*Analisis Pola Traffic Jaringan Menggunakan Algoritma K-Means dengan Penentuan Jumlah Cluster Otomatis*

*Gilang Aldiano*1*.*

*‘Teknik Informatika, universitas komputer indonesia”*

|  |  |
| --- | --- |
| Kata Kunci  K-Means, Auto-K, UNSW-NB15, Clustering Jaringan, Unsupervised Machine Learning. | **Abstrak**  Traffic jaringan modern mengandung potensi ancaman siber yang semakin kompleks, sementara pada banyak kondisi nyata data trafik tidak selalu tersedia dalam bentuk berlabel. Hal ini membatasi penerapan pendekatan supervised learning pada sistem keamanan jaringan. Oleh karena itu, teknik clustering sebagai metode unsupervised learning menjadi relevan untuk mengeksplorasi pola trafik jaringan tanpa ketergantungan terhadap label.Penelitian ini bertujuan untuk menganalisis pola trafik jaringan pada dataset UNSW-NB15 menggunakan algoritma K-Means dengan penentuan jumlah cluster secara otomatis (Auto-K). Berbeda dari penelitian sebelumnya yang menggunakan K-Means terutama untuk tujuan visualisasi data, penelitian ini menerapkan K-Means sebagai metode analisis utama dengan penentuan jumlah cluster berbasis kombinasi Elbow Method dan Silhouette Score.  Hasil eksperimen menunjukkan bahwa jumlah cluster optimal yang diperoleh adalah dua cluster (K = 2). Analisis karakteristik cluster menunjukkan perbedaan pola statistik trafik jaringan yang signifikan dan berkorelasi dengan distribusi trafik normal dan serangan, meskipun label tidak digunakan dalam proses pembentukan cluster. Pendekatan ini diharapkan dapat menjadi dasar eksplorasi pola trafik jaringan dan mendukung pengembangan sistem pengawasan jaringan berbasis unsupervised learning. |

1. **PENDAHULUAN**
   1. **Latar Belakang**

Pertumbuhan teknologi jaringan dan internet telah membawa peningkatan volume dan kompleksitas trafik jaringan. Sementara itu, kegiatan berbahaya seperti serangan siber juga semakin berkembang baik dari segi teknik maupun pola pelaksanaan. Sistem Intrusion Detection System (IDS) berperan penting dalam mendeteksi anomali atau kejanggalan lalu lintas jaringan yang dapat mengindikasikan aktivitas berbahaya. Kebanyakan pendekatan IDS menggunakan model supervised learning yang memerlukan data berlabel, sedangkan banyak kasus nyata tersedianya data trafik tidak berlabel atau mengandung pola baru yang belum pernah dilabeli.

Dalam konteks tersebut, pendekatan unsupervised machine learning menjadi semakin menarik karena mampu menemukan pola atau struktur data tanpa membutuhkan label. Clustering adalah salah satu teknik unsupervised yang umum digunakan untuk mengelompokkan data berdasarkan kemiripan karakteristik, sehingga data dalam satu kelompok memiliki pola yang relatif serupa.

Algoritma K-Means merupakan metode clustering non-hierarkis yang banyak digunakan karena kesederhanaan, efisiensi, dan kemampuannya mengelompokkan data numerik berdimensi tinggi. Namun, kelemahan utama K-Means terletak pada kebutuhan untuk menentukan jumlah cluster (K) sebelum proses clustering, yang sering kali bersifat subjektif dan bergantung pada pengalaman peneliti.

Untuk mengurangi subjektivitas tersebut, penelitian ini menerapkan pendekatan Auto-K, yaitu penentuan jumlah cluster secara otomatis menggunakan kombinasi Elbow Method dan Silhouette Score. Pendekatan ini diharapkan menghasilkan jumlah cluster yang lebih objektif dan dapat dipertanggungjawabkan secara ilmiah.

Dataset yang digunakan adalah UNSW-NB15, salah satu benchmark dataset yang dikembangkan untuk evaluasi IDS, berisi trafik jaringan normal dan berbagai jenis serangan. Dataset ini sangat cocok untuk penelitian clustering karena karakter fitur yang kompleks dan variasi pola trafik.

* 1. **Rumusan Masalah**

Berdasarkan latar belakang tersebut, rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana penerapan algoritma K-Means untuk mengelompokkan data trafik jaringan dari dataset UNSW-NB15 secara unsupervised?
2. Berapa jumlah cluster optimal yang diperoleh dengan menggunakan pendekatan Auto-K?
3. Bagaimana karakteristik cluster trafik jaringan yang terbentuk berdasarkan fitur dataset?

