JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 11, No. 1, March 2026, Pp.458-466

OPTIMALISASI PENANGANAN SPARSITY MENGGUNAKAN RANDOM FOREST, DEEP LEANING, DAN HOT-DECK IMPUTATION

Rafli Banu Satrio*[1)], Sri Lestari[2)], Hendra Kurniawan[3)], Sushanty Saleh[4)]

1. Teknik Informatika, Ilmu Komputer, Institut Informatika dan Bisnis Darmajaya, Indonesia
2. Teknik Informatika, Ilmu Komputer, Institut Informatika dan Bisnis Darmajaya, Indonesia
3. Sains Data, Ilmu Komputer, Institut Informatika dan Bisnis Darmajaya, Indonesia
4. Sistem Informasi, Ilmu Komputer, Institut Informatika dan Bisnis Darmajaya, Indonesia

Article Info

Kata Kunci: Sparsity; Imputation; Missing Value; Random Forest; Deep-Learning; Hot-Deck

Keywords: *Sparsity; Imputation; Missing Value; Random Forest; Deep-Learning; Hot-Deck*

Article history:
Received 20 April 2025
Revised 21 May 2025
Accepted 5 June 2025
Available online 1 May 2026

DOI :
https://doi.org/10.29100/jipi.v11i1.7692

* Corresponding author.
Sri Lestari
E-mail address:
srilestari@darmajaya.ac.id

ABSTRAK

Sparsitas data dalam sistem rekomendasi menjadi tantangan utama yang dapat menurunkan akurasi prediksi serta relevansi saran yang diberikan kepada pengguna. Penelitian ini bertujuan untuk membandingkan efektivitas tiga metode imputasi dalam mengatasi masalah sparsitas, yaitu Random Forest Imputation, Deep Learning-Based Imputation, dan Hot-Deck Imputation. Evaluasi dilakukan dengan menggunakan metrik Root Mean Square Error (RMSE) pada berbagai tingkat sparsitas data. Hasil eksperimen menunjukkan bahwa Random Forest Imputation secara konsisten menghasilkan nilai RMSE terendah di seluruh kondisi sparsitas yang diuji. Pada tingkat sparsitas 20%, metode ini lebih unggul dengan selisih RMSE hingga 0.443 dibandingkan Deep Learning-Based Imputation dan 0.338 dibandingkan Hot-Deck Imputation. Perbedaan performa antar metode semakin signifikan seiring meningkatnya tingkat sparsitas, di mana Random Forest Imputation tetap menunjukkan keunggulan paling jelas. Kesimpulannya, Random Forest Imputation terbukti paling efektif dalam menangani data sparsitas tinggi, sehingga mampu meningkatkan akurasi prediksi dan kualitas rekomendasi dalam sistem rekomendasi berbasis data yang tidak lengkap.

ABSTRACT

Data sparsity in recommendation systems poses a significant challenge that can reduce the accuracy of predictions and the relevance of suggested items to users. This study aims to compare the effectiveness of three imputation methods—Random Forest Imputation, Deep Learning-Based Imputation, and Hot-Deck Imputation—in addressing sparsity issues. The evaluation was conducted using the Root Mean Square Error (RMSE) metric across varying levels of data sparsity. Experimental results demonstrate that Random Forest Imputation consistently achieves the lowest RMSE values under all tested sparsity conditions. At 20% sparsity, this method outperforms Deep Learning-Based Imputation by a margin of up to 0.443 and Hot-Deck Imputation by up to 0.338 in RMSE. The performance gap between methods becomes more pronounced as sparsity increases, with Random Forest Imputation maintaining clear superiority. In conclusion, Random Forest Imputation proves to be the most effective approach for handling high data sparsity, thereby improving prediction accuracy and the overall quality of recommendations in systems reliant on incomplete datasets.

.

I. PENDAHULUAN

Sistem rekomendasi merupakan komponen krusial dalam ekosistem digital kontemporer yang berperan dalam memfasilitasi penyaringan informasi serta pemilihan konten secara efektif oleh pengguna. Sistem ini banyak diterapkan pada berbagai platform, termasuk layanan streaming seperti Netflix dan Spotify, serta platform e-commerce seperti Amazon [1], [2]. Seiring dengan meningkatnya volume data dan jumlah pengguna, sistem rekomendasi menghadapi tantangan signifikan berupa sparsity, yaitu keterbatasan interaksi antara pengguna dan item dalam data. Contohnya, pada Netflix, penurunan kualitas rekomendasi akibat sparsity berpotensi menurunkan

retensi pengguna, yang pada gilirannya berdampak negatif terhadap pendapatan platform [3]. Kondisi sparsity ini menghambat kemampuan sistem dalam mengidentifikasi pola preferensi pengguna, sehingga menurunkan relevansi rekomendasi sekaligus meningkatkan beban komputasi akibat hasil yang kurang tepat sasaran [4].

