Advances in Artificial Intelligence and Machine Learning; Research 2 (4) 516-532 Received 19-10-22; Accepted 01-12-22; Published 10-12-22

Missing Data Recovery in the E-health Context Based on Machine Learning Models

Ines Rahmany ines.rahmani@fstsbz.u-kairouan.tn
Department of Maths and Computer Sciences,
Faculty of Sciences and Technologies of Sidi Bouzid, University of Kairouan,
Sidi Bouzid, 9100, Tunisia

Sami Mahfoudhi* s.mahfoudhi@qu.edu.sa
Department of Management Information Systems and Production Management,
College of Business and Economics, Qassim University,
Buraydah, 51411, Saudi Arabia

Mushira Freihat* mm.freiat@qu.edu.sa
Department of Management Information Systems and Production Management,
College of Business and Economics, Qassim University,
Buraydah, 51411, Saudi Arabia

Tarek Moulahi* t.moulahi@qu.edu.sa
Department of Information Technology,
College of Computer, Qassim University,
Buraydah, 51411, Saudi Arabia

Corresponding Author: Ines Rahmany

Copyright © 2022 Ines Rahmany This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Diabetes mellitus is a set of metabolic illnesses characterized by abnormally high blood sugar levels. In 2017, 8.8% of the world's population had diabetes. By 2045, it is expected that this percentage will have risen to approximately 10%. Missing data, a prevalent problem even in a well-designed and controlled study, can have a major impact on the conclusions that can be derived from the available data. Missing data may decrease a study's statistical validity and lead to erroneous results due to distorted estimations. In this study, we hypothesize that (a) replacing missing values using machine learning techniques rather than the mean value and group mean value and (b) using SVM kernel RBF classifier will result in the highest level of accuracy in comparison to traditional techniques such as DT, RF, NB, SVM, AdaBoost, and ANN. The classification results improved significantly when using regression to replace the missing values over the group median or the mean. This is a 10% improvement over previously developed strategies that have been reported in the literature.

Keywords: E-Health, Missing data, Recovery, Classifications, Machine learning.

* These authors contributed equally to this work

Citation: Ines Rahmany et al. Missing Data Recovery in the E-health Context Based on Machine Learning Models Advances in Artificial Intelligence and Machine Learning. 2022;2(4):35.

1. INTRODUCTION

Recently, the proposal of Artificial Intelligence (AI) has invaded human lives as an intelligent solution to many complex problems [1]. AI is based on analyzing behaviors represented by data to make the next decisions. However, the application of AI to resolve a problem faces many possible drawbacks such as the loss of data linked to certain behaviors or features [2,3]. In fact, missing data produces many problems, particularly during the data analysis stage. The incomplete data minimizes the statistical and analytical power, which leads to the potential refusal of the null hypothesis if the test is wrong. Additionally, missing data may lead to bias in the estimation of the parameters. There is a problem when the sample does not represent the study. All of these issues lead to uncertainty about the truth of the data as well as unreliable outcomes of the studies. A number of statistics deal with missing data. These techniques are frequently used to test classic statistical hypotheses. However, it has not been sufficiently investigated how the analysis of deep observation, which is one of the techniques of machine learning for the new generation, impacts the performance of predicting the missed observation.

Machine learning methodologies, such as support vector machines (SVM), artificial neural networks (ANN), random forest (RF), and principal component analysis (PCA) [4], cannot be used for processing and decision-making if the data are incomplete. Hence, data recovery is of high importance to facilitate the right outcomes.

1.1 Motivation

Missing data (as well as missing value) is generally defined as a data value for a variable in the observation of interest that is not saved. Missing data, a prevalent problem in almost all studies, can have a major impact on the conclusions that can be derived from the data. Missing data may decrease a study's statistical validity and lead to erroneous results due to distorted estimations. In fact, missing data may be quite problematic. To begin with, most statistical processes automatically reject situations with missing data when there is insufficient data to complete the analysis. Second, due to the little amount of input data, the analysis may run but the conclusions may not be statistically significant. Third, if the studied instances are not a random representative sample of all cases, the results may be erroneous and misleading.

1.2 Contribution

In the field of statistics, the handling of missing data is a critical issue. When the data contains missing values, existing machine learning methods [5], have been ineffective primarily because they classify data (a) without providing replacement values for missing values, or (b) by simply replacing missing values with the mean value. We have developed a missing value strategy based on machine learning methodology. We have further optimized the data set by selecting the optimal classification model from a set of nine classification models to improve classification accuracy. In this study, we hypothesize that (a) replacing missing values using machine learning techniques rather than the mean value and group mean value, and (b) using SVM kernel RBF classifier will

result in the highest level of accuracy compared to traditional techniques such as DT, RF, NB, SVM, AdaBoost, and ANN. The following are the contributions of this work:

1. The design of a new framework based on machine learning techniques instead of median and group median methodology;
2. The optimization of the machine learning framework by selecting the best classification model among nine classifiers; and
3. The demonstration of a real improvement of 10% in classification accuracy compared to current state-of-the-art techniques by using SVM kernel-RBF.

