

OPTIMASI ALGORITMA KLASIFIKASI DECISION TREE (CART) DENGAN METODE BAGGING UNTUK DETEKSI WEBSITE PHISHING

Ria Ester¹, Sartika Lina Mulani Sitio²

^{1,2}Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Pamulang, Kota Tangerang Selatan, Indonesia
Email: ¹dosen02665@unpam.ac.id, ²dosen00847@unpam.ac.id

Abstrak

Kemajuan ilmu pengetahuan dan teknologi (IPTEK), khususnya teknologi informasi, semakin memudahkan dalam menunjang aktivitas masyarakat, namun pada saat yang sama membuat berbagai perangkat rentan dieksploitasi oleh pelaku kejahatan. Salah satu aktivitas tersebut adalah mencuri data dari pengguna Internet melalui situs palsu (juga dikenal sebagai situs phishing) yang dirancang agar terlihat seperti aslinya. Situs web phishing menimbulkan ancaman serius terhadap keamanan informasi online dan memerlukan pendekatan efektif untuk mendeteksi dan mencegahnya. Untuk memerangi maraknya website phishing di dunia maya, diperlukan suatu klasifikasi untuk memprediksi website mana yang akan diklasifikasikan sebagai website phishing dengan menggunakan Decision Tree Classification Algorithm (CART). Untuk meningkatkan kinerja Algoritma Decision Tree Classification (CART) dan mencapai akurasi optimal yang lebih baik, diperlukan optimasi menggunakan metode bagging. Teknik bagging yang menggabungkan hasil beberapa model pohon keputusan diterapkan untuk meningkatkan kinerja dan keandalan algoritma CART dalam mendeteksi situs web phishing. Dalam penelitian ini, kami mengumpulkan dataset yang berisi berbagai karakteristik terkait dengan karakteristik website phishing. Data tersebut kemudian diproses dan dibagi menjadi beberapa subset untuk pelatihan dan pengujian model. Tujuan dari penelitian ini adalah untuk mengoptimalkan algoritma klasifikasi pohon keputusan (CART) dengan menerapkan teknik bagging dalam konteks deteksi situs web phishing. Berdasarkan hasil pengujian, penerapan Decision Tree Classification Algorithm (CART) untuk mengklasifikasikan website phishing menghasilkan akurasi sebesar 96,61%, dan jika dipadukan dengan teknik bagging, akurasinya meningkat sebesar 1,13% menjadi 97,74%. Eksperimen ini menunjukkan bahwa optimasi dapat meningkatkan akurasi prediksi situs web phishing dengan menggabungkan Decision Tree Algorithm (CART) dengan teknik bagging.

Kata Kunci: Website Phising, Klasifikasi, Prediksi, Algoritma, Decision Tree.

Abstract

Advances in science and technology (IPTEK), especially information technology, make it easier to support community activities, but at the same time make various devices vulnerable to exploitation by cybercriminals. One such activity is stealing data from Internet users through fake sites (also known as phishing sites) designed to look like the real thing. Phishing websites pose a serious threat to online information security and require an effective approach to detect and prevent them. To combat the proliferation of phishing websites in cyberspace, a classification is needed to predict which websites will be classified as phishing websites using the Decision Tree Classification Algorithm (CART). To improve the performance of the Decision Tree Classification (CART) algorithm and achieve better optimal accuracy, optimization using the bagging method is needed. A bagging technique that combines the results of several decision tree models is applied to improve the performance and reliability of the CART algorithm in detecting phishing websites. In this research, we collected a dataset containing various characteristics related to the characteristics of phishing websites. The data is then processed and divided into subsets for model training and testing. The aim of this research is to optimize the decision tree classification algorithm (CART) by applying bagging techniques in the context of phishing website detection. Based on test results, applying the Decision Tree Classification Algorithm (CART) to classify phishing websites produces an accuracy of 96.61%, and when combined with bagging techniques, the accuracy increases by 1.13% to 97.74%. This experiment shows that optimization can improve the prediction accuracy of phishing websites by combining the Decision Tree Algorithm (CART) with bagging techniques.

Keywords: Phishing Websites, Classification, Prediction, Algorithms, Decision Trees.

