IMPLEMENTASI ALGORITMA DECISION TREE UNTUK KLASIFIKASI POLA SERANGAN PADA LOG FILE

Miftarullah Firdaus[1], Triawan A. C[2]
Jurusan Teknik Informatika Fakultas Teknik
Universitas Muhammdiyah Jember
miftah.ustad@gmail.com[1], triawanac@unmuhjember.ac.id[2]

ABSTRAK

Sistem Deteksi Intrusi merupakan sebuah mekanisme dalam usaha menyediakan keamanan bagi jaringan komputer. Oleh karena itu diperlukan suatu cara proses identifikasi serangan dalam usaha menjaga keamanan jaringan. Algoritma Machine Learning untuk deteksi intrusi jaringan dilakukan, dimana performa harus berada pada level yang dapat diterima untuk berbagai tipe serangan pada jaringan. Dari semakin meningkatnya teknologi komputer data dalam jumlah yang besar dapat dikumpulkan dan disimpan. Akan tetapi data ini baru berguna jika dianalisa dan depedensi korelasinya ditemukan. Hal ini dapat dicapai dengan menggunakan klasifikasi yang tepat, perlu diperhatikan metode klasifikasi yang tepat. Karena adanya pemprosesan data yang besar dan kompleks serta sifat dinamis dari tipe serangan, metode Data Mining diterapkan dalam Sistem Deteksi Intrusi dalam jaringan berdasarkan trafik data . Salah satu algoritma yang diharapkan mampu digunakan dalam proses klasifikasi serangan ini yakni algoritma Decision tree. Performa dari akurasi diukur dari algoritma machine learning dengan menggunakan test KDDCUP 1999 intrusi dataset untuk menemukan klasifikasi serangan. Dalam percobaan ini deteksi intrusi jaringan dievaluasi performa dengan memanfaatkan benchmark KDDCUP 1999 dari 10000 Training Dataset. Melihat lebih jauh sebuah Decision Tree sebagai model intrusi. Dan pada akhirnya percobaan dibantu dengan Classifier J48 yang berasal dari alat perangkat lunak WEKA untuk mendapatkan akurasi dari performa dalam mencapai deteksi serangan. Dari hasil analisis didapatkan tingkat akurasi dicapai dengan sangat baik menggunakan algoritma ini dalam proses klasifikasi serangan dengan tingkat akurasi 99.76 % pada sistem.

Kata Kunci : Sistem Deteksi Intrusi, Klasifikasi, Data Mining, Decision Tree, KDDCUP 1999, Akurasi.

I. PENDAHULUAN

1. LATAR BELAKANG

Klasifikasi adalah mengatur secara sistematis sekaligus memberi arti informasi yang berguna untuk menentukan atau menetapkan kesesuaian gagasan, peristiwa, barang dan orang. Klasifikasi memiliki tujuan untuk mengklasifikasikan suatu data ke dalam kelas-kelas yang sudah ada. Tidak akan ada pembentukan kelas baru. Masalah klasifikasi sering dijumpai pada kehidupan sehari-hari, baik dibidang pendidikan, sosial, industri, kesehatan maupun perbankan. Contoh masalah klasifikasi dalam bidang pendidikan adalah klasifikasi sekolah berdasarkan akreditas sekolah. Dalam bidang kesehatan dilakukan pengklasifikasian penyakit berdasarkan tingkat keseriusan dan bahaya yang ditimbulkan.

