MIND (Multimedia Artificial Intelligent Networking Database) Journal

MIND Journal | ISSN (p): 2528-0015 | ISSN (e)**:** 2528-0902 | Vol. 9 | No. 1 | Halaman 113 - 123
DOI : https://doi.org/10.26760/mindjournal.v9i1.113-123 Juni 2024

Pemanfaatan Metode *Collaborative Filtering* dengan Algoritma *KNN* pada Sistem Rekomendasi Produk

KURNIA RAMADHAN PUTRA , ILHAM FATHUR RAHMAN

Program Studi Sistem Informasi, Institut Teknologi Nasional Bandung
Email: kurniaramadhan@itenas.ac.id

Received 1 November 2023 **|** *Revised* 1 Februari 2024 **|** *Accepted* 15 Februari 2024

ABSTRAK

Salah satu permasalahan customer pada e-commerce adalah sulitnya menemukan produk yang diinginkan untuk dibeli. Sistem rekomendasi mampu menangani permasalahan tersebut dengan cara mengalisis data profil customer untuk menyaring produk yang sesuai dengan profil customer kemudian merekomendasikannya kepada customer tersebut. Untuk mengetahui hubungan antara produk dengan pengguna maka dapat memanfaatkan sistem rekomendasi. Ada beberapa permasalahan pada sistem rekomendasi yaitu sparsity data, missing value, dan duplikasi data yang sering ditemukan pada data berbasis rating seperti pada e-commerce. Untuk menyelesaikan masalah ini, maka diusulkan metode Item-based Collaborative Filtering dan algoritma K-Nearest Neighbor (KNN) dengan hasil evaluasi nilai MAE sebesar 1,05 dan RMSE sebesar 1,36 yang mampu menangani sistem rekomendasi dengan baik dengan tingkat kesalahan yang kecil.

***Kata kunci**: recommendation system, item-based collaborative filtering, KNN, Sparsity Data, Cold-Start.*

ABSTRACT

In e-commerce, one common customer problem is difficulty in finding the product they want to buy. This issue can be addressed through a recommendation system, which analyzes customer profile data to filter products that match the customer's profile and then recommends them. One way to establish the relationship between products and users is by using a recommendation system. However, recommendation systems often encounter problems such as data sparsity, missing values, and data duplication, particularly in rating-based data. To address these issues, the Item-based Collaborative Filtering method and the K-Nearest Neighbor (KNN) algorithm are proposed. Evaluation results show that these methods have MAE values of 1.05 and RMSE of 1.36, indicating their effectiveness in handling the recommendation system with a low error rate.

***Keywords**: recommendation system, collaborative filtering, item-based CF, KNN*

1. PENDAHULUAN

Khususnya di bidang bisnis, teknologi saat ini sangat penting dan berkembang pesat yang mana pelaku bisnis dapat melakukan transaksi dengan lebih cepat tidak terbatas oleh ruang dan waktu serta mampu menjangkau konsumen yang lebih besar **(Februariyant, dkk., 2021)**. Salah satu pemanfaatan teknologi di bidang bisnis adalah *e-commerce* yaitu teknologi yang memungkinkan orang-orang untuk melakukan transaksi seperti aktivitas penjualan, pembelian, dan pemasaran barang maupun jasa menggunakan aplikasi yang terhubung ke jaringan *internet* **(Oktora & Susanty, 2013)**.

