Pemilihan k Optimum *Clustering K-Means* pada Top 100 Film Berbahasa Korea di IMDb

Sakaoni Rofi Pramesthi, Haiva Qurrota A'yun, dan Adatul Mukarromah
Departemen Statistika, Fakultas Sains dan Analitika Data (FSAD), Institut Teknologi Sepuluh Nopember (ITS)
Jl. Arief Rahman Hakim, Surabaya 60111 Indonesia
e-mail: adatulm@gmail.com

***Abstrak*—Menonton film merupakan kegiatan yang sering dilakukan oleh kebanyakan orang untuk mengisi waktu luang ataupun sebuah hobi. Dewasa ini, banyak sekali situs yang menyediakan judul film dengan *rating* dan juga informasi lain dari pengunjung maupun dari film itu sendiri dan salah satu situs tersebut merupakan situs IMDb. Berdasarkan hal tersebut akan dilakukan *clustering* data menggunakan metode *K-Means* dengan variabel numerik *Runtime, Rating,* dan *Votes*. Untuk menentukan banyaknya *cluster* yang optimal digunakan metode *Elbow,* metode *Silhouette,* metode *Gap Statistic* berdasarkan metode-metode tersebut akan digunakan metode *Elbow* untuk penelitian lebih lanjut hal tersebut dikarenakan nilai k yang optimum yang dirasa baik adalah pada saat menggunakan metode tersebut.**

Kata Kunci*— *Cluster, Data, Film, IMDb, K-Means

I. PENDAHULUAN

A. Latar Belakang

Menonton film merupakan kegiatan yang sering dilakukan oleh kebanyakan orang untuk mengisi waktu luang ataupun sebuah hobi. Akhir-akhir ini film berbahasa korea menjadi pilihan kesukaan para penggemar film. Film berbahasa korea memiliki genre dan alur cerita yang menarik dengan cerita asli film itu sendiri maupun film adaptasi dari novel maupun komik atau yang biasa disebut dengan *webtoon*. Menonton film berbahasa korea juga memiliki manfaat salah satunya adalah belajar bahasa korea dan juga budaya yang ada. Seperti yang kita ketahui, setiap negara memiliki budayanya masing-masing dan hal tersebut cukup baik untuk kita mengetahui budaya dari negara selain negara kita sendiri. Selain bahasa dan budaya, kita juga bisa mempelajari karakter berdasarkan karakter pemain film tersebut, meskipun demikian, kita juga perlu selektif dan cermat dalam memahami karakter tersebut. Keberhasilan dari sebuah film yang diproduksi biasanya menggunakan *rating*, *votes*, dan seberapa banyak film tersebut dibahas di media sosial. Saat ini, banyak sekali situs yang menyediakan judul film dengan *rating* dan juga informasi lain dari pengunjung maupun dari film itu sendiri dan salah satu situs tersebut merupakan situs IMDb. Pada kali ini kami akan mencoba meng*cluster*kan film dengan *K-Means Clustering* pada situs IMDb dengan 3 metode untuk penentuan *cluster* optimum dan memilih metode yang tepat untuk menentukan jumlah *cluster* terbaik yang akan digunakan untuk penelitian lebih lanjut.

II. TINJAUAN PUSTAKA

A. Film

Film merupakan serangkaian gambar diam yang ketika ditampilkan pada layar akan menciptakan ilusi gambar bergerak [1]. Ilusi dari rangkaian gambar tersebut menghasilkan gerakan kontinu berupa video. Film sering disebut sebagai *movie* atau *moving picture*. Film merupakan salah satu bentuk seni modern dan popular yang dibuat untuk kepentingan bisnis dan hiburan.

B. IMDb

IMDb adalah sebuah basis data daring informasi yang berkaitan dengan film, acara televisi, video rumahan, permainan video, acara internet, termasuk juga daftar pemeran, biografi kru produksi dan personil, ringkasan alur cerita, trivia, dan ulasan serta penilaian oleh penggemar [2]. Situs web ini sekarang dimiliki oleh Amazon.com. Koleksi informasi film yang ditampilkan cukup lengkap dan dapat juga dilihat informasi film lama maupun film baru yang akan rilis di bioskop. Dalam IMDb terdapat komunitas yang dapat berkontribusi langsung untuk menuangkan ulasan film dan memberikan *rating* pada film yang diinginkan. Tidak hanya kaum awam, para pakar juga memiliki wadah sendiri untuk memberi *rating* dan menuangkan ulasan secara profesional pada film tersebut.

C. Statistika Deskriptif

Statistika deskriptif merupakan metode-metode yang berkaitan dengan pengumpulan dan penyajian suatu gugus data sehingga memberikan informasi yang berguna [3]. Statistika deskriptif adalah metode yang sangat sederhana. Metode ini hanya mendeskripsikan kondisi dari data yang sudah dimiliki dan menyajikannya dalam bentuk tabel dan juga bentuk lainnya yang disajikan dalam uraian singkat dan terbatas.

1) Mean

Rata-rata merupakan suatu ukuran pusat data bila data itu diurutkan dari yang terkecil sampai yang terbesar atau sebaliknya [4].

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1)$$

Keterangan :
x_i = data ke i
n = jumlah data

2) Median

Median dari variabel kuantitatif adalah nilai dari variabel pada suatu set data yang membagi variabel dari data tersebut menjadi dua, jadi setengah dari nilai yang didapat tersebut adalah lebih kecil atau sama dengan nilai median dan

setengah dari nilai yang didapat lebih besar atau sama dengan nilai median [5].

