Partition-Based Technique to Enhance Missing Data Prediction

Mohammad Mahdi Barati Jozan[1], Hamed Tabesh[2]
[1,2] Department of Medical Informatics, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran

Article Info

Article history:

Received Jan 8, 2023
Revised Mar 10, 2024
Accepted Mar 20, 2024

Keyword:

Preprocessing
Missing value imputation
Text Mining
Expert panle

ABSTRACT

Managing missing data is a critical aspect of preprocessing in data mining endeavors, significantly influencing output accuracy during both model development and utilization phases. This study introduces a novel approach to predicting missing values by partitioning data into disjoint subsets based on partitioning measures. The rationale behind this approach is the elimination of unrelated data through partitioning, thereby improving the accuracy of missing value prediction within each subset. Through a combination of expert panel insights and statistical tests (including the Chi-square test and Cramer's V coefficient), the database partitioning measure was determined using operational data from the Mashhad Fire and Safe Services Organization. Models were constructed for each partition, and missing data were segmented accordingly, with the corresponding models employed for prediction. The results revealed that in 44% of cases, models built on partitioned data outperformed those constructed on the entire dataset. The evaluation of this method underscores its capability to predict missing values with heightened accuracy. Notably, this approach is independent of the method employed for missing value prediction, enabling seamless integration into existing methods as an additional step to bolster prediction accuracy.

Corresponding Author:

Mohammad Mahdi Barati Jozan,
Department of Medical Informatics, Faculty of Medicine,
Mashhad University of Medical Sciences,
Medical School, Pardis Daneshghah, Azadi Sq. Mashhad,Khorasan Razavi, Iran
Email: mondheparag@gmail.com

1. INTRODUCTION

Addressing missing values stands as one of the primary challenges in both data mining and machine learning. The manipulating of missing values emerges as a crucial step in data preprocessing for both research in data mining and real-world projects [1,2]. Causes for missing data encompass loss of data, incomplete user data entries, improper data recording, and measurement errors stemming from instruments [3]. Typically, unless data is meticulously and rigorously recorded, around 5% or more of the database may contain missing values [4].

The presence of missing values can pose challenges both during the model creation phase and when utilizing the model [1, 5-7]. The completeness of the data is directly linked to data quality [8] and the accuracy of models built upon said data [1, 9]. In high-risk endeavors such as designing decision support systems in medicine, the existence of missing values undermines the reliability of constructed models [10-12]. Consequently, numerous studies have been conducted to address missing data across various fields, resulting in the proposal of various techniques for this purpose.

The techniques proposed for handling missing values range from very simple to complex methods. One of the simplest techniques involves removing records that contain missing values. Despite its simplicity, this method yields numerous negative consequences. One obvious ramification is the reduction in available data for model construction [13], rendering it inefficient, particularly in areas with high instances of missing data, such as Electronic Health Records (EHR) [14-17]. For instance, in a dataset comprising patients with pancreatic cancer from a New York academic medical center, 48.9% of pathology reports documenting the disease contain missing data [18]. In databases with significant levels of missing values, the removal of such entries drastically diminishes model accuracy, potentially resulting in adverse outcomes when employing these models for decision-making purposes [14, 19]. Another detrimental consequence of removing missing data is the potential bias it introduces to the model [13, 20]. This issue arises because the elimination of missing values may alter the original data distribution [12], thereby reducing statistical power [12, 21]. Techniques such as removing missing values that solely utilize complete records are referred to as complete case (CC) analysis [22]. When the number of missing data points is minimal and their removal doesn't affect the data distribution, omitting some records to maintain data integrity can be deemed acceptable [23, 24].

Overall, imputation techniques for handling missing values are considered superior to complete case analysis in terms of efficiency and validity [25-27]. These techniques can be broadly categorized into two main groups: statistical techniques and machine learning techniques [28]. In statistical techniques, various statistical measures are employed to address missing values. Examples include replacing missing values with a unique value, such as zero, or with the mean, median, or mode of the available data. Additionally, treating missing values as a distinct category is another approach within statistical techniques. While predicting missing values using statistical techniques can be effective in certain scenarios, they are generally less efficient compared to complete case analysis [5, 29, 30].

Machine learning imputation techniques have been introduced to utilize values and patterns in other data to predict missing values [10], rendering them superior in performance and accuracy. Regression techniques stand out as one of the most widely adopted approaches within this domain [31]. Linear regression is commonly employed for numerical variables [32], while logistic regression is suitable for categorical variables [33], enabling the prediction of missing values. Regression, as a statistical method, aims to establish relationships between a target variable (dependent variable) and other variables (independent variables). In imputation, the variable with missing values is treated as the dependent variable, while the other variables serve as independent variables. The regression model is built based on records devoid of missing values and is utilized to estimate values in records containing missing values.

The Expectation Maximization (EM) technique presents another common approach for estimating numerical missing data [3, 34]. This technique employs a maximum likelihood approach involving iterative expectation and maximization steps to estimate parameters of the model [35]. K-nearest neighbor (KNN) imputation leverages the values of neighboring data points to predict missing values [36, 37]. In this technique, for each record with missing values, the k nearest neighbors without missing values are identified, and the missing value is estimated based on these neighbors. For instance, if the missing value is numeric, it can be replaced by the mean value of the neighbors [38]. In [39], the authors propose a modified version of KNN where a weight is assigned to each neighbor based on the Euclidean distance between the instance with missing values and its nearest neighbors. The closer the distance, the higher the weight assigned to the neighbor, and vice versa. This method, known as weighted k-nearest neighbor imputation (WKNNI), outperforms the original KNN.

