DECODE: JURNAL PENDIDIKAN TEKNOLOGI INFORMASI
P-ISSN: 2775-2984 | E-ISSN: 2775-1813

Vol. 5, No 2, Juli 2025, Hal: 789-798
Doi: http://dx.doi.org/10.51454/decode.v5i2.1262
https://journal.umkendari.ac.id/index.php/decode

Perbandingan Penanganan Missing Value pada Data Numerik Survei Kepuasan Pengguna Lulusan

Dikky Praseptian M[1*], Sinawati[2], Kandi Harianto[3]
[1,2] Program Studi Sistem Informasi, STMIK PPKIA Tarakanita Rahmawati, Indonesia.
[3] Program Studi Teknik Informatika, STMIK PPKIA Tarakanita Rahmawati, Indonesia.

Artikel Info

Kata Kunci:

Kepauasan;
MAPE;
Nilai Hilang;
Pengguna Lulusan;
RMSE.

Keywords:

Satisfaction;
MAPE;
Missing Value;
Graduate Users;
RMSE.

Riwayat Artikel:

Submitted: 24 Juli 2025
Accepted: 31 Juli 2025
Published: 31 Juli 2025

Abstrak: Data survei kepuasan pengguna lulusan merupakan cara yang dilakukan perguruan tinggi untuk menilai kualitas perguruan tinggi ditinjau dari aspek kepuasan pengguna lulusan. Data tersebut sering terjadi adanya nilai atribut yang hilang yang disebut dengan missing value. Missing value ini dapat terjadi karena beberapa alasan tetapi paling sering tidak dapat dinilai karena aspek yang dimaksud tidak digunakan dalam bidang pekerjaan lulusan. Penelitian ini menggunakan data survei kepuasan pengguna lulusan dengan jumlah 100 record dan proporsi missing value sebesar 20% pada atribut numerik. Evaluasi kinerja dilakukan menggunakan metode Root Mean Square Error (RMSE) dan Mean Absolute Percentage Error (MAPE) untuk membandingkan empat teknik imputasi missing value yang tersedia di RapidMiner, yaitu mengganti dengan nilai rata-rata, nilai minimum, nilai maksimum dan nilai 0. Pengukuran kinerja menunjukkan model rata-rata memperoleh hasil terbaik dibanding dengan model yang lainnya, dimana nilai error pada RMSE sebesar 0.742 dan pada MAPE sebesar 13.67%. Pengukuran kinerja juga memperlihatkan tiga model lainnya berasa pada nilai error > 1 pada RMSE dan >20% pada MAPE, bahkan pada model nilai 0 nilai error pada MAPE mencapai 100% sehinggan sangat tidak disarankan mengganti nilai hilang numerik dengan 0.

Abstract: *The survey data on graduate user satisfaction is a method used by universities to assess the quality of higher education from the perspective of graduate user satisfaction. This data often has missing values, which are referred to as 'missing values'. These missing values can occur for several reasons, but most frequently, they cannot be assessed because the intended aspects are not used in the graduates' fields of work. This study uses graduate user satisfaction survey data with a total of 100 records and a missing value proportion of 20% in numerical attributes. Performance evaluation is conducted using the Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) methods to compare four available missing value imputation techniques in RapidMiner, namely replacing with the mean, minimum, maximum, and zero value. Performance measurement shows that the average model achieves the best results compared to other models, with an error value of 0.742 on RMSE and 13.67% on MAPE. Performance measurement also shows that three other models have error values > 1 on RMSE and > 20% on MAPE, even in models with a value of 0, the error value on MAPE reaches 100%, thus it is highly discouraged to replace missing numerical values with 0.*

Corresponding Author:
Dikky Praseptian M
Email: dikky@ppkia.ac.id

PENDAHULUAN

Kehilangan data atau informasi dapat terjadi karena berbagai faktor, seperti kerusakan perangkat penyimpanan, kegagalan piksel, keterbatasan kapasitas data, masalah pada peralatan akuisisi, maupun pertanyaan yang tidak terjawab dalam survei, dan sebagainya (Yan et al., 2021). Kualitas data menjadi perhatian utama bagi ilmuwan data dan peneliti yang bekerja di bidang ilmu analisis data (Jadhav et al., 2019), karena data yang hilang dapat menimbulkan ambiguitas dalam analisis serta memengaruhi sifat penaksir statistik. Kondisi ini berpotensi mengakibatkan hilangnya kekuatan uji dan bias dalam kesimpulan, sehingga data menjadi kurang dapat diandalkan (Mandel, 2015; Ngueilbaye et al., 2021). Tiga masalah utama yang terkait dengan nilai yang hilang adalah: berkurangnya efisiensi proses data mining, meningkatnya kompleksitas penanganan dan analisis data, serta timbulnya bias akibat perbedaan antara data yang hilang dan data yang lengkap (Luengo et al., 2012; Ghorbani & Desmarais, 2017). Penanganan nilai yang hilang merupakan tugas yang penting sekaligus menantang, karena memerlukan pemeriksaan menyeluruh terhadap seluruh data untuk mengidentifikasi pola missing value serta pemahaman terhadap berbagai teknik imputasi yang tersedia (Xu et al., 2018). Imputasi sering menjadi pilihan yang lebih disukai dibanding menghapus seluruh baris data, karena penghapusan dapat mengurangi ukuran dataset secara signifikan dan berpotensi menurunkan kualitas hasil analisis (Nanni et al., 2012).

