Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

The critical role of evaluation metrics in handling missing data in machine learning

Ibrahim Atoum *

Department of Artificial Intelligence, Faculty of Science and Information Technology, Al-Zaytoonah University of Jordan, Amman, Jordan

ARTICLE INFO

Article history:
Received 3 September 2024
Received in revised form
25 December 2024
Accepted 5 January 2025

Keywords:
Missing data handling
Machine learning models
Imputation techniques
Data completeness
Model performance evaluation

ABSTRACT

The presence of missing data in machine learning (ML) datasets remains a major challenge in building reliable models. This study explores various strategies to handle missing data and provides a framework to evaluate their effectiveness. The research focuses on commonly used techniques such as zero-filling, deletion, and imputation methods, including mean, median, mode, regression, k-nearest neighbors (KNN), and flagging. To assess these methods, a detailed evaluation framework is proposed, considering factors such as data completeness, model performance, stability, bias, variance, robustness to new data, computational efficiency, and domain-specific needs. This comprehensive approach allows for a thorough comparison of methods, helping to identify the most suitable technique for specific datasets and tasks. The findings highlight the importance of considering the unique features of the dataset and the goals of the analysis when choosing a method. While basic techniques like deletion and zero-filling may be effective in some cases, advanced imputation methods often preserve data quality and improve model accuracy. By applying the proposed evaluation criteria, researchers and practitioners can make better decisions on handling missing data, leading to more accurate, reliable, and adaptable ML models.

1. Introduction

The quality of datasets constitutes a pivotal component in the training, validation, and testing of ML models, ultimately enabling their effective generalization and performance (Alzubaidi et al., 2023; Sharma et al., 2023; Liu et al., 2023). In addition, datasets of superior quality accurately mirror real-world variation, enhancing the model's capacity to generalize beyond its training examples and empowering models to identify patterns and make precise predictions (Shamji et al., 2023; Pagano et al., 2023). Furthermore, datasets play a pivotal role in feature extraction (Kadhim and Radhi, 2023) and are indispensable in addressing biases, ensuring fairness, and preventing the amplification of societal biases (Pagano et al., 2023).

Missing values in datasets can occur due to errors, non-response, or technical difficulties. These missing values may be sporadic or widespread, and their patterns can vary. Ultimately, the characteristics of datasets significantly impact the effectiveness of ML models (Albahri et al., 2023; Buttia et al., 2023; Mitra et al., 2023).

When assessing the potential ramifications of missing data on ML models, it is imperative to comprehend the underlying reasons for its absence. To this end, these data can be classified into three distinct categories (Jaradat et al., 2024): Missing completely at random (MCAR), where the absence of data is entirely random; missing at random (MAR), where the missingness is correlated with observed data; and not missing at random (NMAR), where the missingness is correlated with the unobserved data itself. A comprehensive understanding of these categories is instrumental in devising effective strategies for handling missing data.

The possession of key characteristics is indispensable for ensuring the accuracy and reliability of model training and predictions. Firstly, the relevance and appropriateness of the datasets employed for the given problem or task constitute essential considerations. Secondly, the overall quality and reliability of the datasets, encompassing factors such as accuracy, completeness, and consistency, are paramount. Thirdly, the diversity of

* Corresponding Author.
Email Address: i.atoum@zuj.edu.jo

https://doi.org/10.21833/ijaas.2025.01.011

Corresponding author's ORCID profile:
https://orcid.org/0000-0002-9259-7937

2313-626X/© 2025 The Authors. Published by IASE.

the datasets ensures that they capture a broad spectrum of perspectives and scenarios. Furthermore, the sufficiency of the datasets provides adequate coverage and representation of the problem domain (Parhi and Patro, 2023; Liguori et al., 2023).

Additionally, the representativeness of the datasets guarantees that they accurately reflect the real-world distribution and characteristics of the data (Agarwal, 2023). Finally, the balance of the datasets is crucial, avoiding biases or skewed distributions that could adversely impact model performance. Moreover, the fundamental availability and accessibility of high-quality datasets that can be effectively utilized for model training and evaluation are essential.

Standard techniques employed to impute missing values in datasets encompass zero-filling, deletion, and imputation utilizing the mean, median, or mode. While zero-filling presents a simplistic approach, it can introduce biases and distort distributions. On the other hand, deletion preserves the observed statistics but results in a reduction of sample size. Central tendency imputation methods, such as the mean, median, and mode, capitalize on informational patterns, maintain the structural integrity of the data, and are adaptable to various variables. However, these methods may still introduce biases and distributional shifts.

The limitations inherent in simple replacement methods necessitate the exploration of more advanced calculations, such as regression or KNN, to address complex scenarios. Regression offers a plethora of advantages, including predictive power, flexibility, interpretability, and scalability. However, it has drawbacks, such as data assumptions, sensitivity to outliers, multicollinearity issues, and less effectiveness when dealing with categorical variables. In contrast, KNN represents a non-parametric approach capable of handling numerical and categorical variables. By leveraging local information, KNN offers a relatively simple solution. However, it may encounter challenges when dealing with high-dimensional datasets, be sensitive to the choice of k, and can be computationally complex.

