IMPUTASI *MISSING DATA* DENGAN *K-NEAREST NEIGHBOR* DANALGORITMA GENETIKA

Ucik Mawarsari
Badan Pusat Statistik
Jl. Dr. Sutomo 6-8 Jakarta, ucik@bps.go.id

ABSTRAK

Permasalahan yang sering muncul pada suatu data adalah adanya ketidaklengkapan data pada suatu variabel atau sering disebut dengan *missing data*. Untuk menangani *missing data*, salah satu cara yang dapat digunakan adalah dengan melakukan imputasi. Salah satu metode imputasi yang cukup sederhana adalah *k-Nearest Neighbor* (kNN). Namun salah satu kelemahan metode kNN adalah permasalahan dalam pemilihan nilai *k* karena pemilihan nilai *k* yang kurang tepat dapat menurunkan kinerja kNN. Penelitian ini mengkaji metode gabungan *k-Nearest Neighbor* dan Algoritma Genetika (kNN-GA) yang digunakan untuk mengestimasi nilai *missing*. Algoritma genetika digunakan untuk melakukan optimasi nilai *k* dan untuk melakukan seleksi variabel yang digunakan untuk estimasi nilai *missing*sehingga dapat menghasilkan nilai estimasi yang baik dengan *Root Mean Square Error* (RMSE) minimum. Hasil yang diperoleh menunjukkan bahwa metode imputasi kNN-GA dapat membantu dalam memperoleh nilai *k* optimum dan dapat melakukan seleksi variabel dengan baik sehingga dapat meningkatkan akurasi kNN.

Kata Kunci : Algoritma Genetika, Imputasi, kNN, kNN-GA.

ABSTRACT

Problems often arise in a data is incompleteness on a variable of the data or often called with missing data. Imputation is a method that can be used to handle missing data. One of simple imputation methods are k-Nearest Neighbor (kNN). However, the weakness of kNN method is the problem in selecting the value of *k*, because a mistake in selecting the value of *k* can degrade performance of kNN. This study examines the combined method k-Nearest Neighbor and Genetic Algorithm (kNN-GA) that are used to estimate missing values. Genetic algorithm is used to optimize the value of *k* and for selecting the variables used to estimate missing values in order to produce a good estimation with minimum Root Mean Square Error (RMSE). The results show that the imputation method of kNN-GA can assist in obtaining the optimum value of *k* and can do a good selection of variables so it can improve the accuracy of kNN.

Keywords : Genetic Algorithm, Imputation, kNN, kNN-GA.

Pendahuluan

Permasalahan yang sering muncul pada suatu data adalah adanya ketidaklengkapan data pada suatu variabel atau sering disebut dengan *missing data*. Adanya *missing data* menyebabkan kesulitan dalam melakukan analisis terhadap data tersebut karena analisis statistik hanya dapat diterapkan pada data yang lengkap.

Untuk menangani *missing data*, salah satu metode yang dapat digunakan adalah dengan mengestimasi nilai *missing* dengan suatu nilai tertentu yang

dianggap sesuai, atau sering disebut dengan imputasi. Salah satu metode imputasi adalah metode *k*-N*earest Neighbor*(kNN). "Metode ini merupakan metode yang sederhana dan fleksibel karena dapat digunakan baik pada variabel dengan data kontinu maupun data diskrit (Batista dan Monard 2003)". "Pada penelitian yang dilakukan oleh Batista dan Monard (2003) yang menganalisa penggunaan kNN sebagai sebuah metode imputasi, dengan melakukan simulasi *missing data* sebanyak 10 hingga 60 persen dari total data menunjukkan bahwa metode imputasi kNN memberikan hasil yang sangat baik, bahkan pada saat data memiliki jumlah *missing* yang besar". "Jerez dan Molina (2010) juga membandingkan metode imputasi yang berbasiskan teknik statistik, yaitu *mean*, *hot deck*, *multiple imputation* dengan metode imputasi berbasiskan teknik *machine learning*, yaitu *Multilayer Perceptron* (MLP), *Self Organizing Map* (SOM), dan kNN pada data *breast cancer*. Hasilnya menunjukkan metode imputasi berbasis *machine learning* memiliki tingkat akurasi yang lebih tinggi daripada metode berbasis teknik statistik. Dan dari ketiga metode imputasi berbasis *machine learning* yang digunakan, metode kNN memberikan nilai prediksi yang paling baik karena menghasilkan nilai *Mean Square Error* yang paling kecil".

