SURVEY UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Missing Value Imputation Techniques: A Survey

Wafaa Mustafa Hameed[1,2*], **Nzar A. Ali**[2,3]

[1]*Technical College of Informatics, Sulaimani Polytechnic University, Sulaimani, 46001, Kurdistan Region, Iraq,*
[2]*Department of Computer Science, Cihan University Sulaimaniya, Sulaimaniya, 46001, Kurdistan Region, Iraq,*
[3]*Department of Statistics and informatics, University of Sulaimani, Sulaimani, 46001, Kurdistan Region, Iraq*

ABSTRACT

Numerous of information is being accumulated and placed away every day. Big quantity of misplaced areas in a dataset might be a large problem confronted through analysts due to the fact it could cause numerous issues in quantitative investigates. To handle such misplaced values, numerous methods were proposed. This paper offers a review on different techniques available for imputation of unknown information, such as median imputation, hot (cold) deck imputation, regression imputation, expectation maximization, help vector device imputation, multivariate imputation using chained equation, SICE method, reinforcement programming, non-parametric iterative imputation algorithms, and multilayer perceptrons. This paper also explores a few satisfactory choices of methods to estimate missing values to be used by different researchers on this discipline of study. Furthermore, it aims to assist them to discern out what approach is commonly used now, the overview may additionally provide a view of every technique alongside its blessings and limitations to take into consideration of future studies on this area of study. It can be taking into account as baseline to solutions the question which techniques were used and that is the maximum popular.

Index Terms: Data Preprocessing, Imputation, Mean, Mode, Categorical Data, Numerical Data

1. INTRODUCTION

Data mining has made an amazing development in the past years; however, the main problem is missing data or value. Data mining is the sector wherein experimental facts sets are analyzed to find out thrilling and potentially beneficial relationships [1]. Lacking records or value in a datasets can affect the performance of classifier which ends up in difficulty of extracting beneficial information from datasets. Plentiful of facts is being gathered and saved each day. Those facts can be used to extract interesting patterns. The information that we collect is incomplete normally [2]. Therefor everyone wishing to apply statistical information evaluation or information cleaning of any type could have problems with lacking data. We still land in some missing attribute values in a function dataset. People typically tend to depart the income area empty in surveys, for instance, and members once in a while do not have any information available or cannot answer the question. Plenty facts can also be lost in the technique of collecting information from multiple resources [1]. Using that information to collect some statistics can now yield misleading effects. Hence, to eliminate the abnormalities, we need to pre-method the statistics earlier than the usage of it. Those instances may be omitted within the case of a small percentage of lacking values, but within the case of huge quantities, ignoring them will now not yield the desired outcome. A number of missing spaces in a dataset is a massive problem. Therefore, a few pre-processing of statistics can be accomplished earlier than acting any information mining techniques to extract a few treasured records from a dataset to keep away from such mistakes and as a result enhance statistics first-class. Fittingly managing with misplaced values is crucial and difficult venture since it requires careful examination of all occurrences of information to recognize design of missingness within the

Access this article online

DOI: 10.21928/uhdjst.v7n1y2023.pp72-81
E-ISSN: 2521-4217
P-ISSN: 2521-4209

Corresponding author's e-mail: Technical College of Informatics, Sulaimani Polytechnic University, Department of Computer Science, Cihan University Sulaimaniya, Sulaimaniya, 46001, Kurdistan Region, Iraq. E-mail id: wafaa.mustafa@spu.edu.iq

Received: 09-11-2022 Accepted: 05-03-2023 Published: 28-03-2023

data. Numerous strategies were proposed to address such lacking values considering 1980 [2].

This file illustrates distinct varieties of lacking values and the techniques used to address them.

It is tremendously vital to note that there may be evaluation in purge and lost value. Purge value implies that no value may be doled out though misplaced value implies actual value for that variable exists but not reachable or captured in dataset due to some motives. The information mineworker should separate between purge esteem and lost esteem. Once in a while, each the values may be treated as misplaced values. Lost records may be due to tools glitch, conflicting with different facts so erased, data no longer entered because of false impression, positive facts might not be considered crucial at the time of statistics collection. a few statistics mining calculations do not require substitution of misplaced values as they are planned and created to handle lost values; however, some data mining calculations cannot good buy with lost values. Sometime, these days making use of any strategy of managing with lost values its miles vital to get it why records is misplaced [2], [3].