- Jika jumlah data adalah ganjil

$$Me = nilai\ data\ ke\frac{1}{2}\ (n+1) \quad (2)$$

- Jika jumlah data adalah genap

$$Me = \frac{1}{2}(nilai\ data\ ke\frac{n}{2} + nilai\ data\ ke(\frac{n}{2}+1)) \quad (3)$$

3) Maksimum dan Minimum

Nilai maksimal adalah nilai terbesar dari suatu data. Nilai minimum adalah nilai terkecil dari suatu data. Nilai maksimum dan minimum juga dapat digunakan untuk menghitung *range*, yaitu dengan cara nilai maksimum dikurangkan nilai minimum [4].

D. Web Scraping and Crawling

Web Scraping adalah proses pengambilan sebuah dokumen semi-terstruktur dari internet, umumnya berupa halaman-halaman web dalam bahasa markup seperti HTML atau XHTML, dan menganalisis dokumen tersebut untuk diambil data tertentu dari halaman tersebut untuk digunakan lagi kepentingan lain [6]. Lebih ringkas lagi *Web Scraping* merupakan teknik untuk mengumpulkan data dari website melalui proses ekstraksi informasi menggunakan *Hypertext Transfer Protocol* (HTTP). Untuk penerapannya dapat dilakukan secara manual maupun secara *automation program*. *Web Crawling* merupakan Teknik mengumpulkan data pada sebuah website dengan memasukkan *Uniform Resource Locator* (URL). URL ini menjadi acuan untuk mencari semua hyperlink yang ada pada website. Kemudian dilakukan indexing untuk mencari kata dalam dokumen pada setiap link yang ada.

E. Clustering

Analisis Pengelompokan/*Clustering* merupakan proses membagi data dalam suatu himpunan ke dalam beberapa kelompok yang kesamaan datanya dalam suatu kelompok lebih besar daripada kesamaan data tersebut dengan data dalam kelompok lain [7]. Algoritma Clustering terdiri dari dua bagian yaitu secara hirarkis dan secara partitional. Algoritma hirarkis menemukan cluster secara berurutan dimana cluster ditetapkan sebelumnya, sedangkan algoritma partitional menentukan semua kelompok pada waktu tertentu [8]. Clustering juga bisa dikatakan suatu proses dimana mengelompokkan dan membagi pola data menjadi beberapa jumlah data set sehingga akan membentuk pola yang serupa dan dikelompokkan pada cluster yang sama dan memisahkan diri dengan membentuk pola yang berbeda di cluster yang berbeda [9]. Clustering dapat ditemukan di beberapa aplikasi yang ada di berbagai bidang. Sebagai contoh pengelompokan data yang digunakan untuk menganalisa data statistik seperti pengelompokan untuk menganalisa data statistik seperti data mining, pengenalan pola, dan lain-lain.

F. K-Means

K-means merupakan suatu algoritma dalam pengelompokan secara partisi yang memisahkan data ke dalam kelompok yang berbeda-beda. Algoritma ini mampu meminimalkan jarak antara data ke *cluster*nya [10]. Kemudian algoritma *K-means* akan menguji masing-masing dari setiap komponen dalam populasi data tersebut dan menandai komponen tersebut ke dalam salah satu pusat *cluster* yang telah didefinisikan sebelumnya tergantung dari jarak minimum antar komponen dengan tiap-tiap pusat *cluster*. Selanjutnya posisi pusat *cluster* akan dihitung kembali sampai semua komponen data digolongkan ke dalam tiap-tiap *cluster* dan terakhir akan terbentuk *cluster* baru [11]. Berikut merupakan langkah yang dilakukan.

1. Menentukan k sebagai jumlah *cluster* yang akan dibentuk.
2. Menentukan k *centroid* (titik pusat *cluster*) secara random/acak.

$$v = \frac{\sum_{i=1}^{n} x_i}{n} \quad (4)$$

3. Menghitung jarak setiap objek ke masing-masing *centroid* dari masing-masing *cluster*. Untuk menghitung jarak antara objek dengan centroid dapan digunakan *Euclidian Distance*.

$$d(x,y) = \|x-y\| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (5)$$

4. Mengalokasikan masing-masing objek ke dalam *centroid* yang paling dekat.
5. Melakukan iterasi, kemudian menentukan posisi *centroid* baru dengan menggunakan persamaan (4).
6. Mengulangi langkah 3 jika posisi *centroid* baru tidak sama.

G. Metode Elbow

Metode Elbow merupakan suatu metode yang digunakan untuk menghasilkan informasi dalam menentukan jumlah cluster terbaik dengan cara melihat persentase hasil perbandingan antara jumlah cluster yang akan membentuk siku pada suatu titik [8]. Metode ini memberikan ide/gagasan dengan cara memilih nilai cluster dan kemudian menambah nilai cluster tersebut untuk dijadikan model data dalam penentuan cluster terbaik dan selain itu persentase perhitungan yang dihasilkan menjadi pembanding antara jumlah cluster yang ditambah. Hasil persentase yang berbeda dari setiap nilai cluster dapat ditunjukkan dengan menggunakan grafik sebagai informasinya. Jika nilai cluster pertama dengan nilai cluster kedua memberikan sudut dalam grafik atau nilainya mengalami penurunan paling besar dan membentuk siku maka nilai cluster tersebut yang terbaik [12]. Berikut rumus SEE pada *K-Means*.

$$SSE = \sum_{K=1}^{K} \sum_{x_i \in S_K} \|x_i - C_K\|_2^2 \quad (6)$$

H. Metode Silhouette

Pendekatan dari metode *silhouette* digunakan untuk melihat kualitas dan kekuatan *cluster*, seberapa baik suatu objek ditempatkan dalam suatu *cluster*. Lebar siluet rata-rata yang tinggi menunjukkan pengelompokkan yang baik. Metode ini menghitung siluet rata-rata pengamatan untuk nilai k yang berbeda. Jumlah optimal *cluster* k adalah cluster yang memaksimalkan siluet rata-rata.

