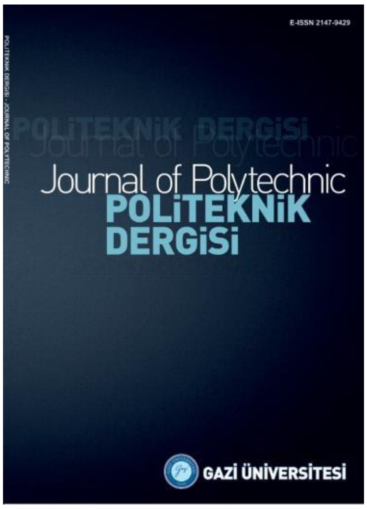


POLİTEKNİK DERGİSİ

JOURNAL of POLYTECHNIC

ISSN: 1302-0900 (PRINT), ISSN: 2147-9429 (ONLINE)
URL: http://dergipark.org.tr/politeknik

A novel data imputation method (m-cbri) for industrial analytics applications

Endüstriyel analitik uygulamaları için eksik verilere değer atama (m-cbri)

Yazar(lar) (Author(s)): *Mehmet Alper ŞAHİN*[1], *Uğur ÜRESİN* [2]

ORCID[1]*:0000-0003-1196-8765*

ORCID[2]*:0000-0002-9100-9697*

To cite to this article: Sahin M.A., Uresin U., "A novel data imputation method (m-cbri) for industrial analytics applications", ***Journal of Polytechnic***, 27(6): 2225-2229, (2024).

Bu makaleye şu şekilde atıfta bulunabilirsiniz: Sahin M.A., Uresin U., "A novel data imputation method (m-cbri) for industrial analytics applications", ***Politeknik Dergisi***, 27(6): 2225-2229, (2024).

Erişim linki (To link to this article): http://dergipark.org.tr/politeknik/archive

DOI: 10.2339/politeknik.1201559

A Novel Data Imputation Method (M-CBRI) for Industrial Analytics Applications

Highlights

- ❖ *Several imputation methods are compared with the proposed method using real datasets from three different processes in a large automotive manufacturer.*
- ❖ *It has been revealed that proposed method is significantly successful in completing the missing data.*

Graphical Abstract

Standard error represents the prediction capability of imputation methods. The smaller the standard error, the more accurate the assigning value is.

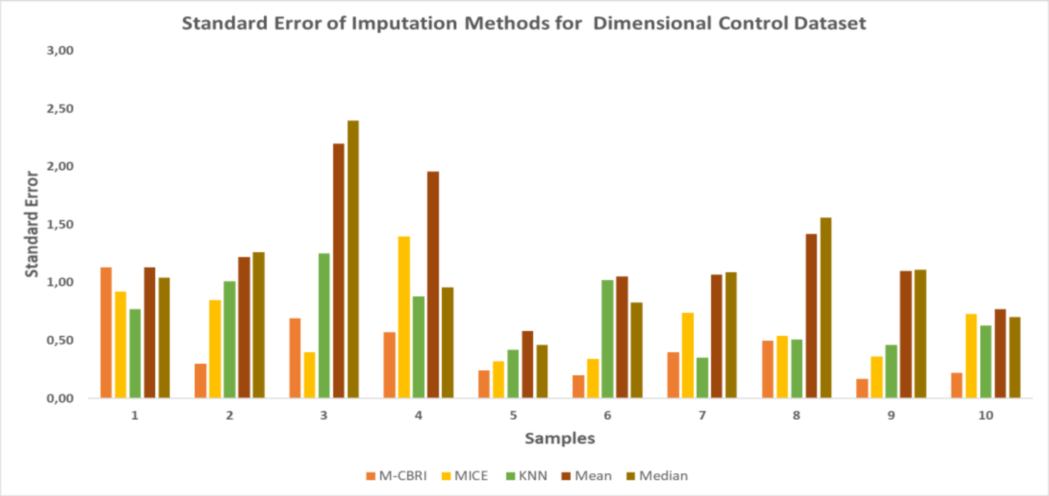


Figure. Standard error of imputation methods for dimensional control dataset

Aim

The aim of this study is to fit correlation-based regression imputation method for missing data imputation by using three different datasets.

