BAB I
PENDAHULUAN

1.1 Latar Belakang

Di era 4.0 ini pertumbuhan data dan informasi yang sangat pesat membuat dibutuhkannya kemajuan teknologi dari database generasi berikutnya dan alat yang digunakan. Sebagian besar bisnis membutuhkan sistem rekomendasi layanan yang sudah digunakan oleh jutaan pengguna. Hari demi hari, banyaknya jumlah konsumen, produk dan informasi sudah berkembang dengan cepat, menghasilkan permasalahan analisa big data kepada layanan sistem rekomendasi. Akibatnya, layanan sistem rekomendasi yang masih konvensional akan mengalami masalah efisiensi dan skala yang terlalu besar saat menganalisa dan memproses data.

Masalah efisiensi ini juga lah yang sering menyebabkan kebimbangan kepada konsumen sehingga mengurungkan niat mereka untuk mencoba produk baru, dan yang akhirnya akan membuat berbagai produk yang mungkin akan disukai oleh konsumen tersebut tidak terkenal akIbat kurangnya rekomendasi yang diberikan. (Xiaoyuan Su dan Taghi M. Khosgoftaar.,2009) menjelaskan bahwa *"Collaborative Filtering"* (CF) adalah cara utama teknik rekomendasi yang telah di adopsi secara luas, walaupun yang merekomendasi dan yang di rekomendasi tidak mengenal satu sama lain, hasil rekomendasi yang di dapat cukup menarik. CF secara fundamental adalah jika user X dan Y menilai n dengan serupa, atau memiliki kebiasaan yang sama (membeli, menonton, mendengarkan) maka mereka akan menilai atau menggunakan barang dengan sedemikian pula.

Teknik CF menggunakan database preferensi kesukaan konsumen untuk memprediksi topik atau produk tambahan yang mungkin juga akan disukai oleh konsumen. Di banyak scenario CF, akan ada daftar *m* user $\{u_1,u_2,\ldots,u_m\}$ dan daftar n item $\{i_1,i_2,\ldots,i_m\}$, dan setiap user, u_i, mempunyai daftar item, Iu_i, yang sudah dinilai oleh user, atau dari preferensi yang disimpulkan dari perilaku mereka. Rating dapat diambil dari indikasi eksplisit, seperti skala 1-5, atau indikasi implisit, seperti pembelian atau pemilihan. Tetapi, ada beberapa kekurangan dalam teknik CF, seperti bahwasannya kemiripan data dasarnya hanya mengambil dari value yang umum yang menyebabkan tidak bisa diandalkan jika data yang dimiliki jarang-jarang (*Sparse*). Data yang hilang akan menyebabkan pendugaan parameter tidak tepat karena berkurangnya ukuran data.

Hal ini disebut *Data Sparsity*, yaitu terjadinya kekosongan data matriks user-item, yang disebabkan karena user merating dalam jumlah kecil dari jumlah item yang tersedia di dalam database.(Teguh Budianto dan Galih Hermawan.2013). Permasalahan *Data sparsity* muncul di berbagai situasi, seperti saat ada user baru yang baru menggunakan layanan, akan sulit untuk menemukan kesukaan yang sama karena informasi yang dimiliki belum begitu banyak. Film terbaru belum bisa untuk direkomendasikan sampai ada user yang merekomendasikannya, dan tidak semua user akan memberikan rekomendasi bagus karena kurangnya histori data rekomendasi mereka. Hal ini dapat mengurangi efektivitas dari layanan rekomendasi yang mengandalkan perbandingan rekomendasi user sehingga dapat mengeluarkan prediksi.

Untuk mendapatkan prediksi performa layanan system rekomendasi yang baik dari data yang kurang melalui teknik CF. Harus dilakukan imputasi, Salah satu metode imputasi yang digunakan adalah mengganti *missing* data dengan nilai rata-rata atau dengan modus tergantung dari jenis datanya. Pada data numerik digunakan untuk cara mengganti *missing*

data dengan nilai rata-rata, sedangkan untuk data kategorik maka digunakan cara mengganti *missing* data dengan nilai nearest neighbor.