Berbagai pendekatan telah dikembangkan untuk mengatasi masalah sparsity, salah satunya adalah melalui teknik imputasi nilai hilang (missing value imputation), yang bertujuan untuk memperkirakan nilai-nilai yang tidak tersedia dalam data. Metode-metode seperti Hot-Deck Imputation, Deep Learning-Based Imputation, dan Random Forest Imputation (RFI) telah banyak diterapkan dalam studi sebelumnya [5], [7]. Secara khusus, Random Forest Imputation (RFI) menonjol karena kemampuannya dalam menangkap pola kompleks serta menghasilkan prediksi yang stabil melalui pendekatan ensemble pohon keputusan [8]. Meskipun demikian, metode ini memiliki keterbatasan intrinsik yang dapat memengaruhi kemampuan generalisasi model, terutama dalam konteks sparsity yang ekstrem. Keterbatasan tersebut meliputi ketergantungan pada pemilihan hyperparameter, seperti jumlah pohon, kedalaman maksimum, dan ukuran subset fitur, potensi overfitting pada data pelatihan yang terbatas atau tidak representatif, serta kurangnya efisiensi pada dataset berskala besar, misalnya MovieLens versi 1M dan 10M [9], [10]. Selain itu, tingkat interpretabilitas model RFI yang rendah juga menjadi tantangan dalam penerapannya pada sistem rekomendasi yang membutuhkan transparansi.

Penelitian ini bertujuan untuk mengevaluasi efektivitas Random Forest Imputation (RFI) dalam mengatasi masalah sparsity pada sistem rekomendasi berbasis rating, dengan menggunakan dataset MovieLens sebagai studi kasus. Dataset ini dipilih karena telah menjadi benchmark standar dalam penelitian sistem rekomendasi dan tersedia dalam beberapa versi dengan tingkat sparsity yang bervariasi, sehingga sangat sesuai untuk menguji ketahanan metode imputasi. Penelitian ini secara khusus membandingkan performa RFI dengan metode Deep Learning-Based Imputation dan Hot-Deck Imputation menggunakan metrik Root Mean Square Error (RMSE) sebagai indikator akurasi [11]. Selain itu, studi ini juga menyoroti trade-off antara akurasi prediksi dan efisiensi komputasi, serta membahas keterbatasan dari masing-masing metode yang diuji.

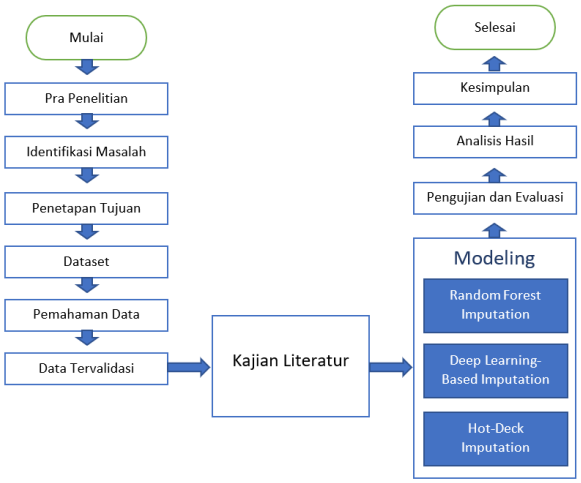
Kontribusi teoretis dari penelitian ini terletak pada pengayaan kerangka konseptual sistem rekomendasi berbasis kolaboratif, dengan menunjukkan bahwa teknik imputasi tidak hanya berfungsi untuk memperbaiki kualitas prediksi, tetapi juga memperkuat ketahanan model terhadap data sparsity. Oleh karena itu, teknik imputasi dapat dipandang bukan sekadar solusi pra-pemrosesan (preprocessing), melainkan juga sebagai strategi augmentasi data yang memperkuat sinyal preferensi pengguna. Dalam konteks sistem rekomendasi hibrid, pendekatan imputasi bahkan dapat dimaknai sebagai bentuk integrasi awal antara data eksplisit dan implisit, sehingga memperluas potensi penerapannya pada berbagai domain.

Dari sudut pandang praktis, hasil penelitian ini diharapkan dapat memberikan wawasan strategis bagi pengembang sistem rekomendasi di sektor hiburan, edukasi, maupun e-commerce. Peningkatan akurasi sistem rekomendasi berpotensi meningkatkan efisiensi platform digital dengan mengurangi pemborosan komputasi (computational waste) akibat rekomendasi yang tidak relevan. Pada akhirnya, hal ini dapat berkontribusi pada keberlanjutan digital (digital sustainability) melalui penghematan sumber daya dan energi [12].

II. METODE PENELITIAN

Dengan menggunakan pendekatan komparatif, penelitian ini menganalisis data kuantitatif yang dikumpulkan secara khusus dari situs web MovieLens selama periode tujuh bulan (September 1997 - April 1998). Tahapan penelitian secara lengkap dapat dilihat pada ilustrasi Gambar 1.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 11, No. 1, March 2026, Pp.458-466



Gambar 1. Alur Penelitian

Kami mengimplementasikan penelitian ini menggunakan lingkungan pengembangan Google Colab dengan bahasa pemrograman Python untuk melakukan eksplorasi data dan pengembangan model secara berulang[11], [12].

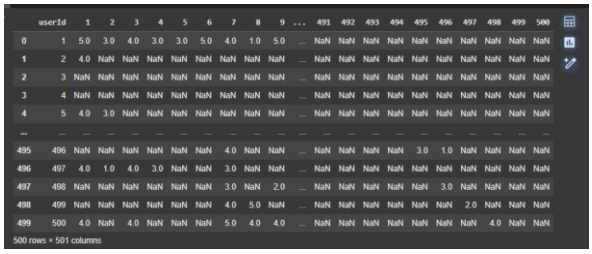
A. Persiapan Data

Data yang digunakan dalam penilitian ini terdiri dari 100.000 penilaian (1-5) yang diberikan oleh 943 pengguna pada 1682 film, di mana setiap pengguna telah memberikan penilaian pada setidaknya 20 film dan atribut yang digunakan untuk dilakukan imputasi berupa Pengguna (*User)*, Film (*Movie*) dan Nilai (*Rating*).

