International Journal of Decision Support System Technology
Volume 14 • Issue 1

Missing Data Imputation:
A Survey

Bhagyashri Abhay Kelkar, KIT's College of Engineering, Kolhapur, India*
https://orcid.org/0000-0001-5847-5795

ABSTRACT

Many real-world datasets may contain missing values for various reasons. These incomplete datasets can pose severe issues to the underlying machine learning algorithms and decision support systems. It may result in high computational cost, skewed output, and invalid deductions. Various solutions exist to mitigate this issue; the most popular strategy is to estimate the missing values by applying inferential techniques such as linear regression, decision trees, or Bayesian inference. In this paper, the missing data problem is discussed in detail with a comprehensive review of the approaches to tackle it. The paper concludes with a discussion on the effectiveness of three imputation methods, namely imputation based on multiple linear regression (MLR), predictive mean matching (PMM), and classification and regression tree (CART), in the context of subspace clustering. The experimental results obtained on real benchmark datasets and high-dimensional synthetic datasets highlight that MLR-based imputation method is more efficient on high-dimensional incomplete datasets.

KEYWORDS

CLUSLINK, High-Dimensional Data, Missing Data, Multiple Imputation, Subspace Clustering

INTRODUCTION

Data storage technology has witnessed evolution from the era of storing 100 bytes on punched cards to the latest nanotechnology based atomic data storage. Storing large volumes of data has become cheaper due to innovations in storage hardware and architectures. Today, most of the automated processes try to record many variables pertaining to an entity. This trend has resulted into increased data sizes in terms of attributes or variables describing the entity. The attributes are also called as dimensions and the objects correspond to vertices in multi-dimensional space described by the attributes. Information hidden inside these valuable data resources have become an essential knowledge-base to aid the decision making process in different sectors such as business development, banking, healthcare, finance and e-commerce. Comprehending such huge data sources is now beyond human capability and requires use of powerful and automated data analysis and mining tools.

Although, increased data size is good for getting deeper insights into "what the data says", at the same time it also increases the chances of degradation in data quality. Even though any data analysis process is mainly driven by factors such as selection of features, sampling methods and algorithms;

DOI: 10.4018/IJDSST.292446 *Corresponding Author

the quality of input data highly affects the fruitfulness of the investigation task. For getting good quality results, it is desirable to fine-tune the input data on which the system will be built and the key dependency lies mostly in efficiently handling incompleteness of the input. Especially in quantitative research, the presence of missing data has become a rule than an exception; and hence analysis of such datasets is a common challenge faced by most of the data scientists (Little & Rubin, 2019). Missing data is defined as values in the data that are not available and if they are observed, would be meaningful. Such values are often stored as blanks, None, Null or NaN. Many real world datasets may have missing values, due to various reasons; like intentional avoidance to feed the values, not knowing the values during the data entry stage, irrelevance of some of the fields for the entity under consideration or a random noise. When less than 1% of the data is missing, it is generally regarded as trivial; 1-5% missing data is manageable; however, the existence of 5-15% of missing data necessitates sophisticated techniques to handle it. If the missing data proportion is still more, it may badly impact any interpretations produced by the analysis process (Acuña & Rodriguez, 2004).

Missing values may pose other severe problems such as increase in computational cost, skewed output, longer preprocessing times and frustration to the researchers. Incomplete records may affect predictions, descriptions and inferences produced by the analysis software. Data analysis process becomes complicated due to the bias resulting from differences between missing and complete data. Especially in case of sampling, if most of the samples contain missing values, the outcome of the analysis algorithm will be misleading. In case of Clinical Research, missing data can have side effects of overestimation or under-estimation of treatment (Kim et al., 2019). Generally, many statistical procedures skip these incomplete records. However, in such cases, very less data remains available to perform further analysis. Non-linear machine learning algorithms may fail completely in the absence of adequate data samples.

LITERATURE SURVEY

The missing data is an ever-present challenge faced by machine learning researchers while working on real-world datasets. Many such examples can be found; the UCI Machine Learning Repository hosts many datasets with missing values (Dua & Karra Taniskidou, 2017). Honeywell, (a well-known company that manufactures and services complex equipments) despite imposing regulatory conditions for data collection, had an industrial database which contained around 50% missing data (Lakshminarayan et al., 1999). The problem is more prominent in medical datasets related to patients' health records, and in most of the cases the data is collected in an unorganized manner resulting into considerable information loss (Cios & William Moore, 2002). Almost every entry in these databases can have important values missing. In the case of wireless sensor networks, due to sensor failures or power outage, incomplete data is unavoidable (Gruenwald et al., 2010).

Rubin (Rubin, 1976) was pioneer in proposing a framework for illustrating the processes that generate missing data and identification of the relationships between missing and observed values. Most of the later research in this area is influenced by his work. Recent commercial software products support missing data imputation. Multiple imputation algorithms are implemented in well known data science and statistics software like Stata, SPSS, and SAS. The fancyimpute library in Python implements various algorithms for missing data imputation. Other software packages that provide multiple imputation are - IVEware, S-Plus, ICE, Amelia II, and SOLAS. The package called XMISS in the LogXact provides machine learning based methods for MAR data values. R packages - miP and VIM support visualization of imputed data and facilitate comparison of observed and imputed data distributions graphically. The Amelia package also has a feature to diagnose "overimputation" by generating cross-validation plots. Such an analysis can be very helpful for knowing the distribution of the newly filled values, and whether the imputation process has produced sensible output. This process can be aided with the help of domain expertise.

A genetic algorithm based approach is suggested by Galán et al. for imputation of missing data (Ordóñez Galán et al., 2017). Wang and Chaib-draa apply Bayesian framework using Gaussian Process Regression for imputation of surface temperature analysis data (Wang & Chaib-draa, 2017). (Morvan et al., 2020) have shown that, in presence of missing values, a machine learning algorithm may not be linear. To address this issue, they propose a novel architecture called NeuMiss Networks which apply multiplication by the missingness indicator. A recent approach for handling missing values is suggested by Smieja et al. by using Gaussian mixture model (GMM) in connection with a discriminative neural network (Smieja et al., 2019). Yi et al. addressed the problem of sparsity variations (the output of the model varies with level of sparsity in the data) by applying sparsity normalization (Yi et al., 2020). In a paper by Khan et al., the authors mention that, the medical data classification models are greatly affected by presence of missing values and the associated prediction risk is also high. To mitigate this issue, they propose a hybrid approach by combining single imputation and multiple imputation methods (S. I. Khan & Hoque, 2020). Bai et al. discuss the challenges posed by incomplete medical datasets. They suggest filling the missing cells by categorical values to produce better results (Bai et al., 2015). Mayer et al. discuss imputation methods in the context of supervised learning and found that mean imputation shows consistent performance (Mayer et al., 2019). (Gómez-Carracedo et al., 2014) used multiple imputation to fill missing values in the air quality dataset. They observed that the imputed values are more dispersed compared to single imputation, specifically when the attribute being imputed is poorly correlated to other variables.

Various remarkable reviews are found in literature on the topic of missing data imputation. In the handbook "Handling Missing Attribute Values"; Grzymala-Busse et al. categorize missing data imputation methods as sequential and parallel methods (Grzymala-Busse & Grzymala-Busse, n.d.). They elaborate some of the parallel imputation techniques such as attribute value pairing, lower and upper approximation and rule induction. In another review, the authors give a detailed study of the imputation methods and categorize them as local, global, knowledge-assisted and hybrid approaches (Armina et al., 2017).

HANDLING MISSING DATA IN DECISION SUPPORT SYSTEMS

The evolution of the decision support system (DSS) can be traced back to 1940s, even before the first computing machine was made functional. A DSS aids the decision making process in complex environments such as business risk assessment, sales projection and optimization, medical diagnosis and agricultural product management etc. Thus the main intention of DSS is to solve semi-structured or unstructured decision problems which are difficult to specify in advance and are dynamic in nature. Current pandemic situation has underlined the need of DSS in every business that needs quick decisions. Following paragraphs discuss about few recent works which integrate missing data imputation into the DSS workflow.

