ANALISIS PERBANDINGAN KORELASI SPEARMAN DAN MAXIMAL INFORMATION COEFFICIENT DALAM SELEKSI FITUR WEBSITE PHISHING MENGGUNAKAN ALGORITMA MACHINE LEARNING

COMPARATIVE ANALYSIS OF SPEARMAN CORRELATION AND MAXIMAL INFORMATION COEFFICIENT IN PHISHING WEBSITE FEATURE SELECTION USING MACHINE LEARNING ALGORITHMS

Jimmy H. Moedjahedy*[1], Arief Setyanto[2], Komang Aryasa[3]
MTI Universitas Amikom Yogyakarta, Universitas Amikom Yogyakarta,
STMIK Dipanegara Makassar
e-mail: *[1]jimmy@unklab.ac.id, [2]arief_s@amikom.ac.id, [3]komang.aryasa@dipanegara.ac.id

Abstrak

Phishing adalah kejahatan yang menggunakan teknik social engineering, baik dalam pernyataan yang menipu maupun secara teknis untuk mencuri data identitas pribadi konsumen dan kredensial akun keuangan. Phishing dirancang untuk mengarahkan konsumen ke website phishing yang menipu penerima untuk membocorkan data keuangan seperti nama pengguna dan kata sandi. Dalam dataset phishing, terdapat fitur-fitur yang bisa mengkategorikan apakah sebuah website adalah website phishing atau bukan. Tujuan dari penelitian ini adalah untuk membandingkan hasil seleksi fitur-fitur yang ada dengan menggunakan dua metode yaitu metode gabungan Maximal Information coefficient dan Total Information Coefficient dengan metode korelasi Spearman. Hasil seleksi diuji dengan lima algoritma machine learning yaitu, Logistic Regression, Naïve Bayes, J48, AdaBoost MI dan *Random Forest. Hasil dari penelitian ini adalah metode gabungan Maximal Information coefficent dan Total Information Coefficient memiliki nilai akurasi 97.25 % dengan menggunakan Random Forest mengungguli metode korelasi Spearman dengan nilai akurasi 95,33%.*

Kata kunci*—Maximal Information Coefficient, Total Information Coefficient,* korelasi *Spearman,* Seleksi fitur, *Phishing*

Abstract

Phishing is a crime that uses social engineering techniques, both in a supportive and technical settlement, to win consumer's identity data and financial account credentials. Phishing is designed to direct consumers to a Phishing website that invites recipients to divulge financial data such as usernames and passwords. In a phishing dataset, some features can categorize whether a website is a phishing website or not. This study compares the results of the selection of existing features using two methods: the combined method of Maximum Information Coefficient and Total Information Coefficient with the Spearman comparison method. The selection results quote five machine learning algorithms, namely, Logistic Regression, Naïve Bayes, J48, AdaBoost MI, and Random Forest. This study's results are the combined method of Maximal Information coefficient and Total Information Coefficient, which has an accuracy value of 97.25% using Random Forest to outperform the Spearman correlation method with an accuracy value of 95.33%.

Keywords*—Maximal Information Coefficient, Total Information Coefficient, Spearman Correlation, Features Selection, Phishing*

1. PENDAHULUAN

Internet telah menjadi sebuah infrastruktur yang banyak memberikan kenyamanan pada manusia. Namun dalam penggunaannya ada beberapa masalah keamanan seperti *phishing* atau penipuan, perangkat lunak berbahaya, kebocoran privasi yang dapat mengancam ekonomi dari pengguna internet. Salah satu asosiasi yang merupakan asosiasi internasional dalam bidang kejahatan di dunia internet yaitu APWG (Anti-Phising Working Group), menuliskan bahwa *phishing* adalah kejahatan yang menggunakan teknik *social engineering*, baik dalam pernyataan yang menipu maupun secara teknis untuk mencuri data identitas pribadi konsumen dan kredensial akun keuangan. Skema *social engineering* memangsa korban yang tidak waspada dengan membodohi mereka agar percaya bahwa mereka berurusan dengan pihak yang terpercaya dan sah, seperti dengan menggunakan alamat *email* dan pesan *email* yang menipu. Ini dirancang untuk mengarahkan konsumen ke *website* palsu yang menipu penerima untuk membocorkan data keuangan seperti nama pengguna dan kata sandi [1] [2].

