TASHKENT STATE UNIVERSITY OF ECONOMICS

International journal of topics

INNOVATION SCIENCE AND TECHNOLOGY

Scopus || Electronic journal specializing in Scopus

ISSUE 5

Acceptance of papers **May, 2026**

Acceptance of papers

Published monthly

Topics

economics, technology, social sciences

ISSN 3060-5229

OpenAIRE Google BASE doi Digital Object Identifier OPEN ACCESS

Visit the website
t.me/scopus_IST2100

Electronic publication, Issue 5.
Approved for publication on May, 2026.

THE SCIENTIFIC-POPULAR ELECTRONIC JOURNAL **"INNOVATION SCIENCE AND TECHNOLOGY"** HAS BEEN REGISTERED UNDER THE NUMBER **C-5669633** BY THE AGENCY FOR INFORMATION AND MASS COMMUNICATIONS (AOKA) OF THE REPUBLIC OF UZBEKISTAN, EFFECTIVE FROM OCTOBER 9, 2024.

CONTACTS

Phone: **+998 50 737 87 88**

Website: https://ist-journal.uz

Email: innovationist2025@gmail.com

The scientific electronic journal “Innovation Science and Technology” has been included in the list of scientific publications recommended for the publication of main scientific results of dissertations for the award of PhD and DSc degrees in economics and technical sciences, in accordance with the Resolution No. 370 of the Presidium of the Higher Attestation Commission of the Republic of Uzbekistan, dated May 8, 2025.

Editorial board:

CONTENTS

CONTENTS

COMPARATIVE ANALYSIS AND EXPERIMENTAL EVALUATION OF ALGORITHMS FOR RECOVERING MISSING (NAN) VALUES IN INFORMATION SYSTEM DATA

Yarmatov Sherzodjon Shokir oglu
PhD, Samarkand State University named after Sharof Rashidov (SamDU)
E-mail: sherzod2601@gmail.com

Orifov Oxunjon Fazliddinzoda
Second-Year Master's Student,
Samarkand State University named after Sharof Rashidov
E-mail: okhunjonorifov555@gmail.com

Abstract. This article investigates the problem of identifying and recovering missing (NaN) values in information system data and examines their influence on analytical results. During the research process, artificial missing values with different proportions were generated based on a complete dataset and subsequently restored using statistical and machine-learning-based imputation methods. The effectiveness of each algorithm was evaluated using error metrics obtained through comparison with the original ground-truth values. The obtained results made it possible to determine the efficiency of different methods depending on the structure of the data and to establish a methodological basis for selecting optimal approaches in the intelligent analysis of information system data. The findings of the study contribute positively to improving data quality and enhancing the reliability of analytical processes.

Keywords: NaN values, missing data, imputation algorithms, KNN imputation, MICE method, data analysis, machine learning, information systems.

Annotatsiya. Ushbu maqolada axborot tizimlari ma'lumotlarida uchraydigan yetishmayotgan (NaN) qiymatlarni aniqlash va tiklash muammosi hamda ularning tahliliy natijalarga ta'siri tadqiq etilgan. Tadqiqot jarayonida to'liq ma'lumotlar to'plami asosida turli ulushdagi sun'iy yetishmayotgan qiymatlar shakllantirilib, ular statistik va mashinaviy o'qitishga asoslangan imputatsiya usullari yordamida qayta tiklandi. Har bir algoritmning samaradorligi tiklangan qiymatlarni original ground-truth ma'lumotlari bilan taqqoslash orqali olingan xatolik ko'rsatkichlari asosida baholandi. Olingan natijalar ma'lumotlar tuzilmasiga qarab turli usullarning samaradorligini aniqlash hamda axborot tizimlari ma'lumotlarini intellektual tahlil qilishda optimal yondashuvlarni tanlash uchun metodologik asos yaratishga xizmat qiladi. Tadqiqot natijalari ma'lumotlar sifati va tahliliy jarayonlarning ishonchliligini oshirishga ijobiy hissa qo'shadi.

Kalit so'zlar: NaN qiymatlar, yetishmayotgan ma'lumotlar, imputatsiya algoritmlari, KNN imputatsiyasi, MICE usuli, ma'lumotlar tahlili, mashinaviy o'qitish, axborot tizimlari.

Аннотация. В данной статье исследуется проблема выявления и восстановления отсутствующих (NaN) значений в данных информационных систем, а также их влияние на результаты аналитической обработки. В ходе исследования на основе полного набора данных были искусственно сформированы пропущенные значения с различной долей, после чего они восстанавливались с использованием статистических методов и алгоритмов импутации, основанных на машинном обучении. Эффективность каждого алгоритма оценивалась на основе показателей ошибок, полученных путем сравнения восстановленных значений с исходными ground-truth данными. Полученные результаты позволили определить эффективность различных методов в зависимости от структуры данных и сформировать методологическую основу для выбора оптимальных подходов при интеллектуальном анализе данных информационных систем. Результаты исследования способствуют повышению качества данных и надежности аналитических процессов.

Ключевые слова: NaN-значения, пропущенные данные, алгоритмы импутации, KNN-импутация, метод MICE, анализ данных, машинное обучение, информационные системы.

