https://doi.org/10.48047/AFJBS.6.12.2024.2927-2939

AFJBS

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

Healthcare Data Quality Enhancement By Identifying and Replacing Missing Values

[1]Suresh Kapare, [2]Dr. V. Maria Anu

[1]Research Scholar, Sathyabama Institute of Science and Technology, Chennai.

[2]Associate Professor, VIT, Chennai.

Email-ID:[1]sureshkapare2012@gmail.com, [2]mariaanu.v@vit.ac.in

Article History
Volume 6, Issue 12, 2024
Received Date: 20May 2024
Acceptance Date: 28 June 2024
Doi:
10.48047/AFJBS.6.12.2024.2927-2939

Abstract

Data science and engineering are two different concerns when working on large volumes of data. Data engineering is the process of organizing, managing, maintaining, and pipelining data, whereas data science is the process of analyzing and manipulating data. The original or raw data generated from various sources are poor in quality and can not be processed or analyzed without preprocessing. Especially in the medical or healthcare industry, poor-quality data leads to wrong diagnosis and treatment. Data availability is also increasing rapidly because of the increasing number of online applications. One of the reasons is because of missing values and outliers. The data quality determines the prediction model's efficiency and accuracy. Though it is impossible to maintain a dataset without missing values, various methods are detected to extract the maximum accuracy possible from the available model. Imputation of duplicate values is widely used in various fields that provide the necessary accuracy for the prediction model. However, duplicating the value during the imputation process should not affect the dataset's quality or the model's performance. This paper discusses the existing imputation methods to tackle missing values, and the dataset's quality is evaluated. A model healthcare dataset is considered, the proposed missing value analysis methods are experimented with, and their performance is verified. This model clearly shows that the proposed model provides better data imputation and improves the model's performance. It is experimented with various classification algorithms, and their results are compared.

Keywords: Data Preprocessing, Missing Value Detection, Data Quality, Duplicate Data, Healthcare Data.

Introduction

The missing data problem should be handled appropriately to improve the data quality; otherwise, the manipulation output becomes wrong. A data science aspirant must be good at missing data prediction for data processing. In order to get good results in any data processing, it is essential to check the data quality. Thus, a Database Management System(DBMS) is required to maintain the datasets. Only complete data can save the results of machine learning and also its accuracy. The current paper gives the missing data's reason, representation, and description.

Along with different categories, the ways of handling missing values are given with examples.Missing data is defined as the value unavailable in the variable. In a dataset, the blank space shows the missing values. In Pandas, usually, missing values are represented by NaN. It stands for Not a Number.Multiple reasons have been found so far for data missing. These reasons affect the approach to handling missing data. A few causes are: The process might get corrupted due to improper data feed, failure in recording the values, the User not providing the values and Item nonresponse. It states that the participant refused to respond. Missing values are categorized as MCAR, MNAR, and MAR.

Missing Completely At Random(MCAR)

Here, the probability of data being lost is the same for all observations.There is no relationship between the missing data and any other values within the given dataset. Missing values are completely independent of other data. There is no pattern. In MCAR, the major reasons for missing data may be human error, equipment failure, loss of sample, and unsatisfactory technicalities while the data is being recorded. If the data in the overdue books in the library are missing, then the reason might also be a human error. So, the missing values of overdue books are separate from any other variable or data in the system. Assuming it to be a rare case must be avoided. The advantage of such data is that the statistical analysis remains unbiased.

Missing at Random(MAR)

From this, the missing values can be explained by variables with complete information. In this case, there is no relationship between the missing data and the variable. The data is found to be unavailable within its sub-samples. Some pattern is also found. As an example, on checking the survey, you may find that whoever has answered their gender is female; their age is found to be missing. This is because most women would not like revealing their age and maintaining it as a secret. Thus, in this phase, the missing data's probability depends only on the observed value. Prediction of missing value and its respective reason becomes a tough task based on two variables. The variables' Gender' and 'Age'are found to be related. If a poll is taken, gender and the number of overdue books are asked. Assuming that women are most likely and men are less likely to answer the poll, the data would be found to be missing based on gender. In such a case, the statistical analysis might result in a bias. An unbiased estimate of the parameters is done only by modeling the missing data.

Missing Not At Random(MNAR)

Here, missing values depend on the observed data. If a pattern is missing, the other observed data cannot explain it. In this case, it is considered Missing, Not At Random(MNAR). If the missing data does not fall under MCAR or MAR, it can be categorized as MNAR. Due to the reluctance of people to provide their information for constructive purposes, it happens. Some respondents may need to answer the questions in the survey also. As an example, the name and number of overdue books in a library, the people having more overdue books are less likely to answer. Thus, the missing value of the number of overdue books depends on the people answering the poll with more overdue books. In such a survey, it was found that people with less income refused to share their financial status, which eventually resulted in missing data. In MNAR also, the statistical analysis might result in bias. It can be seen in Figure-1.