Now-a-days, there is a growing tendency to keep patients' health related documents in digital format. Referred as EHR for Electronic Health Records, it is a collection of large volumes of patients' data comprising medical history, doctor's diagnosis, laboratory reports etc. recorded over a large period of time. These records are an important source of information for research in the field of personalized medicine and clinical decision support systems (CDSS) to produce patient-centered outcomes. The reliability of a CDSS is mainly dependent on the availability of complete patient database. On the other hand, EHR is mainly collected for billing purposes rather than being used for analytics purposes and it may have lots of missing clinical measurements. If the issue is left unaddressed, it can reduce the legitimacy of the conclusions drawn. Hence the development of a CDSS is mainly preceded by missing data imputation step. Research carried out in recent years underlines the need for effective imputation methods in this field.

In (Piri, 2020), the authors propose a framework called 'Missing Care' to tackle the issue of existence of incomplete records in Parkinson's disease data. It employs ensemble and imbalanced

data learning methods to select the most important attributes for developing predictive models under missing data. A hybrid DSS for early diagnosis of heart disease is presented by (Rani et al., 2021). The authors use recursive feature elimination and Genetic Algorithms (GA) for feature selection step and MICE algorithm for the imputation step. In another work on CDSS, the authors propose left-center-right method to fill unobserved biomarkers by using existing biomarker measurements from individual patient's visit records (Gupta et al., 2020). In (McCombe et al., 2021), the authors discuss CDSS for diagnosis of Dementia on extremely missing data (i.e. more than 50% of values are missing in over half of the attributes) in both training and test datasets. They state that, the imputation methods which require high computational power do not necessarily show accurate results; and combination of iterative imputation and reduced-feature classification produce the best results. The authors in (Nijman et al., 2021) use Multiple Imputation methods to fill unobserved predictor variables in cardiovascular dataset for real-time risk prediction. They underline that Multiple Imputations combined with joint modeling imputation (JMI) or conditional modeling imputation (CMI) are useful in real-time environments. A comparison of eight missing data imputation techniques in the context of CDSS is presented in the article by (Altukhova, 2020). They evaluated Mean, Mode, Median, EM (Expectation Maximization), Iterative Imputer, MICE, Fast-KNN and Random imputation methods. The empirical results highlight that, fast-KNN algorithm and Iterative Imputer shows best results in the group. In the paper by (Löw et al., 2019), the authors present a "Multiple Retrieval Case-Based Reasoning (MRCBR)" for datasets with missing numerical and categorical attributes. Case based reasoning (CBR) is used as a tool for artificial intelligence (AI) systems for imitating the decision making process in humans and it is currently getting attention in medical field. The authors also highlight that removal of missing data is a loss to the CBR systems, whereas the imputation step really works well to produce a trustworthy CBR output.

(Ma et al., 2020) introduce a new framework for predicting hypoglycemia risk in type 2 diabetes. They propose "Multiple models for Missing values at Time Of Prediction" (MMTOP) algorithm, that eliminates the need of measuring the missing data elements by constructing several risk models that are predictively equivalent. These models are preserved for future use and one of the models which is most suitable for the available measurements is referred during prediction. For clinical big data, (Dong et al., 2021) propose a machine learning based data imputation method called "Generative Adversarial Imputation Nets" (GAIN). The authors compared GAIN with missForest and MICE and found that GAIN was more effective on highly missing, skewed continuous variables and imbalanced categorical variables. In a research work during the coronavirus pandemic, the authors (Cro et al., 2020) propose a four-step algorithm for highly incomplete data with non-standard causes for missingness. They use Controlled Multiple Imputation (MI) followed by Sensitivity analysis under missing-not-at-random assumptions.

Missing data problem is ubiquitous. Hence the research in this field has got wider applicability in other industrial and scientific fields too. In a recent work on Space-Weather applications, the authors (H. et al., 2021) perform imputation of the missing values in ground electromagnetism datasets. The data can be missing due to sensor failure and non-responsiveness in data transmission. They suggest imputation based on Support Vector Regression (SVR). Imputation of worldwide patent statistical dataset is presented in (de Rassenfosse & Seliger, 2021). A method of imputation based on linear as well as non-linear machine learning models for gas permeability data related to polymer membranes is suggested in (Yuan et al., 2021). In (Guastella et al., 2021), the authors present a AI-based DSS for smart cities to facilitate good quality services to its citizens. The authors suggest Multi-Agent Systems (MAS) approach to impute the missing sensor data during data acquisition step. They apply the concept of Voronoi tessellation in which the computation is distributed among mobile and fixed devices operating in the close vicinity. In another paper on the same topic, the authors propose an ensemble of two GRNNs to predict missing values in a smart city environment and apply extended-input SGTM neural-like structure. A liners regression based strategy for handling missing values in analysis of annual stream temperature is outlined in (Johnson et al., 2021).

APPROACHES TO HANDLE MISSING VALUES

Missing values are typically recognized into three types - MCAR, MAR, and NMAR (Little & Rubin, 2019). When the probability of presence of a missing value in a variable is completely independent of the values that are already known, the missingness pattern is called MCAR. There is no explanation for having a value missing for the given object and given attribute. The missing data can be assumed to be a random subset of the observed data. The hypothesis that the data being MCAR can be tested with Little's test (Jakobsen et al., 2017), where a non-significant p-value shows that the data is probably MCAR. If the data is of MAR type, certain facts about the data origin can be known in advance. For example, males avoid filling the information related to depression compared to females. In such cases, the missing values can be determined from the known instances and can be filled by using prior domain knowledge about the data and by applying statistical analysis methods. An object may contain valid missing values in an attribute depending on values of some other attributes. E.g. spouse name may be missing for an unmarried person. Thus, the value is not missing due to random error. Such values should be handled carefully and should not be replaced by any random value.

Various strategies are suggested in literature for dealing with missing values. The incompleteness in the data can be tackled by ignoring the whole record or attribute, manual entry of the missing values or replacing them by appropriate substitutes, also called as imputation (Han et al., 2012). If more than half of values in a variable/record are missing, it is advisable to remove it from further processing. However, such kind of omission leads to a great information loss and may result in loss of expressiveness of the data. It may also produce biased estimates and wrong investigations of the associations (Rubin, 1987).

To acquire the actual data that is missing is often not possible. However, the only viable solution is to find the most probable substitute using the observed records in the dataset. This process referred to as imputation, also adds to the quality of data and enhances the results which will be otherwise unrealistic in the presence of incomplete records. All The imputation methods produce a complete dataset that can then be further explored using standard software procedures existing for analysis of complex datasets. It is expected that the imputed dataset should aid the analysis process and not worsen the situation. Hence imputation process should be carefully designed by adhering to following aspects:

1. Arbitrarily assigning any value for missing cells even by means of expert judgment does not protect the statistical properties of the data like means, variances, covariances etc. The integrity and statistical transparency of the imputations should be ensured by means of a model-based imputation process.
2. The process of imputation should be stochastic in nature. The uncertainty in imputed values should be reflected by addition of a random error term.
3. The imputation model should be based on all variables that are essential to reflect the associations and correlations in the dataset (multivariate imputation). It should also account for non-linear relationships existing within the attributes. Many recent statistical softwares ensure this by means of a sequence of conditional imputation or by means of iterative Markov chain Monte Carlo (MCMC) methods (Schunk, 2008).
4. The imputation procedure should be executed multiple times independently in order to allow for the estimation of the variance. Ideally the parameter of number of imputations must be set to ∞. However, research shows that 3 to 5 imputations are sufficient to produce satisfactory results (Graham et al., 2007).
5. The Imputation process should be robust against moderate deviations of the data from the underlying assumptions. It should provide satisfactory output even if the missing data pattern is more complex

COMPLETE CASE ANALYSIS

Many data analysis software products default to "complete case analysis" for handling missing values (Henry et al., 2013), in which incomplete observations are removed from further processing and complete cases - i.e. the observations having no missing values are only preserved. The impact of data exclusion is negligible if the proportion of missing data is below 5% or when the missing data is present only in the target variable (Jakobsen et al., 2017). It is also a better choice when multiple imputation may not be fruitful or the uncertainty induced by the multiple imputation may increase the standard error. Complete case analysis is very useful in exploratory studies, for example in the initial phases of drug development. The method also ensures that remaining data is unbiased.

