Buletin Poltanesa Vol. 23 No. 2 (Desember 2022) p-ISSN 2721-5350 e-ISSN 2721-5369
https://doi.org/10.51967/tanesa.v23i2.1736 © 2022 Politeknik Pertanian Negeri Samarinda


# Analisis *Clustering Trafik* Jaringan Menggunakan Metode K-*Means*

**Muhammad Fahmi***
Sistem Informasi, STMIK Widya Cipta Dharma,
Samarinda, 75123
mfahmi@wicida.ac.id
**Corresponding author*

**Ahmad Fajri**
Teknik Informatika, STMIK Widya Cipta Dharma,
Samarinda, 75123
ahmadfajri@wicida.ac.id

***Abstrak*—**Penelitian ini dilakukan untuk membuat sebuah model *clustering* trafik internet dengan menggunakan algoritma K-*means*. Penggunaan internet di wilayah kampus banyak digunakan oleh mahasiswa dan staf pegawai untuk keperluan proses belajar mengajar, ataupun membantu proses bekerja. Penggunaan internet, pada jam-jam sibuk dan perkuliahan yang aktif membuat kecepatan internet menjadi lambat. Hal ini dipengaruhi oleh banyaknya pengiriman paket *header* pada *flow*/arus lalu lintas internet yang membuat koneksi pada internet menjadi semakin berat/lambat. Dalam mengatasi masalah tersebut, maka diperlukan metode *clustring* penggunaan trafik internet dengan algoritma k-*means* yang dapat mengetahui jenis atau trafik internet dengan berdasarkan fitur arus/*flow* paket trafik internet menggunakan metode pengembangan data *mining*. Data yang digunakan dalam penelitian ini yaitu data yang diambil melalui hasil *capture wifi* selama 3 hari di sub bagian Pusat Komputer kampus STMIK WiCiDa menggunakan *wireshark* dan *bettercap* yang akan melakukan serangan *arp spoof*, dimana metode ini akan membuat penulis diposisikan sebagai penengah dan dapat menangkap paket dari semua perangkat di jaringan yang sama. *Tools* ini akan dijalankan di kali *linux*. Data paket yang sudah dicapturing dan difilter kemudian akan di*export* dalam bentuk .pcap. Hasil penelitian ini berupa model algoritma Clustring trafik jaringan dengan metode k-means yang dapat meng-*cluster*kan arus penggunaan trafik internet dengan tiga *cluster* yaitu Web, Video VoIP, *Network*. Pada saat pengujian dengan menggunakan tiga *cluster* menghasilkan nilai akurasi data yang baik Mendapatkan hasil *Clustring* yaitu : *Cluster* 0 = 302638 data, *Cluster* 1 = 331982 data, dan *Cluster* 3 = 451426 data.

***Kata Kunci*—**Machine Learning, Clustring, Jaringan Komputer, Trafik jaringan, K-Means

## I. PENDAHULUAN

Penggunaan internet di wilayah kampus merupakan kebutuhan yang penting untuk mendukung aktivitas di kampus. Hal ini menjadi poin penting untuk kampus memfasilitasi mahasiswanya ataupun staf-staf yang ada dengan memberikan kelancaran pada saat menggunakan internet yang disediakan dan memberikan bandwith yang sangat besar, namun seringkali besar bandwidth yang dirasa kurang, terutama pada jam-jam sibuk dan perkuliahan yang aktif (Prathivi, 2015).

Keterbatasan sumber daya jaringan serta semakin meningkatnya pengguna internet saat ini berdampak pada tingginya trafik yang mengakibatkan menurunnya kecepatan akses pada layanan internet (Tasmi et al., 2021).

Meningkatnya jumlah pengguna internet menyebabkan banyak sektor yang menggunakan jaringan internet untuk menyediakan layanan kepada para pelanggannya. STMIK WiCiDa merupakan salah satu stake holder yang banyak memanfaatkan internet sebagai salah satu fasilitas yang disediakan bagi civitasnya. Jaringan internet merupakan salah bagian dari infrastruktur kampus yang harus dapat dijaga (Anggraeni & Andriani, 2021).

Hal ini dipengaruhi oleh banyaknya pengiriman paket *header* pada flow/aliran lalu lintas internet yang membuat koneksi pada internet menjadi semakin berat/lambat. Sehingga perlu diketahui bagaimana mengidentifikasi trafik internet yang ada selama ini, hal ini dapat berguna untuk dijadikan dasar kebijakan manajemen koneksi internet untuk saat sekarang dan diwaktu yang akan datang (Yasriady, 2022).

Selain itu proses mengidentifikasi trafik internet dapat menunjukkan aktifitas pengguna sehari-hari dalam menggunakan platform tersebut dan aplikasi apa saja yang digunakan oleh mayoritas pengguna selama ini. Hal tersebut berkaitan dengan tujuan utama dan prioritas ketersediaan internet. Sehingga jangan sampai, internet lebih banyak dimanfaatkan untuk hal-hal di luar tujuan utamanya (Amri et al., 2019).

Teknik klasifikasi trafik yang umum digunakan didasarkan *ip address* pada pemeriksaan langsung terhadap konten setiap paket di beberapa titik di jaringan. Paket-paket IP yang berurutan memiliki 5 tupel tipe protokol yang sama, *source address*, *port* dan *destination address* dianggap milik *flow* yang aplikasi pengontrolnya ingin ditentukan . Pendekatan yang lebih baru mengandalkan karakteristik statistik trafik untuk mengidentifikasi aplikasi. Hal yang mendasari metode tersebut adalah bahwa trafik pada jaringan memiliki sifat statistik, seperti distribusi durasi arus, waktu idle arus, waktu antar paket dan panjang paket yang unik untuk kelas aplikasi tertentu dan memungkinkan sumber aplikasi yang berbeda untuk dibedakan satu sama lain (Premitasari, 2019).

https://doi.org/10.51967/tanesa.v23i2.1736 © 2022 Politeknik Pertanian Negeri Samarinda


Analisis kluster bertujuan untuk mengelompokan objek data yang mempunyai kemiripan karakteristik satu sama lain dalam kluster yang sama dan berbeda karakteristiknya terhadap objek yang berbeda kluster (Nagari & Inayati, 2020). K-Means adalah metode yang termasuk dalam algoritma klustering berbasis jarak yang dimulai dengan menentukan jumlah kluster yang diinginkan (Nagari & Inayati, 2020).

Hasil yang diharapkan dari penelitian ini adalah untuk memperoleh informasi tentang arus penggunaan trafik internet atau lalu lintas jaringan penggunaan internet, pada jam-jam sibuk dan perkuliahan yang aktif membuat kecepatan internet menjadi lambat.

