OPTIMASI ALGORITMA *ITERATIVE DICHOTOMISER* 3 KOMBINASI METODE MINMAX - *INFORMATION GAIN* UNTUK DETEKSI WEBSITE *PHISING*

Victor Tarigan[1)], Jeremia Siregar[2),] Adinda Franky Nelwan[3)], Reinhard Komansilan[4)], Ade Yusupa [5)]

[1), 3), 4), 5)]Fakultas Teknik, Universitas Sam Ratulangi,
Jalan Kampus Unsrat Manado, Sulawesi Utara

[2)]Fakultas Teknologi Industri, Institut Sains dan Teknologi TD Pardede
Jl. DR.TD.PardedeNo. 8 Medan, Sumatera Utara

victortarigan@unsrat.ac.id[1)], jeremiasiregar@istp.ac.id, [2]afnelwan@unsrat.ac.id[3] ,
reinhardkomansilan@unsrat.ac.id[4], , ade@unsrat.ac.id[5],

ABSTRAK

Deteksi website phising merupakan tantangan penting dalam keamanan siber, terutama dengan semakin canggihnya teknik-teknik phising yang digunakan untuk mengecoh pengguna dan mencuri informasi pribadi. Seiring dengan meningkatnya jumlah serangan phising, diperlukan metode yang efektif dan efisien untuk mengidentifikasi situs web berbahaya. Algoritma Iterative Dichotomiser 3 (ID3) merupakan salah satu algoritma klasifikasi yang efektif dalam konteks ini, namun kinerjanya dapat ditingkatkan melalui optimasi fitur yang digunakan. Penelitian ini bertujuan untuk mengoptimalkan Algoritma ID3 dalam mendeteksi website phising dengan menggabungkan metode Min-Max Scaling dan Information Gain untuk proses seleksi fitur. Metode ini bertujuan untuk meningkatkan akurasi model dengan hanya mempertahankan fitur-fitur yang paling relevan. Dalam proses seleksi fitur, diperoleh tiga fitur yang memiliki nilai Information Gain sebesar 0, yaitu Favicon, Iframe, dan popUpWidnow. Ketiga fitur ersebut tidak memberikan kontribusi signifikan terhadap prediksi. Fitur-fitur ini kemudian dihilangkan dari total 30 fitur yang ada, menghasilkan model yang lebih ramping dan efisien.

Evaluasi algoritma ID3 dilakukan menggunakan metode k-fold cross-validation dengan 8 fold, yang memberikan gambaran lebih akurat mengenai kinerja model. Hasil pengujian menunjukkan bahwa model ID3 yang menggunakan 27 fitur terbaik menghasilkan nilai rata-rata akurasi sebesar 0.9856, presisi 0.9863, dan recall 0.9878. Di sisi lain, model ID3 yang menggunakan semua 30 fitur menghasilkan nilai rata-rata akurasi sebesar 0.9854, presisi 0.9860, dan recall 0.9878. Meskipun perbedaan akurasi antara kedua model relatif kecil, hasil ini menunjukkan bahwa seleksi fitur yang tepat dapat meningkatkan kinerja model dalam mendeteksi website phising. Penelitian ini memberikan kontribusi terhadap pengembangan teknik deteksi phising yang lebih baik dan dapat diandalkan, serta menjadi referensi untuk penelitian lebih lanjut di bidang keamanan siber.

Kata kunci: Website, *Phising*, *Iterative Dichotomiser* 3, *Information Gain*, Normaliasasi, Seleksi Fitur

ABSTRACT

Detecting phishing websites is a critical challenge in cybersecurity, especially with the increasing sophistication of phishing techniques used to deceive users and steal personal information. As the number of phishing attacks increases, an effective and efficient method is needed to identify malicious websites. The Iterative Dichotomiser 3 (ID3) algorithm is one of the effective classification algorithms in this context, but its performance can be improved through optimization of the features used. This study aims to optimize the ID3 algorithm in detecting phishing websites by combining the Min-Max Scaling

and Information Gain methods for the feature selection process. This method aims to improve model accuracy by retaining only the most relevant features. In the feature selection process, three features were obtained that had an Information Gain value of 0, namely Favicon, Iframe, and popUpWidnow. These three features did not contribute significantly to the prediction. These features were then removed from a total of 30 existing features, resulting in a cleaner and more efficient model. The evaluation of the ID3 algorithm was carried out using the k-fold cross-validation method with 8 folds, which provides a more accurate picture of the model's performance. The test results show that the ID3 model using the 27 best features produces an average accuracy value of 0.9856, precision of 0.9863, and recall of 0.9878. On the other hand, the ID3 model using all 30 features produces an average accuracy value of 0.9854, precision of 0.9860, and recall of 0.9878. Although the difference in accuracy between the two models is relatively small, these results indicate that proper feature selection can improve the model's performance in detecting phishing websites. This study contributes to the development of better and more reliable phishing detection techniques, as well as being a reference for further research in the field of cybersecurity.

Keywords: Website, Phising, Iterative Dichotomiser 3, Information Gain, Normaliasasi, Seleksi Fitur

1. PENDAHULUAN

Internet telah menjadi bagian integral dari kehidupan sehari-hari, memberikan akses cepat dan mudah ke berbagai layanan dan informasi. Berdasarkan hasil survei Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), pengguna internet di Indonesia mencapai 215,63 juta orang pada periode 2022-2023. Jumlah tersebut meningkat 2,67% atau sekitar 5,6 juta orang dibandingkan pada periode sebelumnya yang sebanyak 210,03 juta pengguna. Jumlah pengguna internet tersebut setara dengan 78,19% dari total populasi Indonesia yang sebanyak 275,77 juta jiwa(Annisa Indah Mutiasari, Anggit Dyah Kusumastuti, Rusnandari Retno Cahyani, 2024).

