



Identifikasi Serangan Denial Of Service (Dos) Di Jaringan Dengan Algoritma Decision Tree C4.5

Meirza Pramana ^{a,1}, Endang Setyati ^{b,2}, F.X. Ferdinandus ^{c,3}

^{a, b, c}, Departemen Teknologi Informasi, Institut Sains dan Teknologi Terpadu Surabaya, Surabaya, Indonesia

Email coresponden author : meirza@gmail.com

ABSTRAK

DoS dan DDoS merupakan serangan pada jaringan komputer yang membanjiri lalu lintas jaringan dengan *request* yang terus menerus. Untuk itulah usaha mengamankan jaringan komputer dan tindakan pencegahan perlu dilakukan dengan memasang *firewall*, perangkat IDS / IPS. IDS berfungsi sebagai alarm untuk admin bahwa ada aktivitas yang tidak normal pada jaringan, sehingga admin dapat segera mengambil tindakan pencegahan. Dalam mendeteksi serangan, IDS menggunakan metode atau algoritma untuk mengenali anomali yang terjadi dalam jaringan. Algoritma tersebut diharapkan mampu melakukan klasifikasi antara trafik yang berbahaya dan trafik normal. Data mining cocok diterapkan dalam klasifikasi trafik jaringan karena besarnya ukuran data dan jenis serangan yang beragam. Algoritma *decision tree* C4.5 diharapkan mampu digunakan dalam proses klasifikasi trafik dengan tujuan mengidentifikasi serangan DoS. Hasil uji coba dengan *dataset testing*, C.45 menghasilkan akurasi sebesar 90,68 % dalam melakukan klasifikasi trafik untuk identifikasi serangan DoS, dan menghasilkan akurasi sebesar 74,99 % dalam melakukan klasifikasi untuk semua jenis trafik. Algoritma *Naïve Bayes* digunakan sebagai perbandingan, akurasinya sebesar 86,56 % dalam melakukan klasifikasi trafik identifikasi serangan DoS, dan menghasilkan akurasi sebesar 69,50 % dalam melakukan klasifikasi semua jenis trafik. Algoritma C4.5 lebih unggul dalam hal akurasi tetapi membutuhkan waktu lebih lama untuk membangun model daripada algoritma *Naïve Bayes*.

ABSTRACT

DoS and DDoS are attacks on computer networks that flood network traffic with continuous requests. For this reason, efforts to secure computer networks and preventive measures need to be carried out by installing firewalls, IDS / IPS devices. The IDS acts as an alarm to the admin that there is abnormal activity on the network, so that the admin can take immediate preventive action. In detecting attacks, IDS uses methods or algorithms to identify anomalies that occur in the network. The algorithm is expected to be able to classify between dangerous traffic and normal traffic. Data mining is suitable to be applied in the classification of network traffic because of the large size of the data and the various types of attacks. The

Sejarah Artikel

Diterima : 8 Nopember 2021
Disetujui : 20 Desember 2021

Kata kunci:

DoS, C4.5, NSL-KDD, Klasifikasi, Data Mining

Keywords:

DoS, C4.5, NSL-KDD, Classification, Data Mining

C4.5 decision tree algorithm is expected to be able to be used in the traffic classification process with the aim of identifying DoS attacks. The results of the trial with dataset testing, C.45 yielded an accuracy of 90,68% in classifying traffic for the identification of DoS attacks, and yielded an accuracy of 74,99% in classifying all types of traffic. The Naïve Bayes algorithm is used as a comparison, the accuracy is 86,56% in classifying DoS attack identification traffic, and produces an accuracy of 69,50% in classifying all types of traffic. The C4.5 algorithm is superior in terms of accuracy but takes longer to build the model than the Naïve Bayes algorithm.

Pendahuluan

Dengan semakin berkembangnya teknologi informasi, semua orang semakin mudah berkomunikasi tanpa dibatasi jarak dan waktu. Pemanfaatan teknologi informasi telah digunakan hampir di semua bidang, untuk mengurangi keterbatasan. Kemudahan tersebut menjadikan ketergantungan manusia terhadap internet begitu besar. Internet telah digunakan di hampir setiap aspek kehidupan kita. Pada perkembangan teknologi jaringan internet sekarang ini banyak sekali tipe ancaman pada jaringan internet seperti DoS (Denial of Service), DDoS (Distributed Denial of Service).

Dengan semakin besarnya ketergantungan manusia terhadap internet saat ini, akses terhadap internet memicu terjadinya trafik yang tinggi / anomali trafik di dalam suatu jaringan komputer. Trafik yang tinggi / anomali trafik dapat terjadi akibat dua hal, yaitu *flash crowd* dan serangan *traffic flooding*.

