



MODEL KLASIFIKASI
SITUS WEB BERBAHAYA
MENGGUNAKAN
ALGORITME POHON
KEPUTUSAN C5.0
UNTUK MENDUKUNG
PROGRAM INTERNET CAKAP

Alvin Reinaldo

Restu Triadi

Alif Hilmi Akbar

Muhammad Farid Marzuq

Ahmad Maulvi Alfansuri

1.

PENDAHULUAN

Latar Belakang
Tujuan Penelitian
Manfaat Penelitian

Pada tahun 2016, Indonesia
menempati peringkat

10

negara dengan pengguna
internet terbanyak di dunia.

(ITU 2018)



Pada tahun 2017:

143,26 M orang

Pengguna internet Indonesia

54,86%

Tingkat penetrasi internet

(APJII 2017)

34,02%

Pengguna internet Indonesia tidak sadar datanya dapat diambil melalui internet

16,02%

Pengguna internet Indonesia tidak sadar penipuan dapat terjadi melalui internet

(APJII 2017)

Tingkat cyber crime Indonesia
di posisi



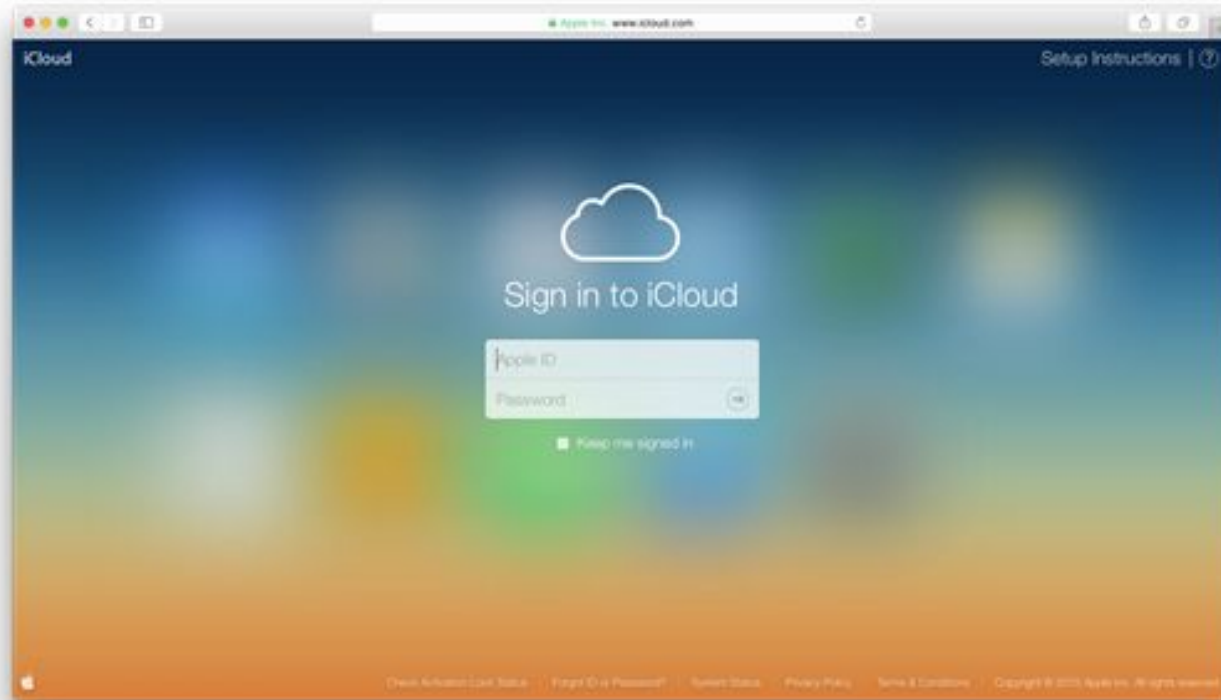
tertinggi di dunia

(Rizki 2018)