**1.3 Tujuan Penelitian**

adapun tujuan dari penelitian ini adalah:

1. Menerapkan algoritma K-Means untuk mengelompokkan trafik jaringan dari dataset UNSW-NB15.
2. Menentukan jumlah cluster optimal secara otomatis menggunakan metode Elbow dan Silhouette Score.
3. Menganalisis karakteristik cluster yang terbentuk sebagai representasi pola trafik jaringan.

**1.4 Manfaat Penelitian**

Manfaat yang diharapkan dari penelitian ini adalah sebagai berikut:

1. **Manfaat Akademis**

Penelitian ini diharapkan dapat menambah referensi dalam bidang machine learning, khususnya penerapan K-Means clustering untuk eksplorasi pola data trafik jaringan.

1. **Manfaat Praktis**

Hasil clustering dapat digunakan sebagai informasi awal untuk mengidentifikasi pola trafik yang mungkin berkaitan dengan aktivitas anomali atau serangan jaringan, mendukung riset lanjutan di bidang keamanan jaringan.

1. **Manfaat bagi Penelitian Selanjutnya**

Penelitian ini dapat menjadi dasar atau referensi bagi penelitian selanjutnya yang berkaitan dengan clustering, unsupervised learning, atau pengembangan sistem pendeteksian anomali jaringan.

**1.5 Metode dan Algoritma yang Digunakan**

Penelitian ini menggunakan pendekatan machine learning unsupervised dengan teknik clustering untuk mengeksplorasi pola data trafik jaringan. Algoritma utama yang digunakan adalah K-Means Clustering, karena efektif dalam mengolah data numerik berdimensi tinggi dan relatif efisien secara komputasi.

Data pertama kali melalui tahap pra-pemrosesan, termasuk pemilihan subset fitur numerik, penanganan nilai yang hilang (jika ada), dan normalisasi data menggunakan StandardScaler agar setiap fitur berada pada skala yang sebanding. Selanjutnya, pendekatan Auto-K diterapkan dengan mengombinasikan Elbow Method dan Silhouette Score untuk menentukan jumlah cluster optimal. Setelah nilai K optimal diperoleh, proses K-Means dijalankan untuk menghasilkan label cluster masing-masing data. Hasil clustering kemudian dievaluasi menggunakan Silhouette Score dan dianalisis untuk mengeksplorasi pola trafik jaringan pada setiap cluster.

**1.6 Solusi yang Diusulkan (Penyelesaian Masalah)**

Permasalahan utama dalam penelitian ini adalah ketergantungan pada label dalam pendekatan supervised dan subjektivitas dalam pemilihan jumlah cluster pada algoritma K-Means. Solusi yang diusulkan adalah penggunaan K-Means clustering dengan penentuan jumlah cluster secara otomatis (Auto-K). Pendekatan ini memungkinkan pengelompokan data trafik jaringan tanpa label, serta pemilihan jumlah cluster yang objektif berdasarkan metrik evaluasi internal.

Dengan demikian, hasil clustering diharapkan membantu dalam pemahaman struktur dasar data trafik jaringan, yang dapat digunakan sebagai langkah awal dalam pengembangan sistem pendeteksian anomali atau serangan jaringan berbasis machine learning.

# TINJAU PUNTAKA

### Machine Learning

Machine learning merupakan salah satu cabang dari kecerdasan buatan (Artificial Intelligence) yang memungkinkan sistem komputer untuk belajar dari data dan membuat keputusan atau prediksi tanpa diprogram secara eksplisit. Berdasarkan cara pembelajarannya, machine learning dibagi menjadi tiga kategori utama, yaitu supervised learning, unsupervised learning, dan reinforcement learning.

Pada supervised learning, model dilatih menggunakan data berlabel, sedangkan pada unsupervised learning, model bekerja tanpa label dan berfokus pada pencarian pola atau struktur tersembunyi dalam data. Penelitian ini menggunakan pendekatan unsupervised learning karena dataset trafik jaringan pada kondisi nyata sering kali tidak memiliki label yang lengkap atau akurat.

### 2.2 Unsupervised Learning

Unsupervised learning adalah metode pembelajaran mesin yang bertujuan untuk menemukan pola, struktur, atau hubungan dalam data tanpa menggunakan label kelas. Teknik ini sangat berguna untuk eksplorasi data awal, pengelompokan, dan deteksi anomali.

