Vol. 4, No. 2, December, 2025, pp. 65-73
e-ISSN: 2987-906X
https://ejournal.unib.ac.id/diophantine

Pananganan Data Hilang pada Data Bangkitan Bivariate Gamma

Muhammad A. Alwansyah[1*], Khaola R. Adzima[2], dan Muhammad R. Wujudi[3]

[1,2] Program Studi S1 Pendidikan Matematika, Universitas Negeri Jakarta, Indonesia
[3] Program Studi S1 Matematika, Universitas Negeri Jakarta, Indonesia

* Corresponding Author Email: muhammadarib@unj.ac.id

Article Information

Article History:
Submitted: 10 December 2025
Accepted: 19 December 2025
Published: 31 December 2025

Key Words:
Bivariate Gamma
Random Forest Imputations
Classification and Regression Trees
Root Mean Square Error
Mean Absolute Percentage Error
Correlation
T Test

DOI:
https://doi.org/10.33369/diophantine.v4i2.46691

Abstract

Missing data is a problem in data processing that can reduce the quality of analysis results if not addressed. This study aims to evaluate the performance of two imputation methods, namely Random Forest Imputation (RF) and Classification and Regression Tree (CART), at various levels of missing value proportions, namely 5%, 10%, 15%, and 20%. The data used in this study are Bivariate Gamma data of 200 observations with two variables, which were generated using RStudio software. The evaluation was carried out based on the correlation value between the imputed data and the original data, as well as the error measures Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE). The results showed that at the missing value levels of 5% and 10%, the CART method produced the smallest MAPE and RMSE values, so that the CART method was the best method, although there was no significant difference between the RF method and the 10% missing value data. At 15% and 20% missing values, the RF method demonstrated superior performance with smaller MAPE and RMSE values compared to CART. Overall, the CART method is more suitable for use with a low proportion of missing values, while the RF method provides more stable performance at a high proportion of missing values. The results of this study provide recommendations for selecting a more appropriate imputation method based on the level of missing data.

1. PENDAHULUAN

Data hilang atau juga sering disebut dengan *missing data* merupakan salah satu permasalahan utama dalam penelitian karena dapat menurunkan kualitas analisis, mengurangi efisiensi estimasi, serta berpotensi menghasilkan inferensi yang bias. Keberadaan data hilang sering kali tidak dapat dihindari, baik dalam penelitian empiris maupun dalam studi simulasi data bangkitan. Dalam distribusi kontinu positif, salah satu model yang secara luas digunakan untuk mempelajari hubungan dua variabel yang saling berkorelasi adalah distribusi Bivariate Gamma. Distribusi Bivariate Gamma banyak diaplikasikan dalam berbagai bidang seperti *reliability engineering*, analisis risiko, hidrologi, aktuaria, dan epidemiologi, karena sifatnya yang fleksibel dalam merepresentasikan hubungan dua variabel acak yang bersifat *skewed* dan *strictly positive* [1]. Akan tetapi, ketika sebagian data Bivariate Gamma mengalami data hilang, maka analisis lanjut seperti estimasi parameter, pengujian hipotesis, dan pemodelan dependensi menjadi terhambat serta rentan terhadap bias.

Metode imputasi sederhana seperti *mean imputation* atau *single regression imputation* umumnya masih sering digunakan, tetapi metode tersebut memberikan estimasi kasar, mengabaikan variabilitas alami data, dan tidak mampu menangkap struktur hubungan non-linear antar variabel. Seiring perkembangan *machine learning*, teknik imputasi modern dikembangkan untuk menghasilkan estimasi nilai hilang yang lebih akurat dan konsisten. Di antara metode yang banyak digunakan dan terbukti efektif adalah *Random Forest Imputation* (RF) dan *Classification and Regression Trees* (CART) [2].

Metode *Random Forest Imputation* (RF) merupakan metode *ensemble learning* yang membangun ratusan hingga ribuan pohon keputusan secara acak dan menggabungkan prediksinya melalui mekanisme agregasi. Karakteristik *bagging* dan pemilihan fitur secara acak pada setiap pohon memungkinkan *Random Forest* menangkap pola kompleks, hubungan non-linear, serta interaksi antarvariabel yang sulit ditangani oleh

