DOI: http://dx.doi.org/10.26483/ijarcs.v16i6.7388

ISSN No. 0976-5697

Volume 16, No. 6, November-December 2025

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

MISSING VALUES HANDLING METHODS IN R FOR MACHINE LEARNING

Dr. Sanjay Gour
Professor, Department of Computer Science & Engineering
Gandhinagar University, Gandhinagar
Gujrat, India

Abstract: The pre-processing section of machine learning, as well as data science, is one of the crucial sections that denotes the basic building block. It means the data preparation is the most significant part of implementing any type of machine learning algorithm, as well as models. The data pre-processing comprises various stages like missing values handling, outliers' detection and correction, data encoding, normalization of numeric values, handling of class imbalance, etc... The missing values handling is one of the elementary and well as most significant sections, which is very necessary to deal with carefully. There is no chance to avoid or ignore the case, as it may affect the final result by imposing ambiguities in the result. Thus, there is always a need to deal with missing values as per the situation and conditions. It is also tackled according to the algorithms and their working nature. Various programming languages and software packages are also providing approaches to deal with the same. In this research study, we are trying to explore the capability of R programming with reference to various machine learning algorithms. This is explored with the nature of machine learning algorithms.

Keywords: Machine Learning, R Programming, Algorithms, Missing values, Pre-processing.

I. INTRODAUCTION

It is always notable that an irregular and biased dataset results in ambiguities in the consequences. That preparation is very vital for properly implementing machine learning models. The pre-processing section of machine learning as well as data science, is one of the crucial sections that denotes to the basic building block. It provides a firm structure to the dataset before data transformation stage. It means the data preparation is most significant part of the implementing any types of machine learning algorithm as well as models. If we are minutely considering the data pre-processing stages, we will find that data pre-processing comprises various stages like missing values handling, outliers' detection and correction, data encoding, normalization of numeric values, handling of class imbalance etc... Among all these activities missing values handling is one of the elementary and well as most significant section which is very necessary to deal with watchfulness. So, it is noted that there is no chance to avoid or ignore the missing value case, as it may disturb the final result through striking ambiguities in consequence. Hence, there is constantly needed to deal with missing values as per situation and conditions. It is also tackled according to the algorithms and their working nature.
The R programming is one of the specific languages for statistics, Data Analytics and Data Science. The R is performing efficiently with machine learning algorithms. It also has mechanism to deal with the machine learning with reference of various machine learning algorithms. The approaches of dealing missing values in datasets depends on the nature of machine learning algorithms. Although some common strategies are also implied to deal the missing values in the datasets. As various machine learning may be classified in the groups thus the patterns of dealing are common for groups; similarly, the approaches for groups of algorithms are the same. We can also say that similar types of algorithms utilize a common pattern to deal with missing values in datasets. This study is basically an attempt to explore such kind of patterns in R programming.

II. LITERATURE REVIEW

The handling of missing values is one of the most urgent jobs before data analysis as well as data mining. There is lot of methods and packages are available in the programming language R. some of the significant approaches are discussed here.

S Gaur, DD Pandya and D Soni (2020), The closest fit approach is phenomenon in which the generated value will be very close in the numerical form of the actual value. The linear interpolation approach is one of the excellent approaches to recover the missing values which will be statistically significant to the data mining. The method used in this experimental study gives excellent outcome at individual place as well as for consolidation result also [3]. S Gaur and M. S. Dulawat (2011) The authors are worked a lot on the approach closest fit values. The improved closest fit method is next step about situational based generated computerised algorithm to tackle the missing values. This method carries some advance then the ordinary closest fit methods. This method provided improved and refined values at the place of missing values by the utilization advance algorithm [8]. DD Pandya and S Gaur (2018), Anomalies values may be available in two major forms in the dataset as inliers or outliers. In both the case these are accountable as the irregular values and are not having linear nature. To find the proper analytics result there is need to remove the irregularity from the dataset. It means anomalous vales gives biased result in the data mining and analysis [1]. Madan Lal Yadav, and Basav Roychoudhury (2018). In this research, the writers accomplish proportional study of the performance of the mutual R packages, that is VIM, MICE, MissForest, and HMISC, cast-off for missing value attribution. The writers assessed the presentations of the supposed packages about their imputation period, imputation competence and the consequence on the alteration [2].

