Journal of Big Data

RESEARCH Open Access

Smoothing target encoding and class center-based firefly algorithm for handling missing values in categorical variable

Heru Nugroho*, Nugraha Priya Utama and Kridanto Surendro

*Correspondence:
33218003@std.stei.itb.ac.id

School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Jl. Ganesha, 10, Bandung 40132, Jawa Barat, Indonesia

Abstract

One of the most common causes of incompleteness is missing data, which occurs when no data value for the variables in observation is stored. An adaptive approach model outperforming other numerical methods in the classification problem was developed using the class center-based Firefly algorithm by incorporating attribute correlations into the imputation process (C3FA). However, this model has not been tested on categorical data, which is essential in the preprocessing stage. Encoding is used to convert text or Boolean values in categorical data into numeric parameters, and the target encoding method is often utilized. This method uses target variable information to encode categorical data and it carries the risk of overfitting and inaccuracy within the infrequent categories. This study aims to use the smoothing target encoding (STE) method to perform the imputation process by combining C3FA and standard deviation (STD) and compare by several imputation methods. The results on the tic tac toe dataset showed that the proposed method (C3FA-STD) produced AUC, CA, F1-Score, precision, and recall values of 0.939, 0.882, 0.881, 0.881, and 0.882, respectively, based on the evaluation using the kNN classifier.

Keywords: Missing data, Encoding, Smoothing, Firefly algorithm, Class center

Introduction

The missing of data or missing value is a frequent issue in real-world data analysis. Missing values in datasets may be shown as "?", "nan," "N/A", or "blank cells". In most studies, missing data is a common and challenging problem because it can lead to biased, inaccurate, and unreasonable conclusions when it is mishandled [1–6]. A complete dataset is required for the current analytical methods to operate, as shown by [7, 8], with related missing variable issues serving as opportunities to obtain the correct problem-solving technique [9]. Missing data is shown to be a common problem in classification tasks, leading to the prediction system's ineffectiveness [10]. The ignorance of this issue has an impact on analytical [1, 11, 12], learning, and predictive outcomes for problems involving collaborative prediction, respectively [13]. Furthermore, it can undermine the validity of results and conclusions [3, 12]. In predictive models, the improper selection of missing data methods often affects the model performance [4, 14] as well as the accuracy and efficiency of the classifiers [15].

Univariate or multivariate, monotone, non-monotone, connected, unconnected, planned, and random is the pattern of missing values can be seen in Fig. 1.

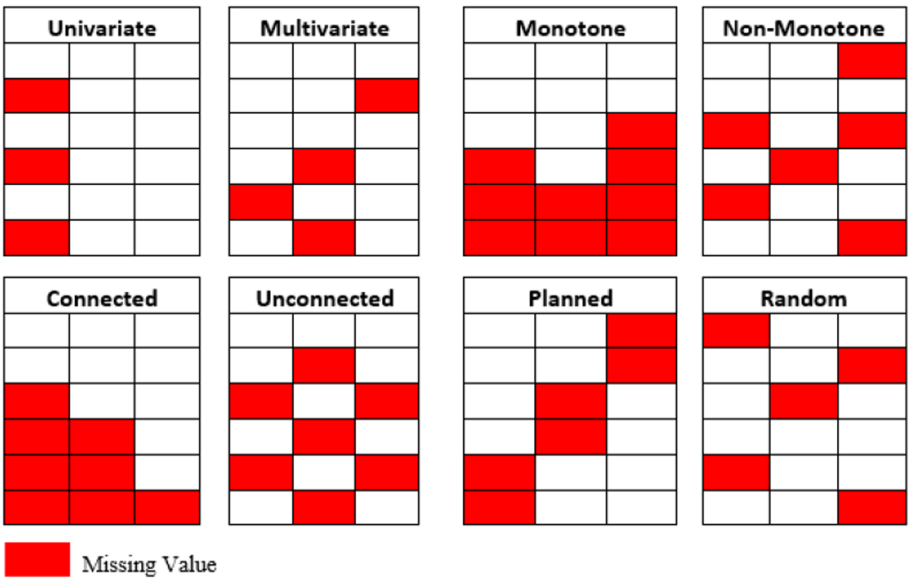

Fig. 1 The pattern of missing values

One of the strategies to deal with missing data is to data imputation. Data imputation is defined as the process of replacing the missing values in a data set through an estimation process with certain values so that a complete data set is produced. Currently, most of the models dealing with missing data use an imputation strategy to complete it [16]. Class center missing value imputation (CCMVI) is develop by [17] and based on the results of experiments carried out for categorical datasets, the classification accuracy values are not better than the approach with mode imputation. Class center-based imputation can obtain an accurate data value when the correlation is considered. However, it does not effectively work on the datasets with high standard deviation attributes [18]. An adaptive search is an alternative to estimate missing data when considering correlations [19]. As a result, it can also estimate the number of missing values [20] in any search problem by maximizing the objective function.

To subsequently implement an adaptive search, Xin-She Yang created the Firefly algorithm at the end of 2007 and the beginning of 2008 [21]. This improved algorithm was inspired by nature and has progressed significantly since its inception a decade ago [22]. The firefly algorithm (FA) is a heuristic optimization algorithm inspired by nature that is based on the luminescence and attraction behavior of fireflies [23]. The FA algorithm is used for a number of reasons, including its effectiveness in solving continuous optimal problems [24], it's a simple and effective swarm intelligence algorithm that has garnered significant scholarly attention [25], and its widespread application to the solution of complex engineering optimization problems [26]. However, the FA algorithm's effectiveness in missing data estimation tasks has not been studied [19]. In the case of missing data imputation, the firefly behavior, in which a bright firefly attracts a firefly with a weaker brightness, can be used. According to reports, this is accomplished by obtaining the closest predicted value to the known variable and then substituting the missing data [27].

According to several previous studies by author, a class center-based firefly algorithm was developed for missing data imputation [28] through the consideration of correlation [18] otherwise known as the C3FA algorithm. The overall architecture framework of C3FA can be seen in Fig. 2.

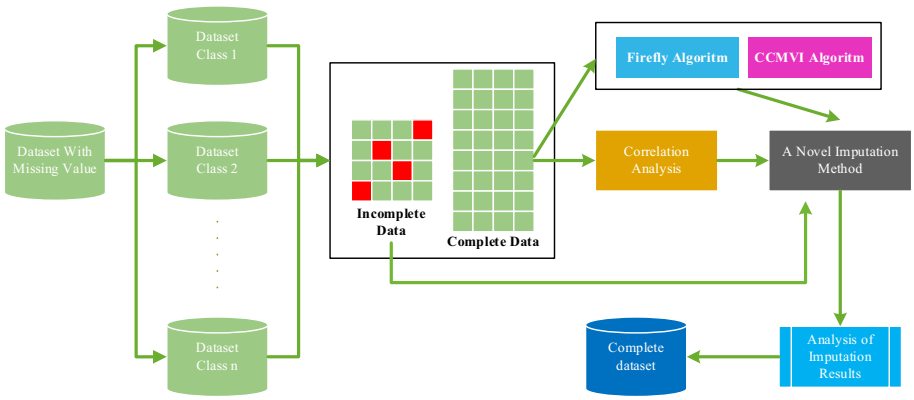

Fig. 2 Architecture framework of C3FA [28]

Further research conducted by the author where at the beginning the imputation method, the author employed a standardization and outlier identification strategy. The result showed combining normalization and outlier removals in C3FA was an efficient technique for obtaining actual data in handling missing values [29]. However, the imputation method with the class center-based firefly algorithm was not analyzed on the categorical datasets in other previous studies.