2. MISSING VALUE PATTERNS

2.1. Missing Completely at Random (MCAR)

MCAR is the most improved degree of randomness and it indicates that the layout of misplaced value is completely arbitrary and does not rely on any variable which may additionally or might not be covered inside the examination [3]. It refers to facts that do not rely on the interest variable or every other parameter observed inside the dataset [4]. While missing values are distributed uniformly across all measurements, then we find the records to be absolutely randomly missing. For this reason, a brief test is to compare pieces of data – one with missing observations and the other without missing observations. On a t-test, if there is no mean difference between the two data units, we will expect that the data are MCAR [5]. Anything that is missing and sometimes because this form of missing facts is not often observed and the best manner to ignore these instances, for example: Water damage to paper forms due to flooding before it enters [1], [2] or in a survey, if we get 5% responses missing randomly, it is MCAR [6], [7]. This type is described by using the equation

$$P\,(\,p_{1}\;|X,Y_{0,l},Y_{m,l})\;=f\,(\,l\,,X\,)$$

Where f is a function, that is, the missing data patterns are determined only by the covariate variables X. Note here that MARX is equivalent to MCAR if there are no covariates in the model [7], [8].

2.2. Missing at Random (MAR)

When missed value does not rely on any given or ignored value [8]. Often information may not be deliberately missing; however, it can be named “missing at random”. If the data meet the requirement that missingness should not rely on X’s value after accounting for some other parameter, we may also find an X entry to be missing at random. Depressed people seem to have less income, as an instance, and the reported earnings now depend on the thing depression. The percentage of lacking records among depressed people could be high as depressed people have lower incomes [1] if we get 10% missing for the male responses in a survey and 5% missing for the woman survey, then it is MAR [6]. this kind is defined through the equation

$$P\left(p_l|X,Y_{0,l},Y_{m,l}\right)=f(l,X,Y_{0,l})$$

In which f is a function, that is, only the covariate variables X and the based variables has been located have an impact on the patterns of lacking statistics. Remember the fact that if there may be most effective one dependent variable Y then there may be best one missing series that does not encompass any found dependent variables. For models with one structured variable, MAR is therefore equal to MARX [7].

2.3. Not Missing at Random (NMAR)

If the data are not missing at random or informatively, it is labeled “not missing at random.” This kind of situation happens while the technique of messiness depends at the actual value of missing statistics [4]. This type is defined by the equation:

$$P\left(p_l|X,y_{0,l},Y_{m,l}\right)=f(l,X,Y_{0,l},Y_{m,l})$$

Where f is a function, that is, all three types of variables have an effect on the missing data patterns. It is well known how full information maximum likelihood (FIML) estimation performs under all of these conditions [7].

2.4. Missing in Cluster (MIC)

Data are regularly more missing in some attributes than in others. In addition, the missing values in the ones attributes can be correlated. It is extremely tough to use statistical techniques to show multi-attribute correlations of lacking values. On this sample of missing values, the exceptional of statistics is much less homogeneous than that with MAR.

The effects of any applications of analytical based on the complete facts set have to be cautious, for the reason that pattern data are biased in the attributes with a big number of missing values [7], [8].

2.5. Systematic Irregular Missing (SIM)

Data can be missing quite irregularly, however systematically. There is probably overly missing correlations among the attributes, but those correlations are extraordinarily tiresome to analyze. An implication of SIM is that the data with complete entities are unpredictably under-representative [7]. The first-class of records with this sample of missing values is minimal homogeneous than the ones in MAR and additionally less controllable than that with MIC. Applications of any analytical results based at the whole data set are enormously questionable [9].

3. STRATEGIES OF HANDLING MISSING DATA

Managing missing data may be carried out in two exclusive strategies for. The first method is definitely ignoring missing values and second approach is to take into account imputation of missing values.

