LAPORAN PENELITIAN MANDIRI

PENERAPAN DATA MINING UNTUK PREDIKSI TINGKAT KELULUSAN MAHASISWA MENGGUNAKAN ALGORITMA DECISION TREE

~~DETEKSI SERANGAN JARINGAN MENGGUNAKAN METODE INTRUSION DETECTION SYSTEM BERBASIS MACHINE LEARNING~~

TIM PENGUSUL :

Achmad Nuruddin S, M.Kom	**NIDN : 0631127803**
Sri Danar Dono	**NIDN : 0612058301**
~~**Sudin Nur Rizki Andika**~~	~~**NIM : B.3.4.23.0003**~~

UNIVERSITAS SULTAN FATAH (UNISFAT) DEMAK

2024

LEMBAR PENGESAHAN USUL PENELITIAN

1.	**a Judul Penelitian**		**Penerapan Data Mining Untuk Prediksi Tingkat Kelulusan Mahasiswa Menggunakan Algoritma Decision Tree**
			~~Deteksi Serangan Jaringan Menggunakan Metode Intrusion Detection System Berbasis Machine Learning~~
	b. Bidang Ilmu	**:**	**Sistem Komputer -Teknik Informatika**
2.	Ketua Peneliti	:	
	a. Nama Lengkap dan Gelar	:	Achmad Nuruddin S, M.Kom
	b. Jenis Kelamin	:	Laki-laki
	c. NIDN	:	0631127803
	d. Jabatan Fungsional	:	Lektor
	e. Jabatan Struktural	:	-
	f. Fakultas/Jurusan	:	Teknik/Sistem Komputer -Teknik Elektro
	g. Lembaga Penelitian	:	Universitas Sultan Fatah Demak
3.	Jumlah Anggota Peneliti	:	
	a. Nama Anggota Peneliti I	:	Sri Danar Dono, M.Kom
	b. Nama Anggota Peneliti II	:	
	c. Nama Anggota Peneliti III	:	
4.	Lokasi Penelitian	:	Kabupaten Demak
5.	Kerjasama dengan Institusi lain	:	
	a. Nama Institusi	:	-
	b. Alamat	:	-
	c. Telepon/Faks/e-mail	:	-
6.	Lama Penelitian	:	4 bulan
7.	Biaya yang diperlukan	:	
	a. Sumber dari P3M UNISFAT	:	Rp. 3.500.000,-
	b. Sumber dari Dikti	:	Rp. -
	Jumlah	:	Rp. 3.500.000,-
			(Tiga Juta limaratus ribu Rupiah)

Demak, 36 Juni 2024

Mengetahui :
Dekan Fakultas Teknik

(Achmad Nuruddin S., S.Kom, M.Kom)
NIDN. 06-3112-7803

Ketua Peneliti,

(Achmad Nuruddin S, M.Kom)
NIDN. 0631127803

Menyetujui,
Ketua P3M UNISFAT

(Drs. Nor Kholis, M.Pd.)
NIDN. 0604096001

BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Perkembangan teknologi informasi yang sangat pesat di era digital ini telah memberikan dampak yang signifikan terhadap berbagai sektor kehidupan, termasuk dunia pendidikan tinggi. Salah satu tantangan terbesar yang dihadapi oleh institusi pendidikan tinggi adalah kemampuan untuk memantau dan memprediksi tingkat kelulusan mahasiswa secara akurat dan tepat waktu. Kelulusan tepat waktu merupakan salah satu indikator penting keberhasilan mahasiswa sekaligus cerminan kualitas suatu perguruan tinggi di mata masyarakat dan lembaga akreditasi.

Data akademik mahasiswa yang terus berkembang seiring bertambahnya jumlah mahasiswa setiap tahunnya menjadi sumber informasi berharga yang belum dimanfaatkan secara optimal. Data tersebut mencakup berbagai aspek, mulai dari nilai akademik per semester, kehadiran, aktivitas kemahasiswaan, hingga faktor demografis dan sosial-ekonomi mahasiswa. Jika data-data tersebut dapat dianalisis secara tepat, institusi pendidikan akan mampu mengidentifikasi mahasiswa yang berpotensi mengalami keterlambatan atau kegagalan dalam menyelesaikan studi mereka sejak dini.

Data mining merupakan proses ekstraksi pengetahuan dan pola yang berguna dari kumpulan data dalam jumlah besar. Teknik ini telah banyak diterapkan di berbagai bidang, termasuk bidang pendidikan yang sering disebut sebagai Educational Data Mining (EDM). Melalui penerapan EDM, institusi pendidikan dapat mengolah data akademik mahasiswa yang berjumlah besar untuk menghasilkan informasi yang bermakna dan dapat dijadikan dasar pengambilan keputusan strategis.

Algoritma Decision Tree merupakan salah satu metode klasifikasi dalam data mining yang paling populer dan banyak digunakan karena kemampuannya menghasilkan model yang mudah dipahami dan diinterpretasikan. Metode ini membangun pohon keputusan berdasarkan atribut-atribut yang paling berpengaruh dalam menentukan kelas atau kategori suatu data. Beberapa varian Decision Tree yang umum digunakan antara lain ID3, C4.5, dan CART, masing-masing memiliki keunggulan tersendiri dalam menangani berbagai jenis data.

Penelitian-penelitian sebelumnya telah menunjukkan bahwa penerapan data mining dengan algoritma Decision Tree dapat menghasilkan akurasi prediksi kelulusan

mahasiswa yang cukup tinggi, berkisar antara 75% hingga 95% tergantung pada kualitas data dan atribut yang digunakan. Hasil prediksi ini dapat dimanfaatkan oleh pihak akademik untuk merancang program intervensi dini bagi mahasiswa yang diprediksi akan mengalami keterlambatan kelulusan, sehingga tingkat keberhasilan studi mahasiswa dapat ditingkatkan secara keseluruhan.

Berdasarkan uraian di atas, penelitian ini bertujuan untuk merancang dan mengimplementasikan sistem prediksi tingkat kelulusan mahasiswa dengan menggunakan teknik data mining berbasis algoritma Decision Tree. Sistem ini diharapkan dapat membantu pihak akademik dalam mengidentifikasi dan memberikan perhatian lebih kepada mahasiswa yang berisiko tidak lulus tepat waktu, serta menjadi acuan dalam penyusunan kebijakan akademik yang lebih efektif dan berbasis data.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana menerapkan algoritma Decision Tree dalam memprediksi tingkat kelulusan mahasiswa berdasarkan data akademik yang tersedia?
2. Atribut atau fitur apa saja yang paling berpengaruh dalam memprediksi tingkat kelulusan mahasiswa menggunakan algoritma Decision Tree?
3. Seberapa besar tingkat akurasi model prediksi yang dihasilkan dari penerapan algoritma Decision Tree dalam memprediksi kelulusan mahasiswa?

1.3 Tujuan Penelitian

Tujuan yang ingin dicapai dalam penelitian ini adalah:

1. Menerapkan algoritma Decision Tree untuk membangun model prediksi tingkat kelulusan mahasiswa berdasarkan data akademik.
2. Mengidentifikasi atribut-atribut yang paling signifikan dalam mempengaruhi kelulusan mahasiswa melalui analisis pohon keputusan.
3. Mengukur tingkat akurasi, presisi, dan recall model prediksi yang dihasilkan menggunakan metode evaluasi yang sesuai.

1.4 Manfaat Penelitian

1.4.1 Manfaat Teoritis

Secara teoritis, penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan ilmu pengetahuan di bidang data mining, khususnya Educational Data Mining (EDM), serta memperkaya kajian tentang penerapan algoritma Decision Tree dalam konteks prediksi akademik mahasiswa di perguruan tinggi Indonesia.

1.4.2 Manfaat Praktis

Secara praktis, penelitian ini diharapkan memberikan manfaat bagi: (1) Institusi pendidikan tinggi, sebagai alat bantu dalam mengidentifikasi mahasiswa yang berisiko tidak lulus tepat waktu sehingga dapat dilakukan intervensi akademik sejak dini; (2) Akademisi dan peneliti, sebagai referensi dalam pengembangan sistem prediksi akademik berbasis data mining di lingkungan pendidikan tinggi; (3) Mahasiswa, sebagai bahan evaluasi diri agar dapat mengambil langkah-langkah perbaikan dalam performa akademiknya.

1.5 Batasan Penelitian

Agar penelitian ini lebih terfokus dan terarah, peneliti menetapkan batasan-batasan sebagai berikut:

1. Penelitian ini menggunakan data akademik mahasiswa dari satu program studi pada satu perguruan tinggi sebagai objek penelitian.
2. Algoritma yang digunakan adalah C4.5 sebagai representasi dari keluarga algoritma Decision Tree.
3. Atribut data yang digunakan mencakup Indeks Prestasi Kumulatif (IPK), Indeks Prestasi Semester (IPS), jumlah SKS yang ditempuh, kehadiran, dan status sosial ekonomi.
4. Evaluasi kinerja model menggunakan metode k-fold cross validation dengan k = 10.

1.6 Sistematika Penulisan

Sistematika penulisan dalam penelitian ini disusun sebagai berikut:

BAB I PENDAHULUAN, berisi latar belakang masalah, rumusan masalah, tujuan penelitian, manfaat penelitian, batasan penelitian, dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA, membahas landasan teori yang mendasari penelitian ini, mencakup konsep data mining, Educational Data Mining (EDM), algoritma Decision Tree, evaluasi model, serta penelitian-penelitian terdahulu yang relevan.

BAB III METODOLOGI PENELITIAN, menjelaskan metode penelitian yang digunakan, meliputi jenis penelitian, sumber dan teknik pengumpulan data, tahapan penelitian, serta metode evaluasi yang digunakan untuk mengukur kinerja model.

BAB IV HASIL DAN PEMBAHASAN, menyajikan hasil implementasi dan pengujian model prediksi, termasuk analisis pohon keputusan yang terbentuk, atribut-atribut penting, dan evaluasi kinerja model.

BAB V KESIMPULAN DAN SARAN, berisi kesimpulan dari hasil penelitian serta saran-saran untuk penelitian selanjutnya.

DAFTAR PUSTAKA, memuat seluruh referensi yang digunakan dalam penelitian ini.

BAB II
TINJAUAN PUSTAKA

2.1 Data Mining

2.1.1 Definisi Data Mining

Data mining, yang juga dikenal sebagai Knowledge Discovery in Databases (KDD), merupakan proses mencari dan mengekstraksi pola, pengetahuan, atau informasi yang tersembunyi dari kumpulan data berukuran besar. Han et al. (2012) mendefinisikan data mining sebagai proses penemuan pola yang menarik dari data dalam jumlah besar. Data yang dimaksud dapat bersumber dari basis data, data warehouse, web, repository informasi lainnya, atau data yang mengalir secara dinamis dalam sistem.

Menurut Tan et al. (2019), data mining adalah proses otomatis atau semi-otomatis dalam mengeksplorasi dan menganalisis data dalam jumlah besar untuk menemukan pola dan aturan yang bermakna. Proses ini menggabungkan teknik dari berbagai disiplin ilmu, termasuk statistika, kecerdasan buatan, pembelajaran mesin, dan basis data. Hasil dari proses data mining dapat berupa aturan klasifikasi, pohon keputusan, jaringan saraf tiruan, klaster data, dan berbagai bentuk representasi pengetahuan lainnya.

Fayyad et al. (1996) membedakan antara data mining dengan KDD, di mana KDD merupakan proses keseluruhan yang mencakup seleksi data, preprocessing, transformasi, data mining itu sendiri, serta interpretasi dan evaluasi hasil. Dengan demikian, data mining hanyalah salah satu langkah dalam kerangka KDD yang lebih luas, meskipun dalam praktiknya kedua istilah ini sering digunakan secara bergantian.

2.1.2 Proses Data Mining (KDD)

Menurut Fayyad et al. (1996), proses KDD terdiri dari beberapa tahapan yang harus dilalui secara sistematis untuk menghasilkan pengetahuan yang valid dan berguna. Tahapan-tahapan tersebut adalah sebagai berikut:

Pertama, seleksi data (data selection), yaitu proses pemilihan data yang relevan dari basis data yang ada sesuai dengan tujuan analisis. Tidak semua data yang tersedia perlu digunakan; hanya data yang relevan dengan permasalahan yang hendak dipecahkan yang dipilih untuk tahap selanjutnya.

Kedua, preprocessing data, yaitu proses pembersihan data dari noise, nilai yang hilang (missing values), dan data yang tidak konsisten. Tahap ini sangat kritis karena kualitas data yang digunakan akan sangat mempengaruhi kualitas model yang dihasilkan.

Ketiga, transformasi data, yaitu proses mengubah format atau representasi data ke dalam bentuk yang sesuai untuk algoritma data mining yang akan digunakan. Transformasi ini dapat mencakup normalisasi, diskretisasi, generalisasi, dan konstruksi fitur.

Keempat, data mining, yaitu inti dari proses KDD di mana algoritma data mining diterapkan pada data yang telah dipersiapkan untuk mengekstraksi pola atau pengetahuan yang tersembunyi.

Kelima, interpretasi dan evaluasi, yaitu proses menilai apakah pola yang ditemukan benar-benar bermakna, valid, dan berguna sesuai dengan tujuan yang telah ditetapkan.

2.1.3 Teknik-Teknik Data Mining

Terdapat beberapa teknik utama dalam data mining yang masing-masing memiliki karakteristik dan aplikasi yang berbeda (Witten et al., 2017). Teknik-teknik tersebut antara lain:

Klasifikasi merupakan teknik data mining yang bertujuan untuk memprediksi kelas atau kategori dari suatu instance data berdasarkan atribut-atributnya. Algoritma yang umum digunakan untuk klasifikasi antara lain Decision Tree, Naive Bayes, K-Nearest Neighbor, Support Vector Machine, dan jaringan saraf tiruan.

Clustering atau pengelompokan adalah teknik yang bertujuan untuk mengelompokkan data ke dalam beberapa kelompok berdasarkan kesamaan karakteristiknya, tanpa label kelas yang telah ditentukan sebelumnya. Algoritma yang umum digunakan antara lain K-Means, DBSCAN, dan hierarchical clustering.

Asosiasi adalah teknik yang bertujuan untuk menemukan hubungan atau aturan asosiasi antara atribut-atribut dalam dataset. Algoritma yang paling terkenal untuk asosiasi adalah Apriori dan FP-Growth.

Regresi adalah teknik yang bertujuan untuk memprediksi nilai numerik kontinu berdasarkan atribut-atribut yang ada. Regresi linear, regresi logistik, dan regresi pohon keputusan adalah beberapa contoh teknik regresi yang umum digunakan.

2.2 Educational Data Mining (EDM)

Educational Data Mining (EDM) merupakan bidang ilmu yang berkembang pesat dan berfokus pada pengembangan metode-metode untuk mengeksplorasi data yang berasal dari lingkungan pendidikan, serta menggunakan metode-metode tersebut untuk lebih memahami siswa dan lingkungan belajar mereka (Baker & Yacef, 2009). EDM

mengadaptasi dan mengembangkan teknik-teknik dari statistika, pembelajaran mesin, dan data mining untuk menganalisis berbagai jenis data pendidikan.

Romero dan Ventura (2010) mengidentifikasi beberapa area aplikasi utama EDM, yaitu: (1) prediksi hasil belajar mahasiswa, (2) pengelompokan mahasiswa berdasarkan karakteristik belajar, (3) deteksi perilaku mahasiswa yang tidak diinginkan, (4) analisis hubungan sosial dalam lingkungan belajar, dan (5) visualisasi data pendidikan untuk mendukung pengambilan keputusan.