Data with categorical variables must be coded into appropriate vectors using feature engineering [30]. In the preprocessing stage, categorical variables are also found to be necessary because the majority of machine learning models only consider numerical values. This indicates that the categorical variable should be numerically converted for the model to recognize and retrieve important information [31]. Furthermore, there are numerous approaches for encoding categorical variables for modeling, with one commonly utilized method being the target encoding (TE). This method encodes categorical data [32] and a variant of the continuity scheme based on the value difference metric [33]. In this method, each category is also coded based on its effect on the target variable [30]. For TE, the average target variable for each category is reportedly calculated, subsequently replacing the categorical value with the mean [31]. However, there is an overfitting risk with target encoding, where the accuracy of the machine learning model is effective and ineffective on training and test data, respectively. One frequent smoothing strategy is to combine the category target with the global target mean for each data point (smoothing target encoding).

The contribution of this study is combination smoothing target encoding (STE) before the performance of the missing data imputation with class center-based firefly algorithm in the imputation method. Other contribution of this study is combination of the previously generated imputation was conducted with the standard deviation value (STD) of each attribute (C3FA $\pm$ STD) gives different results for each type of missing rate of data where previous research has never been done. In this study also, each of these methods is selected from the imputation results that produce the closest distance to the class center of each attribute after the imputation process is carried out. The best results are also compared with the existing methods, mode imputation and decision tree imputation.

Target encoding

The use of encoding techniques is often analyzed on machine learning platforms with many complex datasets containing features with high cardinality. When the number of levels reaches a point where an encoding indicator is present, several unreasonable

features are often provided and orderly mapped to integer values. Another general strategy aims to reduce the number of levels by several methods, such as hierarchical clustering based on the statistics of the target variable. However, these are rarely described in several scientific publications [34].

Target Encoding (TE) is often used in encoding category data [32], where each group is encoded based on its effect on the target variable [30]. This method indicates that the mean target variable for each category is calculated, subsequently replacing the categorical value with the average data [31]. However, target encoding has overfitting risk, where the accuracy of the machine learning model is effective and ineffective on the training and test data, respectively. The statistics calculated on the cluster are also likely to be wildly inaccurate due to the infrequent occurrence of the categorical data. Therefore, the solution to this problem is the addition of smoothing, based on the combination of the categorical and overall averages as follows,

$$\text{Smoothing Target Encoding } (STE) = w \times TE_i + (1 - w) \times \overline{TE} \tag{1}$$

The weight (w) is a value between 0 and 1, calculated from the category frequency with the following formula,

$$w = n/(n + m) \tag{2}$$

where n is the number of categorical occurrences in the data and m is the smoothing factor. In this equation, a larger m-value subsequently provided more weight to the overall estimate. TE_i is average in- category i and $\overline{TE}$ is the overall average from all categories. To determine the difference in the results of TE and STE with several m-values, a trial was carried out using a tic tac toe dataset as follows (Fig. 3). The tic tac toe dataset was used in a previous study by [17] for the categorical dataset.

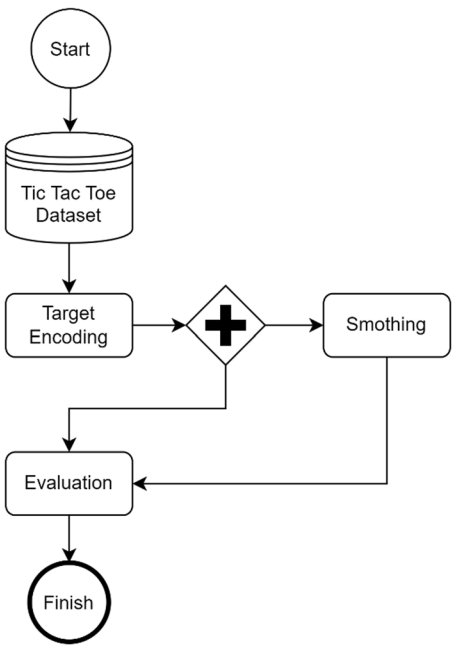

Fig. 3 Flow map of target encoding and smoothing target encoding on tic tac toe dataset

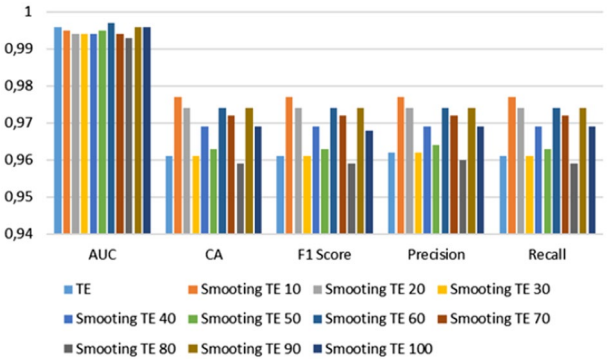

Fig. 4 Comparison of target encoding accuracy and smoothing target encoding on the tic tac toe dataset

The evaluation comparison between both methods (TE and STE) on this dataset is also shown in Fig. 4. For a classification problem, the AUC (Area Under The Curve) ROC (Receiver Operating Characteristics) curve to check or visualize the performance of the classification problem. During the analysis, we evaluate the TE and STE method performance using four metrics: classification accuracy (CA), precision, recall, and F1-score. Smoothing balances category values with overall averages to reduce small-group influence. The value of 10–100 means weight of smoothing, and large values leads to global averages. Based on Fig. 4, the STE method had a better AUC, classification accuracy, FI Score, precision, and recall value than the TE method. This indicated that categorical data were numerically transformed in the next stage using STE method. The AUC value is higher, it indicates that the model is performing better when it comes to differentiating between the positive and negative classes.

Imputation method

Imputation method fill in the missing data to produce a complete data matrix that can be analyzed using standard techniques. The imputation methods used in this study are, mode imputation, decision tree imputation, and class center-based firefly algorithm. The missing data mechanism used in the experiment is Missing Completely at Random (MCAR). Missing Completely at Random is considered to be the simplest type of missing data to understand [35]. This type of missing data has no pattern among missing data values. This approach makes the assumption that the missing data (or missingness) is unrelated to any of the other observed or missing variables [36–40]. The probability (P) that the data variable is missing does not depend on the observed data value or the missing data value ($P(missing|complete\ data) = P(missing)$).