Meningkatnya ketergantungan penduduk Indonesia pada internet, juga meningkatnya ancaman keamanan online, salah satunya adalah serangan *phising*. *Phising* adalah aktivitas kriminal yang menggunakan teknik rekayasa sosial, *phising* dapat menggunakan halaman web palsu (menyamar sebagai situs resmi bank) untuk menipu dan mencuri identitas pengguna(Amin Muftiadi, 2022). Frekuensi aktivitas *phising* semakinmeningkat, terbukti dari data global yang dipublikasikan di situs resmi kelompok kerja anti-*phising* (APWG). Studi bulanan mengungkapkan bahwa skema *phising* mencakup 42% dari seluruh tindakan penipuan yang dilaporkan.Catatan komprehensif telah dikumpulkan dan didokumentasikan dengan total sebanyak 12,845 email *phising* unik, yang diamati bersamaan dengan penyebaran 2,560 situs web palsu sebagai metode untuk melakukan serangan *phising* (Sari & Sutabri, 2023).

Penting untuk dapat mendeteksi situs web *phising* dengan cepat dan akurat guna melindungi pengguna internet dari kehilangan data pribadi dan keuangan. Algoritma *Iterative Dichotomiser* 3 (ID3) adalah sebuah metode yang digunakan untuk membuat pohon keputusan. Pohon keputusan merupakan salah satu metode klasifikasi dengan model prediksi menggunakan struktur pohon(Ferdina et al., 2023). Penelitian yang dilakukan oleh (Lemantara, 2022), Algoritma ID3 adalah salah satu algoritma *Decision Tree* yang digunakan untuk menetukan keputusan dengan berbentuk akar pohon dan hasil dari penelitian ini Algoritma ID3 dapat dengan baik dapat mengklasifikasikan kualitas sebuah objek dengan baik dan sesuai dengan hasil yang diharapkan.

Algoritma ini telah digunakan secara luas untuk tugas-tugas seperti bagaimana mendeteksi, prediksi, ataupun mengklasifikasikan karena kemampuannya untuk membuat model pohon keputusan yang dapat menentukan apakah sebuah situs web cenderung *phising* atau tidak. Namun, ID3 memiliki keterbatasan dalam hal kemampuan adaptasi terhadap variasi data yang luas dan kompleks yang ditemukan dalam set data deteksi *phising*. Dalam konteks ini, optimasi algoritma klasifikasi ID3 dengan kombinasi metode Min-Max dan *Information Gain* menjadi penting untuk meningkatkan akurasi dan kinerja deteksi situs web *phising*.

Metode Min-Max digunakan untuk normalisasi nilai-nilai fitur dalam dataset, sehingga memungkinkan algoritma untuk mengoperasikan data dalam rentang nilai yang konsisten dan dapat dibandingkan (Permana, 2022). Berdasarkan penelitian yang dilakukan oleh (Sukmayadi et al., 2021), min-max normalization yang merupakan metode normalisasi dengan melakukan transformasi

linear pada atribut data asli sehingga menghasilkan keseimbangan nilai perbandingan antar data saat sebelum dan sesudah proses.

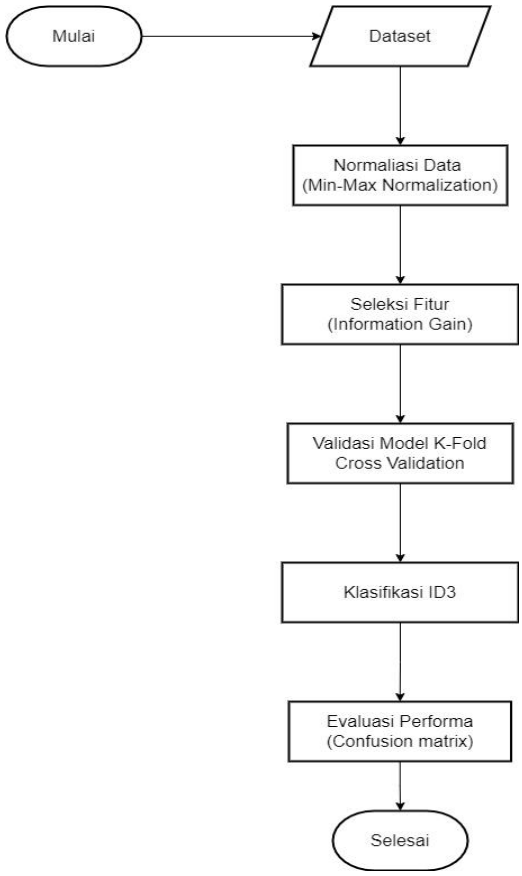
Sementara itu, *Information Gain* merupakan salah satu metode yang digunakan untuk memilih atribut terbaik. *Information Gain* dimanfaatkan untuk merangking kata-kata penting dari hasil reduksi atribut. Hasil dari proses *Information Gain* adalah kata penting yang bersifat informatif. Metode *Information Gain* dapat melihat setiap fitur untuk memprediksi label kelas yang benar yang karena memilih nilai yang tertinggi dan lebih efektif untuk mengoptimalkan hasil deteksi sebuah wesbite terindikasi *Phising* atau tidak(Fajriyan et al., 2022). Penelitian yang dilakukan oleh (Destitus et al., 2020), *Information Gain* adalah salah satu metode seleksi fitur yang sering digunakan oleh para peneliti untuk mengidentifikasi batas kepentingan atribut.

Dengan mengoptimalkan algoritma ID3 menggunakan pendekatan ini, diharapkan dapat meningkatkan akurasi dan efektivitas dalam mendeteksi website *phising*. Hal ini akan memberikan manfaat besar bagi pengguna internet, organisasi, dan lembaga keamanan cyber untuk melindungi informasi sensitif dan mencegah kerugian yang disebabkan oleh serangan *phising*. Selain itu, penelitian ini juga dapat memberikan kontribusi penting dalam pengembangan teknik deteksi ancaman keamanan cyber yang lebih canggih dan efisien di masa depan.

2. TINJAUAN PUSTAKA

Penelitian ini menggunakan beberapa algoritma machine learning untuk dilakukan perbandingan performa dalam pengklasifikasian analisis sentimen, yaitu Support Vector Machine, Random Forest, dan Naïve Bayes. Gambar 1 menunjukan tahapan penelitian perbandingan algoritma klasifikasi yang digunakan

Tahapan penelitian yang dilakukan dalam melakukan penelitian ini dapat dilihat pada gambar 1



Gambar 1. Diagaram Alur penelitian

Dataset

Dataset yang digunakan pada penelitian diambil dari penelitian yang dilakukan oleh. Dataset ini dapat diakses secara bebas dan publik di UCI *Machine Learning Repository*. Dataset ini berisi fitur-fitur yang menjadi patokan apakah sebuah website tersebut terindikasi *phising* atau tidak. Ada 30 fitur yang dijadikan parameter utama dan pada proses berikutnya akan diseleksi fitur mana yang memiliki nilai *Information Gain* yang kecil untuk dihilangkan dalam proses klasifikasi. 30 fitur dapat dilihat pada tabel 1.