1.3 Paper Organisation

The rest of this paper is organized as follows: Section 2 presents the literature review, section 3 presents the theoretical model, section 4 presents results and discussion including dataset description, and the conclusion is presented in section 5.

2. LITERATURE REVIEW

Missing data or absent data is defined as the value of data that is not available (blank) or missing for some variables. This problem has a significant impact on the results that can be extracted from the data. Thus, primarily in medical research, several studies have focused on processing the missing data, determining the effect of missing data, identifying the type of missing data, and finding the mechanism to eschew or handle the missing data [**?**].

2.1 The Effect of Missing Data

For many years it has been known that the rate of missing data is an essential problem in actual clinical experiments. Missing data has always been a strong challenge in clinical studies. The process of decision-making depends on the accuracy of the information. This accuracy is strongly based on the wholeness of the data from the data source. Nevertheless, in actuality, data tends to be both deficient and conflicting. Sometimes, data may be absent or entered incompletely, thereby negatively and passively affecting the data quality [7,8]. Medical records are the most important sources for clinical research, yet it is almost impossible to obtain data sets without losing values in actual clinical databases. For researchers, the major challenges of missing data include decreasing the statistical accuracy of a study, generating biased estimates, and finding incorrect results [8-10].

2.2 Types of Missing Data

According to previous studies, missing data is divided into three types which depend on the cause of the missing data [11,12]: missing completely at random, missing at random, and missing not at random.

1. Missing completely at random (MCAR): When the missing data does not depend on the probability of a particular value that is assumed to be received or the set of observed responses, this is called missing completely at random. Missing completely at random is a perfect but illogical presumption for many studies. Data missing via design due to machine failure, are considered to be missing completely at random. The data that are missing completely at random have the statistical benefit of the analysis remaining fair exemplifies techniques to deal with MCAR imputation K-nearest neighbors, statistical analysis of data set, and the identification of similar assets from a data set population [13].
2. Missing at random (MAR): This is a more effective supposition for many studies. When the missing data does not depend on the probability of a particular value that is assumed to be received but depends on the set of observed responses, this is considered missing at random. As there is no bias in randomness, it can be concluded that there is no trouble from data that is missing at random. Therefore, missing at random does not mean that the missing data can be ignored. If the missing variables are missing randomly, we would suppose the probability of the variable missing in each case to be conditionally independent of the variable which is present or is expected to be obtained in the future, given the history of cases that have previously obtained the variable. It is used to deal with MAR by identifying similar assets from a population of data set. Statistical analyses methods are used to identify data possessing the same characteristics.
3. Missing not at random (MNAR): If the properties of missing data are not like the properties of data missing completely at random or the properties of data missing at random, then they are missing not at random. When data are missing not at random, there are problems. Creating a model for missing data is the only way to obtain a fair approximation of the variable in this case. The model can then be combined with a more complex model to estimate the missing values. The techniques used to deal with MNAR utilize machine learning approaches to imputing data.

2.3 Mechanisms to Eschew or Handle the Missing Data

The best way to deal with missing data is to make improvements in the data collection process so that the problem can be prevented through good study planning and careful data collection [11,14]. There are two ways to handle the missing data: deletion and imputation. Although deletion is simpler than imputation, it is not the ideal solution to deal with the missing data because deletion eliminates the responses with missing values. This causes biased estimation. In addition, when there are many responses that have missing values, this leads to standard errors due to reduced sample size. The imputation solution is used more frequently than deletion. Using easy to complex imputation techniques, this solution exchanges the missing data with alternative values [15,16]. It is irrational to think that this will delete all missing data from the study. Therefore, it is important to understand the methods used to address the type of missing data [17].

Many techniques of handling the missing data have been developed, several of which we examine throughout this study. First and most commonly, the listwise or case deletion technique deletes those cases that contain missing data and analyzes the remaining data. Listwise deletion has become the first choice for analysis in most statistical programs. Many researchers agree that listwise deletion is likely to generate bias in the estimation of the parameters. However, listwise deletion generates

fair estimates of the parameters when the missing completely at random assumption is satisfied. When the missing at random or missing not at random assumptions are satisfied, listwise deletion may produce bias in the estimates [18]. Additionally, in a big sample, listwise deletion may be a logical choice. Otherwise, listwise deletion is not a logical choice of technique.