Mode imputation

One of the most naive and straightforward methods for filling in missing values for categorical variables is mode imputation [41]. This indicated that the non-missing value mode of each variable was used to calculate the missing data [41–43]. When using imputation mode, the value was found not to exceed the minimum or maximum requirements. However, the underlying data or distribution was mainly distorted, with bias observed to any estimate except the mean [44]. It did not also correctly address the uncertainty of the data set, subsequently leading to biased imputation [45]. Additionally, the mode imputation in [17] was superior to other methods, including the type proposed by the study of Tsai (the class center method), based on the MCAR missing data.

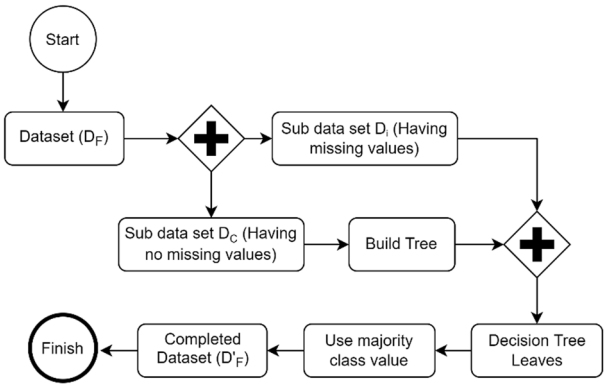

Fig. 5 The overall block diagram of decision tree imputation [51]

Decision tree imputation

This technique was initially and subsequently introduced by Shapiro (1987) and Quinlan (1987), respectively, where each attribute is missing values were determined using a decision tree. This method was then populated using the appropriate tree, with a separate construction produced through known value instances. The unknown values of specific attributes were then determined using these trees [46].

According to Creel and Krotki [47], the decision tree nodes were used to define the imputation class. Subsequently, these nodes were used to apply different imputation methods within the class [47]. This algorithm is found to handle numeric and categorical variables, as well as identify and eliminate most of the data and available remnants, respectively. Meanwhile, decision trees produce complex, time-consuming, and low-bias constructions [48]. Other previous studies using this algorithm were [5, 49–51], with the stages of imputing missing data shown on Fig. 5 [51].

1. Split a complete dataset (D_F) into two sub-datasets, D_C and D_i (only has records without and with missing values).
2. Construct a set of decision trees in D_C based on the attributes which have missing values in D_i as class features.
3. Assign each record from D_i to the leaf where the record's class attribute has a missing value. A record was assigned to more than one leaf with several attributes with missing values.
4. Calculate the categorical missing value using the majority class variable in the leaf.
5. Merge records to form a complete dataset (D'_F).

Class center-based firefly algorithm

The two key components of the Firefly algorithm are the variation in light intensity $I(x)$ and the calculation of attractiveness β. The pattern of fireflies with a lower light intensity is more comparable to the collection with a brighter beam, which imputes missing data. This shows that lower light fireflies are comparable to the properties of missing data, with the complete variable feature being analogous to those with brighter beam intensity. Based on imputation, the class center was used as the objective function $f(x)$, indicating that the value was the prefix in determining $I(x)$. For class center-based missing data, the main steps of the Firefly Algorithm are summarized as follows,

1. Incomplete datasets are subdivided into subsets that are both complete and incomplete.
2. The class center and standard deviation of the complete subset should be calculated for each *i*-class.
3. Use Euclidean distance to calculate the distance between the centDi class center and the remainder of the data samples in *i*.

$$Dis(cent(D_i), j) = \sqrt{(x_i - cent(D_i))^2} \tag{3}$$

4. Calculate the attribute correlation (R) on the complete subset.

$$R_{x_1x_2} = \frac{n\sum x_1x_2 - \left(\sum x_1\right)\left(\sum x_2\right)}{\sqrt{\left(n\sum x_1^2 - \left(\sum x_1\right)^2\right)\left(n\sum x_2^2 - \left(\sum x_2\right)^2\right)}} \tag{4}$$

5. The variable of the class center, $f(x)$, is used to calculate $I(x)$ for each attribute in the incomplete dataset.

$$I(x) = \frac{1}{cent(D_i)} \tag{5}$$

6. Determine the $I(x) = \frac{1}{x_i}$ value greater than $I(x) = \frac{1}{cent(D_i)}$. When data containing the highest $I(x)$ is available, it is necessary to revise the movement $x^k_{i\_new}$. Applying the assumption of the following movement equation $\beta_0 = 1, r = Dis(cent(D_i), j)$ and $\alpha \in [0, 1]$ where β_0 is the attractiveness at $r = 0$, and r is the distance between two fireflies. The parameter γ is the light absorption coefficient and $\alpha = 0.1$ is the step factor.

   a. The following formula is utilized when the class center value (CentDi) of the missing data feature is similar to that of the correlated attribute data,

$$x^k_{i\_new} = x^k_{i\_old} + \beta_0 e^{-\gamma r^2}\left|centD_i - x_{i\_old}\right| + \alpha\left(rand - \frac{1}{2}\right), \text{ with } \gamma = centD_i \tag{6}$$

   b. The following formula is utilized when the CentDi value of the missing data feature is less than that of the correlated attribute data,

$$x^k_{i\_new} = x^k_{i\_old} + \beta_0 e^{-\gamma r^2}\left|centD_i - x_{i\_old}\right| + \alpha\left(rand - \frac{1}{2}\right), \text{ with } \gamma = \left({}^{centD_i}/_{R}\right) + \left|diff\ of\ centD_i\right| \tag{7}$$

   c. The following formula is utilized when the CentDi value of the missing data feature is greater than that of the correlated attribute data,

$$\begin{aligned} x^k_{i\_new} &= x^k_{i\_old} + \beta_0 e^{-\gamma r^2}\left|centD_i - x_{i\_old}\right| + \alpha\left(rand - \frac{1}{2}\right), \\ \text{with } \gamma &= (centD_i \times R) - \left|diff\ of\ centD_i\right| \end{aligned} \tag{8}$$

7. Compare the data distance with the class canter generated from the previous imputation value ± standard deviation to analyze the imputed results. The closest distance is also used to decide the results.

Table 1 Confusion matrix for binary classification

		Actual	
		Positive	**Negative**
Predictions	Positive	TP	FP
	Negative	FN	TN

TP true positive, *FP* false positive, *FN* false negative, TN true negative

Performance evaluation

Imputing missing values is followed by an evaluation of the imputation results. The most popular strategy involves comparing the actual value in the collected data set to the estimated or predicted value in the incomplete data set known as direct evaluation using RMSE. Another method for evaluating the quality of imputation is to look at the classification performance of some classifiers trained on imputed datasets using classification accuracy (CA). Different imputation methods for the same incomplete datasets are likely to yield different imputation results, the classifier with higher classification accuracy is indicated by the higher imputation quality of its training and datasets. As a result, the most effective imputation methods can be identified [52]. The evaluation of machine learning models was utilized to determine the influence of smoothing target encoding on numerous imputation strategies. These included the AUC, Precision, Recall, and F1-Score models, based on the Confusion Matrix popularly used when solving classification problems (Table 1). This was subsequently applied to binary and multiclass classification problems, respectively [53].

The Confusion Matrix represents the machine learning-based data's predictions and actual conditions. The precision is also the ratio of the correct and overall positive predictions, respectively.

$$\text{Precesion} = \frac{TP}{TP + FP} \tag{9}$$

The recall (or sensitivity) is defined as the ratio of true positive predictions to the total data.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{10}$$

The F1-Score is a weighted average precision and recall comparison.