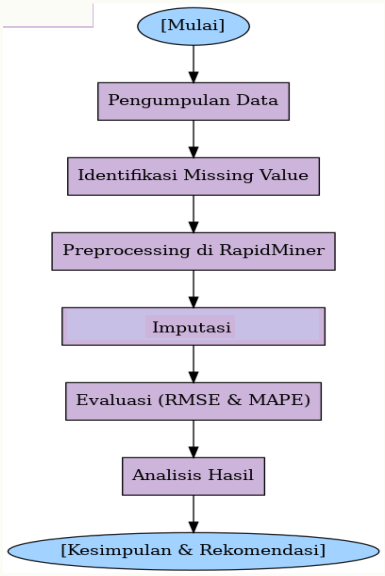
Survei tingkat kepuasan pengguna lulusan merupakan metode yang digunakan perguruan tinggi di Indonesia untuk menilai kualitas lulusan dan menjadi bahan evaluasi peningkatan mutu institusi. Survei ini biasanya diisi oleh atasan langsung di instansi atau perusahaan tempat lulusan bekerja, sehingga penilaian diharapkan lebih objektif (Joshi et al., 2015). Penelitian ini menggunakan data survei kepuasan pengguna lulusan STMIK PPKIA Tarakanita Rahmawati terhadap 100 lulusan periode 2017–2021, dengan tujuh atribut penilaian: C1 = Etika, C2 = Kompetensi Utama, C3 = Kemampuan Bahasa Asing, C4 = Penggunaan Teknologi Informasi, C5 = Kemampuan Komunikasi, C6 = Kerja Sama, dan C7 = Pengembangan Diri, masing-masing menggunakan skala Likert 1–5. Permasalahan yang sering terjadi adalah adanya missing value pada satu atau lebih atribut penilaian. Dalam konteks survei kepuasan, hal ini biasanya disebabkan karena aspek tertentu tidak relevan atau tidak digunakan di bidang pekerjaan lulusan, sehingga responden tidak memberikan penilaian. Missing value yang tidak ditangani dengan baik dapat menghasilkan inferensi analisis, seperti pengelompokan atau prediksi, yang kurang akurat (Luengo et al., 2012)

METODE

Nilai yang hilang merupakan masalah umum dalam berbagai studi ilmiah seperti di bidang medis (Little et al., 2012), biologi (Troyanskaya et al., 2001), ilmu iklim (Dixon, 1979), bahkan pada data lalu lintas (Deb & Liew, 2016). Beberapa penelitian telah dilakukan untuk membandingkan berbagai metode imputasi dalam mengisi nilai yang hilang (Noyunsan et al., 2018). Penelitian yang dilakukan oleh Acuña dan Rodriguez (2004) serta Dixon (1979) menunjukkan hasil yang hampir sama dalam mengevaluasi empat metode penanganan data hilang: penghapusan catatan, serta tiga metode imputasi—imputasi rata-rata, imputasi median, dan imputasi tetangga terdekat. Klasifikasi dilakukan menggunakan dua metode: analisis diskriminan linier (LDA) dan k-nearest neighbor (k-NN). Hasilnya menunjukkan bahwa imputasi memiliki dampak yang kurang signifikan terhadap akurasi klasifikasi. Kelemahan dari kedua penelitian ini adalah pengujian hanya dilakukan pada dataset dengan tingkat kehilangan data di bawah 20%. Penelitian oleh Batista dan Monard (2003) menguji akurasi klasifikasi dari dua metode, yaitu C4.5 dan CN2, serta tiga teknik imputasi: mean imputation, imputasi modus, dan k-nearest neighbor imputation (kNNI). Nilai yang hilang hanya disisipkan pada beberapa atribut tertentu. Hasilnya menunjukkan bahwa imputasi kNN memberikan akurasi yang tinggi, namun hanya efektif jika atribut dalam dataset tidak memiliki korelasi tinggi satu sama lain. Selain itu, distribusi nilai yang hilang dalam studi tersebut cenderung tidak merata di seluruh atribut. Penelitian lain oleh Grzymala-Busse dan Hu (2001) mengevaluasi akurasi klasifikasi pada sepuluh dataset dengan nilai yang hilang menggunakan lima metode imputasi: imputasi modus, algoritma C4.5, LERS, serta dua

