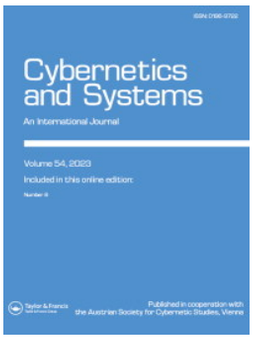
Analysis of Machine Learning Based Imputation of Missing Data



Taylor & Francis
Taylor & Francis Group

Cybernetics and Systems
An International Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ucbs20

Analysis of Machine Learning Based Imputation of Missing Data

Syed Tahir Hussain Rizvi, Muhammad Yasir Latif, Muhammad Saad Amin, Achraf Jabeur Telmoudi & Nasir Ali Shah

To cite this article: Syed Tahir Hussain Rizvi, Muhammad Yasir Latif, Muhammad Saad Amin, Achraf Jabeur Telmoudi & Nasir Ali Shah (2023): Analysis of Machine Learning Based Imputation of Missing Data, Cybernetics and Systems, DOI: 10.1080/01969722.2023.2247257

To link to this article: https://doi.org/10.1080/01969722.2023.2247257

Published online: 09 Sep 2023.

CYBERNETICS AND SYSTEMS: AN INTERNATIONAL JOURNAL
https://doi.org/10.1080/01969722.2023.2247257

Analysis of Machine Learning Based Imputation of Missing Data

Syed Tahir Hussain Rizvi[a], Muhammad Yasir Latif[b], Muhammad Saad Amin[c], Achraf Jabeur Telmoudi[d], and Nasir Ali Shah[e]

[a]Department of Electrical Engineering and Computer Science, University of Stavanger, Stavanger, Norway; [b]Educative, Inc., Islamabad, Pakistan; [c]Dipartimento di Informatica, Universita Degli Studi di Torino, Torino, Italy; [d]LISIER Laboratory, The National Higher Engineering School of Tunis (ENSIT), University of Tunis, Tunis, Tunisia; [e]Dipartimento Di Elettronica E Telecomunicazioni, Politecnico di Torino, Torino, Italy

ABSTRACT

Data analysis and classification can be affected by the availability of missing data in datasets. To deal with missing data, either deletion- or imputation-based methods are used that result in the reduction of data records or imputation of incorrect predicted value. Quality of imputed data can be significantly improved if missing values are generated accurately using machine learning algorithms. In this work, an analysis of machine learning-based algorithms for missing data imputation is performed. The K-nearest neighbors (KNN) and Sequential KNN (SKNN) algorithms are used to impute missing values in datasets using machine learning. Missing values handled using a statistical deletion approach (List-wise Deletion (LD)) and ML-based imputation methods (KNN and SKNN) are then tested and compared using different ML classifiers (Support Vector Machine and Decision Tree) to evaluate the effectiveness of imputed data. The used algorithms are compared in terms of accuracy, and results yielded that the ML-based imputation method (SKNN) outperforms the LD-based approach and KNN method in terms of the effectiveness of handling missing data in almost every dataset with both classification algorithms (SVM and DT).

KEYWORDS

Imputation; imputation using KNN; imputation using SKNN; missing data; statistical imputation

Introduction

In the Modern era where data has much importance and is being analyzed at a broader level for useful purposes, the missing patterns of data can affect the results (Graham 2009). The most famous datasets like Modified National Institute of Standards and Technology (MNIST) and ImageNet are complete, clean, and perfect but most real-time datasets are far from this perfection as they have missing values in them. Data whether it's a voice signal or data related

CONTACT Syed Tahir Hussain Rizvi tahir.rizvi@uis.no Department of Electrical Engineering and Computer Science, University of Stavanger, Stavanger, Norway.

to energy consumptions collected to predict future load, it's image data or other datasets required for research purposes must need to be error/corruption-free and complete. The missing values present in data make it incomplete and can produce less accurate results when applied to real-time applications.

The resources being used for data collection can corrupt data and can produce errors. Reasons like equipment error, incorrect measurements, manually entered data, refusal of participants to answer and data traveling over a long distance can cause the data to be missed from the data source. Missing data is mostly termed as nan/null due to the above-defined reasons. Data missed in any type of data (video, audio, image, signal, etc.) creates a problem in evaluation and analysis.

There have been some studies on handling missing data specifically in the medical field where life depends on predictions and decisions made on collected data (Little et al. 2012; O'Neill and Temple 2012). The best solution to any problem is to avoid the problem, and in terms of missing data it can be avoided by collecting data properly and perfectly (DeSarbo, Green, and Carroll 1986; Wisniewski et al. 2006). Data becomes of much importance in a medical study, and it can be collected cleanly by practicing and instructing all the persons involved in the study, i.e., persons collecting data and persons from whom data is being collected (Wilcox et al. 2001). If data is collected by equipment, the quality of equipment must be good and tested properly. The quality of data has no exact method to get measured but it is generally said the data is of good quality if the defined goal is achieved perfectly in a given context from available data. Completeness, accuracy, coherence, and uniqueness are some attributes that define the quality of data (Juddoo 2015). Now, if the data has already been collected and quality was not an important parameter at the time of data collection, then its quality needs to be enhanced through pre-processing. Besides removing noise and cleaning data through other parameters, the one most important step is handling missing values in the dataset. Missing data is either handled by deletion (Donner 1982; Kim and Curry 1977; Orrawan et al. 2008) of missing records or imputation (Barros et al. 2012; Dempster, Laird, and RubinRubin 1977; Malhotra, 1987; Osman, Abu-Mahfouz, and Page 2018; Quinlan 1986; Wei and Tang, 2003) of missing values.