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{11}$$

AUC is the area under the curve of ROC (Receiver Operating Characteristic), which describes the sensitive and specific probability variables with boundary values between 0 and 1. Since this is a common approach for determining the quality of predicted predictions, it offers an overview of the model's overall appropriateness measurement [54]. Furthermore, the classification accuracy analysis of the complete dataset was carried out through an algorithm, namely k-Nearest Neighbors (kNN). This was in line with previous studies [52], which showed that the widely used classifier to evaluate the performance of imputation accuracy is kNN. In addition to the evaluation described in the previous section, the method proposed in this study will also be tested based on the RMSE and R Square values.

Table 2 Dataset Information

Dataset characteristics	Attribute characteristics	Associated tasks	Number of instances	Number of attributes
Multivariate	Categorical	Classification	958	9

Results

The first stage of this study was based on the selection of the tic-tac-toe dataset, it was accessed via the UC Irvine Machine Learning Repository. Dataset information can be seen in Table 2.

This contained a tic tac toe endgame footage, where the first nine attributes represented the nine fields on the board. However, the tenth attribute was the class feature containing the winning status information of player *x*. The stages of research carried out can be seen in Fig. 6.

1. The tic tac toe dataset was encoded using the STE method by RapidMiner (Fig. 7). RapidMiner is a system that facilitates the design and documentation of a comprehensive data mining procedure. It provides not only a nearly exhaustive set of operators, but also structures that express the process control flow [55].

Algorithm 1. Generating Amputation

```
library (MASS)
library (VIM)
library (mice)
library (lattice)
library (readxl)
ampute_tictactoe <– ampute (tictactoe, prop = 0.4, patterns = NULL,freq =
NULL,
                    mech = "MCAR", weights = NULL, std = TRUE,cont
= TRUE,
                    type = NULL, odds = NULL, bycases = TRUE, run =
TRUE)
```

*Example amputation for tic tac toe dataset with missing rate 40% and missing mechanism MCAR

2. The tic tact toe dataset amputation with missing rate 10–60% and missing mechanism MCAR. The generation of missing values is a crucial step in the assessment of a methodology for missing data. the procedure whereby missingness is introduced as an amputation in complete data [56]. The generating amputation using R programming language can be seen in algorithm 1.
3. Imputation process using mode imputation (MdI) and decision tree imputation (DTI), and the proposed method developed in the previous study by author, the Class Centre-based Firefly algorithm (C3FA) [28].
4. Imputation result by Class Centre-based Firefly algorithm (C3FA) combined in a number of ways.
    a. Combination with the standard deviation (±STD) of each attribute (C3FA±STD).
    b. Comparison of the distance between each data record, the smallest distance being used as a reference for the previously obtained imputation results (C3FA+Dist).
5. Evaluation of the performance of missing data method.

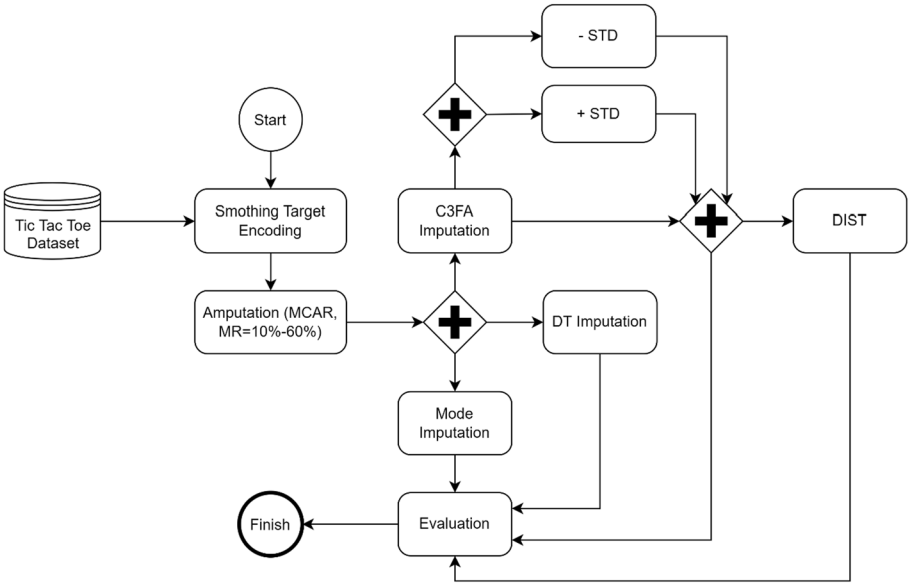

Fig. 6 Research Stages

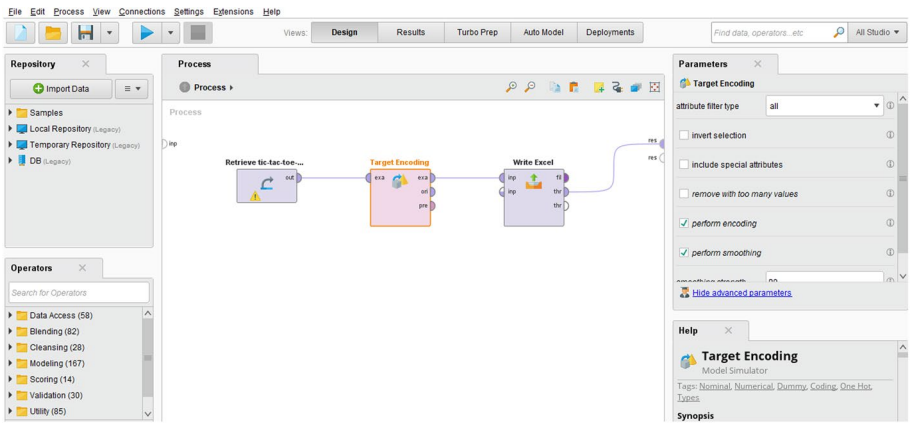

Fig. 7 Smoothing target encoding using RapidMiner

Imputation was subsequently carried out by replacing the missing values of the categorical variables through the MdI, DTI, and C3FA methods. This was conducted after obtaining the tic tac toe dataset with 10–60% missing values in the previous stage through the MCAR mechanism (MCAR_10—MCAR_60). The following are the analytical results based on the AUC, Classification Accuracy (CA), F1-Score, Precision, and Recall values.

Tables 3, 4, 5 showed that the MdI, DTI, and C3FA methods produced AUC, Classification Accuracy (CA), F1-Score, Precision, and Recall values, which decreased with an increase in the percentage of missing data through the MCAR mechanism. The following is a comparison of the performance of the proposed method with the imputation mode and decision tree imputation methods as state-of-art techniques based on the average value of each performance.

