

BAB II

TINJAUAN PUSTAKA

2.1. *Big Data*

Big data merupakan gabungan dari berbagai jenis data, baik yang tidak terstruktur, semi terstruktur, maupun terstruktur (Arief & Saputra, 2019). *Big data* memiliki karakteristik 5V, yaitu *Volume*, *Variety*, *Velocity*, *Veracity* dan *Value* (Aldisa et al., 2022). *Volume* diartikan sebagai jumlah besar data yang dihasilkan setiap detik dari media sosial, dan lain-lain. *Variety* dapat diartikan juga sebagai keragaman data. *Velocity* mengacu pada kecepatan data seperti seberapa cepat data diproses dan dihasilkan untuk memenuhi persyaratan. *Veracity* mengacu pada tingkat akurasi data dan *Value* mengacu pada pengambilan keputusan berdasarkan berbagai data yang ada.

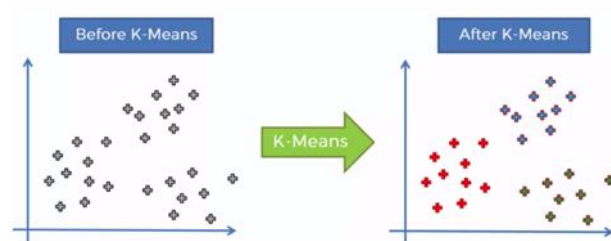
2.2. *Clustering*

Clustering adalah pengelompokan atau klasifikasi suatu data yang sama untuk menjadi suatu kelompok atau *cluster* yang berbeda untuk menjadikan suatu kelompok yang baru (Darmi & Setiawan, 2017). *Clustering* sangat berguna untuk menemukan group atau kelompok yang tidak dikenal dalam suatu data dan untuk mengurangi jumlah data yang besar dengan memberikan kategori atau kelompok data yang memiliki kesamaan yang tinggi (Edy Irwansyah, 2015). Metode *clustering* secara umum dibagi menjadi dua, yaitu *hierarchical clustering* dan *non-hierarchical clustering*. *Hierarchical clustering* merupakan suatu metode pengelompokan data yang dimulai dengan mengelompokkan dua atau lebih objek yang memiliki kesamaan paling dekat. Proses itu dilakukan berulang kali hingga *cluster* akan membentuk semacam pohon yang memiliki *hierarchi* (tingkatan) yang jelas antar objek. Sedangkan *non-hierarchical clustering* merupakan suatu metode pengelompokan data yang dimulai dengan menentukan jumlah cluster terlebih dahulu, setelah itu akan dilakukan proses *clustering* tanpa mengikuti

hierarchi. Dalam penelitian ini metode *clustering* yang digunakan adalah dengan menggunakan metode *K-means Clustering*.

2.2.1 *K-Means Clustering*

K-means Clustering merupakan salah satu algoritma klaster non hirarki. Algoritma klasterisasi ini akan mengelompokkan sebuah data dalam beberapa kelompok untuk memudahkan dalam mengambil suatu keputusan (Rahmah & Antares, 2022). Algoritma ini juga merupakan metode *unsupervised learning* dan merupakan salah satu metode yang melakukan pengelompokan data dengan sistem partisi. Data akan dikelompokkan didalam *cluster* yang memiliki karakteristik yang sama, sehingga data yang sudah dikelompokkan memiliki variasi bentuk yang kecil (Darmawan, 2022). Pembagian kelompok pada suatu data tersebut dilakukan dengan cara melakukan iterasi data dan mencari jarak tiap datanya menggunakan *euclidean distance*. Berikut merupakan ilustrasi pengelompokan *K-means Clustering* yang dapat dilihat pada gambar 2.1.