model parametrik tradisional [3]. Beberapa penelitian menunjukkan bahwa RF *Imputation* sangat *robust* terhadap *outlier*, dapat menangani *mixed-type data*, serta memiliki performa superior pada distribusi yang memiliki skewness tinggi[4]. Sedangkan metode *Classification and Regression Trees* (CART) berfungsi sebagai model pohon keputusan tunggal yang membagi data menjadi node-node homogen berdasarkan aturan pemisahan (*splitting rules*) tertentu. CART memiliki keunggulan berupa interpretasi yang sederhana, performa yang baik pada data non-normal, dan kemampuan menangani data kategorik maupun kontinu tanpa transformasi. Pada proses imputasi, CART memprediksi nilai hilang dengan mempelajari struktur data yang tersedia dan mengisi nilai dengan hasil prediksi dari pohon regresi atau klasifikasi [5]. Meskipun CART tidak sekompleks RF, metode ini tetap menjadi pilihan yang efektif karena mudah diimplementasikan, cepat, dan mampu mengatasi berbagai pola data hilang.

Imputasi pada data bangkitan Bivariate Gamma menggunakan metode RF dan CART memiliki nilai penting karena keduanya mampu menjawab keterbatasan metode parametrik dalam menghadapi ketidaksimetrisan data dan hubungan kompleks antardimensi. Penggunaan dua metode ini juga memungkinkan peneliti melakukan evaluasi dalam menentukan metode mana yang mampu menghasilkan estimasi nilai hilang terbaik terhadap distribusi sebenarnya. Dengan demikian, hasil penelitian ini diharapkan dapat berkontribusi dalam imputasi yang efektif untuk data hilang.

1.1 Distribusi Bivariat Gamma

Distribusi Bivariat Gamma merupakan perluasan dari distribusi univariat Gamma yang digunakan untuk memodelkan dua peubah acak kontinu yang saling berkorelasi dan memiliki sifat *positively skewed*. Distribusi ini banyak digunakan dalam bidang asuransi, keandalan, analisis survival, serta hidrologi, ketika dua variabel respon saling berkaitan namun memiliki sifat sebaran kanan [6]. Misalkan (X, Y) adalah pasangan peubah acak yang mengikuti Bivariat Gamma, maka dependensinya dibangun melalui *shared component model*, yaitu:

$$X = U + V, \qquad Y = W + V$$

Diketahui $U, W, V \sim \text{Gamma}(\alpha, \lambda)$ independen satu sama lain, oleh karena itu komponen V menjadi penyebab korelasi positif antara X dan Y [7].

1.2 Karakteristik Distribusi Bivariat Gamma

Karakteristik distribusi Bivariat Gamma seperti nilai harapan, varian, kovarian dan korelasi. Jika struktur komponen bersama digunakan, maka nilai harapan bivariat gamma adalah sebagai berikut:

$$E(X) = \frac{\alpha_U}{\lambda} + \frac{\alpha_V}{\lambda}, E(Y) = \frac{\alpha_W}{\lambda} + \frac{\alpha_V}{\lambda} \tag{1}$$

Varian dari bivariat gamma adalah sebagai berikut:

$$\text{Var}(X) = \frac{\alpha_U + \alpha_V}{\lambda^2}, \text{Var}(Y) = \frac{\alpha_W + \alpha_V}{\lambda^2} \tag{2}$$

Kovarian dan Korelasi dari bivariat gamma adalah sebagai berikut:

$$\text{Cov}(X, Y) = \frac{\alpha_V}{\lambda^2} \tag{3}$$

Sehingga diperoleh korelasinya adalah sebagai berikut:

$$\rho = \frac{\alpha_V}{\sqrt{(\alpha_U + \alpha_V)(\alpha_W + \alpha_V)}} \tag{4}$$

Korelasi selalu positif karena bergantung pada komponen *shared gamma* (V) [8].

1.3 Uji Beda Rata-rata

Uji t digunakan untuk mengevaluasi apakah terdapat perbedaan yang signifikan antara dua rata-rata populasi berdasarkan data sampel. Secara khusus, *independent sample t-test* digunakan ketika dua kelompok

sampel bersifat saling bebas. Uji ini membandingkan estimasi rata-rata kedua kelompok sambil memperhitungkan variabilitas data di dalam masing-masing sampel [9].
Berikut merupakan persamaan Statistik Uji t:

$$t = \frac{\bar{X} - \mu}{\left(\frac{SD}{\sqrt{n}}\right)} \tag{5}$$

Pengujian dilakukan dengan membandingkan nilai statistik t terhadap tabel t pada derajat kebebasan tertentu dan taraf signifikansi yang telah ditetapkan. Jika nilai $p\text{-}value$ lebih kecil dari α atau nilai statistik uji t lebih besar dari t tabel, maka hipotesis nol yang menyatakan bahwa kedua rata-rata populasi adalah sama ditolak [10].