Dalam beberapa penelitian, sebuah metode sering dikombinasikan dengan teknik optimasi untuk memperoleh hasil yang lebih optimal. "Salah satunya adalah penelitian Patil dan Bichkar (2010) yang mengintegrasikan *decision tree learning* dan algoritma genetika untuk menyusun *decision tree* yang optimal untuk mengestimasi nilai *missing*. Hasil penelitiannya menunjukkan bahwa algoritma yang diusulkan dapat meningkatkan akurasi estimasi nilai *missing*". "Siedlecky dan Sklansky (1989) juga mengkombinasikan *K-Nearest Neighbor* dan Algoritma Genetika, dimana algoritma genetika digunakan untuk melakukan seleksi variabel yang digunakan pada klasifikasi. Hasil penelitiannya menunjukkan bahwa metode tersebut dapat meningkatkan ketepatan klasifikasi. Selain itu, dalam penelitiannya juga menyatakan bahwa algoritma genetika merupakan alat yang *powerfull* dalam melakukan seleksi variabel terutama ketika demensi variabel yang digunakan cukup besar".

Salah satu kelemahan pada metode kNN adalah permasalahan dalam pemilihan nilai *k* karena pemilihan nilai *k* yang kurang tepat dapat menurunkan kinerja kNN. Penelitian ini bertujuan untuk mengatasi permasalahan *missing*

data dengan teknik imputasi menggunakan gabungan kNN dan Algoritma Genetika (kNN-GA) dimana Algoritma Genetika digunakan untuk optimasi nilai *k* pada kNN dan untuk melakukan seleksi variabel yang digunakan untuk mengestimasi nilai *missing*.

Metode Penelitian

k-Nearest Neighbor (kNN)

Algoritma *k-Nearest Neighbor* telah banyak digunakan untuk diiplementasikan dalam imputasi *missing data* terutama dalam dataset dengan *missing data* pada lebih dari satu variabel. Algoritma kNN menggunakan observasi yang mirip/serupa dengan observasi yang memiliki nilai *missing*. Sebuah observasi yang memuat satu atau lebih nilai *missing* yang akan diimputasi disebut sebagai observasi target. Berdasarkan pendekatan ini, ukuran jarak dihitung antara observasi target dengan tiap-tiap observasi lainnya. Jika *k* merepresentasikan jumlah observasi yang mirip/serupa dengan observasi target**,** maka kemudian dipilih *k* observasi yang memiliki jarak minimum dari observasi target.

Algoritma imputasi kNN adalah sebagai berikut :

1. Menentukan nilai *k*, yaitu jumlah observasi terdekat yang diinginkan.
2. Menghitung jarak euclidian antara observasi target dengan observasi yang tidak memuat nilai *missing*, dengan formula:

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{j=1}^{s} (x_j - y_j)^2} \qquad (1)$$

dimana:

$\boldsymbol{x}$ = vektor observasi target dengan variabel sebanyak *s* variabel, $\boldsymbol{x} = [x_1, x_2, \dots, x_s]^T$

$\boldsymbol{y}$ = vektor observasi yang tidak memuat nilai *missing* dengan variabel sebanyak *s* variabel, $\boldsymbol{y} = [y_1, y_2, \dots, y_s]^T$

$d(\boldsymbol{x}, \boldsymbol{y})$ = jarak antara $\boldsymbol{x}$ dan $\boldsymbol{y}$

x_j =nilai variabel ke-*j* pada $\boldsymbol{x}$

y_j =nilai variabel ke-*j* pada $\boldsymbol{y}$

j = 1,2, ..., *s*.

