

Ceng Giap Yo

Data Mining Tomy.pdf



Subclass 5



31 Agustus - 7 September 2026



Universitas Dian Nuswantoro

Document Details

Submission ID

trn:oid::1:3639481411

Submission Date

Sep 3, 2026, 7:29 PM GMT+7

Download Date

Sep 3, 2026, 8:12 PM GMT+7

File Name

Data_Mining_Tomy.pdf

File Size

201.6 KB

5 Pages**2,332 Words****13,446 Characters**

29% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography

Exclusions

- 27 Excluded Sources

Top Sources

- 26%  Internet sources
 - 20%  Publications
 - 14%  Submitted works (Student Papers)
-

Top Sources

- 26% Internet sources
- 20% Publications
- 14% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	jurnal.itpln.ac.id	4%
2	Internet	jti.respati.ac.id	3%
3	Internet	www.researchgate.net	3%
4	Student papers	Universitas Putera Indonesia YPTK Padang	2%
5	Publication	Siti Puspita Hida Sakti MZ, Marzuki, Asslia Johar Latipah. "PENERAPAN TEKNOLO...	2%
6	Internet	stmikelrahma.e-journal.id	2%
7	Internet	digilib.mercubuana.ac.id	2%
8	Publication	Rara Pramita Putri, M. Ilayaraja, K. Shankar, Wahidah Hashim, Andino Maselena, ...	1%
9	Internet	sipeg.univpancasila.ac.id	1%
10	Internet	tunasbangsa.ac.id	1%
11	Student papers	KYUNG HEE UNIVERSITY	<1%

12	Internet	ojs.uniska-bjm.ac.id	<1%
13	Student papers	Universitas Mulawarman	<1%
14	Internet	etheses.uin-malang.ac.id	<1%
15	Publication	Retno Maharesi. "Pembentukan Model Pohon Keputusan pada Database Car Eval...	<1%
16	Internet	stitmultazam.ac.id	<1%
17	Internet	repository.ub.ac.id	<1%
18	Internet	www.kominfo.go.id	<1%
19	Publication	Agusti Frananda Alfonsus Naibaho, Amalia Zahra. "PREDIKSI KELULUSAN SISWA S...	<1%
20	Publication	Zhafira Zafitri, Muhammad Ihsan Jambak. "Karakteristik Pembatalan Reservasi K...	<1%
21	Internet	journal.ikipsiliwangi.ac.id	<1%
22	Publication	Farid Fadli, Belsana Butar Butar. "Penerapan Decision Tree Menggunakan Algorit...	<1%
23	Internet	eprints.unisnu.ac.id	<1%
24	Internet	journal-s.org	<1%
25	Internet	repositorio.uss.edu.pe	<1%

26	Internet	repository.unsulbar.ac.id	<1%
27	Internet	www.scribd.com	<1%
28	Internet	pendidikanmu.com	<1%

DATA MINING IDENTIFIKASI WEBSITE PHISING MENGGUNAKAN ALGORITMA C4.5

Tomy Salim¹⁾ Yo Ceng Giap²⁾

Teknik Informatika Universitas Buddhi Dharma

Jl. Imam Bonjol No. 41 Karawaci Ilir Tangerang Banten

Email : lin_guo_qiang@yahoo.co.id

Teknik Informatika Universitas Buddhi Dharma

Jl. Imam Bonjol No. 41 Karawaci Ilir Tangerang Banten

Email : cenggiap@buddhidharma.ac.id

ABSTRACT

The background of this research is to help internet users around the world to be more careful and avoid phishing websites while surfing in cyberspace. The faster development of information technology and number of big websites is increasing every day, the more likely an internet user to accidentally open a phishing website. With that reason, research on phishing websites that are very harmful to internet users is seemed necessary. To solve this problem, the author uses a data mining model to search for patterns that contains information on a large number of sample website data. Data mining method used in this research is Decision Tree because the result is suitable and satisfying. In this study, the author used sample data from a website named uci dataset, which on that website page there are many data sets that can be used for researching and academic interests. From a large number of data rows, the author managed to find several factors that can be used as references or signs of phishing websites. Based on the evaluation result of data mining, this research has met the provisions of data mining research requirements by the university and this study also shows results as the author expected.