Tabel 1. Fitur Dataset

No	Nama Fitur	Type
1	*having_ip_address*	*Integer*
2	*url_length*	*Integer*
3	*shortining_service*	*Integer*
4	*having_at_symbol*	*Integer*
5	*double_slash_redirecting*	*Integer*
6	*prefix_suffix*	*Integer*

7	*having_sub_domain*	*Integer*
8	*sslfinal_state*	*Integer*
9	*domain_registration_length*	*Integer*
10	*favicon*	*Integer*
11	*port*	*Integer*
12	*https_token*	*Integer*
13	*request_url*	*Integer*
14	*url_of_anchor*	*Integer*
15	*links_in_tags*	*Integer*
16	*sfh*	*Integer*
17	*submitting_to_email*	*Integer*
18	*abnormal_url*	*Integer*
19	*Redirect*	*Integer*
20	*on_mouseover*	*Integer*
21	*rightclick*	*Integer*
22	*popupwindow*	*Integer*
23	*iframe*	*Integer*
24	*age_of_domain*	*Integer*
25	*dnsrecord*	*Integer*
26	*web_traffic*	*Integer*
27	*page_rank*	*Integer*
28	*google_index*	*Integer*
29	*links_pointing_to_page*	*Integer*
30	*statistical_report*	*Integer*

30 fitur tabel yang ada di tabel 1 memiliki tipe data integer dengan nilai inputan : 0,1,-1. 0 artinya mencurigakan, 1 artinya tidak *phising*, dan -1 dipastikan *phising*, 1 Keterangan dari masing-masing atribut dapat dilihat sebagai berikut :

1. *Having_ip_address*
   Jika sebuah halaman web memiliki ip address dalam url nya makan terindikasi *phising*. Di bawah ini adalah *Rule* dari fitur having_ip_address

$$\begin{cases} If\ The\ Domain\ Part\ has\ an\ IP\ Address \rightarrow -1 \\ Otherwise \rightarrow 1 \end{cases}$$

2. url_length
   fitur ini merujuk berapa banyak atau Panjang karakter url. Di bawah ini adalah *Rule* dari fitur url_length :

$$Rule:$$
$$IF$$
$$\begin{cases} URL\ length < 54 \rightarrow feature = 1 \\ else\ if\ URL\ length \geq 54\ and\ \leq 75 \rightarrow feature = 0 \\ otherwise \rightarrow feature = -1 \end{cases}$$

3. shortening_service
   fitur ini merujuk apakah url tersebut memiliki singkatan dalam sebuah URL. *Rule* shortening_service dapat dilihat di bawah ini

$$\underline{Rule}: IF \begin{cases} TinyURL \rightarrow -1 \\ Otherwise \rightarrow 1 \end{cases}$$

4. having_at_symbol
   fitur ini merujuk apakah url tersebut memiliki simbol atau tidak. *Rule* dari fitur ini dapat dilihat sebagai berikut :

$$Rule: IF \begin{cases} Url\ Having\ @\ Symbol \rightarrow -1 \\ Otherwise \rightarrow 1 \end{cases}$$

5. *double_slash_redirecting*
   Fitur ini merujuk apakah sebuah url memiliki *double slash* yang menuju secara langsung ke website yang lain. *Rule* dari fitur ini dapat dilihat sebagai berikut :

$$Rule: \text{IF} \begin{cases} \text{"//" in the URL} > 7 \rightarrow -1 \\ \text{Otherwise} \rightarrow 1 \end{cases}$$

6. *prefix_suffix*
   Fitur ini merujuk apakah url memiliki tanda (-) pada sebuah url. Contohnya adalah sebagai berikut : http://www.Confirme-paypal.com/ *Rule* dari fitur ini dapat dilihat sebagai berikut :

$$Rule: IF \begin{cases} Domain\ Name\ (-)Symbol \rightarrow -1 \\ Otherwise \rightarrow 1 \end{cases}$$

7. *having_sub_domain*
   Fitur ini merujuk apabila sebuah domain memilik tanda dot (.). Tanda ini mengindikasikan bahwa di dalam domain memiliki sub domain. *Rule* dari fitur ini adalah sebagai berikut :

$$Rule: IF \begin{cases} Dots\ In\ Domain\ Part = 1 \rightarrow 1 \\ Dots\ In\ Domain\ Part = 2 \rightarrow 0 \\ Otherwise \rightarrow -1 \end{cases}$$

8. *sslfinal_state*
   Fitur ini merujuk sebuah url memiliki secure socket layer atau tidak. Apabila memiliki SSL kemudian dicari info berapa lama sertifikat SSL tersebut. *Rule* dari fitur ini adalah sebagai berikut :

Rule:

$$\begin{cases} IF \\ Use\ https\ and\ Age\ Of\ Certificate \geq 1\ Years \rightarrow 1 \\ Using\ https\ and\ Issuer\ Is\ Not\ Trusted \rightarrow 0 \\ Otherwise \rightarrow -1 \end{cases}$$

9. *domain_registeration_length*
   Fitur ini merujuk seberapa banyak Panjang domain tersebut sudah terdaftar atau berjalan. *Rule* dari fitur ini adalah sebagai berikut :

$$Rule:\ IF\begin{cases} Domains\ Expires\ on \leq 1\ years \rightarrow -1 \\ Otherwise \rightarrow 1 \end{cases}$$

10. *Favicon*
    Favicon adalah gambar grafis (ikon) yang terkait dengan halaman web tertentu. Banyak agen pengguna yang ada seperti peramban grafis dan pembaca berita menampilkan favicon sebagai pengingat visual identitas situs web di bilah alamat. Jika favicon dimuat dari domain yang berbeda dari yang ditunjukkan di bilah alamat, maka halaman web tersebut kemungkinan akan dianggap sebagai upaya *Phising*. *Rule* dari fitur ini adalah sebagai berikut :