A plethora of studies have delved into the complexities associated with missing data within the realm of ML, offering practical methodologies to address these challenges. These studies underscore the paramount importance of selecting suitable approaches based on the unique characteristics of the data to optimize outcomes in ML tasks (Emmanuel et al., 2021; Palanivinayagam and Damaševičius, 2023; Abidin et al., 2018).

Several publications have employed imputation methods to handle missing values in datasets for ML tasks effectively (Li et al., 2024; Huang et al., 2024). Notably, certain publications have indicated that in specific domains of ML, the adoption of more straightforward imputation methods could potentially yield superior benefits due to the distinctive characteristics involved (Chen and McCoy, 2024; Liu et al., 2024; Sierra-Porta, 2024). Some researchers have posited that simple mean imputation can more effectively address missing data in ML portfolios characterized by cross-sectional predictors than complex methodologies (Liu et al., 2024).

Other studies have explored the intricacies of handling missing data and imbalanced classes within the context of ML to predict consumer preferences. These studies discuss various techniques and emphasize the criticality of selecting the most appropriate approach for specific data and objectives. A noteworthy study proposes a system that synergizes deep learning and dead reckoning to address missing AIS data within maritime traffic monitoring effectively. This innovative approach aims to enhance the accuracy and robustness of real-time vessel tracking systems (Sedaghat et al., 2024).

Imputation techniques are vital in preserving sample size and dataset integrity, facilitating effective analysis. Common approaches include mean/median imputation, regression imputation, and KNN. Although zero-imputation may appear convenient, it has the potential to introduce biases. Deletion methods, while simplifying the process, can also reduce sample size and lead to biases. Imputation helps retain important information but comes with a degree of uncertainty. The selection of an imputation method should be based on the specific dataset and the problem being addressed. Employing a structured evaluation framework is crucial to ensure the selected technique's effectiveness and to validate results through iterative experimentation. The article evaluates each technique based on its simplicity, noise reduction, efficiency, sample size reduction, bias introduction, loss of information, and preservation of statistical properties.

Assessing techniques for handling missing data necessitates a thorough approach. Important metrics include data completeness, model performance, stability, bias, variance, robustness, computational efficiency, and factors specific to the domain. These metrics are instrumental in evaluating the effectiveness of various techniques, considering trade-offs and alignment with project objectives. Although these metrics have their advantages, they also come with limitations. Multiple metrics can yield a more holistic understanding but may introduce computational complexity.

The article seeks to simplify the evaluation metrics of ML models through clear mathematical notations and a comprehensive explanation of how to evaluate their effectiveness—whether positive, negative, or zero. It provides a robust framework for understanding these metrics in practical contexts, supported by compelling real-world examples illustrating their importance. This contribution stands out from others (Nezami et al., 2024; Pagano et al., 2023; Kazemi et al., 2024; Albahri et al., 2023; Munshi, 2024) by clarifying and including the metrics in actual use cases. In doing so, it enables readers to understand their importance in real-world applications, enhances their understanding of model performance evaluation, and establishes itself as an essential resource for researchers and practitioners navigating the complexities of ML.

This article focuses on exploring how missing data can impact model performance and interpretability within the realm of ML. By considering the specific characteristics of the data and the task at hand, this article delves into strategies for evaluating various approaches to handling missing data. Rather than focusing on other areas, the article concentrates on the potential consequences of missing data on model performance and interpretability, offering valuable insights for researchers and practitioners.

This discourse is structured into Three distinct sections: Section One presents the analysis of missing date techniques. Section Two presents a comparative analysis of missing data mitigation techniques tailored explicitly for ML Datasets. Finally, Section Three concludes with a comprehensive summary of the findings and future works.

2. Navigating missing data: A comparative analysis of handling techniques

Imputation techniques, designed to estimate missing values within datasets, offer the significant advantage of preserving sample size and ensuring the integrity of the dataset for subsequent analysis. Common strategies employed for this purpose include mean/median imputation, mode imputation, regression imputation, KNN imputation, and flagging (Aubaidan et al., 2024). The optimal choice among these techniques is contingent upon the specific characteristics of the dataset and the nature of the problem at hand. Therefore, a judicious consideration of the data context and the goals of the analysis is essential to make an informed decision.

While the expedient approach of zero-imputation may initially appear viable, it is imperative to exercise caution as it can inadvertently introduce biases or distortions into the analysis. Before its application, a meticulous examination of the dataset and the underlying causes of missing values is imperative. Zeros may need to accurately represent the true nature of missing data, thereby leading to erroneous conclusions. Moreover, the domain-specific context and significance of the data must be carefully considered. For instance, within medical data (Tahyudin et al., 2024), a zero value for a vital sign such as blood pressure might accurately reflect a critical health concern. However, in the context of financial data, a zero-stock price could signal a significant event, such as a company bankruptcy (Alzyadat et al., 2024). The imputation of zeros in such instances can distort the analysis, failing to capture historical trends accurately.

