

BAB III

METODOLOGI PENELITIAN

3.1 Studi Literatur

Penelitian ini dilakukan dengan melakukan penelitian kepustakaan, yang berarti mengumpulkan literatur yang relevan dengan topik penelitian, seperti buku, artikel, kertas, jurnal, makalah, dan beberapa situs internet yang relevan.

3.2 Pengumpulan Data

Proses metode ini dilakukan untuk mengolah data access log web yang didapat dari politeknik negeri lampung .

1. Pengumpulan Informasi

Daftar data server (log server) yang dipakai, yang mencakup pergi ke polinela.ac.id. sejauh satu bulan (Juli 2022). Preprocessing dimulai setelah mendapatkan data set.

2. Preprocessing

Preprocessing membersihkan data, mengidentifikasi pengguna, dan mengidentifikasi sesi adalah tugas preprocessing.

a. Pembersihan data

Tujuan dari tahap ini adalah untuk menghilangkan data yang tidak diperlukan dari server log. terkait dengan proses mining. Contoh data ini termasuk file multimedia (seperti format *.png, *.gif, dan *.pdf), metode permintaan HTTP selain GET, kode status atau permintaan status selain 200, 301, dan 304, jenis browser, authuser dan bytes.

b. Identifikasi pengguna

Identifikasi proses pengguna yang mengakses situs web. Ini dilakukan dengan memilih IP yang sama, sehingga pengguna dianggap sebagai satu pengguna.

c. Identifikasi sesi

Proses untuk mengelompokkan permintaan dan mengidentifikasi konten dari setiap pengguna yang memiliki akses ke situs web. Jika seorang pengguna mengakses halaman dengan kode status 200, 301, dan 302, pengguna tersebut dianggap sebagai pengguna.

3.2.1 Pengumpulan Alat dan Bahan Pendukung

1. Perangkat Keras (Hardware)

Perangkat keras yang digunakan dalam menganalisa data adalah sebagai berikut:

a. Processor Core- i7

Processor Core-i7 diperlukan agar penggunaan aplikasi implementasi dapat berjalan dengan lancar. Ini karena jenis processor ini memiliki kemampuan untuk menjalankan banyak proses dalam skala kecil.

b. SSD 128 GB

SSD 128 GB memiliki kapasitas yang besar, sehingga dapat digunakan sebagai referensi untuk aplikasi analisis data.

c. RAM 8 GB

RAM 8 GB cukup untuk memenuhi standar untuk menyimpan data set yang cukup besar untuk aplikasi analisis data dan berbagai aplikasi perhitungan, seperti Notebook Jupyter dan Microsoft Excel.

2. Program Perangkat Lunak (Software)

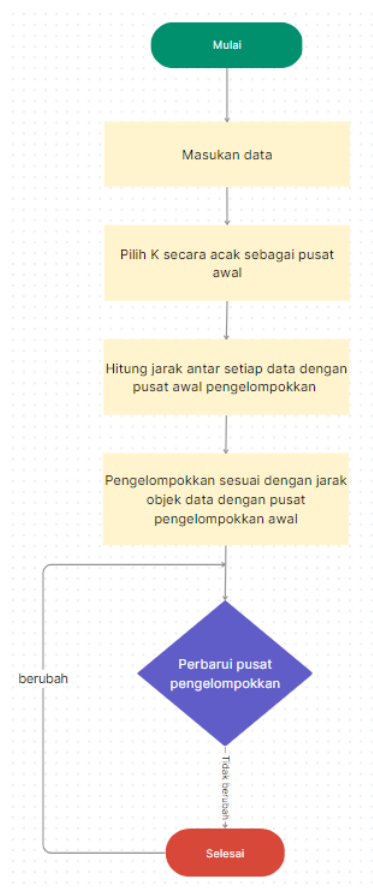
Sistem Operasi Microsoft Windows 10 64 bit yang digunakan dan Anaconda Navigator

3. Bahan

Bahan yang digunakan dalam penentuan ANALISIS ACCESS LOG WEB SERVER MENGGUNAKAN K-MEANS CLUSTERING UNTUK PROSES CYBER PROFILING (Studi Kasus Politeknik Negeri Lampung) yaitu penulis membutuhkan data-data log website yang dalam bentuk mentah dengan format txt. Data yang digunakan yaitu data 1 bulan terakhir yakni juli 2022 dengan berisi variabel yang berkaitan dengan judul penelitian penulis.