Keywords: **Data Mining, Phishing, Website, Internet, Decision Tree, Research, Algorithm, C4.5**

I. PENDAHULUAN

Saat ini internet sudah menjadi bagian penting dalam kehidupan masyarakat terutama pada aktifitas sosial dan finansial. Sebagai contohnya media sosial yang digunakan sebagai sarana berkomunikasi, mencari teman dan juga bisnis online yang digunakan beberapa pihak terutama perusahaan untuk menawarkan perdagangan online melalui e-mail dan memberitahu kepada calon pelanggan tentang website mereka, namun saat ini ada pihak yang tidak bertanggung jawab melakukan tindakan yang merugikan banyak orang yang salah satunya adalah tindakan phishing.

Phishing adalah tindakan penipuan yang dilakukan untuk mencoba mendapatkan informasi penting dari user yang menggunakan internet dengan mengirim sejumlah e-mail palsu kepada para user^[5]. Pada penelitian ini, penulis melakukan salah satu pendekatan untuk mengurangi masalah phishing yaitu mengumpulkan data yang diperlukan untuk mengidentifikasi website phishing dan menganalisa data tersebut dengan menggunakan data mining.

Ada beberapa fungsionalitas dari data mining, antara lain analisis asosiasi antar data, klasifikasi data, klastering data dan lain-lain, dalam penelitian ini, fungsionalitas yang dipakai adalah klasifikasi data. Klasifikasi data adalah proses menemukan model atau fungsi yang menjelaskan dan membedakan kelas data dan konsepnya^[1]. Dari

klasifikasi data memiliki beberapa model yang salah satunya adalah decision tree. Decision tree adalah model yang mirip flowchart berbentuk seperti pohon, dari decision tree ada beberapa algoritma yang dapat dipakai seperti ID3, C4.5 dan CART. Algoritma yang dipakai di dalam penelitian ini adalah C4.5. Dengan ini diharapkan dapat membantu dalam mendeteksi website phishing dengan akurat dan mudah.

II. TINJAUAN PUSTAKA

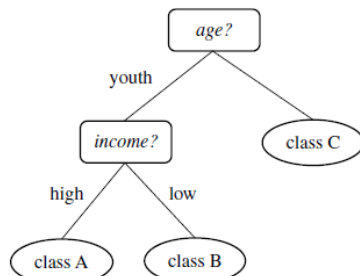
2.1. Data Mining

Data Mining adalah proses penemuan keteraturan, pola, dan hubungan dalam set data berukuran besar. Data yang dapat dianalisa dengan data mining bisa dari database, data warehouse, web, repositori informasi atau data yang dapat dialirkan ke dalam sistem secara dinamis^[1].

2.2. Decision Tree

Decision tree adalah sebuah struktur pohon, dimana setiap node internal (non-leaf) merepresentasikan pengujian atribut, setiap cabang merupakan suatu pembagian hasil uji, dan node daun (leaf) merepresentasikan kelompok kelas tertentu. Tingkat node teratas dari sebuah decision tree adalah node akar (root) yang biasanya berupa atribut yang paling berpengaruh pada suatu kelas tertentu.

Pada umumnya *decision tree* melakukan strategi pencarian secara *top-down* untuk solusinya. Pada proses mengklasifikasi data yang tidak diketahui, nilai atribut akan diuji dengan cara melacak jalur dari *node* akar (*root*) sampai *node* akhir (*leaf*) dan kemudian akan diprediksi kelas yang dimiliki oleh suatu data baru tertentu. *decision tree* dapat dengan mudah dirubah menjadi aturan klasifikasi^[1].



Gambar 1. Contoh *Decision Tree* ^[1]

2.3. Algoritma C4.5

C4.5 adalah sebuah algoritma yang digunakan untuk memproduksi sebuah *decision tree* yang merupakan ekspansi dari pendahulunya yaitu kalkulasi ID3. Algoritma ini meningkatkan algoritma ID3 dengan mengatur properti yang berkelanjutan dan berlainan, *missing value* dan pemotongan tree setelah konstruksi. *Decision tree* yang dibuat dengan C4.5 dapat digunakan untuk pengelompokan dan seringkali cenderung mengarah ke *statistical classifier*.