The second technique is pairwise deletion. This technique deletes data only based on the particular assumption that specific data is needed but missing. Existing data is used in the statistical analysis if data is missing elsewhere in the data set. The pairwise deletion technique keeps more data than the listwise deletion technique because it uses all data set. This technique has the following issues: the model parameters will depend on various sets of data with various statistics such as the size of the sample and standard errors, and it can generate a specific non-positive correlation matrix which is likely to prevent further analysis [19]. When the missing at random or missing not at random assumption is satisfied, the pairwise deletion technique may produce fair estimates, and the compatible techniques are included as covariates. However, when there are many missing data, then the analysis will be incomplete.

The third technique is mean substitution; the losing data value placed by the variable mean for that identical variable. This allows the data to be used even if there is missing data. Using the mean in mean substitution is a reasonable estimate for selecting an observation randomly from a normal distribution. However, for missing data that are not completely random, particularly if there is a large variance in the number of missing data for diverse variables, the mean substitution technique may produce bias in the estimates. In addition, this technique does not add new data but only raises the data set. This results in minimizing the errors [20]. Generally, mean substitution is not accepted.

The fourth technique is regression imputation; by substituting the missing data with estimated values, this technique keeps all situations by predicting missing data from other available data and then filling in the missing data with a possible value. After filling in the missing data by this technique, the data set is analyzed by the standard statistical techniques for whole data. This technique has many advantages because the imputation keeps more data than the listwise or pairwise deletion techniques and eschews major changes in the standard deviation or the distribution shape. However, similar to mean substitution, when missing data are replaced by regression that is predicted from other variables, no new data is actually added while the data set is raised and the standard error is minimized.

The fifth technique is maximum likelihood; this technique has been utilized by many studies to handle the missing data. The maximum likelihood technique uses a spotted sample drawn from a multivariate normal distribution. By using the obtainable data, the parameters are estimated. Then, it is easy to estimate the missing data based on the estimated parameters. If there is missing data but the data is comparatively complete, then the correlations between the variables can be calculated by the maximum likelihood technique. So, by performing the conditional distribution of the other variables, the missing data can be estimated [21].

The sixth technique is expectation maximization. Similar to the maximum likelihood technique, expectation maximization is applied to produce a new data set. By using the maximum likelihood technique, all missing data are filled with estimated data. The expectation maximization technique starts with expectation. First, the parameters are estimated, such as using the listwise deletion. Next, the results are applied to generate a regression equation. Then, this equation is applied to complete the missing data by using maximization. Again, with the new parameters,

the expectation is performed and a new regression equations are generated to handle the missing data. The expectation and maximization steps are repeated until the parameters are stabilized. The expectation maximization technique has both advantages and weaknesses. The main advantage of the expectation maximization technique is the creation of the new sample size with no missing values. For each imputed value, a random disturbance is incorporated to detect the unreliability related to the imputation. In addition, the expectation maximization technique has some weaknesses, as it may take a long time, particularly when there is a massive portion of losing data. As well, it is very difficult for extraordinary statisticians to agree. This technique may produce biased parameter estimates and may estimate the standard error incorrectly [21].

The seventh technique, multiple imputations, is a good technique for dealing with missing data. With the multiple imputations technique, instead of replacing missing data with one value, the missing data are substituted with a set of logical data. This technique starts with a forecast of the missing values by the current values using else variables [22]. Then, the forecasted data complete the missing data to produce an imputed data set from a full sample size. Then, various computed data sets are generated by repeating this process. After the imputed data set is created, it is analyzed by standard statistical analysis tools to obtain complete data. Various analysis results are produced. Thence, all analysis outcomes are merged to produce one overall analysis outcome. The advantage of the multiple imputation technique is that it can resolve the normal disruption of the missing data. Including uncertainty due to missing data leads to a valid statistical conclusion. Multiple imputation has overcome uncertainty associated with the missing data estimation to generate a valid statistical conclusion. In addition, multiple imputation has overcome a small sample size or a large number of missing data to produce adequate results. Multiple imputation can be used easily with the new statistical tools, even though the statistical principles of this technique may be difficult to understand [22].

The last technique is sensitivity analysis. This technique determines how the uncertainty in the form result can be assigned to various sources of uncertainty in its inputs. While analyzing the missing data, extra assumptions about the causes for the missing data are produced. These assumptions can often be applied in standard statistical analysis. However, the assumptions cannot be validated with certainty. Thus, the National Research Council has suggested that sensitivity analysis be performed to assess the strength of the outcomes with the deviations from the missing at random assumption [23].

Finally, after reviewing all these previous studies, we conclude that the optimum solution for missing data is to increase the collection of the data during the study design. Techniques or mechanisms to handle the missing data should only be used after all efforts have been made to minimize the amount of missing data. It is difficult to determine whether the multiple imputations technique or the full maximum likelihood technique is better; however, both are superior to the traditional techniques. Both techniques have optimal performance with big data sets. Multiple imputation is a perfect technique to use when analyzing samples with missing data [11].