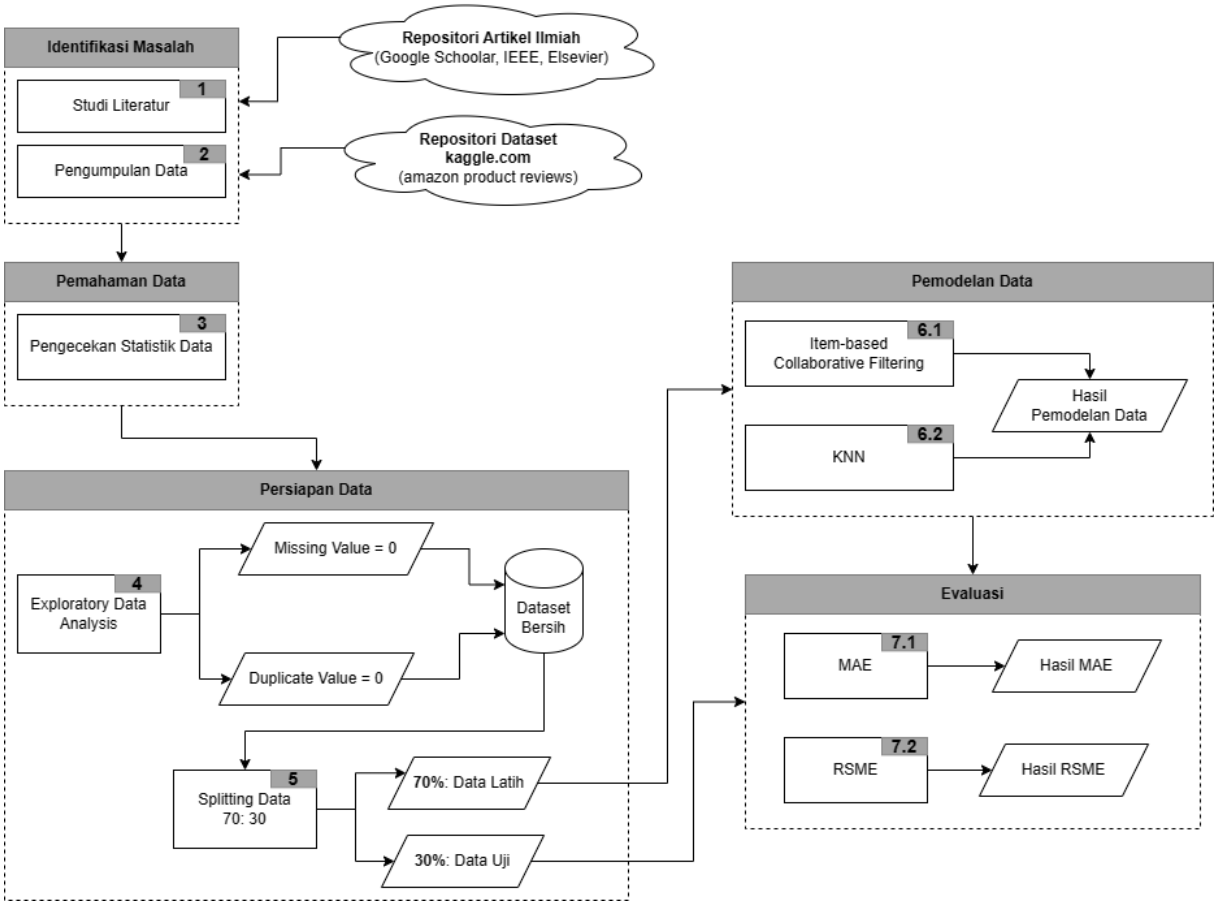
Ketika mengunjungi e-*commerce,* customer sering mengalami kesulitan dan kebingungan untuk menemukan produk yang akan dibeli **(Knijnenburg, dkk., 2012)**. Sistem rekomendasi membantu dalam menganalisis data calon pelanggan kemudian merekomendasikan produk berdasarkan data calon pelanggan tersebut **(C. S. D. Prasetya, 2017)**. Berbagai *platform* bisnis online seperti Amazon dan Ebay telah menggunakan sistem rekomendasi untuk membantu meningkatkan penjualan produk dan membangun loyalitas pelanggannya. **(Brusilovsky & Kobsa, 2007)**. Selain itu juga dapat memberikan rekomendasi produk yang relevan sesuai kebutuhan customer **(Imandoust & Bolandraftar, 2013)**. Ada beberapa data masukan pada sistem rekomendasi yaitu berdasarkan *rating* produk, produk yang paling sering dibeli, produk yang paling sering diklik, dan produk yang memiliki kategori yang sama dengan produk yang sudah pernah dibeli sebelumnya **(Felfernig, dkk., 2007)**.

Content-based filtering, *collaborative filtering*, dan *hybrid filtering* adalah metode-metode yang dipertimbangkan untuk digunakan pada sistem rekomendasi yang mana *collaborative filtering* dianggap paling cocok untuk sistem rekomendasi produk berbasis rating **(Lubis, dkk., 2020)**. Item-based dan User-based adalah dua kategori dari metode collaborative filtering yang sering digunakan dalam sistem rekomendasi, namun memiliki masalah *sparsity data* **(Vozalis & Margaritis, 2003).** *Sparsity* adalah kondisi di mana terdapat banyaknya data atau nilai rating yang kosong **(Ajipradana, 2017)**. Sehingga, *sparsity* membuat sistem rekomendasi tidak mampu memberikan rekomendasi kepada pengguna dengan akurasi yang baik. Masalah *sparsity* tersebut dapat ditangani menggunakan pendekatan *item-based collaborative filtering* untuk mengisi data rating yang kosong **(Sarwar, dkk., 2000)**.

Algoritma KNN dimanfaatkan untuk mengukur kesamaan antar item yang akan diprediksi pada data uji kemudian dibandingkan data latih menggunakan metode perhitungan jarak seperti *Cosine Similarity* atau *Jaccard Similarity* yang mana produk yang direkomendasikan kepada pelanggan adalah produk yang jaraknya dekat satu sama lain sehingga hasil rekomendasi menjadi lebih akurat **(Rahardja, dkk., 2019)**. Algoritma KNN membantu untuk menemukan nilai kesamaan antar item kemudian menghitung nilai prediksi dan membatasi jumlah item yang direkomendasikan sebagai *top-n item*.