3.3 Metode Analisis Data

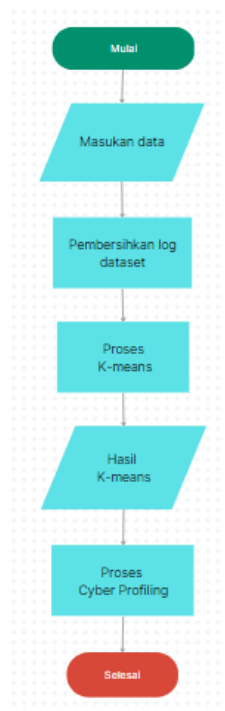
Peneliti akan dengan menggunakan teknik data mining untuk menganalisis informasi yang telah dkumpulkan. Metode clustering K-MEANS akan diterapkan untuk melakukan ini. Ini akan memprofilkan pengguna yang mengakses website pada data yang diperoleh, yang akan memungkinkan peneliti untuk mengidentifikasi fitur penelusuran web. Gambar 3.1 berikut menunjukkan bagaimana metode ini digunakan.



Gambar 3.1 Flowchart K-MEANS Clustering

3.4 Proses Pengujian

Fase pengujian saat itu, hasil dari proses data mining dihasilkan dengan metode K-MEANS digunakan untuk mengidentifikasi halaman akses yang paling sering diakses dan fitur pengguna. Tahap ini menentukan keberhasilan pengklasteran. Informasi tambahan dapat ditemukan dalam keterangan gambar, seperti yang ditunjukkan pada gambar 3.2:



Gambar 3.2 Metode Penelitian yang Diusulkan

Proses preprocessing akan dilakukan pada log data set yang dikumpulkan dari Politeknik Negeri Lampung. Proses ini bertujuan untuk menghilangkan informasi yang bukan dibutuhkan dari data. Setelah itu, semua halaman web yang sering dilihat pengguna akan ditampilkan dalam dataset yang telah dibersihkan. Dilakukan perhitungan untuk menghitung banyaknya halaman web yang sama, dan data yang dikumpulkan. Misalnya, data konten kunjungan web = {2,4,8,9,9,11,14,15,19,21,98,163,211,232,744,754,2510,2740} yang dikumpulkan dengan nama akan dibagi menjadi 3 cluster dengan $k = 3$, dan centroid pertama yang dipilih, dengan nilai $\mu_1=2$, $\mu_2=232$, dan $\mu_3=2740$. Centroid akhir ditentukan dengan menghitung nilai rata-rata dari "Kedekatan" diukur

menggunakan jarak geometris dan menghentikannya jika memenuhi persyaratan.

Hasil clustering sangat dipengaruhi oleh pemilihan centroid awal.

Hasil metode KMeans dengan klaster 3 dan kriteria untuk kelompok banyak, sedang, dan sedikit akan digunakan untuk mengevaluasi proses cyber profiling.

Cluster tersebut didasarkan pada jumlah kunjungan atau pengguna terhadap konten web yang ditunjukkan dalam data yang diproses oleh algoritma K-MEANS. Untuk membandingkan hasil algoritma K-MEANS, Cyber profiling juga memanfaatkan data yang diambil dari kategori konten web dan kategori waktu akses web yang dilakukan tanpa menggunakan algoritma K-MEANS:

- Kategori konten yang paling sedikit dikunjungi (Sedikit) untuk C1:

$$= \sqrt{(764 - 2)^2 + (365 - 0)^2 + (270 - 1)^2 + (399 - 0)^2 + (494 - 0)^2} = 609585$$

Formula Excel : =SQRT((764- 2)^2)+ (365- 0)^2)+ (270- 1)^2)+ (399- 0)^2)+ (494- 0)^2))