B. Pra-pemrosesan Data

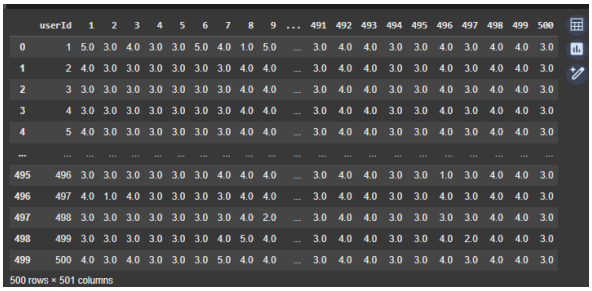
Setelah memperoleh *dataset,* langkah selanjutnya melakukan *preprocessing* dengan tahapan sebagai berikut:

1. Membuat Pivot Table pada dataset MovieLens untuk menghasilkan matriks berukuran [300x300], [500x500], [800x800] dan [943x1682].



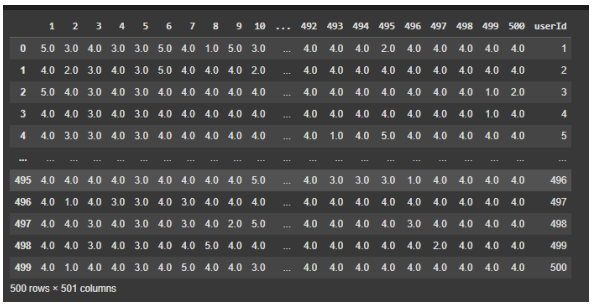
Gambar 2. Table Matrix [500x500]

2. Mengubah nilai string '[NULL]' menjadi NaN agar dapat dilakukan imputasi.
3. Menerapkan tiga model imputasi: Random Forest Imputation, Deep Learning-Based Imputation, dan Hot-Deck Imputation.



Gambar 3. Hasil Model Imputasi Random forest Imputation

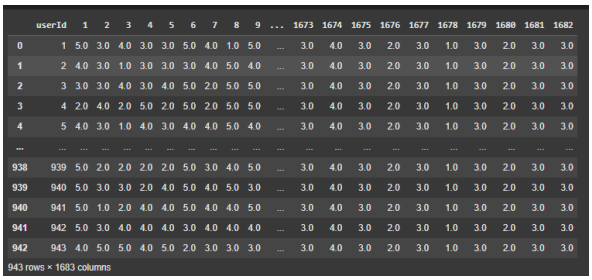
JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 11, No. 1, March 2026, Pp.458-466



	1	2	3	4	5	6	7	8	9	10	...	492	493	494	495	496	497	498	499	500	userId
0	5.0	3.0	4.0	3.0	3.0	5.0	4.0	1.0	5.0	3.0	...	4.0	4.0	4.0	2.0	4.0	4.0	4.0	4.0	4.0	1
1	4.0	2.0	3.0	4.0	3.0	5.0	4.0	4.0	4.0	2.0	...	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	2
2	5.0	4.0	3.0	4.0	3.0	4.0	4.0	4.0	4.0	4.0	...	4.0	4.0	4.0	4.0	4.0	4.0	4.0	1.0	2.0	3
3	4.0	4.0	3.0	4.0	3.0	4.0	4.0	4.0	4.0	4.0	...	4.0	4.0	4.0	4.0	4.0	4.0	4.0	1.0	4.0	4
4	4.0	3.0	3.0	4.0	3.0	4.0	4.0	4.0	4.0	4.0	...	4.0	1.0	4.0	5.0	4.0	4.0	4.0	4.0	4.0	5
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
495	4.0	4.0	4.0	4.0	3.0	4.0	4.0	4.0	4.0	5.0	...	4.0	3.0	3.0	3.0	1.0	4.0	4.0	4.0	4.0	496
496	4.0	1.0	4.0	3.0	3.0	4.0	3.0	4.0	4.0	4.0	...	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	497
497	4.0	4.0	3.0	4.0	3.0	4.0	3.0	4.0	2.0	5.0	...	4.0	4.0	4.0	4.0	3.0	4.0	4.0	4.0	4.0	498
498	4.0	4.0	3.0	4.0	3.0	4.0	4.0	5.0	4.0	4.0	...	4.0	4.0	4.0	4.0	4.0	2.0	4.0	4.0	4.0	499
499	4.0	1.0	4.0	4.0	3.0	4.0	5.0	4.0	4.0	3.0	...	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	500