Berikut Hipotesis dari uji beda rata-rata satu sampel:

1. Pengujian Hipotesis
   $H_0 : \mu = \mu_0$ (Tidak berbeda secara signifikan)
   $H_1 : \mu \neq \mu_0$ (Berbeda secara signifikan)
2. Besaran yang diperlukan
   Taraf Signifikansi $\alpha = 5\%$
3. Statistik uji
   $t = \frac{\bar{X} - \mu}{\left(\frac{SD}{\sqrt{n}}\right)}$
4. Daerah Penolakan
   Tolak H_0 jika nilai $t_{hit} > t_{\frac{\alpha}{2},n-1}$ atau $p - value < \alpha = 0.05$
5. Kesimpulan

1.4 Metode Imputasi

1.4.1 *Random Forest Imputations* (RF)

Random Forest Imputation merupakan pendekatan non-parametrik untuk menggantikan nilai-nilai hilang (*missing values*) dalam himpunan data dengan memanfaatkan kekuatan *ensemble* dari pohon keputusan. Metode ini memodelkan setiap variabel yang mengandung nilai hilang sebagai fungsi dari variabel-variabel lain yang teramati, lalu memprediksi observasi yang hilang menggunakan prediktor yang dibangun dari *random forests*. Algoritma *missForest* merupakan algoritma iteratif yang dimulai dengan inisialisasi nilai-nilai hilang, kemudian untuk setiap variabel dengan *missingness* dilakukan pelatihan *random forest* menggunakan observasi lengkap pada variabel tersebut sebagai respons dan variabel lain sebagai prediktor, selanjutnya nilai-nilai hilang diprediksi dan digantikan, proses ini diulang berulang kali sampai kriteria konvergensi terpenuhi. Metode ini bersifat fleksibel untuk data campuran (kontinu dan kategorik) dan mampu menangkap pola non-linear antar variabel tanpa asumsi distribusi parametrik tertentu [11].
Langkah-langkah algoritma *missForest* adalah sebagai berikut:

1. Tentukan urutan variabel berdasarkan proporsi *missing*
2. Untuk setiap variabel X_j yang memiliki *missing*, bentuk dataset pelatihan $\{(X_{-j}^{(i)}, X_j^{(i)}) : X_j^{(i)}$ teramati$\}$ dan *fit* model *random forest* $\hat{f}_j$ untuk memprediksi X_j dari variabel lain X_{-j}.
3. Gunakan $\hat{f}_j$ untuk mengisi nilai-nilai X_j yang hilang.
4. Ulangi langkah 2–3 untuk seluruh variabel, dan ulangi siklus ini hingga kriteria konvergensi tercapai.

Kelebihan praktis dari pendekatan ini adalah ketersediaan estimasi kesalahan imputasi bawaan melalui mekanisme *out-of-bag* (OOB) pada *random forest*, sehingga peneliti dapat memperoleh taksiran RMSE (*root mean squared error*) untuk variabel kontinu dan PFC (*proportion of falsely classified*) untuk variabel kategorik tanpa set validasi terpisah [12]. *Random Forest imputation* menunjukkan performa unggul

dibandingkan metode parametrik sederhana lainnya dengan kondisi di mana hubungan antarvariabel bersifat non-linear. Studi komparatif pada data observasional dan simulasi melaporkan bahwa *missForest* cenderung menghasilkan RMSE lebih rendah serta akurasi klasifikasi yang lebih tinggi dalam berbagai skenario *missingness* (MCAR/MAR), meskipun biaya komputasi dan waktu eksekusi biasanya lebih besar dibandingkan metode sederhana [13]. Dalam proses imputasi, *Random Forest* membangun sejumlah pohon keputusan menggunakan *bootstrap sampling*, kemudian menghasilkan prediksi berdasarkan agregasi model, yaitu *mode* untuk variabel kategori dan *average* untuk variabel numerik [14]. Secara matematis, suatu *forest* terdiri atas B pohon regresi atau klasifikasi yang dibangun dari sampel *bootstrap* $\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_B$. Untuk prediksi regresi, imputasi nilai hilang $\hat{y}$ diekspresikan sebagai rata-rata prediksi seluruh pohon [15]:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{6}$$

Diketahui $T_b(x)$ adalah pohon ke-b untuk observasi x.