metode imputasi non-tradisional berbasis pembelajaran mesin dan teori himpunan kasar. Hasilnya menunjukkan bahwa proses imputasi sebelum klasifikasi dapat meningkatkan akurasi hasil klasifikasi. Namun, penelitian ini hanya menguji satu algoritma klasifikasi dan pada tingkat kehilangan data yang rendah (1%–13%). Sementara itu, Mundfrom dan Whitcomb (1998) menggunakan dua teknik klasifikasi—fungsi diskriminan linier dan regresi logistik—untuk menguji tiga metode imputasi: mean imputation, hot deck, dan imputasi regresi. Hasilnya menunjukkan bahwa mean imputation memberikan kinerja terbaik. Meski begitu, karena hanya menggunakan satu dataset, kesimpulan dari penelitian ini dianggap kurang meyakinkan.

Berdasarkan kajian tersebut, masih terdapat kesenjangan penelitian pada tiga aspek utama: (1) sebagian besar studi hanya menguji metode imputasi pada tingkat kehilangan data yang rendah, (2) distribusi nilai yang hilang sering kali tidak merata sehingga kurang mewakili kondisi riil data, dan (3) pemilihan metrik evaluasi jarang dibahas secara mendalam. Oleh karena itu, penelitian ini membandingkan beberapa metode penanganan nilai hilang pada data numerik dengan tingkat kehilangan yang bervariasi dan distribusi yang lebih merata. Proses imputasi dilakukan menggunakan RapidMiner Studio dengan memanfaatkan operator Replace Missing Values (Noyunsan et al., 2018). Sebagian besar parameter menggunakan pengaturan default, namun pemilihan metode imputasi diatur secara custom sesuai skenario uji, yaitu penggantian dengan nilai minimum, maksimum, rata-rata, nilai nol, dan nilai acak (random). Evaluasi kinerja masing-masing metode dilakukan dengan dua metrik, yaitu Root Mean Squared Error (RMSE) dan Mean Absolute Percentage Error (MAPE) (Luengo et al., 2012; Xu et al., 2018). RMSE dipilih karena sensitif terhadap perbedaan besar antara nilai hasil imputasi dan nilai sebenarnya, sehingga efektif untuk mendeteksi kesalahan besar. MAPE dipilih karena memberikan ukuran kesalahan dalam bentuk persentase yang memudahkan interpretasi bagi pengguna non-teknis. Namun, kedua metrik ini memiliki keterbatasan: RMSE cenderung memberikan bobot berlebih pada kesalahan besar yang jarang terjadi, sedangkan MAPE tidak dapat digunakan jika terdapat nilai aktual bernilai nol karena akan menyebabkan pembagian dengan nol (Mandel, 2015; Ngueilbaye et al., 2021). Berikut Alur Penelitian pada Gambar 1.



Gambar 1. Alur Penelitian

Dataset tingkat kepuasan pengguna lulusan memiliki tujuh atribut dengan skala likert 1-5. Dataset asli berisi 100 titik data dengan semua atributnya memiliki nilai atau tidak ada nilai yang hilang. Selanjutnya Dataset diberikan 20 titik data dengan missing value. Karena penelitian ini hanya berfokus pada imputasi tunggal, maka dilakukan modifikasi terhadap titik-titik data yang memiliki missing value dengan cara menghilangkan nilai satu atribut saja. Penentuan titik data mana dalam dataset yang akan memiliki nilai mising dan atribut titik data mana yang nilainya akan hilang dilakukan secara acak. Dataset asli yang akan digunakan dalam penelitian ini dapat dilihat pada Tabel 1.

Tabel 1. Dataset

Alumni	P1	P2	P3	P4	P5	P6	P7
Data1	5	5	4	5	5	5	5
Data2	4	4	4	5	4	4	4
Data3	4	4	4	4	4	4	4
Data4	4	4	3	4	5	4	4
Data5	5	4	4	4	4	4	4
..	..	..	..	..	..	..	..
Data96	5	4	3	3	5	5	5
Data97	4	4	3	4	3	3	3
Data98	5	5	3	5	5	5	4
Data99	5	5	3	5	5	5	5
Data 100	5	4	2	3	4	4	4

Pengukuran kinerja dilakukan dengan mengukur tingkat kesalahan menggunakan dua metode, yaitu Root Mean Squared Error (RMSE) dan Mean Absolute Percentage Error (MAPE). RMSE merupakan salah satu metode yang umum digunakan untuk mengevaluasi teknik peramalan atau mengukur akurasi hasil dari suatu model peramalan. RMSE menunjukkan nilai rata-rata dari jumlah kuadrat selisih antara nilai yang diprediksi dan nilai aktual. Nilai RMSE yang kecil mengindikasikan bahwa variasi nilai yang dihasilkan oleh model prediksi mendekati variasi nilai observasinya. Salah satu ukuran kesalahan dalam peramalan adalah nilai rata-rata akar kuadrat dari error tersebut atau Root Mean Squared Error (RMSE), dengan persamaan sebagai berikut (Hodson, 2022; Winarni & Pratiwi, 2025).

