e-ISBN 978-623-93343-3-8
Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)
Jakarta-Indonesia, 20 April 2021

Penerapan *Collaborative Filtering*, PCA dan *K-Means* dalam Pembangunan Sistem Rekomendasi Film

Mu'tashim Billah [1], Muhammad Aidil Zartesya [2], Desta Sandya Prasvita, S Komp, M Kom. [3],
Ilmu Komputer
Jl. TB Simatupang, RT.3/RW.3, Cilandak Timur, Jakarta Selatan, DKI Jakarta 12560, Indonesia
e-mail: m.billah@students.esqbs.ac.id, m.aidil.z@students.esqbs.ac.id, desta.sandya@esqbs.ac.id

Abstrak. Penelitian ini bertujuan untuk mengembangkan sistem rekomendasi film menggunakan kombinasi dari *Collaborative Filtering*, PCA, dan K-Means. Metode PCA diterapkan pada data agar waktu yang dibutuhkan saat proses clustering lebih cepat. Rata-rata kompleksitas waktu yang dihasilkan adalah 1.061282. Proses clustering akan menentukan karakteristik seorang user berdasarkan tingkat kemiripan dengan user lainnya. Didapatkan hasil k terbaik dari pengujian *Silhouette Coefficient* dan pengujian *Elbow* terletak pada *k* = 3. Rekomendasi yang dihasilkan kemudian dihitung dengan *Mean Reciprocal Rank* (MRR) untuk mengetahui tingkat ketepatan sebuah rekomendasi. Rata-rata MRR yang dihasilkan adalah 0.44533417402269865. Dari nilai tersebut dapat dikatakan rekomendasi yang dihasilkan kurang tepat.

Kata Kunci: *Collaborative Filtering*, *K-Means Clustering*, PCA, Sistem Rekomendasi

1 Pendahuluan

Teknologi informasi dan telekomunikasi semakin berkembang dan mengalami peningkatan yang sangat tinggi, dalam hal ini dapat diketahui banyak sekali kegiatan manusia yang membutuhkan teknologi informasi dan komunikasi untuk saat ini, tidak terkecuali dalam bidang musik maupun film. Film merupakan audio visual yang memiliki banyak *genre*, seperti *genre* komedi, drama, *horor*, *action*, dan masih banyak lagi.

Film sudah menjadi salah satu media hiburan yang populer di kalangan masyarakat. Sejak tahun 1874 sampai 2015, sebanyak 3,361,741 judul film telah dikeluarkan oleh industri perfilman. Banyak-nya judul-judul film yang telah beredar memunculkan masalah baru bagi penikmat film untuk menemukan film mana yang selanjutnya akan ditonton. Data - data film yang terdapat dalam suatu website dapat diolah dan dimanfaatkan untuk merekomendasikan film kepada *user* lain. Masalah ini dapat diatasi dengan menyampaikan informasi berupa daftar-daftar film yang menjadi rekomendasi kepada penikmat film tersebut berdasarkan preferensinya sendiri (*user*). Oleh karena itu, diperlukan suatu sistem yang dapat memberikan rekomendasi film kepada *user*.

Dalam memberikan rekomendasi, sistem rekomendasi perlu mengetahui daftar item mana saja yang menjadi ciri dari *user* tersebut agar dapat mengenali dan memberikan rekomendasi terkait item yang disukainya tersebut. Pada penerapannya, sistem rekomendasi dibagi menjadi dua pendekatan antara lain *content-based filtering* dan *collaborative filtering*. Pada penelitian ini, metode yang digunakan adalah *collaborative filtering* yang melibatkan data *user* lain yang memiliki kemiripan dengan *user* yang akan diberikan rekomendasi.

Mengacu kepada penikmat film yang jumlahnya tidak sedikit, maka perlu adanya pengelompokan terlebih dahulu sebelum memberikan rekomendasi agar hasil daftar rekomendasi menjadi lebih akurat. Untuk mengelompokan user menjadi digunakan salah satu metode clustering yaitu K-Means *Clustering*. *User* akan terlebih dahulu dikelompokan berdasarkan daftar *item* yang disukai sebelum diberikan rekomendasi *item*. Namun, karena jumlah film juga tidak sedikit dan mengakibatkan fitur yang dihasilkan semakin banyak, maka pada penelitian ini digunakan metode *Principal Component Analysis* (PCA) guna mengurangi dimensi pada data namun tidak menghilangkan makna dari data tersebut.