1.4.2 *Classification and Regression Trees* (CART)

Metode *Classification and Regression Trees* (CART) memanfaatkan struktur pemisahan data yang dipelajari oleh pohon keputusan untuk memprediksi nilai yang tidak terobservasi. Dalam algoritma CART, ruang fitur dibagi menjadi region–region R_m melalui pemilihan *split* yang meminimalkan ukuran ketidakhomogenan, untuk variabel numerik kriteria yang umum digunakan adalah *mean squared error (MSE)* sehingga pemilihan *split* (s, j) pada variabel X_j dapat dituliskan sebagai berikut:

$$\arg min \sum_{x_i \in R_{\text{left}}(s,j)} (y_i - \bar{y}_{\text{left}})^2 + \sum_{x_i \in R_{\text{right}}(s,j)} (y_i - \bar{y}_{\text{right}})^2 \tag{7}$$

Sedangkan untuk variabel kategorik sering digunakan *Gini impurity*:

$$G = \sum_{k=1}^{K} p_k(1 - p_k) \tag{8}$$

Diketahui p_k proporsi kelas ke-k pada *node*. Setelah pohon dibangun dari observasi lengkap, prediksi untuk observasi baru dilakukan dengan menetapkan observasi ke *leaf* R_m yang sesuai dan menggunakan nilai rata-rata atau modus pada *leaf* tersebut sebagai imputasi. Prediksi pohon dapat ditulis sebagai berikut:

$$\hat{f}(x) = \sum_{m=1}^{M} c_m \, \mathbf{1}(x \in R_m) \tag{9}$$

Diketahui c_m= rata-rata respon pada *region* R_m. Prinsip ini memungkinkan CART menghasilkan imputasi yang tidak bergantung pada asumsi distribusi parametris dan menangkap hubungan non-linear serta interaksi antarvariabel [16]. CART memberikan hasil yang kompetitif terutama pada data dengan variabel campuran dan pola non-linear [17]. Keterbatasan CART yaitu rentan terhadap variabilitas model dan *overfitting* pada sampel kecil, dan apabila variabel-variabel penjelas juga banyak yang hilang maka informasi yang tersedia untuk pelatihan menjadi terbatas sehingga imputasi dapat bias [18].

1.5 *Mean Absolute* Percentage *Error* (MAPE)

Mean Absolute Percentage Error (MAPE) merupakan ukuran kesalahan peramalan yang populer dan mudah diinterpretasikan, yang menyatakan rata-rata persentase deviasi absolut antara nilai observasi dan nilai prediksi. Persamaan MAPE didefinisikan sebagai berikut:

$$MAPE = \frac{\sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{\hat{y}_i} \right|}{n} \times 100 \tag{10}$$

Diketahui y_i adalah nilai aktual pada observasi ke-i, $\hat{y}_i$ adalah nilai prediksi, dan n jumlah pengamatan [19]. MAPE memiliki beberapa keterbatasan seperti menjadi tidak terdefinisi atau sangat besar jika ada nilai aktual y_i yang mendekati nol, sehingga menimbulkan bias penilaian model pada seri dengan nilai kecil atau

nol. MAPE juga memberi bobot relatif lebih besar pada kesalahan yang terjadi pada observasi berukuran kecil [20].

1.6 *Root Mean Square Error* (RMSE)

Root Mean Square Error merupakan salah satu ukuran akurasi prediksi yang digunakan dalam evaluasi kinerja model statistik maupun *machine learning*. Nilai RMSE yang lebih kecil mengindikasikan bahwa hasil prediksi model semakin mendekati nilai aktual, serta menunjukkan variasi nilai prediksi sesuai dengan variasi data observasi. RMSE sering digunakan karena lebih kuat terhadap kesalahan prediksi yang besar [21]. Persamaan RMSE dapat dituliskan sebaga berikut:

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i})^2} \tag{11}$$

Diketahui $\mathrm{y_i}$ adalah nilai aktual, $\hat{\mathrm{y}}_\mathrm{i}$ nilai yang diprediksi, dan n jumlah pengamatan. RMSE memungkinkan evaluasi yang konsisten terhadap performa model, terutama dalam konteks regresi, peramalan deret waktu, dan pemodelan prediktif berbasis *machine learning*.

1.7 Korelasi

Model Analisis korelasi merupakan metode statistik yang digunakan untuk mengukur kekuatan dan arah hubungan linier antara dua variabel. Koefisien korelasi memberikan nilai numerik antara −1 sampai +1. Nilai mendekati 1 menunjukkan hubungan positif yang sangat kuat, mendekati −1 menunjukkan hubungan negatif yang sangat kuat, sedangkan nilai mendekati 0 menunjukkan tidak ada atau sangat lemah hubungan linear. Rumus korelasi dapat dituliskan sebagai berikut:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\ \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{12}$$

Diketahui $x_i$dan $y_i$adalah nilai observasi ke-i, $\bar{x}$ dan $\bar{y}$ adalah rata-rata sampel X dan Y, dan n adalah jumlah observasi. Nilai r menunjukkan derajat kekuatan dan arah asosiasi linear antara X dan Y [22].

