Bidang Ilmu : Ilmu Komputer

LAPORAN PENELITIAN

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

Tim Peneliti:

Ketua : Ahmad Nur Ihsan Purwanto
Anggota : Mu'tashim Billah
M. Aidil Zartesya

PROGRAM STUDI ILMU KOMPUTER

LEMBAGA PENELITIAN DAN PENGABDIAN KEPADA MASYARAKAT
SEKOLAH TINGGI ILMU MANAJEMEN DAN ILMU KOMPUTER
(STIMIK ESQ)
2020

PENGESAHAN

1. Judul Penelitian : Penerapan Collaborative Filtering, PCA dan K-Means dalam Pembangunan Sistem Rekomendasi Film
2. Peneliti
   a. Nama Lengkap : Ahmad Nur Ihsan Purwanto
   b. Jenis Kelamin : Laki-Laki
   c. NIP/NIDN : 0310019102
   d. Jabatan Struktural :
   e. Jabatan fungsional : Dosen Tetap
   f. Pangkat / Golongan : -
   g. Fakultas/Program Studi : Ilmu Komputer
   h. Pusat Penelitian : STIMIK ESQ
   i. Alamat Institusi : Menara 165 Lt.18-19. Jl. TB Simatupang Kav 1 Cilandak
   j. Telpon/Faks/E-mail : +62 857-7698-1248
3. Jangka Waktu Penelitian : 12 bulan (2 semester)
4. Pembiayaan
   a. Jumlah biaya yang diajukan ke STIMIK ESQ : Rp. 3.000.000

Jakarta, 5 Februari 2020

Mengetahui,

Ketua Program Studi Ilmu Komputer	Ketua Peneliti
Ahlijati Nuraminah, S.Kom., M.T.I.	Ahmad Nur Ihsan Purwanto
NIDN: 0317128404	NIDN: 0310019102

Kepala LPPM

Danang Indrajaya, S.Si., M.Si
NIDN: 0311118108

IDENTITAS PENELITIAN

1. Judul Penelitian : Penerapan Collaborative Filtering, PCA dan K-Means dalam Pembangunan Sistem Rekomendasi Film
2. Peneliti
   a. Nama Lengkap : Ahmad Nur Ihsan Purwanto
   b. NIDN : 0310019102
   c. Pangkat / Golongan : -
   d. Jabatan fungsional : Dosen Tetap
   e. Fakultas/Program Studi : Ilmu Manajemen dan Ilmu Komputer/Ilmu Komputer
   f. Pusat Penelitian : LP2M – Menara 165 Lt.18-19
   g. Alamat Institusi : Menara 165 Lt.18-19. Jl. TB Simatupang Kav 1 Cilandak
   h. Telpon/Faks/E-mail : +62 857-7698-1248
3. Anggota Peneliti :

NO	NAMA	KEAHLIAN	ALOKASI WAKTU
1	Mu'tashim Billah	Ilmu Komputer	
2	M. Aidil Zartesya	Ilmu Komputer	

4. Objek Penelitian :
5. Masa Penelitian
   Mulai : September 2019
   Berakhir : Februari 2020
6. Anggaran yang diusulkan : Rp. 3.000.000
7. Lokasi Penelitian : STIMIK ESQ
8. Hasil yang ditargetkan (temuan baru/paket teknologi/hasil lain), beri penjelasan :
9. Institusi lain yang terlibat :

DAFTAR ISI

ABSTRAK

Judul : Penerapan Collaborative Filtering, PCA dan K-Means dalam Pembangunan Sistem Rekomendasi Film.

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

BAB I

PENDAHULUAN

1.1. Latar Belakang

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. Dalam penelitian ini dicoba mereduksi data tersebut agar diharapkan bahwa hasil yang diberikan bisa lebih efektif dibandingkan dengan hasil yang biasa. Logikanya jika data lebih sedikit maka waktu prosesnya lebih cepat, tapi untuk hasilnya belum tentu lebih baik dari data tanpa reduksi.

Pada penelitian ini menggunakan *Principal Component Analysis* yang diperkenalkan oleh Karl Pearson pada 1901. Digunakan untuk menghitung kombinasi linear dan variable baru yang menggambarkan keragaman data asli sebanyak mungkin, dengan dimensi matriks data asli dapat disederhanakan tanpa harus kehilangan informasi penting (Retno, Georgina, & Nurtiti, 2010).

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.

1.2. Rumusan Masalah

Secara garis besar permasalahan pada penelitian ini adalah:

1. Bagaimanakah cara mengelompokan user yang akan diberikan rekomendasi?
2. Bagaimanakah cara melakukan penyeleksian daftar item film yang akan direkomendasikan, dan bagaimana tingkat ketepatannya dengan item film yang *user* nikmati.

1.3. Tujuan Penelitian

Beberapa tujuan dari penelitian ini adalah sebagai berikut:

1. Mengelompokan user yang akan diberikan rekomendasi menggunakan metode K-Means *Clustering*.

2. Menerapkan metode *Principal Component Analysis* (PCA) untuk melakukan penyeleksian daftar item film yang akan direkomendasikan, serta menganalisa bagaimana ketepatan hasil penyeleksian tersebut dengan item film yang *user* nikmati.