INTRODUCTION

In recent years, the rapid development of information systems has led to the generation of large volumes of data across various domains. Such data have become an important resource in management processes, scientific research, and decision-making systems. At the same time, the problem of incomplete data frequently arises during the formation of real-world datasets. Missing values may occur due to the absence of information for certain attributes or as a result of inaccuracies during data collection and recording processes. These values are commonly represented in the form of NaN (Not a Number).

The presence of missing values in a dataset can significantly affect the quality and reliability of analytical processes. Statistical calculations, data-driven models, and machine-learning algorithms generally require complete and consistent datasets. When missing values exist within the data, certain computational procedures may produce less accurate results, while the predictive performance of analytical models may decrease. Therefore, identifying missing values, analyzing their causes, and recovering them using appropriate methods represent essential stages prior to conducting data analysis.

Various methods are applied in practice to handle missing values effectively. The simplest approaches include statistical techniques such as mean, median, and mode imputation. Although these methods are computationally efficient and easy to implement, they do not always account for complex relationships among data attributes. For this reason, machine-learning-based approaches, including KNN imputation, MICE, and regression-based methods, have become increasingly popular in recent years. These algorithms enable more accurate recovery of missing values by considering statistical and functional relationships among data attributes.

The effectiveness of missing-value recovery algorithms largely depends on the structure of the dataset, the number of attributes, and the degree of interdependence between variables. Therefore, conducting experimental comparisons of different algorithms under identical conditions constitutes an important scientific task. Such studies make it possible to determine which algorithms provide higher efficiency and reliability for specific types of data.

This article experimentally evaluates the effectiveness of algorithms designed for recovering missing values in information-system data. During the research process, a complete dataset was selected, and artificial missing values with different percentages were generated within it. These missing values were subsequently restored using various imputation algorithms, and the obtained results were compared with the original ground-truth values. Based on the experimental evaluation results, a comparative analysis of the algorithms was carried out, and their effectiveness was assessed using error metrics.

The results of this study may serve as an important methodological basis for selecting appropriate missing-data-processing techniques in the intelligent analysis of information-system data. Furthermore, the proposed approach contributes to improving data quality, increasing the reliability of analytical processes, and enhancing the accuracy of subsequent predictive models.

LITERATURE REVIEW

The problem of missing values in datasets has long been recognized as one of the significant challenges in the field of data analysis. In various information systems, missing values may occur during data collection processes due to technical failures, human factors, interruptions in data transmission, or incomplete recording of certain attributes. Since statistical and analytical models generally operate on complete and consistent datasets, the presence of missing values may lead to inaccurate analytical results and reduce the overall accuracy and reliability of predictive models. Therefore, the development and improvement of methods for identifying and recovering missing values have become one of the key directions of contemporary scientific research.

Scientific literature extensively discusses the causes of missing values and their impact on data analysis processes. In most studies, missing data are classified into three main categories: MCAR (Missing Completely At Random), MAR (Missing At Random), and MNAR (Missing Not At Random). In the MCAR case, missing values occur randomly and are independent of other attributes within the dataset. In the MAR case, missing values may depend on other observed attributes in the dataset. In the MNAR case, the occurrence of missing values depends directly on the attribute itself. This classification provides an important methodological foundation for selecting appropriate algorithms and approaches for processing missing values effectively (Figure 1).

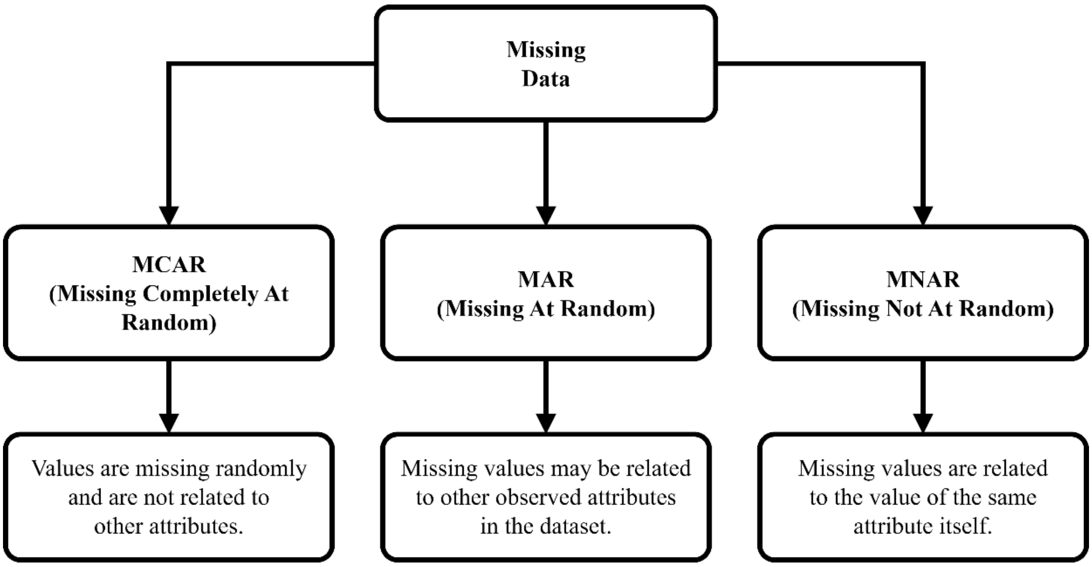

Figure 1. Classification of Missing Data Types.

