Implementasi Metode K-Means, Dbscan, dan Meanshift Untuk Analisis Jenis Ancaman Jaringan Pada Intrusion Detection System

Toga Aldila Cinderatama[1], Rinanza Zulmy Alhamri [2], Yoppy Yunhasnawa [3]
PSDKU Polinema di Kota Kediri, Jl Lingkar Maskumambang, Kec. Mojoroto, Kota Kediri
Jurusan Teknologi Informasi, Politeknik Negeri Malang, Jl. Soekarno Hatta 9, Kec Lowokwaru, Kota Malang[1,2,3]
toga.aldila@polinema.ac.id[1], rinanza.z.alhamri@polinema.ac.id[2], yunhasnawa@polinema.ac.id[3]

Abstract - The implementation of network security infrastructure has been carried out, including the Intrusion Detection System (IDS). However, in its implementation there are still many who have not combined with Data Technology (Data Science) to get a more comprehensive analysis. This study aims to analyze the types and characteristics of network threats using data science. Using the Unsupervised Learning method, an in-depth analysis of the types and characteristics of threats (threats) on the network infrastructure of a government health agency in the Health Sector will be analyzed. As a computational method, the results of 3 algorithms in the unsupervised learning category will be implemented and compared, namely K-Means, Meanshift, and Density-Based Spatial Clustering of Applications with Noise (DBSCAN). From the experimental results as measured by the Silhouette Index (SI) the best cluster of each implemented algorithm is DBSCAN which has the best SI value of 0.3424 with an Eps value of 0.2 and a MinPts value of 3. Meanwhile, from the results of clustering using K-Means, The best SI value was obtained by experiment k=4 with a value of 0.4531. The results of clustering using MeanShift, the best SI value was obtained by experiment bandwidth = 1 with a value of 0.5305.

Keywords - IDS, DBSCAN, K-Means, Network Security, MeanShift, Unsupervised Learning.

Intisari - Implementasi infrastruktur keamanan jaringan telah banyak dilakukan diantaranya adalah Intrusion Detection System (IDS). Namun dalam implementasinya masih banyak yang belum memadukan dengan Teknologi Data (Data Science) untuk mendapatkan analisis yang lebih komprehensif. penelitian ini bertujuan untuk menganalisis jenis-jenis dan karakteristik ancaman jaringan memanfaatkan data science. Dengan menggunakan metode Unsupervised Learning akan dianalisis secara mendalam tentang jenis dan karakteristik ancaman (threat) pada infrastuktur jaringan suatu instansi kesehatan pemerintah Bidang Kesehatan. Sebagai metode komputasi akan diimplementasikan dan dibanding hasil dari 3 algoritma dalam kategori unsupervised learning yaitu K-Means, Meanshift, dan Density-Based Spatial Clustering of Applications with Noise (DBSCAN). Dari hasil percobaan yang diukur dengan Silhouette Index (SI) didapatkan cluster terbaik dari masing-masing Algoritma yang diimplementasikan yaitu DBSCAN yang memiliki nilai SI terbaik yaitu 0.3424 dengan nilai Eps 0,2 dan nilai MinPts 3. Sementara dari hasil klasterisasi menggunakan K-Means, nilai SI terbaik diperoleh percobaan k=4 dengan nilai 0,4531. Hasil klasterisasi menggunakan MeanShift, nilai SI terbaik diperoleh percobaan bandwidth=1 dengan nilai 0,5305.

Kata Kunci - IDS, DBSCAN, K-Means, Keamanan Jaringan, Mean Shift, Unsupervised Learning.

I. PENDAHULUAN

Saat ini penggunaan teknologi informasi pada berbagai bidang telah meningkat pesat, sebagai contoh diantaranya instansi pemerintah pada bidang kesehatan telah memanfaatkan teknologi informasi untuk bertukar data dan informasi melalui jaringan komputer atau internet. Semakin meningkatnya penggunaan jaringan komputer maupun internet tersebut akan beriringan dengan resiko yang dimunculkan. Berdasarkan data dari situs *fireeye* dalam satu hari

terdapat enam ratus lima puluh delapan ribu serangan siber yang terdiri dari berbagai macam jenis serangan diantaranya *Denial of Service, Malware, Phising, Credential Reuse* dan lain sebagainya yang terjadi dari dan ke seluruh dunia [13]. Bahkan menurut situs Checkpoint [14] terdapat 11 juta lebih serangan cyber dalam satu hari. Dari data-data tersebut, dapat dikatakan bahwa pada setiap detik ada serangan siber (*cyber attack*) yang terjadi di seluruh penjuru dunia. Dengan semakin banyaknya digunakan aplikasi dan komunikasi data yang memanfaatkan jaringan komputer dan internet resiko akan terjadinya serangan cyber menjadi meningkat sehingga diperlukan infrastruktur yang memadai sebagai sarana pengamanan dan pencegahan internal sistem terhadap serangan-serangan siber yang berdampak terhadap hilangnya data, kerusakan data bahkan sampai kelumpuhan sistem.

Dengan banyaknya kejadian penyerangan jaringan dan berbagai kemungkinan cara menyerang, dibutuhkan tindakan yang harus dilakukan untuk mencegah serangan terjadi. Salah satu infrastuktur yang terdapat pada suatu sistem komputer atau pada jaringan komputer adalah *Intrusion Detection System (IDS)*. IDS adalah salah satu cara bagaimana mendeteksi sebuah serangan yang terjadi pada sebuah komputer atau server pada jaringan komputer. IDS akan berkerja dengan prinsip mendeteksi serangan-serangan atau percobaan intrusi dari luar sistem pada umumnya internet ke dalam suatu internal sistem. IDS akan memonitor lalu lintas jaringan, namun pada IDS dibutuhkan tindakan lebih lanjut untuk memberitahu serangan dengan karakteristik tersendiri [10].

Sedangkan pada keilmuan lain di Bidang Teknologi Informasi yaitu teknologi data (*data science*) yang merupakan perpanjangan evolusioner dari statistik yang mampu menangani sejumlah besar data yang diproduksi hari ini. Dimana pada teknologi data ini menambahkan metode dari ilmu komputer ke repertoar statistik. Hal utama yang membedakan data science dengan statistik biasa adalah kemampuan untuk bekerja dengan big data dan pengalaman di dalamnya pembelajaran mesin (machine learning), komputasi, dan pembangunan algoritma [6]. Dengan adanya teknologi data tersebut dapat digunakan untuk membantu penerapan infrastuktur jaringan seperti IDS sebagai pertahanan terhadap resiko serangan cyber. Sehingga dapat diketahui pengetahuan-pengetahuan yang baru mengenai jenis dan karakteristik ancaman jaringan apa saja yang dialami khususnya pada studi kasus penelitian yaitu salah satu instansi kesehatan pemerintah sebagai dasar untuk pengambilan keputusan bagaimana menerapkan dan mengembangkan infrastruktur jaringan yang sebaik mungkin sebagai tindakan pencegahan maupun respon akan resiko-resiko dari serangan cyber pada infrastuktur instansi tersebut.