In contrast, deletion methods, designed to address missing values within datasets, involve the removal of observations or variables characterized by missingness. While these methods offer the advantages of simplicity and the preservation of observed data, they can also introduce significant drawbacks. These methods can lead to a reduction in sample size, potentially compromising the analysis's statistical power and generalizability. Furthermore, they can introduce biases into the estimates, mainly if the missingness is not randomly distributed. Additionally, deletion methods may disproportionately affect specific subgroups or conditions characterized by high rates of missing values, leading to biased representation and hindering meaningful subgroup analyses.

The repertoire of deletion methods for addressing missing values within datasets encompasses listwise deletion, pairwise deletion, and column-wise deletion. Listwise deletion involves the removal of rows containing missing values, resulting in a reduction of the dataset without compromising the integrity of the remaining values. However, this approach can lead to a loss of information if the missingness is reasonably random. Pairwise deletion judiciously ignores pairs of missing values, enabling analysis with available data but potentially leading to varying sample sizes. Column-wise deletion involves the removal of columns or variables that exceed a predetermined threshold or percentage of missing values, which can be beneficial for irrelevant variables characterized by high levels of missingness.

Imputation techniques, designed to estimate missing values within datasets, offer the significant advantage of preserving sample size and ensuring the integrity of the dataset for subsequent analysis (Sun et al., 2023). By considering the interrelationships among variables, these techniques reduce bias and capture the underlying patterns of missing data (Sierra-Porta, 2024; Zhou et al., 2023). Compared to deletion methods, which involve the removal of observations characterized by missing values, imputation techniques are generally preferred due to their ability to retain valuable information and mitigate the introduction of biases.

While imputation techniques offer several advantages, it is imperative to acknowledge their inherent limitations. Estimating missing values introduces uncertainty, as it is predicated upon assumptions and statistical models (Blázquez-García et al., 2023). Certain imputation methods, mainly those reliant on regression or predictive models, may inadvertently distort the variability of the imputed values towards the mean or predicted values. This phenomenon, known as regression to the mean, can result in an underestimation of the true variability within the dataset, potentially compromising the accuracy of statistical analyses, such as hypothesis testing or confidence interval estimation (Başakın et al., 2023).

Furthermore, imputation techniques rely heavily on assumptions regarding the missingness mechanism and the interrelationships between observed and missing data (Sierra-Porta, 2024). If these assumptions are violated or the imputation model is misspecified, the imputed values may not accurately reflect the true values for the missing data. Consequently, meticulous consideration and evaluation of the imputation model and its underlying assumptions are essential to ensure the validity and reliability of the results.

Additionally, while imputation aims to mitigate bias, there remains a potential for bias to persist if the imputation model or assumptions are not appropriate for the data. Inaccurately imputing missing values can propagate biases or distort the dataset's characteristics (Li et al., 2023). Therefore, conducting a thorough assessment of the imputation method and performing sensitivity analyses to evaluate its impact on the results is imperative.

The range of imputation strategies includes various methods such as mean/median imputation, mode imputation, regression imputation, KNN imputation, and flagging. In mean/median imputation, missing numerical values are replaced with the mean or median of the available data. Mode imputation is used for categorical variables, replacing missing values with the most common category. Regression imputation is a more advanced method that uses a regression model to predict missing values based on the relationships between variables. KNN imputation estimates missing values by considering the values of the KNN of the data point with missing values, making it effective when an appropriate distance metric is available. Lastly, flagging involves creating a binary feature to indicate whether a value is missing, which can be helpful when the fact that a value is missing holds useful information for the prediction task. Choosing an appropriate imputation method requires careful consideration of the dataset's characteristics and the specific problem being addressed. It is important to select methods that minimize bias and prevent distortions in the data. Additionally, the effect of missing values on the performance of the machine learning algorithm should be evaluated, along with the factors that cause the missing data.

Addressing missing data requires acknowledging that no single solution fits all situations. The most appropriate method depends on the specific analysis being performed. Researchers might choose deletion, imputation, or a combination of both to ensure reliable results. While deletion is a quick method, it can reduce the dataset size and introduce bias if the missing data is not random. Table 1 summarizes the trade-offs between deletion and imputation techniques.

Comparing methods for handling missing values in machine learning (ML) datasets involves carefully assessing their effectiveness in maintaining data quality and minimizing their impact on model performance. The first step is to define the evaluation metrics for measuring ML model performance. These metrics could include accuracy, precision, recall, F1-score, or other problem-specific measures.