I. Metode Gap Statistic

Pendekatan ini dapat diterapkan ke metode pengelompokan apapun termasuk pengelompokan *K-Means*.

Gap Satistic membandingkan total variasi *intracluster* untuk nilai k yang berbeda dengan nilai yang diharapkan di bawah distribusi referensi nol dari data (yaitu distribusi tanpa pengelompokan yang jelas). Dataset referensi dihasilkan menggunakan simulasi Monte Carlo dari proses pengambilan sampel yang artinya setiap variabel (x_i) dalam kumpulan data dihitung jangkauannya dan menghasilkan nilai untuk n poin secara seragam dari interval minimum hingga maksimum.

Untuk data observasi dan data referensi, total variasi *intracluster* dihitung menggunakan nilai k yang berbeda. *Gap statistic* untuk k tertentu didefinisikan sebagai berikut.

$$Gap_n(k) = E_n^* \log(W_k) - \log(W_k) \qquad (7)$$

Dimana E_n^* menunjukkan ekspektasi di bawah ukuran sampel n dari distribusi referensi. E_n^* didefinisikan melalui *Bootstrap* (B) dengan menghasilkan Salinan B dari kumpulan data referensi dan dengan menghitung rata-rata log (W_k^*). *Gap statistic* mengukur deviasi nilai W_k yang diamati dari nilai yang diharapkan di bawah hipotesis nol. Estimasi *cluster* optimal akan menjadi nilai yang memaksimalkan $Gap_n(k)$. ini berarti bahwa struktur pengelompokan jauh dari distribusi titik yang seragam.

III. METODOLOGI

A. *Sumber Data*

Data yang digunakan dalam tulisan ini merupakan 100 film terbaik berbahasa korea data dari web IMDb (https://www.imdb.com/search/title/?title_type=feature&languages=ko&count=100) yang diambil menggunakan *web scrapping* dan *crawling* pada tanggal 4 Januari 2021 pukul 20:00. Variabel yang digunakan dalam data ini yakni:

1) Judul Film
2) Genre
3) *Runtime*
4) *Rating*
5) *Votes*

B. *Metode Pengelompokan Data*

Metode Pengelompokan yang akan digunakan pada tulisan ini merupakan metode *K-Means Clustering* pada variabel numerik dari data yang akan digunakan yaitu runtime, *rating*, dan *votes*.

C. *Langkah Analisis*

Berikut merupakan langkah-langkah analisis yang dilakukan.

1. Menentukan data yang ingin diolah.
2. Melakukan *web Crawling and Scraping* untuk mendapatkan data dengan variabel yang diinginkan.
3. Mengetahui karakteristik dari data yang diperoleh.
4. Menentukan variabel numerik yang akan dilakukan *clustering*.
5. Mengetahui kesamaan setiap pasangan sampel.
6. Menentukan jumlah *cluster* k yang akan digunakan dengan bantuan Rstudio.
7. Melakukan *clustering* sebanyak k menggunakan bantuan Rstudio dan juga dilakukan visualisasi hasil *clustering*.
8. Menggabungkan data hasil *clustering* dengan data set yang ada.

IV. HASIL DAN PEMBAHASAN

A. *Web Crawling and Scraping*

Sebelum melakukan *web crawling and scraping* perlu dilakukan *install packages* xml2 dan rvest, selanjutnya menggunakan xml2 dan rvest sebagai *library* pada *syntax*. Setelah itu dilakukan pengambilan data dengan variabel Judul Film, *Genre*, *Runtime*, *Rating*, dan *Votes* dari web (https://www.imdb.com/search/title/?title_type=feature&languages=ko&count=100) menggunakan bantuan Rstudio. Setelah data sudah terambil, data akan diubah menjadi data frame dan disimpan ke dalam bentuk csv. Dengan demikian proses pengambilan data telah selesai dan data sudah diambil dalam bentuk csv dan juga tersimpan dalam dokumen pada perangkat komputer.

B. *Karakteristik Data*

Untuk mengetahui karakteristik data dari data yang sudah diperoleh dan tersimpan dapat digunakan analisa statistika deskriptif sebagai berikut.

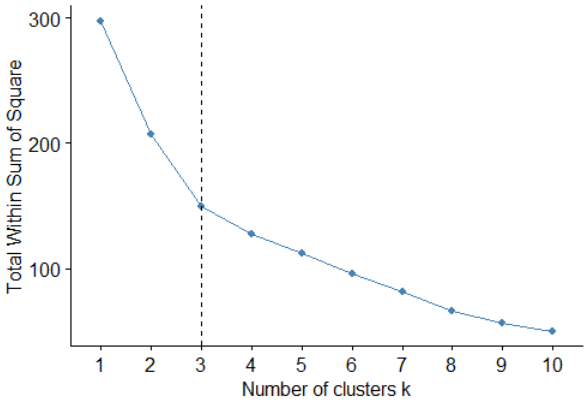
Tabel 1. Statistika deskriptif data

Variabel	Mean	Median	Minimum	Maksimum
Runtime	115,5	116	75	172
Rating	6,868	7	3,1	8,6
Votes	135,036	56,76	1,022	970

Berdasarkan tabel statistika deskriptif dari data tersebut dapat dilihat bahwa untuk setiap variabel numerik memiliki *mean*, median, nilai minimum dan maksimum seperti yang tertera pada tabel. Untuk variabel genre berupa data kategorik dengan 6 kategori genre yaitu *action* sebanyak 52, *Biography* sebanyak 2, *Comedy* sebanyak 11, *Crime* sebanyak 12, *Drama* sebanyak 17, dan *Horror* sebanyak 6.