2 Landasan Teori

Adapun teori-teori yang digunakan dalam penelitian ini adalah sebagai berikut:

2.1 Sistem Rekomendasi

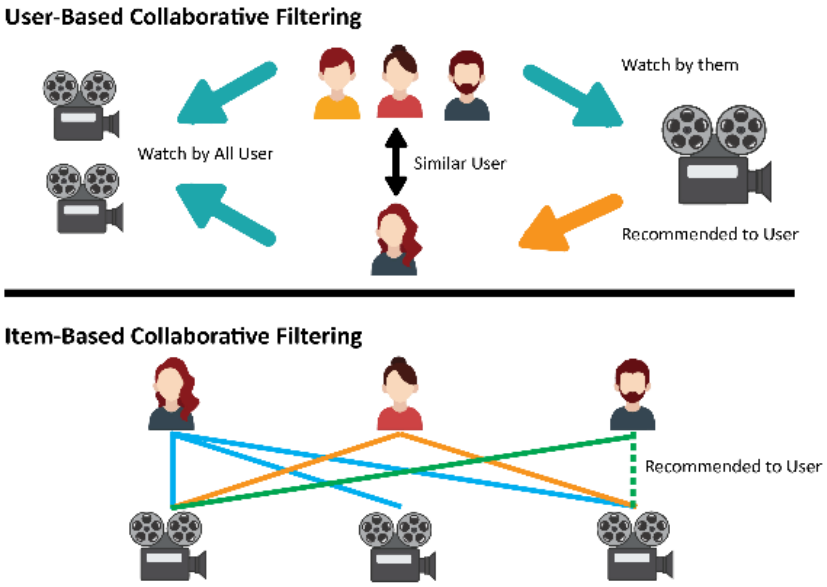
Sistem rekomendasi merupakan sebuah sistem atau program yang dapat membuat keputusan bagi pengguna terkait item yang disukai dan diinginkannya [1]. Sistem rekomendasi dapat digambarkan sebagai daftar kebutuhan atau keinginan pengguna berdasarkan karakteristik dari pengguna itu sendiri [1]. Sistem rekomendasi memiliki output berupa daftar item yang diurutkan berdasarkan rating kemiripan tertinggi hingga terendah.

2.2 *Collaborative Filtering*

Collaborative Filtering merupakan suatu metode atau cara menyaring dan mengevaluasi suatu item berdasarkan opini user lain [2]. Collaborative Filtering dibagi menjadi 2 metode, yaitu *Item-Based Collaborative Filtering* (ICF) dan *User-Based Collaborative Filtering* (UCF). Pada metode ICF, sistem memberikan rekomendasi item yang mirip dengan profil pengguna [3]. Sedangkan UCF memberikan rekomendasi item kepada pengguna berdasarkan pendapat pengguna yang terdekat atau mirip yang memiliki pemikiran yang sama terkait item yang disukai [4]. Visualisasi perbedaan UCF dan ICF dalam memberikan rekomendasi kepada *user* dapat dilihat pada gambar 2 dan rumus UCF dalam memberikan rekomendasi *item* dapat dilihat dibawah ini.

$$S(x,y) = \sum R(z,y) \times Si(x,z)ny \\ = 1 \sum |Si(x,z)|ny = 1 \quad (1)$$

Dimana:
$S(x,y)$ = Nilai rekomendasi *user* (x) terhadap *item* (y)
$R(z,y)$ = Besar *rating* atau tingkat kesukaan *user* (z) terhadap *item* (y)
$Si(x,z)$ = Nilai *similarity* antara *user* (x) dengan *user* (z)



Gambar. 2.1. Visualisasi User-Based dan Item-Based Collaborative Filtering

e-ISBN 978-623-93343-3-8
Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)
Jakarta-Indonesia, 20 April 2021

2.3 *Principal Component Analysis*

Principal Component Analysis (PCA) merupakan metode yang mampu melakukan reduksi dimensi pada suatu data, namun tetap menggambarkan dan mempertahankan pola dan tren data tersebut [5].

2.4 *K-Means Clustering*

K-Means *Clustering* merupakan metode yang mengoptimalkan data yang sensitif dan berdekatan terhadap pemilihan awal dan posisi tengah dari kumpulan data tersebut [6]. K-Means sering digunakan dalam proses pengelompokan data untuk menentukan label atau cirinya.