## II. STUDI PUSTAKA

### *A. Lalu Lintas Jaringan*

Istilah trafik atau lalu lintas dapat diartikan sebagai banyaknya informasi yang melewati saluran komunikasi. Pemantauan lalu lintas jaringan adalah salah satu cara untuk mempelajari dan mengidentifikasi penggunaan jaringan. Pemantauan jaringan atau Monitoring jaringan adalah proses pengumpulan data lalu lintas jaringan untuk dilakukan analisis terhadap data tersebut dengan tujuan memaksimalkan pemanfaatan seluruh sumber daya jaringan. Pemantauan data lalu lintas jaringan dapat digunakan untuk mengetahui aktivitas yang terjadi selama koneksi berlangsung (Hartati & Arie Wijaya, 2022).

### *B. Data Mining*

Data Mining adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam database. Data mining adalah proses yang menggunakan Teknik statistic, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar. Pemilihan tugas Data Mining merupakan pemilihan goal dari proses KDD misalnya karakterisasi, klasifikasi, regresi, clustering, asosiasi, dan lain-lain. Pemilihan teknik, metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan (Darma et al., 2020).

### *C. Anomali Trafik*

Anomali trafik merupakan sebuah kejanggalan alur pada trafik data jaringan. Ini disebabkan oleh adanya aktifitas-aktifitas dalam jaringan yang menyimpang dari batas normal. Anomali yang terjadi bisa dilihat melalui kenaikan lonjakan pengguna internet, melalui serangan pada suatu trafik dan lonjakan yang tidak disengaja. Kenaikan lonjakan dapat dilihat pada saat adanya bencana yang terjadi kejadian yang tidak biasa terjadi. Kenaikan lonjakan yang terjadi menimbulkan penurunan performansi dari suatu jaringan. Untuk itu perlu dilakukan deteksi terhadap anomali yang terjadi (Rizqi utami, A., Purwanto, Y., Anbarsanti, 2017).

### *D. Clustring*

*Clustering* adalah suatu teknik analisis dalam pengelompokan objek berdasarkan informasi yang diperoleh. Pada dasarnya objek akan saling berhubungan satu sama lain untuk memaksimalkan dan meminimalkan kesamaan dari,anggota cluster. Clustering dapat dilakukan pada data yang memiliki beberapa atribut yang dipetakan sebagai ruang multidimensi (Hartati & Arie Wijaya, 2022).

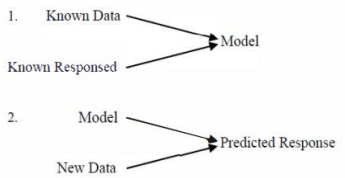
### *E. K-Means*

Metode K-Means adalah metode yang termasuk dalam algoritma clustering berbasis jarak yang membagi data ke dalam sejumlah cluster dan algoritma ini hanya bekerja pada atribut numeric. K-Means adalah suatu metode penganalisaan data atau Data Mining yang melakukan proses pemodelan tanpa supervisi (unsupervised) dan merupakan salah satu metode yang melakukan pengelompokan data dengan sistem partisi (Darma et al., 2020).

### *F. Machine Learning*

Machine Learning atau pembelajaran mesin merupakan salah satu kecerdasan buatan yang memungkinkan mesin melakukan pembelajaran berdasarkan contoh data, pembelajaran mesin memanfaatkan hubungan antar variabel dan probabilitas untuk menghasilkan prediksi (Tasmi et al., 2021).

Berdasarkan masukan dan keluaran yang diharapkan, pembelajaran mesin terbagi menjadi dua kelompok yaitu : Supervised Learning, merupakan pembelajaran yang bertujuan untuk memetakan masukan dan keluaran yang diinginkan seperti pada pengelompokkan, dengan menghasilkan sebuah model yang mampu memetakan masukan yang baru menjadi keluaran yang tepat. Prinsip kerja *Supervised Learning* yaitu dengan mempelajari sekumpulan contoh masukan dan keluaran dan menghasilkan sebuah model yang mampu memetakan masukan yang baru menjadi keluaran yang tepat. Seperti yang rerlihat pada Gambar 1 deskripsi supervised learning.




Gambar 1. Deskripsi Supervised Learning (Shafiq et al., 2017)

Terdapat beberapa algoritma yang dapat diterapkan antara lain, *Naïve Bayes*, SVM dan *Decision Tree*. *Unsupervised Learning*, merupakan pembelajaran yang memodelkan himpunan masukan untuk mempelajari dan mencari pola-pola tertentu pada masukan yang diberikan, *Clustering* atau penggolongan merupakan penerapan dari pembelajaran ini. Beberapa algoritma pada pembelajaran ini antara lain, *Birch*, *Cure* dan *K-Mean* (Shafiq et al., 2017).

https://doi.org/10.51967/tanesa.v23i2.1736 

## III. METODOLOGI

Pada tahapan dalam penelitian ini akan dilakukan langkah-langkah diagram alur berdasarkan SKKNI No. 299 Tahun 2020. Dapat di lihat pada Gambar 2 diagram alur SKKNI.


Gambar 2. Diagram alur SKKNI No. 299 Tahun 2020

Adapun penjelasan tahapan tersebut adalah sebagai berikut:

Dari tahapan penelitian ini, dijelaskan masing-masing Langkah kerja sebagai berikut:

1. Tempat dan waktu penelitian

   Kegiatan penelitian ini dilakukan di STMIK Widya Cipta Dharma berada di jalan M. Yamin, Gn. Kelua, Kec. Samarinda Ulu, Kota Samarinda, Kalimantan Timur 75123.
2. Studi Pustaka

   Studi pustaka pada penelitian ini digunakan untuk mempelajari teori-teori yang dapat mendukung proses penelitian. Teori yang dipelajari antara lain data mining dan metode klasifikasi menggunakan metode K-Means.
3. Capture Data Trafik

   Data yang digunakan dalam penelitian ini yaitu data yang diambil melalui hasil capture wifi selama 5 hari di sub bagian Puskom (Pusat Komunikasi) kampus STMIK Widya Cipta Dharma menggunakan wireshark dan bettercap yang akan melakukan serangan arp spoof, dimana metode ini akan membuat penulis diposisikan sebagai penengah dan dapat menangkap paket dari semua perangkat di jaringan yang sama. tools ini akan dijalankan di kali linux. Data paket yang sudah dicapturing dan difilter kemudian akan diexport dalam bentuk .pcap .
4. Generate Data Trafik

   Data trafik yang berbentuk .pcap akan dilakukan proses generate menggunakan NFStream untuk mendapatkan informasi secara detail mengenai fitur pada paket header trafik internet dan mendeteksi trafik , serta mendapatkan informasi mengenai arus pada paket trafik yang sudah melalui proses stastistika. Proses generate akan menghasilkan sebuah file .csv yang akan dilakukan proses data mining.
5. Bussiness Understanding