DoS dan DDoS merupakan serangan traffic flooding yang membanjiri lalu lintas jaringan dengan request terus menerus menggunakan data yang besar (traffic flooding) terhadap server. Jika *bandwidth* dan *resource server* tidak bisa lagi menampung *request*, maka server down sehingga tidak dapat diakses oleh user yang terdaftar/berhak. Menurut laman imperva.com dalam “2019 Global DDoS Threat Landscape” Report Laporan ini adalah analisis statistik dari 3.643 serangan DDoS pada network layer sepanjang tahun 2019 dan 42.390 serangan DDoS pada *application layer* yang ditanggulangi oleh Imperva dari Mei hingga Desember 2019. Serangan DDoS pada network layer yang mencapai 580 juta PPS (Packet Per Second) (PPS) pada bulan April, dan serangan pada application layer berlangsung selama 13 hari dan puncaknya hingga 292.000 RPS (Request Per Second).

Untuk itulah usaha mengamankan jaringan komputer dan tindakan pencegahan perlu dilakukan dengan memasang *firewall*, anti virus, perangkat IDS (Intrusion Detection System) / IPS (Intrusion Prevention System). IDS adalah sebuah mekanisme untuk mendeteksi aktivitas dalam suatu jaringan. IDS berfungsi sebagai alarm yang member peringatan kepada admin bahwa ada aktivitas yang tidak normal pada jaringan, sehingga admin dapat segera mengambil tindakan pencegahan untuk mengamankan. IPS adalah mekanisme untuk melakukan pencegahan terhadap gangguan pada jaringan. Dalam mendeteksi serangan pada jaringan ini, IDS menggunakan metode atau algoritma untuk mengenali suatu serangan maupun anomali yang terjadi dalam jaringan.

Algoritma tersebut diharapkan mampu melakukan klasifikasi antara trafik yang berbahaya dan trafik normal. Konsep *data mining* cocok untuk diterapkan dalam identifikasi serangan karena besarnya ukuran data dan jenis serangan yang cukup beragam. Salah satu algoritma yang diharapkan mampu digunakan dalam proses klasifikasi trafik ini yakni algoritma *decision tree* J48/C4.5.

Metode

Pada bagian ini akan dipaparkan metode, desain sistem dan analisis data yang digunakan dalam penelitian.

1. DoS (Denial of Service)

Denial of Service adalah serangan pada jaringan atau server yang dilakukan menggunakan satu komputer. Serangan ini bisa dijalankan jika penyerang memiliki sumber daya yang lebih besar dari target dalam hal ini memiliki bandwidth, kapasitas memory dan kecepatan procesor yang lebih besar daripada target. Jika sumber daya lebih kecil dari target, maka penyerang akan mengalami kegagalan koneksi, karena jaringan penuh dengan paket yang dia kirimkan. Secara garis besar serangan DoS dapat dibagi menjadi tiga jenis yaitu:

- a. Volume Based Attack (Serangan Berbasis Volume): serangan yang bertujuan untuk membanjiri target dengan request-request berukuran besar. Serangan ini menargetkan bandwidth situs yang diserang, biasanya diukur dalam bits per detik (bps). Contoh serangan ini adalah UDP flood, ICMP flood.
- b. Protocol Attack (Serangan Berbasis Protokol): Jenis serangan yang menghabiskan sumber daya pada server, atau peralatan komunikasi menengah, seperti router, load balancers, dan

bahkan beberapa firewall. Beberapa contoh serangan protokol termasuk SYN flood, Ping of Death, Smurf dan banyak lagi. Protocol attack biasanya diukur dalam paket per detik (pps).

- c. Application Layer Attack (Serangan pada layer aplikasi): Serangan ini bertujuan untuk menghancurkan server web dengan mengirimkan respon ke HTTP request, tidak menjadi masalah jika hanya melayani satu request. Tetapi akan menjadi masalah jika melayani banyak request secara bersamaan apalagi jika sekaligus menjalankan query pada database. Contoh serangan pada layer aplikasi adalah Slowloris, serangan Zeroday DoS, serangan DoS yang mentargetkan web server Apache, dan kerentanan sistem operasi. Tipe dari serangan ini diukur dalam request per detik (rps).

2. Data Mining

Data mining merupakan bidang dari beberapa bidang keilmuan yang menyatukan teknik dari pembelajaran mesin, pengenalan pola, statistik, database, dan visualisasi untuk penanganan permasalahan pengambilan informasi dari database yang besar. Data mining cocok untuk memecahkan masalah serangan pada jaringan karena memiliki kemampuan untuk melakukan proses dari sejumlah besar data. Menurut Han dan Kamber, (2011, p24), secara garis besar data mining dapat dikelompokkan menjadi 2 kategori utama, yaitu:

- a. Descriptive mining, yaitu proses untuk menemukan karakteristik penting dari data dalam suatu basis data. Teknik data mining yang termasuk dalam descriptive mining adalah clustering, association, dan sequential mining.
- b. Predictive, yaitu proses untuk menemukan pola dari data dengan menggunakan beberapa variabel lain di masa depan. Salah satu teknik yang terdapat dalam predictive mining adalah klasifikasi.