2.5 *Silhouette Coefficient*

Silhouette Coefficient dapat digunakan sebagai metode ukur dari hasil clustering. Metode ini juga dapat memilih jumlah *k* terbaik dalam model K-Means *Clustering*, sehingga model yang dibuat berdasarkan nilai *Silhouette Coefficient* tertinggi dapat menggambarkan struktur data yang telah dikelompokkan [7]. Adapun rumus untuk menghitung nilai *Silhouette Coefficient* adalah sebagai berikut:

$$s(a) = y - x \times \max(x, y) \quad (2)$$

Dimana:
$s(a)$ = Nilai *Silhouette Coefficient*
x = Rata-rata nilai *intra cluster distance*
y = Rata-rata nilai *inter cluster distance*

2.6 *Elbow*

Elbow Method berperan penting dalam proses pengujian jumlah *k* pada model *clustering*. Algoritma K-Means *Clustering* memiliki kelemahan saat menentukan jumlah *k* terbaik dari *n* percobaan [6]. Oleh karena itu, *Elbow Method* dapat mengatasi masalah tersebut sehingga model yang dihasilkan K-Means menjadi lebih baik. Berikut adalah rumus dari *Elbow Method*:

$$d = \sum (x_i - t_x) + (y_i - t_y) \quad (3)$$

Dimana:
d = Nilai *Distortion*
x_i = C*luster* (x) pada perulangan ke (i)
t_x = Titik tengah *cluster* (x)
y_i = *Cluster* (y) pada perulangan ke (i)
t_y = Titik tengah *cluster* (y)

2.7 *Mean Reciprocal Rank*

Saat proses rekomendasi item telah selesai diproses, alat ukur untuk menentukan tingkat akurasi dalam pemberian rekomendasi dapat dilakukan melalui penghitungan *Mean Reciprocal Rank* (MRR). MRR membandingkan dua atau lebih rekomendasi, dimana rekomendasi relevan yang pertama dibandingkan dengan rekomendasi selanjutnya hingga mendapatkan nilai *rank* untuk masing-masing item di rekomendasi yang berbeda [8]. Untuk mengukur nilai MRR pada sebanyak *n* rekomendasi dapat menggunakan rumus dibawah ini:

$$MRR = 1|n| \sum 1R|n|i = 1 \quad (4)$$

e-ISBN 978-623-93343-3-8
Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)
Jakarta-Indonesia, 20 April 2021

Dimana:
MRR = Nilai *Mean Reciprocal Rank*
n = Frekuensi data
R = Ranking *item* dalam data

2.8 Penelitian Terdahulu

Penelitian ini mengacu kepada penelitian-penelitian sebelumnya, khususnya pada penelitian yang dilakukan oleh Ichwanto Hadi, Leo Willyanto Santoso, dan Alvin Nathaniel Tjondrowiguno yang berjudul "Sistem Rekomendasi Film menggunakan *User-based Collaborative Filtering* dan *K-modes Clustering*" yang memiliki kompleksitas waktu yang cukup besar dan memiliki nilai *Mean Reciprocal Rank* yang rendah. Penelitian ini menerapkan metode tambahan yaitu *Principal Component Analysis* (PCA) dengan tujuan agar proses *clustering* yang dilakukan dapat menjadi lebih cepat dan efisien.

3 Analisis dan Desain

Dari himpunan data yang sudah tersedia, perlu adanya proses analisis terlebih dahulu karena ada data-data yang tidak sesuai untuk dimasukkan kedalam model. Selain itu sistem juga memiliki beberapa proses yang terpisah, sehingga desain sistem yang dihasilkan menjadi beberapa bagian. Berikut adalah hasil analisis dan gambaran desain alur sistem yang diterapkan pada penelitian ini.

