INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

MVI-DR: An Efficient Missing Value Imputation Method Using Decision Tree and Regression Analysis

Robindro Singh Khumukcham[1]*, Devi Mayanglambam[1], Boby Clinton Urikhimbam[1], Nazrul Hoque[1]

1 Department of Computer Science, Manipur University, Canchipur, Imphal, 795003, Manipur, India

OPEN ACCESS

Received: 10-08-2023

Accepted: 16-10-2023

Published: 14-11-2023

Citation: Khumukcham RS, Mayanglambam D, Urikhimbam BC, Hoque N (2023) MVI-DR: An Efficient Missing Value Imputation Method Using Decision Tree and Regression Analysis. Indian Journal of Science and Technology 16(43): 3862-3874. https://doi.org/10.17485/IJST/v16i43.1864

* **Corresponding author**.

rbkh@manipuruniv.ac.in

Funding: DST-SERB Start-up- Grant bearing File No: SRG/2022/001692 and UGC Start-up-Grant No: F.30-592/2021(BSR).

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: The main objective of the research work is to estimate the missing values of a dataset that contains both numeric and categorical type attributes and features. Developing a missing value imputation method to handle mixed-type data is an important problem for machine learning researchers. **Methods:** We developed a method called MVI-DR to estimate the missing values of a mixed-type dataset. The proposed MVI-DR method incorporates linear regression (LiR) and Decision Trees (DT) to compute the missing values for numeric and categorical data, respectively. The proposed MVI-DR method is validated using five classifiers viz., Logistic Regression (LoR), Support Vector Machine (SVM), k-Nearest Neighbor (k-NN), DT, and Random Forest (RF) on 9 mixed-type datasets taken from UCI and Kaggle repositories. **Findings:** From the experimental results, we observed that the proposed MVI-DR method effectively estimates the missing values for both numeric and categorical data types. Especially on the Car, Lung Cancer, Thyroid, Melb, and Penguins datasets, the proposed method gives 75.7% accuracy, whereas the traditional method gives 75.6% accuracy using the LR model. Similarly, on the Lung Cancer dataset, MVI-DR yields 66.3%, 57.2%, 61.1%, 51%, 60.7%, and traditional one gives 62.2%, 53.2%, 55.8%, 47.4%, 58.6%, using LR, k-NN, SVM, DT and RF classifiers, respectively. In addition to accuracy, the proposed method yields better results on most of the datasets in terms of MCC. Moreover, we found that the proposed method performed better on high-dimensional mixed-type datasets. **Novelty:** A new missing value imputation method called MVI-DR is developed. The method can handle both numeric and categorical data types. The MVI-DR method is evaluated in terms of Accuracy, F_1-score and MCC.

Keywords: Numeric; Categorical; Imputation; MVIDR; Machine Learning; Mixed type

1 Introduction

Missing value imputation is an important data-handling problem of Machine Learning (ML). The occurrence of missing values on a dataset creates several problems, such as degradation of ML performance, data analysis challenges, and biased results by the prediction models [1]. Researchers find it difficult to apply ML models to datasets with missing values since the severity of the missing values relies on the amount of missing data, the pattern of missing data, and the mechanism causing the missingness of the data. Missing value imputation (MVI) is one of the primary approaches for dealing with missing values in data preprocessing since they may cause ambiguity in data analysis [2]. The MVI method replaces the missing data points with the evaluated value based on the remaining available information, utilizing conventional approaches to maintain the dataset. The performance of learning models can be improved using MVI approaches in ML models [3]. There are different imputation methods for filling up the missing values in a dataset, which are limited to handling only one data type, either numerical or categorical. There are limited research studies on imputing the missing values for mixed data types consisting of numerical and categorical data. It is a challenging field of research in data analysis for handling such mixed data due to the requirement to pay attention to the relationships between variables evaluated on various scales.

We have discovered a significant number of research articles on MVI techniques in the literature. ML-based imputation techniques have become vital for handling missing values during the data preprocessing phases of ML applications. These imputation techniques substantially influence ML models since they preserve information, boost model performance, improve data quality, and increase the model's ability to generalize. We have discussed several studies on ML/Deep Learning (DL) methods to address the missing values in any dataset. Wang, S. et al. [4] proposed an MVI method based on Artificial Neural Network (ANN) for classification problems. The proposed method fills in the missing values with the value provided by the trained ANN model by using the complete records in the dataset as the training dataset. In [5], a novel CNN-based imputation approach is explored. The performance of the method is assessed using the optimizers Rmsprop, Adam, Nadam, Stochastic Gradient Descent (SGD), and Adagrad. The SGD optimizer is discovered to be more reliable in predicting the missing values among them. The k-NN imputation methods effectively address missing data in various contexts because of their flexibility, adaptability, and ability to capture local data links. As a result, the incorporation of k-NN in various MVI approaches in diverse fields has therefore been noticed in several research investigations [6–9]. The authors in [10,11] were inspired to develop denoising autoencoders for the imputation of missing values in a variety of missing data situations, including the missing data in databases for food composition, by the success of deep learning models-based imputation methods. The study discussed above demonstrates that employing ML/DL techniques to address missing data issues has seen considerable success; however, there are still limitations to the research we have looked at in mixed data types.

It is necessary to incorporate multiple imputation techniques into a single framework to handle mixed-type data in a dataset. In order to handle mixed-type data, we have explored the use of multiple imputation methods. When dealing with missing data of mixed type, the reliable and adaptable Multivariate Imputation by Chained Equation (MICE) method efficiently imputes values while also considering the relationships between the variables [12]. In order to boost its ability to adapt, MICE, one of the most successful MVI approaches, has undergone quite a number of improvements. Khan, S.I., et al. [13] established the Single Centre Imputation from Multiple Chained Equation (SICE), an enhancement of the well-known MICE with two forms for numeric and categorical data. The integration of ensemble learning, deep learning, and clustering with MICE has shown fewer biases and enhanced imputation accuracy while dealing with a significant amount of missing data [14,15]. It is common practice in clustering problems in data analysis to combine clustering and imputation techniques when the dataset contains missing or incomplete data. As we have seen in [16,17], several methods are available for incorporating imputation techniques in clustering mixed numerical and categorical problems, such as the k-CMM and k-POD. When addressing missing values and simulating relationships, the Gaussian Copula is a valuable tool because it makes assumptions about the linearity and normality of the data. The shortcomings of the original Gaussian Copula model can be improved by incorporating methods like Expected Maximization [18], Randomized quasi-Monte Carlo [19], and a probabilistic imputation method [20]. These methods help estimate copula parameters from incomplete mixed data, include unordered multinomial variables, and model mixed-type variables' unordered characteristics.

