

Oke Hariansyah, M.Kom
Ines Heidiani Ikasari, M.Kom
Perani Rosyani, M.Kom



TEKNIK INPUTASI DATA

Teori dan Penerapan
pada Dataset Smartphone

Teknik

INPUTASI DATA

Teori dan Penerapan pada Dataset Smartphone

PENULIS

Oke Hariansyah, M.Kom.

Ines Heidiani Ikasari, M.Kom.

Perani Rosyani, M.Kom.

PENERBIT KBM INDONESIA Adalah penerbit dengan misi memudahkan proses penerbitan buku buku penulis di tanah air Indonesia. Serta menjadi media sharing proses penerbitan buku.

TEKNIK IMPUTASI DATA

Teori dan Penerapan pada Dataset Smartphone

Copyright @2026 By Oke Hariansyah, M.Kom., dkk

All right reserved

Penulis

Oke Hariansyah, M.Kom.
Ines Heidiani Ikasari, M.Kom.
Perani Rosyani, M.Kom.

Desain Sampul

Aswan Kreatif

Tata Letak

Sofitahm

Editor

Dr. Muhamad Husein Maruapey, Drs., M.Sc.
Background isi buku di ambil dari <https://www.freepik.com/>

Official

Depok, Sleman-Jogjakarta (Kantor)
Penerbit Karya Bakti Makmur (KBM) Indonesia
Anggota IKAPI/No. IKAPI 279/JTI/2021
081357517526 (Tlpn/WA)

Website

<https://penerbitkbm.com>
www.penerbitbukumurah.com

Email

naskah@penerbitkbm.com

Distributor

<https://penerbitkbm.com/toko-buku/>

Youtube

Penerbit KBM Sastrabook

Instagram

@penerbit.kbmindonesia
@penerbitbukujogja

ISBN: 978-634-278-282-8

Cetakan ke-1, April 2026
15,5 x 23 cm, iv + 126 halaman

Isi buku diluar tanggungjawab penerbit
Hak cipta merek KBM Indonesia sudah terdaftar di DJKI-Kemenkumham
dan isi buku dilindungi undang-undang.
Dilarang keras menerjemahkan, memfotokopi, atau
memperbanyak sebagian atau seluruh isi buku ini
tanpa seizin penerbit karena beresiko sengketa hukum

Sanksi Pelanggaran Pasal 113
Undang-Undang No. 28 Tahun 2014 Tentang Hak Cipta

1. Setiap Orang yang dengan tanpa hak melakukan pelanggaran hak ekonomi sebagaimana dimaksud dalam Pasal 9 ayat (1) huruf i untuk Penggunaan Secara Komersial dipidana dengan pidana penjara paling lama 1 (satu) tahun dan/atau pidana denda paling banyak Rp 100.000.000 (seratus juta rupiah).
2. Setiap Orang yang dengan tanpa hak dan/atau tanpa izin Pencipta atau pemegang Hak Cipta melakukan pelanggaran hak ekonomi Pencipta sebagaimana dimaksud dalam Pasal 9 ayat (1) huruf c, huruf d, huruf f, dan/atau huruf h untuk Penggunaan Secara Komersial dipidana dengan pidana penjara paling lama 3 (tiga) tahun dan/atau pidana denda paling banyak Rp 500.000.000,00 (lima ratus juta rupiah).
3. Setiap Orang yang dengan tanpa hak dan/atau tanpa izin Pencipta atau pemegang Hak Cipta melakukan pelanggaran hak ekonomi Pencipta sebagaimana dimaksud dalam Pasal 9 ayat (1) huruf a, huruf b, huruf e, dan/atau huruf g untuk Penggunaan Secara Komersial dipidana dengan pidana penjara paling lama 4 (empat) tahun dan/atau pidana denda paling banyak Rp 1.000.000.000,00 (satu miliar rupiah).
4. Setiap Orang yang memenuhi unsur sebagaimana dimaksud pada ayat (3) yang dilakukan dalam bentuk pembajakan, dipidana dengan pidana penjara paling lama 10 (sepuluh) tahun dan/atau pidana denda paling banyak Rp 4.000.000.000,00 (empat miliar rupiah).

KATA PENGANTAR



Puji syukur penulis panjatkan ke hadirat Allah SWT atas segala rahmat dan karunia-Nya sehingga buku ini dapat diselesaikan dengan baik. Buku ini disusun sebagai salah satu upaya untuk memberikan pemahaman yang lebih komprehensif mengenai permasalahan *missing value* dalam analisis data serta berbagai teknik yang dapat digunakan untuk mengatasinya.

Dalam proses analisis data, keberadaan data yang tidak lengkap merupakan salah satu permasalahan yang sering ditemui. Missing value dapat muncul karena berbagai faktor seperti kesalahan pengumpulan data, keterbatasan informasi dari sumber data, maupun perbedaan struktur data antar sumber. Apabila tidak ditangani dengan tepat, keberadaan data yang hilang dapat mempengaruhi kualitas dataset serta menurunkan akurasi hasil analisis dan pengambilan keputusan berbasis data.

Melalui buku ini, penulis berusaha menyajikan pembahasan mengenai konsep dasar missing value, tipe-tipe missing data, serta berbagai teknik imputasi data yang umum digunakan dalam analisis data. Selain itu, buku ini juga dilengkapi dengan studi kasus pada dataset spesifikasi smartphone yang digunakan untuk menganalisis dan membandingkan beberapa metode imputasi seperti mean, median, modus, KNN, regression, dan MICE. Pendekatan ini diharapkan dapat memberikan gambaran yang lebih jelas mengenai

kelebihan dan keterbatasan masing-masing metode dalam menangani missing value.

Penulis menyampaikan rasa terima kasih yang sebesar-besarnya kepada semua pihak yang telah memberikan dukungan dan kontribusi dalam proses penyusunan buku ini. Ucapan terima kasih penulis sampaikan kepada keluarga tercinta yang senantiasa memberikan doa, dukungan, serta motivasi sehingga penulis dapat menyelesaikan penulisan buku ini dengan baik.

Penulis juga mengucapkan terima kasih kepada rekan-rekan akademisi dan kolega yang telah memberikan berbagai masukan, diskusi ilmiah, serta inspirasi dalam pengembangan materi yang disajikan dalam buku ini. Diskusi dan pertukaran gagasan tersebut sangat membantu dalam memperkaya perspektif penulis dalam membahas permasalahan missing value dan teknik imputasi data dalam analisis dataset.

Selain itu, penulis juga menyampaikan apresiasi kepada para peneliti dan praktisi di bidang ilmu komputer, data science, dan statistik yang karya-karyanya telah menjadi referensi penting dalam penyusunan buku ini.

Penulis menyadari bahwa buku ini masih memiliki berbagai keterbatasan. Oleh karena itu, penulis sangat terbuka terhadap kritik dan saran yang membangun guna penyempurnaan buku ini di masa yang akan datang. Semoga buku ini dapat memberikan manfaat bagi pengembangan ilmu pengetahuan, khususnya dalam bidang pengolahan data dan analisis missing value.

