http://doi.org/10.35668/2520-6524-2025-3-07
УДК 303.732.4-048.62:004.9

A. O. MELNYK, PhD Student
I. V. ROZORA, D. Sc. in Physics and Mathematics

COMPARATIVE ANALYSIS OF IMPUTATION METHODS IN MACHINE LEARNING MODELS

Abstract. *Missing data is a prevalent issue in machine learning and data analysis that impacts the credibility and performance of predictive models. This article provides a comprehensive study of missing data, its types, consequences, and popular imputation methods. Using real datasets, we compare the performance of Mean/ Median Imputation, K-Nearest Neighbors (KNN) Imputation, Multiple Imputation, Regression Imputation, and Hot Deck Imputation. Furthermore, we study how these imputation techniques affect machine learning models such as Random Forest, Gradient Boosting Machines (GBM), and Support Vector Machines (SVM). Our study emphasizes the need for careful experimentation and model-specific investigation when handling missing data, where an important part is played by the selection of suitable imputation techniques based on dataset attributes and machine learning models. Lastly, our findings underscore the importance of tailored imputation strategies in enhancing model fit and ensuring stable analytical findings.*

Keywords: *missing data, imputation methods, machine learning, evaluation metrics, predictive models.*

INTRODUCTION

Missing data is a common problem in machine learning and data analysis, which can potentially pose serious consequences for the quality and precision of forecasting models. Missing data results from a variety of factors such as error in data entry, instrument malfunction, or participant non-response, among others. Effective handling of missing data is critical to ensuring validity and integrity of analytical results.

This work attempts to characterize the phenomenon of missing data and contrast different imputation approaches to address the issue. We shall discuss the types and impacts of missing data, characterize common imputation techniques, and compare their performance based on evaluation metrics. Further, we shall study the influence of imputation methods on the predictive capability of machine learning algorithms such as Random Forest, Gradient Boosting Machines (GBM), and Support Vector Machines (SVM).

By conducting experiments on real-world datasets and studying the results, we aim to provide insights into how different imputation techniques perform in enhancing prediction performance. We will also cover the importance of sound experimentation and model-specific evaluation in dealing with missing data and pointing to the need for researchers and practitioners to be cautious when selecting effective imputation techniques in relation to the nature of their datasets and the machines learning algorithms used.

UNDERSTANDING MISSING DATA

Missing data is a prevalent issue in any field's research, which implies that there is a lack of observations or measurements for a single or several variables in a data set [1; 2]. Its impact on research validity and reliability is profound.

Researchers are confronted with the problem of potentially biased estimates, less statistical power, and lowered generalizability because of missing data. Its ignoring or improper handling can lead to distorted results and compromised research credibility. It is important to understand its causes, which include participant non-response, equipment failure, and purposeful omission.

Missing data hamper statistical analysis, with the dangers of biased parameter estimates and power loss. Proper handling is required to minimize these dangers and maintain research validity. A variety of methods, ranging from the naive complete case analysis to more advanced techniques like multiple imputation, offer solutions, each with advantages and drawbacks.

Complete case analysis excludes cases with missing data, at risk of bias unless missingness is unrelated to outcome. Imputation methods like mean or regression imputation replace missing values with predictions from the data available. Multiple imputation generates multiple datasets, preserving uncertainty and improving estimates. Maximum likelihood estimation incorporates missing data directly, assuming an ignorable mechanism.

Types of Missing Data

When dealing with datasets, it's typical to come across instances where information is absent. This can happen due to a range of factors, such as mistakes during data gathering or inherent constraints of the data origin. However, it's important to note that not all missing data is the same, and recognizing the various types is essential for making well-informed choices and carrying out precise analyses. There are generally three types of missing data: missing completely at random, missing at random and missing not at random [3].

Missing Completely at Random (MCAR) refers to a situation in which the likelihood of a data point being absent is entirely random and not connected to any observed or unobserved factors [3]. In mathematical terms, MCAR can be represented as follows:

Let Y be the variable of interest, and R be the indicator variable denoting whether a value is observed $(R = 1)$ or missing $(R = 0)$. Then, MCAR can be represented as:

$$P(R = 0|Y, X) = P(R = 0), \quad (1)$$

where $P(R = 0|Y, X)$ is the conditional probability of the data being missing given the observed values of Y and any covariates X, $P(R = 0)$ is the marginal probability of the data being missing, which is constant across all observations.