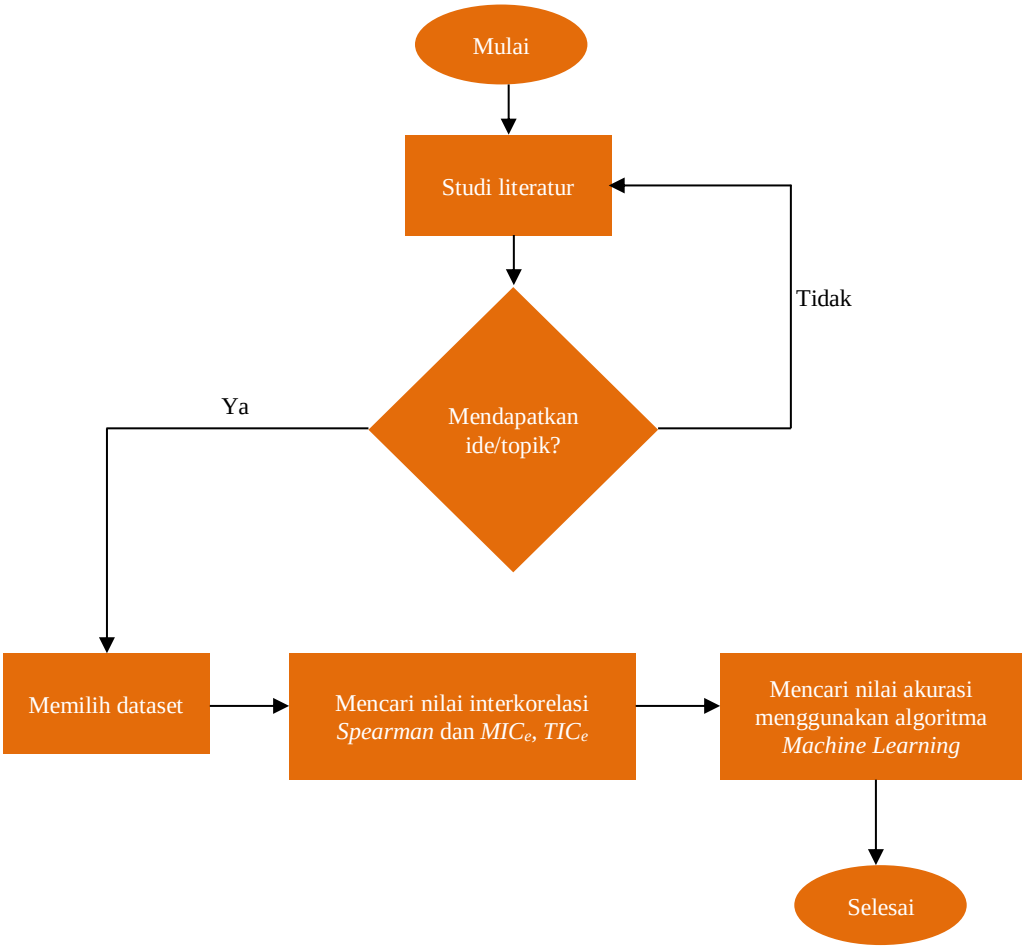
Menurut [3], situs web palsu atau yang disebut dengan *phishing website* sulit untuk diidentifikasi karena *website* tersebut bisa sering berubah atau masa aktif *domain* yang sudah habis waktu. Beberapa penelitian yang terkait dengan topik ini mencoba untuk menyimpulkan apa saja komponen-komponen atau ciri-ciri sehingga *website* tersebut dikategorikan *website phising*. Penelitian yang dilakukan oleh [4], mereka mengategorikan berdasarkan pola URL dan mengekstrak kata kunci dari website tersebut. Selanjutnya penelitian dari [5], dimana mereka megkategorikan berdasarkan *browser extention* yang dikembangkan. Cara yang dilakukan oleh [6] adalah dengan menggunakan dua metode, yaitu seleksi fitur dengan menggunakan pohon keputusan dan *wrapper*. Penelitian berikutnya adalah yang dilakukan oleh [7] yang melakukan seleksi fitur dengan menggunakan *Variance Inflation Factor (VIF)*. Konsep penelitian dengan mengeluarkan nilai berdasarkan sebuah nilai juga dilakukan oleh [8], tetapi mereke mengeluarkan berdasarkan data pencilan.

