Jurnal Sistim Informasi dan Teknologi
https://jsisfotek.org/index.php
2022 Vol. 4 No. 3 Hal: 112-117 e-ISSN: 2686-3154

Klasterisasi Penggunaan Trafik Internet Menggunakan K-Mean Clustering

Dedy Yasriady[1]✉

[1]Dinas Komunikasi Statistik dan Persandian Kota Pekanbaru

yasriady@gmail.com

Abstract

The use of internet traffic in a government office needs to be monitored carefully to obtain efficient and effective use. The internet line that has been provided is an official facility funded by the people's budget, so it needs to be monitored carefully. Domain Name System (DNS) provides rich and interesting data, and can be extracted to reveal information that can be analyzed for various purposes such as security measures, measuring traffic usage levels, bandwidth restrictions, user profiling to other policies implemented in a network. This study aims to make a clustering of the use of internet traffic so as to provide benefits that can be used to Improve Network Services (QoS), make efficiency of bandwidth usage and create user profiles. This research was conducted based on DNS Log which is operated on a network connected to the internet. In this study, it is shown how to consolidate traffic on port 53/udp to collect DNS logs, so that in this way the activities of internet users can be recorded in a centralized server until finally used as a primary data source. The datasets used are information extraction from the DNS Server log file (dnsmasq) which was retrieved for 5 working days in the period of working hours. The total extracted datasets used are 213 records. The available data is then processed to get the target cluster by utilizing the concept of data mining using the K-Mean Clustering method. The results classify the use of internet traffic into 3 clusters, namely high, medium and low. Each cluster consists of Kla1 is 23, Kla2 is 3, and Kla3 is 160. This research can be used as a reference in grouping internet traffic so that communication lines are well and smoothly maintained.

Keywords: K-Mean, Clustering, Dnsmasq, Data Mining, Internet Traffic.

Abstrak

Penggunaan trafik internet dalam suatu kantor pemerintah perlu diawasi secara cermat untuk memperoleh efisiensi pemakaiannya secara baik dan tepat guna. Jalur internet yang telah disediakan merupakan fasilitas resmi dibiayai dari anggaran yang bersumber rakyat sehingga perlu diawasi secara cermat. *Domain Name System (DNS)* menyediakan data yang kaya dan menarik, serta dapat diekstraksi untuk mengungkap informasi yang bisa dianalisis bagi berbagai keperluan seperti tindakan keamanan, mengukur tingkat penggunaan trafik, pembatasan *bandwidth*, *user profiling* hingga kebijakan lain yang diterapkan dalam suatu jaringan. Penelitian ini bertujuan membuat klasterisasi terhadap penggunaan trafik internet sehingga memberikan manfaat yang dapat digunakan untuk meningkatkan layanan jaringan (QoS), melakukan efisiensi terhadap pemakaian *bandwidth* serta membuat *profile* pengguna. Penelitian ini dilakukan berdasarkan *DNS Log* yang dioperasikan pada suatu jaringan yang terhubung ke internet. Pada penelitian ini diperlihatkan bagaimana melakukan konsolidasi trafik pada port 53/udp guna mengumpulkan *DNS log*, sehingga dengan cara ini aktivitas pengguna internet dapat dicatat dalam sebuah *server* terpusat hingga akhirnya digunakan sebagai sumber data primer. *Datasets* yang digunakan merupakan ekstraksi informasi berasal dari *log file* DNS Server (*dnsmasq*) yang diambil selama 5 hari kerja dalam periode jam kerja. Total *datasets* hasil ekstraksi yang digunakan adalah sebanyak 213 *records*. Data-data yang tersedia selanjutnya diolah untuk mendapatkan target klaster dengan memanfaatkan konsep data mining menggunakan metode *K-Mean Clustering*. Hasil mengelompokkan penggunaan trafik internet menjadi 3 klaster yaitu tinggi, sedang dan rendah. Masing-masing klaster terdiri dari Kla1 adalah 23, Kla2 adalah 3, dan Kla3 adalah 160. Penelitian ini dapat dijadikan rujukan dalam pengelompokan trafik internet sehingga jalur komunikasi terjaga dengan baik dan lancar.

Kata kunci: K-Means, Klasterisasi, *Dnsmasq*, *Data Mining*, Trafik Internet

1. Pendahuluan

Klasterisasi terhadap penggunaan trafik jaringan internet selama jam kerja [1] yang dilakukan oleh para pegawai pada Kantor Pemerintahan harus dilakukan untuk menjaga kelancaran dalam komunikasi data. Klasterisasi dapat digunakan untuk menemukan *profile* pengguna sehingga diharapkan memberikan manfaat dalam meningkatkan layanan jaringan (QoS) [2] serta melakukan efisiensi terhadap pemakaian *bandwidth* internet. QoS dapat dimanfaatkan sebagai data primer berupa *datasets* yang bersumber dari *log file* yang dihasilkan oleh Domain Name System (DNS) Server.