Single Imputation

In reality, any imputation method does not assure to replace the missing cells by exact values. Ignoring this fact, the single imputation techniques use complete cases in the data to find the most likely value. The newly imputed value is then used for further processing as if it is the true value. In mean imputation, the missing values are replaced by the average of the values present in a given attribute and the sample mean remains unchanged. For numeric variables, the median of the attribute values can be used and non-numeric variables can be imputed with mode. Mean imputation is somewhat helpful in univariate analysis but creates problems in multivariate analysis. If the context of the data indicates that the observed values have low variance, then imputation with mean produces satisfactory results and the overall mean is not changed. However, in case of large variance, this kind of replacement may change the context of the data. It may also produce distortion in the distribution and standard deviation of the attribute. Single imputation is an easy approach and involves less computation, if the missing data fraction is less. It fills a given missing cell with a specific alternative. However, for a large proportion of missing data, it tends to produce bias that results in incorrect analysis. After imputation, missing (now imputed) and non-missing values are given equal importance. I.e. imputed cases are treated in the same way as other observed cases and thus the uncertainty of the data is underestimated. After single imputation, the validity of the results is mostly dependent on the assumptions about the underlying data. For example in LOCF, it is assumed that a missing value in a variable is likely to be identical with the last observed value. However, in many cases these assumptions may not be true, hence single imputation methods should be applied with caution (Jørgensen et al., 2014).

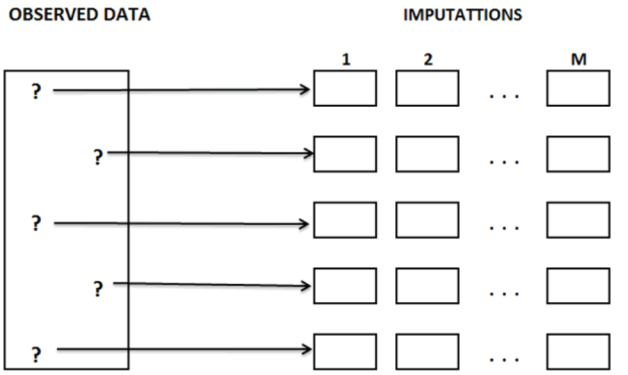
Multiple Imputation

A single pass imputation is unacceptable due to probabilistic characteristics of imputation. Rubin (Rubin, 1987) demonstrated the process of Multiple Imputation (MI) in the context of non-response in censuses and sample surveys. Now, MI has become one of the popular choices for missing data imputation and is available in many recent statistical packages. It uses resampling and Bayesian approach to alleviate the shortcomings of other imputation methods, especially the bias induced by single imputation. While generating the plausible values, it also accounts for uncertainty in the generation of real world data and the associations between input variables. MI is found very effective for imputation of small to medium portions of missingness in the data. Figure 1 outlines the procedure of multiple imputation.

The Multiple Imputation Process

1. In the imputation phase, several copies of the incomplete dataset are created. The number of copies to be created is a user specified parameter denoted by M. Generally five replicas are sufficient for the purpose. In the imputation phase, the missing values in each replica are replaced by applying a separate imputation model with a random variation of the imputation parameters. A careful selection of the imputation model is very essential so that it preserves

Figure 1. Multiple imputation - M specifies the number of imputations



the assumptions about distribution of the data. The imputation model selection is also dependent on the missing data pattern and type of the variables containing missing values. It is desirable to include all important variables while creating the imputation model. Some auxiliary variables are also included wherever required, so that the estimation of the values to be replaced will be the most accurate.

2. In the analysis phase, each dataset created above is analyzed by using complete and standard statistical analysis methods. This phase creates *M* analyses of the imputed datasets.
3. In the final pooling phase, for each missing cell in the original dataset, single point estimation is obtained by combining the parameter estimates (e.g. standard errors and coefficients) by applying Rubin's rule (Rubin, 1987).

Generally 3 to 5 iterations of the imputations produce reliable results; however, modern approaches can go up to 20 to 100 iterations. For moderately missing data, multiple imputation provides a better option by preserving variability in the imputed dataset. The major shortcoming of MI is that it involves complex processing for performing the imputations. For obtaining meaningful results, it is desirable that the user needs to be well aware of the analysis phase and combine (pooling) phase of the multiple imputation. Another disadvantage is that it assumes that the data is missing at random, which may not always be true.

Model Based Imputation

Model-based imputation works by developing a predictive model based on the observed data to find the most accurate estimates for the missing cells. Then the predicted value replaces each corresponding missing value to create a completed dataset. In progressive imputation variant, the values which are imputed in earlier iterations are used to predict remaining missing values. In another variant called imputation with uncertainty, randomness based on other observed values is incorporated in the process. As a rule, the imputation model should be built using a broader and larger set of attributes essential for the analysis. It should not be restricted to the variables having missing values or the variables to be incorporated in later analysis. For example, for an analytic model designed for imputation of diastolic BP, if BMI and age are identified as predictor variables, then for imputation of BMI and diastolic BP, other attributes such as systolic BP, weight, height, race/ethnicity, gender etc. should also be considered. Decision on the variables to be included in the model is generally followed by identification of joint distributional model for the concerned variables. For continuous variables

multivariate normal distribution is preferred and for categorical variables multinomial distribution can be used. For mixed continuous and categorical variables, the location model of distribution can be considered.

Regression Imputation

The most important goal in statistics is to find answer to the question: is the attribute X (or multiple attributes $X_1, ..., X_p$) have any association with another variable Y, and, if the association is present, can the relationship be used to predict Y? This kind of predictive modeling called regression analysis, tries to identify the dependence of a target variable (Y) on independent variable(s) $(X_1, \ldots, X_p)$ also called as predictor(s). The technique is also useful for time series modeling and forecasting. The prediction model tries to fit a line over the given data points such that the variation between the distance of data points from the line or curve is minimized. If several explanatory variables are involved in regression analysis, it is termed as multiple linear regression. The regression based imputation is a promising method if the incomplete dataset contains correlated variables. The regression is performed based on complete case analysis.

Imputation by Multiple Linear Regression (MLR)

The imputation method proposed here is based on multiple linear regression (Gelman & Hill, 2010). It is a two step process described below:

1. In the first step, a correlation matrix is created for the input dataset DS containing attribute set A. The matrix stores correlation coefficients between variables. Two attributes are called highly correlated if the correlation factor is greater than or equal to $\pm 60\%$ (McDowell & Jenkinson, 1996). The correlation matrix is used to identify set X of attributes ($X \subset A$) which are highly correlated to each attribute Y containing missing values. A multiple linear regression model is then developed by using X as predictors and the attribute Y as target variable. Complete cases in the data i.e. the records having no missing values are used as training samples to generate the regression model. For each missing value in the given target attribute Y, the developed model is then used to predict appropriate substitution value.
2. It is quite possible that no other attribute is correlated with the attribute being imputed. In the second step of the imputation process, all such missing cells are filled by values from random selected records.