2. METODE

Metode yang digunakan adalah metode *Random Forest Imputation* dan *Classification and Regression Trees*. Jenis data dalam penelitian ini berupa data numerik yaitu data bivariate gamma yang dibangkitkan pada program R studio versi R 4.5.1 dengan *pacekages* VGAM berjumlah 200 data dengan 2 variabel.
Langkah-langkah analisis data

1. Membangkitkan data bivariate gamma menggunakan program R Studio
2. Mencari statistik deskriptif data
3. Membuat *Missing Value* 5%, 10%, 15%, 20% pada data
4. Imputasi data *missing value* menggunakan 2 Metode
5. Mencari korelasi hasil imputasi 5%, 10%, 15%, 20% pada setiap metode
6. Melakukan uji beda rata-rata pada setiap metode

3. HASIL DAN PEMBAHASAN

3.1 Deskripsi Data

Dalam pelitian ini, data yang digunakan adalah data Bivariate Gamma yang di bangkitkan pada program R studio dengan *pacekages VGAM* berjumlah 200 data dengan 2 variabel.

3.2 Statistik Deskriptif Data

Berikut adalah data bivariate gamma yang di bangkitkan pada program R:

Tabel 1. Data Bivariate Gamma

V1	V2
3.554	9.447
1.818	5.895
⋮	⋮
2.903	11.821

Berikut merupakan statistik deskriptif pada data bivariate gamma yang telah di bangkitkan:

Tabel 2. Statistik Deskriptif Data

Item	V1	V2
Minimum	0.219	2.522
Quartil 1	1.856	6.716
Median	3.085	9.166
Mean	4.110	9.925
Quartil 3	5.601	12.473
Maksimum	23.575	32.948

Berdasarkan Tabel 2, dapat dijelaskan bahwa pada variabel V1 nilai minimum tercatat sebesar 0,219, kuartil pertama sebesar 1,856, nilai median sebesar 3,085, nilai rata-rata sebesar 4,110, kuartil ketiga sebesar 5,601, dan nilai maksimum sebesar 23,575. Sementara itu, pada variabel V2 nilai minimum adalah 2,522, kuartil pertama sebesar 6,716, nilai median sebesar 9,166, nilai rata-rata sebesar 9,925, kuartil ketiga sebesar 12,473, serta nilai maksimum sebesar 32,948.

3.3 *Missing Value* Data

Sebelum melakukan imputasi, akan dilakukan *missing value* pada data sebesar 5%, 10%, 15% dan 20%. Berikut adalah data setelah dilakukan *missing value*:

Tabel 3. Data *Missing Value*

Item	5%		10%		15%		20%	
	V1	V2	V1	V2	V1	V2	V1	V2
Minimum	0.219	2.522	0.219	2.522	0.219	2.522	0.219	2.522
Quartil 1	1.806	6.610	1.730	6.683	1.759	6.601	1.650	6.505
Median	3.044	9.120	3.078	9.175	3.064	9.152	2.941	9.062
Mean	4.092	9.874	4.079	9.940	4.115	9.934	4.024	9.794
Quartil 3	5.601	12.473	5.524	12.491	5.735	12.396	5.477	12.508
Maksimum	23.575	32.948	23.575	32.948	23.575	32.948	23.575	32.948
NA's	12	8	27	13	32	28	50	30

Berdasarkan informasi pada Tabel 3, dapat diketahui bahwa pada skenario data hilang sebesar 5%, jumlah nilai hilang (NA) pada variabel V1 sebanyak 12 data, sedangkan pada variabel V2 sebanyak 8 data. Pada tingkat data hilang 10%, variabel V1 memiliki 27 data hilang, sementara variabel V2 memiliki 13 data hilang. Selanjutnya, pada kondisi data hilang 15%, jumlah NA pada variabel V1 tercatat sebanyak 32 data, dan pada variabel V2 sebanyak 28 data. Adapun pada tingkat data hilang 20%, variabel V1 mengalami 50 data hilang, sedangkan variabel V2 memiliki 30 data hilang