After defining the metrics, the next step is to select suitable methods for addressing missing values. These methods might include using mean, median, mode, or ML-based approaches for imputation.

Table 1: Deletion vs. imputation

Technique	Simplicity	Reduction of noise	Efficiency	Reduction in sample size	Introducing bias	Loss of information	Preservation of statistical properties
Zero-filling	Yes	Yes	Yes	No	Yes	No	No
Deletion approach	Yes	Yes	Depends	Yes	Depends	Yes	No
Advanced approaches	No	Yes	Depends	No	No	No	Yes

The dataset should be partitioned into training and validation/test sets to ensure a reliable evaluation. This division enables the training of models on the modified datasets and the subsequent evaluation of their performance on the validation/test set. If applicable, the selected techniques can be applied to the training set, addressing missing values within both the input features and target variables. This process entails the creation of modified datasets while preserving the original dataset as a baseline for comparative analysis. ML models can be constructed and evaluated upon dataset preparation using modified and original datasets. The performance of these models can be assessed by employing the selected evaluation metrics on the validation/test set, thereby facilitating a comparative analysis of performance across the various techniques. Statistical tests, such as t-tests or ANOVA, can be conducted to delve deeper into model performance and identify any significant disparities among the methods. Beyond the sole consideration of model performance, it is imperative to evaluate other pertinent factors when selecting the most appropriate technique for the given dataset and problem. These factors may encompass the complexity of the method, its ease of implementation, and the underlying assumptions associated with each technique.

It is strongly recommended that the findings be iterated and validated through repeated experimentation to ensure the reliability and generalizability of the results. This can be accomplished by employing different evaluation metrics, train-test splits, or additional datasets. This iterative process reinforces the robustness and applicability of the results to a broader range of scenarios.

3. Evaluation metrics for machine learning techniques

A comprehensive comparison of various techniques for addressing missing data within the realm of ML necessitates a meticulous evaluation of their efficacy in preserving data integrity and minimizing their impact on model performance. To achieve this objective, it is imperative to define evaluation metrics that accurately capture the influence of these techniques on ML models. The subsequent subsections offer a systematic approach for determining such evaluation metrics, as illustrated in Table 2.

3.1. Data completeness metric

The data completeness metric quantifies the percentage of missing values within the dataset prior to and after applying each technique, thereby providing a measure of its effectiveness (Munshi, 2024). This metric indicates the extent to which each technique successfully addresses missing values and ensures the completeness of the data. Let $Missing_{before}$ denote the percentage of missing values in the dataset prior to the application of a technique, and let $Missing_{after}$ denote the percentage of missing values after the technique is applied. These values can be calculated as follows:

$$Missing_{before} = \frac{Number\ of\ missing\ values\ in\ the\ dataset\ before\ applying\ the\ technique}{Total\ number\ of\ data\ points\ in\ the\ datasets} * 100\%$$

$$Missing_{after} = \frac{Number\ of\ missing\ values\ in\ the\ dataset\ after\ applying\ the\ technqiue}{Total\ number\ of\ data\ points\ in\ the\ datasets} * 100\%$$

The effectiveness of the technique can be quantified as follows:

- If $Missing_{after}$<$Missing_{before}$, the technique is deemed effective in addressing missing values.
- If $Missing_{after}$=$Missing_{before}$, the technique is considered to have no impact on missing values.
- If $Missing_{after}$>$Missing_{before}$, the technique indicates the introduction of new missing values or a failure to effectively handle existing ones, suggesting the necessity for reevaluation and exploring alternative approaches.

For instance, consider a healthcare organization implementing an imputation technique to address missing patient records. Initially, the dataset comprised 1,000 records, with 200 entries lacking critical health information, resulting in a $Missing_{before}$ value of 20% (200/1000). Upon applying a mean imputation technique, the organization successfully reduced the missing values to just 5%, yielding a $\text{Missing}_{\text{after}}$ value of 5%. This marked improvement demonstrates the technique's effectiveness, as evidenced by the condition $\text{Missing}_{\text{after}}$<$\text{Missing}_{\text{before}}$ (5%<20%). Such an increase in data completeness reinforces the integrity of patient information and plays a vital role in enhancing clinical outcomes. Consequently, the data completeness metric is a robust indicator of the technique's success, highlighting its crucial role in upholding high-quality data standards in healthcare.

3.2. Model performance

This metric evaluates the impact of missing data handling techniques on the performance of your ML models (Nezami et al., 2024). This can be done by assessing relevant performance metrics such as accuracy, precision, recall, F1-score, mean squared error (MSE), or any other appropriate metrics based on the specific task (classification, regression, etc.). To comprehend the impact of each technique, we compare the performance of models trained on the original dataset (without missed values), the dataset with deleted values, and the imputed dataset (after applying the imputation method). Let's denote the performance of the model trained on the original dataset as $\text{P}_{\text{Original}}$, the performance of the model trained on the deleted dataset as $\text{P}_{\text{Deleted}}$, and the performance of the model trained on the imputed dataset as $\text{P}_{\text{Imputed}}$.