Imke Mayer et al. (2022) The missing values are unescapable at what time employed with data. Their incidence is intensified as additional data from diverse sources turn out to be available. Though, furthermost statistical representations and visualization approaches need comprehensive data, and inappropriate treatment of missing data consequences in evidence loss or biased examines [4]. S Gaur, DD Pandya, MK Sharma (2020), The research aper gives a method based on the applied interpolation to recover the missing values which are scattered randomly in the dataset. The method follows advance interpolation-based algorithm which takes values grom the neighbouring attribute to generate new values for the missing values place. It is observed that the consolidated result is 98-100% accurate and statistically significance [11]. Moritz and T. Bartz-Beielstein (2017), they discussed the imputation od new values in the time series dataset to recover the missing values cases. The whole study is accomplished by the utilization of R language [5]. S Gaur and MS Dulawat (2010), This paper gives complete details about the statistical inference in the data mining with special reference of missing values. The article covers the whole hypothetical and advance method which are utilised for assessing inference are covered with necessary algorithms and implementation guideline [6]. The univariate approach is one of the capable methods which utilised single attribute to recover the missing values efficiently. The statistical significance of the study gives significant result. The result is similar to the advance and bivariate methods [7].

N. Tierney et al. (2021), the study is based on the R package "naniar" which is utilised to resolve the missing values status and problem with support of visualization techniques. The approach is user friendly which take support of R studio for the proper implementation [9]. S Sharma and S Gaur (2013), The research paper gives a solid concept to recover the missing values by applying agility method which is applicable on the odd size missing values blocks. Handling of such kind of missing values is tedious task which need advance algorithm and programming support despite utilising in-built functions. The result of the study is very good to recover bock missing values [12]. Y. Xie et al. (2017), the study gives a different angle to solve the missing values by using diverse package of R language. The data recovery is good with the development of web-based solution approaches **[10].**

III. TYPE OF MISSING VALUES IN R

In the R programming, handling of missing values taken as a vital task which helps to statistical modelling and data analysis. As the real-world datasets frequently comprise incomplete, vague, or inattentive data. R delivers several separate representations to handle diverse classes of missing data. Knowledge of types of missing values in R benefits to analysts to clean data efficiently and evade misleading outcomes.

In general, most mutual kind of missing value in R language is NA "Not Available". This specifies that a worth is missing or unidentified. The keyword "NA" can happen in any data type, with numeric, character, logical, and factor variables like survey defendants might bounce questions, subsequent in NA values. The R likewise supports typed NA values, just like NA_integer_, NA_real_, NA_character_, and NA_complex_. These guarantee that the misplaced value ties the variable's data type, that is particularly valuable while carrying out strict type checks or uniting datasets.

One more significant kind is NaN (Not a Number). It happens while a mathematical procedure produces a vague outcome, like dividing zero by zero (0/0) or captivating the logarithm of a negative number. Though NaN signifies an unacceptable numeric result, R take it as a distinct arrangement of NA. So, the function is.na() proceeds TRUE for NaN, whereas is.nan() is precisely utilised to spot NaN values.

The NULL is diverse since NA and characterizes the comprehensive absence of a value or entity. This is usually come upon in lists, function results, or when an entity not primed. Disparate of NA, NULL has no span and not inhabit space in data frames. Eliminating an element from a list frequently fallouts in the NULL. For the reason NULL is characteristically cast-off in programming logic despite in the statistical analysis.

The R comprises Inf denoted as "positive infinity" and -Inf denoted as "negative infinity". These are non-technically missing ideals, but then characterize values which is surpass the max or min numerical bounds. It frequently rise from operations like division by zero or logarithmic transformations. Whereas Inf and -Inf are lawful numeric ideals in R, so they can misrepresent statistical outcomes if not held correctly and are frequently dried in the same way to missing values throughout preprocessing.