Put more plainly, this equation means that the chance of data being absent is consistent across all observations, regardless of the values of the variable being studied Y or any other related factors X.

Missing at Random (MAR) represents a situation where the likelihood of a data point being absent relies on the observed data but not on the absent data itself [3]. In simpler terms, the probability of data being missing is linked to other variables in the dataset, rather than to the missing values specifically. Mathematically, MAR can be depicted as follows:

Let Y be the variable of interest, R be the indicator variable denoting whether a value is observed $(R = 1)$ or missing $(R = 0)$, and X be a set of observed variables. Then, MAR can be represented as:

$$P(R = 0|Y, X) = P(R = 0|X), \quad (2)$$

where $P(R = 0|Y, X)$ is the conditional probability of the data being missing given the observed values of Y and X, $P(R = 0|X)$ is the conditional probability of the data being missing given the observed values of X.

Put more plainly, this equation indicates that the chance of data being absent relies solely on the observed values of other variables in the dataset and does not relate to the missing values themselves.

Missing Not at Random (MNAR) is a category of missing data mechanism where the probability of a data point being absent is influenced by the missing values themselves, even after considering the observed data. In essence, this implies that the likelihood of data being missing is associated with information not included in the dataset [3]. Mathematically, MNAR can be described as follows:

Let Y be the variable of interest, R be the indicator variable denoting whether a value is observed $(R = 1)$ or missing $(R = 0)$, and X be a set of observed variables. Then, MNAR can be represented as:

$$P(R = 0|Y, X) \neq P(R = 0|X), \quad (3)$$

where $P(R = 0|Y, X)$ is the conditional probability of the data being missing given the observed values of Y and X, $P(R = 0|X)$ is the conditional probability of the data being missing given the observed values of X.

Put more straightforwardly, this equation suggests that the chance of data being absent is not only influenced by the observed values of other variables in the dataset but also by the missing values themselves.

Impact of Missing Data

Missing data can possibly alter analysis outcomes, especially if the missingness is not at random. Such biases have the result of distorting aspects of the target population, compromising validity and applicability of outcomes. For example, if there is an increased prevalence of missing data among certain demographic groups, the results can, albeit unintentionally, prejudice or favor certain subgroups, making incorrect conclusions [4].

Missing data diminishes the power of analysis statistically, and it becomes harder for real effects or associations in the data to be made apparent. The power decrease makes it more probable for Type II errors, as the researchers cannot find patterns or associations that are present because there is inadequate data. Therefore, the ability to form strong conclusions from the analyses is compromised, and scientific progress is hindered.

Inadequate handling of missing data can lead to unwarranted conclusions. If missing data persons are not the same in some respects as those with complete data, then conclusions based on the analysis will lack generalizability. Differences contaminate research results and trust in conclusions drawn from the dataset.

Absence of data results in the loss of vital information, lowering validity and richness of analysis. The absence does not allow researchers to be in a position to establish underlying trends and patterns within the dataset, thereby inhibiting comprehensive understanding of the phenomena being

researched. The greatly significant implications of studies are therefore masked or poorly analyzed due to an absence of vital points of data.

COMMON IMPUTATION METHODS

Imputation statistically involves the substitution of missing data with estimated values, a standard process applied in data analysis and machine learning when working with incomplete data. Ranging from straightforward methods of filling missing values with the mean or median of available data to sophisticated approaches such as k-nearest neighbors imputation or predictive modeling, imputation methods depend on data complexity and underlying distribution. The objective of imputation is to enable the analysis of datasets or modeling while reducing the impact of missing data on the outcome [5].

Mean/Median Imputation

Mean and median imputation are very popular in data preprocessing for handling missing values, which is a frequent phenomenon in data analysis and may cause difficulties for some algorithms and statistical methods [6]. They offer simple but effective solutions by replacing missing values with the mean or median of the available data.

Mean imputation fills missing values with the mean of the observed values for a given feature, thereby maintaining data distribution and central tendency. It is appropriate for normally distributed data or where the presence of outliers is minimal. Median imputation, on the other hand, fills missing values with the median, which is resistant to outliers and is ideal for skewed distributions.

