Analysis of Machine Learning Based Imputation of Missing Data

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], Nasir Ali Shah[e]

[a] *Department of Electrical Engineering and Computer Science, University of Stavanger, Norway, Italy;* [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, Tunisia;* [e]*Dipartimento Di Elettronica E Telecomunicazioni, Politecnico di Torino, Torino, Italy.*

*Corresponding Author: tahir.rizvi@uis.no

Abstract

Data analysis and classification can be affected by the availability of missing data in datasets. To deal with missing data, either deletion-based or imputation-based methods are used that results in the reduction of data records or wrong predicted value imputed by means/median respectively. A significant improvement can be done if missing values are imputed more accurately with less computation cost. In this work, a flow for analysis of machine learning-based algorithms for missing data imputation is proposed. 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 statistical deletion approach (List-wise Deletion) and ML-based imputation methods (KNN and SKNN) is then tested and compared using different ML classifiers (Support Vector Machine and Decision Tree) to evaluate effectiveness of imputed data. The used algorithms are compared in terms of accuracy, and results yielded that the ML-based imputation method (SKNN) outperforms LD-based approach and KNN method in terms of effectiveness of handling missing data in almost every dataset with both classification algorithms (SVM and DT).

Keywords: Missing Data, Imputation, Statistical Imputation, Imputation using KNN, Imputation using SKNN.

Introduction

In the Modern era where data have much importance and are being analyzed at a broader level for useful purposes, the missing patterns of data can affect the results [1]. The most famous datasets like MNIST (modified national institute of standards and technology) and ImageNet are complete, clean and perfect but the most real-time datasets are far from this perfection as they have missing values in it. Data either its voice signal, data related to energy consumptions collected to predict future load, image data or datasets being used 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. The equipment can be damaged (equipment errors), incorrect measurements, data entered manually, participants can refuse to answer and data traveling over a long distance can get weaken and may lose some information. These are all the reasons which can cause the data to be missed from the data source. Missing data is defined as the value of one or more variables is missed and mostly termed as nan/null due to 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 based on data [2, 3]. 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 [4, 5]. 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 [6]. 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 [7]. Now, if data is collected and quality was not an important parameter at time of data collection, then its quality is 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 [8-10] of missing records or imputation [11-16] 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, but the reduction in data records is another problem. This doesn't produce a problem when the dataset is very large as the deletion of few numbers of records doesn't affect results. But if the dataset is very small i.e., IRIS plant dataset which only has 150 records, the reduction in the dataset may affect results on larger scales. So, these types of techniques are not suitable as 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 missing place when a feature or variable has many missing values.

In this paper, 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 Listwise 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 case of classification.

Types of Datasets, Missing Values and Imputation Methods

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

Types of Signals

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

Audio Signals

When it comes to audio signals which consist of time-based data, the patches corrupted by any reason like noise, device's 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 images 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 would be required to reconstruct or fill the gap of the original signal as it plays a much important role in every type of data.

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 the missing data as defined by Rubin [17]: MAR (Missing at Random), MNAR (Missing Not at Random) and MCAR (Missing Completely at Random). Structurally or logically missing data is also a type of missing mechanism which 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 that why the data is missing and it cannot be filled. For example, a participant can't answer either 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 value. The shaded columns represent logically missed data.

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

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 number 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 is difficult to find out the missing value as well. A formal way of testing data if its MCAR or not is by doing MCAR test [18]. Table 2 lists examples of data that is missing completely at random.

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	

Missing at Random (MAR) Data

Missing at random [17] is where missing values are fully dependent on observed values available in dataset. The probability of missing value depends only on observed values not on other missing variables of same dataset. Example is given as, if salary of a person is missing then it can be estimated on qualification, experience and skills of the person. It means that the data missing at random can be predicted based on observed values. Table 3 presents examples which contains missing at random data. It is observed that the persons having age greater than 30 have a handsome salary around 100k and they own a car as well. So, the missing data can either be imputed by analysing the available data or can be imputed by using high level imputation method.

Table 3: Data Missing at Random.

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

Missing not at Random (MNAR) Data

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

Table 4: Data Missing not at Random.