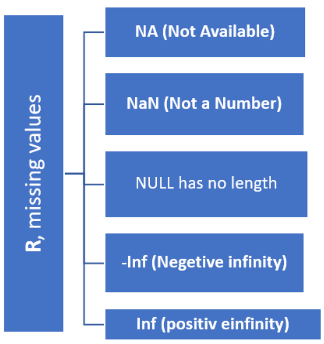

Figure 1(A): Type of missing values in R

As per built-in types, the missing data in R language could be defined abstractly as MCAR "Missing Completely at Random", MAR "Missing at Random", and MNAR "Missing Not at Random". Such classes elucidate wherefor data are missing and guide the assortment of suitable attribution or modelling practices.

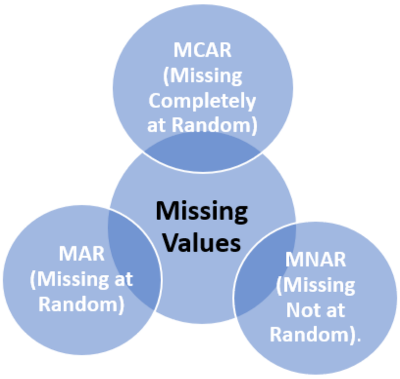

Figure 1(**B**): Type of missing values in R

The R gives manifold machineries to characterize missing / undefined data, with NA, NaN, NULL, and infinite values. Appropriate recognition and treatment of such kinds are vital for precise data analysis, consistent statistical inference, and vigorous machine learning representations.

IV. HANDLING MISSING VALUES FOR MACHINE LEARNING ALGORITHS IN R

The below mentioned table gives summary of approaches for handling of the missing values in R. It is evidently mapping respectively method to its resolution and applicable functions / packages. The Deletion/ Listwise Deletion is a candid method which eliminates rows comprising NA values and is appropriate while the dataset is huge and the number of missing values is little. This is usually applied by utilising na.omit() and complete.cases().

The Mean and Median Imputation is utilised to substitutes missing numerical values by measurement of central tendency, by the median favoured while data comprise outliers. The functions replace(), dplyr::mutate(), and tools as of the imputeTS package are naturally utilised. In order to deal categorical data, Mode Imputation substitutes missing data with the furthermost, frequently applied by the use of plyr package /custom-defined mode functions.

Table1: Techniques for Handling Missing Values, ML Algorithms in R

Method / Technique	Description	R Functions / Packages
Deletion (Listwise Deletion)	Removes rows with NA values. Suitable when dataset is large and missing proportion is small.	na.omit(), complete.cases()
Mean / Median Imputation	Replace missing numeric values with mean or median. Median preferred when data has outliers.	replace (), dplyr::mutate (), imputeTS
Mode Imputation (Categorical)	Replace missing categorical values with most frequent category.	plyr, custom mode function
KNN Imputation	Fills missing values using similarity to nearest Neighbors; more accurate for nonlinear data.	VIM::kNN(), DMwR::knnImputation()
MICE (Multiple Imputation)	Creates multiple imputations to preserve data variability; used in advanced statistical modelling.	mice::mice()
Predictive Model Imputation	Build regression/classification models to predict missing values.	caret, mice, randomForest
Interpolation (Time-series)	Fills missing values based on trends in time-series data.	imputeTS::na_interpolation()

One of the progressive methods is shown as, the KNN Imputation handles the missing data on the basis of similarity to nearest neighbours and accomplishes sound for nonlinear data by utilising functions as VIM::kNN() and DMwR::knnImputation().

The MICE also known as Multiple Imputation is publicized as statistically vigorous technique that makes manifold imputed datasets to reserve data erraticism, executed by function mice::mice(). The predictive Model Imputation utilised classification / regression approach by packages caret, mice, and randomForest to guess missing data. The interpolation is offered as a focused method for time-series data, here the missing data are imputed on the basis of temporal tendencies by the use of function imputeTS::na_interpolation().