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} (A_t - F_t)^2}{n}}$$

Dimana
RMSE = Kesalahan Kuadrat Rata-Rata Akar
n = Jumlah Sampel
At = Nilai Aktual
Ft = Nilai Prediksi

Mean Absolute Percentage Error (MAPE) adalah ukuran dalam bentuk persentase yang digunakan untuk mengukur besarnya kesalahan hasil prediksi. Semakin kecil nilai MAPE, maka semakin kecil pula tingkat kesalahan prediksi; sebaliknya, semakin besar nilai MAPE menunjukkan tingkat kesalahan yang semakin tinggi. Hasil imputasi nilai yang hilang dikategorikan sangat baik jika MAPE < 10%, dan dikategorikan baik jika nilai MAPE berada pada rentang 10% hingga 20%. MAPE dapat dihitung dengan menggunakan rumus berikut (Al-Khowarizmi et al., 2021; Mayni, Manurung, & Nehe, 2024).

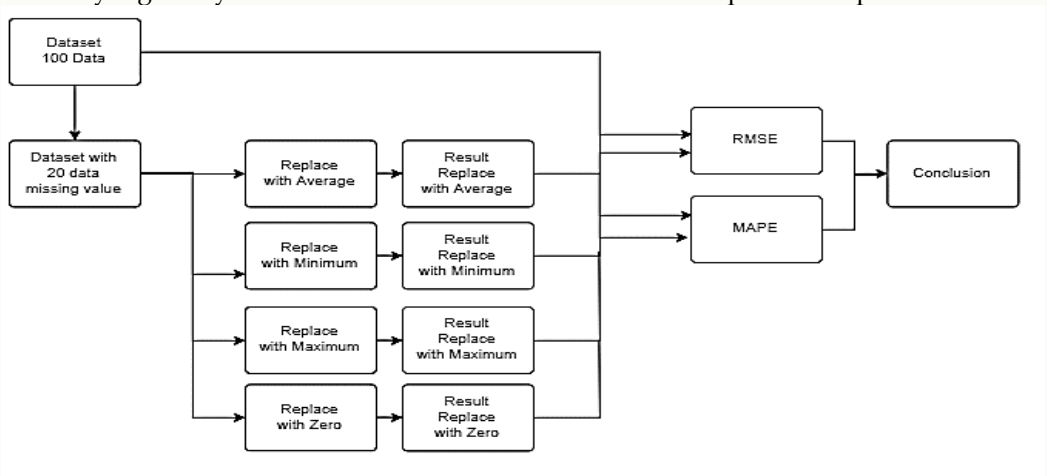
$$MAPE = \frac{\sum_{i=1}^{n} \left| \frac{X_i - F_i}{X_i} \right|}{n} 100\%$$

Dimana
Xi = data sebenarnya
Fi = hasil prediksi
n = Jumlah Sampel

HASIL DAN PEMBAHASAN

Prinsip Analisis

Analisis dilakukan dengan menggunakan 100 data kepuasan pengguna lulusan yang berisi data lengkap. Dataset kemudian diberikan nilai hilang pada salah satu atribut di 20 data, dari data ke-81 sampai dengan 100. Dataset hasil proses pemberian nilai hilang tadi selanjutnya disebut dataset dengan 20% missing value. Nilai hilang pada dataset dengan 20% missing value akan coba diprediksi atau digantikan dengan nilai tertentu. Ada 4 model pergantian nilai yang akan dilakukan yaitu rata-rata, minimum, maximum dan nilai nol. Empat model pergantian akan menghasilkan empat dataset hasil proses pergantian nilai hilang. Keempat dataset tadi akan dihitung nilai errornya jika dibandingkan dengan dataset asli. Perhitungan error dilakukan dengan dua metode yaitu RMSE dan MAPE. Hasil perhitungan error kemudian akan di analisis, model pergantian nilai hilang mana yang memiliki nilai error terendah yang artinya memiliki hasil terbaik. Proses analisis dapat dilihat pada Gambar 2.



