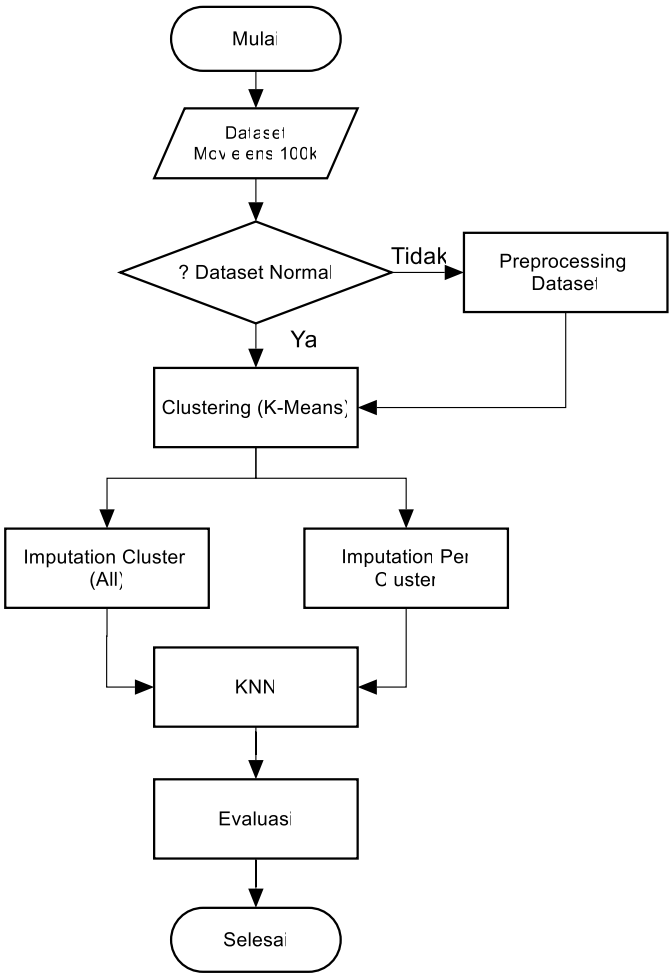
BAB III
METODOLOGI PENELITIAN

Penelitian ini bertujuan untuk mengatasi tantangan *cold start* dan *sparsity* pada *collaborative filtering* menggunakan teknik *clustering* dan *imputation*. Hasil dari proses ini kemudian digunakan sebagai landasan untuk memberikan rekomendasi kepada pengguna yang sesuai dengan preferensinya. Langkah-langkah yang dilakukan dapat dilihat pada gambar tahapan penelitian berikut.



Gambar 3.1 Flowchart Penelitian

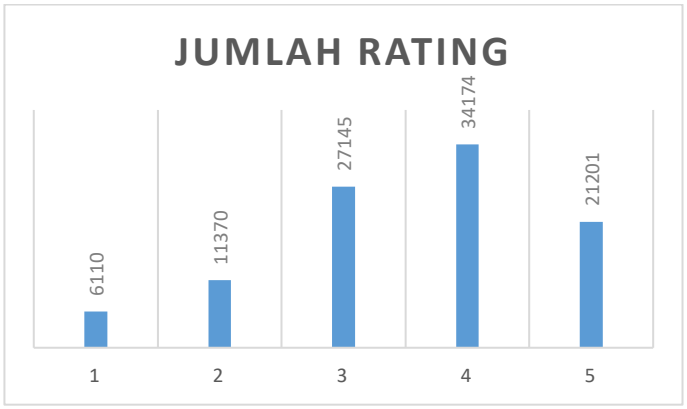
3.1 Pengumpulan Data

Dataset yang akan digunakan dalam penelitian ini adalah Dataset Movielens yang tersedia di situs web https://grouplens.org/datasets/movielens/100k/. Dataset ini terdiri dari 100.000 peringkat dalam rentang skala 1-5, yang diberikan oleh 943 pengguna untuk 1682 film. Setiap pengguna setidaknya memberikan peringkat untuk 20 film, dan dataset juga mencakup informasi demografis seperti usia, jenis kelamin, pekerjaan, dan kode pos pengguna. Dataset ini dirilis pada bulan April 1998 dan mengandung *sparsity* sebesar 93,7%.