Dalam konteks analisis trafik jaringan, unsupervised learning sering digunakan untuk:

1. Mengelompokkan pola trafik normal dan tidak normal
2. Mengidentifikasi anomali atau serangan baru yang belum pernah didefinisikan sebelumnya
3. Mengurangi ketergantungan terhadap data berlabel

Salah satu teknik unsupervised learning yang paling umum digunakan adalah clustering.

### Clustering

Clustering merupakan teknik pengelompokan data berdasarkan tingkat kemiripan tertentu sehingga data dalam satu kelompok (cluster) memiliki karakteristik yang mirip, sedangkan data antar cluster memiliki perbedaan yang signifikan. Clustering banyak digunakan dalam berbagai bidang seperti data mining, pengenalan pola, bioinformatika, dan keamanan jaringan.

Beberapa metode clustering yang umum digunakan antara lain:

1. K-Means
2. Hierarchical Clustering
3. DBSCAN

Dalam penelitian ini, metode yang digunakan adalah K-Means clustering karena efisien untuk data berukuran besar dan fitur numerik.

### Algoritma K-Means

K-Means adalah algoritma clustering non-hierarkis yang bertujuan untuk membagi data ke dalam K cluster berdasarkan jarak terdekat terhadap centroid. Algoritma ini bekerja secara iteratif dengan tahapan sebagai berikut:

* Menentukan jumlah cluster (K)
* Menginisialisasi centroid secara acak
* Menghitung jarak setiap data ke centroid
* Mengelompokkan data ke centroid terdekat
* Memperbarui centroid berdasarkan rata-rata anggota cluster
* Mengulangi proses hingga konvergen
* Kelebihan K-Means antara lain:
* Sederhana dan mudah diimplementasikan
* Efisien secara komputasi
* Cocok untuk data numerik berdimensi tinggi

Namun, K-Means memiliki kelemahan utama yaitu penentuan jumlah cluster (K) yang harus ditentukan di awal dan sering bersifat subjektif.

**2.5 Penentuan Jumlah Cluster (Auto-K)**

Pemilihan jumlah cluster yang tidak tepat dapat menghasilkan pengelompokan yang kurang optimal. Oleh karena itu, penelitian ini menggunakan pendekatan penentuan jumlah cluster otomatis (Auto-K) dengan mengombinasikan dua metode evaluasi internal, yaitu Elbow Method dan Silhouette Score.

**2.5.1 Elbow Method**

Elbow Method digunakan untuk menentukan jumlah cluster optimal berdasarkan nilai Sum of Squared Error (SSE) atau inertia. Nilai SSE akan menurun seiring bertambahnya jumlah cluster, namun penurunan tersebut akan melambat pada titik tertentu yang disebut sebagai “titik siku” (elbow point). Titik inilah yang dianggap sebagai jumlah cluster optimal.

**2.5.2 Silhouette Score**

Silhouette Score digunakan untuk mengukur kualitas clustering dengan mempertimbangkan tingkat kedekatan data dalam satu cluster dan jarak antar cluster. Nilai Silhouette berada pada rentang −1 hingga 1, di mana nilai mendekati 1 menunjukkan hasil clustering yang baik.

Dengan mengombinasikan Elbow Method dan Silhouette Score, pemilihan nilai K menjadi lebih objektif dan tidak bergantung pada intuisi peneliti semata.

**2.6 Dataset UNSW-NB15**

UNSW-NB15 merupakan dataset trafik jaringan yang dikembangkan oleh Australian Centre for Cyber Security (ACCS) di University of New South Wales. Dataset ini dirancang sebagai pengganti dataset lama seperti KDD99 dan NSL-KDD yang dinilai sudah tidak merepresentasikan kondisi serangan jaringan modern.

Dataset UNSW-NB15 berisi:

1. Trafik jaringan normal
2. Berbagai jenis serangan seperti DoS, Exploits, Fuzzers, Reconnaissance, dan lainnya
3. Fitur numerik dan kategorikal yang merepresentasikan karakteristik trafik jaringan

Dataset ini sangat cocok digunakan untuk penelitian clustering karena kompleksitas fitur dan variasi pola trafik yang tinggi.