To compare the impact of each technique, we can calculate the difference in performance between these datasets as follows:

$$Impact_{deleted} = P_{deleted} - P_{Orginal}$$
$$Impact_{Imputed} = P_{Imputed} - P_{Orginal}$$

- If $Impact_{deleted}$<0 indicates information loss or decreased performance due to missing value removal. This means the model's performance on the dataset with deleted values is worse than the original dataset's.
- If $Impact_{deleted}$=0, there is no difference in performance between the model trained on the dataset with deleted values and the model trained on the original dataset (without missing values). In this case, the removal of missing values did not have any impact on the model's performance.
- If $Impact_{deleted}$>0 indicates a positive difference in performance between the model trained on the dataset with deleted values and the model trained on the original dataset without missing values. In other words, removing the missing values has led to an improvement in the model's performance.
- If $Impact_{Imputed}$>0. This indicates that the model's performance on the imputed dataset is better than the model trained on the original dataset without missing values. It suggests that the missing value imputation techniques have improved the model's performance or resulted in a similar performance to the original dataset.
- If $Impact_{Imputed}$=0 means that the performance of the model trained on the imputed dataset is the same as that of the model trained on the original dataset without missing values. It suggests that the missing value imputation techniques have preserved the model's performance, maintaining it at the same level as the original dataset.

- If $Impact_{Imputed}$<0 indicates that the performance of the model trained on the imputed dataset is worse than that of the model trained on the original dataset without missing values. It suggests that the missing value imputation techniques may have introduced noise or incorrect information, leading to a decrease in the model's performance compared to the original dataset.

Consider, for instance, a retail company developing a customer churn prediction model. The initial dataset comprises 10,000 records, with 15% missing values in key features. The model trained on this complete dataset achieves an impressive accuracy of 85% ($P_{Original}$). This baseline performance underscores the critical role that data integrity plays in the effectiveness of predictive modeling. To address the missing values, the company tests two strategies. The first approach involves deleting records with missing values, which results in a reduced dataset of 8,500 records. However, this method yields a lower accuracy of 80% ($P_{Deleted}$). The second strategy employs mean imputation, which slightly mitigates the data loss, leading to an accuracy of 84% ($P_{Imputed}$). The impact calculations reveal significant insights: $Impact_{Deleted}$=80%-85%=-5% decline for the deletion method, and $Impact_{Imputed}$ = 84%-85%=-1% drop for the imputation method. Both techniques result in diminished model performance compared to the original dataset, highlighting the adverse effects of missing data on predictive accuracy. This decline in performance is not merely a numerical loss; it signifies potential information loss or the introduction of noise, which can adversely affect business decisions and strategies. In a competitive retail landscape, even a slight reduction in accuracy can lead to missed opportunities in customer retention and engagement. Thus, maintaining high data quality is paramount, as it directly influences the reliability and effectiveness of predictive models, ultimately impacting a company's bottom line.

3.3. Stability of results

This metric examines the consistency and variability in model performance across different deletion or imputation methods. Determine if there are significant fluctuations or variations in the obtained results from each technique (Santos et al., 2024). Let's denote the variance of model performance for a specific technique as $Tech_{var}$. If we have multiple techniques and want to compare the stability across them, we can calculate the mean variance across all techniques as $Mean_{var}$. Stability can be determined by comparing $Tech_{var}$ with $Mean_{var}$. If $Tech_{var}$ is significantly lower than $Mean_{var}$, it indicates higher stability in model performance for that technique. Additionally, confidence intervals can be calculated to quantify the uncertainty in the model performance estimates as the following:

- $If\ Tech_{var}$<$Mean_{var}$: Higher stability, consistent and reliable results.
- $If\ Tech_{var}$>$Mean_{var}$: Lower stability, more variability, and potential inconsistency.
- $If\ Tech_{var}$=$Mean_{var}$: Similar stability as the average, consistent, and reliable performance.

Let's denote the Confidence Interval (CI) for a specific technique as $Tech_{CI}$. By comparing the size of confidence intervals across different techniques, we can assess the stability of the results. If the width of $Tech_{CI\_i}$, denoted as W ($Tech_{CI\_i}$), is smaller than other techniques, indicating lower variability and higher stability in the model performance for that specific technique. On the other hand, if the width of $Tech_{CI\_i}$ is larger, it suggests higher variability and potentially lower stability in the model performance. If the widths of the confidence intervals for different techniques are equal, it suggests comparable variability and stability in the model performance among those techniques.

Assume a data scientist is evaluating techniques for handling missing values in a dataset used for predicting customer churn. This dataset includes features such as age, income, and purchase history, with some entries missing fundamental values. The scientist applies three methods: Record deletion, mean imputation, and mode imputation.