1.4. Metodologi Penelitian

Secara garis besar, rangkaian tahapan penelitian yang akan dilakukan adalah sebagai berikut:

a. Studi Literatur

Pada permulaan penelitian dilakukan beberapa studi literatur berkaitan dengan metode yang digunakan. Beberapa kajian yang dipelajari adalah mengenai *content-based filtering, collaborative filtering,* K-Means *Clustering, Silhouette Coefficient, Principal Component Analysis* (PCA), serta penelitian-penelitian sebelumnya yang terkait dengan system rekomendasi, *clustering,* dan teknik *Mechine Learning* dalam mereduksi dimensi untuk menginterpretasikan suatu data.

b. Pengumpulan Data

Tahapan selanjutnya adalah proses pengumpulan data penelitian. Data yang digunakan adalah data csv movies serta ratingnya.

c. Implementasi

Setelah melakukan persiapan data, selanjutnya dilakukan implementasi system rekomendasi menggunakan metode *Collaborative Filtering.* Tahapan ini akan menghasilkan sejumlah besar item film rekomendasi yang kemudian direduksi atau dikurangi menggunakan metode *Principal Component Analysis* dan di *clustering* serta dioptimalkan menggunakan *K-Means Clustering* yang selanjutnya akan digunakan pada tahapan analisis.

d. Analisis

Dari hasil implementasi dilakukan analisis untuk mengetahui bagaimana ketepatan hasil penyeleksian daftar item film menggunakan metode yang diusulkan tersebut dengan item film yang dipilih *user.*

e. Penarikan Kesimpulan

Setelah melakukan semua rangkaian tahapan penelitian, terakhir dilakukan penarikan kesimpulan untuk mengetahui bagaimana hasil akhir dari penelitian ini.

1.5. Sistematika Penulisan

Laporan penelitian ini terdiri dari lima bab dengan sistematika penulisan sebagai berikut:

a. BAB 1 PENDAHULUAN.

Terdiri dari penjelasan mengenai latar belakang dilakukannya penelitian, permasalahan yang ingin dipecahkan, tujuan penelitian, ruang lingkup penelitian, dan metodologi yang digunakan untuk melakukan penelitian. Bab ini diakhiri dengan sistematika penulisan laporan hasil penelitian.

b. BAB 2 LANDASAN TEORI

Pada bab ini dibahas mengenai teori tentang system rekomendasi dan *Mechine Learning*.

c. BAB 3 METODOLOGI

Bab ini menjelaskan mengenai metodologi penelitian meliputi penjelasan mengenai rancangan penelitian, data yang digunakan, *content-based filtering, collaborative filtering,* metode K-Means *Clustering, Silhouette Coefficient, Principal Component Analysis* (PCA) yang digunakan serta rangkaian eksperimennya.

d. BAB 4 HASIL DAN PEMBAHASAN

Bab ini memaparkan hasil yang diperoleh dari rangkaian eksperimen beserta analisis terkait hasil eksperimen.

e. BAB 5 KESIMPULAN

Bab ini menyampaikan kesimpulan terhadap hasil dan analisis beserta saran terhadap metode yang digunakan.

BAB II
LANDASAN TEORI

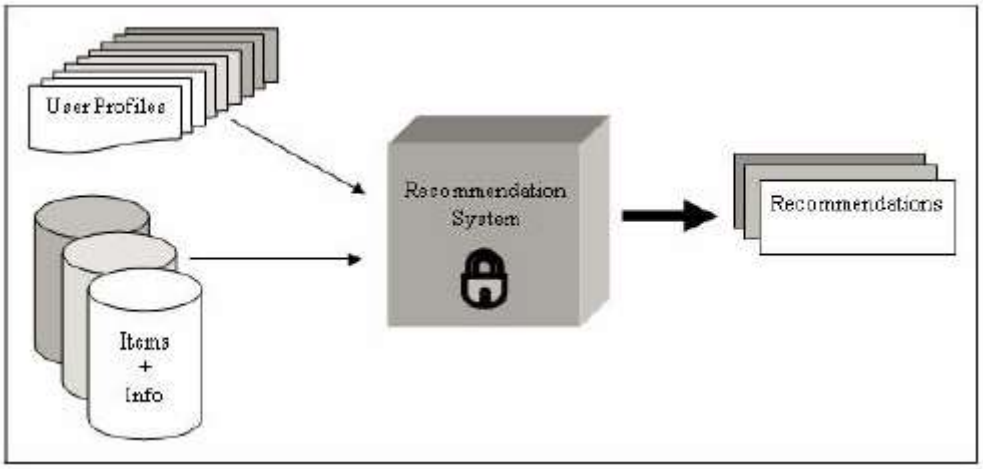
2.1. Dasar Teori

2.1.1 System Rekomendasi

Sistem rekomendasi adalah suatu system yang digunakan oleh para user atau pelanggan untuk mendapatkan produk yang diinginkan. Ide awal dari sistem rekomendasi sendiri adalah untuk menggunakan beberapa sumber informasi, tujuan utama dari sistem rekomendasi adalah untuk meningkatkan penjualan produk.

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.

Proses rekomendasi dilakukan dengan mengidentifikasi sumber informasi yang diperlukan yaitu informasi yang dijadikan sebagai masukan. Untuk lebih jelas peroses tersebut dapat dilihat pada gambar berikut:



Gambar 2.1. Proses Rekomendasi

Dalam hal sistem rekomendasi audio atau video, informasi ini dapat menghasilkan basis data berdimensi besar. Pada gambar diatas kita dapat membedakan bahwa produk akhir dari sistem akan menjadi seperangkat rekomendasi bagi pengguna. Representasi akhir dari rekomendasi ini tergantung pada sistem itu sendiri tetapi dapat berkisar dari daftar item yang dipesan, tangkapan dari item, atau seluruh item. Terdapat beragam metode yang digunakan untuk membuat sistem rekomendasi.

1. *Content-based filtering*

   Memberikan rekomendasi berdasarkan kemiripan atribut dari item atau barang yang disukai. Pada sistem rekomendasi lagu kemiripan berdasarkan atribut yang dimiliki oleh lagu seperti genre, beat, informasi dari artis.

2. *Knowledge-based*

   Memberikan rekomendasi berdasarkan kondisi nilai atribut yang telah ditentukan oleh user. Dalam sistem rekomendasi ini adalah pada awal penggunanaanya user di minta memasukkan item-item yang dia sukai secara eksplisit yang nantinya akan digunakan untuk merekomendasikan berdasarkan atribut dari item-item yang sudah disukai.