Both are straightforward to apply and computationally efficient, which makes them attractive to use in ensuring dataset integrity. Nevertheless, they can introduce bias when missing values are not missing at random and underestimate variability in the data, which can influence downstream analyses. They also might not be appropriate for categorical or time-series data, where other approaches such as mode imputation or forward/backward filling are more appropriate.

K-Nearest Neighbors (KNN) Imputation

K-Nearest Neighbors (KNN) imputation is an approach that aims to handle missing values through the exploitation of data point similarity, in contrast to mean or median imputation, which are based only on summary statistics [7]. In contrast, KNN imputation infers missing values based on neighboring data points.

This approach consists of locating the K most nearest data points (neighbors) to the observation with missing values, generally based on distance metrics like Euclidean distance. The missing value is then imputed by calculating the average or weighted average of these neighboring values.

One of the significant benefits of KNN imputation is its ability to identify complex patterns in data, which is especially useful in high-dimensional or nonlinear data. It is also flexible, supporting both numerical and categorical variables.

But choosing the best K value is a challenge since it controls the trade-off between overfitting and underfitting. In addition, KNN imputation can be computationally expensive, especially on large datasets. Irregularly or sparsely distributed data can also undermine its efficiency since it becomes difficult to find significant neighbors. The selection of the distance metric also greatly affects its performance and differs based on dataset features and missing data.

Multiple Imputation

Multiple imputation is a method used in missing data management in that more than one value for each missing observation is produced, thereby injecting uncertainty into the process. Unlike single imputation methods that produce a single imputed value for a missing data point, multiple imputation produces a number of imputed datasets [8].

The process includes three stages: imputation, analysis, and pooling. During the imputation stage, missing values are repeatedly imputed using model-based techniques such as linear regression or logistic regression. During the analysis stage, each full data set is analyzed individually, yielding multiple sets of parameter estimates. Finally, each imputed data set's results are combined using Rubin's rules to produce complete parameter estimates, standard errors, and p-values.

There are the strengths of multiple imputation in handling uncertainty and preserving relationships among variables. It is able to handle missing data effectively under the missing at random (MAR) assumption. Its effectiveness may rely on the goodness of imputation models and the MAR assumption adherence, however.

Regression Imputation

Regression imputation is a missing data handling technique that imputes missing values based on predicted values from relationships found in the data. It involves the process of model fitting to available data using a regression model and then using it to predict missing values for the variable in question [9].

In regression imputation, the variable containing missing values is the dependent variable, and variables with complete information are independent variables for estimation. The regression model complexity can range from simple linear regression to a more complex model based on the characteristics of the dataset and the relationships between variables.

Regression imputation is successful in picking up inherent patterns in the data and correlations, providing more precise estimates than simpler methods such as mean or median imputation. Further, it is possible to include other predictors, and thus it is appropriate for data sets that have complicated relationships.

Still, regression imputation presumes constant relationships among variables in various subsets of data, which might not always hold true. It can be influenced by outliers and multicollinearity as well, which could undermine the quality of imputation. In instances of nonlinear relationships or limited models, more flexible approaches such as nonparametric regression or machine learning could be more suitable.

Hot Deck Imputation

Hot deck imputation is a technique for handling missing data that involves the replacement of missing values with observed values of similar cases within the data. The technique utilizes the search for complete cases that are close to the case with missing data and imputing the missing value by using observed values from the similar cases [10].

There are several types of hot deck imputation, such as nearest neighbor hot deck imputation and random hot deck imputation. Nearest neighbor imputation uses the observed value from the most similar complete case, while random hot deck imputation imputes by randomly selecting a complete case.

One advantage of hot deck imputation is that it is straightforward and preserves dataset structure and variable relationships. It may be particularly useful for categorical variables or datasets with complex missingness patterns. Hot deck imputation is also efficient at handling missing data in real-time settings.

But it relies on the assumption that similar cases have similar values and may be influenced by the choice of similarity measure. It may not perform optimally when there are few complete cases available for imputation or the mechanism of missing data is non-ignorable.

EVALUATION METRICS FOR IMPUTATION METHODS

Evaluation measures are quantifiable values used to measure the efficiency or performance of a model, algorithm, system, or procedure. In data science, machine learning, and computer science, these metrics play a crucial role in identifying how well a system or model is performing compared to its desired goals [11].