II. SIGNIFIKANSI STUDI

Telah dilakukan berbagai macam penelitian yang memiliki relevansi terhadap infrastruktur jaringan computer, Barbara et al telah menelaah penelitian-penelitian sebelumnya tentang penggunaan data mining pada intrusion detection system [1]. Berdasarkan penelaahan tersebut, mereka mengapati empat empat hal penting yang menarik. Yang pertama adalah bahwa kebanyakan penelitian sebelumnya lebih berfokus pada bagaimana mewujudkan sistem IDS berlandaskan data mining yang operasional. Kedua, terjadinya pengabaian terhadap keseluruhan proses Knowledge Discovery & Data Mining (KDD). Ketiga, selalu tergantung pada data training dengan kualitas yang tinggi. Keempat, kebanyakan penelitian sebelumnya berfokus hanya ke sedikit hal yang penting. Berdasarkan fenomena tersebut, peneliti menyimpulkan rekomendasi fokus penelitian untuk dilakukan di masa depan yaitu, Penelitian di masa depan sebaiknya lebih memperhatikan proses KDD, Penelitian di masa depan sebaiknya mencoba untuk merumuskan teknik pembuatan data training dengan kualitas tinggi secara (semi-)otomatis, atau menemukan cara agar tidak terlalu bergantung kepada data training kualitas tinggi tersebut. Penelitian di masa depan disarankan untuk mengeksplorasi penerapan data mining yang bersifat terobosan, yang tidak hanya mengandalkan feature selection atau

anomaly detection saja. Untuk mengatasi tantangan-tantangan umum yang ada pada data mining, akan lebih baik untuk mengembangkan solusi spesial yang secara khusus dirancang untuk intrusion detection.

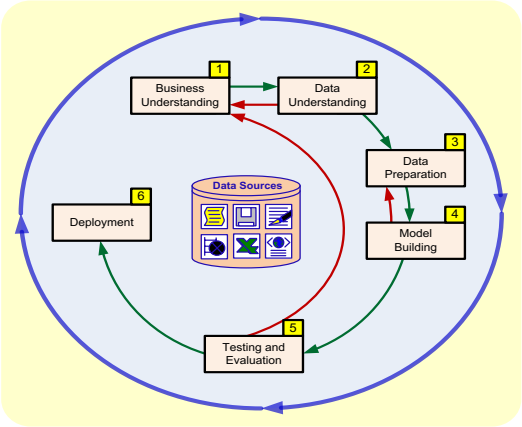
Chandrashekhar Azad et.al, juga melakukan penelaahan terhadap 75 artikel publikasi ilmiah tentang intrusion detection yang dipublikasikan sejak tahun 2000 hingga 2012 [2]. Menurut mereka, Intrusion Detection System terbagi menjadi 2 kategori utama berdasarkan pendekatan deteksinya yaitu deteksi anomali vs deteksi penyalahgunaan (misuse detection). Dari semua paper tersebut, mereka menyimpulkan, sebanyak sekitar 67% dari publikasi tersebut berfokus pada deteksi anomali, 10% berfokus pada misuse detection, dan 23% berfokus pada keduanya. Selain itu, untuk metode-metode yang digunakan, pada deteksi anomali kebanyakan yang digunakan antara lain: Neural Network, statistik, predictive pattern generation, serta sequence matching & learning. Sedangkan untuk pendekatan deteksi penyalahgunaan, kebanyakan paper berfokus pada penggunaan metode diantaranya sistem pakar, pattern matching, dan analisis transasisi state (state transition analysis). Sedangkan dataset yang digunakan meliputi 42% dari KDD cup dataset, 20% dari dataset DARPA, dan 38% menggunakan dataset lain.

G. Yedukondalu, telah merancang sebuah metode untuk mengidentifikasi intruder secara efektif dan pada saat bersamaan mengurangi waktu pencariannya. Pada penelitian tersebut mereka menggunakan algoritma K-Means dan Jaccard Distance Similarity. Dataset yang digunakan adalah dataset DARPA tahun 1998 yang digunakan untuk mengklasifikasikan pengguna apakah mereka intruder atau pengguna normal. Setiap dokumen dari dataset tersebut dikonversi menjadi signature biner 32-bit dengan menggunakan Hashing dan teknik Superimposed Coding. Databes biner tersebut kemudian diklasterkan dan dievaluasi sebagai intrusi-intrusi. IDS yang mereka buat berhasil mendeteksi intruder yang tidak diketahui dengan akurasi sebesar 76,4% sedangkan untuk pengguna yang sudah diketahui sebelumnya, IDS mereka berhasil mendeteksinya hingga tingkat keakuratan sebesar 99.9%. Teknik clustering yang digunakan mampu meningkatkan performa dari IDS yang mereka buat [8].

Nasrin et al. telah melakukan review tentang programmable networks dan mempelajari penggunaan teknologi yang sedang berkembang yaitu SDN (Software-Defined Networking) pada kasus NIDS. Mereka juga mengidentifikasi berbagai macam mekanisme intrusion detection dengan pendekatan ML/DL. Pada penelitian ini mereka menekankan pada SDN sebagai platform yang diatasnya diterapkan berbagaimacam pendekatan ML/DL untuk mendeteksi kelemahan dan memonitor jaringan. Penggunaan deep learning menjadi penting karena kefisienannya dalam mengevaluasi keamanan jaringan. Begitupula, metode-metode baru deep learning semakin meningkat kecepatannya dan semakin efisien dalam data taxation. Menurut mereka, berbagai macam permsalahan juga perlu dipertimbangkan dalam mengimplementasikan NIDS, disebabkan sifat dasar serangan yang selalu dinamis. Oleh karena itu, kemampuan adaptasi dari suatu pendekatan deteksi juga menjadi diperlukan. Lebih jauh enurut mereka, pengembangan metoda pemilihan fitur dengan classifier yang mengurangi dimensi dari dataset juga saat ini masih menjadi tantangan tersendiri. untuk merancang sebuah SDN controller yang terpusat, yang mampu memonitor dan mengimplementasikan IDS secara real-time pada jaringan kecepatan tinggi merupakan sebuah kemungkinan arah penelitian yang menantang kedepannya. Pada kesimpulannya, mereka berpendapat bahwasannya keakurasian dan skalabilitas SDN akan memungkinkan peneliti di masa depan untuk merumuskan suatu NIDS berbasis ML/DL pada infrastuktur penting [12].

III. HASIL DAN PEMBAHASAN

Metode penelitian yang digunakan adalah mengadopsi metode standar pada data mining yaitu (Cross-Industry Standard Process for Data Mining) CRISP-DM seperti yang dijelaskan pada gambar 3.1



Gambar 1. Metodologi CRISP-DM

A. Business Understanding

Merupakan tahap awal yaitu pemahaman penelitian, penentuan tujuan dan rumusan masalah data mining. Pada penelitian ini yang menjadi obyek penelitian adalah mengenai data pada infrastuktur Intrusion Detection System (IDS). Permasalahan yang terjadi yaitu belum dikembangkannya suatu sistem yang berbasis data science untuk menganalisis secara mendalam terhadap jenis serangan atau ancaman pada infrastruktur keamanan jaringan komputer.

B. Data Understanding

Pada data understanding ini meliputi proses-proses:

a. Mengumpulkan data.
b. Menggunakan analisis penyelidikan data untuk mengenali lebih lanjut data dan pencarian pengetahuan awal.
c. Mengevaluasi kualitas data.
d. Jika diinginkan, pilih sebagian kecil kelompok data yang mungkin mengandung pola dari permasalahan.

C. Data Preparation

Tahap ini adalah pekerjaan berat yang perlu dilaksanakan secara intensif. Memilih kasus atau variable yang ingin dianalisis, melakukan perubahan pada beberapa variable jika diperlukan sehingga data siap untuk dimodelkan.

Tahapan pada data preparation ini meliputi :

1. Data cleaning (Pembersihan data): Proses yang dilakukan pada tahap data cleaning ini adalah seperti, mengisi nilai yang hilang, menghilangkan data yang bersifat noise, identifikasi atau hapus outliers, dan mengatasi ketidakkonsistenan data.
2. Data reduction (Pengurangan data): dalam tahap ini dilakukan beberapa proses seperti: pengurangan dimensi, pengurangan numerik, dan kompresi data.
3. Transformasi data dan diskritisasi data: dalam tahap ini dilakukan normalisasi data atau generalisasi hierarki konsep.