Penelitian yang kami lakukan adalah menggunakan metode seleksi fitur berdasarkan nilai korelasi *Spearman* dan kombinasi *Maximal Information Coefficent* (MIC_e), *Total Information Coefficient* (TIC_e) dengan membandingkan nilai akurasi yang dihasilkan oleh kedua metode tersebut menggunakan algoritma *machine learning*. Algoritma *machine learning* yang digunakan adalah algoritma *Random Forest*. Kami memilih *algoritma Logistic Regression (LR), Naïve Bayes (NB), J48, AdaBoost M1 (AM)* dan *Random Forest (RF)* dalam penelitian ini. Dalam beberapa penelitian yang berkaitan dengan deteksi *phishing webiste*, *Random Forest* mengungguli algoritma *machine learning* lainnya seperti penelitian-penelitian yang dilakukan oleh [9], [10], [11], dan [12], maka kami membandingkan lima algoritma *machine learning* tersebut. Adapun dataset yang akan digunakan adalah dataset yang dikembangkan oleh [13], yang terdiri dari 30 fitur dan 11055 data yang sudah memiliki kategori *phishing* dan bukan *phishing*.

2. METODE PENELITIAN

2.1 Tahapan Penelitian

Tahapan penelitian dimulai dengan melakukan studi literatur dari penelitian-penelitian yang berkaitan dengan topik penelitian ini. Kemudian dilakukan pemilihan dataset yang digunakan untuk tujuan penelitian. Selanjutnya melakukan pencarian nilai interkorelasi atau fitur-fitur yang memiliki nilai korelasi yang tinggi satu sama lain. Pencarian nilai digunakan dengan metode korelasi *Spearman* dan MIC_e, TIC_e. Setelah mendapatkan nilai tersebut, dikeluarkan fitur yang memiliki nilai interkorelasi 0.5-1. Langkah selanjutnya adalah membandingan dengan menggunakan lima algoritma *machine learning* yaitu *Logistic Regression, Naïve Bayes, J48, AdaBoost MI* dan *Random Forest* dan membandingan nilai akurasi dari kedua metode tersebut. Alur penelitian dapat dilihat pada gambar 1.