Gambar 2. Prinsip Analisis

Uji Coba

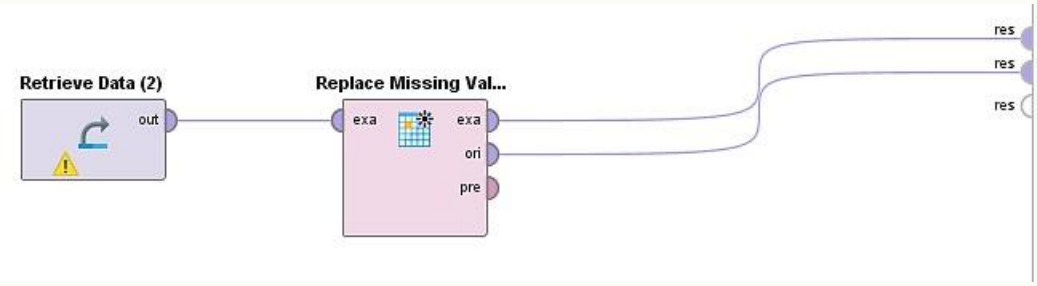
Uji coba dilakukan pada 20 data dari data ke-81 sampai dengan 100. Diawali dengan menghilangkan salah satu nilai atribut. Nilai pada atribut yang akan di hilangkan dilakukan secara acak. Berikut data ke-81 sampai dengan 100 yang telah diberi nilai hilang pada salah satu atributnya tersaji pada Tabel 2. NaN (*Not a Number*) adalah istilah dalam komputasi yang digunakan untuk menandai nilai yang bukan angka valid seperti *missing value*.

Tabel 2. Dataset dengan nilai hilang

Alumni	P1	P2	P3	P4	P5	P6	P7
Data81	4	4	3	3	4	NaN	3
Data82	4	4	3	4	3	3	NaN
Data83	5	5	4	5	5	NaN	5
Data84	4	4	3	4	NaN	4	4
Data85	4	4	4	NaN	4	4	4
…	…	…	…	…	…	…	…
Data96	5	4	3	3	NaN	5	5
Data97	4	4	3	NaN	3	3	3
Data98	5	5	NaN	5	5	5	4
Data99	5	NaN	3	5	5	5	5
Data100	NaN	4	2	3	4	4	4

Dataset dengan data ke-1 sampai dengan 80 memiliki nilai lengkap dan data ke-81 sampa dengan 100 telah diberi nilai hilang pada salah satu atribut ini membentuk satu dataset yang siap diujicoba. Dataset tersebut diujicoba pada tools rapidminer dengan menggunakan model atau fungsi penggantian

nilai hilang. Model penggantian nilai hilang memiliki pilihan yaitu dengan rata-rata, minimum, maximum dan nilai 0. Seluruh pilihan pada model tersebut akan diujicoba untuk mengetahui sederhana mana yang paling baik untuk digunakan pada data numerik pada dataset kepuasan pengguna lulusan. Berikut tampilan desain ujicoba pada tools rapidminer tersaji pada Gambar 3.



Gambar 3. Proses Uji Coba RapidMiner

1. Hasil Replace dengan rata-rata

Penggantian nilai hilang dengan rata-rata merupakan penggantian nilai hilang dengan menghitung rata-rata nilai atribut yang berisi nilai dan akan menjadi nilai pada atribut yang memiliki nilai hilang. Sebagai contoh pada data ke 88 memiliki nilai hilang pada atribut P1 maka untuk mencari pengganti nilainya adalah dengan mencari rata-rata nilai dari atribut P1 pada data ke-1 .. 87 dan ke-89 .. 99, mengapa data ke 100 tidak digunakan karena data tersebut juga memiliki nilai hilang pada atribut P1. Berikut hasil penggantian dengan rata-rata pada Tabel 3.

Tabel 3. Dataset Hasil Replace dengan rata-rata

Alumni	P1	P2	P3	P4	P5	P6	P7
Data81	4	4	3	3	4	4	3
Data82	4	4	3	4	3	3	4
Data83	5	5	4	5	5	4	5
Data84	4	4	3	4	4	4	4
Data85	4	4	4	4	4	4	4
…	…	…	…	…	…	…	…
Data96	5	4	3	3	4	5	5
Data97	4	4	3	4	3	3	3
Data98	5	5	4	5	5	5	4
Data99	5	4	3	5	5	5	5
Data100	5	4	2	3	4	4	4

2. Hasil Replace dengan Minimum

Penggantian nilai hilang dengan minimum merupakan penggantian nilai hilang dengan mencari nilai terkecil pada atribut yang berisi nilai dan akan menjadi nilai pada atribut yang memiliki nilai hilang. Sebagai contoh pada data ke 88 memiliki nilai hilang pada atribut P1 maka untuk mencari pengganti nilainya adalah dengan mencari nilai terkecil dari atribut P1 pada data ke-1 .. 87 dan ke-89 .. 99, mengapa data ke 100 tidak digunakan karena data tersebut juga memiliki nilai hilang pada atribut P1. Berikut hasil penggantian dengan minimum pada Tabel 4.