Terima Kasih

Penulis

DAFTAR ISI



KATA PENGANTAR	i
DAFTAR ISI	iii

BAB 1

PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	5
1.3 Tujuan Penulisan	6
1.4 Manfaat Penulisan	6

BAB 2

KONSEP DASAR DATA DAN MISSING VALUE	9
2.1 Konsep Data dalam Sistem Informasi	9
2.2 Jenis Data (Numerik, Kategorikal, Ordinal)	11
2.3 Pengertian Missing Value	13
2.4 Tipe Missing Data	14
2.5 Dampak Missing Data terhadap Analisis	19

BAB 3

TEKNIK IDENTIFIKASI DAN ANALISIS MISSING VALUE	23
3.1 Eksplorasi Dataset Smartphone	23
3.2 Statistik Deskriptif Awal	24
3.3 Analisis Missing Value per Kolom	25
3.4 Visualisasi Missing Value	29

BAB 4

METODE IMPUTASI DATA / TEKNIK PENANGANAN MISSING

VALUE	33
4.1 Mean Imputation	39
4.2 Median Imputation	40
4.3 Modus Imputation (Untuk Data Kategorikal)	41
4.4 K-Nearest Neighbor (KNN) Imputation	42
4.5 Regression Imputation	45
4.6 Multiple Imputation (MICE)	46
4.7 Panduan Pemilihan Metode Imputasi (<i>Decision Guide</i>)	47

BAB 5

IMPLEMENTASI IMPUTASI MENGGUNAKAN PYTHON ----- 51

5.1 Library yang Digunakan	51
5.2 Tahapan Data Cleaning	53
5.3 Proses Imputasi pada Dataset Smartphone	56
5.4 Hasil Setelah Imputasi	58
5.5 Visualisasi Perbandingan Hasil Imputasi	65
5.6 Kesimpulan	71

BAB 6

EVALUASI KUALITAS DATA SETELAH IMPUTASI----- 73

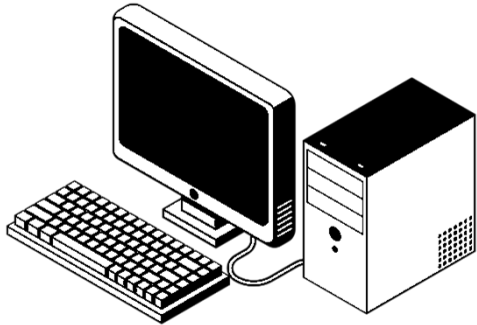
6.1 Perbandingan Statistik Sebelum dan Sesudah Imputasi	74
6.2 Perubahan Distribusi Data	80
6.3 Dampak terhadap Variansi dan Standar Deviasi	82
6.4 Analisis Potensi Bias	86
6.5 Kesimpulan Evaluasi	95

DAFTAR PUSTAKA ----- 97

GLOSARIUM ----- 101

LAMPIRAN ----- 107

PROFIL PENULIS----- 123



PENDAHULUAN

1.1 LATAR BELAKANG

Perkembangan teknologi informasi dan analisis data dalam beberapa dekade terakhir telah menghasilkan berbagai dataset dalam jumlah besar yang berasal dari berbagai sumber, seperti sensor lingkungan, sistem informasi, perangkat mobile, serta platform digital lainnya. Data tersebut menjadi komponen penting dalam berbagai proses analisis, mulai dari pemodusan sistem, prediksi, hingga pengambilan keputusan berbasis data. Dalam konteks ilmu komputer, data science, dan sistem informasi, kualitas data menjadi faktor yang sangat menentukan keberhasilan suatu analisis. Dataset yang berkualitas tinggi mampu menghasilkan modul analisis yang akurat dan dapat diandalkan, sedangkan dataset yang tidak lengkap atau memiliki kualitas rendah dapat menurunkan kinerja modul serta menghasilkan kesimpulan yang bias.

Salah satu permasalahan yang paling umum ditemukan dalam pengolahan data adalah **missing value** atau nilai data yang hilang. Missing value terjadi ketika suatu atribut dalam dataset tidak memiliki nilai yang tercatat akibat berbagai faktor seperti kesalahan pengumpulan data, kerusakan sensor, gangguan komunikasi,

maupun kesalahan input manusia. Dalam berbagai penelitian, missing value dianggap sebagai salah satu tantangan utama dalam proses *data preprocessing* karena dapat mempengaruhi integritas data serta akurasi modul analisis yang dihasilkan. Penelitian mengenai data observasi suhu jangka panjang menunjukkan bahwa keberadaan missing value dapat mengganggu karakteristik spasial dan temporal dari data serta menurunkan keandalan hasil analisis jika tidak ditangani dengan tepat. (Guo & Holy, 2026)

Permasalahan missing data juga ditemukan dalam berbagai domain penelitian lainnya, seperti bidang keuangan, energi, transportasi, dan lingkungan. Pada penelitian mengenai data **Environmental, Social, and Governance (ESG)**, misalnya, ditemukan bahwa hingga sekitar **50% indikator kinerja utama dalam dataset ESG mengandung missing value**, yang berpotensi menyebabkan bias dalam penilaian kinerja perusahaan. Penelitian tersebut menunjukkan bahwa perusahaan dengan data yang lebih lengkap cenderung memperoleh skor evaluasi yang lebih tinggi, bukan semata-mata karena performa yang lebih baik tetapi karena ketersediaan data yang lebih lengkap. Hal ini menunjukkan bahwa missing value dapat menyebabkan distorsi dalam proses evaluasi dan pengambilan keputusan berbasis data. (Yu et al., 2026)

Selain itu, dalam bidang pemodulan energi bangunan, kualitas data juga menjadi faktor penting dalam menentukan akurasi modul prediksi. Penelitian mengenai database performa energi bangunan menunjukkan bahwa atribut penting seperti **nilai U (thermal transmittance)** sering kali tidak tersedia secara lengkap dalam dataset bangunan. Ketidaklengkapan data tersebut dapat menyebabkan ketidakpastian dalam estimasi konsumsi energi bangunan serta menurunkan keandalan modul simulasi energi. Oleh karena itu, teknik penanganan missing value menjadi langkah

penting dalam meningkatkan kualitas dataset sebelum digunakan dalam analisis lebih lanjut. (Li et al., 2026a)

Permasalahan missing value juga banyak ditemukan pada data yang bersumber dari sensor dan sistem pemantauan. Dalam sistem transportasi cerdas, misalnya, data lalu lintas sering mengalami kehilangan data akibat kerusakan sensor, gangguan jaringan komunikasi, atau kegagalan sistem pencatatan. Beberapa studi melaporkan bahwa tingkat kehilangan data pada sistem lalu lintas dapat mencapai lebih dari **20% bahkan hingga 50% dalam beberapa kasus**, sehingga mempengaruhi akurasi modul prediksi lalu lintas yang dibangun berdasarkan dataset tersebut. (Li et al., 2026a)

