Perbandingan Kinerja Metode *Hybrid* KNNI-GA dan *MissForest* Dalam Menangani *Missing Values*

(Comparison of Performance Hybrid KNNI-GA and MissForest Methods In Handling Missing Values)

Lalu Moh. Arsal Fadila[1*], Siti Muchlisoh[2]

[1,2]*Politeknik Statistika STIS*
DKI Jakarta, Indonesia
Email*:* *111911119@stis.ac.id*

ABSTRAK

Suatu proses bisnis survei yang baik dan benar adalah memilih sampel yang representatif agar dapat memperoleh data yang berkualitas. Akan tetapi salah satu masalah yang relevan dalam kualitas data adalah adanya *missing values*. *Missing value* hampir ditemukan di semua pengumpulan data berskala besar. *Missing values* dapat menyebabkan berbagai macam masalah. Oleh karena itu, hal tersebut harus ditangani. Salah satu cara mengatasi *missing values* adalah dengan metode imputasi. *Hybrid* KNNI-GA dan *missForest* merupakan metode imputasi yang dapat digunakan untuk menangani *missing values*. *Hybrid* KNNI-GA menggunakan algoritma genetika untuk pemilihan nilai k optimum dan memerlukan variabel prediktor untuk melakukan imputasi. Sedangkan *missForest* membentuk model untuk melakukan proses imputasi. Penelitian ini membandingkan *hybrid* KNNI-GA dan missForest dalam menangani *missing values* dari segi ketepatan estimator dan performa komputasi. Hasil simulasi yang didapatkan, *hybrid* KNNI-GA lebih baik daripada missForest dari segi ketepatan estimator. Sedangkan performa komputasi missForest lebih stabil dibandingkan *hybrid* KNNI-GA.

Kata kunci: *KNNI-GA, MissForest,* Imputasi, *Missing Values*

ABSTRACT

A good and correct survey business process is to select a representative sample in order to obtain quality data. However, one of the relevant problems in data quality is the presence of missing values. Missing value is found in almost all large-scale data collections. Missing values can cause all sorts of problems. Therefore, it must be addressed. One way to overcome missing values is the imputation method. Hybrid KNNI-GA and missForest are imputation methods that can be used to handle missing values. Hybrid KNNI-GA uses a genetic algorithm to select the optimum k value and requires predictor variables to perform imputation. Meanwhile, missForest forms a model to carry out the imputation process. This study compares the hybrid KNNI-GA and missForest in dealing with missing values in terms of estimator accuracy and computational performance. The simulation results obtained, the KNNI-GA hybrid is better than missForest in terms of estimator accuracy. Meanwhile, missForest's computational performance is more stable than the KNNI-GA hybrid.

Keywords*: KNNI-GA, MissForest, Imputation, Missing Values*

PENDAHULUAN

Suatu proses bisnis survei yang baik dan benar adalah memilih sampel yang representatif agar dapat dilakukan generalisasi terhadap populasi sehingga dapat menyimpulkan keadaan populasi yang sebenarnya dan dapat memperoleh data yang berkualitas. Kualitas data sangat penting dalam menganalisis fenomena di berbagai bidang disiplin ilmu, tak terkecuali statistik (Sidi et al., 2012). Akan tetapi salah satu masalah yang relevan dalam kualitas data adalah adanya data hilang atau yang dikenal dengan *missing values.* Adanya *missing values* menyebabkan data menjadi tidak lengkap atau *incompleteness.*

Missing values adalah kondisi di mana tidak tersedianya data atau fitur pada dataset yang menyebabkan dataset menjadi tidak lengkap (Sallaby et al., 2021). Menurut Kaiser (2014) *missing value* dapat terjadi karena responden tidak menjawab semua pertanyaan pada kuesioner, ketika entri data secara manual, pengukuran yang salah, eksperimen yang salah, beberapa data disensor, dan banyak lainnya. Selain itu *missing values* dapat berupa *outlier* atau nilai yang bermasalah sehingga harus dihapus sebelum dilakukan suatu analisis (Kaiser, 2014).

Menurut Kang (2013) ada empat masalah yang terjadi karena *missing values*. Pertama, hilangnya data mengurangi *statistical power* yang mengacu pada kemungkinan bahwa pengujian akan menolak hipotesis nol ketika hipotesis itu salah. Kedua, dapat menyebabkan bias dalam pendugaan parameter. Ketiga, dapat mengurangi keterwakilan sampel. Keempat, dapat mempersulit analisis penelitian. Masing-masing masalah tersebut dapat menyebabkan kesimpulan yang tidak tepat. Masalah lain yang dapat disebabkan oleh *missing value* yakni terbuangnya dana yang telah digunakan untuk pengumpulan data (Pazanudin, 2017). Oleh sebab itu, perlu dilakukan penanganan untuk permasalahan terkait *missing values.*

Ada tiga cara dalam menangani *missing values* (Cohen, 1996). Pertama, menghapus *missing values* tersebut untuk variabel apapun yang dianalisis. Kedua, menghapus *missing values* hanya jika diperlukan untuk perkiraan tertentu. Ketiga, melakukan imputasi pada *missing values*. Namun, cara pertama dan kedua tersebut jika dilakukan, terlebih pada banyak variabel yang dianalisis akan menyebabkan bias dan ketidakkonsistenan. Sedangkan cara yang ketiga lebih baik digunakan karena dapat dilakukan proses analisis lanjutan tanpa menghapus adanya *missing values*. Little dan Rubin (1988) menyatakan bahwa salah satu cara mengatasi *missing value* adalah dengan metode imputasi. Metode imputasi merupakan suatu metode yang bertujuan untuk mengisi *missing value* dengan nilai yang diperkirakan. Keuntungan dari metode imputasi adalah data dapat diperhitungkan atau dianalisis dari pengumpulan data yang sebelumnya mengandung *missing values* (Cohen, 1996).