3. PROPOSED MODEL

Our main approach is to replace the dataset's missing values that are needed for building the machine learning models using machine learning itself. It is known that replacing missing values is crucial

for obtaining good metric scores in the construction of machine learning models. Therefore, to obtain better results, we need to replace these missing values. So far, some commonly used classical techniques for replacing the missing values are to use the median, the mode, or the mean. However, our approach is to try to use machine learning to replace missing values before building the machine learning models. We compare the built models' metric values with those based on the classical techniques for replacing missing values.

Therefore, our methodology is to replace the missing data values of a specific column or feature based on the data of the other features through classification or regression predictions. Basically, if one feature is missing some values, our methodology uses the other columns or the other features to train a new model to replace the missing data by prediction. This target feature will be used as the Y matrix. On the other hand, the columns will represent the X matrix. Then, we train a machine learning model just on the X values targeting the Y column in order to build a model to predict the missing values later. After replacing the missing values, we build another machine learning model in order to predict the initial target column.

Thus, in order to replace missing values through machine learning models, our approach is based on choosing the best model from the following regressors: Decision Tree, Random Forest, LAssoCV, Ridge, ElasticNetCV, and LinearRegression.

We compare these regressors using the following metrics:

First, the Mean Absolute Error (MAE) is a measure of errors between paired observations expressing the same phenomenon.

$$MAE = \frac{\sum_{i=1}^{n} y_i - x_i}{n} \tag{1}$$

Where:
MAE = mean absolute error,
y_i = prediction,
x_i = true value,
n = total number of data points.

Second, the Root Mean Square Error (RMSE) or the Root Mean Square Deviation (RMSD) is the standard deviation of the residuals (prediction errors). Residuals are a measure of distance from the regression line data points; RMSE is a measure of how spread out these residuals are. In other words, it indicates the level of data concentration around the line of best fit. Root Mean Square Error is commonly used in regression analysis to verify experimental results.

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \widehat{x_i})^2}{N}} \tag{2}$$

Where:
$RMSD$ = root-mean-square deviation,

i = variable i,
n = number of non-missing data points,
x_i = actual observations time series,
$\widehat{x}$ = estimated time series.

Therefore, after trying different regressors and comparing them based on the MAE and the RMSE metrics, we begin the classifiers model-building phase. In this phase, we use the completed filled data to train different models for the initial model-building. After building the classifiers, we use the accuracy metrics to choose the best classifier.

Then, we repeat the classifiers model-building based on the mean and the median techniques in order to compare them with the machine learning technique. We can prove that the machine learning technique is better at replacing the missing values.

The proposed classical techniques were motivated by the principle of statistical measurements integrated into a classification framework. However, we have developed a missing value strategy based on machine learning methodology. We have further optimized the data set by selecting the best classification model from a set of nine classification models to improve classification accuracy.

In this section, the theoretical model of the proposed contribution is shown. TABLE 1 illustrates the four phases of this model. The proposed machine learning scheme for missing data recovery is presented in FIGURE 1.

3.1 Phase 1: Applying Group Median to Replace Missing Values Methodology

In this phase we take the following steps:

1. Replace all the missing values in the diabetic dataset with the group median and create a new dataset of Diabetic1.
2. Apply classifiers to Diabetic1 in which the Outcome column is the target column.

3.2 Phase 2: Applying Mean to Replace Missing Values Methodology

In this phase, we repeat these steps to replace the mean methodology.

3.3 Phase 3: Apply Machine Learning to Replace Missing Values Methodology

In this phase, we apply the machine learning replacing values methodology to the Insulin column through the following steps:

1. Drop the Outcome column.

2. Create a new dataset D1 containing only the lines with non-missing values for the Insulin column. Set the Insulin column as the target column.
3. Build different machine learning regression models based on D1.
4. Choose the most significant model based on the RMSE and MAE metrics. However, considering the fact that the RMSE has the same unit as the predicted values, we have to take a look at the importance of the RMSE in comparison with the predicted values. Thus, we have to estimate the Scatter Index, which is simply the RMSE divided by the average value of the observed values. The same applies to the MAE.
5. Predict the missing values of the Insulin column by using the LassoCV regression model. Create a new dataset D2.
6. Regroup all the lines of the Diabetic dataset from merging D1 and D2. Create a new dataset Diabetic2.
7. Build different machine learning classifiers based on Diabetic2 in which the Outcome column is the target column.

3.4 Phase 4: Compare Metrics Accuracy

In this phase, we compare the metrics (Accuracy) of the classifiers applied on Diabetic1 with the metrics of the classifiers applied on Diabetic2.

Table 1: Steps and phases proposed model.