Keberadaan missing value dalam dataset dapat menimbulkan berbagai dampak negatif, antara lain menurunnya akurasi modul prediksi, munculnya bias dalam analisis statistik, serta berkurangnya reliabilitas sistem yang dibangun berdasarkan data tersebut. Salah satu pendekatan yang sering digunakan untuk mengatasi permasalahan ini adalah dengan melakukan **imputasi data**, yaitu proses mengisi nilai yang hilang dengan nilai estimasi yang diperoleh melalui metode statistik maupun algoritma pembelajaran mesin. Metode imputasi yang umum digunakan meliputi teknik sederhana seperti **mean imputation dan median imputation**, serta metode yang lebih kompleks seperti **k-nearest neighbors (KNN), multiple imputation, regression-based imputation, dan modul berbasis machine learning**. Beberapa penelitian menunjukkan bahwa metode berbasis machine learning sering kali memberikan hasil yang lebih akurat dibandingkan metode statistik sederhana karena mampu menangkap hubungan antar variabel dalam dataset secara lebih baik. (Liu et al., 2025)

Dalam konteks **sistem pendukung keputusan (SPK)**, kualitas data menjadi aspek yang sangat penting karena hasil keputusan yang

dihasilkan sangat bergantung pada kualitas dataset yang digunakan. SPK merupakan sistem berbasis komputer yang dirancang untuk membantu pengambil keputusan dalam menyelesaikan permasalahan kompleks dengan memanfaatkan data dan modul analisis tertentu. Apabila dataset yang digunakan dalam SPK mengandung missing value yang tidak ditangani dengan baik, maka proses evaluasi alternatif dan perankingan keputusan dapat menghasilkan hasil yang tidak akurat.(Febrian et al., 2023)

Salah satu contoh penerapan SPK yang banyak digunakan adalah dalam **pemilihan produk teknologi**, termasuk pemilihan smartphone. Industri smartphone mengalami perkembangan yang sangat pesat dengan berbagai variasi spesifikasi produk yang tersedia di pasar. Konsumen dihadapkan pada berbagai pilihan perangkat dengan karakteristik yang berbeda-beda seperti kapasitas RAM, penyimpanan internal, kapasitas baterai, kualitas kamera, refresh rate layar, serta harga perangkat. Oleh karena itu, sistem pendukung keputusan dapat digunakan untuk membantu pengguna dalam menentukan pilihan smartphone yang paling sesuai dengan kebutuhan berdasarkan berbagai kriteria evaluasi.

Namun demikian, dataset spesifikasi smartphone yang tersedia dari berbagai sumber sering kali mengandung missing value pada beberapa atribut penting, seperti refresh rate layar, kapasitas kartu memori maksimum, kemampuan fast charging, maupun spesifikasi perangkat lainnya. Apabila dataset tersebut digunakan secara langsung dalam proses pengambilan keputusan tanpa penanganan missing value yang tepat, maka hasil perankingan alternatif yang dihasilkan dapat menjadi kurang akurat dan berpotensi menyesatkan pengguna.(Amalia et al., 2024)

Berdasarkan permasalahan tersebut, diperlukan analisis yang komprehensif mengenai teknik penanganan missing value serta pengaruhnya terhadap proses pengambilan keputusan berbasis

data. Oleh karena itu, buku ini membahas **analisis komparatif berbagai teknik imputasi data pada dataset smartphone serta implikasinya terhadap akurasi sistem pendukung keputusan.**

Pembahasan tidak hanya berfokus pada metode pengisian data yang hilang, tetapi juga menganalisis bagaimana perbedaan teknik imputasi dapat mempengaruhi kualitas dataset dan hasil keputusan yang dihasilkan oleh sistem pendukung keputusan. Dengan demikian, buku ini diharapkan dapat memberikan kontribusi akademik dalam memahami pentingnya pengolahan missing value dalam analisis data serta implikasinya terhadap kualitas keputusan berbasis data.

1.2 RUMUSAN MASALAH

Berdasarkan latar belakang yang telah diuraikan sebelumnya, permasalahan utama dalam buku ini berkaitan dengan keberadaan *missing value* pada dataset spesifikasi smartphone yang dapat mempengaruhi kualitas analisis data serta hasil pengambilan keputusan. Ketidaklengkapan data berpotensi menimbulkan bias dalam proses evaluasi alternatif dan perankingan dalam sistem pendukung keputusan. Oleh karena itu, perlu dilakukan kajian mengenai karakteristik *missing value* yang terdapat pada dataset smartphone serta bagaimana teknik imputasi data dapat digunakan untuk mengatasi permasalahan tersebut. Selain itu, penting untuk menganalisis perbandingan kinerja berbagai metode imputasi dalam menghasilkan dataset yang lebih lengkap dan konsisten. Permasalahan lain yang menjadi fokus kajian adalah bagaimana pengaruh penggunaan teknik imputasi data terhadap hasil perankingan dalam sistem pendukung keputusan pemilihan smartphone serta bagaimana implikasi penggunaan metode imputasi yang berbeda terhadap kualitas keputusan yang dihasilkan.

1.3 TUJUAN PENULISAN

Tujuan utama penulisan buku ini adalah untuk memberikan pemahaman teoritis dan praktis mengenai penanganan *missing value* dalam dataset serta dampaknya terhadap proses pengambilan keputusan berbasis data. Secara khusus, buku ini bertujuan untuk mengidentifikasi karakteristik dan distribusi *missing value* pada dataset smartphone, menjelaskan berbagai metode imputasi data yang dapat digunakan untuk menangani data yang hilang, serta mengimplementasikan beberapa teknik imputasi pada dataset tersebut. Selain itu, buku ini juga bertujuan untuk menganalisis perbandingan hasil dari berbagai metode imputasi dalam meningkatkan kualitas dataset serta mengevaluasi pengaruh hasil imputasi terhadap proses pengambilan keputusan dalam sistem pendukung keputusan berbasis multi kriteria.

1.4 MANFAAT PENULISAN

Penulisan buku ini diharapkan dapat memberikan manfaat baik secara teoritis maupun praktis. Secara teoritis, buku ini diharapkan dapat memberikan kontribusi dalam pengembangan ilmu pengetahuan khususnya pada bidang *data science*, *machine learning*, dan sistem pendukung keputusan, terutama dalam kajian mengenai kualitas data dan teknik penanganan *missing value*. Buku ini juga dapat menjadi referensi akademik bagi mahasiswa, peneliti, maupun praktisi yang ingin mempelajari teknik pengolahan data tidak lengkap serta dampaknya terhadap proses analisis dan pengambilan keputusan berbasis data.

Secara praktis, buku ini diharapkan dapat memberikan panduan bagi pengembang sistem, analis data, maupun praktisi teknologi informasi dalam memahami permasalahan *missing value* yang sering muncul dalam dataset nyata. Selain itu, buku ini dapat

membantu dalam memilih teknik imputasi data yang tepat, meningkatkan kualitas dataset sebelum digunakan dalam proses analisis, serta meningkatkan akurasi dan reliabilitas keputusan yang dihasilkan oleh sistem berbasis data. Buku ini juga dapat dimanfaatkan sebagai bahan ajar pada mata kuliah *data mining*, *machine learning*, maupun sistem pendukung keputusan, khususnya dalam pembahasan mengenai kualitas data dan teknik *data preprocessing*.