Dan juga ada metode lain yang digunakan yaitu *Hot-deck* . *Hot- deck* Imputation umumnya mengacu pada Sequential *Hot-deck* Imputation, yang berarti bahwa set data diurutkan dan nilai-nilai yang hilang diimputasikan secara berurutan berjalan melalui observasi demi observasi. Pengurutan data menggunakan variabel prediktor dipilih berdasarkan hubungannya dengan variabel yang akan diimputasi (Grau, 2004). *Stochastic hot deck imputation* adalah metode hotdeck yang melibatkan pemilihan acak record donor dari kasus lengkap dalam data dan menggunakannya untuk mengisi nilai yang hilang untuk sebuah kasus yang tidak lengkap. Imputasi dilakukan secara stokastik, yang berarti bahwa pemilihan record donor bersifat acak dan tergantung pada distribusi record donor. Saat ini, metode dengan sifat teoritis yang lebih baik telah tersedia, tetapi metode Hot-deck Imputation masih cukup popular karena kesederhanaan dan kecepatannya.

Saya terdorong untuk menggunakan teknik imputasi mean dan Hot-Deck terhadap data yang kurang untuk membantu menghadapi masalah *sparsity*. Saya berfokus pada aspek utama kegiatan penelitian ini yakni mengimputasi value yang hilang pada data sistem rekomendasi agar masalah *Data Sparsity* dapat terpecahkan dengan data movielens 100k sebagai dataset penelitian yang digunakan .

Penelitian ini diharapkan dapat membantu sistem rekomendasi agar dapat mengelola informasi data rekomendasi secara efisien dan akurat. Peneliti juga berharap penelitian ini dapat menjadi referensi bagi peneliti lain dimasa mendatang, dan kekurangan dalam penelitian ini dapat di kritik dan diberi saran yang membangun.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang di paparkan, maka dirumuskan masalah yang akan dibahas dalam penelitian ini sebagai berikut :

(1) Bagaimana mengatasi masalah Data Sparsity menggunakan teknik imputation ?
(2) Bagaimana dampak positif dari data yang lengkap terhadap kualitas rekomendasi yang dihasilkan ?

1.3 Batasan Masalah

Dalam penelitian ini, masalah harus memiliki batasan masalah agar masalah yang di teliti tidak terlalu luas dan dapat sesuai dengan tujuan yang diharapkan.

(1) Ruang lingkup program untuk penelitian ini adalah program pengimputasian menggunakan dataset melalui website movielens.org .
(2) Metode imputasi yang digunakan untuk menghadapi masalah Data Sparsity pada sistem rekomendasi film adalah menggunakan metode mean, dan *Hot - Deck*.

1.4 Tujuan Penelitian

Penelitian ini bertujuan untuk:

(1) Menemukan jawaban dari permasalahan *Data Sparsity* untuk sistem rekomendasi.
(2) Memanfaatkan teknik imputasi yang digunakan untuk memberikan layanan rekomendasi yang lebih baik lagi.

1.5 Manfaat Penelitian

(1) Teknik imputasi yang digunakan dapat memperbaharui layanan sistem rekomendasi agar lebih baik lagi.
(2) Memberikan layanan yang lebih baik untuk konsumen agar mendapatkan rekomendasi yang sesuai.

1.6 Sistematika Penulisan

Secara garis besar penelitian ini terdiri dari 5 (lima) bab dengan sistematika penulisan sebagai berikut:

(1) **BAB I Pendahuluan**

Pada bab ini berisi latar belakang masalah, identifikasi masalah, rumusan masalah,batasan masalah, tujuan dan manfaat penelitian serta sistematika penelitian.

(2) **BAB II Tinjauan Pustaka**

Bab ini berisikan teori-teori dasar yang berhubungan dengan penelitian yangdilakukan oleh penulis atau berupa pengertian definisi yang diambil dari kutipan buku dan literatur review yang berhubungan dengan penelitian.

(3) **BAB III Metodologi Penelitian**

Bab ini menjelaskan metode-metode yang digunakan dalam tahap penelitian yang akan digunakan dalam melakukan imputasi terhadap dataset Movielens.org.

(4) **BAB IV Hasil Penelitian dan Pembahasan**

Bab ini berisikan tentang hasil dari imputasi berdasarkan metode yang telah digunakan dan hasil dari perhitungan.

(5) **BAB V Simpulan dan Saran**

Bab ini berisi simpulan dan saran-saran yang terkait dengan pembahasan dalam skripsi ini.

(6) **Daftar Pustaka**

(7) **Lampiran**