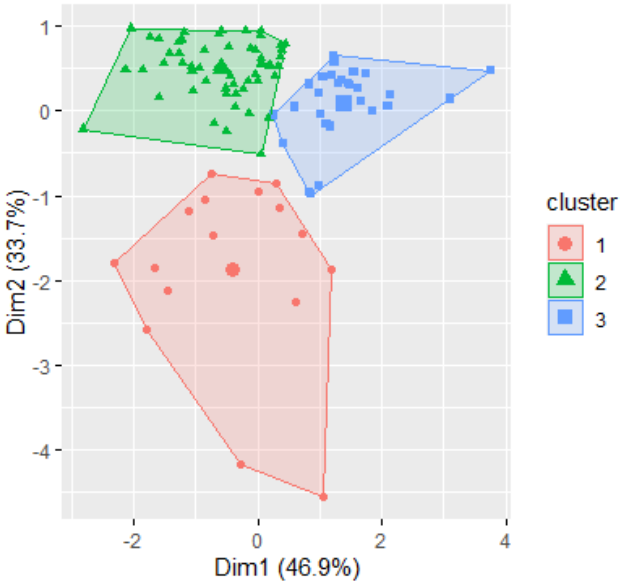
C. *K-Means Clustering dengan Metode Elbow*

Dari pengolahan data yang dilakukan menggunakan bantuan aplikasi Rstudio didapatkan *output* visualisasi penentuan k dengan metode *Elbow* sebagai berikut.



Gambar 1. Nilai k Optimal Metode *elbow*

Berdasarkan gambar tersebut dapat diartikan bahwa nilai k yang optimal untuk pengolahan 100 data dengan variabel numerik *Runtime*, *Rating*, dan *Votes* menggunakan metode *Elbow* adalah sebanyak 3 *cluster* hal tersebut dikarenakan pada saat k sebesar 3 grafik seperti membentuk siku. Maka akan dilakukan *Clustering* sebanyak 3 *cluster*. Dengan bantuan Rstudio didapatkan hasil *clustering* sebagai berikut.

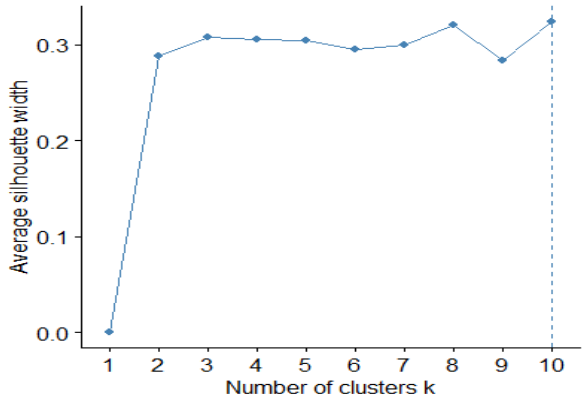


Gambar 2. 3 *Cluster Plot* Data

Berdasarkan gambar tersebut dapat diartikan bahwa 100 data film berbahasa korea sudah terbagi menjadi 3 *cluster* berdasarkan variabel numerik *Runtime*, *Rating*, dan *Votes*. Didapatkan terdapat sebanyak 28 film di *cluster* 1, 56 film di *cluster* 2, dan 16 film di *cluster* 3.

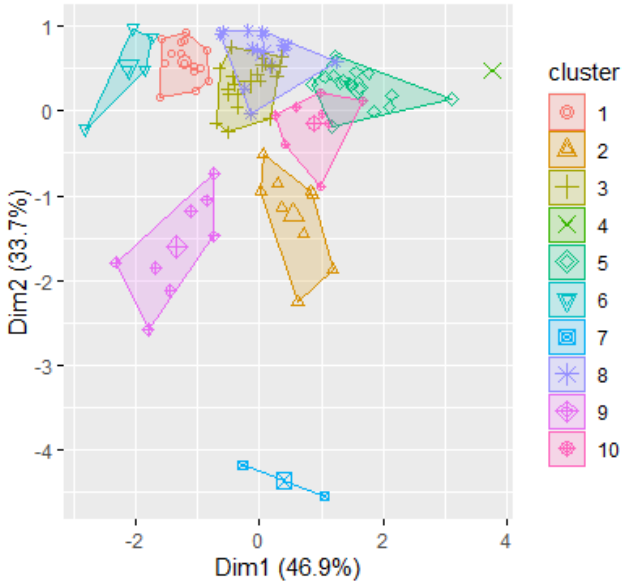
D. K-Means Clustering dengan Metode Silhouette

Dari pengolahan data yang dilakukan menggunakan bantuan aplikasi Rstudio didapatkan *output* visualisasi penentuan k dengan metode *Silhouette* sebagai berikut.



Gambar 3. Nilai k Optimal Metode *Silhouette*

Berdasarkan gambar tersebut dapat diartikan bahwa nilai k yang optimal untuk pengolahan 100 data dengan variabel numerik *Runtime*, *Rating*, dan *Votes* menggunakan metode *Silhouette* adalah sebanyak 10 *cluster* hal tersebut dapat dilihat dari garis yang ditunjukkan oleh gambar. Maka akan dilakukan *Clustering* sebanyak 10 *cluster*. Dengan bantuan Rstudio didapatkan hasil *clustering* sebagai berikut.