Gambar 2.1 *K-Means Clustering*

Adapun langkah-langkah dalam penggunaan algoritma *K-means Clustering*, yaitu sebagai berikut (Fernando et al., 2022):

1. Menentukan jumlah *cluster* (*K*) yang akan dibentuk.
2. Menentukan titik awal dari setiap *cluster*.

3. Hitung jarak setiap data input ke setiap *centroid* menggunakan rumus jarak *Euclidean Distance* hingga ditemukan jarak terdekat dari setiap data dengan *centroid*. Berikut adalah persamaan Jarak *Euclidean*:

$$D_{(i,j)} = \sqrt{(X_{1i} - X_{1j})^2 + \dots + (X_{ki} - X_{kj})^2} \quad (1)$$

Keterangan:

$D_{(i,j)}$: *Euclidean distance* / Jarak data i ke pusat *centroid* j

X_{ki} : Data ke- i pada atribut k

X_{kj} : *Centroid* ke- j pada atribut data k

4. Hasil perhitungan jarak akan dibandingkan dengan tiap *centroid*, data yang memiliki jarak paling dekat, akan dikelompokkan menjadi satu kelas.
5. Memperbaharui nilai *centroid*, nilai *centroid* baru yang diperoleh dari rata-rata *cluster* yang baru saja terbentuk.
6. Ulangi proses hitung jarak pada pengelompokan kelas baru hingga nilai *centroid* tidak berubah lagi dan *cluster* akhir dapat ditentukan.
7. Setelah selesai menentukan *cluster* dari setiap agregat konsentrasi, seluruh hasil klasterisasi dimasukkan kedalam tabel keputusan untuk menjadi dasar keputusan rekomendasi.

Selain itu metode K-means Clustering ini memiliki kelebihan dan kekurangan yang dapat di lihat pada tabel 2.1 berikut:

Tabel 2.1 Kelebihan dan Kekurangan K-Means Clustering

Kelebihan	Kekurangan
Relatif sederhana dan mudah untuk diterapkan.	Perlu menentukan nilai k secara manual
Dapat diskalakan untuk dataset dalam jumlah besar.	Sangat bergantung pada inisialisasi awal. Jika nilai random untuk inisialisasi kurang baik, maka pengelompokan yang dihasilkan

	pun menjadi kurang optimal.
Mudah beradaptasi dengan contoh baru.	Apabila hasil nilai random untuk centroid kurang baik, maka hasil clustering yang didapatkan menjadi tidak optimal
Sangat umum digunakan sebagai teknik clustering.	Cukup sulit jika digunakan untuk mencari jarak dari data yang berdimensi banyak.

2.3. *Davies-Bouldin Index (DBI)*

Davies-Bouldin Index (DBI) adalah suatu ukuran yang digunakan untuk menentukan jumlah kluster yang terbaik setelah proses pengklasteran selesai (Rifai et al., 2022). Pendekatan DBI ini bertujuan untuk memaksimalkan jarak antara kluster yang satu dengan kluster yang lain dan mencoba meminimalkan jarak antar objek dalam suatu kluster. Langkah-langkah untuk menghitung nilai *Davies Bouldin Index* adalah sebagai berikut (Butsianto & Saepudin, 2019) :

1. *Sum of Square Within cluster (SSW)*, rumus yang digunakan untuk mengetahui kohesi dalam sebuah *cluster* ke-*i* seperti persamaan berikut:

$$SSWi = d(X_i, C_j) \quad (2)$$

Keterangan :

$d(X_i, C_j)$ = jarak setiap data ke centroid *i* yang dihitung menggunakan jarak euclidean.

2. *Sum of Square Between cluster (SSB)* bertujuan untuk mengetahui separasi atau jarak antar cluster seperti berikut:

$$SSB_{i,j} = d(X_i, X_j) \quad (3)$$

Keterangan :

$d(X_i, X_j)$ = jarak antara data ke *i* dengan data ke *j* di cluster lain

3. Ratio (Rasio), perhitungan rasio ($R_{i,j}$) ini bertujuan untuk mengetahui nilai perbandingan antara cluster ke-i dan cluster ke-j untuk menghitung nilai rasio yang dimiliki oleh masing-masing cluster.

$$R_{i,j} = \frac{SSW_i + SSW_j + \dots + SSW_n}{SSB_{i,j} + \dots + SSB_{ni,nj}} \quad (4)$$