After training models using each technique, the data scientist records performance metrics, such as accuracy, over multiple iterations. The accuracy across iterations is [0.82, 0.84, 0.83, 0.81, 0.83] for record deletion. For mean imputation, the accuracy is [0.78, 0.76, 0.77, 0.79, 0.75], and for mode imputation, the accuracy is [0.85,0.91,0.87,0.88,0.86]. The variance of model performance for each technique ($Tech_{var}$) is then calculated, yielding 0.00013 for record deletion, 0.00025 for mean imputation, and 0.00053 for mode imputation. The mean variance across all methods is computed as $Mean_{var}$=0.000303.

By comparing each $Tech_{var}$ to $Mean_{var}$, the data scientist finds that record deletion (0.00013) and mean imputation (0.00025) both indicate lower stability, while mode imputation (0.00053) offers higher stability. This suggests that mode imputation provides the most consistent performance. Additionally, the scientist calculates Confidence Intervals (CIs) for each technique: Record deletion has a CI of [0.8118, 0.8402] (width=0.0284), mean imputation has a CI of [0.7504, 0.7896] (width=0.0392), and mode imputation has a CI of [0.8454, 0.9026] (width=0.0572). The comparison of widths shows that record deletion exhibits a smaller width than both mean imputation and mode imputation, which indicates that it is the most stable technique. In contrast, mean imputation has a larger width compared to record deletion, suggesting reduced stability. Meanwhile, mode imputation has the largest width, signifying it is the least stable among the three techniques.

3.4. Bias and variance

This metric measures the influence of missing data handling techniques on the bias and variance of your models (Sedaghat et al., 2024). Evaluate these aspects using techniques such as learning curves or bias-variance trade-off analysis. Learning curves can be used to plot the convergence behavior and final performance of models trained on the original and modified datasets. Comparable convergence rates and performance indicate minimal bias impact. Bias-variance trade-off analysis involves adjusting hyperparameters and architecture to minimize bias and variance. Let the impact of missing data handling techniques on the bias of models represented as $Bias_{impact}$ and the impact on the variance as $Variance_{impact}$. To assess the impact of bias, we can calculate the difference in bias between the models trained on the original dataset and the models trained on the modified datasets:

$$Bias_{impact} = Bias_{modified} - Bias_{original}$$

where, $Bias_{original}$ represents the bias of the model trained on the original dataset, and $Bias_{modified}$ represents the bias of the model trained on the modified dataset (deleted or imputed). A high $Bias_{impact}$ indicates a significant difference in bias, while a low $Bias_{impact}$ suggests a minimal difference. The interpretation depends on the context and desired level of bias in the models. To quantify the level of bias impact as "high" or "low," a threshold or a criterion specific to your problem domain and context must be established. This threshold could be based on domain expertise, performance requirements, or other relevant considerations for your application. For example, you could define a threshold value such as "if $Bias_{impact}$>0.1, then the bias impact is considered high; otherwise, it is considered low." This threshold value of 0.1 is arbitrary and must be determined based on your needs and goals.

To assess the impact on variance, we can calculate the difference in variance between the models trained on the original dataset and the models trained on the modified datasets:

$$Variance_{impact} = Variance_{modified} - Variance_{original}$$

To evaluate the impact of missing data handling techniques, we consider $Variance_{original}$ as the variance of the model trained on the original dataset and $Variance_{modified}$ as the variance of the model trained on the modified dataset (with deleted or imputed data).

It is necessary to specify a threshold to quantify the level of $Variance_{impact}$. If $Variance_{impact}$ exceeds the threshold ($Variance_{impact}$>Threshold), the missing data handling technique can significantly influence the variance of the models, potentially causing increased variability or instability in the predictions. Conversely, if $Variance_{impact}$ falls below the threshold ($Variance_{impact}$<Threshold), the missing data handling technique has a relatively minor effect on model variance, indicating that it effectively preserves the underlying variability or has minimal impact on stability. Metrics for quantifying bias and variance, such as mean squared error (MSE) for regression or accuracy for classification, may vary based on the task. Approaches like k-fold cross-validation or bootstrapping can assess bias and variance and are tailored to dataset requirements.

Suppose, for instance, a data scientist is working with a dataset to predict house prices, which includes features like square footage, number of bedrooms, and age of the property but has missing values. To handle the missing data, apply three different techniques: Record deletion, mean imputation and mode imputation. Your goal is to analyze how these techniques affect the bias and variance of your predictive models.

The models were trained using three different techniques for handling missing values: Record deletion, mean imputation, and mode imputation. In the case of record deletion, any observations with missing values were removed, resulting in a bias of $Bias_{original} = 0.2$. For mean imputation, the missing values were replaced with the mean of the available data, which led to a bias of $Bias_{modified} = 0.25$. When mode imputation was applied, missing values were filled in with the mode, resulting in a bias of $Bias_{modified} = 0.3$. Next, the bias impact for each technique was calculated. For mean imputation, the bias impact was determined by subtracting the original bias from the modified bias: $Bias_{impact} = Bias_{modified} - Bias_{original} = 0.05$. For mode imputation, the calculation was similar: $Bias_{impact} = Bias_{modified} - Bias_{original} = 0.1$.