Design & Methodology

1. *Detection of missing values*
2. *Creating Train & Validation Sets (without missing values) and Test data (includes missing values).*
3. *Determining highly correlated parameters which link to missing value.*
4. *Equating the multi-linear equation to predict missing values.*
5. *Confirm results on a validation set.*
6. *Fitting the acceptable multi-linear equation to test set*
7. *Filling missing values by linear equation.*

Originality

The originality of this study is the comparison of the performance of correlation-based regression imputation method with mean, median, k-nearest neighbor, and multivariate imputation by chained equation.

Findings

Assignment values with the proposed method gives better results compared to mean assignment, median assignment, k-nearest neighbor assignment and multivariate imputation by chained equations for small datasets.

Conclusion

The proposed method creates a multi-linear regression to estimate the missing values based on the linear correlation between the parameters. By comparison with other methods, the prediction performance of the proposed method enables assigning missing values more accurately.

Declaration of Ethical Standards

The authors of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

A Novel Data Imputation Method (M-CBRI) for Industrial Analytic Applications

Research Article / Araştırma Makalesi

Mehmet Alper ŞAHİN[1*], Uğur ÜRESİN[1]

[1]Ford Otosan Research and Development Center, Istanbul, Turkey

(Geliş/Received: 16.11.2022 ; Kabul/Accepted: 03.07.2023 ; Erken Görünüm/Early View: 15.03.2024)

ABSTRACT

Data analysis is mainly based on understanding and preprocessing the data coming from various sources for various applications. Missing values might play a critical role to reflect to characteristic of datasets; thus, imputation of missing values is a valuable process to not only handle reducing deviation but also avoid loss of data. There are different approaches to filling missing values. One of them is correlation-based imputation method. This approach is based on the high correlation between the parameters, these parameters are variables of linear equation, the linear equation enables to predict missing values. In this study, improvements were made to the correlation-based imputation method to predict missing values. The proposed method was performed on three various datasets which are related to the automotive industry. Missing values are handled in a manual process, and these values are picked randomly from the real data. After generating missing values, missing values are predicted using the correlation-based imputation method; furthermore, the margin of error between the estimated value and actual value was calculated. The results were compared to different methods which are arithmetic mean assignment, median value assignment, k- nearest neighbor assignment, and multivariate imputation by chained equations; consequently, much more successful results were obtained with the proposed method for three datasets.

Keywords: Missing data imputation, data preprocessing, missing value, data imputation, industrial data processing.

Endüstriyel Analitik Uygulamaları için Eksik Verilere Değer Atama (M-CBRI)

ÖZ

Veri analitiği çalışmalarının ilk aşamaları, veriyi toplama, veriyi analiz etme ve veriyi temizleme şeklindedir. Toplanan verilerin, farklı kaynaklardan elde edilmesi ve veri kaynaklarındaki kesilmeler, veriseti içerisinde eksik değerlerin oluşmasına sebep olabilmektedir. Bununla birlikte, veriyi temizleme çalışmalarında bazı aykırı değerlerin verisetinden çıkarılması da yine eksik değerlerin oluşmasına yol açmaktadır. Veride yer alan eksik değerler, analitik uygulamalarda elde edilmek istenen çıktılarda sapmalara sebep olabilir. Hem bu sapmayı azaltmak hem de toplanan veride kayıp yaşamamak adına eksik verilerin giderilmesi önemli bir süreçtir. Literatürde, eksik verilerin yerine değer atanması konusunda pek çok yöntem yer almaktadır ama söz konusu yöntemlerden uygun olanın seçilmesi tecrübe ve uzmanlık gerektirmektedir. Bu çalışmada, eksik verileri tahminlemek adına doğrusal korelasyona bağlı değer atama algoritması üzerinden geliştirmeler yapılmıştır. Bu algoritma, bir otomotiv üretecisinin farklı proseslerinden elde edilen üç farklı gerçek veriseti üzerinde test edilmiştir. Verisetlerinden rastgele silinen veriler, geliştirilen yöntemler yardımıyla tahminlenmiştir ve tahminlenen değer ile gerçek değer arasındaki hata payı hesaplanmıştır. Geliştirilen algoritmanın sonuçları, ortalama değer atama, medyan değer atama, en yakın komşuya göre değer atama ve zincir denklemlerle çok değişkenli değer atama yöntemleriyle karşılaştırılmıştır. Üç veriseti için de, geliştirilen yöntemin diğer yöntemlere göre daha başarılı tahminde bulunduğu gözlemlenmiştir.

