

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Perkembangan dan persebaran data informasi pada jaringan internet dari tahun ke tahun mengalami peningkatan yang signifikan. Peningkatan ini menyebabkan kesulitan untuk menentukan informasi yang tepat dan relevan sesuai kebutuhannya. Untuk itu diperlukan adanya sistem yang mampu mengolah dan memilah informasi yang dibutuhkan pengguna[1].

Sistem rekomendasi merupakan sebuah sistem yang mampu mengolah dan memilah informasi sesuai dengan kebutuhan pengguna[2]. Sistem rekomendasi menjadi populer di banyak hal yang dijual di internet seperti buku ,film,artikel penelitian, media promosi seperti iklan atau ads dan banyak lagi, salah satu yang paling populer adalah sistem rekomendasi pilihan film pada layanan Netflix. Netflix membuat rekomendasi berdasarkan film yang telah dilihat pelanggan sebelumnya. Pada tahun 2006 Netflix mengadakan kompetisi kepada orang pertama yang dapat meningkatkan sistem rekomendasi sebesar 10% dengan total hadiah 1 juta dollar[3]. Dapat kita sadari bahwa peningkatan kinerja sistem rekomendasi memiliki kompleksitas dan merupakan hal yang berharga.

Sistem rekomendasi menjadi komponen penting yang digunakan dalam berbagai bidang kehidupan terutama dalam mendapatkan informasi yang tepat dan relevan sesuai kebutuhan . Banyak teknik yang dapat digunakan untuk sistem rekomendasi dan secara umum dikategorikan menjadi *Content-based filtering* dan *Collaborative Filtering*. *Content-based filtering* memberikan rekomendasi

berdasarkan nilai kemiripan dari *feature* dalam item dan pengguna. Sedangkan *Collaborative Filtering* memberikan rekomendasi dengan mengolah informasi yang disukai pengguna dengan melihat rating yang diberikan oleh pengguna[2].

Teknik rekomendasi yang paling populer yaitu *Collaborative Filtering*. Metode *Collaborative Filtering* memiliki dua pendekatan umum, yaitu *item-based* dan *user-based*. *Item based collaborative filtering* menggunakan pola kemiripan antar item-item atau kesamaan antar item. Sedangkan *User based collaborative filtering* menggunakan pola kemiripan antar pengguna atau kesamaan antar pengguna. Meskipun pendekatan *Collaborative Filtering* berbasis pengguna (*user based*) telah berhasil diterapkan di berbagai domain, beberapa tantangan serius tetap ada dalam hal situs *e-commerce* besar, di mana kita harus menangani jutaan pengguna dan jutaan katalog item-item. Secara khusus, kebutuhan untuk memindai sejumlah besar tetangga potensial membuatnya mustahil untuk menghitung prediksi secara *real time*. rekomendasi berbasis item (*Item based*), adalah cara yang lebih tepat untuk pemrosesan *offline* dengan demikian memungkinkan untuk komputasi rekomendasi secara *real time* bahkan untuk matriks peringkat yang sangat besar[2].

Secara umum pencarian pola kemiripan (*Similarity*) guna menemukan hasil evaluasi terbaik memiliki banyak cara, seperti pencarian pola kemiripan berdasarkan perhitungan jarak *vector* atau *cosine*, *euclidean distance*, *pearson* maupun *manhattan* dan lain-lain. Karena setiap pola perhitungan kemiripan ada berbagai macam dan hasil yang beragam, maka penulis melakukan pengujian *item-based collaborative filtering* yang berfokus pada kedekatan *cosine* dan *pearson*

untuk menemukan hasil akurasi terbaik dengan menggunakan algoritma *k-nearest neighbor*. Karena hasil akhir dari pendekatan *Item based Collaborative Filtering* berupa prediksi, maka penulis menggunakan metode pengujian *root mean square error* (RMSE). RMSE termasuk kedalam metode *predictive accuracy metrics* yang dapat mengukur perbedaan nilai error antara nilai prediksi, dan nilai asli rating yang diberikan oleh user.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang disampaikan, maka didapatkan rumusan masalah sebagai berikut :

1. Bagaimana langkah-langkah pengujian *item-based collaborative filtering* dengan algoritma *k-nearest neighbor* menggunakan *cosine similarity* dan *pearson similarity* dalam pengujian *similarity* sebuah sistem rekomendasi?
2. Manakah *similarity* yang lebih baik antara *cosine* dengan *pearson similarity* pada pengujian *item-based collaborative filtering* dengan algoritma *k-nearest neighbor*?
3. Apakah *tuning GridSearchCV* parameter nilai K memberikan hasil evaluasi yang signifikan?
4. Bagaimana hasil evaluasi RMSE sistem rekomendasi yang telah dibangun berdasarkan nilai k yang diberikan?
5. Faktor apa saja yang menyebabkan tingginya hasil skor evaluasi RMSE?