3.1 Analisis Data

Himpunan data dibagi menjadi dua himpunan antara lain himpunan data *movies* dan himpunan data (*user*) *ratings*. Contoh beberapa data pada himpunan data *movies* dapat dilihat pada gambar 3.1 dan contoh himpunan data (*user*) *rating* dapat dilihat pada gambar 3.2.

	movieId	title	genres
0	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
1	2	Jumanji (1995)	Adventure|Children|Fantasy
2	3	Grumpier Old Men (1995)	Comedy|Romance

```
Shape: (9708, 3)
```

Gambar. 3.1. Contoh himpunan data *movies* dengan *atribut movieId, title, dan genres*

	userId	movieId	rating	timestamp
0	1	1	4.0	964982703
1	1	3	4.0	964981247
2	1	6	4.0	964982224

```
Shape: (100836, 4)
```

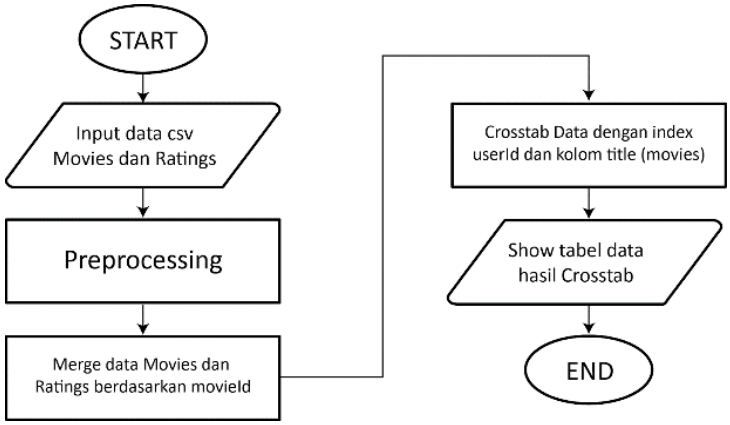
Gambar. 3.2. Contoh himpunan data *(user) ratings* dengan atribut *userId, movieId, rating, dan timestamp*

Pada data yang tersedia beberapa data *movie* masih ada terdapat *missing value* sehingga perlu adanya penghapusan terhadap data-data tersebut. Tentunya setelah data-data *movie* ada yang dihapus, maka perlu menghapus juga data-data (*user*) *ratings* dimana atribut movieId-nya tidak tersedia di himpunan data *movie*.

Selain itu, data genres pada himpunan data *movie* juga perlu dipisahkan menjadi koma agar memudahkan proses *clustering*.

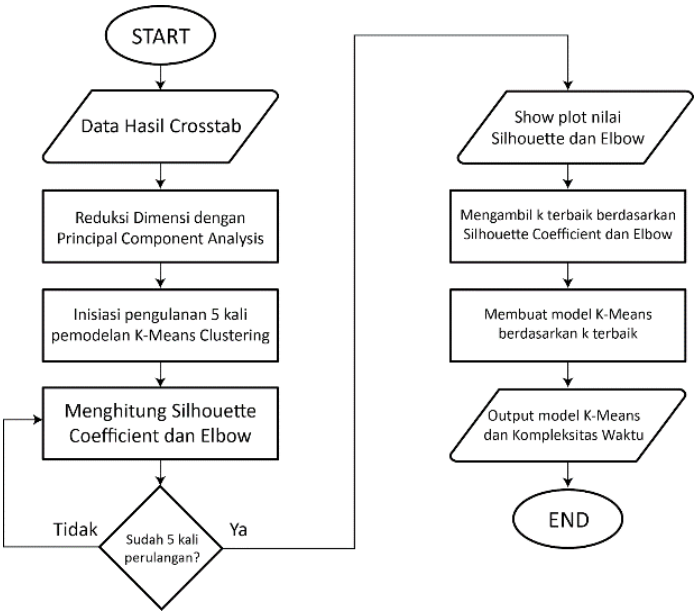
3.2 Desain Alur Sistem

Pada pemrosesan data *movie* dan (*user*) *ratings* yang akan menghasilkan himpunan rekomendasi kepada *user*, maka perlu dilakukan proses persiapan terlebih dahulu agar data yang digunakan menjadi data yang baik. Proses yang terjadi yaitu *preprocessing* dimana data yang tersedia masih ada beberapa field yang kosong. Selain itu, data juga perlu di migrasi berdasarkan atribut movieId dan data juga perlu dilakukan *crosstab* sehingga kita dapat melihat rekomendasi item berdasarkan item yang telah disukai oleh user lainnya. Alur proses pertama dapat dilihat pada gambar 3.3.



