BAB I

PENDAHULUAN

1.1. Latar Belakang

Saat ini, dunia digital telah mengalami kemajuan yang signifikan dengan adanya teknologi yang semakin canggih dan terus berkembang, yang paling menonjol adalah sistem informasi berbasis situs web. Dengan hadirnya teknologi ini, data akan mengalami pertumbuhan pesat di berbagai sumber, termasuk situs web, aplikasi seluler, media sosial, *e-commerce*, dan lain-lain. Perkembangan data yang sangat pesat ini, dapat disebut juga dengan ledakan data. Ledakan data terjadi karena adanya kumpulan data yang berukuran besar dan terus bertambah setiap waktunya atau biasa dikenal dengan big data. Berdasarkan *data International Data Corporation (IDC)* memperkirakan, jumlah kolektif data dunia akan tumbuh dari 33 *Zettabytes* 2018 menjadi 175 *Zettabytes* pada 2025, dengan tingkat pertumbuhan tahunan sebesar 61% (Nurchayadi, 2021). Ketika ledakan terjadi, data akan terus bertambah dan berlipat ganda setiap 12 jam sekali (Gisela, 2021). Hal ini mengakibatkan volume data dalam lingkungan digital mengalami peledakan data. Ledakan data ini berdampak pada infrastruktur teknologi telekomunikasi, kecepatan akses dan volume data. Selain itu, dampak yang akan dirasakan yaitu sulit menentukan data yang dibutuhkan, membagikan data, dan menjaga data agar tetap konsisten. Oleh karena itu, perlu untuk mengeksplorasi skalabilitas pada sistem rekomendasi agar mampu mengatasi volume data yang semakin besar dengan efisien dan memberikan rekomendasi film yang akurat serta tepat waktu.

Skalabilitas merupakan kemampuan suatu sistem untuk tetap berfungsi secara efisien dan efektif ketika menghadapi peningkatan ukuran dataset dan kompleksitas algoritma. Dalam konteks sistem rekomendasi film menggunakan *K-Means Clustering*, tantangan skalabilitas mencakup

beberapa aspek kritis yang perlu dipahami. Pertama, ukuran dataset yang besar dapat menyebabkan kinerja algoritma *K-Means* menjadi lambat. Proses iteratif untuk mengelompokkan data menjadi kluster membutuhkan waktu yang signifikan, terutama ketika terdapat jutaan atau bahkan miliaran data film dan pengguna. Dalam konteks big data, waktu eksekusi yang lama dapat mengurangi responsivitas sistem dan membuat pengguna menjadi tidak sabar dalam menerima rekomendasi. Kedua, kompleksitas algoritma *k-means clustering* juga merupakan faktor kunci dalam menilai skalabilitas sistem rekomendasi. Semakin tinggi dimensi data dan jumlah atribut yang diperhitungkan dalam proses *clustering*, semakin rumit perhitungan dan analisis yang diperlukan. Tingginya kompleksitas dapat menimbulkan beban komputasi yang berat dan mempengaruhi kualitas rekomendasi. Selanjutnya, keandalan dan konsistensi hasil kluster yang dihasilkan oleh algoritma *K-Means* saat menghadapi data yang besar menjadi hal yang penting untuk dipertimbangkan. Dengan data yang sangat besar, ada kemungkinan bahwa beberapa kluster dapat terfragmentasi atau data outlier dapat mempengaruhi pembentukan kluster yang akurat. Oleh karena itu, perlu mempertimbangkan teknik dan strategi khusus untuk menghadapi situasi ini dan menghasilkan kluster yang kohesif dan relevan.

Penelitian yang pernah dilakukan untuk mengatasi masalah penyelesaian skalabilitas pada sistem rekomendasi diantaranya penelitian oleh (Budiman & Anto, 2019) fokus pada penggunaan teknik pemrosesan paralel untuk meningkatkan kinerja sistem rekomendasi saat menghadapi peningkatan jumlah pengguna. Penelitian ini mengusulkan pendekatan dengan menerapkan data mining dengan metode *clustering* mengelompokkan item objek penelitian berdasarkan kemiripan sifatnya. Kedua, penelitian oleh (Billah et al., 2021) 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 mengatasi masalah skalabilitas dengan mempercepat analisis *clustering*.