In machine learning, these evaluation metrics are utilized to gauge the performance of predictive models. Accuracy, precision, recall, F1 score, area under the ROC curve (AUC-ROC), and mean squared error (MSE) are some of the common evaluation metrics used for this purpose, among others. These metrics assist in gauging the predictive power of a model, areas of improvement, and comparing various models or algorithms.

Likewise, for specific areas such as information retrieval, natural language processing, and optimization, there exist bespoke measures of evaluation that are used in assessing the performance of algorithms or systems by the individual goals. Such measures of evaluation give useful feedback regarding the strengths and weaknesses of the systems under test, which can be used to inform further development and tuning.

Mean Absolute Error (MAE)

Mean Absolute Error (MAE) functions as a tool for assessing the effectiveness of either a predictive model or an imputation technique. It calculates the average absolute disparity between the predicted or imputed values and the actual values [12].

The formula for Mean Absolute Error (MAE) is:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \quad (4)$$

where n stands for the number of samples or observations, y_i indicates the actual or true value and $\hat{y}_i$ signifies the predicted or imputed value for the i^{th} sample.

MAE offers an advantage by giving equal weight to all errors, irrespective of their size. It is presented in the same units as the original data, thereby enhancing interpretability and facilitating comprehension.

For instance, when assessing an imputation method for missing data, a lower MAE suggests superior performance, indicating that the imputed values are, on average, closer to the actual values.

Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) is another commonly utilized metric for evaluating the effectiveness of predictive models or imputation methods. It computes the square root of the average of the squared differences between the predicted or imputed values and the actual values [12].

The formula for Root Mean Squared Error (RMSE) is:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \quad (5)$$

where n stands for the number of samples or observations, y_i indicates the actual or true value and $\hat{y}_i$ signifies the predicted or imputed value for the i^{th} sample.

RMSE imposes a heavier penalty on larger errors compared to smaller ones because of the squaring process. It offers an indication of the

dispersion of errors and is notably responsive to outliers.

Like MAE, decreased RMSE values indicate improved performance, suggesting that the predicted or imputed values are generally closer to the actual values. However, RMSE cannot be directly interpreted in the original data units because it involves squaring the errors.

MACHINE LEARNING MODELS

Machine learning, a branch of artificial intelligence (AI), involves the development of algorithms and techniques that allow computers to learn from data and make predictions or decisions independently, without being explicitly programmed to do so for particular tasks [13]. In contrast to traditional programming, in which humans develop code to instruct computers to solve problems or perform tasks, machine learning algorithms learn patterns and relationships directly from data. As they are exposed to additional data, the algorithms improve their performance, making accurate predictions or decisions without requiring tailored programming for every scenario.

Conversely, machine learning models are mathematical or computational algorithms trained on data to identify patterns, relations, or behavior. These models find application in predicting, deciding, or inferring on new or unseen data [14]. Fundamentally, a machine learning model develops from training a machine learning algorithm with a specified dataset, and it tunes its parameters or coefficients to reduce the gap between its predictions and actual results in the training data. After training, the model can generalize its predictive power to new data that it has not seen before.

Random Forest

Random Forest is an ensemble learning method in which a large number of decision trees are constructed during training, and the output is the mode of classes for classification or the mean prediction for regression problems [15]. The "random" in Random Forest comes from two major sources: the random sampling of feature subsets for each tree and the random sampling of training instances with replacement. The randomness works to decorrelate individual trees, preventing overfitting and making the model generalize better. Random Forest also provides a feature importance estimate, facilitating feature selection and understanding data relationships underlying the data. Its scalability also makes it efficient to deal with large, high-dimensional data.

Celebrated for its strength and versatility, Random Forest is useful on a range of machine learning problems. Its ability to generate several decision trees and combine their predictions reduces overfitting, especially useful for complex datasets, and its feature importance estimation makes it easy to gain insights from data and tune models. Random Forest's scalability feature is especially useful in this age of big data where dealing with large volumes of data is typical. Its ability to handle high-dimensional data makes it a popular option for data scientists and machine learning enthusiasts.

Gradient Boosting Machines (GBM)

Gradient Boosting Machines (GBM) is another form of ensemble learning in which a chain of weak learners, typically decision trees, are learned sequentially, and each subsequent learner tries to compensate for the faults of the previously learned ones [16]. GBM's core concept is optimizing a differentiable loss function using greedily added weak learners. Extremely flexible, GBM can optimize various loss functions and can be used with various types of data. One of the strong points of GBM is that it can detect fine-grained patterns and interactions in the data, thus guaranteeing good predictive accuracy.