V. MISSING VALUES OUTLOOK IN R PACKAGE

The M L algorithms and related function to treat the missing values in R language. The below given table is projecting the commonly used R packages. The treatment of missing values differs across M L algorithms in R language, basically depend on the concern function and structure. The linear regression, is accomplished through the lm function under stats package, which reduces the Mean Squared Error. It is not able to treat the missing values straightly so, annotations holding NA, which is removed by the use of functions na.omit() or dried by imputation methods like mean or median swap. Likewise, Logistic Regression, obtainable by the function glm, which optimizes the log loss /cross-entropy and severely needs comprehensive cases earlier to model fitting, which makes erstwhile data cleaning essential.

The algorithms like Naïve Bayes, executed by the package e1071, which depend on maximum likelihood estimation. It proposed partial backing to treat missing values, frequently snubbing them throughout probability approximation. Though, unambiguous imputation is usually suggested to keep prophecy trustworthiness. The k-Nearest Neighbors (kNN) utilised distance-based loss functions which is obtainable by class and caret packages. But still, it is not able to deal missing values due to inadequate data distort distance scheming. Similarly, Support Vector Machines (SVM) which is founded on hinge loss or ε-insensitive loss, needed completely pre-processed datasets.

The tree-based algorithms are much vigorous to deal missing values. The Decision Trees, employed with package rpart, which augment standard like entropy, Gini or MSE and natively treat missing values via substitute splits. The Random Forest algorithm executed with package randomForest and outspreads competence by amassing several trees and internally handling missing values throughout node splitting. The Gradient Boosting Machines employed with package gbm, it takes limited support to treat missing data within reliant on the loss function utilised.

The cutting-edge boosting base like XGBoost, LightGBM, and CatBoost are particularly operative. The XGBoost acquires the ideal way for missing values throughout tree creation, whereas LightGBM proficiently deal missing data deprived of overt imputation. The

CatBoost streamlines pre-processing by auto-treatment of missing data and categorical features. The ANN, PCA and K-Means Clustering are not directly able to deal missing values and need imputation or masking approaches

It is observed that Tree-based models in R (rpart, randomForest, xgboost, lightgbm, catboost) are utmost good to treat the missing values itself. Whereas Linear, distance-based, and neural models need overt imputation by use of packages mice, missForest, or caret.

Table 2: M L Algorithms and concern function for handling missing values in R Type Styles

ML Algorithm	Concern / Objective Function	Handling of Missing Values in R	R Package(s)	Remarks
Linear Regression	Mean Squared Error (MSE)	Cannot handle directly	stats (lm)	Missing values must be removed (na.omit) or imputed
Logistic Regression	Log Loss (Cross-Entropy)	Cannot handle directly	stats (glm)	Requires complete cases before model fitting
Naïve Bayes	Maximum Likelihood Estimation	Limited support	e1071	Missing values often ignored; imputation recommended
k-Nearest Neighbors (kNN)	Distance-based loss	Cannot handle directly	class, caret	Missing values distort distance calculations
Support Vector Machine (SVM)	Hinge Loss / ε-insensitive loss	Cannot handle directly	e1071	Data must be preprocessed
Decision Tree	Gini Index / Entropy / MSE	Can handle natively	rpart	Uses surrogate splits to manage missing values
Random Forest	Aggregated Tree Loss	Can handle natively	randomForest	Internally handles missing values via splits
Gradient Boosting Machine (GBM)	Custom loss (deviance, RMSE)	Partial support	gbm	Handles missing values internally to some extent
XGBoost	Regularized Objective (RMSE, Log Loss)	Can handle natively	xgboost	Learns optimal direction for missing values
LightGBM	Gradient-based objective	Can handle natively	lightgbm	Efficient handling without explicit imputation
CatBoost	Log Loss / RMSE	Can handle natively	catboost	Automatically processes missing values
Artificial Neural Network (ANN)	MSE / Cross-Entropy	Cannot handle directly	nnet, keras	Requires imputation or masking
Principal Component Analysis(PCA)	Reconstruction Error	Cannot handle directly	Stats, FactoMineR	Missing values must be imputed
K-Means Clustering	Within-Cluster SSE	Cannot handle directly	stats (kmeans)	Requires complete numeric data