3. Klasifikasi

Klasifikasi adalah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui (Han dan Kamber, 2006). Klasifikasi merupakan fungsi pembelajaran yang memetakan (mengklasifikasi) sebuah unsur (item) data ke dalam salah satu dari beberapa kelas yang sudah didefinisikan. Data input untuk klasifikasi adalah koleksi dari record. Setiap record dikenal sebagai instance atau contoh, yang ditentukan oleh

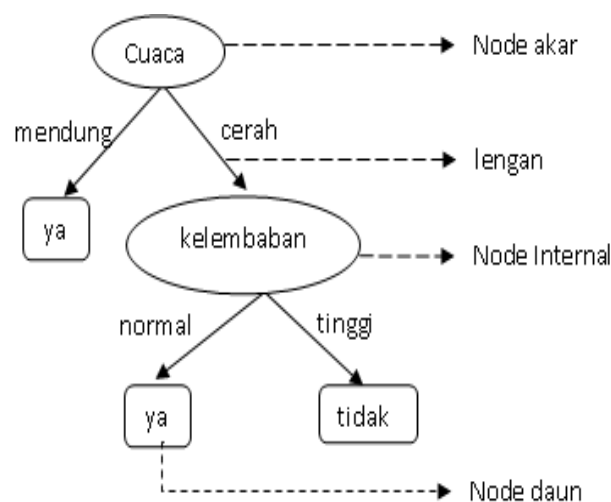
sebuah tuple (x, y) , dimana x adalah himpunan atribut dan y adalah atribut tertentu, yang dinyatakan sebagai label kelas (juga dikenal sebagai kategori atau atribut target).

Beberapa teknik klasifikasi yang digunakan adalah decision tree classifier, rule-based classifier, neural-network, support vector machine, dan naive bayes classifier. Setiap teknik menggunakan algoritma pembelajaran untuk mengidentifikasi model yang memberikan hubungan yang paling sesuai antara himpunan atribut dan label kelas dari data input

4. Konsep Decision Tree

Han & Kamber, (2006) dalam bukunya *Data Mining Concepts and Techniques* menjelaskan bahwa decision tree merupakan metode yang menyimpulkan pernyataan-pernyataan dari sebuah label kelas pelatihan pada data latih (training set), untuk menemukan masalah-masalah yang mengarah dalam membuat sebuah keputusan. Sebuah decision tree secara umum, adalah flowchart seperti struktur pohon yaitu gambaran permodelan dari suatu persoalan, dimana setiap simpul internal menunjukkan data uji pada atribut, setiap cabang menunjukkan dari datang dan keluarnya data uji, dan masing-masing node memegang label kelas. Node yang paling atas merupakan simpul akar. Konsepnya adalah mengubah data menjadi decision tree dan aturan-aturan keputusan (rule).

5. Algoritma C4.5



Gambar 1
Contoh Konsep Pohon Keputusan

Pohon keputusan merupakan pendekatan *divide and conquer* dalam mempelajari masalah dari sekumpulan data independen yang digambarkan dalam bagan pohon (Witten, Frank, & Hall, 2011). Pohon keputusan dianggap sebagai salah satu pendekatan yang paling populer, dalam klasifikasi pohon keputusan terdiri dari sebuah node yang membentuk akar, node akar tidak memiliki inputan. Node lain yang bukan sebagai akar tetapi memiliki tepat satu inputan disebut node internal atau test node, sedangkan node lainnya dinamakan daun. Daun mewakili nilai target yang paling tepat dari salah satu kelas (Maimon & Rokach, 2010).

Menurut Kusriani & Luthfi, (2009) ada beberapa tahap dalam membuat sebuah pohon keputusan dengan, algoritma C4.5, yaitu:

- a. Menyiapkan data training. Data training biasanya diambil dari data histori yang pernah terjadi sebelumnya dan sudah dikelompokkan ke dalam kelas-kelas tertentu.
- b. Menentukan akar dari pohon. Akar akan diambil dari atribut yang terpilih, dengan cara menghitung nilai gain dari masing-masing atribut, nilai gain yang paling tinggi yang akan menjadi akar pertama. Sebelum menghitung gain dari atribut, hitung dahulu nilai entropy yaitu:

$$Entropy(S) = \sum_{i=1}^n -p_i \log_2 p_i \quad (1)$$

Berikut ini penjelasan dari persamaan 1, dimana S adalah himpunan kasus. Variabel n adalah jumlah partisi himpunan S. Sedangkan p_i adalah jumlah sampel kelas i.