But GBM training does come at a computational cost and needs proper hyperparameter selection to prevent overfitting. Techniques such as regularization, learning rate calibration, and early stopping are usually employed to address these concerns. Despite its computational cost, GBM remains very popular in practice and has continued to deliver state-of-the-art results in machine learning competitions and real-world applications.

Essentially, GBM is a robust ensemble learning algorithm that is appreciated for its capacity to learn complex data relationships. Through the refinement of previous errors in successive iterations, GBM ensures optimal performance from a model, offering flexibility and robust predictive abilities across numerous domains.

Support Vector Machines (SVM)

Support Vector Machines (SVM) is an effective supervised learning technique applied in classification, regression, and outlier detection problems. SVM aims to find the best hyperplane in the feature space with a maximum margin between various classes to achieve good generalization performance. When data is not linearly separable, SVM uses kernel functions to transform the input features to a higher dimensional space to enable class separation. This trait enables SVM to be highly effective in dealing with datasets that have many features and performing well in high-dimensional spaces [17].

One of the inherent strengths of SVM is its robustness against overfitting, owing to its capacity to control the margin and enforce appropriate regularization. However, it is worth mentioning here

that SVM training can be computationally demanding, especially with large datasets, and the choice of kernel and regularization parameters has a large impact on model performance.

Despite these considerations, SVM maintains its popularity across diverse classification tasks, especially in scenarios prioritizing interpretability and robust generalization. Its adeptness at navigating complex datasets and furnishing dependable predictions underscores its enduring relevance within the realm of machine learning.

EXPERIMENT

In this section, we delineate the methodology employed to conduct an extensive evaluation of the selected imputation techniques. Subsequently, we delve into the comparison of machine learning models utilizing the imputed datasets. This comprehensive experimental setup aims to elucidate the efficacy of different imputation strategies and the subsequent impact on the predictive performance of various machine learning algorithms.

Dataset

The dataset utilized for this experiment is the "Credit Card Fraud Detection Dataset 2023" [18]. This dataset comprises credit card transactions conducted by European cardholders throughout the year 2023. With over 36,000 records, the dataset offers a substantial volume of transactions for analysis. Notably, the data has undergone anonymization to safeguard the identities of the cardholders.

Here, we have a total of 29 features (predictor variables) and 1 outcome variable, which is the transaction amount. The predictor variables encompass various anonymized transaction attributes, while the outcome variable serves as the target for predictive modeling tasks.

Programming

Python and its libraries were used to write the programs. Python, along with pandas, numpy, and sklearn, is the backbone of modern data analysis and machine learning. Pandas simplifies data manipulation, numpy facilitates fast numerical computing, and sklearn streamlines machine learning model building. Together, they empower users to extract insights and build predictive models efficiently, making Python the language of choice for data-driven tasks [19].

The comparative analysis of imputation methods in machine learning models involves nine programs that explore different techniques to handle missing data in datasets.

The initial program loads the dataset, displays its size and structure, and calculates missing values. It introduces random missing values to simulate incomplete data, ensuring each row retains at least one non-missing value. The data is split into training and validation sets, and both the modified and original datasets are saved for future reference.

Following data download, mean imputation is applied to fill missing values by replacing them with the mean of their respective columns. The imputed datasets are saved for further analysis.

Another program implements K-Nearest Neighbors imputation, which fills missing values by considering neighboring data points. After imputation, the datasets are saved for subsequent analysis and modeling.

Multiple imputed datasets are generated using statistical methods or predictive models to handle missing entries. The resulting datasets are saved for further use in analysis and modeling tasks.

Regression-based imputation is employed to predict missing values using a regression model. After imputation, complete datasets are used to build a linear regression model, which is then utilized to predict and fill missing values in both training and validation sets.

Hot Deck imputation fills missing values by identifying observations with similar characteristics and using their values. The process is applied consistently to both training and validation datasets, ensuring data integrity.

The program loads datasets with imputed values and target values for training a RandomForestRegressor model. A function facilitates model training and metric calculation for each imputation method, enabling performance comparison.