**2.7 Penelitian Terkait**

Beberapa penelitian sebelumnya telah menerapkan metode clustering pada dataset UNSW-NB15. Sebagian besar penelitian menggunakan K-Means atau metode clustering lain untuk mengelompokkan trafik jaringan sebagai tahap awal deteksi anomali. Namun, banyak penelitian masih menentukan jumlah cluster secara manual atau hanya menggunakan satu metode evaluasi.

Penelitian ini berbeda dari penelitian sebelumnya karena:

1. Menggunakan pendekatan Auto-K dengan kombinasi Elbow Method dan Silhouette Score
2. Berfokus pada eksplorasi pola trafik jaringan secara unsupervised
3. Menyediakan analisis cluster sebagai dasar interpretasi struktur data trafik

Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan metode clustering yang lebih objektif dan aplikatif untuk analisis trafik jaringan.

**2.7.1 Penelitian Zoghi dan Serpen (2021)**

Zoghi dan Serpen (2021) melakukan analisis terhadap dataset UNSW-NB15 menggunakan algoritma K-Means yang dikombinasikan dengan teknik reduksi dimensi seperti Principal Component Analysis (PCA) dan t-SNE. Penelitian tersebut bertujuan untuk melakukan eksplorasi dan visualisasi struktur data UNSW-NB15 guna memahami kompleksitas dan distribusi trafik jaringan.W

Dalam penelitian tersebut, K-Means digunakan sebagai alat bantu visualisasi dan eksplorasi data, tanpa pembahasan mendalam mengenai penentuan jumlah cluster secara objektif maupun interpretasi cluster sebagai representasi pola trafik jaringan. Selain itu, penentuan nilai K tidak menjadi fokus utama penelitian.

Berbeda dengan pendekatan tersebut, penelitian ini menempatkan K-Means sebagai metode analisis utama dengan penentuan jumlah cluster secara otomatis (Auto-K) dan fokus pada interpretasi karakteristik cluster sebagai pola trafik jaringan.

# Gap Analisis

**3.1 Penelitian Terdahulu**

Beberapa penelitian sebelumnya telah menerapkan algoritma K-Means pada dataset UNSW-NB15 untuk analisis trafik jaringan dan keamanan siber. Umumnya, penelitian tersebut menggunakan K-Means sebagai metode pendukung untuk eksplorasi data atau sebagai tahap awal sebelum klasifikasi serangan.

Namun demikian, sebagian besar penelitian terdahulu masih memiliki keterbatasan, khususnya pada aspek penentuan jumlah cluster (K) yang umumnya ditentukan secara manual atau berdasarkan asumsi peneliti.

**3.2 Identifikasi Kesenjangan**

Tabel berikut menyajikan perbandingan antara penelitian terdahulu yang relevan dengan penelitian yang diusulkan, khususnya pada penerapan algoritma K-Means pada dataset UNSW-NB15.

|  |  |  |
| --- | --- | --- |
| **Aspek** | **Penelitian Sebelumnya** | **Penelitian Ini** |
| Dataset | UNSW-NB15 | UNSW-NB15 |
| Jenis Pembelajaran | Unsupervised Learning | Unsupervised Learning |
| Metode Clustering | K-Means standar | K-Means dengan pendekatan Auto-K |
| Penentuan Jumlah Cluster | Ditentukan secara manual atau berdasarkan asumsi peneliti | Ditentukan secara otomatis menggunakan Elbow Method dan Silhouette Score |
| Evaluasi Kualitas Cluster | Terbatas atau tidak konsisten, sering kali hanya visualisasi | Menggunakan metrik internal Silhouette Score |
| Pemanfaatan Label | Sering digunakan dalam interpretasi atau pembentukan cluster | Tidak digunakan dalam proses clustering, hanya untuk analisis interpretatif |
| Reproducibility | Implementasi tidak selalu disertakan atau sulit direplikasi | Disediakan dalam bentuk notebook Python yang reproducible |

Penelitian ini menutup kesenjangan metodologis dan empiris melalui penggunaan Auto-K dan data terbaru.

**3.3 Kontribusi dan Novelty Penelitian**

Berdasarkan kesenjangan yang telah diidentifikasi, kontribusi utama dan novelty penelitian ini adalah:

1. Menerapkan pendekatan Auto-K pada dataset UNSW-NB15 untuk menentukan jumlah cluster optimal secara objektif.
2. Mengombinasikan Elbow Method dan Silhouette Score dalam proses evaluasi cluster.
3. Menyediakan pipeline clustering UNSW-NB15 yang reproducible menggunakan Python.
4. Menyajikan analisis karakteristik cluster trafik jaringan tanpa memanfaatkan label serangan.