3. *Hybrid filtering*

   Merupakan kombinasi dari metode rekomendasi yang lain untuk menghasilkan rekomendasi lebih akurat

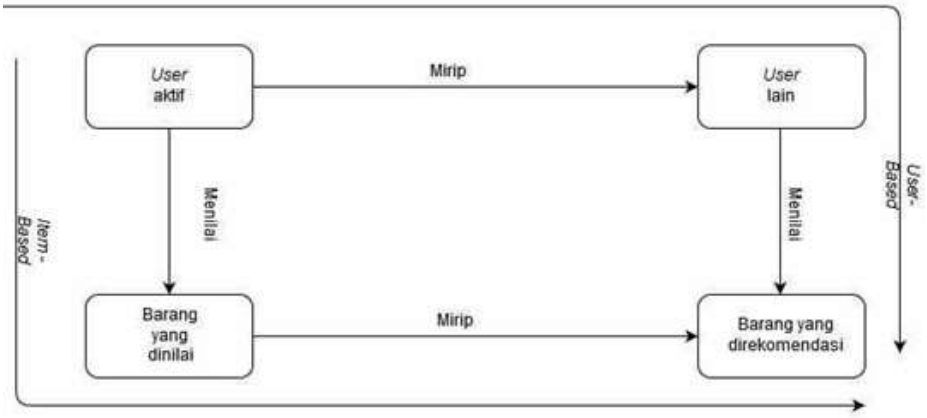
4. *Collaborative-filtering*

   Memberikan rekomendasi berdasarkan feedback dari user yang lain atau dari diri sendiri. Penerapan dalam rekomendasi musik yaitu pembentukan user-matrix yang berisi pereferensi dari user yang di bentuk dari data feedback yang berupa data streaming dan download user yang lain.

   Jika disederhanakan, *collaborative filtering* merupakan proses rekomendasi daftar barang untuk pengguna, berdasarkan pengguna lain yang menyukai hal yang sama, dan merekomendasikan hal yang pengguna itu belum pernah beli (Kim Falk, 2018). Untuk membuat rekomendasi, *collaborative filtering* perlu menghubungkan dua entitas yang berbeda secara mendasar: *item* dan pengguna (Ricci, Rokach, Saphira, & Kantor, 2010).

Untuk mencari *item* dan pengguna yang sama dapat dilihat dari *history* transaksi. Semakin banyak transaksi yang ada maka semakin bagus performa hasil rekomendasinya. Jika data transaksi masih sedikit maka tidak dapat diharapkan untuk mendapatkan hasil rekomendasi yang bagus, ini disebut dengan *cold start problem*.

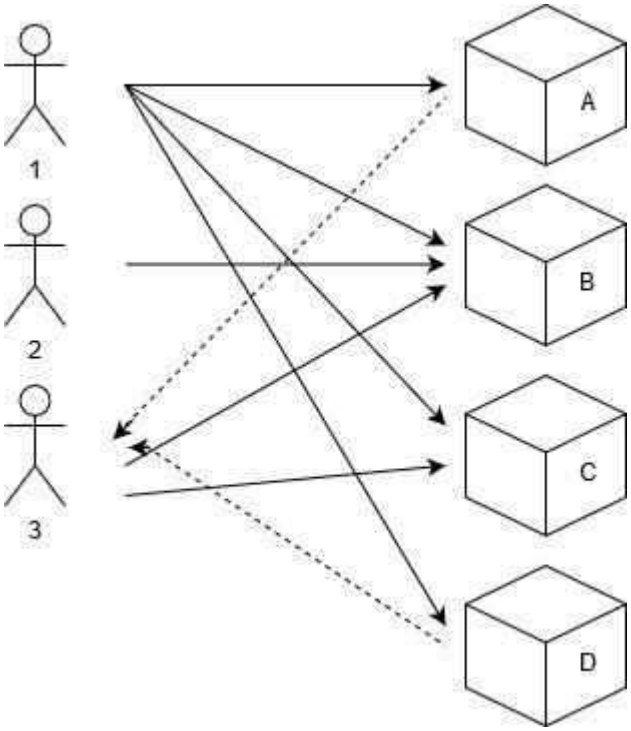
Collaborative filtering ada 2 pendekatan yaitu *user based* dan *item based*, berikutadalah ilustrasi dari kedua pendekatan pada Gambar 2.1.



Gambar 2.2. Metode Pendekatan

a. User-Based Collaborative Filtering

Saat teman meminta rekomendasi film untuk ditonton, dengan asumsi bahwa kami memiliki selera yang sama yang didasari oleh menonton beberapa film yang sama dan memberikan rating yang sama, tetapi teman saya belum menonton film A dan saya sudah. Jika saya menyukai film tersebut maka dapat diasumsikan bahwa teman saya juga menyukai film tersebut. *User-based collaborative filtering* menggunakan metode tersebut dengan merekomendasikan barang dari beberapa user yang mirip ke user meminta rekomendasi. Sedangkan untuk *Item-based Collaborative Filtering,* merupakan metode rekomendasi berdasarkan kemiripan barang.



Gambar 2. 2 Ilustrasi *user-based collaborative filtering*

Gambar 2.2 adalah ilustrasi dari *user-based collaborative filtering*, dimana *user* ke 3 direkomendasikan barang dari *user* 1 karena mereka sama-sama memilih barang C dan B. Sedangkan *user* ke 3 tidak mempunyai barang A dan D seperti *user* pertama.

2.1.2 Similarity

Similarity adalah metode *machine learning* yang menghitung kemiripan antara 2 data atau lebih, menggunakan metode algoritma seperti *pearson correlation, cosine similarity, jaccard similarity,* dan masih banyak lagi. Untuk penelitian ini akan dihitung kemiripan antara *user* melalui data *rating* dari tiap *user* terhadap film. Tujuan *similarity* pada penelitian ini adalah untuk mendapatkan *user* yang mempunyai kemiripan akan film yang sama sehingga dapat direkomendasikan film antara *user* yang mirip.