A threshold for bias impact was defined at 0.1. The results revealed that the bias impact of 0.05 for mean imputation is considered low, whereas the bias impact of 0.1 for mode imputation may be viewed as borderline and potentially high depending on the context. Variance was also assessed for each technique. For record deletion, the variance was found to be $Variance_{original} = 0.04$. In the case of mean imputation, the variance of the modified dataset was $Variance_{modified} = 0.06$, and for mode imputation, it was recorded as $Variance_{modified} = 0.08$.

The variance impact of each technique was calculated. Mean imputation showed a variance impact of 0.02, while mode imputation had a variance impact of 0.04. A threshold of 0.03 was established to assess the variance impact. Based on this threshold, the variance impact of 0.02 for mean imputation is classified as low, whereas the 0.04 variance impact for mode imputation is considered high. These findings highlight the notable effect that different methods for handling missing data can have on the bias and variance of predictive models.

3.5. Robustness to new data

This metric facilitates the evaluation of model generalization by assessing the performance of models trained on datasets where values have been deleted or imputed. The Employ cross-validation or hold-out validation can be used on a separate test set to measure each technique's robustness and generalization capabilities (Mundargi et al., 2024). Cross-validation evaluates ML models by dividing the dataset into subsets or folds (Gorriz et al., 2024). It helps estimate performance, assess generalization, and aid hyperparameter tuning and model selection. Hold-out validation is used to evaluate model performance and generalization (Veetil et al., 2024). The dataset is divided into a training set for model training and a validation set for assessment. It is commonly employed for initial evaluation, hyperparameter tuning, and model comparison.

If the generalization performance of models trained on deleted or imputed datasets is represented as $G_{performance}$, $GPerf_{deleted}$ to represent the $G_{performance}$ for models trained on the deleted dataset and $GPerf_{imputed}$ as the $G_{performance}$ for models trained on the imputed dataset. By using cross-validation or hold-out validation on a separate test set, we can estimate the $G_{performance}$ for both the deleted and imputed datasets. To compare the robustness and generalization capabilities of each technique, we can analyze the difference in Generalization_performance between the deleted and imputed datasets:

$$G_{impact} = GPerf_{imputed} - GPerf_{deleted}$$

If G_{impact} is significantly higher, it indicates that models trained on the imputed dataset have better generalization capabilities and are more robust to new, unseen data. If the G_{impact} it is low, suggesting that models trained on the imputed dataset may have limited generalization capabilities and perform poorly on new, unseen data.

Assume that a professional is working on a customer segmentation project using demographic and purchase history data to evaluate the robustness and generalization capabilities of models trained on datasets with missing values. The dataset has missing values, prompting the scientist to assess two techniques: Record deletion and mean imputation.

The dataset is split into a training set and a test set. Two models are trained: One on the dataset with record deletion and the other on the dataset with mean imputation, while the test set is reserved for evaluating generalization performance. The professional employs k-fold cross-validation on the training set to estimate $G_{performance}$. After training, the model performances are recorded as $GPerf_{deleted}$=0.75 and $GPerf_{imputed}$=0.82. The impact of the techniques on generalization performance is calculated as G_{impact}=0.07. This positive G_{impact} indicates that the imputed model has better generalization capabilities and is more robust to new, unseen data.

3.6. Computational efficiency

It refers to the ability of a technique to process and handle missing values in datasets with minimal time and resource requirements (Koukaras et al., 2024). If the computational efficiency of each method is represented as $Tech_{effeciency}$, the time required for applying a specific technique as $Tech_{time}$ and the resource requirements as $Tech_{resources}$. If there are multiple techniques, their efficiencies can be compared by calculating the ratio of time and resource requirements concerning a reference technique. This can be expressed as:

$$Tech_{effeciency} = Tech_{time} / Time_{ref} + Tech_{resources} / Resources_{ref}$$

Where, $Time_{ref}$ and $Resources_{ref}$ represent the time and resource requirements of the reference technique, respectively. By calculating $Tech_{effeciency}$ we can compare the computational efficiency of each technique. A lower $Tech_{effeciency}$ value signifies that a technique necessitates greater computational resources or time for data processing and handling. This may result in slower execution, higher memory usage, or increased computational complexity. Conversely, a higher $Tech_{effeciency}$ value indicates that a technique is more efficient, requiring fewer resources or less time for processing. Consequently, in most cases, higher computational efficiency values are preferred as they signify more efficient and faster execution of the technique.

