Cumulative bayesian ridge for handling missing data

Samih M. Mostafa[a,*], Abdelrahman S. Mohamed[a] and Safwat Hamad [b]

[a]Mathematics Department, Faculty of Science, South Valley University, Qena, Eygpt
[b]Faculty of Computer and Information Sciences, Ain Shams University, Cairo, 11566, Egypt
*samih_montser@sci.svu.edu.eg

Abstract. Old approaches for functioning with missing values may drive to biased estimations and may also decrease or magnify statistical influence. To each of these misrepresentations may drive to unacceptable conclusions. The conduct of diverse missing value imputation algorithms may fluctuate for dissimilar datasets and may lean on the amount of the missing values in the dataset and the dataset's dimension. In, this paper the authors proposed a new algorithm for handling missing data and implement an unbiased study of some registered practical imputation methods used for handling missing value. The suggested algorithm is based on the Bayesian Ridge technique works in a cumulative order with a gain ratio feature selection in its kernel to select the candidate feature to be imputed; any imputed feature will be incorporated in Bayesian Ridge equation to impute missing values in the next chosen feature. If a missing value gives high imputation precision and requires less imputation time is considered better. On examining eight datasets with several missing values proportions formed from the three mechanisms, it was observed that the performances be different depending on the missingness mechanism, size, and missing proportion.

Keywords: missing value, imputation, missingness mechanism, Bayesian Ridge, Gain ratio.

1. Introduction

Avert missing data is the best means for management incomplete instances. All skilled researchers give a great carefulness in research procedures, in hire informants, and in increasing measures. However, best researchers still come across missing values that could take place for reasons we have not predictable. In the course of the data gathering stage, the researcher has the prospect to make choices about what data to gather, and how to screen data gathering. The distribution and scale of the features in the data and the whys and wherefores for missing data are two acute topics for employ the proper missing data methods[1].

Data preparation is the supreme vital and time consuming mission, which toughly impacts the success of the research. Feature selection lies in detecting a valuable subset of possible predictors from a huge set of candidates. Discarding features with an undue number of missing values (e.g. >50 %) is frequently a good rule of thumb, nevertheless it is not a risk-free route. Discarding a feature may drive to a loss of analytical power

and capability to observe statistically significant differences and it can be a foundation of bias, affecting the representativeness of the outcomes. For these aims, feature selection requests to be custom-made to the missing data mechanism. Imputation can be completed in advance or later of feature selection[2].

1.1. Motivation and novelty

Some imputation methods are unproductive in the imputation of the missing data; others require time for imputation or offers underprivileged attainment. This paper manages these weaknesses by offering a novel imputation method that exploits the powerful features. The precedence of features to be carefully chosen in the imputation was studied based on gain ratio, which will be explained in Section 3.

1.2. Contributions of this paper

The main impacts of this paper are: it offers a summary of the studies related to handling missing data, gives the weaknesses and gains of the earlier studies, indicates how the performance metrics influenced by the volume of the dataset and the proportion of missing data, offers an imputation method which profits from all the features to rise the fineness of data, and compares between the proposed algorithm and these studies.

1.3. Missingness mechanisms

To elect how to manage missing data, it is suitable to recognize why there are missing values in data. We consider three general missingness mechanisms[2]–[5]:

- MCAR: if the prospect of missingness is identical for all elements, e.g., if each survey defendant chooses whether to response the "earnings" question by rolling a die and rejecting to answer if a "5" shows up. In MAR, missing value deletion does not bias your conclusions[6].
- MAR: if the prospect a feature is missing leans only on existing information. Therefore, if race, sex, age, and education are documented for all the persons in the survey, at that moment "earnings" is missing at random if the prospect of not response to this question leans only on these other, totally documented variables. It is frequently rational to model this process as a logistic regression, where the outcome features matches 1 for detected cases and 0 for missing[6].
- MNAR There is two situations first, if it leans on information that has not been noted and this information also guesses the missing values. E.g., assume that "surly" people are less probable to respond to the earnings question, surliness is foretelling of earnings, and "surliness" is undetected[6]. Second, if it leans on the missing value itself. If the prospect of missingness leans on the (potentially missing) feature itself. E.g., assume that people with upper earnings are less probable to make public them. In the risky case (for example, all people earning more than $100,000 reject to answer), this is known as censoring[6].

1.4. Handling missing data

The finest approach for handling missing data is to avert it overall over watchful data gathering and follow-up, along with determining missing data after the fact (for example, by detecting missing forms or re-contacting study members). Though it is commonly impossible to avert missing data in total and hence statistical methods for managing missing data are required. Since missing data are exceptionally complex, statisticians cannot create a universal set of rules that functions for all cases. Rather, they run emulation to expect the finest method[7]. The approaches should be bespoke to the reasons for Missingness and the percentage of missing data. Overall, a method is selected for its ease and its capability to present as tiny bias as potential in the dataset[2]. The modest approach to handle missing data is to remove instances that involve missing values. In common, case deletion methods drive to valid conclusions only for MCAR[8]. Imputation is the other alternative approach for handling missing data and overcome the disadvantages of deletion approach.

As the proposed algorithm leans on Bayesian ridge regression, so it's a regression model plus a regularization parameter for the coefficients. The model satisfies the following:

$$y \sim N(\mu, \alpha^{-1}) \qquad (1)$$

Where:

$$\mu = \beta X = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p$$

$$\beta \sim N(0, \lambda^{-1} Ip)$$

$$\alpha \sim g(\alpha_1, \alpha_2)$$

$$\lambda \sim g(\lambda_1, \lambda_2)$$

2. Literature review

Data-dependent tools (Data mining or big data analysis …) have been known as significant and challenging tasks for various problems in existence. To accomplish any data-dependent tool, a particular dataset for a selected target problem is gathered[9]. Though, in practice, the gathered dataset habitually has some proportion of incomplete missing data.

For coping with data set that contains the missing value one can use deletion, Deletion might be 'complete deletion', 'list-wise deletion', or 'Complete Case Analysis', where all observations containing one or more of their feature values missing are erased or 'specific deletion' where only those observations are erased which contains more than a pre-specified threshold of their feature values that are missing[4], [7], [9], [10]. Also, there are 'feature deletion' or 'pair-wise deletion', in which the observations having missing values in the features incorporates in the in-

progress analysis are erased, these observations being however used for other analyses which do not involve the concerned features contains the missing values; in the risky situation of each feature ensuring missing values through lots of observations may affect in the deletion of the whole dataset[11]. To assess, the missing value Imputation approaches benefit from all the information existing in the selected dataset, in which an applicable value is imputed instead of the missed one[9], [10], [12]. The imputed value might be mean, mode, median or any pre-specified value of the feature that contains missing value[13]–[15], or might be obtained from case substitution. Imputed value can also be computed using KNN, regression models[3][16], cold deck imputation[17], EM (expectation maximization) imputation[18]–[20], hot deck imputation[21], etc. In techniques involving prediction models, a model is developed based on existing information, which is then used to predict appropriate values for the missing data[22]. If the missing values are of MAR or MCAR kind, and if each record or feature in the dataset is very important and a single record does not have missing values across many features Imputation methods are used to cope with it missing values[11]. For MCAR kind, there are no universal approaches to cope MNAR missing values kind, missing values in a dataset can be handled also by using deletion, preferably listwise deletion or maximum likelihood methods[4], [23].