Keterangan :

SSW_i = Sum Of Square Within- Cluster pada centroid i

$SSB_{i,j}$ = Sum of Square Between Cluster data ke i dengan j pada cluster yang berbeda

4. Nilai-nilai *Davies-Bouldin Index* (DBI) didapatkan dari persamaan sebagai berikut:

$$DBI = \frac{1}{K} \times \sum_{i=1}^K \max_{i \neq j} (R_{i,j} \dots k) \quad (5)$$

Dimana, $R_{i,j}$ merupakan ratio dari nilai SSW dan SSB dan K adalah jumlah *cluster* yang digunakan. jika semakin kecil nilai Davies Bouldin Index (DBI) yang diperoleh (non negatif ≥ 0) maka cluster tersebut semakin baik.

2.4. Sistem Rekomendasi

Sistem rekomendasi adalah sistem yang membantu pengguna dalam dan diharapkan rekomendasi ini akan memenuhi keinginan dan kebutuhan pengguna (Jaja et al., 2020). Sistem rekomendasi banyak digunakan untuk membuat prediksi seperti buku, musik, film, tempat wisata dan lain-lain. Sebagai contoh sistem rekomendasi digunakan untuk pemilihan film yang nantinya dapat membantu pengguna memilih mengatasi informasi yang meluap dengan memberikan rekomendasi kepada pengguna film yang akan ditonton. Diperlukan model rekomendasi yang tepat agar rekomendasi yang diberikan sesuai dengan keinginan pengguna, serta mempermudah pengguna mengambil keputusan yang tepat dalam menentukan film yang akan dipilih. Terdapat tiga jenis sistem rekomendasi berdasarkan metode yang digunakannya yakni, *collaborative filtering*, *content-based filtering*, dan

hybrid. Pada penelitian ini menggunakan pendekatan *collaborative filtering* (*user based*).

2.5. Collaborative Filtering

Collaborative filtering adalah proses penyaringan atau penilaian item menggunakan pendapat orang lain. Metode ini menghasilkan rekomendasi untuk pengguna yang menargetkan satu atau beberapa item. Konsep ini muncul dalam bentuk anggapan bahwa jika seseorang menyukai suatu barang, maka barang tersebut juga akan disukai oleh orang lain. *Collaborative Filtering* ini dibagi menjadi dua kelas yaitu *user-based* dan *item-based*.

1. User-Based Collaborative Filtering

User-Based Collaborative Filtering mengasumsikan teknik statistik untuk menemukan hubungan antara pengguna yang berjarak dekat berdasarkan produk yang sedang dievaluasi. Pada User-Based, rekomendasi memanfaatkan penilaian User. Sistem akan menemukan pengguna dengan korelasi erat yang kuat dan kemudian merekomendasikan produk yang tersedia (Khusna et al., 2021).

2. Item-Based Collaborative Filtering

Item-Based Collaborative Filtering menggunakan item yang sudah diberi peringkat oleh pengguna sebagai dasar perhitungan. Metode ini menghitung kesamaan item yang sudah diberi peringkat dengan item lain, lalu memilih sekelompok item yang diberi peringkat memiliki nilai yang sama dengan elemen yang dievaluasi. Nilai kesamaan dicatat dan kemudian digunakan sebagai nilai bobot untuk memprediksi nilai rating item yang ditargetkan (Jaja et al., 2020).

Algoritma yang digunakan untuk memberikan rekomendasi berdasarkan kemiripan user adalah cosine similarity. Adapun rumusnya adalah sebagai berikut :

$$\text{sim}(x, y) = \cos(x, y) = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \sqrt{\sum_{i=1}^n y_i^2}} \quad (6)$$

Keterangan :

$\text{sim}(x, y)$ = kesamaan antara x dan y

x, y = merupakan dua vector

n = mewakili vektor dimensi

2.6. *Movielens*

Movielens adalah situs rekomendasi film personal yang disusun berdasarkan rating yang diberikan pengguna terhadap sebuah film. Dataset *movielens* berisi susunan rating/peringkat yang diberikan pengguna terhadap film yang diminati. *Movielens* yang digunakan adalah data 100K yang berisi dari 943 pengguna dan 1682 film yang telah dirating sebanyak 100.000 kali yang dikumpulkan selama tujuh bulan dari 19 September 1997 hingga 22 April 1998. Data ini diambil dari web *kaggle*.