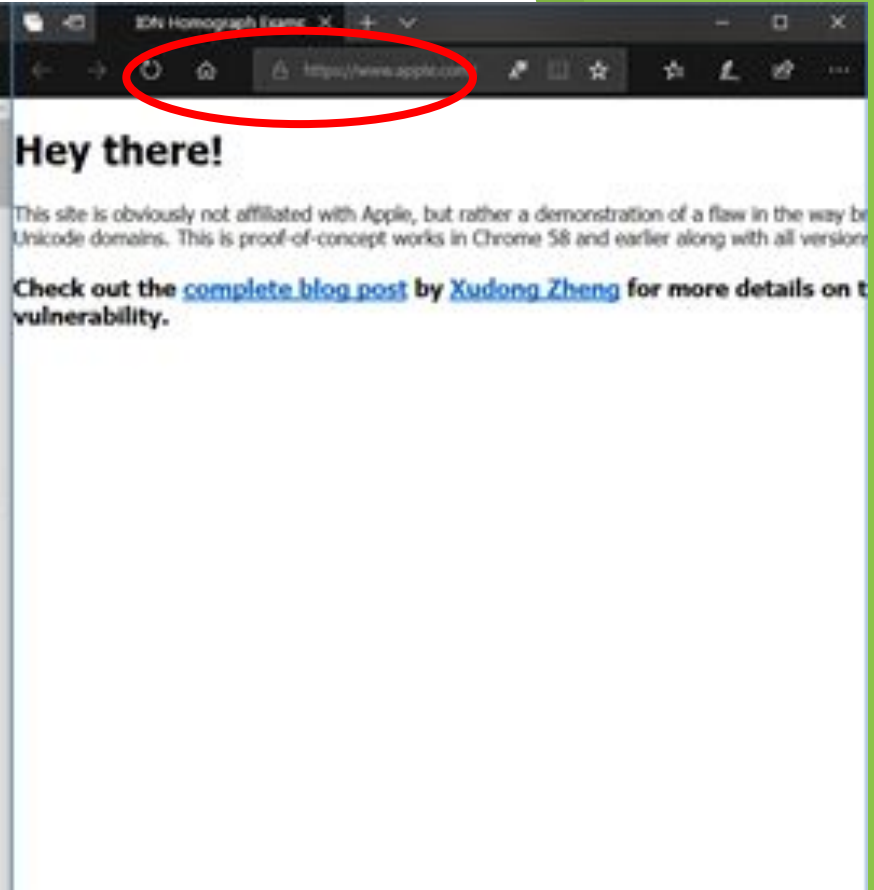
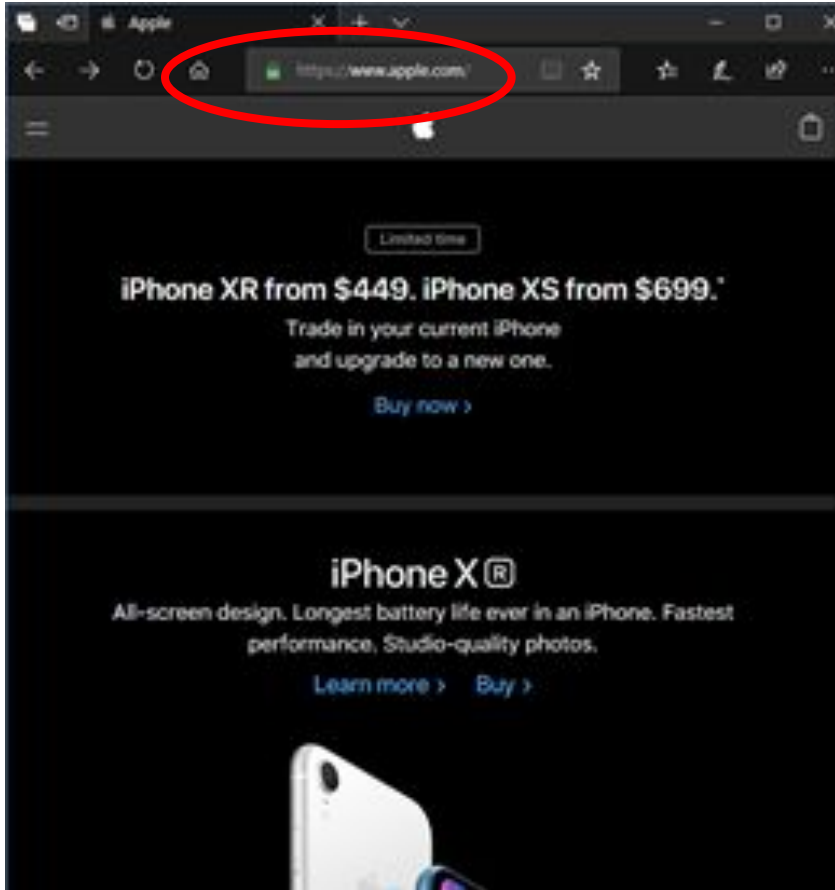
Real or Fake?