4. Data integration: dalam tahap ini dilakukan proses integrasi beberapa basis data atau file jika sumber data lebih dari satu.

D. Modeling

Dalam tahap ini akan dilakukan data modelling dengan cara mengimplementasikan 3 algoritma dalam kategori unsupervised learning [7] yaitu K-Means, Mean Shift dan DBSCAN. Pemilihan algoritma ini dengan maksud sesuai studi kasus IDS yang dipilih.

1. Algoritma K-Means

Algoritma K-Means merupakan algoritma pengelompokan iteratif yang melakukan partisi set data ke dalam sejumlah K cluster yang sudah ditetapkan di awal. Algoritam K-Means sederhana [9] untuk diimplementasikan dan dijalankan, relatif cepat, mudah beradaptasi, umum penggunaannya dalam praktek. K-Means dapat diterapkan pada data yang direpresentasikan dalam r-dimensi ruang tempat. K- means mengelompokan set data r-dimensi, $X= \{x_i \mid i=1,...,N\}$. Algoritma K-Means mengelompokan semua titik data dalam X sehingga setiap titik xi hanya jatuh dalam satu K partisi. Tujuan pengelompokan ini adalah untuk meminimalkan fungsi objek yang diset dalam proses pengelompokan, yang pada umumnya berusaha meminimalkan variasi di dalam suatu kelompok dan memaksimalkan variasi antarkelompok. Parameter yang hrsus dimasukkan ketika menggunakan algoritma K-Means adalah nilai K. Nilai K yang digunkan pada umumnya didasarkan pada informasi yang diketahui sebelumnya mengenai sebenarnya berapa banyak cluster yang muncul dalam X, berapa banyak yang digunakan untuk penerapnnya, atau jenis cluster dicari dengan melakukan percobaan dengan beberapa nilai K. Set representatif cluster dinyatakan $C =\{c_j \mid j=1,\ldots,K\}$. sejumlah K representatif cluster tersebut sebagai cluster centroid (titik pusat cluster). Untuk set data dalam X dikelompokan berdasarkan konsep kedekatan atau kemiripan, namun kuantitas yang digunkan untuk mengukurnya adalah ketidakmiripan. Metrik yang umum digunakan untuk ketidakmiripan tersebut adalah Euclidean.

Secara umum algoritma K-Means memiliki langkah-langkah dalam pengelompokan, diantaranya:

1) Inisilisasi: menentukan nilai K centroid yang diinginkan dan metrik ketidakmiripan (jarak) yang diinginkan.
2) Memilih K data dari set X sebagai centroid. Untuk menentukan centroid dapat menggunakan persamaan.

$$\frac{Jumlah\ Data}{Jumlah\ Class + 1} \tag{1}$$

3) Mengalokasikan semua data ke centroid terdekat dengan metrik jarak yang telah ditetapkan.
4) Menghitung kembali centroid C berdasarkan data yang mengikuti cluster masing – masing.
5) Mengulangi langkah 3 dan 4 hingga kondisi konvergen tercapai.

Berikut ini adalah rumus untuk menentukan jumlah cluster:

$$\sqrt[\kappa]{\frac{N}{2}} \tag{2}$$

Keterangan:
K = klaster
N = jumlah data

Menghitung jarak pada ruang jarak Euclidean menggunakan formula:

$$D(x_2, x_1) = \|x_2 - x_1\|_2 = \sqrt{\sum_{j=1}^{p} |x_{2j} - x_{1j}|^2} \quad (3)$$

Keterangan:
D = euclidean distance
x = banyaknya objek
Σ^p = jumlah data record

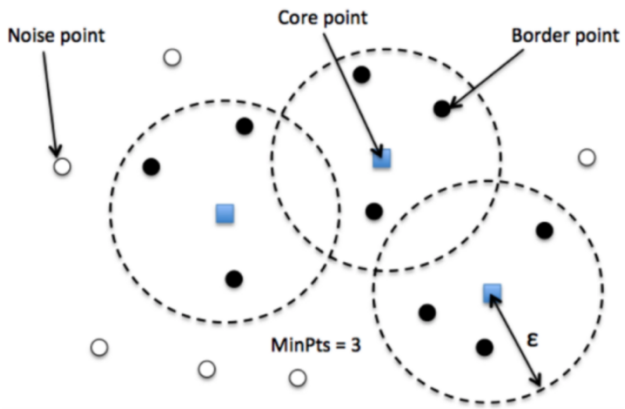
2. Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) adalah algoritma dasar untuk clustering berbasis kepadatan. Ia dapat menemukan cluster dengan berbagai bentuk dan ukuran dari sejumlah besar data, yang mengandung noise dan outlier [3]. Algoritma DBSCAN menggunakan dua parameter:

1) minPts: Jumlah minimum titik (ambang batas) yang dikelompokkan bersama untuk suatu wilayah yang dianggap padat.
2) eps (ε): Ukuran jarak yang akan digunakan untuk menemukan titik-titik di sekitar titik mana pun.

Parameter ini dapat dipahami jika kita mengeksplorasi dua konsep yang disebut Density Reachability dan Density Connectivity. Keterjangkauan dalam hal kepadatan menetapkan titik yang dapat dijangkau dari yang lain jika terletak dalam jarak tertentu (eps) darinya. Konektivitas, di sisi lain, melibatkan pendekatan rantai berbasis transitivitas untuk menentukan apakah titik-titik terletak di cluster tertentu. Sebagai contoh, titik p dan q dapat dihubungkan jika p-> r-> s-> t-> q, di mana a-> b berarti b berada di lingkungan a.

Ada tiga jenis titik setelah pengelompokan DBSCAN selesai:



Gambar 1. Tiga jenis titik DBSCAN

Core - Ini adalah titik yang memiliki setidaknya m titik dalam jarak n dari dirinya.
Border - Ini adalah titik yang memiliki setidaknya satu titik Inti pada jarak n.

Noise - Ini adalah titik yang bukan Inti maupun Perbatasan. Dan itu memiliki kurang dari m titik dalam jarak n dari dirinya sendiri.

Langkah-langkah algoritmik untuk pengelompokan DBSCAN (Chauhan, 2020)

1) Algoritma melanjutkan dengan mengambil titik dalam kumpulan data secara sewenang-wenang (hingga semua titik telah dikunjungi).
2) Jika ada setidaknya titik 'titik kecil' dalam radius 'ε' ke titik tersebut, maka dianggap semua titik ini sebagai bagian dari kelompok yang sama.
3) Cluster tersebut kemudian diperluas dengan mengulangi penghitungan lingkungan secara rekursif untuk setiap titik tetangga.