Multivariate Imputation via Chained Equations (MICE)

Multivariate Imputation via Chained Equations (MICE) uses multiple imputation approach to fill the missing values several times, by applying succession of regression or similar suitable models to create several completed data sets. It operates under the assumption that the missing data are Missing At Random (MAR). The model to be used is specified as one of the parameters for the imputation process. Each data set is then analyzed separately using techniques designed for complete datasets, and then the results are combined in such a way that the variability due to imputation is incorporated. The chained equation imputation process works in following steps:

Step 1: All missing cells in each variable are imputed with the mean value of the corresponding variable. The mean is calculated using available values. Thus the imputed mean is a "place holder" to represent missingness.

Step 2: One attribute that originally had missing values (called "var") is selected for processing. Regression is performed by using remaining variables as predictors and the variable "var" as target. The probability of a value missing in a variable is assumed to be dependent on the observed data. Predictions are performed only for missing cells. Observed values in the variable "var" do

not change. The predicted values are then imputed in corresponding missing cells. The variable "var" is now complete and contains observed and newly imputed values. It can then be used as a predictor for imputation of remaining variables.

Step 3: Iteratively each variable candidate for imputation is processed by applying step 2. When all the variables are processed, one "cycle" is complete.

Step 4: Steps 2 & 3 are executed repeatedly for a number of cycles and the imputations are updated iteratively. The number of cycles to perform is a user specified input and generally it is set to 10 (Azur et al., 2011). At the end of all cycles, the final imputed dataset is retained for further processing.

Step 5: The entire process of imputation is repeated for a user specified parameter M, which represents count of imputations to perform. The default value is set to 5. The process finally converges and the observed and final imputed values form a "complete" data set.

R language MICE package currently supports more than twenty four methods for building the imputation model.

Predictive Mean Matching (PMM)

Predictive Mean Matching (PMM) is preferable when the variables are not normally distributed, especially for missing data in quantitative variables. It avoids induction of bias in the imputed dataset by selecting "real" values sampled from the data. This is accomplished by building a small subset of samples containing the target value matching with the target of the records with missing values. Thus PMM outputs values that are very close to real values present in the attribute. Thus, for skewed variables it produces skewed imputed values and for discrete valued variables, the output will be discrete. If the target variable is bounded by some values, the imputed values will also be bounded by the same bounds. However, PMM based imputation is somewhat expensive as compared to the multiple linear regression based method.

Imputation by Classification and Regression Tree (CART)

The Classification and Regression Tree methodology, also known as the CART was introduced in (Gordon et al., 1984). CART has several characteristics that make it useful for imputation (Burgette & Reiter, 2010). It is flexible to fit nonlinear relations and complex distributions in the data without need for data transformations or parametric assumptions. CART finds optimal partition of the data via recursive binary splits based on predictor variables. Each leaf node has a prediction for the target variable. The method handles the outliers very well. It can deal with skewed distributions and multicollinearity. It is well suited for nonlinear relations. The model fitting can be automated without the need for manual parameter tuning.

Other Imputation Methods

The maximum likelihood method uses observed data to find maximum likelihood estimates of the parameters that best represent the available data making. Hence the method is useful to produce unbiased estimates of the parameters but it is restricted to linear models only. The Expectation Maximization imputation method (Holmes & Rubin, 2002) works in two phases – the Expectation phase and the Maximization phase. If it finds that the imputed value is not the best fit, it tries to re-impute a more suitable value. The iterations proceed until the best fit is found and the output converges. This method tries to preserve correlations between the variables, which are very important for linear regression and factor analysis. The *k*-NN imputation model approximates a missing value by the values in the neighboring objects (Murti et al., 2019). The model is first trained on complete cases and for imputation an actual measured value from the most nearest neighbor (1-NN) is used or it is calculated by averaging *k* measured values from the neighboring records. DataWig is a Deep Neural Network based technique for imputation of data missing in data frames that contain heterogeneous

variables (Biessmann et al., 2018). It automatically identifies all hyper-parameters and thus does not rely on the user's expertise. It is scalable and robust for imputation of non-numerical values such as unstructured text in a variety of languages.

MACHINE LEARNING ON INCOMPLETE DATA

Many recent works discuss the importance of missing data imputation for successful implementation of machine learning algorithms. In general, the data imputation process can be executed as a separate preprocessing step or it can be embedded into the machine learning algorithm itself. (H. Khan et al., 2021) suggest an imputation method based on Fuzzy C-Mean clustering to improve the accuracy of classification.(Shin et al., 2021) recommend a minority oversampling technique based on multiple imputation called as MI-MOTE, to simultaneously overcome the issue of classifying incomplete and imbalanced data. A Conditional Generative Adversarial Networks (CGAN) based data imputation method which utilizes class-specific characteristics in classifying class imbalanced data is proposed by (Awan et al., 2021). (Zhang et al., 2021) outline "Evidence Integration Credal classification Algorithm" (EICA) for multiple classifiers. EICA groups the whole dataset into numerous subsets and then estimates the missing values by analyzing patterns in the subsets. (Faisal & Tutz, 2021) suggest an enhancement to the nearest neighbor imputation technique by utilizing the information of the associations found among the variables.

Clustering is another unsupervised machine learning approach. It aims to identify groups existing in the input data such that members in a group (cluster) are highly similar whereas the clusters themselves are highly dissimilar from each other. In recent years subspace clustering is gaining attention for clustering high-dimensional datasets (the datasets with more than ten dimensions (Han et al., 2012)). Due to the "Curse of Dimensionality", traditional clustering algorithms fail to find meaningful clusters from such high-dimensional datasets. This is the effect of the existence of irrelevant and correlated variables and shortfall of similarity measures such as Euclidean distance in higher dimensions. The solution is to find clusters over subset of attributes and subset of objects, and the corresponding clusters are called subspace clusters. The variables which are part of a subspace cluster signify reasons for grouping the entities together. A single object can be part of multiple subspace clusters, representing multiple characteristics of the same object. Thus subspace clusters represent a useful knowledge-base that cannot be uncovered by traditional clustering algorithms.

Research in subspace clustering has gained momentum in the past two decades due to its widespread applications. (Alghawli, 2022) highlights the use of subspace clustering in anomaly detection. (Kang et al., 2020) propose large-scale "Multi View Subspace Clustering" for big data to reduce the high computing time shown by other algorithms in this category. In (Zhuang et al., 2021), the authors propose S3C2, an efficient framework for sparse subspace clustering and imputation of scRNA-seq dataset. (Niu et al., 2021) overcome the problem of incomplete views by suggesting "One-step multi-view subspace clustering with incomplete views" (OMVSC-IV) technique for computer vision applications. In another similar approach, (Liu et al., 2021) outline a method for jointly exploiting the multi-view information and the cross-view data point relations jointly. (Chen et al., 2018) propose a data representation technique called Low-Rank constrained AutoEncoder (LRAE) for subspace clustering. It takes advantage of capturing global data composition and finds low-rank approximations to promote low rank for underlying neural network. Yao et al. (Yao et al., 2018) highlight that, the noises present in real high-dimensional data have non-Gaussian distribution with complex structures. They modify Expectation Maximization (EM) method to estimate parameter values required by the PMoG-LRR model they propose. A modification of Low-rank Representation-based (LRR) subspace clustering and Sparse Subspace Clustering (SSC) for multimodal data is presented in (Abavisani & Patel, 2018). Struski et al. propose SuMC (Subspace Memory Clustering) which is based on information theory, Minimal Description Length Principle and lossy compression (Struski

et al., 2018). SuMC can estimate optimal dimensionality and count of clusters and the best possible compression ratio by using Bayesian Information Criterion (BIC).