1. PENDAHULUAN

Di era digital saat ini, penggunaan Internet telah menjadi bagian penting dalam kehidupan sehari-hari. Namun, selain manfaat yang diberikan dunia online, terdapat juga tantangan keamanan, seperti ancaman situs web phishing. Phishing adalah kejahatan dunia maya yang bertujuan memperoleh informasi pribadi sensitif secara ilegal. Informasi ini mencakup nama pengguna, kata sandi, dan informasi ID sensitif lainnya yang digunakan oleh situs web phishing. Situs web phishing adalah upaya penipuan untuk memperoleh informasi pribadi pengguna dengan berpura-pura menjadi situs web tepercaya. Tujuannya adalah untuk mengelabui korban agar memasukkan informasi pribadi sensitif yang diminta, sehingga memudahkan pelaku kejahatan siber mendapatkan informasi yang dapat digunakan untuk aktivitas yang tidak diinginkan, sehingga membahayakan informasi dan data keuangan korban sehingga keduanya akan hilang. Ketika serangan phishing meningkat, deteksi dini menjadi semakin penting untuk melindungi pengguna dari potensi kehilangan data sensitif dan risiko keamanan terkait [4]. Dalam konteks ini, algoritma klasifikasi menjadi alat utama untuk membangun sistem yang efektif untuk mendeteksi situs web phishing. Salah satu algoritma klasifikasi yang paling umum digunakan adalah pohon keputusan, khususnya pohon klasifikasi dan regresi (CART) [3][5]. Meskipun pohon keputusan dapat menangani tugas klasifikasi dengan baik, pohon keputusan sering kali cenderung overfit dan sensitif terhadap fluktuasi data. Untuk mengatasi kendala tersebut, penelitian ini bertujuan untuk mengoptimalkan kinerja algoritma pohon keputusan (CART) dengan menerapkan metode bagging. Bagging merupakan teknik pembelajaran ansambel yang menggabungkan hasil beberapa model untuk meningkatkan akurasi dan keandalan prediksi. Dalam konteks pendeteksian website phishing, penggunaan teknik bagging diharapkan dapat meningkatkan akurasi dan ketahanan algoritma dalam mendeteksi website yang mencurigakan [6].

Untuk memerangi maraknya phishing di dunia maya, diperlukan suatu sistem yang menggunakan teknik machine learning untuk mendeteksi website tersebut dalam kategori phishing atau non-phishing. Pada penelitian sebelumnya yang bertajuk “Model Klasifikasi Pendeteksi Situs Phishing di Indonesia” yang dilakukan oleh Agung Susilo Yuda Irawan, dkk dan Aris Tjahyanto, terdapat beberapa model klasifikasi untuk mendeteksi situs phishing. Pada penelitian ini penulis melakukan identifikasi website phishing yang dapat mendeteksi situs phishing dan menguji model klasifikasi yang dibuat. Algoritma klasifikasi yang digunakan dalam percobaan antara lain SMO

(Sequential Minimal Optimization), Naive Bayes, Bagging, dan Multilayer Perceptron [7]. Selain itu, penelitian sebelumnya yang dilakukan oleh Zuhri Halim pada tahun 2017 berjudul “Memprediksi Situs Web Phishing yang Mengandung Informasi Phishing Penting Menggunakan Support Vector Machines (SVM)” menggunakan beberapa algoritma termasuk Support Vector Machines dan Naive Bayes. Beberapa metode klasifikasi digunakan dalam penelitian ini untuk membandingkan beberapa metode tersebut untuk memprediksi situs web phishing [8].

Berdasarkan pembahasan di atas, terdapat beberapa teknik algoritma klasifikasi yang telah digunakan pada penelitian sebelumnya untuk mengklasifikasikan dan memprediksi website phishing. Penelitian selanjutnya yang dilakukan peneliti akan memutuskan menggunakan Algoritma Decision Tree Classification (CART). Algoritma Klasifikasi Pohon Keputusan (CART) ini mempunyai beberapa kelemahan dan kekurangan. Sulit untuk merancang pohon keputusan yang optimal dan jumlah kesalahan pada setiap tingkat pohon terakumulasi. Keputusan ini penting dan mempengaruhi tingkat akurasi terutama melalui penggunaan metode bagging. Bagging dimaksudkan untuk meningkatkan stabilitas dan akurasi algoritma pembelajaran mesin dalam klasifikasi statistik dan regresi.