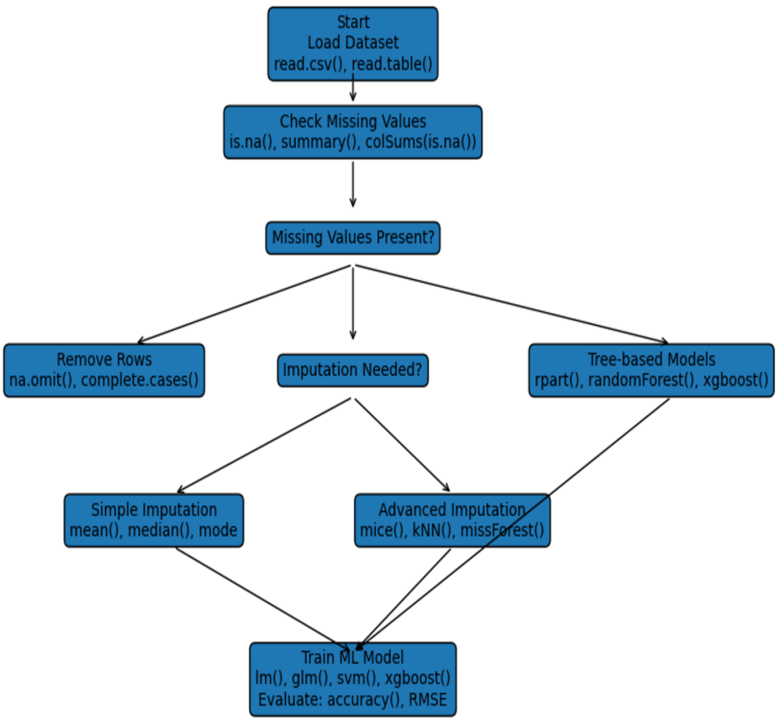

Figure 2: R-based M L workflows, handling missing values

VI. WORKFLOWS, HANDLING MISSING VALUES

In R-based machine learning systems, treatment of missing data instigates thru data loading. The dataset is bring-in by function function read.csv() or read.table(). When the dataset encumbered, subsequent phase is detection of missing value. This will be accomplished with functions is.na() to recognize missing data, summary() to get an summary of missingness crossways to the variables, and colSums(is.na()) to count missing data attribute-wise.

Later-on, the finding of the missing data, analysts might pick row removal option if amount of missing data is little. For sake of this function na.omit() and complete.cases() are utilised to remove observations covering missing case and confirm that complete records are remained. Though, if there is chance of data damage then imputation approaches are favored. The general approach includes substituting missing values by mean(), median(), or mode of the attribute.

In order to deal complex missing case advanced imputation approaches are taken in consideration. Packages like mice() empower manifold imputation by shackled equations, kNN() deals with k-nearest neighbors–based imputation, similarly missForest() utilised to deal random forest algorithm for forecasting of missing values. The direct modeling methods uses algorithms which are natively treat missing values, as decision trees by rpart(), ensemble approach by randomForest(), or boosting approaches as xgboost().

The treated data are utilised for model training and assessment. Functions like lm(), glm(), and svm()are functional through model performance measuring the evaluation metrics like RMSE and accuracy.

VII. R CODE SNIPPETS FOR HANDLING MISSING VALUES

Beforehand put-on M L algorithms in R, it is important to treat missing values. The initial stage is inspection of missing values; this gives assistances in understanding the degree and outline of missingness. The functions like sum(is.na(data)) gives the sum of missing values, whereas colSums(is.na(data)) provides a attribute-wise total, allowing analysts to recognize attributes which need distinct care.

If the ratio of missingness is little with randomness then removal of missing values is utilised. In such order na.omit() function is usually considered to remove rows comprising any NA values. This becomes resultant in a cleaned dataset. Though, this technique led to data loss while missingness is considerable.