The deletion-based method tackles this issue by deleting records where data is missing in the dataset. Though the problem of missing data is eradicated, the reduction in data records is another problem. This does not produce a problem when the dataset is very large as the deletion of a few numbers of records does not affect results. But if the dataset is very small (i.e., the IRIS plant dataset which only has 150 records), the reduction in dataset may affect results on larger scales. So, these types of techniques are not suitable because the amount of data is also an important parameter for good analysis or training of machine learning algorithms.

Imputation-based methods are another and better option to solve this problem. These methods impute missing values in the dataset to make it complete and accurate. The statistical imputation methods are old-fashioned, and values imputed/generated by statistical approaches can be wrong. For example, mean value imputation can impute the wrong value at the missing place when a feature or variable has many missing values.

In this article, statistical (List-wise deletion) and machine learning-based imputation methods (K-nearest neighbors [KNN] and Sequential K-nearest neighbors [SKNN]) are analyzed for the task of handling missing data using deletion and imputation, respectively. The used algorithms are applied on different datasets of different domains of life (i.e., social life, medical, and general datasets of objects) by manually deleting random data. Deletion using statistical List-wise deletion (LD), and imputation using ML-based (KNN and SKNN) algorithms yield different completed datasets for each case. The missing values are imputed, and results are compared based on accuracy. The comparison shows that the SKNN is much more time efficient than simple KNN and yields more accurate results in the case of classification.

Types of Datasets, Missing Values, and Imputation Methods

Datasets can be of different types having different structures. It may contain univariate and multivariate values (having one or several features, respectively), periodic or non-periodic data (time series signals), differently structured data (visual or audio data), etc.

Types of Signals

A real-time dataset can have different types of signals, like audio, image, or video.

Audio Signals

When it comes to audio signals which consist of time-based data, the patches corrupted by any reason like noise, device malfunctions, or something else must be imputed to have the maximum level of understanding. Audio signals have importance in speech recognition, so, the missing audio signal would produce inaccurate results.

Images, Videos, and Other Signals

Similar problems can occur in image and video processing where a partial part of an image like pixels can get corrupted due to any reason or video can lose some of its data either sound or frame. Similarly, the signals traveling on longer distances can be corrupted due to the failure of devices or noise and a chunk of data can be missed. This gap in data is undesirable and it is necessary to reconstruct the original signal and fill the gap.

Categories of Missing Data

Data can be missed in different ways, and it is important to understand the characteristics of the missing pattern before imputation. There are three types of missing data as defined by Little and Rubin (Little and Rubin 2019): missing at random (MAR), missing not at random (MNAR), and missing completely at random (MCAR). Structurally or logically missing data is also a type of missing mechanism that can occur in many cases.

Structurally Missing Data

The data in a dataset can be missed due to logical reasons. In simple words, it can be explained or there is a logical reason why the data is missing and cannot be filled in. For example, a participant cannot answer whether he/she has a son or daughter if he/she has no child. Also, it is not possible to answer the age of the youngest child in the same scenario. Table 1 lists the data which have logically missing values. The shaded columns represent logically missed data.

Missing Completely at Random (MCAR) Data

Missing completely at random is a mechanism where values have a fully independent relationship with other variables of the same dataset. An example is given as; information of a customer, i.e., emails or contact numbers missing from the dataset is independent of all other information available in the dataset. This type of missing data is difficult to impute as have no relationship with others and it would be difficult to find out the missing value as well. A formal way of testing data if its MCAR or not is by doing an MCAR test (Little 1988). Table 2 lists examples of data that is MCAR.

Missing at Random (MAR) Data

Missing at random (Little and Rubin 2019) is a pattern where missing values are fully dependent on observed values available in the dataset.

Table 1. Structurally or logically missing data.

ID	Name	No. of childs	Number of son(s)	Number of daughter(s)	Age of youngest child
1	John	2	1	1	4
2	White	0			
3	David	3	1	2	6
4	Sana	2		2	3
5	Akram	3		3	5

Table 2. Data missing completely at random.

ID	Name	age	Address	Email	Phone No.
1	John	35	Street 4, New York, America	John123@abc.com	+190909789234
2	Adam	50	Walton road, California		+3980988787
3	David	28		David_mook@abc.com	

Table 3. Data missing at random.

ID	Name	Age	Salary (USD)	Owns a car
1	John	35	100 k	Yes
2	White	50	90 k	
3	David	28	20 k	No
4	Sana	40	110 k	Yes
5	Adam	20		

Table 4. Data missing not at random.

