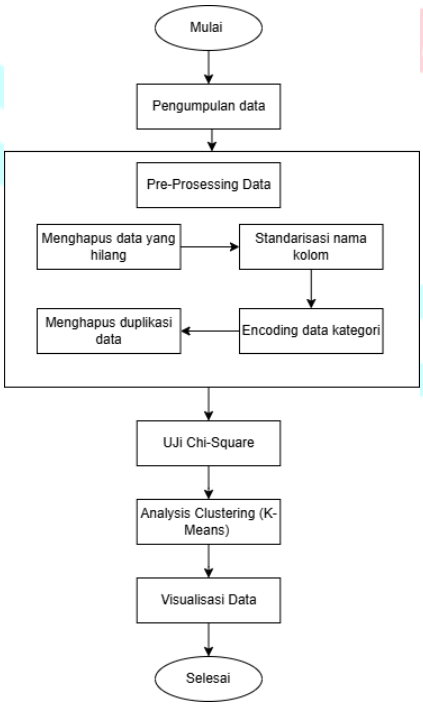
BAB III

METODE PENELITIAN

3.1 Objek Penelitian

Objek penelitian disini adalah data tayangan dari platform streaming Netflix, yang mencakup informasi genre, negara asal produksi, dan kategori rating konten. Penelitian ini berfokus pada klasifikasi genre dan rating konten lebih teknis dengan pendekatan Chi-Square dan K-Means. Lokasi penelitian dilakukan secara virtual menggunakan data yang tersedia dari sumber daring (Kaggle), sehingga tidak terkait pada lokasi fisik tertentu. Penelitian ini mencakup tahapan: pengumpulan data, praproses data (yang meliputi penghapusan data yang hilang, penghapusan data duplikat, standarisasi nama kolom, dan encoding data kategori), pengujian Chi- Square, analisis clustering dengan metode K-Means, serta visualisasi data sebagai tahap akhir. Lingkup penelitian dibatasi pada analisis tayangan yang tersedia di katalog Netflix hingga tahun tertentu sesuai metadata yang tersedia untuk memastikam konsistensi data. Dengan demikian, Penelitian ini memberikan fokus yang terukur pada pola preferensi genre tayangan dalam konteks negara asal dan rating konten.

3.2 Tahapan Penelitian



Gambar 1. Flowchart Tahapan Penelitian

Penjelasan Tahapan penelitian

1. **Pengumpulan Data**

   Data yang digunakan berasal dari website Kaggle yang dapat dilihat di link berikut Netflix popular movies dataset | Kaggle, Dataset Netflix ini memiliki 4.898 data film dengan jumlah kolom atribut yang dimiliki ada 8 atribut yaitu show_id, type, title, country, duration, genre, rating, dan listed_in. Ditahap ini juga, peneliti memahami dan mengecek data yang digunakan apakah terdapat missing values atau data yang tidak konsisten. Untuk penelitian ini, atribut yang akan digunakan ada 2 atribut yaitu listed_in dan rating.

2. **Pre-Prosessing Data**

   Dari 6.901 jumlah dataset, banyak data yang tidak konsisten dan memiliki missingvalue. Data yang missing value tersebut dibersihkan dari dataset, kemudian data yang sudah dibersihkan akan diproses menggunakan metode K-Means Clustering dataset juga dilakukan normalisasi. Untuk penelitian ini, atribut yang digunakan pada model K-Means ada 2 artibut, yaitu genre dan rating konten. Berikut tampilan gambar dataset yang sudah dibersihkan dari missing value dengan cara menghapus baris nilai yang kosong dari setiap atribut, sehingga data yang akan digunakan sebanyak 4.367 data.

	rating	listed_in
1	PG-13	Documentaries
2	TV-MA	International TV Shows, TV Dramas, TV Mysteries
3	TV-MA	International TV Shows, Romantic TV Shows, TV Comedies
4	TV-MA	Dramas, Independent Movies, International Movies
5	TV-14	British TV Shows, Reality TV
6	PG-13	Comedies, Dramas
7	TV-MA	Dramas, International Movies
8	TV-MA	TV Comedies, TV Dramas
9	TV-MA	Crime TV Shows, Spanish-Language TV Shows, TV Dramas
10	TV-14	International TV Shows, TV Action & Adventure, TV Dramas
11	TV-14	Comedies, International Movies, Romantic Movies
12	TV-14	Docuseries, International TV Shows, Reality TV
13	PG-13	Comedies
14	PG-13	Horror Movies, Sci-Fi & Fantasy
15	PG-13	Thrillers
16	TV-MA	British TV Shows, International TV Shows, TV Comedies

Gambar 2. data yang sudah diolah

Disini, Pre-prosessing data yang dilakukan yaitu Mengatasi missing values

$$D' = D - M$$

Dimana:

D adalah dataset asli

M adalah jumlah baris yang memiliki missing values

D′ adalah dataset setelah menghapus missing values

Selanjutnya standarisasi nama kolom :

C_{\text{new}} = \text{lowercase}(\text{replace}(C, \text{" "}, \text{"_" })) Encoding data kategori dengan one-hot encoding

Content Rating={G,PG,PG−13,R,TV−MA} penulis membuat vector biner

$$x_i = \begin{cases} 1, & \text{jika } x \text{ termasuk kategori tersebut} \\ 0, & \text{jika tidak} \end{cases}$$

Contoh encoding untuk kategori PG-13:

[0,0,1,0,0] (vector untuk PG-13)

Penghapusan duplikasi data

$$r_i = r_j \Rightarrow \text{hapus } r_j$$

3. **Uji Chi-Square**

   Uji Chi-Square digunakan untuk menguji hubungan antara genre dan rating konten. Pengolahan dilakukan dengan langkah berikut:

   1. Membuat tabel kontingensi yang menunjukkan distribusi genre berdasarkan kategori rating.
   2. Melakukan perhitungan statistik Chi-Square untuk menentukan apakah terdapat hubungan signifikan antara genre dan rating konten.
   3. Jika nilai p-value yang dihasilkan lebih kecil dari 0.05, maka hubungan tersebut dianggap signifikan.

4. **Analisis Clustering dengan K-Means**

   Setelah memahami hubungan antar variabel, dilakukan clustering analysis untuk mengelompokkan genre berdasarkan pola distribusi rating kontennya. Langkah-langkahnya:

   1. Standarisasi data menggunakan StandardScaler agar semua variabel berada dalam skala yang sama.
   2. Menentukan jumlah cluster optimal menggunakan metode Elbow Method dengan melihat titik optimal pada grafik inertia.

3. Menggunakan K-Means Clustering untuk mengelompokkan genre berdasarkan kemiripan pola distribusi rating konten.

5. Visualisasi Data

Hasil analisis disajikan dalam bentuk grafik agar lebih mudah dipahami:

1. Heatmap digunakan untuk menampilkan distribusi hubungan antara genre dan rating konten.
2. Elbow Method Plot digunakan untuk menentukan jumlah optimal cluster.

3.3 Alat dan Bahan

a) Perangkat Lunak : Google Colab, Python

b) Perangkat Keras : Laptop dengan spesifikasi minimal professor intel i5 dan RAM 8GB

c) Dataset : Dataset Netflix dengan atribut yang relevan untuk penelitian.