Overall, the approach of estimating missing values based on nearby neighbors is being explored by many researchers, with a focus on the efficiency of near neighbors over distant ones [10]. In [10], the authors utilize the Fuzzy C-Means (FCM) clustering algorithm [40] to select records closest to those with missing values, demonstrating its effectiveness in imputing missing values.

The current methods for predicting missing values based on neighborhoods often utilize all available data. For instance, techniques like K-nearest neighbor (KNN) [36, 37] and weighted K-nearest neighbor imputation (WKNNI) [39] typically consider all data points when selecting nearest neighbors or calculating mean values. An alternative approach that holds promise for enhancing results involves partitioning the data into disjoint partitions and independently predicting missing values within each partition. This strategy operates under the assumption that segmenting the data effectively filters out irrelevant information, thereby enabling more accurate predictions for missing values within each partition.

In this study, our objectives are twofold:

- To investigate whether partitioning the data into disjoint partitions enhances the accuracy of missing value predictions.
- To determine the optimal method for partitioning the data.

2. RESEARCH METHOD

In most of the proposed techniques for predicting missing values based on neighborhoods, all data are utilized for model building. However, in the proposed approach, the data is divided into disjoint partitions, and patterns are extracted for each partition. This partitioning strategy is premised on the belief that dividing data into smaller disjoint datasets ensures that related data points are grouped together within each partition, thereby enhancing the accuracy of the models. Following the construction of a model for each partition, the corresponding partition containing missing data is identified, and the model created for that specific partition is utilized to predict the missing values. The proposed method falls within the category of machine learning imputation techniques for predicting missing data.

The flowchart illustrating the model creation and utilization phases of the proposed approach is depicted in Figure 1 and Figure 2.

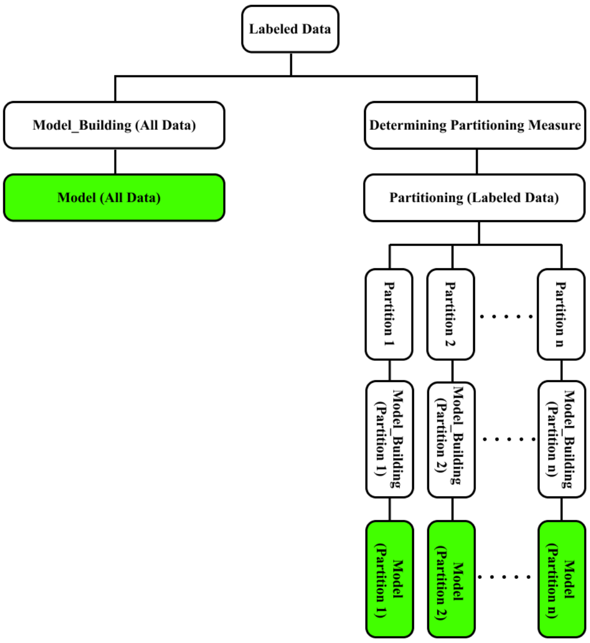

Labeled Data
Model_Building (All Data)
Determining Partitioning Measure
Model (All Data)
Partitioning (Labeled Data)
Partition 1
Partition 2
Partition n
Model_Building (Partition 1)
Model_Building (Partition 2)
Model_Building (Partition n)
Model (Partition 1)
Model (Partition 2)
Model (Partition n)

Figure 1. Model creation phase

As depicted in Figure 1, to create the model, a model was constructed using the entirety of the available labeled data. Simultaneously, employing a partitioning measure, the dataset was divided into disjoint datasets, each possessing unique characteristics. Finally, individualized models were developed for each partition.

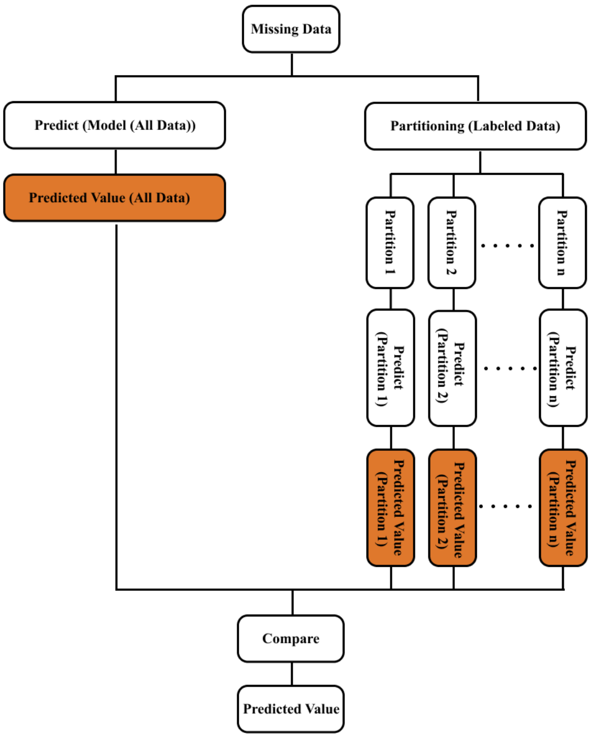

Missing Data
Predict (Model (All Data))
Partitioning (Labeled Data)
Predicted Value (All Data)
Partition 1
Partition 2
Partition n
Predict (Partition 1)
Predict (Partition 2)
Predict (Partition n)
Predicted Value (Partition 1)
Predicted Value (Partition 2)
Predicted Value (Partition n)
Compare
Predicted Value

Figure 2. Model utilization phase

As shown in Figure 2, missing values were predicted using a model trained on the entirety of available labeled data. Simultaneously, employing a partitioning measure, the missing data were stratified into disjoint datasets. Within each partition, a bespoke model was deployed to predict the missing values. To ensure optimal model selection, predicted values from both the overarching model trained on all data and the models for disjoint datasets were compared. Ultimately, the model demonstrating superior predictive performance was chosen for imputing missing values.