Similarly, a GradientBoostingRegressor model is utilized for predictive modeling, assessing performance across different imputation techniques.

Finally, Support Vector Regressor (SVR) model is utilized for predictive modeling, with a focus on handling non-linear relationships within the data. Evaluation metrics are computed to compare the performance of SVR across different imputation methods.

Result Analysis

The tables present the results of evaluating different imputation methods across three machine learning models: Random Forest, Gradient Boosting Machines, and Support Vector Machines. Two metrics, Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE), are used to measure the performance of these combinations.

In **Table 1** (RMSE), Mean Imputation consistently yields the highest RMSE values across all three models. This suggests that replacing missing values with the mean of the observed data leads to less accurate predictions compared to other imputation techniques. On the other hand, Hot Deck Imputation stands out by consistently providing the

Table 1

RMSE for Different Imputation Methods and Models

Imputation / Model	Random Forest	Gradient Boosting Machines	Support Vector Machines
Mean Imputation	7112,92	6985,28	6967,43
KNN Imputation	7209,07	6981,96	6967,34
Multiple Imputation	7088,89	6995,17	6967,31
Regression Imputation	7112,92	6985,28	6967,43
Hot Deck Imputation	7086,02	6983,79	6967,42

Table 2

MAE for Different Imputation Methods and Models

Imputation / Model	Random Forest	Gradient Boosting Machines	Support Vector Machines
Mean Imputation	6138,00	6071,56	6059,31
KNN Imputation	6188,43	6069,24	6059,18
Multiple Imputation	6125,22	6079,80	6059,16
Regression Imputation	6138,00	6071,56	6059,31
Hot Deck Imputation	6116,96	6067,78	6059,27

lowest RMSE values, particularly notable when applied with Support Vector Machines. This indicates that Hot Deck Imputation effectively preserves the underlying structure of the data, resulting in improved predictive accuracy.

The performance of KNN and Multiple Imputation techniques varies across models, with no clear dominance observed. While these methods show mixed performance, they offer viable alternatives, particularly in scenarios where Hot Deck Imputation may not be applicable or feasible. Regression Imputation, meanwhile, displays RMSE values similar to Mean Imputation, implying comparable performance in terms of predictive accuracy.

Moving to **Table 2** (MAE), similar trends emerge. Mean Imputation consistently exhibits higher MAE values across all models, indicating larger average errors in predictions. Conversely, Hot Deck Imputation consistently demonstrates lower MAE values, reaffirming its effectiveness in improving predictive accuracy across different machine learning models.

KNN and Multiple Imputation methods exhibit varied performance across models, emphasizing the need for careful consideration of their suitability based on specific modeling contexts. Regression Imputation, like in RMSE analysis, shows MAE values akin to Mean Imputation, suggesting similar performance.

Overall, these findings underscore the importance of selecting appropriate imputation methods tailored to the characteristics of both the dataset and the machine learning model being employed. While Hot Deck Imputation consistently outperforms other methods in terms of both RMSE and MAE, the variability in performance among alternative techniques highlights the necessity of thoughtful experimentation and model-specific evaluation when handling missing data.

CONCLUSION

Based on the analysis conducted in this study, it is evident that the choice of imputation method significantly influences the performance of machine learning models in handling missing data. Across various evaluation metrics and machine learning algorithms, Hot Deck Imputation consistently emerges as the most effective technique, yielding lower Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) values compared to other methods such as Mean Imputation, KNN, Multiple Imputation, and Regression Imputation.

The superior performance of Hot Deck Imputation can be attributed to its ability to preserve the underlying structure of the data, leading to more accurate predictions across different modeling scenarios. This highlights the importance of considering the characteristics of both the dataset and the machine learning model when selecting an appropriate imputation method.

While KNN and Multiple Imputation methods exhibit mixed performance across models, they still

present viable alternatives, particularly in situations where Hot Deck Imputation may not be feasible or applicable. Regression Imputation, although offering comparable performance to Mean Imputation, may not be the most suitable choice for improving predictive accuracy.

Overall, the findings emphasize the necessity of thoughtful experimentation and model-specific evaluation when dealing with missing data. Researchers and practitioners should carefully assess the suitability of different imputation methods based on the specific characteristics of their datasets and the machine learning algorithms employed. By selecting the most appropriate imputation technique, they can enhance the predictive performance of their models and ensure more robust and reliable results in data analysis and decision-making processes.