In order to solve the issues with missing data in ML applications, many techniques, ranging from traditional to ML-based ways of imputation, have been developed, as we have seen from the aforementioned reviewed works. Dealing with datasets comprising mixed-type data necessitates employing multiple imputation methods to ensure the imputation approach can handle data uncertainty while collaborating with other imputation methods. The use of ML approaches in handling missing values has been the focus of numerous publications in recent years. Numerous methods to effectively impute missing values in the dataset utilize supervised, semi-supervised, and unsupervised ML techniques. There have been a few research efforts on handling mixed-type data despite the fact that there have been numerous research efforts on ML approaches for missing value imputation. This study aims to develop a missing value imputation method for a mixed-type dataset. Let's consider a dataset

D with n instances where some instances may have missing values for both numerical and categorical features. The problem is formulated to impute the best possible values for the corresponding missing value of an instance so that any ML model can yield better predictive accuracy during the learning process.

The main contributions of this paper can be summarized as follows:

1. An effective missing values imputation method is developed.
2. The proposed method can impute missing values for mixed-type data.
3. The proposed MVI-DR combines the DT and Regression method to impute mixed-type missing values.
4. The effectiveness of the proposed method is evaluated using LoR, SVM, k-NN, DT, and RF classifiers on 9 datasets.
5. Performances of the proposed method are compared with traditional methods using Accuracy, F_1-score, and MCC.

The novelty of the proposed MVI-DR method is that it can compute missing values for numeric and categorical features using a single conceptual framework. Many existing methods convert the categorical data to numeric for applying mean or median during missing value imputation. Modern clustering-based methods also convert the mixed-type data to uniform type before computing missing values. The proposed method would be suitable for data science researchers on missing value imputations.

The paper is organized as follows. The background study of this paper is discussed in section 2. The proposed method, algorithm, and conceptual framework are discussed in section 3. The experimental analysis is analyzed in section 4. Lastly, the conclusion and future work are reported in section 5.

2 Methodology

Missing value imputation is an important data preprocessing step of Machine Learning. The presence of missing values might cause a fatal error in the ML model that degrades the performance of the model. ML researchers always face the problem of missing values during model training on real-life data. Many research studies on MVI approaches have been published over the years. ML-based MVI techniques have gained importance for addressing these issues since they replace missing data with predicted values. Several investigations on ML-based missing value imputation techniques have been proposed. Some of these techniques include KNN, SVM, DT, Clustering imputation, etc. The KNN imputation approach is gaining popularity due to its simplicity. Still, it has limitations due to the fact that the big data settings have not been investigated and the need to choose the best imputation parameters. SVM is also one of the more extensively used approaches for missing value imputation, potentially impacting the model's accuracy and usefulness. The use of DT for MVI has also risen significantly because it can handle both categorical and numerical data; however, it has drawbacks, such as building complex trees that take time but have low bias. Despite their rising popularity, clustering techniques are rarely employed since they are inaccurate when addressing missing data.

Despite the fact that there are multiple studies on MVI methods for handling missing values in a given dataset, there are limitations in the methods that deal with mixed-type datasets of numeric and categorical variables. Many popular approaches, such as "mean", "median", and "mode," make it simple to impute missing values but struggle with mixed-type data since they concentrate mostly on single imputation. When imputation is performed on a mixed-type dataset, many conventional methods transform categorical data into numerical to apply mean or median. To overcome these limitations, we developed the "MVI-DR: An Efficient Missing Value Imputation Method using DT and Regression Analysis" method, which effectively imputes missing values in numerical and categorical data. The proposed MVI-DR method makes it possible to impute missing values in mixed-type datasets by computing the missing values simultaneously for numerical and categorical variables.

Initially, the original dataset D is divided into two halves, viz., numerical and categorical, based on the feature types. Next, the method divides the numerical part into two halves: numerical data without missing values and numerical data with missing values. Similarly, the categorical part of the dataset is also divided into categorical data without missing values and categorical data with missing values. Here, the proposed method is applied to two distinct parts: numerical data with missing values and categorical data with missing values. The missing values are computed using regression analysis and a decision tree classifier. To compute the missing values present in the numeric part, we consider the numeric data without having missing values and compute the corresponding missing value for a feature by computing linear regression on the numerical data without having missing values. The method removes the features from the numerical data without missing values and considers them as training feature label data (say, y), whereas the remaining features of the numerical data without missing values are used as training feature sets (say, X). Then, the method computes the missing values using linear regression on X and y. In the same way, the method computes the missing value of the categorical data with missing values using Decision Tress on X and y.

To discuss the proposed method, we used different symbols, and their meanings are shown in Table 1 .