The critical aspect of the proposed method lies in determining the partitioning measure. The partitioning measure consists of one or more fields based on which the database is divided into disjoint partitions. To establish this measure, the following three principles must be taken into account:

- **Power of the measure:** The measure can be determined through various methods. In this study, both qualitative and quantitative methods have been explored to establish the criteria.
- **Accuracy of the measure:** All methods and data utilized to establish the measure must exhibit high accuracy.
- **Comprehensiveness of the measure:** The measure should be applicable to datasets containing missing values. In other words, missing data should not be integral to determining the measure.

2.1. Partitioning measure and Comparing method

The study utilized the expert panel technique as a qualitative method to select the partitioning measure. Panel members were chosen based on their expertise and experience in relevant domains, ensuring a balanced representation of skills and a wide range of expertise to thoroughly discuss and address all pertinent issues [41]. Employing a structured process, the expert panel deliberated on the measures, evaluating them for importance, utility, and potential impact on the disjoint partitions. Through consensus, the panel identified the most crucial measures, prioritizing them for further development and refinement.

Statistical tests, as a quantitative method, will be used to determine the optimal partitioning measure for creating disjoint partitions. The Figure 3 serves as a guide for selecting the appropriate statistical test to determine the partitioning measure. The selection of the best test is based on two indicators of the data type and the number of distinct values in the partitioning measure.

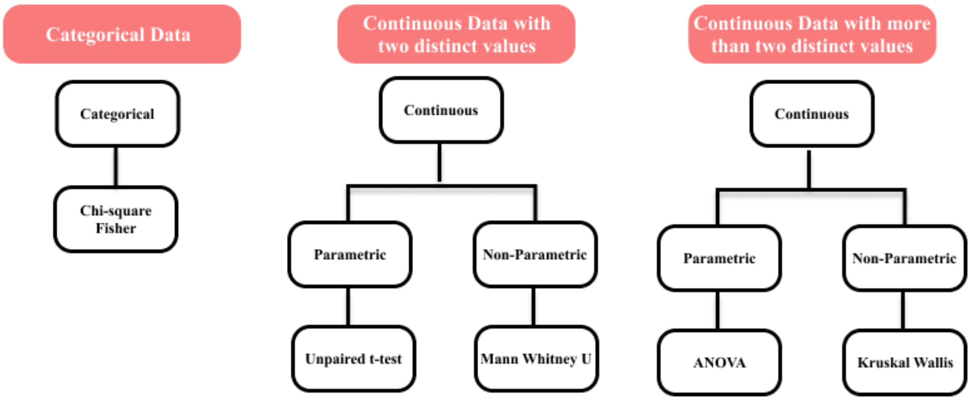


Figure 3. Classification of statistical tests to compare partition measure.

To assess the efficacy of the generated models, we employed the accuracy [42] that is calculated as follows:

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

where True positives (TP) are instances where the model accurately predicts a true positive outcome and false positives (FP) are instances where the model incorrectly predicts a true positive outcome. Similarly, True negatives (TN) are instances where the model correctly predicts ax true negative outcome and False negatives (FN) correspond to instances where the model incorrectly predicts a true negative outcome.

3. RESULT

In this section, we will first introduce the database used in the study. Subsequently, we will explain the details of both qualitative and quantitative methods employed to divide the data into disjoint partitions. Finally, we will report the results of the proposed method at the end of this section.

3.1. Dataset

Mashhad is the capital of Razavi-Khorasan province, which is the second most populous city in Iran [43]. It is divided into four firefighting operational areas, with 49 fire stations distributed throughout the city. According to the latest statistics, there are currently 1,072 firefighters and 195 fire engines in Mashhad. The Mashhad Fire and Safety Services Organization store approximately 50 data fields for each performed operation. The list of the most important of these fields is provided in Table 1.

Table 1. List of the most important fields store for each operation

Variable Name	Variable Description
Operation Type	Type of Operation
Malicious Reporting	Whether the call is a fake report or not.
Announcement Time	Time to send the fire brigade to the incident location
Shift Work	The shift during which operations are performed
Fire Station	The fire station that the fire brigade was sent from there to service.
Incident Location	Address of reported incidents
Phone Number	Phone number of the reporter
Last Name	Last name of the reporter
Redevelopment Code	Identification number assigned to each property in the urban structure
Municipal District	Municipal district of the incident location
Arrival Time	The time of the operational team's arrival at the incident location
Longitude	Longitude of the incident location
Latitude	Latitude of the incident location
Distance	The distance covered by the fire engine to reach the scene
Completion Time	The operation completion time
Return Time	The time of returning to the station
Operational team's arrival at the station	The time of the operational team's arrival at the station
Operational Code (a three-level hierarchical structure)	The first level consists of 9 sectors representing the type of incident
	The second level shows the severity of the incident
	The third level determines the type of operation
Type of incident site (a two-level hierarchical structure)	The first level of incident location
	The second level of incident location
First source of ignition	The material that initially caused the fire (used in fire operations)
Primary combustible material	The main material causing the fire (used in fire operations)
Heat Source	The source of heat generation in the fire (used in fire operations)
Intentional/unintentional	Whether it was intentional or not (used in fire operations)
Description	The important points of the operation recorded in the system by the commander
Additional operation information	Based on the type of service, some additional information is reported e.g. the source of fire
Date and time of operation completion	Date and time of operation completion reported by the commander

Other fields include the number of firefighters involved in the operation, the number of fire engines used in the operation, whether the police were involved in the operation, number of civilians/firefighters died in the operation, number of civilians/firefighters injured in the operation, etc., are also stored.