   Pada penelitian ini berhubungan langsung dengan penggunaan data trafik internet yang terhubung dengan Puskom WiCiDa. Peneliti mengambil data trafik internet secara langsung dengan menggunakan wireshark dan bettercap untuk arp spoofing, agar dapat melihat data trafik internet yang terhubung dengan puskom dalam jaringan yang sama. Serta melihat penggunaan trafik yang padat oleh aplikasi tertentu yang mempengaruhinya dan melihat kategori trafik internet pada Puskom WiCiDa.
6. Data Understanding

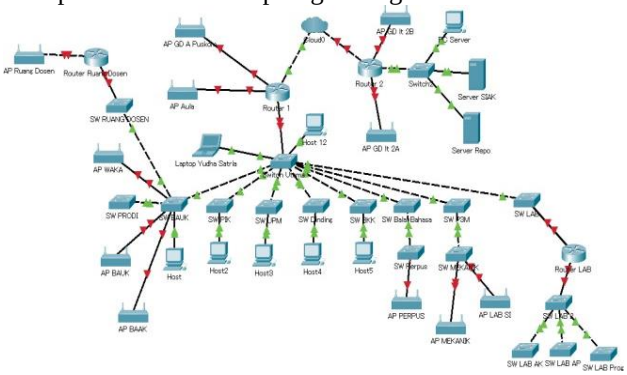
   Tahap ini dimulai setelah mengumpulkan arus data trafik internet yang sudah di generate nfstream dan menghasilkan atribut dengan trafik yang lebih rinci. kemudian dilakukan proses analisis data trafik internet yang digunakan .
7. Data Preparation

   Pada tahap ini akan melakukan proses persiapan data dengan data trafik internet Puskom WiCiDa yang memiliki 45 atribut, dan perlu dilakukan proses pembersihan untuk menghilangkan data duplikat, dan memilih atribut, serta data yang digunakan untuk digunakan dalam tahap pemodelan (modeling).

## IV. HASIL DAN PEMBAHASAN

### *A. Data Trafik Internet*

Proses ini diawali dengan menghubungkan kedalam jaringan Puskom WiCiDa yang dijalankan dengan sistem operasi kali linux didalam virtual box. Agar mendapatkan koneksi internet, terlebih dahulu melakukan konfigurasi pada virtualbox, dengan mengganti network adapter dengan "Bridged Adapter". Proses ini dilakukan agar kali linux host dapat terhubung langsung dengan koneksi internet puskom dan jalankan perintah "nano /etc/network/interfaces" didalam terminal kali linux dan konfigurasi *ip static* pada eth0. Berikut topologi yang bisa lihat pada Gambar 3 Topologi Jaringan Wicida.



Gambar 3. Topologi Jaringan Wicida

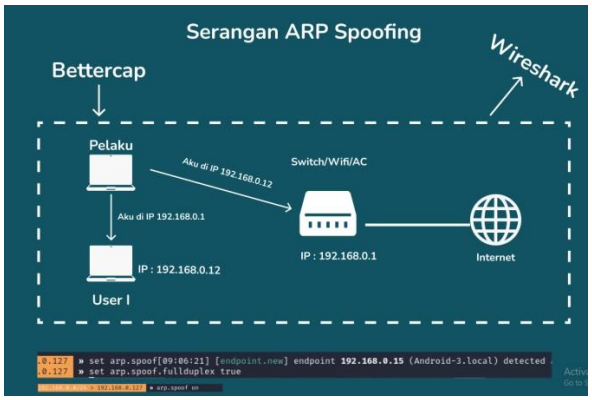
Kemudian menjalankan perintah "service networking restart" pada terminal agar settingan jaringan di kali linux diperbarui dengan konfigurasi terbaru. Kemudian proses MiTM attack dilakukan dengan menggunakan perangkat lunak bettercap dengan menjalankan perintah yang dapat dilihat pada Gambar 4 perintah menjalankan *bettercap.*

https://doi.org/10.51967/tanesa.v23i2.1736 © 2022 Politeknik Pertanian Negeri Samarinda



Gambar 4. Perintah Menjalankan *Bettercap*

Perintah ini akan menampilkan daftar perangkat yang akan dilakukan serangan dan terhubung dalam jaringan yang sama (Agustiara et al., 2022). Kemudian dijalankan proses serangan arp spoof dengan perintah yang dapat dilihat pada Gambar 5 mengaktifkan serangan Arp Spoof secara penuh.




Gambar 5. Mengaktifkan Serangan Arp Spoof Secara Penuh (Kurnia, 2019)

Proses ini akan membuat laptop perangkat penulis berada ditengah-tengah jalur trafik jaringan, dikarenakan arp spoof akan mengirimkan serangan berupa arp protokol dalam jumlah sangat banyak kepada target dan hal ini akan membuat perangkat target mengira pelaku sebagai seorang user dan otomatis data akan dikirimkan melalui computer pelaku, dengan menggunakan metode ini laptop peneliti akan berada di tengah-tengah skema jaringan dan otomatis arus paket yang dikirim oleh user ataupun yang diterima oleh user didalam jaringan yang sama di puskom dapat dilihat (M. Ferdy Adriant & Is Mardianto, 2015).

Arp spoof telah dijalankan, maka dilakukan proses menangkap trafik Internet Puskom dengan menggunakan perangkat lunak wireshark. Data trafik ini ditangkap dari jam 09:00 – 16:00 dari kurun waktu lima hari dan disimpan dalam bentuk .csv. Data trafik trafik internet yang dikumpulkan memiliki jumlah yang beragam, seperti yang terlihat pada tabel 1 jumlah data mentah trafik per hari.

Tabel 1. Jumlah Data Mentah trafik Per Hari

| Tanggal | Jumlah Data |
|---|---|
| 13-06-2022 | 18.085.931 |
| 14-06-2022 | 42.087.977 |
| 15-06-2022 | 28.606.473 |
| 16-06-2022 | 41.802.893 |
| 17-06-2022 | 47.988.765 |

*B. Generate Data Trafik Internet*

Pada tahap ini setiap data trafik internet yang sudah dikumpulkan akan dilakukan proses mengekstrak semua informasi arus paket dengan file berbentuk .csv dan pcap, kemudian dengan menggunakan berbagai library *python* seperti *vaex*, *nfstream*, *dask*, dan *pandas* serta *google collab notebook* dan penyimpanan *google drive*. Peneliti mengawalinya dengan mengupload data trafik internet kedalam google drive. Selanjutnya memanggil setiap data trafik berbentuk pcap dan menggenerate setiap data trafik internet untuk melakukan proses perhitungan terhadap data arus trafik secara statistika serta mengidentifikasi jenis trafik puskom dengan fungsi NFStreamer(). Kemudian nilai kembali akan disimpan kedalam format *vaex* dataframe. Proses ini akan menghasilkan sebuah data trafik yang memiliki atribut sebanyak 45 dan proses ini menghasilkan data baru pada masing-masing data trafik internet. Contoh data trafik internet pada hari senin dapat dilihat pada tabel 2 data hasil generate trafik internet.