The diagram presented in Figure 1 illustrates the main types of missing data. In statistical analysis, missing values are generally classified into three primary categories: MCAR (Missing Completely At Random), MAR (Missing At Random), and MNAR (Missing Not At Random). This classification plays an important role in understanding the mechanisms of data loss and in selecting appropriate imputation methods.

Various approaches have been proposed for handling missing values effectively. Among the simplest methods are statistical imputation techniques. In these approaches, missing values are replaced with the mean, median, or mode of the corresponding attribute. Such methods are computationally efficient and provide rapid results, particularly for small datasets. However, these approaches often fail to consider complex relationships among data attributes, which may artificially reduce data variance and negatively affect analytical accuracy.

In recent years, machine-learning-based imputation methods have become increasingly widespread. One of the most commonly applied approaches is the KNN (K-Nearest Neighbors) imputation method, which restores missing values based on the most similar observations. In this approach, distances between data instances are calculated, and missing values are estimated using the nearest neighbors. Since this method takes local relationships within the data into account, it can provide higher accuracy and reliability compared to simple statistical techniques.

Another widely used method is the MICE (Multiple Imputation by Chained Equations) algorithm. In this approach, a regression model is constructed for each attribute, and missing values are recovered iteratively. The MICE algorithm makes it possible to account for statistical relationships among attributes in multivariate datasets. Therefore, it can produce highly effective results when working with data characterized by complex structures and multidimensional dependencies.

In addition, regression-based imputation methods are also widely utilized in scientific research. In such approaches, missing values are predicted on the basis of existing relationships among other attributes. Furthermore, some studies have employed ensemble-algorithm-based approaches, which have demonstrated high levels of predictive accuracy and analytical stability (Figure 2).

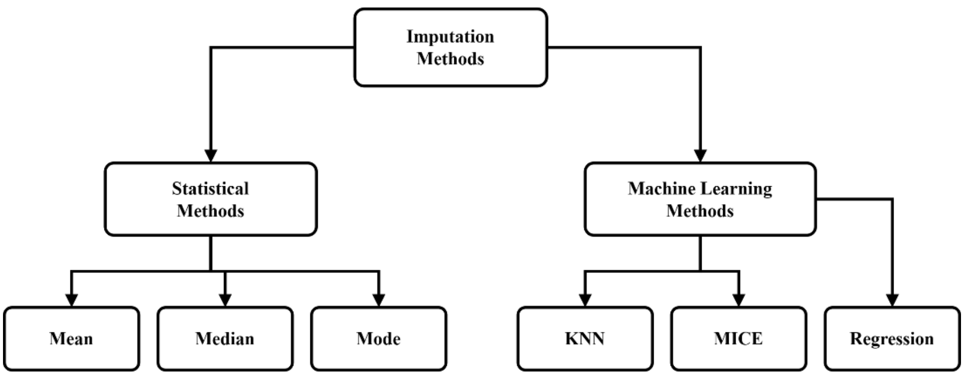

Figure 2. Main Groups of Missing Value Recovery Methods.

The scheme presented in Figure 2 illustrates the main groups of methods used for recovering missing values. Imputation approaches are generally divided into statistical methods and machine-learning-based methods. While statistical methods are computationally simple and efficient, machine-learning approaches make it possible to recover missing values with higher accuracy by taking relationships among data attributes into account.

Numerous scientific studies have experimentally compared the effectiveness of different imputation algorithms. Research findings indicate that the performance of these algorithms may vary depending on the structure of the dataset, the proportion of missing values, and the relationships among variables. Therefore, it is difficult to consider any single algorithm as a universal solution for all types of datasets. In practical applications, determining the most effective method requires testing different algorithms on the same dataset and conducting a comparative analysis of the obtained results.

The present study is also based on this scientific approach. Specifically, artificial missing values are generated using a complete dataset and subsequently recovered through various imputation algorithms. The obtained results are then compared with the original ground-truth values, and the effectiveness of the algorithms is evaluated using error metrics. This approach makes it possible to conduct an objective and reliable comparison of missing-value recovery algorithms and to determine their applicability under different data conditions.

The problem of missing values in datasets has been extensively studied in modern scientific research due to its significant impact on data quality and analytical accuracy. Roderick J. A. Little and Donald B. Rubin emphasized that missing data mechanisms such as MCAR, MAR, and MNAR play an important role in selecting appropriate imputation techniques. According to Stef van Buuren, advanced imputation methods, particularly MICE, allow researchers to preserve statistical relationships among variables in multivariate datasets. Research conducted by Olga Troyanskaya demonstrated that KNN-based approaches can effectively recover missing values in complex datasets by utilizing similarities between observations. Furthermore, studies by Trevor Hastie and Jerome Friedman highlighted the importance of machine learning algorithms in improving predictive accuracy during missing-data processing. These studies provide a strong methodological foundation for evaluating imputation algorithms in information systems.