Karena adanya pemprosesan data yang besar dan kompleks serta sifat dinamis dari tipe serangan, metode *Data Mining* diterapkan dalam Sistem Intrusi Deteksi dalam jaringan berdasarkan trafik data. Dengan semakin canggihnya teknologi komputer maka kumpulan data yang besar mampu dikumpulkan dan disimpan. Tetapi data ini semakin berguna ketika dianalisis dan beberapa dependensi serta korelasinya terdeteksi. Hal ini diharapkan mampu tercapai dengan pemanfaatan teknik *Machine Learning* algoritma *Classifier J48* dalam membangun sebuah model Sistem Intrusi Deteksi dengan *Decision Tree* yang efisien. Dalam proses analisis nantinya akan memanfaatkan standar tes data set KDD CUP 1999 dalam menentukan tingkat performa dari Sistem Intrusi Deteksi dalam menemukan

anomali serangan yang terjadi pada suatu sistem jaringan komputer. Penelitian akan dilakukan selanjutnya menggunakan bantuan dari *Classifier J48* dalam proses pembangunan *Decision Tree* menggunakan alat perangkat lunak WEKA dalam mendapatkan tingkat akurasi dalam mencapai proses pendeteksian serangan dan anomali yang terjadi pada suatu sistem.

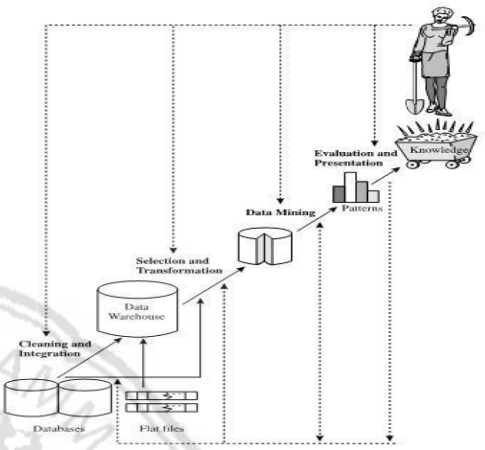
II. LANDASAN TEORI

1. Literature Review

Intrusion Detection System (IDS) (Khaerani dan Handoko. 2018) merupakan sebuah kemampuan yang dimiliki oleh sebuah sistem atau perangkat untuk dapat melakukan deteksi terhadap serangan yang mungkin terjadi dalam jaringan baik lokal maupun yang terhubung dengan internet. Masalah dimulai ketika paket data yang datang sangat banyak dan harus di analisa di kemudian hari. Teknik Data Mining merupakan teknik yang tepat untuk melakukan analisa terhadap sebuah data. Beberapa penelitian telah menggunakan teknik data mining untuk mengatasi masalah serangan IDS seperti analisis frequent itemset, analisis clustering, analisis klasifikasi dan analisis asosiasi. Tujuan dari penelitian ini adalah untuk mengklasifikasikan serangan pada data-data yang diujikan dengan menggunakan metode klasifikasi dan algoritma klasifikasi C4.5. Penelitian ini menggunakan koleksi data dari KDD'99 dan memiliki 41 atribut dimana atribut ini dilakukan fitur seleksi untuk menghapus atribut yang tidak relevan dengan menggunakan teknik evolusi. Hasil yang didapatkan dari fitur seleksi ini adalah 16 atribut dengan akurasi tinggi mencapai 98,67% dari 41 atribut yang ada. Kemudian hasilnya dilakukan pemodelan dengan menggunakan algoritma C4.5 dan menghasilkan sebuah aturan untuk digunakan dalam implementasi sistem analisa klasifikasi data. Aturan yang dihasilkan dapat digunakan dalam sistem untuk mengklasifikasikan data serangan seperti dos, u2r, r2l dan probe serta aktifitas jaringan normal. Data yang dikumpulkan dalam proses seleksi penerima beasiswa memiliki variabel berbentuk nominal dan numerik serta terdiri atas banyak *item*, sehingga diperlukan metode yang tepat untuk melakukan klasifikasi data secara akurat dan memastikan validitas data yang dipergunakan dalam penilaian penerima beasiswa. (Budiman dan Parandani.2018)

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 (Larose, 2005).



