

PERBANDINGAN METODE IMPUTASI DATA KATEGORIK PADA PERILAKU KEUANGAN DIGITAL: SIMULASI MCAR DAN MAR MIKRODATA GLOBAL FINDEX INDONESIA 2025

Dina Eka Putri^{1*}, Adis Tia Juli Agil Asri¹, Arni Nazira¹

¹ Departemen Statistika, Fakultas Sains dan Matematika, Universitas Mataram

*e-mail: dina.putri@unram.ac.id

DOI: 10.14710/j.gauss.15.2.391-401

Article Info:

Received: 2026-05-25

Accepted: 2026-10-13

Available Online: 2026-05-25

Keywords:

Global Findex; Imputasi Data Hilang; Data Survei Kategorikal; MICE; Random Forest; Inklusi Keuangan Digital

Abstract: Large-scale surveys such as the Global Findex Database are crucial for monitoring financial inclusion, yet their microdata often contain missing values that can bias empirical analyses and weaken inference. This study evaluates three imputation methods for three categorical digital financial behavior variables (Fin3, Con12, and Con26) in the 2025 Global Findex microdata for Indonesia. Missing values were artificially introduced into a complete-case reference dataset ($n = 407$) under Missing Completely at Random (MCAR) and Missing at Random (MAR) mechanisms at rates of 10%, 20%, and 30%, with 30 Monte Carlo replications per scenario. Mode imputation, Multiple Imputation by Chained Equations (MICE), and random forest imputation (missForest) were compared using the Proportion of False Classification (PFC). Because all target variables are categorical, the simple baseline was mode imputation; no numerical variable was imputed using the mean. Mode imputation yielded the lowest PFC in 17 of 18 combinations of variable, missing rate, and mechanism, followed generally by random forest and then MICE. These findings suggest that, for categorical and imbalanced survey data, simple distribution-based methods can be competitive with model-based approaches, partly because of majority-class dominance and limited predictive structure. The conclusion is limited to the PFC criterion and the scenarios examined.

1. PENDAHULUAN

Digitalisasi mengubah cara orang mengakses layanan keuangan sehingga semakin inklusif dan mudah dijangkau. Sejak 2011, Global Findex Database menjadi sumber data definitif tentang cara orang dewasa di seluruh dunia menggunakan layanan keuangan, mulai dari pembayaran, menabung, dan meminjam hingga mengelola peristiwa keuangan seperti pengeluaran besar atau kehilangan pendapatan (World Bank, 2025). Inklusi keuangan didefinisikan sebagai ketersediaan, aksesibilitas, dan pemanfaatan efektif layanan keuangan yang terjangkau (tabungan, kredit, asuransi, dan sistem pembayaran), dengan fokus pada populasi yang kurang terlayani (Farah, Nor dan Musse, 2026), serta dipandang sebagai faktor katalitik bagi pemberdayaan ekonomi, pengentasan kemiskinan, dan pembangunan berkelanjutan (Khan et al., 2024).

Data hilang terjadi ketika sebagian variabel tidak memiliki nilai pada sebagian responden. Hal ini menjadi masalah karena sebagian besar metode dan perangkat lunak statistik konvensional mengasumsikan seluruh variabel terukur untuk semua kasus (Allison, 2001), dan ketidaklengkapan data menurunkan ketelitian analisis serta efektivitas inferensi (Li et al., 2024). Penghapusan listwise, yaitu menghapus baris yang mengandung nilai hilang, mudah dilakukan tetapi dapat mengurangi daya statistik dan menghasilkan estimasi parameter yang bias ketika kehilangan terkait faktor lain yang memengaruhi hasil (Wabina et al., 2025).

Penelitian sebelumnya menunjukkan bahwa MICE menghasilkan akurasi lebih tinggi dibandingkan metode lain, sedangkan random forest (RF) mencapai keseimbangan terbaik antara akurasi dan pelestarian varians berdasarkan RMSE dan MAE, sehingga dinilai paling cocok untuk dataset kecil berdimensi tinggi (Abdullah et al., 2026). Namun, temuan tersebut berasal dari data morfometri kraniofasial, sedangkan Wabina et al. (2025) mengevaluasi data klinis campuran numerik-kategorikal. Kedua studi belum menjawab bagaimana mode imputation, MICE, dan RF berperilaku pada data survei kategorikal berskala besar dengan class imbalance parah, seperti mikrodatab Global Findex.

Tujuan penelitian ini adalah membandingkan kinerja mode imputation, MICE, dan random forest (missForest) dalam mengimputasi tiga variabel kategorikal keuangan digital (Fin3, Con12, dan Con26) pada mikrodatab Global Findex 2025 untuk Indonesia, di bawah mekanisme MCAR dan MAR dengan tingkat data hilang 10%, 20%, dan 30%. Kinerja dinilai menggunakan Proportion of False Classification (PFC) dari 30 replikasi Monte Carlo. Hasilnya diharapkan menjadi dasar empiris bagi peneliti dan praktisi survei dalam memilih metode imputasi untuk data kategorikal yang timpang, sekaligus menunjukkan kapan metode yang lebih kompleks tidak memberi keuntungan.