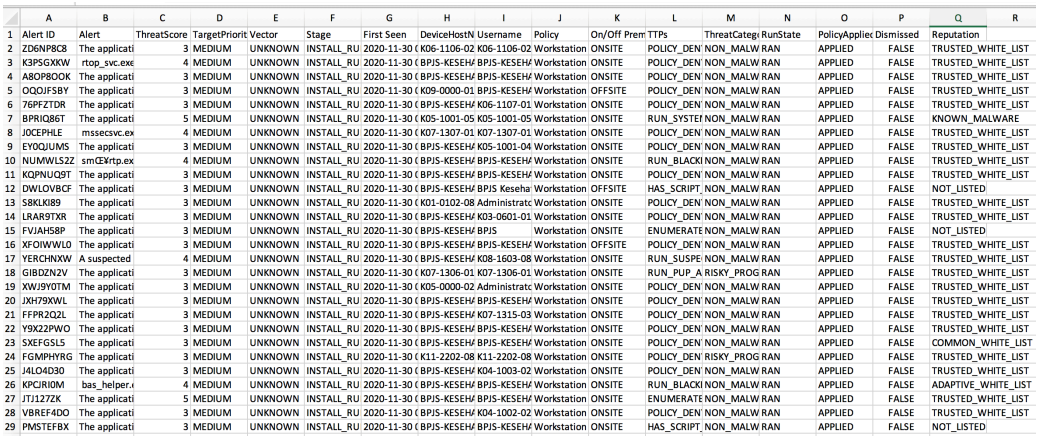
3. Algoritma Mean Shift

Mean shift clustering adalah algoritma berbasis jendela geser yang mencoba menemukan area titik data yang padat. Ini adalah algoritma berbasis centroid yang berarti bahwa tujuannya adalah untuk menemukan titik pusat dari setiap kelompok/kelas, yang bekerja dengan memperbarui kandidat untuk titik pusat menjadi rata-rata titik di dalam jendela geser. Kandidat jendela ini kemudian disaring dalam tahap pasca-pemrosesan untuk menghilangkan duplikat yang hampir sama, membentuk set akhir titik pusat dan grup yang sesuai.

1) Untuk menjelaskan pergeseran rata-rata, dipertimbangkan sekumpulan titik dalam ruang dua dimensi. Dimulai dengan jendela geser melingkar yang berpusat pada titik C (dipilih secara acak) dan memiliki radius r sebagai kernel. Pergeseran rata-rata adalah algoritma hill-climbing yang melibatkan pergeseran kernel ini secara iteratif ke wilayah kepadatan yang lebih tinggi pada setiap langkah sampai konvergen.
2) Pada setiap iterasi, jendela geser digeser ke arah daerah dengan kepadatan lebih tinggi dengan menggeser titik pusat ke rata-rata titik di dalam jendela (karena itu namanya). Kepadatan di dalam jendela geser sebanding dengan jumlah titik di dalamnya. Secara alami, dengan menggeser ke rata-rata titik-titik di jendela itu secara bertahap akan bergerak menuju area dengan kerapatan titik yang lebih tinggi.
3) Kemudian jendela geser terus digeser sesuai rata-rata sampai tidak ada arah di mana pergeseran dapat menampung lebih banyak titik di dalam kernel.
4) Proses langkah 1 sampai 3 ini dilakukan dengan banyak jendela geser sampai semua titik berada di dalam jendela. Ketika beberapa jendela geser tumpang tindih, jendela yang berisi poin terbanyak dipertahankan. Titik-titik data kemudian dikelompokkan sesuai dengan jendela geser di mana mereka berada.

Pada penelitian ini data yang diambil adalah data dari infrastruktur Intrusion Detection System pada salah satu instansi kesehatan pemerintah. Data diambil dari aplikasi Endpoint Detection and Response (EDR), dimana EDR [4] digunakan untuk melindungi titik akhir dari ancaman, gambar 3 merupakan contoh data yang akan dianalisa yang diambil dari EDR, data yang dianalisis pada penelitian ini adalah dipilih sebanyak 10rb data.