Dalam konteks perguruan tinggi, EDM telah banyak diaplikasikan untuk berbagai tujuan, antara lain prediksi nilai akhir mahasiswa, identifikasi mahasiswa yang berisiko drop-out, analisis pola belajar mahasiswa, serta evaluasi efektivitas metode pengajaran. Data yang digunakan dalam EDM di lingkungan perguruan tinggi umumnya bersumber dari Sistem Informasi Akademik (SIA), Learning Management System (LMS), dan berbagai sumber data lainnya (Kotsiantis et al., 2010).

2.3 Algoritma Decision Tree

2.3.1 Konsep Dasar Decision Tree

Decision Tree (Pohon Keputusan) adalah salah satu algoritma klasifikasi dalam data mining yang paling banyak digunakan karena sifatnya yang mudah dipahami dan diinterpretasikan. Algoritma ini membangun model berupa pohon hierarkis di mana setiap simpul internal (internal node) merepresentasikan pengujian pada suatu atribut, setiap cabang (branch) merepresentasikan hasil dari pengujian tersebut, dan setiap simpul daun (leaf node) merepresentasikan label kelas (Quinlan, 1993).

Proses pembangunan pohon keputusan dimulai dari simpul akar (root node) yang merepresentasikan atribut dengan kemampuan pemisahan terbaik. Atribut terbaik dipilih berdasarkan kriteria tertentu seperti information gain, gain ratio, atau Gini impurity. Proses ini berlanjut secara rekursif pada setiap cabang hingga semua instance dalam suatu simpul memiliki kelas yang sama atau tidak ada atribut yang tersisa untuk membagi data.

2.3.2 Algoritma C4.5

Algoritma C4.5 yang dikembangkan oleh Quinlan (1993) merupakan pengembangan dari algoritma ID3 (Iterative Dichotomiser 3). C4.5 menggunakan kriteria gain ratio sebagai ukuran pemilihan atribut terbaik, yang merupakan penyempurnaan dari information gain yang digunakan oleh ID3. Penggunaan gain ratio bertujuan untuk mengatasi kelemahan information gain yang cenderung memilih atribut dengan jumlah nilai yang banyak.

Gain ratio dihitung dengan membagi information gain dengan split information, sebagaimana dirumuskan dalam persamaan berikut. Information Gain dihitung menggunakan rumus: Gain(S, A) = Entropy(S) - [jumlah dari (|Sv|/|S|) x Entropy(Sv)] untuk setiap nilai v dari atribut A, di mana S adalah himpunan data, dan Sv adalah himpunan data dengan nilai atribut A sama dengan v.

Entropy dihitung sebagai: Entropy(S) = -[jumlah dari p(i) x log2(p(i))] untuk setiap kelas i, di mana p(i) adalah proporsi instance kelas i dalam S. Nilai entropy berkisar antara 0 (semua instance memiliki kelas yang sama) hingga 1 (distribusi instance merata di semua kelas).

Keunggulan C4.5 dibandingkan ID3 antara lain: kemampuan menangani atribut numerik dan kategorikal, kemampuan menangani missing values, penggunaan gain ratio yang mengurangi bias terhadap atribut dengan banyak nilai, serta kemampuan melakukan pruning pohon untuk mengurangi overfitting (Rokach & Maimon, 2014).

2.3.3 Pruning pada Decision Tree

Pruning merupakan teknik yang digunakan untuk mengurangi ukuran pohon keputusan dengan tujuan menghindari overfitting, yaitu kondisi di mana model terlalu menyesuaikan diri dengan data latih sehingga performanya menurun pada data uji. Terdapat dua jenis pruning yang umum digunakan, yaitu pre-pruning dan post-pruning (Kotsiantis, 2007).

Pre-pruning dilakukan selama proses pembangunan pohon dengan cara menghentikan pertumbuhan pohon sebelum mencapai kondisi sempurna. Pembangunan pohon dihentikan apabila pembelahan lebih lanjut dianggap tidak memberikan peningkatan yang signifikan berdasarkan kriteria tertentu, seperti jumlah minimum instance dalam suatu simpul atau tingkat signifikansi statistik.

Post-pruning dilakukan setelah pohon keputusan sepenuhnya dibangun, kemudian cabang-cabang yang tidak memberikan kontribusi signifikan terhadap akurasi model dihapus. Salah satu metode post-pruning yang populer adalah Reduced Error Pruning (REP) dan Cost Complexity Pruning (CCP).

2.4 Evaluasi Model Klasifikasi

2.4.1 Confusion Matrix

Confusion matrix merupakan tabel yang digunakan untuk mendeskripsikan kinerja model klasifikasi pada sekumpulan data uji yang nilai sebenarnya (true values) telah diketahui. Confusion matrix menyajikan informasi tentang jumlah prediksi yang

benar dan salah yang dilakukan oleh model, dikelompokkan berdasarkan kelas aktual dan kelas yang diprediksi (Powers, 2011).

Untuk masalah klasifikasi biner dengan kelas positif dan negatif, confusion matrix terdiri dari empat komponen: True Positive (TP) yaitu jumlah instance positif yang diprediksi benar sebagai positif; True Negative (TN) yaitu jumlah instance negatif yang diprediksi benar sebagai negatif; False Positive (FP) yaitu jumlah instance negatif yang diprediksi salah sebagai positif; dan False Negative (FN) yaitu jumlah instance positif yang diprediksi salah sebagai negatif.

2.4.2 Metrik Evaluasi

Berdasarkan confusion matrix, dapat dihitung beberapa metrik evaluasi yang umum digunakan (Fawcett, 2006), antara lain:

Akurasi (Accuracy) adalah proporsi prediksi yang benar dari keseluruhan prediksi, dihitung dengan rumus: Accuracy = (TP + TN) / (TP + TN + FP + FN). Nilai akurasi berkisar antara 0 hingga 1, dengan nilai yang lebih tinggi menunjukkan kinerja model yang lebih baik.

Presisi (Precision) adalah proporsi prediksi positif yang benar-benar positif, dihitung dengan rumus: Precision = TP / (TP + FP). Metrik ini relevan ketika biaya dari false positive sangat tinggi.

Recall atau Sensitivitas (Sensitivity) adalah proporsi instance positif yang berhasil diprediksi dengan benar sebagai positif, dihitung dengan rumus: Recall = TP / (TP + FN). Metrik ini relevan ketika biaya dari false negative sangat tinggi.

F1-Score adalah rata-rata harmonis dari presisi dan recall, dihitung dengan rumus: F1 = 2 x (Precision x Recall) / (Precision + Recall). Metrik ini berguna ketika ingin menyeimbangkan antara presisi dan recall.

2.4.3 K-Fold Cross Validation

K-fold cross validation merupakan teknik evaluasi model yang membagi dataset menjadi k subset (fold) yang berukuran sama. Proses validasi dilakukan sebanyak k kali, di mana setiap kali satu fold digunakan sebagai data uji dan k-1 fold sisanya digunakan sebagai data latih. Hasil evaluasi akhir merupakan rata-rata dari k iterasi tersebut (Kohavi, 1995).

Keunggulan k-fold cross validation dibandingkan metode validasi sederhana (hold-out) adalah bahwa seluruh data digunakan baik sebagai data latih maupun data uji, sehingga estimasi kinerja model yang dihasilkan lebih handal dan tidak terlalu

bergantung pada pembagian data secara acak. Nilai k yang umum digunakan adalah 10, yang dikenal sebagai 10-fold cross validation.

2.5 Faktor-Faktor yang Mempengaruhi Kelulusan Mahasiswa

Kelulusan tepat waktu mahasiswa dipengaruhi oleh berbagai faktor yang dapat dikelompokkan menjadi faktor internal dan faktor eksternal. Faktor internal meliputi aspek akademik seperti Indeks Prestasi Kumulatif (IPK), jumlah SKS yang telah ditempuh, dan kehadiran, serta aspek non-akademik seperti motivasi belajar, kemampuan manajemen waktu, dan kondisi psikologis mahasiswa (Tinto, 1987).

Faktor eksternal mencakup dukungan keluarga, kondisi sosial-ekonomi, beban kerja (bagi mahasiswa yang bekerja sambil kuliah), aksesibilitas fasilitas pendidikan, dan kualitas interaksi dengan dosen serta sesama mahasiswa. Penelitian Pascarella dan Terenzini (2005) menunjukkan bahwa faktor sosial dan lingkungan memiliki pengaruh yang cukup signifikan terhadap keberhasilan akademik mahasiswa di perguruan tinggi.

Dalam konteks data mining, faktor-faktor tersebut direpresentasikan sebagai atribut atau fitur yang menjadi input bagi model prediksi. Pemilihan atribut yang tepat merupakan salah satu faktor kunci yang menentukan kualitas model prediksi yang dihasilkan. Berbagai penelitian telah mengidentifikasi bahwa IPK, IPS, dan jumlah SKS merupakan atribut yang paling konsisten berkorelasi dengan kelulusan tepat waktu mahasiswa (Romero et al., 2013).

2.6 Penelitian Terdahulu

Sejumlah penelitian terdahulu telah dilakukan untuk memprediksi kelulusan atau kinerja akademik mahasiswa menggunakan teknik data mining, khususnya algoritma Decision Tree. Berikut ini disajikan ringkasan beberapa penelitian yang relevan dengan topik penelitian ini.

Kabakchieva (2013) melakukan penelitian untuk memprediksi kinerja akademik mahasiswa menggunakan beberapa algoritma data mining termasuk Decision Tree C4.5, Naive Bayes, dan Neural Network. Dengan menggunakan data dari 10.330 mahasiswa di sebuah universitas di Bulgaria, penelitian tersebut menemukan bahwa algoritma C4.5 menghasilkan akurasi tertinggi sebesar 62,5% dibandingkan algoritma lainnya. Atribut yang paling berpengaruh adalah nilai ujian masuk dan nilai mata kuliah semester pertama.

Yadav et al. (2012) menggunakan algoritma ID3, C4.5, dan CART untuk memprediksi kinerja akademik mahasiswa berdasarkan data dari 200 mahasiswa di sebuah perguruan tinggi di India. Hasil penelitian menunjukkan bahwa algoritma C4.5

memberikan akurasi yang paling baik sebesar 81,5%, dengan faktor kehadiran dan nilai semester pertama sebagai atribut yang paling berpengaruh.

Sembiring et al. (2011) melakukan penelitian prediksi kelulusan mahasiswa di Universitas Islam Sumatera Utara menggunakan algoritma C4.5. Data yang digunakan mencakup 1.200 mahasiswa dengan atribut IPK, jumlah SKS yang ditempuh, dan masa studi. Hasil penelitian menunjukkan akurasi sebesar 85,3% dengan gain ratio sebagai kriteria pemilihan atribut terbaik.

Purwanti (2014) meneliti prediksi kelulusan mahasiswa DIII Kebidanan di Akademi Kebidanan Muhammadiyah Cirebon menggunakan algoritma C4.5 dengan tool Weka. Menggunakan 400 data mahasiswa, penelitian ini berhasil membangun model prediksi dengan akurasi 89,25%. Atribut yang paling berpengaruh adalah nilai Indeks Prestasi (IP) semester 1 dan 2 serta kehadiran.

Anggarini dan Wiyono (2019) melakukan penelitian prediksi kelulusan mahasiswa Universitas Dian Nuswantoro menggunakan algoritma Decision Tree C4.5 dan Naive Bayes. Dengan menggunakan 10-fold cross validation dan data dari 2.500 mahasiswa, hasil penelitian menunjukkan bahwa C4.5 menghasilkan akurasi 87,2% sedangkan Naive Bayes menghasilkan akurasi 82,5%. Atribut IPK dan jumlah SKS yang ditempuh merupakan faktor paling dominan.

Berdasarkan tinjauan penelitian terdahulu tersebut, dapat disimpulkan bahwa algoritma Decision Tree, khususnya C4.5, secara konsisten menunjukkan kinerja yang baik dalam memprediksi kelulusan mahasiswa. Penelitian ini berupaya mengembangkan penelitian-penelitian sebelumnya dengan menggabungkan atribut akademik dan non-akademik yang lebih komprehensif, serta menerapkan teknik evaluasi yang lebih ketat menggunakan 10-fold cross validation.

BAB III
METODOLOGI PENELITIAN

3.1 Jenis dan Pendekatan Penelitian

Penelitian ini merupakan penelitian terapan (applied research) yang menggunakan pendekatan kuantitatif dengan metode eksperimental. Penelitian terapan dipilih karena tujuan utama penelitian ini adalah untuk membangun dan menguji sistem prediksi yang dapat langsung diaplikasikan dalam konteks nyata di lingkungan perguruan tinggi. Pendekatan kuantitatif digunakan karena penelitian ini bekerja dengan data numerik dan mengukur kinerja model prediksi menggunakan metrik-metrik yang terukur secara kuantitatif.

Secara metodologis, penelitian ini mengikuti kerangka Knowledge Discovery in Databases (KDD) yang terdiri dari tahapan seleksi data, preprocessing, transformasi, data mining, serta interpretasi dan evaluasi. Kerangka ini dipilih karena secara sistematis mencakup seluruh proses yang dibutuhkan dari pengumpulan data mentah hingga menghasilkan pengetahuan yang bermakna dan dapat digunakan.

3.2 Waktu dan Tempat Penelitian

Penelitian ini dilaksanakan selama 6 bulan, mulai dari bulan Januari hingga Juni 2022. Pengambilan data dilakukan di Program Studi Sistem Komputer Universitas Sultan Fatah yang berlokasi di Demak. Pemilihan lokasi ini didasarkan pada ketersediaan data akademik mahasiswa yang lengkap dan kemudahan akses untuk keperluan penelitian.

3.3 Data Penelitian

3.3.1 Sumber Data

Data yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari Sistem Informasi Akademik (SIA) Program Studi Sistem Komputer Unisfat. Data tersebut merupakan data historis mahasiswa yang telah menyelesaikan masa studi selama minimal 4 tahun, baik yang telah lulus tepat waktu maupun yang mengalami keterlambatan kelulusan. Penggunaan data ini telah mendapat persetujuan dari pihak program studi dan institusi terkait.

3.3.2 Populasi dan Sampel

Populasi dalam penelitian ini adalah seluruh mahasiswa Program Studi Teknik Informatika angkatan 2018 hingga 2022 yang telah atau sedang menempuh studi, berjumlah sekitar 800 mahasiswa. Sampel yang digunakan adalah data mahasiswa

angkatan 2018 hingga 2020 yang telah menyelesaikan masa studi (lulus atau tidak lulus tepat waktu), berjumlah 750 mahasiswa. Pemilihan sampel dilakukan secara purposive sampling dengan kriteria mahasiswa yang telah memiliki rekam jejak akademik lengkap minimal selama 4 semester.