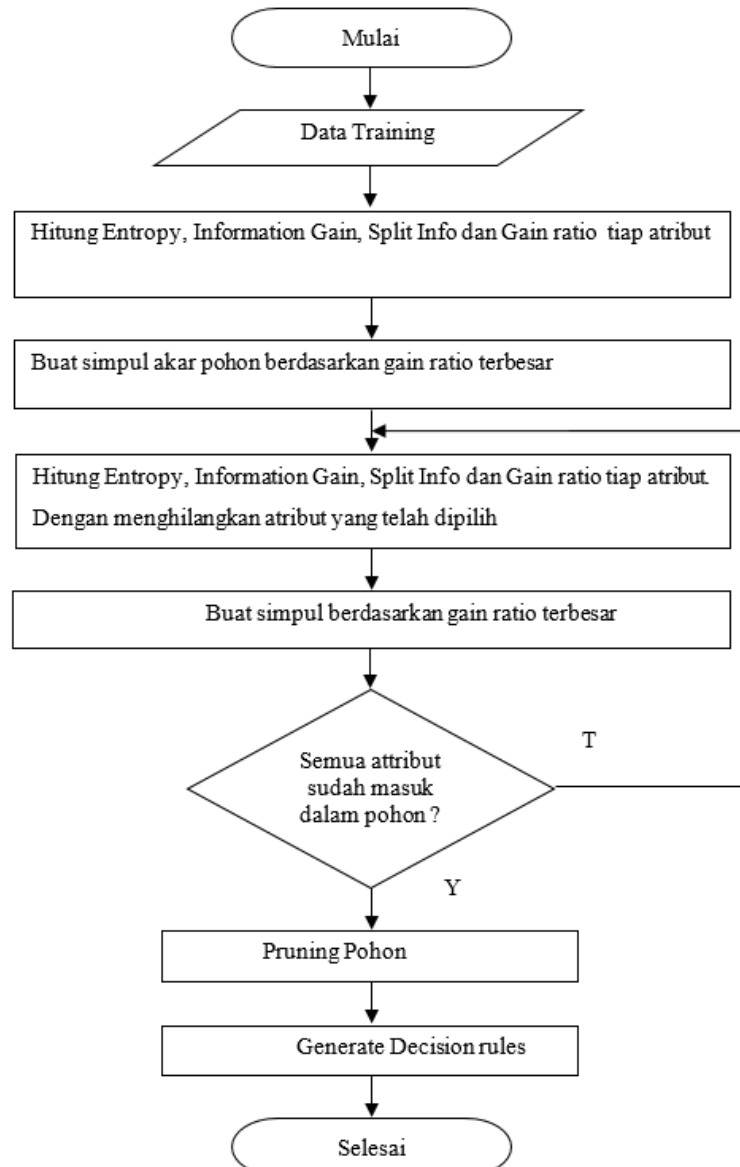
- c. Kemudian hitung nilai gain dengan metode informasi gain

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} Entropy(S_i) \quad (2)$$

Berikut ini penjelasan dari persamaan 2 dimana S adalah himpunan kasus. A adalah atribut dari S, variabel n adalah jumlah partisi atribut A. $|S_i|$ adalah jumlah kasus pada partisi ke-i sedangkan $|S|$ adalah jumlah kasus pada S.

- d. Ulangi langkah ke-2 hingga semua tupel terpartisi.
- e. Proses partisi pohon keputusan akan berhenti saat:
 - Semua tupel dalam node N mendapat kelas yang sama
 - Tidak ada atribut di dalam tupel yang dipartisi lagi.
 - Tidak ada tupel di dalam cabang yang kosong.

Flowchart algoritma C4.5 bisa dilihat pada gambar 2 berikut ini.



Gambar 2
Flowchart algoritma C4.5

6. Desain Sistem

Pembahasan pada bagian ini meliputi tentang analisa permasalahan arsitektur sistem yang hendak dibangun. Terdiri dari beberapa tahap antara lain, pengumpulan dataset, penyusunan algoritma, dan gambaran arsitektur sistem. Penelitian dalam karya ilmiah ini termasuk dalam jenis penelitian kuantitatif, dataset yang digunakan adalah dataset NSL-KDD (Network Security Lab – Knowledge Discovery in Databases) yang merupakan perbaikan

dari dataset KDD CUP 99 yang memiliki masalah dengan data bias, data ganda. NSL-KDD menghapus data redundan pada KDD CUP 99 yang dapat mempengaruhi performa algoritma learning Dataset training dan dataset testing NSL-KDD memiliki 41 fitur seperti yang ditampilkan pada tabel 1, dan 5 kelas yang terdiri dari 1 kelas trafik normal dan 4 kelas jenis serangan. Berikut ini jenis serangan pada dataset NSL-KDD :

