UNIVERSITAS
GADJAH MADA
Metode Single Center Imputation from Multiple Chained Equations (SICE) untuk Penanganan Missing Value Data Nominal
Afrizal Meka Mulyana, Aina Musdholifah, S.Kom., M.Kom., Ph.D.
Universitas Gadjah Mada, 2025 | Diunduh dari http://etd.repository.ugm.ac.id/

INTISARI

METODE *SINGLE CENTER IMPUTATION FROM MULTIPLE CHAINED EQUATIONS* (SICE) UNTUK PENANGANAN *MISSING VALUE* DATA NOMINAL

Oleh

Afrizal Meka Mulyana

23/511773/PPA/06488

Missing value dalam atribut nominal menjadi masalah penting dalam analisis data karena dapat menurunkan kualitas informasi, memicu bias, dan mengurangi akurasi model prediktif. Berbeda dengan data numerik, data nominal tidak memiliki urutan, sehingga metode imputasi konvensional seperti *K-Nearest Neighbors* (*KNN*) dan *Multivariate Imputation by Chained Equations* (*MICE*) sering kali kurang optimal. Penelitian ini bertujuan mengevaluasi efektivitas metode *Single Center Imputation from Multiple Chained Equations* (*SICE*) dalam menangani *missing value* pada atribut nominal.

Eksperimen dilakukan pada dua dataset *benchmark*, *Adult Income* dan *Hypothyroid*. Masing-masing diproses melalui identifikasi *missing value*, *one-hot encoding*, dan imputasi menggunakan *KNN*, *MICE*, dan *SICE*. Hasilnya dievaluasi menggunakan pemodelan klasifikasi dengan *Logistic Regression* dan *Random Forest*. Kinerja model diukur menggunakan *accuracy*, *precision*, *recall*, *F1-score*, serta diuji secara statistik menggunakan *paired t-test*.

Hasil menunjukkan bahwa *SICE* unggul secara signifikan pada dataset *Hypothyroid*, dengan *F1-score* mencapai 74% dan *precision* 73% pada model *Logistic Regression*. Kombinasinya dengan *Random Forest* juga mencatat akurasi tertinggi sebesar 85,01%. *Uji paired t-test* menunjukkan perbedaan performa yang signifikan ($p < 0,05$) antara *SICE* dan metode lain pada dataset *Hypothyroid*. Sebaliknya, pada dataset *Adult Income* tidak ditemukan perbedaan signifikan antar metode imputasi. Performa *SICE* bahkan lebih rendah dibanding *KNN* dan *MICE*, mengindikasikan bahwa metode sederhana lebih sesuai. Pola *missing value* pada *Adult Income* cenderung acak (*MCAR*), sedangkan pada *Hypothyroid* menunjukkan keterkaitan antar fitur (*MAR*), yang memungkinkan pendekatan prediktif seperti *SICE* menghasilkan imputasi lebih akurat. Dengan demikian, *SICE* layak diterapkan pada data nominal berpola *MAR* karena stabil dan efisien.

Kata Kunci: *Missing Values*, Imputasi Data, *SICE*, Data Nominal, *Logistic Regression*, *Random Forest*.

UNIVERSITAS
GADJAH MADA
Metode Single Center Imputation from Multiple Chained Equations (SICE) untuk Penanganan Missing Value Data Nominal
Afrizal Meka Mulyana, Aina Musdholifah, S.Kom., M.Kom., Ph.D.
Universitas Gadjah Mada, 2025 | Diunduh dari http://etd.repository.ugm.ac.id/

ABSTRACT

SINGLE CENTER IMPUTATION FROM MULTIPLE CHAINED EQUATIONS (SICE) FOR HANDLING MISSING VALUES IN NOMINAL DATA

By

Afrizal Meka Mulyana
23/511773/PPA/06488

Missing values in nominal attributes present a major challenge in data analysis, as they can degrade information quality, introduce bias, and reduce predictive model accuracy. Unlike numerical data, nominal data lacks inherent order, making conventional imputation methods such as K-Nearest Neighbors (KNN) and Multivariate Imputation by Chained Equations (MICE) often suboptimal. This study aims to evaluate the effectiveness of the Single Center Imputation from Multiple Chained Equations (SICE) method in addressing missing values specifically in nominal attributes.

Experiments were conducted on two benchmark datasets: Adult Income and Hypothyroid. Each dataset underwent missing value identification, one-hot encoding, and imputation using KNN, MICE, and SICE. The imputed data was then evaluated through classification modeling using Logistic Regression and Random Forest. Model performance was assessed using accuracy, precision, recall, and F1-score, followed by statistical significance testing using paired t-tests.

The results show that SICE significantly outperformed the other methods on the Hypothyroid dataset, achieving an F1-score of 74% and a precision of 73% with Logistic Regression. Combined with Random Forest, it also achieved the highest accuracy of 85.01%. Paired t-tests confirmed that these performance differences were statistically significant ($p < 0.05$). In contrast, no significant differences were found on the Adult Income dataset. In fact, SICE performed worse than KNN and MICE, suggesting that simpler methods may be more suitable. The missing value pattern in Adult Income tends to be random (MCAR), while Hypothyroid exhibits feature dependency (MAR), which predictive approaches like SICE can leverage. Therefore, SICE is considered a stable and efficient imputation method, particularly effective for nominal data with MAR patterns.

Keywords: ***Missing Values, Data Imputation, SICE, Nominal Data, Logistic Regression, Random Forest.***