Imputation types are single imputation or multiple imputation. In single imputation, a single applicable value is imputed rather than missed value[17]. Multiple imputation, in which, 'm' full datasets are gotten by imputing the missing values 'm' epochs, the concluding imputed dataset being the analysis average of these 'm' datasets[24]. Though multiple imputation, requires more resources[25]; it has benefits over other approaches, namely maximum likelihood techniques[26], single imputation, and deletion. Single imputation approaches cope with all data values even imputed ones as the true value which leads to inflated type I error as a result of not accounting uncertainty[4].

Inverse Probability Weighting (IPW) methods also good for managing missing data, which leans on the inverse of the detected probability to weight observed observations in this manner representing the entire data even the missing values. But the performance of imputation approach is better[27]. OLS, PLS, and SVD are also suitable for multiple imputation[4].

For datasets having binary and ordinal features MVNI and FCS approaches are generate similar results and usually less biased. model specification are easily provided by MVNI but as a result of its unrealistic nature more or less people may possibly have problem with it[4]. FCS tends to have complex model specification as a reason for requiring a single regression model for each feature whose missing value is going to be handled[28].

KNN is used by FINNIM which is an effective iterative multiple imputation approach to assess missing value[29]. Sequential regression trees are used as Multiple imputation conditional model which has the facility to detect complex relation and need slight correction by the user[30]. Handling missing value PMM is accomplished using a randomly drawn observation from a set of detected instances whose predictive mean is near to predictive mean of the value that are missing[4]. LRD approach handles the missing value using the predictive mean as PMM, with extra randomly drawn from the residuals of a set of detected instances with predictive means near to that of the missing value[31]. Approaches which depends on Reinforcement Learning like RP gives better performance as contrast to mean per category imputation, zero imputation, GA in terms of sum of square error and computational time[32]. Cumulative linear e regression

which leans on the linear regression algorithm handles missing data in a cumulative works well for small and large datasets[3].

Imputing the missing data in features of concern with the aid of detected values from other features. This method leans on the similarities of the instance values in the donor to cope with the missing data in the recipient. It works well when the proportion of the missing values is larger[33].

3. Experimental implementation

3.1. Datasets

The study performs an analysis on the effect of the proportion of missing values and dataset dimension on imputation time and the accuracy of such imputation. To accomplish this study the authors used eight different datasets that are usually used within the literature (Table 1).

Table 1. Datasets specifications

Dataset name	Instances	Features	References
diabetes	442	11	[34]
graduate admissions	500	8	[35]
profit estimation of companies	1000	6	[36]
red & white wine dataset	4898	12	[37]
California	20640	9	[38]
diamonds	53940	10	[39]
Poker Hand Dataset	1,025,010	11	[40]
BNG_heart_statlog	1,000,000	14	[41]

The missing values were generated from the three missingness mechanisms with different proportions 10%, 20%, 30% and 40% missingness ratios (MRs). Analysis for BNG_heart_statlog and Poker Hand datasets were completed on randomly sampled sub datasets of 10000, 15000, 20000 and 50000 of instances[4].

Table 2. List of terminologies

Terms	Description	Comments
n	number of all variables	Comp + Miss = n
$X^{(comp)}$	set of complete variables	
$X^{(mis)}$	set of incomplete variables	
$X_{imp}^{(mis)}$	imputed variable from X(Miss)	
c	number of complete independents	c +m = n
m	number of variables containing missing variables	
MissObs. y	set of missing observations in the dependent variable y	
MissObs. $X_i^{(mis)}$	set of missing observations in the independent variable Xi	

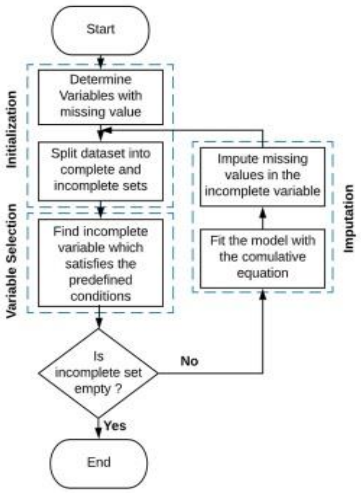

Figure 1. Algorithm flowchart

- ***Initialization***
  - Identify features that contain missing values.
    - $X^{(comp)} = \{ X_1^{(comp)}, X_2^{(comp)}, \cdots, X_c^{(comp)} \}$,and
    - $X^{(mis)} = \{ X_1^{(mis)}, X_2^{(mis)}, \cdots, X_m^{(mis)} \}$.
  - $y \in \begin{cases} X^{(comp)} & if\ MissObs.y = \emptyset \\ X^{(mis)} & otherwise \end{cases}$

- ***Feature selection***
  - From $X^{(mis)}$, find $X_i^{(mis)}, i \in \{1, \ldots, \mathrm{m}\}$, select the feature with:
    - Higher Gain Ratio $(X_i^{(mis)}, y)$

- ***Imputation***
  - For each column in $X^{(mis)}$:
    - Fit the model with cumulative Bayesian Ridge Regression equation:

$$X_g^{(miss)} \sim N\left(\mu_g\ ,\ \alpha^{-1}{}_g\right)$$

$$g = 1,2, \ldots, m$$

Where:

$$\mu_g = \beta_g X_g = \beta_o + \sum_{i=1}^{c} \beta_i X_i^{(comp)} + \beta_{c+1} y + \sum_{imp=1}^{g-1} \beta_{imp+c+1} X_{imp}^{(mis)\,2}$$

$$\beta_g \sim N\left(0, \lambda^{-1}{}_g\, Ip\right)$$

$$\alpha_g \sim \mathcal{g}(\alpha_{1g}, \alpha_{2g})$$

$$\lambda_g \sim \mathcal{g}(\lambda_1 g, \lambda_{2g})$$

    - Impute missing values.
  - Repeat until all missing values in all columns are imputed.

Figure 2. Algorithm: CBRG

3.2. *Feature selection* is an approach commonly used in machine learning, in which a lesser set of the features from the data are selected for use of a learning algorithm. Feature selection attains dimensionality reduction by choosing a lesser set of the main features[42]. The objective of feature selection takes account of building

simpler and more comprehensible models, making data-dependent tools performance better, and making clean, reasonable data[43].

Gain Ratio: Information gain main disadvantage is that it is biased towards picking features with lots of values. This encouraged Quinlan to define the Gain Ratio which reduces this bias[44].

$$IG(Ex,a) = H(Ex) - \sum_{v \in values(a)} \left(\frac{|\{x \in Ex | value(x,a) = v\}|}{|Ex|} . H(\{x \in Ex | value(x,a) = v\}) \right) \quad (2)$$

$$IV(Ex,a) = - \sum_{v \in values(a)} \frac{|\{x \in Ex | value(x,a) = v\}|}{|Ex|} . log_2 \left(\frac{|\{x \in Ex | value(x,a) = v\}|}{|Ex|} \right) \quad (3)$$

$$IGR(Ex,a) = IG/IV \quad (4)$$

In feature selection stage, the proposed algorithm leans on selecting the feature with the top information gain ratio named CBRG (Cumulative Bayesian Ridge regression with Gain-ratio).