ID	Name	IQ score
1	John	135
2	White	
3	David	
4	Sana	140

The probability of missing values depends only on observed values, not on other missing variables of the same dataset. An example is given, if the salary of a person working in a certain company is missing then it can be estimated on the qualification, experience, and skills of the person. It means that the data MAR can be predicted based on observed values. Table 3 presents examples that contain MAR data. It is observed that the persons having age greater than 30 have a handsome salary of around 100 k and they own a car as well. So, the missing data can either be imputed by analyzing the available data or can be imputed by using high-level imputation method.

Missing Not at Random (MNAR) Data

Missing not at random is a pattern when missing values have a direct dependent relationship with the nature of the variable used for data recording. Table 4 represents data MNAR in which IQ scores of those students are missing who obtained less than 100 out of 150. Now, these marks cannot be predicted by simply analyzing data as students can have any number of marks between 0 and 100.

MAR- and MCAR-based data can be ignored instead of imputing as it does not affect significantly, whereas data that is missed with MNAR pattern cannot be ignored as it has a significant effect on results or classifications.

Imputation Techniques/Algorithms

Due to the problem of missing data and its effects on modeling of prediction systems, different techniques have been proposed by researchers to overcome the issue of missing data. There are mainly two types of imputation techniques, the statistical approaches and the machine learning-based approaches for handling missing data.

Statistical Methods

Statistical methods of missing data are categorized into further two categories, mostly known as deletion and imputation. Few of the techniques delete missing data and few techniques impute missing values with different mathematical operations. List-wise deletion is the most traditional method where a row of missing data is deleted or a record is omitted and the remaining data is analyzed (Donner 1982; Orrawan et al. 2008). This mechanism works well where a smaller number of records are missing or where missing data follows the mechanism of MCAR. Otherwise, it will result in the loss of more critical information and reduce the size of the dataset. There is another method named Pairwise deletion (Kim and Curry 1977) that also deletes data but preserves more values than LD and is biased for both MAR and MCAR. Mean or median substitution is a statistical method of imputation where the missing value of a variable is substituted by taking the mean of the available values of that variable (Malhotra 1987). When the number of records that are missing is less, then this technique is helpful in recovering data. There are few other statistical techniques like maximum likelihood and expectation maximization (EM) (Dempster, Laird, and Rubin 1977) that work on the basis of the likelihood of the value's occurrence.

Machine Learning Based Approaches

Missing values must be imputed automatically without deleting the records to have complete datasets. Machine learning methods are best to impute missing data to improve accuracy. There are a few machine-learning approaches that can be used for the imputation of missing data. KNN is a supervised machine learning algorithm that uses its neighbors to impute missing values. Using this algorithm, most likely value of missing data can be calculated using the Euclidean distance of nearest neighbors (Orrawan et al. 2008; Osman, Abu-Mahfouz, and Page 2018). The neural networks (NNs) can also be used for imputation that is usually based on the interconnection of multiple artificial neurons which create a complex structure between input and output to find or predict required data (Wei and Tang, 2003).

Proposed Methodology

Figure 1 presents the complete flow to analyze the performance of machine learning-based imputed data. In this analysis, the data is obtained from well-known repositories having different publicly available datasets. Then the data is checked if it already contains missing values or not as there are many datasets that come with missing values.

If the dataset does not contain any missing values, they are removed manually by deleting some records from the dataset for experimental purposes. Once an incomplete dataset is available, it is then fed to the system