3.4 Imputasi Data

3.4.1 Imputasi *Classification and Regression Trees*

Imputasi dilakukan pada data *missing value* sebesar 5%, 10%, 15% dan 20%, pada tahapan ini di lakukan imputasi sebesar 5% dan diperoleh hasil MAPE, RMSE dan Korelasi sebagai berikut:

Tabel 4. MAPE, RMSE dan CORELASI pada Data

MAPE	RMSE	CORELASI
0.538437	3.107496	0.7086

Berikut adalah uji beda rata-rata korelasi pada data:

1. Pengujian Hipotesis

$H_0 : \mu = \mu_0$ (Tidak berbeda secara signifikan)
$H_1 : \mu \neq \mu_0$ (Berbeda secara signifikan)

2. Besaran yang diperlukan
   Taraf Signifikansi $\alpha = 5\%, n = 200$
3. Statistik uji
   $$t = \frac{\bar{X} - \mu}{\left(\frac{SD}{\sqrt{n}}\right)} = \frac{0.7086 - 0.7288161}{\left(\frac{0.011}{\sqrt{200}}\right)} = -12.987$$
   $p - value = 2.2e - 16$
4. Daerah Penolakan
   Tolak H_0 jika nilai $t_{hit} > t_{\frac{\alpha}{2},n-1}$ atau $p - value < \alpha = 0.05$
5. Kesimpulan
   Oleh karna nilai $p - value = 2.2e - 16 < \alpha = 0.05$, maka H_0 ditolak, artinya nilai rata-rata antara μ dan μ_0 berbeda secara signifikan

3.4.2 *Random Forest Imputations*

Imputasi dilakukan pada data *missing value* sebesar 5%, 10%, 15% dan 20%, pada tahapan ini dilakukan imputasi sebesar 5% dan diperoleh hasil MAPE, RMSE dan Korelasi sebagai berikut:

Tabel 5. MAPE, RMSE dan CORELASI pada Data

MAPE	RMSE	CORELASI
0.567461	3.185404	0.7096

Berikut adalah uji beda rata-rata korelasi pada data:

1. Pengujian Hipotesis
   $H_0 : \mu = \mu_0$ (Tidak berbeda secara signifikan)
   $H_1 : \mu \neq \mu_0$ (Berbeda secara signifikan)
2. Besaran yang diperlukan
   Taraf Signifikansi $\alpha = 5\%, n = 200$
3. Statistik uji
   $$t = \frac{\bar{X} - \mu}{\left(\frac{SD}{\sqrt{n}}\right)} = \frac{0.7096 - 0.7288161}{\left(\frac{0.01107}{\sqrt{200}}\right)} = -12.202$$
   $p - value = 2.2e - 16$
4. Daerah Penolakan
   Tolak H_0 jika nilai $t_{hit} > t_{tabel}$ atau $p - value < \alpha = 0.05$
5. Kesimpulan
   Oleh karna nilai $p - value = 2.2e - 16 < \alpha = 0.05$, maka H_0 ditolak, artinya nilai rata-rata antara μ dan μ_0 berbeda secara signifikan

Berdasarkan metode Imputasi *Classification and Regression Trees* dan Metode *Random Forest Imputations,* maka diperoleh hasil imputasi seperti pada tabel berikut:

Tabel 6. Kesimpulan pada Data

No	Missing Value	Metode Imputasi	Keterangan	Corelasi 0.72881	P-Value	MAPE	RMSE
1	5%	CART	Berbeda secara signifikan	0.7086	$2.2e - 16$	0.538	3.108
		RF	Berbeda secara signifikan	0.7096	$2.2e - 16$	0.568	3.185
2	10%	CART	Berbeda secara signifikan	0.7211	$8.2e - 7$	0.525	3.398
		RF	Tidak berbeda secara signifikan	0.7270	0.3387	0.525	3.370
3	15%	CART	Berbeda secara signifikan	0.7226	0.017	0.599	4.181
		RF	Berbeda secara signifikan	0.7381	0.0001	0.593	3.891

4	20%	CART	Berbeda secara signifikan	0.7793	$2.2e-16$	0.855	4.058
		RF	Berbeda secara signifikan	0.7788	$2.2e-16$	0.827	3.973