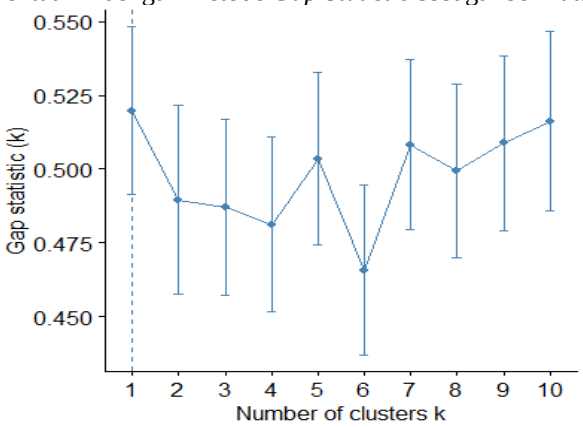


Gambar 4. 10 *Cluster Plot* Data

Berdasarkan gambar tersebut dapat diartikan bahwa 100 data film berbahasa korea sudah terbagi menjadi 10 *cluster* berdasarkan variabel numerik *Runtime*, *Rating*, dan *Votes*. Didapatkan terdapat sebanyak 2 film di *cluster* 1, 5 film di *cluster* 2, 22 film di *cluster* 3, 1 film di *cluster* 4, 9 film di *cluster* 5, 16 film di *cluster* 6, 15 film di *cluster* 7, 8 film di *cluster* 8, 14 film di *cluster* 9, 8 film di *cluster* 10. Dengan demikian *clustering* dengan *cluster* sebanyak 10 terlalu banyak dan mengakibatkan adanya 1 data saja dalam 1 *cluster*.

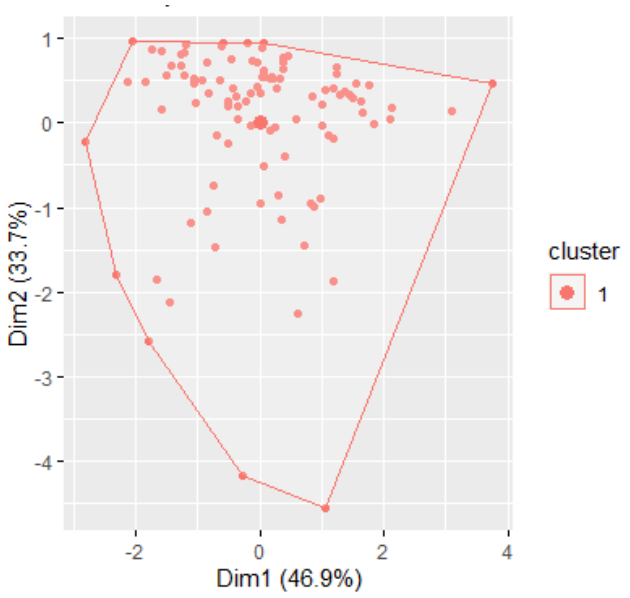
E. K-Means Clustering dengan Metode Gap Statistic

Dari pengolahan data yang dilakukan menggunakan bantuan aplikasi Rstudio didapatkan *output* visualisasi penentuan k dengan metode *Gap Statistic* sebagai berikut.



Gambar 5. Nilai k Optimal Metode *Gap Statistic*

Berdasarkan gambar tersebut dapat diartikan bahwa nilai k yang optimal untuk pengolahan 100 data dengan variabel numerik *Runtime*, *Rating*, dan *Votes* menggunakan metode *Elbow* adalah sebanyak 1 *cluster* hal tersebut ditunjukkan dengan adanya garis lurus vertikal pada saat k sebesar 1. Maka akan dilakukan *K-Means Clustering* sebanyak 1 *cluster*. Dengan bantuan Rstudio didapatkan hasil *clustering* sebagai berikut.



Gambar 6. 1 *Cluster Plot* Data

Berdasarkan gambar tersebut dapat diartikan bahwa 100 data film berbahasa korea tidak terbagi atau dapat dikatakan tidak dilakukan *clustering*.

V. KESIMPULAN

Pada web IMDb 100 data film berbahasa korea terbaik dapat dilakukan *web crawling and scrapping* dan disimpan dalam bentuk csv dan juga dapat dilakukan perhitungan statistika deskriptif pada variabel *runtime*, *rating*, dan *votes* karena variabel tersebut merupakan variabel numerik. Selanjutnya dilakukan perhitungan untuk mendapatkan jumlah *cluster* optimal menggunakan metode *Elbow* dan didapatkan jumlah *cluster* yang optimal untuk data adalah sebanyak 3. Setelah itu dilakukan proses *clustering* data sebanyak 3. *Clustering* dengan metode *silhouette* didapatkan jumlah *cluster* optimal sebanyak 10 dan setelah dilakukan proses *clustering* dapat disimpulkan bahwa jumlah *cluster* tersebut terlalu banyak untuk dijadikan jumlah *cluster* optimum. *Clustering* dengan metode *Gap Statistic* didapatkan jumlah *cluster* optimum sebanyak 1 *cluster* yang sama saja artinya tidak dilakukan clustering. Perbedaan jumlah *cluster* optimum pada ketiga metode tersebut adalah dikarenakan oleh perbedaan rumus yang digunakan pada perhitungan k di setiap metode. Berdasarkan ketiga metode untuk k optimal, akan digunakan k sebanyak 3 untuk penelitian lebih lanjut.

DAFTAR PUSTAKA

[1] https://id.wikipedia.org/wiki/Film

[2] https://id.wikipedia.org/wiki/Internet_Movie_Database

[3] Witten, I. H. & Frank, E., 2005. *Data Mining: Practical Machine Learning Tools and Techniques.* San Francisco: Elsevier Inc.

[4] Walpole, R. E. (1995). Pengantar Statistika. Jakarta: PT Gramedia.

[5] Freund, J. (2001) *Modern Elementary Statistics.* Prentice Hall.

[6] Turland, M. (2010). Php| architect's Guide to Web Scrapping with PHP. Introduction-Web Scrapping Defined, str, 2.

[7] N. Atthina dan L. Iswari, "Klasterisasi Data Kesehatan Penduduk untuk Menentukan Rentang Derajat kesehatan Daerah dengan Metode K-means," Seminar Nasional Aplikasi Teknologi Informasi (SNATI), Vol. %1 dari %2ISSN 1907 - 5022, pp. B52 - B59, 2014.