3.3. Performance evaluation

The performance of the proposed algorithm is evaluated using RMSE, MAE, R2, and the time of imputation in seconds (t)[3].

- RMSE: Given by Given by the equation below

$$RMSE = \sqrt{\frac{\sum_{l=1}^{n} (y_l - \hat{y}_l)^2}{n}} \quad (5)$$

- MAE: Given by the equation below:

$$MAE = \frac{1}{n} \sum_{l=1}^{n} |y_l - \hat{y}_l| \quad (6)$$

- R2: Given by the equation below:

$$R^2 (y_l - \hat{y}_l) = 1 - \frac{\sum_{l=1}^{n} (y_l - \hat{y}_l)^2}{\sum_{l=1}^{n} (y_l - \overline{y})^2} ; \quad (7)$$

$$\overline{y} = \frac{1}{n} \sum_{l=1}^{n} y_l$$

Where y_l and $\hat{y}_l$ are the real value and predicted value of the lth observation, respectively, and n is the number of samples

4. Experimental outcomes and discussion

4.1. Error analysis:

- For admission data set: In all missingness mechanisms, MAE and RMSE of CBRG are better than all packages.
- For diabetes dataset: In MAR and MCAR, MAE and RMSE of CBRG are better than all packages but worse than LeastSquares and MICE. In MNAR, MAE of CBRG is better than all packages but worse than LeastSquares and MICE and its RMSE is better than all packages but worse than LeastSquares, Stochastic and MICE.
- For profit dataset: in MAR and MCAR, MAE of CBRG is better than all packages. In MNAR, MAE of CBRG is worse than MICE. In MAR and MNAR, RMSE of CBRG is better than all packages. In MCAR, RMSE of CBRG is better than all packages but worse than MICE.
- For wine dataset: in all missingness mechanisms the MAE and RMSE of CBRG are better than all packages but worse than LeastSquares and MICE.
- For California dataset: In MAR, MAE of CBRG is better than all packages but worse than MICE and its RMSE is better than all packages but worse than EMI, Fast KNN, LeastSquares and MICE. In MCAR, MAE of CBRG is better than all packages but worse than Fast KNN, LeastSquares and MICE, and its RMSE is better than all packages but worse than EMI, Fast KNN, LeastSquares and MICE. In MNAR, MAE and RMSE of CBRG are better than all packages but worse than LeastSquares and MICE.
- For diamond dataset: In MAR, MAE and RMSE of CBRG are better than all packages but worse than LeastSquares, Stocahstic and MICE. In MCAR, MAE of CBRG is better than all packages but worse than LeastSquares, Stocahstic and MICE and its RMSE is better than all packages but worse than LeastSquares and MICE. In MNAR, MAE and RMSE of CBRG are better than all packages but worse than LeastSquares, Stocahstic and MICE
- For BNG (10000): In MAR, MAE of CBRG is better than all packages but worse than MICE and its RMSE is better than all packages. In MCAR and MNAR, MAE and RMSE of CBRG are better than all packages but worse than LeastSquares and MICE.
- For BNG (15000): In MAR and MNAR, MAE and RMSE of CBRG are better than all packages but worse than LeastSquares and MICE. In MCAR, MAE of CBRG is better than all packages but worse than LeastSquares and MICE, and its RMSE is better than all packages but worse than LeastSquares, Fast KNN, EMI and MICE
- For BNG (20000): In all Missingness mechanisms, MAE and RMSE of CBRG are better than all packages but worse than LeastSquares and MICE.
- For BNG (50000): In all Missingness mechanisms, MAE and RMSE of CBRG are better than all packages but worse than LeastSquares and MICE.

- For Poker (10000): In MAR and MNAR, MAE and RMSE of CBRG are better than all packages. In MCAR, MAE of CBRG is better than all packages, and its RMSE is better than all packages but worse than Fast KNN and EMI.
- For Poker (15000): In MAR and MNAR, MAE and RMSE of CBRG are better than all packages. In MCAR, MAE of CBRG is better than all packages but worse than LeastSquares and MICE, and its RMSE is better than all packages but worse than LeastSquares, MICE, Fast KNN and EMI.
- For Poker (20000): In MAR, MAE and RMSE of CBRG are better than all packages. In MCAR, MAE of CBRG is better than all packages but worse than LeastSquares and MICE, and its RMSE is better than all packages but worse than LeastSquares, MICE, Fast KNN and EMI. In MNAR, MAE of CBRG is better than all packages but worse than MICE and its RMSE is better than all packages.
- For Poker(50000): In MAR, MAE of CBRG is better than all packages but worse than MICE and its RMSE is better than all packages but worse than MICE, Fast KNN and EMI. In MCAR, MAE of CBRG is better than all packages but worse than LeastSquares and MICE, and its RMSE is better than all packages. In MNAR, MAE and RMSE of CBRG are better than all packages.

4.2. Imputation time analysis:

- For admission dataset: In MCAR and MNAR, Imputation times of CBRG are better than all packages. In MAR, Imputation times for CBRG are better than all packages but worse than Norm, Fast KNN and EMI.
- For diabetes dataset: In all missingness mechanisms, Imputation times of CBRG are better than all packages but worse than Norm.
- For profit dataset: In all missingness mechanisms, Imputation times of CBRG are worse than all packages.
- For wine dataset: In all missingness mechanisms, Imputation times of CBRG are better than MICE and EMI but worse than all other packages.
- For California dataset: In all missingness mechanisms, Imputation times of CBRG are worse than all packages.
- For diamond dataset: In all missingness mechanisms, Imputation times of CBRG are better than MICE and EMI but worse than all other packages.
- For BNG (10000): In all missingness mechanisms, Imputation times of CBRG are better than all packages but worse than Norm.
- For BNG (15000): In all missingness mechanisms, Imputation times of CBRG are better than all packages but worse than Norm.
- For BNG (20000): In all missingness mechanisms, Imputation times of CBRG are better than all packages but worse than Norm.
- For BNG (50000): In all missingness mechanisms, Imputation times of CBRG are better than all packages but worse than Norm.
- For Poker (10000): In all missingness mechanisms, Imputation times of CBRG are better than all packages but worse than Norm.
- For Poker (15000): In all missingness mechanisms, Imputation times of CBRG are better than all packages but worse than Norm.

- For Poker (20000): In all missingness mechanisms, Imputation times of CBRG are better than Fast KNN, EMI and MICE, but worse than Norm, Leastsquares and Stochastic.
- For Poker (50000): In all missingness mechanisms, Imputation times of CBRG are better than all packages but worse than Norm.