RESEARCH METHODOLOGY

In this study, an experimental approach was employed to determine the effectiveness of algorithms for recovering missing (NaN) values in information-system data. The research process consisted of several sequential stages, including dataset preparation, artificial generation of missing values, application of various imputation algorithms, and evaluation of the obtained results. Such an approach makes it possible to test algorithms under identical conditions and to conduct an objective and reliable comparative analysis of their performance.

The research process was organized according to the following main stages: dataset preparation, generation of artificial NaN values, application of imputation algorithms, recovery of missing values, and analysis of algorithm accuracy based on evaluation metrics. This methodological framework ensured the consistency of the experimental procedures and enhanced the reliability and scientific validity of the obtained results.

ANALYSIS AND RESULTS

The study utilized a dataset derived from the California Housing dataset, which contains statistical information related to housing in the California region. This dataset consists of attributes representing economic and demographic indicators associated with the real-estate market and is considered one of the most widely used real-world datasets in the fields of machine learning and data analysis.

The dataset includes indicators describing the geographical location of houses, population size within districts, number of households, average age of houses, number of rooms, number of bedrooms, median income, and median house value. These attributes represent important factors reflecting the economic and demographic characteristics of the housing market and provide a reliable basis for analytical research.

At the initial stage, the structure of the dataset was examined, and the data types of the attributes were analyzed. Since the dataset contained the categorical attribute *ocean_proximity*, only numerical attributes were selected for the application of imputation algorithms. As a result, the experimental analysis was conducted using a dataset consisting of nine numerical attributes.

The preliminary inspection of the dataset demonstrated that it did not contain any missing values. This made it possible to artificially generate NaN values and compare the values restored by the algorithms with the original ground-truth values contained in the dataset. Therefore, the selected dataset provided a suitable and reliable experimental environment for evaluating the accuracy and effectiveness of imputation algorithms.

At the next stage of the research process, artificial missing values were generated within the dataset. For this purpose, certain attribute values were randomly selected and replaced with NaN values. This process was carried out to simulate the occurrence of missing values that may arise in real information systems during data collection, transmission, or storage processes.

Since the missing values were generated using a random-selection approach, they correspond to the MCAR (Missing Completely At Random) model. In this model, missing values occur randomly and are independent of other attributes within the dataset. Such an approach provides a reliable experimental basis for objectively evaluating the performance of missing-value recovery algorithms under controlled conditions (Table 1).

Table 1

During the experimental study, missing values were generated at four different levels[1]

Missing Value Ratio	Mean Imputation	**Median Imputation**	**KNN Imputation**	**MICE Imputation**
5%	5%	5%	5%	5%
10%	10%	10%	10%	10%
20%	20%	20%	20%	20%
30%	30%	30%	30%	30%

At each level, values from every attribute in the dataset were randomly selected with equal probability and replaced with NaN values. Such an approach makes it possible to evaluate the performance and stability of the algorithms under different levels of data loss.

To recover the artificially generated missing values, four imputation algorithms based on statistical and machine-learning approaches were applied. The table presents the main characteristics, mathematical formulations, and Python-based implementations of the imputation algorithms used in the study. Since these algorithms cover both statistical and machine-learning approaches, they provide an opportunity to recover missing values from different analytical perspectives and to conduct a comprehensive comparative evaluation of their effectiveness (Table 2).

Table 2

Description of the Imputation Algorithms Used in the Study[2]

№	Algorithm Name	Description (Purpose)	Mathematical Expression	Python (Scikit-learn) Implementation
1	**Mean Imputation**	In this method, missing values are replaced with the mean value of the corresponding attribute. Although this approach is computationally simple and efficient, it may lead to a reduction in data variance.	$x_i = \bar{x}$	`SimpleImputer (strategy='mean')`
2	**Median Imputation**	In this approach, missing values are replaced with the median value of the corresponding attribute. The median method can provide more stable and reliable results, particularly when the data distribution is skewed or when outlier values are present.	$x_i = \tilde{x}$	`SimpleImputer (strategy=’median’)`
3	**KNN Imputation**	In this method, which is based on the K-Nearest Neighbors (KNN) algorithm, missing values are recovered using similar observations. The distances between data instances are calculated, and the estimated value is determined based on the values of the nearest neighbors. This approach makes it possible to utilize relationships among attributes effectively. In the mathematical expression, xjx_jxj represents the corresponding values of the kkk nearest neighbors.	$x_i = \frac{1}{k}\sum_{j=1}^{k} x_j$	`KNNImputer(n_ neighbors=5)`

1 Created by author.
2 Created by author.

4	**MICE Imputation**	The Multiple Imputation by Chained Equations (MICE) method iteratively reconstructs missing values by building a separate regression model for each attribute. This approach allows statistical relationships among attributes to be taken into account during the imputation process. In this expression, f(X−i)f(X_{-i})f(X−i) represents the regression model constructed using the remaining attributes.	$x_i = f(X_{-i})$	`IterativeImputer max_iter=10, random_state=42)`

Before applying the KNN and MICE algorithms, attribute values were standardized using the StandardScaler method. This process reduces the excessive influence of attributes with different value ranges on distance calculations and ensures more accurate and reliable performance of the algorithms. Standardization also improves the stability of machine-learning-based imputation methods and enhances the comparability of attribute values during the recovery process.