Rule:

$$IF\begin{cases} Favicon\ Loaded\ External\ Domain \rightarrow -1 \\ Otherwise \rightarrow 1 \end{cases}$$

11. Port
    Fitur ini berguna untuk memvalidasi apakah layanan tertentu (misalnya HTTP) aktif atau tidak pada server tertentu. Dalam upaya mengendalikan intrusi, jauh lebih baik untuk hanya membuka port yang Anda butuhkan. Beberapa firewall, Proxy, dan server Network Address Translation (NAT) secara default akan memblokir semua atau sebagian besar port dan hanya membuka port yang dipilih. Jika semua port terbuka, phisher dapat menjalankan hampir semua layanan yang mereka inginkan dan akibatnya, informasi pengguna terancam. *Rule* dari fitur ini adalah sebagai berikut :

$$IF\begin{cases} Port\ \#\ is\ of\ the\ Preffered\ Status \rightarrow -1 \\ Otherwise \rightarrow 1 \end{cases}$$

12. *https_token*
    Fitur ini merujuk apabila penyerang mungkin menambahkan kata “HTTPS” ke nama domain untuk menipu korban. Contohnya adalah sebagai berikut :
    http://https-www-paypal-it-webapps-mpp-home.soft-hair.com/.
    Rule dari fitur ini adalah sebagai berikut :

$$IF\begin{cases} HTTP\ Token\ in\ Domain\ Part\ of\ The\ URL \rightarrow -1 \\ Otherwise \rightarrow 1 \end{cases}$$

13. *request_url*
    Fitur ini merujuk pada apakah objek eksternal yang ada di web seperti video, gambar, suara diambil dari domain lain. *Rule* dari fitur ini adalah sebagai berikut :

$$IF\begin{cases} \%\ of\ Request\ URL < 22\% \rightarrow 1 \\ \%of\ Request\ URL \geq 22\%\ and\ 61\% \rightarrow 0 \\ Otherwise \rightarrow feature = -1 \end{cases}$$

14. *url_of_anchor*
    Fitur ini merujuk pada Anchor adalah elemen yang didefinisikan oleh tag <a>. Fitur ini diperlakukan persis seperti "Request URL". *Rule* dari fitur ini adalah sebagai berikut :

$$IF\begin{cases} \%\ of\ URL\ Of\ Anchor < 31\% \rightarrow 1 \\ \%\ of\ URL\ Of\ Anchor \geq 31\%\ And \leq 67\% \rightarrow 0 \\ Otherwise \rightarrow 1 \end{cases}$$

15. *link_in_tags*
    Fitur ini merujuk pada berapa banyak sebuah website menggunakan tag <meta>, tag <script>, dan tag <link>. *Rule* dari fitur ini adalah sebagai berikut :

$$\begin{cases} IF \\ \%\ "<Meta>","<Script>"\ and\ "<Link>" < 17\% \rightarrow 1 \\ \%\ <Meta>","<Script>"\ and\ "<Link>" \geq 17\%\ And \leq 81\% \rightarrow 0 \\ Otherwise \rightarrow -1 \end{cases}$$

16. sfh
    SFH yang mengandung string kosong atau "about:blank" dianggap meragukan karena seharusnya ada tindakan yang diambil terhadap informasi yang dikirimkan. Selain itu, jika nama domain dalam SFH berbeda dari nama domain halaman web, ini mengungkapkan bahwa halaman web tersebut mencurigakan karena informasi yang dikirim jarang ditangani oleh domain eksternal. *Rule* dari fitur ini adalah sebagai berikut :
    theworldinyourhand

$$IF\begin{cases} SFH\ is\ "about:\ blank"\ Or\ Is\ Empty \rightarrow 1 \\ SFH\ Refers\ To\ A\ Different\ Domain \rightarrow 0 \\ Otherwise \rightarrow -1 \end{cases}$$

17. *submitting_to_email*
    Fitur ini merujuk pada sebuah website yang mengizinkan pengunjung websitenya untuk mengirimkan informasi personalnya lewat email dengan menggunakan fungsi mail() di bahasa pemrograman PHP. *Rule* dari fitur ini adalah sebagai berikut :
18. abnormal_url
    Fitur ini dapat diekstrak dari basis data WHOIS. Untuk situs web yang sah, identitas biasanya

merupakan bagian dari URL-nya. *Rule* dari fitur ini adalah sebagai berikut :

$$if\begin{cases}Host\ Name\ isn't\ Included\ In\ URL\ \rightarrow\ -1\\ Otherwise \rightarrow\ 1\end{cases}$$

19. *Redirect*
Fitur ini merujuk berapa kali sebuah nama domain web merujuk ke halaman website yang lain. Lewat penelitian yang dilakukan oleh apabila lebih dari 1 kali nama domain merujuk ke nama domain yang lain dapat diinformasikan nama domain tersebut mengandung *phising*. *Rule* dari fitur ini adalah sebagai berikut :

$$if\begin{cases}of\ Redirect\ Page \leq 1\ \rightarrow\ 1\\ of\ Redirect\ Page \geq 2\ And < 4 \rightarrow\ 0\\ Otherwise\ \rightarrow\ -1\end{cases}$$

20. *on_mouseover*
Fitur ini merujuk jika pelaku mungkin menggunakan javascript untuk menunjukkan URL palsu di status bar kepada pengguna. *Rule* dari fitur ini adalah sebagai berikut :

$$IF\begin{cases}onMouseOver\ Changes\ Status\ Bar \rightarrow\ -1\\ It\ Does't\ Change\ Status\ Bar \rightarrow 1\end{cases}$$

21. *rightclick*
Fitur ini merujuk bahwa pelaku menggunakan javascript untuk menonaktifkan fungsi klik kanan sehingga pengguna tidak dapat melihat dan menyimpan kode halaman web. *Rule* dari fitur ini adalah sebagai berikut

$$IF\begin{cases}Right\ Click\ Disabled\ \rightarrow\ -1\\ Otherwise \rightarrow 1\end{cases}$$