The proposed method is implemented as follows:

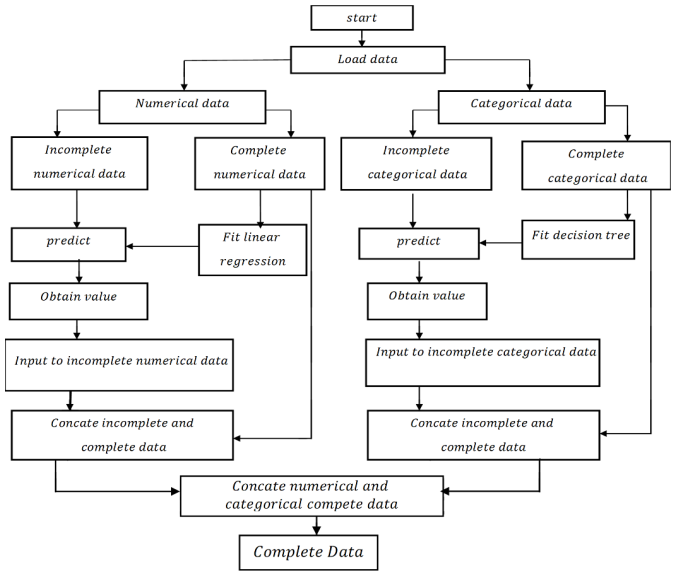


Fig 1. Conceptual Framework of the Proposed Method

• In the first step, the dataset D is divided into two sub-datasets based on the data type: numerical data N_d and categorical data C_d. Further, split N_d into two parts: one with missing values, N_{id} and the other with no missing values, N_{cd}. Similarly, the categorical dataset is also split into two parts: complete categorical data C_{cd} without missing values and incomplete categorical data C_{id} with missing values.

• In the next step Linear Regression (LiR) is fitted in to N_{cd} . Then the predicted values are imputed into the missing objects of N_{id}. Thus, the imputed numerical data $N_d^{'}$ is obtained. Similarly, the decision tree is fitted into C_{cd} to impute each missing value of C_{id}. Each missing attribute is considered a class attribute to construct a DT from the complete categorical dataset C_{cd}. Thus, the imputed categorical data C' is obtained.

• Finally, the imputed numerical data $N_d^{'}$ and imputed categorical data C' are combined to get a complete dataset D with no missing values.

• The conceptual framework and pseudo code of the proposed method are presented in Figure 1 and Algorithm 1, respectively.

Table 1. Symbols and their meanings.

Symbols	Meaning
D	Dataset
N_d	Numerical dataset
N'_d	Numerical dataset after imputed
C_d	Categorical dataset
C'_d	Categorical dataset after imputed
N_{cd}	Numerical complete dataset
N_{id}	Numerical incomplete dataset
C_{cd}	Categorical complete dataset
C_{id}	Categorical incomplete dataset
R_i	Rows of incomplete dataset
f_i	Features of incomplete data
D'	Complete data

2.1 Working example

We have discussed a working example here to understand the proposed method better. Let's look at a dataset with four instances, each of which has twelve features with numerical and categorical data types. Table 2 demonstrates how to impute missing values.

Table 2. Demonstration of imputing missing values with mixed data types of numerical and categorical values.Algorithm 1 Steps of the Proposed MVI-DR Algorithm*D*

Sl. no.	Passenger Id	Survived	P class	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
1	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
2	28	0	1	Fortune, Mr. Charles Alexander	NaN	19.0	3	2	19950	263.0000	C23 C25 C27	S
3	29	1	3	O'Dwyer, Miss. Ellen" Nellie"	female	NaN	0	0	330959	7.8792	NaN	Q
The mixed data is divided into two sub-datasets: numerical and categorical data.												
Numerical data												
0	6	0	3			NaN	0	0	330877	8.4583		
1	7	0	1			54	0	0	17463	51.8625		
2	28	0	1			19	3	2	19950	263.0000		
3	29	1	3			NaN	0	0	330959	7.8792		
Categorical data												
0				Moran, Mr. James	male						NaN	Q
1				McCarthy, Mr. Timothy J	male						E46	S
2				Fortune, Mr. Charles Alexander	NaN						C23 C25 C27	S
3				O'Dwyer, Miss. Ellen "Nellie"	female						NaN	Q
Again, the numerical and categorical data are divided into complete and incomplete data												
Complete numerical data												
1	7	0	1			54	0	0	17463	51.8625		
2	28	0	1			19	3	2	19950	263.0000		
Incomplete numerical data												
0	6	0	3			NaN	0	0	330877	8.4583		
3	29	1	3			NaN	0	0	330959	7.8792		
Complete categorical data												
1				McCarthy, Mr. Timothy J	male						E46	S
Incomplete categorical data												
0				Moran, Mr. James	male						NaN	Q
2				Fortune, Mr. Charles Alexander	NaN						C23 C25 C27	S
3				O'Dwyer, Miss. Ellen "Nellie"	female						NaN	Q
For Numerical Part: Select the first row of the incomplete data with a missing value.												

Continued on next page

Table 2 continued

0	6	0	3		NaN	0	0	330877	8.4583	

A particular feature of the row that has a missing value.

0	NaN

Now X-train is taken as: Complete numerical data – column Age, where, X-train comprises of PassengerID, Survived, Pclass, SibSp, Ticket and Fare, and y-train comprises of 'Age' column.

X-train

1	7	0	1		0	0	17463	51.8625	
2	28	0	1		3	2	19950	263.0000	

y-train

1	54.0
2	19.0

And X-test is the first row of incomplete data – the column with missing values.

X-test

0	6	0	3		0	0	330877	8.4583	

Now, X-train and y-train are trained using the linear regression model. Finally, the predicted output is inserted as the value of the column 'Age' in the first row of the incomplete numerical data; the highlighted cell represents the imputed value.

Imputation of one missing value into the Age column

0	6	0	3	43	0	0	330877	8.4583	

Henceforth, for all the missing values in incomplete numerical data, a predicted value is obtained and is thus inserted into the column which has the missing value.

Imputation of missing values in the Age column

0	6	0	3	**43**	0	0	330877	8.4583	
3	29	1	3	**24**	0	0	330959	7.8792	

For categorical part: Select the first row of the incomplete data which has a missing value

0	Moran, Mr. James	male		NaN	Q

Select a particular feature of the row that has a missing value.

0	NaN

Now X-train is taken as: Complete categorical data – column Age, where, X-train comprises Name, Sex, and Embarked. And y-train comprises of 'Cabin' column.

X-train

1	McCarthy, Mr. Timothy J	male		S

y-train

0	E46

And X-test is the first row of incomplete categorical data with a missing value– missing column.

X-test

0	Moran, Mr. James	male		Q

Now, X-train and y-train are trained using the DT model. Finally, the predicted output is inserted as the value of the column 'cabin' in the first row of the incomplete numerical data; the highlighted cell represents the imputed value.