2. METODOLOGI PENELITIAN

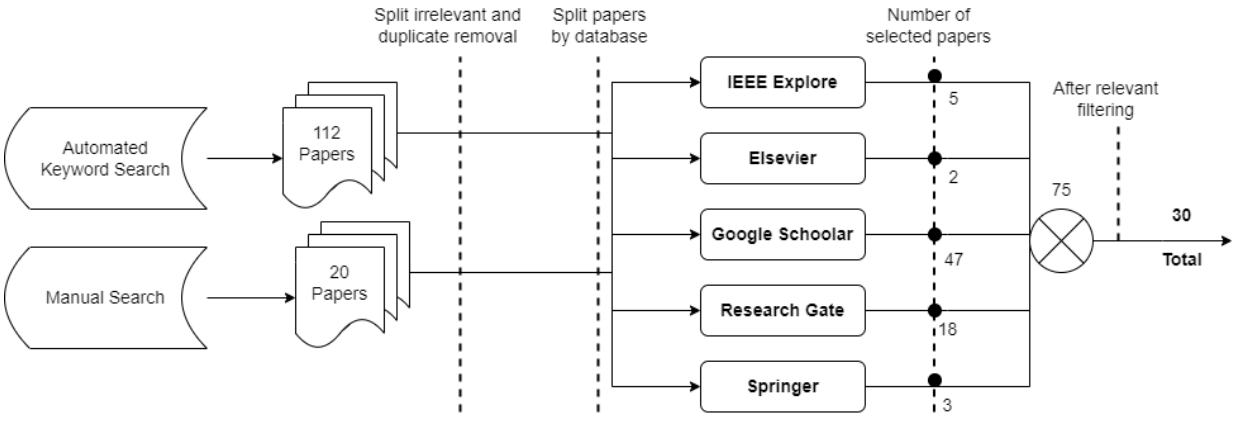
Penelitian dimulai dengan studi litelatur kemudian dilanjutkan dengan mengumpulan data. Setelah data dikumpulkan maka akan dilakukan data understanding dengan teknik *exploratory data analysis* (EDA) untuk mencari nilai data yang hilang atau duplikat. Data yang sudah bersih dilanjutkan pada tahap pemodelan dan terakhir dilakukan evaluasi. Tahapan-tahapan tersebut sebagai berikut: (1) studi litelatur; (2) pengumpulan data; (3) pemeriksaan deskripsi data; (4) *exploratory data analysis*; (5) *data splitting*; (6) pemodelan data menggunakan algoritma KNN; (7) evaluasi kinerja model yang mana alur metodologi penelitiannya ditunjukkan pada Gambar 1.



Gambar 1. Gambaran Alur Metodologi Penelitian

2.1 Studi Literatur

Beberapa repositori jurnal yang digunakan untuk melakukan studi literatur yaitu Google Schoolar, IEEE Explore, Elsevier, dan Research Gate. Ada 30 artikel ilmiah yang paling relevan digunakan pada penelitian ini dengan periode tahun 2015 sampai dengan 2022. Gambar 2 menunjukkan studi literatur yang dilakukan pada repositori jurnal ilmiah.



Gambar 2. Studi Literatur dari Repositori Jurnal Ilmiah

Operator dan kata kunci yang digunakan untuk menemukan referensi penelitian ini adalah "Recommendation System" AND "Item-based" AND "K-nearest neighbors".

2.2 Pengumpulan Data

Dataset yang digunakan yaitu *Amazon Product Reviews* dari bulan Mei tahun 1996 sampai dengan bulan Juli tahun 2014 yang dapat diakses pada kaggle.com. Dataset disajikan dalam format .csv dengan 7.824.482 baris data dan berisikan 3 kolom numerik yaitu userId, productId, dan rating. Tabel 1 menunjukkan bentuk dataset Amazon Product Reviews tersebut.

Tabel 1. *Dataset Amazon Product Reviews*

Nomor	User Id	Product Id	Rating
0	AKM1MP6P0OYPR	0132793040	5.0
1	A2CX7LUOHB2NDG	0321732944	5.0
2	A2NWSAGRHCP8N5	0439886341	1.0
3	A2WNBOD3WNDNKT	0439886341	3.0
4	A1GI0U4ZRJA8WN	0439886341	1.0
....	...	...	...
7824479	A322MDK0M89RHN	BT008UKTMW	5.0
7824480	A1MH90R0ADMIK0	BT008UKTMW	4.0
7824481	A10M2KEFPEQDHN	BT008UKTMW	4.0
7824482	A2G81TMIOIDEQQ	BT008V9J9U	5.0

2.3 Pengecekan Statistik Data

Pengecekan statistik data numerik penting dilakukan agar dapat diproses secara matematis menggunakan operasi mean, standar deviasi, min, quartil, mix, dan max pada fitur rating.