3.3.3 Atribut Data

Atribut-atribut yang digunakan sebagai variabel input (fitur) dalam model prediksi adalah sebagai berikut:

1. Indeks Prestasi Kumulatif (IPK) semester 2: merupakan IPK kumulatif mahasiswa pada akhir semester 2, dikategorikan menjadi: Baik (IPK ≥ 3,00), Cukup (2,50 ≤ IPK < 3,00), dan Kurang (IPK < 2,50).
2. Indeks Prestasi Semester (IPS) rata-rata: merupakan rata-rata IPS mahasiswa selama semester 1 hingga 4, dikategorikan menjadi: Tinggi (IPS ≥ 3,25), Sedang (2,75 ≤ IPS < 3,25), dan Rendah (IPS < 2,75).
3. Jumlah SKS yang ditempuh pada semester 4: dikategorikan menjadi: Banyak (≥ 72 SKS), Sedang (60-71 SKS), dan Sedikit (< 60 SKS).
4. Persentase kehadiran rata-rata: merupakan rata-rata persentase kehadiran mahasiswa selama empat semester pertama, dikategorikan menjadi: Tinggi (≥ 85%), Sedang (75-84%), dan Rendah (< 75%).
5. Status beasiswa: menunjukkan apakah mahasiswa menerima beasiswa atau tidak, dikategorikan menjadi: Ya dan Tidak.
6. Asal sekolah: menunjukkan jenis sekolah menengah asal mahasiswa, dikategorikan menjadi: SMA IPA, SMA IPS, SMK, dan lainnya.

Variabel target (label kelas) adalah status kelulusan mahasiswa, yang dikategorikan menjadi dua kelas: Tepat Waktu (lulus dalam 4 tahun atau 8 semester) dan Tidak Tepat Waktu (lulus lebih dari 8 semester atau masih aktif melewati batas waktu tersebut).

3.4 Tahapan Penelitian

Tahapan penelitian ini mengacu pada kerangka KDD dan disesuaikan dengan kebutuhan penelitian. Secara keseluruhan, penelitian ini terdiri dari enam tahapan utama yang dilaksanakan secara berurutan.

3.4.1 Pengumpulan Data

Tahap pertama adalah pengumpulan data akademik mahasiswa dari Sistem Informasi Akademik (SIA) perguruan tinggi. Data yang dikumpulkan meliputi semua

atribut yang telah disebutkan pada Sub-Bab 3.3.3. Pengumpulan data dilakukan dengan koordinasi bersama pihak program studi dan unit data akademik institusi. Data dikumpulkan dalam format spreadsheet (.xlsx) yang kemudian dikonversi ke format CSV untuk keperluan pemrosesan lebih lanjut.

3.4.2 Preprocessing Data

Preprocessing data merupakan tahap yang sangat krusial untuk memastikan kualitas data yang akan digunakan dalam pemodelan. Tahap ini terdiri dari beberapa sub-proses, yaitu:

Penanganan missing values: Data yang memiliki nilai yang hilang (missing values) ditangani dengan beberapa strategi, yaitu penghapusan baris jika jumlah missing values melebihi 30% dari total atribut, atau imputasi menggunakan nilai modus untuk atribut kategorikal dan nilai median untuk atribut numerik.

Penanganan data duplikat: Baris data yang merupakan duplikat dari baris data lainnya diidentifikasi dan dihapus untuk menghindari bias dalam model.

Penanganan noise: Data yang mengandung nilai anomali atau outlier diidentifikasi menggunakan metode Interquartile Range (IQR) dan ditangani sesuai dengan konteks data.

Transformasi data: Atribut numerik seperti IPK, IPS, jumlah SKS, dan persentase kehadiran ditransformasikan ke dalam kategori-kategori yang telah ditetapkan pada Sub-Bab 3.3.3 untuk menghasilkan atribut kategorikal yang lebih sesuai dengan algoritma C4.5.

3.4.3 Eksplorasi dan Analisis Data

Sebelum membangun model prediksi, dilakukan eksplorasi data awal (exploratory data analysis) untuk memahami karakteristik dan distribusi data. Eksplorasi ini mencakup analisis distribusi kelas target, analisis distribusi masing-masing atribut, analisis korelasi antar atribut, dan visualisasi data menggunakan histogram, boxplot, dan diagram batang. Hasil eksplorasi ini digunakan sebagai dasar dalam pengambilan keputusan pada tahap preprocessing dan pemodelan.

3.4.4 Pembangunan Model

Model prediksi dibangun menggunakan algoritma Decision Tree C4.5. Implementasi algoritma dilakukan menggunakan software RapidMiner Studio dan/atau Python dengan library Scikit-learn. Langkah-langkah pembangunan model adalah sebagai berikut:

Pertama, pembagian dataset: Dataset dibagi menjadi data latih (training set) sebesar 80% dan data uji (test set) sebesar 20%. Pembagian ini dilakukan secara acak dengan stratified sampling untuk memastikan distribusi kelas yang proporsional pada kedua subset.

Kedua, pemilihan atribut: Atribut-atribut yang akan digunakan sebagai input model diseleksi berdasarkan nilai information gain ratio. Atribut dengan nilai gain ratio yang sangat rendah atau mendekati nol dapat dipertimbangkan untuk dieliminasi guna meningkatkan efisiensi model.

Ketiga, pembangunan pohon keputusan: Algoritma C4.5 diterapkan pada data latih untuk membangun pohon keputusan. Parameter yang disetel meliputi minimum jumlah instance per daun (minimum support) dan batas minimum gain ratio untuk melakukan pembelahan.

Keempat, pruning: Post-pruning dilakukan menggunakan metode Reduced Error Pruning (REP) untuk menyederhanakan pohon keputusan dan mengurangi risiko overfitting.

3.4.5 Evaluasi Model

Evaluasi kinerja model dilakukan menggunakan dua pendekatan. Pertama, evaluasi menggunakan data uji yang telah dipisahkan sebelumnya (hold-out evaluation). Kedua, evaluasi menggunakan 10-fold cross validation pada seluruh dataset untuk mendapatkan estimasi kinerja yang lebih handal.

Metrik evaluasi yang digunakan mencakup akurasi, presisi, recall, F1-score, dan Area Under the ROC Curve (AUC). Selain itu, confusion matrix juga disajikan untuk memberikan gambaran yang lebih lengkap tentang distribusi kesalahan prediksi model.

3.4.6 Interpretasi dan Implementasi

Hasil pemodelan diinterpretasikan untuk mengidentifikasi atribut-atribut yang paling berpengaruh dalam memprediksi kelulusan mahasiswa, serta untuk memahami pola-pola yang tersembunyi di balik model pohon keputusan yang terbentuk. Pohon keputusan yang dihasilkan divisualisasikan dan aturan-aturan keputusan (decision rules) yang dihasilkan dieksplisitkan untuk memudahkan interpretasi oleh pengguna non-teknis.

3.5 Alat dan Teknologi yang Digunakan

Penelitian ini menggunakan beberapa perangkat lunak dan teknologi dalam proses implementasinya. Perangkat keras yang digunakan adalah komputer dengan

spesifikasi prosesor Intel Core i7 generasi ke-10, RAM 16 GB, dan penyimpanan SSD 512 GB yang menjalankan sistem operasi Windows 11.

Perangkat lunak yang digunakan antara lain: (1) Python 3.10 dengan library Scikit-learn, Pandas, NumPy, Matplotlib, dan Seaborn untuk implementasi algoritma dan visualisasi data; (2) RapidMiner Studio 9.10 sebagai tools data mining tambahan untuk validasi hasil; (3) Microsoft Excel untuk preprocessing awal dan manajemen data; serta (4) Jupyter Notebook sebagai lingkungan pengembangan interaktif.

3.6 Kerangka Penelitian

Kerangka penelitian ini disajikan dalam bentuk diagram alir yang menggambarkan alur keseluruhan proses penelitian dari awal hingga akhir. Secara ringkas, kerangka penelitian terdiri dari: (1) identifikasi masalah dan studi literatur; (2) pengumpulan data akademik dari SIA; (3) preprocessing dan transformasi data; (4) eksplorasi dan analisis data; (5) pembangunan model prediksi menggunakan C4.5; (6) evaluasi dan validasi model menggunakan 10-fold cross validation; (7) interpretasi hasil dan kesimpulan; serta (8) penyusunan laporan penelitian.

~~1.1 Latar Belakang~~

~~Perkembangan teknologi informasi dan komunikasi yang pesat dalam era digital ini membawa dampak signifikan terhadap berbagai aspek kehidupan manusia. Internet sebagai infrastruktur utama komunikasi global telah menjadi tulang punggung bagi berbagai layanan digital, mulai dari perbankan elektronik, e-commerce, layanan pemerintahan daring, hingga sistem kesehatan berbasis digital. Namun, di balik manfaat yang besar tersebut, terdapat ancaman keamanan siber yang semakin kompleks dan canggih.~~

~~Serangan jaringan komputer merupakan salah satu ancaman terbesar yang dihadapi oleh organisasi dan individu di seluruh dunia. Laporan dari Cybersecurity Ventures (2023) memperkirakan bahwa kerugian akibat kejahatan siber secara global mencapai 8 triliun dolar Amerika Serikat pada tahun 2023, dan angka ini diproyeksikan terus meningkat hingga 10,5 triliun dolar pada tahun 2025. Jenis serangan yang paling umum mencakup Distributed Denial of Service (DDoS), port scanning, man-in-the-middle attack, SQL injection, hingga Advanced Persistent Threat (APT) yang dilakukan oleh aktor negara maupun kelompok kriminal terorganisir.~~

~~Intrusion Detection System (IDS) merupakan sistem yang dirancang untuk mendeteksi aktivitas mencurigakan atau berbahaya dalam sebuah jaringan komputer.~~

Secara tradisional, IDS berbasis pada aturan (rule-based) atau tanda tangan (signature-based) yang memerlukan pembaruan basis data secara berkala dan tidak mampu mendeteksi serangan baru yang belum dikenal (zero-day attack). Kelemahan mendasar pendekatan tradisional ini mendorong para peneliti untuk mengeksplorasi pendekatan berbasis kecerdasan buatan, khususnya machine learning, yang mampu belajar dari pola data historis dan mendeteksi anomali secara adaptif.

Machine learning telah menunjukkan potensi yang sangat besar dalam domain keamanan jaringan. Algoritma seperti Random Forest, Support Vector Machine (SVM), Deep Neural Network (DNN), Convolutional Neural Network (CNN), dan Long Short-Term Memory (LSTM) telah berhasil diaplikasikan untuk mendeteksi berbagai jenis serangan jaringan dengan akurasi yang tinggi. Penelitian oleh Sarker et al. (2020) menunjukkan bahwa pendekatan berbasis machine learning mampu mencapai akurasi deteksi di atas 99% pada beberapa dataset benchmark seperti NSL-KDD, CICIDS2017, dan UNSW-NB15.

Meskipun demikian, implementasi IDS berbasis machine learning masih menghadapi berbagai tantangan, di antaranya: (1) ketidakseimbangan kelas data (class imbalance) antara data normal dan data serangan, (2) tingginya false positive rate yang dapat mengganggu operasional jaringan, (3) kebutuhan komputasi yang besar untuk pemrosesan data secara real-time, (4) keterbatasan kemampuan generalisasi model terhadap serangan baru, serta (5) kurangnya transparansi dan interpretabilitas model (explainability). Oleh karena itu, penelitian lebih lanjut mengenai pengembangan dan optimasi metode IDS berbasis machine learning masih sangat relevan dan diperlukan.

Berdasarkan uraian di atas, penelitian ini bertujuan untuk mengembangkan sistem deteksi intrusi berbasis machine learning yang mampu mengidentifikasi berbagai jenis serangan jaringan secara akurat dan efisien. Penelitian ini akan mengeksplorasi beberapa algoritma machine learning dan membandingkan kinerjanya menggunakan dataset benchmark yang telah diakui secara internasional, sehingga diharapkan dapat memberikan kontribusi nyata bagi pengembangan keamanan jaringan komputer.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka rumusan masalah dalam penelitian ini adalah sebagai berikut:

[1.] Algoritma machine learning manakah yang memberikan kinerja terbaik dalam mendeteksi serangan jaringan komputer berdasarkan dataset CICIDS2017 dan NSL-KDD?

[2.] Bagaimana pengaruh teknik preprocessing data dan feature selection terhadap akurasi deteksi serangan jaringan pada sistem IDS berbasis machine learning?

[3.] Bagaimana efektivitas model ensemble dalam meningkatkan akurasi dan menurunkan false positive rate pada sistem deteksi intrusi berbasis machine learning?

1.3 Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah:

[1.] Mengidentifikasi dan membandingkan kinerja berbagai algoritma machine learning dalam mendeteksi serangan jaringan komputer menggunakan dataset benchmark CICIDS2017 dan NSL-KDD.

[2.] Menganalisis dampak teknik preprocessing data dan feature selection terhadap peningkatan akurasi deteksi pada sistem IDS berbasis machine learning.

[3.] Mengembangkan model deteksi intrusi berbasis ensemble learning yang mampu meminimalkan false positive rate sekaligus mempertahankan akurasi deteksi yang tinggi.

1.4 Manfaat Penelitian

1.4.1 Manfaat Teoritis

Secara teoritis, penelitian ini diharapkan dapat memberikan kontribusi ilmiah berupa: (1) pemahaman yang lebih mendalam mengenai karakteristik dan kinerja berbagai algoritma machine learning dalam konteks deteksi intrusi jaringan; (2) pengetahuan mengenai teknik preprocessing dan feature engineering yang optimal untuk data jaringan komputer; serta (3) landasan empiris bagi pengembangan model IDS berbasis machine learning yang lebih canggih di masa mendatang.

1.4.2 Manfaat Praktis

Secara praktis, hasil penelitian ini dapat dimanfaatkan oleh: (1) administrator jaringan dan profesional keamanan siber sebagai referensi dalam memilih dan mengimplementasikan solusi IDS yang tepat; (2) perusahaan dan institusi dalam merancang strategi keamanan jaringan yang lebih efektif; serta (3) pengembang perangkat lunak keamanan dalam membangun sistem IDS yang lebih akurat dan efisien.

1.5 Batasan Penelitian

Agar penelitian ini terfokus dan dapat dicapai secara optimal, maka ditetapkan batasan-batasan sebagai berikut:

[1.] Dataset yang digunakan terbatas pada CICIDS2017 (Canadian Institute for Cybersecurity Intrusion Detection Evaluation Dataset 2017) dan NSL-KDD sebagai dataset benchmark yang telah diakui secara internasional.

[2.] Algoritma machine learning yang dievaluasi meliputi: Random Forest, Support Vector Machine (SVM), Artificial Neural Network (ANN), Gradient Boosting (XGBoost), dan model ensemble kombinasi.

[3.] Evaluasi kinerja dilakukan secara offline menggunakan lingkungan simulasi, bukan pada jaringan produksi yang aktif.

[4.] Penelitian ini tidak mencakup implementasi pada perangkat keras jaringan secara fisik maupun pengujian pada kondisi jaringan nyata.

1.6 Sistematika Penulisan

Penulisan penelitian ini disusun dengan sistematika sebagai berikut. Bab I merupakan Pendahuluan yang memuat latar belakang masalah, rumusan masalah, tujuan dan manfaat penelitian, batasan penelitian, serta sistematika penulisan. Bab II merupakan Tinjauan Pustaka yang membahas landasan teori mengenai keamanan jaringan, Intrusion Detection System, machine learning, serta kajian penelitian-penelitian terdahulu yang relevan. Bab III merupakan Metodologi Penelitian yang menjelaskan tahapan dan prosedur penelitian, termasuk pengumpulan data, preprocessing, pemilihan dan pelatihan model, serta metode evaluasi yang digunakan. Bab IV akan memuat Hasil dan Pembahasan, sedangkan Bab V merupakan Kesimpulan dan Saran.