Terdapat banyak parameter dari data sumber dan yang akan digunakan untuk modeling adalah 3 parameter utama yaitu Thread Category, Thread Score dan Reputation. Berikut pada tabel 1, data yang digunakan yaitu data dari EDR yang terdiri dari 16 atribut.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Alert ID	Alert	ThreatScore	TargetPriorit	Vector	Stage	First Seen	DeviceHostN	Username	Policy	On/Off Prem	TTPs	ThreatCateg	RunState	PolicyApplie	Dismissed	Reputation	
2	ZD6NP8C8	The applicati	3	MEDIUM	UNKNOWN	INSTALL_RU	2020-11-30 (	K06-1106-02	K06-1106-02	Workstation	ONSITE	POLICY_DEN	NON_MALW	RAN	APPLIED	FALSE	TRUSTED_WHITE_LIST	
3	K3PSGXKW	rtop_svc.exe	4	MEDIUM	UNKNOWN	INSTALL_RU	2020-11-30 (	BPJS-KESEHA	BPJS-KESEHA	Workstation	ONSITE	POLICY_DEN	NON_MALW	RAN	APPLIED	FALSE	TRUSTED_WHITE_LIST	
4	A8OPBOOK	The applicati	3	MEDIUM	UNKNOWN	INSTALL_RU	2020-11-30 (	BPJS-KESEHA	BPJS-KESEHA	Workstation	ONSITE	POLICY_DEN	NON_MALW	RAN	APPLIED	FALSE	TRUSTED_WHITE_LIST	
5	OQOJFSBY	The applicati	3	MEDIUM	UNKNOWN	INSTALL_RU	2020-11-30 (	K09-0000-01	BPJS-KESEHA	Workstation	OFFSITE	POLICY_DEN	NON_MALW	RAN	APPLIED	FALSE	TRUSTED_WHITE_LIST	
6	76PFZTDR	The applicati	3	MEDIUM	UNKNOWN	INSTALL_RU	2020-11-30 (	BPJS-KESEHA	K06-1107-01	Workstation	ONSITE	POLICY_DEN	NON_MALW	RAN	APPLIED	FALSE	TRUSTED_WHITE_LIST	
7	BPRIQ86T	The applicati	5	MEDIUM	UNKNOWN	INSTALL_RU	2020-11-30 (	K05-1001-05	K05-1001-05	Workstation	ONSITE	RUN_SYSTE	NON_MALW	RAN	APPLIED	FALSE	KNOWN_MALWARE	
8	J0CEPHLE	mssecsvc.ex	4	MEDIUM	UNKNOWN	INSTALL_RU	2020-11-30 (	K07-1307-01	K07-1307-01	Workstation	ONSITE	POLICY_DEN	NON_MALW	RAN	APPLIED	FALSE	TRUSTED_WHITE_LIST	
9	EY0QJUMS	The applicati	3	MEDIUM	UNKNOWN	INSTALL_RU	2020-11-30 (	BPJS-KESEHA	K05-1001-04	Workstation	ONSITE	POLICY_DEN	NON_MALW	RAN	APPLIED	FALSE	TRUSTED_WHITE_LIST	
10	NUMWLS2Z	smCẋrtp.ex	4	MEDIUM	UNKNOWN	INSTALL_RU	2020-11-30 (	BPJS-KESEHA	BPJS-KESEHA	Workstation	ONSITE	RUN_BLACKI	NON_MALW	RAN	APPLIED	FALSE	TRUSTED_WHITE_LIST	
11	KQPNUQ9T	The applicati	3	MEDIUM	UNKNOWN	INSTALL_RU	2020-11-30 (	BPJS-KESEHA	BPJS-KESEHA	Workstation	ONSITE	POLICY_DEN	NON_MALW	RAN	APPLIED	FALSE	TRUSTED_WHITE_LIST	
12	DWLOVBCF	The applicati	3	MEDIUM	UNKNOWN	INSTALL_RU	2020-11-30 (	BPJS-KESEHA	BPJS Keseha	Workstation	OFFSITE	HAS_SCRIPT	NON_MALW	RAN	APPLIED	FALSE	NOT_LISTED	
13	S8KLKI89	The applicati	3	MEDIUM	UNKNOWN	INSTALL_RU	2020-11-30 (	K01-0102-08	Administrato	Workstation	ONSITE	POLICY_DEN	NON_MALW	RAN	APPLIED	FALSE	TRUSTED_WHITE_LIST	
14	LRAR9TXR	The applicati	3	MEDIUM	UNKNOWN	INSTALL_RU	2020-11-30 (	BPJS-KESEHA	K03-0601-01	Workstation	ONSITE	POLICY_DEN	NON_MALW	RAN	APPLIED	FALSE	TRUSTED_WHITE_LIST	
15	FVJAH58P	The applicati	3	MEDIUM	UNKNOWN	INSTALL_RU	2020-11-30 (	BPJS-KESEHA	BPJS	Workstation	ONSITE	ENUMERATE	NON_MALW	RAN	APPLIED	FALSE	NOT_LISTED	
16	XFOIWWL0	The applicati	3	MEDIUM	UNKNOWN	INSTALL_RU	2020-11-30 (	BPJS-KESEHA	BPJS-KESEHA	Workstation	OFFSITE	POLICY_DEN	NON_MALW	RAN	APPLIED	FALSE	TRUSTED_WHITE_LIST	
17	YERCHNXW	A suspected	4	MEDIUM	UNKNOWN	INSTALL_RU	2020-11-30 (	BPJS-KESEHA	K08-1603-08	Workstation	ONSITE	RUN_SUSPE	NON_MALW	RAN	APPLIED	FALSE	TRUSTED_WHITE_LIST	
18	GIBDZN2V	The applicati	3	MEDIUM	UNKNOWN	INSTALL_RU	2020-11-30 (	K07-1306-01	K07-1306-01	Workstation	ONSITE	RUN_PUP_A	RISKY_PROG	RAN	APPLIED	FALSE	TRUSTED_WHITE_LIST	
19	XWJ9Y0TM	The applicati	3	MEDIUM	UNKNOWN	INSTALL_RU	2020-11-30 (	K05-0000-02	Administrato	Workstation	ONSITE	POLICY_DEN	NON_MALW	RAN	APPLIED	FALSE	TRUSTED_WHITE_LIST	
20	JXH79XWL	The applicati	3	MEDIUM	UNKNOWN	INSTALL_RU	2020-11-30 (	BPJS-KESEHA	BPJS-KESEHA	Workstation	ONSITE	POLICY_DEN	NON_MALW	RAN	APPLIED	FALSE	TRUSTED_WHITE_LIST	
21	FFPR2Q2L	The applicati	3	MEDIUM	UNKNOWN	INSTALL_RU	2020-11-30 (	BPJS-KESEHA	K07-1315-03	Workstation	ONSITE	POLICY_DEN	NON_MALW	RAN	APPLIED	FALSE	TRUSTED_WHITE_LIST	
22	Y9X22PWO	The applicati	3	MEDIUM	UNKNOWN	INSTALL_RU	2020-11-30 (	BPJS-KESEHA	BPJS-KESEHA	Workstation	ONSITE	POLICY_DEN	NON_MALW	RAN	APPLIED	FALSE	TRUSTED_WHITE_LIST	
23	SXEFGSL5	The applicati	3	MEDIUM	UNKNOWN	INSTALL_RU	2020-11-30 (	BPJS-KESEHA	BPJS-KESEHA	Workstation	ONSITE	POLICY_DEN	NON_MALW	RAN	APPLIED	FALSE	COMMON_WHITE_LIST	
24	FGMPHYRG	The applicati	3	MEDIUM	UNKNOWN	INSTALL_RU	2020-11-30 (	K11-2202-08	K11-2202-08	Workstation	ONSITE	POLICY_DEN	RISKY_PROG	RAN	APPLIED	FALSE	TRUSTED_WHITE_LIST	
25	J4LO4D30	The applicati	3	MEDIUM	UNKNOWN	INSTALL_RU	2020-11-30 (	BPJS-KESEHA	K04-1003-02	Workstation	ONSITE	POLICY_DEN	NON_MALW	RAN	APPLIED	FALSE	TRUSTED_WHITE_LIST	
26	KPCJRI0M	bas_helper.(	4	MEDIUM	UNKNOWN	INSTALL_RU	2020-11-30 (	BPJS-KESEHA	BPJS-KESEHA	Workstation	ONSITE	RUN_BLACKI	NON_MALW	RAN	APPLIED	FALSE	ADAPTIVE_WHITE_LIST	
27	JTJ127ZK	The applicati	5	MEDIUM	UNKNOWN	INSTALL_RU	2020-11-30 (	BPJS-KESEHA	BPJS-KESEHA	Workstation	ONSITE	ENUMERATE	NON_MALW	RAN	APPLIED	FALSE	TRUSTED_WHITE_LIST	
28	VBREF4DO	The applicati	3	MEDIUM	UNKNOWN	INSTALL_RU	2020-11-30 (	BPJS-KESEHA	K04-1002-02	Workstation	ONSITE	POLICY_DEN	NON_MALW	RAN	APPLIED	FALSE	TRUSTED_WHITE_LIST	
29	PMSTEFBX	The applicati	3	MEDIUM	UNKNOWN	INSTALL_RU	2020-11-30 (	BPJS-KESEHA	BPJS-KESEHA	Workstation	ONSITE	HAS_SCRIPT	NON_MALW	RAN	APPLIED	FALSE	NOT_LISTED	

Gambar 3. Data EDR

TABEL I
DATA UNDERSTANDING

NAMA ATRIBUT	DESKRIPSI	TIPE DATA
“NON_MALWARE”, “RISKY_PROGRAM”,"NEW_MALWARE”, ”MALWARES	Thread category: merupakan kategori dari ancaman yang masuk ke IDS	Data Numerik
“3”,”4”,”5”,”6”,”7”,”8”	Thread score: merupakan skor dari ancaman yang masuk ke IDS	Data Numerik
“NOT_LISTED”,” KNOWN_MALWARE”,”SUSPECT_MALWARE”, ”PUP”,” COMMON_WHITE_LIST”,” ADAPTIVE_WHITE_LIST”,” TRUSTED_WHITE_LIST”	Reputation: merupakan reputasi dari ancaman yang masuk ke IDS	Data Numerik

Tahapan selanjutnya adalah melakukan normalisasi data mentah sebelum diterapkan ke dalam masing-masing algoritma yang telah ditentukan untuk kemudian diclustering, dengan mengikuti parameter-parameter yang telah dimasukan di tiap-tiap implementasi algoritma. Berdasarkan range yang telah ditentukan pada pengelompokan data, maka dapat Disimpulkan bahwa pengelompokan data EDR ke dalam data yang akan dikelompokan. Pada data tersebut, tidak semua kriteria yang dapat menjadi patokan. Dalam hal ini, kriteria yang dinormalisasi adalah Thread Category diubah menjadi X1, Thread Score diubah menjadi X2, dan Reputation diubah menjadi X3.