Real or Fake?



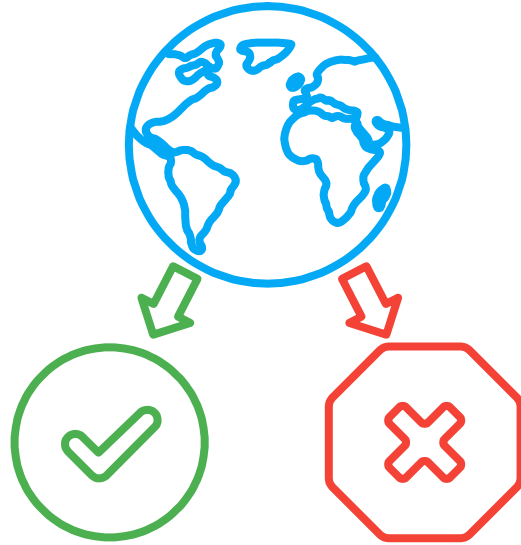
Real or Fake?





~~KEAMANAN INTERNET~~

TUJUAN KAMI



Membuat model klasifikasi situs web berbahaya dengan menerapkan algoritme pohon keputusan C5.0

Hasil dari penelitian ini diharapkan dapat digunakan sebagai acuan dalam mengklasifikasikan situs web berbahaya, sehingga dapat mendukung program Internet Cakap dengan membuat internet lebih aman bagi penggunaanya.



2.

METODE DAN HASIL PENELITIAN

Lingkungan Pengembangan
Dataset

Tahapan Penelitian

Eksplorasi Data -

Praproses Data -

Pembagian Data -

Pemodelan Klasifikasi Pohon Keputusan -

Pengujian Model Klasifikasi -



LINGKUNGAN PENGEMBANGAN

- ▶ Perangkat Keras
 - ▷ Intel® Core™ i5-7200U CPU @2.50GHz
 - ▷ 8192 MB RAM
 - ▷ NVIDIA GeForce 940MX
- ▶ Perangkat Lunak



Eksplorasi



+



Encoding



+



Processing

DATASET

- ▶ MALICIOUS AND BENIGN WEBSITES: CLASSIFY BY APPLICATION AND NETWORK FEATURES
 - ▷ Versi ketiga, diunggah 9 April 2018 di Kaggle
 - ▷ Format .csv
 - ▷ 1781 *instances*
 - ▷ 20 fitur + 1 fitur kelas
 - ▷ Data penelitian yang dilakukan Xu et al. (2013) yang telah melalui proses reduksi fitur pada penelitian yang dilakukan oleh Urcuqui et al. (2017)