Additionally, future research endeavors could explore the development of novel imputation techniques or the integration of advanced machine learning algorithms specifically tailored to handle missing data, aiming to further improve predictive accuracy and enhance the robustness of predictive models across diverse domains and datasets.

REFERENCES

1. Auria, L., & Moro, R. A. (2008). Support vector machines (SVM) as a technique for solvency analysis. *German Institute for Economic Research.* Berlin, 19 p. Retrieved from: https://www.econstor.eu/bitstream/10419/27334/1/576821438.PDF.
2. Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual review of psychology, 60*, 549-576.
3. Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581-592.
4. Groenwold, R. H., & Dekkers, O. M. (2020). Missing data: the impact of what is not there. *European journal of endocrinology, 183* (4), E7-E9.
5. Chhabra, G., Vashisht, V., & Ranjan, J. (2017). A comparison of multiple imputation methods for data with missing values. *Indian Journal of Science and Technology, 10* (19), 1-7.
6. Engels, J. M., & Diehr, P. (2003). Imputation of missing longitudinal data: a comparison of methods. Journal of clinical epidemiology, 56(10), 968-976.
7. Malarvizhi, R., & Thanamani, A. S. (2012). K-nearest neighbor in missing data imputation. *International Journal of Engineering Research and Development, 5* (1), 5-7.
8. Feelders, A. (1999, September). *Handling missing data in trees: Surrogate splits or statistical imputation?* In European Conference on Principles of Data Mining and Knowledge Discovery (pp. 329-334). Berlin, Heidelberg.
9. Templ, M., Kowarik, A., & Filzmoser, P. (2011). Iterative stepwise regression imputation using standard and robust methods. *Computational Statistics & Data Analysis, 55* (10), 2793-2806.
10. Sullivan, D., & Andridge, R. (2015). A hot deck imputation procedure for multiply imputing nonignorable missing data: The proxy pattern-mixture hot deck. *Computational statistics & data analysis, 82*, 173-185.
11. Jäger, S., Allhorn, A., & Bießmann, F. (2021). A benchmark for data imputation methods. *Frontiers in big Data, 4*, 693674.
12. Hodson, T. O. (2022). Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development Discussions, 2022,* 1-10.
13. Alpaydin, E. (2021). Machine learning. Mit Press.
14. Mosavi, A., Ozturk, P., & Chau, K. W. (2018). Flood prediction using machine learning models: Literature review. *Water, 10* (11), 1536.
15. Rigatti, S. J. (2017). Random forest. *Journal of Insurance Medicine, 47* (1), 31-39.
16. Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics, 7,* 21.
17. Auria, L., & Moro, R. A. (2008). Support vector machines (SVM) as a technique for solvency analysis.
18. (2023) Credit Card Fraud Detection Dataset 2023. Kaggle. Retrieved from: https://www.kaggle.com/datasets/nelgiriyewithana/*credit*-card-fraud-detection-dataset-2023.
19. Saabith, A. S., Vinothraj, T., & Fareez, M. (2020). Popular python libraries and their application domains. *International Journal of Advance Engineering and Research Development, 7* (11).

А. О. МЕЛЬНИК, аспірантка
І. В. РОЗОРА, д-р фіз.-мат. наук

ПОРІВНЯЛЬНИЙ АНАЛІЗ МЕТОДІВ ІМПУТАЦІЇ В МОДЕЛЯХ МАШИННОГО НАВЧАННЯ

Резюме. *Відсутність даних є поширеною проблемою в аналізі даних і машинному навчанні, що впливає на точність і надійність прогнозних моделей. У цій статті здійснено комплексний аналіз проблеми відсутніх даних, досліджено їхні типи, наслідки та поширені методи імпутації. Використовуючи реальні набори даних, здійснено оцінювання ефективності серединної / медіанної імпутації, імпутації методом k-найближчих сусідів, множинної імпутації, регресійної імпутації та імпутації методом "hot deck" ("гаряча колода"). Окрім того, досліджено те, як ці методи імпутації впливають на моделі машинного навчання, зокрема метод випадкового лісу, градієнтно-бустерні машини та метод опорних векторів. У статті підкреслено необхідність ретельного експериментування та оцінки конкретної моделі під час обробки відсутніх даних, наголошено на критичній ролі вибору відповідних методів імпутації залежно від характеристик набору даних та алгоритмів машинного навчання. Зрештою, висновки підкреслюють важливість адаптованих стратегій імпутації для підвищення продуктивності моделі та забезпечення надійних аналітичних результатів.*