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

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

Imputation Techniques/Algorithms

With arise in 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 of handling missing data.

Statistical Methods

Statistical methods of missing data are categorized to further two categories mostly named as deletion and imputation. Few of the techniques are for deleting missing data and few techniques impute missing values with different mathematical operations. Listwise deletion is the most traditional method where a row of missing data is deleted or record is omitted and the remaining data is analyzed [8, 10]. 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 loss of more critical information and reduce the size of dataset. There is another method named as Pairwise deletion [9] that

also delete data but preserves more values than listwise deletion and is biased for both MAR and MCAR. Mean or median substitution is statistical method of imputation where missing value of a variable is substituted by taking the mean of the available values of that variable [13]. When the number of records that are missing are less, then this technique is helpful in recovering data. There are few other statistical techniques like Maximum likelihood and expectation maximization (EM) [12] that work on the basis of likelihood of 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 imputation of missing data. KNN is the supervised machine learning algorithm that uses its neighbors to impute missing values. The missing value is substituted by using distance function namely the Euclidean distance of nearest neighbors and impute the most likely one [10, 14]. A Neural Networks (NN) can also be used for imputation that is based on interconnection of multiple artificial neurons which create a complex structure between input and output to find data [16].

Proposed Methodology

Figure 1 presents the complete flow to analyze the performance of machine learning based imputed data. In this analysis, at the start, data is either selected from publicly available datasets from well-known repositories or collected according to requirements. Then the data is checked if it already contains missing values or not as there are many datasets that comes with missing values and most of the time creators of datasets indicate this problem.