Pada penelitian ini, mengusulkan algoritma *clustering* yaitu *K-means Clustering*. Cara kerjanya adalah dengan melakukan pengelompokan atau pembagian data menjadi beberapa kelompok data berdasarkan kriteria yang dibutuhkan (Hartatik & Rosyid, 2020). Penelitian yang dilakukan oleh (Ahuja et al., 2019), skalabilitas menjadi masalah karena jumlah data rating film yang digunakan dari dataset *kagle* sangat besar (Dataset terdiri dari 100.000 film dengan rating dari 1-5). Namun, penelitian ini berhasil mengatasi masalah skalabilitas tersebut dengan menggabungkan algoritma *silhouette*, *K-means Clustering*, dan *K-Nearest Neighbor* dalam menghasilkan sistem rekomendasi film. Dengan menggunakan metode ini, penelitian ini mencapai hasil yang baik dengan nilai RMSE yang rendah (1.081648) dan akurasi sebesar 98,67%.

Seiring dengan perkembangan teknologi, film telah menjadi media hiburan yang mudah dijangkau dan diakses melalui berbagai platform. Saat ini, bisa dikatakan perfilman Indonesia semakin berkembang. Meski teknologinya tidak secanggih film-film *Hollywood*, industri perfilman Indonesia telah mencapai tingkat perkembangan tertinggi dalam sejarahnya. Ada banyak film dan genre yang populer dan diminati oleh penonton. Dengan begitu, membuat banyak penonton atau pengguna kebingungan dalam memilih dan menentukan rekomendasi film apa yang akan ditonton, sehingga diperlukan sistem rekomendasi film. Adanya informasi demografis dan rating film dari pengguna sebelumnya dapat digunakan untuk merekomendasikan film kepada pengguna baru.

Terdapat peneliti terdahulu yang membahas tentang sistem rekomendasi film menggunakan algoritma *K-means Clustering*. Hasil dari penelitian ini, mempermudah melakukan perekomendasi film sesuai dengan genre dan rating (Mutiasari et al., 2021). Ada juga peneliti terdahulu yang membahas

tentang penerapan data mining terhadap minat siswa dalam mata pelajaran matematika menggunakan algoritma *K-means*. Hasil dari penelitian ini, mampu mengelompokkan data menjadi 2 *cluster*, *cluster* 1 yaitu (Minat) dan *cluster* 2 (Tidak Minat). Disertai penggunaan DBI (*Davies Bouldin Index*) untuk mendapatkan nilai yang optimal (Butsianto & Saepudin, 2020).

Implementasi *K-Means Clustering* pada *Big Data* di sistem rekomendasi film berdasarkan usia dan rating menghadapi beberapa permasalahan yaitu banyaknya data film, menjadikan adanya kendala dalam mengelola dan mengelompokkan *big data* (skalabilitas) untuk menentukan rekomendasi film. Dan banyak pengguna yang kebingungan dalam memilih film. Untuk meringankan masalah tersebut, peneliti berupaya membuat sistem rekomendasi film berdasarkan usia dalam bentuk website dengan mengelompokkan usia pengguna, yang diharapkan dapat memberikan informasi film dan merekomendasikan beberapa film sesuai usia dengan membandingkan kemiripan data usia pengguna sebelumnya dengan menggunakan algoritma *K-Means Clustering*. Sehingga, penulis melakukan penelitian dengan judul ***“IMPLEMENTASI K-MEANS CLUSTERING PADA BIG DATA DI SISTEM REKOMENDASI FILM”***.

1.2. Batasan Masalah

Adapun batasan masalah dalam penelitian ini yang dapat dilihat pada poin-poin berikut:

1. Objek pada penelitian ini berupa data film.
2. Penelitian ini menggunakan dataset movielens 100K yang bersifat publik sebagai sampel data untuk eksperimen, yang di ambil dari *website kaggle.com*.
3. Penelitian ini menggunakan algoritma *K-means Clustering* untuk melakukan partisi (pengelompokkan) berdasarkan usia pengguna yaitu 17 hingga 59 tahun dan merekomendasikan film berdasarkan demografis pengguna yaitu usia, *gender*, *occupation*.