DAFTAR PUSTAKA



- Al Islami, H., & Rosyani, P. (2024). Penerapan Metode Preference Selection Index (PSI) Untuk Menentukan Maklon Skincare Terjangkau. *KLIK: Kajian Ilmiah Informatika Dan Komputer*, 4(4), 2139–2147. <https://doi.org/10.30865/klik.v4i4.1720>
- Amalia, R., Ikasari, I. H., & Rosyani, P. (2024). Sistem Pendukung Keputusan Pemilihan Maklon SkinCare Terbaik dengan Metode VIKOR. In *Journal of Decision Support System Research* (Vol. 1, Number 2).
- Angulo, D., & Murray, S. (2026a). Multiple imputation of censored bivariate event-times via inverse transform and nonparametric Gibbs sampling. *Statistics and Probability Letters*, 228, 110564. <https://doi.org/10.1016/j.spl.2025.110564>
- Angulo, D., & Murray, S. (2026b). Multiple imputation of censored bivariate event-times via inverse transform and nonparametric Gibbs sampling. *Statistics and Probability Letters*, 228, 110564. <https://doi.org/10.1016/j.spl.2025.110564>
- Bhandari, A., & Tyagi, S. (2026). A comparative evaluation of handling missing data points and modalities in electronic health records. In *International Journal of Medical Informatics* (Vol. 210). Elsevier Ireland Ltd. <https://doi.org/10.1016/j.ijmedinf.2026.106302>
- Caruso, C. M., Soda, P., & Guarrasi, V. (2026). Not another imputation method: A transformer-based model for missing values in

- tabular datasets. *AI Open*. <https://doi.org/10.1016/j.aiopen.2026.02.005>
- de Sousa Guilhon Araujo, M. A., Cyrino Oliveira, F. L., & Souza, R. C. (2026). Missing data imputation using climate data reanalysis in solar energy generation time series. *Energy Conversion and Management: X*, 30(August 2025), 101653. <https://doi.org/10.1016/j.ecmx.2026.101653>
- Downing, N. J. (2025). Missing value imputation in environmental, social, and governance data: an impact on emissions scores. *Finance Research Letters*, 85, 1–10. <https://doi.org/10.1016/j.frl.2025.107818>
- Febrian, A., Oktavio Selan, M., Purnama, R., Surya Bernada, R., & Rosyani, P. (2023). Penerapan Metode SMART Dalam Menentukan Aplikasi Pesan-Antar Terbaik. *Jurnal Artificial Intelligent Dan Sistem Penunjang Keputusan*, 1(3). <https://jurnalmahasiswa.com/index.php/aidanspk>
- Guo, Y., & Holy, T. E. (2026). Recovering Missing Features in Nonnegative Matrix Factorization via Generalized Singular Value Decomposition. *IScience*, 114708. <https://doi.org/10.1016/j.isci.2026.114708>
- Kim, S., Lee, T. H., & Lee, J. (2025). TMF-GNN: Temporal matrix factorization-based graph neural network for multivariate time series forecasting with missing values. *Expert Systems with Applications*, 275. <https://doi.org/10.1016/j.eswa.2025.127001>
- Li, W., Subašić, G., Korolija, I., Guida, R., & Hong, S. M. (2026a). Data imputation methods for missing U-values of building envelopes in building performance database. *Journal of Building Engineering*, 118. <https://doi.org/10.1016/j.jobe.2025.115046>
- Li, W., Subašić, G., Korolija, I., Guida, R., & Hong, S. M. (2026b). Data imputation methods for missing U-values of building envelopes in building performance database. *Journal of*

- Building Engineering*, 118, 1–6. <https://doi.org/10.1016/j.jobe.2025.115046>
- Liu, X., Wang, B., Jin, S., & Song, Z. (2025). Missing value interpolation algorithm for long-term temperature observation data based on data augmentation multiple interpolation method. *Results in Engineering*, 27. <https://doi.org/10.1016/j.rineng.2025.106211>
- Patel, D., Loyola-González, O., Trivedi, A., Rajgor, H., Mehta, T., Patel, S., Vyas, P., Ganatra, N., & Patel, H. I. (2022). Single and Multiple Imputation Techniques to Treat Missing Numerical Variables (MNV) in Perspectives of Data Science Project - A Case Study. *International Journal of Engineering Trends and Technology*, 70(5), 9–14. <https://doi.org/10.14445/22315381/IJETT-V70I5P202>
- Pereira, R. C., Abreu, P. H., & Rodrigues, P. P. (2024). Siamese Autoencoder Architecture for the Imputation of Data Missing Not at Random. *Journal of Computational Science*, 78. <https://doi.org/10.1016/j.jocs.2024.102269>
- Pereira, R. C., Abreu, P. H., Rodrigues, P. P., & Figueiredo, M. A. T. (2024). Imputation of data Missing Not at Random: Artificial generation and benchmark analysis. *Expert Systems with Applications*, 249. <https://doi.org/10.1016/j.eswa.2024.123654>
- Permadi, Y., & Rosyani, P. (2025). BULLETIN OF COMPUTER SCIENCE RESEARCH Penerapan Integrasi Metode AHP dan Entropy dalam Pengambilan Keputusan Seleksi Beasiswa Perguruan Tinggi. *Media Online*, 6(1), 539–547. <https://doi.org/10.47065/bulletincsr.v6i1.936>
- Rácz, A., & Gere, A. (2025). Comparison of missing value imputation tools for machine learning models based on product development cases studies. *LWT*, 221. <https://doi.org/10.1016/j.lwt.2025.117585>

- Raj, S., Raj, S., Namdeo, V., & Srivastava, A. (2024). Decoding the gene-disease associations in type 2 diabetes: A curated dataset for text mining-based classification. *Data in Brief*, 54. <https://doi.org/10.1016/j.dib.2024.110418>
- Shi, Y., Zhong, H., Rogalski, J. C., & Foster, L. J. (2025). Optimizing imputation strategies for mass spectrometry-based proteomics considering intensity and missing value rates. *Computational and Structural Biotechnology Journal*, 27, 1818–1826. <https://doi.org/10.1016/j.csbj.2025.04.041>
- Umar, N., & Gray, A. (2023). Comparing Single and Multiple Imputation Approaches for Missing Values in Univariate and Multivariate Water Level Data. *Water (Switzerland)*, 15(8), 1–21. <https://doi.org/10.3390/w15081519>
- Yu, Y., Guo, L., Chen, D., Jiang, G., Wang, B., Tong, Y., & Dong, J. (2026). Dynamic Dual-Channel Asynchronous Graph Neural Network for Traffic Prediction with Missing Values. *Fundamental Research*. <https://doi.org/10.1016/j.fmre.2026.01.019>

GLOSARIUM



Algoritma

Serangkaian langkah atau prosedur sistematis yang digunakan untuk menyelesaikan suatu masalah atau melakukan proses komputasi dalam pengolahan data.

Analisis Data

Proses pengolahan, interpretasi, dan evaluasi data untuk memperoleh informasi yang bermakna serta mendukung pengambilan keputusan.

Analisis Statistik

Metode analisis yang menggunakan teknik statistik untuk memahami karakteristik data, mengidentifikasi pola, dan menarik kesimpulan dari suatu dataset.

ANOVA (Analysis of Variance)

Metode statistik yang digunakan untuk menguji apakah terdapat perbedaan signifikan antara rata-rata dari beberapa kelompok data.

Atribut

Karakteristik atau variabel yang menggambarkan suatu objek dalam dataset.

Bias Data

Kondisi dimana hasil analisis atau model dipengaruhi oleh distribusi data yang tidak representatif sehingga menghasilkan kesimpulan yang tidak akurat.