2. TINJAUAN PUSTAKA

Inklusi keuangan pada dasarnya merujuk pada kemampuan seseorang untuk mengakses dan memanfaatkan berbagai layanan keuangan resmi secara optimal, seperti menabung, melakukan pembayaran, mengakses pinjaman, hingga memiliki perlindungan asuransi (Sarma, 2008). Global Findex Database yang dikembangkan oleh World Bank merupakan salah satu sumber data internasional utama untuk mengukur tingkat inklusi keuangan masyarakat di berbagai negara. Data ini tidak hanya mencatat siapa saja yang memiliki rekening di lembaga keuangan formal, tetapi juga mencakup kebiasaan bertransaksi secara digital, menabung, meminjam uang, serta kemudahan akses terhadap layanan keuangan pada umumnya. Laporan Global Findex 2021 mencatat bahwa penggunaan pembayaran digital meningkat tajam pasca pandemi COVID-19, terutama di negara berkembang seperti Indonesia (Demirgüç-Kunt et al., 2022).

Selain itu, akses terhadap internet dan perangkat seluler juga terbukti ikut mempengaruhi pola perilaku keuangan digital masyarakat. Karena itu, berbagai variabel yang tersedia dalam Global Findex dinilai sangat cocok dipakai untuk menggali lebih dalam mengenai perilaku keuangan digital serta tingkat inklusi keuangan pada tataran individu. Edukasi literasi keuangan juga berperan penting dalam mendorong inklusi keuangan. Inklusi keuangan adalah upaya untuk memastikan bahwa semua lapisan masyarakat memiliki akses yang memadai terhadap produk dan layanan keuangan. Dengan literasi keuangan yang baik, masyarakat dapat lebih mudah memahami dan memanfaatkan berbagai produk keuangan yang tersedia, seperti rekening tabungan, kredit mikro, dan asuransi. Hal ini dapat meningkatkan partisipasi masyarakat dalam sistem keuangan formal dan mengurangi ketergantungan pada lembaga keuangan informal yang mungkin tidak diatur dengan baik (Prihatni et al., 2024).

Karena survei seperti Global Findex mengandalkan jawaban responden, ketidaklengkapan data merupakan persoalan metodologis yang umum. Data hilang, baik yang bersifat sistematis maupun sporadis, lazim dijumpai pada data kategorikal dan survei multilevel (Jolani, 2025). Rubin (1976) menunjukkan bahwa jika data yang hilang tidak dikelola dengan tepat, dapat timbul bias dalam hasil estimasi sekaligus menurunnya validitas analisis statistik, dan tingkat keparahan masalah ini bergantung pada mekanisme yang menyebabkan data hilang. Rubin mengklasifikasikan mekanisme tersebut menjadi tiga

kategori. Pertama, Missing Completely at Random (MCAR), yaitu kejadian data hilang independen dari variabel yang teramati maupun yang tidak teramati didefinisikan pada persamaan (1):

$$P(R|x_{comp}) = P(R) \quad (1)$$

kedua, *Missing at Random* (MAR), di mana kehilangan hanya bergantung pada data yang teramati didefinisikan pada persamaan (2):

$$P(R|x_{comp}) = P(R|x_{obs}) \quad (2)$$

ketiga, *Missing Not at Random* (MNAR), di mana kehilangan bergantung pada nilai data yang tidak teramati maupun nilai yang hilang itu sendiri didefinisikan pada persamaan (3):

$$P(R|x_{comp}) = P(R|x_{obs}, x_{mis}) \quad (3)$$

berdasarkan ketiga mekanisme, sejumlah pendekatan untuk menangani data yang hilang telah diciptakan dengan tingkat validitas yang berbeda-beda. Dalam penelitian ini, diterapkan tiga metode, di mana metode yang paling sederhana adalah mean/mode imputation, yaitu menggantikan nilai yang hilang dengan rata-rata dari nilai yang telah diamati untuk data numerik didefinisikan pada persamaan (4):

$$\hat{x}_{mis,j} = \hat{x}_{obs,j} = \frac{1}{n_{obs}} \sum_{i \in obs} x_{ij} \quad (4)$$

atau dengan modus untuk data kategorikal:

$$\hat{x}_{mis,j} = (x_{obs,j}) \quad (5)$$

karena seluruh variabel target dalam penelitian ini bersifat kategorikal, hanya mode imputation (Persamaan 5) yang diterapkan; Persamaan (4) disajikan sebagai pembandingan umum untuk data numerik dan tidak digunakan dalam simulasi.

Meskipun penerapannya tergolong sederhana, metode ini hanya valid secara asimtotik dengan syarat MCAR dan cenderung menurunkan varians dari estimasi karena mengabaikan berbagai ketidakpastian akibat proses imputasi (Little dan Rubin, 2019). Untuk mengatasi keterbatasan ini, (Rubin, 1987) memperkenalkan suatu kerangka imputasi ganda yang menyuplai setiap nilai yang hilang dengan beberapa nilai yang mungkin untuk mencerminkan ketidakpastian estimasi secara lebih tepat. Kerangka ini kemudian diperluas melalui pendekatan *Multiple Imputation by Chained Equations* (MICE), yang mengestimasi distribusi kondisional dari setiap variabel berdasarkan variabel lain dalam kumpulan data (van Buuren, 2018).

$$x_j^{mis} \sim P(x_j | x_{-j}^{obs}, x_{-j}^{mis(t)}; \theta_j) \quad (6)$$

diiterasi untuk $t = 1, 2, \dots, T$ hingga konvergen, menghasilkan m dataset imputasi berbeda yang kemudian digabung menggunakan aturan Rubin (1987) didefinisikan pada persamaan (7):

$$\bar{Q} = \frac{1}{m} \sum_{l=1}^m \hat{Q}_l, \quad T = \bar{W} + \left(1 + \frac{1}{m}\right) B \quad (7)$$

dimana $\bar{W}$ adalah rata-rata varians dalam-imputasi dan B adalah varians antar-imputasi (van Buuren, 2018). MICE sangat berguna untuk dataset dengan nilai yang hilang di berbagai variabel. Alih-alih mendefinisikan distribusi gabungan, MICE secara iteratif mengambil sampel dari kepadatan kondisional setiap variabel. Pada penelitian ini, MICE dijalankan dengan Predictive Mean Matching (PMM), yaitu metode yang mengisi nilai hilang dengan nilai teramati dari responden yang prediksinya paling mirip (van Buuren, 2018), melalui paket mice (van Buuren dan Groothuis-Oudshoorn, 2011). Karena evaluasi dilakukan

langsung pada nilai hasil imputasi, bukan pada estimasi parameter, digunakan satu dataset imputasi ($m = 1$) sehingga aturan penggabungan Rubin pada Persamaan (7) tidak diterapkan.

