

Application of machine learning technics to similar case imputation for missing values in Demographic and Health Survey (DHS) data

Mr. Cyprien HABINSHUTI

University of Rwanda

Prof. François NIRAGIRE

fniragire@gmail.com

University of Rwanda

Systematic Review

Keywords: Machine learning, Imputation, SVM, KNN, MICE, Random Forest, Decision tree

Posted Date: April 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9380937/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

Background: The study sought to assess the performance of machine learning algorithms for handling missing data in demographic and Health Survey (DHS) dataset using similar case imputation method. In quantitative data analysis, imputation has the potential to greatly improve the knowledge available for mining high-quality compounds by supplying accurate predictions to fill in the gaps.

Methods: This study used data from three rounds of the Rwanda Demographic and Health Survey (RDHS) conducted between 2015 and 2020. The study was conducted using three datasets with a total of 400,656 to 459,102 observations, each including 9,002, 7,856, and 8,092 rows with 51 columns. Twenty numerical variables and thirty-one categorical variables made up the merged dataset after concatenation.

The Multiple Imputation Regression for 'm' iteration, Decision Tree based, Support Vector Machine (SVM) as classification algorithm, K-nearest neighbor (KNN) as Clustering algorithms, and Random Forest (RF) machine learning algorithms were applied and compared using performance metrics to identify the best algorithm to impute missing values.

Results: It was found that Support Vector Machine (SVM) ranked first for imputing categorical variables, with an accuracy of 100% and precision of 100%. It was followed by Decision Tree, which achieved an accuracy of 79.9% and precision of 100%, Random Forest came in third with an accuracy of 78.0% and precision of 99.2%, while KNN ranked last with an accuracy of 67.1% and precision of 69.4%.

For imputing numerical data, Random Forest performed the best, with both MSE and MAE values of 0. It was followed by Multiple Imputation by Chained Equations (MICE), which had an MSE of $1.53e^09$ and an R^2 of 0.81. Finally, Support Vector Machine (SVM) regression ranked last among the machine learning models used for numerical data imputation.

Conclusions: Performance measurements revealed that Random Forest is the optimal algorithm model for numerical variables, while Support Vector Machine classification, and Decision Tree, Random Forest are the best for categorical variables. It was concluded that the best machine learning algorithm for managing missing values, both categorically and numerically, is Random Forest.

1. Introduction

Without data, nothing is possible in the contemporary digital era, and enormous volumes are used every second. When data is complete, it can be processed to yield the expected results. However, incomplete data makes it challenging to achieve accurate outcomes. This incomplete data is often referred to as missing data or unknown values. Consequently, researchers, academicians, industry professionals, and those in the medical field strive to avoid incomplete data. To address this, it must be converted into complete data before processing. Several scholars have investigated diverse techniques for managing absent data in datasets (1).

In research, controlling missing values is crucial since their existence might introduce bias, diminish the reliability of results, and call into question the accuracy of conclusions. Missing data can cause statistical studies to be distorted, concealing the underlying relationships between variables and leading to inaccurate findings if not handled correctly (2). The research sample loses important information due to missing values. The efficiency and clarity of study findings are reduced the more missing data there are since more information is lost. Missing data can impair the validity and dependability of findings and, in certain situations, cause disruptions to the entire data structure, making analysis more difficult (2).

Missing data is a significant challenge in both research and practical applications involving real-world datasets. For example, in the fields of educational and psychological research, Peugh and Enders discovered that, on average, 9.7% of the data was missing, with some datasets having up to 67% missing values. Similarly, Rombach et al. reported that missing data percentages varied from 1% to over 70%, with a median of 25%. Additionally, missing values are prevalent in over 45% of datasets, highlighting the widespread nature of this issue even in the University of California Irvine machine learning repository (UCI repository), one of the most widely used sources of datasets for machine learning and data science researchers (3).

Addressing missing values in a dataset has been a persistent challenge across various disciplines. A key issue is that many algorithms cannot work effectively with incomplete datasets (4). The issue of missing values negatively impacts analysis results, particularly by causing biased parameter estimates (5). Because of things like human error, malfunctioning instruments, and other problems, real-world datasets are frequently noisy, redundant, incomplete, and inconsistent. In order to manage incomplete data efficiently, sophisticated algorithms are suggested to fill in the missing information. Numerous machine learning methods have been utilized to approximate and supplement these absent variables with logical substitutes (6).

Ensuring data quality has become a significant difficulty in modern software programs as data pipelines become more sophisticated and important. Maintaining stringent standards for data quality is essential for machine learning (ML) applications to provide accurate prediction outcomes and thoughtful automated judgment. A prevalent problem in data quality is the absence of values. If incomplete datasets are not adequately treated, they can cause havoc with data pipelines and negatively affect downstream machine learning applications (7).