Anahtar Kelimeler: Eksik verilere değer atama, veri ön işleme, eksik veri, değer atama, endüstriyel veri işleme.

1. INTRODUCTION

Within development in the computer field, data is a valuable key concept to observe and simulate the real world. From finance to the automotive industry, almost in all sectors, data plays a significant role.[1] In machine learning applications, creating added value from data depends on its quality of it; consequently, data gathering, data storing, data wrangling, and data analysis are prioritized by organizations. In an industrial environment, there are many data sources such as sensors, databases, flat files, PLC, and industrial controllers etc. [2] However, not always these data resources yield complete datasets. Even though the majority of missing values are caused by manual data entry, missing data might have occurred due to several factors such as human error, noise generation during transformation, equipment and measurement error, lack of response [3, 4, 5, 6]. The first approach is eliminating missing values and using the rest of the data as a complete dataset [7]. Even though this approach is quite simple compared to other approaches, eliminating missing values means not only loss of useful information but also decreasing performance of models [8]. Thus, the outputs of the model may not be beneficial for the use case. Data imputation is preferred because of the many reasons, which are referred above, rather than removing missing

**Sorumlu Yazar (Corresponding Author)*

e-mail : msahin42@ford.com.tr

data points [9]. Data imputation means handling missing values to preserve the data and prepare the raw data for the analysis by imputation missing values with plausible values [7]. In addition, the meaning of the original characteristic of dataset is preserved without destroying the underlying characteristic [10]. One of the popular methods to assigned missing values are statistical principles. Statistical behavior depends upon the observation dataset; consequently, statistical inference and interpolation might not be effective imputation technique [11]. In the mean assignment method, missing values of the variable are replaced with the mean of the variable. In the median assignment method, missing values of the variable are filled with median of the variable. Correlation between the features is ignored by using these methods; thus, mean, and median imputation might lead to poor imputation [12].

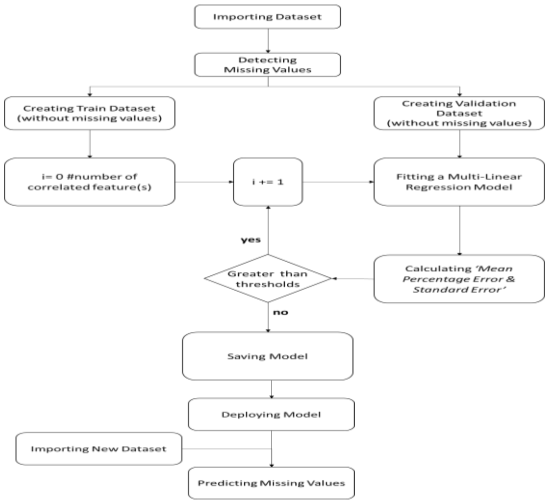
These methods can be implemented easily to handle missing values. K-nearest neighbor method is based on finding the k closest neighbors to the observation with missing data and then predict it based on the non-missing values in the neighbors [13, 14]. Another method is multivariate imputation by chained equation (MICE) which is set under the assumption that the missing data are Missing at Random (MAR). MAR means that missing values are only linked to observed variables. [15] In the first step, missing values are assigned with statistical methods such as imputing the mean. These methods can be thought of as ‘place holders. Missing data is dependent variable and other variables in the dataset are independent variables for regression model. Assigned data is replaced with prediction of regression model [16, 17]. These methods make possible to both keep all data points and explore all values by data engineer, basic statistical methods, such as replacing missing values to mean, the median, nearest neighbor of features and m imputation by chained equations (MICE), these methods might not handle missing values efficiently since these methods are sensitive to outliers [16]. For instance, if datasets are not large enough to represent characteristics of features, filling missing values with either mean or median of features may occur to change characteristic of data [18]. Statistical methods (mean, median assignment) directly affect whether the parameter is a significant parameter for the datasets. Therefore, mean, and median imputation does not seem to be the best way where the collected data is scarce. [19] In a small dataset, these methods could make it difficult to reach optimum results. In addition to that, mean and median imputation assignment is used especially in relatively large datasets to fill in missing values. Even if there is an improvement in the prediction of missing values with mean or median imputation when the datasets are saturated, missing values may not reflect the general behavior of the parameters.[20] Standardizing of missing values might lead to misunderstanding between the features or prevent the model from running at maximum metrics. Similar to statistical principles, predicting the missing values with methods such as k-nearest neighbor might impact the score of models negatively. As the number of neighboring samples increases, the result of the k-nearest neighbor will differ from the previous results. To obtain satisfying results with high precision and low variability, datasets should be enlarged. Nonetheless, data generation and data collection can take much time, time is the main constraint on many industrial projects in general.