Sebagai alternatif non-parametrik, metode missForest memanfaatkan algoritma random forest untuk memprediksi nilai yang hilang berdasarkan variabel lain secara iteratif, didefinisikan pada persamaan (8) (Stekhoven dan Bühlmann, 2012):

$$x_j^{(t)} = f_{RF}(x_{-j}^{(t-1)}) \quad (8)$$

hingga mencapai kriteria berhenti berdasarkan perubahan estimasi antar iterasi, didefinisikan pada persamaan (9):

$$\Delta = \frac{\sum_j (x_j^{(t)} - x_{-j}^{(t-1)})^2}{\sum_j (x_j^{(t)})^2} \quad (9)$$

Stekhoven dan Bühlmann (2012) melaporkan bahwa missForest unggul terutama pada interaksi kompleks dan hubungan non-linear, serta menangani data campuran. Namun keunggulan RF tidak selalu konsisten karena bergantung pada skenario klasifikasi dan struktur data (Rácz dan Gere, 2025), dan menonjol khususnya pada dataset kecil berdimensi tinggi (Abdullah et al., 2026). Belum jelas apakah keunggulan itu berlaku pada data survei kategorikal timpang dengan hubungan antarvariabel lemah, yang menjadi fokus penelitian ini.

Kinerja imputasi variabel kategorikal umumnya dievaluasi dengan PFC, yaitu proporsi nilai hasil imputasi yang berbeda dari nilai sebenarnya (Thurow et al., 2021), yang juga standar pada evaluasi missForest (Stekhoven dan Bühlmann, 2012). PFC hanya mengukur ketepatan reproduksi nilai asli dan tidak menilai kemiripan distribusi: PFC atau NRMSE rendah tidak selalu berarti selisih distribusi kecil (Thurow et al., 2021). Keterbatasan ini lebih nyata pada distribusi kelas yang sangat timpang, karena metode yang selalu memilih kelas mayoritas dapat memperoleh PFC rendah tanpa merekonstruksi kelas minoritas. Karena itu, PFC di sini ditafsirkan sebagai ukuran kesesuaian keseluruhan, bukan bukti bahwa kelas minoritas terimputasi dengan baik, dan dibandingkan dengan acuan proporsi di luar kelas mayoritas. Penelitian lanjutan disarankan menambahkan metrik terkoreksi peluang seperti Cohen's Kappa (Cohen, 1960) serta evaluasi pelestarian distribusi dan struktur asosiasi sebelum dan sesudah imputasi.

3. METODE PENELITIAN

Penelitian ini merupakan studi simulasi menggunakan R versi 4.5.3 melalui RStudio (R Core Team, 2026) dan dilaporkan mengikuti kerangka ADEMP (Aims, Data-generating mechanisms, Estimands, Methods, Performance measures) (Morris, White dan Crowther, 2019). Tahapannya: (1) penyiapan data dan pembentukan dataset acuan lengkap, (2) simulasi data hilang di bawah MCAR dan MAR, (3) penerapan tiga metode imputasi, dan (4) evaluasi kinerja dengan PFC. Seed awal 123 digunakan agar seluruh proses acak dapat direproduksi.

Data yang digunakan adalah mikrodta The Global Findex Database 2025 (World Bank, 2025) untuk Indonesia, tanpa perbandingan lintas negara, dengan 1068 responden dan 199 variabel awal. Unit observasi adalah individu berusia 15 tahun ke atas dari populasi sipil non-institusi yang dipilih melalui probability sampling; individu di bawah 15 tahun, anggota militer aktif, dan penghuni institusi berada di luar populasi sasaran. Dataset analisis dibentuk dalam tiga tahap: (1) memilih 18 dari 199 variabel yang relevan dengan inklusi keuangan digital, menyesuaikan tipenya (numerik, ordinal, atau biner), dan mengeliminasi variabel bervariasi mendekati nol dengan nearZeroVar() (Kuhn, 2008); (2) memeriksa kualitas data,

termasuk proporsi data hilang tiap variabel; (3) mempertahankan responden yang lengkap pada seluruh variabel (Bagian 3.2).

Variabel respons adalah variabel target imputasi, yaitu Fin3, Con12, dan Con26, yang dipilih karena proporsi data hilangnya terbesar dan relevan dengan inklusi keuangan digital. Variabel prediktor adalah seluruh variabel lain yang lengkap dan menjadi informasi bantu bagi MICE dan random forest, termasuk inc_q dan internet_use yang juga menentukan mekanisme MAR. Kedua istilah ini tidak merujuk pada model regresi. Tabel 1 merangkum variabel yang digunakan.