Gambar 1. Alur Penelitian

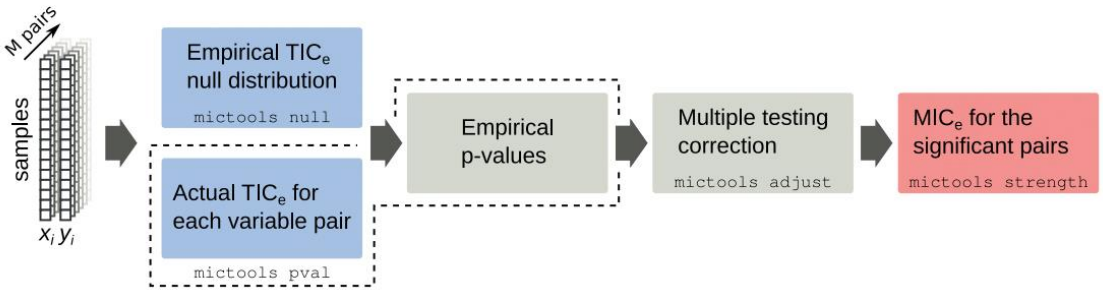
2.2 Korelasi Spearman

Menurut [14], korelasi *Spearman* merupakan metode korelasi non parametrik yang didesain untuk mengukur hubungan yang kuat antara dua set rangking data. Jenis korelasi ini digunakan jika data yang akan diuji berjenis data ordinal dan digunakan juga jika ingin mengukur konsistensi dari hubungan antar variabel X dan Y. Dalam buku yang ditulis oleh [15], keuntungan dari korelasi *Spearman* adalah analisis berbasis rangking komputasinya lebih mudah dibandingkan dengan analisis berbasis angka numerik, selain itu juga dapat digunakan untuk mencari tahu korelasi non linear. Metode korelasi ini menghitung rangking korelasi mulai dari -1 yang artinya korelasi sempurna dalam derajat kemiringan negatif dan +1 yaitu korelasi sempurna dalam derajat kemiringan sempurna. Selain nilai -1 dan +1, nilai diantara kedua angka tersebut, menurut buku yang dikutip dari [16] jika angka di atas 0,5 atau di bawah 0,5 maka itu disebut dengan hubungan yang moderat atau cukup kuat. Dalam penelitian ini, kami akan mengeluarkan fitur dalam data set yang nilainya lebih besar sama dengan 0.5 sehingga tersisa fitur-fitur yang tidak saling berhubungan.

2.3 MIC$_e$ *dan TIC*$_e$

Pada tahun 2011, [17] mengajukan sebuah metode korelasi antar dua variabel yang diberi nama *Maximal Information Coefficient (MIC)* yang merupakan bagian dari sebuah kelas yang lebih besar lagi yaitu *maximal information-based nonparametric exploration (MINE).* Konsep *MIC* adalah mendapatkan berbagai asosiasi baik fungsional maupun bukan fungsional, dan untuk hubungan fungsional memberikan skor yang kira-kira sama dengan koefisien determinasi (R^2). Namun menurut beberapa penelitian sesudah itu, dijelaskan bahwa MIC tidak menghasilkan sesuatu yang baru atau

state of the art. Karena kekurangan yang didapat oleh peneliti-peneliti lainnya, maka muncul konsep *MIC* baru yaitu kombinasi MIC_e dan TIC_e [18]. Kombinasi dari kedua metode ini dikatakan sangat baik dimana TIC_e digunakan untuk melakukan secara efisien dalam hal penyaringan hasil yang tinggi dari semua hubungan berpasangan yang mungkin dan menilai signifikansinya, sementara MIC_e digunakan untuk memberi peringkat subset dari asosiasi yang signifikan dalam hal kekuatan.



Gambar 2. proses pencarian nilai MIC_e [19]

Untuk mencari nilai MIC_e dari setiap pasangan fitur digunakan *Mictools* yang dikembangkan oleh [19], berikut adalah prosesnya:

1. Memperkirakan nili distribusi nol TIC_e empiris dengan 200.000 permutasi.
2. Menghitung statistik TIC_e dan nilai p empiris untuk setiap pasangan variabel.
3. Menerapkan beberapa strategi koreksi pengujian untuk mengontrol tingkat kesalahan berdasarkan *Family-Wise Error Rate (FWER)*
4. Menggunakan MIC_e untuk memperkirakan kekuatan hubungan antar variabel yang disebut signifikan.

3. HASIL DAN PEMBAHASAN

Pada tabel 1, merupakan 30 fitur yang ada dalam dataset yang diteliti, kami membandingkan hasil nilai akurasi yang didapatkan dengan menggunakan fitur yang diekstrak dari hasil korelasi *Spearman* dan MIC_e TIC_e.

Tabel 1. Fitur yang ada pada dataset

Nomor	Nama fitur	Penjelasan
1	*having_IP_Address*	Memiliki alamat IP
2	*URL_Length*	Panjang alamat URL
3	*Shortining_Service*	Menggunakan penyingkat URL
4	*having_At_Symbol*	Memiliki simbol @
5	*double_slash_redirecting*	Memiliki dua tanda /
6	*Prefix_Suffix*	Domain memiliki awalan, akhiran
7	*having_Sub_Domain*	Mempunyai subdomain
8	*SSLfinal_State*	Memiliki SSL
9	*Domain_registeration_length*	Panjang alamat domain
10	*Favicon*	Memiliki favicon atau logo tab
11	*port*	Memiliki port di domain
12	*HTTPS_token*	Memiliki token aman HTTPS