Table 3 AUC, CA, F1-Score, Precision, And Recall Result with Mode Imputation

Mechanism_missing rate	Performance evaluation of mode imputation (MdI)				
	AUC	CA	F1-Score	Precision	Recall
MCAR_10	**0.959**	**0.894**	**0.893**	**0.893**	**0.894**
MCAR_20	0.937	0.888	0.887	0.887	0.888
MCAR_30	0.925	0.880	0.879	0.879	0.880
MCAR_40	0.911	0.876	0.874	0.875	0,876
MCAR_50	0.891	*0.839*	*0.838*	*0.838*	*0,839*
MCAR_60	*0.877*	0.844	0.843	0.843	0.844

Italic is lower result, Bold is higher result

Table 4 AUC, CA, F1-Score, Precision, And Recall Result with Decision Tree Imputation

Mechanism_missing rate	Performance evaluation decision tree imputation (DTI)				
	AUC	CA	F1-Score	Precision	Recall
MCAR_10	**0.964**	**0.906**	**0.906**	**0.906**	**0.906**
MCAR_20	0.960	0.902	0.901	0.901	0.902
MCAR_30	0.934	0.887	0.886	0.886	0.887
MCAR_40	0.917	0.871	0.869	0.869	0.871
MCAR_50	0.902	0.862	0.861	0.861	0.862
MCAR_60	*0.895*	*0.838*	*0.836*	*0.836*	*0.838*

Italic is lower result, Bold is higher result

Table 5 AUC, CA, F1-Score, Precision, And Recall Result with C3FA Imputation

Mechanism_Missing rate	Performance evaluation (C3FA)				
	AUC	CA	F1-Score	Precision	Recall
MCAR_10	**0.965**	**0.904**	**0.904**	**0.904**	**0.904**
MCAR_20	0.965	0.889	0.888	0.889	0.889
MCAR_30	0.933	0.878	0.877	0.877	0.878
MCAR_40	0.943	0.874	0.873	0.873	0.874
MCAR_50	0.933	*0.870*	*0.868*	*0.868*	*0.870*
MCAR_60	*0.920*	0.871	0.868	0.870	0.871

Italic is lower result, Bold is higher result

Table 6 shows that the proposed method produces an average value of AUC, Classification Accuracy (CA), F1-Score, Precision, and Recall than the imputation mode and decision tree imputation methods. Theoretical and empirical work demonstrates that when comparing two measures for learning algorithms, AUC is superior to classification accuracy based on formal criteria [57]. Our proposed method is 2.6% superior to the imputation mode method and 1.4% from the Decision tree imputation for the AUC value. Mode imputations are easy to implement, but they fail to account for relationships between variables and thus underestimate variance. The decision tree imputation (DTI) method and class center-based Firefly algorithm by incorporating attribute correlations into the imputation process (C3FA) are better than the mode imputation (MdI) methods because they consider attribute correlation. This result is in line with the fact that the performance of the missing data imputation algorithm is significantly affected by the correlation in the data [1, 11, 58–60]. Another advantage of the proposed method over other methods is the use of the firefly algorithm in the data imputation process to produce an optimal imputation value.

Table 6 Comparison of Average Performance Evaluation

Method	Performance evaluation				
	AUC	CA	F1-Score	Precision	Recall
Mode imputation (Mdl)	*0.917*	*0.870*	*0.869*	*0.869*	*0,870*
Decision tree imputation (DTI)	0.929	0.878	0.877	0.877	0,878
C3FA imputation (Proposed method)	**0.943**	**0.881**	**0.880**	**0.880**	**0,881**

Italic is lower result, Bold is higher result

Another contribution of this research is the use of the standard deviation of each attribute in the data on the imputation results. The combination of imputation results with the standard deviation of the imputation method in previous studies has never been used. Using the proposed C3FA method, an imputation process was also carried out based on the combination of the previous results with the standard deviation (±STD) of each attribute. Based on the results of the imputation of C3FA, C3FA + STD, and C3FA—STD, a comparison of the distance between each imputed data to the class center was carried out. The smallest distance in the data will be used as a reference for the imputation results (C3FA + Dist). In addition, Figs. 8, 9, 10, 11, 12 are the performance evaluation of each combination,

Figures 8, 9, 10, 11, 12 shows that each combination of imputed results using the standard deviation and the resulting distance for each imputation shows a different pattern in each evaluation result. Previous studies related to data imputation methods have not used standard deviation as a consideration for data imputation results. Therefore, the findings in this study can be tested on the standard imputation method which is widely used in previous studies.

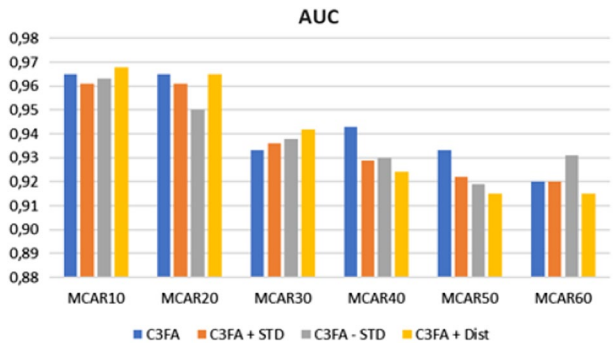

Fig. 8 Comparison of AUC results with C3FA imputation in the tic tac toe dataset MR:10–60%, MCAR

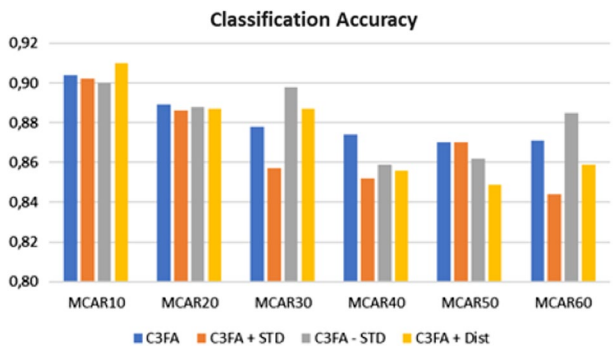

Fig. 9 Comparison of CA results with C3FA imputation in the tic tac toe dataset MR:10–60%, MCAR

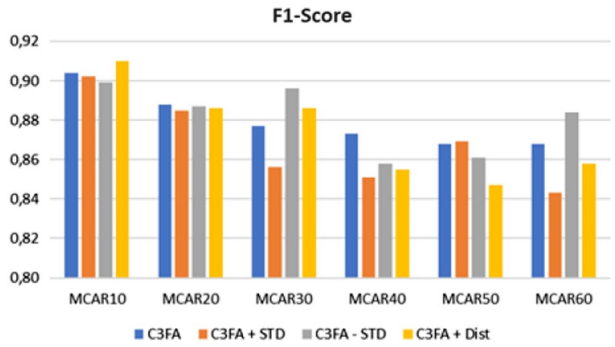

Fig. 10 Comparison of F1-Score results with C3FA imputation in the tic tac toe dataset MR:10–60%, MCAR