2.7. *Data Preprocessing*

Data *preprocessing* adalah proses mengubah data mentah menjadi format yang lebih mudah dipahami. Proses ini penting karena data mentah seringkali tidak memiliki bentuk yang teratur. *Preprocessing* biasanya dilakukan dengan mengeliminasi data yang tidak sesuai. Adapun tahapan *preprocessing* data yang digunakan dalam penelitian ini adalah sebagai berikut (Zai, 2022) :

1. *Data Cleaning*

Dalam tahap *cleaning*, data mentah akan dibersihkan dimana data-data tidak lengkap, mengandung *error* dan tidak konsisten dibuang dari koleksi data, sehingga data yang telah bersih relevan dapat digunakan untuk diproses ulang untuk penggalian pengetahuan (*discovery knowledge*).

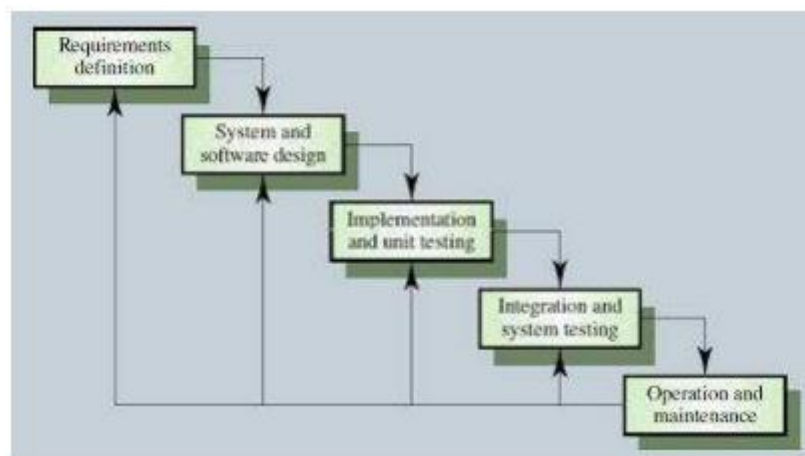
2. *Data Transformation*

Pada tahap ini, data-data yang telah terpilih, ditransformasikan kedalam bentuk-bentuk yang cocok untuk prosedur penggalian (*meaning*

procedure) dengan cara melakukan normalisasi (mengubah data ke dalam skala yang teratur).

2.8. Pengembangan Perangkat Lunak *Waterfall*

Metode pengembangan perangkat lunak sistematis dan sekuensial. Disebut juga “*Classic Life Cycle*”. Disebut *waterfall* (berarti air terjun) karena memang diagram tahapan prosesnya mirip dengan air terjun yang bertingkat (Nawawi, 2017), berikut tahapan *waterfall* yang dapat dilihat pada gambar 2.2



Gambar 2.2 Metode *Classic Life Cycle*

Tahapan dalam melakukan metode *Waterfall* adalah:

1. *Requirements analysis and definition*

Dalam tahap ini, mengumpulkan informasi data secara lengkap. Kemudian dianalisis dan didefinisikan kebutuhan yang harus dipenuhi oleh program yang akan dibangun.

2. *System and software design*

Pada tahap ini, desain dikerjakan setelah kebutuhan selesai dikumpulkan secara lengkap. Tujuan perencanaan desain adalah untuk memberikan gambaran lengkap tentang apa yang perlu dilakukan.

3. *Implementation and unit testing*

Desain program diartikan ke dalam kode-kode dengan menggunakan bahasa pemrograman yang sudah ditentukan. Setelah itu, program yang dibangun langsung diuji.