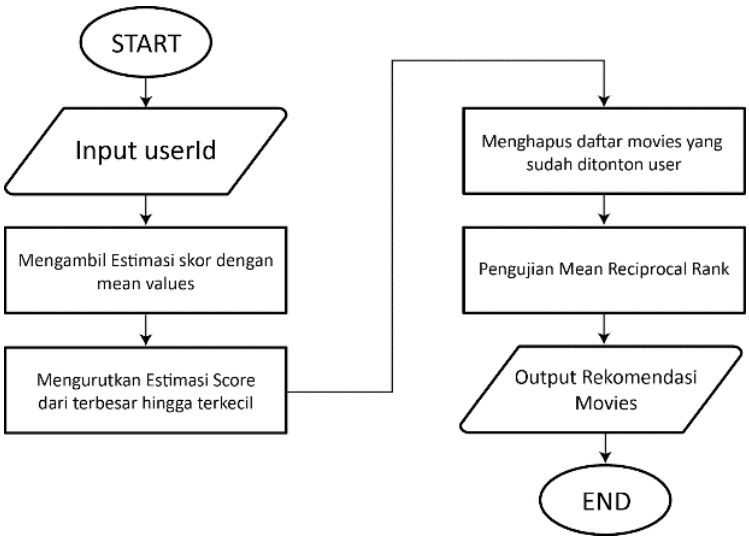
Gambar. 3.3. Desain alur kerja sistem dalam memberikan rekomendasi (Tahap 1)

Setelah dilakukan proses persiapan, maka data siap melakukan proses clustering. Pada penelitian ini proses *clustering* terlebih dahulu dilakukan penilaian untuk mencari *k* terbaik agar hasil pengelompokan dapat menjadi lebih akurat. Pengujian dilakukan dengan membandingkan hasil nilai *Silhouette Coefficient* dan *Elbow*. Alur kerja kedua dalam penelitian ini dapat dilihat pada gambar 3.4.



Gambar. 3.4. Desain alur kerja sistem dalam memberikan rekomendasi (Tahap 2)

Proses clustering akan menentukan karakteristik seorang user berdasarkan tingkat kemiripan dengan *user* lainnya. Sehingga proses selanjutnya adalah memberikan rekomendasi berdasarkan kemiripan user pada *cluster*-nya. *User* akan diberikan 15 rekomendasi *movie* dengan *mean values* tertinggi berdasarkan *movie* yang belum pernah ditonton sebelumnya. Pada proses ini juga akan mengukur seberapa besar tingkat keakuratan hasil rekomendasi dengan menggunakan pengujian *Mean Reciprocal Rank*. Alur kerja sistem saat memberikan rekomendasi movie kepada *user* dapat dilihat pada gambar 3.5.



Gambar. 3.5. Desain alur kerja sistem dalam memberikan rekomendasi (Tahap 3)

4 Hasil Pengujian

Penelitian ini diuji dengan pengujian *Silhouette Coefficient* dan *Elbow* untuk menemukan jumlah *cluster* terbaik, serta pengujian *Mean Reciprocal Rank* untuk menguji ketepatan hasil rekomendasi. Pengujian dilakukan pada laptop dengan rincian sistem Windows 10, CPU Intel Core i5-8250 CPU @1.60 GHz (8 CPUs) dan pada kondisi malam hari.

4.1 Pengujian Kompleksitas Waktu

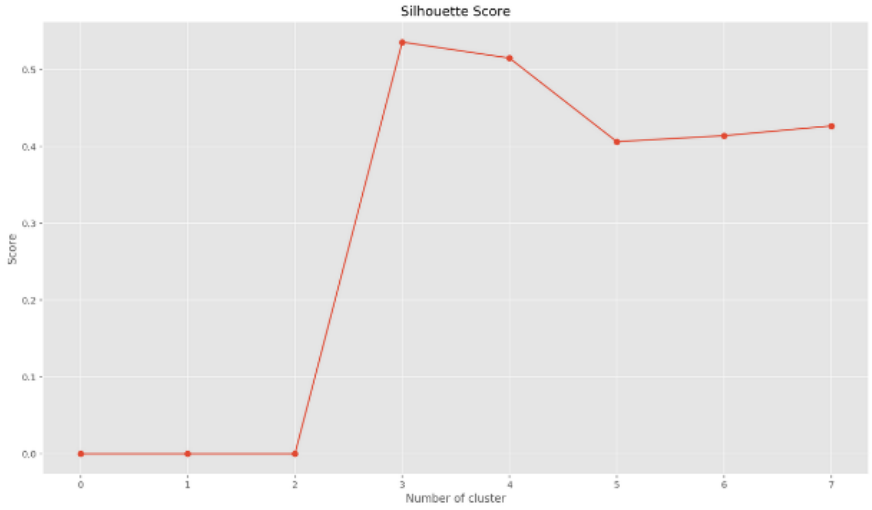
Pengujian kompleksitas waktu yang dihasilkan dalam 5 kali iterasi dalam melakukan *Clustering* dengan metode K-Means yang sebelumnya dilakukan reduksi dimensi menggunakan *Principal Component Analysis*. Hasil yang didapatkan adalah kompleksitas waktu paling cepat terletak pada k = 3 dengan kompleksitas waktu sebesar 1.0532, sedangkan kompleksitas tertinggi terletak pada k = 7 yaitu sebesar 1.07056. Sehingga dihasilkan rata-rata dari kelima perulangan tersebut sebesar 1.061282. Hasil ini dapat dikatakan bahwa kompleksitas waktu yang dibutuhkan tergolong cepat dan singkat. Rincian hasil pengujian kompleksitas waktu dapat dilihat pada tabel 4.1.