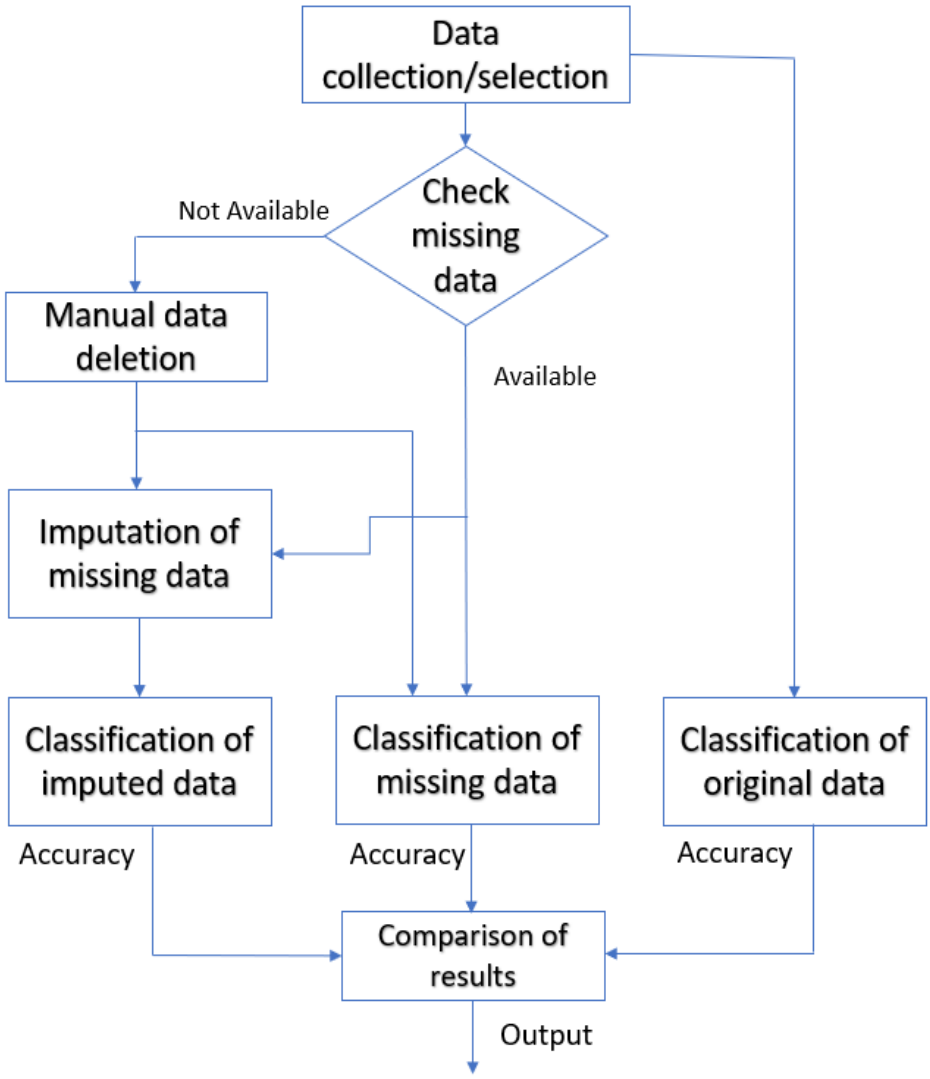

Figure 1: Proposed Flow for Analyzing Machine Learning based Data Imputation.

If the dataset doesn't contain any missing value, they are generated manually by deleting some records from the dataset for experiment purposes only. Once an incomplete dataset is available, it is then fed to the system where 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 forms of datasets available i.e., original dataset, datasets with missing values, dataset with statistical approach-based deleted values and dataset 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 along with other features is the parameter for comparison of results of the used techniques. If output results of the classification of the new datasets are better, then it indicates or proves the 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 proposed flow.

Classification Learner App of MATLAB

The classification learner app is used to perform these experiments and analysis. The classification learner app is a built-in tool/application of MATLAB which is normally used for classification purposes. It provides a user-friendly access to selection of specific dataset, required predictor, response variables and validation methods (cross-validation, hold out validation and no validation). After classification, it provides different visualization tools like, scatter plot, confusion matrix, ROC curve and parallel coordinates plot.

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

Used Datasets

The datasets used to evaluate the proposed flow are IRIS, WISCONSIN and HAR (Human activity recognition) using smartphone. These datasets are downloaded from UCI repository. IRIS dataset contains data related to IRIS plants according to three different classes named: Iris Setosa, Iris Versicolour and Iris Virginica defined by length and width of petal and sepal. 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 hospital in 90s. It consists of 32 different features out of them 1 is for classification containing two classes of cancer named: Benign and Malignant. There are total of 569 records available in this dataset which are real-time. The HAR is one of the recent collected datasets to monitor or classify human activities using mobile sensors. It contains data collected using two sensors (accelerometer and gyroscope) installed in smartphone. This dataset is consisted of 561 different attributes having time and frequency domain parameters. 30 volunteers aged 19-48 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 values are missed by deleting some records randomly and manually and saved 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 fed to two different ML-based algorithms (KNN and SKNN) programmed/trained in MATLAB. The manually missed values for evaluation are imputed by KNN and SKNN and are saved as different dataset files. List-wise statistical deletion approach is also used to tackle missing data. After completion of these process, 5 different datasets would be available, original dataset, the dataset having missing values, the dataset having LD-based approach, the dataset with imputed values using KNN and dataset with imputed values using SKNN. These datasets are then used for evaluation of ML-based imputed values by 2 different classifiers (SVM

and DT).

Results

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

Results of IRIS dataset

The different copies of 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 form of accuracy and the confusion matrix are obtained and analyzed. The accuracy is listed in table 5.

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	

The confusion matrix of IRIS dataset for both classifiers (SVM and DT) on four different datasets can be visualized from Figures 2 to 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 1-3 instead of

classes name on graphs which represent class names (IRIS-Setosa, IRIS-Versicolor and IRIS-Virginica) respectively.