0	Moran, Mr. James	male	E30	Q

Imputation of missing values in the Cabin column

0	Moran, Mr. James	male	E30	Q
2	Fortune, Mr. Charles Alexander	male	C23 C25 C27	S
3	O'Dwyer, Miss. Ellen "Nellie"	female	C43	Q

Combination of the imputed incomplete and complete numerical data

0	6	0	3	43	0	0	330877	8.4583	
1	7	0	1	54.0	0	0	17463	51.8625	
2	28	0	1	19.0	3	2	19950	263.0000	

Continued on next page

Table 2 continued

3	29	1	3			24	0	0	330959	7.8792		
Combination of the imputed categorical incomplete data with the complete categorical data											
0				Moran, Mr. James	male						E30	Q
1				McCarthy, Mr. Timothy J	male						E46	S
2				Fortune, Mr. Charles Alexander	male						C23 C25 C27	S
3				O'Dwyer, Miss. Ellen "Nellie"	female						C43	Q
The final dataset was obtained after imputing the missing values into the original dataset.											
0	6	0	3	Moran, Mr. James	male	43	0	0	330877	8.4583	E30	Q
1	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
2	28	0	1	Fortune, Mr. Charles Alexander	male	19.0	3	2	19950	263.0000	C23 C25 C27	S
3	29	1	3	O'Dwyer, Miss. Ellen "Nellie"	female	23	0	0	330959	7.8792	C43	Q

Output: A dataset D' with all missing values imputed.
Step 1:
$N_d \rightarrow$ All records with numerical data;
$C_d \rightarrow D - N_d$;
$N_{id} \rightarrow$ All records with numerical missing values;
$N_{cd} \rightarrow N_d - N_{id}$;
$C_{id} \rightarrow$ All records with categorical missing values;
$C_{cd} \rightarrow C_d - C_{id}$;
Step 2:
for numerical data **do**
forR_i in N_{id}**do**
$f_i \rightarrow$ select column name which has missing value in R_i;
for each $f_i \in N_{id}$ do
Call linear regression algorithm L.R from N_{cd};
$X \rightarrow N_{cd} - N_{cd}[f_i]$;
$Y \rightarrow N_{cd}[f_i]$;
Fit LR (X, Y);
Imputed value $\rightarrow$ Predict $(R_i - f_i)$;
$f_i \rightarrow$ imputed value;
end for
end for
Return $N_d^{'}$
end for
for categorical data **do**
forR_i in C_{id}**do**
$f_i \rightarrow$select column name which has missing value in Ri;
for each $f_i \in C_{id}$**do**
Calling D.T. from C_{cd} considering f_i as the target variable;
$X \rightarrow C_{cd}$ $C_{cd}[f_i]$;
Y $\rightarrow C_{cd}[f_i]$;

Fit DT (X, Y);
Imputed value $\rightarrow$ Predict $(R_i \ f_i)$;
$f_i \rightarrow$ imputed value;
end for
end for
Return $C_d^{'}$
end for
Step 3:
$D^{'} \rightarrow N_d^{'} + C_d^{'}$;
Return $D^{'}$
Complexity of the algorithm:
The time complexity of the proposed method depends on the number of missing values of the instances. Let us consider,
n = number of instances
m = number of features
Thus, the time complexity is: $O(n \times m)$

3 Results and Discussion

The experiments in this study were carried out on a computer with 8 GB main memory, i5 11th Gen Intel processor having 64 bits Windows 10 home operating system. The proposed method is implemented using Python programming language, and various Python packages are used in our implementation.

Description of the datasets used

To evaluate the proposed MVI-DR method, we employed 9 datasets available in the UCI and Kaggle repositories. The datasets contain both numerical and categorical data types with missing values. The description of the dataset and their descriptions are summarized in Table 3.

Table 3. Dataset Description

Sl.no.	Dataset	No. of instance	No. of attributes	Data type	No. of the class label
1	Titanic [21]	891	12	Real, Integer, Categorical	2
2	Car [22]	11914	16	Real, Integer, Categorical	5
3	Lung Cancer [23]	442	22313	Real, Integer, Categorical	2
4	Spine [24]	310	14	Real, Categorical	2
5	Bands [25]	540	40	Real, Categorical	2
6	Student Study [26]	63	10	Real, Integer, Categorical	9
7	Thyroid [27]	9172	31	Real, Integer, Categorical	2
8	Melb [28]	18396	22	Real, Integer, Categorical	3
9	Penguin [29]	344	17	Real, Integer, Categorical	3

We have observed from the experimental analysis that the proposed method outperforms the traditional missing value imputers such as mean and median in terms of Accuracy, F_1-score and MCC in datasets such as "Titanic", "Car", "Lung Cancer", "Bands", "Student study", "Thyroid", "Melb" and "Penguins" using the LoR, k-NN, SVM, DT, and RF classifiers. There are exceptions in the datasets where MVI-DR performs similarly to k-NN and SVM in terms of Accuracy and F_1-score in the "Titanic" dataset, k_NN performs poorly in the "Bands" dataset, and RF performs similarly in the "Student Study" dataset, MVI-DR performs comparably to LiR and SVM in terms of Accuracy and F_1-score in the "Thyroid" dataset, whereas LiR and k-NN perform identically in the "Penguins" dataset as demonstrated in Figures 2, 3, 4, 5, 6, 7, 8, 9 and 10.