Tabel 2. Data hasil generate trafik Internet

| | No | src_ip | : | Bidirectional_bytes | application_category_name |
|---|---|---|---|---|---|
| Senin | 0 | 192.168.0.133 | : | 1614 | Download |
| | 1 | 172.217.194.188 | : | 3156 | Web |
| | 2 | 77.43.132.126 | : | 384 | Download |
| | … | … | : | … | … |
| | *1,089,578* | 192.168.0.210 | | 132 | System |
| | *1,089,579* | 192.168.0.210 | | 204 | Web |

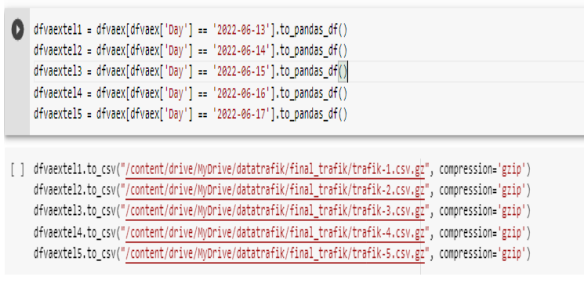
Data pada trafik internet yang sudah dilakukan proses generate tidak memiliki atribut waktu oleh karena itu diperlukannya proses penggabungan/join dengan data mentah trafik, proses ini dilakukan agar data trafik internet mendapatkan informasi mengenai kapan arus trafik itu terjadi. Proses join dilakukan dengan masing-masing kedua tabel terlebih dahulu membuat atribut baru yang bernama src_dst_address_&src_dst_port, yang dimana nilai pada atribut ini adalah hasil penggabungan string dengan 4 atribut yaitu *Source_Address*, *Destination_Address*, *Source_Port* dan *Destination_Port*, setelah kedua tabel membuat atribut tersebut maka proses join trafik internet dengan data mentah dilakukan.

Karena proses join kedua tabel mengakibatkan data menjadi berduplikat oleh karena itu proses menghapus data duplikat harus dilakukan, serta melakukan proses pembuatan atribut baru Time_Hours, Day, dan Day_Time dengan nilai yang didapatkan dari hasil proses join dengan data mentah trafik internet, kode dapat dilihat pada Gambar 6 membuat atribut time hourse dan menghapus data duplikat.

https://doi.org/10.51967/tanesa.v23i2.1736 © 2022 Politeknik Pertanian Negeri Samarinda


Gambar.6 Membuat Atribut Time_Hours dan Menghapus Data Duplikat

Kemudian masing-masing dataframe trafik internet dari tanggal 13-17 juni akan diexport dalam format hdf5, yang akan membuat pemrosesan data pada tahap selanjutnya menjadi lebih cepat dengan bantuan *library vaex*. Data yang sudah diubah dengan format hdf5 akan dipanggil kelima data tersebut dan digabungkan menjadi satu menjadi format vaex dataframe. Data kelima hari trafik internet memiliki jumlah data sebesar 14.409.876 data terlebih dahulu harus dilakukan pemeriksaan berdasarkan kategori data trafik yang tidak diketahui, hal ini dilakukan agar dapat mempermudah proses data mining yang akan dilakukan pada tahap selanjutya, dan data kategori trafik berjenis *System* sebanyak 12.543.120 harus dihapus dan Unspecified sebanyak 898.230. Karena data trafik yang berkategori System memiliki data aplikasi ARP yang dilakukan oleh peneliti pada saat melakukan serangan ditahap sebelumnya, dan tujuan dari penelitian ini adalah mengetahui jenis trafik internet yang pada puskom, maka data trafik berkategori System harus dihapus, proses ini menyisakan data trafik internet menjadi 1.086.046. Kemudian peneliti memisahkan data berdasarkan tanggalnya dan mengubah data vaex dataframe menjadi format pandas, lalu data diexport dalam bentuk .csv.gz. Kode dapat dilihat pada Gambar 7 memanggil data berdasarkan tanggalnya.



Gambar 7. Mengambil Data Berdasarkan Tanggalnya

## *C. Tahapan Data Mining*

1. Bussiness Understanding

Berdasarkan hasil generate data yang dilakukan dengan library NFStream pada tahap sebelumnya, menghasilkan data trafik internet sebanyak 1.086.046 dengan 45 atribut. Peneliti menggunakan data tersebut untuk menampilkan visualisasi mengenai data arus trafik internet Puskom yang ada selama 5 hari dari mulai jam 09:00 – 16:00, hal ini dilakukan agar dapat membantu manajemen internet di puskom untuk mengetahui jenis trafik internet yang padat pada jam tertentu dan membangun model klasifikasi dengan menggunakan K-Means yang dapat mengkategorikan arus paket trafik internet yaitu Download, Game, SocialNetwork, Web.

2. Data Understanding

Tahap data understanding dilakukan setelah menentukan tujuan dari penelitian, pada tahap ini peneliti menggunakan data trafik internet yang berasal dari hasil menangkap trafik data internet pada Puskom WiCiDa selama 5 hari dan sudah dilakukan proses labeling dengan menggunakan nfstream. Menghasilkan data dengan jumlah 1.086.046 dengan 45 atribut yang digunakan dalam eksplorasi data serta menggunakan fitur *Application_category_name* yang akan peneliti gunakan untuk memprediksi variabel target. Atribut dan fitur juga disebut sebagai variabel independen dan variabel target dapat disebut sebagai variabel dependen. Peneliti menggunakan variabel independen untuk memprediksi variabel dependen. Data trafik internet yang sudah diolah menggunakan NFStream mengandung informasi mengenai paket-paket data yang sudah dilakukan proses perhitungan statistika dan paket-paket data ini mendekripsikan jumlah paket data yang dikirim oleh user dalam mengakses internet. Adapun atribut dan kelas yang digunakan dalam proses penelitian ini, dapat dilihat pada tabel 3 deskripsi fitur trafik Internet.