2. METODOLOGI PENELITIAN

a. Pengumpulan Data

Metode penelitian yang digunakan adalah pendekatan kuantitatif dengan menggunakan metode bagging yang merupakan bagian dari algoritma penggabungan machine learning. Dataset website phishing dan non-phishing diperoleh dari UCI Machine Learning Repository.

b. Pencarian Literatur

Literatur diperoleh dari jurnal, buku dan referensi lainnya mengenai algoritma CART, Teknik Bagging.

c. Algoritma Chart

Classification and Regression Trees (CART) merupakan metode atau algoritma yang merupakan teknik pencarian pohon keputusan. CART dikembangkan sekitar tahun 1980 oleh Leo Bryman, Jerome H. Friedman, Richard A. Olshen, dan Charles J. Stone. CART adalah metode statistik nonparametrik yang dimaksudkan untuk dikembangkan untuk analisis topik dalam klasifikasi, baik untuk variabel respon kategorikal maupun kontinu. CART menghasilkan pohon klasifikasi bila variabel respon bersifat kategorikal, dan pohon regresi bila variabel respon kontinu [9].

Langkah algoritma cart adalah :

1. Pembentukan Pohon Klasifikasi
2. Pemangkas Pohon Klasifikasi
3. Penentuan Pohon Klasifikasi Optimal

d. Bagging

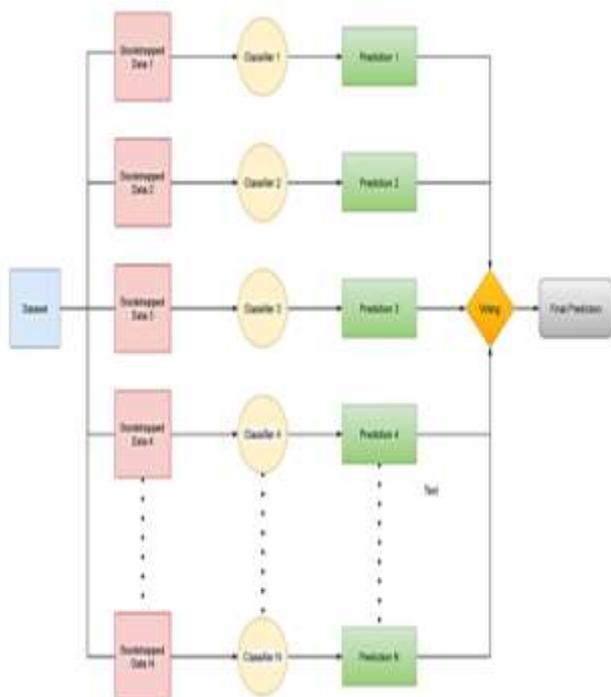
Bagging adalah singkatan dari Bootstrap Aggregating. Metode bagging ini menggunakan subdataset untuk membuat set pelatihan pembelajaran (L). Prediksi tunggal yang diperoleh dari algoritma bagging ini adalah yang merupakan hasil dari bobot suara terbanyak yang serupa. Algoritme bagging digunakan dalam klasifikasi dan regresi. Jika regresi kuat, Anda dapat menghitung rata-ratanya saat menggabungkan prediksi [1][10]. Bagging merupakan algoritma pembelajaran yang stabil terhadap perubahan kecil pada set pelatihan yang menyebabkan perbedaan besar pada pembelajar yang dihasilkan

Gambar 1. Alur Diagram Metode Bagging

3. HASIL DAN PEMBAHASAN

3.1 Hasil Pengujian Decision Tree (Cart)

Hasil pengujian algoritma decision tree (CART) yang dilakukan pada penelitian ini yaitu untuk mengukur kinerja akurasi dari hasil training dan testing dataset yang telah melewati beberapa tahap preprocessing data. Dataset yang digunakan pada penelitian ini berjumlah 11.055 baris terdiri dari barisan angka biner tiap atribut



kolom, dalam pengujiannya dataset yang digunakan training sebesar 80% (8844 data) dan testing 20%

(2211 data) yang mana ini merupakan pemisahan dataset training dan testing yang ideal. Hasil diagnosa yang muncul digunakan sebagai pemberian keputusan apakah website tersebut termasuk dalam website phishing atau non-phishing. Dibawah ini merupakan detail hyperparameter value yang digunakan dan hasil pengujian algoritma decision tree (CART) yang menggunakan library scikit-learn pada paket anaconda aplikasi jupyter notebook dengan bahasa Python berdasarkan confusion matrix.