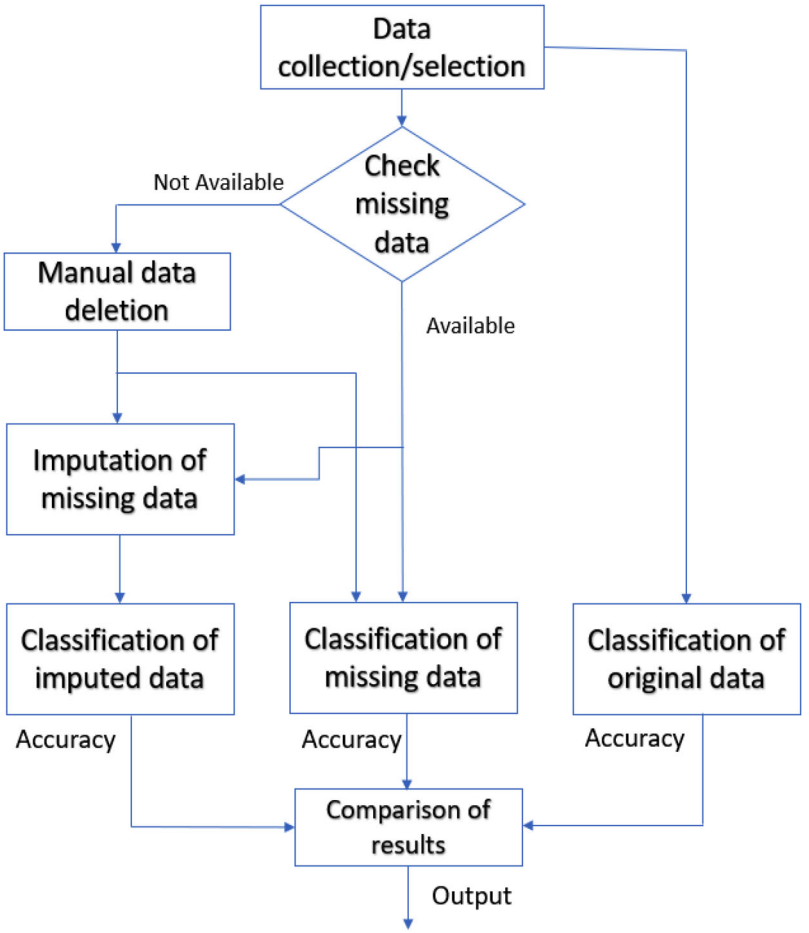


Figure 1. Proposed flow for analyzing machine learning based data imputation.

where the imputation of missing data is done. The statistical and ML-based algorithms are applied to handle missing values in the dataset. After this stage, there are different versions of datasets available, i.e., original datasets, datasets with missing values, datasets with statistical approach-based deleted values, and datasets with ML-based imputed values. All these datasets are then fed to classification algorithms one by one and classification is performed. The accuracy of classification algorithms is the main parameter for comparison of results of the used techniques. If the classification results on the new datasets are better, then it indicates or proves the effectiveness or better performance of used algorithms in handling missing data.

Experimental Setup and Implementation

The implementation and analysis are completely done using MATLAB software. Different datasets are used to analyze the performance of imputation techniques using the proposed flow.

Classification Learner App of MATLAB

The classification learner app is used to perform these experiments and analyses.

The classification learner app is a built-in tool/application of MATLAB which is normally used for classification purposes. It provides user-friendly access to the selection of specific datasets, required predictors, response variables, and validation methods (cross-validation, hold-out validation, and no validation). After classification, it provides different visualization tools like scatter plots, confusion matrices, ROC curves, and parallel coordinates plots.

In this analysis, this classification learner app is used for the imputation of missing values by training KNN and SKNN on the missing datasets. This classification/imputation provides a dataset imputed using a selected ML-based approach. This new imputed dataset is then used to train the classifiers (Support Vector Machine and Decision Tree) to evaluate the effectiveness of ML-based imputed data. List-wise deletion is also used to remove missing data from the dataset, so this statistical method can be compared with ML-based imputation methods.

Used Datasets

The datasets that are used to evaluate the proposed flow are IRIS, WISCONSIN, and Human activity recognition (HAR) using smartphones. These datasets are downloaded from the UCI repository. IRIS dataset contains data related to IRIS plants according to three different classes named: Iris Setosa, Iris Versicolor, and Iris Virginica defined by the length and width of petals and sepals. It consists of 150 records having 4 attributed only. The WISCONSIN is a well-known breast cancer diagnosis dataset of different patients collected from different hospitals in the 1990s. It consists of 32 different features out of them 1 is for classification containing two classes of cancer named: Benign and Malignant. There are a total of 569 records available in this dataset which are real-time. The HAR is one of the recently collected datasets that record human activities using mobile sensors. It contains data collected using two sensors (accelerometer and gyroscope) installed in the smartphone. This dataset consists of 561 different attributes having time and frequency domain parameters. Thirty volunteers (between the age group of 19 to 48 years) participated in collecting this dataset.

These datasets do not contain any missing values originally. These datasets are processed in such a way that the missing values are produced by deleting some random records manually and saving these new reduced data entries as a separate dataset. Every dataset is processed in such a way that the values are missed by deleting some records randomly/manually and saved as a

separate dataset. Every dataset is then fed to two different ML-based algorithms (KNN and SKNN) programmed/trained in MATLAB. The manually missed values for evaluation are imputed by using KNN and SKNN algorithms and are saved as different versions of dataset. List-wise statistical deletion approach is also used to tackle missing data. After completion of these steps, five different versions of a single dataset are obtained (original dataset, the dataset having missing values, the dataset having values computed using LD-based approach, the dataset with imputed values using KNN and dataset with imputed values using SKNN). These datasets are then used for the evaluation of ML-based imputed values by 2 different classifiers (SVM and DT).

Results

The results of these experiments are analyzed and compared using an accuracy and confusion matrix. Accuracy is given in the tabular form which shows the results in percentage.

Results of IRIS Dataset

The different copies of the IRIS dataset are used as mentioned earlier (original dataset, the dataset which contains missing values, Dataset having LD-based handled missing data, datasets containing/imputed data using KNN, and datasets containing complete/imputed data using SKNN). On each dataset, both classifiers (SVM and DT) are trained, and result in the form of accuracy, and the confusion matrix is obtained and analyzed. The accuracy is listed in Table 5.

The confusion matrix of the IRIS dataset for both classifiers (SVM and DT) on four different datasets can be visualized in Figures 2–9. There are three classes in the IRIS dataset named IRIS-Setosa, IRIS-Versicolor, and IRIS-Virginica which are represented in the confusion matrix graph. There are numbers from 1 to 3 instead of class name on graphs which represent class names (IRIS-Setosa, IRIS-Versicolor, and IRIS-Virginica), respectively.