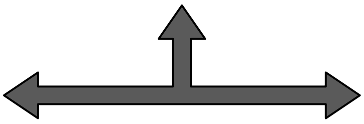

Figure-1. Different Types Of Missing Values

Is handling the missing data a major concern?

Yes, Much care should be given to handling the missing data appropriately. Precision gets abandoned due to missing data in statistical analysis. Every machine learning algorithm needs to be revised in determining the datasets containing missing values. Sometimes, algorithms such as K-nearest and Naïve Bayes support data with missing values. It might lead to building a biased machine learning model, which paves the way for incorrect results if the missing values are not handled properly. The complete dataset should be analyzed for missing values carefully. How to deal with the missing data without knowing the missing values? Two primary ways are found in handling the missing values. They are Deleting the missing values and Imputing the missing values. Each column must be analyzed to find the cause of missing those values, which is important to choose the strategy for handling the missing values.

Deleting the missing value

This is the quick and dirtiest technique that one could ever follow to deal with missing values. If the missing value is of the type Missing Not At Random(MNAR), then it should not be deleted. If that is of MAR or MCAR, it can be deleted after proper analysis, followed by pairwise deletion. Useful data might also get deleted in this process which is a disadvantage.This action can be performed in two ways. They are Deleting the entire row and Deleting the entire column.The entire row can be dropped if it is found to have many missing values.The entire column can also be dropped if it is found to have too much of missing values.

Imputing the missing value

Multiple imputation methods are found for replacing the missing values. Python libraries such as Pandas and Sci-kit Learn can be used for this purpose. The process involved in imputing missing values for categorical features Two ways can be followed to impute missing values for categorical features. They are: Impute the most frequent value and Impute the missing value. In both of the above approaches, data is needed to be encoded. Missing values can also be imputed using the sci-kit library by structuring a model to predict the observed value of a variable based on another variable which is known as regression imputation. The following are the two approaches.

Univariate approach

Only a single feature is taken into consideration. You can use the class SimpleImputer and replace the missing values with mean, mode, median, or some constant value.

Multivariate approach

More than one feature is taken into consideration. Two ways are used to impute missing values when it comes to the case of a multivariate approach. They are KNN imputer and Iterative imputer classes.

Can "missingness" be used as a feature?

While imputing the values, the information can be preserved and added as an additional feature. It is because there may also be a relationship between the missing values and the target variable. The figure-2 shows the sub-classification in the missing values.

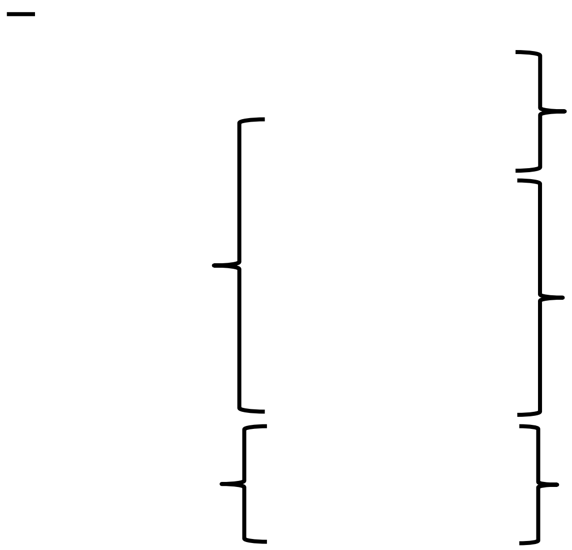

Figure-2. Classification Of The Missing Valu

Literature Survey

Missing data is a commonly occurring issue in various d
have greatly overlooked the topic to a great extent. The reliab
analysis is found to be reduced by the missing values. A series
action. But, their appropriateness depends on the patterns
missing patterns and distribution of missing values may b
insight into potential problems and the processes giv
Visualization has its unique way of supporting the inv
differences in the results show the performance of t
patterns. Recommendations for futuristic designs of m
frequently encountered while collecting data. Its pre
analyzing the data in the statistical power of the stud
crucial role in the estimation process and statistical
be dependent on how the missing values are proc
values and identifying outliers [2]. Data preproc
ensuring data quality. Data preprocessing tas
imputation, smothering out of noisy data, and r
be followed to impute the missing values a
expectation-maximization algorithm are used
horizontal partition of data and whole data
algorithm is better for bigger data sets. The
segments by correlating among the attribute
oncomparing their performances. Missing ra
of evaluation criteria such as coefficient of
root mean squared error(RMSE) is used. In
operations better than EMI [3].