Biemer dan Lyberg (2003) menyatakan bahwa *missing value* hampir ditemukan di semua pengumpulan data berskala besar. Salah satu pengumpulan data dengan skala besar rutin dilakukan oleh Politeknik Statistika Sekolah Tinggi Ilmu Statistik (Polstat STIS) dalam bentuk praktik kerja lapangan (PKL). Beberapa tantangan yang dihadapi ketika mengumpulkan data pada PKL yaitu responden tidak mau informasi tentangnya dicatat, tidak mengingat jawaban atas pertanyaan yang diajukan, tidak mau menjawab pertanyaan karena menganggap pertanyaan tersebut memalukan, dan pewawancara terlalu lama melakukan proses wawancara (Longford, 2005). Hal-hal tersebut yang dapat memengaruhi terjadinya *missing values*. Selain itu beberapa moda pengumpulan data yang digunakan seperti *computer assisted personal interviewing* (CAPI), *computer assisted telephone interviewing* (CATI), dan *computer assisted web interviewing* (CAWI) tidak terlepas dari risiko kesalahan dalam pengumpulan data yang dilakukan (Silva-Ramírez et al., 2011). Oleh karena itu, pada dasarnya data awal PKL masih dalam keadaan mentah dan memerlukan proses *cleaning* data. Data tersebut perlu melewati proses pengolahan terlebih dahulu. Salah satu tahapan yang dilakukan yakni dengan melakukan imputasi data. Imputasi dilakukan pada data yang hilang dan pada isian yang tidak wajar atau tidak konsisten. Selain itu, imputasi juga dilakukan untuk melakukan *treatment* terhadap *outlier* pada data. Dengan demikian penting untuk menentukan metode imputasi terbaik yang akan digunakan.

Jerez et al. (2010) menyatakan bahwa metode imputasi terbagi menjadi dua, yakni metode imputasi berbasis statistik dan *machine learning*. Metode imputasi berbasis statistik melakukan proses imputasi menggunakan kaidah-kaidah analisis statistik. Beberapa contoh metode imputasi berbasis statistik yaitu *mean imputation, hot-deck imputation, cold-deck imputation,* metode *regression,* dan *multiple imputation*. Sedangkan metode imputasi berbasis *machine learning* melakukan imputasi data dengan memanfaatkan metode pembelajaran/*learning* untuk memprediksi nilai yang hilang. Beberapa contoh metode imputasi berbasis *machine learning* yaitu *K-Nearest Neighbour Imputation*, C4.5, CN2, dan *missforest*.

K-Nearest Neighbour (KNN) merupakan metode imputasi berbasis *machine learning* yang memanfaatkan pembelajaran mesin. Metode KNN dapat dikombinasikan dengan imputasi untuk mengatasi masalah *missing values,* biasanya disebut dengan *K-Nearest Neighbour Imputation* (KNNI). KNNI melakukan proses imputasi dengan cara menemukan pola berdasarkan jarak ke tetangga terdekat. Beberapa penelitian tentang KNNI telah banyak dilakukan, seperti penelitian oleh Troyanskaya et al. (2001), Batista dan Monard, (2002), Malarvizhi (2012), Sartika (2018), dan Fadillah dan Muchlisoh (2020) menyatakan bahwa metode KNNI merupakan salah satu metode imputasi terbaik dalam menangani *missing values*. Kelebihan dari metode KNNI yakni dapat digunakan untuk imputasi data berskala numerik maupun kategorik.

Akan tetapi salah satu kelemahan pada metode KNNI adalah permasalahan dalam pemilihan nilai k yang kurang tepat dapat menurunkan kinerja KNNI. Salah satu cara untuk meningkatkan akurasi kinerja pada sebuah metode adalah dengan menggunakan teknik optimasi. Algoritma genetika atau *genetic algorithm* (GA) merupakan algoritma yang dapat digunakan untuk menangani masalah optimasi dengan cara mereplikasi proses genetika pada makhluk hidup. Pada penelitian yang dilakukan oleh Irhamah (2012) didapatkan hasil bahwa metode KNNI yang dioptimasi dengan GA dapat memperoleh nilai *k* yang optimum dan dapat meningkatkan akurasi dari KNNI. Penelitian lain yang dilakukan oleh Hayatin (2012) menyatakan bahwa GA yang digunakan untuk mendapatkan nilai *k* yang optimum pada KNNI, terbukti dapat meningkatkan akurasi dari metode KNNI tersebut. Oleh karena itu, KNNI dapat dikombinasikan dengan GA atau dapat disebut dengan *hybrid* KNNI-GA.

Selain metode *hybrid* KNNI-GA, metode imputasi berbasis *machine learning* lainnya adalah *missForest*. *MissForest* memanfaatkan algoritma *random forest* untuk melakukan imputasi *missing values*. Metode ini membuat model *random forest* untuk setiap variabel pada dataset dan menggunakannya untuk memprediksi *missing values* untuk setiap variabel penelitian. Beberapa penelitian tentang *missForest* telah banyak dilakukan, seperti penelitian oleh Stekhoven dan Bühlmann (2012), Misztal (2013), Pazanudin (2017), dan Zhang et al. (2021) menyatakan bahwa *missForest* merupakan salah satu metode imputasi terbaik yang dapat digunakan untuk melakukan imputasi pada kasus *missing values*. Keuntungan menggunakan metode ini adalah dapat diterapkan pada data bertipe kontinu dan kategorik, membutuhkan sedikit penyesuaian, dan menyediakan estimasi kesalahan yang divalidasi silang secara internal. Namun, kekurangan dari metode ini adalah membutuhkan waktu yang lama dalam proses imputasi karena dipengaruhi oleh pembentukan model dibandingkan penggantian *missing values* (Pazanudin, 2017).

Algoritma dan cara penanganan *missing values* yang berbeda pada metode *missForest* dan *hybrid* KNNI-GA tentunya akan menghasilkan estimasi yang berbeda. Namun, kedua metode tersebut sama-sama memanfaatkan pembelajaran mesin (*machine learning*) dan dapat diterapkan pada data bertipe numerik/kontinu dan kategorik dalam proses imputasi, Oleh karena itu peneliti tertarik untuk membandingkan kedua metode imputasi berbasis *machine learning* tersebut yaitu *hybrid* KNNI-GA dengan *missForest* dalam menangani *missing values* serta mengidentifikasi kelebihan serta kekurangan dari kedua metode imputasi tersebut.

METODE

Landasan Teori

KNNI melakukan proses imputasi dengan cara menemukan pola berdasarkan jarak ke tetangga terdekat. Untuk mengukur jarak tetangga tersebut dapat berdasarkan jarak *Euclidean,* jarak *Manhattan,* atau jarak *Gower*. Pada prinsipnya, variabel yang mengandung *missing values* akan diisi oleh nilai-nilai yang sama dari variabel lain yang memiliki nilai yang lengkap. Variabel yang memiliki nilai lengkap disebut dengan variabel prediktor (Longford, 2005). Sebelum menjalankan algoritma KNN, ditentukan terlebih dahulu jumlah tetangga terdekatnya (k). Agar mendapatkan nilai k yang optimal dan tepat, algoritma GA dapat digunakan sebagai alternatif. GA dapat digunakan seleksi variabel prediktor. Pada KNNI-GA untuk seleksi variabel, sebagai contoh variabel prediktor yang digunakan sebanyak 5 variabel. Maka, akan terbentuk sebuah vektor biner berukuran 5 x 1 yang terdiri dari nilai 1 dan 0 dan dapat diartikan sebagai vektor indikator yang menunjukkan variabel-variabel tersebut diikutsertakan, kode 1 menunjukkan keikutsertaan variabel, sedangkan kode 0 tidak ikut. Contoh kromosom individu yang terbentuk: [1 1 0 1 1]. Selanjutnya nilai k optimum didapatkan dari pengkodean dari vektor biner yang terbentuk. Adapun nilai k direpresentasikan dalam 4 gen biner. Sebagai contoh, kromosom individu yang terbentuk adalah: [1 0 1 0]. Maka nilai k yang digunakan adalah $k = (1010)_2$, ditransformasi ke nilai desimal menjadi $k = (10)_{10}$. Selanjutnya, imputasi *missing values* dengan menghitung nilai estimasi rata-rata bobot pada pengamatan k tetangga terdekat yang tidak mengandung *missing values* dengan rumus matematis sebagai berikut:

$$\bar{x}_j = \frac{\sum_{l=1}^{k} w_k \, v_{jk}}{\sum_{l=1}^{k} w_k} \quad (1)$$

$$w_k = \frac{1}{d_{(i,j)_k}} \quad (2)$$

Di mana :

$\bar{x}_j$ = estimasi rata-rata tertimbang pada observasi ke-j
v_{jk} = nilai data yang lengkap pada variabel observasi ke-j dari tetangga ke-k
k = jumlah observasi terdekat yang digunakan
l = observasi dari k
w_k = bobot dari observasi tetangga terdekat ke-k
$d_{(i,j)_k}$ = jarak observasi i ke j tetangga ke-k

MissForest merupakan metode imputasi nonparametrik menggunakan *random forest*. *MissForest* memanfaatkan algoritma *random forest* untuk melakukan imputasi *missing values*. Metode ini membuat model *random forest* untuk setiap variabel pada dataset dan menggunakannya untuk memprediksi *missing values*

untuk setiap variabel penelitian. Hal tersebut dilakukan dalam model *cyclic* untuk semua variabel dan seluruh proses diulang secara iteratif sampai mencapai kriteria maksimum. Selain itu menurut Penone et al. (2014), kelebihan dari *missForest* adalah dapat diterapkan pada tipe data numerik maupun kategorik dan sedikit pengujian asumsi.

Data dan Sumber Data

Pada penelitian ini data yang digunakan ada tiga, yakni data bangkitan, data *Hepatitis C Virus for Egyptian Patients* yang diambil UCI *repository machine learning*, dan data PKL Politeknik Statistika STIS Tahun 2022. Data bangkitan akan dibangkitkan menggunakan *software* Rstudio. Data bangkitan tersebut terdiri dari empat variabel. Variabel prediktor yang digunakan pada data tersebut masing-masing berdistribusi binomial, distribusi normal, dan distribusi *uniform*. Sedangkan variabel imputasinya berdistribusi normal dan bertipe numerik. Keempat variabel tersebut dibangkitkan dengan secara *random* di *software* Rstudio dengan sejumlah 1000 observasi. Diperoleh untuk variabel prediktor pertama $X_1 \sim BIN(100, 0{,}5)$, variabel prediktor kedua $X_2 \sim N(15, 7)$, dan variabel prediktor ketiga $X_3 \sim UNIF(-1{,}1)$. Sedangkan untuk variabel imputasinya diperoleh $Y \sim N(10, 5)$. Kemudian nilai-nilai dari variabel-variabel tersebut akan digunakan dalam proses analisis.

Data berikutnya adalah data *Hepatitis C Virus for Egyptian Patients* yang diambil UCI *repository machine learning*. Data ini memiliki 29 variabel dengan sejumlah 1385 observasi. Data tersebut pertama kali digunakan oleh Sanaa Kamal, dkk pada tahun 2017. Adapun data tersebut berisi tentang kondisi pasien di Mesir yang menjalani pengobatan dengan dosis *hepatitis C Virus* selama 18 bulan. Selain itu, variabel *body mass index* digunakan sebagai variabel imputasi dan bertipe numerik, sedangkan variabel *age*, *gender*, *headache,* dan *hemoglobin* digunakan sebagai variabel prediktor.

Data selanjutnya yakni data PKL Polstat STIS Tahun 2022. Variabel yang digunakan adalah lama waktu wawancara, penerimaan awal responden, jenis kelamin responden, ijazah tertinggi responden, kemampuan responden berbahasa Indonesia, dan situasi lingkungan pada saat wawancara. Adapun variabel lama waktu wawancara akan digunakan sebagai variabel imputasi dan bertipe numerik, sedangkan yang lainnya sebagai variabel prediktor. Variabel-variabel yang digunakan merupakan variabel-variabel yang merupakan instrumen dalam mengukur paradata pada PKL Polstat STIS Tahun 2022 dengan sejumlah 5269 observasi. Variabel prediktor tersebut diperoleh dari penelitian Krosnick (1991), Barry et al. (2002), Kreuter et al. (2010), Lau et al. (2017), dan Darcy dan Hoeta (2021).

Metode Analisis

Simulasi dilakukan pada penelitian ini untuk melakukan analisis. Simulasi dilakukan dengan tiga tahap. Tahap pertama adalah membentuk suatu dataset yang mengandung *missing values*. Dataset dibentuk dengan cara menghapus data secara *random* dengan pola *univariate* dan mekanisme MCAR dengan tingkat *missing* sebesar 10 persen, 20 persen, 30 persen, 40 persen, 50 persen, dan 60 persen. Penghapusan dengan tingkat *missing* tersebut berdasarkan penelitian yang pernah dilakukan oleh Batista dan Monard (2002) dan Fadillah dan Puspita (2022). Tahap kedua adalah melakukan imputasi data. Imputasi data dilakukan pada setiap dataset yang mengandung *missing values* menggunakan metode *K-Nearest Neighbour Imputation* dan *missForest*. *Package VIM* dan *MissForest* pada RStudio digunakan untuk melakukan imputasi pada kedua metode tersebut. Tahap yang ketiga adalah analisis. Sebelum dilakukan perbandingan hasil imputasi pada kedua metode tersebut, terlebih dahulu dilakukan pemilihan nilai k optimum pada KNNI menggunakan GA. Hasil imputasi dengan pemilihan nilai k optimum dengan GA tersebut akan digunakan untuk dibandingkan dengan metode *missForest*. Adapun RMSE digunakan untuk mengukur akurasi kinerja atau ketepatan estimator dari kedua metode imputasi. Selain itu, *running time* atau waktu yang diperlukan untuk melakukan proses imputasi digunakan untuk mengukur performa komputasi. Proses pada tahapan ini dilakukan sebanyak 10 kali percobaan. Kemudian nilai rata-rata dari RMSE dan *running time* dari kedua metode tersebut dibandingkan dan dianalisis. Berikut rumus matematis dari RMSE:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{3}$$