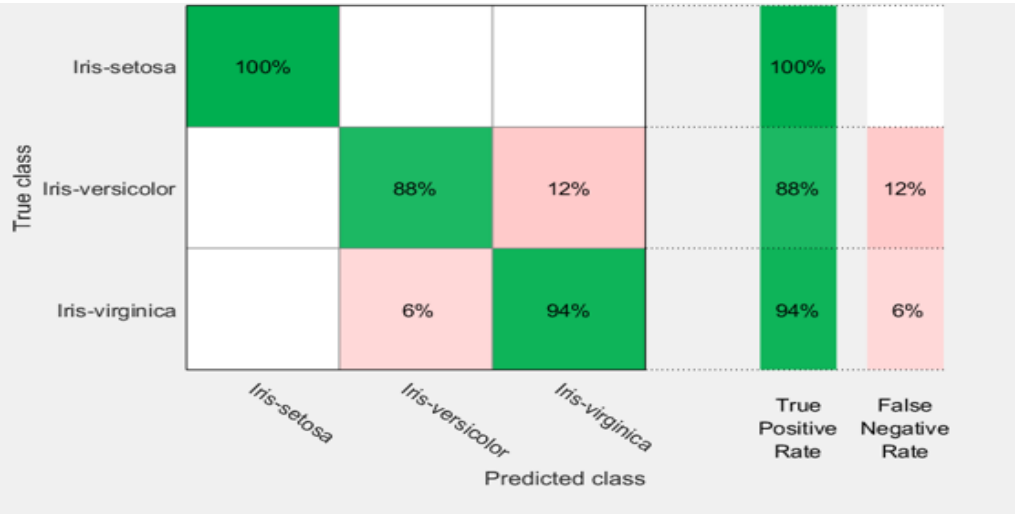

Figure 2: DT classification in terms of confusion matrix of original IRIS dataset.

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

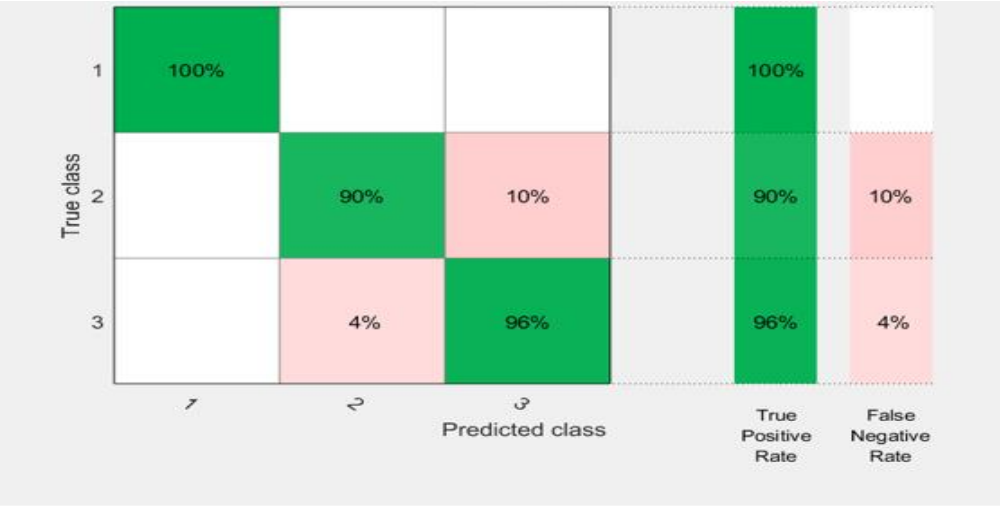

Figure 4: DT classification in terms of confusion matrix of imputed IRIS dataset with KNN.