Tabel 1. Data *Dictionary*

Kode	Variabel (definisi operasional)	Skala	Peran
Fin3	Akses rekening melalui kartu atau ponsel	Ordinal	Target imputasi
Con12	Frekuensi penggunaan ponsel dalam aktivitas sehari-hari	Ordinal	Target imputasi
Con26	Frekuensi penggunaan internet	Ordinal	Target imputasi
inc_q	Kuintil pendapatan rumah tangga	Ordinal	Prediktor & penentu MAR
internet_use	Penggunaan internet oleh responden	Biner	Prediktor & penentu MAR
age, educ	Usia (tahun); tingkat pendidikan	Numerik; ordinal	Prediktor
emp_in, urbanicity	Termasuk angkatan kerja; wilayah tinggal (perdesaan/perkotaan)	Biner	Prediktor
con1	Kepemilikan ponsel untuk panggilan pribadi	Biner	Prediktor
account, account_mob	Memiliki akun pada lembaga keuangan formal; akun uang seluler	Biner	Prediktor
anydigpayment, merchantpay_dig	Melakukan/menerima pembayaran digital; pembayaran digital ke pedagang	Biner	Prediktor
borrowed, saved	Meminjam dari, dan menabung di, bank formal atau lembaga sejenis	Biner	Prediktor
receive_wages, receive_transfers	Cara menerima upah; cara menerima transfer dana	Nominal	Prediktor

Nilai sebenarnya dari data yang hilang pada dataset asli tidak diketahui. Oleh karena itu, evaluasi metode imputasi memerlukan dataset acuan yang lengkap. Penelitian ini membentuk dataset acuan melalui prosedur **listwise deletion** dengan mengeliminasi responden yang memiliki minimal satu nilai hilang pada variabel yang dianalisis. Prosedur tersebut menyisakan 407 dari 1.068 responden (38,1%) yang memiliki data lengkap pada seluruh variabel, termasuk Fin3, Con12, dan Con26. Selanjutnya, penelitian ini menginduksi *missingness* secara buatan pada dataset lengkap tersebut. Prosedur ini memungkinkan nilai sebenarnya dari setiap sel yang dihilangkan tetap diketahui sehingga hasil imputasi dapat dibandingkan secara langsung dengan nilai aslinya (Morris, White, dan Crowther, 2019). Dalam penelitian ini, **listwise deletion** hanya berfungsi sebagai prosedur untuk membentuk dataset acuan dan tidak digunakan sebagai metode penanganan data hilang yang dibandingkan. Oleh karena itu, hasil evaluasi terutama mencerminkan karakteristik responden dengan data lengkap, yang mungkin berbeda dari karakteristik keseluruhan sampel.

Nilai hilang diinduksi pada ketiga variabel target secara bersamaan dengan proporsi nominal $p = 10\%$, 20% , dan 30% di bawah MCAR dan MAR. Setiap kombinasi mekanisme dan proporsi diulang 30 kali (replikasi Monte Carlo), dan ketiga metode diterapkan pada dataset hilang yang sama sehingga perbandingan bersifat berpasangan. Hasilnya adalah 180 dataset simulasi ($2 \text{ mekanisme} \times 3 \text{ tingkat} \times 30 \text{ replikasi}$) dan 18 skenario hasil (2 mekanisme

$\times 3$ tingkat $\times 3$ variabel target). Pada MCAR, untuk setiap variabel target dipilih $[p \times n]$ observasi secara acak tanpa pemulihan, terpisah untuk tiap variabel, sehingga peluang hilang seragam. Dengan $n = 407$, jumlah nilai yang dihilangkan adalah 40, 81, dan 122 observasi (9,83%, 19,90%, dan 29,98%); selisih dari proporsi nominal hanya berasal dari pembulatan ke bawah sehingga tidak diperlukan kalibrasi. Pada MAR, peluang hilang bergantung pada variabel teramati: responden dengan $\text{inc_q} = 1$ (kuintil pendapatan terendah) atau $\text{internet_use} = 0$ memiliki peluang hilang 1,8p, sedangkan lainnya 0,8p, dengan $p = 0,10; 0,20; \text{ atau } 0,30$ (kelompok berisiko tinggi 2,25 kali lebih besar). Peluang diterapkan independen pada setiap observasi dan variabel target melalui bilangan acak seragam, tanpa model regresi dan tanpa kalibrasi, sehingga proporsi missing aktual bergantung pada komposisi kelompok berisiko tinggi, dengan rata-rata 9,31%, 18,60%, dan 27,80%.

Mode imputation mengganti nilai hilang dengan kategori paling sering pada data teramati; tidak ada nilai yang diimputasi dengan rata-rata karena seluruh target kategorik. MICE dijalankan dengan PMM, satu dataset imputasi ($m = 1$), dan 5 iterasi bawaan paket mice (van Buuren dan Groothuis-Oudshoorn, 2011). Random forest dijalankan melalui missForest dengan 100 pohon ($\text{ntree} = 100$) dan kriteria konvergensi bawaan, yaitu iterasi berhenti ketika perubahan hasil imputasi antar iterasi mulai meningkat (Stekhoven dan Bühlmann, 2012). Seluruh variabel lain digunakan sebagai prediktor pada MICE dan missForest.