Boxplot

Metode visualisasi statistik yang digunakan untuk menggambarkan distribusi data berdasarkan kuartil, median, serta keberadaan outlier.

Cleaning Data (Data Cleaning)

Proses pembersihan dataset dari kesalahan, data duplikat, atau data yang tidak lengkap sebelum dilakukan analisis lebih lanjut.

Clipping

Teknik pembatasan nilai hasil prediksi agar tetap berada dalam rentang nilai yang realistis sesuai dengan karakteristik data asli.

Dataset

Kumpulan data yang tersusun secara terstruktur dalam bentuk tabel yang terdiri dari baris (observasi) dan kolom (atribut).

Distribusi Data

Pola penyebaran nilai dalam suatu dataset yang menunjukkan frekuensi kemunculan nilai dalam rentang tertentu.

Euclidean Distance

Metode pengukuran jarak antara dua titik dalam ruang multidimensi yang sering digunakan dalam algoritma KNN.

Evaluasi Model

Proses penilaian kinerja suatu metode atau model analisis untuk mengetahui tingkat akurasi dan keandalannya.

Feature (Fitur)

Variabel atau atribut yang digunakan sebagai input dalam proses analisis data atau machine learning.

Gradient Boosted Trees (GBT)

Algoritma machine learning berbasis ensemble yang menggabungkan beberapa decision tree untuk meningkatkan akurasi prediksi.

Heatmap

Visualisasi data berbentuk matriks warna yang digunakan untuk menunjukkan intensitas atau pola tertentu dalam dataset.

Imputasi Data

Proses pengisian nilai yang hilang dalam dataset menggunakan metode statistik atau algoritma tertentu.

Iterasi

Proses pengulangan langkah perhitungan secara berulang hingga mencapai kondisi tertentu atau konvergensi.

K-Nearest Neighbor (KNN)

Algoritma machine learning yang menentukan nilai suatu data berdasarkan kemiripan dengan sejumlah data terdekat.

KNN Imputation

Metode imputasi yang menggunakan algoritma KNN untuk memperkirakan nilai yang hilang berdasarkan data yang memiliki karakteristik serupa.

Korelasi

Ukuran statistik yang menunjukkan hubungan antara dua variabel dalam dataset.

Label Encoding

Teknik transformasi data kategorikal menjadi bentuk numerik agar dapat diproses oleh algoritma machine learning.

Machine Learning

Cabang kecerdasan buatan yang memungkinkan komputer mempelajari pola dari data untuk melakukan prediksi atau klasifikasi.

MAR (Missing At Random)

Jenis missing data dimana kemungkinan terjadinya missing value bergantung pada variabel lain yang dapat diamati dalam dataset.

MCAR (Missing Completely At Random)

Jenis missing data dimana kehilangan data terjadi secara acak dan tidak dipengaruhi oleh variabel lain dalam dataset.

Mean (Rata-rata)

Nilai rata-rata yang diperoleh dari penjumlahan seluruh nilai data kemudian dibagi dengan jumlah data.

Mean Absolute Error (MAE)

Metrik evaluasi yang mengukur rata-rata selisih absolut antara nilai asli dan nilai hasil prediksi atau imputasi.

Mean Imputation

Metode imputasi yang menggantikan nilai yang hilang dengan nilai rata-rata dari variabel tersebut.

Median

Nilai tengah dari data yang telah diurutkan.

Median Imputation

Metode imputasi yang menggantikan nilai yang hilang dengan nilai median dari suatu variabel.

Missing Data

Data yang tidak tersedia atau tidak tercatat pada suatu atribut dalam dataset.

Missing Value

Nilai yang hilang atau kosong pada suatu variabel dalam dataset.

MNAR (Missing Not At Random)

Jenis missing data dimana kemungkinan terjadinya missing value bergantung pada nilai data itu sendiri.

Model Machine Learning

Representasi matematis yang dibangun dari data untuk melakukan prediksi atau klasifikasi.

Modus (Mode)

Nilai yang paling sering muncul dalam suatu dataset.

Modus Imputation

Metode imputasi yang menggantikan nilai yang hilang dengan nilai yang paling sering muncul dalam suatu variabel.

Multiple Imputation

Teknik imputasi yang menghasilkan beberapa dataset lengkap untuk memperhitungkan ketidakpastian dari nilai yang hilang.

MICE (Multiple Imputation by Chained Equations)

Metode multiple imputation yang menggunakan regresi berantai untuk memperkirakan nilai yang hilang.

Normalisasi Data

Proses penyesuaian skala nilai data agar berada dalam rentang tertentu sehingga memudahkan proses analisis.

Observasi

Satu baris data dalam dataset yang merepresentasikan satu objek atau entitas tertentu.

Outlier

Nilai data yang berada jauh di luar rentang mayoritas data dalam dataset.

Pooling

Proses penggabungan hasil analisis dari beberapa dataset hasil multiple imputation.

Preprocessing Data

Tahapan awal dalam pengolahan data yang meliputi pembersihan, transformasi, dan penanganan missing value.

Regression Imputation

Metode imputasi yang menggunakan model regresi untuk memprediksi nilai yang hilang berdasarkan variabel lain.

Root Mean Squared Error (RMSE)

Metrik evaluasi yang mengukur tingkat kesalahan prediksi dengan memberikan penalti lebih besar pada kesalahan yang besar.

Statistik Deskriptif

Teknik statistik yang digunakan untuk menggambarkan karakteristik dasar suatu dataset.

Standar Deviasi

Ukuran statistik yang menunjukkan seberapa jauh penyebaran data terhadap nilai rata-rata.

Structural Missing

Missing value yang terjadi karena atribut tersebut memang tidak relevan untuk objek tertentu.

Variabel

Atribut yang nilainya dapat berubah atau berbeda antar observasi dalam dataset.

Variansi

Ukuran statistik yang menunjukkan tingkat penyebaran data terhadap nilai rata-ratanya.

Visualisasi Data

Proses penyajian data dalam bentuk grafik atau gambar untuk memudahkan pemahaman pola dan hubungan dalam data.

LAMPIRAN



```
Import library yang diperlukan
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.impute import SimpleImputer,
KNNImputer
from sklearn.experimental import
enable_iterative_imputer
from sklearn.impute import IterativeImputer
from sklearn.linear_model import BayesianRidge
from sklearn.preprocessing import StandardScaler,
LabelEncoder
import warnings
warnings.filterwarnings('ignore')

# Load dataset
df = pd.read_csv('/kaggle/input/smartphone-
specifications-dataset/cleaned_data.csv') #
Sesuaikan dengan path file Anda

print("="*80)
print("ANALISIS PERBANDINGAN 6 METODE IMPUTASI")
print("="*80)
print(f"Dataset shape: {df.shape}")
print(f"Total missing values awal:
{df.isnull().sum().sum()}")
```

```

#
=====
# 1. ANALISIS MISSING VALUE AWAL
#
=====
print("\n" + "="*80)
print("1. ANALISIS MISSING VALUE SEBELUM
IMPUTASI")
print("="*80)

# Hitung missing value per kolom
missing_before = df.isnull().sum()
missing_percent = (df.isnull().sum() / len(df)) *
100

missing_df = pd.DataFrame({
    'Kolom': missing_before.index,
    'Jumlah Missing': missing_before.values,
    'Persentase (%)': missing_percent.values
})
missing_df = missing_df[missing_df['Jumlah
Missing'] > 0].sort_values('Jumlah Missing',
ascending=False)

print("\nKolom dengan missing value:")
print(missing_df.to_string(index=False))

# Visualisasi missing value sebelum imputasi
fig, axes = plt.subplots(2, 3, figsize=(18, 12))
fig.suptitle('Visualisasi Missing Value Sebelum
Imputasi', fontsize=16, fontweight='bold')

# Heatmap missing value
sns.heatmap(df.isnull(), cbar=False,
yticklabels=False, cmap='viridis', ax=axes[0,0])
axes[0,0].set_title('1 Heatmap Missing Value
(Sebelum)')