Gambar 2.1 *Data mining* sebagai tahapan dalam proses *KDD*

Proses KDD secara garis besar dapat dijelaskan sebagai berikut (Fayyad,dan Smyth 1996) :

1. Data Selection
   Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang akan digunakan, disimpan dalam suatu berkas, terpisah dari basis data operasional.
2. Pre-processing/Cleaning
   Sebelum proses data mining, perlu dilakukan proses cleaning pada data yang menjadi fokus KDD. Proses cleaning mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak.
3. Transformation Data
   Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining. Proses coding dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola

informasi yang akan dicari dalam basis data.

4. Data mining
   Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik tertentu. Teknik, metode, atau algoritma dalam data mining sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.
5. Interpretation Evaluation
   Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang disebut interpretation. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.

3. Decision Tree

Decision tree adalah sebuah diagram alir yang berbentuk seperti struktur pohon yang mana setiap *internal node* menyatakan pengujian terhadap suatu atribut, setiap cabang menyatakan output dari pegujian tersebut dan *leaf node* menyatakan kelas–kelas atau distribusi kelas. *Node* yang paling atas disebut sebagai *root node* atau *node* akar. *(Gewehr, Szugat, and Zimmer.2007).*

Proses *decision tree* adalah mengubah bentuk data (tabel) menjadi model pohon, mengubah model pohon menjadi *rule* dan menyederhanakan *rule* (Basuki, Achmad dan Syarif. 2003). Dalam membentuk pohon keputusan dengan algoritma C5.0 digunakan entropy dan information gain untuk menentukan akar node. *Gain* dengan nilai tertinggi akan menjadi node akar dari *entropy* terkecil tiap atribut. Berikut persamaan untuk menghitung nilai entropy:

$$Entropy(S) = \sum_{i=1}^{n} - pi.Log_2 pi$$

Dengan :

S : Himpunan kasus
n : Jumlah partisi S
pi : Proporsi dari S_i terhadap S

Sementara itu untuk menghitung gain atribut dapat dilihat pada persamaan sebagai berikut :

$$\text{Gain}(S.A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Si) *$$

Dengan:

S : Himpunan kasus
A : Atribut
n : Jumlah partisi atribut A
|Si| : Jumlah kasus pada partisi ke i
|S| : Jumlah kasus dalam S

4. Weka

WEKA adalah kumpulan algoritma pembelajaran mesin dan pemrosesan data alat yang mencakup hampir semua algoritma yang dijelaskan dalam buku kami. Ini dirancang agar Anda dapat dengan cepat mencoba metode yang ada pada dataset baru dengan cara yang fleksibel. Ini menyediakan luas dukungan untuk seluruh proses penambangan data eksperimental, termasuk menyiapkan data input, mengevaluasi skema pembelajaran secara statistik, dan memvisualisasikan data input dan hasil pembelajaran. Serta berbagai macam algoritma pembelajaran, itu termasuk berbagai alat preprocessing. Toolkit yang beragam dan komprehensif ini diakses melalui antarmuka umum sehingga penggunanya dapat membandingkan berbagai metode dan mengidentifikasi metode yang paling sesuai untuk masalah tersebut tangan. WEKA dikembangkan di University of Waikato di Selandia Baru; nama singkatan Lingkungan Waikato untuk Analisis Pengetahuan. Di luar universitas, WEKA diucapkan untuk berima dengan Mekah, adalah burung yang tidak bisa terbang dengan sifat ingin tahu hanya ditemukan di pulau – pulau Selandia Baru. Sistem ini ditulis dalam Java dan didistribusikan di bawah ketentuan GNU General Lisensi Publik. Ini berjalan di hampir semua platform dan telah diuji di Linux, Windows, dan Sistem operasi Macintosh. (Frank, Hall, and Witten.2016).

<u>Beberapa kelebihan utama Weka antara lain:</u>

a Merupakan perangkat lunak gratis yang dapat disebarluaskan dan digunakan yang memiliki naungan lisensi dibawah GNU General Public License.

b Bersifat sangat *portable* karena dapat diimplementasikan dalam pemrograman Java dan dapat berjalan diberbagai platform sistem komputer saat ini .

c Berisikan koleksi yang meliputi berbagai teknik *pre-processing* dan teknik permodelan data.

d Mudah digunakan oleh pemula karena terdapat antar muka grafis yang mudah dipahami bagi orang awam sekalipun.

