eISSN : 2621-4970
pISSN : 2621-4962

Volume 8 | No 1
Januari | **2025**

Mengoptimalkan Kinerja Naïve Bayes Pada Ancaman Modern Dengan Menggunakan PCA Pada Data Intrusion Detection System (IDS)

Kevin Salsabil Arlandy[1], Ahmad Faqih[2], Ade Rizki Rinaldi[3]
[1, 2]Program Studi Teknik Informatika
[3]Program Studi Rekayasa Perangkat Lunak
STMIK IKMI Cirebon
kevinarlandyprimary@gmail.com

Abstrak

Intrusion Detection System (IDS) digunakan untuk mendeteksi serangan atau aktivitas mencurigakan dalam jaringan. Dengan meningkatnya ancaman siber modern, penelitian ini mengusulkan kombinasi metode Naïve Bayes dan Principal Component Analysis (PCA) untuk meningkatkan akurasi dan efisiensi deteksi. Metode tambahan PCA dapet mereduksi dimensi dataset menjadi 30 komponen utama tanpa kehilangan informasi penting, menggunakan dataset UNSW-NB15. Proses melibatkan standarisasi data dengan Standard-Scaler, reduksi dimensi menggunakan PCA, serta evaluasi model Naïve Bayes pada dataset dengan dan tanpa PCA. Analisis ini menggunakan program Python yang di eksekusi dengan Google Collab, dengan hasil menunjukkan bahwa model dengan PCA mencapai akurasi sebesar 96.65% dengan recall 1.00 untuk kelas ancaman, meskipun presisi masih rendah (0.49). Sebaliknya, tanpa PCA, akurasi hanya mencapai 92.72% dengan presisi 0.31 untuk kelas yang sama. Selain itu, penggunaan PCA berhasil mengurangi waktu komputasi dari 1 menit menjadi 30 detik. Kombinasi dengan teknik reduksi dimensi Principal Component Analysis (PCA) menunjukkan kinerja yang lebih baik dalam mengklasifikasikan data pada sistem Intrusion Detection System (IDS). PCA dan Naïve Bayes terbukti menjanjikan dalam mendeteksi ancaman modern, meskipun masih diperlukan perbaikan untuk mencapai kinerja yang lebih optimal.

Kata kunci: **Intrusion Detection System, Naïve Bayes, PCA, Keamanan Jaringan**

Abstract

An Intrusion Detection System (IDS) is used to detect attacks or suspicious activities in the network. With the increase of modern cyber threats, this research proposes a combination of Naïve Bayes and Principal Component Analysis (PCA) methods to improve detection accuracy and efficiency. The additional PCA method can reduce the dataset dimension to 30 principal components without losing important information, using the UNSW-NB15 dataset. The process involves data standardization with Standard-Scaler, dimensionality reduction using PCA, and Naïve Bayes model evaluation on the dataset with and without PCA. This analysis used a Python program executed with Google Collab, with the results showing that the model with PCA achieved an accuracy of 96.65% with a recall of 1.00 for the threat class. However, the precision was still low (0.49). In contrast, without PCA, the accuracy only reached 92.72% with a precision of 0.31 for the same class. In addition, the use of PCA successfully reduced the computation time from 1 minute to 30 seconds combination with the Principal Component Analysis (PCA) dimension reduction technique shows better perfor-

mance in classifying data in the Intrusion Detection System (IDS). PCA and Naïve Bayes proved promising in detecting modern threats, although improvements are still needed to achieve more optimal performance.
Keywords: ***Intrusion Detection System, Naïve Bayes, PCA, Network Security***

PENDAHULUAN

Di era digital yang terus berkembang, jaringan komputer telah menjadi tulang punggung berbagai aktivitas, mulai dari bisnis, pendidikan, hingga pemerintahan. Namun, pesatnya perkembangan ini juga diiringi dengan meningkatnya ancaman keamanan siber, seperti Distributed Denial of Service (DDoS), malware, dan network intrusion. Serangan-serangan ini dapat menyebabkan dampak yang signifikan, baik secara finansial maupun terhadap reputasi suatu organisasi. Oleh karena itu, pengamanan jaringan menjadi isu yang semakin krusial.

Salah satu teknologi yang dikembangkan untuk menangani ancaman tersebut adalah Intrusion Detection System (IDS). IDS adalah sistem keamanan yang dirancang untuk memantau aktivitas jaringan atau sistem komputer dengan tujuan mendeteksi perilaku mencurigakan atau indikasi adanya intrusi. IDS berfungsi sebagai lapisan pertahanan tambahan yang dapat mengidentifikasi ancaman secara real-time. IDS secara umum terbagi menjadi dua kategori utama, yaitu Network-based Intrusion Detection System (NIDS) yang menganalisis lalu lintas jaringan, dan Host-based Intrusion Detection System (HIDS) yang memantau aktivitas pada perangkat individu. Metode seperti Naïve Bayes menawarkan solusi klasifikasi yang sederhana namun efektif, sedangkan Principal Component Analysis (PCA)[1]

Naïve Bayes adalah salah satu algoritma klasifikasi yang sering digunakan dalam sistem deteksi intrusi. Algoritma ini bekerja berdasarkan prinsip probabilitas statistik, di mana ia mengasumsikan bahwa setiap atribut dalam dataset bersifat independen satu sama lain. Dalam konteks IDS, Naïve Bayes dapat digunakan untuk mengklasifikasikan jenis serangan berdasarkan pola yang terdeteksi dalam data. Penelitian menunjukkan bahwa penggunaan Naïve Bayes dalam IDS dapat meningkatkan akurasi deteksi, terutama ketika dikombinasikan dengan teknik pemilihan atribut seperti correlation-based feature selection.[2]