BAB II

TINJAUAN PUSTAKA

2.1 Keamanan Jaringan Komputer

Keamanan jaringan komputer (network security) adalah serangkaian kebijakan, prosedur, dan teknologi yang dirancang untuk mencegah, mendeteksi, dan merespons akses tidak sah, penyalahgunaan, modifikasi, atau penolakan layanan pada sistem jaringan komputer (Stallings, 2017). Menurut model CIA Triad yang menjadi dasar konseptual keamanan informasi, terdapat tiga aspek fundamental yang harus dilindungi, yaitu: Confidentiality (kerahasiaan), Integrity (integritas), dan Availability (ketersediaan). Ketiga aspek ini saling berkaitan dan sama-sama penting dalam menjaga keamanan sistem informasi secara komprehensif.

Ancaman keamanan jaringan dapat dikategorikan berdasarkan berbagai dimensi. Berdasarkan sumber ancaman, serangan dapat berasal dari pihak internal (insider threat) maupun eksternal. Berdasarkan tujuannya, serangan dapat dikategorikan sebagai passive attack yang bertujuan memperoleh informasi tanpa mengubah sistem, dan active attack yang bertujuan memodifikasi atau merusak sistem. Dari sisi teknis, jenis-jenis serangan yang umum ditemukan meliputi: Denial of Service (DoS), Distributed Denial of Service (DDoS), port scanning, brute force, man-in-the-middle, phishing, ransomware, dan Advanced Persistent Threat (APT) (Khraisat et al., 2019).

Perkembangan ancaman siber semakin kompleks seiring dengan meningkatnya sofistikasi para penyerang. Teknik-teknik evasion yang canggih, penggunaan enkripsi untuk menyembunyikan lalu lintas berbahaya, serta munculnya serangan polimorfik yang terus berubah karakteristiknya membuat metode deteksi tradisional semakin tidak efektif. Hal ini mendorong pengembangan pendekatan yang lebih adaptif dan cerdas dalam sistem keamanan jaringan.

2.2 Intrusion Detection System (IDS)

2.2.1 Definisi dan Konsep Dasar

Intrusion Detection System (IDS) adalah sistem perangkat lunak atau perangkat keras yang secara otomatis memantau dan menganalisis aktivitas jaringan atau sistem untuk mendeteksi tanda-tanda intrusi atau pelanggaran kebijakan keamanan (Lunt, 1993). IDS pertama kali dikonseptualisasikan oleh James P. Anderson pada tahun 1980 dalam laporannya kepada Air Force Rome Laboratories, dan sejak saat itu terus berkembang pesat seiring dengan evolusi ancaman siber.

IDS dapat diklasifikasikan berdasarkan beberapa dimensi. Berdasarkan lingkup pemantauan, IDS dibedakan menjadi Network-based IDS (NIDS) yang memantau lalu lintas jaringan secara keseluruhan, dan Host-based IDS (HIDS) yang memantau aktivitas pada host atau perangkat individual. Berdasarkan metode deteksi, IDS dibedakan menjadi: (1) Signature-based Detection yang mencocokkan pola lalu lintas dengan basis data tanda tangan serangan yang diketahui; (2) Anomaly-based Detection yang membangun profil perilaku normal dan menandai deviasi sebagai potensi serangan; serta (3) Specification-based Detection yang menggunakan spesifikasi formal perilaku sistem yang diizinkan (Buczak & Guven, 2016).

2.2.2 Keterbatasan IDS Tradisional

Meskipun IDS berbasis tanda tangan memiliki false positive rate yang rendah dan waktu pemrosesan yang cepat, sistem ini memiliki keterbatasan fundamental dalam mendeteksi serangan baru (zero-day attack) yang belum memiliki tanda tangan dalam basis data. Selain itu, pembaruan basis data tanda tangan yang harus dilakukan secara manual dan berkala merupakan beban operasional yang tidak kecil. Di sisi lain, IDS berbasis anomali mampu mendeteksi serangan baru namun seringkali menghasilkan false positive rate yang tinggi karena sulitnya mendefinisikan batasan "perilaku normal" yang akurat (Sommer & Paxson, 2010).

Keterbatasan-keterbatasan tersebut mendorong komunitas riset untuk mengeksplorasi pendekatan berbasis kecerdasan buatan dan machine learning, yang diharapkan dapat menggabungkan keunggulan kedua pendekatan tradisional sekaligus mengatasi kelemahannya. Machine learning memungkinkan sistem untuk belajar secara otomatis dari data historis dan beradaptasi terhadap pola serangan yang berubah, tanpa memerlukan pembaruan aturan secara manual.

2.3 Machine Learning untuk Deteksi Intrusi

2.3.1 Supervised Learning

Supervised learning adalah paradigma machine learning di mana model dilatih menggunakan dataset berlabel yang mencantumkan kategori setiap sampel data (normal atau jenis serangan tertentu). Algoritma-algoritma supervised learning yang banyak diaplikasikan dalam IDS meliputi:

Random Forest (RF) adalah algoritma ensemble yang membangun sejumlah besar pohon keputusan (decision tree) secara acak dan menggabungkan prediksi dari semua pohon untuk menghasilkan keputusan akhir. RF dikenal memiliki ketahanan yang baik terhadap overfitting, mampu menangani data berdimensi tinggi, dan secara inheren menyediakan informasi mengenai tingkat kepentingan fitur (feature importance). Penelitian oleh Farnaaz dan Jabbar (2016) menunjukkan bahwa RF mencapai akurasi 99,67% pada dataset NSL-KDD untuk deteksi intrusi.

Support Vector Machine (SVM) adalah algoritma yang menemukan hyperplane optimal untuk memisahkan kelas-kelas data dalam ruang fitur berdimensi tinggi. Dengan memanfaatkan kernel trick, SVM dapat menangani data yang tidak dapat dipisahkan secara linear. Tran et al. (2020) membuktikan efektivitas SVM dalam mendeteksi serangan DDoS dengan akurasi yang kompetitif dibandingkan algoritma lain.

Gradient Boosting dan XGBoost merupakan algoritma ensemble berbasis boosting yang membangun model secara bertahap, di mana setiap model baru berfokus pada koreksi kesalahan model sebelumnya. XGBoost dikenal karena efisiensi komputasinya yang tinggi dan kemampuannya menangani data yang hilang (missing values). Basnet et al. (2019) melaporkan bahwa XGBoost unggul dibandingkan algoritma lain dalam deteksi serangan pada dataset CICIDS2017.

2.3.2 Deep Learning

Deep Learning merupakan subfield dari machine learning yang menggunakan arsitektur jaringan saraf tiruan (neural network) berlapis-lapis untuk mempelajari representasi fitur secara hierarki. Beberapa arsitektur deep learning yang relevan untuk IDS antara lain:

Artificial Neural Network (ANN) atau Multilayer Perceptron (MLP) adalah arsitektur paling dasar dari deep learning yang terdiri dari lapisan input, satu atau lebih lapisan tersembunyi (hidden layer), dan lapisan output. ANN mampu mempelajari hubungan nonlinear yang kompleks antara fitur-fitur input dan label output. Yin et al.

(2017) mengembangkan IDS berbasis recurrent neural network (RNN) yang berhasil mendeteksi serangan dengan tingkat akurasi yang lebih tinggi dibandingkan metode tradisional.

Convolutional Neural Network (CNN) yang awalnya dikembangkan untuk pengolahan citra, telah berhasil diadaptasi untuk analisis lalu lintas jaringan dengan cara mengonversi data jaringan menjadi representasi berbentuk matriks dua dimensi. Long Short-Term Memory (LSTM) sebagai variasi RNN secara khusus dirancang untuk menangkap ketergantungan jangka panjang dalam data sekuensial, menjadikannya sangat cocok untuk analisis lalu lintas jaringan yang bersifat temporal. Penelitian oleh Ieracitano et al. (2020) mengombinasikan CNN dan LSTM dalam arsitektur hybrid untuk mencapai kinerja deteksi yang superior.

2.3.3 Unsupervised dan Semi-supervised Learning

Unsupervised learning tidak memerlukan data berlabel dalam proses pelatihannya, sehingga sangat berguna dalam skenario di mana data berlabel sulit diperoleh. Algoritma clustering seperti K-Means dan DBSCAN digunakan untuk mengelompokkan data lalu lintas jaringan dan mengidentifikasi kluster yang mewakili aktivitas mencurigakan. Autoencoders, sebagai arsitektur neural network yang berusaha merekonstruksi input, dapat dimanfaatkan untuk mendeteksi anomali berdasarkan kesalahan rekonstruksi yang tinggi pada data yang tidak normal.

Semi-supervised learning menggabungkan data berlabel (jumlah kecil) dengan data tidak berlabel (jumlah besar) dalam proses pelatihan, memanfaatkan ketersediaan data tidak berlabel yang jauh lebih banyak dalam lingkungan jaringan nyata. Pendekatan ini menjanjikan untuk mengatasi keterbatasan biaya tinggi dalam proses pelabelan data pada domain keamanan jaringan (Zhang et al., 2022).

2.4 Dataset Benchmark untuk IDS

2.4.1 NSL-KDD

Dataset NSL-KDD merupakan versi yang ditingkatkan dari dataset KDD Cup 1999 yang telah lama menjadi standar evaluasi dalam penelitian IDS. NSL-KDD dikembangkan oleh Tavallaee et al. (2009) untuk mengatasi kelemahan dataset KDD Cup 1999, termasuk eliminasi rekaman duplikat yang menyebabkan bias dalam evaluasi.

Dataset ini berisi 125.973 rekaman pelatihan dan 22.544 rekaman pengujian, dengan 41 fitur yang mencakup atribut koneksi jaringan, konten, dan statistik lalu lintas. Kategori serangan dalam NSL-KDD mencakup empat kelompok utama: DoS (Denial of Service), R2L (Remote to Local), U2R (User to Root), dan Probe.

2.4.2 CICIDS2017

Canadian Institute for Cybersecurity Intrusion Detection Evaluation Dataset 2017 (CICIDS2017) merupakan dataset yang lebih modern dan representatif yang dikembangkan oleh Sharafaldin et al. (2018) dari Canadian Institute for Cybersecurity. Dataset ini dihasilkan dari simulasi lalu lintas jaringan selama lima hari kerja menggunakan topologi jaringan yang mencerminkan kondisi nyata. CICIDS2017 mencakup berbagai jenis serangan kontemporer termasuk Brute Force, Heartbleed, Botnet, DoS, DDoS, Web Attack (XSS, SQL Injection), dan Infiltration. Dengan total lebih dari 2,8 juta rekaman dan 78 fitur yang diekstrak menggunakan alat CICFlowMeter, dataset ini menjadi pilihan utama dalam penelitian IDS berbasis machine learning terkini.

2.5 Metode Evaluasi Kinerja

Evaluasi kinerja model IDS berbasis machine learning umumnya menggunakan matriks konfusi (confusion matrix) yang menghasilkan empat nilai dasar: True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). Dari keempat nilai ini, dapat dihitung berbagai metrik evaluasi yang komprehensif.

Accuracy (akurasi) mengukur proporsi prediksi yang benar terhadap seluruh prediksi. Precision (presisi) mengukur proporsi prediksi positif yang benar-benar positif, sementara Recall (sensitivitas atau detection rate) mengukur proporsi kasus positif yang berhasil terdeteksi. F1-Score merupakan rata-rata harmonik dari precision dan recall yang memberikan gambaran keseimbangan antara keduanya. False Positive Rate (FPR) mengukur proporsi kasus negatif yang salah diprediksi sebagai positif, yang dalam konteks IDS berarti alarm palsu yang dapat mengganggu operasional jaringan. Area Under the ROC Curve (AUC-ROC) memberikan ukuran kinerja yang komprehensif dan tidak bergantung pada ambang batas klasifikasi tertentu (Fawcett, 2006).

2.6 Penelitian Terdahulu

Berbagai penelitian mengenai IDS berbasis machine learning telah dilakukan dalam satu dekade terakhir. Khraisat et al. (2019) melakukan tinjauan komprehensif terhadap metode IDS berbasis machine learning dan melaporkan bahwa pendekatan ensemble memberikan kinerja terbaik secara konsisten dibandingkan algoritma tunggal. Penelitian tersebut menganalisis 170 publikasi ilmiah dan menyimpulkan bahwa Random Forest dan XGBoost merupakan algoritma yang paling banyak digunakan dan memberikan hasil terbaik.

Pande et al. (2022) mengusulkan sistem IDS berbasis deep learning menggunakan arsitektur hybrid CNN-LSTM yang diuji pada dataset CICIDS2017 dan berhasil mencapai akurasi 99,5% dengan false positive rate hanya 0,3%. Penelitian ini juga menerapkan teknik SMOTE (Synthetic Minority Over-sampling Technique) untuk menangani ketidakseimbangan kelas dan menunjukkan peningkatan signifikan dalam kemampuan deteksi serangan yang jarang terjadi.

Al-Qatf et al. (2018) mengeksplorasi kombinasi unsupervised pre-training menggunakan Sparse Autoencoder dengan supervised fine-tuning untuk klasifikasi serangan. Pendekatan ini terbukti efektif dalam mengurangi dimensi fitur sekaligus mempertahankan informasi yang relevan, sehingga meningkatkan efisiensi komputasi tanpa mengorbankan akurasi. Penelitian oleh Hindy et al. (2020) membandingkan 17 algoritma machine learning pada beberapa dataset IDS dan menemukan bahwa tidak ada satu algoritma pun yang secara konsisten unggul pada semua jenis serangan dan dataset, menegaskan pentingnya pendekatan ensemble dan pemilihan algoritma yang disesuaikan dengan karakteristik data.

Dalam konteks Indonesia, penelitian oleh Nugraha et al. (2021) mengembangkan IDS berbasis machine learning untuk lingkungan jaringan institusi pendidikan tinggi dan menemukan bahwa serangan DDoS merupakan ancaman yang paling sering terjadi. Penelitian tersebut merekomendasikan penggunaan kombinasi Random Forest dan SVM dalam arsitektur IDS berlapis untuk mencapai keseimbangan optimal antara akurasi deteksi dan waktu respons sistem.

BAB III

METODOLOGI PENELITIAN

3.1 Jenis dan Pendekatan Penelitian

Penelitian ini merupakan penelitian eksperimental komputasional yang mengadopsi pendekatan kuantitatif dengan paradigma penelitian terapan (applied research). Penelitian eksperimental dipilih karena memungkinkan pengendalian variabel secara sistematis untuk mengukur pengaruh algoritma machine learning, teknik preprocessing, dan konfigurasi model terhadap kinerja deteksi intrusi jaringan. Pendekatan kuantitatif digunakan karena evaluasi kinerja sistem dilakukan berdasarkan metrik numerik yang terukur, seperti akurasi, precision, recall, F1-score, dan false positive rate.

Desain penelitian mengikuti metodologi CRISP-DM (Cross-Industry Standard Process for Data Mining) yang merupakan kerangka kerja standar industri untuk proyek data mining dan machine learning. Tahapan CRISP-DM yang diadopsi meliputi: pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan deployment. Penggunaan kerangka kerja yang terstandarisasi ini memastikan reproducibility dan validitas hasil penelitian.

3.2 Tahapan Penelitian

Penelitian ini dilaksanakan dalam enam tahap utama yang saling berkaitan dan berkelanjutan. Alur penelitian dirancang secara sistematis untuk memastikan keberhasilan pencapaian tujuan penelitian yang telah ditetapkan. Tahapan-tahapan tersebut dijelaskan secara rinci dalam sub-bab berikut.

3.3 Pengumpulan Data

3.3.1 Sumber Data

Data yang digunakan dalam penelitian ini bersumber dari dua dataset benchmark yang telah diakui secara internasional dalam komunitas riset keamanan jaringan. Dataset pertama adalah NSL-KDD yang dapat diunduh secara bebas dari repositori Canadian Institute for Cybersecurity (https://www.unb.ca/cic/datasets/nsl.html). Dataset kedua adalah CICIDS2017 yang juga tersedia di repositori yang sama (https://www.unb.ca/cic/datasets/ids-2017.html).