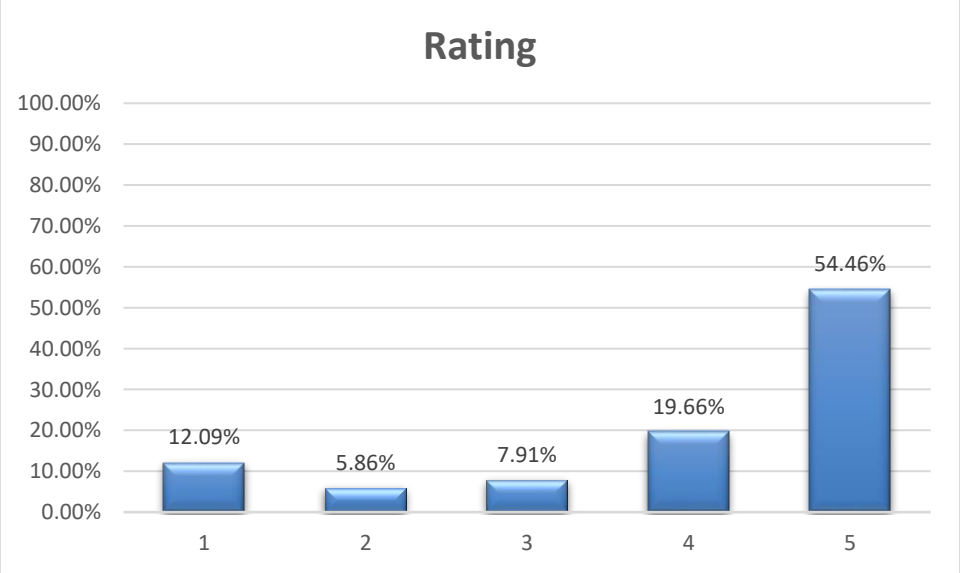
Tabel 2. Ringkasan Statistik Data Numerik

Fitur	Mean	Min	Q1	Q2	Q3	Max
Rating	3.97338	1.00000	3.00000	5.00000	5.00000	5.00000

Penjelasan dari Tabel 2 adalah sebagai berikut.

a. *Mean* atau rata-rata dari jumlah *rating* dengan nilai 3,97338
b. *Minimum* atau nilai terendah *rating* ialah 1.00000
c. *Maximum* atau nilai terbesar *rating* 5.00000
d. Kuartil membagi seluruh distribusi frekuensi kedalam empat bagian yang sama besar dengan nilai:
   - Q1 = 25% = 3.00000
   - Q2 = 50% = 5.00000
   - Q3 = 75% = 5.00000

Gambar 3 menunjukkan distribusi rating dataset *Amazon Product Reviews* yang mana distribusi nilai rating paling banyak adalah 5 dengan persentase sebesar 54,46% dan nilai rating paling sedikit adalah nilai 2 dengan persentase sebesar 5,86%



Gambar 3. Distribusi Rating pada Dataset Amazon Product Reviews

2.4 *Exploratory Data Analysis*

EDA merupakan langkah dalam proses analisis data, dimana sejumlah teknik digunakan untuk lebih memahami kumpulan data yang akan digunakan apakah ada data yang hilang atau data duplikat.

Missing Value

Missing value adalah keadaan suatu kolom yang tidak memiliki nilai di dalamnya. *Missing value* tidak sama dengan nol, karena nol juga termasuk sebuah nilai. Biasanya pada data, jika ada *missing value* maka disimbolkan dengan "NaN" atau karakter spesial "?". Tidak jarang juga bahwa data yang tidak terdeteksi akan ditampilkan sebagai sel kosong yang tidak ada nilainya sama sekali. Hasil pengecekan *missing value* pada dataset *Amazon Product Reviews* dapat dilihat pada Tabel 3, dapat dijelaskan bahwa pada dataset tersebut tidak ditemukan adanya *missing value* pada fitur User id, Product Id, dan Rating.

Tabel 3. Pemeriksaan *Missing Value* pada Dataset

Nomor	Fitur	Jumlah Data Hilang	Total Data	Persentase Missing Value
1	User Id	0	7.824.482	0%
2	Product Id	0	7.824.482	0%
3	Rating	0	7.824.482	0%

Data Duplikat

Untuk mengidentifikasi data duplikat menggunakan fungsi *duplicated()* yang mengembalikan nilai *boolean* yaitu *true* atau *false* untuk setiap item. Kemudian dengan menggunakan fungsi *any()* maka dapat mengetahui apakah ada nilai yang duplikat pada *dataframe* yang diuraikan pada Tabel 4.