Kinerja metode imputasi dievaluasi hanya pada sel yang sengaja dibuat hilang dalam proses simulasi. Akurasi imputasi untuk variabel kategorik diukur menggunakan *Proportion of False Classification* (PFC), yang didefinisikan sebagai berikut:

$$\text{PFC} = \frac{1}{M} \sum_{i=1}^M I(y_i \neq \hat{y}_i),$$

dengan M menyatakan jumlah sel yang sengaja dihilangkan, y_i merupakan nilai sebenarnya pada sel ke- i , $\hat{y}_i$ merupakan nilai hasil imputasi, dan $I(\cdot)$ adalah fungsi indikator yang bernilai 1 apabila nilai hasil imputasi berbeda dari nilai sebenarnya dan bernilai 0 apabila keduanya sama. Nilai PFC dihitung secara terpisah untuk setiap variabel target pada setiap replikasi simulasi. Selanjutnya, kinerja pada setiap skenario dirangkum menggunakan rata-rata dan simpangan baku (SD) dari nilai PFC yang diperoleh selama 30 replikasi. Dengan demikian, nilai rata-rata yang dilaporkan merupakan rata-rata PFC antarreplikasi, bukan rata-rata dari kode atau nilai kategori variabel. Pendekatan ini sesuai untuk mengevaluasi ketepatan imputasi pada variabel kategorik, dengan nilai PFC yang lebih kecil menunjukkan kinerja imputasi yang lebih baik.

4. HASIL DAN PEMBAHASAN

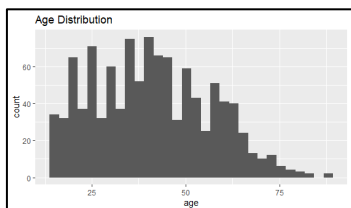
Penelitian ini melakukan diagnostik awal untuk mengidentifikasi tingkat ketidaklengkapan data pada setiap variabel berdasarkan proporsi nilai yang hilang. Tabel 2 menyajikan hasil diagnostik tersebut:

Tabel 2. Diagnostik Missing Data

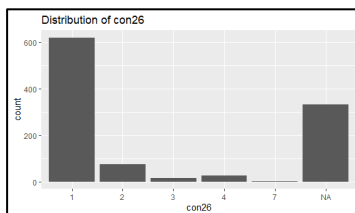
Variabel	Fin3	Con26	Con12	Educ
Missing	52,81%	31,09%	20,32%	0,09%

berdasarkan Tabel 2, proporsi data hilang tertinggi terdapat pada Fin3 (52,81%), diikuti Con26 (31,09%) dan Con12 (20,32%). Proporsi sebesar ini berpotensi menimbulkan bias bila observasi bermissing dihapus langsung (listwise deletion) dan mengurangi ukuran sampel secara signifikan, sehingga teknik imputasi diperlukan agar informasi tetap

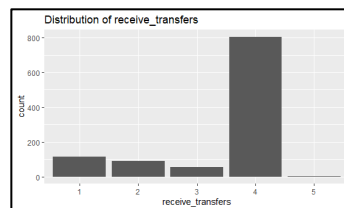
dipertahankan. Ketiganya dipilih sebagai target imputasi karena proporsi data hilangnya terbesar dan relevan dengan indikator inklusi keuangan digital.



Gambar 1. Distribusi usia

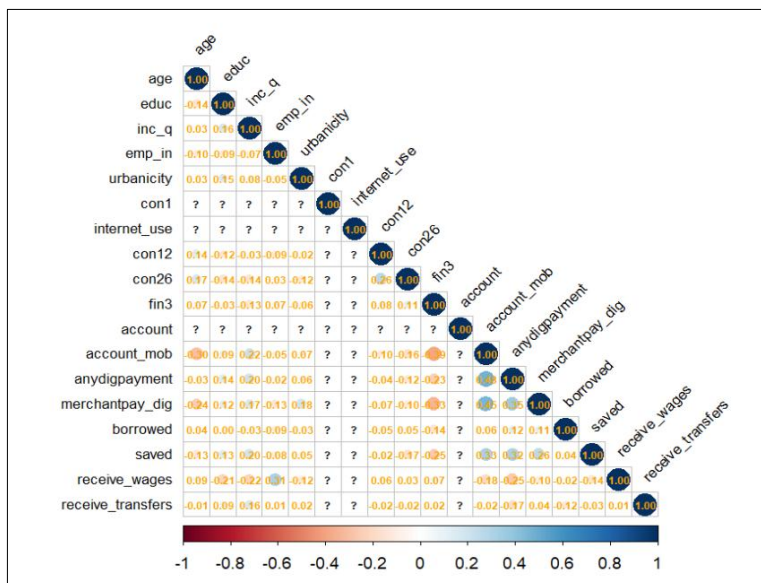


Gambar 2. Distribusi frekuensi penggunaan internet



Gambar 3. Distribusi penerimaan transfer

Gambar 1 menunjukkan distribusi usia yang cenderung miring ke kanan (right-skewed), dengan kelompok usia 30–50 tahun terbanyak dan frekuensi menurun di atas 60 tahun, sehingga mayoritas responden berada pada usia aktif secara ekonomi. Gambar 2 memperlihatkan distribusi penggunaan internet (Con26) yang sangat timpang dan didominasi kategori 1; beberapa kategori sangat sedikit sehingga berpotensi sparsity. Bersama proporsi data hilang 31,09% (Tabel 2), hal ini menjadikan Con26 menantang untuk diimputasi. Gambar 3 menunjukkan receive_transfers terdiri atas lima kategori dengan kategori 4 dominan, sedangkan kategori 1, 2, 3, dan 5 jauh lebih sedikit, yang mengindikasikan class imbalance. Ketimpangan ini penting karena pada data yang didominasi satu kategori, metode berbasis modus berpotensi mereproduksi nilai asli dengan baik.



Gambar 4. Korelasi antar variabel

Gambar 4 menyajikan matriks korelasi antarvariabel dengan variabel kategorik dikodekan numerik. Pengguna internet lebih mungkin memiliki akun di lembaga keuangan formal, sedangkan kelompok berpendapatan lebih tinggi cenderung lebih banyak memanfaatkan pembayaran digital. Uji Cramér's V antara kepemilikan akun (account) dan penggunaan internet (internet_use) menguatkan temuan ini dengan nilai 0,3174, yang menunjukkan asosiasi moderat.