To reserve the data imputation approach is extensively utilised. The simplest and unbiased way is to mean and median imputation. As missing data for age attribute can be substituted by using mean values as mean(data$age, na.rm = TRUE), whereas skewed attributes like salary is well imputed by use of median. Intended for categorical variables, mode substitution is utilised by recognizing the utmost class by ensuring consistency in qualitative data.

The erudite treatment, kNN imputation is hired by use of VIM package. This technique substitutes missing values on the basis of resemblances with neighbouring notes. This makes dataset suitable for multifaceted associations. One more advanced technique is Multiple Imputation by use of Chained Equations (MICE). This is applied by the mice package which produces numerous imputed datasets by means of predictive mean matching (pmm) and associated them in a comprehensive dataset

Check Missing Values
sum(is.na(data)) colSums(is.na(data))
Remove Missing Values
data_clean <- na.omit(data)
Mean / Median Imputation
data$age[is.na(data$age)] <- mean(data$age, na.rm = TRUE) data$salary[is.na(data$salary)] <- median(data$salary, na.rm = TRUE)
Mode Imputation (Categorical)
mode_value <- names(sort(table(data$gender), decreasing = TRUE)) data$gender[is.na(data$gender)] <- mode_value
kNN Imputation
library(VIM) data_knn <- kNN(data, k = 5)
Multiple Imputation (MICE)
library(mice) imputed_data <- mice(data, m = 5, method = "pmm") data_complete <- complete(imputed_data)
Random Forest Imputation
library(missForest) data_rf <- missForest(data)$ximp

Table 3: R Code Snippets for Handling Missing Values

The erudite treatment, kNN imputation is hired by use of VIM package. This technique substitutes missing values on the basis of resemblances with neighbouring notes. This makes dataset suitable for multifaceted associations. One more advanced technique is Multiple Imputation by use of Chained Equations (MICE). This is applied by the mice package which produces numerous imputed datasets by means of predictive mean matching (pmm) and associated them in a comprehensive dataset.

The Random Forest–based imputation is employed by missForest package. It forecasts missing values by utilising an ensemble of DT. This technique is extremely strong and works fine for mixed data types. These methods guarantee superior data grounding previously to put on machine learning models.

VIII. R METHODS FOR HANDLING MISSING VALUES AND THEIR PROS AND CONS

The treatment missing values in R is vital for precise data analysis and modelling. The language R delivers numerous approaches, since basic finding by use of is.na() and complete.cases() to omission approaches such as na.omit() and na.exclude(), which eliminate imperfect tuples but might upshot in sloss of data. The modest imputation methods, like missing data whereas upholding data structure. The picturing tools as VIM and naniar gives assistance to recognize the missingness forms and provide guide for substituting missing data with the central tendency reserve dataset magnitude but it can slightly misrepresent associations. In the advancement form it includes mice and missForest which used to predictive modelling to guess improved attribution approaches for consistent analyses and vigorous models. The pros and cons are as given in the table.

Table 4: R methods for handling missing values and their pros and cons

Method	Description	Pros	Cons
`is.na()`	Detects missing values in a dataset.	Simple, fast, helps locate missing data.	Does not handle or impute missing values; only identifies them.
`complete.cases()`	Returns rows with no missing values.	Easy to filter complete observations; useful for quick checks.	Removes entire rows with any missing value; can lead to data loss.
`na.omit()` / `na.exclude()`	Deletes rows containing missing values.	Simple implementation; avoids errors in functions that do not allow NAs.	May discard large portions of data; can bias results if missingness is not random.
Mean/Median/Mode Imputation	Replaces missing values with column mean, median, or mode.	Easy to implement; preserves dataset size.	Can distort variance and relationships; not suitable for complex datasets.
`mice` **(Multivariate Imputation by Chained Equations)**	Predictive imputation using other variables iteratively.	Preserves relationships in data; suitable for multivariate datasets.	Computationally intensive; requires careful parameter tuning.
`missForest`	Random forest–based imputation for mixed data types.	Handles non-linear relationships; works with both categorical and numeric data.	Slower on very large datasets; may overfit in small datasets.
`VIM` / `naniar`	Visualization of missingness patterns.	Provides insight into missing data structure; guides imputation strategy.	Does not impute by itself; primarily exploratory.