To address these issues, we proposed a novel method which considers multiple relationships between the features that allow better predictions of the missing values. This proposed method was analyzed for three different datasets and its results were compared with common methods in literature. Data from different lines of an automotive manufacturer form the basis of this study. The first datasets were collected from the robots in the welding line, the second dataset was taken from the dimensional control station in the quality line and the final dataset was gathered from the CNC machines.

2. THE PROPOSED METHOD

The proposed methods in this paper are based on strong correlations between features. First of all, the missing values in the datasets are detected. Detecting missing values are removed from the original datasets. Pearson’s pairwise correlations are calculated. The datasets which do not contain any missing values are split randomly into two subsets. One of the generated subsets was used to establish a multi-linear equation between the parameters to be estimated and the other parameters associated with the missing value. Another subset enables to test of multi-linear equation performance. If the prediction ability of the equation is not smaller than thresholds, the number of related parameters as input is increased. Multi-linear equation is formulated again, and mean percentage error and standard error are calculated, if still mean percentage error and standard error do not a desired level, previous steps are repeated until the mean percentage error and standard error are much smaller than thresholds. In the light of targeting thresholds, the optimum parameters, and formulation of the linear equation are obtained to predict missing values.

Flow chart 1: M-CBRI Methodology


Importing Dataset
Detecting Missing Values
Creating Train Dataset (without missing values)
Creating Validation Dataset (without missing values)
i= 0 #number of correlated feature(s)
i += 1
Fitting a Multi-Linear Regression Model
yes
Greater than thresholds
Calculating *‘Mean Percentage Error & Standard Error’*
no
Saving Model
Deploying Model
Importing New Dataset
Predicting Missing Values

M-CBRI method construct on correlation between columns which have missing values. Although there are many different methods to calculate correlation coefficient, Pearson product-moment correlation coefficient is used. M-CBRI methodology which is given in *"Flow chart 1"* is described step by step as a formulation in below.

1. **Calculating Pearson Correlation Coefficient**

Symbol	Definition
n	Sample Size
x_i, y_i	Individual sample points indexed with i
$\bar{x}, \bar{y}$	Mean of observing parameters
r_{xy}	Correlation coefficient between x and y

$$r_{xy} = \frac{\sum_{i=0}^{n} (x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i-1}^{n} (x_i - \bar{x})^2}\,\sqrt{\sum_{i-1}^{n} (y_i - \bar{y})^2}}$$

2. **Elimination of the self-correlations.**
3. **Build a Multiple Regression Model**
    - **Set the multiple regression model and predict the result of test data.**
    - **Calculate the Model Metrics which reflect to performance of model.**
        i. **Percentage Error**

$$\frac{1}{n}\sum_{i=0}^{n} \frac{|\hat{y_\iota} - y_i|}{y_i}$$

Symbol	Definition
$\hat{y}$	Predicted value of the missing value
y	Exact Value of the missing value

ii. **Standard Error**

$$\frac{1}{n}\sum_{i=0}^{n} \frac{|\widehat{y_\iota - y_i}|}{\sigma_y}$$

Symbol	Definition
$\hat{y}$	Predicted value of the missing value
y	Exact Value of the missing value
σ_y	Standard deviation of y in the training set.