Missing values represent one of the numerous difficulties using data from the real world. Since the quality of the data has a major impact on the accuracy and efficacy of machine learning models, researchers and data analysts need to use pertinent methods to deal with these inescapably missing values (8, 9).

This study is impactful to the different researchers in the domain of health. It plays an important role in improving the research findings for students, scholars, and scientific researchers.

2. Related works and Materials

2.1. Related work

The results from the first paper on a Classifier Ensemble Machine Learning Approach to Improve Efficiency for Missing Value Imputation settled the most accurate model was Support Vector Machine (SVM), and the stacking strategy utilizing GLM and to combine predictions from many models, such as Random Forest with svmRadial, rpart, glm, lda, and kNN was applied. The correlation between the two sub-models with the highest correlation, kNN and Logistic Regression (GLM), was just 0.517, which is still considered low because it falls significantly below the 0.75 criteria, even though it was the strongest (10).

The second papers on a survey on missing data in machine learning, on the Iris dataset, it was found that KNN imputation performed better than RF imputation when RMSE was used as the assessment metric at two different missingness ratios. On the ID fan dataset, RF imputation, however, performed better across all missingness ratios (11).

And the last paper, on Proper Imputation Techniques for Missing Values concluded that the HD imputation greatly improves prediction accuracy, especially for large datasets; C5.0 achieves strong classification accuracy across different types of data, even though it does not use all of the dataset's attributes; Expectation Maximization (EM) takes less time than KNN, especially when working with huge datasets. Mean imputation also violates assumptions about normality and weakens connections with other variables, but it might be appropriate in situations when fewer than 5% of the data are missing (12). Referring to the results from the above results researcher came to decide on the use of some of the algorithms in this study in order to apply them for handling missing data. More specifically, the researcher used KNN, SVM, Random Forest (RF), decision Tree and Multiple imputation algorithms.

2.2. Methods

The study used the longitudinal study design. It used 3 cross-section data from RDHS that were combined to make one dataset (DHS 2015–2020). The specific dataset used was children or kids record (KR) as the one with almost all variables of interest to respond to child mortality. The combined dataset was used as standardized and allow comparability according to the variables of interest. The datasets were accessed from the official website for Demographic and Health Survey (DHS) for different countries¹. The population of the study was all children whose 5 years at the time of interview as indicated in children or kids record (KR).

The study variables are identified in the dataset of children/kids that are related to child mortality. The Table 1 gives variable name and its code in demographic health Survey kids dataset.

Table 1
List of study variables

S/N	Children or kids record (KR)	Variable code
1	Smokes/uses other	V463X
2	Age at death	B6
3	Age at death (months, imputed)	B7
4	Age at first birth	V212
5	Age at first cohabitation	V511
6	Age of household head	V152
7	Age of respondent at 1st birth	V212
8	Age of respondent in 5-years group	V013
9	Antenatal care visits	M14
10	Birth order	BORD
11	Birth weight in kilograms	M19
12	Births in past year	V209
13	Breastfeeding status (Whether the child currently breastfeed at the time of survey)	M4
14	Child is twin	B0
15	Cluster number	V001
16	Current breastfeeding	V404
17	Current contraceptive method	V312
18	Current marital status	V501
19	Currently residing with husband/partner	V504
20	Date of birth	B3
21	De jure region of residence	V139
22	De jure type of place of residence	V140
23	Delivery by caesarean section	M17
24	Does not use tobacco	V463Z
25	Highest education level	S105
26	Highest educational level	V106
27	Household has: electricity	V119
28	Household head	V151
29	Household number	V002
30	Husband/partner's education	V701
31	Husband's desire for children	V621
32	live with father (Living with one of the lists)	B9
33	Number of household members	V136
34	Number of living children	V218

S/N	Children or kids record (KR)	Variable code
35	Place of delivery (health facility, resp home, elsewhere)	M15
36	Preceding birth interval	B11
37	Region	V101, V024
38	Religion	V130
39	Respondent currently working (employment)	V714
40	Respondent slept under mosquito bed net	V461
41	Respondent's year of birth	V010
42	Selected for Domestic Violence module	V044
43	Sex of child	B4
44	Sex of household head:	V151
45	Size of child at birth	M18
46	Survival status (Child is alive, 1 = Yes, 0 = No)	B5
47	Total children ever born	V201
48	Type of place of residence	V025
49	Type of place of residence	V102
50	Wealth index combined: Wealth index combined	V190
51	Wealth index for urban/rural	V190A
52	When child put to breast	V426
53	Women's individual sample weight (6 decimals)	V005
54	Year of first cohabitation	V508
55	Year of interview	V007

[Table 1: List of study variables]

The methods of dealing with missing data is the matching pattern and imputation method. The data is separated into groups based on similarities, and the missing value is imputed by choosing an algorithm-based set of observed values from each category. A number of procedures were used during the data analysis process, including data splitting, which divides the data into two halves. The training and test part where they are used to train machine learning model whereas the test part is used to evaluate the performance of machine learning model.