4.3. Accuracy analysis:

- For admission dataset: R2 of CBRG is better than all packages but worse than LeastSquares and MICE in all missingness mechanisms.
- For diabetes dataset: R2 of CBRG is better than all packages but worse than LeastSquares and MICE in all missingness mechanisms.
- For Profit dataset: in MAR and MNAR, R2 of CBRG is better than all packages but worse than MICE. In MCAR, R2 of CBRG is better than all packages but worse than LeastSquares and MICE.
- For Wine dataset: R2 of CBRG is better than all packages but worse than LeastSquares and MICE in all missingness mechanisms.
- For California dataset: In MAR and MNAR, R2 of CBRG is better than all packages but worse than LeastSquares, Stochastic and MICE. In MCAR, R2 of CBRG is better than all packages but worse than LeastSquares and MICE.
- For diamond dataset: in MAR and MNAR, R2 of CBRG are better than all packages.in MCAR, R2 of CBRG is better than all packages but worse than LeastSquares, Fast KNN and MICE.
- For BNG (10000) dataset: R2 of CBRG is better than all packages but worse than LeastSquares and MICE in all missingness mechanisms.
- For BNG (15000) dataset: R2 of CBRG is better than all packages but worse than LeastSquares and MICE in all missingness mechanisms.
- For BNG (20000) dataset: R2 of CBRG is better than all packages but worse than LeastSquares and MICE in all missingness mechanisms.
- For BNG (50000) dataset: R2 of CBRG is better than all packages but worse than LeastSquares and MICE in all missingness mechanisms.
- For Poker (10000) dataset: R2 of CBRG is better than all packages in all missingness mechanisms.
- For Poker (15000) dataset: In MAR, R2 of CBRG is better than all packages but worse than MICE. In MCAR, R2 of CBRG is better than all packages but worse than LeastSquares and MICE. In MNAR, R2 of CBRG is better than all packages.
- For Poker (20000) dataset: In MAR and MNAR, R2 of CBRG is better than all packages. In MCAR, R2 of CBRG is better than all packages but worse than LeastSquares and MICE.
- For Poker (50000) dataset: In MAR, R2 of CBRG is better than all packages but worse than MICE. In MCAR, R2 of CBRG is better than all packages but worse than LeastSquares and MICE. In MNAR, R2 of CBRG is better than all packages.

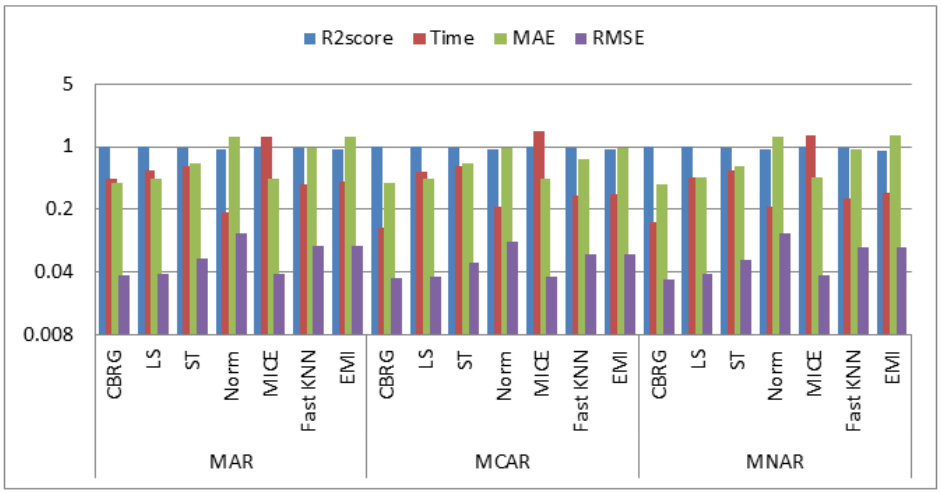

Fig. 3. Comparison between the proposed method, leastSquares, Stochastic, Norm, MICE, Fast KNN and Expectation-Maximization Imputation (graduate admission dataset)

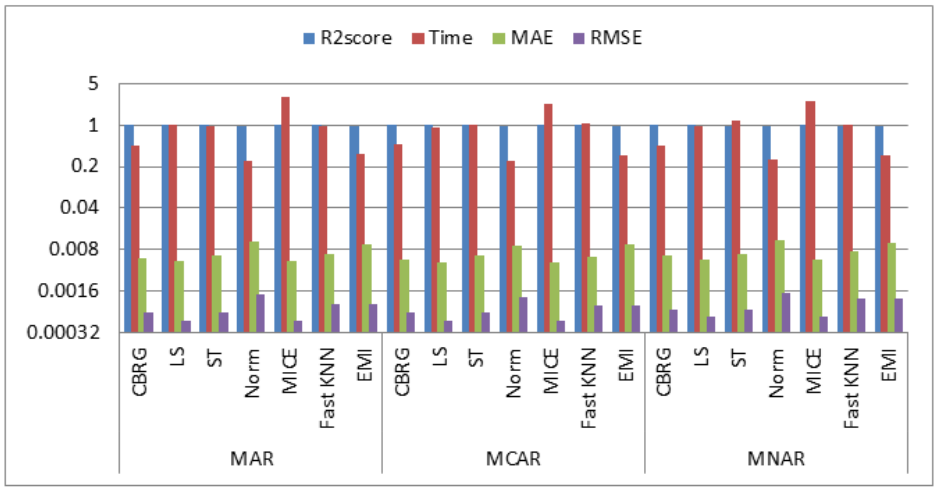

Fig. 4. Comparison between the proposed method, leastSquares, Stochastic, Norm, MICE, Fast KNN and Expectation-Maximization Imputation (diabetes dataset)

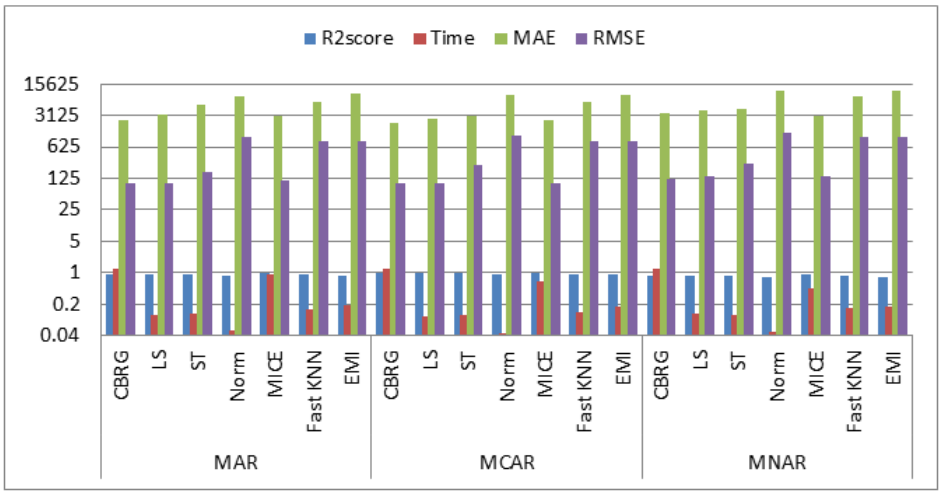

Fig. 5. Comparison between the proposed method, leastSquares, Stochastic, Norm, MICE, Fast KNN and Expectation-Maximization Imputation (Profit dataset)