500 rows × 501 columns

Gambar 4. Hasil Model Imputasi Deep Learning-Based Imputation



	userId	1	2	3	4	5	6	7	8	9	...	1673	1674	1675	1676	1677	1678	1679	1680	1681	1682
0	1	5.0	3.0	4.0	3.0	3.0	5.0	4.0	1.0	5.0	...	3.0	4.0	3.0	2.0	3.0	1.0	3.0	2.0	3.0	3.0
1	2	4.0	3.0	1.0	3.0	3.0	3.0	4.0	5.0	4.0	...	3.0	4.0	3.0	2.0	3.0	1.0	3.0	2.0	3.0	3.0
2	3	3.0	3.0	4.0	3.0	4.0	5.0	2.0	5.0	5.0	...	3.0	4.0	3.0	2.0	3.0	1.0	3.0	2.0	3.0	3.0
3	4	2.0	4.0	2.0	5.0	2.0	5.0	2.0	5.0	5.0	...	3.0	4.0	3.0	2.0	3.0	1.0	3.0	2.0	3.0	3.0
4	5	4.0	3.0	1.0	4.0	3.0	4.0	4.0	5.0	4.0	...	3.0	4.0	3.0	2.0	3.0	1.0	3.0	2.0	3.0	3.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
938	939	5.0	2.0	2.0	2.0	2.0	5.0	3.0	4.0	5.0	...	3.0	4.0	3.0	2.0	3.0	1.0	3.0	2.0	3.0	3.0
939	940	5.0	3.0	3.0	2.0	4.0	5.0	4.0	5.0	3.0	...	3.0	4.0	3.0	2.0	3.0	1.0	3.0	2.0	3.0	3.0
940	941	5.0	1.0	2.0	4.0	4.0	5.0	4.0	4.0	5.0	...	3.0	4.0	3.0	2.0	3.0	1.0	3.0	2.0	3.0	3.0
941	942	5.0	3.0	4.0	4.0	4.0	3.0	4.0	4.0	4.0	...	3.0	4.0	3.0	2.0	3.0	1.0	3.0	2.0	3.0	3.0
942	943	4.0	5.0	5.0	4.0	5.0	2.0	3.0	3.0	3.0	...	3.0	4.0	3.0	2.0	3.0	1.0	3.0	2.0	3.0	3.0

943 rows × 1683 columns

Gambar 5. Hasil Model Imputasi Hot-Deck Imputation

4. Evaluasi RMSE dari ketiga model imputasi untuk menentukan model yang paling efektif dalam mengatasi missing value.

C. *Pembentukan Model*

Penelitian ini menggunakan 3 model dalam memperoses imputasi, yaitu Random Forest Imputation, Deep Learning-Based Imputation, dan Hot-Deck Imputation.

- **Random Forest Imputation**
  Metode Random Forest adalah teknik pelatihan berbasis ensemble learning yang memanfaatkan algoritma Decision Tree[13], [9]. Dalam prosesnya, beberapa model Decision Tree dilatih dengan data yang sama, dan prediksi dilakukan pada data uji untuk masing-masing model. Hasil prediksi kemudian digabungkan melalui voting mayoritas dengan menggunakan metode modus, sehingga kelas yang diperoleh dari modus tersebut menjadi kelas prediksi akhir[14]. Selain itu, terdapat rumus untuk menghitung Entropy dan Gain dalam konteks Random Forest yaitu pada persamaan 1 dan persamaan 2. Teknik Random Forest imputation juga digunakan untuk mengisi nilai yang hilang dalam dataset dengan memanfaatkan algoritma yang sama[15]. Dalam pendekatan ini, model dibangun berdasarkan data yang lengkap, kemudian digunakan untuk memprediksi nilai yang hilang dengan mempertimbangkan hubungan antar variabel. Pendekatan ini efektif karena mampu menangkap pola kompleks dalam data dan biasanya menghasilkan imputasi yang akurat.

$$E = -\sum_{i=1}^{n} pi * log2\ (pi) \tag{1}$$

*P*i adalah peluang terpilihnya contoh secara acak di kelas *i.*

$$Gain = E_{parent} - E_{childern} \tag{2}$$

E_{parent} adalah entropi dari node induk dan
$E_{childern}$ adalah entropi rata-rata dari node anak

- **Deep Learning-Based Imputation**
  Deep Learning-Based Imputation adalah teknik yang menggunakan jaringan saraf dalam untuk mengisi nilai yang hilang dalam dataset[16]. Pendekatan ini memanfaatkan kekuatan model deep learning dalam menangkap pola kompleks dan hubungan antar fitur yang mungkin tidak terdeteksi oleh metode imputasi tradisional. Dalam proses ini, setiap entitas, seperti pengguna atau item, direpresentasikan sebagai vektor fitur yang mencakup semua informasi yang tersedia, termasuk fitur dengan nilai yang hilang. Model deep learning terdiri dari beberapa lapisan, termasuk lapisan input, beberapa lapisan tersembunyi, dan lapisan

output, di mana setiap lapisan menggunakan fungsi aktivasi untuk memperkenalkan non-linearitas, memungkinkan model untuk belajar representasi yang lebih kompleks. Perhitungan pada deep leaning menggunakan persamaan 3, 4, dan 5.

$$X = [x_{1,} x_{2,}.....,x_n]$$
$$h^l = f(W^l + \cdot h^{l\ 1} + b^l) \quad (3)$$

h^l = output dari lapisan ke-lll
W^l = bobot antara lapisan (l−1)(l-1)(l−1) dan lapisan lll
b^l = bias untuk lapisan lll
f = fungsi aktivasi (contoh: ReLU, sigmoid)

$$y^{\wedge} = W^o \cdot h^L + b^o \quad (4)$$

W^o = bobot dari lapisan terakhir ke output
h^L = output dari lapisan terakhir (L)
b^o = bias untuk output

$$\text{MSE} = \frac{1}{M} \sum_{i=1}^{m} (yi - y^i)\, 2 \quad (5)$$

m adalah jumlah contoh dalam dataset.