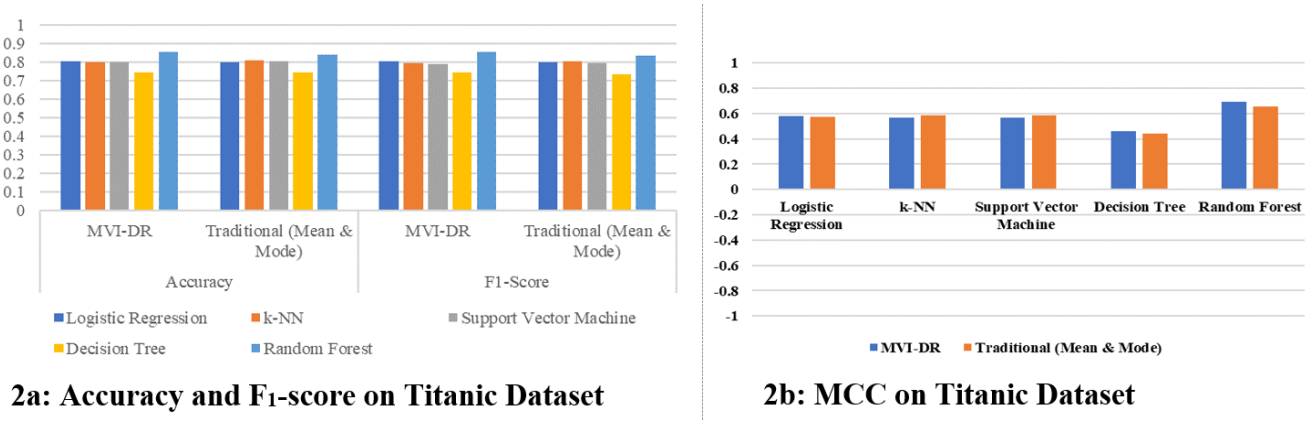


Fig 2. Accuracy, F_1-score and MCC on Titanic Dataset

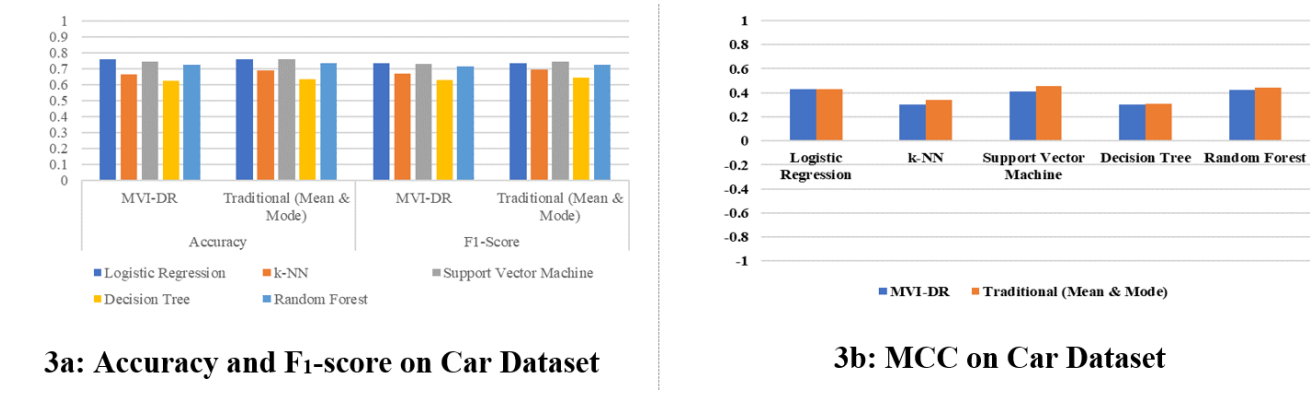


Fig 3. Accuracy, F_1-score and MCC on Car Dataset

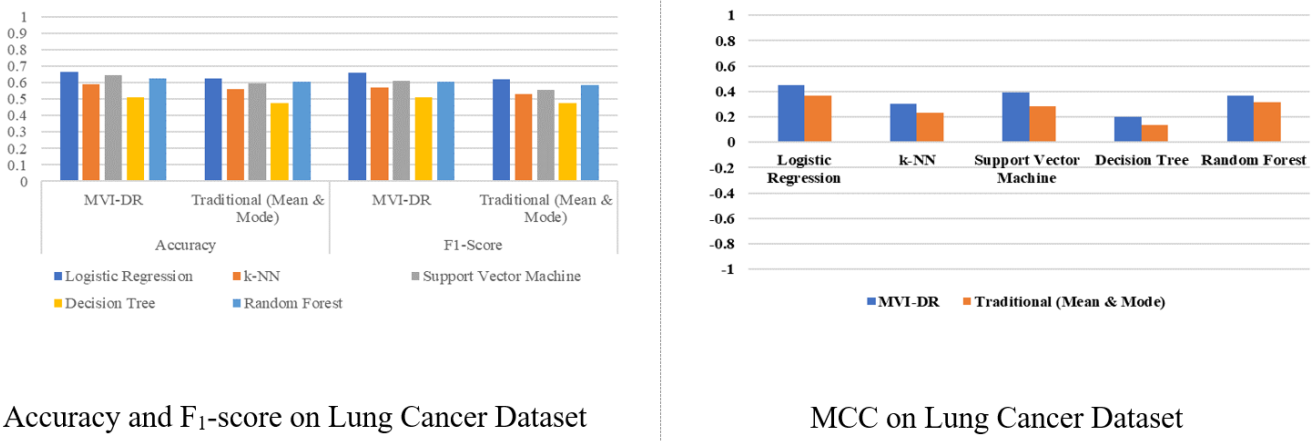
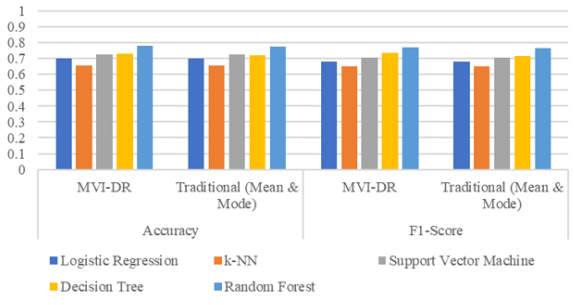
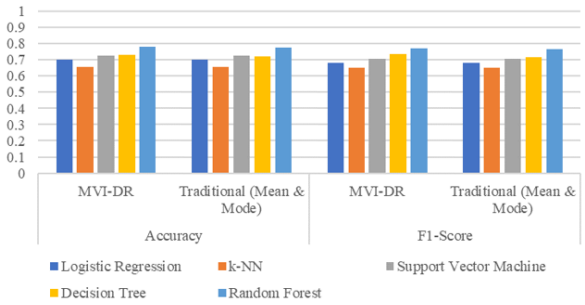