Dengan kontribusi tersebut, penelitian ini memperkuat pendekatan unsupervised learning pada domain keamanan jaringan.

1. **METODOLOGI PENELITIAN**

**4.1 Desain Penelitian**

Penelitian ini menggunakan pendekatan eksperimental kuantitatif dengan metode unsupervised machine learning. Fokus utama penelitian adalah melakukan pengelompokan data trafik jaringan menggunakan algoritma K-Means tanpa memanfaatkan label kelas.

**4.2 Dataset Penelitian**

**4.2.1 Sumber Dataset**

Dataset yang digunakan adalah UNSW-NB15, yang dikembangkan oleh Australian Centre for Cyber Security (ACCS), University of New South Wales.

Sumber resmi dataset:

https://research.unsw.edu.au/projects/unsw-nb15-dataset

Dataset ini dipilih karena:

* 1. Mewakili trafik jaringan modern
  2. Banyak digunakan dalam penelitian internasional
  3. Cocok untuk eksperimen clustering

**4.2.2 Deskripsi Dataset**

Dataset UNSW-NB15 terdiri dari data trafik jaringan dengan berbagai fitur numerik yang merepresentasikan karakteristik koneksi jaringan. Pada penelitian ini, label serangan tidak digunakan karena fokus penelitian adalah clustering secara unsupervised.

**4.3 Pra-pemrosesan Data**

Tahapan pra-pemrosesan dilakukan untuk memastikan kualitas data sebelum diterapkan algoritma clustering. Tahapan ini meliputi:

1. **Seleksi fitur numerik** yang relevan untuk algoritma K-Means
2. **Pembersihan data**, termasuk penanganan nilai kosong atau tidak valid menggunakan imputasi median
3. **Normalisasi data** menggunakan *StandardScaler*

Normalisasi dilakukan untuk memastikan setiap fitur berada pada skala yang seragam sehingga tidak terjadi dominasi fitur tertentu dalam perhitungan jarak berbasis Euclidean.

Secara matematis, proses normalisasi menggunakan *StandardScaler* dirumuskan sebagai berikut:

dengan:

* = nilai fitur asli
* = rata-rata fitur
* = standar deviasi fitur
* = nilai fitur hasil normalisasi

**4.4 Penentuan Jumlah Cluster Otomatis (Auto-K)**

Untuk menghindari subjektivitas dalam pemilihan jumlah cluster, penelitian ini menerapkan pendekatan Auto-K, yang mengombinasikan Elbow Method dan Silhouette Score.

**4.4.1 Elbow Method**

Elbow Method digunakan untuk menentukan jumlah cluster optimal berdasarkan nilai Sum of Squared Error (SSE) atau inertia. SSE mengukur total jarak kuadrat antara setiap data dan centroid cluster-nya.

Rumus SSE adalah sebagai berikut:

dengan:

* = jumlah cluster
* = himpunan data pada cluster ke-
* = data ke-
* = centroid cluster ke-

Nilai SSE akan menurun seiring bertambahnya jumlah cluster. Jumlah cluster optimal ditentukan pada titik di mana penurunan SSE mulai melandai (*elbow point*).

**4.4.2 Silhouette Score**

Silhouette Score digunakan untuk mengevaluasi kualitas pemisahan cluster dengan mempertimbangkan jarak intra-cluster dan inter-cluster. Nilai Silhouette untuk setiap data dirumuskan sebagai berikut:

dengan:

* = rata-rata jarak data ke- terhadap data lain dalam cluster yang sama
* = jarak rata-rata minimum data ke- terhadap cluster terdekat lainnya

Nilai Silhouette berada pada rentang . Nilai yang mendekati 1 menunjukkan kualitas clustering yang baik, sedangkan nilai mendekati 0 menunjukkan data berada pada batas antar cluster.

Jumlah cluster optimal dipilih berdasarkan **konsistensi hasil Elbow Method dan Silhouette Score**.

**4.5 Proses Clustering Menggunakan K-Means**

Setelah jumlah cluster optimal diperoleh, proses clustering dilakukan menggunakan algoritma K-Means. Algoritma ini bertujuan untuk meminimalkan jarak antara data dan centroid cluster berdasarkan jarak Euclidean.