The Description and the Operational Code field are two important fields stored for each operation. The Description field is a free text field filled by the operation commander. In addition to describing how the operation is performed, the commander must report important points that occurred in each operation in this field.

Another crucial field in the database is the Operational Code. The code of each operation is determined by the operation commander based on a coding system. The coding system was a self-developed one by a team of experienced operation commanders and used from March 2015 to September 2018. Due to emerging needs in the organization, it was necessary to change the coding system. Finally, the organization developed a new personalized coding system based on the National Fire Protection Association (NFPA) coding system. The self-developed coding system had a linear structure, but the new coding system has a three-level hierarchical structure as shown in Figure 4.

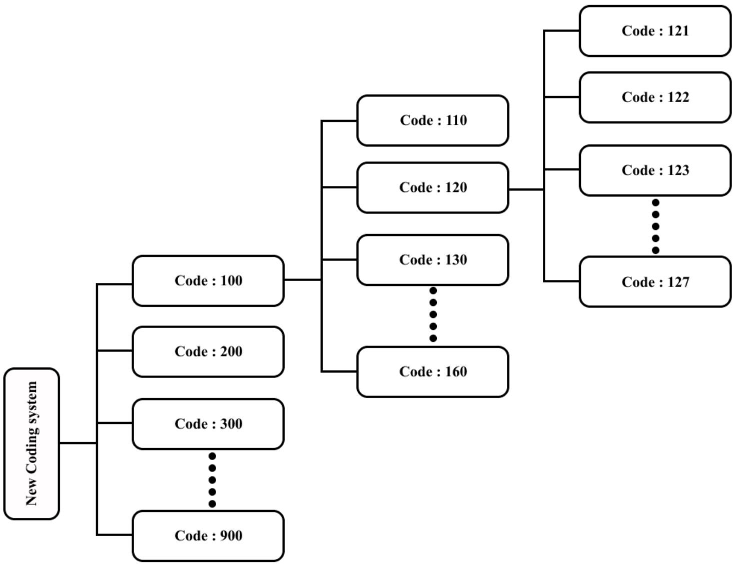
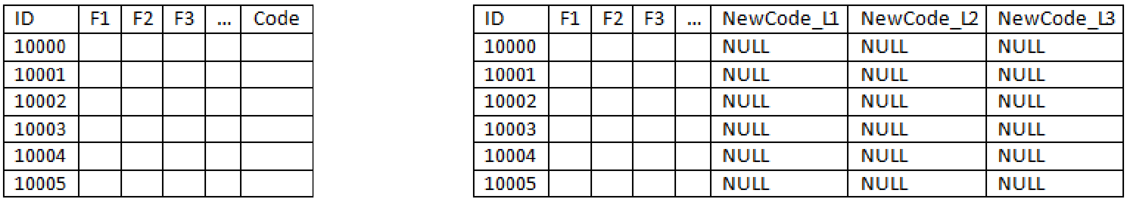


Figure 4. The three-level hierarchical structure of new coding system

3.2. Problem

As mentioned, from March 2015 to September 2018, the code system structure was one-level. During this period, 53,411 operations were performed. After developing a three-level coding system, it replaced the previous one. Therefore, the NULL value was assigned to the new operation code field for the 53,411 operations that took place from March 2015 to September 2018. Because the new coding system was three-level, three missing values appeared for each operation in the database, as shown in Figure 5.



ID	F1	F2	F3	...	Code
10000					
10001					
10002					
10003					
10004					
10005					

ID	F1	F2	F3	...	NewCode_L1	NewCode_L2	NewCode_L3
10000					NULL	NULL	NULL
10001					NULL	NULL	NULL
10002					NULL	NULL	NULL
10003					NULL	NULL	NULL
10004					NULL	NULL	NULL
10005					NULL	NULL	NULL

Figure 5. Data structure before and after using the new coding system

To determine the new code for the operations, three solutions were proposed:

- **Solution 1:** Building a map between the old and new coding systems involved holding several meetings with experts. Despite efforts, no acceptable mapping was achieved, making this solution practically impossible.
- **Solution 2:** Given that information about the operation is stored in the Description field, which is the only field that can be used to determine the new code, operation commanders would need to read this field for the 53,411 operations and determine the new code. However, this solution is very costly and time-consuming, making it practically impossible.
- **Solution 3:** Using a machine learning approach to predict missing values based on labeled data. Building a classifier using the Description field based on operations coded with the new coding system, and then predicting the missing values using the classifier for the operations performed before September 2018.

Ultimately, solution 3 was selected as the only one compatible with the organization's conditions. The managers and decision makers of the organization determined the accuracy rates of 95%, 90%, and 85% as the minimum accuracies for determining the missing values for NewCode_L1, NewCode_L2, and NewCode_L3 fields, respectively.