DNS menyediakan data yang kaya dan menarik, serta dapat dilakukan ekstraksi untuk mengungkap informasi yang menjadi dasar analisis bagi berbagai keperluan seperti tindakan keamanan [3], [4] *bandwidth management* hingga kebijakan lain yang diterapkan dalam suatu jaringan [5]. Data dapat diidentifikasi

Diterima: 22-08-2022 | Revisi: 30-08-2022 | Diterbitkan: 30-09-2022 | doi: 10.37034/jsisfotek.v4i3.141

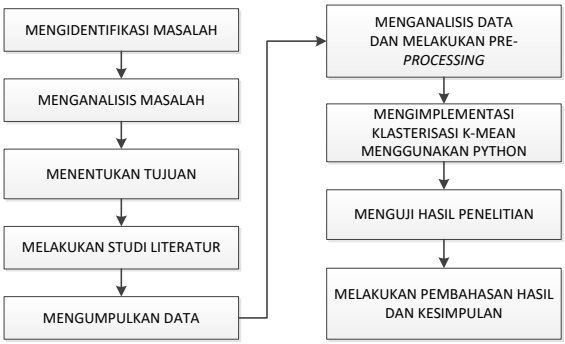
dengan memanfaatkan Knowledge Discovery in Database (KDD).

KDD merupakan suatu proses dalam melakukan identifikasi pola yang *valid*, baru dan berguna serta dapat dipahami dari sekumpulan data yang besar dan kompleks [6]. Inti proses yang dilakukan dalam KDD adalah Data Mining [7]. Data Mining melibatkan kesimpulan dari algoritma yang mengeksplorasi data, mengembangkan model serta menemukan pola yang sebelumnya tidak diketahui [8]. Proses penambangan atau mining dapat dilakukan dengan metode statistik, matematika hingga teknologi kecerdasan buatan (Artificial Intelligent) dan Machine Learning yang bertujuan untuk mengekstrak serta mengidentifikasi informasi dan pengetahuan potensial terkandung dalam suatu database besar [9].

Network Traffic Profiling dapat diolah dengan menggunakan Algoritma K-Mean pada Data Mining. *Trend* pola jaringan yang diakses pengguna internet serta menghasilkan profil trafik jaringan pada volume lalu lintas data yang tinggi menghasilkan tiga klaster berdasarkan penggunaan trafik, yaitu tinggi, sedang dan rendah sesuai dengan protokol layanan dan IP Address yang berbeda [10]. Dalam mendeteksi domain berbahaya dengan proses klasterisasi terhadap sejumlah besar data trafik DNS dalam menemukan domain berbahaya juga tepat dengan menggunakan metode K-Mean [11]. Penelitian yang memanfaatkan DNS log dalam membuat klasterisasi penggunaan trafik internet sangat diperlukan dalam menjaga kelancaran trafik [12], [13] maka penelitian ini bertujuan untuk mengklasterisasi penggunaan trafik internet menggunakan K-Mean Clustering.

2. Metodologi Penelitian

Metodologi yang dilakukan dalam penelitian ini disajikan pada Gambar 1.



Gambar 1. Metodologi Penelitian

Metodologi yang digunakan dalam penelitian ini dimulai dengan mengidentifikasi masalah, menganalisa masalah, menentukan tujuan, melakukan studi literatur hingga membuat pembahasan hasil dan kesimpulan dijelaskan sebagai berikut.

a. Mengidentifikasi Masalah

Identifikasi dilakukan guna menemukan akar permasalahan sehingga mencapai hasil sesuai dengan yang diharapkan. Dalam penelitian ini, masalah yang berhasil diidentifikasi adalah tidak tersedianya suatu mekanisme dalam memantau dan mengawas penggunaan trafik internet.

b. Menganalisis Masalah

Hal ini dilakukan agar permasalahan dapat dipahami secara baik untuk kemudian diselesaikan dengan langkah-langkah yang sesuai. Setelah melakukan observasi dan analisis, ditemukan beberapa hal yang menjadi penyebab, diantaranya tidak tersedia *log file* yang mencatat setiap aliran *DNS query*, sehingga bisa dikatakan tidak tersedia data awal yang menjadi landasan untuk memulai penelitian. Selama proses observasi lebih lanjut, juga ditemukan tidak adanya konsolidasi pencatatan *query* tersebut sehingga sulit untuk mendapatkan data primer dalam sebuah log file.