4. **Repeating the operations described above by increasing the number of independent variables.**

5. **Comparing the error metrics with the previous model.**

6. **Saving the model which has a lower error metrics.**

3. RESULTS AND DISCUSSION

In this study, 10 different values which were randomly selected from the datasets were deleted from the data. To estimate these deleted values, linear equations which are based on the ' Multiple Correlation Based Regression Imputation' method (M-CBRI) given in flow chart 1, were used. The prediction results of the proposed method (M-CBRI) were also compared with other methods in the literature. The reason for observing the results on 3 different datasets collected from the production lines is that each dataset has unique characteristics.

3.1. Results on Quality Control Data

In the first dataset, the data is from an automobile company's dimensional control line, it is difficult to collect data from the line hence there is a discontinuous flow in the line. In a situation where the dataset is so small, eliminating the missing values will lead to a further reduction of the dataset. As for assigning values instead of missing data, if these values cannot be predicted with a low deviation, it may affect the statistical characteristics of the dataset misleadingly and mislead machine learning models. Therefore, the standard error metric has been analyzed to measure M-CBRI method against other methods. The differences in results between the M-CBRI method and other methods were compared regarding dimensional measurement data (328 rows & 66 columns).

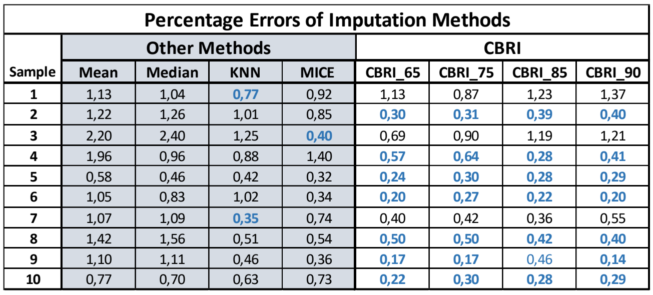
The standard error is calculated by equation 1.

Equation 1:

$$Percentage\ Error = \frac{|Actual\ Value - Predicted\ Value|}{|Actual\ Value|}$$

Obtained results are shown in table 1.

Table 1. Missing value imputation on quality control dataset



Percentage Errors of Imputation Methods								
	Other Methods				CBRI			
Sample	Mean	Median	KNN	MICE	CBRI_65	CBRI_75	CBRI_85	CBRI_90
1	1,13	1,04	0,77	0,92	1,13	0,87	1,23	1,37
2	1,22	1,26	1,01	0,85	0,30	0,31	0,39	0,40
3	2,20	2,40	1,25	0,40	0,69	0,90	1,19	1,21
4	1,96	0,96	0,88	1,40	0,57	0,64	0,28	0,41
5	0,58	0,46	0,42	0,32	0,24	0,30	0,28	0,29
6	1,05	0,83	1,02	0,34	0,20	0,27	0,22	0,20
7	1,07	1,09	0,35	0,74	0,40	0,42	0,36	0,55
8	1,42	1,56	0,51	0,54	0,50	0,50	0,42	0,40
9	1,10	1,11	0,46	0,36	0,17	0,17	0,46	0,14
10	0,77	0,70	0,63	0,73	0,22	0,30	0,28	0,29

When the table is examined, it is seen that using the M-CBRI method to estimate missing values gives more reliable prediction results.

3.2. Results on Welding Data

Unlike the first dataset, the second dataset has a continuous flow of welding robots, and these data are overwritten in a database. Therefore, it was possible to not only work with a large dataset but also observe the results of imputation methods compared to the first dataset. The second dataset consists of 10000 rows and 7 columns in total.