3.1. Ignoring Missing Values

The missing records ignoring technique absolutely releases the state that includes missing data. They are mightily used for handling lacking facts. The earnest problem with this method is that it decreases the dataset size. This is handy whilst the dataset has small amount of lacking values. There are two common methods for ignoring missing data:

3.1.1. Listwise deletion

Complete case analysis approach excludes all observations with missing values for any variable of interest. This approach thus limits the analysis to those observations for which all values are observed. This techniques is simple to use but cause loss of huge data, loss of precision, high effect on variability, and induce bias.

3.1.2. Pairwise deletion

For all the instances, we perform analysis with in which the variables of interest are present. It does no longer exclude complete unit but uses as lots data as feasible from every unit. This method is straightforward, keeping all available values, that is, best missing values are deleted but motive the loss of data, no longer a higher solution compared to other techniques. The pattern size for every individual evaluation is better than the entire case analysis [2], [10].

3.2. Single Imputation

Single imputation procedures produce a precise value for a dataset's missing real value. This method necessitates a lower computing cost. Researchers have proposed a variety of single imputation strategies. The typical strategy is to analyze other responses and select the greatest possible response. The value can be calculated using the mean, median, or mode of the variable's available values. Single imputation can also be done using other methods, such as machine learning-based techniques. Imputed values are considered actual values in single imputation. Single imputation ignores the reality that no imputation method can guarantee the true value. Single imputation approaches ignore the imputed values' uncertainty. Instead, in future analysis, they recognize the imputed values as actual values [11], [12].

3.3. Multiple Imputations

The use of distinct simulation models, multiple imputation methods yield several values for the imputation of single missing records. Those strategies use imputed data's variability to generate a diffusion of credible responses. Multiple imputation strategies are sophisticated in nature, but in contrast to single imputation, they do no longer suffer from bias values. In multiple imputations, every missing facts point is replaced with m values obtained through m iterations (wherein m > 1 and m generally sits between 3 and 10) [6]. In this technique, a statistical approach used for coping with missing values, it performs through three stages:

- Imputation: Generate *m* imputed data sets from a distribution which results in *m* complete data sets. The distribution can be different for each missing entry.
- Analysis: In this stage each *m* imputed data Sets the analysis is performed, it is known as complete data analysis.
- Pooling: Use simple rules the output obtained after data analysis is pooled to get final result.

The resulting inferences form this stage is statistically valid if the methods to create imputations are "decent."

For substituting missing values with possible solutions, the multiple imputation method is used. The missing data set is transformed into complete data set using suitable imputation methods that can then be analyzed by any standard analysis method.

Therefore, multiple imputations have become popular in the handling of missing data. In this method, the process is repeated multiple times for all variables having missing values as the name indicates and then analyzed to combine m number of imputed data set into one imputed data set [7], [11].

4. MISSING VALUE IMPUTATION TECHNIQUE

4.1. Mean Imputation

Using this technique, calculate the mean of missing value through using the corresponding attribute value. This technique is easy to apply; it is built in maximum of the statistical bundle and quicker comparing with other techniques. It introduces precise result when facts is small, but it provides not proper result for large facts, this technique is appropriate for only MAR but no longer beneficial for MCAR [8], [13].

$$\hat{x}_{ij} = \sum_{i:x_{ij} \in c_k} \frac{x_{ij}}{n_k}$$

Wherein n_k represents the number of non-missing values within the j-th feature of the *k-th* class C_k, is missing [7], [8].

4.2. Hot (Cold) Deck Imputation

The concept, in this case, is to use some criteria of similarity to cluster the data earlier than executing the data imputation. This is one of the most used strategies.

Hot deck strategies impute missing values inside a data matrix by way of the usage of available values from the equal matrix. The item, from which these available values are taken for imputation within some other, is referred to as the donor. The replication of values ends in the trouble, that a single donor might be selected to accommodate multiple recipients. The inherent risk posed through that is that too many, or even all, missing values can be imputed with the values from a single donor. To mitigate this chance, a few hot deck variants restrict the amount of times anyone donor may be selected for donating its values. The similar techniques of hot deck are cold deck imputation method which takes other data source than current dataset. Using hot deck, the missing values are imputed by realistically obtained values which avoids distortion in distribution, but bit empirical work for accuracy estimation, creates problem if any other sample has no close relation in entire manner of the dataset [8], [10], [11].