Step	**Step description**	**Phase**
Step 1.1	Apply group median to replace missing values and build different classifiers accordingly.	Phase 1
Step 2.1	Apply mean to replace missing values and build different classifiers accordingly.	Phase 2
Step 3.1	Drop the Outcome column.	*Phase 3(1)
Step 3.2	Create a new dataset D1 containing only the lines with non-missing values for the Insulin column.	
Step 3.3	Build different machine learning regression models.	
Step 3.4	Choose the most accurate model.	
Step 3.5	Predict the missing values of the Insulin column.	
Step 3.6	Regroup all the lines of the diabetic dataset	
Step 3.7	Build different machine learning classifiers.	
Step 4.1	Compare the metric accuracy of the different built classifiers.	Phase 4

(1): Apply machine learning to replace missing values methodology.

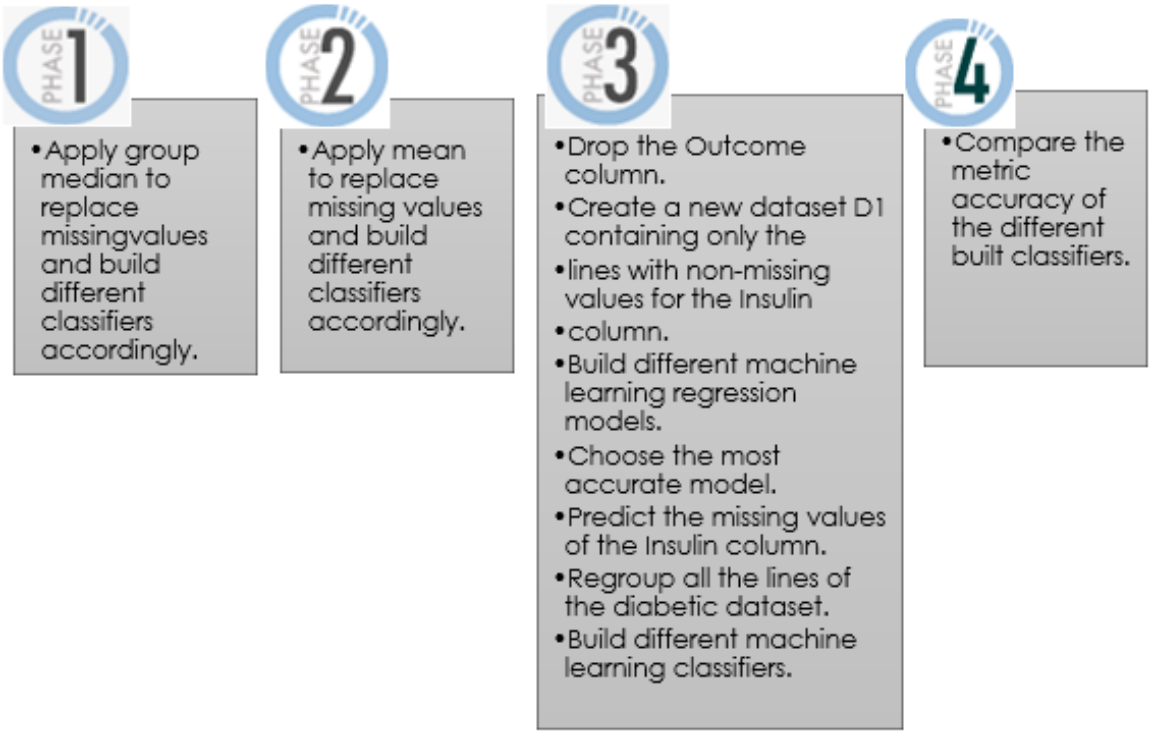

Figure 1: Machine Learning scheme for missing data recovery.

3.5 Dataset

To validate our approach and to demonstrate the higher performance of the machine learning replacing missing data technique, the diabetes dataset (2) was obtained from the UCI Repository at the University of California, Irvine. There are 768 female patients in this dataset, all of whom are at least 21 years old and of Pima Indian ancestry, with 268 diabetic cases and 500 controls. In this dataset, 5 patients have a zero-glucose level, 35 patients have zero diastolic blood pressure, 27 patients have zero body mass index, 227 patients have zero skinfold thickness, and 374 patients have zero serum insulin level. These zero values are interpreted as missing values since they have no meaning.

FIGURE 2 contains descriptions of the attributes as well as a statistical summary.

SN	Attributes	Descriptions	Attributes type's	Mean ± SD
1	Pregnant	Number of times pregnant	Continuous	3.84 ± 3.36
2	Glucose	Plasma glucose (2-h)	Continuous	121.67 ± 30.46
3	Pressure	Diastolic blood pressure (mm Hg)	Continuous	72.38 ± 12.10
4	Triceps	Triceps skin fold thickness (mm)	Continuous	29.08 ± 8.89
5	Insulin	Two hours serum-insulin (μ U/ml)	Continuous	141.76 ± 89.10
6	Mass	Body mass index (weight in kg/ (height in m)^2)	Continuous	32.43 ± 6.88
7	Pedigree	Diabetes pedigree function	Continuous	0.47 ± 0.33
8	Age	Age (years)	Continuous	33.24 ± 11.76
9	Class	Diabetic (500) vs. control (268)	Categorical	–

Figure 2: Description - Demographics of the diabetic patient cohort.

