BAB 3
METODE PENELITIAN

Metode yang digunakan pada penelitian ini yaitu klasifikasi dengan menggunakan algoritma C4.5 dan membuat pohon keputusan (*decision tree*) untuk menentukan tingkat akurasi dalam mengklasifikasikan serangan pada dataset. Data *log event, packet capture*, dan data lainnya yang menggambarkan serangan pada *intrusion detection system* dikumpulkan dari sumber sumber yang sesuai. Dataset NSL-KDD yang diperoleh dari *Kaggle.com*, dipilih untuk digunakan dalam penelitian ini. Data ini memiliki 41 atribut dan berisi informasi tentang serangan pada sistem IDS.

3.1 BAHAN PENELITIAN

Data yang digunakan dalam penelitian ini yaitu data sekunder NSL-KDD *(Network Security Layer – Knowledge Discovery in Database)*. Data ini diambil dari *Kaggle.com* dan merupakan modifikasi dari dataset KDD Cup'99, yang berasal dari University of California, Irvine (UCI). Para peneliti di Institut Keamanan Siber Kanada (CIC) di Universitas New Brunswick (UNB) membuat kumpulan dataset NSL-KDD pada tahun 2009 dan telah mengalami beberapa pembaruan terbaru yang dirilis pada tahun 2019. Mereka mengambil data KDD Cup'99 dan mengatasi kekurangannya, termasuk redundasi kelas yang tidak seimbang, sehingga menghasilkan kumpulan data yang lebih baik untuk digunakan. Dataset ini digunakan untuk mengklasifikasikan serangan pada *intrusion detection system*. Ada beberapa alasan mengapa dataset NSL-KDD digunakan untuk algoritma pohon keputusan C4.5 dalam penelitian *intrusion detection system*:

1. Kesesuaian untuk algoritma C4.5

   Pohon keputusan C4.5 baik digunakan untuk pengambilan keputusan yang selaras dengan karakteristik kumpulan dataset NSL-KDD. Format terstruktur yang berisi berbagai fitur, menyediakan pola yang sesuai dengan algoritma C4.5 untuk membangun aturan keputusan.

2. Format Standar

   Kumpulan dataset NSL-KDD memiliki format yang terstruktur dengan baik dan terstandarisasi, sehingga mudah digunakan dengan berbagai alat dan algoritma pembelajaran mesin.

3. Data berlabel

   Kumpulan dataset NSL-KDD diberi label, artinya setiap rekaman koneksi jaringan dikategorikan sebagai jenis normal atau jenis serangan. Data berlabel ini sangat penting untuk melatih model pembelajaran mesin untuk mengidentifikasi serangan.

Dataset NSL-KDD dan algoritma C4.5 membentuk kombinasi yang baik untuk penelitian *intrusion detection*. Kesesuaian kumpulan data untuk pohon keputusan, kemampuan interpretasi, dan kemampuan pemilihan fitur menjadikan pilihan yang sangat baik untuk mengevaluasi model IDS berbasis C4.5. Dengan pemrosesan awal data yang cermat dan penyesuaian parameter, C4.5 dapat secara efektif mengekstrak pola dari kumpulan dataset NSL-KDD dan membangun aturan keputusan yang kuat untuk mengidentifikasi serangan. Dataset NSL-KDD memiliki atribut yang menunjukkan perilaku dari setiap koneksi jaringan. Jenis layanan, protokol, jumlah byte yang dikirim, dan informasi lainnya dapat ditemukan di atribut ini. Data ini mencoba mensimulasikan skenario serangan jaringan komputer dengan menggabungkan berbagai macam serangan. Variasi serangan dan koneksi jaringan normal disimpan dalam data ini yang dapat digunakan sebagai referensi untuk deteksi serangan. Dalam dataset NSL-KDD setiap koneksi termasuk dalam salah satu dari dua kategori koneksi normal atau mencurigakan. Data tersebut mencakup jenis serangan yang dibagi menjadi lima kategori, yaitu:

1. Normal adalah lalu lintas jaringan yang normal dan tidak menunjukkan perilaku anormal.
2. *Denial of Services (DoS)* adalah serangan komputer dengan tujuan untuk menghentikan dan mengganggu layanan yang telah disediakan.
3. *Probe* adalah serangan yang dimaksudkan untuk mengumpulkan data dari jaringan atau sistem komputer seperti *scanning port* yang mencoba menemukan kelemahan sistem yang ada.

4. *User to Root (U2R)* adalah serangan dengan tujuan untuk meningkatkan akses user menjadi super user ketika pelaku mempunyai akun dalam sistem pada level user.
5. *Remote to Local (R2L)* adalah serangan jarak jauh ke lokasi dengan tujuan untuk mengakses sistem dalam jaringan menggunakan komputer lain.