C4.5 membuat *decision tree* dari sebuah set data sampel sama seperti algoritma ID3. Sebagai algoritma pembelajaran yang terkelola, algoritma ini membutuhkan sebuah set dari data sampel yang dapat terlihat seperti data pasangan : objek masukan dan nilai keluaran yang diinginkan (*class*).

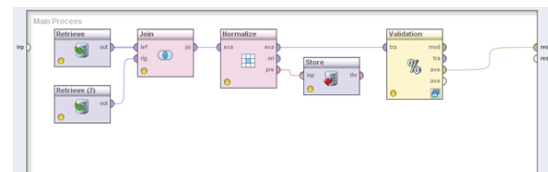
Algoritma ini menganalisa set data sampel dan membangun sebuah *classifier* yang harus mempunyai sebuah kapasitas untuk mengatur kasus latihan dan kasus percobaan dengan akurat. Sebuah kasus percobaan itu seperti sebuah objek masukan dan algoritma harus memprediksi sebuah nilai keluaran. Mempertimbangkan sampel dari training dataset $S=S_1, S_2, \dots, S_n$ yang dimana sudah diklasifikasikan. Masing-masing sampel memiliki vector $(x_{1,i}, x_{2,i}, \dots, x_{n,i})$, dimana x merepresentasikan atribut-atribut atau fitur pada sampel dan kelas dimana S_1 berada.

Pada tiap *node* pohon, C4.5 memilih satu atribut data yang paling efisien untuk membagi set tersebut menjadi subset-subset yang menghasilkan *output* pada satu kelas atau yang lain^[4].

2.4. Rapidminer 5

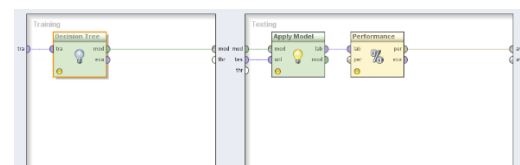
RapidMiner adalah perangkat lunak *open-source* untuk *data mining* yang digunakan secara meluas. Proyek ini dimulai di Universitas Dortmund pada tahun 2001 dan dikembangkan lebih lanjut oleh Rapid-I GmbH sejak 2007. Dengan latar belakang akademis ini, RapidMiner tidak hanya ditujukan kepada bisnis klien, namun juga kepada universitas-universitas dan para peneliti dari bidang studinya yang berbeda-beda.

RapidMiner memiliki antarmuka yang nyaman, dimana analisa dikonfigurasi dalam sebuah *process view*. Dalam konsep modular untuk *process view* ini, setiap langkah analisis digambarkan dengan sebuah operator dalam proses analisis. Operator-operator ini memiliki *port* untuk *input* dan *output* dimana operator tersebut dapat berkomunikasi dengan operator lain untuk mendapatkan *input data* atau mengirim data yang telah diubah dan menggenerasi model, sebagai contohnya dapat dilihat pada gambar 2



Gambar 2. Proses sederhana dengan contoh pengisian data, praproses dan produksi data^[2]

Situasi dan kebutuhan analisis yang sangat kompleks dapat ditangani oleh yang disebut sebagai *super-operator*, dimana dapat berisi sub proses yang lengkap^[2]. Sebagai contoh yang paling dikenal adalah *cross-validation* pada gambar 3



Gambar 3. Sub proses di dalam *cross-validation* ^[2]

III. Materi Penelitian

Penelitian ini didasari oleh *Predicting Phishing Websites based on Self-Structuring Neural Network*^[3]. Sumber data yang dipakai dalam penelitian ini adalah data sekunder yang didapat dari lampiran jurnal. Data ini berjumlah 11055 baris data dimana data ini dipakai 3000 baris data. Berikut tabel penjelasan tentang atribut-atribut yang terdapat pada data sekunder ini