Results of WISCONSIN Dataset

The different copies of the WISCONSIN dataset are used (original dataset, the dataset which contains missing values, the dataset having LD-based handled missing data, datasets containing/imputed data using KNN, and

Table 5. Accuracy of classification algorithms on IRIS dataset.

	SVM	DT
Original dataset	96.0	94.0
Missing dataset	89.3	89.3
LD-based dataset	90.9	91.4
Imputed-KNN dataset	94.0	95.3
Imputed-SKNN dataset	94.7	96.7

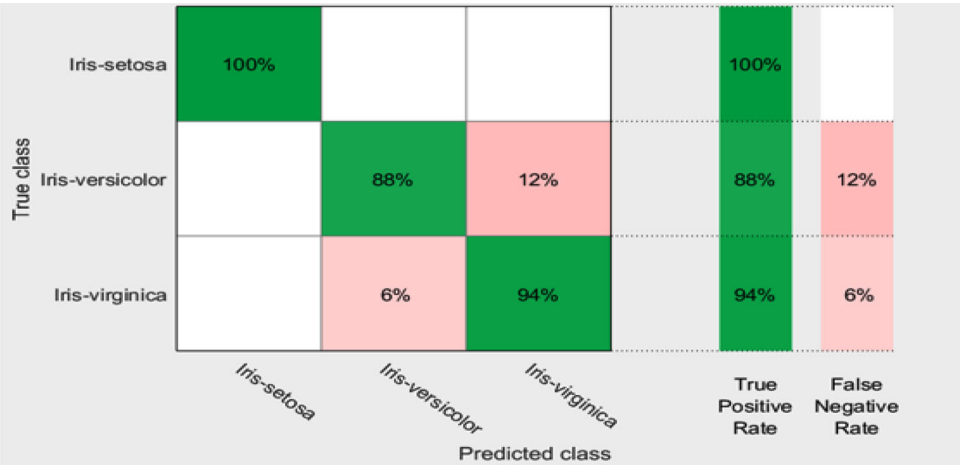


Figure 2. DT classification in terms of confusion matrix of original IRIS dataset.

Figure 3. DT classification in terms of confusion matrix of missing IRIS dataset.

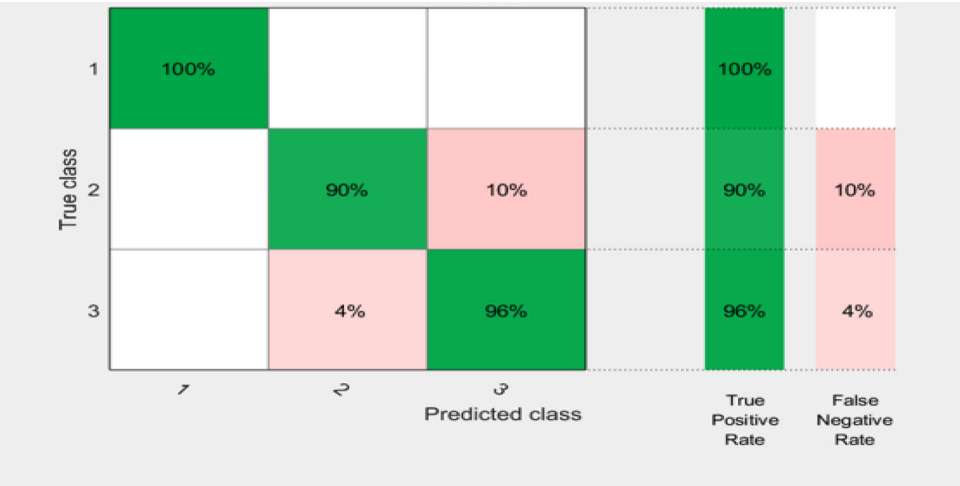


Figure 4. DT classification in terms of confusion matrix of imputed IRIS dataset with KNN.

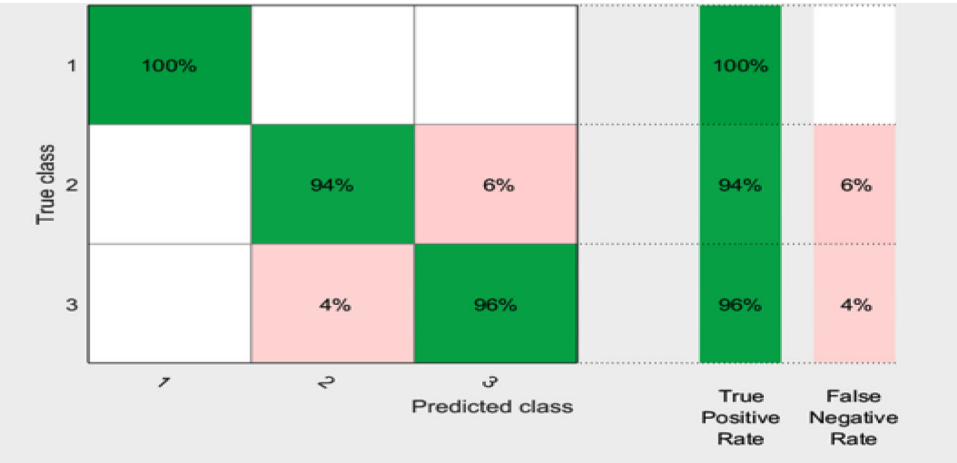


Figure 5. DT classification in terms of confusion matrix of imputed IRIS dataset with SKNN.