Dataset NSL-KDD memiliki 41 fitur yang dapat digunakan untuk klasifikasi serangan pada IDS. Data ini bersumber dari *www.researchgate.net* yang ditunjukkan pada tabel 3.1

Tabel 3.1 41 Fitur Dataset NSL-KDD

No	Nama Atribut	Deskripsi
1	*Duration*	Panjang waktu
2	*Protocol type*	Penggunaan protokol koneksi seperti TCP, UDP, dan ICMP
3	*Service*	Layanan tujuan yang digunakan dalam jaringan
4	*Src_byte*	Jumlah *byte* data yang dikirim dari satu sumber ke tujuan dalam satu koneksi
5	*Dst_byte*	Jumlah *byte* data yang dikirim dari tujuan untuk sumber di satu koneksi
6	*Flag*	Status koneksi (normal atau *error*)
7	*Land*	Variabel ini bernilai 1 jika alamat IP sumber dan *port* tujuan sama. Jika tidak, variabel ini bernilai 0
8	*Wrong_fragment*	Jumlah nomor untuk fragment yang salah dalam koneksi
9	*Urgent*	Jumlah paket yang diperlukan untuk koneksi
10	*Hot*	Banyaknya indikator konten panas seperti: mengakses sistem direktori, membuat dan menjalankan program

No	Nama Atribut	Deskripsi
11	*Num_failed_logins*	Menghitung jumlah upaya login yang tidak berhasil
12	*Logged_in*	Login status ditandai dengan 1 berhasil dan 0 sebaliknya
13	*Num_compromised*	Nomor kondisi yang dapat dipromosikan
14	*Root_shell*	Jika nilainya 1, *shell root* dapat diakses, dan jika 0 sebaliknya
15	*Su_attempted*	Jika nilainya 1, percobaan perintah *su root* dapat digunakan, dan jika 0 sebaliknya
16	*Num_root*	Mengakses *root* yang signifikan atau jumlah kegiatan yang dilakukan sebagai *root* dalam koneksi
17	*Num_file_creations*	Jumlah operasi yang dilakukan untuk membuat file dalam satu koneksi
18	*Num_shells*	Banyaknya jumlah *shell prompt*
19	*Num_access_shells*	Nomor yang terlibat dalam operasi akses kontrol file
20	*Num_outbound_cmds*	Banyaknya perintah outbound yang ada dalam sesi ftp
21	*Is_hot_login*	Jika nilainya 1, login termasuk daftar *hot,* tetapi jika nilainya 0 maka tidak
22	*Is_guest_login*	Jika nilainya 1 maka login *guest*, tetapi jika nilainya 0 maka tidak
23	*Count*	Jumlah koneksi ke *host* tujuan yang sama dalam dua detik sebelumnya
24	*Serror_rate*	Persentase koneksi pada *flag* seperti s0, s1, s2 atau s3 yang dikumpulkan ke dalam *count*

No	Nama Atribut	Deskripsi
25	*Rerror_rate*	Persentase koneksi yang telah diaktifkan pada koneksi *flag* REJ yang dikumpulkan ke dalam *count*
26	*Same_srv_rate*	Persentase koneksi ke *service* yang sama, yang dikumpulkan pada koneksi *count*
27	*Diff_srv_rate*	Persentase koneksi ke berbagai *service* yang dikumpulkan pada koneksi *count*
28	*Srv_count*	Jumlah koneksi ke *service* yang sama dalam 2 detik terakhir
29	*Srv_serror_rate*	Persentase koneksi yang terhubung diantara koneksi *flag* s0, s1, s2, dan s3 yang dikumpulkan pada *srv_count*
30	*Srv_rerror_rate*	Jumlah koneksi yang telah mengaktifkan *flag* REJ, dihitung diantara koneksi *srv_count*
31	*Srv_diff_host_rate*	Persentase koneksi ke berbagai mesin tujuan dan dikumpulkan diantara koneksi *srv_count*
32	*Dst_host_count*	Nomor yang terhubung ke tujuan *host* IP *address*
33	*Dst_host_srv_count*	Nomor koneksi dengan *port* yang sama
34	*Dst_host_same_srv_count*	Jumlah koneksi ke layanan yang sama, dikumpulkan diantara koneksi *dst_host_count*
35	*Dst_host_diff_srv_count*	Persentase koneksi ke berbagai *services*, dikumpulkan diantara koneksi *dst_host_count*