Tabel 2 Jumlah data untuk masing-masing kelas

Kelas	Jumlah Data
0	1565
1	216

FITUR

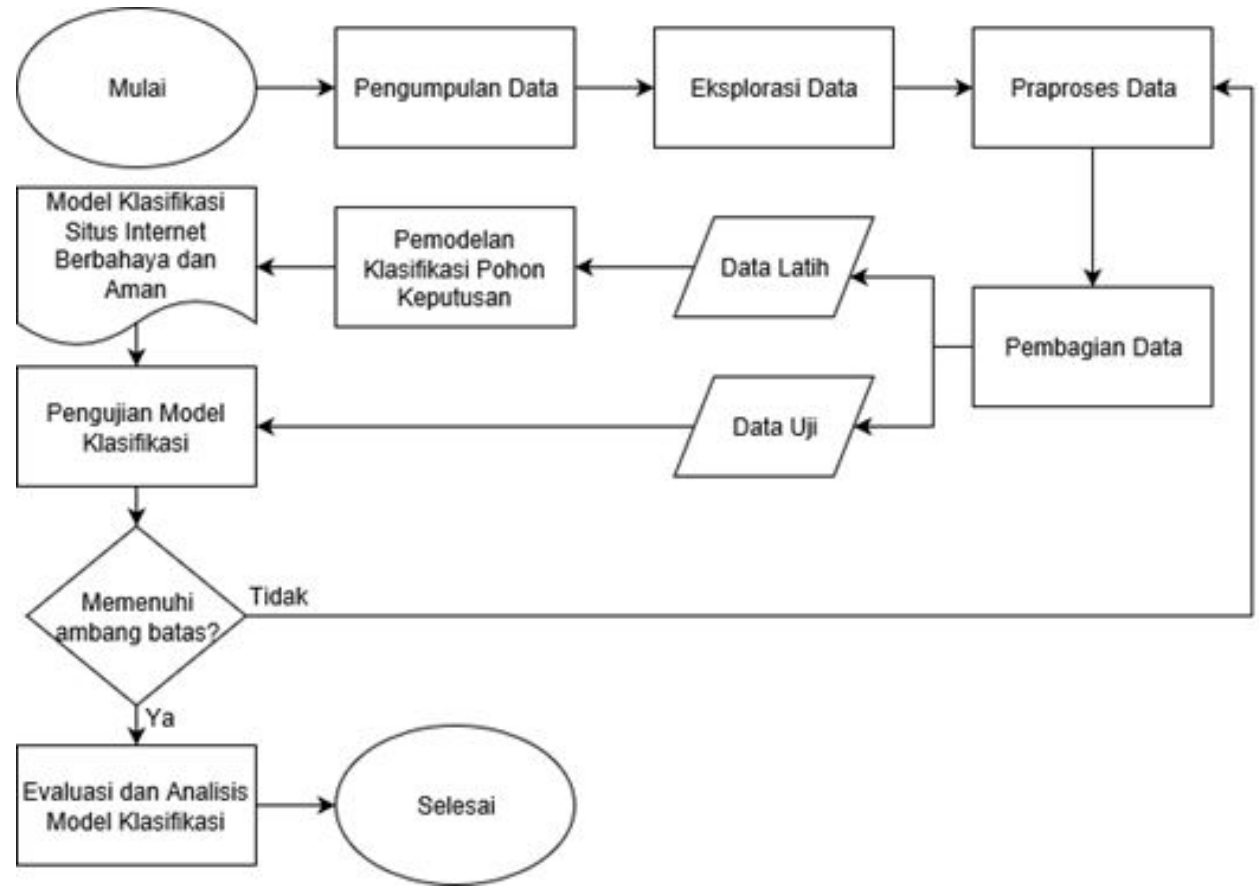
No	Nama Fitur	Keterangan
1	URL	Identifikasi anonim dari alamat URL yang dianalisis dalam penelitian ini
2	URL_LENGTH	Jumlah karakter pada alamat URL
3	NUMBER_SPECIAL_CHARACTERS	Jumlah <i>special character</i> pada alamat URL, misal "?", "-", "_", "=", dan "%"
4	CHARSET	Standar pengkodean karakter (<i>character set</i>) yang digunakan
5	SERVER	Sistem operasi <i>server</i> yang diterima dari respons paket
6	CONTENT_LENGTH	Ukuran konten dari <i>header</i> HTTP

No	Nama Fitur	Keterangan
7	WHOIS_COUNTRY	Nama negara yang didapat dari respons <i>server</i>
8	WHOIS_STATEPRO	Nama negara bagian atau provinsi yang didapat dari respons <i>server</i>
9	WHOIS_REGDATE	Tanggal registrasi <i>server</i> yang dianalisis, berformat DD/MM/YYYY HH:MM
10	WHOIS_UPDATED_DATE	Tanggal pembaharuan terakhir dari <i>server</i> yang dianalisis, berformat DD/MM/YYYY HH:MM
11	TCP_CONVERSATION_EXCHANGE	Jumlah total paket TCP yang dikirim ke <i>server</i> oleh <i>crawler</i>
12	DIST_REMOTE_TCP_PORT	Jumlah total <i>port</i> TCP yang berbeda yang digunakan <i>server</i> ketika berkomunikasi dengan <i>crawler</i>
13	REMOTE_IPS	Jumlah total alamat IP yang berbeda yang terhubung oleh <i>crawler</i> tidak termasuk alamat IP dari <i>server</i> DNS