Tabel 3. Deskripsi Fitur Trafik Internet

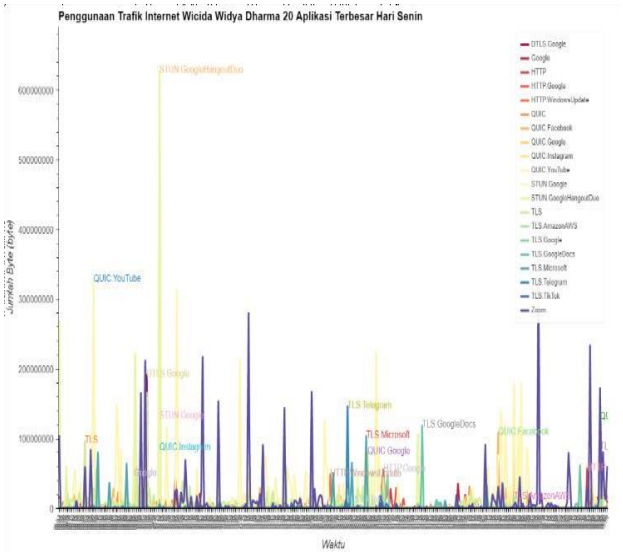
| Tipe Atribut | Deskripsi Atribut | Jumlah |
|---|---|---|
| Paket Src2Dst | Arus paket total, byte, mean,max, min dan standar deviasi yang dikirim dari alamat asal ke alamat tujuan | 6 |
| Paket Dst2Src | Arus paket total, byte, mean, max, min dan standar deviasi yang dikirim dari alamat tujuan ke alamat asal | 6 |
| Bidirectional Paket | Arus paket total, byte, mean, max,min dan standar deviasiyang dikirim dari dua arah | 6 |
| Paket Interval Src2Dst | Waktu antar paket min, mean, max dan standar deviasi yang dikirim dari alamat asal ke alamattujuan | 4 |
| Paket Interval Dst2Src | Waktu antar paket min, mean,max dan standar deviasi yang dikirim dari alamat tujuan kealamat asal | 4 |
| Bidirectional Interval | Waktu antar paket min, mean, max dan standar deviasi yang dikirim dari dua arah | 4 |
| Src Alamat | Alamat asal yang terdiri dari (ip, mac, oui, dan port) | 4 |
| Dst Alamat | Alamat tujuan yang terdiri dari(ip, mac, oui, dan port) | 4 |
| Protocol | Nomor idenfikasi transport layer protokol | 1 |
| Applicatio | Nama aplikasi | 1 |

Proses visualisasi data trafik dilakukan untuk mendapatkan gambaran mengenai jumlah pengiriman paket terbanyak berdasarkan aplikasinya, data yang

https://doi.org/10.51967/tanesa.v23i2.1736 

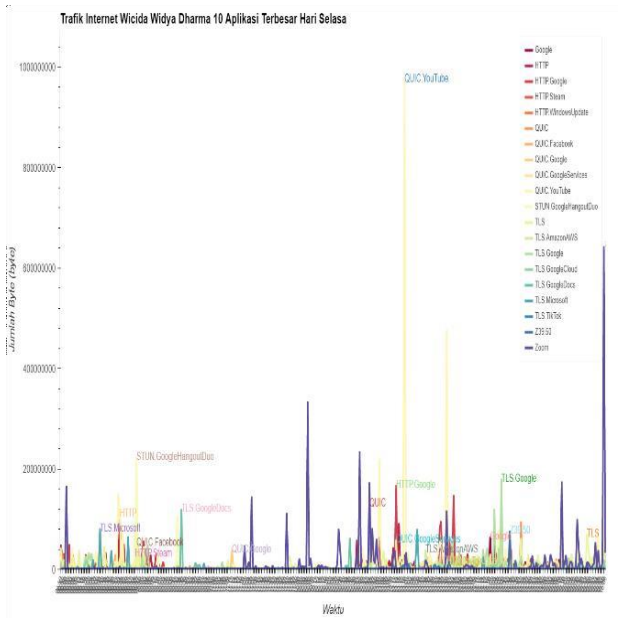
ditampilkan perharinya mulai dari tanggal 13-06-2022 s/d 17-06-2022.

Visualisasi data pada hari senin tanggal 13-06-2022 menampilkan aplikasi STUN.GoogleHangoutDuo sebagai pengirim paket data dengan ukuran byte terbesar yaitu 628.874.797 byte atau 628 mb, yang dilakukan pada pukul 10:13. Visualisasi dapat dilihat pada gambar 8 data trafik di ambil pada tanggal 13-06-2022.



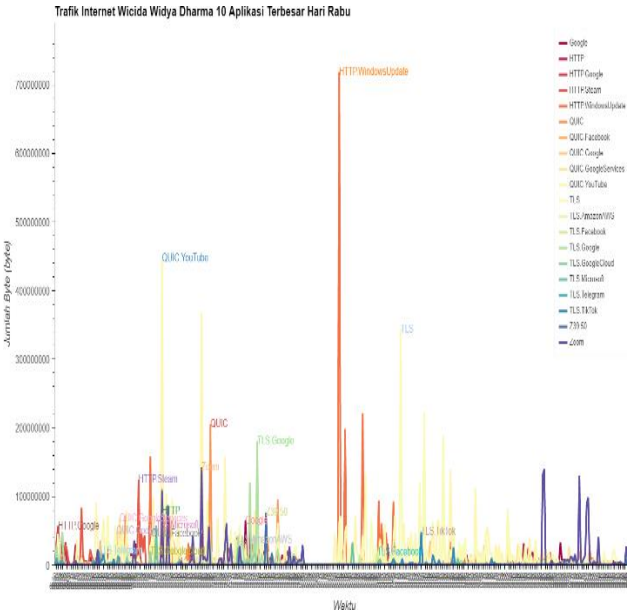

Gambar 8. Data Trafik Tanggal 13-06-2022

Visualisasi data pada hari selasa tanggal 14-06-2022. Menampilkan aplikasi Quic.Youtube menggunakan data terbesar diantara yang lain, yaitu sebesar 976.538.316 byte atau 976 mb yang dilakukan pada jam 13:19. Visualisasi dapat dilihat pada gambar 9 data trafik di ambil pada tanggal 14-06-2022.