No	Nama Atribut	Deskripsi
36	*Dst_host_same_src_port_rate*	Jumlah koneksi ke sumber *port* yang sama, yang dikumpulkan diantara koneksi *dst_host_srv_count*
37	*Dst_host_srv_diff_host_rate*	Jumlah koneksi ke mesin tujuan yang berbeda, dan dikumpulkan diantara koneksi *dst_host_srv_count*
38	*Dst_host_serror_rate*	Jumlah koneksi yang telah diaktifkan *flag* s0, s1, s2, dan s3 dikumpulkan diantara koneksi *dst_host_count*
39	*Dst_host_srv_serror_rate*	Jumlah koneksi yang telah diaktifkan *flag* s0, s1, s2, dan s3 dikumpulkan diantara koneksi *dst_host_srv_count*
40	*Dst_host_rerror_rate*	Jumlah koneksi yang telah diaktifkan *flag* REJ, yang dikumpulkan diantara koneksi *dst_host_count*
41	*Dst_host_srv_rerror_rate*	Jumlah koneksi yang telah diaktifkan diantara koneksi *dst_host_srv_count*

3.2 ALAT PENELITIAN

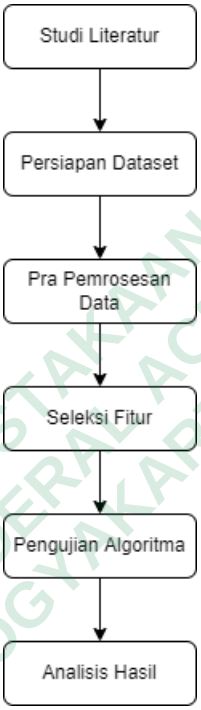
Untuk mendukung pelaksanaan penelitian, diperlukan alat-alat berupa piranti-piranti yang berguna untuk kebutuhan analisis dan keberhasilan penelitian untuk memproses atau memperlakukan bahan penelitian. Komputer dengan spesifikasi cukup digunakan dalam penelitian ini untuk menjalankan sistem operasi serta software yang dapat terhubung ke internet. Berikut adalah sistem operasi dan aplikasi yang penulis gunakan meliputi:

1. Sistem Operasi: Windows 10
2. Processor: Intel Celeron tipe N4000 CPU@2.8 Ghz
3. RAM: 4 GB
4. Storage: SSD 512 GB

5. Software: Weka, Google Colab

3.3 JALAN PENELITIAN

Jalan penelitian mencakup penjelasan lengkap dan mendalam tentang proses yang diambil selama pelaksanaan penelitian dan pengumpulan data. Jalan penelitian diperlihatkan pada gambar 3.1



Gambar 3.1 Jalan Penelitian

Berikut adalah penjelasan tentang alur jalannya penelitian:

1. Studi Literatur

   Dalam tahap ini dilakukan literatur *review* untuk mendapatkan informasi dan acuan dari buku atau jurnal penelitian yang relevan mengenai implementasi *data mining* untuk klasifikasi serangan pada sistem deteksi intrusi.

2. Persiapan dataset

   Pengumpulan data *intrusion detection system* yang akan digunakan untuk penelitian. Data ini berupa log kegiatan yang terjadi dalam jangka waktu tertentu pada sistem jaringan. Dataset yang digunakan adalah NSL-KDD yang berupa file CSV yang telah berlabel.

3. Pra Pemrosesan Data

   Setelah data dikumpulkan, proses *preprocessing* harus dilakukan untuk menghilangkan data yang tidak diperlukan atau tidak valid, mengubah format data. Setelah itu, data difilter dengan mengubah atribut yang akan digunakan.

4. Seleksi Fitur

   Setelah *preprocessing,* dilakukan pemilihan fitur yang relevan untuk klasifikasi serangan. Fitur ini biasanya disebut atribut dan merupakan input yang digunakan untuk algoritma *data mining*. Dalam penelitian sebelumnya, atribut seperti *protocol type, service, flag, src bytes,* dan *dst bytes* adalah atribut yang berkaitan dengan klasifikasi serangan pada *intrusion detection system* yang sering digunakan.

5. Pengujian Algoritma C4.5

   Algoritma ini merupakan pembagian kelas yang menggunakan pohon keputusan untuk mengklasifikasikan data ke dalam kelas tertentu. Setelah data diproses, maka data akan dibuat menjadi input. Algoritma C4.5 akan membuat pohon keputusan untuk mengklasifikasikan serangan pada *intrusion detection system*. Setelah pohon keputusan dibuat, maka perlu dilakukan pengujian akurasi. Akurasi algoritma C4.5 dapat dihitung dengan menggunakan persamaan *precision, recall,* dan *accuracy.*

6. Analisis Hasil

   Setelah pengujian selesai, maka perlu dilakukan penilaian dan penyelesaian tentang hasil pengujian. Penilaian terdiri atas bagaimana kinerja dan tingkat akurasi klasifikasi dengan algoritma C4.5 pada *intrusion detection system*.