Kombinasi Principal Component Analysis (PCA) dengan Naive Bayes Classifiers (NBC) secara signifikan meningkatkan akurasi deteksi intrusi, mencapai 100% untuk Gaussian dan Complement Naive Bayes, serta 99% untuk Multinomial Naive Bayes pada dataset UNSW-NB15. PCA efektif mengurangi dimensi data, mempercepat pelatihan, dan meningkatkan akurasi. Namun, keterbatasan penelitian ini meliputi asumsi independensi fitur pada Naive Bayes, ketiadaan uji real-time, dan belum mencakup ancaman kompleks seperti IoT dan zero-day attacks. Selain itu, hubungan temporal data diabaikan, membuka peluang penggabungan PCA dengan deep learning. Penelitian lanjutan dapat mengeksplorasi teknik reduksi dimensi alternatif untuk mengurangi overhead komputasi.[3]

Beban kerja manual yang tinggi dalam menganalisis log alert yang dihasilkan oleh IDS, khususnya pada sistem berbasis Host-based (HIDS) dan Signature-based (SIDS). Sistem ini memiliki keterbatasan dalam mendeteksi serangan baru dan sering menghasilkan banyak alert yang tidak rel-

evan. Penelitian menggunakan metodologi dari SKKNI (Standar Kompetensi Kerja Nasional Indonesia) dalam bidang data science yang meliputi pemahaman data, persiapan data, pembuatan model, dan evaluasi model. Penerapan algoritma Naïve Bayes secara signifikan meningkatkan efisiensi dan efektivitas analisis log alert IDS. Hal ini memberikan keuntungan bagi tim keamanan perusahaan untuk fokus pada data kritis, mengurangi risiko, serta menghemat waktu dan tenaga. [4]

Penelitian ini akan menggunakan pendekatan eksperimental dengan menerapkan algoritma Naïve Bayes dan Principal Component Analysis (PCA) pada dataset UNSW-NB15. Teknik seleksi fitur berbasis korelasi akan diterapkan untuk mengidentifikasi atribut yang paling relevan dalam mendeteksi ancaman, dengan tujuan meningkatkan akurasi klasifikasi dan menurunkan tingkat kesalahan positif. Model yang dihasilkan akan dievaluasi menggunakan metrik performa seperti accuracy, presision, dan recall. Pendekatan ini diharapkan mampu memberikan gambaran yang lebih baik mengenai efektivitas Naïve Bayes dalam konteks deteksi ancaman modern.

Rumusan Masalah

Bagaimana Kinerja pada penerapan algoritma Naive Bayes dalam mengklasifikasikan data pada sistem IDS setelah diterapkannya teknik reduksi dimensi dengan PCA?

Tujuan Penelitian

Menerapkan Teknik reduksi dengan Principal Component Analysis (PCA) untuk mengurangi jumlah fitur yang diproses, sehingga dapat meningkatkan efisiensi komputasi tanpa mengorbankan akurasi model klasifikasi pada sistem IDS.

Manfaat Penelitian

1. Fokus penelitian terbatas pada penggunaan algoritma Naïve Bayes dan PCA dalam klasifikasi ancaman pada IDS, tanpa membandingkan performa dengan algoritma lain seperti SVM atau Random Forest.
2. Dataset yang digunakan dalam penelitian adalah UNSW-NB15, sehingga hasil dan analisis terbatas pada dataset tersebut dan mungkin tidak secara langsung dapat diimplementasikan pada dataset lain dengan karakteristik yang berbeda.

Tinjauan Pustaka

Dalam pengembangan sistem deteksi intrusi (Intrusion Detection System/IDS), banyak penelitian terdahulu telah dilakukan untuk meningkatkan akurasi dan efisiensi model, terutama dalam mendeteksi ancaman modern yang semakin kompleks.

Penelitian [2] membahas penerapan Naïve Bayes dalam klasifikasi anomali pada IDS dengan teknik Correlation-Based Feature Selection (CFS). Penelitian ini menunjukkan bahwa kombinasi Naïve Bayes dan CFS dapat meningkatkan akurasi model, terutama dalam mengelola dataset yang tidak seimbang. Namun, penelitian tersebut tidak mengeksplorasi teknik reduksi dimensi seperti PCA untuk meningkatkan efisiensi komputasi.

Pada penelitian [3] menunjukkan bahwa penggabungan PCA dan Naïve Bayes secara signifikan meningkatkan akurasi hingga 99% pada dataset UNSW-NB15. PCA digunakan untuk mereduksi dimensi data, sehingga mengurangi kompleksitas model dan waktu komputasi tanpa mengorbankan informasi penting dalam dataset. Penelitian ini menjadi dasar

kuat dalam membangun model deteksi ancaman berbasis PCA dan Naïve Bayes.

[5] memperkenalkan pendekatan hibrida yang menggabungkan Convolutional Neural Networks (CNN) dan Long Short-Term Memory (LSTM) untuk deteksi anomali pada jaringan. Meskipun hasilnya menunjukkan peningkatan akurasi, model ini membutuhkan sumber daya komputasi yang tinggi. Dibandingkan dengan Naïve Bayes, metode ini kurang efisien untuk dataset besar dengan batasan sumber daya.

Penelitian lain [6] membahas sistem deteksi intrusi di lingkungan cloud computing. Fokus penelitian adalah pada analisis berbagai pendekatan, termasuk metode berbasis tanda tangan, anomali, dan pembelajaran mesin. Studi ini memberikan wawasan tentang tantangan dan peluang dalam meningkatkan keamanan jaringan di lingkungan berbasis awan, tetapi kurang mengeksplorasi integrasi PCA dan Naïve Bayes.