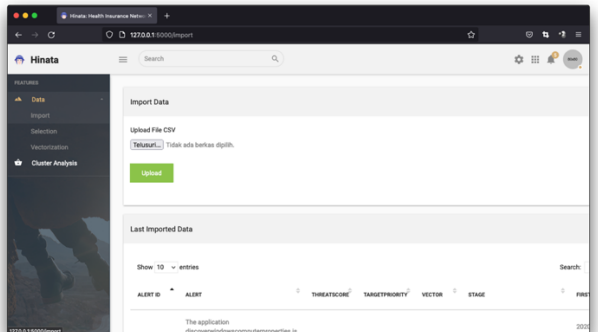
TABEL II
HASIL NORMALISAI DATA

Alternatif	**Kriteria**		
	X1	X2	X3
A1	-0.410	-0.310	0.720
A2	1.693	-0.310	0.720
A3	-0.410	-0.310	0.720
A4	-0.410	-0.310	0.720
A5	-0.410	-0.310	0.720

A6	3.796	-0.310	-1.288
A7	1.693	-0.310	0.720
A8	-0.410	-0.310	0.720
A9	1.693	-0.310	0.720
A10	-0.410	-0.310	0.720
A11	-0.410	-0.310	-1.511
A12	-0.410	-0.310	0.720
A13	-0.410	-0.310	0.720
A14	-0.410	-0.310	-1.511
A15	-0.410	-0.310	0.720
A16	-0.410	1.653	0.720
…	…		…
A998	-0.410	-0.310	0.274
A999	1.693	-0.310	0.720
A1000	-0.410	-0.310	0.720

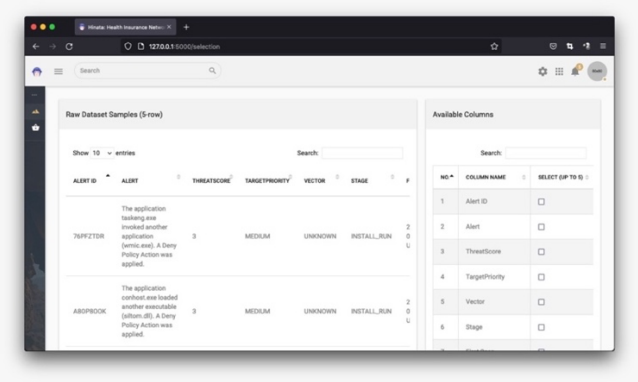
4.2 Implementasi dan Pembahasan

Pada bagian ini akan dibahas tentang Implementasi aplikasi. Untuk implementasi dari beberapa algoritma unsupervised learning yang dipilih, telah dikembangkan aplikasi berbasis web yaitu HINATA - Health Insurance Network Advanced Traffic Analyzer, yaitu merupakan aplikasi yang didalamnya menerapkan data science yaitu implementasi metode unsupervised learning terutama clustering untuk menganalisa jenis-jenis dan karakteristik serangan jaringan pada Intrusion Detection Sistem. Terdapat 3 algoritma yang diimplementasikan yaitu K-Means, DBSCAN, dan Mean Shift. Berikut implementasi aplikasi berbasis website HINATA.



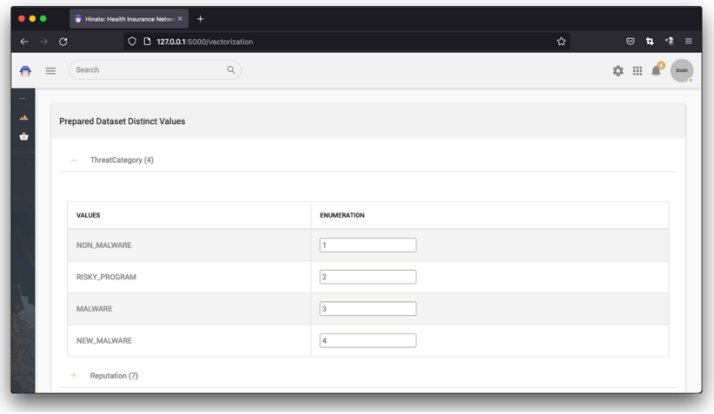
Gambar 4. Import Data

Untuk dapat menggunakan aplikasi ini, langkah awal adalah mengupload data source yang akan dianalisa (file sumber seperti pada contoh berformat .csv) seperti pada gambar 4 Kemudian oleh aplikasi akan dibaca data tersebut dan ditampilkan semua atribut hasil data sumber.



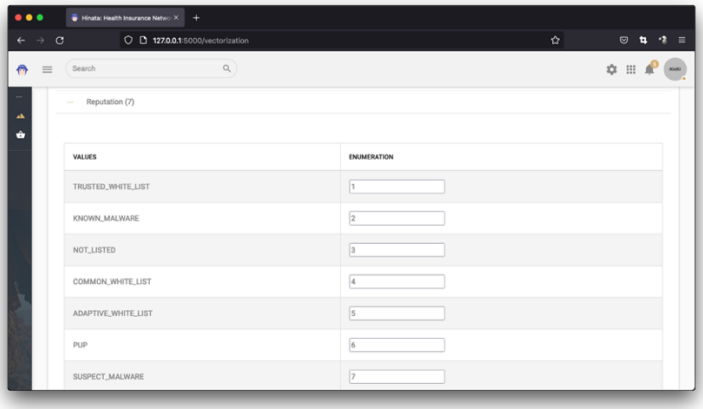
Gambar 5. Data Raw

Langkah selanjutnya adalah memilih atribut-atribut terpilih yang akan dianalisa ditunjukan gambar 5 kemudian melakukan enumerasi atau data processing misal direpresentaikan dari data non numerik menjadi data numerik seperti pada gambar 6



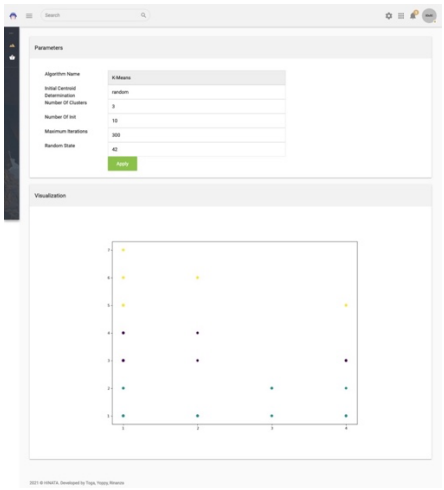
Gambar 6. Contoh Enumerasi Thread Category

Pada proses yang ditunjukan gambar 6 adalah contoh implementasi persiapan data pada parameter thread category yang dienumerasi menjadi integer 1 sampai dengan 4.



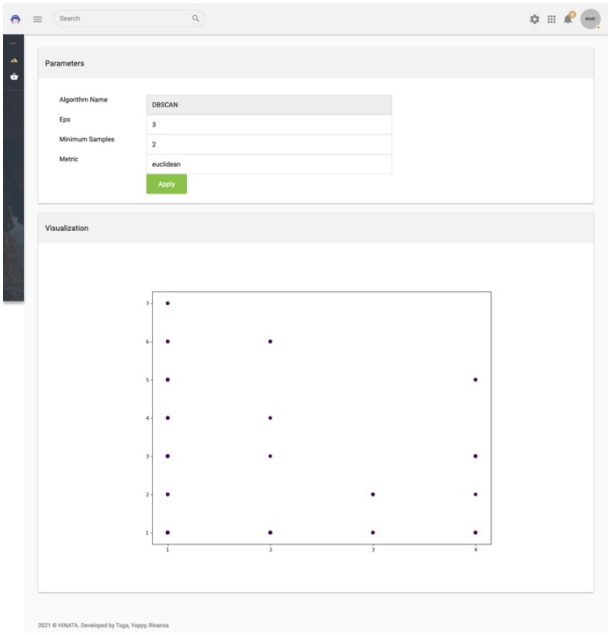
Gambar 7. Contoh Enumerasi Reputation

Pada proses yang ditunjukan gambar 7 adalah contoh implementasi persiapan data pada parameter reputation yang dienumerasi menjadi integer 1 sampai dengan 7. Langkah selanjutnya adalah memilih metode clustering yang akan digunakan dan input parameter sesuai algoritma terpilih, sebagai contoh pada pengujian pertama menggunakan algoritma K-means, user perlu input data Number of cluster, Number of Init, Maximum Iterations dan Random State. Selanjutnya sekaligus dimodelkan clustering data seperti ditunjukan gambar 8.