```

```

axes[0,0].set_xlabel('Kolom')

# Barplot jumlah missing
missing_before[missing_before >
0].sort_values().plot(kind='barh', ax=axes[0,1],
color='coral')
axes[0,1].set_title('2 Jumlah Missing per Kolom
(Sebelum)')
axes[0,1].set_xlabel('Jumlah Missing')

# Pie chart persentase missing
sizes = [df.isnull().sum().sum(), df.size -
df.isnull().sum().sum()]
labels = ['Missing', 'Tidak Missing']
colors = ['#ff6b6b', '#51cf66']
axes[0,2].pie(sizes, labels=labels,
colors=colors, autopct='%1.1f%%', startangle=90)
axes[0,2].set_title('3 Proporsi Missing Value')

# Matriks missing value
def plot_missing_matrix(data, ax):
    missing_data = data.isnull()
    ax.imshow(missing_data.T, aspect='auto',
cmap='RdBu_r', interpolation='nearest')
    ax.set_yticks(range(len(data.columns)))
    ax.set_yticklabels(data.columns, fontsize=8)
    ax.set_xlabel('Baris Data')
    ax.set_title('4 Missing Value Matrix')

plot_missing_matrix(df, axes[1,0])

# Distribusi data sebelum imputasi (sample kolom
numerik)
numeric_cols =
df.select_dtypes(include=[np.number]).columns[:5]
# Ambil 5 kolom numerik pertama
for i, col in enumerate(numeric_cols):
    if i < 2: # Hanya plot 2 kolom untuk contoh

```

```

        df[col].dropna().hist(bins=30, alpha=0.7,
ax=axes[1, i+1], color='steelblue',
edgecolor='black')
        axes[1, i+1].set_title(f'Distribusi {col}
(Sebelum)')
        axes[1, i+1].set_xlabel(col)
        axes[1, i+1].set_ylabel('Frekuensi')

```

```

plt.tight_layout()
plt.show()

```

```

#
=====
=====
# 2. FUNGSI UNTUK IMPLEMENTASI 6 METODE IMPUTASI
#
=====
=====
print("\n" + "="*80)
print("2. IMPLEMENTASI 6 METODE IMPUTASI")
print("="*80)

```

```

# Pisahkan kolom numerik dan kategorikal
numeric_cols =
df.select_dtypes(include=[np.number]).columns.tolist()
categorical_cols =
df.select_dtypes(include=['object']).columns.tolist()

```

```

# Buat dictionary untuk menyimpan hasil imputasi
imputed_datasets = {}
imputation_stats = []

```

```

# Metode 1: Mean Imputation (untuk numerik)
def mean_imputation(df, numeric_cols,
categorical_cols):
    df_mean = df.copy()

```

```

# Mean imputation untuk numerik
for col in numeric_cols:
    if df_mean[col].isnull().sum() > 0:
        mean_val = df_mean[col].mean()
        df_mean[col].fillna(mean_val,
inplace=True)

# Untuk kategorikal tetap seperti aslinya
(atau bisa juga diisi dengan modus nanti)

return df_mean

# Metode 2: Median Imputation
def median_imputation(df, numeric_cols,
categorical_cols):
    df_median = df.copy()

    for col in numeric_cols:
        if df_median[col].isnull().sum() > 0:
            median_val = df_median[col].median()
            df_median[col].fillna(median_val,
inplace=True)

    return df_median

# Metode 3: Modus Imputation (untuk kategorikal)
def modus_imputation(df, numeric_cols,
categorical_cols):
    df_modus = df.copy()

    # Untuk numerik, isi dengan modus (untuk
perbandingan)
    for col in numeric_cols:
        if df_modus[col].isnull().sum() > 0:
            modus_val = df_modus[col].modus()[0]
if not df_modus[col].modus().empty else 0
            df_modus[col].fillna(modus_val,
inplace=True)

```

```

        # Untuk kategorikal, isi dengan modus
        for col in categorical_cols:
            if col in df_modus.columns and
df_modus[col].isnull().sum() > 0:
                modus_val = df_modus[col].modus()[0]
if not df_modus[col].modus().empty else 'Unknown'
                df_modus[col].fillna(modus_val,
inplace=True)

        return df_modus

# Metode 4: KNN Imputation
def knn_imputation(df, numeric_cols,
categorical_cols, n_neighbors=5):
    df_knn = df.copy()

    # Untuk KNN, kita hanya imputasi kolom
numerik dulu
    if numeric_cols:
        # Normalisasi data
        scaler = StandardScaler()
        df_numeric_scaled = pd.DataFrame(

scaler.fit_transform(df_knn[numeric_cols]),
            columns=numeric_cols
        )

        # KNN Imputer
        knn_imputer =
KNNImputer(n_neighbors=n_neighbors)
        df_numeric_imputed = pd.DataFrame(

knn_imputer.fit_transform(df_numeric_scaled),
            columns=numeric_cols
        )

        # Kembalikan ke skala asli
        df_numeric_imputed = pd.DataFrame(

```

```

scaler.inverse_transform(df_numeric_imputed),
    columns=numeric_cols
)

# Update dataset
for col in numeric_cols:
    df_knn[col] = df_numeric_imputed[col]

return df_knn

# Metode 5: Regression Imputation (menggunakan
IterativeImputer)
def regression_imputation(df, numeric_cols,
categorical_cols):
    df_reg = df.copy()

    if numeric_cols:
        # Gunakan IterativeImputer dengan
Bayesian Ridge sebagai estimator
        imputer = IterativeImputer(
            estimator=BayesianRidge(),
            max_iter=10,
            random_state=42,
            sample_posterior=True
        )

        df_numeric_imputed = pd.DataFrame(
imputer.fit_transform(df_reg[numeric_cols]),
            columns=numeric_cols
        )

        for col in numeric_cols:
            df_reg[col] = df_numeric_imputed[col]

    return df_reg

```

```

# Metode 6: MICE (Multiple Imputation by Chained
Equations)
def mice_imputation(df, numeric_cols,
categorical_cols, n_imputations=5):
    """
    Implementasi MICE sederhana - menghasilkan
beberapa dataset
    dan mengambil rata-ratanya
    """
    df_mice = df.copy()

    if numeric_cols:
        imputed_datasets_list = []

        for i in range(n_imputations):
            imputer = IterativeImputer(
                estimator=BayesianRidge(),
                max_iter=20,
                random_state=42 + i,
                sample_posterior=True
            )

            df_imputed = pd.DataFrame(
imputer.fit_transform(df_mice[numeric_cols]),
                columns=numeric_cols
            )

            imputed_datasets_list.append(df_imputed)

            # Rata-ratakan hasil dari multiple
imputations
            df_avg = sum(imputed_datasets_list) /
n_imputations

            for col in numeric_cols:
                df_mice[col] = df_avg[col]

    return df_mice