- **Hot-Deck Imputation**

Metode Hot-Deck Imputation digunakan untuk mengimputasi nilai yang hilang dalam suatu dataset dengan mengganti nilai yang hilang tersebut dengan nilai dari observasi lain yang memiliki karakteristik serupa. [10]. Metode ini berfungsi dengan mencari pengamatan yang memiliki karakteristik serupa dengan pengamatan yang hilang, sehingga nilai yang diambil dapat mencerminkan pola dan tren yang ada dalam data. Proses Hot-Deck dimulai dengan mengidentifikasi pengamatan yang memiliki nilai yang hilang[17]. Selanjutnya, kriteria kesamaan ditentukan, biasanya berdasarkan atribut atau fitur tertentu, dan jarak antar pengamatan dihitung menggunakan metrik seperti jarak Euclidean. Dari pengamatan yang lengkap, model kemudian memilih pengamatan terdekat yang akan digunakan untuk mengisi nilai yang hilang. Setelah pengamatan terdekat diidentifikasi, nilai yang hilang digantikan dengan nilai dari pengamatan tersebut. Perhitungan Euclidian menggunakan persamaan 6 dan 7.

$$d(i,j) = \sqrt{(X1i - X1j)2 + (X2j)2} \quad (6)$$

d (i, j) adalah jarak antara pengamatan i dan j, dan $X1i$, $X2i$ adalah nilai fitur dari pengamatan i.

j = arg min $d(i, j)$ dengan $j \in$ pengamatan yang lengkap

$$X_{ih} = X_{jh} \quad (7)$$

X_{ih} adalah nilai yang hilang yang akan diisi, dan X_{jh} adalah nilai dari fitur yang sama pada pengamatan j.

- **Root Mean Squared Error (RMSE)**

RMSE merupakan akar kuadrat dari rata-rata kesalahan kuadrat dalam peramalan. Semakin kecil nilai RMSE, semakin kecil pula rata-rata selisih antara nilai prediksi dengan nilai aktual, sehingga model peramalan dapat dianggap lebih akurat.[3], [18], dengan menggunakan persamaan 8.

$$RMSE = \sqrt{\frac{1}{|E|} \sum_{i=1}^{|E|} |P_{ut,i} - r_{ui}|} \quad (8)$$

Keterangan:
n = banyaknya data
i = (1,2,3,..., l), l merupakankeseluruhan data
yi = nilai actual

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 11, No. 1, March 2026, Pp.458-466

$\overline{yi}$= nilai hasil prediksi

III. HASIL DAN PEMBAHASAN

Evaluasi eksperimen dalam penelitian ini dilakukan menggunakan empat skenario dataset, yaitu 300 pengguna × 300 film, 500 pengguna × 500 film, 800 pengguna × 800 film, dan 943 pengguna × 1682 film. Setiap dataset memiliki tingkat sparsity yang bervariasi, yaitu 20%, 40%, 60%, 80%, dan 83,37% untuk dataset [300×300], 85,1% untuk dataset [500×500], 88,58% untuk dataset [800×800], serta 93,7% untuk dataset [943×1682].

Kinerja masing-masing metode imputasi dievaluasi menggunakan metrik Root Mean Square Error (RMSE), dengan membandingkan tiga teknik yang berbeda, yaitu Random Forest Imputation, Deep Learning-Based Imputation, dan Hot-Deck Imputation. Evaluasi difokuskan pada analisis nilai bias, varians, dan total error yang dihasilkan oleh masing-masing metode. Hasil perbandingan disajikan pada Tabel 1 hingga Tabel 4, serta divisualisasikan melalui diagram untuk mendukung analisis secara lebih komprehensif.

TABLE I
HASIL EVALUASI METODE IMPUTATION MATRIK [300X300]

Method	RMSE				
	Sparsity 20	Sparsity 40	Sparsity 60	Sparsity 80	Sparsity 83.37
Random Forest Imputation	**1.187**	**1.748**	**2.160**	**2.642**	**2.847**
Deep Learning-Based Imputation	1.616	2.367	2.918	3.509	3.703
Hot-Deck Imputation	1.500	2.141	2.669	3.240	3.373

TABLE II.
HASIL EVALUASI METODE IMPUTATION MATRIK [500X500]

Method	RMSE				
	Sparsity 20	Sparsity 40	Sparsity 60	Sparsity 80	Sparsity 85.1
Random Forest Imputation	**1.084**	**1.633**	**2.015**	**2.389**	**2.690**
Deep Learning-Based Imputation	1.482	2.214	2.730	3.227	3.564
Hot-Deck Imputation	1.352	2.020	2.503	2.990	3.270

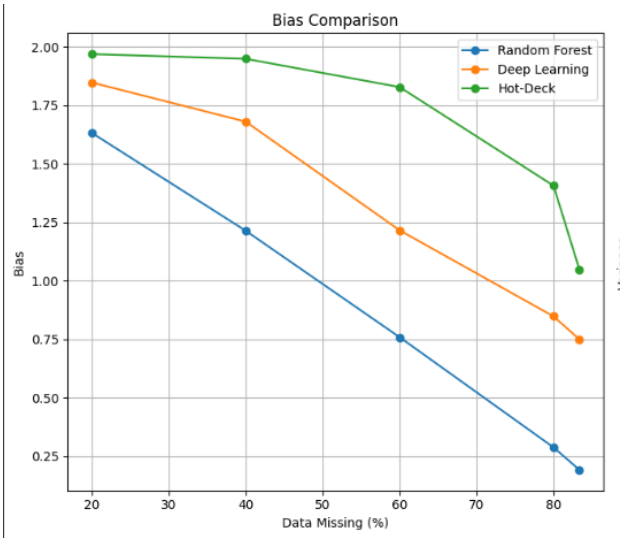
TABLE III.
HASIL EVALUASI METODE IMPUTATION MATRIK [800X800]