13	*Request_URL*	Ada request url
14	*URL_of_Anchor*	url yang memiliki link ke lain
15	*Links_in_tags*	Memiliki link pada tag
16	*SFH*	Memiliki string kosong atau about blank
17	*Submitting_to_email*	Mengharuskan untuk kirim email
18	*Abnormal_URL*	*URL* yang tidak normal
19	*Redirect*	Mengarahkan ke link lain
20	*on_mouseover*	Ada fitur mouse over
21	*RightClick*	Bisa klik kanan
22	*popUpWindow*	Ada pop up windows
23	*Iframe*	Memiliki iframe yang ditempelkan
24	*Age_of_domain*	Umur dari domain
25	*DNSRecord*	Memiliki record DNS
26	*Web_traffic*	Lalulintas web yang padat
27	*Page_Rank*	Halaman yang memiliki rangking
28	*Google_Index*	Terindeks dimesin pencari google
29	*Links_pointing_to_page*	Link yang mengarah ke halaman web
30	*Statistical_report*	Laporan statisik seperti pengunjung, akses.

Setelah menjalankan uji korelasi *Spearman*, maka fitur yang tersisa dan akan diuji nilai akurasinya dengan algoritma Random Forest adalah sebanyak 20 fitur yaitu *having_IP_Address, URL_Length, having_At_Symbol, Prefix_Suffix, having_Sub_Domain, Request_URL Domain_registeration_length , URL_of_Anchor, Links_in_tags, SFH, Abnormal_URL, Redirect, Iframe, Age_of_domain, DNSRecord, Web_traffic, Page_Rank, Google_Index, Links_pointing_to_page, Statistical_report*. Fitur-fitur tersebut didapatkan dari data yang ditunjukkan pada tabel 2, ada 10 fitur yang memiliki nilai korelasi lebih besar dari 0.5 atau moderat sehingga harus dikeluarkan. Kemudian fitur yang tersisa setelah menjalankan uji MIC_e dan TIC_e adalah sebanyak 29 fitur yaitu *URL_Length, Shortining_Service, having_At_Symbol, double_slash_redirecting, Prefix_Suffix, having_Sub_Domain, SSLfinal_State, HTTPS_token Domain_registeration_length, port, , Request_URL, URL_of_Anchor, Links_in_tags, SFH, Submitting_to_email, Abnormal_URL, Redirect, on_mouseover, RightClick, popUpWindow, Iframe, Age_of_domain, DNSRecord, Web_traffic, Page_Rank, Google_Index, Links_pointing_to_page, Statistical_report.* Fitur-fitur tersebut didapatkand ari data yang ditunjukkan pada tabel 3, hanya ada 1 fitur yang memilikai nilai lebih besar dari 0.5 atau moderat sehingga harus dikeluarkan.

Tabel 2. Hasil korelasi *Spearman*

Variabel 1	Variabel 2	SpearmanRho
double_slash_redirecting	*Abnormal_URL*	0.723724
double_slash_redirecting	*HTTPS_token*	0.760799
Favicon	*Iframe*	0.627607
Favicon	*on_mouseover*	0.706179
Favicon	*popUpWidnow*	0.939633
Favicon	*port*	0.803834
Favicon	*Submitting_to_email*	0.668317

HTTPS_token	*Abnormal_URL*	0.716287
on_mouseover	*Iframe*	0.659478
on_mouseover	*popUpWidnow*	0.733629
popUpWidnow	*Iframe*	0.629406
port	*Iframe*	0.687044
port	*on_mouseover*	0.623298
port	*popUpWidnow*	0.748517
port	*Submitting_to_email*	0.799088
RightClick	*Iframe*	0.655863
Shortining_Service	*Abnormal_URL*	0.73929
Shortining_Service	*double_slash_redirecting*	0.842796
Shortining_Service	*HTTPS_token*	0.757838
SSLfinal_State	*URL_of_Anchor*	0.564536
Submitting_to_email	*Iframe*	0.57749
Submitting_to_email	*on_mouseover*	0.531656
Submitting_to_email	*popUpWidnow*	0.629462

Tabel 3. Hasil MIC_e dan TIC_e

Variabel 1	Variabel 2	MICe TICe
Favicon	*popUpWindow*	0.575613

Hasil dari tabel 2 dan tabel 3 didapatkan dengan menjalankan *Mictools* yang dikembangkan oleh [19]. *Mictools* dijalankan lewat *docker desktop* di sistem operasi *macOS High Sierra*. Berikut adalah langkah-langkahnya:

1. Menghubungkan direktori lokal *mictools* dikomputer dengan *docker mictools*, seperti pada gambar 3. Terlihat ada 30 fitur dan data yang ada sebanyak 11055.

```
jimmy — docker run --rm -t -i -v ~/Documents/TEACHING/S2 AMIKOM YOGYAKARTA/SEMESTER 2/THESIS PROPOSAL/mictools:/mi...
Last login: Thu Jul  2 21:20:38 on console
^[[A^[[A^[[A(base) Jimmys-MacBook-Pdocker run --rm -t -i -v /Users/jimmy/Documen
ts/TEACHING/S2\ AMIKOM\ YOGYAKARTA/SEMESTER\ 2/THESIS\ PROPOSAL/mictools:/mictoo
ls -w /mictools minepy/mictools /bin/bash
root@ea6db0d9968e:/mictools# cd examples/
root@ea6db0d9968e:/mictools/examples# ls
JM-dataset.txt  journal_result  paperdata.txt
datasaurus.txt  paper_results   thesis_results
root@ea6db0d9968e:/mictools/examples# X=datasaurus.txt
root@ea6db0d9968e:/mictools/examples# X=paperdata.txt
root@ea6db0d9968e:/mictools/examples# ODIR=paper_result
root@ea6db0d9968e:/mictools/examples# mictools null $X $ODIR/null_dist.txt
X: 30 variables
Samples per class:
* None: 11055
```

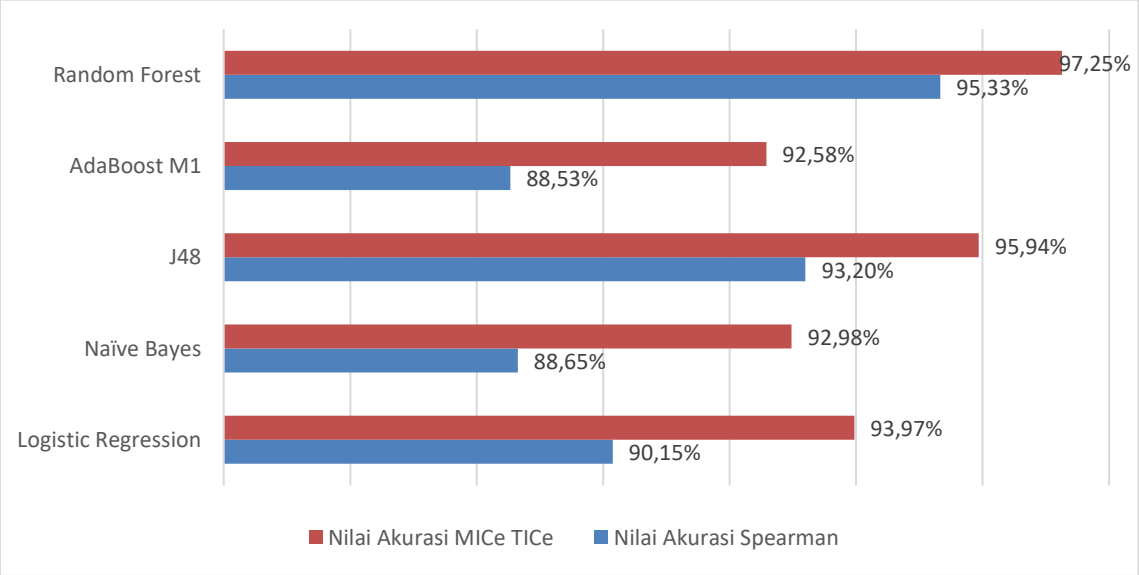
Gambar 3. Mencari nilai null_dist dari TIC_e