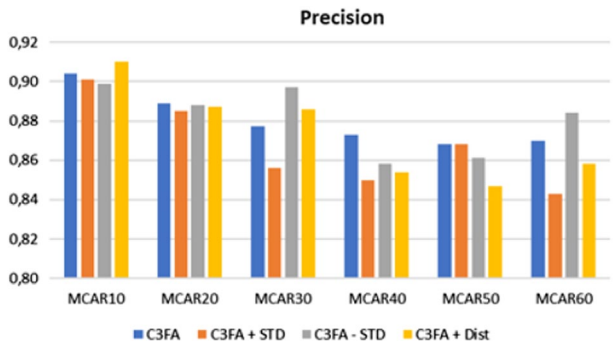

Fig. 11 Comparison of Precision results with C3FA imputation in the tic tac toe dataset MR:10–60%, MCAR

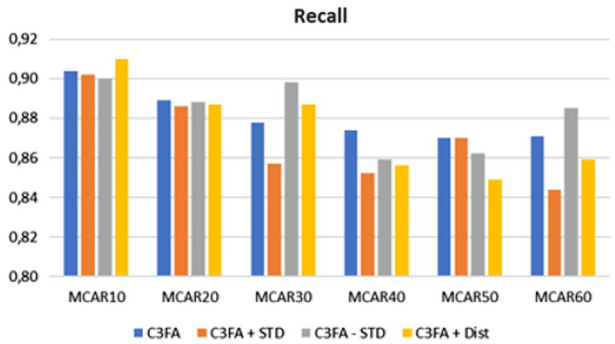

Fig. 12 Comparison of Recall results with C3FA imputation in the tic tac toe dataset MR:10–60%, MCAR

The method suggested in this study (C3FA ± STD) will also be tested based on the RMSE and coefficient of determination (R^2) values in addition to the evaluation outlined in the previous section. Root Mean Square Error (RMSE) is widely recognized as the primary metric for comparing the performance of forecasting methods, as it measures the difference between the imputed value and the original value for a given feature. In this instance, a value closer to zero yields superior imputation [61, 62]. The correlation coefficient (r) is one of the most common ways to measure imputation ability and its square is the coefficient of determination (R^2), which is the amount of variation that can be explained and is between 0 and 1. An efficient imputation technique must have an value R^2 close to 1 [61–64].

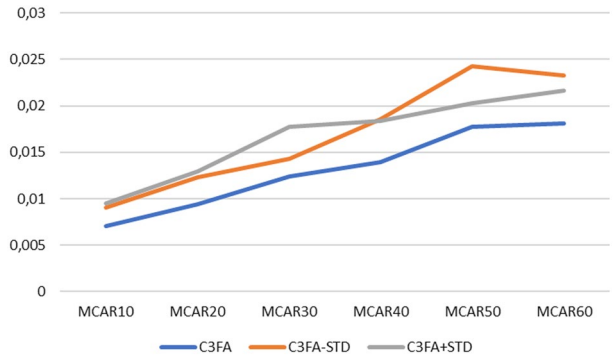

Fig. 13 Comparison of RMSE with C3FA imputation in the tic tac toe dataset MR:10–60%, MCAR

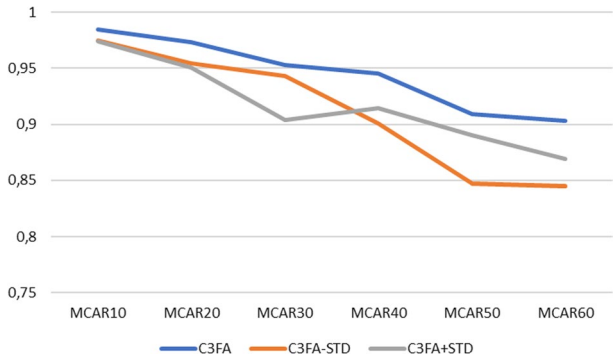

Fig. 14 Comparison of R Square with C3FA imputation in the tic tac toe dataset MR:10–60%, MCAR

Based on the results of RMSE and R^2, in general the C3FA method is efficiency of an imputation technique base one RMSE values closer to 0 and R^2 values closer to 1 compared to the combination of C3FA ± STD methods as can be seen in Figs.13 and 14. The Predictive Accuracy (PAC) relates to the efficiency of an imputation technique to retrieve the true values in data that can be measures by R^2 and Root Mean-Squared Error (RMSE).

Analysis and discussion

At the preprocessing stage, these categorical variables were converted to numeric data for the model to understand and retrieve useful information [31]. One method of numerically converting categorical data is observed through target encoding, whose limitations also entail overfitting risks and the inaccuracy of infrequent categorical data. Furthermore, the results of the simulation showed that the smoothing target encoding technique produced better classification accuracy values than the TE. Using the tic tac toe dataset, differences were observed in the values of AUC, Classification Accuracy, Precision, F1-score, and Recall, based on the C3FA method with several imputation combinations, namely C3FA + STD, C3FA-STD, and C3FA + Dist can be seen di Figs. 8, 9, 10, 11, 12. This is indicated that the C3FA + Dist method produced the best evaluation value for the total missing data of 10%. Meanwhile, the C3FA method produced the best values for the total missing data of 20%, 40%, and 50%, with C3FA-STD yielding optimized parameters at 30% and 60%. According to the average results of the AUC, Classification Accuracy, Precision, F1-score, and Recall values, the C3FA and C3FA-STD methods had better advantages than other combinations, as shown in Fig. 15.

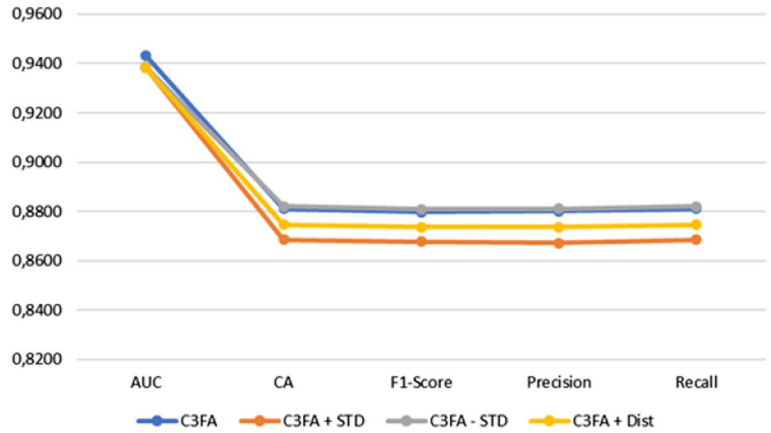

Fig. 15 Comparison of the performance of the C3FA imputation method in the tic tac toe dataset MR:10–60%, MCAR

The Smoothing Target Encoding method was used in the preprocessing stage before the imputation process within the C3FA method, which produced several patterns through the several missing data of each model. This showed a new result, where the imputation analysis did not produce better performance values by comparing the C3FA + Dist, C3FA + STD, and C3FA-STD methods. It also showed that C3FA + Dist had better results at a low missing data < 20%. The C3FA method had a good performance on the missing data of 40% through the MCAR mechanism which is in line with previous research [28, 29]. Other experimental results showed that the C3FA-STD method produced the best performance evaluation when the dataset had a reasonably high amount of missing data (60%). However, based on the evaluation using RMSE and R Square values, the C3FA method showed better performance than C3FA ± STD.