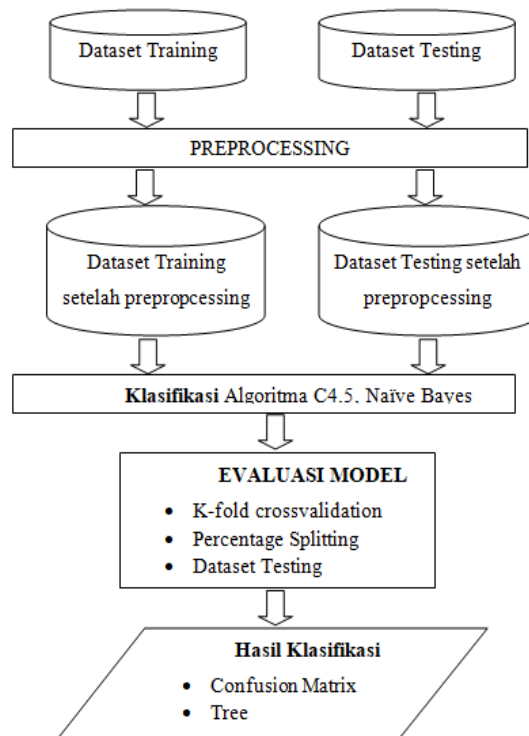
- a. DoS (Denial of Service) adalah tipe serangan yang membebani sumberdaya komputer (misalnya dengan synflood atau ping of death) sehingga komputer target mengalami system crash dan tidak mampu untuk memproses koneksi normal bahkan mengakibatkan user tidak dapat mengakses komputer tersebut.
- b. R2L (remote to local) adalah tipe serangan yang bertujuan untuk mendapatkan akses sebagai pengguna sistem. R2L dilakukan oleh penyerang yang memiliki akses ke sistem dan melakukan eksploitasi untuk mendapatkan akses lokal.
- c. Serangan Probe bertujuan untuk mendapatkan informasi tentang status jaringan komputer dengan cara melakukan pemindaian terhadap komputer - komputer dalam jaringan tersebut. Informasi ini dapat digunakan oleh penyerang untuk memetakan jaringan yang berguna dalam melakukan penyerangan berikutnya.
- d. U2R (user to root) adalah tipe serangan yang berusaha untuk mendapatkan akses root/admin pada komputer target dengan melakukan eksploitasi celah keamanan sistem. Serangan U2R umumnya dilakukan setelah penyerang mendapatkan akses user normal ke sistem (baik melalui sniffing, social engineering, ataupun dictionary attack).

Tabel 1
Atribut Dataset NSL-KDD

No	Atribut	Type
1	duration	Numeric
2	protocol_type	Nominal
3	service	Nominal
4	flag	Nominal
5	src_bytes	Numeric
6	dst_bytes	Numeric
7	land	Biner
8	wrong_fragment	Numeric
9	urgent	Numeric
10	hot	Numeric
11	num_failed_logins	Numeric

12	logged_in	Biner
13	num_compromised	Numeric
14	root_shell	Biner
15	su_attempted	Biner
16	num_root	Numeric
17	num_file_creations	Numeric
18	num_shells	Numeric
19	num_access_files	Numeric
20	num_outbound_cmds	Numeric
21	is_hot_login	Biner
22	is_guest_login	Biner
23	count	Numeric
24	srv_count	Numeric
25	serror_rate	Numeric
26	srv_serror_rate	Numeric
27	rerror_rate	Numeric
28	srv_rerror_rate	Numeric
29	same_srv_rate	Numeric
30	diff_srv_rate	Numeric
31	srv_diff_host_rate	Numeric
32	dst_host_count	Numeric
33	dst_host_srv_count	Numeric
34	dst_host_same_srv_rate	Numeric
35	dst_host_diff_srv_rate	Numeric
36	dst_host_same_src_port_rate	Numeric
37	dst_host_srv_diff_host_rate	Numeric
38	dst_host_serror_rate	Numeric
39	dst_host_srv_serror_rate	Numeric
40	dst_host_rerror_rate	Numeric
41	dst_host_srv_rerror_rate	Numeric

Secara garis besar pada gambar 3 berikut ini adalah blok diagram dari model klasifikasi serangan DoS dengan algoritma C4.5 yang digunakan dalam penelitian ini.



Gambar 3
Blok Diagram Klasifikasi Serangan DoS dengan Algoritma C4.5

a. Preprocessing

Dalam data mining, tahap yang paling awal adalah tahap praproses atau lebih dikenal data preprocessing. Pada tahap praproses, dataset yang akan dijadikan sebagai data latih dianggap sebagai data mentah, artinya dataset tersebut harus diperiksa terlebih dahulu kelengkapannya dan proporsi jumlah label tiap kelasnya.

b. Pemilihan Atribut Dataset

Setelah dataset NSL-KDD dinilai sudah bersih, lengkap dan matang, tahap berikutnya adalah memeriksa tingkat kontribusi tiap variabel prediktor pada dataset NSL-KDD. Tingkat kontribusi suatu variabel prediktor (atribut) sebenarnya cukup memiliki banyak kriteria pilihannya, namun salah satu metode yang dapat digunakan adalah Information Gain. Information Gain akan meranking semua fitur variabel prediktor yang memiliki banyak informasi berdasarkan kelas tertentu.

c. Klasifikasi

Pada bagian ini akan dibahas skenario uji coba pembentukan model prediktif dataset training NSL-KDD menggunakan algoritma Decision Tree C4.5 dan algoritma naïve bayes. Beberapa teknik *validation sampling* antara lain menggunakan K-Fold, Percentage Split dan yang terakhir adalah menggunakan dataset testing.

d. Evaluasi Kinerja

Evaluasi dilakukan untuk menguji hasil klasifikasi dengan mengukur nilai kinerja dari sistem yang dibuat. Jumlah yang benar dan prediksi yang salah dirangkum dengan dipecah ke dalam kelas masing-masing. Confusion matriks pada table 2 menjadi dasar untuk memeriksa keakuratan dan kredibilitas model yang diusulkan. Beberapa parameter pengujian yang digunakan untuk evaluasi adalah:

- Accuracy (akurasi) adalah Informasi seberapa baik algoritma klasifikasi dalam mengidentifikasi / melakukan deteksi aktivitas serangan DoS dengan benar dari suatu dataset. Atau sebagai persentase dari jumlah total prediksi yang benar. Itu dapat dihitung menggunakan persamaan 3.