```

```

# Terapkan semua metode
print("\nMenerapkan 6 metode imputasi...")
df_mean = mean_imputation(df, numeric_cols,
categorical_cols)
print("✓ Mean Imputation selesai")

df_median = median_imputation(df, numeric_cols,
categorical_cols)
print("✓ Median Imputation selesai")

df_modus = modus_imputation(df, numeric_cols,
categorical_cols)
print("✓ Modus Imputation selesai")

df_knn = knn_imputation(df, numeric_cols,
categorical_cols, n_neighbors=5)
print("✓ KNN Imputation selesai")

df_reg = regression_imputation(df, numeric_cols,
categorical_cols)
print("✓ Regression Imputation selesai")

df_mice = mice_imputation(df, numeric_cols,
categorical_cols, n_imputations=5)
print("✓ MICE Imputation selesai")

# Simpan hasil dalam dictionary
imputed_datasets = {
    'Mean': df_mean,
    'Median': df_median,
    'Modus': df_modus,
    'KNN': df_knn,
    'Regression': df_reg,
    'MICE': df_mice
}

```

```

#
=====
=====
# 3. ANALISIS PERBANDINGAN HASIL IMPUTASI
#
=====
=====
print("\n" + "="*80)
print("3. ANALISIS PERBANDINGAN 6 METODE
IMPUTASI")
print("="*80)

# Tabel perbandingan missing value setelah
imputasi
comparison_data = []
for name, dataset in imputed_datasets.items():
    missing_count = dataset.isnull().sum().sum()
    missing_cols =
dataset.isnull().sum()[dataset.isnull().sum() >
0].index.tolist()

    comparison_data.append({
        'Metode': name,
        'Total Missing': missing_count,
        'Status': ' Lengkap' if missing_count
== 0 else f' {missing_count} missing',
        'Kolom Tersisa': ', '.join(missing_cols)
if missing_cols else 'Semua lengkap'
    })

comparison_df = pd.DataFrame(comparison_data)
print("\nPerbandingan Hasil Imputasi:")
print(comparison_df.to_string(index=False))

# Statistik deskriptif perbandingan
print("\n" + "="*80)
print("4. STATISTIK DESKRIPTIF PERBANDINGAN")
print("="*80)

```

```

# Pilih beberapa kolom numerik kunci untuk
dianalisis
key_columns = ['price', 'ram_gb', 'battery_mah',
'fast_charge_w', 'refresh_rate_hz']
available_key_cols = [col for col in key_columns
if col in numeric_cols]

for col in available_key_cols:
    print(f"\n--- Perbandingan Statistik untuk
{col} ---")

    stats_data = []

    # Data asli (sebelum imputasi)
    orig_data = df[col].dropna()
    stats_data.append({
        'Metode': 'Original (sebelum)',
        'Mean': orig_data.mean(),
        'Median': orig_data.median(),
        'Std': orig_data.std(),
        'Min': orig_data.min(),
        'Max': orig_data.max()
    })

    # Data setelah imputasi
    for name, dataset in
imputed_datasets.items():
        stats_data.append({
            'Metode': name,
            'Mean': dataset[col].mean(),
            'Median': dataset[col].median(),
            'Std': dataset[col].std(),
            'Min': dataset[col].min(),
            'Max': dataset[col].max()
        })

stats_df = pd.DataFrame(stats_data)

```

```

print(stats_df.round(2).to_string(index=False))

#
=====
=====
# 5. VISUALISASI PERBANDINGAN
#
=====
=====
print("\n" + "="*80)
print("5. VISUALISASI PERBANDINGAN HASIL
IMPUTASI")
print("="*80)

# Visualisasi 1: Heatmap Missing Value After
Imputation untuk setiap metode
fig, axes = plt.subplots(2, 3, figsize=(18, 10))
fig.suptitle('1 Heatmap Missing Value Setelah
Imputasi', fontsize=16, fontweight='bold')

for ax, (name, dataset) in zip(axes.flat,
imputed_datasets.items()):
    sns.heatmap(dataset.isnull(), cbar=False,
yticklabels=False, cmap='viridis', ax=ax)
    ax.set_title(f'{name} Imputation')
    ax.set_xlabel('Kolom')

plt.tight_layout()
plt.show()

# Visualisasi 2: Perbandingan Distribusi Data
Before vs After
fig, axes = plt.subplots(len(available_key_cols),
2, figsize=(16, 5*len(available_key_cols)))
fig.suptitle('2 Perbandingan Distribusi Data
Sebelum vs Setelah Imputasi', fontsize=16,
fontweight='bold')

```

```

colors = {'Original': '#2ecc71', 'Mean':
'#3498db', 'Median': '#9b59b6',
          'Modus': '#e74c3c', 'KNN': '#f39c12',
'Regression': '#1abc9c', 'MICE': '#e67e22'}

for i, col in enumerate(available_key_cols):
    # Plot distribusi original
    axes[i, 0].hist(df[col].dropna(), bins=30,
alpha=0.7, color=colors['Original'],
                    edgecolor='black',
label='Original', density=True)
    axes[i, 0].set_title(f'{col} - Distribusi
Original')
    axes[i, 0].set_xlabel(col)
    axes[i, 0].set_ylabel('Density')
    axes[i, 0].legend()

    # Plot distribusi semua metode
    for name, dataset in
imputed_datasets.items():
        axes[i, 1].hist(dataset[col], bins=30,
alpha=0.5, color=colors.get(name, 'gray'),
                    label=name, density=True)
        axes[i, 1].set_title(f'{col} - Perbandingan
Semua Metode')
        axes[i, 1].set_xlabel(col)
        axes[i, 1].set_ylabel('Density')
        axes[i, 1].legend(loc='upper right')

plt.tight_layout()
plt.show()

# Visualisasi 3: Boxplot perbandingan
fig, axes = plt.subplots(2, 3, figsize=(18, 12))
fig.suptitle('3 Boxplot Perbandingan Metode
Imputasi', fontsize=16, fontweight='bold')

```

```

for ax, col in
zip(axes.flat[:len(available_key_cols)],
available_key_cols):
    data_to_plot = [df[col].dropna()]
    labels = ['Original']

    for name, dataset in
imputed_datasets.items():
        data_to_plot.append(dataset[col])
        labels.append(name)

    bp = ax.boxplot(data_to_plot, labels=labels,
patch_artist=True)

    # Warna boxplot
    colors_box = ['#2ecc71', '#3498db',
'#9b59b6', '#e74c3c', '#f39c12', '#1abc9c',
'#e67e22']
    for patch, color in zip(bp['boxes'],
colors_box):
        patch.set_facecolor(color)

    ax.set_title(f'Perbandingan {col}')
    ax.set_ylabel('Nilai')
    ax.tick_params(axis='x', rotation=45)

plt.tight_layout()
plt.show()

# Visualisasi 4: Barplot perbandingan mean dan
median
fig, axes = plt.subplots(1, 2, figsize=(16, 6))
fig.suptitle('4 Perbandingan Mean dan Median
Antar Metode', fontsize=16, fontweight='bold')

# Plot perbandingan mean
x = np.arange(len(available_key_cols))
width = 0.1

