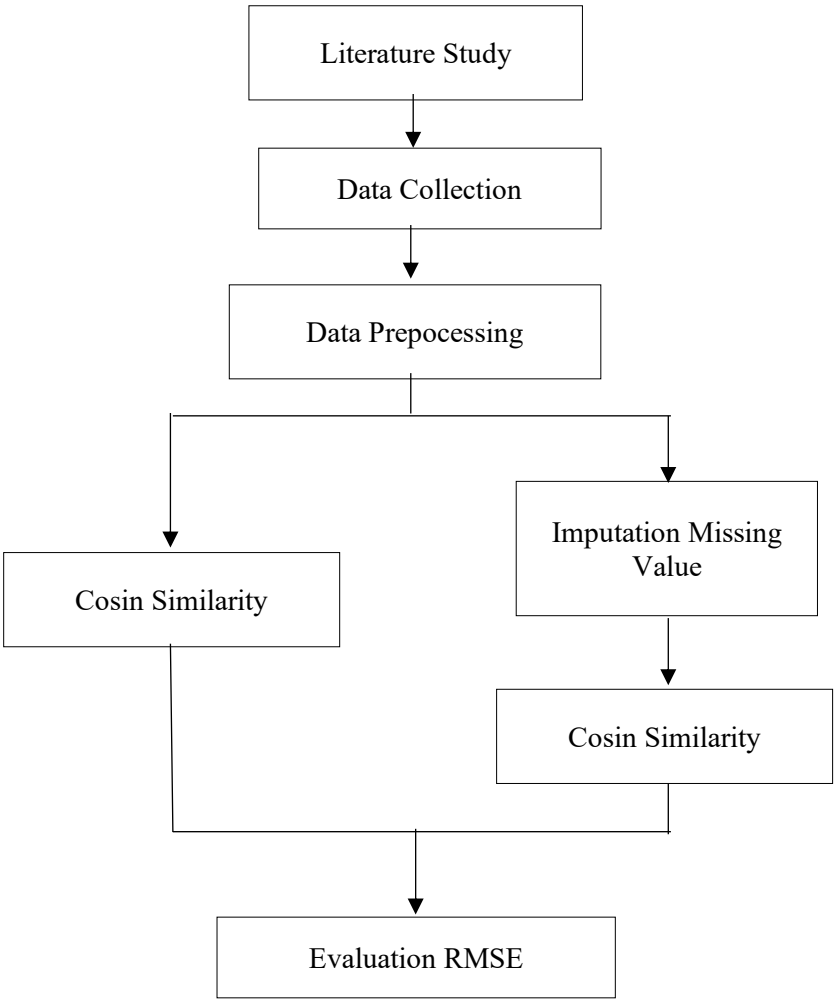
BAB III
METODOLOGI PENELITIAN

Penelitian ini dilakukan dengan beberapa tahapan kegiatan dari literatur study, data collection, data preprocessing, dilanjutkan dengan teknik imputation dan Cosin Similarity, serta dilakukan evaluasi menggunakan Root Mean Square Error (RMSE), serperti pada gambar 1.



Gambar 3.1. Flowchart Tahapan Penelitian

3.1 Pengumpulan Data

Data yang kami pakai dalam penelitian ini diperoleh dari website Kaggle, website yang menyediakan data-data yang di butuhkan oleh data scientis dan data diperoleh dalam bentuk file csv (comma-separated values file).

Dataset yang kami gunakan adalah dataset movielens yang saya unduh dari link ini https://grouplens.org/datasets/movielens/100k/. Movielens merupakan satu kumpulan data yang paling sering digunakan penelitian, movielens menyimpan

informasi terkait user dan movie[24]. Data set MovieLesn 100K memiliki 100.000 rating yang diberikan oleh 943 user terhadap 1682 movie dengan scala 1-5 [25] [26] . Data movielens terdiri dari 3 file yang terpisah meliputi file data user, file data movie dan file data ratting.

a. Data User
   Data user mencangkup atribut userId, gender, age, occupation, zip. Berikut contoh penggalan data user.

Tabel 3.1 Data User

userId	gender	age	occupation	zip
1	F	25	10	48067
2	M	56	16	70072
3	M	25	15	55117
4	M	45	7	2460
5	M	25	20	55455

Ket :

a. Id Pengguna (User Id)
b. Umur User (Age)
c. Jenis Kelamin User (Pria / Male dan Wanita (Female)
d. Pekerjaan User (Occupation) 21 macam pekerjaan
e. Zipcode (Kode Pos Rumah)

b. Data Movie
   Data Movie mencangkup atribut movieId, movie_names, genres. Berikut contoh penggalan data movie.

Tabel 3.2 Data Movie

movieId	movie_names	genres
1	Toy Story (1995)	Animation|Children's|Comedy
2	Jumanji (1995)	Adventure|Children's|Fantasy
3	Grumpier Old Men (1995)	Comedy|Romance
4	Waiting to Exhale (1995)	Comedy|Drama
5	Father of the Bride Part II (1995)	Comedy

Ket :

a. Id Movie (movieId)
b. Nama Movie (movie_names)
c. Jenis / Tipe (genres)

c. Data Ratting

Data ratting mencangkup atribut userId, MovieId, dan rating. Berikut contoh penggalan data ratting.