The following sections are dedicated to empirical evaluation of three missing data imputation techniques i) imputation method based on Multiple Linear Regression (MLR) ii) MICE using Predictive Mean Matching (MICE-PMM) and iii) MICE using Classification And Regression Tree (MICE-CART). Although the evaluation is restricted to the subspace clustering context, the findings in general can be applied to any of the supervised and unsupervised machine learning algorithms.

EXPERIMENTAL SETUP

One recent subspace clustering algorithm data called CLUSLINK presented in (Kelkar et al., 2019) is used for comparison of output produced by various imputation methods. CLUSLINK is a parameter light subspace clustering algorithm for numerical and structured data. The algorithm requires only one input parameter that specifies the desired granularity of the resulting clusters, however if not specified, the default values set in the algorithm are used. It outputs a set of identified subspace clusters, which then can be compared with already known distribution of the subspace clusters to evaluate the efficiency of the imputation process.

For comparison of the imputation methods mentioned above, synthetic datasets and real datasets namely - Iris, Ecoli, Glass, Liver, Pima and Vowel were used. The real datasets can be downloaded from UCI machine learning repository and they originally do not contain missing values. Synthetic datasets are also used in the experiments because the structure and dimensions of the subspace clusters to be embedded can be controlled as per the requirements. The synthetic data generator is programmed in R. A short description of the synthetic and real datasets used in the experiments is mentioned in Table 1 and Table 2 respectively.

Table 1. Description of synthetic datasets

Count of attributes	100, 200, 300, 400, 500
Count of objects	1000
Range of each attribute	1.0 to 100.0
Standard Deviation	0.01
Percentage of outliers	10
Size of subspace clusters	10 objects and 10 attributes
Count of embedded subspace Clusters	5

Table 2. Description of UCI machine learning repository datasets

Dataset	Instances	Attributes	Classes
Iris	140	4	3
Ecoli	336	7	8
Glass	214	9	6
Liver	345	6	2
Pima(Diabetes)	768	8	2
Vowel	990	13	11

In the first phase of the experiments, CLUSLINK algorithm was executed on each of the synthetic/real complete datasets and the results were recorded. In the second phase, for experiment purposes, these datasets were damaged to contain missing values at random positions across rows and columns. This process of induction of missing values is called amputation. The proportion of missing values was varied to 5%, 10% and 20% and are placed at random positions. These partially incomplete datasets were then preprocessed to fill the missing values by each of the MLR, MICE-PMM & MICE-CART methods; followed by execution of CLUSLINK algorithm on imputed datasets. The results obtained were recorded again. All the experiments were conducted on a personal computer having Intel(R) Pentium® P6200 CPU @ 2.13 GHz, 2.00 GB RAM, Windows 7 Operating System and R version 3.4.3. Figure 2 shows a snapshot of a small portion of a synthetic dataset damaged to contain missing values. The missing values are indicated as 'NA'.

Evaluation of Output Quality

For conventional clustering algorithms, clustering quality is evaluated under the assumption that all attributes belonging to a record in the dataset are part of the output clusters. However, the evaluation measures for subspace clustering check for subset of attributes relevant to a given subspace cluster. Hence these measures are called subspace and object based evaluation measures (Parsons et al., 2004). The popular subspace clustering evaluation measures are: accuracy, F-measure, Clustering Error (CE) and Relative Non Intersecting Area (RNIA). The desirable value of accuracy and F-measure is 1.0 and for the measures RNIA and CE, it is expected to be 0.0.

Experimental Results

This section highlights the effect of imputation on the output quality. Table 3, Table 4, Table 5, Table 6 respectively show Accuracy, F1-value, CE and RNIA of output produced on synthetic datasets described in Table 1. The tables show comparison of the results on original complete data and imputed data (containing 5%, 10% and 20% missing values). Figure 3 shows the quality of clustering output on original real datasets and corresponding imputed datasets. Table 7 and Table 8 respectively show the time of imputation on synthetic and real datasets by each of the imputation methods.

ANALYSIS OF EXPERIMENTAL RESULTS

1. The results of the experiments show that, MLR based imputation, MICE-PMM and MICE-CART imputation methods effectively fill the missing values by appropriate substitutes.
2. The subspace clustering output quality (expressed in terms of Accuracy, F1-value, CE and RNIA) is approximately the same on original (complete) and imputed datasets for all imputation methods.

Figure 2. Example of a synthetic dataset containing missing values at random positions

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
1	37.087	81.175	11.689	41.272	65.395	65	77.206	74.426	91.46	57.294
2	14.847	92.811	NA	31.47	82.235	97.897	32.407	35.261	81.567	19.943
3	46.545	51.688	74.557	NA	55.098	93.535	95.499	10.309	77.151	67.52
4	71.772	91.663	98.848	NA	1.652	50.915	15.048	31.251	43.38	38.6
5	20.116	62.08	61.139	23.821	25.695	89.909	87.282	85.249	84.708	66.309
6	23.249	43.431	89.113	55.166	35.846	18.927	81.137	76.371	NA	62.593
7	27.026	61.036	53.079	65.095	32.562	9.774	20.377	45.975	20.727	90.478
8	43.62	45.91	19.544	91.364	72.533	NA	88.108	15.292	94.952	50.006
9	67.23	98.804	58.564	67.077	55.405	NA	6.705	89.541	95.871	39.63
10	35.706	NA	30.208	NA	97.683	64.117	NA	39.345	36.217	28.52

Table 3. Accuracy on imputed synthetic data with 1000 records

Imputation method	Missing %	Accuracy on varied attribute count				
		#100	#200	#300	#400	#500
	0%	1	1	1	1	1
MLR	5%	1	0.98	0.96	1	1
	10%	1	0.98	0.96	1	0.998
	20%	0.96	1	0.96	1	0.98
MICE-PMM	5%	0.98	1	1	0.96	0.94
	10%	1	1	0.98	0.98	0.95
	20%	0.96	1	0.98	0.94	0.96
MICE-CART	5%	0.98	0.98	0.94	0.94	0.94
	10%	0.98	0.94	0.94	0.96	0.97
	20%	0.98	0.94	0.94	0.94	0.96

Table 4. F1-value on imputed synthetic data with 1000 records

Imputation method	Missing %	F1-Value on varied attribute count				
		#100	#200	#300	#400	#500
	0%	1.000	1.000	1.000	1.000	1.000
MLR	5%	1.000	0.989	0.978	1.000	1.000
	10%	1.000	0.989	0.979	1.000	0.999
	20%	0.979	1.000	0.979	1.000	0.989
MICE-PMM	5%	0.989	1.000	1.000	0.979	0.989
	10%	1.000	1.000	0.98	0.98	0.989
	20%	0.978	1.000	0.989	0.968	0.989
MICE-CART	5%	0.989	0.98	0.989	0.989	0.989
	10%	0.98	0.989	0.989	0.989	0.989
	20%	0.98	0.989	0.989	0.989	0.989

3. It is also observed that, there is not much difference in the output subspace clustering quality produced by the three imputation methods – MLR based imputation, MICE-PMM and MICE-CART on synthetic as well as real datasets.
4. The selection of imputation method mainly affects the imputation time. MICE is an iterative approach; and due to multiple iterations, it needs large imputation time. MICE-CART requires the highest imputation time.
5. The MLR based imputation method is more suitable for imputation of high dimensional numerical datasets. The implications of the results are more relevant to the decision support systems as a DSS is mainly built using high dimensional datasets that are mostly incomplete.