Method	RMSE				
	Sparsity 20	Sparsity 40	Sparsity 60	Sparsity 80	Sparsity 88.58
Random Forest Imputation	**1.183**	**1.726**	**2.140**	**2.492**	**2.758**
Deep Learning-Based Imputation	1.619	2.352	2.900	3.366	3.639
Hot-Deck Imputation	1.489	2.121	2.625	3.102	3.321

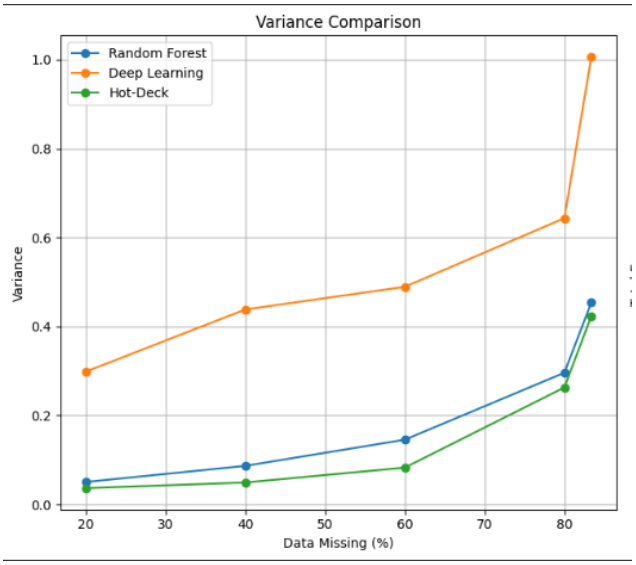
TABLE IV.
HASIL EVALUASI METODE IMPUTATION MATRIK [943X1682]

Method	RMSE				
	Sparsity 20	Sparsity 40	Sparsity 60	Sparsity 80	Sparsity 93.7
Random Forest Imputation	**1.156**	**1.632**	**2.006**	**2.332**	**2.559**
Deep Learning-Based Imputation	1.599	2.255	2.769	3.212	3.359
Hot-Deck Imputation	1.494	2.095	2.578	3.004	3.111

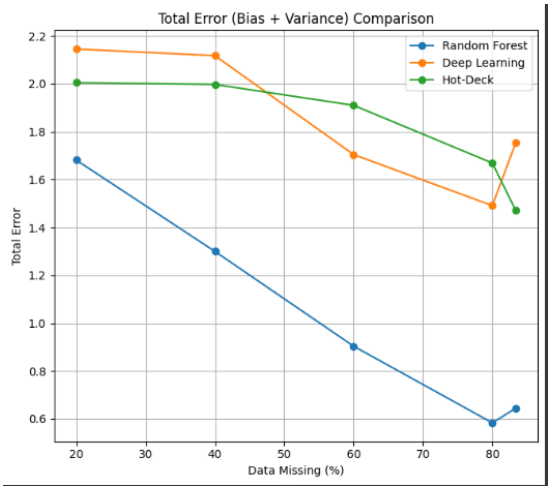
JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 11, No. 1, March 2026, Pp.458-466



Gambar 6. Perbandingan Bias



Gambar 7. Perbandingan Variance



Gambar 8. Total Eror Perbandingan (Bias + Variance)

Eksperimen ini membandingkan kinerja tiga metode imputasi, yaitu Random Forest Imputation, Deep Learning-Based Imputation, dan Hot-Deck Imputation, pada dataset dengan ukuran [300×300], [500×500], [800×800], dan

[943×1682], serta tingkat sparsitas yang bervariasi. Evaluasi performa dilakukan menggunakan metrik Root Mean Square Error (RMSE), bias, dan varians.

Hasil evaluasi menunjukkan bahwa Random Forest Imputation secara konsisten menghasilkan nilai RMSE terendah dibandingkan metode lain pada semua skenario. Contohnya, pada dataset [300×300], RMSE meningkat dari 1,187 (sparsitas 20%) menjadi 2,847 (sparsitas 83,37%). Pola serupa juga terlihat pada dataset [500×500] (RMSE 1,084–2,690), [800×800] (1,183–2,758), dan [943×1682] (1,156–2,559). Meskipun nilai RMSE meningkat seiring dengan bertambahnya sparsitas, Random Forest tetap menunjukkan stabilitas performa yang lebih baik dibandingkan Deep Learning-Based dan Hot-Deck Imputation [3], [7], [12].

Analisis bias dan varians memperkuat temuan tersebut. Random Forest memiliki bias rendah dan varians moderat yang tetap terkendali, bahkan pada tingkat sparsitas tinggi. Sebaliknya, Deep Learning-Based Imputation cenderung mengalami overfitting, yang terlihat dari peningkatan varians pada sparsitas tinggi. Sedangkan Hot-Deck Imputation, meskipun memiliki varians kecil, menunjukkan bias yang cukup tinggi, yang mengindikasikan hasil imputasi cenderung menyimpang dari nilai sebenarnya [4], [9].