Tabel 4. Dataset Hasil Replace dengan nilai minimum

Alumni	P1	P2	P3	P4	P5	P6	P7
Data81	4	4	3	3	4	3	3
Data82	4	4	3	4	3	3	3
Data83	5	5	4	5	5	3	5
Data84	4	4	3	4	3	4	4
Data85	4	4	4	2	4	4	4
…	…	…	…	…	…	…	…

Data96	5	4	3	3	3	5	5
Data97	4	4	3	2	3	3	3
Data98	5	5	2	5	5	5	4
Data99	5	4	3	5	5	5	5
Data100	4	4	2	3	4	4	4

3. Hasil Replace dengan Maximum

Penggantian nilai hilang dengan mixiumum merupakan penggantian nilai hilang dengan mencari nilai terbesar pada atribut yang berisi nilai dan akan menjadi nilai pada atribut yang memiliki nilai hilang. Sebagai contoh pada data ke 88 memiliki nilai hilang pada atribut P1 maka untuk mencari pengganti nilainya adalah dengan mencari nilai terbesar dari atribut P1 pada data ke-1 .. 87 dan ke-89 .. 99, mengapa data ke 100 tidak digunakan karena data tersebut juga memiliki nilai hilang pada atribut P1. Berikut hasil penggantian dengan maximum pada Tabel 5.

Tabel 5. Dataset Hasil Replace dengan nilai maximum

Alumni	P1	P2	P3	P4	P5	P6	P7
Data81	4	4	3	3	4	5	3
Data82	4	4	3	4	3	3	5
Data83	5	5	4	5	5	5	5
Data84	4	4	3	4	5	4	4
Data85	4	4	4	5	4	4	4
…	…	…	…	…	…	…	…
Data96	5	4	3	3	5	5	5
Data97	4	4	3	5	3	3	3
Data98	5	5	5	5	5	5	4
Data99	5	5	3	5	5	5	5
Data100	5	4	2	3	4	4	4

4. Hasil Replace dengan nilai 0

Penggantian nilai hilang dengan nilai 0 merupakan penggantian nilai hilang dengan mengganti nilai hilang dengan angka 0. Cara ini tidak terpengaruh dengan keberadaan nilai hilang pada atribut mana saja, artinya nilai hilang terletak pada atribut apapun akan diberikan nilai 0 sebagai penggantinya. Cara ini merupakan cara paling mudah dilakukan dibanding cara lainnya. Berikut hasil penggantian dengan nilai 0 pada Tabel 6.

Tabel 6. Dataset Hasil Replace dengan nilai 0

Alumni	P1	P2	P3	P4	P5	P6	P7
Data81	4	4	3	3	4	5	3
Data82	4	4	3	4	3	3	5
Data83	5	5	4	5	5	5	5
Data84	4	4	3	4	5	4	4
Data85	4	4	4	5	4	4	4
…	…	…	…	…	…	…	…
Data96	5	4	3	3	5	5	5
Data97	4	4	3	5	3	3	3
Data98	5	5	5	5	5	5	4
Data99	5	5	3	5	5	5	5
Data100	5	4	2	3	4	4	4

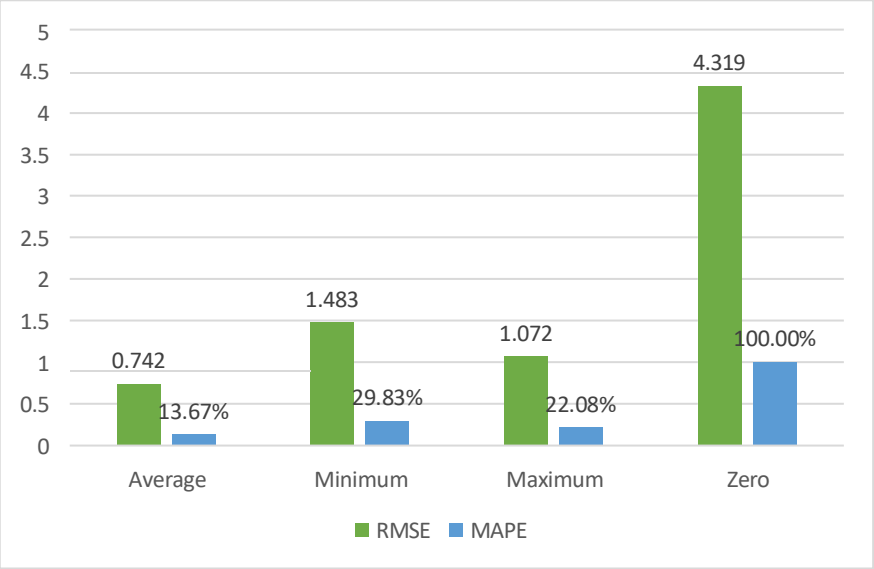
Pengukuran Kinerja

Pengukuran kinerja dilakukan dengan dua metode, yaitu RMSE dan MAPE, yang keduanya digunakan untuk mengukur tingkat error antara dataset hasil imputasi dengan dataset asli yang memiliki nilai lengkap. Nilai error dihitung dengan membandingkan hasil imputasi menggunakan

empat model sederhana, yaitu rata-rata, minimum, maksimum, dan nilai 0. Nilai error terkecil dianggap sebagai hasil terbaik. Hasil pengukuran (Gambar 4) menunjukkan bahwa model rata-rata memperoleh kinerja terbaik dibandingkan dengan model lainnya, dengan nilai RMSE sebesar 0,742 dan MAPE sebesar 13,67%. Tiga model lainnya menunjukkan performa yang lebih rendah, dengan nilai RMSE > 1 dan MAPE > 20%. Khusus pada model pengisian dengan nilai 0, nilai error MAPE mencapai 100%, sehingga metode ini sangat tidak disarankan untuk menangani nilai hilang numerik. Selain itu, hasil pengukuran RMSE dan MAPE konsisten—tidak ditemukan perbedaan tren (kontra)—sehingga kedua metrik dapat saling mendukung dalam mengevaluasi performa metode imputasi.