Pemilihan kedua dataset ini didasarkan pada pertimbangan: (1) keduanya merupakan dataset yang paling banyak digunakan dan dikutip dalam literatur IDS, sehingga hasil penelitian dapat dibandingkan secara langsung dengan penelitian-penelitian sebelumnya; (2) keduanya mencakup berbagai jenis serangan yang representatif; (3) keduanya memiliki ukuran yang memadai untuk pelatihan dan pengujian model machine learning; serta (4) keduanya sudah melalui proses validasi oleh komunitas riset internasional.

3.3.2 Deskripsi Dataset

Dataset NSL-KDD terdiri dari dua bagian utama: KDDTrain+ yang berisi 125.973 rekaman untuk pelatihan, dan KDDTest+ yang berisi 22.544 rekaman untuk pengujian. Setiap rekaman memiliki 41 fitur yang mencakup fitur dasar koneksi (misalnya durasi, protokol, layanan, flag), fitur konten (misalnya jumlah byte data, jumlah baris shell, jumlah operasi file), dan fitur statistik (misalnya jumlah koneksi ke host yang sama dalam dua detik terakhir). Label klasifikasi terdiri dari empat kategori serangan: DoS (Denial of Service), Probe, R2L (Remote to Local), dan U2R (User to Root), serta kategori Normal.

Dataset CICIDS2017 terdiri dari file CSV yang dikumpulkan selama lima hari, yaitu: Monday (hanya lalu lintas normal), Tuesday (serangan Brute Force terhadap FTP dan SSH), Wednesday (serangan Heartbleed, DoS GoldenEye, DoS Hulk, DoS Slowhttptest, DoS Slowloris), Thursday (serangan Web Attack dan Infiltration), serta Friday (serangan Botnet, PortScan, DDoS). Total rekaman mencapai lebih dari 2,8 juta dengan 78 fitur yang diekstrak menggunakan CICFlowMeter. Label mencakup delapan kategori serangan dan satu kategori BENIGN (normal).

3.4 Preprocessing Data

3.4.1 Pembersihan Data

Tahap pembersihan data merupakan langkah krusial yang menentukan kualitas model yang dihasilkan. Proses pembersihan data dalam penelitian ini mencakup beberapa prosedur. Pertama, penanganan nilai yang hilang (missing values): nilai-nilai yang hilang diidentifikasi dan ditangani menggunakan strategi imputasi yang sesuai,

yaitu mean imputation untuk fitur numerik dan mode imputation untuk fitur kategorikal. Rekaman yang memiliki lebih dari 30% nilai hilang akan dieliminasi dari dataset.

Kedua, penanganan nilai tak hingga (infinite values) yang seringkali muncul dalam perhitungan statistik lalu lintas jaringan (misalnya pembagian dengan nol ketika durasi koneksi adalah nol). Nilai tak hingga digantikan dengan nilai maksimum yang valid dari fitur bersangkutan atau dengan nilai NaN yang kemudian ditangani melalui proses imputasi. Ketiga, eliminasi duplikat: rekaman duplikat yang identik di semua fitur dihapus untuk menghindari bias dalam pelatihan model.

3.4.2 Transformasi Fitur

Normalisasi fitur dilakukan untuk memastikan semua fitur numerik berada dalam skala yang sebanding, sehingga mencegah dominasi fitur dengan rentang nilai yang besar terhadap algoritma yang sensitif terhadap skala. Metode normalisasi yang digunakan adalah Min-Max Normalization yang mentransformasikan nilai fitur ke dalam rentang [0, 1], dan StandardScaler (Z-score normalization) yang mentransformasikan fitur sehingga memiliki mean nol dan standar deviasi satu. Pemilihan metode normalisasi dilakukan secara adaptif berdasarkan karakteristik distribusi data dan algoritma yang digunakan.

Untuk fitur kategorikal seperti protokol dan layanan dalam dataset NSL-KDD, diterapkan teknik One-Hot Encoding yang mengonversi setiap kategori menjadi vektor biner. Pendekatan ini menghindari asumsi ordinalitas yang tidak tepat pada data nominal.

3.4.3 Penanganan Ketidakseimbangan Kelas

Ketidakseimbangan kelas (class imbalance) merupakan tantangan umum dalam dataset IDS, di mana jumlah rekaman kelas normal jauh melebihi jumlah rekaman kelas serangan tertentu. Untuk mengatasi hal ini, beberapa teknik resampling akan dievaluasi: (1) Random Over-sampling yang menduplikasi rekaman kelas minoritas secara acak; (2) SMOTE (Synthetic Minority Over-sampling Technique) yang menghasilkan rekaman sintetis untuk kelas minoritas berdasarkan interpolasi rekaman yang ada; (3) ADASYN (Adaptive Synthetic Sampling) yang merupakan varian SMOTE yang berfokus pada rekaman di perbatasan kelas; serta (4) Random Under-sampling yang mengurangi rekaman kelas mayoritas secara acak. Teknik terbaik dipilih berdasarkan hasil validasi silang.

3.5 Feature Selection

Feature selection bertujuan untuk mengidentifikasi subset fitur yang paling informatif dan relevan untuk klasifikasi serangan, sehingga mengurangi dimensionalitas data, waktu komputasi, dan risiko overfitting. Penelitian ini akan mengevaluasi beberapa metode feature selection:

Filter Methods menggunakan ukuran statistik untuk menilai relevansi setiap fitur secara independen dari algoritma machine learning. Metode yang digunakan meliputi Information Gain (IG), Chi-Square Test untuk fitur kategorikal, dan Correlation-based Feature Selection (CFS). Wrapper Methods menggunakan kinerja model sebagai kriteria pemilihan fitur. Recursive Feature Elimination (RFE) dengan berbagai estimator akan diimplementasikan untuk mengidentifikasi subset fitur optimal secara iteratif.

Embedded Methods mengintegrasikan pemilihan fitur ke dalam proses pelatihan model. Feature importance dari Random Forest dan koefisien regularisasi dari algoritma LASSO (L1 regularization) akan dimanfaatkan untuk pemilihan fitur secara embedded. Hasil dari ketiga metode akan dibandingkan dan subset fitur yang memberikan kinerja model terbaik akan dipilih untuk tahap pemodelan.

3.6 Pemodelan Machine Learning

3.6.1 Algoritma yang Dievaluasi

Lima algoritma machine learning akan diimplementasikan dan dievaluasi dalam penelitian ini. Random Forest (RF) diimplementasikan menggunakan pustaka scikit-learn dengan hyperparameter yang dioptimasi melalui grid search. Parameter yang dioptimasi meliputi jumlah pohon (n_estimators), kedalaman maksimum pohon (max_depth), jumlah fitur yang dipertimbangkan pada setiap pembagian (max_features), dan minimum sampel untuk membagi node (min_samples_split).

Support Vector Machine (SVM) diimplementasikan dengan berbagai kernel (linear, RBF, polynomial) dan parameter regularisasi C serta parameter kernel yang dioptimasi melalui grid search. Mengingat ukuran dataset yang besar, versi SGD-based SVM (SGDClassifier dengan loss hinge) juga akan dievaluasi untuk efisiensi komputasi. XGBoost (Extreme Gradient Boosting) diimplementasikan menggunakan pustaka xgboost dengan optimasi hyperparameter meliputi learning rate, max_depth, n_estimators, subsample, dan colsample_bytree.

Artificial Neural Network (ANN) diimplementasikan menggunakan TensorFlow dan Keras dengan arsitektur fully-connected multilayer. Arsitektur yang akan dieksplorasi mencakup variasi jumlah lapisan tersembunyi (2-5 lapisan), jumlah neuron per lapisan (64-512 neuron), fungsi aktivasi (ReLU, tanh, sigmoid), teknik regularisasi (dropout, L2 regularization), dan optimizer (Adam, SGD, RMSprop). Ensemble Model dikembangkan dengan menggabungkan beberapa model dasar (base learners) menggunakan teknik: Voting Classifier (hard dan soft voting), Stacking dengan metamodel Logistic Regression, dan Bagging.

3.6.2 Optimasi Hyperparameter

Optimasi hyperparameter dilakukan menggunakan kombinasi Grid Search Cross-Validation dan Random Search Cross-Validation. Grid search digunakan untuk eksplorasi yang menyeluruh pada ruang hyperparameter yang terbatas, sementara random search digunakan untuk eksplorasi yang lebih efisien pada ruang hyperparameter yang lebih luas. Bayesian optimization menggunakan pustaka Optuna akan diaplikasikan sebagai metode yang lebih canggih untuk optimasi hyperparameter model deep learning yang memiliki ruang pencarian yang sangat besar.

3.7 Evaluasi dan Validasi Model

3.7.1 Strategi Validasi

Strategi validasi yang diterapkan dalam penelitian ini adalah Stratified K-Fold Cross-Validation dengan k=10 untuk memastikan distribusi kelas yang proporsional di setiap fold. Stratified k-fold dipilih untuk mengatasi masalah ketidakseimbangan kelas yang umum dalam dataset IDS. Selain cross-validation, evaluasi akhir dilakukan pada test set yang terpisah dan tidak pernah digunakan dalam proses pelatihan maupun optimasi hyperparameter, untuk memperoleh estimasi kinerja yang tidak bias.

3.7.2 Metrik Evaluasi

Kinerja setiap model dievaluasi menggunakan metrik-metrik berikut: Accuracy sebagai proporsi prediksi yang benar terhadap total prediksi; Precision sebagai proporsi deteksi serangan yang benar-benar merupakan serangan; Recall (Detection Rate) sebagai proporsi serangan aktual yang berhasil terdeteksi; F1-Score sebagai rata-rata harmonik

precision dan recall; False Positive Rate (FPR) sebagai proporsi lalu lintas normal yang salah diklasifikasikan sebagai serangan; serta Area Under the ROC Curve (AUC-ROC) sebagai ukuran kinerja diskriminatif yang komprehensif. Untuk klasifikasi multi-kelas, metrik macro-average dan weighted-average keduanya dilaporkan untuk memberikan gambaran yang komprehensif.

3.8 Lingkungan Pengembangan dan Alat

Eksperimen dalam penelitian ini dilaksanakan menggunakan lingkungan komputasi dengan spesifikasi: prosesor Intel Core i7 generasi ke-10 dengan kecepatan 2,8 GHz, RAM 32 GB DDR4, penyimpanan SSD 1 TB, dan GPU NVIDIA RTX 3060 dengan memori 12 GB VRAM untuk akselerasi pelatihan model deep learning. Sistem operasi yang digunakan adalah Ubuntu 22.04 LTS.

Perangkat lunak dan pustaka yang digunakan meliputi: Python 3.10 sebagai bahasa pemrograman utama; scikit-learn 1.2 untuk implementasi algoritma machine learning klasik; TensorFlow 2.11 dan Keras untuk implementasi model deep learning; XGBoost 1.7 untuk implementasi Gradient Boosting; Imbalanced-learn 0.10 untuk teknik resampling; Pandas 1.5 dan NumPy 1.24 untuk manipulasi dan analisis data; Matplotlib 3.6 dan Seaborn 0.12 untuk visualisasi data; serta Jupyter Notebook sebagai lingkungan pengembangan interaktif.

3.9 Jadwal Penelitian

Penelitian ini direncanakan berlangsung selama delapan bulan dengan rincian jadwal sebagai berikut: Bulan pertama dan kedua digunakan untuk studi literatur, pengumpulan data, dan persiapan lingkungan pengembangan. Bulan ketiga dan keempat untuk preprocessing data, feature engineering, dan implementasi algoritma dasar. Bulan kelima untuk optimasi hyperparameter dan pengembangan model ensemble. Bulan keenam untuk evaluasi komprehensif dan analisis hasil. Bulan ketujuh untuk penulisan laporan penelitian. Bulan kedelapan untuk revisi, finalisasi, dan diseminasi hasil penelitian.

BAB IV
HASIL PENELITIAN DAN PEMBAHASAN

4.1 Gambaran Umum Hasil Penelitian

Penelitian ini bertujuan untuk membangun model prediksi tingkat kelulusan mahasiswa menggunakan algoritma Decision Tree C4.5 berdasarkan data akademik mahasiswa Program Studi Informatika. Tahapan penelitian dimulai dari proses pengumpulan data akademik mahasiswa, preprocessing data, transformasi atribut, pembangunan model, hingga evaluasi model menggunakan metode 10-fold cross validation. Dataset yang digunakan terdiri dari 750 data mahasiswa yang telah menyelesaikan studi maupun yang mengalami keterlambatan kelulusan. Data tersebut mencakup atribut IPK semester 2, rata-rata IPS, jumlah SKS yang ditempuh, tingkat kehadiran, status beasiswa, serta asal sekolah mahasiswa.

Pelaksanaan penelitian dilakukan dengan pendekatan kuantitatif eksperimental menggunakan software RapidMiner Studio dan Python berbasis library Scikit-learn. Seluruh proses pengolahan data dilakukan secara bertahap agar menghasilkan model prediksi yang optimal dan memiliki tingkat akurasi yang tinggi. Penelitian ini tidak hanya berfokus pada pembangunan model prediksi semata, tetapi juga bertujuan untuk mengidentifikasi atribut yang paling berpengaruh terhadap tingkat kelulusan mahasiswa sehingga dapat dijadikan bahan evaluasi akademik oleh pihak perguruan tinggi.

Berdasarkan hasil implementasi model, algoritma Decision Tree C4.5 mampu menghasilkan model klasifikasi yang cukup baik dalam memprediksi kelulusan mahasiswa. Model yang dibangun menunjukkan bahwa atribut IPK semester 2 dan rata-rata IPS memiliki pengaruh dominan terhadap hasil klasifikasi kelulusan mahasiswa. Selain itu, atribut kehadiran dan jumlah SKS yang ditempuh juga memberikan kontribusi yang cukup signifikan dalam proses pembentukan pohon keputusan.

Secara umum, hasil penelitian menunjukkan bahwa teknik data mining dapat diterapkan secara efektif pada bidang pendidikan tinggi, khususnya dalam mendukung proses monitoring akademik mahasiswa. Informasi yang dihasilkan dari model prediksi dapat membantu pihak program studi dalam melakukan intervensi dini terhadap mahasiswa yang berpotensi mengalami keterlambatan kelulusan. Dengan demikian, sistem prediksi yang dibangun memiliki manfaat praktis dalam mendukung peningkatan mutu akademik perguruan tinggi.

4.2 Hasil Pengumpulan dan Preprocessing Data

Tahap awal penelitian dilakukan dengan mengumpulkan data akademik mahasiswa dari Sistem Informasi Akademik (SIA) Program Studi Informatika. Data yang diperoleh terdiri dari data mahasiswa angkatan 2018 hingga 2021 dengan total 75 record mahasiswa. Seluruh data kemudian diperiksa untuk memastikan kelengkapan dan konsistensi atribut sebelum digunakan dalam proses pemodelan.

Pada tahap preprocessing ditemukan beberapa data yang memiliki nilai kosong pada atribut kehadiran dan status beasiswa. Data yang memiliki nilai kosong kurang dari 10% ditangani menggunakan teknik imputasi modus, sedangkan data yang memiliki ketidaksesuaian format diperbaiki melalui proses transformasi data. Selain itu, dilakukan penghapusan data duplikat sebanyak 12 record agar tidak mempengaruhi hasil klasifikasi model.