Tabel 4.1. Hasil pengujian Kompleksitas Waktu

Jumlah Cluster	Waktu Proses
3	1.0532
4	1.05449
5	1.06057
6	1.06759
7	1.07056
Rata-rata	1.061282

4.2 Pengujian *Silhouette Coefficient*

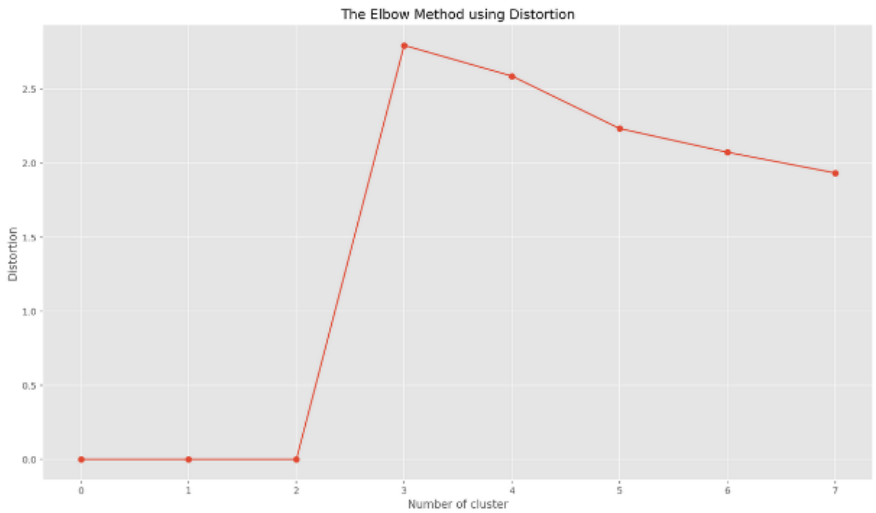
Pengujian *Silhouette Coefficient* dilakukan sebanyak 5 kali, mulai dari k = 3 hingga 7. Hasil pengujian tersebut mendapatkan cluster terbaik berada pada k = 3. Hasil ini didapatkan karena semakin tinggi hasil *Silhouette Coefficient* maka data-data yang berada pada cluster-cluster tersebut juga sudah sesuai. Grafik perbandingan nilai *Silhouette Coefficient* tiap k-nya dapat dilihat pada gambar 4.1.



Gambar. 4.1. Hasil Pengujian *Silhouette Coefficient*

4.3 Pengujian *Elbow*

Metode kedua dalam menentukan jumlah k terbaik untuk melakukan proses *clustering* adalah dengan menguji nilai *Elbow*. Dari hasil yang sudah diperoleh, didapatkan bahwa jumlah *cluster* terbaik terletak pada k = 3. Hasil ini diperoleh berdasarkan grafik pada gambar 4.2 yang menunjukan penurunan yang landai dan signifikan pada jumlah *cluster* setelah 4.