1.3 Batasan Masalah

Permasalahan dalam penelitian ini akan dibatasi dengan beberapa hal berikut :

1. Perbandingan Sistem rekomendasi pada penelitian ini dibangun menggunakan metode pendekatan *item-based collaborative filtering* dengan algoritma *k-nearest neighbor*.
2. Pengujian perbandingan nilai kedekatan yang diujikan adalah *Cosine similarity* dan *Pearson similarity*
3. Evaluasi algoritma sistem rekomendasi menggunakan metode RMSE (*Root mean square error*).
4. Pembuatan sistem rekomendasi menggunakan *google colab*.
5. Dalam pengujian menggunakan 2 dataset yaitu *Goodbooks-10k* yang diperoleh dari *Kaggle*, dan *Book Crossing* yang diperoleh dari *Grouplens*
6. Nilai *k* untuk uji coba dan evaluasi sistem rekomendasi pada penelitian ini menggunakan nilai ganjil dari interval 1-100[4].

1.4 Tujuan dan Manfaat Penelitian

Tujuan dari penelitian ini adalah untuk menguji hasil evaluasi terbaik pada sistem rekomendasi dengan membandingkan tingkat kesamaan *cosine* dan *pearson similarity*. Selain itu, terdapat beberapa manfaat sebagai berikut :

1. langkah-langkah dalam pengujian perbandingan similarity antara cosine dan pearson menggunakan *item-based collaborative filtering*

dengan algoritma *K-Nearest-Neighbor*

2. Mencari nilai *similarity* yang lebih baik antara pengujian *cosine* dan *pearson similarity*.
3. Memberikan hasil pengaruh dari pengujian nilai k dari 1 – 100 menggunakan bilangan ganjil pada 2 dataset yang berbeda.
4. Menjelaskan hasil evaluasi RMSE pada nilai k yang diberikan.
5. Menjelaskan penyebab tingginya hasil evaluasi RMSE yang tinggi.

1.5 Metode Penelitian

Metodologi penelitian yang digunakan dalam penelitian ini adalah dengan kuantitatif, yaitu penelitian yang mengikuti prosedur-prosedur yang telah ditentukan sebelumnya, dan melibatkan perhitungan angka dan matematis dalam implementasi maupun pengujiannya. Macam-macam metode penelitian yang digunakan adalah sebagai berikut :

1.5.1 Metode Studi Literatur

Penulis mengumpulkan data penelitian terkait dari berbagai sumber literatur seperti buku-buku, pustaka, *internet*, jurnal-jurnal ilmiah baik nasional maupun internasional yang berkaitan dengan masalah yang dibahas untuk menunjang proses penelitian.

1.5.2 Metode Data Mining

Metode data mining yang digunakan dalam penelitian ini adalah forecasting. Di dalam metode data mining, terdapat proses-proses sebagai berikut :

1. Pengumpulan Data

Dataset yang digunakan dalam penelitian ini adalah data sekunder berupa rating user terhadap suatu buku dan judul buku. Data bersifat terbuka yang bersumber dari Goodbooks dan juga Book Crossing [13] berjenis data time series, yaitu data yang dikumpulkan dalam kurun waktu tertentu. Pada langkah ini data yang didapatkan di filter menggunakan query tertentu berdasarkan kategori sehingga didapatkan dataset buku yang ideal dan siap untuk dilatih.

2. Pemilihan Data

Pada tahap ini penulis melakukan pemilihan dataset yang telah dikumpulkan. Dataset yang terpilih, nantinya akan digunakan untuk menjadi data latih sistem rekomendasi.

3. Pembersihan Data

Pada tahap ini penulis melakukan pembersihan data terhadap dataset yang telah terpilih. Pembersihan yang dimaksud disini adalah penghapusan data-data yang tidak diperlukan pada saat pembersihan data sehingga data tersebut dapat dinyatakan ideal dan layak untuk dilatih.

4. Pengolahan Data

Pengolahan data merupakan proses utama data mining. Pada tahap ini data akan diolah menggunakan pendekatan user based collaborative filtering dengan algoritma KNN (K-nearest neighbor) sehingga mencapai hasil rekomendasi.

1.5.3 Metode Pengujian

Metode pengujian yang digunakan untuk mengukur akurasi hasil prediksi adalah RMSE (root mean square error). Metode pengujian ini digunakan untuk mengukur pengaruh nilai k terhadap akurasi hasil prediksi nilai rating sistem rekomendasi.

1.6 Sistematika Penulisan

Pada dasarnya penyusunan sistematika penulisan bertujuan untuk memudahkan pembaca dalam mengikuti apa yang dipaparkan dalam laporan skripsi ini. Sistematika penulisan skripsi ini adalah sebagai berikut :

BAB I PENDAHULUAN

Pada bab ini terdiri dari latar belakang masalah, perumusan masalah, batasan masalah, tujuan dan manfaat penelitian, metode penelitian, dan sistematika penulisan yang digunakan untuk menyusun penelitian.

BAB II LANDASAN TEORI

Pada bab ini membahas mengenai tinjauan pustaka, yang berisi dasar teori-teori yang menunjang dalam pembuatan sistem.

BAB III METODOLOGI PENELITIAN

Dalam bab ini tahapan yang dilakukan adalah implementasi dari metodologi penelitian serta pembahasan dari hasil evaluasi algoritma

BAB V **PENUTUP**

Bab ini adalah penutupan dari penulisan skripsi ini. Pada bab ini terdapat kesimpulan dari hasil yang telah diperoleh serta saran untuk pengembangan sistem rekomendasi ini untuk peneliti selanjutnya.