Tabel 1. Hyperparameter Value Train Test Split Algorithm Decision Tree (Cart)

Metode	Hyperparameter	Hyperparameter Value
Decision Tree	test_size	0.2
	train_size	none
	random_state	7
	shuffle	true
	stratify	none

Tabel 2. Hyperparameter Algorithm Decision Tree (Cart)

Metode	Hyperparameter	Hyperparameter Value
Decision Tree	criterion	Gini
	Splitter	Best
	Max_depth	none
	Min_samples_split	2
	Min_samples_leaf	1
	Min_weight-Fraction_leaf	0.0
	Max_features	None
	Random_state	27
	Max_leaf_nodes	None
	Min_impurity_decrease	0.0
	Min_impurity_split	None
	Class_weight	None
	Ccp_alpha	0.0

Tabel 3. Confusion Matrix Algorithm Decision Tree (Cart)

Prediksi	Hasil	
	Phishing	Non - Phishing
Phishing	934	40
Non - Phishing	35	1202

Pada Tabel 3, bahwa algoritma decision tree (CART) dapat memprediksi 35 website phishing sebagai non-phishing (FP), 40 website non-phishing sebagai website phishing (FN), 934 prediksi benar terhadap

website phishing (TP), dan 1202 prediksi benar terhadap website non-phishing (TN). Selain itu algoritma decision tree (CART) menghasilkan kinerja akurasi dari hasil pengujian sebesar 96,61%, jika dihitung secara manual dengan rumus berdasarkan hasil dari tabel confusion matrix adalah sebagai berikut:

$$\begin{aligned} \text{Accuracy} &= \frac{TP+TN}{TP+TN+FP+FN} \\ &= \frac{934+1202}{934+1202+35+40} \\ &= 2136/2211 = 96.61\% \end{aligned}$$

Sedangkan hasil kinerja algoritma decision tree (CART) berdasarkan precision, recall, dan f-measure menggunakan library scikit-learn terdapat pada Tabel 4 di bawah ini.

Tabel 4. Kinerja Algoritma Decision Tree (Cart)

Kelas	Precision	Recall	F - Measure
Phishing	0.9639	0.9589	0.9614
Non - Phishing	0.9678	0.9717	0.9697

Berdasarkan tabel 4, untuk mendapatkan hasil dari precision, recall dan F-Measure dapat dihitung sebagai berikut:

$$\begin{aligned} 1. \text{ Precision Phishing} &= \frac{TP}{TP+FP} \\ &= \frac{934}{934+35} = 0.9639 \end{aligned}$$

$$\begin{aligned} 2. \text{ Precision Non - Phishing} &= \frac{TN}{TN+FN} \\ &= \frac{1202}{1202+40} = 0.9678 \end{aligned}$$

$$\begin{aligned} 3. \text{ Recal Phishing} &= \frac{TP}{TP+FN} \\ &= \frac{1202}{1202+35} = 0.9717 \end{aligned}$$

$$\begin{aligned} 4. \text{ F - Measure Phishing} &= \frac{2 \times \text{Precision} + \text{Recall}}{\text{Precision} + \text{Recall}} \\ &= \frac{2 \times 0.9639 + 0.9589}{0.9639 + 0.9589} = 0.9614 \end{aligned}$$

$$\begin{aligned} 5. \text{ F - Measure Non Phishing} &= \frac{2 \times \text{Precision} + \text{Recall}}{\text{Precision} + \text{Recall}} \\ &= \frac{2 \times 0.9678 + 0.9717}{0.9678 + 0.9717} = 0.9697 \end{aligned}$$

3.2 Hasil Pengujian Decision Tree (Cart) Bagging

Hasil pengujian algoritma decision tree (CART) bagging yang dilakukan pada penelitian ini yaitu untuk mengukur kinerja akurasi dari hasil training dan testing dataset yang telah melewati beberapa tahap preprocessing data dan perbaikan metode dengan merubah hyperparameter value. Dataset yang digunakan pada penelitian ini berjumlah 11.055 baris terdiri dari barisan angka biner tiap atribut kolom, dalam pengujiannya dataset yang digunakan training sebesar 80% (8844 data) dan testing 20% (2211 data) yang mana ini merupakan pemisahan dataset training dan testing yang ideal. Hasil diagnosa yang muncul digunakan sebagai pemberian keputusan apakah website tersebut termasuk dalam kategori website phishing atau non-phishing. Tabel di bawah merupakan detail dari hyperparameter value yang digunakan dan hasil pengujian algoritma decision tree (cart) bagging yang menggunakan library sklearn pada paket anaconda aplikasi jupyter notebook dengan bahasa python.