3. Mencari *k* observasi yang memiliki nilai $d(\boldsymbol{x}, \boldsymbol{y})$ minimum.
4. Melakukan imputasi *missing data* dengan menggunakan prosedur *weighted mean imputation*, dengan formula:

$$\hat{x}_j = \frac{1}{W} \sum_{k=1}^{K} w_k y_{kj} \qquad (2)$$

dimana:

$\hat{x}_j$ =nilai imputasi

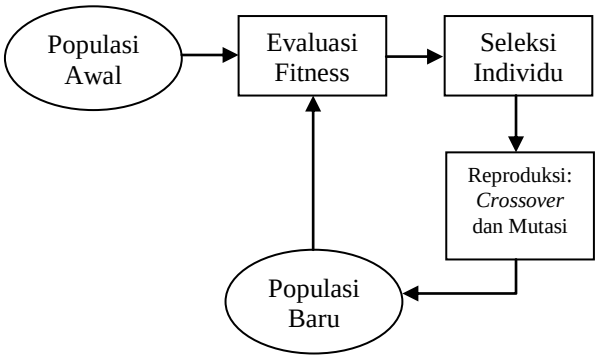
y_{kj} =nilai variabel ke-*j* pada observasi ke-*k*, k = 1,2, …, *K*

w_k=bobot observasi ke-*k*, $w_k = 1/d(\boldsymbol{x}, \boldsymbol{y}_k)$

$W = \sum_{k=1}^{K} w_k$

Algoritma Genetika

Algoritma genetika merupakan algoritma untuk memecahkan suatu pencarian nilai dalam masalah optimasi dengan meniru proses genetik pada makhluk hidup. Algoritma genetika bekerja dengan mempertahankan sebuah populasi yang terdiri dari individu-individu, dimana tiap individu merepresentasikan sebuah solusi pada permasalahan yang dihadapi. Individu dikodekan dalam bentuk kromosom yang terdiri dari komponen genetik terkecil, yaitu gen. Dari individu yang ada, dihitung nilai *fitness* yang digunakan sebagai kriteria solusi terbaik.



Gambar 1. Siklus Algoritma Genetika

Individu-individu yang lolos proses seleksi kemudian melakukan reproduksi dengan perkawinan silang (*crossover*) dan mutasi dengan probabilitas *crossover*(P_c) dan probabilitas mutasi (P_m) sehingga menghasilkan keturunan. Keturunan-keturunan ini kemudian dievaluasi untuk membentuk populasi baru yang memiliki kriteria yang lebih baik. "Setelah beberapa generasi terbentuk, algoritma akan konvergen pada individu terbaik yang diharapkan merepresentasikan solusi optimal dari permasalahan yang dihadapi (Gen dan Cheng 1999)".

kNN-GA

Metode kNN-GA merupakan metode gabungan antara kNN dan Algoritma Genetika. Algoritma imputasi *missing data* dengan metode kNN-GAadalah sebagai berikut:

1. Menentukan inisial populasi, meliputi jumlah gen dalam kromosom dan jumlah kromosom dalam individu.
2. Melakukan proses algoritma kNN untuk setiap kromosom dalam populasi. Setiap kromosom dikodekan menjadi nilai *k* dan representasi variabel yang digunakan oleh kNN dalam melakukan imputasi. Kemudian menghitung RMSE dengan formula:

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

dimana:

n = jumlah observasi

y_i = nilai aktual pada observasi ke-*i*

$\hat{y}_i$ = nilai estimasi pada observasi ke-*i*

3. Menghitung nilai *fitness* dari tiap kromosom dalam populasi dengan formula:

$$f = \frac{1}{(RMSE + h)} \qquad (4)$$

dimana:

h=nilai yang kecil ($0 \leq h \leq 1$)

4. Memilih kromosom berdasarkan nilai *fitness*.
5. Melakukan perkawinan silang (*crossover*) dan mutasi untuk mendapatkan keturunan (*offspring*)
6. Melakukan *elitism* dan *replacement* sehingga diperoleh populasi baru.
7. Kembali ke tahap nomor 2 hingga kriteria yang ditentukan terpenuhi.

Metodologi

Data yang digunakan pada penelitian ini adalah data vertebral column, seeds, dan computer hardware yang diambil dari UCI *Machine Learning Repository* yang merupakan data lengkap atau tidak terdapat nilai *missing*. Jumlah data dan jumlah variabel dari masing-masing dataset dapat dilihat pada Tabel 1.