4. *Integration and system testing*

Penyatuan unit-unit program, kemudian diuji secara keseluruhan (*system testing*) untuk mengidentifikasi kemungkinan adanya kegagalan dan kesalahan sistem.

5. *Operation and maintenance*

Pada tahap terakhir, mengoperasikan program dilingkungannya dan melakukan pemeliharaan. Pemeliharaan memungkinkan pengembang untuk melakukan perbaikan atas kesalahan yang tidak terdeteksi pada tahap-tahap sebelumnya dan peningkatan dan penyesuaian sistem sesuai dengan kebutuhan. Metode pengembangan perangkat lunak *Waterfall* juga memiliki kelebihan dan kekurangan yang dapat dilihat pada tabel 2.2.

Tabel 2.2 Kelebihan dan Kekurangan Metode Waterfall

Kelebihan	Kekurangan
Software yang dikembangkan menghasilkan kualitas yang baik	Sulit untuk kembali lagi dan mengubah sesuatu yang tidak terdokumentasi dengan baik dalam tahap konsep sebelumnya
Dokumen pengembangan sistem terorganisasi dengan baik, karena setiap fase harus diselesaikan sepenuhnya sebelum melanjutkan ke fase berikutnya	Waktu pengerjaan relatif lebih lama, karena harus menunggu tahap sebelumnya selesai
Dinilai sangat cocok untuk menjalankan pembuatan aplikasi berskala besar yang melibatkan banyak sumber daya manusia dan prosedur kerja yang kompleks.	Model ini cocok ketika kebutuhan dikumpulkan secara lengkap sehingga perubahan bisa diminimalkan. Tapi nyatanya jarang sekali konsumen/pengguna yang bisa memberikan kebutuhan secara lengkap, perubahan kebutuhan adalah sesuatu yang wajar terjadi

2.9. **Black-Box Testing**

Blackbox testing adalah metode pengujian yang berfokus pada spesifikasi fungsional perangkat lunak (A. Naufal et al., 2022). Pengujian yang dilakukan untuk mengetahui apakah fungsi-fungsi, masukan, dan keluaran dari perangkat lunak sesuai dengan spesifikasi yang dibutuhkan.

2.10. **PHP**

Hypertext Preprocessor atau yang disingkat dengan PHP adalah sebuah bahasa pemrograman yang menyusun data di dalam server. Pembuatan isian atas halaman web yang diinginkan, menyebabkan PHP harus berfokus pada dokumen HTML (Muarif & Winarno, 2022)

2.11. **MYSQL**

MYSQL bersifat relational. Artinya dapat memalsukan data jauh lebih cepat karena database dikelola dan disimpan di beberapa tabel yang terpisah. MYSQL berfungsi untuk mengatur database dari yang kecil hingga yang sangat besar (Muarif & Winarno, 2022)

2.12. **Penelitian Terkait**

Sebelumnya, penelitian telah dilakukan dengan mengadopsi pendekatan studi literatur yang mengacu pada jurnal atau penelitian sebelumnya yang relevan dengan masalah yang diangkat dalam penelitian ini. Namun, dalam penelitian tersebut, masih terdapat beberapa masalah yang belum terselesaikan secara optimal, terutama dalam menghadapi skalabilitas dalam metode *K-means Clustering* pada data yang berskala besar (*big data*). Oleh karena itu, penelitian ini bertujuan untuk mengatasi keterbatasan tersebut dengan menghadirkan solusi yang lebih efektif dan efisien dalam mengelola big data dalam konteks *K-means Clustering*. Penelitian terkait dapat dilihat pada tabel 2.3 dibawah ini.