3.3. Create Model

Given that the Description field is the sole source for determining missing values, a common approach is to construct a text classification model based on operations coded with the new system. The missing values are then predicted using this model. In this process, the data without missing values (labeled data) is divided into training and test sets. The model is built on the training data, and its accuracy is assessed using the test data. If the model's accuracy meets predefined acceptable accuracies, it is used to predict missing values.

As outlined in the introduction section, the field containing missing values is referred to as the dependent variable. In classifier construction, the goal is to predict the dependent variable based on other variables without missing values, known as independent variables. In this study, there are three dependent variables: NewCode_L1, NewCode_L2, and NewCode_L3 (Operational Code). Therefore, three separate models are required. The sole independent variable is the Description field, which contains free text.

Before model construction, preprocessing operations were conducted on the Description field. Key operations included removing stop words, punctuation, and bad characters [44]. The text in the Description field was then converted into a matrix of token counts, which was further transformed into a normalized TF/TF-IDF representation [45]. Subsequently, several classifiers from the Scikit-Learn Python library [46] were trained, including the Naive Bayes classifier, Linear Support Vector Machine classifier [47], and Logistic Regression classifier [47]. Default parameter values for each algorithm were utilized in this research.

For classifier training, operations recorded with the new coding system (labeled data) were employed. A total of 87,119 operations were registered with the new coding system from September 2018 to August 2021. Seventy percent of this data was allocated as training data, while the remaining 30% served as test data. The accuracies of the classifiers at each level are presented in Table 2.

Table 2. Accuracy of classifiers in each of the levels

Classifier	Accuracy Level 1	Accuracy Level 2	Accuracy Level 3
Naive Bayes	0.97	0.84	0.76
Linear Support Vector Machine	0.98	0.90	0.84
Logistic Regression	0.96	0.87	0.81

As demonstrated in Table 2, the Linear Support Vector Machine classifier exhibits the highest accuracy. The accuracy achieved for NewCode_L1 and NewCode_L2 surpassed the predetermined threshold, meeting the standards of the Mashhad Fire and Safety Services Organization. However, addressing the challenge with NewCode_L3 remains imperative.

3.4. Determining the partitioning measure

As mentioned in the previous section, two methods were utilized to determine the optimal measure for data separation: 1) an expert panel (Qualitative method) and 2) the use of statistical tests (Quantitative method).

In the first method, five experts from the Mashhad Fire and Safety Services organization participated in an expert panel to determine the measure for database splitting. Demographic and job information of experts is shown in Table 3.

Table 3. Demographic and job information of experts in the Delphi method

Expert	**Position**	**Gender**	**Academic degree**	**Age**	**Work Experience**
$Expert_1$	Director of Education Department	Male	Doctorate	41	15 years
$Expert_2$	Director of Industrial Education	Male	Master	52	24 years
$Expert_3$	Firefighter/industry training instructor	Male	Bachelor	33	10 years
$Expert_4$	Firefighter/industry training instructor	Male	Doctorate	36	12 years
$Expert_5$	Firefighter/industry training instructor	Male	Master	46	23 years
	Average			42	17 years

Eventually, Operation Type was selected as the measure. The available values for this field are Fire_Operation, Accident_Operation, and Prevention_Operation. Experts believe that there is a significant relationship between the coding system (Operational Code) and the type of operation (Operation Type). Therefore, partitioning the database based on Operation Type can segment it into three smaller databases, allowing for the creation of a more accurate classifier for each category.

To assess this perspective, operations from September 2018 to August 2021 were divided into three partitions based on Operation Type. The number of operations in each partition is shown in Table 4. A classifier was developed for each partition, with 70% of the data utilized as training data and 30% as test data. The accuracy obtained from the split database is presented in Table 5.

Table 4. Number of operations in each Operation Type

Type	Number of operations	Percentage of total
Fire_Operation	34,492	39.59%
Accident_Operation	46,783	53.70%
Prevention_Operation	5,844	6.71%

Table 5. The accuracy of classifiers for each Operation Type

Classifier	Accuracy Level 1	Accuracy Level 2	Accuracy Level 3
All_Operation	0.98	0.90	0.84
Fire_Operation	0.99	0.87	0.81
Accident_Operation	0.98	0.97	0.94
Prevention_Operation	0.94	0.85	0.80

To determine the best classifier, if the accuracy of the classifier built based on all the data, known as the All_Operation classifier, is superior to the classifier built for each of the partitions, then the All_Operation classifier is employed; otherwise, the classifier corresponding to that partition is utilized.Table 6 outlines the final classifier for predicting the missing values of each partition. Cells marked in green indicate that the accuracy of the classifier built for that partition surpasses that of the All_Operation classifier.

Table 6. Selected classifier for each partition

Database	Accuracy Level 1	Accuracy Level 2	Accuracy Level 3
Fire_Operation	Fire_Operation	All_Operation	All_Operation
Accident_Operation	Accident_Operation	Accident_Operation	Accident_Operation
Prevention_Operation	All_Operation	All_Operation	All_Operation

Building a model based on disjoint databases has notably enhanced the results, particularly in the Accident_Operation partition. In the Accident_Operation partition, the accuracy of the classifier at Level 1, Level 2, and Level 3 has improved from 0.98%, 0.91%, and 0.87% to 0.98%, 0.97%, and 0.94%, respectively.

According to experts, there is no significant relationship between the coding system (Operational Code) and other fields stored in the database. In other words, partitioning the database based on other independent variables will not yield improved results.