Tabel 1. Jumlah Data dan Jumlah Variabel pada Dataset

Dataset	Jumlah data	Jumlah variabel
Vertebral column	310	6
Seeds	210	7
Computer hardware	209	8

Penelitian ini menggunakan data lengkap karena data aktual dan data hasil imputasi akan dibandingkan untuk mengukur kinerja metode kNN-GA. Langkah berikutnya adalah menghilangkan secara random beberapa nilai pada salah satu variabel pada masing-masing dataset untuk mendapatkan nilai *missing* sebanyak 10% dan 20% dari jumlah data. Kemudian dilakukan imputasi *missing data* dengan kNN-GA. Untuk mencari nilai *k* optimum pada metode kNN-GA, nilai *k* direpresentasikan dalam 4 gen biner. Sebagai contoh kromosom individu yang terbentuk untuk nilai $k = 10$ adalah:

1	0	1	0

Gambar 2. Kromosom Individu untuk Nilai *k*

String biner selanjutnya merepresentasikan kromosom untuk seleksi variabel. Misalkan terdapat 5 variabel dalam dataset, maka sebuah vektor biner berukuran 5×1 yang berunsurkan nilai 1 dan 0 diartikan sebagai vektor indikator yang menunjukkan kode seleksi variabel, kode 1 menunjukkan variabel yang terpilih. Sebagai contoh kromosom individu yang terbentuk adalah:

1	1	0	1	1

Gambar 3. Kromosom Individu untuk Variabel

String tersebut akan dikodekan ke dalam metode kNN menjadi, variabel yang terpilih dalam proses imputasi dengan kNN adalah variabel ke-1, ke-2, ke-4, dan ke-5. Variabel ke-3 tidak terpilih karena berkode 0.

Langkah-langkah metode kNN-GA selanjutnya adalah sebagai berikut:

1. Membangkitkan secara random sebanyak 50 individu dalam populasi.
2. Menghitung nilai *fitness* untuk setiap individu dalam populasi sesuai dengan persamaan (4).
3. Membentuk individu baru dengan melakukan seleksi *roullete wheel*, *crossover* satu titik dengan probabilitas (P_c) sebesar 0,8 dan mutasi dengan probabilitas (P_m) sebesar 0,2. Kemudian melakukan *elitism* dan *replacement* sehingga diperoleh populasi baru.
4. Memilih individu terbaik dari populasi yang merupakan solusi terbaik setelah kriteria yang ditentukan terpenuhi, yaitu ketika mencapai generasi maksimum 50 generasi atau selisih nilai *fitness* terbaik dalam 5 generasi terakhir tidak lebih dari 1×10^{-8}.

Hasil dan Pembahasan

Sebelum melakukan proses imputasi, terlebih dahulu dilakukan eksplorasi data untuk mengetahui gambaran awal dari data yang digunakan. Pada penelitian ini, pola *missing data* yang digunakan adalah *missing data* univariat atau *missing data* hanya terjadi pada satu variabel, yaitu variabel yang menempati urutan terakhir pada masing-masing dataset. Statistik deskriptif dari variabel yang mengandung nilai *missing* dapat dilihat pada Tabel 2.

Tabel 2. Statistik Deskriptif dari Variabel yang mengandung Nilai *Missing*

Dataset	Urutan Variabel	Mean	Min	Maks
Vertebral column	Ke-6	26.30	-11.06	418.54
Seeds	Ke-7	5.41	4.52	6.55
Computer hardware	Ke-8	99.3	15.0	1238.0

Untuk melakukan analisa hasil seleksi variabel pada metode kNN-GA, maka perlu dilihat korelasi antar variabel pada masing-masing dataset untuk mengetahui apakah hasil seleksi variabel pada metode kNN-GA dipengaruhi oleh korelasi antar variabel atau tidak. Matriks korelasi dari masing-masing dataset adalah sebagai berikut:

Vertebral column

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & & & & & \\ 0.63 & 1 & & & & \\ 0.72 & 0.43 & 1 & & & \\ 0.82 & 0.06 & 0.59 & 1 & & \\ -0.25 & 0.03 & -0.08 & -0.34 & 1 & \\ 0.64 & 0.39 & 0.53 & 0.52 & -0.03 & 1 \end{bmatrix}$$

Dari matriks korelasi tersebut diketahui bahwa terdapat tiga variabel yang mempunyai korelasi kuat dengan variabel ke-6 atau variabel yang mengandung nilai *missing*, yaitu variabel ke-1 (ρ_{16} = 0.64), variabel ke-3 (ρ_{36} =

0.53), dan variabel ke-4 (ρ_{46} = 0.52). Sedangkan variabel ke-2 dan ke-5 tidak mempunyai korelasi yang kuat dengan variabel ke-6 karena memiliki nilai korelasi kurang dari 0.5.