Fig 4. Accuracy, F_1-score and MCC on Lung Cancer Dataset

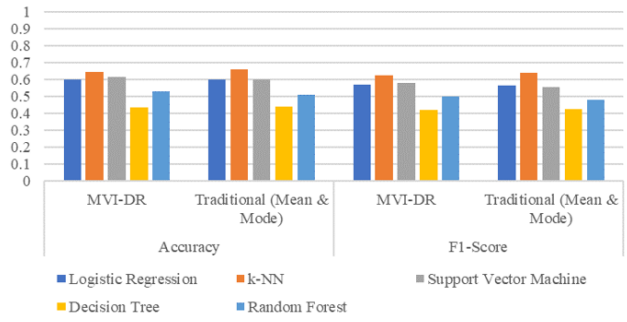


Accuracy and F_1-score on Spine Dataset

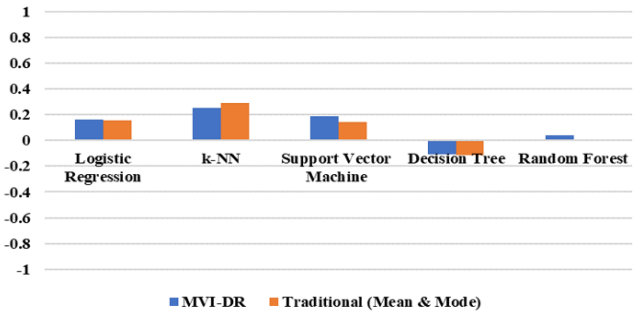


5b: MCC on Spine Dataset

Fig 5. Accuracy, F_1-score and MCC on Spine Dataset

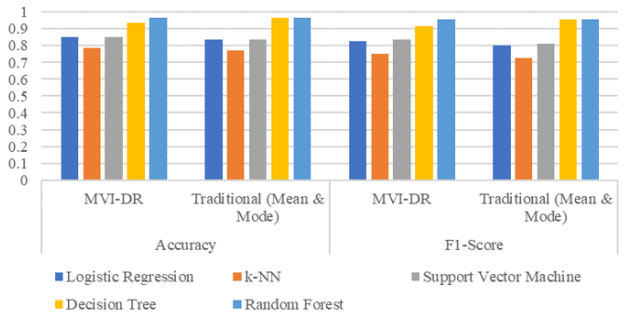


6a: Accuracy and F_1-score on Bands Dataset

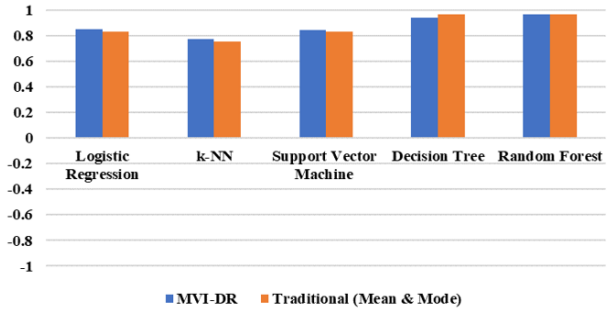


6b: MCC on Bands Dataset

Fig 6. Accuracy, F_1-scoreand MCC on Bands Dataset

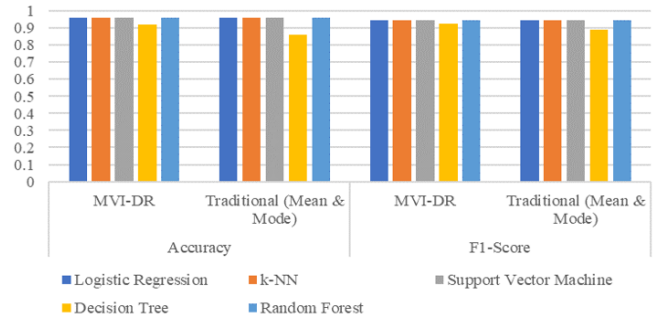


7a: Accuracy and F_1-score on Student Study Dataset

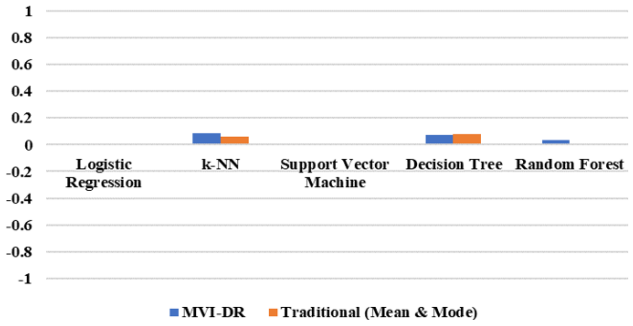


7b: MCC on Student Study Dataset

Fig 7. Accuracy, F_1-scoreand MCC on Student Study Dataset

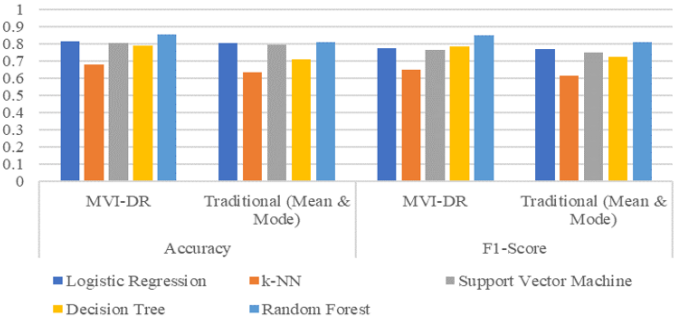


8a: Accuracy and F_1-score on Thyroid Dataset

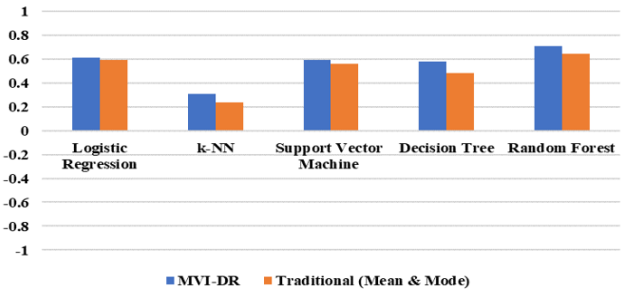


8b: MCC on Thyroid Dataset

Fig 8. Accuracy, F_1-scoreand MCC on Thyroid Dataset

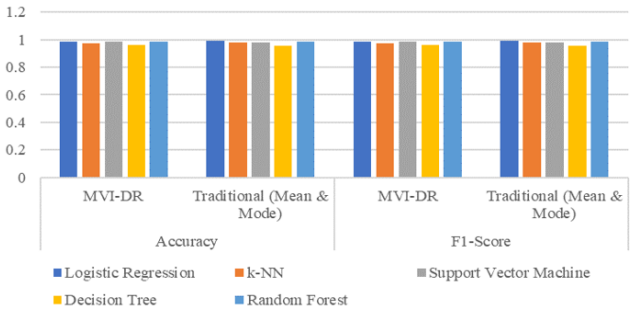


9a: Accuracy and F_1-score on Melb-Data

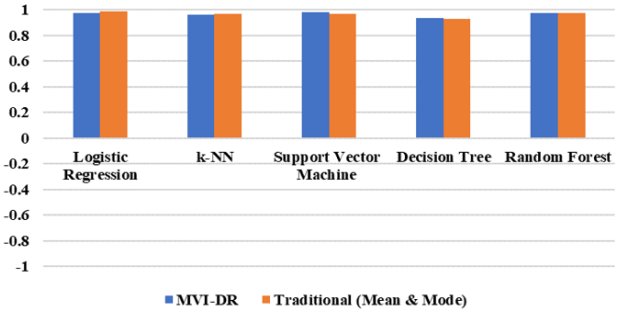


9b: MCC on Melb-Data

Fig 9. Accuracy, F_1-scoreand MCC on Melb-Data



10a: Accuracy and F_1-score on Penguin Dataset



10b: MCC on Penguin Dataset

Fig 10. Accuracy, F_1-scoreand MCC on Penguin Dataset

The proposed method performs similarly to the traditional imputation methods in terms of Accuracy, F_1-score and MCC in the "Spine" dataset. However, from the experimental evaluation, we observed that the proposed method outperforms the traditional methods using k-NN, SVM, DT, and RF classifiers in terms of MCC. Similarly, the DT and RF classifiers outperform the conventional methods in the "Spine" dataset, as demonstrated in Figure 5.

Missing value imputation is one of the essential data preprocessing steps of ML applications. Due to the increasing growth of big data applications, preprocessing enormous data is still a significant research problem. Imputation of missing values on a high-dimensional dataset is another challenging task. This work discusses an MVI method incorporating regression and classification techniques. From the experimental analysis, we observed that the proposed method performs better on most datasets. However, on some other datasets, the MVI-DR method performs similarly to the mode and median methods. The MVI technique handles the missing values of numerical and categorical data types. The method's performance is evaluated regarding prediction accuracy, F_1-score, and MCC metric on 9 datasets from the Kaggle and UCI repositories using well-known standard classifiers, viz., LoR, SVM, k-NN, DT and RF. It has been revealed that the proposed MVI-DR method can combine the two different imputation strategies to compute the missing values in mixed-type datasets concurrently.