Microarray gene expression data f
various technical causes. Numerous algor
missing values. But, this can affect the miss
of existing missing value imputation alg
techniques and perspective performance in
are expressed. The performance of various
based on futuristic research options availab

the paper. An upgraded version of the imputation algorithm is expected from the aspirants [4]. When it comes to the case of bio-informatics, the data required for the prediction and diagnosis of diseases such as Cancer plays a delicate role. Having missing values in that data creates a huge impact. This dismays many analysts. Missing value imputation algorithms should be developed to resolve this issue, which should be able to acquire the missed data. This paper presents the best imputation algorithms to find and retrieve the missing values. Much focus is given to analyzing the ability of those algorithms. Mentioned algorithms are categorized into various approaches such as global, local, hybrid, and knowledge-assisted approaches. The above methods are evaluated based on their performance.

The possibility of improving the algorithm's function is expressed in this paper [5]. In the biomedical research field, the collection of phenomic data always contains missing values. In bioinformatics, this issue is overcome through an imputation process that takes values from practical examples. SG. Liao et al. investigated all the imputation methods in practice and developed a self-training selection scheme that helped detect the best way of imputing data. They used the K-nearest neighbor for the selection process. The algorithms were divided into four methods, each concentrating on a specific parameter like a variable, subject, hybrid, and weighted hybrid. They also considered imputation methods like MICE and Miss Forest. They compared the proposed model with the MICE and missForest and found that it provided better imputation and improved the data quality [6]. The impact of known and unknown data on periodic biases is high in the case of Shotgun proteomics. Normalization is done when an attempt is made to remove periodic biases from the data. It is done before statistical inference, followed by missing value imputation. The above process is done to obtain a complete matrix of the intensities. A few approaches to normalization are also discussed, which is an added advantage to dealing with the missing values.

Some approaches are initially developed for micro-array data, and some are designed and developed specifically for mass spectrometry-based data. Yuliya V Karpievitch et al. have expressed it amazingly to impute the algorithms for the above-stated purpose[7]. Md. Shahjaman has developed new techniques to impute the algorithms to find the missing values. Missing values in data fall under the category of transcriptomics and metabolomics. It happens because of a lack of resources and the respective data acquisition techniques. Almost every statistical method requires complete datasets to perform downstream analysis. Several methods are developed for missing value imputation. The presence of outliers deteriorates the performance of these methods. Precise imputation is required to handle the outliers. The robust approach using robust estimators is developed in the paper. The performance of the proposed method is investigated with comparison. The above-suggested approach maintains the consistency of performance in the absence of outliers. The above method seems to be simple and accurate. The author recommends applying the above procedure for large-scale transcriptomics and metabolomics. The analysis of the above method is said to be implemented in the R package[8].

According to Nishith Kumar, the sophisticated study based on entire metabolites is Metabolomics. The study is known to be the connector of genotypes and phenotypes. Bio-marker identification from the study is found to be hampered. Although it is a high-dimensional data matrix, it contains missing values and outliers. The author has proposed a new bio-marker identification technique to identify the dataset's biomarkers. The performance of the above technique is compared with those of the conventional techniques. This method uses upregulated and downregulated metabolites as metabolomics biomarkers for HCC disease[9].The authors P. Arumugam and R. Saranya have come up with their ideas to simplify the process with accuracy. Analysis of the statistical data is required with utmost precision. Prediction of rainfall is quite a big task when the data collected is found to be missing or varying at the last minute. In this model, the method is found to suit the data well. Stochastic seasonal variation was found to be overcome by this method successfully. Based on the above study, this seasonal ARIMA model is suitable for rainfall forecasting [10].

Limitations and motivation

The missing value identification can be made through various methods. Widely machine learning algorithms are used for the prediction process. The dataset in all machine learning applicationsconsists of various missing values and outliers removed to improve the model's accuracy. Most datasets are divided into two halves, one for training and another for testing. Usually, the training model consists of data mainly concentrating on the ground truth, while the testing dataset is in the prediction process. While processing the dataset, one of the key things to be noticed is feature selection and extraction. Feature selection helps to reduce the computational complexity and reduces the size of the data. Feature selection can be made differently, like adopting machine learning algorithms, AI models, and statistical applications. Some of the widely used methods are RF, LR, FDR, MI, PCA, and ANOVA. The models are trained with these features maintained in the database. These ML models are called classifiers, like NB, LDA,QDA,SVM, LR,DT,RF, and ANN. These algorithms can extract features and fill in the missing values that need to be filled. Otherwise, the accuracy of the model is less.