Tabel 2.3 Penelitian Terkait

No	Judul, Penulis dan Tahun	Algoritma	Dataset	Perbedaan Penelitian	Hasil
1	<i>Application of data mining with the K-means clustering method and Davies Bouldin index for grouping IMDB movies</i> (Ashari et al., 2022)	K-Means	1000	Parameter yang digunakan adalah movie time, imdb_rating, meta score, no of votes dan gross. Sedangkan penelitian ini menggunakan data usia user, id user, gender, id movie dan rating.	Hasil pengelompokan data menggunakan indeks Davies Bouldin menunjukkan empat kelompok cluster. Cluster 1 memiliki skor rating tertinggi (48,74), diikuti oleh Cluster 2 (44,60) dan Cluster 3 (44,3). Cluster 0 memiliki skor rating terendah (42,15). Algoritma K-means digunakan dengan akurasi pengelompokan mencapai 100%. Cluster 1 menunjukkan hubungan yang paling baik antara variabel runtime, IMDB_Rating, Meta_score, No_of_votes, dan Gross, sementara Cluster 0 memiliki skor rating terendah.
2	<i>PENGARUH USER PROFILING PADA REKOMENDASI SISTEM MENGGUNAKAN K MEANS DAN KNN</i> (Hartatik	Silhouette, k-means, K-Nearest Neighbour	700 rating yang di crawling melalui web traveloka	Pada penelitian sebelumnya menggunakan dataset dari traveloka. Sedangkan penelitian ini menggunakan dataset movielens.	Algoritma K-Means memiliki performa baik untuk klustering data dengan waktu dan memori yang efisien. Average Silhouette Method juga membantu dalam menentukan jumlah kluster secara optimal. Algoritma KNN juga memiliki performa cukup baik dengan rata-rata error 1,33% dan akurasi 98,67% pada pengujian data sebanyak 5.

No	Judul, Penulis dan Tahun	Algoritma	Dataset	Perbedaan Penelitian	Hasil
	& Rosyid, 2020)				
3	Movie Recommen- der System Using K- Means Clusterin- g AND K-Nearest Neighbor (Ahuja et al., 2019)	K-means dan K- Nearest Neighbor	1000	Pada penelitian sebelumnya menggunakan penggabungan algoritma K- Means Clustering dan K-Nearest Neighbor. Sedangkan penelitian ini menggunakan K-Means Clustering.	Hasilnya menunjukkan teknik yang diusulkan memiliki nilai RMSE lebih baik daripada teknik sebelumnya. Selain itu, sistem yang diusulkan mencapai nilai RMSE optimal dengan jumlah cluster yang lebih sedikit. Penelitian dapat diperluas dengan menggunakan lebih banyak dataset dan Sentimental Analysis untuk meningkatkan efisiensi sistem rekomendasi film. Menghapus karakteristik individu pengguna dalam rekomendasi dapat meningkatkan kualitas rekomendasi di masa depan.
4	Collabora- tive Filtering Recommen- dation Model Based on k-means Clusterin- g (AL- Bakri & Hashim, 2019)	Collaborati- ve Filtering dan K- means Clustering	100.00 0	Penelitian sebelumnya menggunakan collaborative filtering dan K-Means untuk menemukan tetangga terdekat pengguna dan mengatasi masalah skalabilitas. Fokus penelitian ini	Metode K-means Clustering dieksplorasi untuk mengatasi masalah skalabilitas dalam sistem rekomendasi. Dengan menerapkan K-means secara offline pada matriks rating pengguna-film, sparsitas dapat dikurangi dan masalah skalabilitas dari model dapat diatasi karena perhitungan tetangga pengguna target hanya dilakukan dalam cluster yang sama, mengurangi jumlah tetangga yang perlu dihitung.