Table 2. Missing value imputation on welding dataset (Size = 10000)

Percentage of Imputation Methods								
	Other Methods				CBRI			
Sample	Mean	Median	KNN	MICE	CBRI_20	CBRI_40	CBRI_60	CBRI_85
1	0,15	0,12	0,07	**0,06**	**0,06**	**0,06**	0,07	**0,06**
2	0,10	0,07	**0,03**	0,04	0,04	0,04	0,04	0,04
3	0,10	0,08	**0,02**	**0,02**	0,04	0,04	0,04	0,04
4	0,11	0,12	**0,01**	**0,01**	0,02	**0,01**	**0,01**	**0,01**
5	0,15	0,14	**0,03**	0,05	0,06	0,06	0,06	0,06
6	0,12	0,13	0,02	**0,01**	0,02	0,02	0,02	0,02
7	0,21	0,25	**0,06**	0,19	0,17	0,17	0,18	0,17
8	0,27	0,28	**0,08**	0,15	0,45	0,24	0,25	0,45
9	0,08	0,10	0,08	**0,05**	0,06	0,06	0,06	0,06
10	0,10	0,07	**0,05**	0,06	**0,05**	**0,05**	**0,05**	**0,05**

The KNN method was estimated with less standard error in assigning values to replace the missing data. When table 2 is examined, it is seen that the standard errors of predicted values by M-CBRI method are remarkably close to the standard errors of KNN.

A small dataset was created with 200 randomly selected values from the welding data.

Table 3. Missing value imputation on welding dataset (Size =200)

Percentage Errors of Imputation Methods								
	Other Methods				CBRI			
Sample	Mean	Median	KNN	MICE	CBRI_20	CBRI_40	CBRI_60	CBRI_85
1	0,13	0,11	0,10	**0,03**	0,04	0,04	0,04	0,04
2	0,09	0,08	0,15	**0,06**	**0,06**	**0,06**	**0,06**	**0,06**
3	0,20	0,24	0,24	0,21	0,17	0,17	0,17	**0,16**
4	0,14	0,13	**0,05**	0,06	0,06	0,06	0,06	**0,05**
5	0,08	0,06	**0,04**	**0,04**	0,05	0,05	0,05	0,05
6	0,09	0,09	0,04	**0,01**	**0,01**	**0,01**	**0,01**	**0,01**
7	0,15	0,13	0,09	**0,02**	**0,02**	0,03	0,03	0,03
8	0,10	0,07	0,08	0,09	0,04	**0,03**	**0,03**	0,07
9	0,06	0,07	0,05	**0,01**	**0,01**	**0,01**	**0,01**	**0,01**
10	0,16	0,15	0,07	**0,03**	**0,03**	**0,03**	**0,03**	**0,03**

Considering the standard errors for the welding data, successful results are obtained with MICE and M-CBRI methods. Although the standard errors of both are close, the prediction of the M-CBRI method is more precise.

3.3. Results on Machining Data

The data that creates the final dataset contains information about the process on the CNC machines. Similar to the second dataset, the volume of the last dataset is also large (10000 rows & 5 columns). According to table 4, it is seen that the best results are obtained with the KNN method, since 10000 rows enable explore the dataset deeply to data scientist, it is not surprising that the best results are achieved with KNN.

Table 4. Missing Value Imputation on machining (CNC) dataset

Percentage Errors of Imputation Methods								
	Other Methods				CBRI			
Sample	Mean	Median	KNN	MICE	CBRI_10	CBRI_50	CBRI_65	CBRI_85
1	0,45	0,37	**0,01**	1,02	0,75	0,75	0,75	0,75
2	0,46	0,47	**0,01**	0,37	0,57	0,57	0,57	0,57
3	0,67	0,76	**0,62**	1,04	1,36	1,36	1,36	1,36
4	0,62	0,62	**0,04**	0,21	0,30	0,31	0,31	0,31
5	0,63	0,65	**0,18**	0,19	0,35	0,35	0,35	0,35
6	0,70	0,68	1,03	**0,50**	0,77	0,77	0,77	0,77
7	0,55	0,62	**0,00**	0,50	1,64	1,64	1,64	0,74
8	0,47	0,50	**0,02**	0,82	0,34	0,34	0,34	0,34
9	0,38	0,37	**0,01**	0,15	0,48	0,48	0,48	0,48
10	0,46	0,45	**0,02**	0,27	0,44	0,44	0,44	0,44

It has been pointed out in previous paragraphs that the proposed method in this article is useful for small datasets. If the data from CNC machines could not be collected efficiently due to any reason, handling missing values with the KNN method may not reflect the reality or may less reflect the actual results compare to the M-CBRI method.