No	Nama Fitur	Keterangan
14	APP_BYTES	Jumlah ukuran (dalam Byte) dari data pada <i>application layer</i> yang dikirim oleh <i>crawler</i> ke <i>server</i> , tidak termasuk data yang dikirim ke <i>server</i> DNS
15	SOURCE_APP_PACKET	Jumlah paket yang dikirim oleh <i>crawler</i> ke <i>server</i>
16	REMOTE_APP_PACKET	Jumlah paket yang dikirim oleh <i>server</i> ke <i>crawler</i>
17	SOURCE_APP_BYTES	Jumlah ukuran (dalam Byte) dalam komunikasi dari <i>crawler</i> ke <i>server</i>
18	REMOTE_APP_BYTES	Jumlah ukuran (dalam Byte) dalam komunikasi dari <i>server</i> ke <i>crawler</i>
19	APP_PACKET	Jumlah total paket IP yang dihasilkan untuk mendapatkan konten yang sesuai dengan <i>input</i> URL, termasuk pengalihan dan permintaan DNS
20	DNS_QUERY_TIMES	Jumlah permintaan DNS yang dikirim oleh <i>crawler</i>
21	TYPE	Representasi tipe situs web yang dianalisis, 1 berarti situs web berbahaya dan 0 berarti situs web aman



TAHAPAN PENELITIAN



PRAPROSES DATA

Encoding: Ordinal Encoder

Ordinal encoder mengonversi setiap label data kategorik menjadi nilai integer dari 1 sampai k sesuai dengan banyaknya label data pada atribut tersebut (Hale J 2018).

► Why?

- Banyak *instance* yang memiliki jumlah karakter terlalu panjang
- Dataset banyak mengandung karakter yang tidak dapat diolah oleh *package* C50, seperti titik dua (:) dan titik koma (;)

Selain itu

Atribut `URL` merupakan atribut identifikasi unik dari data sehingga atribut ini dihilangkan.

HASIL ENCODING


URL	URL_LENGTH	NUMBER	CHARSET	SERVER	CONTENT	WHOIS_COUNTRY	WHOIS_STATE	WHOIS_REGDATE	WHOIS_UPDATED_DATE	TCP_CONNECTION
M0_109	16	7	iso-8859-1	nginx	263	None	None	10/10/2015 18:21	None	7
B0_2314	16	6	UTF-8	Apache/2	15087	None	None	None	None	17
B0_911	16	6	us-ascii	Microsoft	324	None	None	None	None	0
B0_113	17	6	ISO-8859-1	nginx	162	US	AK	7/10/1997 4:00	12/9/2013 0:45	31
B0_403	17	6	UTF-8	None	124140	US	TX	12/5/1996 0:00	11/4/2017 0:00	57
B0_2064	18	7	UTF-8	nginx	NA	SC	Mahe	3/8/2016 14:30	3/10/2016 3:45	11
B0_462	18	6	iso-8859-1	Apache/2	345	US	CO	29/07/2002 0:00	1/7/2016 0:00	12
B0_1128	19	6	us-ascii	Microsoft	324	US	FL	18/03/1997 0:00	19/03/2017 0:00	0
M2_17	20	5	utf-8	nginx/1.1	NA	None	None	8/11/2014 7:41	None	0
M3_75	20	5	utf-8	nginx/1.1	NA	None	None	8/11/2014 7:41	None	0