III. METODOLOGI PENELITIAN

1. Analisa Kebutuhan

a. Pengumpulan Data

Dataset yang digunakan berasal dari University of California , Irvine Knowledge Discovery and Data Mining (UCI KDD) website . Data atau record yang digunakan 10000 data. ini memberikan informasi penting bagaimana membuat dan melatih algoritma yang akan digunakan. Data kddcup 1999 berisikan type *class,* yang terdiri dari 23 class yang berbeda. termasuk data *normal*, dimana adanya kondisi tidak adanya serangan yang terjadi. Tipe *serangan* yakni *back, buffer_overlflow, ftp_write, guess_passwd, imap, ipsweep, land , loadmodule, mulithop, neptune, nmap, normal, perl, phf, pod, portsweep, rootkit, satan, smurf, spy, teardrop, warezclient* dan *warezmaster*. (Stolfo,fan,lee,Prodomidis dan K.chan.1999).

b. Algoritma yang digunakan

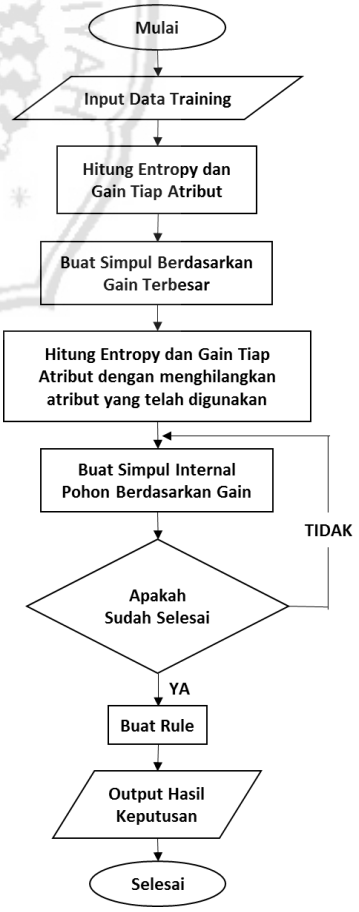
Algoritma yang digunakan yakni Decision Tree . Decision tree merupakan salah satu teknik yang dapat digunakan untuk melakukan klasifikasi terhadap sekumpulan objek atau record. Teknik ini terdiri dari kumpulan decision node, dihubungkan oleh cabang, bergerak ke bawah dari root node sampai berakhir di leaf node. Pengembangan decision tree dimulai dari root node, berdasarkan konvensi ditempatkan di bagian atas diagram decision tree, semua atribut dievaluasi pada decision node, dengan tiap outcome yang mungkin menghasilkan cabang. Tiap cabang dapat baik ke decision node yang lain ataupun ke leaf node(Witten dan Frank. 2005).

c. Program yang digunakan

Program yang digunakan yakni Weka. Walaupun kekuatan Weka terletak pada algoritma yang makin lengkap dan canggih, kesuksesan *data mining* tetap terletak pada faktor pengetahuan manusia implementornya. Tugas pengumpulan data yang berkualitas tinggi dan pengetahuan pemodelan dan penggunaan algoritma yang tepat diperlukan untuk menjamin keakuratan formulasi yang diharapkan.

2. Desain Sistem

Secara umum langkah penelitian dapat dilihat dari flowchart berikut (Iwan, Rosiyah dan Rahmanita.2018) :



3. Perhitungan Algoritma C 4.5

a. Data training

Data yang digunakan yaitu 10000 record, yang mana dalam pemilihan data dilakukan secara acak(random).

b. Pilih atribut sebagai akar

Untuk memilih atribut akar, didasarkan pada nilai Gain tertinggi dari atribut-atribut yang ada. Sedangkan untuk mendapatkan nilai Gain, harus ditentukan terlebih dahulu nilai Entropy.

c. Buat cabang untuk tiap-tiap nilai

Setelah semua akar dan cabang telah dihitung dan semua kasus pada cabang memiliki kelas yang sama maka langkah selanjutnya adalah merancang pohon keputususan (Tree).