4.3. Median Imputation (MDI)

Due to the affected of the mean through the presence of outliers, it seems better to use the median rather simply to make certain robustness. In this situation, the missing data are changed through the median of all recognized values of that attribute within the class where the instance with the missing characteristic belongs. This method is likewise a considered as a choice whilst the distribution of the values of is skewed. Assume that the value x_{ij} of the *k-th* class, C_k, is missing. It will get replaced by means of Singh and Prasad [7].

$$\hat{u}\ddot{y}_{ij}\ddot{u} = \left(i:x_{ij} \in c_k\right)\left\{x_{ij}\right\}$$

4.4. Regression Imputation

This approach may be apply by the use of known values for the construction of model after which calculates the regression between variables ends with applying that technique to calculate the missing values. The outcomes from applying this technique give greater accurate than mean imputation. The calculated data saves deviations from mean and distribution shape but the degree of freedom gets distort and can increases relationship [10].

$$Y = \alpha 0 + \alpha 1 X$$

4.5. Expectation Maximization Imputation (EMI)

There are forms of clustering algorithms. One is soft clustering and other is hard clustering:-

- *Soft clustering:* Clusters may overlap that is with unique degree of belief the factors belong to multiple clusters at the identical time
- *Hard clustering:* Clusters do now not overlap that's mean the element either belong to a cluster or not.
- *Mixture models:* The use of a probabilistic manner for doing soft clustering. Every cluster corresponds to a generative model this is usually Gaussian or multinomial, MVs are imputed by realistically obtained values which avoids distortion in distribution, in this technique, bit empirical work for accuracy estimation creates problem if any other sample has no close relation in entire manner of the dataset [2].

4.6. K-nearest Neighbor Imputation (KNN)

Specifying the similarity between two values and replace the missing value with similar one using Euclidean distance. The advantage of this technique that for the datasets which having both qualitative and quantitative attributes values KNN is suitable. There is no need for creating a predictive model for each attribute of missing data and helpful for multiple missing values.

The KNN looks for the most similar instances, the algorithm searches through all of the data set and that consider as an obstacle for that approach [12].

4.7. Fuzzy K-means Clustering Imputation (FKMI)

In this method, the membership characteristic plays an important position. It is assigned with every data item that describes in what degree the data object is belonging to the precise cluster. data items might not get assign to concrete cluster which is stated using centroid of cluster (i.e., the case of k means), that is due to the various membership degrees of every data with entire k clusters. Unreferenced attributes for every uncompleted data are changing by FKMI on the premise of membership degrees and cluster centroid values. The pros of this approach is that it offers quality outcome for overlapping data, higher than k manner imputation and records objects may be a part of multiple cluster middle but the high computation time and noise sensitive, that is, low or no membership degree for noisy objects considered as a cones for the usage of this technique [10].

4.8. Support Vector Machine Imputation (SVMI)

Its regression primarily based technique to impute the missing values. It takes condition attributes (output) and decision attributes. Then, the SVMI would be carried out for prediction of values of missed condition features. Advantages of this approach are the efficient in massive dimensional areas and efficient memory consumption; however, additionally, there may be a cons for using this technique which it is the bad performance if number of samples are plenty lesser than number of features [10], [14].

4.9. Most Common Imputation (MCI)

On this imputation method, clustered are first shaped by applying k-means clustering method. Like in k-NN, on this method, the nearest neighbors are found using clusters. All the instances in every cluster are referred as nearest neighbor of each other. Then, the missing value is imputed the usage of the same technique as is employed through KNNI imputation approach. This procedure is fast and therefore is ideal for applying in big datasets. This algorithm reduces the intra cluster variance to minimum. Here, too value of k parameter is an important factor and is difficult to predict its value. In addition, this algorithm does no longer assure global minimal variance [2], [15], [16].