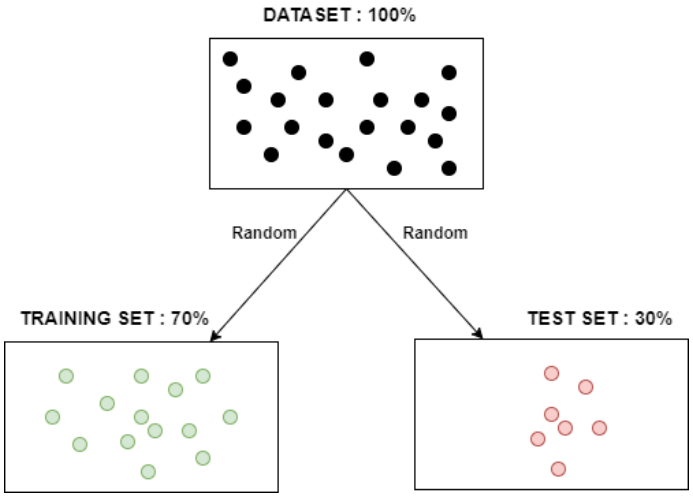
Tabel 4. Pemeriksaan Data Duplikat pada Dataset

Nomor	Fitur	Jumlah Data Duplikat
1	User Id	0
2	Product Id	0
3	Rating	0

Berdasarkan Tabel 4 tersebut, dapat dilihat bahwa pada *dataset* tidak ditemukan adanya duplikasi data pada User id, Product Id, dan Rating.

2.5 *Data Splitting*

Untuk melakukan validasi, dilakukan splitting data dengan membagi data secara acak menjadi dua bagian yaitu data latih untuk membangun model dan data uji untuk mengevaluasi model **(Dwi Untari, dkk., 2010)**. Penelitian ini menggunakan teknik pembagian data 70% untuk data latihan dan 30% untuk data uji, seperti yang ditunjukkan pada Gambar 4.



Gambar 4. *Data Splitting* Antara Data Latih dengan Data Uji

2.6 Pemodelan Data

K-Nearest Neighbor

Untuk menangani kasus klasifikasi, regresi, atau prediksi, digunakan metode metode *item-based collaborative filtering* dan algoritma KNN. Proses yang dilakukan oleh algoritma KNN adalah sebagai berikut:

1. Pertama perlu ditentukan nilai K, yang merupakan jumlah tetangga terdekat digunakan untuk membuat prediksi.
2. Tentukan jarak antara data pada data uji dengan data latih menggunakan metode Cosine Similarity sebagai metrik untuk menghitung jarak dengan formula yang ditunjukkan pada Persamaan (1).

$$Cos\ Sim = cos(\theta) = \frac{\mathbf{A}\ .\ \mathbf{B}}{||\mathbf{A}||\ ||\mathbf{B}||} = \frac{\sum_{i=1}^{n} AiBi}{\sqrt{\sum_{i=1}^{n} Ai}\sqrt{\sum_{i=1}^{n} Bi}} \quad (1)$$

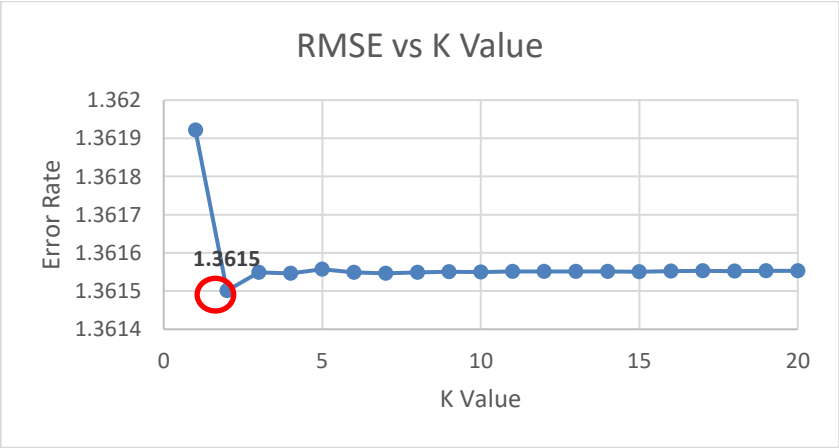
Dimana:
$Cos\ Sim = cos(\theta) = Cosine\ Similarity$
$A = Vektor\ A$
$B = Vektor\ B$

$Ai = Jumlah\ Vektor\ A$
$Bi = Jumlah\ Vektor\ B$
$n = Banyak\ Data$

3. Identifikasi tetangga terdekata dengan memilih titik K titik yang terdekat dengan data uji.
4. Tentukan kelas dan nilai prediksi yaitu mayoritas suara dari tetangga-tetangga terdekat.
5. Setelah menentukan kelas atau nilai prediksi, kemudian dapat mengklasifikasikan data uji ke dalam kelas tersebut untuk menghasilkan nilai prediksi.
6. Evaluasi model menggunakan MAE dan RSME.