4. Perhitungan Klasifikasi dan Kriteria

Kriteria	TP	FP	TN	FN	Total
	2207	6	7781	6	10000

Tabel diatas adalah hasil perhitungan kriteria yang terdapat dari data hasil klasifikasi.

Hasil Dari Prediksi :

- Nilai Klasifikasi [Ya] dan Kriteria [Ya] = True Positive / TP
- Nilai Klasifikasi [Tidak] dan Kriteria [Tidak] = True Negative / TN

Accuracy	99.72887483
Precision	99.88

Accuracy = (TP+TN)/(TP+TN+FP+FN) x100

= (2207+7781/ (2207+7781+6+6) x100

= 99.72887483

Presisi = TP/(TP+FP) x100

= 2207/(2207+6) x100

= 99.88

IV. HASIL DAN PEMBAHASAN

1. Analisis Data

Pada penelitian ini peneliti menggunakan dataset KDD CUP 1999. Dataset ini digunakan karena terdapat beberapa penelitian yang menggunakan data tersebut, sehingga membuktikan bahwa data KDD CUP 1999 layak unttuk digunakan. akan tetapi data yang digunakan pada pengujian menggunakan software Weka Eksplore hanya 10000 data yang di pilih secara acak. serangan-serangan tersebut diklasifikasikan berdasarkan sasaran dan tujuan serangan menjadi lima kelas kategori, yakni : DoS, Probe, U2R, dan U2L, dan Normal.

2. Lingkungan Hasil

Proses uji coba percobaan dataset KDD CUP 1999 dalam proses klasifikasi serangan ini dilakukan dalam lingkungan / platform Windows, di mana Sistem Operasi yang digunakan adalah Microsoft Windows 7.

3. Perangkat Keras

Spesifikasi perangkat keras komputer yang digunakan dalam proses uji coba percobaan dataset KDD CUP 1999 dalam proses klasifikasi serangan ini adalah sebagai berikut :

a. Prosesor Intel Core i3 2310M 2,1 Ghz (3 Core).
b. Memory (RAM) 2 GB.
c. Hard Disk 250 GB.
d. VGA Nvidia 525M .
e. Monitor LCD 14” dengan resolusi layar 1366 x 768 pixel.

4. Perangkat Lunak

Perangkat lunak yang digunakan dalam percobaan ini adalah Weka versi 3.9 dengan Java SDK (Standard Development Kit) versi 1.8.0_152. Perangkat lunak Weka digunakan karena sangat membantu saat proses klasifikasi serangan pada data set KDD CUP 1999 karena telah tersedia algoritma Decision Tree J48 dalam bentuk bahasa Java yang siap pakai sehingga memudahkan proses percobaan. Serta Weka juga mampu menampilkan keluaran dalam bentuk visual sehingga dapat dimanfaatkan oleh pengguna dalam analisis proses klasifikasi.

5. **Pengujian Algoritma C 4.5 Dengan Software WEKA**
   Setelah dataset testing selesai diolah, tahap selanjutnya adalah pengujian menggunakan software Weka Eksplore.

V. KESIMPULAN DAN SARAN

1. Kesimpulan

Berdasarkan hasil-hasil analisis dan percobaan yang dilakukan pada bab sebelumnya, maka kesimpulan yang dapat diambil adalah sebagai berikut :

a. Dari hasil pengujian menggunakan data testing dengan opsi percentage split, pola yang dibentuk memiliki tingkat presisi 99.8%.
b. Tingkat akurasi sebesar 99.76%. yang mana hal tersebut dapat dijadikan acuan sebagai tingkat kelayakan algoritma Decision tree sebagai algoritma yang tepat untuk meng klasifikasi data serangan.