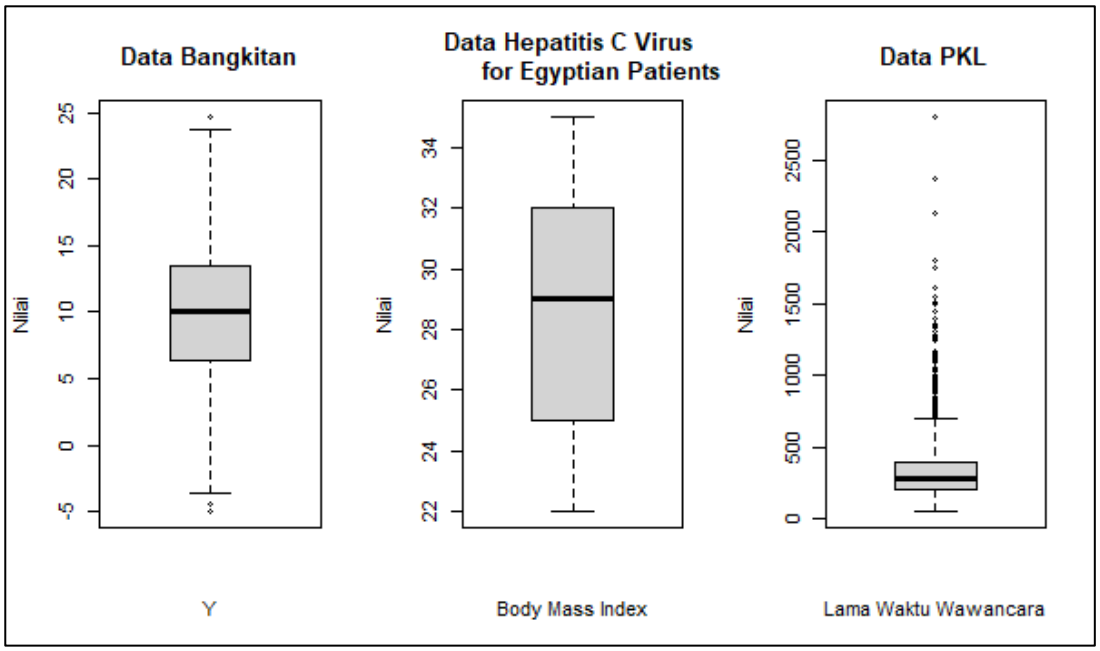
Di mana:

$\hat{y}_i$ = nilai prediksi observasi ke-i
y_i = nilai aktual observasi ke-i
n = jumlah observasi pada estimasi *missing values*

HASIL DAN PEMBAHASAN

Sebaran Data Masing-masing Variabel Imputasi

Berdasarkan Gambar 1 terlihat masing-masing sebaran data dari masing-masing variabel imputasi yang digambarkan melalui *boxplot*. Pada data bangkitan dan data *Hepatitis C Virus for Egyptian Patients* terlihat bahwa sebaran data variabel imputasinya tidak mengandung *outlier* yang ekstim, sedangkan pada data PKL terlihat bahwa sebaran data variabel imputasinya mengandung *outlier* dan juga terdapat *outlier* yang ekstrim.



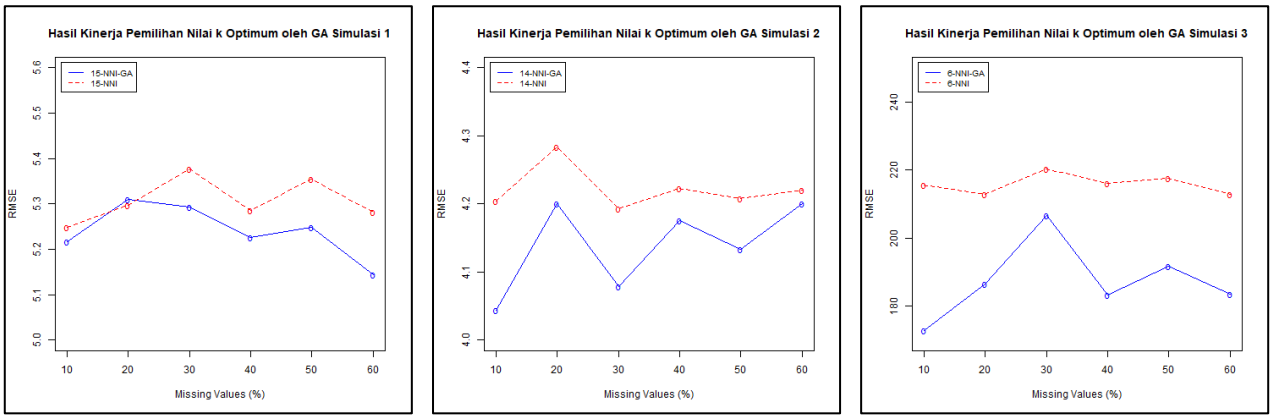
Gambar 1. *Box Plot* Masing-Masing Variabel Imputasi

Pemilihan Nilai K Optimum

Sebelum dilakukan proses algoritma KNNI, ditentukan terlebih dahulu nilai k optimum oleh GA seleksi variabel untuk masing-masing simulasi yang terlihat pada Gambar 2. Pada simulasi 1, *hybrid* KNNI-GA dengan seleksi variabel memberikan nilai k optimum sebesar 15 dengan kombinasi variabel yang terbentuk [1 1 1 1 1 1 0]. Artinya bahwa 4 gen biner pertama yang terbentuk adalah $\boldsymbol{k = (1111)_2}$ yang kemudian dikodekan menjadi nilai desimalnya sebesar $\boldsymbol{k = (15)_{10}}$. Adapun 3 gen biner terakhir terbentuk vektor biner [1 1 0] yang artinya variabel pertama dan kedua ikut serta dalam proses imputasi, sedangkan variabel ketiga tidak diikutsertakan.

Selanjutnya pada simulasi 2, *hybrid* KNNI-GA dengan seleksi variabel memberikan nilai k optimum sebesar 14 dengan kombinasi variabel yang terbentuk [1 1 1 0 0 0 1 0]. Artinya bahwa 4 gen biner pertama yang terbentuk adalah $k = (1110)_2$ yang kemudian dikodekan menjadi nilai desimalnya sebesar $k = (14)_{10}$. Adapun 4 gen biner terakhir terbentuk vektor biner [0 0 1 0] yang artinya hanya variabel ketiga yang ikut serta dalam proses imputasi, sedangkan variabel lainnya tidak diikutsertakan.