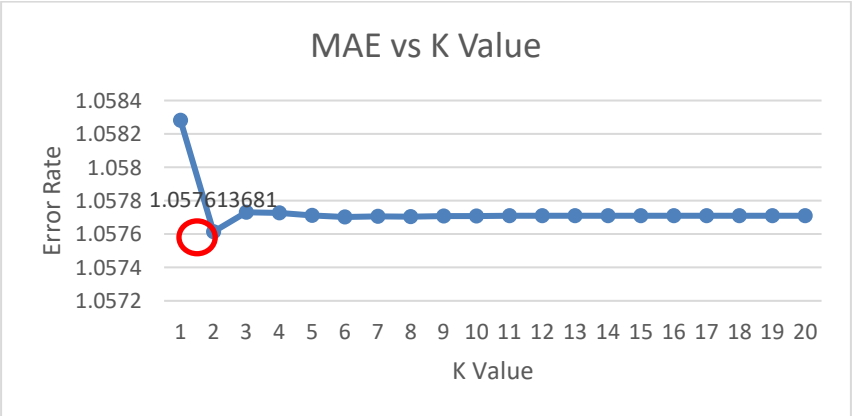
Menentukan K Optimal

Perhitungan dilakukan untuk menemukan nilai K terbaik dengan menggunakan metrik RMSE dan MAE. Nilai K berkisar dari 1 hingga 20, dan hasilnya ditunjukkan pada Gambar 5 dan Gambar 6.



Gambar 5. Menentukan Nilai K Optimal Menggunakan RSME

Berdasarkan hasil metrik RSME diambil nilai yang paling kecil yaitu sebesar 1.3615, sehingga seperti yang terlihat pada Gambar 5 bahwa nilai tersebut berada pada k = 2.



Gambar 6. Menentukan Nilai K Optimal Menggunakan MAE

Berdasarkan hasil metrik MAE juga terlihat bahwa nilai terkecil yaitu sebesar 1.057613681 berada pada k = 2. Dari hasil RSME dan MAE dapat disimpulkan bahwa nilai K Optimal yang digunakan yaitu 2.

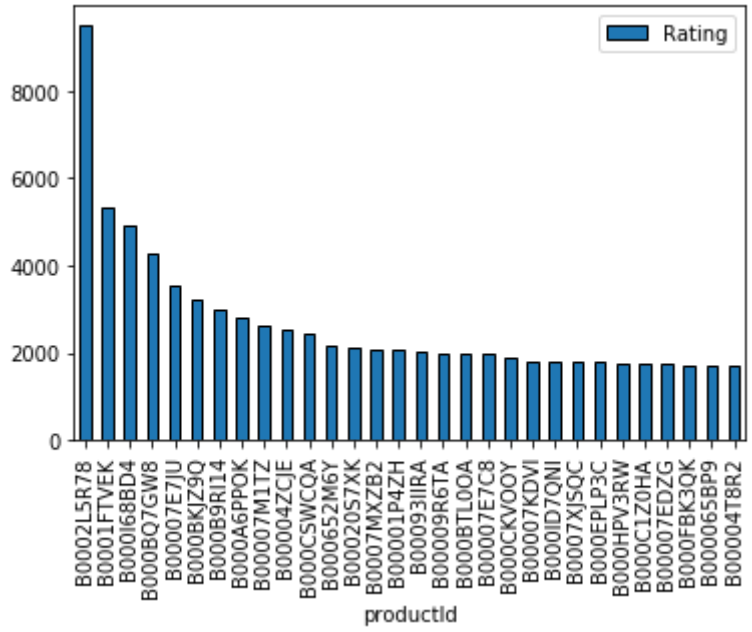
Menangani *Sparsity Data*

Pada penelitian ini *sparsity data* ditangani dengan *Popularity Based Recommendation* yaitu mengisi nilai rating pada data rating yang kosong berdasarakan jumlah rating paling banyak yang telah diberikan oleh user sebelumnya. Langkah pertama yang dilakukan adalah memeriksa rating setiap produk kemudian diurutkan berdasarkan rating paling banyak menggunakan fungsi *head()*. Data dari hasil pengembalian fungsi tersebut diuraikan pada Tabel 5.