Ключові слова: *відсутні дані, методи імпутації, машинне навчання, оціночні метрики, прогнозні моделі.*

INFORMATION ABOUT THE AUTHORS

Melnyk A. O. — PhD Student, Taras Shevchenko National University of Kyiv, 4d Akademika Glushkova Avenue, Kyiv, Ukraine, 02000; anastasiia.melnyk@knu.ua; ORCID: 0000-0002-3167-4353

Rozora I. V. — D. Sc. in Physics and Mathematics, Taras Shevchenko National University of Kyiv, 4d Akademika Glushkova Avenue, Kyiv, Ukraine, 02000; +38 (044) 521-35-35; irozora@knu.ua; ORCID: 0000-0002-8733-7559

ІНФОРМАЦІЯ ПРО АВТОРІВ

Мельник Анастасія Олександрівна — аспірантка, Київський національний університет імені Тараса Шевченка, просп. Академіка Глушкова 4д, м. Київ, Україна, 02000; anastasiia.melnyk@knu.ua; ORCID: 0000-0002-3167-4353

Розора Ірина Василівна — д-р фіз.-мат. наук, Київський національний університет імені Тараса Шевченка, просп. Академіка Глушкова 4д, м. Київ, Україна, 02000; +38 (044) 521-35-35; irozora@knu.ua; ORCID: 0000-0002-8733-7559

Надійшла до редакції 18.08.2025

http://doi.org/10.35668/2520-6524-2025-3-08
УДК 001.18; 001.89; 002.513.5; 330.3

Т. К. КВАША, заввідділу
О. І. КОВАЛЕНКО, с. н. с.

ПІДХОДИ ДО ІДЕНТИФІКАЦІЇ ПЕРСПЕКТИВНИХ НАПРЯМІВ І ПРОГНОЗНІ ТЕНДЕНЦІЇ У СФЕРІ БРОНЕМАТЕРІАЛІВ

Резюме. *Воєнні конфлікти, що супроводжуються масовими кульовими, вогневими та уламковими пораненнями, а також стрімкий розвиток у галузі вибухових речовин, боєприпасів, стрілецької зброї зумовлюють зростання балістичної загрози внаслідок недостатнього ступеня захисту військового контингенту та цивільного населення. Виникає нагальна потреба у створенні нових видів високоміцних матеріалів із розширеними функціональними властивостями, що забезпечать високий рівень бронестійкості елементів засобів індивідуального захисту, військової техніки, особливо під час війн.*

Мета дослідження полягає у визначенні наукових трендів у сфері броне- та захисних матеріалів, що формують нові стандарти мобільності, надійності й адаптивності захисних рішень у реальних бойових умовах.

У статті викладено результати дослідження щодо глобальних наукових трендів в цій сфері на основі аналізу публікацій міжнародної бази Web of Science. Результати підкреслюють наростальний інтерес до досліджень, що пов'язані з технологіями створення броньових і захисних матеріалів. Зокрема визначено тенденції наукових публікацій, продуктивні країни/регіони та джерела публікацій, основні теми досліджень. Дослідження охоплює систематизовану добірку актуальних напрямів досліджень щодо технологій бронезахисту та оцінку ключових наукових викликів, пов'язаних із розробленням майбутніх матеріалів.

Підсумовано, що перспективи розвитку бронезахисних матеріалів полягають у глибокій інтеграції багатошарових, гібридних матеріалів, які поєднують кераміку, полімери, метали та інтелектуальні рішення. Інтеграція оптичних і телекомунікаційних технологій у бронезахист підвищує загальну ефективність бойових систем, забезпечуючи надійний зв'язок, оперативний контроль і точне керування. Такий підхід забезпечує оптимальний баланс між захистом, вагою і гнучкістю.

Ключові слова: *критичні технології, бронезахисні матеріали, тренди майбутнього, короткостроковий прогноз, термостійка кераміка, інтелектуальні матеріали, композити, високотехнологічні сплави, нанопокриття.*