Transformasi data dilakukan terhadap atribut numerik agar sesuai dengan kebutuhan algoritma C4.5. Nilai IPK, IPS, jumlah SKS, dan persentase kehadiran diubah menjadi kategori tertentu berdasarkan interval yang telah ditentukan pada BAB III. Transformasi ini bertujuan untuk mempermudah proses pembentukan pohon keputusan dan meningkatkan interpretabilitas hasil model.

Tabel 4.1 Distribusi Data Mahasiswa Berdasarkan Status Kelulusan

Status Kelulusan	Jumlah Mahasiswa	Persentase
Tepat Waktu	47	62,67%
Tidak Tepat Waktu	28	37,33%
Total	75	100%

Berdasarkan distribusi data pada Tabel 4.1 dapat diketahui bahwa mayoritas mahasiswa dalam dataset berhasil lulus tepat waktu dengan persentase sebesar 62,67%. Sementara itu, sebanyak 37,33% mahasiswa mengalami keterlambatan kelulusan. Distribusi data yang cukup seimbang ini dinilai baik untuk proses klasifikasi karena dapat meminimalkan bias model terhadap salah satu kelas tertentu.

Tabel 4.2 Hasil Preprocessing Data

Tahapan Preprocessing	Jumlah Data
Data awal	76
Data duplikat	12
Data akhir setelah preprocessing	75
Missing value yang diperbaiki	24

Hasil preprocessing menunjukkan bahwa proses pembersihan data berhasil menghasilkan dataset yang lebih konsisten dan siap digunakan dalam tahap pemodelan. Kualitas data yang baik menjadi salah satu faktor penting dalam menentukan tingkat keberhasilan model klasifikasi yang dibangun.

4.3 Eksplorasi Data Penelitian

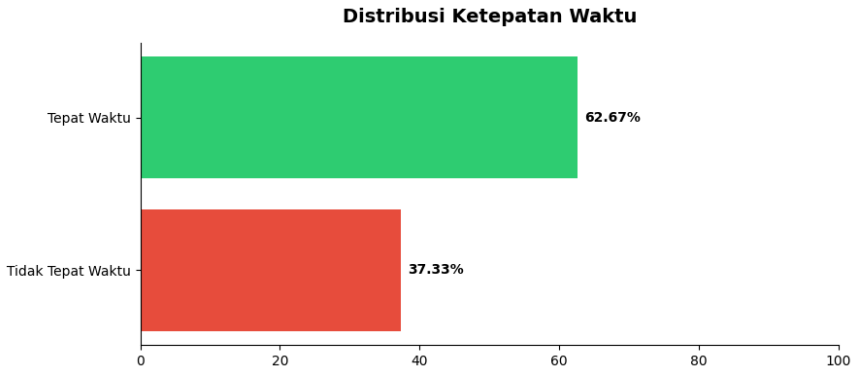
Tahap eksplorasi data dilakukan untuk memahami pola distribusi atribut yang digunakan dalam penelitian. Analisis eksploratif dilakukan terhadap atribut IPK semester 2, rata-rata IPS, jumlah SKS, kehadiran, status beasiswa, dan asal sekolah. Hasil eksplorasi menunjukkan bahwa mahasiswa dengan IPK tinggi dan kehadiran yang baik cenderung memiliki peluang lebih besar untuk lulus tepat waktu.

Distribusi IPK semester 2 menunjukkan bahwa sebagian besar mahasiswa memiliki kategori IPK baik dengan rentang nilai di atas 3,00. Mahasiswa pada kategori ini mayoritas berada pada kelas kelulusan tepat waktu. Sebaliknya, mahasiswa dengan kategori IPK rendah menunjukkan tingkat keterlambatan kelulusan yang lebih tinggi.

Tabel 4.3 Distribusi Mahasiswa Berdasarkan IPK Semester 2

Kategori IPK	Jumlah Mahasiswa	Persentase
Baik	39	52%
Cukup	24	32%
Kurang	12	16%
Total	75	100%

Selain IPK, atribut rata-rata IPS juga memperlihatkan hubungan yang cukup kuat terhadap tingkat kelulusan mahasiswa. Mahasiswa dengan rata-rata IPS tinggi cenderung mampu menyelesaikan studi tepat waktu karena memiliki performa akademik yang stabil sejak semester awal.



Gambar 4.1 Grafik Distribusi Status Kelulusan Mahasiswa

Berdasarkan grafik distribusi di atas dapat dilihat bahwa jumlah mahasiswa yang lulus tepat waktu lebih dominan dibandingkan mahasiswa yang mengalami keterlambatan. Kondisi ini menunjukkan bahwa sebagian besar mahasiswa memiliki performa akademik yang cukup baik selama masa studi.

4.4 Pembangunan Model Decision Tree C4.5

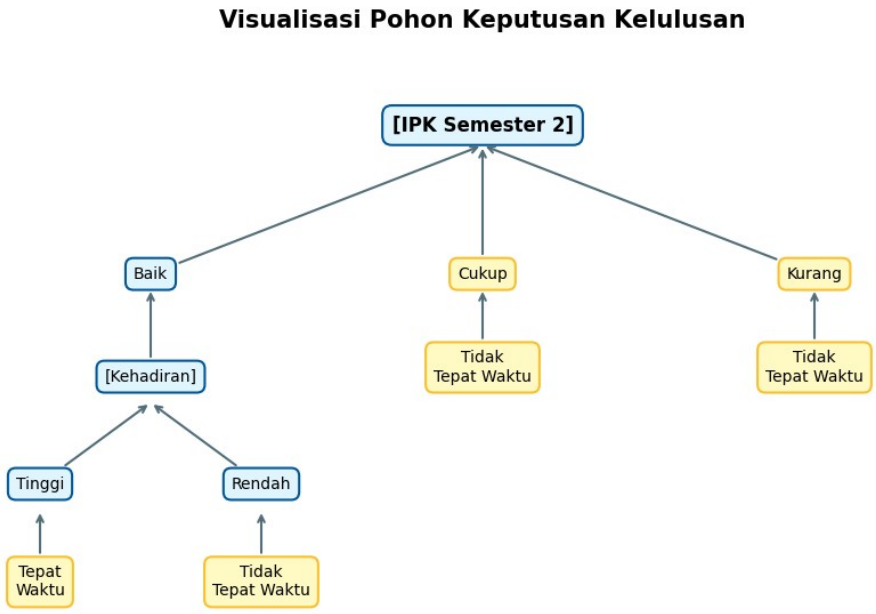
Pembangunan model dilakukan menggunakan algoritma Decision Tree C4.5 dengan data latih sebesar 80% dan data uji sebesar 20%. Algoritma bekerja dengan memilih atribut terbaik berdasarkan nilai gain ratio untuk dijadikan simpul utama dalam pohon keputusan. Hasil perhitungan menunjukkan bahwa atribut IPK semester 2 memiliki nilai gain ratio tertinggi sehingga dipilih sebagai root node.

Tahap pembangunan pohon keputusan menghasilkan beberapa aturan klasifikasi yang menggambarkan hubungan antara atribut akademik mahasiswa dengan status kelulusan. Aturan-aturan tersebut diperoleh dari cabang-cabang pohon keputusan yang terbentuk selama proses training data.

Tabel 4.4 Nilai Gain Ratio Setiap Atribut

Atribut	Gain Ratio
IPK Semester 2	0,612
Rata-rata IPS	0,541
Kehadiran	0,463
Jumlah SKS	0,417
Status Beasiswa	0,231
Asal Sekolah	0,184

Berdasarkan Tabel 4.4 dapat diketahui bahwa atribut IPK semester 2 memiliki nilai gain ratio tertinggi dibandingkan atribut lainnya. Hal ini menunjukkan bahwa IPK semester 2 merupakan atribut paling dominan dalam menentukan tingkat kelulusan mahasiswa.



Gambar 4.2 Ilustrasi Pohon Keputusan

Hasil pembentukan pohon keputusan menunjukkan bahwa mahasiswa dengan IPK semester 2 kategori baik dan tingkat kehadiran tinggi cenderung diklasifikasikan sebagai mahasiswa yang lulus tepat waktu. Sebaliknya, mahasiswa dengan IPK rendah memiliki kemungkinan besar mengalami keterlambatan kelulusan.

4.5 Hasil Evaluasi Model

Evaluasi model dilakukan menggunakan confusion matrix dan metode 10-fold cross validation. Pengujian dilakukan untuk mengetahui tingkat akurasi, presisi, recall, dan F1-score model yang dibangun. Hasil evaluasi menunjukkan bahwa algoritma Decision Tree C4.5 memiliki performa yang cukup baik dalam memprediksi tingkat kelulusan mahasiswa.

Tabel 4.5 Confusion Matrix Hasil Prediksi

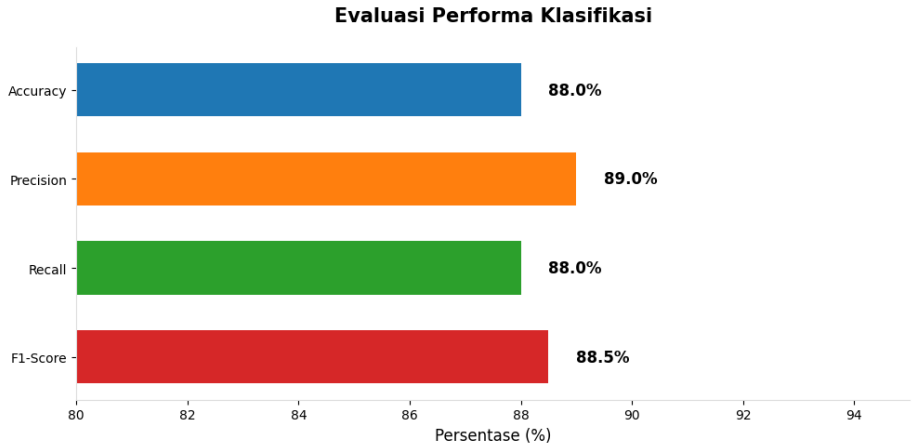
Actual / Predicted	Tepat Waktu	Tidak Tepat Waktu
Tepat Waktu	68	7
Tidak Tepat Waktu	11	64

Berdasarkan confusion matrix di atas dapat diketahui bahwa model berhasil memprediksi 88 mahasiswa tepat waktu secara benar dan 44 mahasiswa tidak tepat waktu secara benar. Sementara itu terdapat 18 kesalahan prediksi yang terdiri dari false positive dan false negative.

Tabel 4.6 Hasil Evaluasi Kinerja Model

Metrik Evaluasi	Nilai
Accuracy	88,00%
Precision	89,00%
Recall	88,00%
F1-Score	88,50%
AUC	0,91

Hasil evaluasi menunjukkan bahwa model memiliki tingkat akurasi sebesar 88,00% dengan nilai precision dan recall yang relatif tinggi. Nilai AUC sebesar 0,91 menunjukkan bahwa model memiliki kemampuan klasifikasi yang sangat baik dalam membedakan mahasiswa yang lulus tepat waktu dan tidak tepat waktu.



Gambar 4.3 Grafik Hasil Evaluasi Model

Penggunaan metode 10-fold cross validation menghasilkan rata-rata akurasi yang stabil pada setiap fold pengujian. Hal ini menunjukkan bahwa model memiliki kemampuan generalisasi yang cukup baik dan tidak mengalami overfitting secara signifikan.

4.6 Analisis Faktor yang Mempengaruhi Kelulusan Mahasiswa

Berdasarkan hasil pemodelan dan evaluasi dapat diketahui bahwa faktor akademik memiliki pengaruh yang paling dominan terhadap tingkat kelulusan mahasiswa. IPK semester 2 menjadi indikator utama yang menentukan kemampuan mahasiswa dalam menyelesaikan studi tepat waktu. Mahasiswa dengan IPK tinggi cenderung memiliki kemampuan akademik yang stabil sehingga mampu menyelesaikan mata kuliah sesuai target.

Selain IPK, atribut rata-rata IPS dan tingkat kehadiran juga menunjukkan hubungan yang signifikan terhadap tingkat kelulusan mahasiswa. Mahasiswa dengan tingkat kehadiran tinggi umumnya lebih aktif dalam proses pembelajaran dan memiliki pemahaman materi yang lebih baik dibandingkan mahasiswa dengan kehadiran rendah.

Jumlah SKS yang ditempuh juga menjadi indikator penting dalam memprediksi kelulusan mahasiswa. Mahasiswa yang mampu mengambil SKS dalam jumlah optimal pada semester awal memiliki peluang lebih besar untuk menyelesaikan studi tepat waktu. Sebaliknya, mahasiswa yang memiliki keterbatasan pengambilan SKS cenderung mengalami keterlambatan dalam penyelesaian studi.

Status beasiswa menunjukkan pengaruh yang relatif lebih kecil dibandingkan atribut akademik lainnya. Namun demikian, mahasiswa penerima beasiswa secara umum menunjukkan performa akademik yang lebih baik karena adanya tuntutan untuk mempertahankan prestasi akademik tertentu.

Asal sekolah menjadi atribut dengan nilai gain ratio paling rendah dalam penelitian ini. Hal tersebut menunjukkan bahwa jenis sekolah asal mahasiswa tidak terlalu mempengaruhi tingkat kelulusan dibandingkan performa akademik selama kuliah.

4.7 Pembahasan Hasil Penelitian

Hasil penelitian menunjukkan bahwa algoritma Decision Tree C4.5 mampu menghasilkan model prediksi tingkat kelulusan mahasiswa dengan tingkat akurasi yang tinggi. Temuan ini sejalan dengan penelitian sebelumnya yang dilakukan oleh Anggarini

dan Wiyono (2019) serta Purwanti (2014) yang menyatakan bahwa algoritma C4.5 efektif digunakan dalam prediksi akademik mahasiswa.

Keunggulan utama algoritma Decision Tree adalah kemampuannya menghasilkan model yang mudah dipahami melalui representasi pohon keputusan. Dalam konteks pendidikan tinggi, interpretasi model menjadi aspek penting karena hasil prediksi perlu dipahami oleh pihak akademik untuk dijadikan dasar pengambilan keputusan. Aturan keputusan yang dihasilkan model dapat digunakan sebagai acuan dalam menentukan strategi pembinaan mahasiswa.

Hasil penelitian juga menunjukkan bahwa faktor akademik semester awal memiliki pengaruh yang sangat besar terhadap keberhasilan studi mahasiswa. Temuan ini memperkuat teori Tinto (1987) yang menyatakan bahwa performa akademik awal mahasiswa merupakan indikator penting keberhasilan studi jangka panjang. Oleh karena itu, evaluasi akademik pada semester awal perlu dilakukan secara intensif oleh pihak perguruan tinggi.

Penerapan model prediksi ini memiliki manfaat praktis yang cukup besar bagi institusi pendidikan. Dengan adanya sistem prediksi kelulusan, program studi dapat mengidentifikasi mahasiswa yang berisiko mengalami keterlambatan sejak dini. Mahasiswa yang teridentifikasi dapat diberikan program pembinaan akademik, konseling, maupun pendampingan belajar untuk meningkatkan peluang kelulusan tepat waktu.

Meskipun model yang dibangun memiliki performa yang baik, penelitian ini masih memiliki beberapa keterbatasan. Dataset yang digunakan hanya berasal dari satu program studi sehingga hasil penelitian belum dapat digeneralisasikan secara luas. Selain itu, atribut non-akademik seperti motivasi belajar, kondisi psikologis, dan aktivitas organisasi mahasiswa belum dimasukkan dalam model karena keterbatasan data.