c. Menentukan Tujuan

Setelah memahami masalah yang ada, pada tahap ini dilanjutkan dengan menetapkan tujuan dari penelitian. Tujuan yang hendak dicapai dari penelitian ini adalah menemukan klaster penggunaan trafik internet berdasarkan *domain* yang menjadi target kunjungan menjadi 3 kelompok yang terdiri klaster tinggi, menengah dan rendah.

d. Melakukan Studi Literatur

Kegiatan ini dilakukan untuk mendapatkan referensi ilmiah serta teori yang mendasari penelitian, di samping untuk meningkatkan pemahaman dalam mengembangkan ide dan gagasan yang mendukung pelaksanaan penelitian.

e. Mengumpulkan Data

Pada tahap ini, dilakukan pengumpulan data yang akan menjadi objek penelitian.

f. Menganalisis Data dan Melakukan *Pre-Processing*

Tahap ini laksanakan untuk menganalisis data sudah diperoleh dari objek penelitian serta mempersiapkan data-data tersebut sehingga layak untuk diproses.

g. Mengimplementasikan Klasterisasi K-Mean

Bagian terakhir dari penelitian ini adalah melakukan pembahasan terhadap hasil yang diperoleh. Pada tahap ini juga dilakukan perumusan kesimpulan yang diperoleh selama melakukan penelitian.

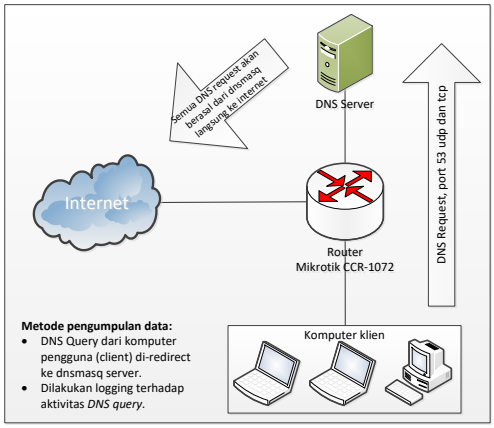
3. Hasil dan Pembahasan

3.1. Pengumpulan Data

Pengolahan data dimulai dari proses mengumpulkan data, *Pre-Processing*, proses Klasterisasi hingga mendapatkan hasil keluaran dan visualisasi berupa grafik yang menunjukkan keberadaan klaster yang

dihasilkan. Pada tahap *Pre-Processing* terdiri dari beberapa tahap seperti seleksi data, *cleaning* dan transformasi data.

Proses pengumpulan data dilakukan dalam jaringan dengan melakukan *redirect* seluruh DNS Traffic (port 53/udp) pada satu *server* yang terpusat. Topologi jaringan dalam mengumpulkan DNS Log disajikan pada Gamabr 2.



Gambar 2. Topologi Jaringan untuk Mengumpulkan *DNS Log*

Langkah-langkah dan metode yang digunakan dalam mengumpulkan data adalah sebagai berikut:

a. Seluruh trafik internet dari komputer klien mengalir melewati Router utama.
b. DNS *Traffic* (port 53/udp dan 53/tcp) yang berasal dari pengguna di-*redirect* ke *Server DNS*, *dnsmasq* kemudian meneruskan *DNS request* tersebut secara ke internet secara normal.
c. Dengan adanya *redirect* ini, maka layanan DNS *query* menjadi terpusat pada *single server*. Konfigurasi *logging* pada server kemudian diaktifkan untuk mencatat seluruh *DNS request* yang terjadi.

Langkah terakhir yang dilakukan dalam pengumpulan data ini adalah membuat tabel *datasets* yang menyimpan informasi *DNS Query* untuk setiap akses yang dilakukan oleh pengguna terhadap domain tujuan berdasarkan periode yang telah ditentukan. Dalam hal ini, data yang digunakan berasal dari *log file* selama 5 hari kerja mulai dari 13-Juli-2022 hingga 19-Juli-202, selama jam kerja dan tidak termasuk hari libur Sabtu dan Minggu dan waktu istirahat sekitar pukul 12:00 s/d 13:00. Untuk setiap harinya, kelompok data dibagi menjadi 2 sesi, yaitu sesi Pagi sebelum istirahat dan sesi Siang setelah jam istirahat. Hal ini bertujuan untuk memudahkan dalam melakukan *profiling* atau analisis lebih detail terhadap waktu-waktu pengguna dalam melakukan aktivitas *online*. Proses yang dilakukan dalam mempersiapkan *datasets* adalah sebagai berikut:

a. Memilih kolom *ClientIP* secara *distinct* (unik) dari tabel *dnslog* dan memasukkannya kedalam tabel *datasets*.
b. Menghitung jumlah *DNS Request* yang dilakukan tiap pengguna (*ClientIP*) untuk setiap hari yang dibagi menjadi 2 sesi pagi dan siang. Sesi pagi dibatasi pada waktu 08:00 s/d 12:00, sedangkan sesi siang adalah pukul 13:00 s/d 17:00.