Proposed

In research and analysis, missing values play an important role, as most of the dataset consists of missing values. Handling missing values is crucial, and the accuracy of the proposed model depends on the method's efficiency in handling the missing values. The missing values may arrive due to various causes: attrition, missing by design, and Item nonresponse. These are some of the widely seen reasons for missing values. Attrition can also be defined as the partial completion of data. This type of missing value occurs when the questionnaires must be completed appropriately. It may be due to the carelessness while answering the questions. It can be mostly seen in situations that involve questionnaires. It can be seen in longitudinal studies, interviews, and online surveys. This kind of missing values surely affects the performance of the model. Another type of missing value is the values that miss by design. It is the problem of the questioner and the type of questions they set. If the question does not apply to a candidate, then it can be termed a question that is missing by design. Otherwise, different questions are set for different people to show variety and variability. In such a case, the missingness mechanism evaluates such features and missing values. Though all the questions belong to everyone, it does not affect the model's performance.

Another problem is the nonresponse of the respondents to certain questions. The response is not expected if it deals with private items like salary, family, and personal traits. The Item nonresponse category can be classified into three types, they are not provided, useless, and lost. The non-provided is the decision of the candidate, and it may be due to the above-said reasons. At the same time, useless data can be found when the answer does not suit the research or analysis carried out with the data. The data lost can be due to corruption, hardware fault, missed processing, etc. The not provided data and useless data are from the candidate, while the data loss is due to the data collector. However, all these types belong to questionnaire-type data and are used in other applications. It applies to applications like IoT, WSN, Medical imaging, and others. If a sensor stops working, it is the partially collected data. If installed in areas out of the coverage, they are defined as missing by design. Suppose the sensors are affected due to wear and tear and weather conditions. Due to this, there is a data loss in the network and a problem with the collection mechanism.

Different mechanisms are used to track the data to improve the traceability of the model; some of the widely used modes are MCAR, MAR, and MNAR. 1) the MCAR represents the types of data that represent the missed data which does not depend on the observed and unobserved parameters. In the case of independent data, statistical methods are used to treat the data. 2) Missing At Random occurs when the missing value is dependent on the data observed but independent of the unobserved data. MAR can be widely seen in surveys like political surveys. MAR can be more realistic compared to the MCAR. 3) Missing Not At Random- This missing value mostly depends on the unobserved data. It is difficult to identify data that is noticed at random. Simple solutions cannot alter it. Identifying the reason for missing the value is important, and an alternative is included.

All these types of missing values are present in a single dataset, and the preprocessing of the data helps improve the model's accuracy. The filling of values on MAR and MNAR can be made through assumption, giving relatively better accuracy. However, it cannot be confirmed as the data is unavailable. Various techniques are adopted to handle the missing values by various data scientists and data enthusiasts. Some techniques are row or listwise deletion, column deletion, Labelled category, and IM.

Row-wise deletion done with the overall row does not contain all the data, and if that specific row consists of many missing values, the entire row is removed. The row in the dataset represents the individuals, and if the individuals' response is incorrect, then the entire data about the individual is removed. It also does not affect the accuracy of the model. The column deletion is different from the row deletion. It may be due to the wrong selection of variables. These variables are not filled if it consists of numerous missing values. It can affect the quality of the data and prediction accuracy. Hence, the overall column deletion helps in improving the quality of the data and improving accuracy. Certain missing values can be labeled and classify them as a specific category, which helps in further analysis and will give a real-time view of the data. It may increase the complexity of the data but provides a better view of the exact situation. Imputation is replacing a missing value with an average or a proper replacement value, which does not affect the data quality or prediction accuracy.

In the case of determining the source of the missing values, to replace these values, the fairness of the missing data is difficult to evaluate, and the analysis is very difficult. The decision made should be precise to improve the fairness of the data. The data's fairness depends on the attribute's quality and parity. Machine learning and deep learning algorithms are needed to make the proper decision.