Tabel 5. Hyperparameter Value Train Test Split Algorithm Decision Tree (Cart) Bagging

Metode	Hyperparameter	Hyperparameter Value
Decision Tree (Cart) Bagging	test_size	0.2
	train_size	none
	random_state	7
	shuffle	true
	stratify	none

Tabel 6. Hyperparameter Algorithm Decision Tree (Cart) Bagging

Metode	Hyperparameter	Hyperparameter Value
Decision Tree	Base_estimator	100
	Max_samples	1.0
	Max_features	1.0
	bootstrap	True
	Bootstrap_features	False
	Oob_score	false
	Warm_start	False
	N_jobs	None
	Random_state	27
	verbose	0
	Min_impurity_split	None
	Class_weight	None
Ccp_alpha	0.0	

Tabel 7. Confusin Matrix Algorithm Decision Tree (Cart) Bagging

Prediksi	Hasil	
	Phishing	Non - Phishing
Phishing	943	31

Non - Phishing	19	1218
----------------	----	------

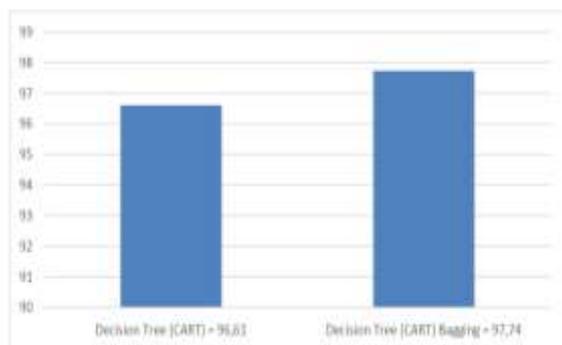
Pada Tabel 7 bahwa algoritma decision tree (CART) bagging dapat memprediksi 19 website phishing sebagai non-phishing (FP), 31 website nonphishing sebagai website phishing (FN), 943 prediksi benar terhadap website phishing (TP), dan 1218 prediksi benar terhadap website non-phishing (TN). Selain itu algoritma decision tree (CART) menghasilkan kinerja akurasi dari hasil pengujian sebesar 97,74%.

Berdasarkan pengujian yang sudah dilakukan dapat terlihat bahwa algoritma *decision tree* (CART) *bagging* memiliki akurasi tertinggi yaitu sebesar 97,74% setelah dilakukan optimasi, sedangkan *decision tree* (CART) sebesar 96,61%. Sehingga peningkatan akurasi setelah dilakukan optimasi yaitu sebesar 1,13% dari hasil pengujian dataset. Berikut hasil perbandingan hasil akurasi dari algoritma *decision tree* (CART) dan *decision tree* (CART) *bagging* digambarkan pada grafik dibawah ini:

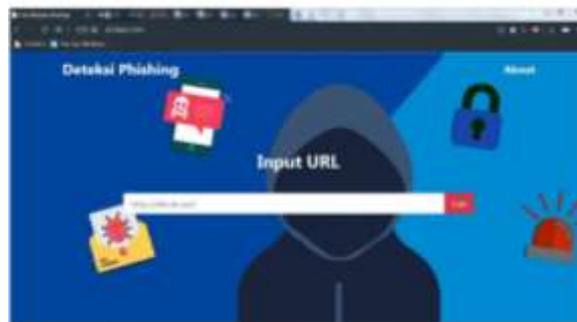
Gambar 2. Perbandingan Akurasi Algoritma Decision Tree (Cart) dan Decision Tree (Cart) Bagging

3.3 Sistem Deteksi Web Phishing

Sistem deteksi *website phishing* merupakan hasil dari implementasi dari algoritma *decision tree* (CART) *bagging* yang telah melewati tahap perbaikan metode dan evaluasi sehingga dihasilkan kinerja yang lebih baik dari sebelumnya. Sistem dibangun dengan bahasa pemrograman Python menggunakan *framework* flask yang hanya memiliki

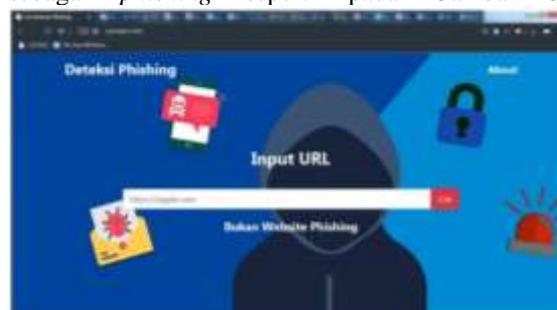


satu buah *interface* utama yaitu *form input* URL seperti gambar berikut:



Gambar 3. Halaman Utama Web Phishing

Jika dilakukan *input* URL dari sebuah *website* kemudian dihasilkan deteksi berupa kalimat “Bukan Website Phishing” maka *website* diprediksi sebagai *nonphishing* seperti pada Gambar 4, namun jika dihasilkan deteksi berupa kalimat “Terindikasi Website Phishing” maka *website* tersebut diprediksi sebagai *phishing* seperti pada Gambar 5.