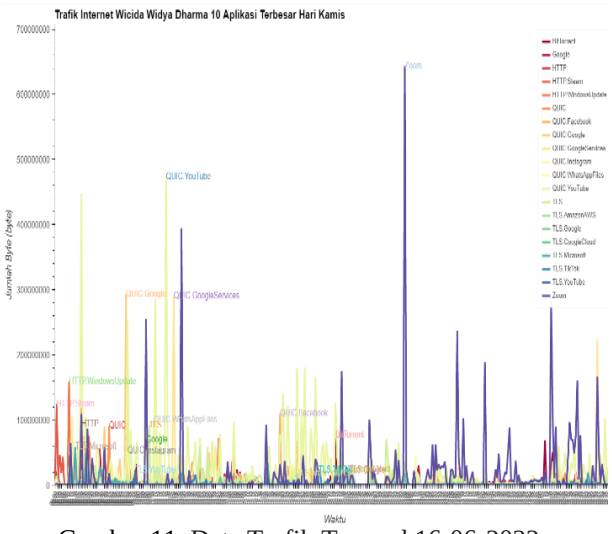

Gambar 9. Data Trafik Tanggal 14-06-2022

Visualisasi data pada hari rabu tanggal 15-06-2022. Menampilkan aplikasi HTTP.Windows Update menggunakan data terbesar diantara yang lain, yaitu sebesar 718.601.470 byte atau 718,6 mb yang dilakukan pada jam 12:32-12:38. Visualisasi dapat dilihat pada gambar 10 data trafik di ambil pada tanggal 15-06-2022.




Gambar 10. Data Trafik Tanggal 15-06-2022

Visualisasi data pada hari kamis tanggal 16-06-2022. Menampilkan aplikasi Zoom menggunakan data terbesar , yaitu sebesar 643.303.985 byte atau 643,3 mb yang dilakukan pada jam 12:59. Visualisasi dapat dilihat pada gambar 11 data trafik di ambil pada tanggal 16-06-2022.

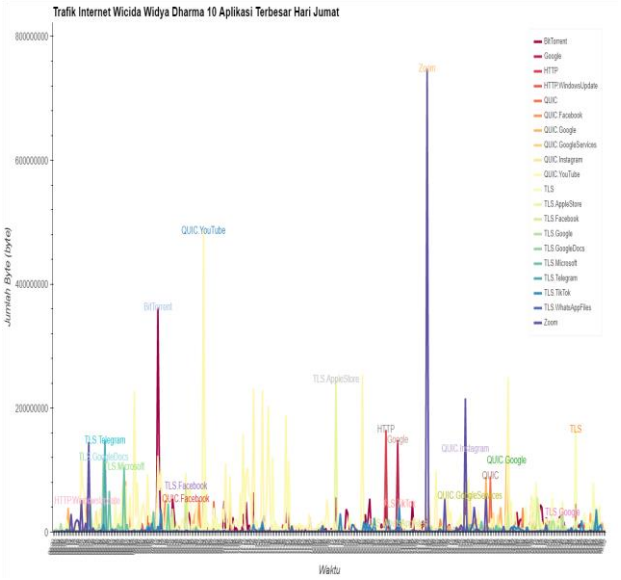



Gambar 11. Data Trafik Tanggal 16-06-2022

Visualisasi data pada hari Jumat tanggal 17-06-2022. Menampilkan aplikasi Zoom menggunakan paket data terbesar, yaitu sebesar 747.524.399 byte atau 747,5 mb yang dilakukan pada jam 13:55 dan penggunaan aplikasi zoom dilakukan mulai dari pukul 13:55 – 14:41, serta banyak aplikasi lainnya yang melakukan aktivitas dipagi

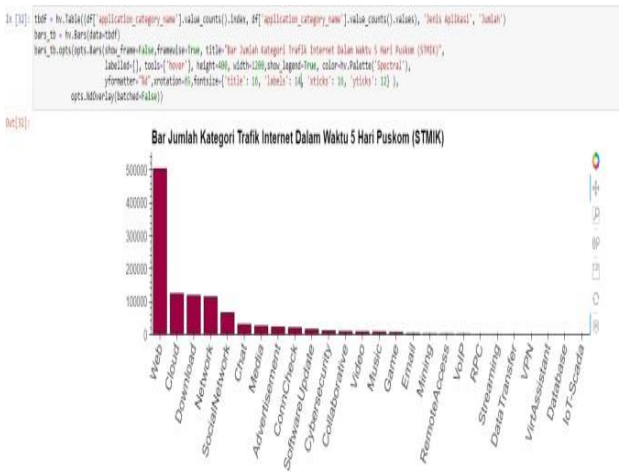
https://doi.org/10.51967/tanesa.v23i2.1736 © 2022 Politeknik Pertanian Negeri Samarinda


hari pada pukul 09:00 – 11:00. Visualisasi dapat dilihat pada gambar 12 data di ambil pada tanggal 17-06-2022.




Gambar12. Data Trafik Tanggal 17-06-2022

Data kategori trafik internet selama 5 hari memiliki 27 kategori, dari keseluruhan kategori, yang memiliki jumlah terbanyak adalah trafik internet dengan kategori Web yang berjumlah 503.131 dan jumlah terkecil dimiliki oleh kategori IoT-Scada dengan 17 data. Hal ini dapat disimpulkan bahwa penggunaan trafik internet di Puskom banyak menggunakan jenis trafik internet yang berjenis Web. Pada gambar 13 menunjukkan data kategori trafik Tanggal 13-06-2022 s/d 17-062022 dapat dilihat pada Gambar 13 hasil data kategori trafik.




Gambar 13. Data Kategori Trafik Tanggal 13-06-2022 s/d 17-062022

3. Data Preparation

Pada tahap ini, peneliti melakukan proses pemilihan atribut sebanyak 11 fitur arus trafik dua arah (bidirectional) dan pemilihan data sebanyak 6000 data yang mana jumlah data tersebut adalah hasil data trafik internet yang didistribusikan berdasarkan 4 kategori yaitu Web, Download, SocialNetwork dan Game dengan masing-masing kategori mendapatkan data 1500 serta dan jumlah distribusi data menjadi sama, data dapat dilihat pada tabel 4 distribusi data.

Tabel 4 Distribusi Data

| No | bidirectional_packets | bidirectional_bytes | : | bidirectional_stddev_piat_ms | bidirectional_max_piat_ms | application_category_name |
|---|---|---|---|---|---|---|
| 1 | 501 | 662364 | : | 383.819371 | 4869 | Web |
| 2 | 4 | 326 | : | 0.577350 | 1 | Web |
| 3 | 128 | 86030 | : | 25.72096 | 242 | Web |
| 4 | 67743 | 58382512 | : | 168.556705 | 42901 | Web |
| 5 | 3 | 1056 | : | 17.677670 | 25 | Web |
| … | … | … | … | … | … | … |
| 5995 | 22 | 9276 | : | 9360.205000 | 36484 | Game |
| 5996 | 4 | 384 | : | 2.309401 | 4 | Game |
| 5997 | 14 | 8068 | : | 66.340614 | 240 | Game |
| 5998 | 4 | 344 | : | 4.041452 | 7 | Game |
| 5999 | 4 | 264 | : | 23.094011 | 40 | Game |