Seeds

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & & & & & & \\ 0.99 & 1 & & & & & \\ 0.61 & 0.53 & 1 & & & & \\ 0.95 & 0.97 & 0.34 & 1 & & & \\ 0.97 & 0.94 & 0.76 & 0.86 & 1 & & \\ -0.23 & -0.22 & -0.33 & -0.17 & -0.26 & 1 & \\ 0.86 & 0.89 & 0.23 & 0.93 & 0.75 & -0.01 & 1 \end{bmatrix}$$

Pada data seeds, sebagian besar variabel memiliki korelasi yang kuat dengan variabel-ke-7 atau variabel yang mengandung nilai *missing*, yaitu variabel ke-1 (ρ_{17} = 0.86), variabel ke-2 (ρ_{27} = 0.86), variabel ke-4 (ρ_{47} = 0.93), dan variabel ke-5 (ρ_{57} = 0.75). Sedangkan variabel-ke-3 dan variabel ke-6 tidak mempunyai korelasi yang kuat dengan variabel ke-7.

Computer hardware

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & & & & & & & \\ -0.34 & 1 & & & & & & \\ -0.38 & 0.76 & 1 & & & & & \\ -0.32 & 0.54 & 0.54 & 1 & & & & \\ -0.30 & 0.52 & 0.56 & 0.58 & 1 & & & \\ -0.25 & 0.27 & 0.53 & 0.49 & 0.55 & 1 & & \\ -0.31 & 0.79 & 0.86 & 0.66 & 0.61 & 0.61 & 1 & \\ -0.29 & 0.82 & 0.90 & 0.65 & 0.61 & 0.59 & 0.97 & 1 \end{bmatrix}$$

Hampir semua variabel pada data computer hardware juga mempunyai korelasi yang kuat dengan variabel ke-8 atau variabel yang mengandung nilai *missing*, yaitu variabel ke-2 (ρ_{28} = 0.82), variabel ke-3 (ρ_{38} = 0.90), variabel ke-4 (ρ_{48} = 0.65), variabel ke-5 (ρ_{58} = 0.61), variabel ke-6 (ρ_{68} = 0.59), dan variabel ke-7(ρ_{78} = 0.97). Hanya variabel-ke-1 yang tidak mempunyai korelasi yang kuat dengan variabel ke-8 (ρ_{18} = -0.29).

Langkah selanjutnya adalah melakukan imputasi *missing data*. Hasil imputasi *missing data* dengan kNN-GA pada data vertebral column, seeds, dan computer hardware dapat dilihat pada Tabel 3.

Tabel 3. Hasil Imputasi *Missing Data* dengan metode kNN-GA

Dataset	% Mising Data	Seleksi Variabel	Nilai *k*	RMSE
Vertebral column	10%	0 1 1 0 1	2	13.762
	20%	1 1 0 0 1	9	15.220
Seeds	10%	0 0 0 1 1 0	3	0.130
	20%	0 1 1 1 1 0	15	0.157
Computer hardware	10%	0 1 1 1 1 1 0	2	4.389
	20%	0 1 1 1 0 0 0	10	10.598

Pada proporsi *missing data* sebesar 10%, nilai *k* optimum yang diperoleh pada data vertebral column adalah *k* = 2 dengan hasil seleksi variabel ke-2, ke-3 dan ke-5 yang terpilih dalam proses imputasi kNN-GA. Pada data seeds nilai *k* optimum yang diperoleh adalah *k* = 3 dengan hasil seleksi variabel ke-4 dan ke-5 yang terpilih dalam proses imputasi kNN-GA. Sedangkan pada data computer hardware nilai *k* optimumnya adalah *k* = 2 dengan hasil seleksi variabel ke-2, ke-3, ke-4, ke-5 dan ke-6.