Kemudian, pada simulasi 3, *hybrid* KNNI-GA dengan seleksi variabel memberikan nilai k optimum sebesar 6 dengan kombinasi variabel yang terbentuk [0 1 1 0 1 0 0 0 0]. Artinya bahwa 4 gen biner pertama yang terbentuk adalah $k = (0110)_2$ yang kemudian dikodekan menjadi nilai desimalnya sebesar $k = (6)_{10}$. Adapun 5 gen biner terakhir terbentuk vektor biner [1 0 0 0 0] yang artinya hanya variabel pertama yang ikut serta dalam proses imputasi, sedangkan variabel lainnya tidak diikutsertakan. Proses optimasi dari GA menunjukkan bahwa ketepatan estimator pada masing-masing simulasi meningkat. Hal ini dibuktikan dari nilai RMSE-nya yang lebih kecil dibandingkan hasil KNNI sebelum dilakukan optimasi.

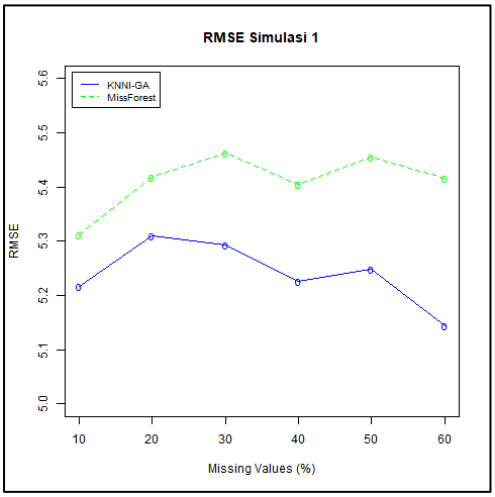


Simulasi 1 Simulasi 2 Simulasi 3

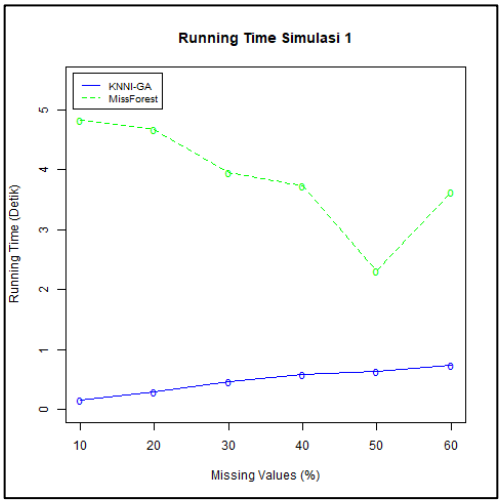
Gambar 2. Hasil Kinerja Pemilihan Nilai k Optimum Oleh GA

Simulasi Pada Analisis Data Bangkitan

Data yang digunakan untuk simulasi pertama adalah data bangkitan. Data tersebut dibangkitkan dengan *software* Rstudio sebanyak empat variabel. Adapun variabel imputasi dibangkitan dengan mengikuti distribusi normal, dan tiga variabel lainnya yang merupakan variabel prediktor mengikuti distribusi binomial, normal, dan *uniform*. Data bangkitan tersebut sejumlah 1000 observasi.



Gambar 3. RMSE Simulasi 1



Gambar 4. *Running Time* Simulasi 1

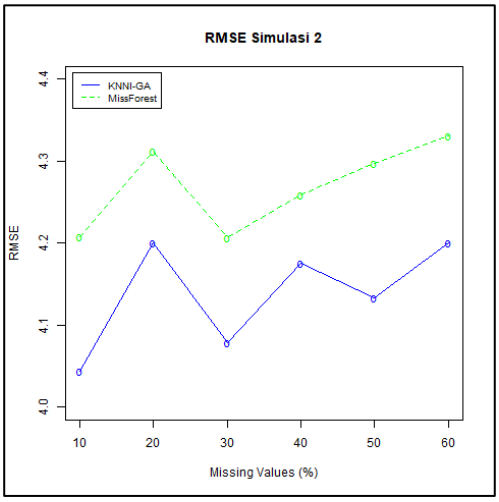
Gambar 3 menunjukkan kinerja imputasi yang dihasilkan pada simulasi 1. Hasil tersebut menunjukkan bahwa metode *hybrid* KNNI-GA menghasilkan ketepatan estimator yang lebih baik daripada metode *missForest*. Hal ini ditunjukkan dari nilai RMSE yang lebih rendah pada masing-masing tingkat *missing values*. Selain itu, pada setiap peningkatan persentase *missing values* pada dataset, menyebabkan akurasi dari *hybrid* KNNI-GA dan *missForest* semakin baik. Hal ini dilihat dari nilai RMSE yang semakin menurun seiring bertambahnya persentase *missing values*. Sementara itu, Gambar 4 menunjukkan performa komputasi yang dihasilkan pada simulasi 1. Hasilnya, metode *hybrid* KNNI-GA lebih baik daripada *missForest* dari sisi akurasi waktu. Hal ini ditandai melalui nilai *running time hybrid* KNNI-GA yang lebih kecil daripada *missForest*.

Hasil simulasi 1 ini menunjukkan bahwa semakin bertambahnya tingkat *missing* maka menyebabkan *running time* pada metode *hybrid* KNNI-GA semakin meningkat. Hal tersebut karena metode *hybrid* KNNI-GA tidak membutuhkan pembentukan model, akan tetapi hanya melakukan penyusunan observasi dan mengisi nilai yang *missing* dengan nilai k tetangga terdekatnya. Kemudian pada metode *missForest*, performa komputasinya menunjukkan bahwa semakin bertambahnya tingkat *missing values*, maka waktu untuk melakukan proses imputasi seiring menurun. Hal tersebut menunjukkan bahwa pembentukan model pada *missForest* menghabiskan waktu yang lebih banyak dibandingkan dengan imputasi pada nilai yang *missing*. Dengan demikian pada simulasi 1, *hybrid* KNNI-GA pada data bangkitan menghasilkan ketepatan estimator dan performa komputasi yang lebih baik daripada *missForest.*