[8] Madhulatha, T.S., 2012. *An Overview on Clustering Methods*. IOSR Journal of Engineering, II (4), pp.719-725

[9] HUNG, C.M., WU, J., CHANG, J.H. & YANG, D.L., 2005. *An Efficient k-Means Clustering Algorithm Using Simple Partitioning*. JOURNAL OF INFORMATION SCIENCE AND ENGINEERING, XXI (1), pp.1157-77

[10] A. P. Windarto, "Implementation of Data Mining on Rice Imports by Major Country of Origin Using Algorithm using K-Means Clustering Method", Int. J. Artif. Intell. Res., vol. 1, no. 1, pp. 60-33,2017.

[11] M. G. Sadewo, A. P. Windarto, and D. Hartama, "PENERAPAN DATA MINING PADA POPULASI DAGING AYAM RAS PEDAGING DI INDONESIA BERDASARKAN PROVINSI MENGGUNAKAN K-MEANS", *InfoTekJar (Jurnal Nas. Inform. Dan Teknol. Jaringan)*, vol. 2, no. 1, pp. 60-67, 2017.

[12] Bholowalia, Purnima & Kumar, Arvind, 2014. *EBK-Means: A Clustering Techniques based on Elbow Method and K-Means in WSN. International Journal of Computer Application* (0975-8887), IX (105), PP. 17-24

Lampiran 1. Syntax yang digunakan di RStudio

```
library(xml2)
library(rvest)
alamatweb<-
'https://www.imdb.com/search/title/?title_type=feature&languages=ko&count=100'
lamanweb<-read_html(alamatweb)
lamanweb
#Title
film_data_laman=html_nodes(lamanweb,'.lister-item-header a')
film_data_laman
film_data=html_text(film_data_laman)
film_data
#RUNTIME
runtime_data_laman<-
html_nodes(lamanweb,'.runtime')
runtime_data_laman
runtime_data<-html_text(runtime_data_laman)
head(runtime_data)
runtime_data<-gsub("min","",runtime_data)
runtime_data
runtime_data<-as.numeric(runtime_data)
runtime_data
#GENRE
genre_data_laman=html_nodes(lamanweb,'.genre')
genre_data_laman
genre_data=html_text(genre_data_laman)
genre_data
genre_data=gsub("\n","",genre_data)
genre_data=gsub(" ","",genre_data)
genre_data=gsub(",.*","",genre_data)
genre_data
genre_data=as.factor(genre_data)
head(genre_data)
#RATING
rating_data_laman=html_nodes(lamanweb,'.ratings-imdb-rating strong')
rating_data_laman
rating_data=html_text(rating_data_laman)
rating_data
rating_data=as.numeric(rating_data)
rating_data
#Votes
votes_data_laman=html_nodes(lamanweb,'.sort-num_votes-visible span:nth-child(2)')
votes_data=html_text(votes_data_laman)
votes_data
votes_data=gsub(",",".",votes_data)
votes_data=as.numeric(votes_data)
votes_data
#Mengumpulkan data menjadi data frame
kumpulan_data_film=data.frame(Titles=film_data,Genre = genre_data,Runtime = runtime_data,Rating = rating_data,Votes=votes_data)
str(kumpulan_data_film)
view(kumpulan_data_film)
summary(kumpulan_data_film)
write.csv(kumpulan_data_film,file='dataEAS.csv',row.names = FALSE)
#Clustering Kmeans
library(ggplot2)
library(factoextra)
data=read.csv("C:/Users/USER/Documents/dataEAS.csv")
data
data_numerik=data[3:5]
data_numerik
#standardisasi untuk menyamakan satuan
data_stds=scale(data_numerik)
data_stds
#Metode Elbow
fviz_nbclust(data_stds,kmeans,method="wss")+
  geom_vline(xintercept = 3, linetype = 2)
clustering=kmeans(data_stds,centers = 3,nstart = 25)
fviz_cluster(clustering,geom = "point",data = data_stds)
#metode silhouette
fviz_nbclust(data_stds,kmeans,method="silhouette")
clustering=kmeans(data_stds,centers = 10,nstart = 25)
fviz_cluster(clustering,geom = "point",data = data_stds)
#Metode gap stat
gap_stat =clusGap(data_stds, FUN=kmeans, nstart = 25,K.max = 10)
print(gap_stat, method = "firstmax")
fviz_gap_stat(gap_stat)
clustering=kmeans(data_stds,centers = 1,nstart = 25)
fviz_cluster(clustering,geom = "point",data = data_stds)
#Menggabungkan data final dengan hasil clustering dengan metode elbow
data_final=data.frame(data,clustering$cluster)
data_final
```

Lampiran 2. Data Hasil *Clustering K-Means* Metode *Elbow*

Titles	Genre	Runtime	Rating	Votes	cluster
Gisaengchung	Comedy	132	8.6	533.468	3
Black Panther	Action	134	7.3	620.825	3
Snowpiercer	Action	126	7.1	316.522	3
Kol	Crime	112	7.1	12.303	2
Dark Waters	Biography	126	7.6	57.151	2
Avengers: Age of Ultron	Action	141	7.3	736.654	3
Oldeuboi	Action	120	8.4	512.172	3
Ah-ga-ssi	Drama	145	8.1	111.543	2
Cloud Atlas	Action	172	7.4	343.91	2
Salinui chueok	Action	131	8.1	137.097	2
Train to Busan 2	Action	116	5.4	18.429	1
Lucy	Action	89	6.4	444.18	3
Busanhaeng	Action	118	7.6	163.626	2