Tabel 1. Atribut dan Label

Atribut dan Label	Deskripsi
having_IP_Address	Adanya <i>IP address</i> sebagai domain pada url (biner) -1 (tidak), 1 (iya)
URL_Length	Panjang <i>url</i> (polinomial) -1 (kurang dari 54 karakter), 0 (antara 54 sampai 75 karakter), 1 (lebih dari 75 karakter)
Shortening_Service	Penggunaan layanan penyingkatan <i>url</i> (biner) -1(tidak), 1 (iya)
having_At_Symbol	Penggunaan simbol "@" pada <i>url</i> (biner) -1 (tidak), 1 (iya)
double_slash_redirecting	Penggunaan simbol "/" pada <i>url</i> untuk mengalihkan <i>website</i> (biner) -1 (tidak), 1 (iya)
Prefix_Suffix	Penggunaan simbol "-" pada domain dalam <i>url</i> (biner) -1 (tidak), 1 (iya)
having_Sub_Domain	Penggunaan subdomain (polinomial) -1 (tidak punya), 0 (1 subdomain), 1 (lebih dari 1 subdomain)
SSLfinal_State	Memiliki sertifikat SSL dimana sertifikat yang dipercaya berasal dari penyedia ternama seperti "GeoTrust, GoDaddy, Network Solutions, Thawte, Comodo, Doster dan VeriSign" (polinomial) -1 (punya sertifikat yang dipercaya), 0 (punya sertifikat yang belum dipercaya), 1 (tidak memiliki sertifikat)
Domain_registration_length	Batas berlakunya domain (biner) -1 (lebih dari 1 tahun), 1 (kurang dari 1 tahun)
Favicon	Memiliki <i>favicon</i> dari <i>link</i> eksternal (biner) -1(tidak), 1 (iya)
port	Penggunaan <i>port</i> seperti 21,22,23,445 dan lainnya (biner) -1 (tidak), 1 (iya)
HTTPS_token	Penggunaan <i>https</i> ke dalam bagian domain pada url (biner) -1 (tidak), 1 (iya)
Request_URL	Persentase permintaan <i>url</i> eksternal dari keseluruhan (polinomial) -1 (kurang dari 22%), 0 (antara 22% sampai 61%), 1 (lebih dari 61%)
URL_of_Anchor	Persentase penggunaan tag <a> yang mengarah selain ke domain yang sama dari keseluruhan (polinomial) -1 (kurang dari 31%), 0 (antara 31% sampai 67%), 1 (lebih dari 67%)
Links_in_tags	Persentase penggunaan tag <link>, <meta>, dan <script> yang mengarah selain ke domain yang sama dari keseluruhan (polinomial) -1 (kurang dari 17%), 0 (antara 17% sampai 81%), 1 (lebih dari 81%)
SFH	Domain pemrosesan <i>Server Form Handler</i> (polinomial) -1 (pada domain yang sama), 0 (pada domain yang berbeda), 1 (kosong)
Submitting_to_email	Penggunaan fungsi "mail()" atau "mailto" dalam php untuk mengirim informasi user (biner) -1 (tidak), 1 (iya)
Abnormal_URL	Kecocokan <i>website</i> dengan catatannya yang ditunjukkan pada basis data WHOIS (biner) -1 (cocok), 1 (tidak cocok)
Redirect	Jumlah pengalihan <i>website</i> yang dilakukan (polinomial) -1 (kurang dari 2 kali), 0 (2,3, atau 4 kali), 1 (lebih dari 4 kali)
on_mouseover	Perubahan <i>status bar</i> ketika <i>event</i> onMouseOver aktif (biner) -1 (tidak), 1 (iya)
RightClick	Keadaan klik kanan pada <i>website</i> (biner) -1 (diaktifkan), 1 (dinonaktifkan)
popUpWidnow	Penggunaan popUpWindow untuk meminta <i>user</i> mengisi data mereka (biner) -1 (tidak), 1 (iya)
Iframe	Penggunaan fungsi <i>iframe</i> (biner) -1 (tidak), 1 (iya)
age_of_domain	Umur domain (biner) -1 (lebih dari atau sama dengan 6 bulan), 1 (kurang dari 6 bulan)
DNSRecord	Adanya catatan dns pada domain (biner) -1 (ada), 1 (tidak ada)
web_traffic	Rank lalu lintas <i>website</i> dalam basis data Alexa (polinomial) -1 (diatas 100,000), 0 (dibawah 100,000), 1 (tidak terdaftar)
Page_Rank	Nilai PageRank <i>website</i> (biner) -1 (lebih dari atau sama dengan 0.2) 1 (kurang dari 0.2)
Google_Index	Adanya <i>website</i> dalam indeks pencarian Google (biner) -1 (iya), 1 (tidak)
Links_pointing_to_page	Jumlah <i>link</i> eksternal yang menunjuk ke <i>website</i> (polinomial) -1 (lebih dari 2), 0 (1 atau 2), 1 (tidak ada)
Statistical_report	<i>Host</i> berasal dari <i>Top Phishing IPs</i> atau <i>Top Phishing Domains</i> yang dibuat oleh beberapa pihak seperti StopBadware dan PhishTank (biner) -1 (tidak), 1(iya)
Result (Label)	Hasil identifikasi <i>website</i> (biner) -1 (bukan phishing), 1 (phishing)