4. Modeling

Pada tahapan ini dari data trafik jaringan yang sudah di kumpulkan selama 5 hari akan di olah berdasarkan data "*application_category_name*" yang dapat di lihat pada gambar 14 menjelaskan hasil data kategori trafik.

https://doi.org/10.51967/tanesa.v23i2.1736 © 2022 Politeknik Pertanian Negeri Samarinda


```
In [135]: df['application_category_name'].value_counts()
Out[135]: Web               503131
          Cloud             123483
          Download          118035
          Network           114068
          SocialNetwork      65593
          Chat               30053
          Media              25528
          Advertisement      22436
          ConnCheck          20309
          SoftwareUpdate     15389
          Cybersecurity      10675
          Collaborative       8872
          Video               7840
          Music               7432
          Game                6246
          Email               2310
          Mining              1433
          RemoteAccess        1254
          VoIP                1085
          RPC                  266
          Streaming            261
          DataTransfer         226
          VPN                   56
          VirtAssistant         25
          Database              23
          IoT-Scada             17
          Name: application_category_name, dtype: int64
```

Gambar 14. Data Kategori Trafik

Selanjutnya Data kategori trafik akan dipisah berdasar kan data latih sebanyak 1086046 data. Akan digunakan 4 fiture berdasarkan data :

| | | |
|---|---|---|
| application_category_name | 1086046 non-null | object |
| Time_H ours | 1086046 non-null | object |
| Day | 1086046 non-null | object |
| Day_Time | 1086046 non-null | object |

Yang bisa di lihat pada gambar 15 data di ambil dari 4 fiture yai tu : Web, VideoIP, dan Network.

```
In [ ]:
In [154]: df2=df.iloc[:,[40,42,43,44]]
In [155]: df2
Out[155]:
```

| | application_category_name | Time_Hours | Day | Day_Time |
|---|---|---|---|---|
| 0 | Web | 09:03 | 2022-06-13 | 2022-06-13 09:03:00 |
| 1 | Video | 09:03 | 2022-06-13 | 2022-06-13 09:03:00 |
| 2 | Web | 09:03 | 2022-06-13 | 2022-06-13 09:03:00 |
| 3 | Web | 09:03 | 2022-06-13 | 2022-06-13 09:03:00 |
| 4 | VoIP | 09:03 | 2022-06-13 | 2022-06-13 09:03:00 |
| ... | ... | ... | ... | ... |
| 1086041 | Network | 16:02 | 2022-06-17 | 2022-06-17 16:02:18 |
| 1086042 | Web | 16:02 | 2022-06-17 | 2022-06-17 16:02:18 |
| 1086043 | Network | 16:02 | 2022-06-17 | 2022-06-17 16:02:18 |
| 1086044 | Network | 16:02 | 2022-06-17 | 2022-06-17 16:02:18 |
| 1086045 | Network | 16:02 | 2022-06-17 | 2022-06-17 16:02:20 |

1086046 rows × 4 columns

```
In [156]: from sklearn.preprocessing import LabelEncoder
```

Gambar 15. Di Ambil 4 Fiture : Web, Video VoIP, Network

Kemudian peneliti melakukan proses normalisasi MinMax pada label dengan LabelEncoder. Hasil bisa dilihat pada Gambar.16 dan Gambar.17 Proses normalisasi MinMax dan Proses normalisasi MinMax pada label dengan LabelEncoder

```
In [166]: from sklearn.preprocessing import MinMaxScaler
          ms = MinMaxScaler()
          X2 = ms.fit_transform(X2)
In [167]: X2 = pd.DataFrame(X2, columns=[cols])
In [168]: X2
Out[168]:
```

| | application_category_name | Day_Time |
|---|---|---|
| 0 | 1.00 | 0.0 |
| 1 | 0.88 | 0.0 |
| 2 | 1.00 | 0.0 |
| 3 | 1.00 | 0.0 |
| 4 | 0.96 | 0.0 |
| ... | ... | ... |
| 1086041 | 0.60 | 1.0 |
| 1086042 | 1.00 | 1.0 |
| 1086043 | 0.60 | 1.0 |
| 1086044 | 0.60 | 1.0 |
| 1086045 | 0.60 | 1.0 |

1086046 rows × 2 columns

Gambar 16. Proses normalisasi MinMax

```
In [156]: from sklearn.preprocessing import LabelEncoder
          #merubah kolom application_name dan applocation_category_name mennjadi bentuk angka
          le = LabelEncoder()
          #df2['application_name'] = le.fit_transform(df2['application_name'])
          df2['application_category_name'] = le.fit_transform(df2['application_category_name'])
          df2
Out[156]:
```

| | application_category_name | Time_Hours | Day | Day_Time |
|---|---|---|---|---|
| 0 | 25 | 09:03 | 2022-06-13 | 2022-06-13 09:03:00 |
| 1 | 22 | 09:03 | 2022-06-13 | 2022-06-13 09:03:00 |
| 2 | 25 | 09:03 | 2022-06-13 | 2022-06-13 09:03:00 |
| 3 | 25 | 09:03 | 2022-06-13 | 2022-06-13 09:03:00 |
| 4 | 24 | 09:03 | 2022-06-13 | 2022-06-13 09:03:00 |
| ... | ... | ... | ... | ... |
| 1086041 | 15 | 16:02 | 2022-06-17 | 2022-06-17 16:02:18 |
| 1086042 | 25 | 16:02 | 2022-06-17 | 2022-06-17 16:02:18 |
| 1086043 | 15 | 16:02 | 2022-06-17 | 2022-06-17 16:02:18 |
| 1086044 | 15 | 16:02 | 2022-06-17 | 2022-06-17 16:02:18 |
| 1086045 | 15 | 16:02 | 2022-06-17 | 2022-06-17 16:02:20 |

1086046 rows × 4 columns

Gambar 17. Proses normalisasi MinMax pada label dengan LabelEncoder

5. Model Evaluasi K-Means

Peneliti menggunakan visual confusion matrix sebagai cara untuk menunjukkan dimana model membuat prediksi yang benar dan membuat prediksi yang salah. Hasil model K-Meansh dapat dilihat pada gambar 18 menentukan cluster dan gambar 19 hasil uji clustering.

```
In [78]: X2['Cluster_Id'] = kmeans.labels_
         X2.head()
Out[78]:
```