To observe the consequences of this scenario, two small datasets were created with 300 randomly selected values out of 10000 data. The assignment of the missing values is estimated in the direction of the methods in the literature and the M-CBRI method.

Table 5. CNC machines Small Datasets Missing Value Imputation

(a: First Small Dataset, b: Second Small Dataset)

Percentage Errors of Imputation Methods								
	Other Methods				CBRI			
Sample	Mean	Median	KNN	MICE	CBRI_10	CBRI_30	CBRI_50	CBRI_85
1	0,37	0,37	**0,29**	0,30	0,30	0,30	0,32	0,31
2	0,37	0,38	0,27	0,12	**0,07**	**0,07**	**0,09**	**0,08**
3	0,37	0,36	**0,18**	**0,18**	0,21	0,21	0,24	0,24
4	0,44	0,43	0,20	0,20	**0,19**	**0,19**	0,24	0,21
5	**0,54**	0,56	0,78	0,85	0,68	0,68	0,61	0,65
6	0,28	0,29	0,33	0,17	**0,13**	**0,15**	**0,16**	**0,15**
7	0,53	0,56	0,31	**0,17**	0,21	0,22	0,22	0,34
8	0,42	0,43	0,40	0,23	**0,17**	**0,19**	**0,20**	**0,15**
9	0,58	**0,39**	1,00	1,76	1,52	1,50	2,70	0,73
10	0,39	0,41	0,50	0,27	**0,25**	**0,26**	**0,15**	**0,20**

(a)

Percentage Errors of Imputation Methods								
	Other Methods				CBRI			
Sample	Mean	Median	KNN	MICE	CBRI_10	CBRI_30	CBRI_50	CBRI_85
1	0,35	0,32	0,86	0,17	**0,11**	**0,12**	**0,13**	0,41
2	0,43	0,41	0,69	0,43	**0,24**	**0,24**	**0,24**	0,69
3	0,29	0,34	0,43	0,12	**0,11**	**0,11**	0,12	0,19
4	0,44	0,35	0,79	**0,25**	0,27	0,28	0,27	0,50
5	0,57	**0,51**	2,38	0,58	0,90	0,86	0,86	0,87
6	0,36	0,37	0,39	**0,29**	0,43	0,33	0,33	0,39
7	0,75	0,75	0,60	0,37	**0,29**	**0,30**	**0,30**	0,51
8	0,54	0,56	0,58	**0,21**	0,24	0,26	0,26	0,46
9	**0,45**	**0,45**	0,80	1,01	0,87	0,74	0,74	0,78
10	0,49	0,46	0,37	0,35	**0,33**	**0,33**	**0,34**	0,38

(b)

In consideration of table 5, it has been observed that the KNN algorithm cannot make accurate predictions for the small dataset although it was the best method for the big dataset in table 4. In contrast with other methods, the M-CBRI method can predict missing values with a low standard error.

4. CONCLUSION

In this paper, the M-CBRI method was tested on three different datasets consisting of not only the manufacturing processes but also the quality process of the automobile company. The performance of the proposed method was compared to the most common approaches in the data science field. These approaches are mean assignment, median assignment, k-nearest neighbor assignment, and multivariate imputation by chained equations.

When the standard error obtained for three small datasets is observed, assignment values with the M-CBRI method give better results. Filling missing values with a low standard error enables the preservation of the character of data. A more accurate assignment allows working with less biased data while running machine learning and deep learning models. As a result of this, the M-CBRI method might affect the performance metrics of models positively.

In the results of two relatively large datasets, it was seen that even though the M-CBRI method could not give as satisfactory results as the KNN algorithm.