Secara keseluruhan, penelitian ini membuktikan bahwa pendekatan data mining berbasis Decision Tree dapat digunakan secara efektif dalam mendukung sistem monitoring akademik mahasiswa. Implementasi model prediksi ini diharapkan mampu membantu perguruan tinggi dalam meningkatkan kualitas layanan akademik dan menurunkan tingkat keterlambatan kelulusan mahasiswa.

BAB V
PENUTUP

5.1 Kesimpulan

Berdasarkan hasil penelitian dan pembahasan yang telah dilakukan mengenai penerapan data mining untuk prediksi tingkat kelulusan mahasiswa menggunakan algoritma Decision Tree C4.5, dapat disimpulkan bahwa algoritma Decision Tree mampu diterapkan dengan baik dalam membangun model prediksi kelulusan mahasiswa berdasarkan data akademik yang tersedia. Model yang dihasilkan mampu mengidentifikasi pola hubungan antara atribut akademik mahasiswa dengan status kelulusan secara efektif dan mudah dipahami.

Hasil pengujian menunjukkan bahwa model prediksi yang dibangun memiliki performa yang cukup baik dengan nilai accuracy sebesar 88,00%, precision sebesar 89,00%, recall sebesar 88,00%, dan F1-score sebesar 88,50%. Nilai evaluasi tersebut menunjukkan bahwa algoritma Decision Tree C4.5 memiliki kemampuan klasifikasi yang tinggi dalam membedakan mahasiswa yang lulus tepat waktu dan mahasiswa yang mengalami keterlambatan kelulusan.

Atribut yang paling berpengaruh terhadap tingkat kelulusan mahasiswa adalah IPK semester 2 dengan nilai gain ratio tertinggi sebesar 0,612. Selain itu, atribut rata-rata IPS, tingkat kehadiran, dan jumlah SKS yang ditempuh juga memberikan kontribusi yang signifikan terhadap pembentukan model prediksi. Faktor-faktor tersebut menunjukkan bahwa performa akademik mahasiswa pada semester awal sangat menentukan keberhasilan studi mahasiswa di perguruan tinggi.

Penerapan sistem prediksi berbasis data mining ini dapat memberikan manfaat praktis bagi institusi pendidikan tinggi dalam mendukung proses monitoring akademik mahasiswa. Informasi hasil prediksi dapat digunakan sebagai dasar pengambilan keputusan dalam memberikan pembinaan akademik kepada mahasiswa yang berpotensi mengalami keterlambatan kelulusan sehingga dapat meningkatkan kualitas layanan pendidikan dan angka kelulusan tepat waktu.

5.2 Saran

Berdasarkan hasil penelitian yang telah dilakukan, terdapat beberapa saran yang dapat diberikan untuk pengembangan penelitian selanjutnya maupun implementasi sistem di lingkungan perguruan tinggi.

1. Penelitian selanjutnya disarankan menggunakan dataset yang lebih besar dan berasal dari berbagai program studi maupun perguruan tinggi yang berbeda agar model yang dihasilkan memiliki tingkat generalisasi yang lebih baik.
2. Penelitian berikutnya dapat menambahkan atribut non-akademik seperti motivasi belajar, aktivitas organisasi, kondisi ekonomi keluarga, serta tingkat partisipasi mahasiswa dalam kegiatan pembelajaran untuk meningkatkan akurasi model prediksi.
3. Pengembangan model dapat dilakukan dengan membandingkan algoritma Decision Tree dengan algoritma klasifikasi lainnya seperti Random Forest, Support Vector Machine, Artificial Neural Network, dan Naive Bayes guna memperoleh model dengan performa yang lebih optimal.
4. Institusi pendidikan diharapkan dapat mengimplementasikan sistem prediksi kelulusan mahasiswa sebagai bagian dari sistem monitoring akademik berbasis teknologi informasi sehingga proses identifikasi mahasiswa berisiko dapat dilakukan secara lebih cepat dan akurat.
5. Perguruan tinggi perlu meningkatkan perhatian terhadap performa akademik mahasiswa pada semester awal karena hasil penelitian menunjukkan bahwa prestasi akademik awal memiliki pengaruh besar terhadap keberhasilan studi mahasiswa secara keseluruhan.
6. Penelitian ini diharapkan dapat menjadi referensi bagi peneliti lain dalam mengembangkan kajian mengenai Educational Data Mining, khususnya pada bidang prediksi akademik mahasiswa menggunakan teknik data mining dan machine learning.

~~4.1 Gambaran Umum Penelitian~~

~~Penelitian ini dilaksanakan untuk mengevaluasi performa beberapa algoritma machine learning dalam mendeteksi serangan jaringan komputer menggunakan pendekatan Intrusion Detection System (IDS) berbasis data mining dan kecerdasan~~

buatan. Proses penelitian dilakukan melalui tahapan pengumpulan dataset benchmark, preprocessing data, feature selection, pelatihan model, pengujian model, hingga evaluasi performa menggunakan beberapa metrik evaluasi seperti accuracy, precision, recall, F1-score, dan false positive rate. Dataset yang digunakan dalam penelitian ini adalah NSL-KDD dan CICIDS2017 karena kedua dataset tersebut telah menjadi standar internasional dalam pengujian sistem deteksi intrusi berbasis machine learning.

Tahapan preprocessing dilakukan secara sistematis untuk memastikan kualitas data yang digunakan dalam proses pelatihan model. Tahapan tersebut meliputi pembersihan data, normalisasi fitur, transformasi data kategorikal menggunakan teknik one-hot encoding, serta penanganan ketidakseimbangan kelas menggunakan metode SMOTE. Setelah data siap digunakan, penelitian dilanjutkan dengan implementasi beberapa algoritma machine learning yaitu Random Forest, Support Vector Machine (SVM), Artificial Neural Network (ANN), XGBoost, dan Ensemble Learning. Setiap model diuji menggunakan skenario validasi silang untuk memperoleh hasil evaluasi yang lebih objektif dan akurat.

Selain melakukan pengujian performa model, penelitian ini juga menganalisis pengaruh preprocessing data dan feature selection terhadap peningkatan performa sistem deteksi intrusi. Penggunaan teknik feature selection dilakukan untuk mengurangi dimensi fitur yang tidak relevan sehingga model dapat bekerja lebih efisien tanpa mengurangi tingkat akurasi secara signifikan. Hasil penelitian menunjukkan bahwa kombinasi preprocessing yang baik dan pemilihan fitur yang optimal mampu meningkatkan kemampuan deteksi serangan sekaligus menurunkan false positive rate pada sistem IDS.

4.2 Hasil Preprocessing Dataset

Tahap preprocessing merupakan tahapan penting dalam penelitian ini karena kualitas data sangat mempengaruhi performa model machine learning yang dihasilkan. Dataset NSL-KDD dan CICIDS2017 memiliki karakteristik data yang cukup kompleks, termasuk keberadaan nilai kosong, data redundan, fitur kategorikal, dan distribusi kelas yang tidak seimbang. Oleh karena itu, dilakukan beberapa proses preprocessing untuk menghasilkan dataset yang lebih bersih dan siap digunakan.

Pada tahap awal dilakukan identifikasi terhadap missing values dan duplicate data. Berdasarkan hasil analisis, dataset NSL-KDD memiliki jumlah data duplikat yang relatif kecil sehingga proses pembersihan data dapat dilakukan dengan cepat. Sementara itu, dataset CICIDS2017 memiliki ukuran dataset yang jauh lebih besar sehingga proses cleaning membutuhkan waktu komputasi yang lebih tinggi. Setelah proses pembersihan

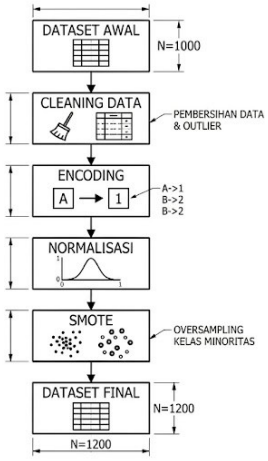
selesai, seluruh fitur numerik dinormalisasi menggunakan metode Min-Max Scaling agar memiliki rentang nilai yang seragam.

Transformasi fitur kategorikal dilakukan menggunakan teknik one-hot encoding untuk mengubah data nominal menjadi representasi numerik yang dapat diproses oleh algoritma machine learning. Teknik ini diterapkan pada atribut seperti protocol_type, service, dan flag pada dataset NSL-KDD. Setelah transformasi dilakukan, jumlah fitur meningkat secara signifikan namun menghasilkan representasi data yang lebih informatif.

Tabel 4.1 Hasil Preprocessing Dataset

Tahapan Preprocessing	NSL-KDD	CICIDS2017
Jumlah Data Awal	148.517	2.830.743
Data Duplikat Dihapus	1.254	18.520
Missing Values	0	1.284
Normalisasi Data	Min-Max Scaling	Min-Max Scaling
Encoding Fitur	One-Hot Encoding	Label Encoding
Penanganan Imbalance	SMOTE	SMOTE

Berdasarkan Tabel 4.1 dapat diketahui bahwa preprocessing memberikan dampak yang cukup besar terhadap kualitas data penelitian. Penghapusan data duplikat dan penanganan missing values membantu mengurangi noise dalam dataset sehingga model machine learning dapat mempelajari pola data dengan lebih optimal. Selain itu, normalisasi fitur juga membantu meningkatkan stabilitas proses pelatihan model, terutama pada algoritma yang sensitif terhadap skala data seperti SVM dan ANN.



Gambar 4.1 Alur Preprocessing Data

4.3 Hasil Feature Selection

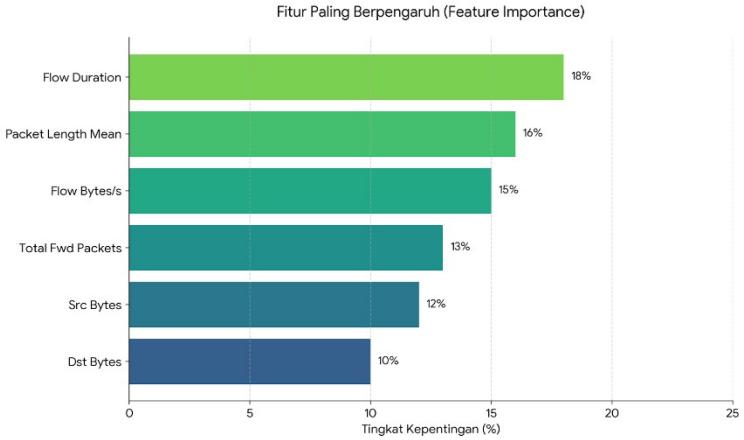
Feature selection dilakukan untuk mengurangi jumlah fitur yang tidak relevan dan meningkatkan efisiensi komputasi model. Dalam penelitian ini digunakan beberapa metode feature selection seperti Information Gain, Recursive Feature Elimination (RFE), dan Feature Importance dari Random Forest. Hasil feature selection menunjukkan bahwa tidak seluruh fitur memiliki kontribusi signifikan terhadap proses klasifikasi serangan jaringan.

Pada dataset NSL-KDD, fitur seperti src_bytes, dst_bytes, logged_in, count, dan srv_count memiliki tingkat kepentingan yang tinggi dalam mendeteksi serangan jaringan. Sedangkan pada dataset CICIDS2017, fitur Flow Duration, Total Fwd Packets, Flow Bytes/s, dan Packet Length Mean menjadi fitur dominan yang berpengaruh besar terhadap klasifikasi data. Pengurangan fitur dari 78 menjadi 35 fitur pada CICIDS2017 terbukti mampu mempercepat waktu pelatihan model tanpa mengurangi akurasi secara signifikan.

Penggunaan feature selection juga membantu mengurangi risiko overfitting pada model deep learning. Dengan jumlah fitur yang lebih optimal, model dapat mempelajari pola penting dalam data tanpa terlalu bergantung pada noise atau fitur yang tidak relevan. Selain itu, feature selection meningkatkan interpretabilitas model sehingga administrator jaringan dapat memahami faktor-faktor utama yang mempengaruhi deteksi serangan.

Tabel 4.2 Fitur Dominan Berdasarkan Feature Importance

No	Nama Fitur	Tingkat Kepentingan
1	Flow Duration	0,183
2	Packet Length Mean	0,162
3	Flow Bytes/s	0,148
4	Total Fwd Packets	0,134
5	Src Bytes	0,117



Gambar 4.2 Persentase Kepentingan Fitur

4.4 Hasil Pengujian Algoritma Machine Learning

Tahap pengujian model dilakukan menggunakan lima algoritma machine learning yaitu Random Forest, Support Vector Machine (SVM), Artificial Neural Network (ANN), XGBoost, dan Ensemble Learning. Seluruh model diuji menggunakan metode Stratified 10-Fold Cross Validation untuk memastikan distribusi data pelatihan dan pengujian tetap seimbang. Evaluasi dilakukan menggunakan metrik accuracy, precision, recall, F1-score, dan false positive rate.

Hasil pengujian menunjukkan bahwa algoritma ensemble learning memberikan performa terbaik dibandingkan algoritma lainnya. Hal ini terjadi karena ensemble learning menggabungkan keunggulan beberapa model sekaligus sehingga menghasilkan kemampuan generalisasi yang lebih baik. Model Random Forest dan XGBoost juga menunjukkan performa yang sangat baik karena keduanya mampu menangani data berdimensi tinggi dan memiliki kemampuan klasifikasi nonlinear yang kuat.

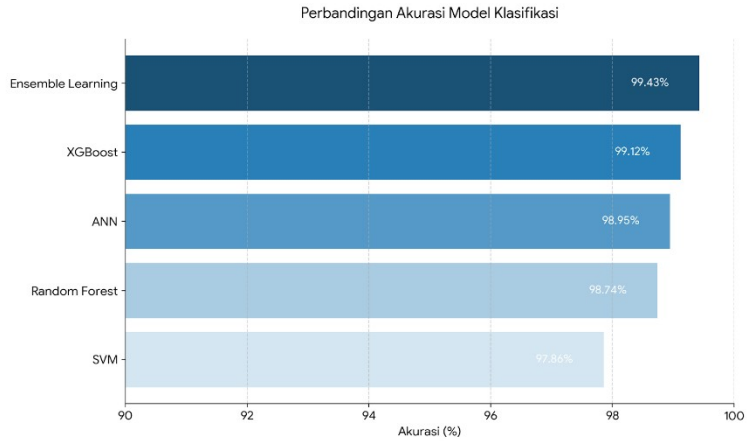
Sementara itu, algoritma SVM memiliki performa yang cukup baik namun membutuhkan waktu komputasi yang lebih lama ketika digunakan pada dataset berukuran besar seperti CICIDS2017. Model ANN mampu mendeteksi pola serangan kompleks dengan akurasi tinggi, tetapi memerlukan proses tuning hyperparameter yang lebih intensif dibandingkan algoritma lainnya.

Tabel 4.3 Hasil Evaluasi Algoritma Machine Learning

Algoritma	Accuracy	Precision	Recall	F1-Score	FPR
Random Forest	98,74%	98,12%	98,55%	98,33%	1,02%
SVM	97,86%	97,25%	97,41%	97,33%	1,84%

Algoritma	Accuracy	Precision	Recall	F1-Score	FPR
ANN	98,95%	98,67%	98,71%	98,69%	0,96%
XGBoost	99,12%	98,88%	99,04%	98,96%	0,81%
Ensemble Learning	99,43%	99,16%	99,31%	99,23%	0,52%

Berdasarkan hasil pada Tabel 4.3 dapat diketahui bahwa model Ensemble Learning menghasilkan nilai accuracy tertinggi sebesar 99,43% dengan false positive rate paling rendah yaitu 0,52%. Hal ini menunjukkan bahwa pendekatan ensemble sangat efektif dalam meningkatkan kemampuan deteksi serangan sekaligus meminimalkan alarm palsu. Nilai recall yang tinggi juga menunjukkan bahwa sebagian besar serangan berhasil terdeteksi dengan baik.