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

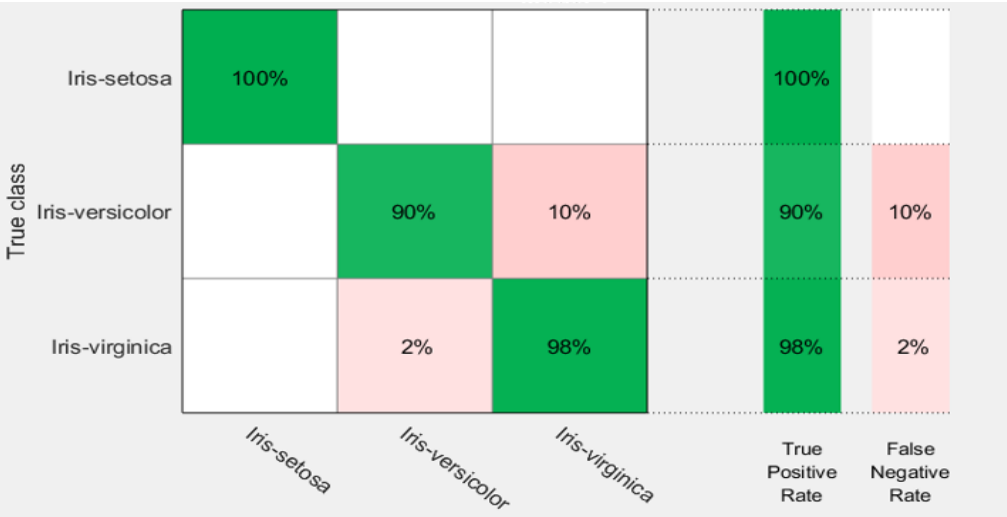

Figure 6: SVM classification in terms of confusion matrix of original IRIS dataset.

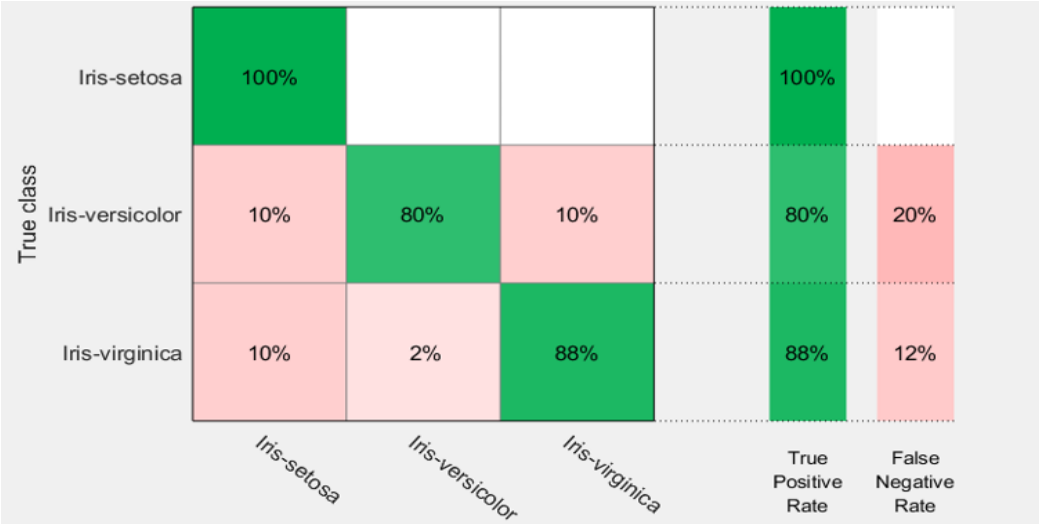

Figure 7: SVM classification in terms of confusion matrix of missing IRIS dataset.

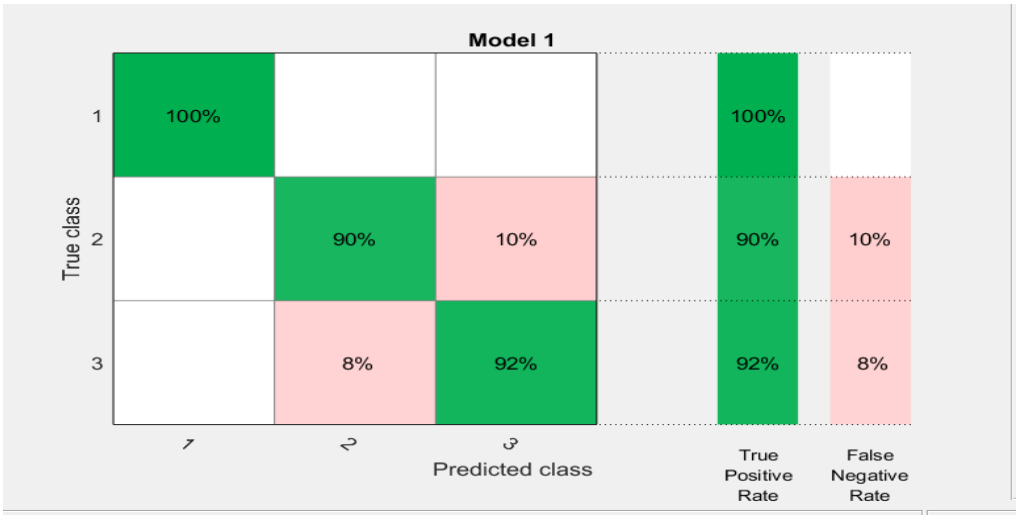

Figure 8: SVM classification in terms of confusion matrix of imputed IRIS dataset with KNN.