The model was evaluated based on performance metrics. The ones for classification models are accuracy, precision, recall or sensitivity, and F-score while the performance metrics for numerical variables were mean absolute error (MAE), mean square error (MSE), R-squared value, and adjusted R². The performance metrics for the model are compared and decide the best model accordingly. The algorithms used in the study are Multiple Imputation regression, Decision Tree based), classification (SVM), Clustering (KNN), and Random Forest (RF) and their performance metrics were compared to rank the top performing algorithm to impute missing values.

3. Results

3.1. Missingness in the dataset

The Table 2 below gives the percentage of missing per the variable in the dataset.

Table 2
Percentage of missing values per variables

SN	Variables	Missing %
1.	Women's individual sample weight (6 decimals)	0.000000
2.	Date of interview (CMC)	0.000000
3.	Age	0.000000
4.	Age group	0.000000
5.	Region	0.000000
6.	Type of place of residence	0.000000
7.	Highest education level	0.000000
8.	Source of drinking water	0.020295
9.	Type of toilet facility	0.069002
10.	Household has: electricity	0.028413
11.	Religion	0.064943
12.	Number of household members (listed)	0.000000
13.	Number of living children	0.000000
14.	Sex of household head:	0.000000
15.	Type of cooking fuel	0.000000
16.	Wealth index combined: Wealth index combined	0.000000
17.	Total children ever born	0.000000
18.	Births in last five years	0.000000
19.	Births in past year	0.000000
20.	Age at first birth	0.000000
21.	Number of living children	0.000000
22.	Current contraceptive method	0.000000
23.	Contraceptive use and intention	0.000000
24.	Wanted last child	0.020295
25.	Current breastfeeding	0.000000
26.	When child put to breast	1.286683
27.	Respondent slept under mosquito bed net	0.000000
28.	Does not use cigarettes and tobacco	0.004059
29.	Current marital status	0.000000
30.	Currently residing with husband/partner	16.118034
31.	Age at first cohabitation	7.768803
32.	Husband's desire for children	24.909689
33.	Husband/partner's educational attainment	10.687178
34.	Respondent currently working	0.056825

SN	Variables	Missing %
1.	Women's individual sample weight (6 decimals)	0.000000
35.	Birth order	0.000000
36.	Child is twin	0.000000
37.	Date of birth (CMC)	0.000000
38.	Sex of child	0.000000
39.	Survival status (Child is alive, 1 = Yes, 0 = No)	0.000000
40.	Age at death	95.575760
41.	Age at death (months, imputed)	95.575760
42.	Current age of child	4.424240
43.	Currently residing with husband/partner	4.424240
44.	Preceding birth interval (months)	26.748387
45.	Live birth between births	26.427731
46.	Duration of breastfeeding	0.227300
47.	Number of antenatal visits during pregnancy	26.147664
48.	Place of delivery (health facility, resp home, elsewhere)	0.069002
49.	Delivery by caesarean section	0.069002
50.	Size of child at birth	0.560133
51.	Baby postnatal check within 2 months	52.628161
52.	Birth to current child age	0.000000

[Table 2: Percentage of missing values per variables]

3.2. Machine learning algorithms for categorical variables

The 4 types of machine learning algorithms were used in this study to treat missingness from categorical variables. These are algorithms for classification like K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, and Random Forest. The Table 3 gives the summary of performance metrics for each machine learning algorithm to treat missing values from the categorical variables.

[Table 3: Machine learning algorithms for categorical data]

Table 3 Performance of machine learning algorithms for categorical data				
Performance metrics				
Algorithms	Accuracy	Precision	Recall	F1 Score
SVM	1.000	1.000	1.000	1.000
KNN	0.671	0.694	0.671	0.668
Random forest	0.780	0.992	0.780	0.992
Decision tree	0.800	1.000	0.800	0.889

3.3. Machine learning algorithms for numerical variables

The 4 machine learning algorithms used to treat missingness from numerical variables are Multivariate Imputation by Chained Equations (MICE), Random Forest, and Support Vector Machines (SVM) for regression. The Table 4 gives the performance for each machine learning model.