Setelah selesai membuat *datasets*, dilanjutkan dengan pra proses lain yang diperlukan sehingga data menjadi *ready* untuk diproses. Dari hasil *cleaning* dan seleksi data yang dilakukan pada *datasets* tersebut, diperoleh sebanyak 213 *records* yang siap untuk diproses menggunakan K-Mean.

3.2. Normalisasi Data

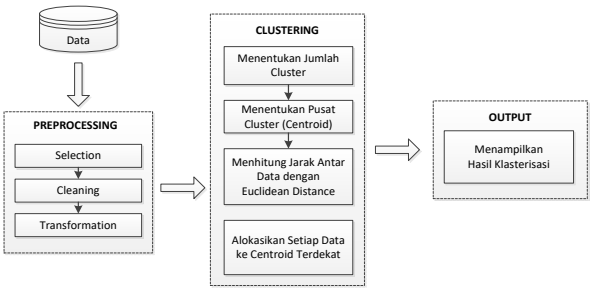
Data untuk penelitian yang diterima langsung dari sumbernya cenderung memiliki rentang dengan skala yang berbeda cukup jauh. Begitu juga pada penelitian ini, data yang terdapat pada kolom atribut terlihat cenderung memiliki rentang yang sangat jauh. Suatu atribut bisa mengandung nilai min-max sangat rendah sementara atribut lain bisa jadi memilki rentang min-max yang sangat tinggi. Untuk beberapa algoritma, tingginya perbedaan nilai (range) tersebut dapat menyebabkan kecenderungan variabel dengan rentang yang lebih besar memiliki pengaruh tidak semestinya pada hasil yang diperoleh [6], untuk itu maka perlu dilakukan normalisasi terhadap data yang diproses. Terdapat beberapa metode yang dapat digunakan, pada penelitian ini normalisasi data dilakukan dengan menggunakan metode Min-Max Normalization, sesuai dengan Persamaan 1.

$$X^*_{mm} = \frac{X - \min(X)}{range(X)} = \frac{X - \min(X)}{\max(X) - \min(X)} \quad (1)$$

Dimana X^*_{mm}adalah nilai hasil normalisasi. Sedangkan X adalah adalah nilai setiap atribut pada masing-masing data yang akan dinormalisasi.

3.3. Proses Klasterisasi K-Mean

Klasterisasi menggunakan algoritma K-Mean merupakan proses yang sederhana untuk digunakan serta memberikan hasil yang efektif dalam melakukan pengelompokan data. Tahapan proses disajikan pada Gamabr 3.



Gambar 3. Tahapan Proses Klasterisasi

Algoritma ini melaksanakan proses klasterisasi sebagai berikut [6]:

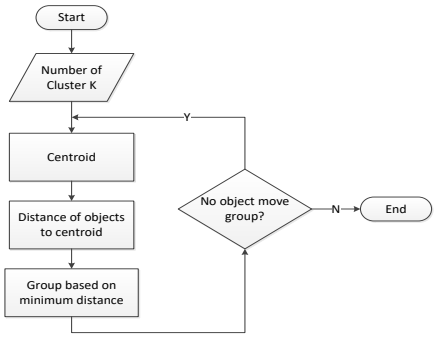
a. Langkah 1: Tentukan jumlah klaster (k).

b. Langkah 2: Secara random, tentukan sejumlah k *record* yang menjadi pusat klaster (*centroid*) untuk masing-masing klaster.
c. Langkah 3: Untuk masing-masing *record*, temukan pusat klaster terdekat. Sehingga masing-masing pusat klaster "memiliki" sebuah *subset* dari kumpulan data. Pada tahap ini kita memiliki klaster yang terdiri dari Kla1, Kla2, …, Kla3.
d. Langkah 4: Untuk setiap klaster k, temukan kembali *centroid*, dan update lokasi masing-masing pusat klaster menjadi nilai *centroid* yang baru.
e. Langkah 5: Ulangi langkah 3 hingga 5 hingga terjadi konvergensi.