Tabel 3.3 Data Ratting

userId	movieId	rating
196	242	3
186	302	3
22	377	1
244	51	2
166	346	1

3.2 Preprosessing Data

Pada tahap preprosessing data akan dilakukan install modul pada python dan melakukan beberapa import modul yang akan di gunakan dalam melakukan penelitian dikarena penelitian ini mengunakan python. Setelah mengimpor data dari csv kita dapat menindentifikasi masalah sparsity pada data ratting yang terdiri dari 100.000 penilaian dengan retang nilai 1 hingga 5 dari 943 userId dan 1682 movieId. Dari data ratting akan dirubah kedalam format pivot table

Tabel 3.4 Pivot Table Ratting

movieId	1	2	3	...	1682
userId					
1	5.0	3.0	4.0	...	NaN
2	4.0	NaN	NaN	...	NaN
3	NaN	NaN	NaN	...	NaN
4	NaN	NaN	NaN	...	NaN
5	4.0	3.0	NaN	...	NaN
…	…	…	…	…	…
943	NaN	5.0	NaN	…	NaN

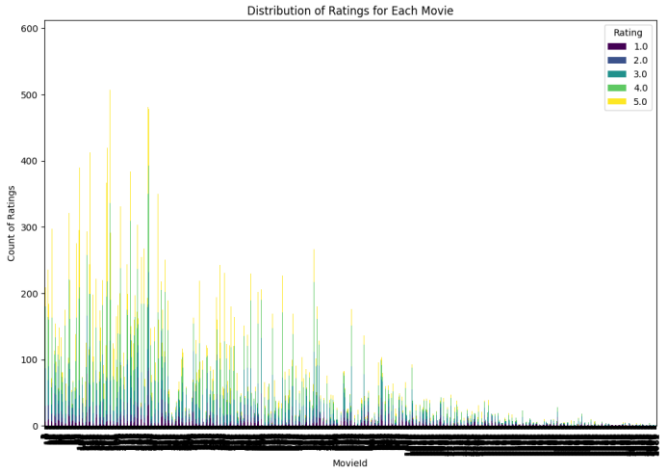
Setelah dilakukan proses perubahan format table kedalam bentuk pivot table terlihat data yang belum memilik nilai ratting atau disebut juga kekosongan.

Kemudian Setelah dilakuan perhitungan Jumlah nilai masing-masing rating dari 1 hingga 5 pada setiap Movie Id dapat di lihat di tabel berikut.

Tabel 3.5 Jumlah nilai masing-masing rating pada movie

Movie Id	**1**	**2**	**3**	…	1682
1	8	8	11	…	
2	27	17	20	…	
3	96	55	25	…	1
4	202	42	23	…	
5	9	11	29	…	

Melihat dari jumlah masing-masing rating yang didapat pada setiap movie id dapat di simpulkan bahwa banyak movie id yang tidak mendapatkan ratting dan didapatkan grafik untuk rating 1 hingga 5 sebagai berikut.



Gambar 3.1 Grafik Sebarang Rating pada Movie

jika dihitungn terdapat 1486126 nilai yang kosong. sehingga data ini mengadung sparsity 93.7%. Data yang kosong ini akan dilakukan imputation. Sehingga memiliki nilai data yang dapat di gunakan dalam metode collaborative filtering.

3.3 Teknik Imputasi

Pada penelitian ini Teknik imputasi menggunakan library pada pyhton yaitu pandas dan KNN Inputation. dan Teknik imputation tersebut melipuit:

- Imputation dengan nilai Rata-Rata (Mean)
- Imputation dengan nilai Tertinggi (Max)
- Imputation dengan nilai Terendah (Min)
- Imputation dengan nilai Observasi Terdekat K-Nearest Neighbors (KNN Imputer)

3.4 Pengujian

Pada tahap ini dilakukan pengujian terdapat performa model menggunakan Root Mean Squared Error (RMSE) dan akan dilakukan perhitungan cosin similarity atau nilai kedekatan pada semua metode inputatuin yang akan di lakukan.

3.5 Jadwal Penelitian

Penelitian ini dilaksanakan melalui serangkaian kegiatan. Rincian jadwal pelaksanaan penelitian yang telah direncanakan dapat dilihat pada table berikut ini:

Tabel. 3.6 Jadwal Penelitian

No	Kegiatan	Bulan ke-1				Bulan ke-2				Bulan ke-3			
		1	2	3	4	1	2	3	4	1	2	3	4
1	Studi Literatur	✓	✓										
2	Review & Preprocessing			✓	✓	✓							
3	Desain Model						✓	✓	✓				
4	Implementasi Model									✓	✓		
5	Pengujian											✓	✓