[Table 4: Machine learning algorithms for numerical data]

Table 4
Performance of machine learning algorithms for numerical data

Performance metrics			
Algorithm	Mean Squared Error (MSE)	Mean Absolute Error (MAE)	R2 Score
	best value = 0; worst value = $+\infty$	best value = 0; worst value = $+\infty$	[0–1]
Multivariate Imputation	1.53E + 09	3248.788	0.810034
Random Forest	0	0	-
Support Vector Machines	8.586746	677.0923	0.186535

3.4. The evaluation of machine learning algorithms

Based on performance metrics, the machine learning algorithm to treat missingness from categorical variables is the SVM classification with the accuracy of 100% and precision of 100%, and the Decision Tree with the accuracy of 79.9% and precision of 100%, and Random forest with 0.78 of accuracy and 0.992 precision while the one to treat missing values from numerical variables is Random Forest with MSE and MAE value of 0 and the Multivariate Imputation by Chained Equations (MICE) with MSE of 1.53. e^+09 and R^2 of 0.81. In addition, the machine learning algorithm to treat missingness in both numerical and categorical variables is Random Forest.

4. Discussion

In this study, it was revealed that the best algorithm model for imputing categorical variables is Support Vector Machine classification, Decision Tree and Random Forest while the one for numerical variables is Random Forest. It was also revealed that Random Forest is the most effective machine learning algorithm for handling both categorical and numerical missing values.

According to the research titled “Machine Learning Based Missing Values Imputation in Categorical Datasets” by Muhammad Ishaq and collaborators, machine learning algorithms demonstrated strong performance in predicting and imputing missing values. The effectiveness of these algorithms varied depending on the dataset and the pattern of missing values (13). The study evaluated a range of machine learning models, including ensemble models that employ the Error Correcting Output Codes (ECOC) architecture. These models included ensembles based on SVM and KNN, as well as an ensemble classifier that integrated both SVM and KNN. The results validate the efficaciousness of various machine learning techniques in handling incomplete data.

Another study by Dimitris Bertsimas titled "From Predictive Methods to Missing Data Imputation: An Optimization Approach" lends support to this, framing the traditional missing data problem as a non-convex optimization problem and utilizing a variety of prediction models. Using effective first-order optimization approaches, the research presents a new family of imputation algorithms dubbed Opt. Impute, which offers high-quality solutions to this challenge. To validate these techniques, numerous computational experiments were carried out (14).

The findings in the study are supported by others research in the same field. One of the documented on categorical variables revealed that the random forest was the effective machine learning algorithm to impute the missing values (13). The researches indicated that the performance metrics indicated that random forest is the best methods for predicting comprehensive strength with the numerical dataset. it is evident that machine learning can help to impute missing values.

The researcher suggests to use machine learning algorithms to impute missing values before proceeding with other data analysis steps for further future research to improve the prediction accuracy.

The study was limited to a slight number of machine learning algorithms and specific dataset related to data generated for responding to health conditions. The researcher would suggest to other researchers to explore other machine learning algorithms using the dataset from the domain of their interest.

5. Conclusions

The research concludes that the best algorithm to treat missing for categorical values is Support Vector Machine for classification, Decision Tree and Random Forest while the best to treat numerical values is Random Forest and the Multivariate Imputation by Chained Equations (MICE) algorithms. It is evidence that the Random Forest is the most effective machine learning algorithm for handling both categorical and numerical missing values.

Abbreviations

RDHS

Rwanda Demographic and Health Survey

SVM

Support Vector Machine

KNN

K-nearest neighbor

RF

Random Forest

MICE

Multiple Imputation by Chained Equations

MSE

Mean Absolute Error

MAE

Mean Square Error

UCI

University of California Irvine

GLM

Logistic Regression

DBMSs

Database Management systems

ML

Maximum Learning

ECOC

Correcting Output Codes

Declarations

6.2. Ethical and approval and consent to participate

I hereby declare that this research was conducted in full accordance with the University's established academic, ethical, and professional standards required for the completion of a Master's degree in Data Science. All methods, procedures, and analytical techniques employed in this study adhere to the principles of integrity, transparency, and responsible research conduct.

The datasets used in this study were ethically cleared and accessed from the DHS Program (dhsprogram.com) under authorized use. No attempt was made to identify individual respondents, and all data handling complied with applicable data protection regulations and best practices for confidentiality, privacy, and responsible data stewardship.

This manuscript represents my original work and has not been submitted, in whole or in part, toward any other journal. All sources of information, data, and prior research have been properly acknowledged through accurate citation and referencing. Any form of

plagiarism, data falsification, data manipulation, or unauthorized reuse of materials has been strictly avoided.