Pada langkah 3, untuk mencari kriteria jarak terdekat yang dimaksud, dapat menggunakan persamaan *Euclidean Distance*:

$$d_{euclidean}(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (2)$$

Tahapan klasterisasi dilakukan dalam beberapa sub proses seperti menentukan jumlah klaster, menentukan *centroids*, menghitung jarak setiap data menggunakan *Euclidean Distance*, mengalokasikan tiap data ke centroid terdekat, serta melakukan pengulangan hingga anggota tiap klaster tidak lagi berubah, sebagaimana diperlihatkan pada *flow chart* pada Gambar 4.



Gambar 4. Algoritma *K-Mean Clustering*

Jumlah klaster yang dihasilkan merupakan hasil yang hendak dicapai. Dalam penelitian ini, jumlah klaster yang akan dibentuk adalah 3 kelompok berdasarkan jumlah akses terhadap situs yang dikunjungi, yaitu klaster tinggi, sedang dan rendah.

3.4. Menentukan Nilai Centroid Awal dan Atribut

Centroid awal perlu ditetapkan terlebih dahulu sebelum melakukan perhitungan lebih lanjut. Nilai untuk centroid awal dapat ditentukan secara acak dengan mengambil sebanyak 3 baris sesuai dengan jumlah klaster yang akan dibuat. Kemudian dibuat perhitungan jarak masing-masing data terhadap pusat klaster ini dengan persamaan *Euclidean Distance*.

Proses klasterisasi dilakukan dengan mengikuti konsep awal yang ditetapkan sebagai berikut:

a. Jumlah klaster yang akan dibentuk adalah sebanyak 3 klaster (tinggi, sedang dan rendah).
b. Data yang dijadikan sebagai *centroid* awal adalah:
   i. Centroid 1 (Kla1): diambil dari IP 10.110.1.46
   ii. Centroid 2 (Kla2): diambil dari IP 10.110.1.10
   iii. Centroid 3 (Kla3): diambil dari IP 10.110.1.106
c. Total jumlah data sebanyak 213.
d. Jumlah variabel atau atribut yang digunakan adalah 10 yang terdiri dari sample data 5 hari kerja. Masing-masing hari kerja dibagi menjadi 2 sesi pagi dan siang:
   i. A1 = Tanggal 13 Jul 2022, sesi Pagi
   ii. A2 = Tanggal 13 Jul 2022, sesi Siang
   iii. B1 = Tanggal 14 Jul 2022, sesi Pagi
   iv. B2 = Tanggal 14 Jul 2022, sesi Siang
   v. C1 = Tanggal 15 Jul 2022, sesi Pagi
   vi. C2 = Tanggal 15 Jul 2022, sesi Siang
   vii. D1 = Tanggal 18 Jul 2022, sesi Pagi
   viii. D2 = Tanggal 18 Jul 2022, sesi Siang
   ix. E1 = Tanggal 19 Jul 2022, sesi Pagi
   x. E2 = Tanggal 19 Jul 2022, sesi Siang
e. Normalisasi Min-Max terhadap data yang tersedia.
f. Data yang dijadikan sebagai *centroid* awal dapat dilihat pada Tabel 1.

Setelah penentuan nilai *centroid* awal, selanjutnya dilakukan penghitungan nilai jarak setiap data ke titik pusat (*centroid*) awal yang diselesaikan menggunakan formula *Euclidean Distance* (2). Nilai Centroid disajikan pada Tabel 1.

Tabel 1. Nilai Centroid Awal

	ClientIP	A1	A2	B1	B2	C1	C2	D1	D2	E1	E2
Kla1	10.110.1.46	0.486	0.496	0.358	0.220	0.005	0.014	0.968	0.295	0.023	0.097
Kla2	10.110.1.10	0.708	0.005	0.032	0.000	0.008	0.087	0.211	0.148	0.016	0.000
Kla3	10.110.1.106	0.079	0.003	0.017	0.005	0.670	0.000	0.005	0.000	0.027	0.101

3.5. Iterasi Pertama

Setelah penentuan centroid awal, jarak minimum setiap data ke pusat klaster ini dapat dilakukan menggunakan formula Euclidean Distance (D). Perhitungan jarak masing-masing data terhadap pusat klaster Kla1 dilakukan sesuai dengan proses berikut ini:

a. Data pertama (D1), dengan IP 10.110.1.10 mempunyai jarak ke pusat klaster Kla1 sebagai berikut:

D1 = $((0.71 - 0.49)^2 + (0.01 - 0.50)^2 + (0.03 - 0.36)^2 + (0.00 - 0.22)^2 + (0.01 - 0.01)^2 + (0.09 - 0.01)^2 + (0.21 - 0.97)^2 + (0.15 - 0.30)^2 + (0.02 - 0.02)^2 + (0.00 - 0.10)^2) \wedge 0.5$

= 1.03

b. Data kedua (D2), dengan IP 10.110.1.100 dapat dihitung jarak ke pusat klaster Kla1 sebagai berikut:

D2 = ((0.04 - 0.49)2 + (0.03 - 0.50)2 + (0.03 - 0.36)2 + (0.00 - 0.22)2 + (0.04 - 0.01)2 + (0.16 - 0.01)2 + (0.05 - 0.97)2 + (0.38 - 0.30)2 + (0.02 - 0.02)2 + (0.01 - 0.10)2) ^ 0.5

= 1.20

c. Perhitungan untuk data ketiga dan seterusnya menggunakan cara yang sama.

Selanjutnya untuk pusat klaster Kla2, penghitungan jarak masing-masing data dilakukan dengan cara yang sama dengan langkah sebelumnya, yaitu:

a. Data pertama (D1), dengan IP 10.110.1.10 mempunyai jarak ke pusat klaster Kla2 sebagai berikut:

$D1 = ((0.71 - 0.71)^2 + (0.01 - 0.01)^2 + (0.03 - 0.03)^2 + (0.00 - 0.00)^2 + (0.01 - 0.01)^2 + (0.09 - 0.09)^2 + (0.21 - 0.21)^2 + (0.15 - 0.15)^2 + (0.02 - 0.02)^2 + (0.00 - 0.00)^2) \wedge 0.5$

= 0.00

Nilai D2 adalah 0 karena data ini merupakan centroid awal untuk klaster Kla2 yang sedang dihitung.

b. Data kedua (D1), dengan IP 10.110.1.100 dapat dihitung jarak ke pusat klaster Kla2 sebagai berikut:

$D2 = ((0.04 - 0.71)^2 + (0.03 - 0.01)^2 + (0.03 - 0.03)^2 + (0.00 - 0.00)^2 + (0.04 - 0.01)^2 + (0.16 - 0.09)^2 + (0.05 - 0.21)^2 + (0.38 - 0.15)^2 + (0.02 - 0.02)^2 + (0.01 - 0.00)^2) \wedge 0.5$

= 0.73

c. Perhitungan untuk data ketiga dan seterusnya dapat dilakukan dengan cara yang sama.

Dari pengelompokan yang dilakukan pada iterasi pertama diperoleh hasil klaster sementara dengan anggota sebagai berikut:

a. Anggota Klaster 1 (Kla1) terdiri dari 14 ClientIP sesuai nomor urut: 13, 23, 29, 41, 63, 66, 81, 127, 128, 160, 186, 200, 211, 213.
b. Klaster 2 (Kla2) mempunyai anggota sementara sebanyak 52 ClientIP terdiri dari: 1, 2, 4, 5, 7, 9, 12, 17, 19, 20, 21, 24, 26, 32, 38, 40, 43, 51, 62, 85, 93, 107, 108, 111, 113, 119, 125, 130, 132, 133, 135, 138, 141, 145, 147, 150, 151, 154, 157, 158, 162, 165, 178, 180, 185, 188, 189, 190, 195, 203, 206, 207.
c. Sedangkan pada Klaster 3 (Kla3) pada iterasi pertama ini mempunyai anggota sebanyak 147 ClientIP yaitu: 3, 6, 8, 10, 11, 14, 15, 16, 18, 22, 25, 27, 28, 30, 31, 33, 34, 35, 36, 37, 39, 42, 44, 45, 46, 47, 48, 49, 50, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 64, 65, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 82, 83, 84, 86, 87, 88, 89, 90, 91, 92, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 109, 110, 112, 114, 115, 116, 117, 118, 120, 121, 122, 123, 124, 126, 129, 131, 134, 136, 137, 139, 140, 142, 143, 144, 146, 148, 149, 152, 153, 155, 156, 159, 161, 163, 164, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 179, 181, 182, 183, 184, 187, 191, 192, 193, 194, 196, 197, 198, 199, 201, 202, 204, 205, 208, 209, 210, 212.

Dari beberapa kali iterasi yang dilakukan, posisi anggota masing-masing klaster tidak lagi mengalami perubahan terdeteksi pada iterasi ke-10. Artinya kondisi *final* sudah diperoleh pada iterasi ke-9. Dengan pusat klaster ditampilkan pada Tabel 2.