2. Hitung statistik TIC_e dan nilai-p empiris terkait untuk semua pasangan variabel dengan perintah *mictools pval $X $ODIR/null_dist.txt $ODIR*
3. Perbaiki nilai p untuk multiplisitas untuk mengontrol *False Discovery Rate (FDR)* dengan perintah *mictools adjust $ODIR/pval.txt $ODIR*
4. Langkah terakhir, kekuatan hubungan yang disebut signifikan dihitung menggunakan MIC_e dengan perintah *mictools strength $X $ODIR/pval_adj.txt $ODIR/strength.txt*
5. Dalam file strength.txt yang dihasilkan, berisi hasil korelasi antar variabel yaitu *Pearson*, *Sperman* dan MIC_e TIC_e

Langkah selanjutnya yang dilakukan adalah melihat nilai akurasi dari masing-masing hasil pemilihan fitur berdasarkan lima *algoritma machine learning*. Nilai akurasi didapatkan menggunakan aplikasi *Weka*. Uji dilakukan dengan *cross validation 10 folds*. Hasil perbandingan nilai akurasi dapat dilihat pada tabel 4.

Tabel 4. Tabel perbandingan nilai akurasi

Algoritma *Machine Learning*	Nilai Akurasi *Spearman*	Nilai Akuasi MIC_e TIC_e
Logistic Regression	90.15%	93.97%
Naïve Bayes	88.65%	92.98%
J48	93.20	95.94$
AdaBoost M1	88.53%	92.58%
Ranom Forest	95.33%	97.25%



Gambar 4. Grafik perbandingan nilai akurasi

Hasil pada Tabel 4 dan visualisasi dalam gambar 4 menunjukkan bahwa untuk kelima algoritma *machine learning*, seleksi fitur dengan menggunakan metode MIC_e TIC_e memiliki nilai akurasi yang paling tinggi. Perbandingannya untuk Random forest, MIC_e TIC_e lebih tinggi 1.92% dari korelasi *Spearman*. Adaboost M1, MIC_e TIC_e lebih tinggi 4.05% dari korelasi *Spearman*. J48, MIC_e TIC_e lebih tinggi 2.74% dari korelasi *Spearman. Naïve Bayes*, MIC_e TIC_e lebih tinggi 4.33% dari korelasi *Spearman.* Logistic regression, MIC_e TIC_e lebih tinggi 3.82% dari korelasi *Spearman*. Nilai akurasi yang paling tinggi dari hasil pengujian adalah kombinasi nilai metode MIC_e TIC_e dengan menggunakan algortima *Random Forest* yaitu sebesar 97.25%.

4. KESIMPULAN

Berdasarkan hasil yang telah didapatkan dari perbandingan nilai akurasi, seleksi fitur yang dihasilkan oleh metode MIC_e TIC_e mengungguli hasil akurasi dari metode korelasi *Spearman*. Hasil nilai akurasi yang didapatkan tersebut sama untuk semua *machine learning* yang diuji yaitu MIC_e TIC_e mengungguli metode korelasi *Spearman*. Dari segi algoritma *machine learning* yang digunakan, *Random Forest* mengungguli 4 algoritma lainnya seperti hasil yang ditunjukkan pada tabel 4 dan gambar 4.

5. SARAN

Untuk peneliti lainnya yang tertarik dengan topik yang kami teliti dapat mencoba untuk membandingkan menggunakan lebih dari satu dataset dengan algoritma *machine learning* yang tidak digunakan dalam penelitian ini. Kemudian bisa melihat hasil apakah hasilnya sama dengan dataset yang kami gunakan.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Tuhan Yang Maha Esa atas berkat dan rahmatNya dalam menyelesaikan peneletian ini. Terima kasih juga kepada Universitas Amikom Yogyakarta, Universitas Klabat Airmadidi, STMIK Dipanegara Makassar atas dukungan yang diberikan sehingga penelitian ini bisa diselesaikan dengan baik dan tepat waktu.

DAFTAR PUSTAKA

[1] P. Yang, G. Zhao, and P. Zeng, "Phishing website detection based on multidimensional features driven by deep learning," *IEEE Access*, vol. 7, pp. 15196–15209, 2019, doi: 10.1109/ACCESS.2019.2892066.

[2] "Phishing Activity Trends Report Q1 2020," 2020. https://docs.apwg.org/reports/apwg_trends_report_q1_2020.pdf.