Tabel 3 membandingkan mode imputation, random forest (missForest), dan MICE pada Fin3, Con12, dan Con26 dengan tingkat data hilang 10%, 20%, dan 30% di bawah MCAR dan MAR. Setiap sel memuat rata-rata PFC dan SD dari $R = 30$ replikasi Monte Carlo;

karena ketiga metode diterapkan pada dataset hilang yang sama, perbandingan bersifat berpasangan.

Tabel 3. Evaluasi Metode Imputasi

Variabel	Missing	Mode Imputation		Random Forest		MICE	
		MCAR (SD)	MAR (SD)	MCAR (SD)	MAR (SD)	MCAR (SD)	MAR (SD)
Fin3	10%	0.296 (0.061)	0.311 (0.086)	0.325 (0.056)	0.331 (0.093)	0.343 (0.060)	0.344 (0.097)
	20%	0.309 (0.050)	0.314 (0.038)	0.314 (0.042)	0.341 (0.042)	0.353 (0.060)	0.365 (0.053)
	30%	0.304 (0.050)	0.324 (0.038)	0.320 (0.042)	0.328 (0.042)	0.354 (0.060)	0.356 (0.053)
Con12	10%	0.025 (0.029)	0.038 (0.028)	0.044 (0.029)	0.036 (0.028)	0.046 (0.034)	0.056 (0.036)
	20%	0.025 (0.017)	0.034 (0.018)	0.031 (0.015)	0.039 (0.019)	0.054 (0.023)	0.060 (0.031)
	30%	0.027 (0.012)	0.030 (0.013)	0.032 (0.012)	0.035 (0.014)	0.059 (0.022)	0.064 (0.022)
Con26	10%	0.105 (0.054)	0.140 (0.053)	0.149 (0.055)	0.187 (0.056)	0.208 (0.068)	0.222 (0.063)
	20%	0.116 (0.028)	0.134 (0.036)	0.189 (0.029)	0.158 (0.035)	0.216 (0.050)	0.242 (0.040)
	30%	0.116 (0.022)	0.134 (0.024)	0.159 (0.024)	0.189 (0.024)	0.216 (0.030)	0.208 (0.039)

Kenaikan persentase data hilang tidak diikuti kenaikan PFC yang monoton pada semua metode. Kenaikan berurutan 10%–20%–30% hanya tampak pada empat kombinasi: MICE pada Con12 (0.046, 0.054, 0.059 pada MCAR dan 0.056, 0.060, 0.064 pada MAR), MICE pada Fin3 MCAR (0.343, 0.353, 0.354), dan Mode pada Fin3 MAR (0.311, 0.314, 0.324). Pada kombinasi lain PFC relatif stabil atau berfluktuasi sempit, misalnya Mode pada Fin3 MCAR (0.296, 0.309, 0.304), dan sebagian bahkan menurun, seperti Mode pada Con12 MAR (0.038, 0.034, 0.030) dan Random Forest pada Con12 MCAR (0.044, 0.031, 0.032). Fluktuasi sekecil ini perlu dibaca bersama SD antarreplikasi, yang pada Mode MCAR 10% mencapai 0.060 untuk Fin3 dan 0.054 untuk Con26, sehingga belum tentu mencerminkan perbedaan kinerja nyata. Perbedaan PFC antarvariabel jauh lebih besar daripada antartingkat missing: Fin3 tertinggi (0.296–0.365), disusul Con26 (0.105–0.242), dan Con12 terendah (0.0251–0.0640) pada semua metode. Variabel yang diimputasi lebih menentukan besarnya PFC daripada banyaknya data yang dihilangkan pada rentang 10–30%.

Pada Fin3, Mode terendah di keenam skenario (0.296–0.324), diikuti Random Forest (0.314–0.341) dan MICE (0.343–0.365). Selisih Mode dan Random Forest kecil: 0.005 pada MCAR 20% dan 0.004 pada MAR 30%, dan 0.016–0.029 pada skenario lain; pada setiap skenario selisih itu lebih kecil daripada SD (0.027–0.093), sehingga keunggulan Mode atas Random Forest perlu ditafsirkan hati-hati. Selisih Mode dan MICE lebih besar (0.032–0.051) dan melampaui SD Mode pada MCAR 30% (0.050 dengan SD 0.028) dan MAR 20% (0.051 dengan SD 0.038), tetapi tidak pada skenario lain. Pada Con12, Mode terendah pada lima dari enam skenario. Satu-satunya pengecualian adalah MAR 10%, ketika Random Forest (0.0367) sedikit lebih rendah daripada Mode (0.0382); selisih 0.0015 itu jauh di bawah SD (0.028). Pada skenario lain selisih Mode dan Random Forest hanya 0.005–0.019 dan juga lebih kecil daripada SD (0.012–0.029). MICE tertinggi pada semua skenario; pada 20% dan 30%, PFC MICE berkisar 1,7 hingga 2,2 kali PFC Mode (misalnya 0.0590 dibanding 0.0270 pada MCAR 30%), dengan selisih umumnya melampaui SD.