Tabel 5. Daftar Produk dengan Rating Terbanyak

Nomor	Product Id	Jumlah Rating
1	B0002L5R78	9487
2	B0001FTVEK	5345
3	B000I68BD4	4903
4	B000BQ7GW8	4275
5	B00007E7JU	3523

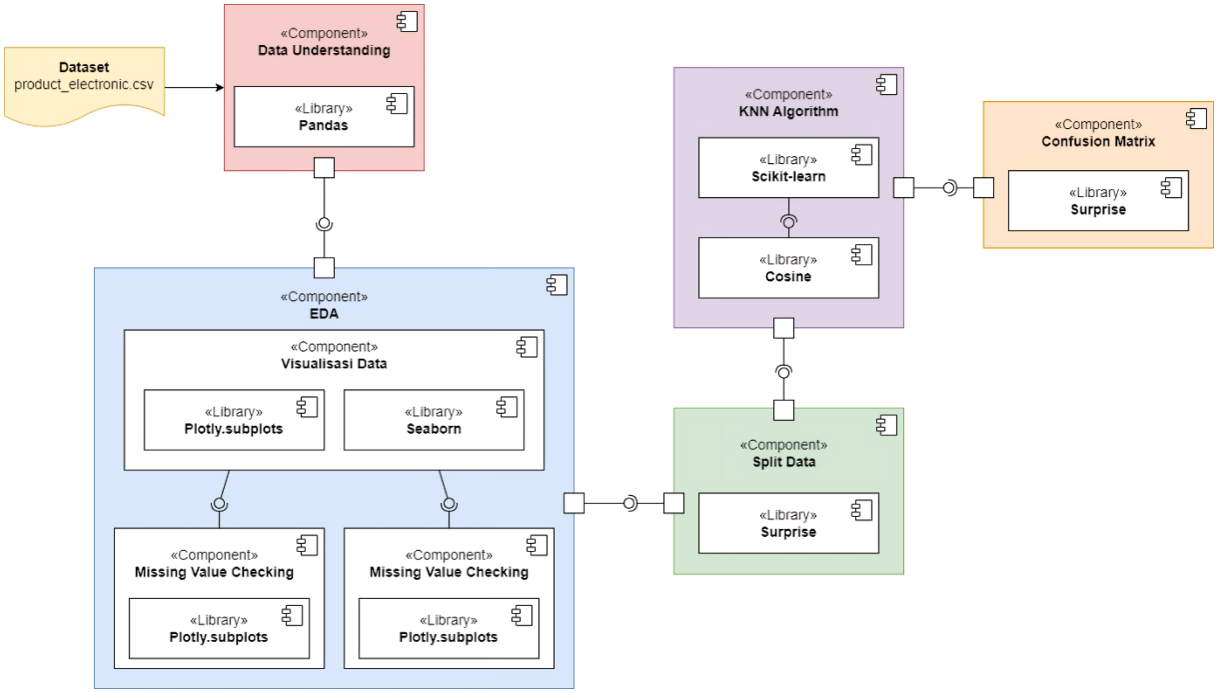
Selanjutnya dilakukan pengecekan secara visual menggunakan fungsi *subplots()* untuk menampilkan data rating seperti yang ditunjukkan pada Gambar 7. Dari hasil ploting, ditunjukkan bahwa *rating* terbanyak dijadikan sebagai acuan untuk memberikan rekomendasi produk kepada customer yang baru atau belum memberikan rating terhadap sebuah produk.



Gambar 7. Komponen Sistem Rekomendasi

Komponen Sistem Rekomendasi

Gambar 8 adalah *component diagram* yang menunjukkan keterkaitan antar komponen di dalam sistem rekomendasi.



Gambar 8. Komponen Sistem Rekomendasi

2.7 Evaluasi Model

Untuk menilai setiap parameter tuning, nilai metrik MAE dan RMSE dari model KNN dibandingkan dengan pemilihan nilai K = 2 dan perbandingan data latih dengan data uji sebesar *splitting* 70:30 seperti yang ditunjukkan pada Tabel 6.

Tabel 6. Hasil Evaluasi Model KNN Menggunakan MAE dan RSME

MAE	RSME
1.05791368103754	1.36150105197105

Jumlah data latih dapat mempengaruhi nilai MAE dan RMSE, karena semakin besar data latihnya maka akan memberikan hasil prediksi yang lebih dekat dengan hasil aktualnya dengan memberikan nilai MAE dan RMSE yang lebih rendah, sehingga menunjukkan hasil prediksi yang lebih baik.