In future studies, researchers can focus on estimating missing values by examining non-linear relationships between features. We believe that the correlation-based regression imputation method can be used to handle missing values in small datasets compared to the other methods.

DECLARATION OF ETHICAL STANDARDS

The authors of this article declare that the methods used in this study do not require ethical committee permission and/or legal-special permission.

AUTHORS' CONTRIBUTIONS

Mehmet Alper ŞAHİN: Analyzed the data, and compared proposal method and other common techniques, wrote the manuscript.

Uğur ÜRESİN: Analyzed the data, and compared proposal method and other common techniques, wrote the manuscript.

CONFLICT OF INTEREST

There is no conflict of interest in this study.

REFERENCES

[1] Tole A. A., "The Importance of Data Warehouses in the Development of Computerized Decision Support Solutions. A Comparison between Data Warehouses and Data Marts", ***Database Systems Journal***, Academy of Economic Studies - Bucharest, Romania, (2016).

[2] Foidl, H.& Felderer, M., "An Approach for Assessing Industrial IoT Data Sources to Determine Their Data Trustworthiness."

[3] Fouad, K. M., Ismail, M. M., Azar, A. T., & Arafa, M. M. "Advanced methods for missing values imputation based on similarity learning", ***PeerJ Computer Science***, 7, (2021).

[4] Rahman MG, Islam MZ. **"**Data quality improvement by imputation of missing values", ***International Conference on Computer Science and Information Technology.*** Yogyakarta, Indonesia, 82–88, (2013).

[5] Srivastava, A. K., Kumar, Y., & Singh, P. K, "Hybrid diabetes disease prediction framework based on data imputation and outlier detection techniques", ***Expert Systems***, (2022).

[6] Lakshminarayan, K., Harp, S.A. & Samad, T., "Imputation of Missing Data in Industrial Databases.", ***Applied Intelligence*** 11, 259–275, (1999).

[7] Jadhav, A., Pramod, D., & Ramanathan, Kr., "Comparison of Performance of Data Imputation Methods for Numeric Dataset. Applied Artificial Intelligence.", (2019).

[8] Armina, R., Mohd Zain, A., Ali, N. A., & Sallehuddin, R. "A Review on Missing Value Estimation Using Imputation Algorithm.", ***Journal of Physics: Conference Series***, 892, (2017).

[9] ***www.stat.columbia.edu***, *"*Missing-data imputation".

[10] Bania, R. K., Halder, A., "R-ensembler: A greedy rough set based ensemble attribute selection algorithm with KNN imputation for classification of Medical Data.", ***Computer Methods and Programs in Biomedicine***,184, (2020).

[11] Buuren, S. *"Flexible Imputation of Missing Data*," Second Edition, (2018).

[12] Little, R. J. A., & Rubin, D. B. "Statistical Analysis with Missing Data." Third Edition, Wiley, (2019).

[13] Troyanskaya, O., et all., "Missing value estimation Methods for DNA microarrays.*"* ***Bioinformatics***, 520–525, (2001).

[14] Zhang, S., "Nearest neighbor selection for iteratively kNN imputation*.", **Journal of Systems and Software***, 2541–2552, (2012).

[15] Rubin, D.B, **"**Inference and missing data", ***Biometrika***, (1976).

[16] Azur, M.J., Stuart, E.A., Frangakis, C. and Leaf, P.J., "Multiple imputation by chained equations: what is it and how does it work?", ***International Journal of Methods in Psychiatric Research***, 40–49, (2011).

[17] Van Buuren S, K Groothuis-Oudshoorn, Leerstoel Van Buuren, & And, M., "mice: Multivariate Imputation by Chained Equations:", 259-268, (2012).

[18] Üresin, U., *"Correlation based regression imputation (CBRI) method for missing data imputation.", **Turkish Journal of Science and Technology**.,* (2021).

[19] Uttley J., "Power Analysis, Sample Size, and Assessment of Statistical Assumptions—Improving the Evidential Value of Lighting Research", 143-162 (2019).

[20] Gu, Y., Wei, H.-L., "A robust model structure selection method for small sample size and multiple datasets problems.", ***Information Sciences***, (2018).