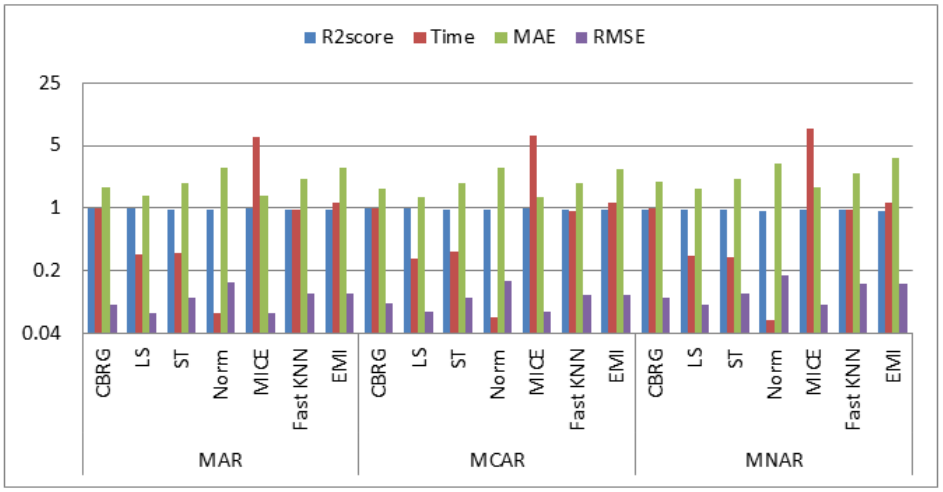

Fig. 6. Comparison between the proposed method, leastSquares, Stochastic, Norm, MICE, Fast KNN and Expectation-Maximization Imputation (wine dataset)

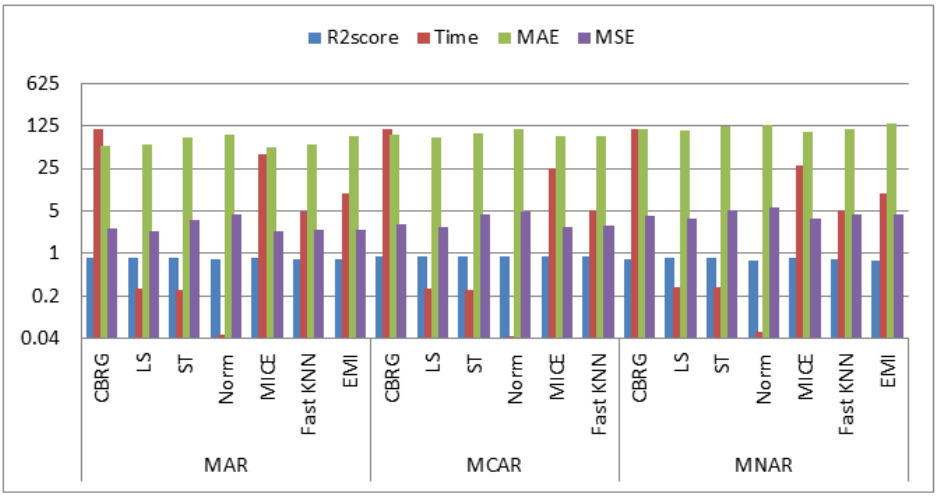

Fig. 7. Comparison between the proposed method, leastSquares, Stochastic, Norm, MICE, Fast KNN and Expectation-Maximization Imputation (California dataset)

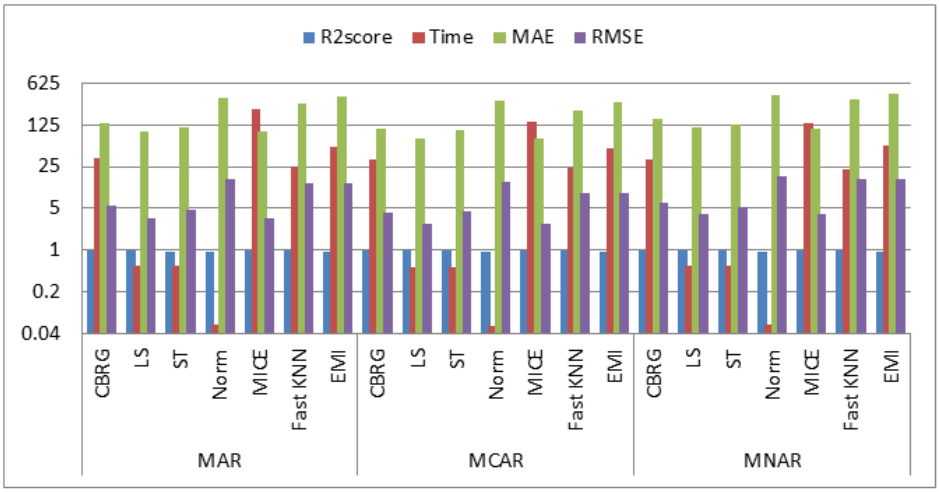

Fig. 8. Comparison between the proposed method, leastSquares, Stochastic, Norm, MICE, Fast KNN and Expectation-Maximization Imputation (diamond dataset)

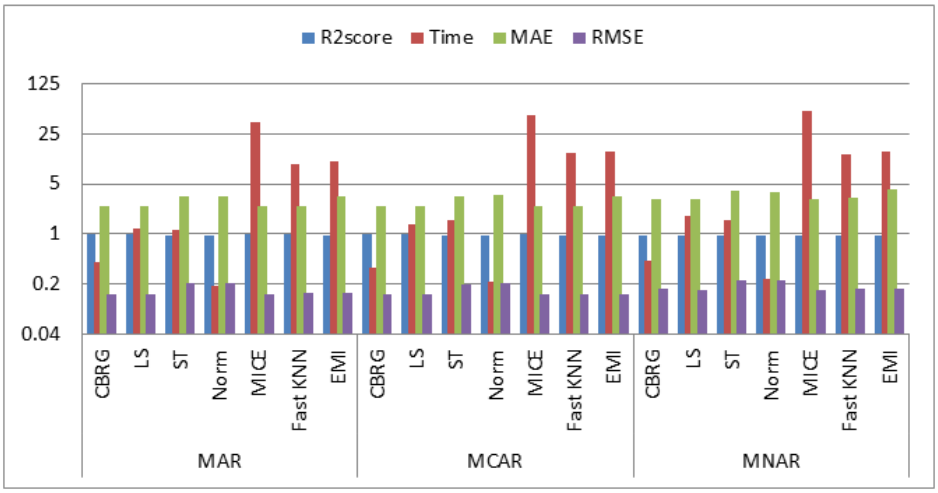

Fig. 9. Comparison between the proposed method, leastSquares, Stochastic, Norm, MICE, Fast KNN and Expectation-Maximization Imputation (BNG (10000) dataset)

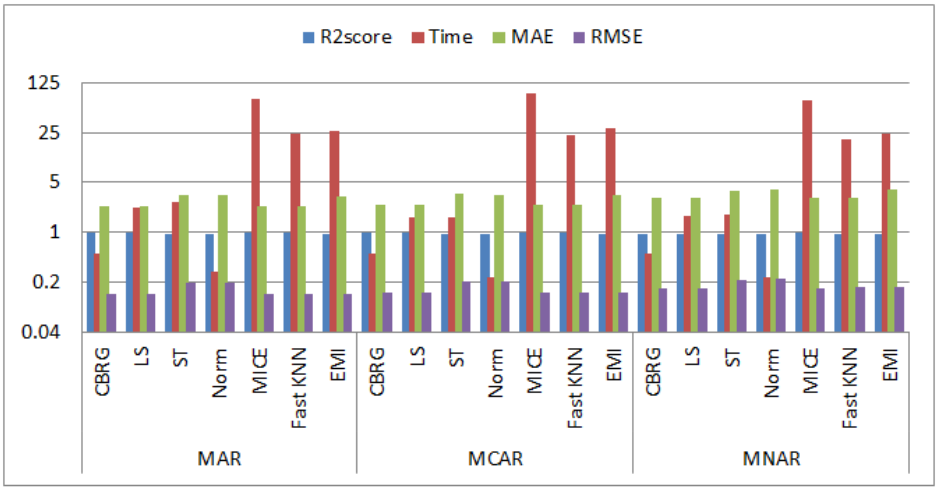

Fig. 10. Comparison between the proposed method, leastSquares, Stochastic, Norm, MICE, Fast KNN and Expectation-Maximization Imputation (BNG (15000) dataset)