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

- Recall (Sensitivitas): Sebuah model mampu melakukan deteksi sebagian besar serangan tetapi ada kalanya disertai dengan kesalahan dalam klasifikasi data sehingga beberapa aktifitas normal terdeteksi sebagai serangan. Bisa dihitung menggunakan persamaan 4

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

- Precision (Presisi): Merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif. Precision menjawab pertanyaan “Berapa persen trafik yang benar merupakan serangan DoS dari keseluruhan trafik yang yang diprediksi sebagai serangan DoS?” Precision dihitung menggunakan persamaan 5.

$$Precision = \frac{TP}{FP+TP} \quad (5)$$

- F-measure atau F_1 score adalah perbandingan rata-rata presisi dan recall yang dibobotkan. F-measure Bisa dihitung menggunakan persamaan 6.

$$F_1 \text{ Score} = 2 * \frac{precision * recall}{precision + recall} \quad (6)$$

- Nilai TP, TN, FP, dan FN diperoleh dengan menggunakan confusion matrix dalam tabel 2.

Tabel 2
Confusion Matrix

Aktual	Prediksi (hasil deteksi)	
	Trafik Lain	DoS
Trafik Lain	TN	FP
DoS	FN	TP

Dari hasil prediksi di atas, hanya ada kemungkinan 4 kasus yang terjadi:

- True Positive (TP): kasus dimana trafik diprediksi (Positif) DoS, dan memang benar (True) merupakan trafik serangan DoS
- True Negative (TN): kasus dimana trafik diprediksi tidak (Negatif) DoS dan sebenarnya trafik tersebut memang (True) bukan DoS.
- False Positive (FP): kasus dimana trafik yang diprediksi positif DoS, ternyata merupakan trafik lainnya. Prediksinya salah (False).
- False Negative (FN): kasus dimana trafik yang diprediksi bukan DoS (Negatif), tetapi ternyata sebenarnya (TRUE) DoS.

Hasil dan Pembahasan

Adapun hasil dari klasifikasi C4.5 secara keseluruhan dapat dilihat pada tabel 3. Terlihat perbedaan akurasi yang cukup signifikan antara training set, validation sampling dan testing, dimana rata-rata akurasi pada validation sampling mencapai $\geq 99,97\%$ namun pada testing hanya mencapai $\geq 90,68\%$. Hal ini biasa terjadi dalam bidang *machine learning* yang dikenal *overfitting*, artinya model prediktif yang terbentuk pada fase pelatihan mampu merepresentasikan prediksi label data latih secara general, namun belum mampu merepresentasikan secara baik terhadap prediksi label data uji baru yang mungkin mempunyai karakteristik berbeda dengan data latih sebelumnya.

Tabel 3

Hasil Uji Coba C4.5 Menggunakan Training Set, Validation Sampling dan Testing Set

Pengujian	Accuracy %	Precision %	Recall %	F-Measure %	Time taken to build model (sec)
Training set	99,99	99,9	99,84	99,9	09,82
K-Fold = 5	99,98	99,96	99,96	99,96	10,33
Percentage Split 70:30	99,97	99,93	99,96	99,94	13,76
Testing set	90,68	95,70	75,23	84,23	14,26

Uji coba selanjutnya menggunakan aplikasi WEKA untuk pembentukan model prediktif dengan training data dataset NSL-KDD menggunakan algoritma naïve bayes. Menggunakan beberapa teknik validation sampling antara lain menggunakan K-Fold, Percentage Split dan yang terakhir adalah testing data. Adapun hasil dari uji coba dengan naïve bayes dapat dilihat pada tabel 4.

Tabel 4

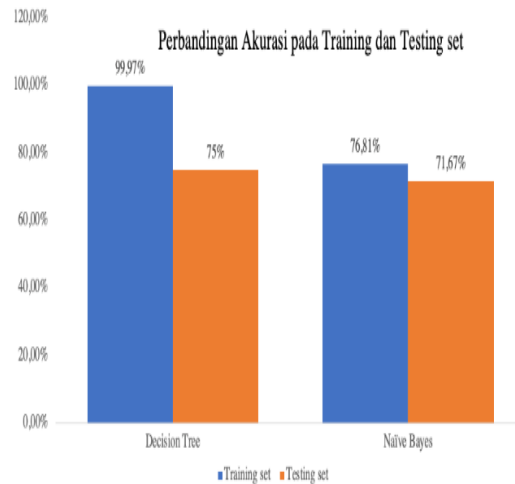
Hasil Uji Algoritma Naïve Bayes, Berdasarkan Training Set, Validation Sampling & Testing Set

Pengujian	Accuracy %	Precision %	Recall %	F-Measure %	Time taken to build model (sec)
Training set	97,81	94,45	94,64	94,55	1,12
K-Fold = 5	97,84	94,57	94,64	94,6	1,19
Percentage Split 70:30	97,77	94,69	94,28	94,49	2
Testing set	86,56	85,99	70,92	77,73	0,93

Secara keseluruhan uji coba, keunggulan dari algoritma naïve bayes adalah pembentukan model yang relatif singkat yakni hanya membutuhkan waktu ≤ 2 detik. Tetapi tingkat akurasi, presisi dan recall dalam melakukan klasifikasi trafik serangan DoS dengan trafik lainnya, lebih rendah dari algoritma C4.5.