2. Saran

a. Sebaiknya pelajari lagi Statistik untuk bisa benar-benar mendukung penguasaan ilmu-ilmu Data Mining, dan Decision Tree pada khususnya lebih banyak mencoba dengan berbagai macam model data dan kasus.
b. Mencoba untuk menggunakan software selain Weka dalam menganalisa data dan mencoba menggunakan metode lain selain algoritma C 4.5.
c. Menggunakan jumlah record dan atribute yang lebih banyak dalam pemrosesan data.

Daftar Pustaka

1. Basuki, Achmad dan Syarif, Iwan. 2003. "Modul Ajar Decision Tree", Surabaya : PENS-ITS.
2. Budiman,Ade Surya.,Anty Adhi Parandani.2018." Uji Akurasi Klasifikasi Dan Validasi Data Pada Penggunaan Metode Membership Function Dan Algoritma C4.5 Dalam Penilaian Penerima Beasiswa". Jakarta : AMIK BSI.
3. Fayyad, U. (1996). Advances in Knowledge Discovery and Data Mining. MIT Press.
4. Hermanto,Bambang.,Azhari SN, Fajri Profesio Putra. 2017" Analisis Perbandingan Algoritma ID3 Dan C4.5 Untuk Klasifikasi Penerima Hibah Pemasangan Air Minum Pada PDAM Kabupaten Kendal".Lampung : Universitas Lampung.
5. J. E. Gewehr, M. Szugat, and R. Zimmer.2007."BioWeka—Extending theWeka
6. Khaerani,Izza, dan Lekso Budi Handoko. 2018. " Implementasi Dan Analisa Hasil Data Mining Untuk Klasifikasi Serangan Pada Intrusion Detection System (Ids) Dengan Algoritma C4.5".Semarang: Universitas Dian Nuswantoro.
7. Larose, D. T. (2005). Discovering Knowledge in Data: An Introduction to Data mining. John Willey and Sons, Inc.
8. RHRDNWK.2016. Contoh Perhitungan Decision Tree dengan Algoritma C45. [online], (https://www.ilmuskripsi.com/2016/05/contoh-perhitungan-decision-tree-dengan.html).
9. Santosa,Iwan., Hammimatur Rosiyah, and Eza Rahmanita.2018. Implementasi Algoritma Decision Tree C.45 Untuk Diagnosa Penyakit Tubercolusis (Tb). Madura : Universitas Trunojoyo.
10. Silberschatz, A. and Tuzhilin, A. 1995. On Subjective Measures of lnterestingness in Knowledge Discovery. In Proceedings of KDD-95: First International Conference on Knowledge Discovery and Data Mining. USA : New York University.
11. S.J. Stolfo, W. Fan, W. Lee, A. Prodromidis, and P. Chan.1999.Cost-based Modeling for Fraud and Intrusion Detection: Results from the JAM Project. New York : Columbia University.
12. Tavallaee, Mahbod., Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani.2009. A Detailed Analysis Of The Kdd Cup 99 Data Set. Proceeding Of the IEEE Symposium on Computational Intelligence.
13. Witten, I. dan Eibe Frank. 2006. Data mining: Practical machine learning

tools and Techniques. USA : San Francisco.
14. Witten, I. dan Eibe Frank. 2005. Data mining: Practical machine learning tools and Techniques. USA : San Francisco.
15. Witten, I., Eibe Frank, Len Trigg, Mark Hall, Geoffrey Holmes, and Sally Jo Cunningham. 1999. Weka: Practical machine learning tools and techniques with java implementations.
16. Witten, I., Eibe Frank, Mark A. Hall. 2016. Data Mining : Practical Machine Learning Tools and Techniques".Morgan Kaufmann: Fourth Edition.