Table 5. Clustering Error on imputed synthetic data with 1000 records

Imputation method	Missing %	Clustering Error on varied attribute count				
		#100	#200	#300	#400	#500
	0%	0.00	0.00	0.00	0.00	0.00
MLR	5%	0.00	0.02	0.04	0.00	0.00
	10%	0.00	0.02	0.04	0.00	0.02
	20%	0.04	0.00	0.04	0.00	0.02
MICE-PMM	5%	0.02	0.00	0.00	0.04	0.04
	10%	0.00	0.00	0.02	0.02	0.04
	20%	0.04	0.00	0.02	0.06	0.04
MICE-CART	5%	0.02	0.02	0.02	0.02	0.04
	10%	0.02	0.02	0.02	0.04	0.04
	20%	0.02	0.02	0.02	0.06	0.06

Table 6. RNIA on imputed synthetic data with 1000 records

Imputation method	Missing %	Relative Non-intersecting Area on varied attribute count				
		#100	#200	#300	#400	#500
	0%	0.00	0.00	0.00	0.00	0.00
MLR	5%	0.00	0.02	0.04	0.00	0.00
	10%	0.00	0.02	0.04	0.00	0.02
	20%	0.04	0.00	0.04	0.00	0.02
MICE-PMM	5%	0.02	0.00	0.00	0.04	0.02
	10%	0.00	0.00	0.02	0.02	0.02
	20%	0.04	0.00	0.02	0.06	0.04
MICE-CART	5%	0.02	0.02	0.02	0.02	0.06
	10%	0.02	0.02	0.02	0.04	0.06
	20%	0.02	0.02	0.02	0.06	0.06

CONCLUDING REMARKS AND FUTURE WORK

A discussion on various missing data imputation methods is presented in this paper. The importance of missing data imputation in the context of decision support systems and machine learning algorithms is also discussed. The process of imputation is attractive and also dangerous. It is attractive because after imputation, the user gets into a state of pleasure as the data seems to be complete. It is also dangerous because, if the assumptions about the data are not realistic it may result in the introduction of biases. Dropping records vs. imputing data are not two mutually exclusive choices. In some situations, simple methods that do not involve complicated processing are found to perform better. It may also happen that data imputation and complete case analysis both show almost similar results. Despite the good performance shown in terms of other quality evaluation metrics, a significant and often disregarded criterion in the evaluation

Table 7. Time taken for imputation of synthetic datasets

Imputation method	Missing %	Imputation Time in Seconds on varied attribute count				
		#100	#200	#300	#400	#500
MLR	5%	2.29	0.43	0.31	1.02	1.01
	10%	0.21	0.41	0.56	0.73	0.97
	20%	0.17	0.35	0.59	0.75	7674.27
MICE-PMM	5%	93.5	522.45	1712.84	3793.55	3982.30
	10%	76.58	555.94	1807.21	3799.36	4439.65
	20%	62.45	541.15	1630.68	3486.94	4857.27
MICE-CART	5%	2878.25	13543.77	23769.00	44438.20	64675.54
	10%	2952	14678.44	25574.65	46987.54	67224.25
	20%	3077.89	15975.69	26846.76	47656.36	69763.67

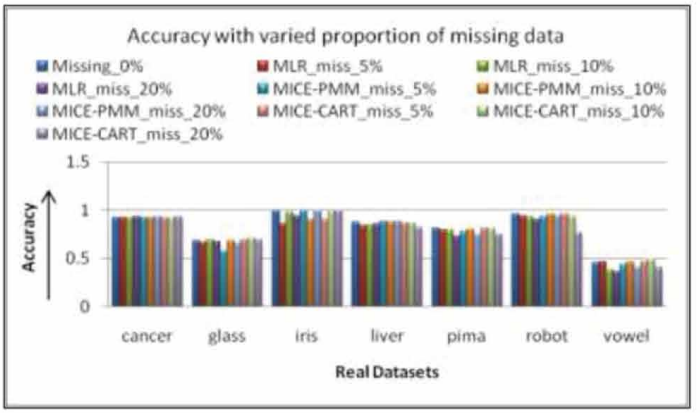
Table 8. Time taken for imputation of real datasets

Imputation method	Missing %	Imputation Time in Seconds on real datasets						
		Cancer	Glass	Iris	Liver	Pima	Robot	Vowel
MLR	5%	0.13	0.14	0.14	0.13	0.17	0.13	0.11
	10%	0.19	0.19	0.14	0.14	0.13	0.20	0.20
	20%	0.10	0.14	0.09	0.06	0.09	0.17	0.19
MICE-PMM	5%	0.63	0.46	0.16	0.31	0.55	2.73	0.68
	10%	1.19	0.70	0.22	0.39	0.86	5.65	1.45
	20%	0.70	0.39	0.13	0.47	1.00	5.48	1.65
MICE-CART	5%	12.03	6.31	1.92	4.09	16.26	20.64	21.23
	10%	20.95	9.28	2.58	6.58	17.47	45.6	38.89
	20%	10.98	4.76	1.49	6.41	11.89	31.84	35.15

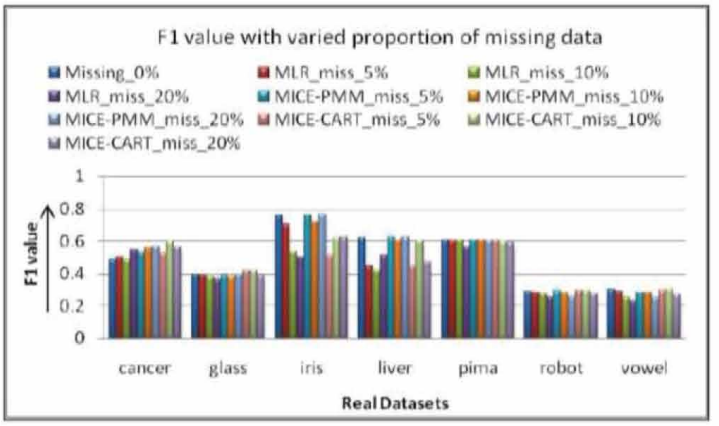
of imputation procedures is the consequence of imputation on structure of data and resulting distortion of estimates. Imputed data may not be necessarily usable if change in underlying distribution impacts drastically on the decision making process. Though missing data imputation is a common preprocessing step, it is not recommended in all situations, specifically when more than half of the records are incomplete. Each data imputation strategy has some pros and cons; performance may be better on certain datasets, whereas it may worsen on different data and it largely depends on the missing pattern. It is also recommended to provide original incomplete dataset along with the imputed dataset to evaluate various alternatives available. Future research in this area can be extended for developing efficient but simple methods that can handle all sorts of missingness.

Declaration of Interest: None
Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

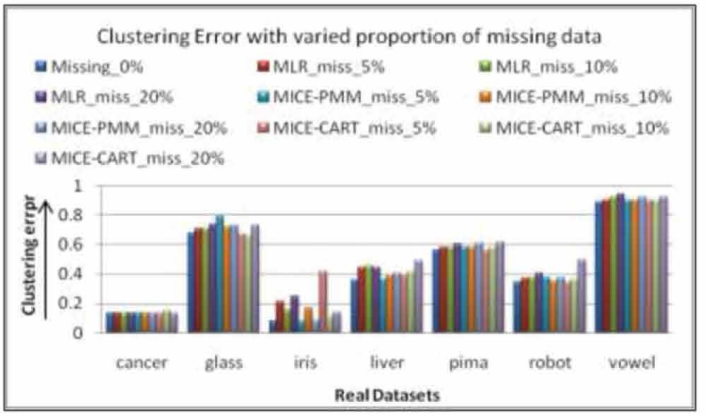
Figure 3. Subspace clustering quality on real datasets containing varied percentage of missing values



(a)Accuracy



(b)F1-value



(c)Clustering Error

REFERENCES

Abavisani, M., & Patel, V. M. (2018). Multimodal sparse and low-rank subspace clustering. *Information Fusion*, *39*, 168–177. doi:10.1016/j.inffus.2017.05.002

Acuña, E., & Rodriguez, C. (2004). The Treatment of Missing Values and its Effect on Classifier Accuracy. In Classification, Clustering, and Data Mining Applications (pp. 639–647). Springer Berlin Heidelberg. doi:10.1007/978-3-642-17103-1_60