As experts may not always be available, we propose another method based on statistical tests (Quantitative method). For this purpose, statistical tests were utilized to investigate the relationship between independent variables and the dependent variable (Operational Code). The fields selected for this analysis include MunicipalDistrict, ShiftWork, FireStation, SiteType, and Operation Type. The relationship between each of these variables and the Operational Code field was examined separately using the Chi-square test. The Chi-square test is a hypothesis test used to assess the statistical significance of the relationship between two categorical variables, determining whether they are independent of each other. The null hypothesis (H_0) and the alternative hypothesis (H_1) for each of the independent variables and the dependent variable are as follows:

Operation Type Field:

H_0 (Operation Type): There is no relationship between Operation Type and Operational Code in operations.

H_1 (Operation Type): There is a relationship between Operation Type and Operational Code in operations.

MunicipalDistrict Field:

H_0 (Municipal District): There is no relationship between Municipal District and Operational Code in operations.

H_1 (Municipal District): There is a relationship between Municipal District and Operational Code in operations.

Shift Work Field:
H_0 (Shift Work): There is no relationship between Shift Work and Operational Code in operations.
H_1 (Shift Work): There is a relationship between Shift Work and Operational Code in operations.

Fire Station Field:
H_0 (Fire Station): There is no relationship between Fire Station and Operational Code in operations.
H_1 (Fire Station): There is a relationship between Fire Station and Operational Code in operations.

Site Type Field:
H_0 (Site Type): There is no relationship between SiteType and Operational Code in operations.
H_1 (Site Type): There is a relationship between Site Type and Operational Code in operations.

First, the relationship between the two variables Operation Type and Operational Code is examined. All statistical tests were performed using SPSS software ver 22.0 [49].

A chi-square test of independence was performed to examine the relation between Operational Code and Operation Type. The relation between these variables was significant, X^2 (6, N = 87119) = 142701.636.146, p < .001 (Table 7).

Table 7. The result of Chi-square test to examine the relationship between Operational Code and Operation Type

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	142701.636[a]	6	.000
Likelihood Ratio	136349.372	6	.000
Linear-by-Linear Association	2359.406	1	.000
N of Valid Cases	86845		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 524.95.

A chi-square test of independence was performed to examine the relation between Operational Code and Shift Work. The relation between these variables was significant, X^2 (6, N = 87119) = 13.929, p = .030 (Table 8).

Table 8. The result of Chi-square test to examine the relationship between Operational Code and Shift Work

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	13.929[a]	6	.030
Likelihood Ratio	13.904	6	.031
Linear-by-Linear Association	.114	1	.735
N of Valid Cases	86845		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 1591.25.

A chi-square test of independence was performed to examine the relation between Operational Code and Municipal District. The relation between these variables was significant, X^2 (39, N = 87119) = 5380.536, p < .001 (Table 9).

Table 9. The result of Chi-square test to examine the relationship between Operational Code and Municipal District

Chi-Square Tests

	Value	Df	Asymp. Sig. (2-sided)
Pearson Chi-Square	5380.536[a]	39	.000
Likelihood Ratio	5540.383	39	.000
Linear-by-Linear Association	8.574	1	.003
N of Valid Cases	86845		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 126.91.

A chi-square test of independence was performed to examine the relation between Operational Code and Site Type. The relation between these variables was significant, X^2 (24, N = 87119) = 50830.404, p < .001 (Table 10).

Table 10. The result of Chi-square test to examine the relationship between Operational Code and Site Type

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	50830.404[a]	24	.000
Likelihood Ratio	47679.284	24	.000
Linear-by-Linear Association	11413.296	1	.000
N of Valid Cases	86845		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 62.23.

A chi-square test of independence was performed to examine the relation between Operational Code and Fire Station. The relation between these variables was significant, X^2 (156, N = 87119) = 9210.949, p < .001 (Table 11).

Table 11. The result of Chi-square test to examine the relationship between Operational Code and Fire Station

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	9210.949[a]	156	.000
Likelihood Ratio	9604.003	156	.000
Linear-by-Linear Association	134.883	1	.000
N of Valid Cases	86845		

a. 3 cells (1.4%) have expected count less than 5. The minimum expected count is 2.63.

Since the Chi-square test is sensitive to the number of observations (operations), and considering the large number of operations, even the slightest difference in the data can lead to a significant p-value. Therefore, we anticipated significant p-values when examining the relationship between each independent variable and the dependent variable. To delve deeper into this relationship, we utilized Cramer's V coefficient as a post-test to ascertain the strength of association after the Chi-square test has determined significance. Cramer's V coefficient falls within the range of [0,1]. To interpret Cramer’s V, the following approach is often used (Table 12):

Table 12. Interpretation value of Cramer’s V coefficient

Cramer’s V	Interpretation
Cramer’s V ≤ 0.2	The result is weak. Although the result is statistically significant, the variables are only weakly associated.
0.2 < Cramer’s V ≤ 0.6	The result is moderate. The variables are moderately associated.
Cramer’s V > 0.6	The result is strong. The fields are strongly associated.

The value obtained for the Cramer's V coefficient for each of the independent variables is shown in Table 13.

Table 13. The result of Cramer’s V coefficient to examine strengths of association between independent variables and Operational Code

Depended variable	Independed variable	Cramer's V coefficient	Association
Code_Level1	Operation Type	0.91	Strong association
Code_Level1	Shift Work	0.01	Weak association
Code_Level1	Municipal District	0.14	Weak association
Code_Level1	Site Type	0.44	Moderate association
Code_Level1	Fire Station	0.19	Weak association

The Table 13 confirms the expert panel result. Finally, the operations from March 2015 to September 2018 were divided based on Operation Type field and the corresponding classifier based on Table 6 was used to determine the missing values.