4. RESULTS AND DISCUSSION

The appraisal of our work is discussed in detail in this section. Please keep in mind that the suggested solution was tested on Google Colab and was written in the Python scripting language. Pandas, Scikit-learn, and Keras are well-known Python packages that were used in the development of the preprocessing, training, and testing of the different machine learning model-building phases.

Following the collection of managed data, data preparation occurs. In this step, the data are cleaned and formatted for the next stage of the information process. This is frequently referred to as "preprocessing." The purpose of this preparation is to eliminate low-quality data that may be missing, redundant, or wrong, and to begin creating data that will assure the high quality of the intelligent model chosen. In fact, the raw dataset is meticulously scrutinized for errors of any type. We turn the input data into a vectorizable matrix in this stage. Thus, replacing the missing data is crucial for the model-building process.

Using our approach, we applied six regressor models to extract the missing values. When comparing the different regressors, we concluded that the LassoCV regressor model outperforms the rest of the models (Decision Tree, Random Forest, Ridge, ElasticNetCV, and LinearRegression) in terms of scatter index for both the MAE and the RMSE metrics. We obtained a score of 0.42 and 0.66 respectively for the MAE scatter index and the RMSE scatter index for the LassoCV regressor (TABLE 2).

Table 2: Regression model results in term of scatter index for RMSE and MAE values.

Regression model	**Scatter index for MAE**	**Scatter index for RMSE**
Decision Tree	0.541	0.829
Random Forest	0.425	0.676
LAssoCV	0.420	0.661
Ridge	0.420	0.664
LinearRegression	0.424	0.661

Hence, to evaluate our machine learning technique of replacing missing data, we use a number of machine/deep learning classifiers in the context of predicting diabetes. Numerous metrics can be used in this test, such as accuracy (Ac), recall, detection rate (DR), positive predictive value (PP), and negative predictive value (NP). These factors are evaluated with the help of true positive (TP), true negative (TN), false positive (FP), and false negative (FN). However, we rely essentially on two main metrics: accuracy and the F-measure which represents the harmonic mean of recall and precision [24,25].

In this phase, we compare the metrics (accuracy) of the built classifiers based on the different techniques of median, mean, and machine learning (TABLE 3 and FIGURE 3). The accuracy metric is used when the true positives and true negatives are more important, while the F1-score is used when the false negatives and false positives are crucial. Accuracy can be used when the class distribution is similar, while F1-score is a better metric to use when there are imbalanced classes as in the above case. Therefore, in our case, the value counts for the Outcome target column is 500 for the 0 class and 268 for the 1 class. Thus, we can conclude with confidence that the accuracy metric is more suitable in comparing the classifiers since we are dealing with balanced classes. The classification results improve significantly when using regression to replace the missing values over the group median or the mean.

$$Accuracy(ACC) = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

Where:

- TP : True Positive
- TN : True Negative
- FP : False Positive
- FN : False Negative