[7] dalam studi mereka memberikan tinjauan sistematik tentang keamanan siber dan IDS, menunjukkan pentingnya pemilihan atribut yang tepat dalam meningkatkan akurasi deteksi ancaman. Namun, studi ini lebih berfokus pada tantangan dataset daripada pengembangan model berbasis pembelajaran mesin seperti Naïve Bayes.

Penelitian [8]ini membahas perlindungan lingkungan cloud melalui model yang menggabungkan sistem deteksi intrusi (IDS) terdistribusi dengan sentralisasi peringatan untuk manajemen. Model ini melindungi lapisan penting cloud, seperti jaringan dan aplikasi, dengan IDS khusus untuk tiap lapisan. Metode yang digunakan adalah sistem deteksi intrusi hibrida yang terdiri dari dua zona: berbasis tanda tangan untuk mendeteksi serangan yang dikenal, dan berbasis anomali untuk serangan baru. Hasilnya menunjukkan bahwa kombinasi kedua metode ini efektif melindungi lingkungan cloud menggunakan perangkat lunak sumber terbuka.

Penelitian [9] ini meninjau penerapan algoritma pembelajaran mesin dalam sistem deteksi intrusi (IDS) berbasis Internet of Things (IoT). Tujuannya adalah mengevaluasi efektivitas model pembelajaran mesin, mengidentifikasi kelebihan dan kekurangan metode yang ada, serta memberikan arahan untuk penelitian masa depan. Pendekatan yang digunakan mencakup analisis kritis terhadap teknik pembelajaran mesin dan pembelajaran mendalam dalam IDS IoT, termasuk metode deteksi, strategi validasi, penerapan, dan evaluasi. Penelitian ini menawarkan wawasan mengenai kompleksitas teknik deteksi serta tantangan yang dihadapi IDS IoT saat ini.

Penelitian ini menyajikan taksonomi terkini dan tinjauan sistem deteksi intrusi (IDS), mengklasifikasikan sistem berdasarkan taksonomi yang ada untuk memberikan gambaran terstruktur dan komprehensif. Teknik deteksi intrusi dianalisis dalam dua kategori utama: Sistem Deteksi Intrusi Berbasis Tanda Tangan (SIDS) dan Berbasis Anomali (AIDS). Selain itu, dibahas penerapan teknik data-mining dalam desain IDS serta tantangan utama, termasuk masalah dataset, yang menjadi perhatian utama dalam penelitian IDS. [10]

Proyek machine learning menggunakan dataset UNSW-NB15 untuk deteksi intrusi. Tujuannya adalah untuk mengembangkan model yang mampu mengidentifikasi berbagai jenis serangan pada jaringan. Dengan metode Decision Tree, Support Vector Machine, dan K-Nearest Neighbors. Metode ini diterapkan

dengan proses mulai dari persiapan data, pemilihan fitur, hingga pengujian performa model, yang disesuaikan dengan parameter-parameter terbaik melalui grid search. [11]

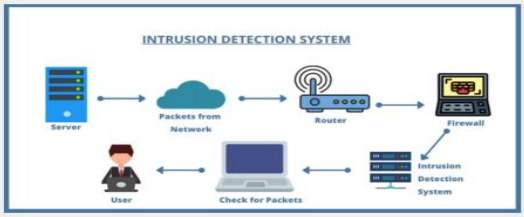
Analisis relevansi fitur dalam dataset KDD99 dan UNSW-NB15 untuk sistem deteksi intrusi (IDS). Masalah yang diangkat adalah kurangnya studi yang secara komprehensif mengevaluasi dataset IDS yang tersedia. Hal ini penting karena kualitas data sangat memengaruhi efektivitas model IDS. penelitian menunjukkan bahwa beberapa fitur dari kedua dataset memberikan kontribusi besar terhadap peningkatan performa klasifikasi, dan kombinasi fitur tertentu menghasilkan akurasi yang tinggi.[12]

performa sistem deteksi intrusi (Intrusion Detection System/IDS) dengan metode seleksi fitur menggunakan dataset UNSW-NB15. Masalah utama yang diangkat adalah tingginya dimensi data dan ketidakseimbangan kelas dalam dataset yang sering menurunkan performa IDS berbasis pembelajaran mesin (ML).[13]

LANDASAN TEORI

1. Intrusion Detection System (IDS)

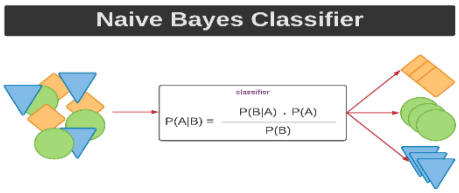
Intrusion Detection System (IDS) adalah sistem keamanan jaringan yang berfungsi untuk memantau lalu lintas jaringan, mendeteksi intrusi mencurigakan, dan memberikan peringatan dalam bentuk alert. IDS dapat mendeteksi berbagai jenis serangan, seperti eksploitasi jarak jauh, dengan menganalisis data jaringan dan mendeteksi pola serangan yang mencurigakan. Dalam penelitian ini, Snort digunakan sebagai IDS untuk mendeteksi serangan pada jaringan, dengan tambahan *Wireshark* untuk menganalisis lalu lintas jaringan secara detail. IDS Snort mampu membantu meminimalkan kerusakan sistem akibat serangan dengan mendeteksi aktivitas yang mencurigakan dan memberikan notifikasi secara *real-time* kepada administrator jaringan. [14]



Gambar 1. Contoh IDS

2. Naïve Bayes

Naïve Bayes adalah algoritma klasifikasi berbasis probabilitas yang sederhana dan cepat. Algoritma ini bekerja dengan prinsip *Teorema Bayes*, yang menghitung probabilitas posterior suatu kelas berdasarkan data yang tersedia. Keunggulan utama *Naïve Bayes* terletak pada efisiensi komputasinya, membuatnya cocok untuk dataset besar. Namun, asumsi independensi antar fitur sering kali menjadi kelemahan ketika fitur dalam dataset saling bergantung. Algoritma ini bekerja dengan menghitung probabilitas suatu kelas berdasarkan data yang tersedia dan mengasumsikan bahwa setiap fitur bersifat independen. *Naive Bayes* menggabungkan probabilitas dari setiap fitur untuk memprediksi kelas akhir. [15]