Gambar. 4.2. Hasil Pengujian *Elbow Method* menggunakan *Distortion*

4.4 Pengujian *Mean Reciprocal Rank*

Pengujian *Mean Reciprocal Rank* melibatkan seluruh data himpunan 15 tertinggi dari rekomendasi tiap-tiap *user*. Tabel dibawah ini merupakan rincian dari hasil *Mean Reciprocal Rank* tiap-tiap *user*, kemudian menghasilkan rata-rata yaitu sebesar 0.44533417402269865. Berdasarkan kaidah dari metode pengujian *Mean Reciprocal Rank*, apabila hasilnya berjarak dari 1 hingga 0.5 dapat dikatakan rekomendasi sudah tepat. Namun, apabila nilainya berada di jarak 0.5 sampai dengan 0, maka rekomendasi yang dihasilkan dapat dikatakan kurang tepat, sehingga dilihat dari hasil yang diperoleh dari penelitian ini, dikatakan hasil rekomendasi *movie* kepada user kurang tepat. Pada tabel 4.2 dapat dilihat beberapa rincian pengujian *Mean Reciprocal Rank* dari beberapa *user*.

Tabel 4.2. Hasil pengujian *Mean Reciprocal Rank*

userId	*Mean Reciprocal Rank*
1	0.00
2	0.06666666666666667
3	0.5333333333333333
4	0.13333333333333333
5	0.00
10	0.06666666666666667
20	0.6
50	0.00
100	0.06666666666666667
500	0.06666666666666667
Rata-rata keseluruhan	1.061282

5 Penutup

5.1 Simpulan

Penelitian ini telah menghasilkan sebuah sistem rekomendasi film dengan menggunakan algoritma K-Means *Clustering* dan *User-Based Collaborative Filtering* yang telah dikembangkan lagi dengan menggunakan metode *Principal Component Analysis*. Kompleksitas waktu yang dihasilkan setelah dilakukan reduksi menggunakan *Principal Component Analysis* yaitu sebesar 1.061282. Tingkat akurasi pada hasil rekomendasi telah dihitung dengan menggunakan nilai MMR. Nilai rata-rata MMR dari tersebut adalah sebesar 0.44533417402269865 yang disimpulkan bahwa rekomendasi yang dihasilkan kurang tepat.

5.2 Saran Pengembangan Lanjutan

Penelitian ini dapat dikembangkan lebih lanjut agar nilai yang dihasilkan dapat lebih baik dan lebih akurat lagi. Peneliti memberikan saran agar menerapkan beberapa Preprocessing tambahan dan menerapkan GridSearchCV agar dapat menemukan jumlah *cluster* yang lebih baik lagi karena menggunakan *hyper parameter* yang lebih lengkap dan rinci.

Referensi

[1] B. T. W. Utomo and A. W. Anggriawan, "Sistem Rekomendasi Paket Wisata Se-Malang Raya Menggunakan Metode *Hybrid Content Based* Dan *Collaborative*," J. Ilm. Teknol. Inf. Asia, vol. 9, no. 1, pp. 6–13, 2015.

[2] M. Nilashi, K. Bagherifard, O. Ibrahim, H. Alizadeh, L. A. Nojeem, and N. Roozegar, "*Collaborative filtering recommender systems*," Res. J. Appl. Sci. Eng. Technol., vol. 5, no. 16, pp. 4168–4182,

[3] F. XUE, X. HE, X. WANG, J. XU, K. LIU, and R. HONG, "*Deep item-based collaborative filtering for top-n recommendation*," arXiv, vol. 37, no. 3, 2018.

[4] Z. Tan and L. He, "*An Efficient Similarity Measure for User-Based Collaborative Filtering Recommender Systems Inspired by the Physical Resonance Principle*," IEEE Access, vol. 5, pp. 27211–27228, 2017, doi: 10.1109/ACCESS.2017.2778424.

[5] J. Lever, M. Krzywinski, and N. Altman, "*Points of Significance: Principal component analysis*," Nat. Methods, vol. 14, no. 7, pp. 641–642, 2017, doi: 10.1038/nmeth.4346.

[6] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, "*Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster*," IOP Conf. Ser. Mater. Sci. Eng., vol. 336, no. 1, 2018, doi: 10.1088/1757-899X/336/1/012017.

[7] R. Lletí, M. C. Ortiz, L. A. Sarabia, and M. S. Sánchez, "*Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes*," Anal. Chim. Acta, vol. 515, no. 1, pp. 87–100, 2004, doi: 10.1016/j.aca.2003.12.020.

[8] Y. Shi, A. Karatzoglou, L. Baltrunas, M. Larson, N. Oliver, and A. Hanjalic, "CLiMF," p. 139, 2012, doi: 10.1145/2365952.2365981.