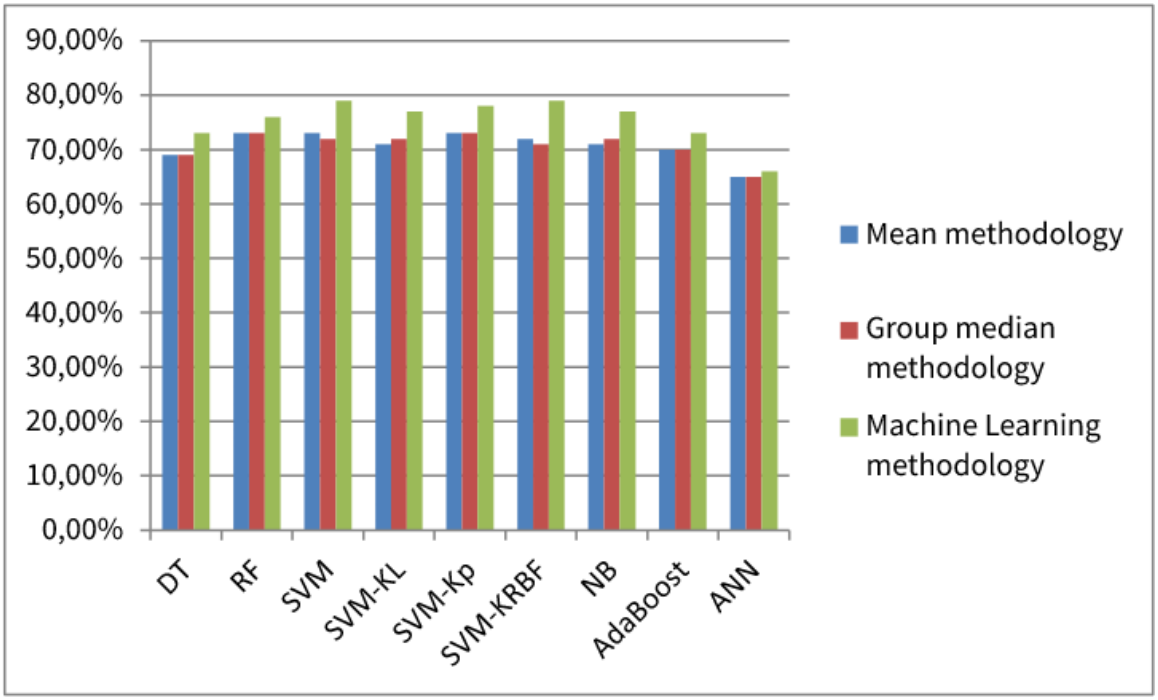

Figure 3: Comparisons of all classifiers and Machine Learning methodology over the the mean and the group median methodologies.

Therefore, in our research, the main objective is to show that the prediction accuracy scores of the chosen classifiers differ positively when using the machine learning technique to replace the missing data in the training dataset. In fact, the machine learning methodology outperforms the mean and the median techniques in all of the nine experimental classifiers. In fact, the accuracy improvement percentage, as exposed in TABLE 4, goes up to 10.87% in the SVM (kernel RBF) classifier which obtained the highest accuracy score of 0.797.

5. CONCLUSION

Diabetes Mellitus (DM) is a set of metabolic illnesses characterized by abnormally high blood sugar levels. Our hypothesis was that filling missing values using machine learning regression techniques would result in greater accuracy when utilized in an ML framework rather than the classical statistical methodologies such as mean and median values. We demonstrated our hypothesis by showing up to 10.87% improvement in term of accuracy. Five regression models, nine classifiers, three different missing data replacement methodologies were used in a comprehensive data analysis. Methodologies' comparison was used to examine the situation based on three different metrics and showed that there had been significant improvements.

Table 3: The overall classifications results.

Classifier	**Mean methodology**	**Group median methodology**	**Machine Learning methodology**	**Accuracy Improvement percentage rate compared to the Mean methodology**	**Accuracy Improvement percentage rate compared to the Median methodology**
Decision Tree	69.270	69.270	73.958	6.766	6.766
Random Forest	73.958	73.958	76.041	2.816	2.816
SVM	73.437	72.916	79.166	7.801	8.571
SVM kernel linear	71.875	72.395	77.604	7.971	7.194
SVM kernel poly	73.4375	73.958	78.645	7.092	6.338
SVM kernel RBF	72.395	71.875	**79.687**	10.071	10.869
Naïve Bayes	71.875	72.395	77.083	7.246	6.474
AdaBoost	70.833	70.833	73.958	4.411	4.411
ANN	65.104	65.96	66.16	1.622	0.303

In the future, we want to compare additional contributions to the same context in order to make more informed decisions. We think it would be fascinating to see other sorts of medical data categorized in this manner, which might result in a cost-effective and time-saving solution for both diabetic patients and physicians.

6. AUTHOR CONTRIBUTIONS

Conceptualization, T. Moulahi; methodology, S. Mahfoudhi; software, S. Mahfoudhi; validation, S. Mahfoudhi; formal analysis, I. Rahamany; investigation, I. Rahamany; resources, S. Mahfoudhi; data curation, S. Mahfoudhi; writing—original draft preparation, M. Freihat; writing—review and editing, I. Rahamany; visualization, M. Freiha; supervision,T. Moulahi; project administration, T. Moulahi; All authors have read and agreed to the published version of the manuscript.

Table 4: Improvement in classification results when using machine learning to replace the missing values over the group median or the mean.

Classifier	Mean methodology	Group median methodology	Machine learning methodology	Accuracy improvement percentage rate compared to the Mean methodology	Accuracy Improvement percentage rate compared to the Median methodology
Decision Tree	0.692	0.692	0.739	6.767	6.767
Random Forest	0.739	0.739	0.76	2.817	2.817
SVM	0.734	0.729	0.791	7.802	8.572
SVM kernel linear	0.718	0.723	0.776	7.972	7.194
SVM kernel poly	0.734	0.739	0.786	7.092	6.338
SVM kernel RBF	0.723	0.71875	0.797	10.072	10.87
Naive Bayes	0.718	0.724	0.77	7.246	6.475
AdaBoost	0.708	0.708	0.739	4.412	4.412
ANN	0.651	0.659	0.661	1.622	0.303

7. ACKNOWLEDGMENTS

The researchers would like to thank the Deanship of Scientific Research, Qassim University for funding the publication of this project

8. DATA AVAILABILITY

The used dataset is downloaded from (`https://www.kaggle.com/uciml/pima-indians-diabetes-database`). No ethics approval is required for this dataset.