Tabel 2. Pusat Klaster Iterasi Ke-9

	ClientIP	A1	A2	B1	B2	C1	C2	D1	D2	E1	E2
Kla1	10.110.1.46	0.085	0.058	0.166	0.104	0.114	0.118	0.413	0.317	0.465	0.421
Kla2	10.110.1.10	0.358	0.439	0.267	0.169	0.171	0.111	0.148	0.095	0.129	0.048
Kla3	10.110.1.106	0.035	0.050	0.068	0.057	0.103	0.072	0.096	0.075	0.074	0.056

Nilai centroid masing-masing pusat klaster ke-9 ini diperoleh dari rata-rata anggota klaster pada iterasi sebelumnya (iterasi ke-8). Berdasarkan nilai-nilai ini, maka analisis jarak masing-masing data terhadap pusat klaster (iterasi ke-9) adalah sebagai berikut:

a. Data pertama (D1), dengan IP 10.110.1.10 mempunyai jarak ke pusat klaster Kla1 sebagai berikut:

$D1 = ((0.71 - 0.08)^2 + (0.01 - 0.06)^2 + (0.03 - 0.17)^2 + (0.00 - 0.10)^2 + (0.01 - 0.11)^2 + (0.09 - 0.12)^2 + (0.21 - 0.41)^2 + (0.15 - 0.32)^2 + (0.02 - 0.46)^2 + (0.00 - 0.42)^2) \wedge 0.5$

= 0.94

b. Data kedua (D2), dengan IP 10.110.1.100 dapat dihitung jarak ke pusat klaster Kla1.

$D2 = ((0.04 - 0.08)^2 + (0.03 - 0.06)^2 + (0.03 - 0.17)^2 + (0.00 - 0.10)^2 + (0.04 - 0.11)^2 + (0.16 - 0.12)^2 + (0.05 - 0.41)^2 + (0.38 - 0.32)^2 + (0.02 - 0.46)^2 + (0.01 - 0.42)^2) \wedge 0.5$

= 0.73

c. Perhitungan untuk data ketiga dan seterusnya menggunakan cara yang sama.

Selanjutnya untuk pusat klaster Kla2, pada iterasi ke-9 ini penghitungan jarak masing-masing data dilakukan dengan cara yang sama, yaitu:

a. Data pertama (D1), dengan IP 10.110.1.10 mempunyai jarak ke pusat klaster Kla2 sebagai berikut:

$D1= ((0.71 - 0.36)^2 + (0.01 - 0.44)^2 + (0.03 - 0.27)^2 + (0.00 - 0.17)^2 + (0.01 - 0.17)^2 + (0.09 - 0.11)^2 + (0.21 - 0.15)^2 + (0.15 - 0.09)^2 + (0.02 - 0.13)^2 + (0.00 - 0.05)^2) \wedge 0.5$

$= 0.67$

Nilai D2 adalah 0 karena data ini merupakan centroid awal untuk klaster Kla2 yang sedang dihitung ini.

d. Data kedua (D1), dengan IP 10.110.1.100 dapat dihitung jarak ke pusat klaster Kla2 sebagai berikut:

$D2 = ((0.04 - 0.36)^2 + (0.03 - 0.44)^2 + (0.03 - 0.27)^2 + (0.00 - 0.17)^2 + (0.04 - 0.17)^2 + (0.16 - 0.11)^2 + (0.05 - 0.15)^2 + (0.38 - 0.09)^2 + (0.02 - 0.13)^2 + (0.01 - 0.05)^2) \wedge 0.5$

$= 0.69$

e. Perhitungan untuk data ketiga dan seterusnya dapat dilakukan dengan cara yang sama.

3.6. Mengelompokkan Data Berdasarkan Klaster

Berdasarkan hasil kalkulasi yang dilakukan pada iterasi ke-9, maka penelitian yang dilakukan telah berhasil mengelompokkan data pada klaster masing-masing:

Kla1 = 23, Kla2= 3, Kla3 =160.

3.7. Pengujian Dengan Software Weka

Hasil analisis dan perancangan yang telah dilakukan dapat diuji menggunakan aplikasi yang tersedia saat ini. Pada penelitian ini, pengujian dilakukan dengan menggunakan software Weka, Gambar 5 menunjukkan hasil pengujian tersebut.

```
Final cluster centroids:
                        Cluster#
Attribute    Full Data          0          1          2
               (213.0)    (23.0)    (30.0)   (160.0)
=============================================================
A1           151.1972  151.1739        619   63.4875
A2                217   70.0435   929.5333   104.525
B1           322.1174  453.1739      834.8    207.15
B2           248.2629  278.1304   567.5333  184.1062
C1           332.7934   280.913      525.5  304.1187
C2           242.7981  294.3478   359.4667  213.5125
D1           477.7136      1386        557  332.2813
D2           232.8498  676.3913      233.2   169.025
E1           391.6667 1424.5652      442.9  233.5813
E2           268.4085 1261.3913   144.9667  148.8125



Time taken to build model (full training data) : 0.03 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       23 ( 11%)
1       30 ( 14%)
2      160 ( 75%)
```