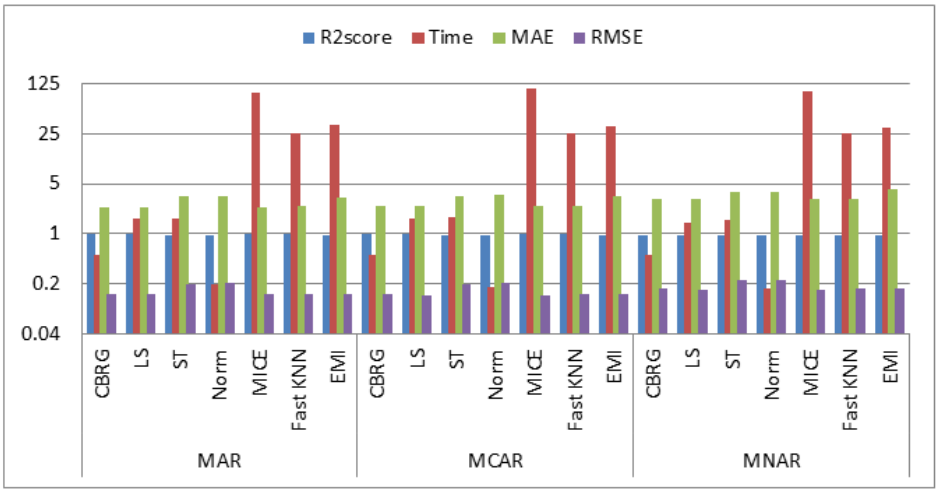

Fig.11. Comparison between the proposed method, leastSquares, Stochastic, Norm, MICE, Fast KNN and Expectation-Maximization Imputation (BNG (20000) dataset)

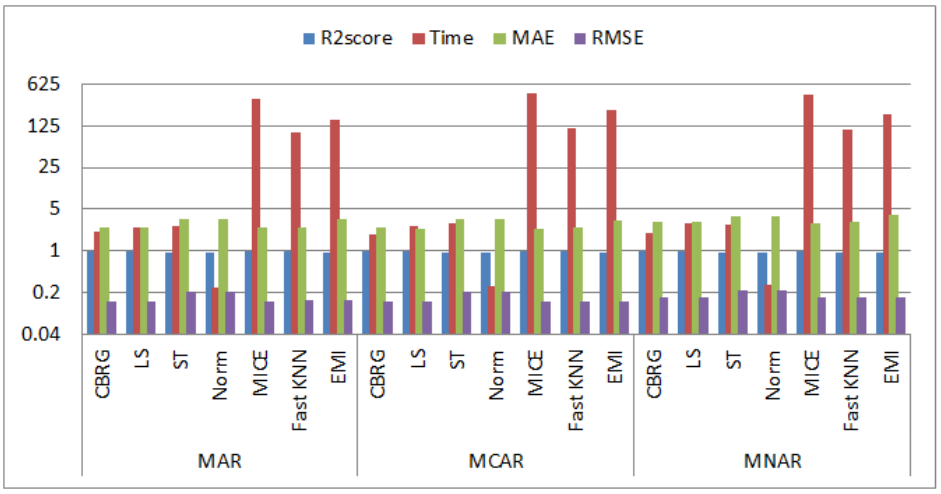

Fig. 12. Comparison between the proposed method, leastSquares, Stochastic, Norm, MICE, Fast KNN and Expectation-Maximization Imputation (BNG (50000) dataset)

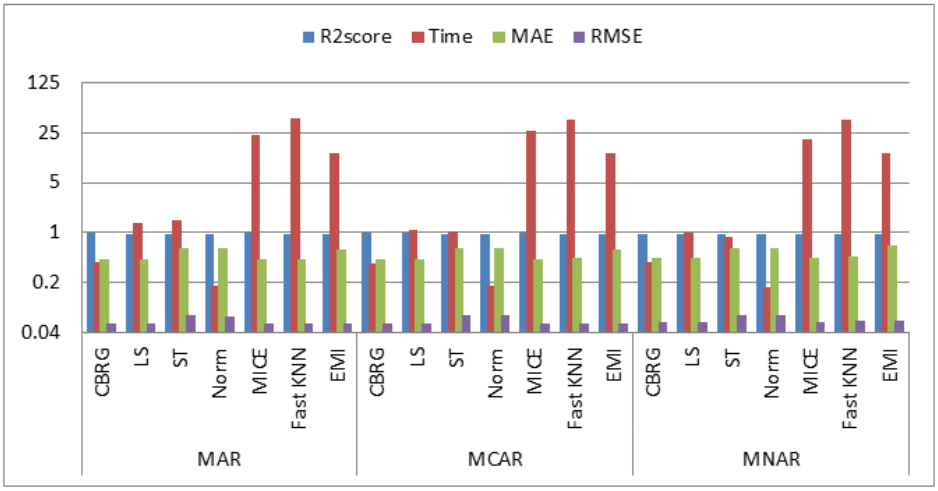

Fig. 13. Comparison between the proposed method, leastSquares, Stochastic, Norm, MICE, Fast KNN and Expectation-Maximization Imputation (Poker (10000) dataset)

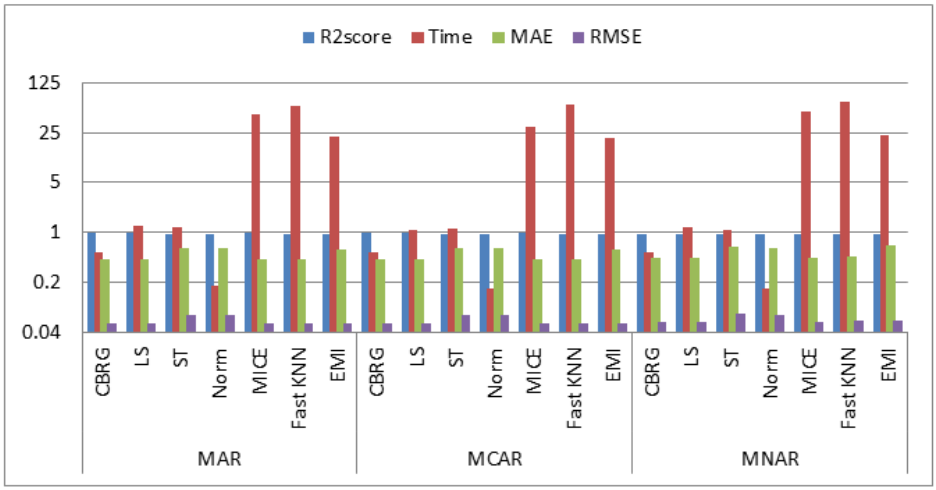

Fig. 14. Comparison between the proposed method, leastSquares, Stochastic, Norm, MICE, Fast KNN and Expectation-Maximization Imputation (Poker (15000) dataset)