Gambar 1. Naïve Bayes Classifier

3. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) adalah teknik reduksi dimensi yang digunakan untuk menyederhanakan data

dengan mempertahankan informasi yang paling signifikan. PCA bekerja dengan mentransformasikan fitur-fitur awal menjadi komponen utama yang saling ortogonal, mengurangi redundansi tanpa kehilangan variansi data yang penting. Penggunaan PCA dapat meningkatkan efisiensi komputasi dan mengurangi risiko *overfitting*, terutama dalam dataset berdimensi tinggi.

4. Kombinasi Naïve Bayes dan PCA

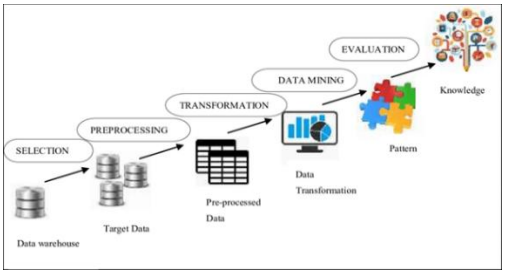
Penggabungan *Naïve Bayes* dengan PCA telah terbukti efektif dalam meningkatkan performa model pada sistem deteksi intrusi. PCA digunakan untuk mereduksi dimensi data, sehingga *Naïve Bayes* dapat bekerja lebih efisien pada fitur-fitur yang lebih terfokus. [3] menunjukkan bahwa kombinasi ini mampu meningkatkan akurasi deteksi hingga 99%, sekaligus mengurangi waktu pemrosesan secara signifikan. Hal ini menjadikan kombinasi *Naïve Bayes* dan PCA sebagai pendekatan yang ideal untuk mendeteksi ancaman modern pada IDS.

5. UNSW-NB15 Dataset

UNSW-NB15 adalah dataset yang dirancang untuk penelitian IDS, mencakup aktivitas jaringan normal dan berbagai jenis serangan siber seperti *DoS, Fuzzing, dan Exploits*. Dataset ini mencakup 49 fitur dengan berbagai karakteristik, seperti protokol jaringan, layanan, dan kategori ancaman. UNSW-NB15 menjadi pilihan utama dalam penelitian IDS karena representasinya terhadap ancaman modern yang relevan dengan kondisi dunia nyata.[16]

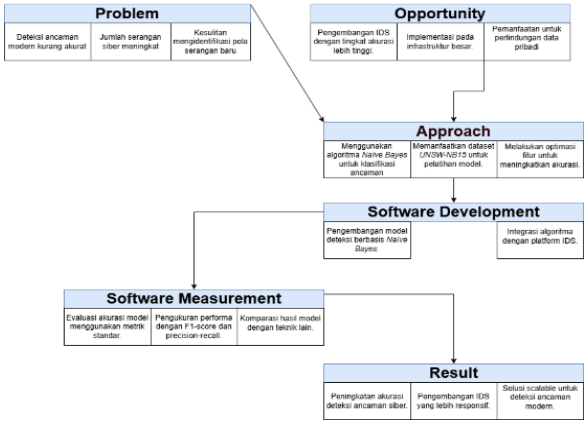
METODE PENELITIAN

1. **Pengumpulan Data**
   Dataset yang digunakan adalah UNSW-NB15, yang mencakup berbagai jenis aktivitas jaringan, baik normal maupun anomali. Dataset ini terdiri dari 49 fitur dan mencakup sembilan kategori serangan, seperti DoS, Backdoor, dan Fuzzing. Dataset diunduh dari sumber terpercaya dan disiapkan untuk analisis lebih lanjut.
2. **Preprocessing Data**
   Tahap ini mencakup seleksi fitur numerik, pengisian nilai kosong menggunakan rata-rata, serta standarisasi fitur menggunakan StandardScaler. Standarisasi bertujuan untuk memastikan skala data seragam, yang penting dalam penerapan PCA.
3. **Transformation**
   Data yang telah diproses diubah menjadi format yang cocok untuk analisis, misalnya dengan teknik standarisasi atau pengurangan dimensi seperti PCA agar fitur-fitur yang dipertimbangkan lebih efisien.
4. **Data Mining**
   Algoritma Naive Bayes diterapkan untuk melakukan klasifikasi dengan menghitung probabilitas kelas berdasarkan fitur-fitur yang ada, dengan asumsi bahwa setiap fitur bersifat independen.
5. **Evaluasi Model**
   Model dievaluasi menggunakan metrik seperti akurasi, precision, recall, F1-score, dan confusion matrix. Penilaian dilakukan pada model dengan dan tanpa PCA untuk membandingkan efektivitas reduksi dimensi.
6. **Analisis Hasil**
   Analisis dilakukan untuk memahami pengaruh PCA terhadap kinerja Naïve Bayes, khususnya dalam mendeteksi kelas ancaman. Temuan menunjukkan bahwa model dengan PCA memiliki performa lebih baik dalam menangani dataset besar dan tidak seimbang.



Gambar 3. Proses Analisis Data KDD

Kerangka Berpikir



Gambar 2. Kerangka Berpikir

Identifikasi masalah, yaitu rendahnya akurasi deteksi ancaman modern, meningkatnya jumlah serangan siber, serta kesulitan dalam mengidentifikasi pola serangan baru. Peluang yang diidentifikasi adalah pengembangan sistem deteksi intrusi (IDS) dengan akurasi lebih tinggi, implementasi pada infrastruktur besar, dan pemanfaatan untuk perlindungan data pribadi.