22. *popupwindow*
Fitur ini merujuk pada sebuah website yang mengizinkan user untuk mengirimkan informasi personal lewat pop up window. Hal ini tidak lazim dilakukan untuk sebuah website yang resmi dan merupakan Tindakan yang mencurigakan. *Rule* dari fitur ini adalah sebagai berikut :

$$IF\begin{cases}Popoup\ Window\ Contains\ Text\ Fields \rightarrow\ -1\\ Otherwise\ \rightarrow\ 1\end{cases}$$

23. *iframe*
IFrame adalah tag HTML yang digunakan untuk menampilkan halaman web tambahan ke dalam halaman yang sedang ditampilkan. Phisher dapat menggunakan tag "iframe" dan membuatnya tidak terlihat, misalnya tanpa bingkai frame. Dalam hal ini, phisher menggunakan atribut "*frameBorder*" yang menyebabkan browser menampilkan batas visual. *Rule* dari fitur ini adalah sebagai berikut :

$$if\begin{cases}Using\ iframe \rightarrow\ -1\\ Otherwise\ \rightarrow\ -1\end{cases}$$

24. *age_of_domain*
Fitur ini merujuk berapa lama umur domain website. Biasanya kebanyakan website *phising* tidak berumur Panjang seperti website resmi atau website non *phising*. *Rule* dari fitur ini adalah sebagai berikut :

$$if\begin{cases}Using\ iframe \rightarrow\ -1\\ Otherwise\ \rightarrow\ -1\end{cases}$$

25. *dnsrecord*
Jika rekaman DNS kosong atau tidak ditemukan maka sebuah website dapat diklasifikasikan *phising* dan sebaliknya website tersebut bersih dari *phising*. *Rule* dari fitur ini adalah sebagai berikut :

$$if\begin{cases}no\ DNS\ Recordn\ \rightarrow\ -1\\ Otherwise\ \rightarrow\ -1\end{cases}$$

26. *web_traffic*
Fitur ini merujuk pada rangking dari sebuah website. Ranking website dapat diambil dari website alexa. Website Alexa adalah website yang menampilkan rangking dari sebuah website. *Rule* dari fitur ini adalah sebagai berikut :

$$if\begin{cases}Website\ Rank < 100{,}000\ \rightarrow\ 1\\ Website\ Rank > 100{,}000\ \rightarrow 0\\ Otherwise\ \rightarrow\ -1\end{cases}$$

27. *page_rank*
PageRank adalah nilai yang berkisar dari "0" hingga "1". *PageRank* bertujuan untuk mengukur seberapa penting sebuah halaman web di Internet. Semakin besar nilai *PageRank*, semakin penting halaman web tersebut. Pada penelitian ditemukan bahwa sekitar 95% halaman web *phising* tidak memiliki PageRank. Selain itu, ditemukan bahwa 5% sisanya dari halaman web *phising* dapat mencapai nilai PageRank hingga "0.2. *Rule* dari fitur ini adalah sebagai berikut :

$$if\begin{cases}PageRank < 0.2\ \rightarrow\ -1\\ Otherwise\ \rightarrow\ -1\end{cases}$$

28. *google_index*
Fitur ini memeriksa apakah sebuah situs web terdaftar dalam indeks Google atau tidak.. Biasanya, halaman web *phising* hanya dapat diakses untuk jangka waktu yang singkat dan akibatnya, banyak halaman web *phising* mungkin tidak ditemukan dalam indeks Google. *Rule* dari fitur ini adalah sebagai berikut :

$$if\begin{cases}Webpage\ Indexed\ by\ Google\ \rightarrow\ 1\\ Otherwise\ \rightarrow\ -1\end{cases}$$

29. *links_pointing_to_page*

Fitur ini mengacu pada bahwa biasanya halaman web yang resmi atau normal biasanya ada minimal 2 tautan yang mengarah ke website resmi, sedangkan website *phising* tidak karena website *phising* tidak bertahan lama atau berumur pendek dan tidak dapat diakses.

$$if \begin{cases} Pointing\ to\ The\ Webpage = 0 \rightarrow -1 \\ Pointing\ to\ The\ Webpage > 0\ and \leq 2 \rightarrow 0 \\ Otherwise \rightarrow -1 \end{cases}$$

30. statistical_report

Rule dari fitur ini adalah sebagai berikut :

$$if \begin{cases} Top\ Phising\ Domains \rightarrow -1 \\ Otherwise \rightarrow -1 \end{cases}$$

Normalisasi Data (Min-Max)

Proses normalisasi data adalah proses mengubah data sehingga konsistensi data dapat tetap baik dan dapat meningkatkan efisiensi dan nilai akurasi dalam proses pembelajaran. Pada penelitian ini untuk memperkecil range data perlu dilakukan proses normalisasi data dengan menggunakan min-max normalization dengan melakukan transformasi linear dan menghasilkan keseimbangan perbandingan antara data sebelum dan sesudah proses normalisasi. Normalisasi ini dapat dilakukan dengan menggunakan persamaan 1 (Yanasari & Arifin, 2023).

$$x' = \frac{x - nilai_{min}}{nilai_{max} - nilai_{min}} \ldots (1)$$

Seleksi Fitur *Information Gain*

Seleksi fitur adalah tahapan penting dalam proses pembelajaran mesin yang memiliki fungsi untuk memilih fitur terbaik dan paling relevan untuk digunakan dalam proses klasifikasi. Tujuan dari memilih fitur ini adalah mengurangi data yang tidak memiliki peran yang penting dalam suatu dataset, meningkatkan kinerja model, mengurangi waktu komputasi. Dalam penelitian metode seleksi fitur yang digunakan adalah metode *Information Gain*(Rahmanita et al., 2023).

Information Gain bekerja dengan memilih fitur-fitur yang memiliki bobot tertinggi sesuai dengan jumlah fitur yang diinginkan. *Information Gain* menggunakan entropi untuk menentukan term terbaik. Adapun Langkah dalam pembobotan fitur dengan information gian adalah sebagai berikut :

a. Menentukan entropy dengan menggunakan persamaan 2

$$Entropy(S) = \sum_{i}^{c} - P_i \log_2 P_i \ldots (2)$$

Dimana :

c = Total nilai yang ada pada kelas klasifikasi

P_i = Total sampel untuk kelas i

Algoritma *Iterative Dichotomiser* 3

Proses klasifikasi yang digunakan untuk menguji apakah proses penghilangan fitur dapat berdampak pada akurasi adalah dengan menggunakan algoritma ID3. Algoritma ID3 (*Iterative Dichotomiser* 3) adalah salah satu algoritma pembelajaran mesin yang digunakan untuk membangun pohon keputusan (decision tree). Algoritma ini dikembangkan oleh Ross Quinlan pada tahun 1986 (Kapri et al., 2021).