Perbedaaan proporsi *missing data* sebesar 10% dan 20% memberikan hasil

yang cukup bervariasi pada masing-masing dataset. Hasil penelitian menunjukkan bahwa nilai *k* optimum yang diperoleh berbanding lurus dengan proporsi *missing data*, yang artinya semakin banyak jumlah *missing data* maka nilai *k* optimum yang dihasilkan juga semakin besar. Hasil seleksi variabel juga menunjukkan perbedaan antara proporsi *missing data* 10% dan 20%.

Pada proporsi *missing data* sebesar 20%, nilai *k* optimum yang diperoleh pada data vertebral column adalah $k = 9$ dengan hasil seleksi variabel ke-1, ke-2 dan ke-5 yang terpilih dalam proses imputasi kNN-GA. Pada data seeds nilai *k* optimum yang diperoleh adalah $k = 15$ dengan hasil seleksi variabel ke-2, ke-3, ke-4 dan ke-5 yang terpilih dalam proses imputasi kNN-GA. Sedangkan pada data computer hardware nilai *k* optimumnya adalah $k = 10$ dengan hasil seleksi variabel ke-2, ke-3, dan ke-4.

Berdasarkan matriks korelasi, variabel yang memiliki korelasi kuat dengan variabel ke-6 atau variabel yang mengandung nilai *missing* pada data vertebral column adalah variabel ke-1, ke-3, dan ke-4. Pada proporsi *missing data* 10%, hasil seleksi variabel adalah variabel ke-2, ke-3, dan ke-5. Jika dilihat nilai korelasinya, variabel ke-3 mempunyai korelasi kuat dengan variabel ke-6, akan tetapi variabel ke-2 dan ke-5 tidak mempunyai korelasi yang kuat dengan variabel ke-6. Selain variabel ke-3, variabel ke-1 dan ke-4 juga mempunyai korelasi yang kuat dengan variabel ke-6, akan tetapi tidak terpilih dalam hasil seleksi variabel dengan metode kNN-GA.

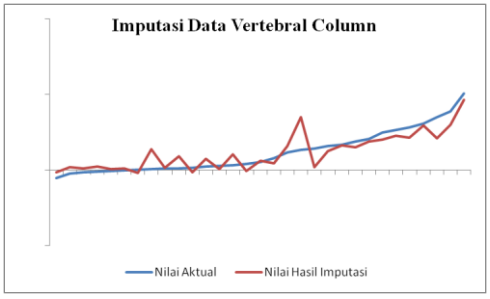
Sedangkan pada data seeds, hasil seleksi variabel dengan metode kNN-GA adalah variabel ke-4 dan ke-5. Jika dilihat dari nilai korelasinya, kedua variabel tersebut mempunyai korelasi yang kuat dengan variabel ke-7. Selain variabel ke-4 dan ke-5, variabel ke-1 dan ke-2 juga mempunyai korelasi yang kuat dengan variabel ke-7, akan tetapi tidak terpilih dalam hasil seleksi variabel dengan metode kNN-GA.

Pada data computer hardware, hasil seleksi variabel adalah variabel ke 2, ke-3, ke-4, ke-5, dan ke-6. Jika dilihat nilai korelasinya, semua variabel hasil seleksi tersebut mempunyai korelasi yang kuat dengan variabel ke-8, namun variabel ke-7 yang juga mempunyai korelasi yang kuat dengan variabel ke-8 tidak terpilih dalam hasil seleksi variabel dengan metode kNN-GA.

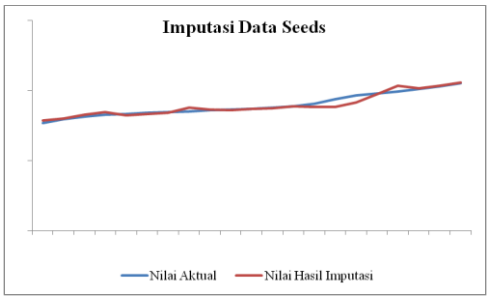
Hasil seleksi variabel pada proporsi *missing data* 20% juga menunjukkan hasil yang random jika dilihat hubungannya dengan nilai korelasi variabel. Hal ini menunjukkan bahwa

hasil seleksi variabel dengan metode kNN-GA tidak sepenuhnya dipengaruhi oleh besarnya nilai korelasi antar variabel dari dataset tersebut.