Pada Con26, Mode terendah di keenam skenario (0.105–0.140), dengan Random Forest 0.149–0.189 dan MICE 0.208–0.242. Selisih Mode dan Random Forest di sini paling besar di antara ketiga variabel (0.024–0.073) dan melampaui SD pada MCAR 20% (0.073 dengan SD 0.028 dan 0.029), MCAR 30% (0.043 dengan SD 0.022 dan 0.024), dan MAR 30% (0.055 dengan SD 0.024), tetapi tidak pada MCAR dan MAR 10% serta MAR 20%. Dengan demikian keunggulan Mode atas Random Forest paling jelas pada Con26. Secara keseluruhan, Mode memiliki PFC terendah pada 17 dari 18 kombinasi variabel, tingkat missing, dan mekanisme. Random Forest kedua dan MICE terakhir pada 17 kombinasi

tersebut; pada Con12 MAR 10% urutannya Random Forest, Mode, lalu MICE. MICE yang selalu terakhir kemungkinan berkaitan dengan konfigurasi PMM satu dataset imputasi ($m = 1$) yang kurang sesuai untuk data kategorik timpang, sedangkan Random Forest yang selalu lebih baik daripada MICE kemungkinan karena kemampuannya menangkap hubungan non-linear, tetapi belum cukup untuk mengungguli Mode.

Pada imputasi modus, galat pada dasarnya sama dengan proporsi nilai di luar kategori terbanyak, sehingga PFC Mode memberi gambaran tidak langsung tentang ketimpangan kelas. PFC Mode pada MCAR sekitar 0.025–0.027 untuk Con12, 0.105–0.116 untuk Con26, dan 0.296–0.309 untuk Fin3, yang mengisyaratkan kategori terbanyak menyusun kira-kira 97% data Con12, 88–90% Con26, dan 69–70% Fin3. Angka ini adalah perkiraan yang diturunkan dari Tabel 3, bukan proporsi yang dihitung langsung dari data. Urutannya sejalan dengan Gambar 2 dan menjelaskan mengapa PFC Con12 paling kecil dan Fin3 paling besar pada semua metode. Perbedaan PFC antarvariabel dengan demikian lebih mencerminkan komposisi kelas daripada kompleksitas struktur data, dan sejalan dengan PFC Mode yang hampir tidak berubah antartingkat missing pada Con12 dan Con26. PFC rendah pada kondisi ini tidak berarti kelas minoritas terimputasi dengan baik.

PFC pada MAR lebih tinggi daripada MCAR pada 24 dari 27 kombinasi metode, variabel, dan tingkat missing; pada Mode, MAR lebih tinggi di seluruh sembilan kombinasi. Tiga pengecualiannya adalah Random Forest pada Con12 10% dan Con26 20%, serta MICE pada Con26 30%. Selisih MAR dikurangi MCAR umumnya kecil: 0.001–0.027 pada Fin3, –0.008 hingga 0.013 pada Con12, dan –0.031 hingga 0.038 pada Con26. MAR karenanya cenderung, tetapi tidak selalu, menghasilkan kesalahan lebih besar daripada MCAR. Proporsi missing aktual pada MAR (9,31%, 18,60%, dan 27,80%) sedikit lebih rendah daripada MCAR (9,83%, 19,90%, dan 29,98%), sehingga PFC yang lebih tinggi pada MAR tidak berasal dari jumlah data hilang yang lebih banyak. Urutan antarmetode juga hampir tidak berubah antarmekanisme; satu-satunya perubahan adalah Con12 MAR 10%.

Hasil simulasi menunjukkan tiga hal. Pertama, jenis variabel dan komposisi kelasnya memengaruhi PFC jauh lebih besar daripada tingkat data hilang 10–30%. Kedua, Mode terendah pada 17 dari 18 kombinasi, Random Forest umumnya kedua, dan MICE terakhir; keunggulan Mode atas Random Forest tergolong kecil dibanding SD pada Fin3 dan Con12, dan paling jelas pada Con26. Ketiga, MAR cenderung menghasilkan PFC sedikit lebih tinggi daripada MCAR tanpa mengubah urutan metode. Seluruh temuan ini berlaku untuk metrik PFC dan skenario yang diuji.

5. KESIMPULAN

Penelitian ini membandingkan mode imputation, MICE, dan random forest (missForest) pada tiga variabel kategorik Global Findex Indonesia 2025 (Fin3, Con12, dan Con26) di bawah MCAR dan MAR dengan tingkat data hilang 10%, 20%, dan 30%. Berdasarkan PFC, mode imputation terendah pada 17 dari 18 kombinasi variabel, tingkat missing, dan mekanisme; satu-satunya pengecualian adalah Con12 pada MAR 10%, ketika random forest sedikit lebih rendah. Random forest umumnya kedua dan MICE terakhir. Selisih mode imputation dan random forest lebih kecil daripada SD pada Fin3 dan Con12, sehingga keunggulan mode paling jelas pada Con26. Besarnya PFC lebih ditentukan oleh variabel yang diimputasi daripada tingkat data hilang, dan MAR cenderung menghasilkan PFC sedikit lebih tinggi daripada MCAR tanpa mengubah urutan metode. Hasil ini sejalan dengan data yang kategorik, timpang, dan berhubungan lemah antarvariabel, sehingga metode sederhana berbasis distribusi dapat bersaing dengan metode berbasis model. Bagi praktisi survei, mode imputation dapat menjadi acuan dasar yang layak sebelum metode yang lebih

rumit, sepanjang tujuannya merekonstruksi nilai individual dan bukan menjaga distribusi atau hubungan antarvariabel.

Kesimpulan ini hanya berlaku pada skenario, variabel, dan metrik PFC yang diuji, dan tidak dapat digeneralisasi sebagai “mode imputation merupakan metode terbaik”. PFC yang rendah pada data timpang dapat mencerminkan dominasi kelas mayoritas, bukan keberhasilan merekonstruksi kelas minoritas. Keterbatasan lain: MNAR belum dianalisis; PFC tidak mengoreksi kesesuaian karena peluang dan tidak menilai pelestarian distribusi maupun hubungan antarvariabel (Thurrow et al., 2021); dataset acuan hanya 407 responden lengkap (38,1%) sehingga rentan bias seleksi; dan cakupan terbatas pada satu negara, konfigurasi metode tetap, dan satu rancangan MAR. Penelitian selanjutnya disarankan menambahkan skenario MNAR, Cohen’s Kappa (Cohen, 1960), serta evaluasi distribusi marginal dan struktur asosiasi, misalnya Cramér’s V sebelum dan sesudah imputasi.