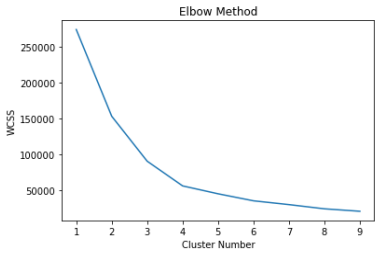
| | protocol | application_name | application_category_name | Day_Time | Cluster_Id |
|---|---|---|---|---|---|
| 0 | 0.0 | 0.339768 | 1.00 | 0.0 | 1 |
| 1 | 1.0 | 0.992278 | 0.88 | 0.0 | 0 |
| 2 | 0.0 | 0.683398 | 1.00 | 0.0 | 1 |
| 3 | 0.0 | 0.683398 | 1.00 | 0.0 | 1 |
| 4 | 1.0 | 0.656371 | 0.96 | 0.0 | 0 |

```
In [79]: kmeans = KMeans(n_clusters=3, random_state=0)
         kmeans.fit(X2)
Out[79]: KMeans(n_clusters=3, random_state=0)
```

Gambar 18. Menentukan Cluster

https://doi.org/10.51967/tanesa.v23i2.1736 

```
In [169]: from matplotlib import pyplot as plt
          from sklearn.cluster import KMeans
          wcss = []
          for i in range(1, 10):
              kmeans = KMeans(n_clusters = i, init = 'k-means++',  random_state = 0)
              kmeans.fit(X2)
              wcss.append(kmeans.inertia_)
          plt.plot(range(1, 10), wcss)
          plt.title('Elbow Method')
          plt.xlabel('Cluster Number')
          plt.ylabel('WCSS')
          plt.show()
```




Gambar 19. Hasil Uji Clustring = 3

## V. KESIMPULAN

Kesimpulan yang bisa diperoleh dari penelitian ini sudah menghasilkan sebuah model clustering trafik jaringan K-Means dengan banyak cluster = 3. Kategori trafik internet dihasilkan setelah diseleksi yaitu Web, Video VoIP, Network.

Dalam Clustering trafik internet dengan menggunakan arus/*flow* paket dua arah dan Data yang dinormalisasi dapat mempengaruhi hasil akurasi dari clustering. Pada saat pengujian dengan menggunakan 3 cluster menghasilkan nilai akurasi yang baik Mendapatkan hasil Clustring yaitu : Cluster 0 = 302638 data, Cluster 1 = 331982, dan Cluster 3 = 451426

## DAFTAR PUSTAKA

Agustiara, W., Pratama, A., Junaidi, S., PGRI Sumatera Barat Jl Gn Pangilun, S., Pangilun, G., Padang Utara, K., Padang, K., & Barat, S. (2022). Analisis Keamanan Protokol Secure Socket Layer Terhadap Serangan Packet Sniffing Pada Website Portal Berita Harian Umum Koran Padang. *Jurnal Teknik Informatika Kaputama (JTIK)*, *6*(1).

Amri, M. A., Windarto, A. P., Wanto, A., & Damanik, I. S. (2019). Analisis Metode K-Means Pada Pengelompokan Perguruan Tinggi Menurut Provinsi Berdasarkan Fasilitas Yang Dimiliki Desa. *KOMIK (Konferensi Nasional Teknologi Informasi Dan Komputer)*, *3*(1), 674–679. https://doi.org/10.30865/komik.v3i1.1677

Anggraeni, I., & Andriani, S. (2021). Implementasi Algoritma C.45 Untuk Klasifikasi Deteksi Serangan Pada Protokol Jaringan. *Komputasi: Jurnal Ilmiah Ilmu Komputer Dan Matematika*, *18*(2), 62–68. https://doi.org/10.33751/komputasi.v18i2.3562

Darma, S., Defit, S., Hartama, D., & Robiansyah, W. (2020). Penerapan Metode K-Means Dalam Pengolompokan Jumlah Wisatawan Asing Di Indonesia. *Prosiding Seminar Nasional Riset Dan Information Science (SENARIS) 2020*, *2*, 255–261.

Hartati, T., & Arie Wijaya, Y. (2022). Analisis Data Lalu Lintas Jaringan di Kantor Canghegar Cyber Operation Center Menggunakan Algoritma K-Means. *Jurnal Ilmiah NERO*, *7*(1), 2022.

Kurnia, D. (2019). Pemanfaatan Bettercap Sebagai Teknik Sniffing Pada Paket Trafik Jaringan Wifi. *Seminar Nasional Teknik (SEMNASTEK) UISU*, *2*(1), 83–85. www.olx.co

M. Ferdy Adriant, & Is Mardianto. (2015). Implementasi Wireshark Untuk Penyadapan (Sniffing) Paket Data Jaringan. *Seminar Nasional Cendekiawan*, 224–228.

Nagari, S. S., & Inayati, L. (2020). Implementation of Clustering Using K-Means Method To Determine Nutritional Status. *Jurnal Biometrika Dan Kependudukan*, *9*(1), 62. https://doi.org/10.20473/jbk.v9i1.2020.62-68

Prathivi, R. (2015). Klasifikasi Data Trafik Internet Menggunakan Metode Bayes Network (Studi Kasus Jaringan Internet Universitas Semarang). *Jurnal Transformatika*, *12*(2), 42. https://doi.org/10.26623/transformatika.v12i2.81

Premitasari, M. (2019). Volume Trafik IP-Based dengan Pemodelan Jam Sibuk. *MIND Journal*, *3*(1), 1–14. https://doi.org/10.26760/mindjournal.v3i1.1-14

Rizqi utami, A., Purwanto, Y., Anbarsanti, A. (2017). PENGELOMPOKAN TRAFIK BERDASARKAN WAKTU DENGAN ALGORITMA CLUSTREAM UNTUK DETEKSI ANOMALI PADA ALIRAN TRAFIK TIME BASED TRAFFIC CLUSTERING USING CLUSTREAM ALGORITHM FOR ANOMALY DETECTION ON STREAMING TRAFFIC. *E-Proceeding of Engineering*, *4*(1), 848–854.

Shafiq, M., Yu, X., Laghari, A. A., Yao, L., Karn, N. K., & Abdessamia, F. (2017). Network Traffic Classification techniques and comparative analysis using Machine Learning algorithms. *2016 2nd IEEE International Conference on Computer and Communications, ICCC 2016 - Proceedings*, *October*, 2451–2455. https://doi.org/10.1109/CompComm.2016.7925139

Tasmi, T., Maulana, R., & Husnawati, H. (2021). Visualisasi Trafik Jaringan Dengan Metode Support Vector Machine (SVM)(Studi Kasus: Universitas Indo Global Mandiri). *Jurnal Informatika Global*, *12*(2), 65–74. http://ejournal.uigm.ac.id/index.php/IG/article/view/1939

Yasriady, D. (2022). Klasterisasi Penggunaan Trafik Internet Menggunakan K-Mean Clustering. *Jurnal Sistim Informasi Dan Teknologi*, *4*(3), 112–117. https://doi.org/10.37034/jsisfotek.v4i3.141