2.1.3 Pearson *Correlation Coefficient*

Pearson correlation adalah 1 dari 2 metode untuk menghitung persamaan pada data kuantitif, metode lainnya adalah *cosine similarity*. Untuk menghitung persamaan *Pearson* perlu untuk menghitung rata-rata *rating*, menormalisasikan *rating*, dan memasukannya ke dalam rumus (Isinkaye, Folajimi, & Ojokoh, 2015).

$$sim(i, j) = \frac{\sum_{e \in U} (r_{i,u} - \bar{r_i})(r_{j,u} - \bar{r_j})}{\sqrt{\sum_{e \in U} (r_{i,u} - \bar{r_i})^2 (r_{j,u} - \bar{r_j})^2}} \quad (2.1)$$

Dimana:

1) Sim(i,j) adalah angka kemiripan antara *item* atau pengguna.
2) $r_{i,u}$ dan $r_{j,u}$ adalah *rating* pengguna i dan j pada item u pada perhitungan persamaan pengguna.
3) $\bar{r}$ dan $\bar{r}$ adalah rata-rata pengguna i dan j pada perhitungan persamaan pengguna.

2.1.4 Prediksi *Rating*

Rating biasanya disimbolkan dengan bintang, yang digunakan untuk penilai untuk menilai sesuatu seperti film, hotel, restoran, dan masih banyak lagi. *Rating* ini memiliki skala dari 0.5 sampai 5, 5 adalah nilai terbaik dan 0.5 merupakan nilai terendah. Untuk merekomendasikan film ke *user* harus diprediksi *rating* yang akan diberikan *user* ke film yang akan direkomendasikan. Jika hasil prediksi *rating* tinggi maka akan direkomendasikan film tersebut kepada *user* tadi. Untuk memprediksi rating ada rumus yang harus digunakan yaitu (Isinkaye, Folajimi, & Ojokoh, 2015):

$$p(a, i) = \bar{r_a} + \frac{\sum_{i=1}^{n} (r_{u,i} - \bar{r_i}) \times s(a, u)}{\sum_{i=1}^{n} s(a, u)} \quad (2.2)$$

Dimana:

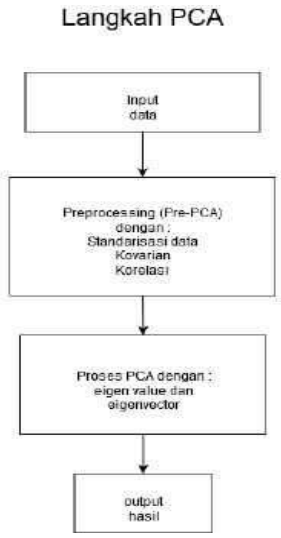
1) p(a,i) adalah hasil prediksi antara *user* a dan *item* i.
2) $\bar{r_a}$ adalah rata-rata *rating user.*
3) $r_{u,i}$ adalah *rating user* u kepada *item* i.
4) $\bar{r_u}$ adalah rata-rata *rating user* terdekat.
5) $s(a,u)$ adalah nilai *similarity* antara *user* a *dan item* i.

2.1.5 Dimensionality Reduction

Pada penelitian ini, data disajikan dalam bentuk matriks, dan matriks tersebut dapat diperkecil ukuran matriks tanpa kehilangan inti dari matriks tersebut. Sebuah matriks yang diperkecil ukurannya hanya dapat dilakukan pada bagian kolom atau baris saja. Proses memperkecil matriks ini disebut *dimensionality reduction.* Beberapa metode dari reduksi adalah PCA, Kernel PCA, Isomap, dan masih banyak metode reduksi dimensi lainnya.

2.1.6 Principal Component Analysis

Pada penelitian ini digunakan metode reduksi dimensi yaitu PCA. PCA mengambil beberapa variabel dan mengurangi mereka menjadi 1 atau lebih komponen yang mewakili varian variable tanpa kehilangan informasi penting (Retno Mayapada, 2019). PCA adalah teknik statistik yang melakukan pengurangan relasi variabel (korelasi / kovarian) menjadi beberapa komponen (dimensi) baru. PCA menggunakan korelasi matriks atau varian-kovarian matriks untuk mendapatkan hasil dari kombinasi linear.



Gambar 2. 3 Langkah PCA

Tujuan dari PCA adalah untuk mereduksi variabel data tanpa kehilangan informasi aslinya, variabel sebanyak n direduksi menjadi sebanyak k yang lebih sedikit dari n dan mempunyai nilai yang sama dengan n. Variabel hasil reduksi disebut sebagai *principal component.*

2.1.7 K-Means Clustering

K-means merupakan algoritma *clustering*. ***K-means Clustering*** adalah salah satu "*unsupervised machine learning algorithms*" yang paling sederhana dan populer. *K-Means Clustering* adalah suatu metode penganalisaan data atau metode *Data Mining* yang melakukan proses pemodelan tanpa supervisi (*unsupervised*) dan merupakan salah satu metode yang melakukan pengelompokan data dengan sistem partisi.

K-means clustering merupakan salah satu metode *cluster* analysis non hirarki yang berusaha untuk mempartisi objek yang ada kedalam satu atau lebih *cluster* atau kelompok objek berdasarkan karakteristiknya, sehingga objek yang mempunyai karakteristik yang sama dikelompokan dalam satu cluster yang sama dan objek yang mempunyai karakteristik yang berbeda dikelompokan ke dalam *cluster* yang lain.

Metode *K-Means Clustering* berusaha mengelompokkan data yang ada ke dalam beberapa kelompok, dimana data dalam satu kelompok mempunyai karakteristik yang sama satu sama lainnya dan mempunyai karakteristik yang berbeda dengan data yang ada di dalam kelompok yang lain.