Let us consider $\boldsymbol{X}$ to be the attributes, and its subsets are defined as $\boldsymbol{S}$, which refers to the protected attributes. This protected attribute is capable of classifying the data into different groups. The protected attribute is chosen randomly and represented as $\boldsymbol{S_i}$, and each value in the group is represented as $\boldsymbol{V_i}$. A group may contain more than one protected attribute. The attribute with a specific character can be termed privileged, which applies to all domains and applications. These metrics help identify missing values and outliers that do not merge with the rest of the data. Fairness metrics help evaluate the features and groups and find the privileged ones. Let the class attributed be defined as $\boldsymbol{Y}$, and it takes the value $\boldsymbol{c}$, and the possible and favorable outcome is represented as $\boldsymbol{C}^{+}$. The unfavorable and impossible outcome is represented as $\boldsymbol{C}^{-}$. While considering the unlabeled data in the dataset as $\boldsymbol{x}$, it is chosen from the subset $\boldsymbol{V_i}$. After identifying the missing value, an unlabeled instance $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$, is unlabelled, and$\boldsymbol{V_i}$ is taken as a tuple. The value $\boldsymbol{x}$ is mapped to$\boldsymbol{\hat{y}}$, $\boldsymbol{\hat{y}}$ is considered the answer, while the ground truth is considered $\boldsymbol{y}$. The datasets are classified into different groups based on the variables. Let us consider a dataset as $\boldsymbol{D}$, and the instance considered in the dataset as $\boldsymbol{D_{X_i=a}}$, and let a be the value selected, for instance $\boldsymbol{X_i}$. In this process, the unlabeled datasets are addressed, but in the same way, labeled datasets are also addressed. And $\boldsymbol{a \epsilon V_i}$And the labeled datasets are denoted as $\boldsymbol{D_{y=c^*}}$. The labeled datasets are denoted as positive examples, while the unlabelled ones are called negative ones. Based on the features in the dataset, the positive and negative values are compared, and the TP, TN, FP, and FN are chosen. The favorable outcome's probability in distribution $\boldsymbol{D}$ is $\boldsymbol{p(y = c^{+})}$, and $\boldsymbol{x \sim D}$ is represented as $\boldsymbol{p^{+}(D)}$. To replace a missing value with a duplicate one of the favorable class with an attribute $\boldsymbol{X_i}$, then,

$$\boldsymbol{SPD_i^{+}(D) = p^{+}\left(D_{X_i=a}\right) - p^{+}\left(D_{X_i!=a}\right)}$$

If the output of the above equation is equal to zero, then the variables considered are at equal terms, and less or more than 0 leads to privileged or unprivileged. The SPD will change based on the nature of the data, and the favorable and unfavorable classes will be swapped,

$$\boldsymbol{SPD_i^{-}(D) = p^{-}\left(D_{X_{i=a}}\right) - p^{-}\left(D_{X_{i!=a}}\right) = 1 - p^{+}\left(D_{X_{i=a}}\right) - \left(1 - p^{+}\left(D_{X_{i!=a}}\right)\right) = -SPD_i^{+}(D)}$$

Figure-3. The Architecture Of The Proposed Model

Figure-3 shows the architecture of the proposed model and provides a better view of how the missing values are removed from the datasets. The overall architecture for the missing value analysis involves processing the data and replacing the none and unprocessable values, which may affect the algorithm's performance.

Results and Discussion

Various techniques used in handling missed values and outliers are discussed in this work. A sample healthcare dataset is considered, various missing value-handling methods are experimented with, and the results are discussed in detail. The dataset and its features are analyzed, and the missing values and outliers are handled by considering the nature of the variable. An attribute-based imputation is followed to replace the missing values and is then experimented with various machine learning algorithms, and the results are verified. The proposed model is experimented with in Python in a Kaggle open-source environment, and the performance obtained is compared with similar methods.

Table-1. Model Dataset With Variable And Data Types

Feature	Variable Type	Variable	Value Type
Age	Objective Feature	age	int(days)
Height	Objective Feature	height	int(cm)
Weight	Objective Feature	weight	float(kg)
Gender	Objective Feature	gender	categorical code
Systolic blood pressure	Examination Feature	ap_hi	int
Diastole blood pressure	Examination Feature	ap_lo	int
Cholesterol	Examination Feature	cholesterol	1: normal, 2: above normal, 3: well above normal
Glucose	Examination Feature	gluc	1: normal, 2: above normal, 3: well above normal
Smoking	Subjective Feature	smoke	binary
yAlcohol intake	Subjective Feature	alco	binary
Physical activity	Subjective Feature	active	binary
Presence or absence of cardiovascular diease	Target Variable	cardio	binary

The table-1 shows the different features, variables, types, and data types that can be seen. Each feature is tagged with the type of variable and datatype. Objective, subjective, examination, and target variables are specified as it helps determine the methods to be adopted to fill the missing values. The subjective features depend on the patients, and the feature may vary depending on the patient type. For example, smoking, alcohol, and physical. The examination type features depend on the instrument used to measure the value. However, only after the examination is the range of the value or the possible range defined, and if any value that is more than that or within that range helps fill the missing value and remove the outliers. Likewise, various other features and their types help in the imputation of the data.