4. Data pada penelitian ini dikelompokkan menjadi tiga *cluster* yaitu C0, C1, dan C2.
5. Variabel yang digunakan dalam penelitian ini yaitu *user id*, usia, *gender*, *occupation*, *movie id*, dan rating.
6. Informasi yang disajikan meliputi list data film dengan informasi tambahan genre dan rating dari film tersebut.

1.3. Rumusan Masalah

Berdasarkan latar belakang yang telah dibahas, rumusan masalah yang diperoleh, adalah sebagai berikut:

1. Bagaimana melakukan partisi (pengelompokkan) data film yang besar (skalabilitas) berdasarkan demografis pengguna yang diwakili oleh usia, menggunakan algoritma *K-Means Clustering*?
2. Bagaimana membangun sistem rekomendasi film berdasarkan hasil *clustering* untuk memberikan rekomendasi film yang sesuai dengan demografis pengguna?
3. Bagaimana menyajikan daftar film yang terkelompokkan berdasarkan usia pengguna, beserta informasi genre dan ratingnya sebagai output dari sistem rekomendasi yang dikembangkan?

1.4. Tujuan Penelitian

Berdasarkan rumusan masalah yang diberikan maka tujuan dari penelitian ini adalah :

1. Melakukan partisi (pengelompokkan) data film yang besar (skalabilitas) berdasarkan demografis pengguna yang diwakili oleh usia menggunakan algoritma *K-Means Clustering*.
2. Membangun sistem rekomendasi film berdasarkan hasil *clustering* untuk memberikan rekomendasi film yang sesuai dengan demografis pengguna.
3. Mempresentasikan daftar film yang terkelompokkan berdasarkan usia pengguna, beserta informasi genre dan ratingnya sebagai output dari sistem rekomendasi yang dikembangkan.

1.5. Manfaat Penelitian

Manfaat dari penelitian ini antara lain :

1. Penelitian ini diharapkan memberikan manfaat dalam pengembangan sistem rekomendasi film yang lebih personal dan akurat berdasarkan demografis pengguna.
2. Dengan mengimplementasikan metode *K-Means Clustering* pada big data film, diharapkan dapat meningkatkan pengalaman pengguna dalam menemukan film-film yang sesuai dengan usia mereka.
3. Mempermudah pengguna untuk mendapatkan rekomendasi film sesuai dengan usianya.
4. Diperolehnya hasil dari pembuatan model *clustering* dan hasil sistem rekomendasi film berdasarkan usia pengguna menggunakan *K-Means Clustering*.

1.6. Sistematika Penulisan

Sistematika penulisan dibuat untuk mempermudah dalam penyusunan penelitian ini maka perlu ditentukan sistematika penulisan yang baik. Berikut adalah sistematika penulisan penelitian ini:

BAB I PENDAHULUAN

Bab ini berisi latar belakang masalah mengenai kondisi, perkembangan dan ledakan dari big data serta dampak dari ledakan yang terjadi di dunia digital khususnya dalam hal big data dan mengapa perlu dibuatnya sistem rekomendasi film yang sesuai untuk dikonsumsi atau ditonton oleh pengguna yang sesuai dengan usianya menggunakan metode *K-means Clustering*, batasan masalah, rumusan masalah, tujuan penelitian, manfaat penelitian dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA

Bab ini berisi tentang teori – teori yang mendukung penelitian yang dilakukan oleh penulis antara lain big data, clustering, *K-means Clustering*, *Davies-Bouldin Index (DBI)*, *Cosine Similarity* metode pengembangan perangkat

lunak, serta literatur terkait penelitian-penelitian terdahulu untuk menunjang penelitian ini.

BAB III METODOLOGI PENELITIAN

Bab ini berisi tentang metode yang digunakan dalam penelitian. Dalam penelitian ini juga membahas tentang alur penelitian dan perancangan sistem rekomendasi film.

BAB IV HASIL PENELITIAN DAN PEMBAHASAN

Bab ini berisi tentang tampilan interface, pengujian dan hasil dari website rekomendasi film.

BAB V KESIMPULAN DAN SARAN

Bab ini memberikan kesimpulan dari penelitian yang telah dilakukan dan saran yang berguna untuk penelitian selanjutnya.