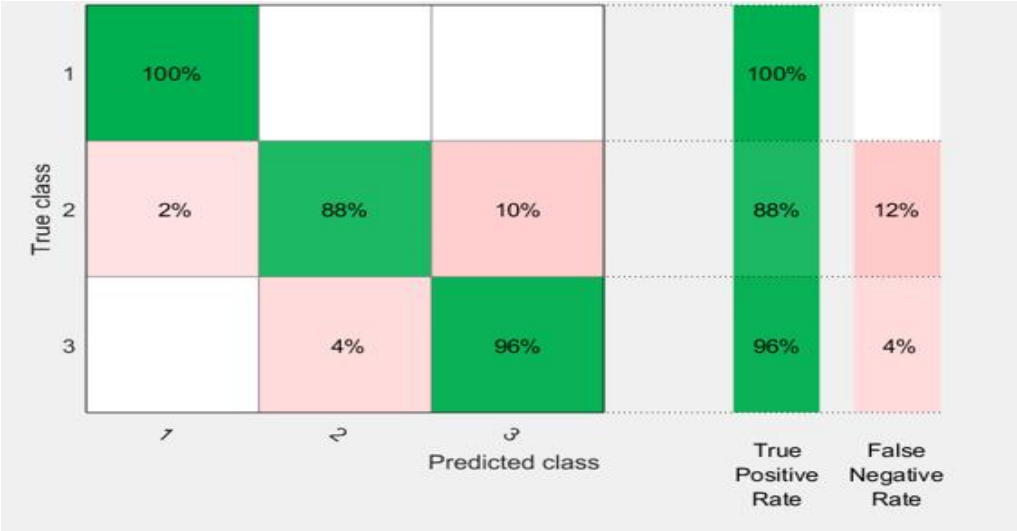

Figure 9: SVM classification in terms of confusion matrix of imputed IRIS dataset with SKNN.

Results of WISCONSIN dataset

The different copies of WISCONSIN 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 each dataset, both classifiers (SVM and DT) are trained as done in previous experiment and result in form of accuracy are obtained and analyzed. The accuracy is listed in table 6.

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		92.3

Results of HAR dataset

Like previous experiments, different copies of 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 each dataset, both classifiers (SVM and DT) are trained as done in previously and result in form of accuracy are obtained and analyzed. The accuracy is listed in table 7.

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

Discussion on Results

Implementation results for the IRIS dataset show that for the original dataset, 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 Listwise deletion yielded that accuracy is decreased with a reduction in the length of data. After that, implementation of 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 same procedure with both algorithms. The results yielded by this data are even more accurate than simple KNN-based imputed data. Moreover, DT again produced more classification accuracy than SVM classifier. For IRIS dataset, it can be concluded that, SKNN is best imputation technique and DT is best classification algorithm. Moreover, true positive and false negative rates of confusion matrices describe the classification in much better way to understand. WISCONSIN is one of the most concerned datasets as it is used for breast cancer diagnosis in patients. Implementation results for the Wisconsin dataset shows that for

the original dataset, the classification accuracy of SVM is proved better than DT. For dataset 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 Listwise deletion yielded that the accuracy is decreased with a reduction in the length of data. After that, implementation of KNN 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 same procedure with both clasifiers. The results yielded by this data are even more accurate than simple KNN-based imputed data in terms of KNN and DT. Moreover, SVM produced more classification accuracy than DT. For WISCONSIN dataset, it can be concluded that, either KNN or SKNN is the best imputation technique than traditional and SVM is best classification algorithm out of all three. Implementation results for the HAR dataset also shows the SKNN is best imputation technique than other traditional as well as machine learning technique.

Conclusion

In this paper, a flow is proposed to evaluate machine learning based imputation techniques and analyze 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 problem by imputation. Classification results of statistical and ML-based approaches to handle missing data were compared on three open 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) shows a definite increase in classification accuracy.