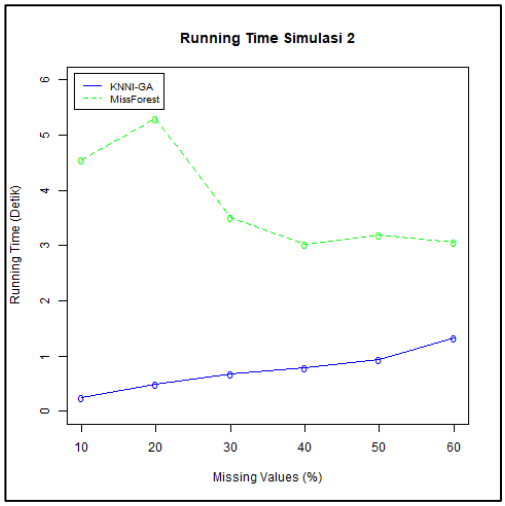
Simulasi Pada Analisis *Data Hepatitis C Virus (HCV) for Egyptian Patients*

Data yang digunakan untuk simulasi kedua adalah data *Hepatitis C Virus for Egyptian Patients* dengan unit observasi sebanyak 1385. Adapun variabel yang digunakan pada data ini dalam proses imputasi adalah *body mass index* (BMI), *age, gender, headache,* dan *hemoglobin.* Adapun variabel BMI digunakan sebagai variabel imputasi, sedangkan yang lainnya sebagai variabel prediktor. Adapun simulasi kedua ini dilakukan pada data yang tidak berdistribusi normal. Hal ini ditandai dari hasil uji *Kolmogorov-Smirnov* yang menunjukkan bahwa pada variabel imputasi dari data *Hepatitis C Virus for Egyptian Patients* tidak memenuhi asumsi kenormalan *(p-value = 3,935-12)*.

Gambar 5 menunjukkan kinerja imputasi yang dihasilkan pada simulasi 2. Hasil tersebut menunjukkan bahwa metode *hybrid* KNNI-GA menghasilkan ketepatan estimator yang lebih baik daripada metode *missForest*. Hal ini ditunjukkan dari nilai RMSE yang lebih rendah pada masing-masing tingkat *missing values*. Adapun nilai RMSE dari *hybrid* KNNI-GA dan *missForest* cenderung mengalami fluktuatif seiring bertambahnya persentase tingkat *missing values*. Hal tersebut berarti nilai akurasi dari RMSE dari kedua metode tersebut naik turun seiring bertambahnya tingkat *missing values*. Sementara itu, Gambar 6 menunjukkan performa komputasi yang dihasilkan pada simulasi 2. Hasilnya tidak jauh beda dengan simulasi 1, metode *hybrid* KNNI-GA lebih baik daripada *missForest* dari sisi akurasi waktu. Hal ini ditandai melalui nilai *running time hybrid* KNNI-GA lebih kecil daripada *missForest*.



Gambar 5. RMSE Simulasi 2

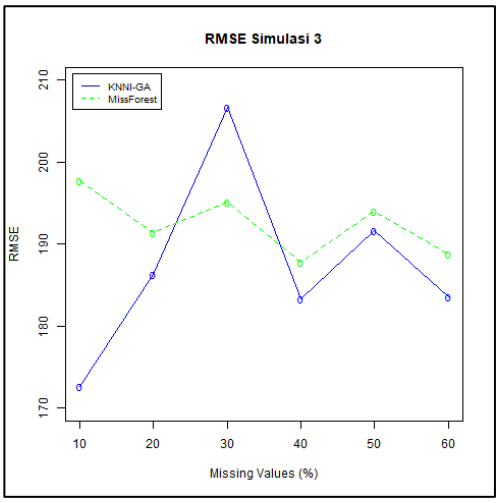


Gambar 6. *Running Time* Simulasi 2

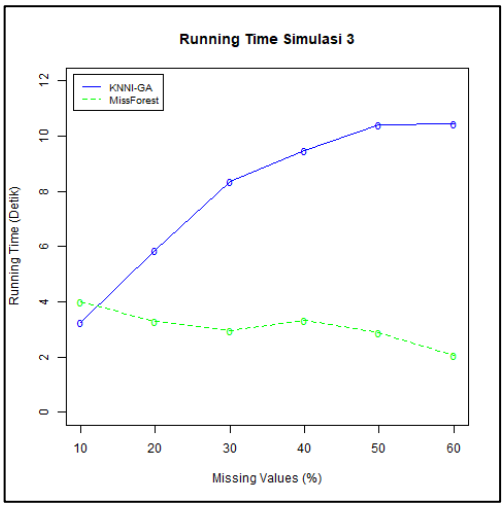
Selain itu, hasil simulasi 2 ini menunjukkan bahwa semakin bertambahnya tingkat *missing values* maka menyebabkan *running time* pada metode *hybrid* KNNI-GA semakin meningkat. Hal tersebut dikarenakan metode *hybrid* KNNI-GA tidak membutuhkan pembentukan model, akan tetapi hanya melakukan penyusunan observasi dan mengisi nilai yang *missing* dengan nilai k tetangga terdekatnya. Pun juga tidak jauh beda dengan simulasi 1, seiring bertambahnya persentase tingkat *missing values*, waktu melakukan imputasi pada metode *missForest* semakin menurun. Hal ini dikarenakan metode *missForest* lebih banyak menghabiskan waktu dalam pembentukan model daripada melakukan imputasi pada nilai yang *missing.* Dengan demikian pada simulasi 2, *hybrid* KNNI-GA pada data *Hepatitis C Virus for Egyptian Patients* menghasilkan ketepatan estimator dan performa komputasi yang lebih baik daripada *missForest.*

Simulasi Pada Analisis Data PKL Polstat STIS Tahun 2022

Data yang digunakan untuk simulasi ketiga adalah data PKL Polstat STIS tahun 2022 dengan unit observasi sebanyak 5269. Adapun variabel yang digunakan pada data ini dalam proses imputasi adalah lama waktu wawancara, penerimaan awal responden, jenis kelamin responden, ijazah terakhir responden, kemampuan berbahasa Indonesia responden, dan situasi lingkungan saat wawancara. Adapun variabel lama waktu wawancara digunakan sebagai variabel imputasi, sedangkan yang lainnya sebagai variabel prediktor. Adapun simulasi ketiga ini dilakukan pada data yang tidak berdistribusi normal. Hal ini ditandai dari hasil uji *Kolmogorov-Smirnov* yang menunjukkan bahwa pada variabel imputasi dari data PKL Polstat STIS tahun 2022 tidak memenuhi asumsi kenormalan *(p-value = 2,2e-16)*.