[3] H. Y. A. Abutair and A. Belghith, "Using Case-Based Reasoning for Phishing Detection," *Procedia Comput. Sci.*, vol. 109, pp. 281–288, 2017, doi: 10.1016/j.procs.2017.05.352.

[4] C. L. Tan, K. L. Chiew, K. S. Wong, and S. N. Sze, "PhishWHO: Phishing webpage detection via identity keywords extraction and target domain name finder," *Decis. Support Syst.*, vol. 88, pp. 18–27, 2016, doi: 10.1016/j.dss.2016.05.005.

[5] G. Varshney, M. Misra, and P. Atrey, "Browshing a new way of phishing using a malicious browser extension," 2017, doi: 10.1109/IPACT.2017.8245147.

[6] M. Babagoli, M. P. Aghababa, and V. Solouk, "Heuristic nonlinear regression strategy for detecting phishing websites," *Soft Comput.*, vol. 23, pp. 4315–4327, 2019, doi: 10.1007/s00500-018-3084-2.

[7] I. Tyagi, J. Shad, S. Sharma, S. Gaur, and G. Kaur, "A Novel Machine Learning Approach to Detect Phishing Websites," *2018 5th Int. Conf. Signal Process. Integr. Networks, SPIN 2018*, pp. 425–430, 2018, doi: 10.1109/SPIN.2018.8474040.

[8] J. Moedjahedy, H. Zein, I. B, E. Tongalu, K. Kusrini, and M. S. Mustafa, "Analisis Seleksi Tingkat Kecocokan Gambar pada MDID Multimedia Database Dengan Menggunakan Metode ImageDNA," *CogITo Smart J.*, 2020, doi: 10.31154/cogito.v6i1.223.50-59.

[9] K. L. Chiew, C. L. Tan, K. S. Wong, K. S. C. Yong, and W. K. Tiong, "A new hybrid ensemble feature selection framework for machine learning-based phishing detection system," *Inf. Sci. (Ny).*, vol. 484, pp. 153–166, 2019, doi: 10.1016/j.ins.2019.01.064.

[10] I. Salihovicd, H. Serdavic, and J. Kevric, "The Role of Feature Selection in Machine Learning for Detection of Spam and Phishing Attacks," in *International Symposium on Innovative and Interdisciplinary Applications of Advanced Technologies (IAT)*, 2019, vol. 3, pp. 476–483, doi: 10.1007/978-3-030-02577-9.

[11] R. S. Rao and A. R. Pais, "Detection of phishing websites using an efficient feature-based machine learning framework," *Neural Comput. Appl.*, vol. 31, no. 8, pp. 3851–3873, 2019, doi: 10.1007/s00521-017-3305-0.

[12] J. Mao *et al.*, "Detecting Phishing Websites via Aggregation Analysis of Page Layouts," *Procedia Comput. Sci.*, vol. 129, pp. 224–230, 2018, doi: 10.1016/j.procs.2018.03.053.

[13] D. Dua and C. Graff, *UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]*. 2019.

[14] F. J. Gravetter and L. B. Wallnau, *Essentials of Statistics for the Behavioral Science*, Sixth. Belmon: Thomson Learning, 2008.

[15] U. Nehmzow, *Scientific methods in mobile robotics: Quantitative analysis of agent behaviour*. Springer Verlag, 2006.

[16] W. Lowie and B. Seton, *Essential Statistics for Applied Linguistics*. Red Globe Press, 2013.

[17] D. N. Reshef *et al.*, "Detecting novel associations in large data sets," *Science (80-.).*, vol. 334, no. 6062, pp. 1518–1524, 2011, doi: 10.1126/science.1205438.

[18] Y. A. Reshef, D. N. Reshef, H. K. Finucane, P. C. Sabeti, and M. Mitzenmacher, "Measuring dependence powerfully and equitably," *J. Mach. Learn. Res.*, vol. 17, pp. 1–63, 2016.

[19] D. Albanese, S. Riccadonna, C. Donati, and P. Franceschi, "A practical tool for maximal information coefficient analysis," *Gigascience*, vol. 7, no. 4, pp. 1–8, 2018, doi: 10.1093/gigascience/giy032.