The proposed measure (Operation Type) had three principles for the selection of the measure:

- **Power of measure:** The selection criteria were determined using a qualitative method involving a panel of experts, where there was a high degree of consensus among all experts. Additionally, in the quantitative method employed, the statistical significance of the results was complemented by an assessment of the strength of the relationship.
- **Measure accuracy:** Both the expert panel and statistical tests utilized in this research demonstrated high accuracy.
- **Measure comprehensiveness:** The selected measure is recorded for all operations conducted.

4. CONCLUSION

Missing data represent one of the most critical challenges in data mining projects. Given its direct impact on data quality, considerable efforts are dedicated to devising effective techniques for predicting missing values. Most proposed methods utilize all available data to estimate missing values. In this study, we present an innovative approach that addresses this challenge by partitioning the database into disjoint partitions, each equipped with its own predictive model tailored to its specific context within the database.

Central to our approach is the identification of a partitioning measure characterized by three essential attributes. Firstly, it must ensure the creation of distinct partitions within the database. Secondly, the accuracy of all processes and data involved in determining this measure is paramount. Finally, the measure should be computable for records with missing parts.

Our evaluation, conducted using the operational database of Mashhad Fire and Safety Services, incorporated expert panel insights and statistical analyses to establish the partitioning measure. By comparing models built on all available data with those constructed on disjoint partitions, we observed that the disjoint partitions model exhibited superior performance in 44% of cases, significantly enhancing the accuracy of missing value prediction.

Simple measures for data separation were employed in this research. The introduction of more complex measures is likely to improve model accuracy. Notably, since the proposed approach is independent of the method used to impute missing values, it can complement existing methods as an additional step in the process of determining missing values.

REFERENCES

[1] Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent Neural Networks for Multivariate Time Series with Missing Values. Scientific Reports. 2018 Apr 17;8(1).

[2] Han J, Kamber M, Computer P. Data mining : concepts and techniques. Amsterdam ; Boston: Elsevier/Morgan Kaufmann; 2012.

[3] Rahman MdG, Islam MZ. Missing value imputation using a fuzzy clustering-based EM approach. Knowledge and Information Systems. 2015 Feb 25;46(2):389–422.

[4] Wang H, Wang S. Mining incomplete survey data through classification. Knowledge and Information Systems. 2009 Aug 20;24(2):221–33.

[5] Fletcher Mercaldo S, Blume JD. Missing data and prediction: the pattern submodel. Biostatistics. 2018 Sep 6;21(2):236–52.

[6] Köpcke F, Trinczek B, Majeed RW, Schreiweis B, Wenk J, Leusch T, et al. Evaluation of data completeness in the electronic health record for the purpose of patient recruitment into clinical trials: a retrospective analysis of element presence. BMC Medical Informatics and Decision Making. 2013 Mar 21;13(1).

[7] Pivovarov R, Albers DJ, Sepulveda JL, Elhadad N. Identifying and mitigating biases in EHR laboratory tests. Journal of Biomedical Informatics. 2014 Oct;51:24–34.

[8] Horvath MM, Rusincovitch SA, Richesson RL. Clinical Research Informatics and Electronic Health Record Data. Yearbook of Medical Informatics. 2014 Aug;23(01):215–23.

[9] Zhang X, Yan C, Gao C, Malin BA, Chen Y. Predicting Missing Values in Medical Data Via XGBoost Regression. Journal of Healthcare Informatics Research. 2020 Aug 3;4(4):383–94.

[10] Khan H, Wang X, Liu H. Missing value imputation through shorter interval selection driven by Fuzzy C-Means clustering. Computers & Electrical Engineering. 2021 Jul;93:107230.

[11] Ngueilbaye A, Wang H, Mahamat DA, Junaidu SB. Modulo 9 model-based learning for missing data imputation. Applied Soft Computing. 2021 May;103:107167.

[12] Xu D, Hu PJH, Huang TS, Fang X, Hsu CC. A deep learning–based, unsupervised method to impute missing values in electronic health records for improved patient management. Journal of Biomedical Informatics. 2020 Nov;111:103576.

[13] Beaulieu-Jones BK, Lavage DR, Snyder JW, Moore JH, Pendergrass SA, Bauer CR. Characterizing and Managing Missing Structured Data in Electronic Health Records: Data Analysis. JMIR Medical Informatics. 2018 Feb 23;6(1):e11.

[14] Ancker JS, Witteman HO, Hafeez B, Provencher T, Van de Graaf M, Wei E. The Invisible Work of Personal Health Information Management Among People With Multiple Chronic Conditions: Qualitative Interview Study Among Patients and Providers. Journal of Medical Internet Research. 2015 Jun 4;17(6):e137.

[15] Forster AJ, Kyeremanteng K, Hooper J, Shojania KG, van Walraven C. The impact of adverse events in the intensive care unit on hospital mortality and length of stay. BMC Health Services Research . 2008 Dec;8(1).

[16] Hu Z, Du D. A new analytical framework for missing data imputation and classification with uncertainty: Missing data imputation and heart failure readmission prediction. Kaderali L, editor. PLOS ONE. 2020 Sep 21;15(9):e0237724.

[17] Kohli R, Tan SSL. Electronic Health Records: How Can IS Researchers Contribute to Transforming Healthcare? MIS Quarterly. 2016;40(3):553–74.

[18] Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PRO, Bernstam EV, et al. Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research. Medical Care. 2013 Aug;51:S30–7.