Tabel 3.1 Deskripsi Dataset

Nama *File Dataset*	Deskripsi *File Dataset*	Atribut	Jumlah *Record*
u.data	Kumpulan data 100.000 *rating*film oleh 943 pengguna pada 1682 film	Id Pengguna (*User Id*) 1 - 943 Id Film (*Movie Id*) 1 - 1682 Rating Film 1 – 5	943 1.682 100.000
u.user	Informasi demografis tentang 943 pengguna yang memberikan rating terhadap 1682 film	Id Pengguna (*User Id*) 1 - 943 Umur User (*Age*) 7 - 73 Jenis Kelamin User (Pria / *Male* dan Wanita (*Female*) Pekerjaan User (*Occupation*) 21 macam pekerjaan *Zipcode* (Kode Pos Rumah)	943 943 943 943 943 943

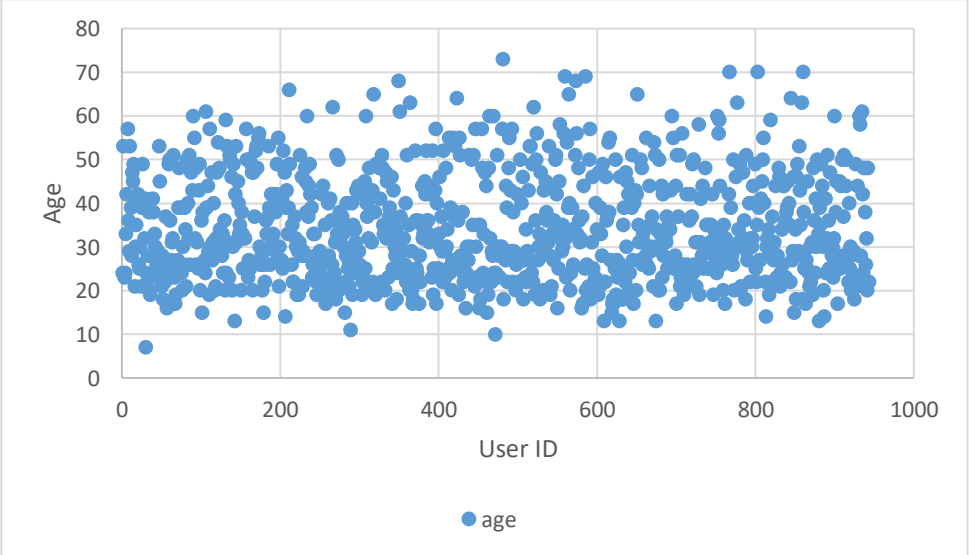
Dari dataset ini, setiap rating memiliki variasi jumlah yang berkisar antara 6.110 hingga 34.174. Jumlah rating terendah yaitu 6.110 pada rating 1, sementara jumlah rating tertinggi yaitu sejumlah 34.174 pada rating 4. Untuk lebih jelasnya dapat dilihat pada gambar berikut



Gambar 3.2 Grafik Sebaran Rating Dataset Movielens 100k

3.2 *Preprocessing* Data Demografi

Collaborative Filtering berbasis pengguna *(User-based)* beroperasi dengan mengasumsikan bahwa setiap pengguna dapat dianggap sebagai anggota dari suatu kelompok yang memiliki kesamaan dengan pengguna lainnya. Oleh karena itu, pengguna yang memiliki kesamaan dalam hubungan atau atribut tertentu cenderung tertarik pada item yang sama[20]. Pada penelitian ini, data demografi pengguna melibatkan faktor-faktor seperti usia (age), jenis kelamin (gender), pekerjaan (occupation), dan kode pos tempat tinggalnya (zipcode). Akan tetapi, atribut kode pos tidak memiliki pengaruh sehingga tidak akan digunakan pada proses selanjutnya. Pada atribut umur juga dilakukan pembersihan terhadap pencilan data untuk memperoleh data dengan kualitas yang lebih baik. Pencilan data pada atribut umur dapat dilihat pada gambar berikut



