*, vol. 1, no. 2, pp. 68–81, 2021.

[2] G. P. Riyanto, "Jumlah Pengguna Internet Indonesia 2021 Tembus 202 Juta." [Online]. Available: https://tekno.kompas.com/read/2021/02/23/16100057/jumlah-pengguna-internet-indonesia-2021-tembus-202-juta. [Accessed: 02-Jul-2021].

[3] M. H. Wibowo and N. Fatimah, "Ancaman phishing terhadap pengguna sosial media dalam dunia cyber crime," *JOEICT(jurnal Educ. Inf. Commun. Technol.*, vol. 1, pp. 1–5, 2017.

[4] K. L. Chiew, K. S. C. Yong, and C. L. Tan, "A survey of phishing attacks:


DOI : https://doi.org/10.35316/jimi.v6i2.1342

P-ISSN : 2549-7480
P. Subarkah dkk/ JIMI 6 (2) pp. 127-136 E-ISSN : 2549-6301

Their types, vectors and technical approaches," *Expert Syst. Appl.*, vol. 106, pp. 1–20, 2018.

[5] S. S. M. Motiur Rahman, T. Islam, and M. I. Jabiullah, "PhishStack: Evaluation of Stacked Generalization in Phishing URLs Detection," *Procedia Comput. Sci.*, vol. 167, no. 2019, pp. 2410–2418, 2020.

[6] F. Eka Purwiantono and A. Tjahyanto, "Model Klasifikasi Untuk Deteksi Situs Phising Di Indonesia," p. 156, 2017.

[7] M. Al-diabat, "Detection and Prediction of Phishing Websites using Classification Mining Techniques," vol. 147, no. 5, pp. 5–11, 2016.

[8] Z. Halim, "Prediksi Website Pemancing Informasi Penting Phising Menggunakan Support Vector Machine (SVM)," *Inf. Syst. Educ. Prof.*, vol. 2, no. 1, pp. 71–82, 2017.

[9] A. S. Sunge, "Optimasi Algoritma C4.5 Dalam Prediksi Web Phishing Menggunakan Seleksi Fitur Genetic Algoritma," *Paradigma*, vol. 10, no. 2, pp. 27–32, 2018.

[10] P. Subarkah, E. P. Pambudi, S. Oktaviani, and N. Hidayah, "Perbandingan Metode Klasifikasi Data Mining untuk Nasabah Bank Telemarketing," vol. 20, no. 1, 2020.

[11] R. M. A. Mohammad, L. M. Cluskey, and F. Thabtah, "Dataset Website Phishing," 2015. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Phishing+Websites. [Accessed: 02-Jun-2021].

[12] A. Moro *et al.*, "Prognostic factors differ according to KRAS mutational status: A classification and regression tree model to define prognostic groups after hepatectomy for colorectal liver metastasis," *Surg. (United States)*, vol. 168, no. 3, pp. 497–503, 2020.

[13] R. Timofeev, *Classification and Regression Trees (CART) Theory and Aplications*. Berlin: Humboldt University, 2004.

[14] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*, Third Edit., vol. 5. USA: Elsevier, 2012.

[15] M. Han, J., & Kamber, *Data Mining Concepts, Model and Techniques 2nd Edition*. San Fransisco: Elsevier, 2006.

[16] F. Gorunescu, *Data mining Concepts, Models and Techniques*. Verlen Berlin: Springer, 2011.

[17] P. Subarkah, "Penerapan Algoritma Klasifikasi Classification And Regression Trees ( CART ) Untuk Diagnosis Penyakit Diabetes Retinopathy," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 19, no. 2, pp. 294–301, 2020.


DOI : https://doi.org/10.35316/jimi.v6i2.1342 