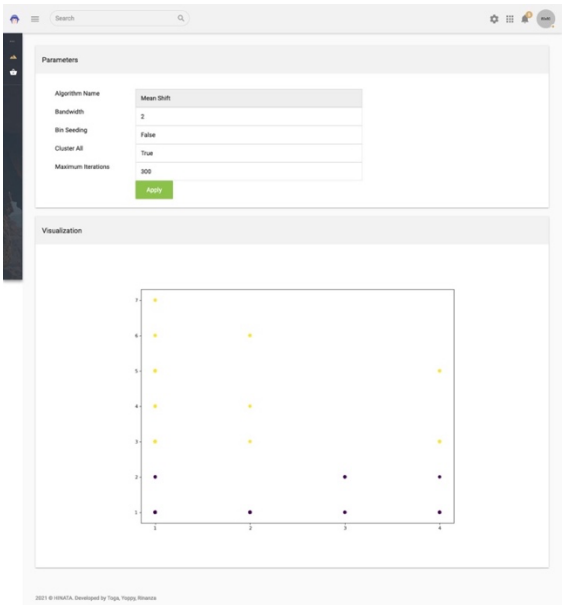
Gambar 8. Implementasi algoritma K-Means

Pilihan selanjutnya adalah memilih metode clustering yang akan digunakan dan input parameter sesuai algoritma terpilih, sebagai contoh pada pengujian kedua menggunakan algoritma DBSCAN, user perlu input data Eps, Minimum Samples, dan Metric. Selanjutnya sekaligus dimodelkan clustering data seperti ditunjukan gambar 9



Gambar 9. Implementasi algoritma DBSCAN

Pilihan selanjutnya adalah memilih metode clustering yang akan digunakan dan input parameter sesuai algoritma terpilih, sebagai contoh pada pengujian ketiga menggunakan algoritma mean Shift, user perlu input data Bandwidth, Bin Seeding, Cluster All, dan Maximum Iteration. Selanjutnya sekaligus dimodelkan clustering data seperti ditunjukan gambar 10

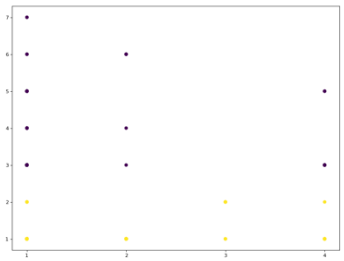
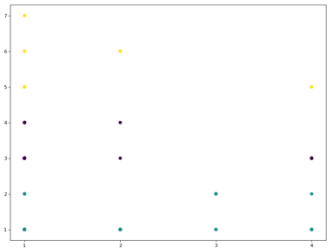
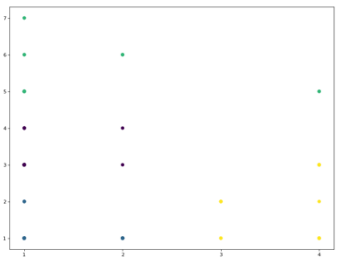
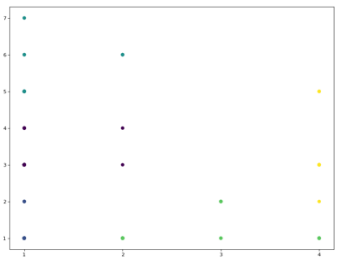


Gambar 10. Implementasi algoritma Mean Shift

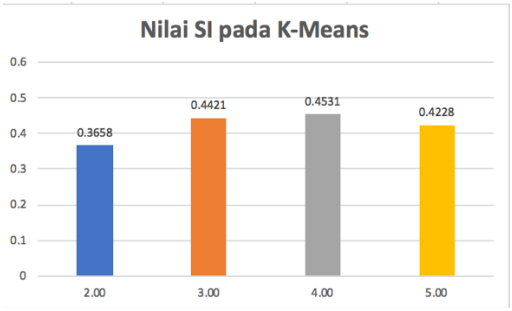
- Analisis Klasterisasi dengan Algoritma K-Means

Proses klasterisasi menggunakan algoritma K-Means dengan percobaan cluster 2, 3, 4, dan 5 cluster. Menurut pengamatan visual, cluster yang paling baik dihasilkan ketika jumlah clusternya 4.

TABEL III
PERBANDINGAN HASIL KLASTERISASI DENGAN K-MEANS

 2 cluster	 3 cluster
 4 cluster	 5 cluster

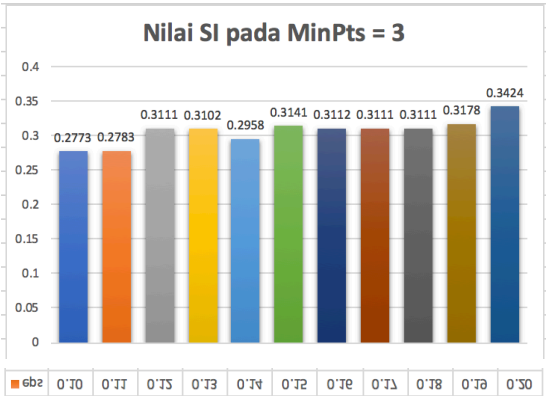
Kemudian juga dilakukan validitas cluster menggunakan Silhouette Index (SI). Nilai SI berdasarkan hasil klasterisasi data menggunakan algoritma K-Means dapat dilihat Gambar 11 berikut.



Gambar 11. Nilai SI hasil klasterisasi K-Means

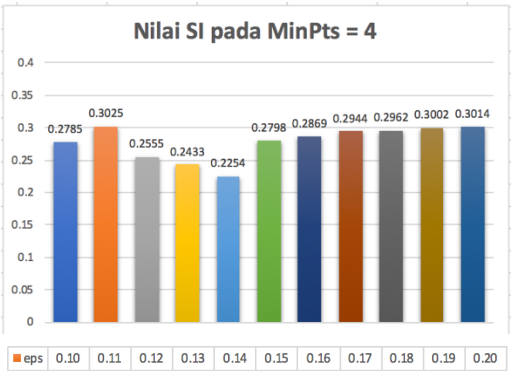
- Klasterisasi dengan Algoritma DBSCAN

Proses klasterisasi menggunakan algoritma DBSCAN dengan percobaan cluster sebanyak 22 kali dimana setiap percobaan menggunakan nilai epsilon (Eps) dan MinPoints (MinPts) yang berbeda. Nilai Eps yang digunakan pada penelitian ini yaitu pada rentang 0,1 hingga 0,2, sementara nilai MinPts 3 dan 4, selanjutnya dilakukan uji validitas cluster menggunakan SI. Nilai SI berdasarkan hasil klasterisasi menggunakan DBSCAN dapat dilihat pada Gambar 12



Gambar 12. Nilai SI pada MinPts = 3

Berdasarkan nilai SI terbaik, cluster terbaik pada algoritma DBSCAN terletak pada percobaan dengan nilai Eps 0,2 dan nilai MinPts 3 dengan nilai SI diperoleh 0,3424.



Gambar 13. Nilai SI pada MinPts = 4

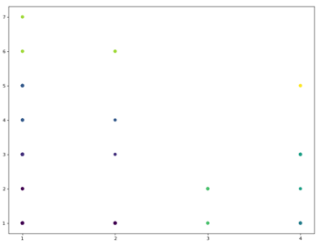
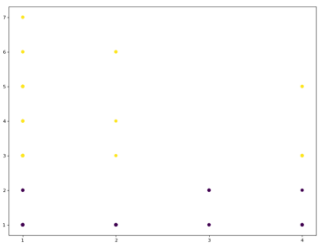
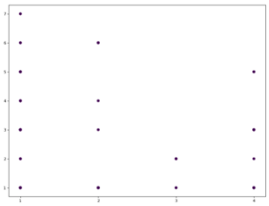
Percobaan selanjutnya dengan nilai Eps 0,15 dan MinPts 3 diperoleh nilai SI 0,3141.