Dengan kata lain, metode *K-Means Clustering* bertujuan untuk meminimalisasikan *objective function* yang diset dalam proses *clustering* dengan cara meminimalkan variasi antar data yang ada di dalam suatu cluster dan memaksimalkan variasi dengan data yang ada di *cluster* lainnya juga bertujuan untuk menemukan grup

dalam data, dengan jumlah grup yang diwakili oleh variabel K. Variabel K sendiri adalah jumlah *cluster* yang diinginkan. Membagi data menjadi beberapa kelompok. Algoritma ini menerima masukan berupa data tanpa label kelas. Hal ini berbeda dengan *supervised learning* yang menerima masukan berupa vektor (*x1 , y1*) , (*x2 , y2*) , …, (*xi , yi*), di mana *xi* merupakan data dari suatu data pelatihan dan *yi* merupakan label kelas untuk *xi* .

Pada algoritma pembelajaran ini, komputer mengelompokkan sendiri data-data yang menjadi masukannya tanpa mengetahui terlebih dulu target kelasnya. Pembelajaran ini termasuk dalam *unsupervised learning.* Masukan yang diterima adalah data atau objek dan *k* buah kelompok (*cluster*) yang diinginkan. Algoritma ini akan mengelompokkan data atau objek ke dalam *k* buah kelompok tersebut. Pada setiap *cluster* terdapat titik pusat (*centroid*) yang merepresentasikan *cluster* tersebut.

K-means ditemukan oleh beberapa orang yaitu Lloyd (1957, 1982), Forgey (1965), Friedman and Rubin (1967), and McQueen (1967). Ide dari *clustering* pertama kali ditemukan oleh Lloyd pada tahun 1957, namun hal tersebut baru dipublikasi pada tahun 1982. Pada tahun 1965, Forgey juga mempublikasi teknik yang sama sehingga terkadang dikenal sebagai Lloyd-Forgy pada beberapa sumber.

Terdapat dua jenis data clustering yang sering dipergunakan dalam proses pengelompokan data yaitu *Hierarchical* dan *Non-Hierarchical*, dan *K-Means* merupakan salah satu metode data *clustering non-hierarchical* atau *Partitional Clustering*.

Data clustering menggunakan metode *K-Means Clustering* ini secara umum dilakukan dengan algoritma dasar sebagai berikut:

1. Tentukan jumlah *cluster*

2. Alokasikan data ke dalam *cluster* secara *random*
3. Hitung *centroid*/rata-rata dari data yang ada di masing-masing *cluster*
4. Alokasikan masing-masing data ke centroid/rata-rata terdekat
5. Kembali ke Step 3, apabila masih ada data yang berpindah cluster atau apabila perubahan nilai *centroid*, ada yang di atas nilai *threshold* yang ditentukan atau apabila perubahan nilai pada *objective function* yang digunakan di atas nilai *threshold* yang ditentukan

Secara umum metode K-Means Cluster menggunakan algoritma sebagai berikut:

- Menentukan k sebagai jumlah cluster yang ingin dibentuk.
- Membangkitkan nilai random untuk pusat cluster awal (centroid) sebanyak k.
- Menghitung jarak setiap data input terhadap masingmasing centroid menggunakan rumus Euclidean Distance. Euclidean Distance adalah perhitungan jarak antar dua buah titik dalam ruang Euclidean (Euclidean Space). Pada ruang 2 dimensi yang melibatkan 2 titik (misal: titik dan titik), maka perhitungan Euclidean Distance menjadi sebagai berikut:

$$d(x,y) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

- Alokasikan masing-masing objek ke dalam centroid yang paling terdekat.
- Penentuan centroid dilakukan secara acak dari objek-objek yang tersedia sebanyak k cluster, kemudian menghitung centroid cluster ke-i berikutnya diperoleh dari rata-rata cluster yang bersangkutan dengan menggunakan rumus:

$$C = \frac{1}{p}\sum_{i=1}^{p} x_i, i = 1,2,3,4..p$$

dimana:

C : *centroid* pada *cluster*
x : data variabel lokasi ke– i (i = 1,2,3, … , p)
p : banyaknya objek/ jumlah anggota yang menjadi *cluster*

- Lakukan iterasi hingga anggota tiap cluster tidak ada yang berubah.

- Apabila anggota tiap cluster tidak ada yang berubah, maka nilai rata-rata pusat cluster pada iterasi terakhir sudah konvergen

2.1.8 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*

2.1.9 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 =

t_x = C*luster* (x) pada

y_i = perulangan ke (i)

Titik tengah
cluster (x)

Cluster (y)
pada perulangan
ke (i) t_y = Titik
tengah *cluster* (y)

2.2. 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.

BAB III
METODE PENELITIAN

Penelitian yang diajukan terdiri atas beberapa tahapan, yaitu pengumpulan data judul film, rating film, dan *user* dalam bentuk format data .csv. 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