To evaluate the effectiveness of the imputation algorithms, the recovered values were compared with the original ground-truth values contained in the dataset. The following error metrics were used during the evaluation process to determine the accuracy and reliability of the algorithms (Table 3).

Table 3

Evaluation Metrics[3]

Evaluation Metric	Mathematical Expression	Description
MAE (Mean (Absolute Error	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$	Average absolute error between the recovered values and the original ground-truth values
RMSE (Root Mean (Square Error	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$	Average error calculated based on the square root of squared differences

These metrics make it possible to quantitatively evaluate the accuracy and effectiveness of imputation algorithms. The MAE metric represents the average absolute difference between the recovered values and the original ground-truth values, while the RMSE metric indicates the average value calculated based on the square root of squared errors. The smaller the metric value, the higher the accuracy and reliability of the algorithm in recovering missing values.

All computational processes in this study were carried out using the Python programming environment. Data preprocessing, application of imputation algorithms, and visualization of the obtained results were performed on the Google Colab platform. During the experimental process, the NumPy, Pandas, Scikit-learn, and Matplotlib libraries of Python were utilized extensively. The experimental results were presented through tables and graphical representations, and the effectiveness of different imputation algorithms was evaluated by means of comparative analysis.

The general block diagram of the research process reflects the sequential stages of dataset preparation, generation of artificial missing values, application of imputation algorithms, recovery of missing values, and evaluation of algorithm performance using error metrics. Such an integrated methodological approach ensures the reliability, consistency, and scientific validity of the experimental results while providing a solid foundation for the comparative assessment of missing-value recovery algorithms (Figure 3).

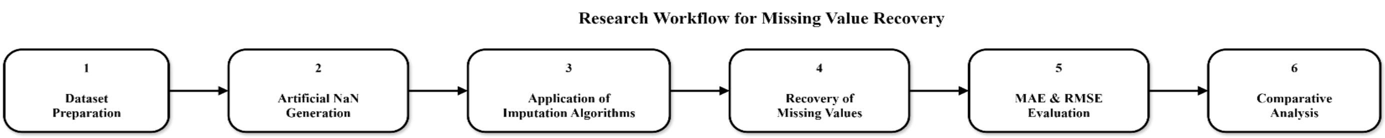

Figure 3. Evaluation Process of Missing Value Recovery Algorithms

The block diagram presented in Figure 3 illustrates the main stages of the research process in a sequential and systematic manner. Initially, the dataset is prepared, and artificial missing values are generated within the data. In the subsequent stages, imputation algorithms are applied, the recovered values are evaluated using MAE and RMSE metrics, and a comparative analysis of the algorithms is conducted. Such an approach ensures

3 Created by author.

the consistency of the experimental process and enables an objective evaluation of algorithm performance.

In this study, an experimental analysis was carried out to evaluate the effectiveness of missing-value recovery algorithms. During the experiment, artificially generated NaN values in the dataset were restored using different imputation algorithms, and the recovered values were compared with the original ground-truth values contained in the dataset. The accuracy and reliability of the algorithms were evaluated using MAE and RMSE metrics.

Within the experimental process, missing values were generated at four different levels: 5%, 10%, 20%, and 30%. At each level, the corresponding percentage of the dataset was randomly selected and replaced with NaN values. Subsequently, the missing values were recovered using the Mean, Median, KNN, and MICE algorithms, and the restored values were compared with the original values. This approach made it possible to assess the effectiveness of the algorithms under different levels of data loss and to determine their relative performance in recovering missing values accurately.

The MAE results obtained from the experiment for different percentages of NaN values are presented in Table 4.

Table 4
MAE Results of the Imputation Algorithms[4]

Algorithm	5%	10%	20%	30%
Mean	10637	10446	10561	10565
Median	10263	10128	10299	10281
KNN	5452	6459	8905	10291
MICE	5963	6099	6545	7717

The results demonstrate that when the proportion of missing values was relatively low (5%), the KNN algorithm achieved the smallest error value and showed high recovery accuracy. However, as the proportion of NaN values increased, the error indicators of the KNN algorithm also increased noticeably.

At the 10%, 20%, and 30% NaN levels, the MICE algorithm demonstrated the lowest MAE values and produced more stable and reliable results compared to the other methods. This indicates that algorithms capable of considering statistical relationships among attributes perform more effectively, particularly when the proportion of missing values increases and the data structure becomes more complex.

In addition, the Mean and Median imputation methods produced relatively higher error values in all experimental cases. This finding suggests that simple statistical methods are not fully capable of recovering complex data structures and multidimensional relationships among attributes. Nevertheless, these methods remain computationally efficient and may still be useful in cases involving small datasets or low levels of missing data.

The RMSE results obtained from the experiment are presented in Table 5.

Table 5
RMSE Results of the Imputation Algorithms[5]

Algorithm	5%	10%	20%	30%
Mean	39185	38628	39083	39012
Median	40102	39575	40309	40202
KNN	23608	26670	34301	38764
MICE	24455	25190	26711	31999

The RMSE results demonstrated a trend consistent with the MAE indicators. At the 5% NaN level, the KNN algorithm achieved the lowest RMSE value and showed high recovery accuracy. However, as the proportion of NaN values increased, the MICE algorithm produced significantly more stable and reliable results compared to the other methods.