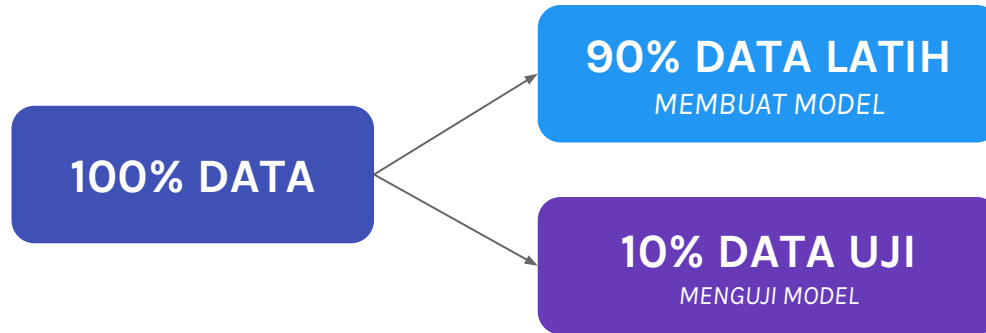
Encode

URL_LENGTH	NUMBER	CHARSET	SERVER	CONTENT	WHOIS_COUNTRY	WHOIS_STATE	WHOIS_REGDATE	WHOIS_UPDATED_DATE	TCP_CONNECTION
16	7	4	200	263	29	98	59	593	7
16	6	3	61	15087	29	98	889	593	17
16	6	5	115	324	29	98	889	593	0
17	6	1	200	162	42	4	806	68	31
17	6	3	124	124140	42	137	93	42	57
18	7	3	200		34	70	644	442	11
18	6	4	17	345	42	24	607	10	12
19	6	5	115	324	42	35	258	202	0
20	5	6	210		29	98	845	593	0
20	5	6	210		29	98	845	593	0

Encoded attributes: CHARSET, SERVER, WHOIS_COUNTRY, WHOIS_STATEPRO, WHOIS_REGDATE, WHOIS_UPDATED_DATE

Implemented with pandas library on 

PEMBAGIAN DATA



With ***K-Fold Cross Validation*** and **$k = 10$** (Refaeilzadeh et al. 2009)

Pembagian data diulang sebanyak sepuluh kali sampai semua *fold* pernah menjadi data uji

DISTRIBUSI DATA LATIH DAN DATA UJI

<i>Fold</i>	Distribusi Kelas Data Latih			Distribusi Kelas Data Uji		
	0	1	Jumlah	0	1	Jumlah
1	1408	194	1602	157	22	179
2	1409	194	1603	156	22	178
3	1408	195	1603	157	21	178
4	1409	194	1603	156	22	178
5	1408	195	1603	157	21	178
6	1409	194	1603	156	22	178
7	1409	195	1604	156	21	177
8	1408	194	1602	157	22	179
9	1409	195	1604	156	21	177
10	1408	194	1602	157	22	179