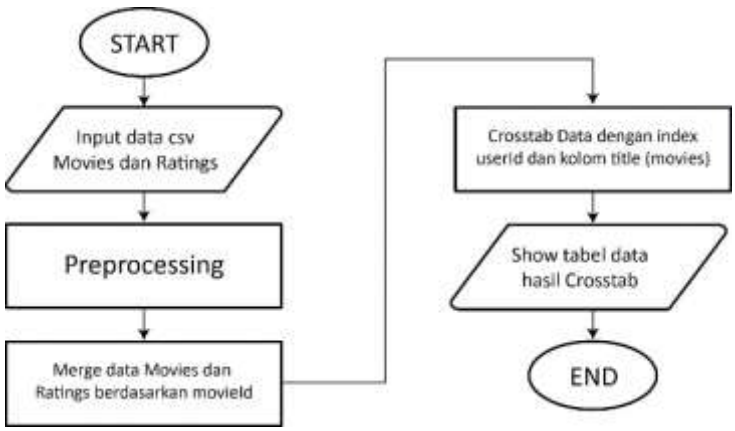
```
Shape: (100836, 4)
```

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

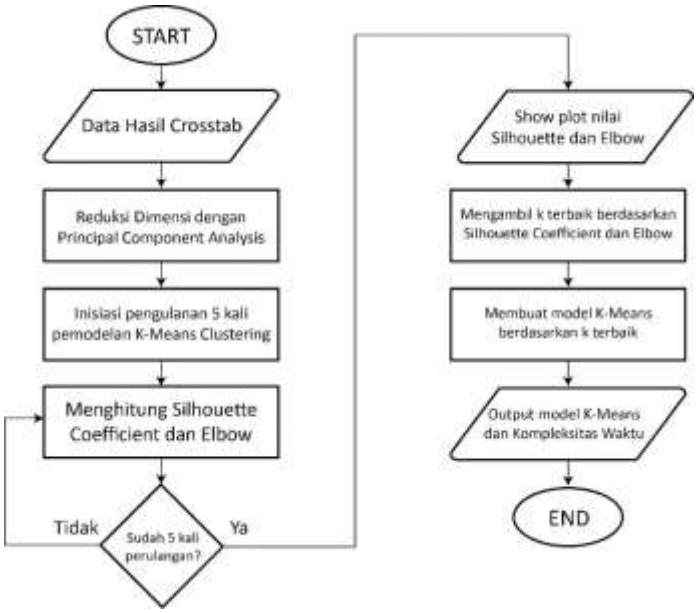
Pada pemrosesan data *movie* dan (*user*) *ratings* yang akan menghasilkan himpunan rekomendasi kepada *user*, maka perlu dilakukan proses persiapan terlebih dahulu agar data yang digunakan menjadi data yang baik. Proses yang terjadi yaitu *preprocessing* dimana data yang tersedia masih ada beberapa 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)

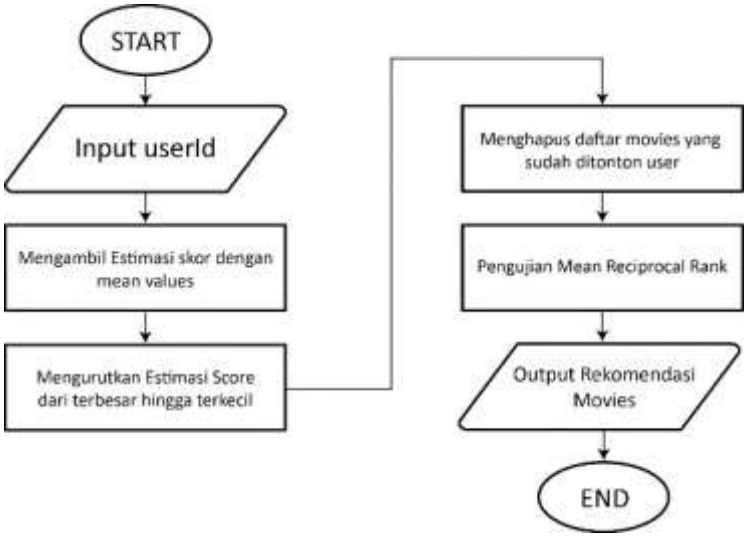
Setelah dilakukan proses persiapan, maka data siap melakukan proses clustering. Pada penelitian ini proses *clustering* terlebih dahulu dilakukan

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)

3.3. Mereduksi dengan Principal Component Analysis (PCA)

PCA adalah teknik statistik yang melakukan pengurangan relasi variabel (korelasi / kovarian) menjadi beberapa komponen (dimensi). PCA menggunakan korelasi matriks atau varian-kovarian matriks untuk mendapatkan hasil dari kombinasi linear. Diketahui tabel sebagai berikut:

Tabel 3. 3 Contoh Data

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2.0	1.6
1.0	1.1
1.5	1.6
1.1	0.9

X dan Y merupakan 2 variabel sembarang dan akan dicarivektor rata-ratasebagai berikut :

$$\bar{x} = \frac{2.5 + 0.5 + 2.2 + 1.9 + 3.1 + 2.3 + 2.2 + 1.0 + 1.5 + 1.1}{}$$

$$= 1.81$$

10

$$\bar{Y} = \frac{2.4 + 0.7 + 2.9 + 2.2 + 3.0 + 2.7 + 1.6 + 1.1 + 1.6 + 0.9}{10} = 1.$$

Akan dicari variansi dan kovariansi dari data tersebut :

S_{11}

$$= \frac{(2.5 - 1.81)^2 + (0.5 - 1.81)^2 + (2.2 - 1.81)^2 + (1.9 - 1.81)^2 + (3.1 - 1.81)^2 + (2.3 - 1.81)^2 + (2.0 - 1.81)^2 + (1.0 - 1.81)^2 + (1.5 - 1.81)^2 + (1.1 - 1.81)^2}{10 - 1}$$

$= 0.616$

$S_{12} = S_{21} =$

$$= \frac{(2.5-1.81)\times(2.4-1.91)\times(0.5-1.81)\times(0.7-1.91)\times(2.2-1.81)\times(2.9-1.91)\times(1.9-1.81)\times(2.2-1.91)\times(3.1-1.81)\times(3.0-1.91)\times(2.3-1.81)\times(2.7-1.91)\times(2.0-1.81)\times(1.6-1.91)\times(1.1-1.81)\times(0.9-1.91}{10-1} = 0.615$$

S_{22}

$$= \frac{(2.4 - 1.91)^2 + (0.7 - 1.91)^2 + (2.9 - 1.91)^2 + (2.2 - 1.91)^2 + (3.0 - 1.91)^2 + (2.7 - 1.91)^2 + (1.6 - 1.91)^2 + (1.1 - 1.91)^2 + (1.6 - 1.91)^2 + (0.9 - 1.91)^2}{10 - 1}$$

$= 0.716$

Dari hasil diatas dapat diperoleh matriks kovarians sebagai berikut:

$$S = \begin{pmatrix} 0.616 - \lambda & 0.615 \\ 0.615 & 0.716 - \lambda \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = 0$$