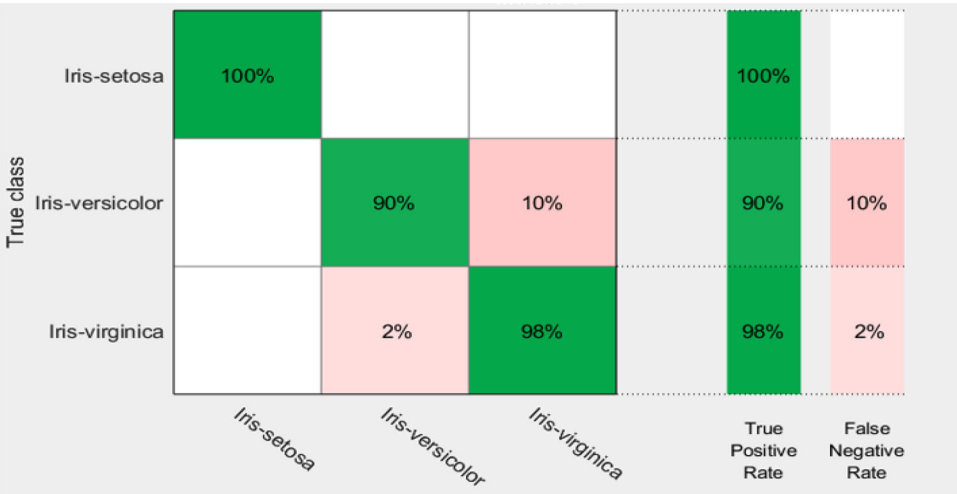


Figure 6. SVM classification in terms of confusion matrix of original IRIS dataset.

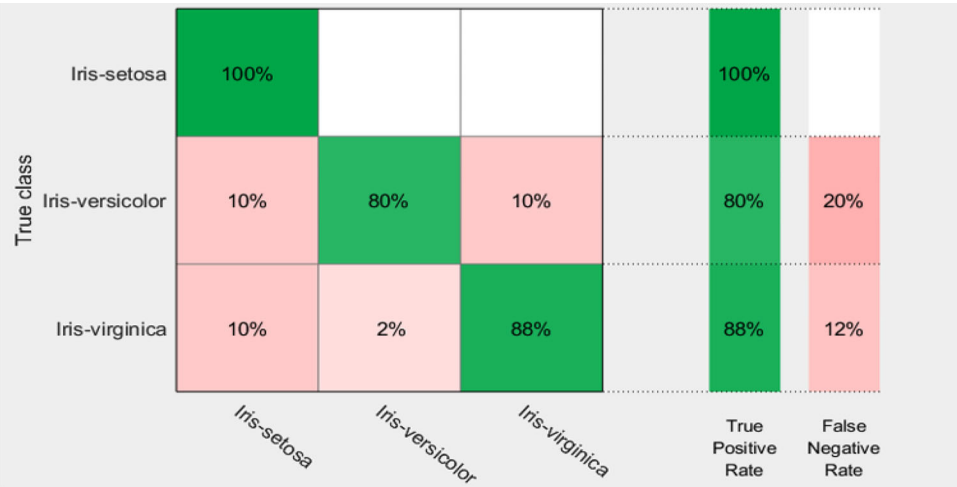


Figure 7. SVM classification in terms of confusion matrix of missing IRIS dataset.

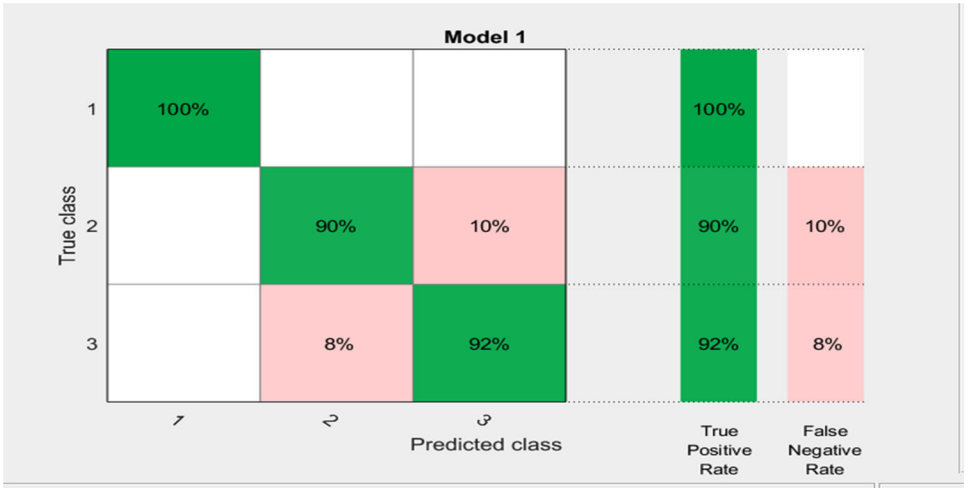


Figure 8. SVM classification in terms of confusion matrix of imputed IRIS dataset with KNN.