Rumus jarak Euclidean antara data dan centroid didefinisikan sebagai:

Centroid cluster diperbarui menggunakan rata-rata data dalam cluster:

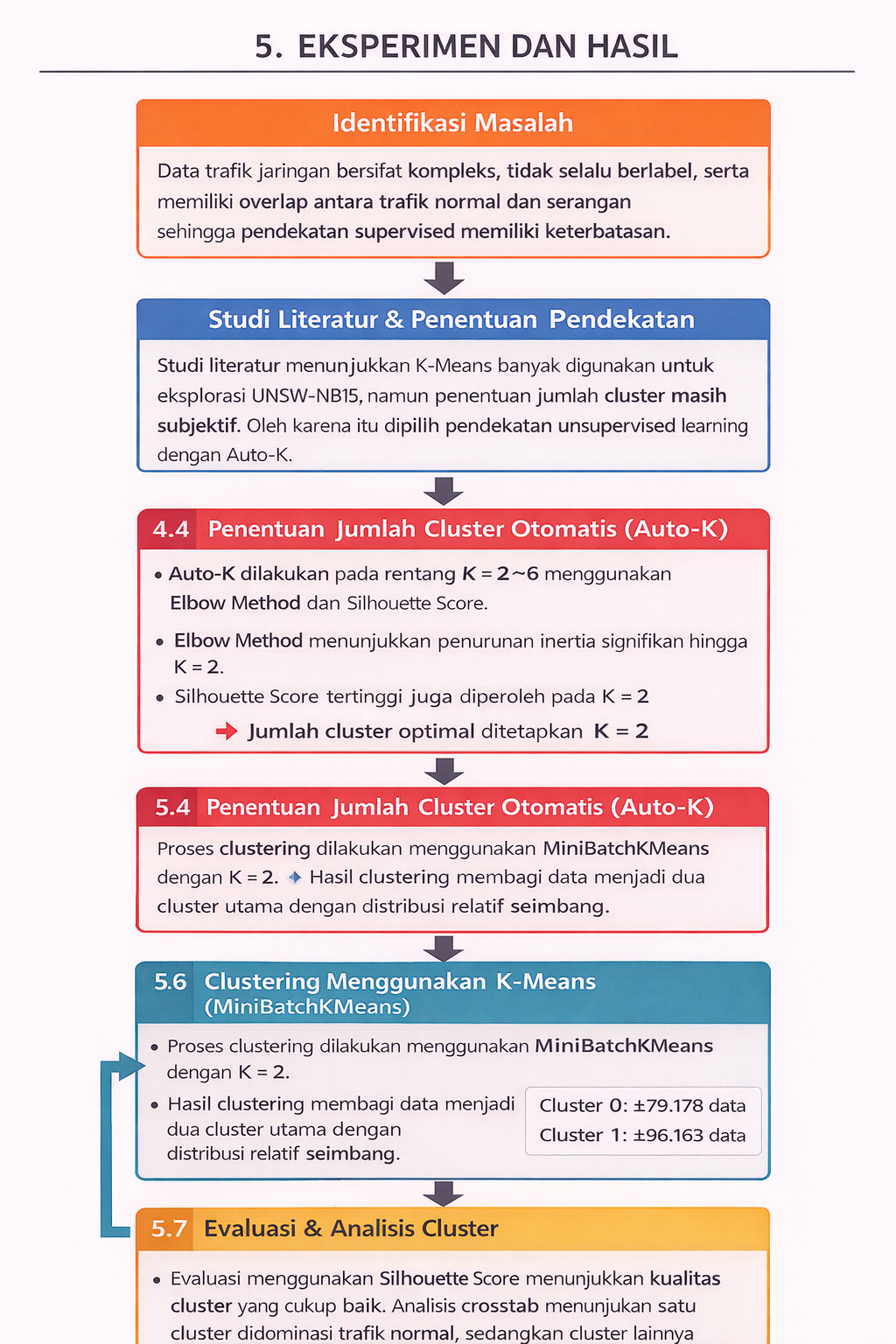
Proses iterasi dilakukan hingga posisi centroid stabil atau perubahan nilai centroid sangat kecil.

Pada penelitian ini digunakan varian MiniBatchKMeans untuk meningkatkan efisiensi komputasi, mengingat ukuran dataset UNSW-NB15 yang besar. MiniBatchKMeans memproses data dalam batch kecil sehingga mempercepat konvergensi tanpa mengurangi kualitas hasil clustering secara signifikan.

**4.6 Evaluasi Hasil Clustering**

Evaluasi hasil clustering dilakukan menggunakan Silhouette Score. Selain itu, dilakukan analisis deskriptif terhadap masing-masing cluster untuk mengidentifikasi pola trafik jaringan yang terbentuk.