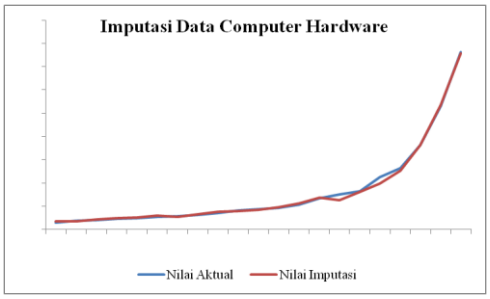
Nilai RMSE yang dihasilkan metode kNN-GA pada Tabel 3 menunjukkan nilai yang relatif kecil. Hal ini menjelaskan bahwa teknik optimasi Algoritma Genetika pada kNN-GA dapat meningkatkan kinerja metode kNN dengan melakukan seleksi variabel dan optimasi nilai *k* dalam melakukan imputasi *missing data*. Nilai *k* optimum dan seleksi variabel dapat dihasilkan secara otomatis tanpa harus melakukan uji coba secara manual satu persatu.



Gambar 4. Perbandingan Nilai Aktual dan Nilai Hasil Imputasi pada Data Vertebral Column



Gambar 5. Perbandingan Nilai Aktual dan Nilai Hasil Imputasi pada Data Seeds



Gambar6. Perbandingan Nilai Aktual dan Nilai Hasil Imputasi pada Data Computer Hardware

Gambar 4, Gambar 5, dan Gambar 6 menunjukkan perbandingan nilai aktual dan nilai hasil imputasi pada data vertebral column, seeds, dan computer hardware dengan proporsi *missing data* 10%. Dari gambar tersebut dapat dilihat bahwa secara umum metode kNN-GA dapat menghasilkan nilai imputasi yang mendekati nilai aktual.

Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, dapat diambil kesimpulan sebagai berikut:

1. Metode imputasi kNN-GA dapat membantu dalam memperoleh nilai *k* optimum. Hasil penelitian pada data vertebral column, seeds, dan computer hardware menunjukkan bahwa nilai *k* optimum yang diperoleh berbanding lurus dengan jumlah *missing data*.

2. Metode imputasi kNN-GA dapat melakukan seleksi variabel dengan baik sehingga dapat meningkatkan akurasi kNN. Hasil seleksi variabel yang diperoleh dengan metode kNN-GA tidak sepenuhnya dipengaruhi oleh besarnya nilai korelasi antar variabel pada dataset.

Pustaka

Analoui, M. dan Amiri, M.F., 2006, "Feature Reduction of Nearest Neigbor Classifier using Genetic Algorithm", *World Academy of Science, Engineering and Technology***17**, 36-39.

Batista G. dan Monard M.C., 2003, *A Study of K-Nearest Neighbour as an Imputation Method*, Working Paper, University Sao Paulo, Brazil.

Gen, M. dan Cheng, R., 1999, *Genetic Algorithm and Optimization Engineering*, John Wiley & Sons, Inc, Japan.

Jerez, J.M., dan Molina, I., 2010, "Missing Data Imputation Using Statistical And Machine Learning Methods In A Real Breast Cancer Problem", *Artificial Intelligence in Medicine***50**, 105-115.

Little, R.J., dan Rubin, D.B., 1987, *Statistical Analysis with Missing Data,* John Wiley & Sons, Inc, New York.

Patil, D.V. dan Bichkar, R.S., 2010, "Multiple Imputation of Missing Data with Genetic Algorithm based Techniques", *Evolutionary Computation for Optimization Techniques*, hal. 74-78.

Siedlecki,W. dan Sklansky, J., 1989, "A Note on Genetic Algorithms for Large-Scale Feature Selection", *Pattern Recognition Letters***10**, 335-347.

Wasito, I. dan Mirkin, B., 2005, "Nearest Neighbor Approach in the Least Square Data Imputation Algorithms", *Information Sciences***169**, 1-25.