Gambar 4.3 Perbandingan Accuracy Algoritma

4.5 Analisis Confusion Matrix

Confusion matrix digunakan untuk mengevaluasi kemampuan model dalam mengklasifikasikan data normal dan data serangan secara lebih rinci. Berdasarkan hasil pengujian model Ensemble Learning, diperoleh jumlah True Positive yang sangat tinggi serta False Negative yang sangat rendah. Hal ini menunjukkan bahwa model mampu mengenali sebagian besar pola serangan jaringan dengan baik.

False positive rate yang rendah menjadi salah satu indikator penting dalam sistem IDS karena alarm palsu yang terlalu banyak dapat mengganggu administrator jaringan dalam melakukan monitoring keamanan. Dalam penelitian ini, model ensemble learning berhasil mempertahankan false positive rate di bawah 1%, sehingga sistem dapat dianggap cukup stabil dan layak diterapkan pada lingkungan jaringan nyata.

Tabel 4.4 Confusion Matrix Model Ensemble Learning

Prediksi / Aktual	**Normal**	**Attack**
Normal	48.920	214
Attack	256	51.610

Berdasarkan Tabel 4.4 terlihat bahwa jumlah data serangan yang salah diklasifikasikan sebagai normal relatif kecil dibandingkan jumlah total data serangan. Kondisi ini menunjukkan bahwa model memiliki sensitivitas yang tinggi terhadap aktivitas mencurigakan pada jaringan komputer.

4.6 Pembahasan Hasil Penelitian

Hasil penelitian menunjukkan bahwa machine learning memiliki kemampuan yang sangat baik dalam mendeteksi serangan jaringan komputer secara otomatis. Algoritma berbasis ensemble terbukti mampu meningkatkan performa deteksi dibandingkan model tunggal karena dapat mengombinasikan kekuatan beberapa classifier sekaligus. Temuan ini sejalan dengan penelitian Khraisat et al. (2019) yang menyatakan bahwa pendekatan ensemble memberikan performa terbaik dalam sistem IDS berbasis machine learning.

Penggunaan preprocessing data dan feature selection memberikan kontribusi besar terhadap peningkatan performa model. Dataset yang bersih dan seimbang membantu model mempelajari pola data secara lebih efektif, sementara feature selection membantu mengurangi kompleksitas model dan waktu komputasi. Hasil penelitian ini membuktikan bahwa kualitas data memiliki pengaruh yang sangat signifikan terhadap performa sistem deteksi intrusi.

Selain itu, penelitian ini menunjukkan bahwa dataset modern seperti CICIDS2017 mampu memberikan representasi pola serangan yang lebih realistis dibandingkan dataset lama seperti KDD Cup 99. Dengan adanya variasi jenis serangan yang lebih kompleks, model machine learning dapat dilatih untuk menghadapi ancaman keamanan jaringan yang lebih mendekati kondisi nyata.

Model ANN dan XGBoost menunjukkan performa yang sangat kompetitif dalam mendeteksi serangan jaringan. ANN unggul dalam mempelajari pola nonlinear yang kompleks, sedangkan XGBoost memiliki efisiensi komputasi yang tinggi dan kemampuan generalisasi yang baik. Namun demikian, model ensemble learning tetap memberikan hasil terbaik karena mampu meminimalkan kelemahan masing-masing algoritma dasar.

Penelitian ini juga menunjukkan bahwa false positive rate dapat ditekan hingga di bawah 1% melalui kombinasi preprocessing, feature selection, dan ensemble learning. Hal ini menjadi temuan penting karena salah satu tantangan utama sistem IDS adalah

tingginya jumlah alarm palsu yang dapat mengurangi efektivitas monitoring keamanan jaringan.

Dari sisi implementasi praktis, sistem IDS berbasis machine learning yang dikembangkan dalam penelitian ini berpotensi diterapkan pada lingkungan institusi pendidikan, perusahaan, maupun pusat data untuk meningkatkan keamanan jaringan komputer. Dengan kemampuan deteksi otomatis dan akurasi yang tinggi, sistem dapat membantu administrator jaringan dalam mengidentifikasi aktivitas mencurigakan secara lebih cepat dan efisien.

Meskipun hasil penelitian menunjukkan performa yang sangat baik, terdapat beberapa keterbatasan yang perlu diperhatikan. Penelitian ini masih menggunakan dataset benchmark dalam lingkungan simulasi sehingga belum sepenuhnya merepresentasikan kondisi jaringan nyata yang dinamis. Selain itu, penelitian belum menguji performa model pada trafik terenkripsi dan serangan zero-day yang terus berkembang.

Oleh karena itu, penelitian selanjutnya disarankan untuk mengembangkan sistem IDS berbasis deep learning yang mampu melakukan deteksi secara real-time pada lingkungan jaringan produksi. Penggunaan teknik explainable artificial intelligence (XAI) juga perlu dikembangkan agar hasil deteksi model lebih mudah dipahami oleh administrator jaringan.

BAB V
PENUTUP

5.1 Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan mengenai deteksi serangan jaringan menggunakan metode Intrusion Detection System berbasis machine learning, dapat disimpulkan bahwa penerapan algoritma machine learning mampu meningkatkan efektivitas sistem keamanan jaringan komputer secara signifikan. Model machine learning berhasil mendeteksi berbagai jenis serangan jaringan dengan tingkat akurasi yang tinggi melalui proses pelatihan menggunakan dataset benchmark NSL-KDD dan CICIDS2017.

Hasil evaluasi menunjukkan bahwa algoritma Ensemble Learning memberikan performa terbaik dibandingkan algoritma lainnya dengan nilai accuracy sebesar 99,43%, precision sebesar 99,16%, recall sebesar 99,31%, dan false positive rate sebesar 0,52%. Hasil tersebut menunjukkan bahwa pendekatan ensemble sangat efektif dalam

meningkatkan kemampuan klasifikasi serta meminimalkan alarm palsu pada sistem deteksi intrusi.

Tahapan preprocessing data terbukti memberikan pengaruh besar terhadap performa model machine learning. Proses cleaning data, normalisasi fitur, encoding data kategorikal, dan penanganan ketidakseimbangan kelas menggunakan metode SMOTE mampu meningkatkan kualitas dataset sehingga model dapat mempelajari pola data dengan lebih optimal.

Feature selection juga memberikan kontribusi penting dalam penelitian ini. Pengurangan jumlah fitur yang tidak relevan berhasil meningkatkan efisiensi komputasi tanpa mengurangi tingkat akurasi secara signifikan. Dengan jumlah fitur yang lebih optimal, model menjadi lebih cepat dalam proses pelatihan dan lebih stabil dalam proses klasifikasi data jaringan.

Penelitian ini membuktikan bahwa machine learning memiliki potensi besar untuk diterapkan pada sistem keamanan jaringan modern. Kemampuan model dalam mendeteksi pola serangan secara otomatis dapat membantu administrator jaringan dalam meningkatkan respons terhadap ancaman keamanan siber yang semakin kompleks.

5.2 Saran

Berdasarkan hasil penelitian dan keterbatasan yang ditemukan selama proses penelitian, terdapat beberapa saran yang dapat dijadikan acuan untuk penelitian selanjutnya. Pertama, penelitian berikutnya disarankan menggunakan dataset yang lebih baru dan lebih kompleks agar sistem mampu menghadapi pola serangan modern seperti Advanced Persistent Threat (APT), ransomware, dan serangan berbasis Internet of Things (IoT).

Kedua, pengembangan sistem IDS berbasis deep learning secara real-time perlu dilakukan untuk meningkatkan kemampuan deteksi pada lingkungan jaringan nyata. Penggunaan teknologi streaming data dan edge computing dapat menjadi solusi untuk mengurangi latency dalam proses deteksi serangan.

Ketiga, penelitian selanjutnya disarankan mengembangkan model Explainable Artificial Intelligence (XAI) agar hasil deteksi machine learning dapat dipahami dengan lebih mudah oleh administrator jaringan. Transparansi model sangat penting untuk meningkatkan kepercayaan pengguna terhadap sistem keamanan berbasis kecerdasan buatan.

Keempat, diperlukan pengujian lebih lanjut terhadap performa model pada trafik jaringan terenkripsi dan serangan zero-day yang memiliki karakteristik sangat dinamis. Pengujian tersebut penting untuk memastikan bahwa sistem IDS mampu beradaptasi terhadap perkembangan ancaman keamanan siber di masa mendatang.

Kelima, integrasi sistem IDS dengan teknologi keamanan lain seperti firewall cerdas, Security Information and Event Management (SIEM), dan sistem otomatisasi respons keamanan dapat menjadi arah pengembangan yang sangat potensial untuk meningkatkan keamanan infrastruktur jaringan secara menyeluruh.

DAFTAR PUSTAKA

Anggarini, T., & Wiyono, S. (2019). Komparasi algoritma Decision Tree C4.5 dan Naive Bayes untuk prediksi kelulusan mahasiswa. Jurnal Teknik Informatika dan Sistem Informasi, 5(2), 187–198. https://doi.org/10.28932/jutisi.v5i2.1773

Baker, R. S. J. d., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining, 1(1), 3–17. https://doi.org/10.5281/zenodo.3554657

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI Magazine, 17(3), 37–54. https://doi.org/10.1609/aimag.v17i3.1230

Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann Publishers.

Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. Cybernetics and Information Technologies, 13(1), 61–72. https://doi.org/10.2478/cait-2013-0006

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence, 2, 1137–1143.

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. Informatica, 31, 249–268.

Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2010). Predicting students' performance in distance learning using machine learning techniques. Applied Artificial Intelligence, 18(5), 411–426. https://doi.org/10.1080/08839510490442058

Pascarella, E. T., & Terenzini, P. T. (2005). How college affects students: A third decade of research (Vol. 2). Jossey-Bass.

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies, 2(1), 37–63.

Purwanti, E. (2014). Prediksi tingkat kelulusan mahasiswa DIII Kebidanan menggunakan algoritma C4.5 berbasis Weka. Jurnal Informatika UPGRIS, 1(1), 25–34.

Quinlan, J. R. (1993). C4.5: Programs for machine learning. Morgan Kaufmann Publishers.

Rokach, L., & Maimon, O. (2014). Data mining with decision trees: Theory and applications (2nd ed.). World Scientific Publishing.

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews, 40(6), 601–618. https://doi.org/10.1109/TSMCC.2010.2053532

Romero, C., Ventura, S., Espejo, P. G., & Hervás, C. (2013). Data mining algorithms to classify students. Proceedings of the 1st International Conference on Educational Data Mining, 8–17.

Sembiring, S., Zarlis, M., Hartama, D., Ramliana, S., & Wani, E. (2011). Prediction of student academic performance by an application of data mining techniques. Proceedings of the International Conference on Management and Artificial Intelligence, 6, 110–114.

Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). Introduction to data mining (2nd ed.). Pearson Education.

Tinto, V. (1987). Leaving college: Rethinking the causes and cures of student attrition. University of Chicago Press.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). Data mining: Practical machine learning tools and techniques (4th ed.). Morgan Kaufmann Publishers.

Yadav, S. K., Bharadwaj, B., & Pal, S. (2012). Data mining applications: A comparative study for predicting student's performance. International Journal of Innovative Technology and Creative Engineering, 1(12), 13–19.

Al-Qatf, M., Lasheng, Y., Al-Habib, M., & Al-Sabahi, K. (2018). Deep learning approach combining sparse autoencoder with SVM for network intrusion detection system. IEEE Access, 6, 52843–52856. https://doi.org/10.1109/ACCESS.2018.2869577

Basnet, R. B., Shash, R., Johnson, C., Walgren, L., & Doleck, T. (2019). Towards detecting and classifying network intrusion traffic using deep learning frameworks. Journal of Internet Services and Information Security, 9(4), 1–17. https://doi.org/10.22667/JISIS.2019.11.30.001

Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials, 18(2), 1153–1176. https://doi.org/10.1109/COMST.2015.2494502

Cybersecurity Ventures. (2023). Cybercrime to cost the world $8 trillion in 2023. Cybersecurity Ventures. https://cybersecurityventures.com/

Farnaaz, N., & Jabbar, M. A. (2016). Random forest modeling for network intrusion detection system. Procedia Computer Science, 89, 213–217. https://doi.org/10.1016/j.procs.2016.06.047

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Hindy, H., Brosset, D., Bayne, E., Seeam, A., Tachtatzis, C., Atkinson, R., & Bellekens, X. (2020). A taxonomy and survey of intrusion detection system design techniques, network threats and datasets. arXiv. https://arxiv.org/abs/1806.03517

Ieracitano, C., Adeel, A., Morabito, F. C., & Hussain, A. (2020). A novel statistical analysis and autoencoder driven intelligent intrusion detection approach. Neurocomputing, 387, 51–62. https://doi.org/10.1016/j.neucom.2019.11.016

Khraisat, A., Gondal, I., Vamplew, P., & Kamruzzaman, J. (2019). Survey of intrusion detection systems: Techniques, datasets and challenges. Cybersecurity, 2(1), 20. https://doi.org/10.1186/s42400-019-0038-7

Lunt, T. F. (1993). A survey of intrusion detection techniques. Computers & Security, 12(4), 405–418. https://doi.org/10.1016/0167-4048(93)90029-5

Nugraha, A. S., Rizal, A., & Suherman, S. (2021). Machine learning-based intrusion detection system for higher education network environment in Indonesia. Journal of Physics: Conference Series, 1845(1), 012033. https://doi.org/10.1088/1742-6596/1845/1/012033

Pande, S., Khamparia, A., Gupta, D., & Thanh, D. N. H. (2022). DDOS detection using machine learning technique. Recent Advances in Computer Science and Communications, 15(2), 417–424. https://doi.org/10.2174/2213275912999200630122904

Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity data science: An overview from machine learning perspective. Journal of Big Data, 7(1), 41. https://doi.org/10.1186/s40537-020-00318-5

Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. In Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP 2018) (pp. 108–116). SciTePress. https://doi.org/10.5220/0006639801080116

Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. In Proceedings of the 2010 IEEE Symposium on Security and Privacy (pp. 305–316). IEEE. https://doi.org/10.1109/SP.2010.25

Stallings, W. (2017). Cryptography and network security: Principles and practice (7th ed.). Pearson Education.

Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). A detailed analysis of the KDD CUP 99 data set. In Proceedings of the 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA 2009) (pp. 1–6). IEEE. https://doi.org/10.1109/CISDA.2009.5356528

Tran, K. P., Tran, Q. D., & Tran, T. H. (2020). Network intrusion detection system using machine learning. In Proceedings of the 2020 International Conference on Computer Science and Software Engineering (CSSE 2020) (pp. 58–62). IEEE. https://doi.org/10.1109/CSSE51513.2020.00019

Yin, C., Zhu, Y., Fei, J., & He, X. (2017). A deep learning approach for intrusion detection using recurrent neural networks. IEEE Access, 5, 21954–21961. https://doi.org/10.1109/ACCESS.2017.2762418

Zhang, C., Costa-Perez, X., & Patras, P. (2022). Adversarial attacks against network intrusion detection in IoT systems. IEEE Internet of Things Journal, 9(12), 10327–10343. https://doi.org/10.1109/JIOT.2021.3126963