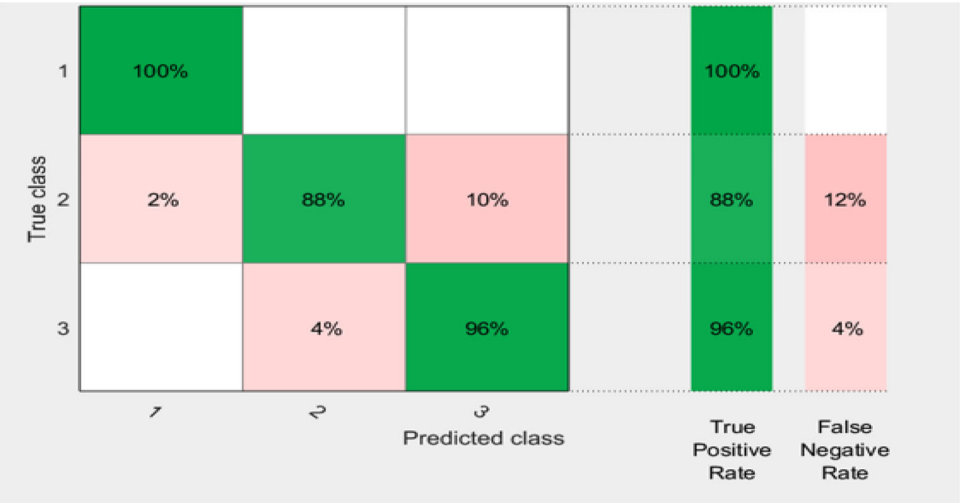


Figure 9. SVM classification in terms of confusion matrix of imputed IRIS dataset with SKNN.

datasets containing complete/imputed data using SKNN). On each dataset, both classifiers (SVM and DT) are trained as done in the previous experiment, and results in the form of accuracy are obtained and analyzed. The accuracy is listed in Table 6.

Results of HAR Dataset

Like previous experiments, different copies of the HAR dataset are used (original dataset, the dataset which contains missing values, Dataset having LD-based handled missing data, datasets containing/imputed data using KNN, and datasets containing complete/imputed data using SKNN). On

Table 6. Accuracy of classification algorithms on WISCONSIN dataset.

	SVM	DT
Original dataset	97.7	91.9
Missing dataset	91.4	92.4
LD-based dataset	91.8	89.3
Imputed-KNN dataset	97.0	91.5
Imputed-SKNN dataset	97.0	92.3

Table 7. Accuracy of classification algorithms on HAR dataset.

	SVM	DT
Original dataset	63.8	74.2
Missing dataset	62.7	73.5
LD-based dataset	63.0	73.6
Imputed-KNN dataset	63.6	74.0
Imputed-SKNN dataset	63.7	74.0

each dataset, both classifiers (SVM and DT) are trained as done previously and results in the form of accuracy are obtained and analyzed. The accuracy is listed in Table 7.

Discussion on Results

Implementation results for the IRIS dataset show that for the original datasets, the classification accuracy of SVM is better than DT. For dataset containing manual missing data, both classification algorithms yielded the same accuracy. After that, statistical as well as machine learning-based missing data-solving techniques are used. Implementation of the LD-based approach yielded that the accuracy is decreased with a reduction in the length of data. After that, the implementation of the KNN imputation method yielded significantly improved results than data with missing values. The decision tree produced more accurate results out of both algorithms for KNN-based imputed data. Thirdly, SKNN-based imputed data was classified using the same procedure with both algorithms. The results yielded by this data are even more accurate than simple KNN-based imputed data. Moreover, DT again achieved more classification accuracy than the SVM classifier. For the IRIS dataset, it can be concluded that SKNN is the best imputation technique and DT is the best classification algorithm.

WISCONSIN is one of the most concerned datasets as it is used for breast cancer diagnosis in patients. Implementation results for the Wisconsin dataset show that for the original dataset, the classification accuracy of SVM is proved better than DT. For datasets containing manual missing data, both classification algorithms yielded different classification accuracy. After that, statistical as well as machine learning-based missing data-solving techniques are used. Implementation of LD yielded that the

accuracy is decreased with a reduction in the length of data. After that, the implementation of the KNN-based imputation method yielded significantly improved results than data with missing values. The SVM produced more accurate results out of both algorithms for KNN-based as well as SKNN-based imputed data. Thirdly, SKNN-based imputed data was classified using the same procedure with both classifiers. The results yielded by this data are even more accurate than simple KNN-based imputed data. Moreover, SVM obtained better classification accuracy than DT. For the WISCONSIN dataset, it can be concluded that, both KNN and SKNN are the best imputation techniques than the traditional and SVM is best classification algorithm.