Berdasarkan Tabel 6 dapat diketahui bahwa pada imputasi dengan *missing value* 5%, metode imputasi CART dan RF memiliki nilai rata-rata korelasi 0.7086 dan 0.7096, nilai $p-value$ yang lebih kecil dari $\alpha = 0.05$ artinya nilai rata-rata berbeda secara signifikan dengan rata-rata awalnya, pada data *missing value* 5% metode CART adalah metode imputasi yang memiliki nilai MAPE dan RMSE terkecil yaitu sebesar 0.538 dan 3.108. Data dengan *missing value* 10% pada metode imputasi RF memiliki nilai rata-rata korelasi 0.7270 dan nilai $p-value$ yang lebih besar dari $\alpha = 0.05$ artinya nilai rata-rata tidak berbeda secara signifikan dengan rata-rata awalnya. Sata *missing value* 10% metode CART adalah metode imputasi yang memiliki nilai MAPE dan RMSE terkecil yaitu sebesar 0.525 dan 3.398. Data dengan *missing value* 15% pada metode imputasi CART dan RF memiliki nilai rata-rata korelasi 0.7226 dan 0.738, nilai $p-value$ yang lebih kecil dari $\alpha = 0.05$ artinya nilai rata-rata berbeda secara signifikan dengan rata-rata awalnya. Data *missing value* 15% metode RF adalah metode imputasi yang memiliki nilai MAPE dan RMSE terkecil yaitu sebesar 0.593 dan 3.891. Data dengan *missing value* 20% pada metode imputasi CART dan RF memiliki nilai $p-value = 2.2e-16 < \alpha = 0.05$ artinya nilai rata-rata berbeda secara signifikan dengan rata-rata awalnya. Data *missing value* 20% metode RF adalah metode imputasi yang memiliki nilai MAPE dan RMSE terkecil yaitu sebesar 0.827 dan 3.973.

4. SIMPULAN

Berdasarkan hasil analisis dan pembahasan mengenai metode imputasi, maka dapat diambil kesimpulan bahwa metode CART dan *Random Forest* menunjukkan variasi ketepatan yang berbeda pada setiap banyaknya data hilang. *Missing value* sebesar 5%, kedua metode memiliki nilai rata-rata korelasi yang relatif mirip, yaitu 0,7086 untuk CART dan 0,7096 untuk *Random Forest*. Namun, nilai *p*-v*alue* yang lebih kecil dari $\alpha = 0{,}05$ pada keduanya menunjukkan bahwa terdapat perbedaan signifikan antara nilai rata-rata hasil imputasi dan nilai rata-rata awal. Metode CART adalah metode terbaik dengan nilai MAPE dan RMSE terkecil, yaitu 0,538 dan 3,108. Selanjutnya, pada missing value 10%, metode *Random Forest* memiliki rata-rata korelasi terbesar, yaitu 0,7270, dengan p-$value$ yang lebih besar dari $\alpha = 0{,}05$ sehingga tidak terdapat perbedaan signifikan terhadap nilai rata-rata awal. Sedangkan, metode CART tetap metode terbaik dengan nilai MAPE dan RMSE terkecil yaitu sebesar 0,525 dan 3,398.

Data *missing value* sebesar 15%, metode CART dan *Random Forest* menunjukkan p-$value < \alpha = 0{,}05$, yang berarti hasil imputasi keduanya berbeda signifikan dengan nilai awal. Metode *Random Forest* memberikan lebih baik dibandingkan CART dengan nilai MAPE dan RMSE terkecil, yaitu 0,593 dan 3,891. Sementara itu, pada *missing value* 20%, kedua metode menghasilkan p-$value$ sebesar $2{,}2 \times 10^{-16}$, yang menunjukkan perbedaan signifikan terhadap nilai rata-rata awal. Pada kondisi data hilang terbesar ini, *Random Forest* menjadi metode terbaik dengan nilai MAPE $= 0{,}827$ dan RMSE $= 3{,}973$. Secara keseluruhan, hasil penelitian menunjukkan bahwa CART lebih unggul pada tingkat *missing value* rendah, sementara *Random Forest* lebih stabil dan memberikan performa lebih baik pada tingkat *missing value* yang lebih tinggi.

UCAPAN TERIMA KASIH

Penulis ingin menyampaikan terimakasih kepada Universitas Negeri Jakarta yang telah memberikan dukungan komputasi dan pemrosesan data selama proses penelitian. Penulis juga berterima kasih kepada rekan-rekan akademisi atas masukan dan diskusi berharga mengenai penerapan Imputasi data hilang pada data bangkitan Bivariat Gamma.

REFERENSI

[1] S. Hong and H. S. Lynn, "Accuracy of random-forest-based imputation of missing data in the presence interaction," *J. BMC Med. Res. Methodol.*, vol. 1, pp. 1–12, 2020.