Pada bagian ini akan dibahas bagaimana analisis perbandingan performa algoritma Decision Tree dan Naïve Bayes dalam menyelesaikan klasifikasi dataset NSL-KDD dari hasil pengujian training set, validation sampling dan testing set dalam bentuk representasi grafik, agar lebih mudah untuk dipahami kondisi perbandingannya. Pada gambar 4 terlihat perbandingan performa akurasi algoritma C4.5 mengungguli performa algoritma Naïve Bayes baik dalam fase training dan testing set. Perbedaan yang cukup signifikan di antara keduanya dimana Decision Tree mencapai akurasi $\geq 99,97\%$, sedangkan algoritma Naïve Bayes mencapai akurasi $\geq 97,77\%$ pada fase training set. Pada fase testing set, algoritma

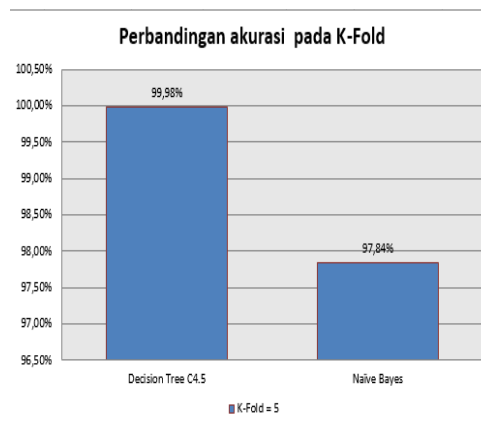
Decision Tree mencapai akurasi 90,68% sedangkan algoritma Naïve Bayes hanya mencapai 86,56%.



Gambar 4

Grafik Perbandingan Performa Akurasi Algoritma Decision Tree dan Naïve Bayes Pada Training dan Testing Set

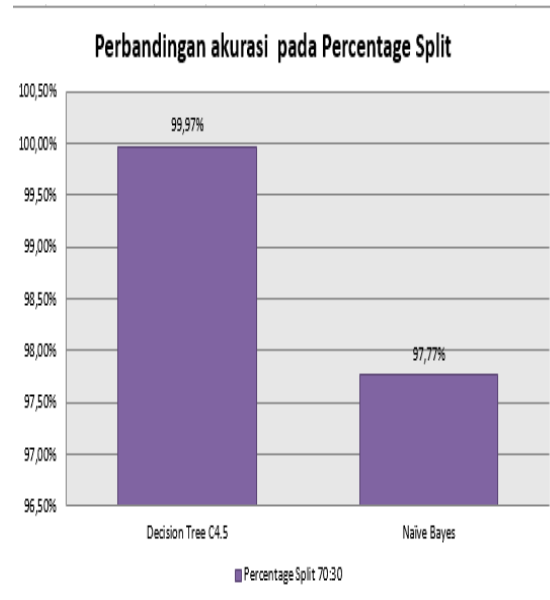
Pada gambar 5 terlihat perbandingan akurasi yang cukup signifikan antara algoritma Decision Tree dan Naïve Bayes saat pengujian menggunakan validation sampling K-Fold. Algoritma C4.5 dapat mencapai akurasi 99,98% pada K-Fold=5, sedangkan algoritma Naïve Bayes mencapai akurasi 97,84 %



Gambar 5

Grafik Perbandingan Performa Akurasi Algoritma Decision Tree dan Naïve Bayes Pada Training dan Testing Set

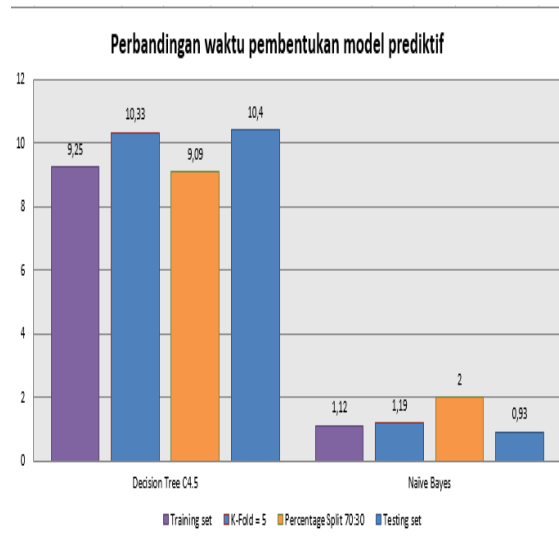
Gambar 6 adalah perbandingan akurasi antara algoritma Decision Tree C4.5 dan Naïve Bayes saat pengujian menggunakan validation sampling Percentage Split 70:30. Algoritma Decision Tree dapat mencapai akurasi $\geq 99,97\%$, sedangkan algoritma Naïve Bayes mencapai akurasi $\geq 97,77\%$.