Algoritma ID3 memiliki cara kerja sebagai berikut (Hikmatulloh, Nurajizah, 2021): Ambil semua atribut yang tidak terpakai dan hitung entropinya yang berhubungan dengan test sample. Pilih atribut dimana nilai entropinya minimum. Buat simpul yang berisi atribut tersebut.

Cross Validation

Cross validation atau dapat disebut estimasi rotasi adalah sebuah teknik validasi model untuk menilai bagaimana hasil statistik analisis akan menggeneralisasi kumpulan data independent. Tujuannya adalah memastikan bahwa model yang dibangun tidak hanya bekerja baik pada data pelatihan tetapi juga pada data yang belum pernah dilihat sebelumnya. Ada beberapa metode cross validation yang dapat digunakan, untuk penelitian ini metode yang digunakan adalah K-Fold Cross Validation (Azis et al., 2020).

Metode ini memecah data menjadi K bagian set data dengan ukuran yang sama. Contoh: Jika k = 5, dataset dibagi menjadi 5 folds, dan model dilatih dan diuji 5 kali, setiap kali menggunakan 4 folds sebagai pelatihan dan 1 fold sebagai uji (Irawan et al., 2024).

Confusion Matrix

Untuk mengukur evaluasi dari performa ketiga algoritma machine learning, teknik yang yang digunakan adalah teknif Confusion Matrix. Confusion Matrix, adalah cara tabel untuk memvisualisasikan kinerja model prediksi pada pembelajaran supervised learning. Setiap data dari masing-masing kelas dalam tabel confusion matrix menunjukkan jumlah prediksi yang dibuat guna untuk mengklasifikasikan kelas yang benar atau salah (Tarigan, 2023). Teknik ini digunakan untuk menghitung nilai akurasi, presisi, dan *recall*. Tabel Confusion Matrix dapat dilihat pada table 2.

Tabel 2. *Confusion Matrix*

	Prediksi	
Aktual	True	False
True	TP	FP

False	FN	TN

Keterangan :

- True Posiitive (TP), ini adalah jumlah dari satu kelas TRUE yang bisa di prediksi dengan benar pada kelas TRUE.
- True Negative (TN), adalah jumlah dari satu kelas FALSE yang bisa di prediksi dengan benar pada kelas FALSE
- False Positive (FP), ini adalah kondisi dimana kelas TRUE yang prediksinya salah pada kelas FALSE, sedangkan
- False Negatif (FN), adalah dimana kondisi pada kelas FALSE yang di prediksi salah pada kelas TRUE.

1. *Accuracy,* ini adalah ukuran kinerja yang akan memberikan tingkat keakuratan dari keseluruhan model atau dalam penjelasan lain adalah menghitung semua prediksi yang benar dari total jumlah data. Berikut persamaan *accuracy* yang dapat dilihat pada persamaan 3 *:*

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} x100 \quad (3)$$

2. Presisi, adalah ukuran kinerja yang akan memberi informasi dari prediksi sebagai kelas positif yang sebenarnya positif. Berikut persamaan presisi yang dapat dilihat pada persamaan 4 :

$$\text{Presisi} = \frac{TP}{(TP + FP)} x100 \quad (4)$$

3. *Recall*, adalah ukuran kinerja yang memberi iinformasi dari prediksi kelas positif yang di prediksi negatif. Berikut persamaan *recall* yang dapat dilihat pada persamaan 5 :

$$Recall = \frac{TP}{(TP + FN)} x\mathbf{100} \quad (\mathbf{5})$$

3. Hasil Dan Pembahasan

Sudah dijelaskan pada subbab sebelumnya sumber data diambil dari UCI Repository dengan jumlah data sebanyak 22110 data dengan kategori hasil dibagi menjadi 2, yaitu -1 dan 1. -1 hasil website tersebut pasti merupakan website *phising* dan 1 bukan website *phising*. Dari dataset yang digunakan jumlah data yang bernilai -1 ada sebanyak 12.313 data dan bernilai 1 sebanyak 9.800 data.

Dari dataset tersebut akan dilakukan proses normalisasi data dengan menggunakan metode min-max dengan tujuan dataset tersebut dapat ternormalisasi dan dapat menghasilkan nilai akurasi yang lebih baik. Proses normalisasi dapat ditampilkan pada koding pyhton yang ditampilkan pada gambar 2.

```
# Normalisasi data menggunakan Min-Max Scaler
scaler = MinMaxScaler()
X_normalized = scaler.fit_transform(X)
```

Gambar 2. Koding Pyhton Untuk Normalisasi Min-Max.

Setelah dilakukan proses normalisasi, selanjutnya proses seleksi fitur dapat dilakukan dengan menggunakan *Information Gain* Pembobotan Menggunakan *Information Gain Dataset* yang digunakan untuk perhitungan *Information Gain* adalah dipakai sebanyak 22.110 data dengan 30 fitur. Pencarian nilai *Information Gain* untuk masing-masing fitur dilakukan dengan menggunakan kode pyhton seperti yang ditampilkan pada gambar 3.

```
# Diskritisasi data menggunakan KBinsDiscretizer
discretizer = KBinsDiscretizer(n_bins=10, encode='ordinal', strategy='uniform')
X_discretized = discretizer.fit_transform(X_normalized)

# Menghitung Information Gain
information_gain = mutual_info_classif(X_discretized, y, discrete_features=True)
```

Gambar 3. Kode Pyhton Untuk Metode *Information Gain*

Hasil perhitungan nilai 30 fitur dengan *Information Gain* dari yang tertinggi hingga terendah seperti yang ditunjukkan pada tabel 2.