No	Judul, Penulis dan Tahun	Algoritma	Dataset	Perbedaan Penelitian	Hasil
				adalah penggunaan K-Means Clustering untuk menyelesaikan masalah big data dan membangun sistem rekomendasi film.	Penggunaan koefisien korelasi Pearson dalam hasil prediksi juga menunjukkan hasil yang relatif baik dalam menemukan tetangga terdekat untuk pengguna target.
5	Impact of Similarity Measures in K-means Clustering Method used in Movie Recommender Systems (Rahul et al., 2021)	K-means dan Collaborative-based Filtering	100.000	Fokus penelitian adalah penggunaan K-Means Clustering untuk menyelesaikan masalah big data dan membangun sistem rekomendasi film dengan analisis standar deviasi (SD), MAE, RMSE, nilai-t, Dunn Matrix, kesamaan rata-rata, dan waktu komputasi.	Penelitian ini mengevaluasi dengan menggunakan metrik standar deviasi, MAE, RMSE, nilai t, Dunn Matrix, rata-rata kesamaan, dan waktu komputasi. Hasilnya menunjukkan bahwa pengukuran kesamaan kosinus efisien dan andal dalam sebagian besar metrik yang diukur, namun membutuhkan waktu komputasi lebih lambat dibandingkan dengan metode lain seperti Euclidean dan Minkowski. Di masa depan, penelitian ini dapat menggabungkan konsep-konsep machine learning lainnya untuk meningkatkan akurasi dan waktu pemrosesan, serta mengaplikasikannya dalam konteks big data.
6	Implementasi K-	K-means dan	100.000	Penelitian ini membahas	Hasil penelitian menunjukkan bahwa

No	Judul, Penulis dan Tahun	Algoritma	Dataset	Perbedaan Penelitian	Hasil
	Means Dalam Mengatasi Masalah Cold Star Pada Collaborative Filtering (Sylvia & Lestari, 2022)	Collaborative-based Filtering		masalah pengguna baru (cold start) yang mempengaruhi kinerja sistem rekomendasi. Sistem menghadapi kesulitan menganalisis minat pengguna baru karena kurangnya data rating yang menyebabkan kesulitan dalam memberikan rekomendasi. Fokus penelitian ini adalah menyelesaikan masalah big data dan membangun sistem rekomendasi film.	penggunaan algoritma K-means dalam pendekatan clustering untuk mengatasi masalah cold start pada sistem rekomendasi dapat menghasilkan 15 cluster yang optimal berdasarkan data demografi dari Movielens.com. Jumlah item dalam masing-masing cluster bervariasi, di antaranya: Cluster 0 (140 items), Cluster 1 (134 items), Cluster 2 (2 items), Cluster 3 (21 items), Cluster 4 (33 items), Cluster 5 (55 items), Cluster 6 (152 items), Cluster 7 (215 items), Cluster 8 (4 items), Cluster 9 (48 items), Cluster 10 (2 items), Cluster 11 (90 items), Cluster 12 (1 item), Cluster 13 (14 items), dan Cluster 14 (27 items). Hasil ini dapat digunakan sebagai dasar untuk memberikan rekomendasi kepada pengguna baru yang belum memberikan rating terhadap produk, sehingga mengurangi dampak negatif dari masalah cold start pada sistem rekomendasi.
7	Sistem Rekomendasi Penawaran Produk Pada	K-means	Tidak disebutkan	Pada peneliti sebelumnya, sampel yang digunakan adalah online shop,	Penelitian ini menciptakan perangkat lunak untuk mengelompokkan data transaksi penjualan berdasarkan karakteristik pembeli, seperti rentang

No	Judul, Penulis dan Tahun	Algoritma	Dataset	Perbedaan Penelitian	Hasil
	Online Shop Menggunakan K-Means Clustering (F. Naufal et al., 2022)			sedangkan pada penelitian ini menggunakan data movielens.	umur, pada toko online. Hasilnya membentuk 3 cluster dengan karakteristik penjualan yang berbeda. Cluster 1 memiliki penjualan sedang dari pembeli berumur 36-50 tahun, cluster 2 memiliki penjualan terbanyak dari pembeli berumur 18-26 tahun, dan cluster 3 memiliki penjualan rendah dari pembeli berumur 27-35 tahun. Produk yang direkomendasikan merupakan produk terpopuler dari masing-masing cluster. Penelitian ini melakukan pengujian dengan 7 variasi jumlah titik pusat cluster, dan ditemukan bahwa cluster yang berjumlah 3 memiliki silhouette coefficient yang mendekati nilai maksimum yaitu 1, dengan nilai 0.5564
8	Penerapan Collaborative Filtering, PCA dan K-Means dalam Pembangunan Sistem Rekomendasi Film	Collaborative Filtering, PCA, dan K-Means	Tidak disebutkan	Pada penelitian sebelumnya, fokus untuk memperkecil waktu proses dengan menggabungkan algoritma K-Means, Collaborative Filtering	Studi ini menghasilkan sistem rekomendasi film yang mengkombinasikan algoritma K-Means Clustering dan User-Based Collaborative Filtering dengan penerapan metode Principal Component Analysis (PCA). Setelah mengurangi dimensi data menggunakan PCA, waktu proses berkurang menjadi 1.061282, membantu