No additional institutional ethical approval was required for this study. All analyses were conducted under academic supervision and in alignment with recognized standards of fairness, accountability, and transparency in Data Science research.

I affirm that this study upholds the principles of academic integrity, respect for data subjects, and responsible knowledge production, as expected within a Master's degree program in Data Science.

6.3. Consent for publication

Not applicable

6.4. Availability of data and materials

The data used in this study is publicly available at The DHS Program. The specific dataset analyzed in this study can be accessed directly through the DHS dataset portal. The code for data analysis are available on: google colab.

6.5. Competing interest

The researcher declares that there are no competing interests associated with this study.

6.6. Funding Declaration

The researcher declares that this study was conducted without any financial support.

6.7. Authors' contributions

Mr. HABINSHUTI Cyprien accessed the study data and led the proposal development. He contributed to the data collection strategy, methodological design, data cleaning, data analysis, and interpretation of the results, and drafted the manuscript.

Email address: nshuticyprien@gmail.com

Prof. François NIRAGIRE contributed to the methodological development and conceptualization of the study and reviewed and edited the manuscript.

Email address: fniragire@gmail.com

6.8. Acknowledgements

The author sincerely acknowledges the administrative support of the University of Rwanda and expresses appreciation to all staff members of the African Center of Excellence in Data Science (AC-DS) for their continuous guidance and support during this study and the Master's program in Data Science.

References

1. Sivakani R, Ansari GA (2020) Imputation Using Machine Learning Techniques. 4th International Conference on Computer, Communication and Signal Processing, ICCSP 2020. ;0–5
2. Setiawan I, Gernowo R, Warsito B (2023) A Systematic Literature Review on Missing Values: Research Trends, Datasets, Methods and Frameworks. E3S Web of Conferences. ;448
3. Dong Y, Peng CYJ (2013) Principled missing data methods for researchers (Expectation Maximization explained). Springerplus 2(1):1–17
4. Platias C, Petasis G (2020) A comparison of machine learning methods for data imputation. ACM International Conference Proceeding Series. ;150–9
5. Nadzurah ZA, Amelia Ritahani I, Nurul A (2018) Performance Analysis of Machine Learning Algorithms for Missing Value Imputation. Int J Adv Comput Sci Appl. ;9(6)

6. Ismail AR, Abidin NZ, Maen MK (2022) Systematic Review on Missing Data Imputation Techniques with Machine Learning Algorithms for Healthcare. *J Rob Control (JRC)* 3(2):143–152
7. Jäger S, Allhorn A, Bießmann F (2021) A Benchmark for Data Imputation Methods. *Front Big Data* 4(July):1–16
8. Cappelletti L, Fontana T, Di Donato GW, Di Tucci L, Casiraghi E, Valentini G (2020) Complex data imputation by auto-encoders and convolutional neural networks—A case study on genome gap-filling. *Computers*. ;9(2)
9. Oluwaseye L, J* A, Doorsamy W, Paul S (2022) A Review of Missing Data Handling Techniques for Machine Learning. *International Journal of Innovative Technology and Interdisciplinary Sciences wwwIJITIS.org [Internet]*. ;5(3):971–1005. Available from: <https://doi.org/10.15157/IJITIS.2022.5.3.971-1005>
10. Chhabra G, Vashisht V, Ranjan J (2018) A classifier ensemble machine learning approach to improve efficiency for missing value imputation. 2018 International Conference on Computing, Power and Communication Technologies, GUCON. 2019;23–7
11. Emmanuel T, Maupong T, Mpoeleng D, Semong T, Mphago B (2021) A survey on missing data in machine learning [Internet]. *Journal of Big Data*. Springer International Publishing; Available from: <https://doi.org/10.1186/s40537-021-00516-9>
12. Aljuaid T, Sasi S (2016) Proper imputation techniques for missing values in data sets. *Proceedings of the 2016 International Conference on Data Science and Engineering, ICDSE*. 2017
13. Ishaq M, Zahir S, Iftikhar L, Bulbul MF, Rho S, Lee MY (2024) Machine Learning Based missing data Imputation in Categorical Datasets. *IEEE Access*. ;1–29
14. Bertsimas D, Pawlowski C, Zhuo YD (2018) From predictive methods to missing data imputation: An optimization approach. *J Mach Learn Res* 18:1–39

Footnotes

1. https://dhsprogram.com/data/dataset_admin/login_main.cfm;jsessionid=A8E4992909AED729EABAB7C9D5579E85.cfusion?CFID=306774449&CFTOKEN=c153a0a8d77f8b19-F7DAA648-D7AD-AB05-AAB740663C775C2B