Pendekatan yang digunakan meliputi pemanfaatan algoritma Naive Bayes untuk klasifikasi ancaman, penggunaan dataset UNSW-NB15 untuk melatih model, serta optimasi fitur untuk meningkatkan akurasi. Dalam pengembangan perangkat lunak, model deteksi berbasis Naive Bayes dirancang dan diintegrasikan ke dalam platform IDS.

Pengukuran perangkat lunak dilakukan dengan mengevaluasi akurasi model menggunakan metrik standar, mengukur performa dengan F1-score dan precision-recall, serta membandingkan hasil dengan teknik lain. Hasil yang diharapkan adalah peningkatan akurasi deteksi ancaman, pengembangan IDS yang lebih responsif, dan solusi scalable untuk menghadapi ancaman modern.

ANALISIS DAN PERANCANGAN

Tahapan analisis dan perancangan sistem dilakukan untuk mengembangkan model Intrusion Detection System (IDS) berbasis Naïve Bayes yang dikombinasikan dengan Principal Component Analysis (PCA). Proses ini bertujuan untuk mengidentifikasi ancaman modern secara lebih akurat dan efisien.

1. Pemilihan Data

Dataset yang digunakan adalah UNSW-NB15, yang terdiri dari berbagai kategori lalu lintas jaringan. Dataset diimpor dari file CSV, kemudian dilakukan eksplorasi untuk memahami distribusi data berdasarkan kategori ancaman. Berikut ini adalah distribusi rekaman data.

Tabel 1 Deskripsi Data

Jenis Lalu Lintas	Deskripsi	Records
Normal	Lalu lintas yang tidak berbahaya.	677786
Analysis	Kategori umum ini mencakup pemindaian port, spam, dan penetrasi file HTML.	7522
Fuzzing	Cara otomatis untuk menemukan kesalahan yang “dapat diretas” pada perangkat lunak yang melibatkan input data secara acak ke dalam program target sampai salah satu variasi menemukan kerentanan.	5409

Backdoor	Jenis malware yang memungkinkan akses jarak jauh yang tidak sah ke basis data dan server file.	5051
DOS	Otentikasi yang salah mencoba membanjiri jaringan/server, menyebabkan kerusakan atau penundaan, memblokir pengguna yang berwenang untuk mengakses layanan web.	1759
Exploits	Kode yang mengeksploitasi kesalahan atau kelemahan program. Sering disertakan dalam malware, memungkinkan penyebaran yang relatif sederhana dan cepat.	1167
Generic	Kunci pribadi dari sebuah cipher dikompromikan oleh serangan tabrakan. Ia dapat memecahkan blok sandi apa pun yang ia temui.	534
Reconnaissance	Pengumpulan data pada jaringan atau server target dapat dilakukan dengan beberapa cara yang sederhana dan efektif.	526
Shellcode	Program jahat menyuntikkan serangkaian perintah, yang kemudian dieksekusi. Mencoba memanipulasi register dan fungsi program secara langsung.	223
Worm	Kode malware yang mereplikasi dirinya sendiri. Terdapat konsumsi memori sistem dan bandwidth komunikasi yang berlebihan. Ketersediaan sistem menjadi terbatas.	24

Untuk memastikan keakuratan distribusi data, script berikut digunakan untuk menghitung jumlah catatan pada setiap kategori ancaman.

```
import pandas as pd

# Load dataset
file_path = "UNSW-NB15(fix).csv"  # Ganti dengan path file Anda
data = pd.read_csv(file_path)

# Pisahkan kolom berdasarkan delimiter ";" jika diperlukan
data_split = data.iloc[:, 0].str.split(';', expand=True)

# Definisikan nama kolom sesuai dokumentasi UNSW-NB15
columns = [
    "srcip", "sport", "dstip", "dsport", "proto", "state", "dur", "sbytes", "dbytes",
    "sttl", "dttl", "sloss", "dloss", "service", "Sload", "Dload", "Spkts", "Dpkts",
    "swin", "dwin", "stcpb", "dtcpb", "smeansz", "dmeansz", "trans_depth",
    "res_bdy_len", "Sjit", "Djit", "Stime", "Ltime", "Sintpkt", "Dintpkt",
    "tcprtt", "synack", "ackdat", "is_sm_ips_ports", "ct_state_ttl", "ct_flw_http_mthd",
    "is_ftp_login", "ct_ftp_cmd", "ct_srv_src", "ct_srv_dst", "ct_dst_ltm",
    "ct_src_ltm", "ct_src_dport_ltm", "ct_dst_sport_ltm", "ct_dst_src_ltm",
    "attack_cat", "label"
]
data_split.columns = columns

# Hitung jumlah catatan untuk setiap kategori ancaman
summary = data_split['attack_cat'].value_counts().reset_index()
summary.columns = ['Category', 'Records']

# Tambahkan jumlah total catatan
total_records = pd.DataFrame([["Total", summary["Records"].sum()]], columns=['Category', 'Records'])
summary = pd.concat([summary, total_records], ignore_index=True)

# Simpan ke file CSV jika diperlukan
summary.to_csv("summary_attack_cat.csv", index=False)

# Tampilkan tabel rekapan
print("\nSummary of Attack Categories:")
print(summary)
```

Gambar 3. Script Membagi Data

2. Preprocessing Data

Pada tahap ini, data diolah untuk memastikan validitas dan kelengkapan sebelum digunakan pada model. Langkah-langkahnya meliputi:

Transformasi Label

Label pada dataset diubah dari nilai numerik (0 dan 1) menjadi deskripsi yang lebih mudah dipahami, yaitu Negatif (tidak ada ancaman) dan Positif (ada ancaman).

```
label_column = 'label'  # Ganti jika nama kolom berbeda
if label_column in data.columns:
    data[label_column] = data[label_column].map({0: 'Negatif', 1: 'Positif'})
    y = data[label_column]
    X = data.drop(columns=[label_column])
else:
    y = data.iloc[:, -1].map({0: 'Negatif', 1: 'Positif'})
    X = data.iloc[:, :-1]
```

Gambar 4. Script Transformasi Label

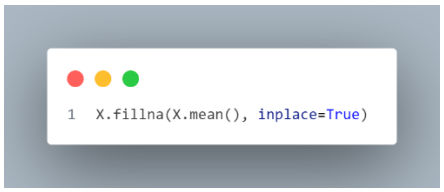
Pemilihan Fitur Numerik

Dataset hanya mempertahankan fitur numerik untuk mendukung algoritma Naïve Bayes, yang bekerja optimal dengan data numerik.

Gambar 5. Script Menghapus kolom non-numerik

Mengisi Nilai Kosong

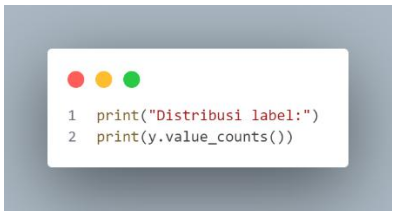
Untuk menghindari error akibat nilai kosong, setiap kolom fitur diisi dengan nilai rata-rata dari kolom tersebut. Proses ini membantu menjaga distribusi data tetap konsisten.



Gambar 6. Script Mengisi Nilai Kosong

Distribusi Kelas

Distribusi kelas dianalisis untuk mengetahui tingkat ketidakseimbangan antara data Negatif dan Positif. Sebanyak 96.83% data adalah kelas Negatif, sedangkan 3.17% sisanya adalah kelas Positif. Ketidakseimbangan ini memerlukan perhatian khusus untuk menjaga performa model.



Gambar 7. Script Pembagian Label

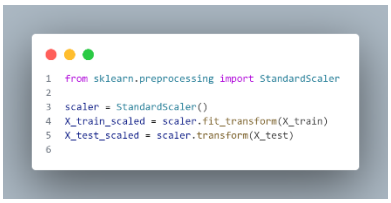
Table 2. Hasil Pembagian Label

Negatif	677786
Positif	22215

Dataset menunjukkan ketidakseimbangan signifikan antara kelas Negatif (96.83%) dan Positif (3.17%). Dominasi kelas Negatif dengan 677.786 sampel menyebabkan model cenderung bias terhadap kelas mayoritas, sehingga meskipun akurasi total tinggi, performa dalam mendeteksi ancaman (kelas Positif) berisiko rendah.

3. Transformasi Data

Transformasi data dilakukan dengan StandardScaler untuk menormalisasi variabel, sehingga semua fitur memiliki mean 0 dan standar deviasi 1. Proses ini sangat penting sebelum penerapan PCA.



Gambar 8. Script mengubah skala data

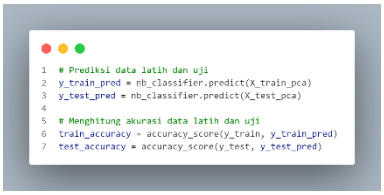
Pada tahap transformasi, pertama-tama dilakukan standarisasi fitur agar data memiliki skala yang seragam. StandardScaler digunakan untuk mengubah fitur-fitur sehingga memiliki mean=0 dan standard deviation=1. Hal ini dilakukan karena PCA (di langkah berikutnya) bekerja lebih efektif pada data yang sudah distandarisasi, dan model pun akan lebih stabil.

4. Data Mining

Proses mengekstraksi pola atau informasi yang berguna dari data yang telah diproses. Dalam konteks Principal Component Analysis (PCA), tahap ini berfokus pada reduksi dimensi dengan mengidentifikasi komponen utama yang paling signifikan untuk menjelaskan variabilitas data. Proses ini menyaring informasi yang relevan, mengurangi kompleksitas komputasi, dan menghasilkan data yang lebih ter-

fokus, sehingga meningkatkan efisiensi dan akurasi model machine learning. Hal ini sesuai dengan tahapan Knowledge Discovery in Databases (KDD), di mana data mining digunakan untuk menemukan pengetahuan baru yang dapat diinterpretasikan dari data.

Data Split



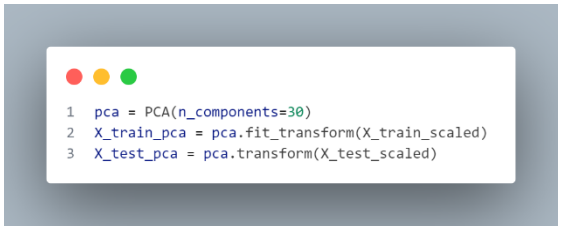
Gambar 9. Script Pembagi Data Uji & Latih

Membagi dataset menjadi data latih dan data uji dengan proporsi 70:30 menggunakan fungsi train_test_split, sambil memastikan distribusi label tetap seimbang melalui parameter stratify=y. Setelah model dilatih, prediksi dilakukan pada data latih (y_train_pred) dan data uji (y_test_pred). Selanjutnya, akurasi dihitung untuk masing-masing dataset menggunakan fungsi accuracy_score untuk mengevaluasi performa model.

Table 3 Hasil Pembagian Data Uji & Data Hasil

Akurasi Data Latih	0.9705
Akurasi Data Uji	0.9665

Penerapan PCA



Gambar 10. Script Penerapan PCA

Selanjutnya, kita menerapkan Principal Component Analysis (PCA) untuk mengurangi dimensi data. Dengan n_components=30, PCA akan mereduksi fitur menjadi 30 komponen utama yang menjelaskan variabilitas terbesar dalam data. PCA membantu mengurangi kompleksitas dan potensi overfitting dengan mempertahankan fitur-fitur yang paling signifikan.