3. KESIMPULAN

Metode *Item-based Collaborative Filtering* dengan algoritma KNN memberikan hasil yang baik, dengan nilai K = 2 dan *data splitting* 70:30. Penelitian ini menghasilkan nilai MAE sebesar 1.05791368103754 dan nilai RSME sebesar 1.36150105197105, yang menunjukkan bahwa semakin kecil nilai error prediksi atau mendekati nol, semakin akurat model yang dihasilkan. Oleh karena itu, penelitian ini dapat dijadikan rujukan dalam penelitian yang akan datang tentang sistem rekomendasi.

UCAPAN TERIMA KASIH

Banyak pihak bekerja sama dan memberikan dukungan untuk memastikan bahwa penelitian ini berjalan dengan baik dan lancar. Kami berterima kasih kepada Itenas, terutama Program Studi Sistem Informasi, LPPM Itenas, dan Teknik Elektro, karena telah menyelenggarakan Seminar Nasional Energi dan Otomasi (SNETO) Tahun 2023.

DAFTAR RUJUKAN

Abdallah, Z. S., & Webb, G. I. (2017). Encyclopedia of Machine Learning and Data Mining. *Encyclopedia of Machine Learning and Data Mining*, *September 2018*. https://doi.org/10.1007/978-1-4899-7687-1

Ajipradana, F. A. (2017). Sistem Rekomendasi Film Menggunakan Algoritma Item-based Collaborative Filtering dan Basis Data Graph. *Universitas Diponegoro*.

Brusilovsky, P., & Kobsa, A. (2007). The Adaptive Web. In *The Adaptive Web* (Vol. 4321, Issue January 2007). https://doi.org/10.1007/978-3-540-72079-9

Dwi Untari, Hastuti, K., Hidayat, E. Y., Dwi Untari, Limão, N., & Gaol, N. Y. L. (2010). Data Mining untuk Menganalisa Prediksi Mahasiswa Berpotensi Non-Aktif Menggunaka Metode Decision Tree C4.5. *Fakultas Ilmu Komputer Universitas Dian Nuswantoro*, *2013*(November), 31–48.

Februariyanti, H., Laksono, A. D., Wibowo, J. S., & Utomo, M. S. (2021). Implementasi Metode Collaborative Filtering Untuk Sistem Rekomendasi Penjualan Pada Toko Mebel. *Jurnal Khatulistiwa Informatika*, *IX*(I), 43–50.

Felfernig, A., Friedrich, G., & Schmidt-Thieme, L. (2007). Introduction to the IEEE Intelligent Systems Special Issue: Recommender Systems. *IEEE Intelligent Systems*, *22*(3), 18–21.

Imandoust, S. B., & Bolandraftar, M. (2013). Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events : Theoretical Background. *Int. Journal of Engineering Research and Applications*, *3*(5), 605–610.

Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., & Newell, C. (2012). Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, *22*(4–5), 441–504. https://doi.org/10.1007/s11257-011-9118-4

Lankford, S. (2020). Effective Tuning of Regression Models using an Evolutionary Approach: A Case Study. *ACM International Conference Proceeding Series*, 102–108. https://doi.org/10.1145/3442536.3442552

Lubis, Y. I., Napitupulu, D. J., & Dharma, A. S. (2020). Implementation of Hybrid Filtering (Collaborative and Content-based) Methods for the Tourism Recommendation System. *12th Conference on Information Technology and Electrical Engineering*, 6–8.

Neighbor, M. M. K. (2018). *Penerapan data mining untuk prediksi penjualan produk elektronik terlaris menggunakan metode k-nearest neighbor*.

Oktora, R., & Susanty, W. (2013). Perancangan Aplikasi E-Commerce Dengan Sistem Rekomendasi Item-Based Collaborative Filltering. *EXPERT: Jurnal Manajemen Sistem Informasi Dan Teknologi*, *3*(1). https://doi.org/10.36448/jmsit.v3i1.477

Prasetya, C. S. D. (2017). Sistem Rekomendasi Pada E-Commerce Menggunakan K-Nearest Neighbor. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *4*(3), 194. https://doi.org/10.25126/jtiik.201743392

Rahardja, C. A., Juardi, T., & Agung, H. (2019). Implementasi Algoritma K-Nearest Neighbor Pada Website Rekomendasi Laptop. *Jurnal Buana Informatika*, *10*(1), 75. https://doi.org/10.24002/jbi.v10i1.1847

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2000). Application of Dimensionality Reduction in Recommender System A Case Study Technical Report CS-TR 00-043. *Computer Science and Engineering Dept., University of Minnesota*, *December 2012*.

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, *181*(2019), 526–534.

Vozalis, E., & Margaritis, K. (2003). Analysis of Recommender Systems' Algorithms. *Hercma*, *September 2003*, 1–14. http://lsa-svd-application-for-analysis.googlecode.com/svn-history/r72/trunk/LSA/Other/LsaToRead/hercma2003.pdf