Training Day	Crime	122	7.7	388.073	3
Collateral	Crime	120	7.5	358.687	3
Olympus Has Fallen	Action	119	6.5	256.16	2
Falling Down	Action	113	7.6	170.628	2
#Saraitda	Action	98	6.2	23.263	1
Crash	Crime	112	7.7	418.476	3
Colossal	Comedy	109	6.2	58.651	1
Die Another Day	Action	133	6.1	204.395	2
Okja	Action	120	7.3	103.725	2
Beoning	Drama	148	7.5	48.331	2
Pineapple Express	Action	111	6.9	315.158	3
Ang-ma-reul bo-at-da	Action	144	7.8	110.271	2
Outbreak	Action	127	6.6	119.301	2
Yes Man	Comedy	104	6.8	334.399	3
Gi-eok-ui bam	Horror	108	7.4	17.953	2
Do the Right Thing	Comedy	120	8	88.746	2
Downsizing	Drama	135	5.7	98.51	2
Gokseong	Horror	156	7.5	50.882	2
Gwoemul	Action	120	7.1	106.476	2
RED 2	Action	116	6.6	159.511	2
The Interview	Action	112	6.5	305.504	3
Money Monster	Crime	98	6.5	93.952	1
Salt	Action	100	6.4	294.486	1
The Pacifier	Action	95	5.6	87.601	1
Liu lang di qiu	Action	125	6	26.533	2
Madeo	Crime	129	7.8	51.963	2
Greta	Drama	98	6	25.106	1
MASH	Comedy	116	7.4	67.275	2
The VelociPastor	Action	75	5.1	3.398	1
Crank	Action	88	6.9	233.609	1
Manyeo	Action	125	7.1	6.164	2
Red Dawn	Action	93	5.4	73.198	1
Starsky & Hutch	Comedy	101	6.1	138.014	1
Team America: World Police	Action	98	7.2	159.064	1
Hoo-goong: Je-wang-eui cheob	Drama	122	6.1	1.581	2
Domangchin yeoja	Drama	77	6.7	953	3
Bom Yeoareum Gaeul Gyeoul Geurigo Bom	Drama	103	8	77.163	2
Ajeossi	Action	119	7.8	62.442	2
Sanyangeui sigan	Action	134	6.3	3.347	2
Chinjeolhan geumjassi	Crime	115	7.6	71.629	2
The Kentucky Fried Movie	Comedy	83	6.5	17.062	1
Geom-gaek	Action	100	6.5	390	3
7-beon-bang-ui seon-mul	Comedy	127	8.2	16.488	2
Chugyeokja	Action	125	7.8	57.974	2
What Happened to Mr Cha?	Comedy	102	4.9	78	1
Bin-jip	Crime	88	8	50.259	1
Bakjwi	Drama	134	7.1	42.455	2
Stir of Echoes	Horror	99	7	75.105	1
Janghwa, Hongryeon	Drama	114	7.2	56.369	2
Akinjeon	Action	109	6.9	7.52	2
Playing It Cool	Comedy	94	6	25.845	1
Taxi	Action	86	7	76.6	1
Deep Rising	Action	106	6.1	33.284	1
Boksuneun naui geot	Crime	129	7.6	62.319	2
Harsh Times	Action	116	6.9	62.972	2
Dalkomhan insaeng	Action	119	7.6	36.232	2
Gamgi	Action	122	6.7	11.194	2
Jipuragirado japgo sipeun jimseungdeul	Crime	108	7.1	1.952	2
Taeksi woonjunsa	Action	137	7.9	15.827	2
Sin-gwa ham-kke: Jwi-wa beol	Action	139	7.3	11.917	2
Namsanui bujangdeul	Drama	114	7	1.629	2
Rollerball	Action	98	3.1	25.673	1
Ayla: The Daughter of War	Biography	125	8.4	33.681	2
Columbus	Drama	104	7.2	14.33	2
Daman akeseo goohasoseo	Action	108	6.8	970	3
Baekdusan	Action	130	6.2	2.885	2
Stealth	Action	121	5.1	51.764	1
Svaha: The Sixth Finger	Horror	122	6.2	2.765	2
Pan-dola	Action	136	6.7	5.71	2
Aknyeo	Action	124	6.7	12.114	2
Geukhanjikeob	Action	111	7.1	6.313	2
Do-ga-ni	Drama	125	8.1	11.185	2
Jigeum mannareo gabmida	Drama	132	7.7	3.923	2
Eye for an Eye	Crime	101	6.2	13.433	1
Gon-ji-am	Horror	95	6.3	4.975	1
Flandersui gae	Comedy	110	7	6.249	2
The Closet	Horror	97	5.7	1.022	1
In-gan-jung-dok	Drama	132	6.1	1.116	2
Gongdong gyeongbi guyeok JSA	Action	110	7.8	26.322	2
Hyeob-sang	Action	114	6.6	2.232	2
Dodookdeul	Action	135	6.8	9.137	2
Sanglyusahoe	Drama	120	5.4	627	3
Sorido Eopsi	Crime	99	6.5	286	1

Sinsegye	Action	135	7.6	20.346	2
Cradle 2 the Grave	Action	101	5.8	41.159	1
Seom	Drama	90	7	12.791	1
The Sun Is Also a Star	Drama	100	5.8	6.012	1