Alghawli, A. S. (2022). Complex methods detect anomalies in real time based on time series analysis. *Alexandria Engineering Journal*, *61*(1), 549–561. doi:10.1016/j.aej.2021.06.033

Altukhova, O. (2020). Choice of method imputation missing values for obstetrics clinical data. *Procedia Computer Science*, *176*, 976–984. doi:10.1016/j.procs.2020.09.093

Armina, R., Mohd Zain, A., Ali, N. A., & Sallehuddin, R. (2017). A Review On Missing Value Estimation Using Imputation Algorithm. *Journal of Physics: Conference Series*, *892*, 012004. doi:10.1088/1742-6596/892/1/012004

Awan, S. E., Bennamoun, M., Sohel, F., Sanfilippo, F., & Dwivedi, G. (2021). Imputation of missing data with class imbalance using conditional generative adversarial networks. *Neurocomputing*, *453*, 164–171. doi:10.1016/j.neucom.2021.04.010

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research*, *20*(1), 40–49. doi:10.1002/mpr.329 PMID:21499542

Bai, B. M., Mangathayaru, N., & Rani, B. P. (2015). An Approach to Find Missing Values in Medical Datasets. *Proceedings of the The International Conference on Engineering & MIS 2015 - ICEMIS '15*, 1–7. doi:10.1145/2832987.2833083

Biessmann, F., Salinas, D., Schelter, S., Schmidt, P., & Lange, D. (2018). "Deep" Learning for Missing Value Imputationin Tables with Non-Numerical Data. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2017–2025. doi:10.1145/3269206.3272005

Burgette, L. F., & Reiter, J. P. (2010). Multiple Imputation for Missing Data via Sequential Regression Trees. *American Journal of Epidemiology*, *172*(9), 1070–1076. doi:10.1093/aje/kwq260 PMID:20841346

Chen, Y., Zhang, L., & Yi, Z. (2018). Subspace clustering using a low-rank constrained autoencoder. *Information Sciences*, *424*, 27–38. doi:10.1016/j.ins.2017.09.047

Cios, K. J., & William Moore, G. (2002). Uniqueness of medical data mining. *Artificial Intelligence in Medicine*, *26*(1–2), 1–24. doi:10.1016/S0933-3657(02)00049-0 PMID:12234714

Cro, S., Morris, T. P., Kahan, B. C., Cornelius, V. R., & Carpenter, J. R. (2020). A four-step strategy for handling missing outcome data in randomised trials affected by a pandemic. *BMC Medical Research Methodology*, *20*(1), 208. doi:10.1186/s12874-020-01089-6 PMID:32787782

de Rassenfosse, G., & Seliger, F. (2021). Imputation of missing information in worldwide patent data. *Data in Brief*, *34*, 106615. doi:10.1016/j.dib.2020.106615 PMID:33354599

Dong, W., Fong, D. Y. T., Yoon, J., Wan, E. Y. F., Bedford, L. E., Tang, E. H. M., & Lam, C. L. K. (2021). Generative adversarial networks for imputing missing data for big data clinical research. *BMC Medical Research Methodology*, *21*(1), 78. doi:10.1186/s12874-021-01272-3 PMID:33879090

Dua, D., & Karra Taniskidou, E. (2017). *UCI Machine Learning Repository*. http://archive.ics.uci.edu/ml

Faisal, S., & Tutz, G. (2021). Imputation methods for high-dimensional mixed-type datasets by nearest neighbors. *Computers in Biology and Medicine*, *135*, 104577. doi:10.1016/j.compbiomed.2021.104577 PMID:34216892

Gelman, A., & Hill, J. (2010). Missing-data imputation. In *Data Analysis Using Regression and Multilevel/Hierarchical Models* (pp. 529–544). Cambridge University Press. doi:10.1017/CBO9780511790942.031

Gómez-Carracedo, M. P., Andrade, J. M., López-Mahía, P., Muniategui, S., & Prada, D. (2014). A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets. *Chemometrics and Intelligent Laboratory Systems*, *134*, 23–33. doi:10.1016/j.chemolab.2014.02.007

Gordon, A. D., Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and Regression Trees. *Biometrics*, *40*(3), 874. doi:10.2307/2530946

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory. *Prevention Science*, *8*(3), 206–213. doi:10.1007/s11121-007-0070-9 PMID:17549635

Gruenwald, L., Yang, H., Sadik, M. S., & Shukla, R. (2010). Using data mining to handle missing data in multi-hop sensor network applications. *Proceedings of the Ninth ACM International Workshop on Data Engineering for Wireless and Mobile Access* - MobiDE'10, 9. doi:10.1145/1850822.1850825

Grzymala-Busse, J. W., & Grzymala-Busse, W. J. (n.d.). Handling Missing Attribute Values. In *Data Mining and Knowledge Discovery Handbook* (pp. 37–57). Springer-Verlag. doi:10.1007/0-387-25465-X_3

Guastella, D. A., Marcillaud, G., & Valenti, C. (2021). Edge-Based Missing Data Imputation in Large-Scale Environments. *Information (Basel)*, *12*(5), 195. doi:10.3390/info12050195

Gupta, A., Liu, T., & Shepherd, S. (2020). Clinical decision support system to assess the risk of sepsis using Tree Augmented Bayesian networks and electronic medical record data. *Health Informatics Journal*, *26*(2), 841–861. doi:10.1177/1460458219852872 PMID:31195874

H., M. A., K.A., N. D., Md Tahir, N., Iffah Abd Latiff, Z., Huzaimy Jusoh, M., & Akimasa, Y. (2021). Missing data imputation of MAGDAS-9's ground electromagnetism with supervised machine learning and conventional statistical analysis models. *Alexandria Engineering Journal*. Advance online publication. doi:10.1016/j.aej.2021.04.096

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. doi:10.1016/C2009-0-61819-5

Henry, A. J., Hevelone, N. D., Lipsitz, S., & Nguyen, L. L. (2013). Comparative methods for handling missing data in large databases. *Journal of Vascular Surgery*, *58*(5), 1353–1359.e6. doi:10.1016/j.jvs.2013.05.008 PMID:23830314

Holmes, I., & Rubin, G. M. (2002). An expectation maximization algorithm for training hidden substitution models. Journal of Molecular Biology, 317(5), 753–764. doi:10.1006/jmbi.2002.5405

Jakobsen, J. C., Gluud, C., Wetterslev, J., & Winkel, P. (2017). When and how should multiple imputation be used for handling missing data in randomised clinical trials - A practical guide with flowcharts. *BMC Medical Research Methodology*, *17*(1), 162. doi:10.1186/s12874-017-0442-1 PMID:29207961

Johnson, Z. C., Johnson, B. G., Briggs, M. A., Snyder, C. D., Hitt, N. P., & Devine, W. D. (2021). Heed the data gap: Guidelines for using incomplete datasets in annual stream temperature analyses. *Ecological Indicators*, *122*, 107229. doi:10.1016/j.ecolind.2020.107229

Jørgensen, A. W., Lundstrøm, L. H., Wetterslev, J., Astrup, A., & Gøtzsche, P. C. (2014). Comparison of Results from Different Imputation Techniques for Missing Data from an Anti-Obesity Drug Trial. *PLoS One*, *9*(11), e111964. doi:10.1371/journal.pone.0111964 PMID:25409438

Kang, Z., Zhou, W., Zhao, Z., Shao, J., Han, M., & Xu, Z. (2020). Large-Scale Multi-View Subspace Clustering in Linear Time. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(04), 4412–4419. doi:10.1609/aaai.v34i04.5867

Kelkar, B. A., Rodd, S. F., & Kulkarni, U. P. (2019). Estimating distance threshold for greedy subspace clustering. *Expert Systems with Applications*, *135*, 219–236. doi:10.1016/j.eswa.2019.06.011

Khan, H., Wang, X., & Liu, H. (2021). Missing value imputation through shorter interval selection driven by Fuzzy C-Means clustering. *Computers & Electrical Engineering*, *93*, 107230. doi:10.1016/j.compeleceng.2021.107230

Khan, S. I., & Hoque, A. S. M. L. (2020). SICE: An improved missing data imputation technique. *Journal of Big Data*, *7*(1), 37. doi:10.1186/s40537-020-00313-w PMID:32547903

Kim, M., Merrill, J. T., Wang, C., Viswanathan, S., Kalunian, K., Hanrahan, L., & Izmirly, P. (2019). SLE clinical trials: Impact of missing data on estimating treatment effects. *Lupus Science & Medicine*, *6*(1), e000348. Advance online publication. doi:10.1136/lupus-2019-000348 PMID:31649825

Lakshminarayan, K., Harp, S. A., & Samad, T. (1999). Imputation of Missing Data in Industrial Databases. *Applied Intelligence*, *11*(3), 259–275. doi:10.1023/A:1008334909089

Little, R. J. A., & Rubin, D. B. (2019). *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc.