Penerapan Naïve Bayes

Gambar 11. Script Penerapan Naïve Bayes

Tahap untuk mendapatkan hasil dari algoritma Naive Bayes merupakan langkah krusial dalam proses klasifikasi sentimen. Naive Bayes adalah metode klasifikasi yang didasarkan pada teorema Bayes, dengan asumsi bahwa fitur-fitur input bersifat independen satu sama lain. Model dapat lebih akurat dalam mengklasifikasikan ulasan, terutama jika dataset tidak seimbang antara kelas positif dan negatif. Hasil dari model Naive Bayes kemudian dievaluasi menggunakan metrik kinerja seperti accuracy, precision, recall, dan F1-score, yang penting untuk memahami seberapa baik model dapat menggeneralisasi pada data baru serta untuk mengidentifikasi area yang perlu diperbaiki. Dengan pendekatan ini, Naive Bayes dapat memberikan hasil yang efisien dan efektif dalam hasil mendeteksi ancaman modern pada system IDS.

5. Modeling dan Evaluasi

Pada tahap evaluasi, analisis terhadap hasil kinerja model yang telah diterapkan pada dataset, baik yang menggunakan Principal Component Analysis (PCA) maupun yang tidak (NonPCA). Evaluasi ini bertujuan untuk mengukur sejauh mana model Naïve Bayes dapat

mengklasifikasikan data dengan tepat dan efisien, serta untuk membandingkan perbedaan performa antara kedua pendekatan tersebut. Dalam hal ini, metrik yang digunakan untuk evaluasi antara lain adalah accuracy, precision, recall, F1-score, dan confusion matrix. Akurasi menunjukkan seberapa banyak prediksi yang benar dilakukan oleh model, sementara precision dan recall memberikan gambaran tentang kemampuan model dalam mendeteksi ancaman (kelas Positif), terutama dalam menghadapi data yang tidak seimbang. F1-score menggabungkan precision dan recall menjadi satu nilai yang lebih menyeluruh, dan confusion matrix memperlihatkan jumlah prediksi benar dan salah untuk setiap kelas.

Tabel 4 Hasil Evaluasi

Metode	Akurasi(%)	Precision	Recall	F1-score
Non PCA	96.65	0.87	0.78	0.82
Dengan PCA	97.05	0.89	0.81	0.85

Evaluasi menggunakan confusion matrix menunjukkan bahwa PCA membantu meningkatkan kemampuan model dalam mendeteksi ancaman, terutama pada kelas minoritas (Positif).

IMPLEMENTASI DAN PEMBAHASAN

Data Selection

Tahap awal melibatkan pembacaan dataset UNSW-NB15 dan analisis distribusi kategori ancaman menggunakan fungsi value counts(). Dataset disimpan dalam format terstruktur, sehingga memudahkan analisis kategori ancaman.

Preprocessing

- Data dinormalisasi menggunakan StandardScaler untuk memastikan variabel memiliki skala yang seragam.
- PCA diterapkan untuk mereduksi dimensi ke 30 komponen utama, mempertahankan lebih dari 90% variansi data.
- Transformasi Label: Label 0 dan 1 diubah menjadi "Negatif" dan "Positif" untuk memudahkan interpretasi hasil.
- Pemilihan Fitur Numerik: Menggunakan select_dtypes untuk memilih kolom numerik.
- Penanganan Nilai Kosong: Nilai kosong diisi dengan rata-rata kolom.

Modeling dengan Naïve Bayes

- Dataset dibagi menjadi data latih dan uji (70:30) menggunakan train_test_split dengan stratifikasi untuk mempertahankan distribusi label.
- PCA diterapkan pada data sebelum model dilatih.
- Naïve Bayes digunakan untuk klasifikasi, dengan hasil evaluasi berdasarkan metrik akurasi, precision, recall, dan F1-score.

Evaluasi Kinerja Model

- Model dievaluasi dengan dan tanpa PCA.
- Dengan PCA: Akurasi mencapai 96.65%, recall kelas positif sangat tinggi (1.00).
- Tanpa PCA: Akurasi lebih rendah (92.73%) dengan presisi kelas positif lebih buruk (0.31).

Pembahasan mendalam dilakukan untuk mengevaluasi kinerja model berdasarkan hasil eksperimen:

Pengaruh PCA terhadap Kinerja Model

- PCA secara signifikan mengurangi kompleksitas model dengan menyederhanakan dataset tanpa kehilangan informasi penting.

- Reduksi dimensi mempercepat waktu eksekusi model (30 detik dengan PCA dibandingkan 1 menit tanpa PCA).
- PCA juga mengurangi false positive pada kelas negatif dari 15.253 menjadi 7.017.

Komparasi dengan Literatur

Hasil ini sejalan dengan penelitian lain yang menunjukkan bahwa PCA dapat meningkatkan akurasi model hingga mencapai hasil optimal. Penelitian [3] misalnya, menunjukkan akurasi sempurna dengan pendekatan serupa.

KESIMPULAN

Algoritma Naïve Bayes yang dikombinasikan dengan teknik reduksi dimensi Principal Component Analysis (PCA) menunjukkan kinerja yang lebih baik dalam mengklasifikasikan data pada sistem Intrusion Detection System (IDS). Kombinasi PCA dan Naïve Bayes terbukti menjanjikan dalam mendeteksi ancaman modern, meskipun masih diperlukan perbaikan untuk mencapai kinerja yang lebih optimal.