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

References

[1] J. W. Graham, "Missing data analysis: Making it work in the real world," Annual review of psychology, vol. 60, pp. 549-576, 2009.

[2] R. J. Little et al., "The prevention and treatment of missing data in clinical trials," New England Journal of Medicine, vol. 367, no. 14, pp. 1355-1360, 2012.

[3] R. O'Neill and R. Temple, "The prevention and treatment of missing data in clinical trials: an FDA perspective on the importance of dealing with it," Clinical Pharmacology & Therapeutics, vol. 91, no. 3, pp. 550-554, 2012.

[4] S. DeSarbo, P. E. Green, and J. D. Carroll, "AN ALTERNATING LEAST-SQUARES PROCEDURE FOR ESTIMATING MISSING PREFERENCE DATA IN PRODUCT-CONCEPT TESTING," Decision Sciences, vol. 17, no. 2, pp. 163-185, 1986.

[5] S. R. Wisniewski, A. C. Leon, M. W. Otto, and M. H. Trivedi, "Prevention of missing data in clinical research studies," Biological psychiatry, vol. 59, no. 11, pp. 997-1000, 2006.

[6] S. Wilcox et al., "Promoting adherence and retention to clinical trials in special populations: a women's health initiative workshop," Controlled clinical trials, vol. 22, no. 3, pp. 279-289, 2001.

[7] S. Juddoo, "Overview of data quality challenges in the context of Big Data," in 2015 International Conference on Computing, Communication and Security (ICCCS), 4-5 Dec. 2015 2015, pp. 1-9, doi: 10.1109/CCCS.2015.7374131.

[8] A. Donner, "The Relative Effectiveness of Procedures Commonly Used in Multiple Regression Analysis for Dealing with Missing Values," The American Statistician, vol. 36, no. 4, pp. 378-381, 1982, doi: 10.2307/2683092.

[9] J.-O. Kim and J. Curry, "The treatment of missing data in multivariate analysis," Sociological Methods & Research, vol. 6, no. 2, pp. 215-240, 1977.

[10] K. Orrawan, R. Panrasee, B. Thongchai, and C. Wichit, "Dealing with missing values for effective prediction of NPC recurrence," in 2008 SICE Annual Conference, 20-22 Aug. 2008 2008, pp. 1290-1294, doi: 10.1109/SICE.2008.4654856.

[11] R. C. Barros, M. P. Basgalupp, A. C. P. L. F. d. Carvalho, and A. A. Freitas, "A Survey of Evolutionary Algorithms for Decision-Tree Induction," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 42, no. 3, pp. 291-312, 2012, doi: 10.1109/TSMCC.2011.2157494.

[12] A. Dempster, N. Laird, and D. Rubin, "Maximal likelihood form incomplete data via the em algorithm," J. Roy. Statist. Soc, pp. 1-38.

[13] N. K. Malhotra, "Analyzing Marketing Research Data with Incomplete Information on the Dependent Variable," Journal of Marketing Research, vol. 24, no. 1, pp. 74-84, 1987, doi: 10.2307/3151755.

[14] M. Osman, A. Abu-Mahfouz, and P. Page, "A Survey on Data Imputation Techniques: Water Distribution System as a Use Case," IEEE Access, vol. 6, pp. 63279-63291, 10/22 2018, doi: 10.1109/ACCESS.2018.2877269.

[15] J. R. Quinlan, "Induction of decision trees," Machine learning, vol. 1, no. 1, pp. 81-106, 1986.

[16] W. Wei and Y. Tang, "A generic neural network approach for filling missing data in data mining," in SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme - System Security and Assurance (Cat. No.03CH37483), 8-8 Oct. 2003 2003, vol. 1, pp. 862-867 vol.1, doi: 10.1109/ICSMC.2003.1243923.

[17] R. J. Little and D. B. Rubin, Statistical analysis with missing data. John Wiley & Sons, 2019.

[18] R. J. A. Little, "A Test of Missing Completely at Random for Multivariate Data with Missing Values," Journal of the American Statistical Association, vol. 83, no. 404, pp. 1198-1202, 1988/12/01 1988, doi: 10.1080/01621459.1988.10478722.