Conclusion

Based on the preprocessing stage, categorical variables were significant because the machine learning models mostly considered numerical values. This indicated that the categorical variables were numerically converted for the model to understand and retrieve helpful information. Using the tic tac toe dataset and three (3) imputation methods, the proposed C3FA-STD method produced the AUC, CA, F1-Score, Precision, and Recall values of 0.939, 0.882, 0.881, 0.881, 0.882, respectively. However, the value outperformed the MdI and DTI methods when using the kNN classifier. This value outperforms the mode imputation method, the best method in previous studies [17] for categorical data, and the imputation method with a decision tree.

Standard deviation is a statistical measure in data other than the class center, and correlation that has been used in previous studies can be used as one of the considerations in the imputation results of missing data. When using the class center-based method, the correlation of attributes in the data was utilized in the imputation process because each data has a relationship with one another. Meanwhile, the imputation process was expected to produce an optimal or closest value to the actual rate. This is indicated that when considering the correlation in the imputation process, a Firefly Algorithm (FA) was used. However, the use of the FA algorithm which has been developed in other studies for optimization such as [23–26] can be tried in further research to handle missing data with a class center approach.

Based on the class center-based method, statistical measures such as standard deviation was used to combine the imputed results. This was due to the standard deviation being used to determine the closeness of a statistical data sample to the average variables. In the simulation results, differences were also observed in the performance evaluation of the proposed method through the combination of the imputation outputs and the standard deviation.

In order to validate the performance of the missing data imputation technique and arrive at a definitive conclusion, evaluation is an extremely important step that must first be taken. The missing data imputation technique can be evaluated in several different ways, the most important of which are the direct evaluation method, the classification accuracy of the classifiers, and the consideration of the computational time. However, this has not been the case, and all three evaluation metrics have not been used together in any of the related studies [52]. One of the future challenges of this research is evaluation based on computational time. Most imputation methods in previous studies were only tested on one missing data mechanism (MCAR, MAR, or MNAR). Therefore, further research that will be carried out is to conduct tests by grouping datasets based on the percentage of missing rate with the mechanism of not only MCAR, but also MAR and MNAR.

Abbreviations

TE	Target encoding
STE	Smoothing target encoding
STD	Standard deviation
C3FA	Class center-based firefly algorithm
DTI	Decision tree imputation
MdI	Mode imputation
kNN	K-nearest neighbor
AUC	Area under the curve
ROC	Receiver operating characteristic

Acknowledgements
We would like to thank Institut Teknologi Bandung and Telkom University for supporting this research.

Author contributions
The authors confirm contribution to the paper as follows: study conception and design: HN and KS; data collection: HN; analysis and interpretation of results: HN and N P U; draft manuscript preparation: HN. All authors reviewed the results and approved the final version of the manuscript.

Funding
Not applicable. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Availability of data and materials
The original dataset used for this study is available in UCI Machine Learning Repositoryhttps://archive.ics.uci.edu/ml/datasets/Tic-Tac-Toe+Endgame.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The author reports no potential competing interest.

Received: 18 May 2022 Accepted: 25 December 2022
Published online: 28 January 2023

References

1. Beretta L, Santaniello A. Nearest neighbor imputation algorithms: a critical evaluation. BMC Med Info Decis Mak. 2016. https://doi.org/10.1186/s12911-016-0318-z.
2. Hayati Rezvan P, Lee KJ, Simpson JA. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. BMC Med Res Methodol. 2015. https://doi.org/10.1186/s12874-015-0022-1.
3. Kang H. The prevention and handling of the missing data. Korean J Anesthesiol. 2013;64:402–6.
4. Pampaka M, Hutcheson G, Williams J. Handling missing data: analysis of a challenging data set using multiple imputation. Int J Res Method Edu. 2016;39:19–37.
5. Rahman MdG, Islam MZ. Missing value imputation using a fuzzy clustering-based EM approach. Knowledge Info Sys. 2016;46:389–422.
6. Zhao Y, Long Q. Multiple imputation in the presence of high-dimensional data. Stat Methods Med Res. 2016;25:2021–35.
7. Nishanth KJ, Ravi V. Probabilistic neural network based categorical data imputation. Neurocomputing. 2016;218:17–25.
8. Van Hulse J, Khoshgoftaar TM. Incomplete-case nearest neighbor imputation in software measurement data. Inf Sci. 2014;259:596–610.
9. Nugroho H, Surendro K. Missing Data Problem in Predictive Analytics. 8th International Conference on Software and Computer Applications - ICSCA '19. Penang, Malaysia: ACM Press; 2019. p. 95–100.
10. Jugulum R. Importance of data quality for analytics. In: Sampaio P, Saraiva P, editors. Quality in the 21st Century. Cham: Springer International Publishing; 2016. p. 23–31.
11. Deb R, Liew AW-C. Missing value imputation for the analysis of incomplete traffic accident data. Info Sci. 2016;339:274–89. https://doi.org/10.1016/j.ins.2016.01.018.
12. Pedersen A, Mikkelsen E, Cronin-Fenton D, Kristensen N, Pham TM, Pedersen L, et al. Missing data and multiple imputation in clinical epidemiological research. Clin Epidemiol. 2017;9:157–66.
13. García-Laencina PJ, Sancho-Gómez J-L, Figueiras-Vidal AR, Verleysen M. K nearest neighbours with mutual information for simultaneous classification and missing data imputation. Neurocomputing. 2009;72:1483–93.
14. Dong Y, Peng C-YJ. Principled missing data methods for researchers. SpringerPlus. 2013;2:222. https://doi.org/10.1186/2193-1801-2-222.
15. Bhati S, Kumar Gupta MKG. 2016 Missing Data Imputation for Medical Database: Review. International Journal of Advanced Research in Computer Science and Software Engineering.
16. Wilmots B, Shen Y, Hermans E, Ruan D. 2011 Missing data treatment : Overview of possible solutions. Uitgave: Steunpunt Mobiliteit & Openbare Werken–Spoor Verkeersveiligheid.
17. Tsai C-F, Li M-L, Lin W-C. A class center based approach for missing value imputation. Knowl-Based Syst. 2018;151:124–35.
18. Nugroho H, Utama NP, Surendro K. 2020 Performance Evaluation for Class Center-Based Missing Data Imputation Algorithm. Proceedings of the 2020 9th International Conference on Software and Computer Applications. Langkawi Malaysia: ACM; 36–40.
19. Leke CA, Marwala T. Deep Learning and Missing Data in Engineering Systems. Cham: Springer International Publishing; 2019.
20. Abdella M, Marwala T. 2005 The use of genetic algorithms and neural networks to approximate missing data in database. Mauritius: IEEE; 207–12.
21. Yang X-S. Nature-Inspired Metaheuristic Algorithms. 2nd ed. United Kingdom: Luniver Press; 2010.
22. Yang X-S, He X-S. Why the Firefly Algorithm Works? In: Yang X-S, editor. Nature-Inspired Algorithms and Applied Optimization. Cham: Springer International Publishing; 2018. p. 245–59.
23. Peng H, Zhu W, Deng C, Wu Z. Enhancing firefly algorithm with courtship learning. Inf Sci. 2021;543:18–42.
24. Cao L, Ben K, Peng H, Zhang X. Enhancing firefly algorithm with adaptive multi-group mechanism. Appl Intell. 2022;52:9795–815.
25. Peng H, Qian J, Kong F, Fan D, Shao P, Wu Z. Enhancing firefly algorithm with sliding window for continuous optimization problems. Neural Comput Applic. 2022. https://doi.org/10.1007/s00521-022-07193-6.
26. Peng H, Xiao W, Han Y, Jiang A, Xu Z, Li M, et al. Multi-strategy firefly algorithm with selective ensemble for complex engineering optimization problems. Appl Soft Comput. 2022;120:108634.
27. Agbehadji IE, Millham RC, Fong SJ, Yang H. Bioinspired computational approach to missing value estimation. Math Probl Eng. 2018;2018:1–16.
28. Nugroho H, Utama NP, Surendro K. Class center-based firefly algorithm for handling missing data. J Big Data. 2021;8:37.
29. Nugroho H, Utama NP, Surendro K. Normalization and outlier removal in class center-based firefly algorithm for missing value imputation. J Big Data. 2021;8:129.
30. Cerda P, Varoquaux G, Kégl B. Similarity encoding for learning with dirty categorical variables. Mach Learn. 2018;107:1477–94.
31. Dahouda MK, Joe I. A deep-learned embedding technique for categorical features encoding. IEEE Access. 2021;9:114381–91.
32. Micci-Barreca D. A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. SIGKDD Explor Newsl. 2001;3:27–32.
33. Duch W, Grudzi K, Stawski G. 2000 Symbolic Features In Neural Networks. 2000.
34. Pargent F. A Benchmark Experiment on How to Encode Categorical Features in Predictive Modeling. München: Ludwig-Maximilians-Universität München; 2019.
35. Graham JW. Missing data analysis: making it work in the real world. Annu Rev Psychol. 2009;60:549–76.
36. Breiman L. Statistical modeling: the two cultures. Stat Sci. 2001;16:199–215.
37. Lang KM, Little TD. Principled missing data treatments. Prev Sci. 2018;19:284–94.
38. Peng L, Lei L. 2005 A Review of Missing Data Treatment Methods. Int J Intel Inf Manag Syst Tech. 8. https://scholar.google.com/scholar_lookup?title=A+review+of+missing+data+treatment+methods&author=Peng,+L.&