Liu, J., Teng, S., Fei, L., Zhang, W., Fang, X., Zhang, Z., & Wu, N. (2021). A novel consensus learning approach to incomplete multi-view clustering. *Pattern Recognition*, *115*, 107890. doi:10.1016/j.patcog.2021.107890

Löw, N., Hesser, J., & Blessing, M. (2019). Multiple retrieval case-based reasoning for incomplete datasets. *Journal of Biomedical Informatics*, *92*, 103127. doi:10.1016/j.jbi.2019.103127 PMID:30771484

Ma, S., Schreiner, P. J., Seaquist, E. R., Ugurbil, M., Zmora, R., & Chow, L. S. (2020). Multiple predictively equivalent risk models for handling missing data at time of prediction: With an application in severe hypoglycemia risk prediction for type 2 diabetes. *Journal of Biomedical Informatics*, *103*, 103379. doi:10.1016/j.jbi.2020.103379 PMID:32001388

Mayer, I., Josse, J., Tierney, N., & Vialaneix, N. (2019). *R-miss-tastic: A unified platform for missing values methods and workflows*. Academic Press.

McCombe, N., Liu, S., Ding, X., Prasad, G., Bucholc, M., Finn, D. P., Todd, S., McClean, P. L., & Wong-Lin, K. (2021). Practical Strategies for Extreme Missing Data Imputation in Dementia Diagnosis. MedRxiv. 10.1101/2020.07.13.20146118

McDowell, I., & Jenkinson, C. (1996). Development standards for health measures. Journal of Health Services Research & Policy, 1(4), 238–246. doi:10.1177/135581969600100410

Morvan, M. Le, Josse, J., Moreau, T., Scornet, E., & Varoquaux, G. (2020). *NeuMiss networks: Differentiable programming for supervised learning with missing values*. Academic Press.

Murti, D. M. P., Pujianto, U., Wibawa, A. P., & Akbar, M. I. (2019). K-Nearest Neighbor (K-NN) based Missing Data Imputation. *2019 5th International Conference on Science in Information Technology (ICSITech)*, 83–88. doi:10.1109/ICSITech46713.2019.8987530

Nijman, S. W. J., Groenhof, T. K. J., Hoogland, J., Bots, M. L., Brandjes, M., Jacobs, J. J. L., Asselbergs, F. W., Moons, K. G. M., & Debray, T. P. A. (2021). Real-time imputation of missing predictor values improved the application of prediction models in daily practice. *Journal of Clinical Epidemiology*, *134*, 22–34. doi:10.1016/j.jclinepi.2021.01.003 PMID:33482294

Niu, G., Yang, Y., & Sun, L. (2021). One-step multi-view subspace clustering with incomplete views. *Neurocomputing*, *438*, 290–301. doi:10.1016/j.neucom.2021.01.080

Ordóñez Galán, C., Sánchez Lasheras, F., de Cos Juez, F. J., & Bernardo Sánchez, A. (2017). Missing data imputation of questionnaires by means of genetic algorithms with different fitness functions. *Journal of Computational and Applied Mathematics*, *311*, 704–717. doi:10.1016/j.cam.2016.08.012

Parsons, L., Haque, E., & Liu, H. (2004). Evaluating Subspace Clustering Algorithms. *Proceedings of the Fourth SIAM International Conference Data Mining, Workshop Clustering High Dimenisional Data and Its Applications*, 9.

Piri, S. (2020). Missing care: A framework to address the issue of frequent missing values;The case of a clinical decision support system for Parkinson's disease. *Decision Support Systems*, *136*, 113339. doi:10.1016/j.dss.2020.113339

Rani, P., Kumar, R., Ahmed, N. M. O. S., & Jain, A. (2021). A decision support system for heart disease prediction based upon machine learning. *Journal of Reliable Intelligent Environments*, *7*(3), 263–275. Advance online publication. doi:10.1007/s40860-021-00133-6

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, *63*(3), 581–592. doi:10.1093/biomet/63.3.581

Rubin, D. B. (Ed.). (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Inc., doi:10.1002/9780470316696

Schunk, D. (2008). A Markov chain Monte Carlo algorithm for multiple imputation in large surveys. *AStA. Advances in Statistical Analysis*, *92*(1), 101–114. doi:10.1007/s10182-008-0053-6

Shin, K., Han, J., & Kang, S. (2021). MI-MOTE: Multiple imputation-based minority oversampling technique for imbalanced and incomplete data classification. *Information Sciences*, *575*, 80–89. doi:10.1016/j.ins.2021.06.043

Smieja, M., Struski, L., Tabor, J., Zielinski, B., & Spurek, P. (2019). *Processing of missing data by neural networks*. Academic Press.

Struski, Ł., Tabor, J., & Spurek, P. (2018). Lossy compression approach to subspace clustering. *Information Sciences*, *435*, 161–183. doi:10.1016/j.ins.2017.12.056

Wang, Y., & Chaib-draa, B. (2017). An online Bayesian filtering framework for Gaussian process regression: Application to global surface temperature analysis. *Expert Systems with Applications*, *67*, 285–295. doi:10.1016/j.eswa.2016.09.018

Yao, J., Cao, X., Zhao, Q., Meng, D., & Xu, Z. (2018). Robust subspace clustering via penalized mixture of Gaussians. *Neurocomputing*, *278*, 4–11. doi:10.1016/j.neucom.2017.05.102

Yi, J., Lee, J., Kim, K. J., Hwang, S. J., & Yang, E. (2020). *Why Not to Use Zero Imputation?* Correcting Sparsity Bias in Training Neural Networks.

Yuan, Q., Longo, M., Thornton, A. W., McKeown, N. B., Comesaña-Gándara, B., Jansen, J. C., & Jelfs, K. E. (2021). Imputation of missing gas permeability data for polymer membranes using machine learning. *Journal of Membrane Science*, *627*, 119207. doi:10.1016/j.memsci.2021.119207

Zhang, Z., Liu, Z., Ma, Z., He, J., & Zhu, X. (2021). Evidence integration credal classification algorithm versus missing data distributions. *Information Sciences*, *569*, 39–54. doi:10.1016/j.ins.2021.04.008

Zhuang, J., Cui, L., Qu, T., Ren, C., Xu, J., Li, T., Tian, G., & Yang, J. (2021). A Streamlined scRNA-Seq Data Analysis Framework Based on Improved Sparse Subspace Clustering. *IEEE Access: Practical Innovations, Open Solutions*, *9*, 9719–9727. doi:10.1109/ACCESS.2021.3049807

Bhagyashri Kelkar is currently working as Associate Professor in Dept. of CSE. KIT's College of Engineering, Kolhapur, Maharashtra, India.