```

```

for i, (name, dataset) in enumerate([('Original',
df)] + list(imputed_datasets.items())):
    means = [dataset[col].mean() for col in
available_key_cols]
    offset = (i - 3) * width
    axes[0].bar(x + offset, means, width,
label=name, alpha=0.8)

axes[0].set_xlabel('Kolom')
axes[0].set_ylabel('Mean')
axes[0].set_title('Perbandingan Mean')
axes[0].set_xticks(x)
axes[0].set_xticklabels(available_key_cols,
rotation=45)
axes[0].legend(loc='upper right', fontsize=8)

# Plot perbandingan median
for i, (name, dataset) in enumerate([('Original',
df)] + list(imputed_datasets.items())):
    medians = [dataset[col].median() for col in
available_key_cols]
    offset = (i - 3) * width
    axes[1].bar(x + offset, medians, width,
label=name, alpha=0.8)

axes[1].set_xlabel('Kolom')
axes[1].set_ylabel('Median')
axes[1].set_title('Perbandingan Median')
axes[1].set_xticks(x)
axes[1].set_xticklabels(available_key_cols,
rotation=45)
axes[1].legend(loc='upper right', fontsize=8)

plt.tight_layout()
plt.show()

#
=====
=====

```

```

# 6. KESIMPULAN DAN REKOMENDASI
#
=====
=====
print("\n" + "="*80)
print("6. KESIMPULAN DAN REKOMENDASI")
print("="*80)

print("\nBerdasarkan analisis perbandingan 6
metode imputasi:")
print("\n📊 Ringkasan:")
for name, dataset in imputed_datasets.items():
    changes = []
    for col in available_key_cols[:3]: # Ambil 3
kolom pertama untuk contoh
        orig_mean = df[col].dropna().mean()
        imp_mean = dataset[col].mean()
        diff_pct = ((imp_mean - orig_mean) /
orig_mean) * 100
        changes.append(f"{col}: {diff_pct:.2f}%")

    print(f"\n{name} Imputation:")
    print(f"    - Perubahan mean: {'',
'.join(changes)}")

```

PROFIL PENULIS



Oke Hariansyah, S.Kom., M.Kom., adalah seorang profesional di bidang teknologi informasi dimana focus yang di dalam adalah Cloud Computing dan Cloud Security.

Beliau meraih gelar Sarjana Komputer (S.Kom.) dan Magister Komputer (M.Kom.) dari Universitas Pamulang, di mana saat ini beliau juga mengajar sebagai dosen di Program Studi Teknik Informatika.

Sebagai akademisi dan peneliti, Oke Hariansyah memiliki minat dalam bidang kecerdasan buatan (Artificial Intelligence), pengolahan citra (image processing), dan machine learning. Salah satu karya ilmiahnya adalah "Pengenalan Citra Bunga Menggunakan Segmentasi Otsu Threshold dan Naïve Bayes", yang dipublikasikan dalam Jurnal Sistem dan Informatika (JSI).

Selain itu, beliau juga aktif dalam kegiatan pengabdian kepada masyarakat. Misalnya, beliau terlibat dalam program "Pengenalan Kecerdasan Buatan kepada Para Remaja di Komunitas Perpus Jungle Parung Panjang", yang bertujuan untuk memberikan edukasi mengenai teknologi AI kepada generasi muda.

Dengan kombinasi peran sebagai praktisi industri dan akademisi, Oke Hariansyah berkontribusi signifikan dalam pengembangan teknologi informasi di Indonesia, khususnya dalam penerapan dan edukasi teknologi terkini.



Ines Heidiani Ikasari, S.Si., M.Kom., adalah seorang akademisi dan peneliti di bidang ilmu komputer yang saat ini menjabat sebagai dosen pada Program Studi Teknik Informatika di Universitas Pamulang (Unpam).

Beliau meraih gelar Sarjana Sains (S.Si.) dan Magister Komputer (M.Kom.), dengan spesialisasi dalam ilmu komputer.

Sebagai peneliti aktif, Ines memiliki h-index 10 dan telah mendapatkan lebih dari 500 sitasi atas karya-karya ilmiahnya.

Salah satu publikasi terbarunya adalah "Classification of Pneumonia Medical Images with Convolutional Neural Networks," yang berfokus pada penggunaan jaringan saraf konvolusional untuk klasifikasi citra medis pneumonia.

Melalui perannya sebagai pendidik dan peneliti, Ines Heidiani Ikasari berperan signifikan dalam pengembangan ilmu komputer, khususnya dalam penerapan teknologi terkini untuk memecahkan berbagai permasalahan di masyarakat.



Perani Rosyani, S.Kom, M.Kom, adalah seorang dosen dan peneliti di Universitas Pamulang (Unpam), khususnya di Program Studi Teknik Informatika. Beliau lahir di Jakarta dan saat ini berdomisili di Tangerang, Banten.

Beliau meraih gelar Sarjana Komputer (S.Kom) dari Universitas Pamulang pada tahun 2013 dan gelar Magister Komputer (M.Kom) dari STMIK Eresha pada tahun 2016.

Sejak tahun 2014, Perani aktif mengajar di Fakultas Teknik Informatika Unpam. Beliau juga terlibat dalam berbagai penelitian, terutama di bidang pengolahan citra (image processing), computer vision, kecerdasan buatan (artificial intelligence), dan machine learning.

Selain itu, Perani telah berkontribusi dalam kegiatan pengabdian masyarakat, seperti pelatihan Microsoft Office di Pondok Pesantren Rahmatuttoyibah Al-Iflahah.

TEKNIK IMPUTASI DATA

Teori dan Penerapan pada Dataset Smartphone

Dalam analisis data modern, kualitas dataset menjadi faktor penting yang menentukan akurasi hasil analisis dan model yang dibangun. Salah satu permasalahan yang sering muncul dalam dataset adalah keberadaan missing value atau data yang hilang. Jika tidak ditangani dengan tepat, missing value dapat menyebabkan bias dalam analisis serta menurunkan kualitas pengambilan keputusan berbasis data.

Buku ini membahas secara sistematis konsep missing value serta berbagai teknik imputasi data yang digunakan untuk mengatasi permasalahan data yang tidak lengkap. Melalui studi kasus pada dataset spesifikasi smartphone, buku ini menyajikan analisis komparatif beberapa metode imputasi seperti *mean*, *median*, *modus*, *KNN*, *regression*, dan *MICE* untuk melihat pengaruhnya terhadap kualitas dataset.

Ditulis dengan pendekatan teoritis sekaligus praktis, buku ini dapat menjadi referensi bagi mahasiswa, peneliti, dan praktisi di bidang ilmu komputer, data science, serta sistem pendukung keputusan yang ingin memahami teknik pengolahan missing value secara lebih mendalam.



PENERBIT KBM INDONESIA
Anggota IKAPI
0813 5751 7526 / 0353 3234 874
Kantor : Maguwoharjo, Depok, Sleman-Yogyakarta
@penerbitbukujojja
@penerbit.kbm

ISBN 978-634-278-282-8