4.10. Multivariate Imputation by Chained Equation (MICE)

MICE expect that data are lost arbitrarily (damage). It imagines the likelihood of a missing variable depends on the watched facts. MICE offers numerous values in the put of one lost esteem through making an arrangement of relapse (or other reasonable) models, tallying on its “method” parameter. In MICE, every lost variable is treated as a variable, and other information inside the record is treated as an independent variable. At to begin with, MICE foresee missing values utilizing the winning information of other factors. At that point, it replaces missing values utilizing the predicted values and makes a dataset known as ascribed dataset. By cycle, it makes numerous ascribed datasets. Every dataset is at that factor analyzed utilizing standard measurable investigation techniques, and numerous investigation comes about are given [17], [18].

4.11. SICE Technique

It pretends the probability of a missing variable depends on the determined data. It gives multiple values within the place of one missing value through creating a sequence of regression models, each missing variable is treated as a dependent variable, and different data in the record are treated as an independent variable, it predicts missing data using the existing data of other variables. Then, it replaces missing values using the predicted values and creates a dataset known as imputed dataset. It achieves 20% higher F-measure for binary data imputation and 11% less errors for numeric data imputations than its competitors with similar execution time. It imputes binary, ordinal and numeric data. It performed well for the imputation of binary and numeric data and fantastic preference for missing data imputation, especially for massive datasets where MICE is impractical to use because of its complexity but it could not show better overall performance than MICE for the case of ordinal data [6].

4.12. Reinforcement Programming

Impute missing data using learning a policy to impute data thru an action-reward-based totally experience imputes missing values in a column by operating best on the identical column (similar to univarite single imputation) however imputes the missing values within the column with different values thus keeping the variance in the imputed values. It is usually used for dynamic approach for the calculation of missing values using machine learning procedures. It has functionality of convergence and to solving imputation problem through using exploration and exploitation [19], [20].

4.13. Utilizing Uncertainty Aware Predictors And Adversarial Learning MIP UA-Adv.

Impute the missing values so that the adversarial neural network cannot distinguish real values from imputed ones. In addition, to account for the uncertainty of imputed values, the usage of confidence scores acquired from the adversarial module. The adversarial module objectives to discriminate imputed values from real ones the resulting imputer in addition to estimating a missing entry with high accuracy, it

Table 1: Short review with mentioning to the advantage and disadvantage of different techniques to handle missing value

Techniques	Note	Advantages	Limitations	References
Leastwise deletion technique	- Deletion of cases containing missing values (complete row is deleted) high missing information because of deletion of entire row high impact on variability loss of precision and induce bias.	- Simple to use.	- Loss of precision, - Loss of enormous data - High effect on variability, - Induce bias	[2], [10]
Pair- wise deletion technique	- Deletion of records best from column containing missing values much less lack of information by using keeping all available values less impact on variability less loss of precision and induce bias.	- Keeping all available values only missing values are deleted. - Simple to use.	- Not a better solution as compared to other methods. - Loss of data,	[2], [10]
Mean Imputation technique	- Calculate the mean of missing value through using the corresponding attribute value. Replace MVs with the mean of facts Resultant may be better than that of original.	- It is built in maximum of the statistical bundle and quicker comparing with other techniques. - It introduce precise result when facts is small	- It provides not proper result for large facts this technique is appropriate for only MAR but no longer beneficial for MCAR - Affected by the presence of outliers.	[3], [8]
Median imputation (MDI) technique	- Missing data replaced by the median of all observed values of that attribute in the class where the features belongs.	- Good choice when the distribution of the values is skewed.	- Not affect by the presence of outlier	[7]
Hot (cold) deck imputation technique	- Cluster the data earlier than executing the data imputation. - Impute missing values inside a data matrix by way of the usage of available values from the equal matrix	- Avoid distortion in distribution.	- Empirical for accuracy estimation. - creates problem if any other sample has no close relation in entire manner of the dataset.	[8], [10], [11]
Regression imputation technique	- Use the known values for the construction between variables then applying the technique to calculate the missing values	- Very easy and simple technique. - Calculated data saves deviations from mean and distribution shape	- Only applicable if data is linearly separable that is there is linear relation between attributes. - Degree of freedom gets distort and may raises relationship.	[10]
Expectation maximization (EM) technique	- Iterative method, finds maximum likelihood Two steps: Expectation (E step), Maximization (M step) using three models soft, hard and mixture clustering Iteration goes on until algorithm converges	- MVs are imputed by realistically obtained values which avoids distortion in distribution	- Bit empirical work for accuracy estimation, creates problem if any other sample has no close relation in entire manner of the dataset	[2]
Fuzzy K- means clustering Imputation (FKMI) technique	- It is assigned with every data item that describes in what degree the data object is belonging to the precise cluster. - Unreferenced attributes for every uncompleted data are substituted by FKMI on the basis of membership degrees and cluster centroid values.	- Best outcome for overlapping data, better than k means imputation. Data objects may be part of more than one cluster center	- High computation time. - Noise sensitive, that is, low or no membership degree for noisy objects	[10]
- Support Vector Machine Imputation (SVMI) technique	- Takes condition attributes (here, decision attribute i.e., output) and decision attributes (here, conditional attributes) SVMI then would be applied for prediction of values of missed condition attribute	- Efficient in large dimensional spaces. - Efficient memory consumption	- Poor performance if number of samples are much less than number of feature	[10], [14]