Gambar 4. Hasil Deteksi Website Non Phishing



Gambar 5. Hasil Deteksi Website Phishing

4. KESIMPULAN

Dari hasil penelitian yang telah dilakukan dapat diambil kesimpulan yaitu Optimasi algoritma *decision tree* (CART) yang dilakukan dengan menggunakan metode penggabungan pada *machine learning* yaitu *bagging* dapat meningkatkan kinerja *accuracy*, *precision*, *recall*, dan *fmeasure* secara keseluruhan. Dalam pengujianya dihasilkan akurasi sebesar 97,74%, hasil ini meningkat dari pengujian algoritma *decision tree* (CART) yang belum dilakukan dioptimasi sebelumnya yaitu sebesar 96,61%. Sehingga peningkatan kinerja akurasi yang diperoleh setelah dilakukan optimasi dengan metode *bagging* yaitu sebesar 1,13%. Dan Sistem deteksi

website phishing dengan implementasi algoritma decision tree (CART) bagging dilakukan uji coba deteksi menggunakan 20 sampe URL yang terdiri 10 website phishing dan 10 non-phishing menghasilkan 17 prediksi benar terhadap prediksi phishing dan nonphishing.

5. REFERENCES

- [1] A. R. Arrahimi, M. K. Ihsan, D. Kartini, M. R. Faisal, and F. Indriani, "Teknik Bagging Dan Boosting Pada Algoritma CART Untuk Klasifikasi Masa Studi Mahasiswa," *J. Sains dan Inform.*, vol. 5, no. 1, pp. 21–30, 2019, doi: 10.34128/jsi.v5i1.171.
- [2] M. I. U. Rosyidi and N. Rochmawati, "Teknik Bagging Pada Algoritma Klasifikasi Decision Tree dan SVM Untuk Klasifikasi SMS Berbahasa Indonesia," vol. 05, pp. 265–271, 2023.
- [3] Y. A. Jatmiko, S. Padmadisastra, and A. Chadidjah, "Analisis Perbandingan Kinerja Cart Konvensional, Bagging Dan Random Forest Pada Klasifikasi Objek: Hasil Dari Dua Simulasi," *Media Stat.*, vol. 12, no. 1, p. 1, 2019, doi: 10.14710/medstat.12.1.1-12.
- [4] R. T. Prasetio and Pratiwi, "Penerapan Teknik Bagging Pada Algoritma Klasifikasi Untuk Mengatasi Ketidakseimbangan Kelas Dataset Medis," *J. Inform.*, vol. II, no. 2, pp. 395–403, 2015, [Online]. Available: <https://ejournal.bsi.ac.id/ejurnal/index.php/ji/article/view/118>
- [5] A. H. Nasrullah, "Implementasi Algoritma Decision Tree Untuk Klasifikasi Produk Laris," *J. Ilm. Ilmu Komput.*, vol. 7, no. 2, pp. 45–51, 2021, doi: 10.35329/jiik.v7i2.203.
- [6] P. Subarkah and A. N. Ikhsan, "Identifikasi Website Phishing Menggunakan Algoritma Classification And Regression Trees (CART)," *J. Ilm. Inform.*, vol. 6, no. 2, pp. 127–136, 2021, doi: 10.35316/jimi.v6i2.1342.
- [7] A. S. Y. Irawan, N. Heryana, H. S. Hopipah, and D. Rahma, "Identifikasi Website Phishing dengan Perbandingan Algoritma Klasifikasi," *Syntax J. Inform.*, vol. 10, no. 01, pp. 57–67, 2021, doi: 10.35706/syji.v10i01.5292.
- [8] Zuhri 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] D. K. Widiyati, M. Wati, and H. S. Pakpahan, "Penerapan Algoritma ID3 Decision Tree Pada Penentuan Penerima Program Bantuan Pemerintah Daerah di Kabupaten Kutai Kartanegara," *J. Rekayasa Teknol. Inf.*, vol. 2, no. 2, p. 125, 2018, doi: 10.30872/jurti.v2i2.1864.
- [10] T. Arifin, "Implementasi Algoritma PSO Dan Teknik Bagging Untuk Klasifikasi Sel Pap Smear," *J. Inform.*, vol. 4, no. 2, pp. 155–162, 2017.