Namun, terdapat trade-off antara akurasi dan efisiensi komputasi. Random Forest, meskipun paling akurat, memerlukan waktu pelatihan dan prediksi yang lebih lama. Hot-Deck Imputation relatif cepat dan sederhana, namun dengan kompromi pada akurasi. Deep Learning-Based Imputation menawarkan efisiensi waktu yang lebih baik dibandingkan Random Forest dalam beberapa kasus, namun hasil prediksinya kurang stabil. Oleh karena itu, pemilihan metode imputasi perlu mempertimbangkan kebutuhan sistem, apakah lebih mengutamakan akurasi atau kecepatan komputasi, khususnya dalam konteks sistem rekomendasi real-time [5], [14].

Hasil ini konsisten dengan penelitian sebelumnya yang menunjukkan Random Forest efektif dalam menangani data yang tidak normal dan non-linear dengan interaksi kompleks [6]. Studi lain juga menegaskan efektivitas Random Forest dalam konteks big data dan precision medicine [11].

Temuan kami juga didukung oleh penelitian yang mengkategorikan metode Hot-Deck sebagai cepat dan sederhana, namun kurang akurat dibandingkan metode lain [2], [8].

Selain itu, penelitian terkait sistem rekomendasi menekankan pentingnya metode imputasi untuk mengatasi sparsitas [16]. Studi lanjutan menyoroti perlunya pendekatan hybrid untuk mengatasi masalah cold-start dan sparsity, memperkuat bahwa pemilihan metode imputasi harus disesuaikan dengan karakteristik data dan konteks penggunaannya [17].

Dengan demikian, kontribusi utama penelitian ini adalah memberikan evaluasi komprehensif terhadap tiga metode imputasi populer pada berbagai ukuran dan tingkat sparsitas dataset. Penelitian ini juga memperkuat bukti empiris bahwa Random Forest merupakan metode yang efektif dan stabil dalam menangani missing values, khususnya dalam sistem rekomendasi berbasis collaborative filtering [1], [10].

IV. KESIMPULAN

Penelitian ini berhasil mengevaluasi efektivitas Random Forest Imputation dalam mengatasi masalah sparsity pada sistem rekomendasi. Temuan menunjukkan bahwa metode ini menghasilkan nilai RMSE yang lebih rendah dibandingkan dengan pendekatan lainnya, seperti Deep Learning-Based Imputation dan Hot-Deck Imputation, serta menunjukkan kinerja yang konsisten pada berbagai ukuran dataset yang diuji.

Keunggulan Random Forest Imputation terletak pada kemampuannya menangani berbagai jenis data, baik numerik maupun kategorikal, dengan fleksibilitas yang lebih tinggi dibandingkan metode lain. Berbeda dengan Deep Learning, yang meskipun mampu menangani data beragam, memerlukan pre-processing yang rumit dan waktu pelatihan yang lebih lama. Sementara itu, Hot-Deck Imputation lebih terbatas pada imputasi data numerik atau yang memiliki struktur tertentu, sehingga kurang efektif untuk dataset kompleks dengan berbagai tipe variabel. Keunggulan metode ini dalam mengatasi sparsity tidak hanya meningkatkan kualitas rekomendasi, tetapi juga berpotensi meningkatkan kepuasan dan kepercayaan pengguna terhadap sistem. Rekomendasi yang lebih akurat dan relevan dapat memperkuat keterlibatan dan loyalitas pengguna terhadap platform.

Secara keseluruhan, hasil penelitian ini menekankan pentingnya pemilihan metode imputasi yang tepat dalam pengembangan sistem rekomendasi yang efektif. Random Forest Imputation menawarkan solusi untuk mengatasi masalah sparsity dan membuka jalan untuk penelitian lebih lanjut dalam mengembangkan algoritma rekomendasi yang lebih inovatif. Untuk itu, penelitian berikutnya diharapkan dapat mengeksplorasi dan membandingkan metode ini dengan teknik lain dalam konteks yang lebih beragam untuk memperoleh pemahaman yang lebih mendalam mengenai kinerjanya.

ACKNOWLEDGEMENTS

Penulis mengucapkan terima kasih kepada Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi, serta Pendidikan Tinggi (Kemendikbud-Ristek DIKTI) yang telah memberikan dukungan dana penelitian ini berdasarkan Surat Keputusan Nomor: 0459/E5/PG.02.00/2024. Dukungan dana ini sangat penting untuk memperlancar berbagai tahapan penelitian, mulai dari pengumpulan data hingga analisis dan dokumentasi akhir. Selain itu, penulis mengucapkan terima kasih kepada Institut Informatika dan Bisnis Darmajaya, khususnya Fakultas Ilmu Komputer, atas dukungan dan dorongan yang terus menerus dalam penelitian ini.

DAFTAR PUSTAKA

[1] Abdillah, A. F., Purwitasari, D., Junita, S., & Purnomo, M. H. (2023). *Pengenalan Entitas Biomedis Dalam Teks Konsultasi Kesehatan Online Berbahasa Indonesia Berbasis Arsitektur Transformers*. *10*(1), 131–140. https://doi.org/10.25126/jtiik.2023106337

[2] Fadillah, I. J., Muchlisoh, S., Statistika, P., & Stis, P. S. (n.d.). *Perbandingan Metode Hot-Deck Imputation dan Metode KNNI dalam Mengatasi Missing Values (Fadillah dan Muchlisoh) Perbandingan Metode Hot-Deck Imputation Dan Metode Knni Dalam Mengatasi Missing Values Penerapan Pada Data Susenas Maret Tahun 2017 (Comparison of Hot-deck Imputation Method and KNNI Method in Overcoming Missing Values)*.