Tabel 3. Hasil Pembobotan *Information Gain*

Peringkat	Nama Fitur	Nilai
1	SSLfinal_State	0.3462
2	URL_of_Anchor	0.3308
3	Prefix_Suffix	0.0856
4	web_traffic	0.0794
5	having_Sub_Domain	0.0761
6	Links_in_tags	0.0326
7	Request_URL	0.0323
8	SFH	0.0260
9	Domain_registeration_length	0.0255
10	Google_Index	0.0083
11	age_of_domain	0.0074

12	Page_Rank	0.0055
13	having_IP_Address	0.0044
14	Statistical_report	0.0032
15	Links_pointing_to_page	0.0030
16	DNSRecord	0.0029
17	URL_Length	0.0026
18	Shortining_Service	0.0023
19	Abnormal_URL	0.0019
20	having_At_Symbol	0.0014
21	on_mouseover	0.0009
22	double_slash_redirecting	0.0008
23	HTTPS_token	0.0008
24	port	0.0007
25	Redirect	0.0002
26	Submitting_to_email	0.0002
27	RightClick	0.0001
28	Favicon	0.0000
29	Iframe	0.0000
30	popUpWidnow	0.0000

Dari tabel 3, Hasil perhitungan *Information Gain* untuk 30 fitur dalam dataset menunjukkan bahwa beberapa fitur memiliki tingkat informatif yang jauh lebih tinggi dibandingkan fitur lainnya. Tiga fitur teratas dengan nilai *Information Gain* tertinggi adalah **SSLfinal_State**, **URL_of_Anchor**, dan **Prefix_Suffix**, dengan nilai masing-masing 0.3462, 0.3308, dan 0.0856. Fitur-fitur ini memiliki kontribusi signifikan dalam memprediksi variabel target.

Sebagian besar fitur lainnya memiliki nilai *Information Gain* yang jauh lebih rendah, bahkan beberapa fitur seperti **Favicon**, **Iframe**, dan **popUpWidnow** memiliki nilai *Information Gain* 0.0000, menunjukkan bahwa fitur-fitur ini tidak memberikan informasi yang berguna untuk prediksi dalam konteks dataset ini.

Ketiga fitur yang memiliki nilai *Information Gain* sebesar 0 selanjutnya akan dihilangkan dari proses klasifikasi kemudian akan diuji nilai akurasi apabila dihilangkan ketiga fitur tersebut apakah lebih baik dibandingkan tanpa menghilangkan ketiga fitur yang memiliki nilai *Information Gain* sebesar 0.

Setelah mendapatkan fitur mana saja yang dihilangkan, selanjutnya akan dilakukan proses klasifikasi dengan menggunakan algoritma ID3. Proses klasifikasi akan dilakukan dengan menggunakan keseluruhan fitur dan menghilangkan 3 fitur, yaitu **Favicon**, **Iframe**, dan **popUpWidnow.** Proses evaluasi ini akan menggunakan proess Cross Validation sebanyak 8 Fold Dimana akan menghasilkan nilai confusion matrix berupa nilau akurasi, presisi, dan recall untuk masing-masing pengujian baik itu dengan menggunakan 27 fitur dan 30 fitur. Pada tabel 4 dan 5 akan diperlihatkan hasil pengujian dengan menggunakan algoritma ID3

Tabel 4. Kinerja Algoritma ID3 Dengan Semua Fitur Sebanyak 8 K-Fold

K-Fold	Akurasi	Presisi	Recall
1	0.9844	0.9769	0.9954
2	0.9859	0.9858	0.9890
3	0.9863	0.9846	0.9909
4	0.9826	0.9868	0.9817
5	0.9848	0.9845	0.9883
6	0.9895	0.9914	0.9895
7	0.9826	0.9890	0.9801
8	0.9870	0.9891	0.9878
Rata-Rata	0.9854	0.9860	0.9878

Dari tabel 3, dapat dianalisis sebagai berikut Fold 1 memiliki akurasi yang sedikit lebih rendah (0.9844) dibandingkan fold lainnya, tetapi recall sangat tinggi (0.9954), menunjukkan model sangat efektif dalam mendeteksi instance positif di fold ini. Fold 6 menunjukkan kinerja terbaik dengan akurasi tertinggi (0.9895), presisi tertinggi (0.9914), dan recall yang sangat baik (0.9895), menandakan performa model yang sangat kuat pada subset data ini. Fold 4 dan Fold 7 memiliki akurasi yang sedikit lebih rendah dibandingkan fold lainnya, tetapi masih dalam rentang yang sangat baik, menunjukkan konsistensi model secara keseluruhan.

Untuk hasil pengujian dengan menggunakan 27 fitur dapat dilihat pada tabel 4 sebagai berikut :

Tabel 5. Kinerja Algoritma ID3 Dengan 27Fitur Sebanyak 8 K-Fold

K-Fold	Akurasi	Presisi	Recall
1	0.9866	0.9806	0.9954

2	0.9873	0.9870	0.9903
3	0.9863	0.9858	0.9896
4	0.9823	0.9868	0.9810
5	0.9848	0.9845	0.9883
6	0.9888	0.9901	0.9895
7	0.9823	0.9877	0.9895
8	0.9862	0.9878	0.9878
Rata-Rata	0.9856	0.9863	0.9878

Dengan menggunakan 27 fitur menghasilkan rata-rata untuk akurasi sebesar 0.9856, presisi 0.9863, dan recall sebesar 0.9878. Hasil ini menunjukkan terdapat peningkatan sedikit dibandingkan dengan menggunakan semua fitur. Meskipun perbedaan kinerja ini sangat kecil, hasil ini menunjukkan bahwa pemilihan fitur yang tepat berdasarkan *Information Gain* dapat membantu dalam meningkatkan performa model. Menggunakan 27 fitur terbaik memberikan akurasi dan presisi yang sedikit lebih tinggi dibandingkan dengan menggunakan semua 30 fitur. Ini mengindikasikan bahwa beberapa fitur mungkin tidak memberikan kontribusi signifikan terhadap kinerja model atau bahkan bisa menjadi sumber *noise*.