Under conditions with 20% and 30% missing values, the MICE algorithm recorded the lowest RMSE values, demonstrating its strong capability to recover missing data accurately even when the proportion of missing values became relatively high. These results confirm that algorithms capable of modeling statistical relationships among attributes can maintain higher performance and stability in complex missing-data scenarios

4 Created by author.
5 Created by author.

(Table 6).

Table 6
MAE and RMSE Results of Imputation Algorithms at Different NaN Percentages[6]

NaN (%)	Algorithm	MAE	RMSE
5	Mean	10637.9365	39185.8703
5	Median	10263.4206	40102.2168
5	KNN	5452.1356	23608.7052
5	MICE	5963.5255	24455.5731
10	Mean	10446.0700	38628.4517
10	Median	10128.4719	39575.3168
10	KNN	6459.0324	26670.4592
10	MICE	6099.4758	25190.9254
20	Mean	10561.7147	39083.2854
20	Median	10299.7700	40309.8616
20	KNN	8905.1554	34301.6908
20	MICE	6545.1929	26711.7006
30	Mean	10565.6010	39012.0422
30	Median	10281.2938	40202.0045
30	KNN	10291.0296	38764.3383
30	MICE	7717.8526	31999.4132

Table 6 summarizes the MAE and RMSE results of the imputation algorithms under different NaN-percentage conditions. The obtained results indicate that when the proportion of missing values is relatively low, the KNN algorithm achieves the smallest error values. In particular, at the 5% NaN level, this algorithm demonstrated significantly lower MAE and RMSE values compared to the other methods, indicating its high recovery accuracy under limited missing-data conditions.

However, as the proportion of NaN values increased, the error indicators of the KNN algorithm also increased considerably. This can be explained by the fact that the performance of distance-based algorithms depends on the availability of a sufficient number of complete observations. As the number of missing values within the dataset grows, identifying similar observations becomes more difficult, which consequently reduces recovery accuracy.

The results also confirm that the MICE algorithm demonstrates relatively stable and reliable performance. Under conditions where the proportion of NaN values reached 10%, 20%, and 30%, this algorithm achieved lower error values compared to the other methods. This finding highlights the importance of considering statistical relationships among attributes in multivariate datasets and demonstrates the effectiveness of iterative regression-based approaches in recovering missing data.

The simple statistical approaches, namely Mean and Median imputation algorithms, produced relatively high error values under all experimental conditions. This indicates that such methods are unable to fully reconstruct complex data structures because they do not account for relationships among attributes. Nevertheless, these methods remain computationally efficient and may still provide acceptable performance in datasets with limited complexity or relatively low proportions of missing data.

Overall, the obtained results demonstrate that the selection of an appropriate missing-value recovery algorithm depends on the characteristics of the dataset and the proportion of NaN values. While the KNN algorithm can provide effective results when the amount of missing data is relatively small, the MICE algorithm ensures more stable, reliable, and accurate performance as the proportion of missing values increases.

Based on the experimental results, the overall performance of the algorithms was comparatively analyzed. After calculating the average error values, the MICE algorithm demonstrated the lowest overall error indicators. This confirms that the algorithm is highly effective in recovering missing values within multivariate datasets by taking relationships among attributes into account.

The results further showed that although simple statistical methods such as Mean and Median imputation are computationally efficient, they exhibit relatively limited performance when reconstructing complex data

6 Created by author.

structures. The distance-based KNN algorithm produced effective results when the proportion of missing data was small; however, its accuracy decreased as the proportion of NaN values increased.

In contrast, the MICE algorithm operates on the basis of iterative regression models, allowing it to account for dependencies among attributes. As a result, it provides relatively stable and reliable performance even under conditions involving a high proportion of missing data.

To provide a clearer interpretation of the experimental findings, the MAE and RMSE indicators of the algorithms were visualized graphically. The graphical analysis demonstrated that the error values increased for all algorithms as the proportion of NaN values grew. However, the MICE algorithm showed a slower growth trend in error values compared to the other methods and demonstrated the most stable overall performance throughout the experimental process (Figure 4).

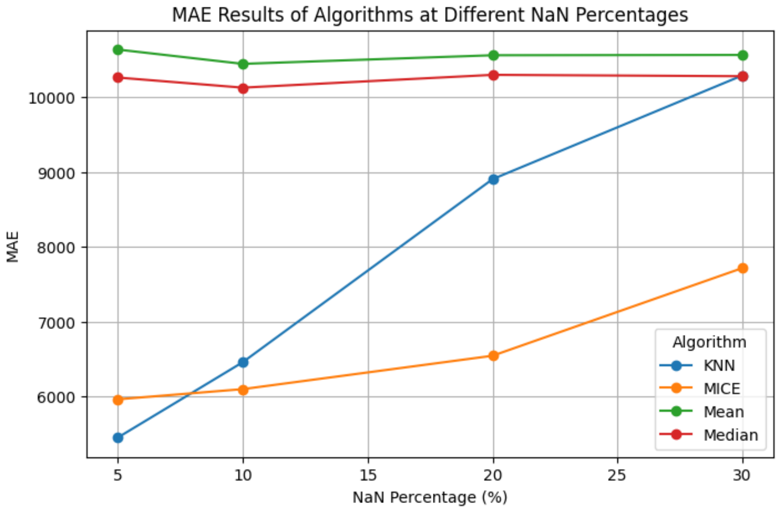

Figure 4. MAE Results of Imputation Algorithms at Different NaN Percentages.