IV. Hasil Penelitian

Berdasarkan gambar *tree* yang dihasilkan dapat dilihat bahwa atribut SSLfinal_State dapat dikatakan sebagai atribut yang paling berpengaruh dikarenakan merupakan *node* yang paling pertama dalam *tree* yang dibuat.

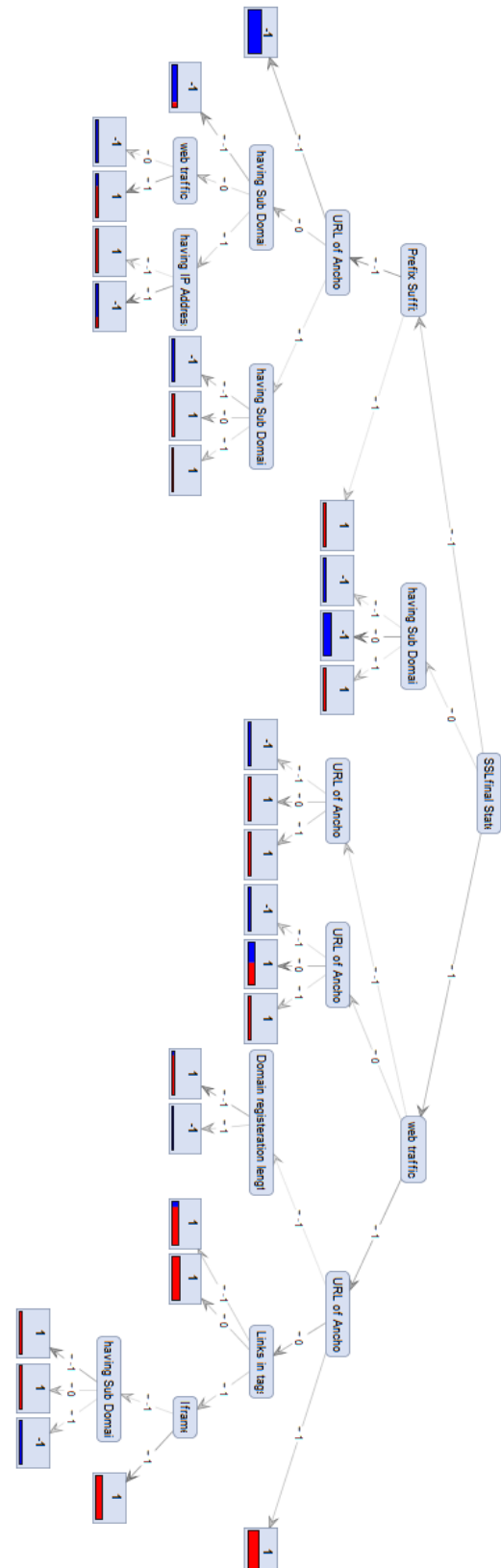
Di sini dapat diketahui bahwa dari *tree* ini, atribut data pertama yang diperiksa adalah nilai SSLfinal_State, jika bernilai -1 maka diperiksa Prefix_Suffix, jika bernilai -1, maka akan diperiksa lagi URL_of_Anchor, jika bernilai -1 maka *tree* akan memprediksi bahwa nilai *result* dari data yang diperiksa bernilai -1, jika URL_of_Anchor bernilai 0 maka diperiksa having_Sub_Domain, jika bernilai -1 maka *result* akan bernilai -1, jika having_Sub_Domain bernilai 0, maka diperiksa web_traffic, jika bernilai 0, maka *result* akan bernilai -1, jika web_traffic bernilai 1, maka *result* akan bernilai 1.