[3] Fathirachman Mahing, N., Gunawan, A. L., Foresta, A., Zen, A., Abdurrachman Bachtiar, F., Wicaksono, S. A., Brawijaya, U., & Korespondensi, P. (n.d.). *Klasifikasi Tingkat Stres Dari Data Berbentuk Teks Dengan Menggunakan Algoritma Support Vector Machine (SVM) Dan Random Forest*. https://doi.org/10.25126/jtiik2024118010

[4] Hong, S., & Lynn, H. S. (2020). Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. *BMC Medical Research Methodology*, *20*(1). https://doi.org/10.1186/s12874-020-01080-1

[5] Hu, J., & Szymczak, S. (2022). *A review on longitudinal data analysis with random forest in precision medicine*. https://doi.org/10.1093/bib/bbad002

[6] Im, J., Cho, I. H., & Kim, J. K. (n.d.). *FHDI: An R Package for Fractional Hot Deck Imputation*.

[7] Khusnul Khotimah, B., Syarief, M., Suprajitno, H., & Korespondensi, P. (2021). *Optimasi Bobot K-Means Clustering Untuk Mengatasi Missing Value Dengan Menggunakan Algoritma Genetika Optimization Weight Of K-Means Clustering To Overcome Missing Value Using Genetic Algorithm*. *8*(4). https://doi.org/10.25126/jtiik.202184912

[8] Korespondensi, P. (2021). *Peningkatan Kinerja Metode Svm Menggunakan Metode Knn Imputasi Dan K-Means-Smote Untuk Klasifikasi Kelulusan Mahasiswa Universitas Bumigora*. *8*(4). https://doi.org/10.25126/jtiik.202183428

[9] Oktafiani, R., & Rianto, R. (2023). Perbandingan Algoritma Support Vector Machine (SVM) dan Decision Tree untuk Sistem Rekomendasi Tempat Wisata. *Jurnal Nasional Teknologi Dan Sistem Informasi*, *9*(2), 113–121. https://doi.org/10.25077/teknosi.v9i2.2023.113-121

[10] Potluri, V., Singanamalla, S., Tieanklin, N., & Mankoff, J. (2023, October 22). Notably Inaccessible - Data Driven Understanding of Data Science Notebook (In)Accessibility. *ASSETS 2023 - Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*. https://doi.org/10.1145/3597638.3608417

[11] Sejati, P., Pilliang, M., & Akbar, H. (2022). *Studi Komparasi Naive Bayes, K-Nearest Neighbor, Dan Random Forest Untuk Prediksi Calon Mahasiswa Yang Diterima Atau Mundur*. *9*(7), 1341–1348. https://doi.org/10.25126/jtiik.202296737

[12] Selle, N., Yudistira, N., & Dewi, C. (2022). *Perbandingan Prediksi Penggunaan Listrik Dengan Menggunakan Metode Long Short Term Memory (Lstm) Dan Recurrent Neural Network (Rnn)*. *9*(1), 155–162. https://doi.org/10.25126/jtiik.202295585

[13] Seto Arifianto, A., Dewi Safitri, K., Agustianto, K., Wiryawan, G., & Jember, P. N. (n.d.). *Pengaruh Prediksi Missing Value Pada Klasifikasi Decision Tree C4.5*. https://doi.org/10.25126/jtiik.202294778

[14] Sza, S., Larasati, A., Nuraida, E., Dewi, K., Farhansyah, B. H., Abdurrachman Bachtiar, F., Pradana, F., Brawijaya, U., & Korespondensi, P. (n.d.). *Penerapan Decision Tree Dan Random Forest Dalam Deteksi Tingkat Stres Manusia Berdasarkan Kondisi Tidur*. https://doi.org/10.25126/jtiik.2024117993

[15] Yang, Y., Kim, J. K., & Cho, I. H. (2022). Parallel Fractional Hot-Deck Imputation and Variance Estimation for Big Incomplete Data Curing. *IEEE Transactions on Knowledge and Data Engineering*, *34*(8), 3912–3926. https://doi.org/10.1109/TKDE.2020.3029146

[16] Lestari, S., Afdila, M. E., & Pratama, Y. A. (2023). Imputation Missing Value to Overcome Sparsity Problems in The Recommendation System. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 7(6), 1285-1291.

[17] Lestari, S., Yulmaini, A., Ma'ruf, S. Y., & Sulyono, R. R. N. F. (2024). Alleviating cold start and sparsity problems in the micro, small, and medium enterprises marketplace using clustering and imputation techniques. International Journal of Electrical and Computer Engineering (IJECE), 14(3), 3220-3229.

[18] Hasibuan, M. S., Firmansyah, Y. A., Lestari, S., Yusendra, M. A. E., Rahmawati, L., & Nugroho, Y. S. D. A. (2023, July). Strategy for Community Empowerment Through an Integrated Waste Bank. In 2023 International Conference on Smart Applications, Communications and Networking (SmartNets) (pp. 1-4). IEEE.

[19] Ifada, N., & Pratama, R. A. (2024). Studi Literatur Penerapan Clustering Data Numerik Untuk Sistem Rekomendasi Berbasis Collaborative Filtering. Jurnal CoSciTech (Computer Science and Information Technology), 5(2), 418-428.

[20] Samosir, A., Justino, W. E., & Hariyono, T. (2021). *Komparasi Algoritma Random Forest, Naïve Bayes dan K- Nearest Neighbor Dalam klasifikasi Data Penyakit Jantung*.