Despite the fact that there are many research studies on methods for imputation of missing values in mixed-type data, there are few studies on methods for mixed-type data that include both numerical and categorical data. Studies in [18,20] support mixed-type data with continuous and ordinal values, while [30] can only handle mixed-type data with binary and ordinal values. The research in [19] enables the impute of mixed-type data containing ordinal, binary, and continuous variables, while [31] is restricted to the imputation of mixed-type data containing just binary and continuous variables; whereas the proposed method is the applicability of imputing numerical and categorical mixed-type data. The proposed method, however, has the drawback of not being applicable if the dataset has many missing values in a row. We intend to improve this method in future work by making it effective on any dataset that contains multiple missing values in a row.

4 Conclusion

This study introduces an efficient missing value imputation technique called "MVI-DR: An Efficient Missing Value Imputation Method using DT and Regression Analysis". Our proposed method, MVI-DR, is unique because it simultaneously computes the missing values in mixed-type datasets using DT and Regression Analysis, two distinct imputation approaches. We examined the performance of the linear regression method as an MVI technique for incomplete numerical data to predict the value for a missing field and impute the predicted value to make the dataset complete. A decision tree classifier method is also used as an MVI technique for incomplete categorical data. The proposed method's performance is evaluated using standard ML classifiers viz., LoR, SVM, k-NN, DT, and RF on 9 datasets from the UCI repository and Kaggle. Compared to the standard imputation method (i.e., mean, mode), our method gives better results in terms of Accuracy, F_1-score, and MCC. Especially the proposed method gives 75.7% accuracy, whereas the traditional method gives 75.6% accuracy using the LR model on the Car dataset. We observed that MVI-DR yields 66.3%, 57.2%, 61.1%, 51%, 60.7%, whereas traditional one gives 62.2%, 53.2%, 55.8%, 47.4%, 58.6%, using LR, k-NN, SVM, DT and RF classifiers, respectively on Lung Cancer dataset. In future work, we plan to improve the proposed method by enabling it to work well on any dataset with more than one missing value in a row.

Acknowledgement

Part of the work is funded by DST-SERB Start-up- Grant bearing File No: SRG/2022/001692 and UGC Start-up-Grant No: F.30-592/2021(BSR).