4. KESIMPULAN

Kesimpulan dari penelitian ini adalah terdapat 3 fitur yang memiliki nilai *Information Gain* yang bernilai 0 yaitu *Favicon, Iframe, dan popUpWidnow*. Dengan menghilangkan ketiga fitur ini dapat meningkatkan hasil akurasi, presisi, dan *recall*. Model ID3 dengan 27 fitur terbaik menunjukkan kinerja yang sangat baik dan konsisten dalam pengujian. Rata-rata akurasi, presisi, dan recall yang tinggi menegaskan bahwa model ini efektif dalam melakukan klasifikasi dengan tingkat kesalahan yang sangat rendah. Proses seleksi fitur yang tepat berkontribusi pada pencapaian kinerja optimal ini, menjadikan model ID3 alat yang sangat efektif untuk klasifikasi dalam konteks dataset yang digunakan. Hal ini menunjukkan pentingnya proses seleksi fitur dalam membangun model yang efisien dan efektif, serta pentingnya mengidentifikasi fitur-fitur yang paling relevan terhadap target prediksi.

DAFTAR PUSTAKA

Amin Muftiadi. (2022). Studi kasus keamanan jaringan komputer: analisis ancaman *phising*terhadap layanan online banking. *Hexatech: Jurnal Ilmiah Teknik*, *1*(2), 60–65.

Annisa Indah Mutiasari, Anggit Dyah Kusumastuti, Rusnandari Retno Cahyani, F. H. S. A. H. (2024). Analisis Customer Rating dan Customer Review terhadap Keputusan Pembelian Produk Makanan Secara Online. *Jurnal Simki Economic*, *7*(2), 357–366.

Azis, H., Purnawansyah, P., Fattah, F., & Putri, I. P. (2020). Performa Klasifikasi K-NN dan Cross Validation Pada Data Pasien Pengidap Penyakit Jantung. *ILKOM Jurnal Ilmiah*, *12*(2), 81–86. https://doi.org/10.33096/ilkom.v12i2.507.81-86

Destitus, C., Wella, W., & Suryasari, S. (2020). Support Vector Machine VS *Information Gain*: Analisis Sentimen Cyberbullying di Twitter Indonesia. *Ultima InfoSys : Jurnal Ilmu Sistem Informasi*, *11*(2), 107–111. https://doi.org/10.31937/si.v11i2.1740

Fajriyan, F. N., Ahsan, M., & Harianto, W. (2022). Komparasi Tingkat Akurasi *Information Gain* Dan Gain Ratio Pada Metode K-Nearest Neighbor. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *6*(1), 386–391. https://doi.org/10.36040/jati.v6i1.4694

Ferdina, F., Satyahadewi, N., & Kusnandar, D. (2023). Penerapan Algoritma *Iterative Dichotomiser* 3 (Id3) Dalam Klasifikasi Faktor Risiko Penyakit Diabetes Melitus. *VARIANCE: Journal of Statistics and Its Applications*, *5*(2), 139–146. https://doi.org/10.30598/variancevol5iss2page139-146

Hikmatulloh, Nurajizah, S. (2021). Peningkatan Akurasi Pada Algoritma ID3 Menggunakan Operator Bagging Dalam Mendiagnosa Kesehatan Kehamilan. *IJCIT (Indonesian Journal on Computer and Information Technology) 6*, *6*(2), 92–96.

Irawan, R. N., Hindrayani, K. M., & Idhom, M. (2024). Penerapan Cross Validation sebagai Analisis Sentimen Pelayanan Publik Kereta Api Lokal Daop 8 Menggunakan Metode Multinomial Naïve Bayes. *G-Tech: Jurnal Teknologi Terapan*, *8*(2), 954–963. https://doi.org/10.33379/gtech.v8i2.4117

Kapri, T., Nasir, M., & Agustini, E. P. (2021). Implementasi Algoritma *Iterative Dichotomiser* 3 (ID3) Untuk Penentuan Jumlah Dana Bantuan Perbaikan Rumah Di Bappeda. *Journal of Computer and Information Systems*

Ampera, *2*(1), 58–67. https://doi.org/10.51519/journalcisa.v2i1.70

Lemantara, J. (2022). Penerapan Algoritma Naïve Bayes dan ID3 untuk Memprediksi Segmentasi Pelanggan pada Penjualan Mobil. *Journal of Technology and Informatics (JoTI)*, *4*(1), 31–40. https://doi.org/10.37802/joti.v4i1.265

Permana, I. (2022). The Effect of Data Normalization on the Performance of the Classification Results of the Backpropagation Algorithm Pengaruh Normalisasi Data Terhadap Performa Hasil Klasifikasi Algoritma Backpropagation. *Indonesian Journal of Informatic Research and Software Engineering*, *2*(1), 67–72. https://media.neliti.com/media/publications/485639-pengaruh-normalisasi-data-terhadap-perfo-e19e3a00.pdf

Rahmanita, E., Negara, Y. D. P., Kustiyahningsih, Y., Sasmeka, V., & Khotimah, B. K. (2023). Implementasi Metode Naïve Bayes dan *Information Gain* Untuk Klasifikasi Penyakit dan Hama Tanaman Jagung. *Teknika*, *12*(3), 198–204. https://doi.org/10.34148/teknika.v12i3.684

Sari, P., & Sutabri, T. (2023). Analisis kejahatan online *phising* pada institusi pemerintah/pendidik sehari-hari. *Jurnal Digital Teknologi Informasi*, *6*(1), 29. https://doi.org/10.32502/digital.v6i1.5620

Sukmayadi, C., Primajaya, A., & Maulana, I. (2021). Penerapan Algoritma K-Medoids dalam Menentukan Daerah Rawan Banjir di Kabupaten Karawang. *INFORMAL: Informatics Journal*, *6*(3), 187. https://doi.org/10.19184/isj.v6i3.25423

Tarigan, V. (2023). Pembuatan Aplikasi Data Mining Untuk Memperediksi Masa Studi Mahasiswa Menggunakan Algoritma Naive Bayes. *Informatika*, *11*(1), 54–62. https://doi.org/10.36987/informatika.v11i1.3847

Yanasari, C., & Arifin, T. (2023). Implementasi Algoritma K-Nearest Neighbor Untuk Klasifikasi Penerimaan Beasiswa Program Indonesia Pintar. *Jurnal Sistem Informasi Dan Ilmu Komputer*, *1*(4), 178–194. https://doi.org/10.59581/jusiik-widyakarya.v1i4.1862