No	Judul, Penulis dan Tahun	Algoritma	Dataset	Perbedaan Penelitian	Hasil
	(Billah et al., 2021)			dan PCA.	mengatasi masalah skalabilitas dengan mempercepat analisis clustering. Meskipun perlu meningkatkan akurasi rekomendasi, PCA terbukti efektif dalam menjaga kinerja sistem di skala yang lebih besar, mengatasi tantangan skalabilitas.
9	Analisa dan Penerapan Metode Algoritma K-Means Clustering Untuk Mengidentifikasi Rekomendasi Kategori Baru Pada List Movie IMDb (Situmorang et al., 2022)	K-Means Clustering	9998	Pada penelitian sebelumnya datasets yang digunakan sebanyak 9998, sedangkan pada penelitian ini menggunakan datasets 100k.	Dalam penelitian ini, penggunaan metode Mini Batch K-Means Clustering berhasil mengatasi masalah skalabilitas. Setelah menghitung Davies-Bouldin Index (DBI) dan melakukan sembilan kali iterasi, optimasi jumlah kluster tercapai pada K=2 dengan akurasi 0.456. Hasil ini membuktikan bahwa Mini Batch K-Means membantu dalam mengelompokkan data dengan lebih efisien, Walaupun sistem rekomendasi yang telah dibangun belum dapat memberikan rekomendasi film yang lebih tepat/akurat kepada calon penonton di IMDb.
10	Perancangan Sistem Rekomendasi Pemantau	K-Means Clustering	Tidak disebutkan	Objek pada penelitian sebelumnya berupa nilai UTS dan UAS mahasiswa,	Dalam penelitian ini, penggunaan Metode K-Means menghasilkan kluster dari data nilai mahasiswa. Pemecahan masalah skalabilitas dilakukan

No	Judul, Penulis dan Tahun	Algoritma	Dataset	Perbedaan Penelitian	Hasil
	an Studi Mahasiswa dengan Metode K-Means (Hassolthine et al., 2023)			sedangkan penelitian ini berupa film.	dengan membagi pekerjaan analisis menjadi beberapa tugas yang dapat dikerjakan secara paralel. Dengan memanfaatkan pendekatan ini, waktu eksekusi dan penggunaan sumber daya dapat dioptimalkan. Hasil ini diperlihatkan dalam grafik distribusi nilai serta jumlah cluster yang dipengaruhi oleh parameter-parameter tertentu. Sistem rekomendasi yang dirancang dengan Metode K-Means mampu memberikan rekomendasi yang diharapkan dan berpotensi meningkatkan kinerja mahasiswa dalam berbagai mata kuliah.

Keunggulan penelitian sebelumnya yang berjudul *Impact of Similarity Measures in K-means Clustering Method used in Movie Recommender Systems* adalah menggabungkan semua ukuran kesamaan untuk mengetahui efeknya dan hasilnya, bahwa *Cosine Similarity* teknik terbaik dalam hal akurasi, efisiensi, dan kecepatan pemrosesan.

Sedangkan keunggulan dalam penelitian ini adalah mampu membuat model clustering dengan mempartisi data yang besar menjadi beberapa *cluster*, dari hasil cluster tersebut mampu merekomendasikan film, dan mampu menyelesaikan masalah skalabilitas pada sistem rekomendasi dengan hasil kinerja cepat dan baik.