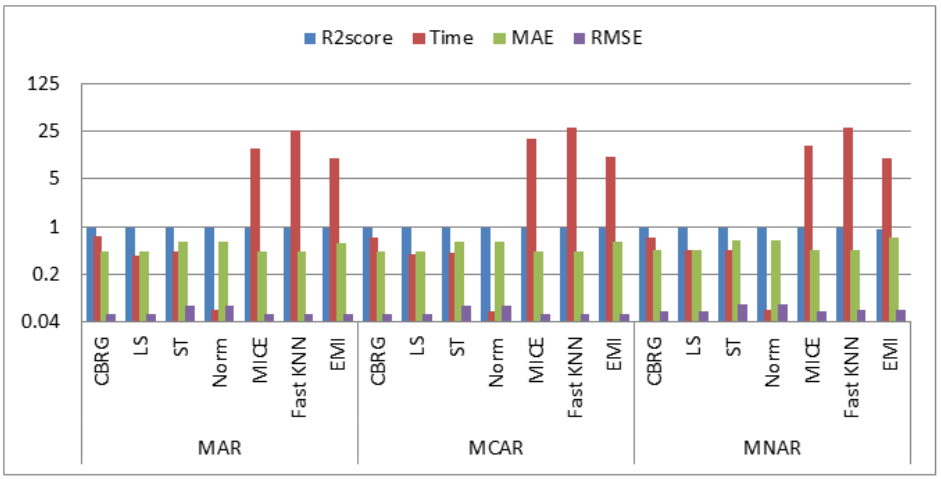

Fig. 15. Comparison between the proposed method, leastSquares, Stochastic, Norm, MICE, Fast KNN and Expectation-Maximization Imputation (Poker (20000) dataset)

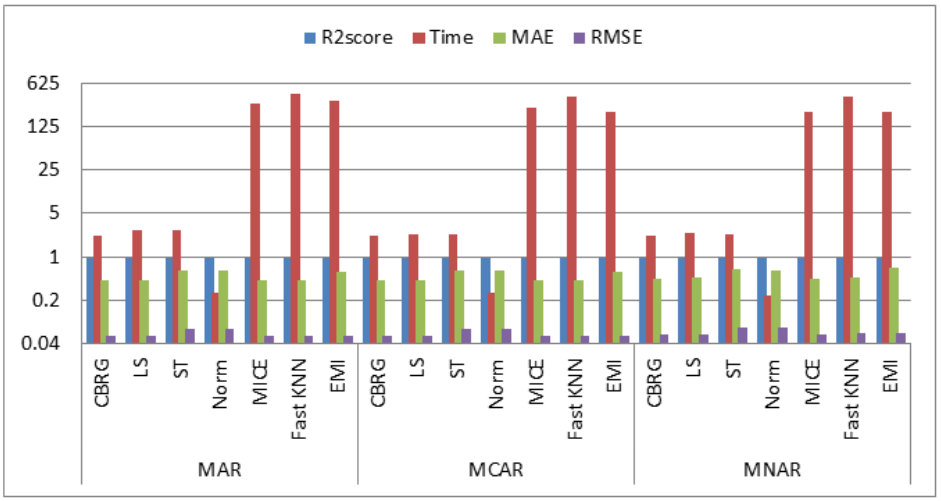

Fig. 16. Comparison between the proposed method, leastSquares, Stochastic, Norm, MICE, Fast KNN and Expectation-Maximization Imputation (Poker (50000) dataset)

5. Conclusion, findings, and future work

It is conclusive essential to handle missing data, as it happens in nearly all real-world research. Handling incomplete instances is very significant for observational analyses with several predictors.

In this paper, we have for a short time studied a set of principled approaches that are used to handle missing data reviewed their implementation on different datasets and different proportions of missing values generated from the three missingness mechanisms and proposing a new algorithm Cumulative Bayesian Ridge Regression works in a cumulative order to impute all missing values in all features one after one.

The candidate feature to be imputed is selected based on the information gain ratio. The proposed al algorithm gives a good performance against the stated approaches even when its accuracy is sometimes worse than some packages but not too much it's very close to them and some times better than all of them. The proposed algorithm shows an acceptable running time and considered a fast method but not the fastest as calculating the gain ratio is computationally expensive

In future research, the proposed imputation algorithm will be studied in additional datasets; additional units of standard error (e.g. T-value and P-value) will be taken into mined when picking the candidate feature. The best important future trend is to take the help of algorithms that cope with optimization problems with mixed features such as GSA-GA algorithm[45].

References

[1] T. D. Pigott, "A review of methods for missing data," *Int. J. Phytoremediation*, vol. 21, no. 1, pp. 353–383, 2001, doi: 10.1076/edre.7.4.353.8937.

[2] C. M. Salgado, C. Azevedo, H. Proença, and S. M. Vieira, "Missing data," *Secondary Analysis of Electronic Health Records*. pp. 143–162, 2016, doi: 10.1007/978-3-319-43742-2_13.

[3] S. M. Mostafa, "Imputing missing values using cumulative linear regression," *CAAI Trans. Intell. Technol.*, vol. 4, no. 3, pp. 182–200, 2019, doi: 10.1049/trit.2019.0032.

[4] M. L. Yadav and B. Roychoudhury, "Handling missing values: A study of popular imputation packages in R," *Knowledge-Based Syst.*, vol. 160, pp. 104–118, 2018, doi: https://doi.org/10.1016/j.knosys.2018.06.012.

[5] M. B. Albayati and A. M. Altamimi, "An empirical study for detecting fake facebook profiles using supervised mining techniques," *Inform.*, vol. 43, no. 1, pp. 77–86, 2019, doi: 10.31449/inf.v43i1.2319.

[6] A. Gelman, J. Hill, A. Gelman, and J. Hill, "Missing-data imputation," *Data Anal. Using Regres. Multilevel/Hierarchical Model.*, pp. 529–544, 2010, doi: 10.1017/cbo9780511790942.031.

[7] K. L. Sainani, "Dealing With Missing Data," *PM R*, vol. 7, no. 9, pp. 990–994, 2015, doi: 10.1016/j.pmrj.2015.07.011.

[8] J. L. Schafer, "Multiple imputation: a primer," *Stat. Methods Med. Res.*, vol. 8, no. 1, pp. 3–15, 1999, doi: 10.1177/096228029900800102.

[9] W. C. Lin and C. F. Tsai, "Missing value imputation: a review and analysis of the literature (2006–2017)," *Artif. Intell. Rev.*, vol. 53, no. 2, pp. 1487–1509, 2020, doi: 10.1007/s10462-019-09709-4.

[10] P. Croiseau, E. Génin, and H. J. Cordell, "Dealing with missing data in family-based association studies: A multiple imputation approach," *Hum. Hered.*, vol. 63, no. 3–4, pp. 229–238, 2007, doi: 10.1159/000100481.

[11] P. Royston, "Multiple imputation of missing values," *Stata J.*, vol. 4, no. 3, pp. 227–241, 2004, [Online]. Available: https://econpapers.repec.org/RePEc:tsj:stataj:v:4:y:2004:i:3:p:227-241.

[12] G. Batista and M. C. Monard, "A study of k-nearest neighbour as an imputation method," *Front. Artif. Intell. Appl.*, vol. 87, pp. 251–260, 2002.