Gambar 6

Grafik perbandingan performa akurasi C4.5 dan Naïve Bayes pada validation sampling Percentage Split

Gambar 7 adalah perbandingan waktu pembentukan model, terdapat perbedaan hasil yang sangat signifikan, karena algoritma Decision Tree C4.5 cukup banyak membutuhkan waktu dengan rentang waktu 9,25 hingga 10,4 detik dibandingkan algoritma Naïve Bayes yang hanya membutuhkan waktu 0,93 hingga 2 detik.



Gambar 7
Grafik Perbandingan Waktu Pembentukan Model Prediktif pada Algoritma Decision Tree dan Naïve Bayes

Simpulan

Dari hasil seluruh uji coba yang telah dilakukan dapat diambil beberapa kesimpulan. Algoritma Decision Tree C4.5 memberikan performa akurasi, presisi, sensitivitas yang sangat baik dalam membangun sebuah model prediktif pada dataset NSL-KDD namun harus membutuhkan waktu komputasi yang lebih lama. Terjadi perbedaan akurasi yang signifikan saat melakukan proses dengan algoritma C4.5 menggunakan dataset training yang akurasinya dalam melakukan prediksi serangan DoS mencapai $\geq 99,8\%$ dibandingkan menggunakan dataset testing yang akurasinya $90,68\%$. Hal ini biasa terjadi dalam bidang machine learning yang dikenal overfitting, artinya model prediktif yang terbentuk pada fase pelatihan mampu merepresentasikan prediksi label data latih secara general, namun belum mampu merepresentasikan secara baik terhadap prediksi label data uji baru yang mungkin mempunyai karakteristik berbeda dengan data latih sebelumnya. Algoritma Naïve Bayes memberikan performa akurasi, presisi, sensitivitas dibawah algoritma C4.5 dalam membangun sebuah model prediktif namun memiliki kelebihan dengan waktu komputasi yang relatif singkat

Referensi

- Andri Andri, Yesi Novaria Kunang, Sri Murniati. (2013), Implementasi Teknik Data Mining Untuk Memprediksi Tingkat Kelulusan Mahasiswa Pada Universitas Bina Darma Palembang, *Seminar Nasional Informatika*, ISSN: 1979-2328, pp. A-57
- A. R. Yusof, N. I. Udzir, A. Selamat, H. Hamdan and M. T. Abdullah. (2017). Adaptive feature selection for denial of services (DoS) attack. *IEEE Conference on Application, Information and Network Security (AINS)*, Miri, 2017, pp. 81-84.
- Eko Prasetyo. (2014). *Data Mining: Mengolah Data Menjadi Informasi Menggunakan Matlab*. Yogyakarta: Andi Offset
- Han, J. & M. Kamber. (2006). *Data Mining Concepts and Techniques Second Edition*. San Francisco: Morgan Kaufmann.
- Imperva.com. Nadav, Avishay, Johnathan, Kim. 2019 Global DDoS Threat Landscape Report. 04 Februari 2020. [Diakses 23 Maret 2020]. Diakses dari <https://www.imperva.com/blog/2019-global-ddos-threat-landscape-report/>
- Kawelah, Wathq & Abdala, Ahmed. (2019). A Comparative Study on Machine Learning Tools Using WEKA and Rapid Miner with Classifier Algorithms C4.5 and Decision Stump for Network Intrusion Detection. *European Academic Research*. 7. 852-861.
- Khan, Suleman & Gani, Abdullah & Wahid, Ainuddin & Singh, Prem. (2017). Feature Selection of Denial-of-Service Attacks Using Entropy and Granular Computing. *Arabian Journal for Science and Engineering*. 43. 499-508. 10.1007/s13369-017-2634-8.
- Kusrini, luthfi taufiq Emha. (2009). *Algoritma Data Mining*. Yogyakarta: Andi Offset
- Maimon, O., & Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook*. London: Springer Science + Business Media
- S. Das, A. M. Mahfouz, D. Venugopal and S. Shiva. (2019). DDoS Intrusion Detection Through Machine Learning Ensemble. *2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C)*, Sofia, Bulgaria, pp. 471-477, doi: 10.1109/QRS-C.2019.00090
- Santosh Kumar Pydipalli, Srikanth Kasthuri, Jinu. (2018). DDoS Detection System Using C4.5 Decision Tree Algorithm. *International Research Journal of Engineering and Technology (IRJET)*
- T. Mehmood and H. B. M. Rais. (2016). Machine learning algorithms in context of intrusion detection. *2016 3rd International Conference on Computer and Information Sciences (ICCOINS)*, Kuala Lumpur, pp. 369-373.
- Unb.ca. unb.ca. NSL-KDD dataset. [Diakses 03 Maret 2020]. Diakses dari <https://www.unb.ca/cic/datasets/nsl.html>
- Wahba, Y., ElSalamouny, E., & ElTaweel, G. (2015). Improving the Performance of Multi-class Intrusion Detection Systems using Feature Reduction. ArXiv, abs/1507.06692.