Gambar 3.3 Pencilan Data pada Atribut Age

Selain itu, dilakukan juga *preprocessing* pada atribut *occupation* dengan melakukan transformasi tipe data dari tipe teks menjadi numerik. Hal ini dikarenakan KNN tidak dapat menerima masukan data dengan tipe teks, sehingga dilakukan transformasi data untuk dapat memproses tahapan selanjutnya dan mendapatkan hasil yang optimal. Transformasi data pada atribut *occupation* dapat dilihat pada tabel berikut

Tabel 3.2 Transformasi Data pada Atribut *Occupation*

Occupation (Teks)	**Occupation (Numerik)**
administrator	1
artist	2
doctor	3
educator	4
engineer	5
entertainment	6
executive	7
healthcare	8
homemaker	9
lawyer	10
librarian	11
marketing	12
none	13
other	14
programmer	15
retired	16
salesman	17
scientist	18
student	19
technician	20
writer	21

Pada tabel di atas, data pada atribut *occupation* sudah ditransformasikan ke dalam tipe data numerik dan digantikan dengan angka. Selanjutnya, data dengan format numerik tersebut yang akan digunakan pada proses berikutnya.

3.3 *Preprocessing* Data Rating

Data rating terdiri dari 100.000 penilaian (rating) dengan rentang nilai antara 1 hingga 5 dari 943 pengguna terhadap 1682 film. Jika setiap pengguna (943 pengguna) memberikan penilaian terhadap setiap film (1682 film), seharusnya terdapat 1.586.126 penilaian. Angka ini dihitung dengan

mengalikan jumlah pengguna dan jumlah film, yaitu 943 x 1682 = 1.586.126 penilaian. Namun, yang berhasil dikumpulkan hanya 100.000 penilaian. Dengan demikian, terdapat kekurangan penilaian sebanyak 1.486.126, yang dihitung sebagai selisih antara 1.586.126 dan 100.000. Jika dihitung dalam persentase, jumlah penilaian yang terisi hanya sebesar 6,0346% dari total yang seharusnya. Sementara itu, persentase penilaian yang kosong adalah 93,695%. Ketidaklengkapan ini dalam data disebut sebagai *sparsity*. Dataset rating mencakup atribut user id, movie id, dan rating. Contoh data pada data rating dapat dilihat pada tabel berikut

Tabel 3.3 Potongan Data Rating Movielens 100k (Bagian Atas)

User Id	*Movie Id*	*Rating*
298	474	4
115	265	2
253	465	5
305	451	3
6	86	3

Langkah selanjutnya yaitu mentranspose data tersebut atau mengubahnya menjadi format kolom dengan User Id, movie1, movie2, dan seterusnya pada suatu tabel baru. User Id menyimpan nomor pengguna dari 1 hingga 943. Kolom movie1 hingga movie1682 berisi nilai rating yang telah diberikan oleh pengguna terhadap masing-masing film. Gambaran data yang sudah ditranspose dapat dilihat pada tabel di bawah

Tabel 3.4 *Preprocessing* Data Rating

User Id	*movie1*	*movie2*	*movie3*	*...*	*...*	*movie1681*	*movie1682*
1	5	3	4	…	…	0	0
2	4	0	0	…	…	0	0
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
909	0	0	0	…	…	0	0
910	0	5	0	…	…	0	0

Setelah itu, dilakukan penggabungan antara data demografi pengguna dengan data rating. Gambaran data demografi dan rating yang sudah digabungkan dapat dilihat pada tabel berikut