PEMODELAN KLASIFIKASI

POHON KEPUTUSAN

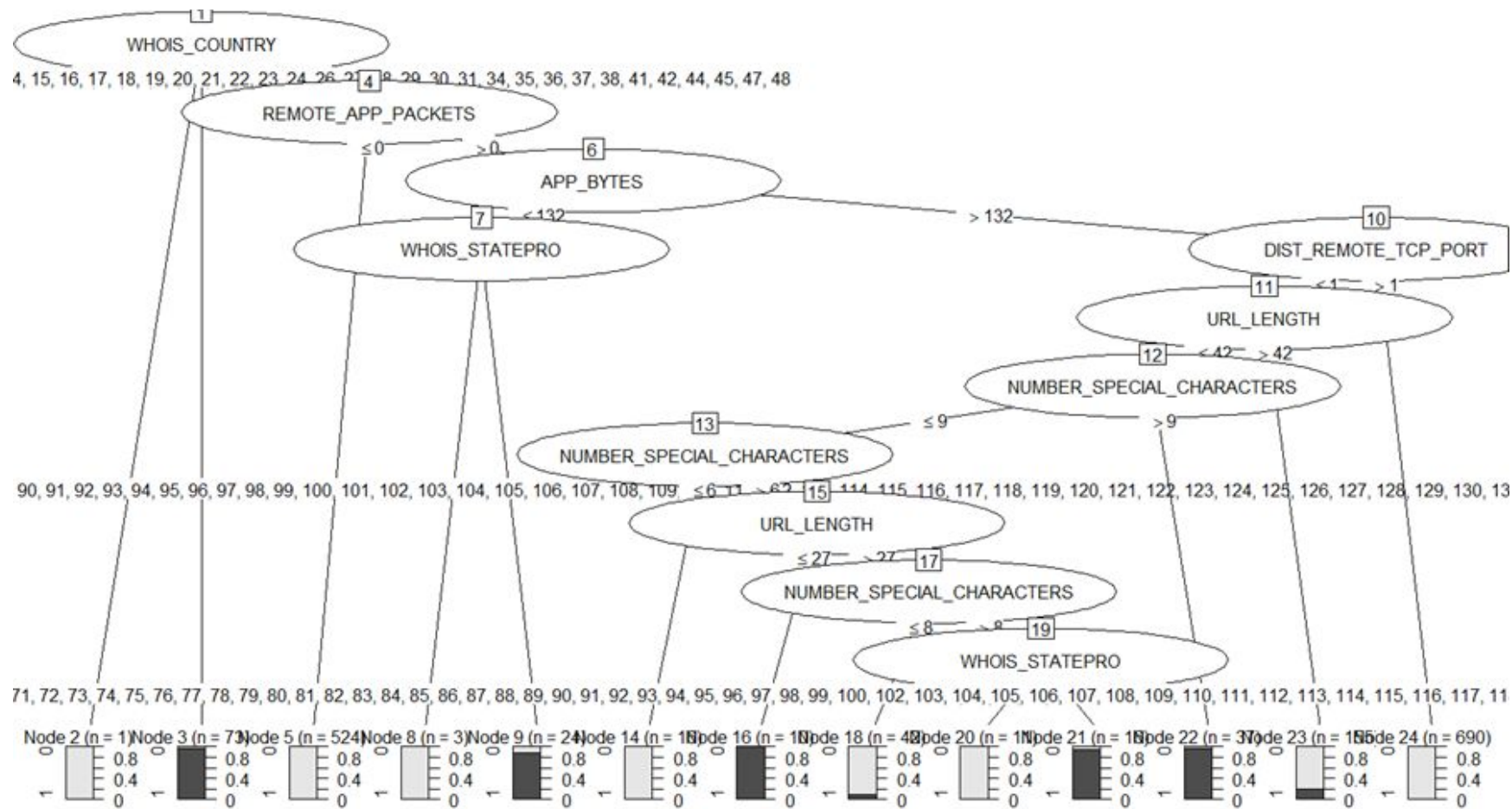
```
treeModel <- C5.0.default(x = trainData[,vars], y =  
  trainData$Type)  
assign(paste0("treeModel",i), treeModel)
```

Implemented with C50 package on 

Pohon keputusan yang ditampilkan merupakan pohon keputusan dengan akurasi terbaik dari model-model yang terbentuk.