(Contd...)

Table 1: (*Continued*)

Techniques	Note	Advantages	Limitations	References
K nearest neighbour imputation (KNN) technique	- Determining the similarity between two values and replace the missing data with similar one using Euclidean.	- Avoids distortion in distribution as missing values are imputed by realistically obtained values - No need for creating a predictive model. - Helpful for multiple missing value	- Obstacle approach since the algorithm search all of the data set - Prediction of value of k is quite a difficult task.	[12]
Most Common Imputation (MCI) technique	- It replaces the missing value by the most common attribute or by the mode. - While the numerical attribute missing value replaced by the average of the mean corresponding attribute	- Fast and good for applying in big dataset. - Reduce the intra cluster variance to minimum.	- Difficult to predict the value if the number of elements too big. - Dose not guarantee global minimum variance.	[2], [15], [16]
Multivariate Imputation by Chained Equation (MICE) technique	- It pretends the probability of a missing variable depends on the observed data. it provides multiple values in the place of one missing value by creating a series of regression models, - Each missing variable is treated as a dependent variable, and other data in the record are treated as an independent variable - Predict missing data using the existing data of other variables. Then it replaces missing values using the predicted values and creates a dataset called imputed dataset	- Flexibility: each variable can be modeled using a model tailored to its distribution. - Can manage imputation of variables defined only on a subset of the data, - Can also incorporate variables that are functions of other variables, - It does not require monotone missing- data patterns.	- Lacking a theoretical rationale - Difficulties encountered when specifying the different imputation models	[17], [18]
SICE technique:	It is an extension of the popular MICE algorithm. Two variants of SICE presented: SICE- Categorical and SICE- Numeric to impute binary, ordinal, and numeric data. Twelve existing Performance of algorithms implemented to predict house prices imputation methods and compare their performance with SICE.	- Achieves 20% higher F- measure for binary data imputation and 11% less error for numeric data imputations than its competitors with similar execution time. Impute binary, ordinal, and numeric data. - Performed better for the imputation of binary and numeric data. - Excellent choice for missing data imputation, especially for massive datasets where MICE is impractical to use because of its complexity	- It could not show better performance than MICE for the case of ordinal data.	[6]
Reinforcement programming technique	Impute data through an action-reward- based experience imputes missing values in a column by working only on the same column but imputes the missing values in the column with different values thus keeping the variance in the imputed values. It is generally used for dynamic approach for the calculation of missing values by using machine learning approaches.	- Performs well compared to other univarite single imputation and ML- based imputation approaches.	- Use of numeric data variables only	[19], [20]

(Contd...)

Table 1: (*Continued*)

Techniques	Note	Advantages	Limitations	References
Utilizing uncertainty aware predictors and adversarial learning MLP UA- Adv Imputer	- Train well with small and large datasets and utilizes a novel adversarial strategy to estimate the uncertainty of imputed data - Proposed a novel hybrid loss function that enforces the imputers to generate values for missing data that on the one hand, obey the underlying data distribution so that it can confuse the well- trained adversarial module, and on the other hand, predict existing non- missing values accurately - The run time of the methods shows that they are efficient and have less execution time in comparison with that of peer imputer models.	- Plays an important role in the overall performance - Less runtime compared to other imputers - Has a very simple structure, can work with any feature type and small and large data set	- It did not consider the imbalanced nature of the imputation task.	[19], [21]