Implementation results for the HAR dataset also show the SKNN is the best imputation technique than others.

Conclusion

In this article, a flow is proposed to evaluate machine learning-based imputation techniques and analyze the effectiveness of using ML-based approaches and their effect on classification accuracy. Machine learning-based techniques named KNN and SKNN were applied to solve missing data problems with the help of imputation. Classification results of statistical and ML-based approaches to handle missing data were compared on three datasets from the UCI repository. Results showed that the Like-wise statistical technique reduces both the data and classification accuracy while ML-based imputation techniques (KNN and SKNN) show a definite increase in classification accuracy.

However, this model is only limited to numeric data, future work can be on the imputation of strings or other types of data.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Barros, R. C., M. P. Basgalupp, A. C. P. L. F. D. Carvalho, and A. A. Freitas. 2012. A survey of evolutionary algorithms for decision-tree induction. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42 (3):291–312. doi: 10.1109/TSMCC.2011.2157494.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximal likelihood form incomplete data via the em algorithm. *Journal of the Royal Statistical Society* 39:1–38. doi: 10.1111/j.2517-6161.1977.tb01600.x.

DeSarbo, S., P. E. Green, and J. D. Carroll. 1986. An alternating least-squares procedure for estimating missing preference data in product-concept testing. *Decision Sciences* 17 (2): 163–85. doi: 10.1111/j.1540-5915.1986.tb00219.x.

Donner, A. 1982. The relative effectiveness of procedures commonly used in multiple regression analysis for dealing with missing values. *The American Statistician* 36 (4):378–81. doi: 10.2307/2683092.

Graham, J. W. 2009. Missing data analysis: Making it work in the real world. *Annual Review of Psychology* 60:549–76. doi: 10.1146/annurev.psych.58.110405.085530.

Juddoo, S. 2015. Overview of data quality challenges in the context of Big Data. Paper presented at the 2015 International Conference on Computing, Communication and Security (ICCCS), 1–9. 4–5 Dec 2015, Mauritius. doi: 10.1109/CCCS.2015.7374131.

Kim, J. O., and J. Curry. 1977. The treatment of missing data in multivariate analysis. *Sociological Methods & Research* 6 (2):215–40. doi: 10.1177/004912417700600206.

Little, R. J. A. 1988. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association* 83 (404):1198–202. doi: 10.1080/01621459.1988.10478722.

Little, R. J., and D. B. Rubin. 2019. *Statistical analysis with missing data*. Hoboken, NJ: John Wiley & Sons.

Little, R. J., R. D'Agostino, M. L. Cohen, K. Dickersin, S. S. Emerson, J. T. Farrar, C. Frangakis, J. W. Hogan, G. Molenberghs, S. A. Murphy, et al. 2012. The prevention and treatment of missing data in clinical trials. *The New England Journal of Medicine* 367 (14):1355–60. doi: 10.1056/NEJMsr1203730.

Malhotra, N. K. 1987. Analyzing marketing research data with incomplete information on the dependent variable. *Journal of Marketing Research* 24 (1):74–84. doi: 10.2307/3151755.

O'Neill, R., and R. Temple. 2012. The prevention and treatment of missing data in clinical trials: An FDA perspective on the importance of dealing with it. *Clinical Pharmacology and Therapeutics* 91 (3):550–4. doi: 10.1038/clpt.2011.340.

Orrawan, K., R. Panrasee, B. Thongchai, and C. Wichit. 2008. Dealing with missing values for effective prediction of NPC recurrence. Paper presented at the 2008 SICE Annual Conference, 1290–1294. 20–22 Aug 2008. doi: 10.1109/SICE.2008.4654856.

Osman, M., A. Abu-Mahfouz, and P. Page. 2018. A survey on data imputation techniques: Water distribution system as a use case. *IEEE Access* 6:63279–91. doi: 10.1109/ACCESS.2018.2877269.

Quinlan, J. R. 1986. Induction of decision trees. *Machine Learning* 1 (1):81–106. doi: 10.1007/BF00116251.

Wei, W. and Y. Tang. 2003. A generic neural network approach for filling missing data in data mining, Paper presented at the SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme - System Security and Assurance (Cat. No.03CH37483), 862–867. 8–8 Oct 2003. doi: 10.1109/ICSMC.2003.1243923.

Wilcox, S., S. A. Shumaker, D. J. Bowen, M. J. Naughton, M. C. Rosal, S. E. Ludlam, E. Dugan, J. R. Hunt, and S. Stevens. 2001. Promoting adherence and retention to clinical trials in special populations: A women's health initiative workshop. *Controlled Clinical Trials* 22 (3):279–89. doi: 10.1016/s0197-2456(00)00130-6.

Wisniewski, S. R., A. C. Leon, M. W. Otto, and M. H. Trivedi. 2006. Prevention of missing data in clinical research studies. *Biological Psychiatry* 59 (11):997–1000. doi: 10.1016/j.biopsych.2006.01.017.