- Klasterisasi dengan Algoritma MeanShift

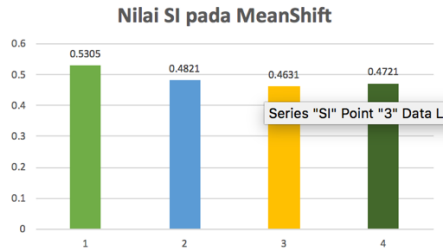
Pada MeanShift digunakan parameter yang sama dengan kedua jenis algoritma yang lainnya. Parameter dipilih agar hasil clustering memungkinkan untuk divisualisasikan dalam

scatter plot 2D. Pada algoritma ini dlakukan 3 percobaan dengan parameter Bandwidth yang berbeda-beda. Dengan bandwidth = 1, akan dihasilkan 6 cluster yang proporsional jumlah anggota tiap-tiap clusternya. Ketika bandwidth diganti menjadi = 2, akan menghasilkan 2 cluster. Untuk bandwidth >= 3, akan dihasilkan 1 cluster saja.

TABEL IV
PERBANDINGAN HASIL KLASTERISASI DENGAN MEANSHIFT

 Bandwidth = 1	 Bandwidth = 2
 Bandwidth >= 3	

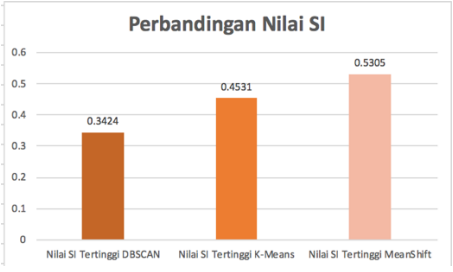
Kemudian juga dilakukan validitas cluster menggunakan Silhouette Index (SI). Nilai SI berdasarkan hasil klasterisasi data menggunakan algoritma MeanShift dapat dilihat Gambar 14 berikut.



Gambar 14. Nilai SI hasil klasterisasi MeanShift

- Perbandingan Algoritma K-Means, DBSCAN dan Mean Shift

Perbandingan algoritma K- Means, DBSCAN dan Mean Shift dilakukan dengan menentukan jumlah cluster yang paling optimal dilihat dari nilai Silhouette Index (SI). Grafik perbandingan nilai SI tertinggi pada setiap algoritma dapat dilihat pada Gambar 15 berikut.



Gambar 15. Hasil perbandingan Nilai SI

Hasil Pengujian Validitas Cluster Terhadap Algoritma K-Means, DBSCAN Dan Meanshift Menggunakan Nilai Silhouette Index (SI) Telah Dilakukan, Klasterisasi Menggunakan DBSCAN Yang Memiliki Nilai SI Terbaik Yaitu 0.3424 Dengan Nilai Eps 0,2 Dan Nilai Minpts 3. Sementara Dari Hasil Klasterisasi Menggunakan K-Means, Nilai SI Terbaik Diperoleh Percobaan K=4 Dengan Nilai 0,4531. Hasil Klasterisasi Menggunakan Meanshift, Nilai SI Terbaik Diperoleh Percobaan Bandwidth=1 Dengan Nilai 0,5305. Maka Pada Penelitian Ini, Algoritma Meanshift Memiliki Nilai Validitas Cluster Lebih Baik Dibandingkan Algoritma Lainnya. Dengan Demikian, Pada Diperoleh Cluster Paling Optimal Yaitu Percobaan Menggunakan Algoritma Meanshift Dengan Bandwidth=1.

IV. KESIMPULAN

Dalam penelitian ini telah berhasil dirancang dan diimplementasikan Aplikasi Berbasis Web Dengan Menerapkan Beberapa Metode Unsupervised Learning Yaitu K-Means, Dbscan Dan Mean Shift. Dari Hasil Percobaan Dan Pengolahan Data Didapatkan Hasil Klasterisasi Terbaik Yaitu Menggunakan Nilai Silhouette Index (Si) Dimana Klasterisasi Menggunakan Dbscan Yang Memiliki Nilai Si Terbaik Yaitu 0.3424 Dengan Nilai Eps 0,2 Dan Nilai Minpts 3. Sementara Dari Hasil Klasterisasi Menggunakan K-Means, Nilai Si Terbaik Diperoleh Percobaan K=4 Dengan Nilai 0,4531. Hasil Klasterisasi Menggunakan Meanshift, Nilai Si Terbaik Diperoleh Percobaan Bandwidth=1 Dengan Nilai 0,5305. Maka Pada Penelitian Ini, Algoritma Meanshift Memiliki Nilai Validitas Cluster Lebih Baik Dibandingkan Algoritma Lainnya. Dengan Demikian, Pada Diperoleh Cluster Paling Optimal Yaitu Percobaan Menggunakan Algoritma Meanshift Dengan Bandwidth=1

REFERENSI

[1] Barbará, D. et al. Applications of Data Mining in Computer Security. Kluwer Academic Publishers. 2002: 33-76.

[2] Chandrashekhar Azad, V. K. Data Mining in Intrusion Detection: A Comparative Study of Methods, Types and Data Sets. *International Journal of Information Technology and Computer Science*. 2013; 5(8): 75-90.

[3] Chauhan, N. S. *An introduction to the DBSCAN algorithm and its Implementation in Python.* Retrieved from https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html, 2020-04-01.

[4] Cynet. *EDR Security and Protection for the Enterprise*. Retrieved from Cynet: https://www.cynet.com/endpoint-protection-and-edr/top-6-edr-tools-compared/. 2019-09-29.

[5] D. F. Pramesti, M. T. "Implementasi Metode K-Medoids Clustering Untuk Pengelompokan Data Potensi Kebakaran Hutan / Lahan Berdasarkan Persebaran Titik Panas (Hotspot). *J-ptiik. 2017; 1(9)*: 723-732.

[6] Davy Cielen, A. D. Introducing Data Science: Big Data, Machine Learning, And More, Using Python Tools. Manning Publications. 2016: 1-30.

[7] Education, I. C. *Unsupervised Learning.* Retrieved from https://www.ibm.com/cloud/learn/unsupervised-learning. 2020-09-21.

[8] G. Yedukondalu, B. R. Intrusion Detection System Using Data Mining Techniques. *International Journal of Advanced Science and Technology*. *2020; 9(15)*: 1687-1695.

[9] H. Zayuka, S. M. Design and Analysis of Data Clustering Using K-Medoids Method For English News. *e-Proceeding Eng*, 2017; 2182-2190.

[10] Hyunseung Choi, M. K. Unsupervised learning approach for network intrusion detection system using autoencoders, . *The Journal of Supercomputing*. 2019; Vol.75: 5597-5621.

[11] I. Parlina, A. P. Memanfaatkan Algoritma K-Means Dalam Menentukan Pegawai Yang Layak Mengikuti Asessment Center. *Journal of Computer Engineering, System and Science*, 2018; 3(1): 87-93.

[12] S. Nasrin, et. al. Survey on SDN based network intrusion detection system using machine learning approaches. Peer-to-Peer Networking and Applications.2019; 12(1-2):1-9.

[13] FireEye Cyber Threat Map, https://www.fireeye.com/cyber-map/threat-map.html, diakses pada Tanggal 11 Oktober 2021.

[14] Live Cyber Attack Map, Live Cyber Attack Map, diakses pada tanggal 11 Oktober 2021.

UCAPAN TERIMA KASIH

Ucapan terimakasih diberikan kepada UPT P2M Polinema yang telah memberikan dukungan secara penuh baik moril maupun materiil sehingga penelitian ini bisa terlaksana dengan baik