Kembali ke *node* having_Sub_Domain, jika bernilai 1, maka diperiksa having_IP_Address, jika bernilai -1, maka *result* akan bernilai 1, sebaliknya jika having_IP_Address bernilai 1, maka *result* akan bernilai -1. Kembali ke *node* URL_of_Anchor, jika bernilai 1 maka akan diperiksa having_Sub_Domain, jika bernilai -1 maka *result* akan bernilai -1, jika tidak maka *result* akan bernilai 1.

Kembali ke *node* Prefix_Suffix, jika bernilai 1 maka *result* akan bernilai 1. Kembali lagi ke *node* SSLfinal_State, jika bernilai 0 maka diperiksa having_Sub_Domain, jika bernilai -1 atau 0, maka *result* akan bernilai -1, jika tidak maka *result* bernilai 1. Kembali ke *node* SSLfinal_State, jika bernilai 1, maka diperiksa web_traffic, jika bernilai -1 atau 0 maka diperiksa URL_of_Anchor, jika bernilai -1 maka *result* akan bernilai -1, jika tidak maka *result* akan bernilai 1.

Kembali ke *node* web_traffic, jika bernilai 1 maka diperiksa URL_of_Anchor, jika bernilai -1 maka akan diperiksa atribut Domain_registration_length, jika bernilai -1 maka *result* akan bernilai 1, sebaliknya jika bernilai 1 maka *result* akan bernilai -1. Kembali ke *node* URL_of_Anchor, jika bernilai bernilai 0 maka diperiksa Link_in_tags, jika bernilai -1 atau 0 maka *result* bernilai 1, jika tidak maka diperiksa IFrame, jika bernilai 1 maka *result* akan bernilai 1, jika tidak akan diperiksa having_Sub_Domain, jika

bernilai -1 dan 0 maka *result* akan bernilai 1, jika tidak maka *result* akan bernilai -1. Kembali ke *node* URL_of_Anchor, jika bernilai 1 maka akan *result* akan bernilai 1.



Gambar 4. Hasil Decision Tree

V. Kesimpulan

Dapat decision tree diatas dapat ditarik kesimpulan bahwa:

Atribut SSL_final_State merupakan atribut yang paling berpengaruh. Hal ini dibuktikan Atribut SSLfinal_State berada pada *node* akar dalam decision tree yang dibuat.

Dari 30 atribut data yang diolah ke dalam *decision tree*, yang berpengaruh dalam logika *decision tree* adalah 9 atribut yaitu SSLfinal_State, Prefix_Suffix, URL_of_Anchor, having_Sub_Domain, web_traffic, having_IP_Address, Domain_registration_length, Links_in_tags, dan iframe

Referensi

- [1] Han, Jiawei., Kamber, Micheline dan Pei, Jian. (2012). *Data Mining Concepts and Techniques Third Edition*. Elsevier Inc; Amsterdam.
- [2] Land, Sebastian dan Fischer, Simon. (2012). *Rapid Miner 5*. Rapid-I GmbH; Dortmund.
- [3] Mohammad, Rami M., Thabtah, Fadi dan McCluskey, Lee. (2014). *Predicting Phishing Websites based on Self-Structuring Neural Network*. University of Huddersfield; Huddersfield. ISSN 0941-0643.
- [4] Nikam, Sagar S. (2015). *A Comparative Study of Classification Techniques in Data Mining Algorithms*. Techno Research Publishers; Bhopal. ISSN 0974-6471.
- [5] Parthasarathy, G., Tomar, D. C. dan Praisys, K. Christina. (2016). *An Enhancement Of Association Classification Algorithm For Identifying Phishing Websites*. Indian Journal of Computer Science and Engineering; Chennai. ISSN : 0976-5166