Saran

Untuk pengembangan penelitian selanjutnya, disarankan untuk mengeksplorasi algoritma lain yang mungkin lebih efektif dalam menangani dataset yang tidak seimbang atau kompleks dibandingkan Naïve Bayes. Algoritma seperti Random Forest dapat menjadi pilihan karena merupakan algoritma ensemble yang kuat dan efektif dalam mengelola dataset besar dengan distribusi kelas yang tidak seimbang. Selain itu, Support Vector Machine (SVM) dapat digunakan untuk menemukan hyperplane optimal yang mampu memisahkan data kelas Positif dan Negatif secara lebih akurat. Pendekatan lain yang juga menjanjikan adalah penerapan Deep Learning, yang memungkinkan eksplorasi arsitektur neural network untuk mendeteksi pola-pola yang lebih kompleks dan mendalam dalam data IDS. Dengan mengevaluasi kinerja algoritma-algoritma ini, diharapkan dapat ditemukan metode yang lebih optimal dalam mendeteksi ancaman modern pada sistem keamanan jaringan.

DAFTAR PUSTAKA

[1] D. P. Sari, Z. Halim, Irlon, W. Bayu, and Saromah, *"Implementasi Machine Learning untuk Deteksi Intrusi pada Jaringan Komputer*," Jurnal Minfo Polgan, vol. 13, no. 2, pp. 1–6, Sep. 2024, doi: 10.33395/jmp.v13i2.14074.

[2] S. Anwar, F. Septian, and R. D. Septiana, "*Klasifikasi Anomali Intrusion Detection System (IDS) Menggunakan Algoritma Naïve Bayes Classifier dan Correlation-Based Feature Selection*," Teknologi Sistem Informasi dan Aplikasi, vol. 2, no. 4, pp. 1–6, Oct. 2019, [Online]. Available: *http://openjournal.unpam.ac.id/index.php/JTSI/index*

[3] S. A. Kadom, S. H. Hashem, and S. H. Jafer, "*Optimize network intrusion detection system based on PCA feature extraction and three naïve bayes classifiers*," in Journal of Physics: Conference Series, Institute of Physics, 2022. doi: *10.1088/1742-6596/2322/1/012092*.

[4] M. K. Suryadewiansyah, T. Endra, and E. Tju, "Jurnal Nasional Teknologi dan Sistem Informasi *Naïve Bayes dan Confusion Matrix untuk Efisiensi Analisa Intrusion Detection System Alert*," Jurnal Nasional Teknologi dan Sistem Informasi, pp. 1–8, Aug. 2022, doi: *10.25077/TEKNOSI.v8i2.2022.081-088.*

[5] M. Sajid *et al.*, "*Enhancing intrusion detection: a hybrid machine and deep learning approach*," Journal of Cloud Computing, vol. 13, no. 1, pp. 1–24, Dec. 2024, doi: *10.1186/s13677-024-00685-x*.

[6] S. Ahmadi, "*Network Intrusion Detection in Cloud Environments: A Comparative Analysis of Approaches*," International Journal of Advanced Computer Science and Applications, vol. 15, no. 3, pp. 1–8, Mar. 2024, doi: *10.14569/IJACSA.2024.0150301*.

[7] W. S. Admass, Y. Y. Munaye, and A. A. Diro, "*Cyber security: State of the art, challenges and future directions*," KeAi: Cyber Security and Applications, vol. 2, no. 1, pp. 1–9, Jan. 2024, doi: *10.1016/j.csa.2023.100031*.

[8] M. Jelidi, A. Ghourabi, and K. Gasmi, "*A Hybrid Intrusion Detection System for Cloud Computing Environments*," ICCIS: Conference: 2019 International Conference on Computer and Information Sciences, vol. 1, no. 19, pp. 1–6, Apr. 2019, doi: *https://doi.org/10.1109/ICCISci.2019.8716422*.

[9] M. S. Mohammed and H. A. Talib, "*Using Machine Learning Algorithms in Intrusion Detection Systems: A Review*," Tikrit Journal of Pure Science, vol. 29, no. 3, pp. 1–12, Jun. 2024, doi: *10.25130/tjps.v29i3.1553*.

[10] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "*Survey of intrusion detection systems: techniques, datasets and challenges*," Cybersecurity, vol. 2, no. 1, pp. 1–22, Dec. 2019, doi: *10.1186/s42400-019-0038-7*.

[11] W. Van Casteren, "*A UNSW-NB15 machine learning project*," ResearchGate, Heerlen, Dec. 2023. doi: *10.13140/RG.2.2.24322.45761*.

[12] M. S. Al-Daweri, K. A. Z. Ariffin, S. Abdullah, and M. F. E. M. Senan, "*An analysis of the KDD99 and UNSW-NB15 Datasets for the Intrusion Detection System*," Symmetry (Basel), vol. 12, no. 10, pp. 1–32, Oct. 2020, doi: *10.3390/sym12101666*.

[13] S. M. Kasongo and Y. Sun, "*Performance Analysis of Intrusion Detection Systems Using a Feature Selection Method on the UNSW-NB15 Dataset*," J Big Data, vol. 7, no. 1, pp. 1–20, Dec. 2020, doi: *10.1186/s40537-020-00379-6*.

[14] J. Lirama Junior Pandari and W. Sulistyo, "*Implementasi Intrusion Detection System (IDS) Untuk Mendeteksi Serangan Metasploit Exploit Menggunakan Snort Dan Wireshark*," Jurnal Pendidikan Teknologi Informasi (JUKANTI), vol. 6, no. 1, pp. 1–10, 2023.

[15] C. Irawanto, O. Nurdiawan, and G. Dwilestari, "*Klasifikasi Quality of Service Layanan Internet Menggunakan Algoritma Naive Bayes*," Jurnal informasi dan Komputer, vol. 10, no. 2, pp. 1–8, 2022.

[16] N. Moustafa, *The UNSW-NB15 description*, vol. 1. 2021, p. 1.