DAFTAR PUSTAKA

- Abdullah, N.A. et al. (2026) “Assessing imputation techniques for missing data in small and multicollinear datasets: insights from craniofacial morphometry,” *BMC Medical Research Methodology*, 26(1), hal. 54. Tersedia pada: <https://doi.org/10.1186/s12874-025-02762-4>.
- Allison, P.D. (2001) *Missing Data*. Thousand Oaks, CA: Sage Publications.
- van Buuren, S. (2018) *Flexible Imputation of Missing Data*. 2 ed. Boca Raton, FL: Chapman & Hall/CRC.
- van Buuren, S. dan Groothuis-Oudshoorn, K. (2011) “mice: Multivariate imputation by chained equations in R,” *Journal of Statistical Software*, 45(3), hal. 1–67. Tersedia pada: <https://doi.org/10.18637/jss.v045.i03>.
- Cohen, J. (1960) “A coefficient of agreement for nominal scales,” *Educational and Psychological Measurement*, 20(1), hal. 37–46. Tersedia pada: <https://doi.org/10.1177/001316446002000104>.
- Demirgüç-Kunt, A. et al. (2022) *The Global Findex Database 2021: Financial Inclusion, Digital Payments, and Resilience in the Age of COVID-19*. Washington, DC: World Bank. Tersedia pada: <https://doi.org/10.1596/978-1-4648-1897-4>.
- Farah, A.A., Nor, B.A. dan Musse, O.S.H. (2026) “Financial inclusion and green economic growth a systematic review of SCOPUS indexed studies from 2015 to 2025,” *Discover Environment*, 4, hal. 37. Tersedia pada: <https://doi.org/10.1007/s44274-026-00550-5>.
- Jolani, S. (2025) “Hierarchical imputation of categorical variables in the presence of systematically and sporadically missing data,” *Research Synthesis Methods*, 16, hal. 729–757. Tersedia pada: <https://doi.org/10.1017/rsm.2025.10017>.
- Khan, N. et al. (2024) “Financial inclusion as a catalyst for economic growth: evidence from selected developing countries,” *Cogent Social Sciences*, 10(1), hal. 2387907. Tersedia pada: <https://doi.org/10.1080/23311886.2024.2387907>.

- Kuhn, M. (2008) "Building predictive models in R using the caret package," *Journal of Statistical Software*, 28(5), hal. 1–26. Tersedia pada: <https://doi.org/10.18637/jss.v028.i05>.
- Li, J. et al. (2024) "Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets," *BMC Medical Research Methodology*, 24, hal. 41. Tersedia pada: <https://doi.org/10.1186/s12874-024-02173-x>.
- Little, R.J.A. dan Rubin, D.B. (2019) *Statistical Analysis with Missing Data*. 3 ed. Hoboken, NJ: Wiley.
- Morris, T.P., White, I.R. dan Crowther, M.J. (2019) "Using simulation studies to evaluate statistical methods," *Statistics in Medicine*, 38(11), hal. 2074–2102. Tersedia pada: <https://doi.org/10.1002/sim.8086>.
- Prihatni, R. et al. (2024) *Analisis Literasi Keuangan dan Inklusi Keuangan di Indonesia: Strategi untuk Meningkatkan Kesejahteraan Finansial Masyarakat*. 1 ed. Diedit oleh N. Mayasar. Bandung: Penerbit Widina Media Utama.
- R Core Team (2026) *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Tersedia pada: <https://www.R-project.org/>.
- Rubin, D.B. (1976) "Inference and missing data," *Biometrika*, 63(3), hal. 581–592. Tersedia pada: <https://doi.org/10.1093/biomet/63.3.581>.
- Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
- Rácz, A. dan Gere, A. (2025) "Comparison of missing value imputation tools for machine learning models based on product development cases studies," *LWT - Food Science and Technology*, 221, hal. 117585. Tersedia pada: <https://doi.org/10.1016/j.lwt.2025.117585>.
- Sarma, M. (2008) *Index of Financial Inclusion*. New Delhi. Tersedia pada: <https://icrier.org/publications/index-of-financial-inclusion/>.
- Stekhoven, D.J. dan Bühlmann, P. (2012) "MissForest — non-parametric missing value imputation for mixed-type data," *Bioinformatics*, 28(1), hal. 112–118. Tersedia pada: <https://doi.org/10.1093/bioinformatics/btr597>.
- Thurow, M., Dumpert, F., Ramosaj, B. dan Pauly, M. (2021) "Goodness (of fit) of imputation accuracy: the GoodImpact analysis," *arXiv preprint arXiv:2101.07532*. Tersedia pada: <https://doi.org/10.48550/arXiv.2101.07532>.
- Wabina, R.S. et al. (2025) "Uncertainty-aware approach for multiple imputation using conventional and machine learning models: a real-world data study," *Journal of Big Data*, 12. Tersedia pada: <https://doi.org/10.1186/s40537-025-01136-3>.
- World Bank (2025) *The Global Findex Database 2025: Connectivity and Financial Inclusion in the Digital Economy*, World Bank. Tersedia pada: <https://www.worldbank.org/en/publication/globalfindex>.