Secara teoretis, keunggulan metode rata-rata pada dataset ini dapat dijelaskan oleh sifat distribusi data yang cenderung tidak terlalu miring (skewness rendah), sehingga rata-rata menjadi representasi sentral yang baik untuk menggantikan nilai hilang. Menurut Little & Rubin (2019), ketika data hilang secara acak (Missing Completely at Random, MCAR) dan distribusi data relatif simetris, imputasi menggunakan nilai rata-rata dapat meminimalkan varian error karena nilai pengganti tidak terlalu menyimpang dari nilai asli pada mayoritas kasus. Sebaliknya, metode minimum atau maksimum cenderung menghasilkan bias besar karena memasukkan nilai ekstrem, sedangkan penggantian dengan 0 akan memperbesar deviasi absolut pada atribut numerik yang tidak memiliki nilai mendekati nol. Temuan ini sejalan dengan penelitian Acuña & Rodriguez (2004) serta Dixon (1979) yang melaporkan bahwa imputasi sederhana berbasis nilai pusat (mean/median) dapat memberikan akurasi klasifikasi yang cukup baik pada dataset dengan tingkat kehilangan data rendah hingga sedang, dan dalam beberapa kasus mengungguli metode ekstrem seperti nilai minimum/maksimum. Namun, hasil ini berbeda dengan studi Batista & Monard (2003) yang menemukan bahwa k-nearest neighbor imputation (kNNI) menghasilkan akurasi lebih tinggi dibandingkan imputasi rata-rata, terutama pada dataset dengan korelasi atribut yang rendah. Perbedaan ini dapat disebabkan oleh karakteristik dataset yang digunakan dalam penelitian ini, di mana atribut-atributnya memiliki hubungan yang cukup kuat sehingga metode sederhana seperti rata-rata sudah cukup untuk menjaga representasi distribusi data tanpa memerlukan pembobotan kompleks antar tetangga terdekat.

Dengan demikian, meskipun imputasi rata-rata terbukti unggul pada penelitian ini, hasilnya tidak serta merta dapat digeneralisasikan untuk semua jenis dataset, khususnya pada data yang memiliki distribusi sangat miring atau korelasi atribut rendah—di mana metode imputasi berbasis model (misalnya kNNI atau regresi) berpotensi lebih unggul..



Gambar 3. Pengukuran Nilai Error

KESIMPULAN

Penelitian ini membandingkan empat metode imputasi sederhana—rata-rata, minimum, maksimum, dan nilai 0—untuk menangani *missing value* pada data numerik tingkat kepuasan pengguna lulusan. Berdasarkan evaluasi menggunakan dua metrik, RMSE dan MAPE, metode imputasi rata-rata terbukti memberikan hasil terbaik dengan nilai RMSE sebesar 0,742 dan MAPE sebesar 13,67%. Tiga metode lainnya menunjukkan kinerja yang lebih rendah, terutama penggantian dengan nilai 0 yang menghasilkan MAPE hingga 100%, sehingga tidak direkomendasikan untuk konteks serupa. Konsistensi hasil RMSE dan MAPE juga menunjukkan bahwa kedua metrik dapat saling melengkapi dalam mengevaluasi kinerja metode imputasi. Kontribusi utama penelitian ini terhadap bidang data science dan pengolahan data numerik adalah memberikan bukti empiris bahwa metode imputasi sederhana berbasis nilai pusat (*mean imputation*) dapat menjadi solusi yang efektif, efisien, dan mudah diimplementasikan untuk dataset dengan distribusi relatif simetris dan kehilangan data acak (MCAR). Hasil ini memperkuat literatur mengenai pemilihan teknik imputasi, khususnya untuk data survei di bidang pendidikan tinggi.

Secara praktis, temuan ini memiliki implikasi penting bagi peneliti dan pengelola data di sektor pendidikan maupun survei serupa. Pada studi yang mengukur kepuasan atau persepsi pengguna—seperti *tracer study*, survei kepuasan mahasiswa, atau evaluasi alumni—imputasi rata-rata dapat digunakan untuk menjaga kelengkapan data tanpa menambah kompleksitas komputasi, sekaligus mempertahankan keakuratan analisis. Namun, penerapan metode ini perlu mempertimbangkan karakteristik distribusi data dan pola *missing value* sebelum digunakan secara luas, serta tidak menutup kemungkinan untuk menguji metode imputasi yang lebih kompleks apabila kondisi data berbeda.