Table 2: Comparing different techniques according to the dataset used in the application

Datasets	Techniques	Notes	References
Iris	Mean Regression Imputation; Reinforcement Programming technique	A comparison of different approaches of MICE methods on iris datasets. Efficiency gain with multiple imputations combined with regression is that it can better use the available information by accommodating non- linarites	[3], [8], [10], [18], [19], [20]

Iris
Credits
Adults
	Mean/Mode;
	Hot Deck;
	Expectation Maximization;
	K- nearest neighbor
		In this paper, the authors compare C5.0 with this newly developed technique known as IITMV and show its performance on different data sets	[3], [8], [10], [11], [12], [22]

Cleveland
Heart
Zoo
Buhl1- 300
Glass
Ionosphere
Iris
Pima
Sonar
WaveForm2
Wine
Hayes- Roth
Led7
Solar
Soybean
	Mean/mode;
	Regression;
	Hot deck;
	MLP UA- Adv
		The result shows that multilayer perceptions (MLP) with different learning rules show better results with quantitative datasets than classical imputation methods. In this paper, the type of missing value is missing completely at random (MCAR)	[3], [8], [10], [11], [19], [21], [22]

Iris *Escherichia coli* Breast cancer 1 Breast cancer 2	Mean K- nearest neighbors (KNN) Fuzzy K- means (FKM) Multiple imputations by chained equations (MICE) MLP UA- Adv	The results show that different techniques are best for different datasets and sizes. MICE are useful for small datasets, but, for big ones and FKM are better, the MLP UA- Adv is better for both small and big datasets	[3], [8], [10], [12], [17], [18], [19], [21], [23]

be able to confuse the adversarial module, it neural network based totally architecture that can train properly with small and large datasets and to estimate the uncertainty of imputed data [19], [21].

5. REVIEW ON MISSING VALUE IMPUTATION METHODS

Table 1.

6. CONCLUSION

The finding of this article summarized in Tables 1 and 2, the article shows that the most popular techniques (mean, KNN, and MICE) are not necessarily the most efficient. It isn't always surprising for mean in regards to the simplicity of the method: The technique does not make use of the underlying correlation structure of the information and for that reason plays poorly. KNN represents a natural improvement of mean that exploits the observed facts structure. MICE are complex algorithm and its behavior seems to be related to the size of the dataset: Rapid and efficient on the small datasets, its overall performance decreases and it becomes time-intensive when carried out to the massive datasets. The more than one imputation combined with Bayesian Regression gives better performance than other strategies, which includes mean, KNN. However, they only taken into consideration the great of imputation based totally on category strategies without worrying of the execution time that may be an exclude criterion. Consequently, FKM may additionally represent the technique of choice but its execution time may be a drag to its use and we take into account bPCA as a more adapted solution to high-dimensional data, the article also shows that the MLP UA-Adv consider a good choice for large and small data set also with different data type. Table 2 shows comparison between the techniques according applications and the dataset used in each one. The strength of this paper that its cover most of the missing value imputation techniques that can be taken into consideration as a reference for other researcher to pick out the most appropriate techniques or make combination from a couple of for imputing the missing values.

REFERENCES

[1] B. Doshi. Handling Missing Values in Data Mining. Rochester Institute of Technology, Rochester, New York, U S A, 2010. Available from: https://www.pdfs.semanticscholar.org/3817/b208fe1f40891cc661ea0db80c8fccc56b70.pdf [Last accessed on 2023 Mar 27].

[2] S. Gupta and M. K. Gupta. "A survey on different techniques for handling missing values in dataset". *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 4, no. 1, pp. 2456-3307, 2018.

[3] A. Jadhav, D. Pramod and K. Ramanathan. "Comparison of performance of data imputation methods for numeric dataset". *Applied Artificial Intelligence*, vol. 33, no. 10, pp. 913-933, 2019.

[4] J. Scheffer. "Dealing with missing data". *Research Letters in the Information and Mathematical Sciences*, vol. 3, pp. 153-160, 2002.