Akan dicari nilai eigen sebagai berikut :

$$\det(S) = (0.616 - \lambda)(0.716 - \lambda) - (0.615)(0.615) = 0$$

$$= 0.441 - 1.332\lambda + \lambda^2 - 0.378 = 0$$

$$= \lambda^2 - 1.332\lambda + 0.063 = 0$$

Dari persamaan diatas, diperoleh: $a = 1, b = -1.332, \ c = 0.063$

$$\lambda_{1.2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

$$= \frac{1.332 \pm \sqrt{1.332^2 - 4.1.0.063}}{2.1}$$

$$= \frac{1.332 \pm \sqrt{1.522}}{2}$$

$$\lambda_1 = \frac{1.332 + \sqrt{1.522}}{2} = 1.282$$

$$\lambda_2 = \frac{1.332 - \sqrt{1.522}}{2} = 0.049$$

Setelah menemukan λ_1 dan λ_2,lalu masukkan ke dalam matriks S untuk dicari *eigen vector:*

Untuk:

$$\lambda_1: 1.282$$

$$R1 = \begin{pmatrix} 0.616 - 1.282 & 0.615 \\ 0.615 & 0.716 - 1.282 \end{pmatrix} = \begin{pmatrix} -0.666 & 0.615 \\ 0.615 & -566 \end{pmatrix}$$

$$\begin{pmatrix} -0.666 & 0.615 \\ 0.615 & -566 \end{pmatrix} \begin{pmatrix} x_1 \\ 2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$= \begin{pmatrix} -0.666 & 0.615 \\ 0 & 0 \end{pmatrix}$$

$$= -0.666x_1 + 0.615x_2$$

Misalkan $x_1 = 1$, maka diperoleh x_2 sebagai berikut:

$$-0.666(1) + 0.615x_2 = 0$$

$$x_2 = 1.082$$

Sehingga: $\begin{matrix} x_1 \\ * \\ x_2 \end{matrix} \begin{matrix} 1 \\ + = * \\ 1.082 \end{matrix} +$

Untuk

$$\lambda_2: 0.049$$

$$R2 = \begin{pmatrix} 0.616 - 0.049 & 0.615 \\ 0.615 & 0.716 - 0.049 \end{pmatrix} = \begin{pmatrix} 0.567 & 0.615 \\ 0.615 & 0.667 \end{pmatrix}$$

$$= \begin{pmatrix} 0.567 & 0.615 \\ 0.615 & 0.667 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$= \begin{pmatrix} 0 & 0 \\ 0.615 & 0.667 \end{pmatrix}$$

$$= 0.615x_1 + 0.667x_2$$

Misalkan $x_1 = 1$, maka diperoleh x_2 sebagai berikut:

$$-0.615(1) + 0.667x_2 = 0$$

$$x_2 = 0.922$$

Sehingga: $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0.922 \end{bmatrix}$

Jadi dari perhitungan diatas diperoleh vektor eigen sebagai berikut:

$$\begin{pmatrix} 1 & 1 \\ 1.082 & 0.922 \end{pmatrix}$$

Dimana elemen pada vektor eigen di atas merupakan koefisien komponenutama pada PCA.

Setelah menemukan λ_1 dan λ_2 maka selanjutnya adalah mencari varian darikedua komponen tersebut:

$$var_1 = \frac{\lambda_1}{n-1}$$
$$= \frac{1.282}{9} = 0.142$$
$$var_2 = \frac{\lambda_2}{n-1}$$
$$= \frac{0.049}{9} = 0.005$$

Lalu untuk mendapatkan PCA menggunakan rumus berikut ini:

$$PCA = \frac{var_{1.2}}{var\ total} \times 100\%$$

$$PCA_1 = \frac{0.142}{0.147} \times 100\% = 96{,}5\%$$
$$PCA_2 = \frac{0.005}{0.147} \times 100\% = 3{,}4\%$$

Jadi variabel baru PCA_1 dapat mewakilkan seluruh data awal sebesar 96.5%.

3.4. Pengelompokan dengan K-Means Clustering

Table 1. Nilai Centroid Awal

	Nilai *Centroid* Awal	
	genre	*rating*
cluster 1	5	3
cluster 2	2	3

Table 2. Data Iterasi 1

No	Data film			Hasil Iterasi 1			
	NamaFilm	*Genre*	*Rating*	C1	C2	Jarak Terpendek	*Cluster*
1	film 1	5	3	2.5	2.4	0.00	C1
2	film 2	2	3	0.5	0.7	0.00	C2
3	film 3	2	1	2.2	2.9	2.00	C2
4	film 4	5	4	1.9	2.2	1.00	C1
5	film 5	5	5	3.1	3.0	2.00	C1
6	film 6	4	5	2.3	2.7	2.24	C1
7	film 7	2	1	2.0	1.6	2.00	C2
8	film 8	3	1	1.0	1.1	2.24	C2
9	film 9	3	1	1.5	1.6	2.24	C2
10	film 10	3	2	1.1	0.9	1.41	C2

Table 3. Nilai Centroid Baru

Nilai *Centroid* 2		
	genre	*rating*
cluster 1	5	3
cluster 2	2	2

Table 4. Data Iterasi 2

Data film			Hasil Iterasi 2			
Nama Film	*Genre*	*Rating*	C1	C2	Jarak Terpendek	*Cluster*
film 1	5	3	0.61	3.02	0.61	C1
film 2	2	3	2.61	1.27	1.27	C2
film 3	2	1	3.54	0.79	0.79	C2
film 4	5	4	0.71	3.55	0.71	C1
film 5	5	5	1.63	4.26	1.63	C1
film 6	4	5	1.67	3.69	1.67	C1
film 7	2	1	3.54	0.79	0.79	C2
film 8	3	1	2.89	1.06	1.06	C2
film 9	3	1	2.89	1.06	1.06	C2
film 10	3	2	2.12	0.79	0.79	C2

BAB IV
HASIL DAN ANALISIS

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 kondisimalam 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.