DAFTAR PUSTAKA

Acuña, E., & Rodriguez, C. (2004). The treatment of missing values and its effect on classifier accuracy. *Classification, Clustering, and Data Mining Applications, 639-647*. https://doi.org/10.1007/978-3-642-17103-1_60

Al-Khowarizmi, R., Syah, R., Nasution, M. K. M., & Elveny, M. (2021). Sensitivity of MAPE using detection rate for big data forecasting crude palm oil on k-nearest neighbor. *International Journal of Electrical and Computer Engineering*, 11(3). https://doi.org/10.11591/ijece.v11i3.pp2696-2703

Batista, G. E. A. P. A., & Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, 17(5–6). https://doi.org/10.1080/713827181

Deb, R., & Liew, A. W. C. (2016). Missing value imputation for the analysis of incomplete traffic accident data. *Information Sciences*, 339. https://doi.org/10.1016/j.ins.2016.01.018

Dixon, J. K. (1979). Pattern recognition with partly missing data. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(10). https://doi.org/10.1109/TSMC.1979.4310090

Ghorbani, S., & Desmarais, M. C. (2017). Performance comparison of recent imputation methods for classification tasks over binary data. *Applied Artificial Intelligence*, 31(1). https://doi.org/10.1080/08839514.2017.1279046

Grzymala-Busse, J. W., & Hu, M. (2001). A comparison of several approaches to missing attribute values in data mining. *In Lecture Notes in Computer Science*. https://doi.org/10.1007/3-540-45554-X_46

Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development*, 15(14). https://doi.org/10.5194/gmd-15-5481-2022

Jadhav, A., Pramod, D., & Ramanathan, K. (2019). Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, 33(10), 913–933. https://doi.org/10.1080/08839514.2019.1637138

Joshi, A., Kale, S., Chandel, S., & Pal, D. (2015). Likert Scale: Explored and explained. *British Journal of Applied Science & Technology*, 7(4). https://doi.org/10.9734/BJAST/2015/14975

Little, R. J., & Rubin, D. B. (2012). The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine,* 367(14). https://doi.org/10.1056/nejmsr1203730

Luengo, J., García, S., & Herrera, F. (2012). On the choice of the best imputation methods for missing values considering three groups of classification methods. *Knowledge and Information Systems*, 32(1). https://doi.org/10.1007/s10115-011-0424-2

Mandel, S. P., & Jadhav, J. (2015). A comparison of six methods for missing data imputation. *Journal of Biom. Biostat*., 06(01). https://doi.org/10.4172/2155-6180.1000224

Mayni, N., Manurung, N., & Nehe, N. K. (2024). Penerapan metode Single Exponential Smoothing prediksi stok sembako pada Toko Suci Berkah di Sei Piring Kecamatan Pulau Rakyat. *DECODE: Jurnal Pendidikan Teknologi Informasi*, 4(3), 748–763. https://doi.org/10.51454/decode.v4i3.495

Mundfrom, D., & Whitcomb, A. (1998). Imputing missing values: The effect on the accuracy of classification. *General Linear Model Journal*, *25(1).*

Nanni, L., Lumini, A., & Brahnam, S. (2012). A classifier ensemble approach for the missing feature problem. *Artificial Intelligence in Medicine*, 55(1). https://doi.org/10.1016/j.artmed.2011.11.006

Ngueilbaye, A., Wang, H., Mahamat, D. A., & Junaidu, S. B. (2021). Modulo 9 model-based learning for missing data imputation. *Applied Soft Computing*, 103. https://doi.org/10.1016/j.asoc.2021.107167

Noyunsan, C., Katanyukul, T., & Saikaew, K. (2018). Performance evaluation of supervised learning algorithms with various training data sizes and missing attributes. *Engineering and Applied Science Research,* 45(3). https://doi.org/10.14456/easr.2018.28

Troyanskaya, O., et al. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics, 17(6).* https://doi.org/10.1093/bioinformatics/17.6.520

Winarni, I., & Pratiwi, N. (2024). Prediksi harga saham menggunakan metode Long Short-Term Memory: Studi kasus saham Intel Corporation. *DECODE: Jurnal Pendidikan Teknologi Informasi*, 5(2), 380–390. https://doi.org/10.51454/decode.v5i2.1192

Xu, X., Chong, W., Li, S., Arabo, A., & Xiao, J. (2018). MIAEC: Missing data imputation based on the evidence Chain. *IEEE Access*, 6. https://doi.org/10.1109/ACCESS.2018.2803755

Yan, Y., Wu, Y., Du, X., & Zhang, Y. (2021). Incomplete data ensemble classification using imputation-revision framework with local spatial neighborhood information. *Applied Soft Computing*, 99. https://doi.org/10.1016/j.asoc.2020.106905