[19] Little RJ, D’Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, et al. The Prevention and Treatment of Missing Data in Clinical Trials. New England Journal of Medicine. 2012 Oct 4;367(14):1355–60.

[20] Nakagawa S, Freckleton RP. Missing inaction: the dangers of ignoring missing data. Trends in Ecology & Evolution. 2008 Nov;23(11):592–6.

[21] Groenwold RHH, White IR, Donders ART, Carpenter JR, Altman DG, Moons KGM. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. Canadian Medical Association Journal. 2012 Feb 27;184(11):1265–9.

[22] Roderick, Rubin DB. Statistical Analysis with Missing Data. John Wiley & Sons; 2014.

[23] Myrtveit I, Stensrud E, Olsson UH. Analyzing data sets with missing data: an empirical evaluation of imputation methods and likelihood-based methods. IEEE Transactions on Software Engineering. 2001;27(11):999–1013.

[24] Khosravi H, Das S, Al-Mamun A, Ahmed I. Binary Gaussian Copula Synthesis: A Novel Data Augmentation Technique to Advance ML-based Clinical Decision Support Systems for Early Prediction of Dialysis Among CKD Patients. arXiv.org. 2024. Available from: https://arxiv.org/abs/2403.00965

[25] MARIMONT RB, SHAPIRO MB. Nearest Neighbour Searches and the Curse of Dimensionality. IMA Journal of Applied Mathematics. 1979 ;24(1):59–70.

[26] Purwar A, Singh SK. Hybrid prediction model with missing value imputation for medical data. Expert Systems with Applications. 2015 Aug;42(13):5621–31.

[27] White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Statistics in Medicine. 2010 Sep 13;29(28):2920–31.

[28] García-Laencina PJ, Sancho-Gómez JL, Figueiras-Vidal AR. Pattern classification with missing data: a review. Neural Computing and Applications. 2009 Sep 3;19(2):263–82.

[29] Willem S, T. Katrien J. Groenhof, Hoogland J, Bots ML, Menno Brandjes, John J.L. Jacobs, et al. Real-time imputation of missing predictor values improved the application of prediction models in daily practice. Journal of Clinical Epidemiology. 2021 Jun 1;134:22–34.

[30] Peng D, Zou M, Liu C, Lu J. RESI: A Region-Splitting Imputation method for different types of missing data. Expert Systems with Applications. 2021 Apr;168:114425.

[31] Yang K, Li J, Wang C. Missing Values Estimation in Microarray Data with Partial Least Squares Regression. Lecture Notes in Computer Science. 2006 Jan 1;662–9.

[32] Zhao P, Tang X. Imputation based statistical inference for partially linear quantile regression models with missing responses. Metrika. 2016 Jun 9;79(8):991–1009.

[33] Sentas P, Angelis L. Categorical missing data imputation for software cost estimation by multinomial logistic regression. Journal of Systems and Software. 2006 Mar;79(3):404–14.

[34] Malan L, Smuts CM, Baumgartner J, Ricci C. Missing data imputation via the expectation-maximization algorithm can improve principal component analysis aimed at deriving biomarker profiles and dietary patterns. Nutrition Research. 2020 Mar;75:67–76.

[35] Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data Via the EM Algorithm. Journal of the Royal Statistical Society: Series B (Methodological). 1977 Sep;39(1):1–22.

[36] Rubul Kumar Bania, Halder A. R-Ensembler: A greedy rough set based ensemble attribute selection algorithm with kNN imputation for classification of medical data. Computer Methods and Programs in Biomedicine. 2020 Feb 1;184:105122–2.

[37] Zhang S. Nearest neighbor selection for iteratively kNN imputation. Journal of Systems and Software. 2012 Nov;85(11):2541–52.

[38] Batista GEAPA, Monard MC. An analysis of four missing data treatment methods for supervised learning. Applied Artificial Intelligence [Internet]. 2003 May;17(5-6):519–33. Available from: http://conteudo.icmc.usp.br/pessoas/gbatista/files/aai2003.pdf

[39] Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001 Jun 1;17(6):520–5.

[40] Aydilek IB, Arslan A. A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. Information Sciences. 2013 Jun;233:25–35.

[41] Loes C.M. Bertens, Broekhuizen BDL, Naaktgeboren CA, Rutten FH, Hoes AW, Yvonne van Mourik, et al. Use of Expert Panels to Define the Reference Standard in Diagnostic Research: A Systematic Review of Published Methods and Reporting. PLOS Medicine. 2013 Oct 15;10(10):e1001531–1.

[42] Allen DM. The Relationship Between Variable Selection and Data Agumentation and a Method for Prediction. Technometrics. 1974 Feb;16(1):125–7.

[43] www.amar.org.ir, Iran Statistics Center Portal

[44] Jianqiang Z, Xiaolin G. Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis. IEEE Access. 2017;5:2870–9.

[45] SALTON G. Developments in Automatic Text Retrieval. Science. 1991 Aug 30;253(5023):974–80.

[46] Fabian P. Scikit-learn: Machine learning in Python. Journal of machine learning research 12. 2011;2825.

[47] Suthaharan S. Support Vector Machine. Machine Learning Models and Algorithms for Big Data Classification. 2016;36:207–35.

[48] Mccullagh P, Nelder JA. Generalized Linear Models. Boca Raton Crc Press Llc Ann Arbor, Michigan Proquest; 1989.

[49] Statistics, I. S. "Ibm corp. released 2013. ibm spss statistics for windows, version 22.0. armonk, ny: Ibm corp.".