Selain evaluasi internal menggunakan Silhouette Score, dilakukan pula analisis distribusi label serangan terhadap cluster sebagai interpretasi hasil clustering tanpa mempengaruhi proses pembentukan cluster.

**4.7 Skema Model**

1. **EKSPERIMEN DAN HASIL**
   1. **Eksperimen**

Pada tahap eksperimen, dilakukan pengolahan data trafik jaringan yang bersumber dari dataset UNSW-NB15. Dataset ini merepresentasikan lalu lintas jaringan yang terdiri dari trafik normal dan berbagai jenis serangan siber. Setiap baris data merepresentasikan satu koneksi jaringan, sedangkan kolom data menggambarkan karakteristik statistik dan perilaku trafik jaringan.

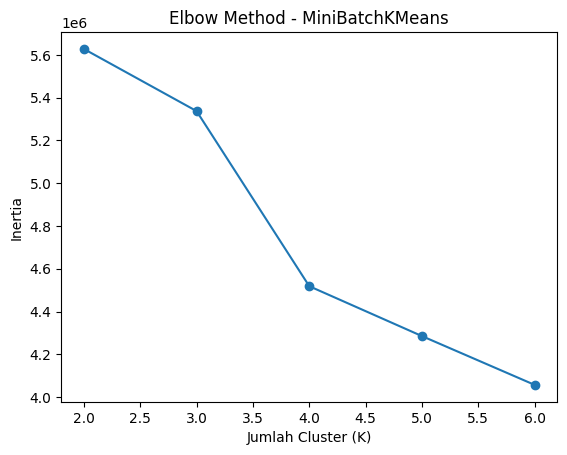
Sebelum dilakukan proses clustering, data terlebih dahulu melalui tahap pra-pemrosesan yang meliputi seleksi fitur numerik dan normalisasi data menggunakan StandardScaler. Normalisasi ini bertujuan untuk memastikan bahwa setiap fitur memiliki skala yang seragam sehingga tidak terjadi dominasi fitur tertentu dalam proses clustering berbasis jarak.

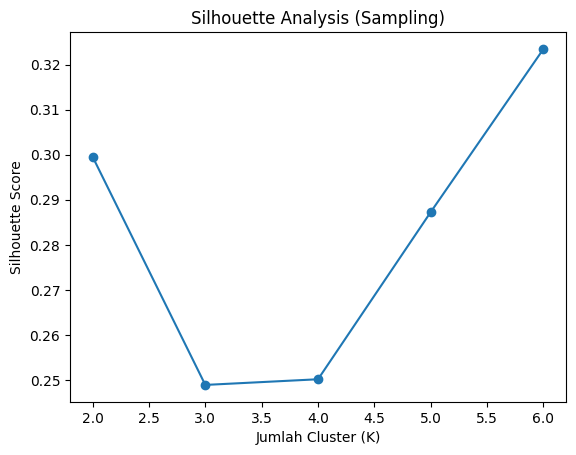
Setelah proses normalisasi, pengelompokan data dilakukan menggunakan algoritma K-Means dengan pendekatan MiniBatchKMeans untuk meningkatkan efisiensi komputasi mengingat ukuran dataset yang besar. Proses clustering bertujuan untuk mengelompokkan data trafik jaringan berdasarkan kemiripan pola karakteristiknya tanpa memanfaatkan label serangan.

**5.2 Penentuan Jumlah Cluster Optimal**

Penentuan jumlah cluster optimal dilakukan menggunakan dua metode evaluasi, yaitu **Elbow Method** dan **Silhouette Score.**

1. **Elbow Method**

Elbow Method diterapkan dengan menghitung nilai inertia pada rentang jumlah cluster K = 2 hingga K = 6. Hasil visualisasi menunjukkan bahwa penurunan nilai inertia signifikan hingga K = 2 dan mulai melandai setelahnya. Pola ini mengindikasikan bahwa K = 2 merupakan jumlah cluster yang optimal.

1. **Silhouette Score**  
   Untuk mengurangi waktu komputasi, evaluasi Silhouette Score dilakukan menggunakan data sampling sebanyak 2.000 record. Hasil perhitungan menunjukkan bahwa nilai Silhouette tertinggi diperoleh pada K = 2, yang menandakan kualitas pemisahan cluster terbaik dibandingkan nilai K lainnya.Berdasarkan hasil Elbow Method dan Silhouette Score, jumlah cluster optimal ditetapkan sebesar dua cluster (K = 2).
2. **Hasil Clustering**  
   Setelah jumlah cluster optimal ditentukan, proses clustering dilakukan menggunakan MiniBatchKMeans dengan K = 2 pada seluruh dataset yang telah dinormalisasi. Setiap data kemudian diberikan label cluster.