Gambar 7. RMSE Simulasi 3



Gambar 8. *Running Time* Simulasi 3

Gambar 7 menunjukkan hasil kinerja imputasi yang dihasilkan pada simulasi 3. Hasil tersebut menunjukkan bahwa metode *hybrid* KNNI-GA menghasilkan ketepatan estimator yang lebih baik daripada metode *missForest*. Akan tetapi, sedikit berbeda dengan simulasi 1 dan simulasi 2, pada tingkat *missing* 30 persen, ketepatan estimator dari *missForest* lebih baik daripada *hybrid* KNNI-GA. Hal tersebut ditunjukkan pada nilai RMSE *missForest* yang lebih rendah daripada *hybrid* KNNI-GA pada kondisi tingkat *missing* 30 persen. Sementara itu, kinerja dari *missForest* cenderung fluktuatif seiring bertambahnya tingkat *missing values,* yang terlihat pada nilai RMSE. Sementara itu, Gambar 8 menunjukkan performa komputasi yang dihasilkan simulasi 3. Hasilnya berbeda dengan simulasi 1 maupun simulasi 2, *missForest* menghasilkan performa komputasi yang lebih baik dari *hybrid* KNNI-GA. Hasil tersebut ditandai dari nilai *running time* dari *missForest* lebih rendah daripada *hybrid* KNNI-GA.

Pada *hybrid* KNNI-GA, seiring bertambahnya tingkat *missing* maka menyebabkan waktu melakukan proses imputasi semakin bertambah. Hal tersebut karena *hybrid* KNNI-GA hanya melakukan penyusunan observasi dan mengisi nilai yang *missing* dengan nilai k tetangga terdekat tanpa pembentukan model. Selain itu, pada *missForest*, seiring bertambahnya tingkat *missing* maka proses melakukan imputasi semakin menurun, hal ini karena *missForest* lebih banyak menghabiskan waktu dalam pembentukan model daripada melakukan imputasi pada nilai yang *missing*. Dengan demikian pada simulasi 3, menunjukkan bahwa *hybrid* KNNI-GA pada data PKL menghasilkan ketepatan estimator yang lebih baik daripada *missForest*, namun pada suatu kondisi tertentu *missForest* dapat lebih baik daripada *hybrid* KNNI-GA. Sementara itu juga, performa komputasi dari *missForest* lebih baik daripada *hybrid* KNNI-GA pada simulasi 3 ini.

Perbandingan Metode *Hybrid* KNNI-GA dan *MissForest*

Pada hasil ketiga simulasi yang didapatkan sebelumnya, dapat dilakukan perbandingan hasil imputasi pada *hybrid* KNNI-GA dan *missForest*. Dari ketiga simulasi tersebut menunjukkan bahwa *hybrid* KNNI-GA menghasilkan ketepatan estimator yang lebih baik dibandingkan *missForest*. Selain itu, pada setiap bertambahnya tingkat *missing values*, hasil kinerja imputasi semakin baik. Hal tersebut dilihat dari nilai RMSE-nya yang lebih kecil dan menurun. Adapun hasil imputasi juga semakin membaik seiring bertambahnya observasi pada data. Hal ini dikarenakan banyak observasi yang mirip dengan karakteristik data yang mengandung *missing values* sehingga donor yang digunakan mendekati karakteristik data tersebut, serta pembentukan model nya menjadi lebih baik. Hasil kinerja imputasi pada data bangkitan menunjukkan hasil yang lebih baik dibandingkan data *Hepatitis C Virus for Egyptian Patients* dan data PKL. Hal tersebut ditunjukkan oleh hasil RMSE pada masing-masing metode untuk setiap simulasi yang dilakukan.

Akan tetapi, performa komputasi tidak selalu berbanding lurus dengan ketepatan estimator-nya. Satu dari tiga simulasi tersebut menunjukkan bahwa performa komputasi dari *missForest* lebih baik dibandingkan *hybrid* KNNI-GA. Pada metode *missForest* menunjukkan bahwa performa komputasi yang semakin pendek seiring dengan bertambahnya jumlah data yang digunakan. Hal tersebut ditunjukkan pada penurunan *running time* untuk simulasi pada data bangkitan, data *Hepatitis C Virus (HCV) for Egyptian Patients*, dan data PKL. Sedangkan pada *hybrid* KNNI-GA menunjukkan performa komputasi yang semakin lama seiring bertambahnya jumlah data yang digunakan. Hal tersebut ditunjukkan pada peningkatan *running time* untuk

simulasi pada data bangkitan, data *Hepatitis C Virus for Egyptian Patients*, dan data PKL. Berikut ringkasan hasil rata-rata imputasi dari ketiga simulasi yang dilakukan.

Tabel 1. Ringkasan Hasil Imputasi

Metode	Simulasi	RMSE	*Running Time* (detik)
Hybrid KNNI-GA	Data Bangkitan	5,239	0,478
	Data HCV	4,139	0,743
	Data PKL	187,367	7,968
MissForest	Data Bangkitan	5,412	3,853
	Data HCV	4,269	3,777
	Data PKL	192,482	3,080

Pada Tabel 1 menunjukkan bahwa tidak selalu ketepatan estimator dan performa komputasi berbanding lurus. Imputasi yang dilakukan pada data yang berdistribusi normal menghasilkan ketepatan estimator dan performa komputasi yang lebih baik. Selain itu, data mentah (*raw data*) yang dalam tahap *pre-processing* biasanya tidak memenuhi asumsi kenormalan sehingga jika ingin meningkatkan ketepatan estimator, dapat melakukan transformasi data dengan risiko terdapat penambahan analisis yang dilakukan. Fadillah dan Muchlisoh (2020) menyatakan bahwa cara lain untuk meningkatkan ketepatan estimator adalah melakukan imputasi pada data dengan jumlah observasi yang besar dan diasumsikan normal dengan *central limit theorem* (CLT). Namun, cara tersebut dapat menurunkan performa komputasi. Hal tersebut ditunjukkan pada proses imputasi yang membutuhkan waktu yang lama. Kendati demikian, *missForest* menghasilkan performa komputasi yang stabil pada setiap penambahan jumlah observasi pada dataset. Oleh karena itu, jika melakukan imputasi pada beberapa variabel, *hybrid* KNNI-GA lebih baik digunakan dibandingkan *missForest*. Selain itu, jika melakukan imputasi pada berbagai macam variabel, terlebih bersifat mikro, *missForest* lebih baik digunakan daripada *hybrid* KNNI-GA.