[13] D. Bertsimas, C. Pawlowski, and Y. D. Zhuo, "From predictive methods to missing data imputation: An optimization approach," *J. Mach. Learn. Res.*, vol. 18, pp. 1–39, 2018.

[14] Y. Li and L. E. Parker, "Nearest neighbor imputation using spatial–temporal correlations in wireless sensor networks," *Inf. Fusion*, vol. 15, pp. 64–79, Jan. 2014, doi: 10.1016/J.INFFUS.2012.08.007.

[15] L. Beretta and A. Santaniello, "Nearest neighbor imputation algorithms: A critical evaluation," *BMC Med. Inform. Decis. Mak.*, vol. 16, no. Suppl 3, 2016, doi: 10.1186/s12911-016-0318-z.

[16] P. D. Allison, *Handling missing data by maximum likelihood*. 2012.

[17] A. Farhangfar, L. A. Kurgan, and W. Pedrycz, "A Novel Framework for Imputation of Missing Values in Databases," *IEEE Trans. Syst. Man, Cybern. - Part A Syst. Humans*, vol. 37, no. 5, pp. 692–709, 2007, doi: 10.1109/TSMCA.2007.902631.

[18] R. Razavi-Far, B. Cheng, M. Saif, and M. Ahmadi, "Similarity-learning information-fusion schemes for missing data imputation," *Knowledge-Based Syst.*, vol. 187, p. 104805, 2020, doi: 10.1016/j.knosys.2019.06.013.

[19] M. G. Rahman and M. Z. Islam, "Missing value imputation using a fuzzy clustering-based EM

approach," *Knowl. Inf. Syst.*, vol. 46, no. 2, pp. 389–422, 2016, doi: 10.1007/s10115-015-0822-y.
[20] C. B. Do and S. Batzoglou, "What is the expectation maximization algorithm?," *Nat. Biotechnol.*, vol. 26, no. 8, pp. 897–899, 2008, doi: 10.1038/nbt1406.
[21] Z. Ma and G. Chen, "Bayesian methods for dealing with missing data problems," *J. Korean Stat. Soc.*, vol. 47, no. 3, pp. 297–313, 2018, doi: 10.1016/j.jkss.2018.03.002.
[22] H. Jiang, "Defect features recognition in 3D Industrial CT Images," *Inform.*, vol. 42, no. 3, pp. 477–482, 2018, doi: 10.31449/inf.v42i3.2454.
[23] F. Sciences, "Working With Missing Values," *J. Marriage Fam.*, vol. 67, no. November, pp. 1012–1028, 2005.
[24] N. J. Horton and S. R. Lipsitz, "Multiple imputation in practice : Comparison of software packages for regress ...," *Sci. York*, no. April 2013, pp. 37–41, 2001.
[25] M. Fichman and J. N. Cummings, "Multiple Imputation for Missing Data: Making the Most of What You Know," *Organ. Res. Methods*, vol. 6, no. 3, pp. 282–308, 2003, doi: 10.1177/1094428103255532.
[26] J. W. Graham, "Missing Data Analysis: Making It Work in the Real World," *Annu. Rev. Psychol.*, vol. 60, no. 1, pp. 549–576, 2009, doi: 10.1146/annurev.psych.58.110405.085530.
[27] Q. Chen, M. C. Paik, M. Kim, and C. Wang, "Using link-preserving imputation for logistic partially linear models with missing covariates," *Comput. Stat. Data Anal.*, vol. 101, pp. 174–185, Sep. 2016, doi: 10.1016/J.CSDA.2016.03.004.
[28] K. J. Lee and J. B. Carlin, "Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation," *Am. J. Epidemiol.*, vol. 171, no. 5, pp. 624–632, 2010, doi: 10.1093/aje/kwp425.
[29] Z. Sahri, R. Yusof, and J. Watada, "FINNIM: Iterative imputation of missing values in dissolved gas analysis dataset," *IEEE Trans. Ind. Informatics*, vol. 10, no. 4, pp. 2093–2102, 2014, doi: 10.1109/TII.2014.2350837.
[30] L. F. Burgette and J. P. Reiter, "Multiple imputation for missing data via sequential regression trees," *Am. J. Epidemiol.*, vol. 172, no. 9, pp. 1070–1076, 2010, doi: 10.1093/aje/kwq260.
[31] N. Schenker and J. M. G. Taylor, "Partially parametric techniques for multiple imputation," *Comput. Stat. Data Anal.*, vol. 22, no. 4, pp. 425–446, Aug. 1996, doi: 10.1016/0167-9473(95)00057-7.
[32] I. E. W. Rachmawan and A. R. Barakbah, "Optimization of missing value imputation using Reinforcement Programming," *Proc. - 2015 Int. Electron. Symp. Emerg. Technol. Electron. Information, IES 2015*, pp. 128–133, 2016, doi: 10.1109/ELECSYM.2015.7380828.
[33] S. M. Mostafa, "Missing Data Imputation by the Aid of Features Similarities," *Int. J. Big Data Manag.*, vol. 1, no. 1, pp. 81–103, 2020, doi: 10.1504/ijbdm.2019.10025856.
[34] I. J. and R. T. (2004) Bradley Efron, Trevor Hastie, "Diabetes Data." https://www4.stat.ncsu.edu/~boos/var.select/diabetes.html (accessed Jun. 01, 2019).
[35] M. S. Acharya, "Graduate admissions-1-6-2019." https://www.kaggle.com/mohansacharya/graduate-admissions (accessed Jun. 01, 2019).
[36] B. Stephen, "profit estimation of companies." https://github.com/boosuro/profit_estimation_of_companies (accessed Aug. 08, 2019).
[37] P. Kartik, "Red & White wine Dataset." https://www.kaggle.com/numberswithkartik/red-white-wine-dataset (accessed Feb. 11, 2019).
[38] N. Cam, "California Housing Prices." https://www.kaggle.com/camnugent/california-housing-prices (accessed Jul. 06, 2019).
[39] S. Magrawal, "Diamonds." https://www.kaggle.com/shivam2503/diamonds (accessed Aug. 30, 2019).
[40] F. O. Robert Cattral, "Poker Hand Dataset." https://archive.ics.uci.edu/ml/datasets/Poker+Hand (accessed Nov. 24, 2019).
[41] J. V. Geoffrey Holmes,Bernhard Pfahringer,Jan van Rijn, "BNG_heart_statlog." https://www.openml.org/d/267 (accessed Sep. 11, 2019).
[42] S. ASonawale and R. Ade, "Dimensionality Reduction: An Effective Technique for Feature Selection," *Int. J. Comput. Appl.*, vol. 117, no. 3, pp. 18–23, 2015, doi: 10.5120/20535-2893.
[43] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM Comput. Surv.*, vol. 50, no. 6, 2017, doi: 10.1145/3136625.
[44] R. L. De Mántaras, "A Distance-Based Attribute Selection Measure for Decision Tree Induction," *Mach. Learn.*, vol. 6, no. 1, pp. 81–92, 1991, doi: 10.1023/A:1022694001379.
[45] H. Garg, "A hybrid GSA-GA algorithm for constrained optimization problems," *Inf. Sci. (Ny).*, vol. 478, pp. 499–523, Apr. 2019, doi: 10.1016/J.INS.2018.11.041.