Tabel 3.5 Dataset Movielens 100k *After Preprocessing*

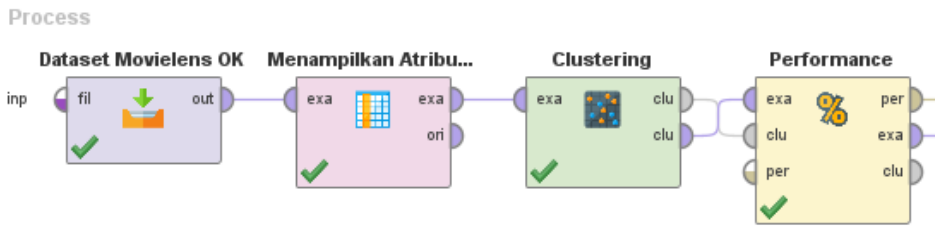
Id	*age*	*gender*	*occupation*	*movie1*	*movie2*	*...*	*...*	*movie1681*	*movie1682*
1	24	2	20	5	3	…	…	0	0
2	53	1	14	4	0	…	…	0	0
…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…
909	48	1	11	0	0	…	…	0	0
910	22	2	19	0	5	…	…	0	0

Kemudian, dataset ini diberikan penamaan sebagai **Dataset Movielens OK**. Dataset tersebut akan digunakan dalam tahap pengolahan data berikutnya.

3.4 Clustering Dataset Movielens OK menggunakan Algoritma K-Means

Proses ini terdiri dari beberapa tahap, yaitu:

1. Lakukan klasterisasi pada Dataset Movielens OK menggunakan algoritma K-Means, mulai dari k = 2 hingga k = 5. Proses klasterisasi ini akan dijalankan menggunakan perangkat lunak Rapidminer. Desain prosesnya dapat dilihat pada gambar di bawah ini.



Gambar 3.4 Desain Model K-Means Clustering pada Dataset Movielens OK

Dataset Movielens OK merupakan dataset movielens yang telah diproses pada tahap-tahap sebelumnya. Operator Select Attributes digunakan untuk menampilkan dan memproses atribut-atribut yang memang diperlukan untuk proses klastering. Operator Clustering digunakan untuk melakukan klasterisasi data menggunakan algoritma K-Means berdasarkan atribut yang sudah dipilih sebelumnya. Lalu operator Performance digunakan untuk mendapatkan informasi mengenai jumlah anggota setiap klaster, anggota di setiap klaster, nilai Davies Bouldin Index (DBI) untuk masing-masing klaster, visualisasi klaster yang ada, dan elemen-elemen lainnya. Beberapa metode yang dapat digunakan untuk menentukan jumlah klaster yang sesuai dalam algoritma K-Means melibatkan Silhouette Method[21], Sum of Squared Errors (SSE) [22], Elbow Method [23], dan Davies-Bouldin Index [24]. Dalam penelitian

ini, Davies-Bouldin Index digunakan untuk menentukan jumlah klaster yang optimal. Penggunaan DBI melibatkan optimalisasi jarak antar klaster serta meminimalkan jarak antar titik di dalam satu klaster. Apabila jarak antar klaster besar, ini menunjukkan perbedaan karakteristik yang lebih jelas di antara klaster karena kesamaan antar klaster lebih rendah. Sebaliknya, jarak antar klaster yang kecil mengindikasikan tingkat kemiripan yang tinggi di antara objek dalam satu klaster[25].

2. Melakukan pengecekan nilai Davies Bouldin Index (DBI) pada k = 2.
3. Lakukan kembali langkah 1, namun ubah nilai k-nya dari k = 3 hingga k = 5, dan catat nilai Davies Bouldin Index untuk setiap percobaan.
4. Identifikasi dan tentukan nilai Davies Bouldin Index terendah dari klaster k = 2 hingga k = 5.
5. Setelah itu, pada hasil klaster dengan Davies Bouldin Index terendah, anggota dari setiap klaster disatukan dengan Dataset Movielens OK. Penggabungan dataset ini diberi penamaan Dataset K-Means Cluster All.
6. Selanjutnya, dataset ini yang akan diolah menggunakan 2 pendekatan imputation, yakni imputation pada data klaster secara menyeluruh dan imputation per klaster secara terpisah.