Table-2. Column-Wise Example Of The Dataset

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	2	18857	1	165	64.0	130	70	3	1	0	0	0	1
3	3	17623	2	169	82.0	150	100	1	1	0	0	1	1
4	4	17474	1	156	56.0	100	60	1	1	0	0	0	0

The table-2 shows the example data for all the features, and from the dataset, it can be seen that each feature has different values. There it can be seen that there are zeros and ones, representing the binary type of variable that helps in faster and more efficient processing. In most of the cases, they may be integer or a binary type of feature. This helps in the processing, and this helps to calculate the missing value. In terms of missing values, it helps in faster replacement and does not affect the model's performance.

Table-3. Preprocessing and checking the dataset

RangeIndex: 70000 entries, 0 to 69999 Data columns (total 13 columns):			
id	70000	not-null	int64
age	70000	not-null	int64
gender	70000	not-null	int64
height	70000	not-null	int64
weight	70000	not-null	float64
ap_hi	70000	not-null	int64
ap_lo	70000	not-null	int64
cholesterol	70000	not-null	int64
gluc	70000	not-null	int64
smoke	70000	not-null	int64
alco	70000	not-null	int64
active	70000	not-null	int64
cardio	70000	not-null	int64
dtypes: float64(1), int64(12) memory usage: 6.9 MB			

After processing and subjecting to the missing value methods shown in table-3, the total dataset consists of missing values, and it is seen that they carry different types of datasets. However, most of the features are int64, and a few of the features with float64. The int64 values may increase the computational time. However, they are difficult to be replaced or impute, as the range will be higher and are mostly subjective and examination type of features.

Table-4. Statistics After Processing The Dataset

	id	age	gender	height	weight	cholesterol
count	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000
mean	49972.419900	19468.865814	1.349571	164.359229	74.205690	1.366871
std	28851.302323	2467.251667	0.476838	8.210126	14.395757	0.680250
min	0.000000	10798.000000	1.000000	55.000000	10.000000	1.000000
25%	25006.750000	17664.000000	1.000000	159.000000	65.000000	1.000000
50%	50001.500000	19703.000000	1.000000	165.000000	72.000000	1.000000
75%	74889.250000	21327.000000	2.000000	170.000000	82.000000	2.000000
max	99999.000000	23713.000000	2.000000	250.000000	200.000000	3.000000

Table-4 shows the statistics of the important features in the dataset, and it gives a better view of the type of data and the range within which the missing value or outlier may fall. It also shows the maximum, minimum, and standard links. The mean value and count of the datasets can find the average value.