IX. Conclusions

Treatment of missing data is a serious phase in data pre-processing, as inappropriate handling can put negotiation the accurateness, trustworthiness, and interpretability of statistical studies and machine learning models. The R gives an all-inclusive bionetwork of tools to discourse the missing data, fluctuating from modest detection and deletion approaches to cutting-edge imputation methods. The functions as is.na() and complete.cases() permits for effectual recognition of missing data, whereas na.omit() and na.exclude() allow upfront elimination of imperfect cases. Meant for preserving data integrity, attribution approaches utilising mean, median, or prophetic modelling via packages as mice and missForest which proposed sophisticated results that uphold the construction and associations within datasets. The visualization tools like VIM and naniar gives assistance to recognise pattern for missingness. Through leveraging these abilities, assessments can guarantee vigorous data treatment foremost to reliable perceptions. The enhanced model performance, and sturdier decision-making in study and industry submissions.

X. References

[1] DD Pandya and S Gaur (2018), Detection of Anomalous value in Data Mining, kalpa in Engineering, Kalpa Publication, Vol.-2, Pp-1-6, 2018

[2] Madan Lal Yadav, and Basav Roychoudhury, Handling missing values: A study of popular imputation packages in R, Knowledge-Based Systems, Vol.- 160, 15 November 2018, Pages 104-118.

[3] S Gaur, DD Pandya and D Soni (2020), Closest Fit Approach through Linear Interpolation to Recover Missing Values in Data Mining, ICICT-2019, Advances in Intelligent Systems and Computing, AISC, Vol. 1041, Pp-513-521 Springer, Singapore. doi.org/ 10.1007/978-981-15-0637-6_44.

[4] Imke Mayer, Aude Sportisse, Julie Josse, Nicholas Tierney and Nathalie Vialaneix (2022), R-miss-tastic: a unified platform for missing values methods and workflows, The R Journal, Vol. 14, No-/2, 244-266, June 2022

[5] Moritz and T. Bartz-Beielstein (2017), imputeTS: Time Series Missing Value Imputation in R, The R Journal, vol-9, no-1, pp-207–218, 2017. doi: 10.32614/RJ-2017-009. [p246]

[6] S Gaur and MS Dulawat (2010), A Perception of Statistical Inference in Data Mining, Int. Journal of Computer Science & Communication, Vol.-1, No-II, Pp-635-658, (2010)

[7] S Gaur, MS Dulawat (2010) , Univariate Analysis for Data Preparation in Context of Missing Values, Journal of Computer & Mathematical Sciences, vol-1, no-5, Pp-628-635.

[8] S Gaur and M. S. Dulawat (2011), Improved Closest Fit Techniques to Handle Missing Attribute Values, Journal of Computer & Mathematical Sciences, Vol-2, No-2, Pp-170-198, april 2011

[9] N. Tierney, D. Cook, M. McBain, and C. Fay.(2021) NANIAR: Data Structures, Summaries, and Visualisations for Missing Data, 2021. URL https://CRAN.R-project.org/package=naniar. R package version 0.6.1. [p244, 252]

[10] Y. Xie, A. Presmanes Hill, and A. Thomas (2017), Blogdown: Creating Websites with R Markdow, The R Series, Chapman and Hall/CRC, 2017. ISBN 978-0815363729. [p245]

[11] S Gaur, DD Pandya, MK Sharma (2020), Applied N F Interpolation Method for Recover Randomly Missing values in Data Mining, Conference proceeding of ICICT-2019, Advances in Intelligent Systems and Computing, AISC, Vol. 1027, Pp-475-485 Springer, Singapore. doi.org/10.1007/978-981-32-9343-4_38

[12] S Sharma and S Gaur (2013), Contiguous Agile Approach to Manage Odd size Missing Block in Data Mining, International Journal of Advanced Research in Computer Science, Vol-4, no-11, pp 214-217, Nov-Dec, 2013.