KESIMPULAN

Berdasarkan simulasi, hasil serta pembahasan yang telah dilakukan, maka kesimpulan dari penelitian ini adalahmMetode *hybrid* KNNI-GA menghasilkan ketepatan estimator yang lebih baik baik dibandingkan *missForest*. Kedua metode tersebut menghasilkan ketepatan estimator yang baik pada data berdistribusi normal dan ketepatan estimatornya meningkat walaupun jumlah observasinya bertambah. Sedangkan jika dilihat dari performa komputasi, *missForest* menghasilkan performa yang lebih stabil dibandingkan *hybrid* KNNI-GA. Selain itu, pada setiap penambahan jumlah observasi pada dataset, maka performa komputasi *hybrid* KNNI-GA semakin bertambah dan *running time* menjadi lebih lama. Sedangkan performa komputasi *missForest* cenderung lebih stabil dibandingkan *hybrid* KNNI-GA. Selain itu, performa komputasi *hybrid* KNNI-GA dipengaruhi oleh tingkat *missing values*, sedangkan performa komputasi *missForest* tidak terlalu dipengaruhi oleh tingkat *missing values*.

DAFTAR PUSTAKA

Barry, L. C., Kasl, S. V., & Prigerson, H. G. (2002). Psychiatric disorders among bereaved persons: The role of perceived circumstances of death and preparedness for death. *American Journal of Geriatric Psychiatry*, *10*(4), 447–457. https://doi.org/10.1097/00019442-200207000-00011

Batista, G. E. A. P. A., & Monard, M. C. (2002). A study of k-nearest neighbour as an imputation method. *Frontiers in Artificial Intelligence and Applications*, *87*(May 2014), 251–260.

Biemer, P. B., & Lyberg, L. E. (2003). *Introduction to Survey Quality*. Wiley-Interscience.

Cohen, M. . (1996). A new approach to imputation. *American Statistical Association Proceding of the Section on Survey Research Methods*, 293–298.

Darcy, A., & Hoeta, M. (2021). The Overlooked Transition From Undergraduate To Postgraduate Study. *University of Otago*.

Fadillah, I. J., & Muchlisoh, S. (2020). Perbandingan Metode Hot-Deck Imputation Dan Metode Knni Dalam Mengatasi Missing Values. *Seminar Nasional Official Statistics*, *2019*(1), 275–285. https://doi.org/10.34123/semnasoffstat.v2019i1.101

Fadillah, I. J., & Puspita, C. D. (2022). Application of The Sequential Hot-deck Imputation Method for Identification of Indonesian Standard Classification of Business Fields (KBLI). *Proceedings of The*

International Conference on Data Science and Official Statistics, *2022*(1), 734–741. https://doi.org/10.34123/icdsos.v2021i1.70

Hayatin, A. I. N. (2012). IMPUTASI MISSING DATA MENGGUNAKAN METODE K - NEAREST NEIGHBOUR Pendahuluan Metodologi. *Universitas Muhammadiyah Malang*.

Irhamah, U. M. (2012). IMPUTASI MISSING DATA DENGAN K-NEAREST NEIGHBOR DAN ALGORITMA GENETIKA. *Institut Teknologi Sepuluh September*, *19*(March), 7–8. http://dx.doi.org/10.1007/978-981-10-7826-2_9

Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine*, *50*(2), 105–115. https://doi.org/10.1016/j.artmed.2010.05.002

Kaiser, J. (2014). Dealing with Missing Values in Data. *Journal of Systems Integration*, 258–263. https://doi.org/10.1109/ICET.2008.4777511

Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, *64*(5), 402–406. https://doi.org/10.4097/kjae.2013.64.5.402

Kreuter, F., Couper, M., & Lyberg, L. (2010). The use of paradata to monitor and manage survey data collection. *Measurement*, 282–296.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236. https://doi.org/10.1002/acp.2350050305

Lau, C. Q., Baker, M., Fiore, A., Greene, D., Lieskovsky, M., Matu, K., & Peytcheva, E. (2017). Bystanders, noise, and distractions in face-to-face surveys in Africa and Latin America. *International Journal of Social Research Methodology*, *20*(5), 469–483. https://doi.org/10.1080/13645579.2016.1208959

Little, R. J. A., & Rubin, D. B. (1988). Statistical Analysis with Missing Data. In *Population (French Edition)* (Vol. 43, Issue 6). https://doi.org/10.2307/1533221

Malarvizhi, M. . (2012). K-NN Classifier Performs Better Than K-Means Clustering in Missing Value Imputation. *IOSR Journal of Computer Engineering*, *6*(5), 12–15. https://doi.org/10.9790/0661-0651215

Misztal, M. (2013). SOME REMARKS ON THE DATA IMPUTATION USING "MISSFOREST" METHOD. *ACTA UNIVERS I TATIS LODZIE NSIS*, 169–179.

Pazanudin, A. F. (2017). Kajian Missing Data : Perbandingan Metode Hot-Deck dan MissForest dalam Imputasi Data. *Sekolah Tinggi Ilmu Statistik*.

Penone;, C., Davidson, A., Shoemaker, K. T., Marco, M. Di, Rondinini, C., Brooks, T., Young, B. E., Graham, C. H., & Costa, G. C. (2014). Imputation of missing data in life-history trait datasets which approach performs the.pdf. *Methods in Ecology AndEvolution*, *5*, 961–970. https://doi.org/10.1111/2041-210X.12232

Sallaby, A., (International, A. A.-T. I., & 2021, undefined. (2021). Analysis of Missing Value Imputation Application with K-Nearest Neighbor (K-NN) Algorithm in Dataset. *Ejurnal.Stmik-Budidarma.Ac.Id*, *5*(2), 141–144. https://doi.org/10.30865/ijics.v5i2.3185

Sartika, E. (2018). Analisis metode k nearest neighbor imputation (knni) untuk mengatasi data hilang pada estimasi data survey. *Tedc*, *12*(3), 219–227.

Sidi, F., Shariat Panahy, P. H., Affendey, L. S., Jabar, M. A., Ibrahim, H., & Mustapha, A. (2012). Data quality: A survey of data quality dimensions. *Proceedings - 2012 International Conference on Information Retrieval and Knowledge Management, CAMP'12*, *August*, 300–304. https://doi.org/10.1109/InfRKM.2012.6204995

Silva-Ramírez, E. L., Pino-Mejías, R., López-Coello, M., & Cubiles-de-la-Vega, M. D. (2011). Missing value imputation on missing completely at random data using multilayer perceptrons. *Neural Networks*, *24*(1), 121–129. https://doi.org/10.1016/j.neunet.2010.09.008

Stekhoven, D. J., & Bühlmann, P. (2012). Missforest-Non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597

T.Longford, N. (2005). *Missing data and small-area estimation: Modern analytical equipment for the survey statistician*. Springer.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, *17*(6), 520–525. https://doi.org/10.1093/bioinformatics/17.6.520

Zhang, S., Gong, L., Zeng, Q., Li, W., Xiao, F., & Lei, J. (2021). Imputation of GPS coordinate time series using missforest. *Remote Sensing*, *13*(12), 1–17. https://doi.org/10.3390/rs13122312