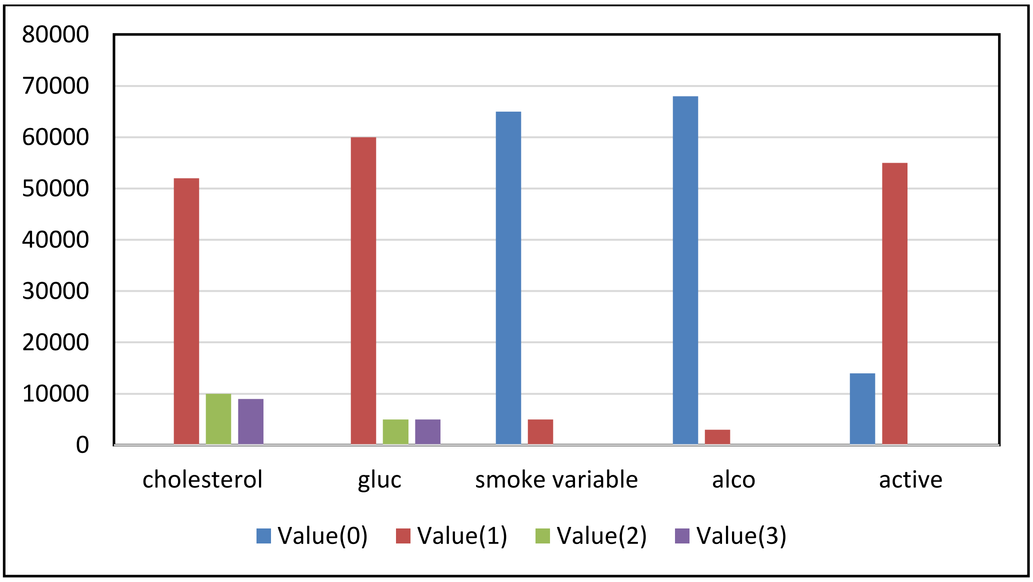

Figure-4. Comparing The Cardio Of The Patients In A Univariate Analysis

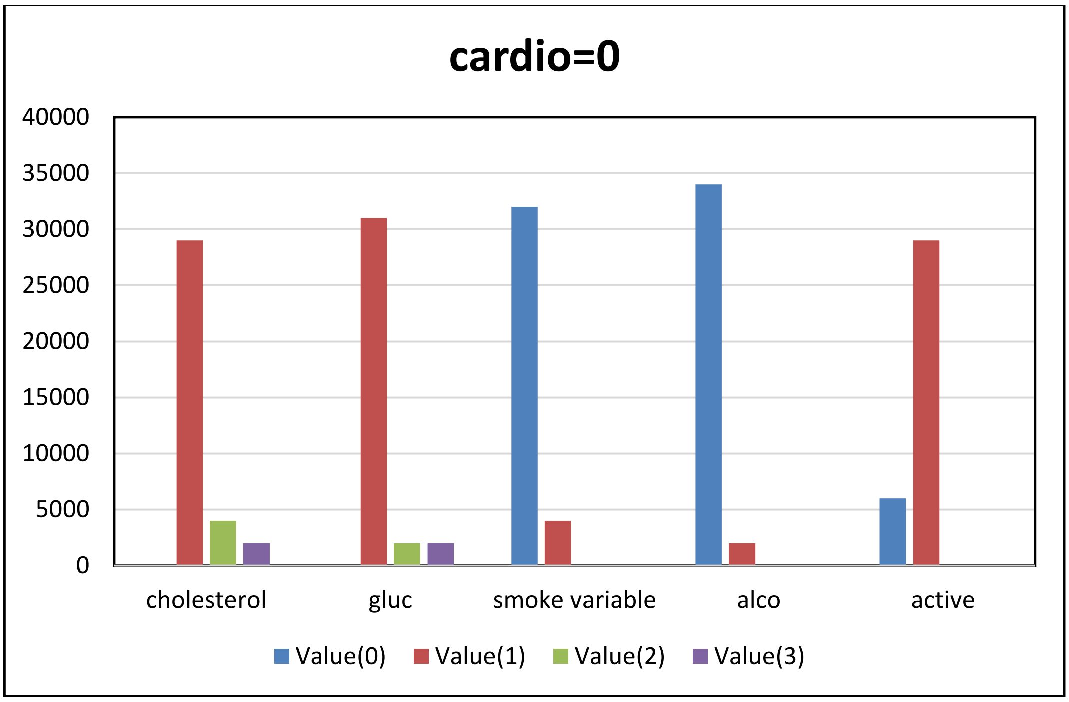

Figure-5. Comparing The Value Of The Various Parameters In A Univariate Analysis

Bivariate analysis

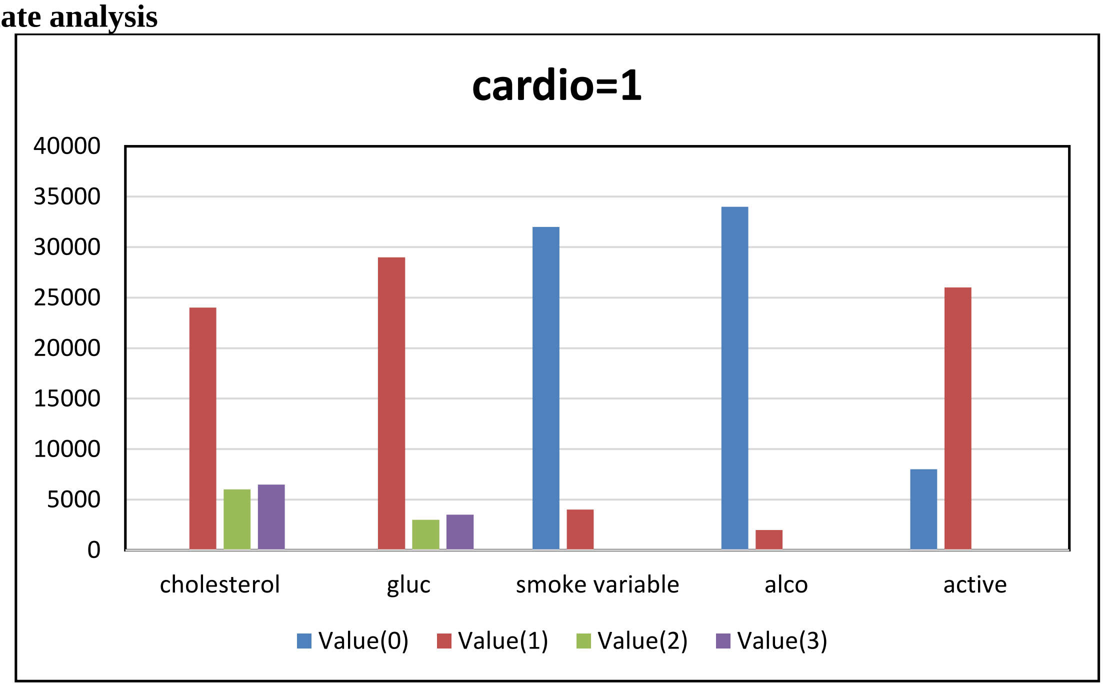

Figure-6. Comparing The Cardio With Different Parameters In A Bivariate Analysis

In Figure-4, the age of the patients is compared, and it can be seen that with the increase in the number of age, the cardio diseases are increasing, and with younger people, the heart and cardiac diseases are very less. It also shows that these distributions are random and, in most cases, discrete. However, it is simple to understand that with minimum age, cardiac diseases are very less. Figure-5 shows the important parameters and the comparison between different values and the count of the persons. From these comparisons, it can be seen that most cardiac arrested are from cholesterol, glucose, smoking, and alcohol. Those actively practicing all these habits and health conditions led to increased subjection to cardiac diseases.Figure-6shows the bivariate comparison of the cardiac patients, and the graphs show the comparison of the important parameters. The bivariate analysis can see in the figures, which provided a better view of cardiac patients.

Conclusion

The proposed missing value analysis technique is implemented with a healthcare dataset, and the obtained result is discussed in detail. The missing values and outliers in the dataset are analyzed and removed. The different models proposed for finding missing values and outliers have experimented with the results obtained compared with the existing ones. The proposed missing value analysis methods suit datasets with varied features and parameters. The dataset for the experiment consists of age, weight, height, cholesterol, etc., which is used to analyze the cardiac risks in the patients. Each field has its own nature of addressing the missing values, like the patient's age may not exceed a certain limit, and the weight cannot be off the charts. Each parameter and its relations need to be analyzed and combined, which is carried out through the proposed model. The dataset cleaned with different missing value imputation methods is fed as input to different machine learning algorithms, and then the accuracy value of each ML algorithm is compared. The efficiency of the data imputation and the quality of the datasets are evaluated. From the proposed model, it is clear that the dataset's quality can be improved by using data imputation techniques, missing value analysis methods, and the accuracy of the prediction model.