author=Lei,+L.&publication_year=2005&journal=Intell.+Inf.+Manag.+Syst.+Technol&volume=1&pages=412%E2%80%93419.
39. Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods. 2002;7:147–77.
40. Singh M. 2017 Implications Of Missing Data Designs With The Use Of A Longitudinal Dataset. University Muncie
41. Xu X, Xia L, Zhang Q, Wu S, Wu M, Liu H. The ability of different imputation methods for missing values in mental measurement questionnaires. BMC Med Res Methodol. 2020;20:42.
42. Mir AA, Kearfott KJ, Çelebi FV, Rafique M. 2022 Imputation by feature importance (IBFI) A methodology to envelop machine learning method for imputing missing patterns in time series data. In: Shahid S, (ed). PLoS ONE. 17: e0262131
43. Jadhav A, Pramod D, Ramanathan K. Comparison of performance of data imputation methods for numeric dataset. Appl Artif Intell. 2019;33:913–33.
44. van Buuren S. Flexible Imputation of Missing Data. US: CRC Press Taylor & Francis Group; 2012.
45. Khan SI, Hoque ASML. SICE: an improved missing data imputation technique. Journal of Big Data. 2020;7:37.
46. Twala B. An empirical comparison of techniques for handling incomplete data using decision trees. Appl Artif Intell. 2009;23:373–405.
47. Krotki K, Creel DV. 2006 Creating imputation classes using classification tree methodology. roceedings of the Survey Research Methods Section (ASA), 2884-2887. https://ww2.amstat.org/meetings/jsm/2006/PDFs/JSM06AbstractBook.pdf.
48. Rokach L. Decision forest: twenty years of research. Information Fusion. 2016;27:111–25.
49. Ghazanfar MA, Prügel-Bennett A. the advantage of careful imputation sources in sparse data-environment of recommender systems: generating improved SVD-based recommendations. Informatica (Slovenia). 2013;37:61–92.
50. Gimpy, Vohra DR, Minakshi. Estimation of Missing Values Using Decision Tree Approach. International Journal of Computer Science and Information Technologies,. 2014;5:5216–5220.
51. Rahman G, Islam Z. 2011 A Decision Tree-Based Missing Value Imputation Technique for Data Pre-Processing. Proceedings of the Ninth Australasian Data Mining Conference—Volume 121. AUS: Australian Computer Society, Inc. 41–50.
52. Lin W-C, Tsai C-F. Missing value imputation: a review and analysis of the literature (2006–2017). Artif Intell Rev. 2020;53:1487–509.
53. Kulkarni A, Chong D, Batarseh FA. Foundations of data imbalance and solutions for a data democracy. In: Batarseh FA, Yang R, editors. data democracy. Amsterdam: Elsevier; 2020.
54. Yuliansyah H, Othman ZA, Bakar AA. Taxonomy of link prediction for social network analysis: a review. IEEE Access. 2020;8:183470–87.
55. Hofmann M, Klinkenberg R, editors. RapidMiner Data Mining Use Cases and Business Analytics Applications. Boca Raton: CRC Press Taylor & Francis Group; 2014.
56. Schouten R. Generating missing values for simulation purposes: a multivariate amputation procedure. J Stat Comput Simul. 2018;88:2909–30.
57. Jin Huang, Ling CX. 2005 Using AUC and accuracy in evaluating learning algorithms. IEEE Trans Knowl Data Eng. 17: 299–310.
58. Armina R, Mohd Zain A, Ali NA, Sallehuddin R. A Review on Missing Value Estimation Using Imputation Algorithm. J Phys: Conf Ser. 2017;892:012004.
59. Tutz G, Ramzan S. Improved methods for the imputation of missing data by nearest neighbor methods. Comput Stat Data Anal. 2015;90:84–99.
60. Zahin SA, Ahmed CF, Alam T. An effective method for classification with missing values. Appl Intell. 2018;48:3209–30.
61. Pompeu Soares J, Seoane Santos M, Henriques Abreu P, Araújo H, Santos J. 2018 Exploring the Effects of Data Distribution in Missing Data Imputation. Advances in Intelligent Data Analysis XVII. Springer International Publishing; New York City. 251–63.
62. Santos MS, Soares JP, Henriques Abreu P, Araújo H, Santos J. 2017 Influence of Data Distribution in Missing Data Imputation. Artificial Intelligence in Medicine. Springer International Publishing. New York City 285–94.
63. Junninen H, Niska H, Tuppurainen K, Ruuskanen J, Kolehmainen M. Methods for imputation of missing values in air quality data sets. Atmos Environ. 2004;38:2895–907.
64. Harel O. The estimation of R^2 and adjusted R^2 in incomplete data sets using multiple imputation. J Appl Stat. 2009;36:1109–18.

Publisher's Note