Lampiran 3. Data Hasil *Clustering K-Means* Metode *Silhouette*

Titles	Genre	Runtime	Rating	Votes	cluster
Gisaengchung	Comedy	132	8.6	533.468	10
Black Panther	Action	134	7.3	620.825	10
Snowpiercer	Action	126	7.1	316.522	10
Kol	Crime	112	7.1	12.303	3
Dark Waters	Biography	126	7.6	57.151	7
Avengers: Age of Ultron	Action	141	7.3	736.654	10
Oldeuboi	Action	120	8.4	512.172	10
Ah-ga-ssi	Drama	145	8.1	111.543	2
Cloud Atlas	Action	172	7.4	343.91	2
Salinui chueok	Action	131	8.1	137.097	7
Train to Busan 2	Action	116	5.4	18.429	6
Lucy	Action	89	6.4	444.18	5
Busanhaeng	Action	118	7.6	163.626	3
Training Day	Crime	122	7.7	388.073	10
Collateral	Crime	120	7.5	358.687	10
Olympus Has Fallen	Action	119	6.5	256.16	5
Falling Down	Action	113	7.6	170.628	3
#Saraitda	Action	98	6.2	23.263	6
Crash	Crime	112	7.7	418.476	10
Colossal	Comedy	109	6.2	58.651	6
Die Another Day	Action	133	6.1	204.395	9
Okja	Action	120	7.3	103.725	3
Beoning	Drama	148	7.5	48.331	2
Pineapple Express	Action	111	6.9	315.158	5
Ang-ma-reul bo-at-da	Action	144	7.8	110.271	2
Outbreak	Action	127	6.6	119.301	9
Yes Man	Comedy	104	6.8	334.399	5
Gi-eok-ui bam	Horror	108	7.4	17.953	3
Do the Right Thing	Comedy	120	8	88.746	7
Downsizing	Drama	135	5.7	98.51	9
Gokseong	Horror	156	7.5	50.882	2
Gwoemul	Action	120	7.1	106.476	3
RED 2	Action	116	6.6	159.511	3
The Interview	Action	112	6.5	305.504	5
Money Monster	Crime	98	6.5	93.952	8
Salt	Action	100	6.4	294.486	5
The Pacifier	Action	95	5.6	87.601	6
Liu lang di qiu	Action	125	6	26.533	9
Madeo	Crime	129	7.8	51.963	7
Greta	Drama	98	6	25.106	6
MASH	Comedy	116	7.4	67.275	3
The VelociPastor	Action	75	5.1	3.398	6
Crank	Action	88	6.9	233.609	8
Manyeo	Action	125	7.1	6.164	3
Red Dawn	Action	93	5.4	73.198	6
Starsky & Hutch	Comedy	101	6.1	138.014	6
Team America: World Police	Action	98	7.2	159.064	8
Hoo-goong: Je-wang-eui cheob	Drama	122	6.1	1.581	9
Domangchin yeoja	Drama	77	6.7	953	1
Bom Yeoareum Gaeul Gyeoul Geurigo Bom	Drama	103	8	77.163	3
Ajeossi	Action	119	7.8	62.442	7
Sanyangeui sigan	Action	134	6.3	3.347	9
Chinjeolhan geumjassi	Crime	115	7.6	71.629	3
The Kentucky Fried Movie	Comedy	83	6.5	17.062	8
Geom-gaek	Action	100	6.5	390	5

7-beon-bang-ui seon-mul	Comedy	127	8.2	16.488	7
Chugyeokja	Action	125	7.8	57.974	7
What Happened to Mr Cha?	Comedy	102	4.9	78	6
Bin-jip	Crime	88	8	50.259	8
Bakjwi	Drama	134	7.1	42.455	7
Stir of Echoes	Horror	99	7	75.105	8
Janghwa, Hongryeon	Drama	114	7.2	56.369	3
Akinjeon	Action	109	6.9	7.52	3
Playing It Cool	Comedy	94	6	25.845	6
Taxi	Action	86	7	76.6	8
Deep Rising	Action	106	6.1	33.284	6
Boksuneun naui geot	Crime	129	7.6	62.319	7
Harsh Times	Action	116	6.9	62.972	3
Dalkomhan insaeng	Action	119	7.6	36.232	3
Gamgi	Action	122	6.7	11.194	9
Jipuragirado japgo sipeun jimseungdeul	Crime	108	7.1	1.952	3
Taeksi woonjunsa	Action	137	7.9	15.827	7
Sin-gwa ham-kke: Jwi-wa beol	Action	139	7.3	11.917	7
Namsanui bujangdeul	Drama	114	7	1.629	3
Rollerball	Action	98	3.1	25.673	4
Ayla: The Daughter of War	Biography	125	8.4	33.681	7
Columbus	Drama	104	7.2	14.33	3
Daman akeseo goohasoseo	Action	108	6.8	970	1
Baekdusan	Action	130	6.2	2.885	9
Stealth	Action	121	5.1	51.764	9
Svaha: The Sixth Finger	Horror	122	6.2	2.765	9
Pan-dola	Action	136	6.7	5.71	9
Aknyeo	Action	124	6.7	12.114	9
Geukhanjikeob	Action	111	7.1	6.313	3
Do-ga-ni	Drama	125	8.1	11.185	7
Jigeum mannareo gabmida	Drama	132	7.7	3.923	7
Eye for an Eye	Crime	101	6.2	13.433	6
Gon-ji-am	Horror	95	6.3	4.975	6
Flandersui gae	Comedy	110	7	6.249	3
The Closet	Horror	97	5.7	1.022	6
In-gan-jung-dok	Drama	132	6.1	1.116	9
Gongdong gyeongbi guyeok JSA	Action	110	7.8	26.322	3
Hyeob-sang	Action	114	6.6	2.232	3
Dodookdeul	Action	135	6.8	9.137	9
Sanglyusahoe	Drama	120	5.4	627	5
Sorido Eopsi	Crime	99	6.5	286	5
Sinsegye	Action	135	7.6	20.346	7
Cradle 2 the Grave	Action	101	5.8	41.159	6
Seom	Drama	90	7	12.791	8
The Sun Is Also a Star	Drama	100	5.8	6.012	6