References

1) Ayilara OF, Zhang L, Sajobi TT, Sawatzky R, Bohm E, Lix LM. Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry. *Health and Quality of Life Outcomes*. 2019;17(1). Available from: https://doi.org/10.1186/s12955-019-1181-2.
2) Lin WCC, Tsai CFF. Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review*. 2020;53(2):1487–1509. Available from: https://doi.org/10.1007/s10462-019-09709-4.
3) Hasan MK, Alam MA, Roy S, Dutta A, Jawad MT, Das S. Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021). *Informatics in Medicine Unlocked*. 2021;27(2):100799. Available from: https://doi.org/10.1016/j.imu.2021.100799.
4) Wang S, Li B, Yang M, Yan Z. Missing Data Imputation for Machine Learning. In: Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering;vol. 4. Springer International Publishing. 2019;p. 67–72. Available from: https://doi.org/10.1007/978-3-030-14657-3_7.
5) Gad I, Hosahalli D, Manjunatha BR, Ghoneim OA. A robust deep learning model for missing value imputation in big NCDC dataset. *Iran Journal of Computer Science*. 2021;4(2):67–84. Available from: https://doi.org/10.1007/s42044-020-00065-z.
6) Choudhury A, Kosorok MR. Missing Data Imputation for Classification Problems. 2020. Available from: https://doi.org/10.48550/arXiv.2002.10709.

7) Fouad KM, Ismail MM, Azar AT, Arafa MM. Advanced methods for missing values imputation based on similarity learning. *PeerJ Computer Science*. 2021;7:e619. Available from: https://doi.org/10.7717/peerj-cs.619.
8) Li D, Zhang H, Li T, Bouras A, Yu X, Wang T. Hybrid Missing Value Imputation Algorithms Using Fuzzy C-Means and Vaguely Quantified Rough Set. *IEEE Transactions on Fuzzy Systems*. 2022;30(5):1396–1408. Available from: https://doi.org/10.1109/TFUZZ.2021.3058643.
9) Wu H, Li S, Shi W, Du S. FUSAIN: Combining Functional Dependencies and Clustering for Missing Values Imputation. *Engineering Letters*. 2022;(2):30. Available from: https://www.engineeringletters.com/issues_v30/issue_2/EL_30_2_15.pdf.
10) Abiri N, Linse B, Edén P, Ohlsson M. Establishing strong imputation performance of a denoising autoencoder in a wide range of missing data problems. *Neurocomputing*. 2019;365:137–146. Available from: https://doi.org/10.1016/j.neucom.2019.07.065.
11) Gjorshoska I, Eftimov T, Trajanov D. Missing value imputation in food composition data with denoising autoencoders. *Journal of Food Composition and Analysis*. 2022;112:104638. Available from: https://doi.org/10.1016/j.jfca.2022.104638.
12) Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. 2020. Available from: https://doi.org/10.18637/jss.v045.i03.
13) Khan SI, Hoque ASML. SICE: an improved missing data imputation technique. *Journal of Big Data*. 2020;7(1):37. Available from: https://doi.org/10.1186/s40537-020-00313-w.
14) Hannah S, Laqueur AB, Shev, Rose MC, Kagawa. SuperMICE: An Ensemble Machine Learning Approach to Multiple Imputation by Chained Equations. *American Journal of Epidemiology*. 2022;191(3):516–525. Available from: https://doi.org/10.1093/aje/kwab271.
15) Samad MD, Abrar S, Diawara N. Missing value estimation using clustering and deep learning within multiple imputation framework. *Knowledge-Based Systems*. 2022;249:108968. Available from: https://doi.org/10.1016/j.knosys.2022.108968.
16) Dinh DTT, Van-Nam N Huynh, Sriboonchitta S. Clustering mixed numerical and categorical data with missing values. *Information Sciences*. 2021;571:418–442. Available from: https://doi.org/10.1016/j.ins.2021.04.076.
17) Aschenbruck R, Szepannek G, Wilhelm AFX. Imputation Strategies for Clustering Mixed-Type Data with Missing Values. *Journal of Classification*. 2023;40(1):2–24. Available from: https://doi.org/10.1007/s00357-022-09422-y.
18) Zhao Y, Udell M. Missing Value Imputation for Mixed Data via Gaussian Copula. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020;p. 636–646. Available from: https://doi.org/10.1145/3394486.3403106.
19) Christoffersen B, Clements M, Humphreys K, Kjellström H. Asymptotically exact and fast Gaussian copula models for imputation of mixed data types. *Asian Conference on Machine Learning*. 2021;p. 870–885. Available from: https://doi.org/10.48550/arXiv.2102.02642.
20) Zhao Y, Townsend A, Udell M. Probabilistic Missing Value Imputation for Mixed Categorical and Ordered Data. *Advances in Neural Information Processing Systems*. 2022;35:22064–22077. Available from: https://doi.org/10.48550/arXiv.2210.06673.
21) . . Available from: https://www.kaggle.com/datasets/azeembootwala/titanic.
22) . . Available from: https://www.kaggle.com/datasets/gagandeep16/car-sales.
23) . . Available from: https://www.kaggle.com/datasets/josemauricioneuro/lung-cancer-patients-mrna-microarray.
24) . . Available from: https://www.kaggle.com/datasets/samuelcortinhas/rsna-2022-spine-fracture-detection-metadata.
25) Evans B. Cylinder Bands. UCI Machine Learning Repository. 1995. Available from: https://doi.org/10.24432/C50C7B.
26) Nabila. Student Study performance. 2022. Available from: https://doi.org/10.34740/KAGGLE/DSV/3873615.
27) . . Available from: https://www.kaggle.com/datasets/pavlokoliada/thyroid-diseases-dataset?select=train.csv.
28) . . Available from: https://www.kaggle.com/datasets/gunjanpathak/melb-data.
29) Gorman KB, Williams TD, Fraser WR. Ecological Sexual Dimorphism and Environmental Variability within a Community of Antarctic Penguins (Genus Pygoscelis). *PLoS ONE*. 2014;9(3):e90081–e90081. Available from: https://doi.org/10.1371/journal.pone.0090081.
30) Feng H, Ning Y. High-dimensional mixed graphical model with ordinal data: Parameter estimation and statistical inference. *The 22nd international conference on artificial intelligence and statistics*. 2019;p. 654–663. Available from: https://proceedings.mlr.press/v89/feng19a.html.
31) Yoon G, Carroll RJ, Gaynanova I. Sparse semiparametric canonical correlation analysis for data of mixed types. *Biometrika*. 2020;107(3):609–625. Available from: https://doi.org/10.1093/biomet/asaa007.