[2] B. O. Petrazzini, H. Naya, F. Lopez-bello, G. Vazquez, and L. Spangenberg, "Evaluation of different approaches for missing data imputation on features associated to genomic data," *BioData Min.*, pp. 1–13, 2021.

[3] M. Kokla, J. Virtanen, M. Kolehmainen, J. Paananen, and K. Hanhineva, "Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data : a comparative study," *J. BMC Med. Res. Methodol.*, pp. 1–11, 2019.

[4] Y. Ge, Z. Li, and J. Zhang, "OPEN A simulation study on missing data imputation for dichotomous variables using statistical and machine learning methods," *Sci. Rep.*, no. 0123456789, pp. 1–13, 2023.

[5] W. Agwil, D. Agustina, H. Fransiska, and I. A. Hasani, "Meningkatkan Kinerja Model Klasifikasi Curah Hujan Melalui Penanggulangan Missing Value Dengan Imputasi Berbasis Model," *Innov. J. Soc. Sci. Res.*, vol. 4, pp. 11773–11783, 2024.

[6] M. Franco and J. Vivo, "A Generator of Bivariate Distributions : Properties , Estimation , and Applications," *Math. Artic.*, vol. 8, no. 1776, pp. 1–30, 2020.

[7] C. Caamaño-Carrillo and J. E. Contreras-Reyes, "A Generalization of the Bivariate Gamma Distribution Based on Generalized Hypergeometric Functions," *Mathematics*, no. 3, pp. 1–17, 2022.

[8] C. K. Amponsah, T. J. Kozubowski, and A. K. Panorska, "A general stochastic model for bivariate episodes driven by a gamma sequence," *J. ofStatistical Distrib. Appl.*, vol. 8, 2021.

[9] D. Curtis, "Welch's t test is more sensitive to real world violations of distributional assumptions than student's t test but logistic regression is more robust than either," *Springer, Stat. Pap.*, pp. 3981–3989, 2024.

[10] H. Mustafidah, A. Imantoyo, and S. Suwarsito, "Pengembangan Aplikasi Uji-t Satu Sampel Berbasis Web," *JUITA J. Inform.*, vol. 8, no. 2, p. 245, 2020, doi: 10.30595/juita.v8i2.8786.

[11] D. J. Stekhoven and P. Bühlmann, "MissForest — non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 2012.

[12] D. J. Stekhoven, "Nonparametric Missing Value Imputation using Random Forests," *CRAN (manual). (terbaru; dokumentasi paket).*, 2025.

[13] A. D. Shah, J. W. Bartlett, J. Carpenter, O. Nicholas, and H. Hemingway, "Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE : A CALIBER Study," *Am. J. Epidemiol.*, vol. 179, no. 7, pp. 764–774, 2014.

[14] R. Rachmawati, N. Afandi, and M. A. Alwansyah, "Survival Analysis on Data of Students Not Graduating on Time Using Weibull Regression, Cox Proportional Hazards Regression, and Random Survival Forest Methods," *Barekeng*, vol. 19, no. 3, pp. 2111–2126, 2025.

[15] M. A. Alwansyah, "Survival Analysis of Students Not Graduated on Time Using Cox Proportional Hazard Regression Method and Random Survival Forest Method," *J. Stat. Data Sci.*, vol. 2, no. 1, pp. 13–21, 2023.

[16] E. Slade and M. G. Naylor, "A fair comparison of tree-based and parametric methods in multiple imputation by chained equations," *HHS Public Access*, vol. 39, no. 8, pp. 1156–1166, 2022.

[17] J. Li *et al.*, "Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets," *BMC Med. Res. Methodol.*, no. 24:41, pp. 1–9, 2024.

[18] M. Afkanpour, E. Hosseinzadeh, and H. Tabesh, "Identify the most appropriate imputation method for handling missing values in clinical structured datasets : a systematic review," *BMC Med. Res. Methodol.*, no. 24:188, 2024.

[19] T. Iida, "Identifying causes of errors between two wave-related data using performance metrics," *Appl. Ocean Res.*, vol. 148, no. March, p. 104024, 2024.

[20] K. Warneke, S. D. Siegel, J. Afonso, and S. Wallot, "What the mean absolute percentage error (MAPE) should adopt from Bland – Altman analyses," *Ger. J. Exerc. Sport Res.*, 2025.

[21] T. O. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not," *Geosci. Model Dev.*, no. 2, pp. 5481–5487, 2022.

[22] P. Schober, C. Boer, and L. A. Schwarte, "Correlation Coefficients: Appropriate Use and Interpretation," *aournal Anesth.*, vol. 126, no. 5, pp. 1763–1768, 2018.