Figure 4 illustrates the MAE results of the imputation algorithms under different NaN percentage conditions. The graphical analysis shows that the error values increase for all algorithms as the proportion of NaN values grows. However, the MICE algorithm demonstrates more stable performance compared to the other methods (Figure 5).

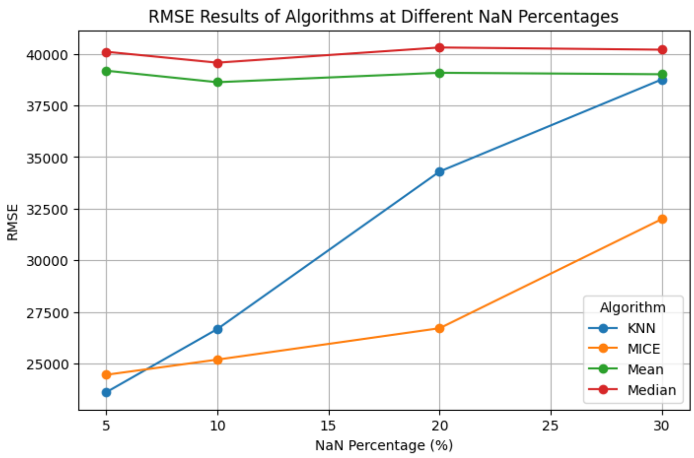

Figure 5. RMSE Results of Imputation Algorithms at Different NaN Percentages.

Figure 5 presents the RMSE results of the imputation algorithms under different NaN-percentage conditions. As illustrated in the graph, the KNN algorithm provides effective performance when the proportion of NaN values is relatively small; however, its error indicators increase rapidly as the amount of missing data grows. In contrast, the MICE algorithm ensures relatively stable and reliable performance across different levels of missing data. These results once again confirm the importance of applying algorithms that take statistical relationships among attributes into account during the process of recovering missing values.

The results of the conducted experimental study demonstrated that the effectiveness of missing-value recovery algorithms is significantly dependent on the proportion of NaN values within the dataset. Different algorithms produced different outcomes under varying conditions, which can be explained by the operational principles of the imputation methods and their level of adaptation to the structure of the data.

According to the experimental findings, when the proportion of missing values was relatively small (5%), the KNN algorithm achieved the lowest error values. The main reason for this is that the algorithm operates on the basis of similar observations. When the amount of missing data in the dataset is limited, the degree of similarity among available observations remains relatively high, enabling the KNN algorithm to recover missing values with greater accuracy. Therefore, distance-based algorithms were found to be particularly effective in situations involving low proportions of missing data.

However, as the proportion of missing values increased, the effectiveness of the KNN algorithm gradually decreased. This can be explained by the increasing difficulty of identifying similarities among observations as the number of missing values grows. When the dataset does not contain a sufficient number of complete observations, determining the nearest neighbors becomes more challenging, which ultimately reduces the accuracy of the recovered values.

The experimental findings also revealed that when the proportion of NaN values reached 10% or higher, the MICE algorithm demonstrated the most stable and reliable performance. This algorithm is based on the iterative recovery of missing values through the construction of regression models for each attribute. As a result, the MICE algorithm takes statistical relationships among attributes into account and is capable of producing relatively accurate recovery results even under conditions involving a high proportion of missing values.

The simple statistical methods, namely Mean and Median imputation, produced relatively high error values under all experimental conditions. These approaches replace missing values solely on the basis of the statistical characteristics of a single attribute and do not consider relationships among variables. Consequently, in multivariate datasets characterized by complex structures, such methods are unable to fully reconstruct the actual structure of the data.

The experimental results further demonstrated that the error indicators increased for all algorithms as the proportion of missing values grew. This is a natural outcome because the reduction in available information within the dataset decreases the possibility of accurately reconstructing missing values. Nevertheless, the MICE algorithm demonstrated considerably more stable behavior in terms of the rate of error growth compared to the other methods.

Overall, the analysis indicates that the selection of an appropriate missing-value recovery algorithm depends on the characteristics of the dataset and the proportion of missing values. In cases where only a small amount of data is missing, the KNN algorithm can provide effective and accurate results. However, as the proportion of missing values increases, the MICE algorithm, which considers statistical relationships among attributes, provides more reliable, stable, and accurate performance.

The findings of this study are important for ensuring data quality in information systems and demonstrate the necessity of considering dataset structure when selecting imputation algorithms for real-world data analysis. Furthermore, the obtained results provide a valuable methodological basis for improving analytical reliability and enhancing the effectiveness of intelligent data-processing systems.

CONCLUSIONS AND RECOMMENDATIONS

In this study, the effectiveness of algorithms for recovering missing (NaN) values in information-system data was analyzed using an experimental approach. During the research process, a statistical housing dataset related to the California region was utilized, and artificial missing values were generated in order to evaluate the performance of different imputation algorithms under controlled conditions.