Gambar 5. Hasil Pengujian Menggunakan Weka

4. Kesimpulan

Dari penelitian yang telah dilakukan dapat ditarik kesimpulan bahwa klasterisasi menggunakan K-Mean mampu menghasilkan 3 klaster penggunaan trafik internet berdasarkan *datasets* yang berasal dari *dnsmasq log*. Sebelum diproses lebih lanjut, data mentah (*raw data*) dari *log file* tersebut terlebih dahulu ditransformasikan menjadi *datasets* yang sesuai. Analisis metode K-Mean dapat dilakukan secara manual menggunakan aplikasi Microsoft Excel, dengan hasil pengujian pada aplikasi Weka menunjukkan hasil dengan klaster yang sama yaitu Kla1 sebanyak 23, Kla2 sebanyak 30, dan Kla3 terdiri dari 160.

Daftar Rujukan

[1] Hadi, H. A., Dwilestari, G., Faqih, A., & Nuris, N. D. (2022). Manajemen Authentifikasi User Menggunakan Metode Radius Server pada RS Jantung Hasna Medika. *KOPERTIP: Jurnal Ilmiah Manajemen Informatika dan Komputer*, *6*(2), 34-41. DOI: https://doi.org/10.32485/kopertip.v6i2.133

[2] Rubangiya, R., Hartati, T., & Wijaya, Y. A. (2022). Analisis Data Lalu Lintas Jaringan Di Kantor Cangehgar Cyber Operation Center Menggunakan Algoritma K-Means. *Network Engineering Research Operation*, *7*(1), 75-84. DOI: http://dx.doi.org/10.21107/nero.v7i1.327

[3] Salam, N. N. (2013). Analisis Implementasi Kebijakan Publik: Studi Kasus Pemblokiran Konter Pornografi di Internet (Doctoral dissertation, Fisipol UGM Politik dan Pemerintahan dh. Ilmu Pemerintahan).

[4] Amiel, T., & Sargent, S. L. (2004). *Individual differences in Internet usage motives*. *Computers in Human Behavior*, 20(6), 711–726. https://doi.org/10.1016/j.chb.2004.09.002

[5] Dakhgan, A., Hadi, A., Al Saraireh, J., & Alrababah, D. (2017, October). Passive DNS Analysis Using Bro-IDS. In 2017 *International Conference on New Trends in Computing Sciences (ICTCS)* (pp. 121-126). IEEE. https://doi.org/10.1109/ICTCS.2017.47

[6] Maimon, O., & Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook (2nd ed. 2010 ed.)*. Springer.

[7] Larose, D. T. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining.* Wiley.

[8] Ginantra, N. L. W. S. R., Arifah, F. N., Wijaya, A. H., Septarini, R. S., Ahmad, N., Ardiana, D. P. Y., ... & Negara, E. S. (2021). Data mining dan penerapan algoritma. Yayasan Kita Menulis.

[9] Turban, E., Aronson, J. E., Liang, T., & McCarthy, R., V. (2004). *Decision Support Systems and Intelligent Systems (7th ed.)*. Prentice Hall.

[10] Mohd Ariffin, M. A. (2020). *Network Traffic Profiling Using Data Mining Technique in Campus Environment. International Journal of Advanced Trends in Computer Science and Engineering*, 9(1.3), 422–428. https://doi.org/10.30534/ijatcse/2020/6691.32020

[11] Wang, Q., Li, L., Jiang, B., Lu, Z., Liu, J., & Jian, S. (2020). *Malicious Domain Detection Based on K-means and SMOTE. Lecture Notes in Computer Science*, 468–481. https://doi.org/10.1007/978-3-030-50417-5_35

[12] Cui, H., Yang, J., Liu, Y., Zheng, Z., & Wu, K. (2014). *Data Mining-based DNS Log Analysis*. Annals of Data Science, 1(3–4), 311–323. https://doi.org/10.1007/s40745-014-0023-7

[13] Grzinic, T., Perhoc, D., Maric, M., Vlasic, F., & Kulcsar, T. (2014). CROFlux — *Passive DNS method for detecting fast-flux domains.* 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). https://doi.org/10.1109/mipro.2014.6859782