Best decision tree: model dari *fold* kedelapan

Attribute with highest information gain: WHOIS_COUNTRY



PENGUJIAN MODEL KLASIFIKASI

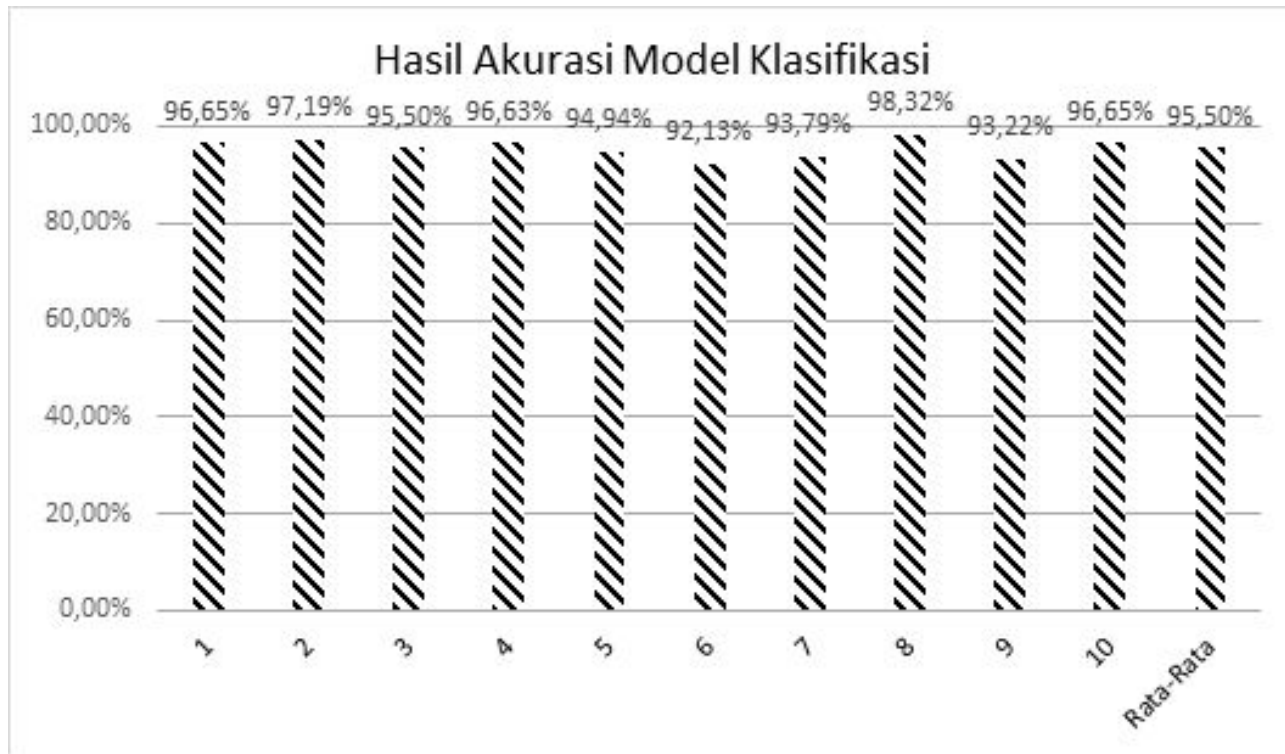
$$Akurasi = \frac{\sum \text{data uji yang diklasifikasikan benar}}{\sum \text{data uji keseluruhan}} \times 100\% \quad (4)$$

(Han et al. 2012)

Previous works accuracy (Urcuqui et al. 2017)

- ▶ Support vector machine (SVM) : 85.46%
- ▶ Regresi logistik : 84.51%
- ▶ Naïve Bayes : 85.46%
- ▶ J48 : 96.05%

Pengujian akurasi pada penelitian ini menggunakan nilai rata-rata dari akurasi pada penelitian-penelitian sebelumnya yaitu sebesar 87.87%.



Pada tahap ini tingkat akurasi model terbaik dibandingkan dengan ambang batas yang didapat dari penelitian sebelumnya (87.87%). Model terbaik memiliki akurasi sebesar 98.32% sehingga penelitian ini tidak perlu diiterasi.

3.

PENUTUP

Simpulan
Saran

SIMPULAN

Penelitian ini berhasil mengklasifikasikan situs web berbahaya dan aman menggunakan algoritme pohon keputusan C5.0.

Model terbaik dari pohon keputusan memiliki tingkat akurasi sebesar 98.32% pada model ke 8 sehingga penelitian tidak perlu diiterasi.



SIMPULAN

Atribut yang memiliki pengaruh/informasi terbanyak adalah:

1. WHOIS_COUNTRY (100%)
2. REMOTE_APP_PACKETS (95.26%)
3. APP_BYTES (62.17%)
4. DIST_REMOTE_TCP_PORT (60.61%)
5. URL_LENGTH (17.98%)
6. NUMBER_SPECIAL_CHARACTERS (8.11%)
7. WHOIS_STATEPRO (3.18%).



SARAN

- ▶ Pemerintah dapat menggunakan model klasifikasi ini untuk menyaring situs-situs yang berbahaya.
- ▶ Perbarui data yang akan diteliti di penelitian selanjutnya dengan data yang lebih relevan dengan kondisi di Indonesia.

THANK YOU! 🍌