Throughout the experiment, missing values were generated at different levels (5%, 10%, 20%, and 30%), and these values were restored using Mean, Median, KNN, and MICE algorithms. The recovered values were compared with the original ground-truth values contained in the dataset, and the accuracy of the algorithms was evaluated using the MAE and RMSE metrics.

The experimental results demonstrated that the KNN algorithm produced relatively accurate results when the proportion of missing values was low. However, as the proportion of missing values in the dataset increased, the effectiveness of this algorithm gradually decreased. In contrast, the MICE algorithm, which considers statistical relationships among attributes, provided relatively stable and reliable results even under conditions involving a high proportion of NaN values.

In addition, although the simple statistical methods, namely Mean and Median imputation, are computationally efficient, they were unable to provide high accuracy in datasets characterized by complex structures. Since these methods do not account for relationships among attributes, they produced relatively high error values during the recovery of missing data. Nevertheless, these approaches may still be useful in cases involving limited data complexity or relatively small proportions of missing values due to their simplicity and computational efficiency.

Based on the obtained results, it can be concluded that the selection of an appropriate algorithm for recovering missing data in information systems should depend on the structure of the dataset and the proportion of missing values. Distance-based algorithms may provide effective performance when the amount of missing data is relatively small, whereas under conditions involving larger proportions of missing values, it is more appropriate to apply the MICE algorithm, which takes relationships among attributes into account.

The findings of this study have important practical significance for improving data quality in information systems and enhancing the process of handling missing values. Furthermore, the results demonstrate the necessity of applying more advanced imputation algorithms and conducting broader experimental studies using real-world datasets in future research.

Based on the experimental findings, it is recommended to apply imputation algorithms according to the structure of the dataset and the proportion of missing values. For datasets with a relatively small amount of missing data, the KNN algorithm can be effectively utilized due to its high recovery accuracy. However, for datasets containing larger proportions of NaN values, the MICE algorithm is recommended because of its stable, reliable, and accurate performance. Furthermore, future studies should focus on deep-learning-based imputation approaches and broader experiments using real-world datasets in order to further improve data quality, analytical reliability, and the efficiency of intelligent information systems.

REFERENCES

1. Little, R. J. A., Rubin, D. B. Statistical Analysis with Missing Data. 3rd ed. — Hoboken, NJ: Wiley, 2019.
2. Van Buuren, S. Flexible Imputation of Missing Data. 2nd ed. — Boca Raton: CRC Press, 2018.
3. Batista, G. E. A. P. A., Monard, M. C. An analysis of four missing-data treatment methods for supervised learning // Applied Artificial Intelligence. — 2003. — Vol. 17. — pp. 519–533.
4. Troyanskaya, O., Cantor, M., Sherlock, G., et al. Missing-value estimation methods for DNA microarrays // Bioinformatics. — 2001. — Vol. 17(6). — pp. 520–525.
5. Zhang, S. Nearest-neighbor selection for iteratively KNN imputation // Journal of Systems and Software. — 2012. — Vol. 85. — pp. 2541–2552.
6. Rubin, D. B. Multiple Imputation for Nonresponse in Surveys. — New York: John Wiley & Sons, 1987.
7. García-Laencina, P. J., Sancho-Gómez, J. L., Figueiras-Vidal, A. R. Pattern classification with missing data: A review // Neural Computing and Applications. — 2010. — Vol. 19. — pp. 263–282.
8. Jerez, J. M., Molina, I., García-Laencina, P. J., et al. Missing-data imputation using statistical and machine-learning methods in a real breast-cancer problem // Artificial Intelligence in Medicine. — 2010. — Vol. 50. — pp. 105–115.
9. Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning. 2nd ed. — New York: Springer, 2009.
10. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. — 2011. — Vol. 12. — pp. 2825–2830.
11. Kuhn, M., Johnson, K. Applied Predictive Modeling. — New York: Springer, 2013.
12. Aggarwal, C. C. Data Mining: The Textbook. — Cham: Springer, 2015.
13. Fayzullo, N., Akbar, R., Yarmatov, S. Determining the number of effective distributions based on neural-network ensemble // International Journal of Intelligent Systems and Applications. — 2025. — Vol. 17(4). — pp. 69–77.
14. Yarmatov, S., Bustanov, X., Safarova, L. Optimization and improvement of reliability of machine-learning algorithms based on regularization methods // 2025 International Russian Smart Industry Conference (SmartIndustryCon). — Sochi, Russian Federation, 2025. — pp. 394–398.

Proofreader: Zokir ALIBEKOV

Layout and Designer: Oloviddin Sobir ugli

2026. № 5

To publish articles in this journal, you may submit articles, advertisements, stories, and other creative materials through the following links. Materials and advertisements are published on a paid basis.

You may subscribe to the journal at any time using the following details. Once subscribed, please send a screenshot or photo of your payment confirmation to our Telegram page @iqtisodiyot_77. Based on this, we will send the latest issue of the journal to your address each month.

“The journal “INNOVATION SCIENCE AND TECHNOLOGY” has been registered by the Agency for Information and Mass Communications under the Administration of the President of the Republic of Uzbekistan from 09.10.2024 under the registration number №390637. License number: C-5669633. PNFL: 30407832680027

Our address: Tashkent city, Yunusobod district, 19th block, House 17.