Reference

1. Fernstad, S. J. (2019). To identify what is not there: A definition of missingness patterns and evaluation of missing value visualization. *Information Visualization*, *18*(2), 230-250.
2. Kwak, S. K., & Kim, J. H. (2017). Statistical data preparation: management of missing values and outliers. *Korean Journal of Anesthesiology*, *70*(4), 407-411.
3. Rahman, M. G., & Islam, M. Z. (2011). A decision tree-based missing value imputation technique for data preprocessing. In *The 9th Australasian Data Mining Conference: AusDM 2011* (pp. 41-50). Australian Computer Society Inc.
4. Liew, A. W. C., Law, N. F., & Yan, H. (2011). Missing value imputation for gene expression data: computational techniques to recover missing data from available information. *Briefings in Bioinformatics*, *12*(5), 498-513.
5. Moorthy, K., Saberi Mohamad, M., & Deris, S. (2014). A review of missing value imputation algorithms for microarray gene expression data. *Current Bioinformatics*, *9*(1), 18-22.
6. Liao, S. G., Lin, Y., Kang, D. D., Chandra, D., Bon, J., Kaminski, N., ... & Tseng, G. C. (2014). Missing value imputation in high-dimensional phenomic data: imputable or not, and how?. *BMC Bioinformatics*, *15*(1), 1-12.
7. Karpievitch, Y. V., Dabney, A. R., & Smith, R. D. (2012). Normalization and missing value imputation for label-free LC-MS analysis. *BMC Bioinformatics*, *13*, 1-9.
8. Shahjaman, M., Rahman, M. R., Islam, T., Auwul, M. R., Moni, M. A., & Mollah, M. N. H. (2021). rMisbeta: A robust missing value imputation approach in transcriptomics and metabolomics data. *Computers in Biology and Medicine*, *138*, 104911.
9. Kumar, N., Hoque, M., Shahjaman, M., Islam, S. M., Mollah, M., & Haque, N. (2017). Metabolomic biomarker identification in the presence of outliers and missing values. *BioMed Research International*, *2017*.
10. Arumugam, P., & Saranya, R. (2018). Outlier detection and missing value in seasonal ARIMA model using rainfall data. *Materials Today: Proceedings*, *5*(1), 1791-1799.