Distribusi jumlah data pada masing-masing cluster ditunjukkan sebagai berikut:

|  |  |
| --- | --- |
| Cluster | Jumlah Data |
| 0 | 79.178 |
| 1 | 96.163 |

Hasil ini menunjukkan bahwa data trafik jaringan terbagi relatif seimbang ke dalam dua cluster utama.

1. **Analisis Hubungan Cluster dan Label Asli**

Untuk memahami karakteristik cluster yang terbentuk, dilakukan analisis silang antara hasil clustering dan label serangan menggunakan tabel kontingensi (pd.crosstab). Analisis ini bertujuan untuk mengevaluasi distribusi trafik normal dan serangan pada masing-masing cluster.

Hasil analisis menunjukkan bahwa salah satu cluster memiliki proporsi trafik normal yang lebih dominan, sementara cluster lainnya memiliki proporsi trafik serangan yang lebih tinggi. Meskipun label tidak digunakan dalam proses clustering, hasil ini mengindikasikan bahwa pola yang terbentuk memiliki keterkaitan dengan karakteristik trafik jaringan.

1. **Interpretasi Pola Cluster**

Cluster yang terbentuk merepresentasikan pola statistik trafik jaringan berdasarkan kemiripan fitur, bukan secara langsung kelas serangan. Dengan demikian, clustering ini dapat dimanfaatkan sebagai:

* 1. Tahap eksplorasi awal data trafik jaringan
  2. Pendukung deteksi anomali berbasis unsupervised learning
  3. Preprocessing sebelum penerapan metode supervised learning

Pendekatan ini memungkinkan pemahaman struktur dasar data trafik jaringan tanpa ketergantungan pada label.

1. **Diskusi**
   1. **Implikasi pada Keamanan Jaringan**

Berdasarkan hasil clustering, beberapa implikasi dalam konteks keamanan jaringan dapat diidentifikasi:

1. Cluster dengan dominasi trafik normal dapat dijadikan representasi baseline pola trafik jaringan yang wajar.
2. Cluster dengan proporsi trafik serangan lebih tinggi dapat dimanfaatkan sebagai indikasi awal adanya aktivitas anomali atau mencurigakan.

Dengan demikian, hasil clustering dapat digunakan sebagai tahap awal dalam sistem Intrusion Detection System (IDS) berbasis unsupervised learning..

* 1. **Keterbatasan**

Penelitian ini memiliki beberapa keterbatasan, antara lain:

* Proses clustering hanya menggunakan fitur numerik tanpa mempertimbangkan fitur kategorikal.
* Evaluasi clustering masih terbatas pada metrik internal dan analisis distribusi label.
* Penelitian belum membahas performa clustering pada data real-time.
  1. **Reproducibility**

Seluruh proses analisis dilakukan menggunakan bahasa pemrograman Python dan terdokumentasi dalam bentuk notebook Jupyter. Dataset UNSW-NB15 bersifat terbuka dan dapat diakses secara publik, sehingga penelitian ini dapat direplikasi oleh peneliti lain.

1. **Kesimpulan**

Penelitian ini berhasil menerapkan algoritma K-Means dengan pendekatan Auto-K untuk mengelompokkan data trafik jaringan pada dataset UNSW-NB15 secara unsupervised. Hasil evaluasi menunjukkan bahwa jumlah cluster optimal adalah dua cluster (K = 2). Setiap cluster memiliki karakteristik trafik jaringan yang berbeda dan berkorelasi dengan distribusi trafik normal dan serangan, meskipun label tidak digunakan dalam proses clustering.

* 1. **Rekomendasi Penelitian Selanjutnya**

Penelitian selanjutnya disarankan untuk:

* Mengombinasikan hasil clustering dengan metode supervised learning sebagai tahap preprocessing.
* Menggunakan algoritma clustering lain seperti DBSCAN atau GMM untuk perbandingan.
* Mengembangkan sistem deteksi anomali jaringan berbasis clustering secara real-time.
  1. **Lampiran**

Link github berisi code beserta outputnya:

https://github.com/Alexaldi/dm