9. CONFLICTS OF INTEREST

The authors declare no conflict of interest.

References

[1] Zidi S, Moulahi T, Alaya B. Fault Detection in Wireless Sensor Networks Through SVM Classifier. IEEE Sens J. 2017;18:340-347.

[2] Moulahi T. Joining Formal Concept Analysis to Feature Extraction for Data Pruning in Cloud of Things. Comput J. 2022;65:2484–2492.

[3] Moulahi T, El Khediri S, Khan RU, Zidi S. A Fog Computing Data Reduce Level to Enhance the Cloud of Things Performance. Int J Commun Syst. 2021;34:e4812.

[4] Mahfoudhi S, Frehat M, Moulahi T. Enhancing Cloud of Things Performance by Avoiding Unnecessary Data Through Artificial Intelligence Tools. In: 15th International Wireless Communications & Mobile Comput Conference (IWCMC). Vol. 2019. IEEE Publications; 2019: 1463-1467.

[5] Bashir S, Qamar U, Khan FH. Intellihealth: A Medical Decision Support Application Using a Novel Weighted Multi-Layer Classifier Ensemble Framework. J Biomed Inform. 2016;59:185-200.

[6] Graham JW. Missing Data Analysis: Making It Work in the Real World. Annu Rev Psychol. 2009;60:549-76.

[7] Kang H. The Prevention and Handling of the Missing Data. Korean J Anesthesiol. 2013;64:402-406.

[8] Wells BJ, Nowacki AS, Chagin K, Kattan MW. Strategies for Handling Missing Data in Electronic Health Record Derived Data. EGEMS (Wash DC).2013;1:1035.

[9] Leke CA, Marwala T. Introduction to Missing Data Estimation. Deep Learning and Missing Data in Engineering Systems. 2019:1-20.

[10] Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, et al. Multiple Imputation for Missing Data in Epidemiological and Clinical Research: Potential and Pitfalls. BMJ. 2009;338:b2393.

[11] O'Neill RT, Temple R. The Prevention and Treatment of Missing Data in Clinical Trials: An FDA Perspective on the Importance of Dealing With It. Clin Pharmacol Ther. 2012;91:550-554.

[12] Kose T, Ozgur S, Cosgum E, Keskinoglu A, Keskinoglu P. Effect of Missing Data Imputation on Deep Learning Prediction Performance for Vesicoureteral Reflux and Recurrent Urinary Tract Infection Clinical Study, BioMed Research International. 2020.

[13] McMahon P, Zhang T, Dwight RA. Approaches to Dealing With Missing Data in Railway Asset Management. IEEE Access. 2020;8:48177-94.

[14] Rubin DB. Inference and Missind Data. Biometrika. 1976;63:581-592. doi: 10.1093/biomet/63.3.581.

[15] Xue Z, Wang H. Effective Density-Based Clustering Algorithms for Incomplete Data. Big Data Mining and Analytics. 2021;4:183-194.

[16] Alharbi H, Kimura M. Missing Data Imputation Using Data Generated by GAN. ICCBD. 2020, August 05-07. Taichung, Taiwan. 2020:73-77.

[17] Wisniewski SR, Leon AC, Otto MW, Trivedi MH. Prevention of Missing Data in Clinical Research Studies. Biol Psychiatry. 2006;59:997-1000.

[18] Brian J, Kevin M, Amy S, Michael W. Strategies for Handling Missing Data in Electronic Health Record Derived Data. EGEMs (Gener Evid Methods Improve Patient Outcomes). 2013;1:1035.

[19] Donner A. The Relative Effectiveness of Procedures Commonly Used in Multiple Regression Analysis for Dealing With Missing Values. Am Stat. 1982;36:378-381.

[20] Kim JO, Curry J. The Treatment of Missing Data in Multivariate Analysis. Sociol Methods Res. 1977;6:215-240.

[21] Malhotra NK. Analyzing Marketing Research Data With Incomplete Information on the Dependent Variable. J Mark Res. 1987;24:74-84.

[22] Dempster AP, Laird NM, Rubin DB. Maximum Likelihood From Incomplete Data via the EM Algorithm. J R Stat Soc Series B Stat Methodol. 1997;39:1-38.

[23] Sinharay S, Stern HS, Russell D. The Use of Multiple Imputation for the Analysis of Missing Data. Psychol Methods. 2001;6:317-329.

[24] Ben Daoud W, Mahfoudhi S. SIMAD: Secure Intelligent Method for IoT- Fog Environments Attacks Detection. CMC Comput Mater Continua. 2022;70:2727-2742.

[25] Panel on Handling Missing Data in Clinical Trials. The Prevention and Treatment of Missing Data in Clinical Trials. 2nd ed. Washington, DC: National Academies Press. 2010:107-114.