- Kategori konten yang sedang dikunjungi (Sedang) untuk C2:

$$= \sqrt{(764 - 232)^2 + (365 - 110)^2 + (270 - 78)^2 + (399 - 112)^2 + (494 - 154)^2} = 300390$$

Formula Excel : =SQRT((764- 232)^2)+ (365-11 0)^2)+ (270- 78)^2)+ (399- 112)^2)+ (494- 154)^2))

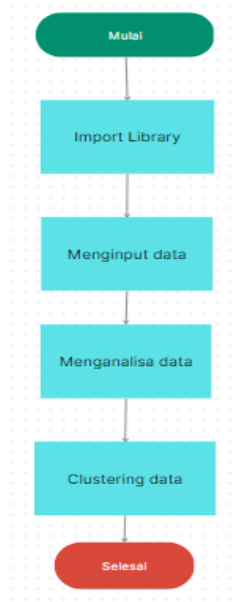
- Kategori konten yang paling terbanyak dikunjungi (Terbanyak) untuk C3:

$$= \sqrt{(764 - 2740)^2 + (365 - 1283)^2 + (270 - 891)^2 + (399 - 1457)^2 + (494 - 1849)^2} = 4185730$$

Formula Excel : =SQRT((764- 2740)^2)+ (365- 1283)^2)+ (270- 891)^2)+ (399- 1457)^2)+ (494- 1849)^2))

3.5 Desain Proses Jupyter Python

Gambar 3.3 Menunjukkan alur kerja untuk penggunaan clustering K-MEANS dalam program yang menggunakan python, bahasa pemrograman:



Gambar 3.3 Flowchart Clustering K-MEANS

Penentuan jumlah kunjungan konten

Menunjukkan proses pengklasteran menggunakan metode algoritma K-MEANS dalam programing dengan Python. Proses ini ada beberapa hal yang dapat dijelaskan sebagai berikut:

- 1) Mengimport beberapa perpustakaan (library) atau paket untuk mendukung pengolahan program pengklasteran K-MEANS
- 2) Gunakan perintah atau source code Python untuk mengklasterkan dataset yang ingin dimasukkan.
- 3) Pada tahap evaluasi, ada banyak langkah untuk mengevaluasi kumpulan data ini, seperti membaca dan memahami data, membersihkan data, dan pengobatan anomali.
- 4) Setelah fase analisis, dataset diklasterkan. Proses pengklasteran terdiri dari beberapa langkah, seperti transformasi frame data diubah menjadi array, melakukan scalling data, menghitung jumlah cluster, dan melihat nilai pusat dari setiap cluster. Selanjutnya, kolom yang menghasilkan pengklasteran dapat dimasukkan ke dalam data frame.

3.6 Panduan untuk Proses Cyber Profiling

Skenario ini dibahas tentang proses cyber profiling yang dilakukan dengan memanfaatkan log web yang dikumpulkan dari Politeknik Negeri Lampung. Data yang telah dikumpulkan akan dibersihkan dengan menggunakan Microsoft Excel, dan digunakan untuk clustering dengan menggunakan Python, aplikasi Jupiter Notebook. Langkah-Langkah dalam menentukan kmeans untuk cyber profiling:

- a) Data mentah.txt dipindah ke Microsoft Excel untuk dibersihkan, yang dirancang untuk menghilangkan data yang tidak digunakan dalam proses clustering, dalam data access log yang tidak digunakan yaitu . Setelah pembuangan data, data yang dihasilkan menjadi 7574 log yang akan digunakan pada proses clustering.
- b) Jupyter Notebook adalah aplikasi web sumber terbuka yang memungkinkan anda membuat dan berbagi dokumen dengan teks naratif, persamaan, visualisasi, dan kode langsung. Pembersihan dan transformasi data, simulasi numerik, pemodelan statistik, visualisasi data, dan pembelajaran mesin dapat digunakan untuk membantu dalam proses pengelompokan data kategori yang telah diproses Microsoft Excel.