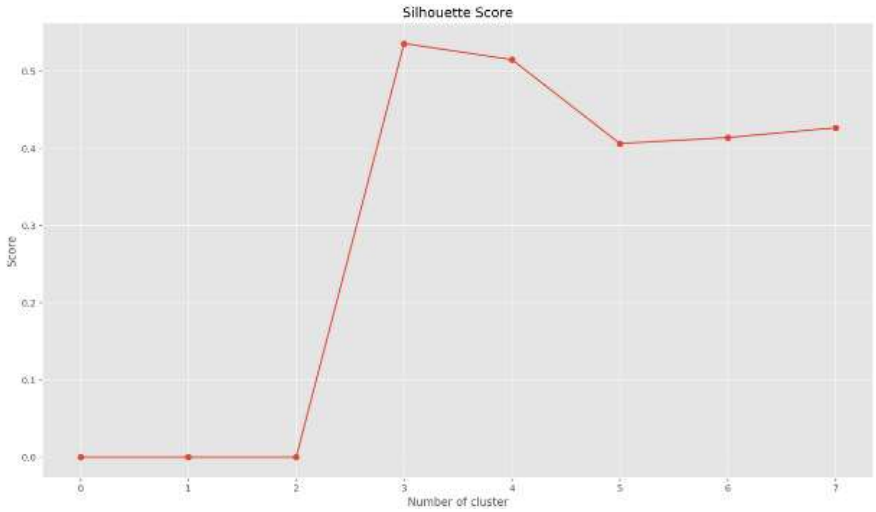
Tabel 4.1. Hasil pengujian Kompleksitas Waktu

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

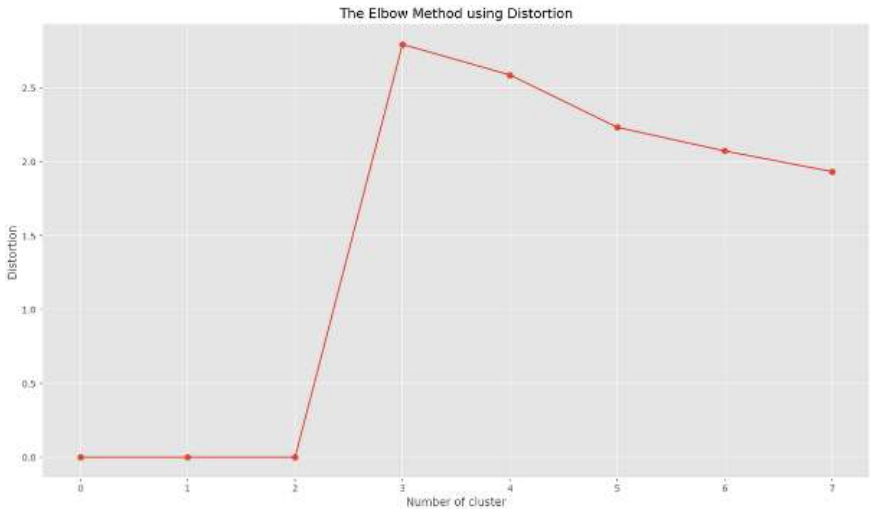
data-data yang berada pada cluster-cluster tersebut juga sudah sesuai. Grafik perbandingan nilai *Silhouette Coefficient* tiap *k*-nya dapat dilihat pada gambar 4.1.



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

BAB V
KESIMPULAN DAN SARAN

Pada bab 5 ini akan dipaparkan kesimpulan dari keseluruhan hasil penelitian dan juga saran-saran untuk penelitian lebih lanjut.

5.1 Kesimpulan

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 yangdisimpulkan bahwa rekomendasi yang dihasilkan kurang tepat.

5.2 Saran

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.

DAFTAR PUSTAKA

Camps-Valls, G., *et al*. "Composite Kernels for Hyperspectral Image Classification". IEEE. Geoscience and Remote Sensing Letters 3.1 (2006): 93 – 97.

Freund, Y. dan Robert E. Sapire. "A short introduction to boosting". Journal of Japanese Society for Artificial Intelligence 14.5 (1999): 771-780.

Habiburrahman. "Modifikasi Pembangunan Minimum Spanning Forest Pada Segmentasi Citra Hyperspectral Dalam Domain Spektral-Spasial". Tesis. Universitas Indonesia, Depok. Juli 2014.

Kuncheva, L. I. (2004). *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience.

K'egl, Bal'azs. "The return of ADABOOST.MH: multi-class hamming trees". In International Conference on Learning Representations (2013).

Li, Wei. "Local Binary Patterns and Extreme Learning Machine for Hyperspectral Imagery Classification". IEEE. Transactions on Geoscience and Remote Sensing 53.7 (2015): 3681-3693.

Lodha, Suresh K., et al. "Aerial lidar data classification using Support Vector Machines (SVM)". Third International Symposium on 3D Data Processing, Visualization, and Transmission (2006): 567 - 574.

Lodha, Suresh K., Darren M. Fitzpatrick, dan David P. Helmbold. "Aerial LiDAR data classification using Expectation-Maximization". In Proceedings of the SPIE Conference on Vision Geometry 14 (2007).

Lodha, Suresh K., Darren M. Fitzpatrick, dan David P. Helmbold. "Aerial LiDAR data classification using AdaBoost". IEEE. Sixth International Conference on 3-D Digital Imaging and Modeling (2007): 435-442.

Menzata, Remmy A. “Perbandingan metode *maximum likelihood*, *random forest*, dan *neural network* untuk klasifikasi pada fusi citra LiDAR dengan *aerial optical images*”. Skripsi. Universitas Indonesia, Depok. Juli 2013.

Moser, G., *et al*. “Combining Support Vector Machines and Markov Random Fields in an Integrated Framework for Contextual Image Classification”. IEEE. Transactions on Geoscience and Remote Sensing 51.5 (2013): 2734 – 2752.

Rokach, Lior. *Ensemble*-based *classifier*s. Artificial Intelligence Review - Springer 33:1–39 (2010).