[5] D. V. Patil. "Multiple imputation of missing data with genetic algorithm based techniques". *IJCA Special Issue on Evolutionary Computation*, vol. 2, pp. 74-78, 2010.

[6] S. I. Khan and A. S. Hoque. "SICE: An improved missing data imputation technique." *Journal of Big Data*, vol. 7, no. 1, p. 37, 2020.

[7] S. Singh and J. Prasad. "Estimation of missing values in the data mining and comparison of imputation methods." *Mathematical Journal of Interdisciplinary Sciences*, vol. 1, no. 2, pp. 75-90, 2013.

[8] I. Pratama, A. E. Permanasari, I. Ardiyanto and R. Indrayani. A Review of Missing Values Handling Methods on Time Series Data, in: International Conference on Information Technology Systems and Innovation (ICITSI). Bandung, Bali, IEEE, 2016, p. 6.

[9] S. Wang and H. Wang. Mining Data Quality in Completeness. University of Massachusetts Dartmouth, United States of America, 2007. Available from: https://www.pdfs.semanticscholar.org/347c/f73908217751c8d5c617ae964fdcb87674c3.pdf [Last accessed on 2023 Mar 27].

[10] R. L. Vaishnav and K. M. Patel. "Analysis of various techniques to handling missing value in dataset". *International Journal of Innovative and Emerging Research in Engineering*, vol. 2, no. 2, pp. 191-195, 2015.

[11] A. Raghunath. Survey Sampling Theory and Applications. Academic Press, Cambridge, 2017.

[12] Holman and C. A. Glas. "Modelling non-ignorable missing-data mechanisms with item response theory models". *British Journal of Mathematical and Statistical Psychology*, vol. 58, no. 1, pp. 1-17, 2005.

[13] A. Puri and M. Gupta. "Review on missing value imputation techniques in data mining. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 2, no. 7, pp. 35-40, 2017.

[14] S. Van Buuren and K. Groothuis-Oudshoorn. "MICE: Multivariate imputation by chained equations in R". *Journal of Statistical Software*, vol. 45, no. 3, pp. 1-67, 2010.

[15] A. S. Kumar and G. V. Akrishna. "Internet of things based clinical decision support system using data mining techniques". *Journal of Advanced Research in Dynamical and Control Systems*, vol. 10, no. 4, pp. 132-139, 2018.

[16] J. W. Grzymala-Busse, L. K. Goodwin, W. J. Grzymala-Busse and X. Zheng. Handling Missing Attribute Values in Preterm Birth Data Sets. Vol. 3642. United Nations Academic Impact, New York, 2005, pp. 342-351.

[17] J. Han, M. Kamber and J. Pei. Data Mining: Concepts and Techniques. 3rd ed. Morgan Kaufmann Publishers, San Francisco, CA, USA, 2012.

[18] G. Chhabra, V. Vashisht and J. Ranjan. "A comparison of multiple imputation methods for data with missing values". *Indian Journal of Science and Technology*, vol. 10, no. 19, pp. 1-7, 2017.

[19] S. E. Awan, M. Bennamoun, F. Sohel, F. Sanfilippo and G. Dwivedi. "A reinforcement learning-based approach for imputing missing data". *Neural Computing and Applications*, vol. 34, pp. 9701-9716, 2022.

[20] I. E. W. Rachmawan and A. R. Barakbah. Optimization of Missing Value Imputation using Reinforcement Programming, in:

International Electronics Symposium (IES). Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, 2015, pp. 128-133.

[21] W. M. Hameed and N. A. Ali. “Enhancing imputation techniques performance utilizing uncertainty aware predictors and adversarial learning”. *Periodicals of Engineering and Natural Sciences*, vol. 10, no. 3, pp. 350-367, 2022.

[22] T. Aljuaid and S. Sasi. Intelligent Imputation Technique for Missing Values, in: Conference on Advances in Computing, Communications and Informatics (ICACCI). Jaipur, India, pp. 2441-2445, 2016.

[23] P. Schmitt, J. Mandel and M. Guedj. “A comparison of six methods for missing data imputation”. *Journal of Biometrics and Biostatistics*, vol. 6, no. 1, pp. 1, 2015.