Consider an individual working on a large dataset of customer transactions with missing entries in several fields. The individual compares three techniques: Record deletion, mean imputation, and KNN imputation. The first step involves measuring the time and resource requirements for each technique. For record deletion, the time taken is 2 seconds, and the resources required are 100 MB. Mean imputation requires 1 second and 50 MB of resources, while KNN imputation takes 5 seconds and uses 200 MB. The data scientist selects mean imputation as the reference technique, which has a time of 1 second and resource requirements of 50 MB. Next, the computational efficiency for each technique is calculated using the formula:

$$Tech_{effeciency} = Tech_{time} / Time_{ref} + Tech_{resources} / Resources_{ref}$$

For record deletion, the efficiency is calculated as 2/1+100/50=4. For mean imputation, the efficiency is 1/1+50/50=2. KNN imputation yields an efficiency of 5/1+200/50=9. These calculations show that mean imputation has the highest computational efficiency (the least time and resources), followed by record deletion, while KNN imputation has the lowest efficiency.

3.7. Domain-specific metrics (DSMs)

They refer to performance measures or evaluation criteria that are specifically designed to align with the requirements and constraints of a particular problem domain (Park et al., 2024). These metrics consider the domain's unique characteristics, context, and goals. DSMs are crucial for meaningful insights into model or system performance and success. DSMs can be tailored to evaluate data quality, accuracy, completeness, and reliability within a specific problem domain. In certain domains, stringent regulatory frameworks necessitate adherence to particular metrics. These metrics ensure compliance with regulations, ethical guidelines, and legal requirements.

Let the evaluated techniques be used for T1, T2, T3, ..., Tn. To make an informed decision, we evaluate and compare these techniques using the following metrics: Data Quality (DQ), which assesses the quality of data for each technique Ti as DQ_{T_i}; Accuracy (ACC), which measures the accuracy of each technique Ti as ACC_{T_i}; Data Completeness (DC), which quantifies the completeness of data for each technique Ti as DC_{T_i}; and Reliability (REL), which assesses the reliability of each technique Ti as REL_{T_i}. The decision-making process based on these metrics can be achieved using a weighted sum approach. Let's denote the weights assigned to each metric as W_{dq} for data quality (DQ), W_{acc} for accuracy (ACC), W_{dc} for data completeness (DC), and W_{rel} for reliability (REL). The decision can be made by calculating a composite score for each technique (Ti) using the following formula:

$$SCORE_{T_i} = W_{dq} * DQ_{T_i} + W_{acc} * ACC_{T_i} + W_{dc} * DC_{T_i} + W_{rel} * REL_{T_i}$$

The technique with the highest composite score calculated using the weighted sum formula is considered the most suitable/optimal choice. The technique with the lowest composite score is considered the least suitable/optimal choice. The weights W_{dq}, W_{acc}, W_{dc}, and W_{rel} should be determined based on the specific goals and requirements of the project, reflecting the relative importance of each metric in the decision-making process.

To evaluate Domain-Specific Metrics (DSMs) in a healthcare context, consider a data scientist tasked with developing a predictive model to identify patients at risk of developing diabetes. Given the critical nature of healthcare data, the scientist decides to compare three techniques for handling missing data: Record deletion, mean imputation, and multiple imputation. The goal is to assess these techniques using metrics that align with the healthcare sector's unique requirements.

Initially, relevant metrics are defined as Data Quality (DQ), Accuracy (ACC), Data Completeness (DC), and Reliability (REL). Then, performance data for each technique is collected. For record deletion, the metrics are DQ_{T1}=0.70, ACC_{T1}=0.75, DC_{T1}=0.60, and REL_{T1}=0.80. For mean imputation, the values are DQ_{T2}=0.85, ACC_{T2}=0.80, DC_{T2}=0.75, and REL_{T2}=0.70. And, for multiple imputation, the metrics are DQ_{T3}=0.90, ACC_{T3}=0.85, DC_{T3}=0.90, and REL_{T3}=0.95. Next, the data scientist assigns weights to each metric based on the project's specific goals. For example, the weights might be W_{dq}=0.4 for Data Quality, W_{acc}=0.3 for Accuracy, W_{dc}=0.2 for Data Completeness, and W_{rel}=0.1 for Reliability. Using these weights, the scientist calculates a composite score for each technique with the formula:

$$SCORE_{T_i} = W_{dq} * DQ_{T_i} + W_{acc} * ACC_{T_i} + W_{dc} * DC_{T_i} + W_{rel} * REL_{T_i}$$

The calculations yield $SCORE_{T1}$=0.705 for record deletion; for mean imputation, $SCORE_{T2}$=0.8; and for multiple imputation, $SCORE_{T3}$=0.89. Based on the scores obtained, the most suitable technique for handling missing data is Multiple Imputation, which achieved a score of 0.89. This method demonstrates the best performance according to the evaluated metrics. In contrast, Record Deletion is the least suitable technique, with a score of 0.705, indicating that it does not perform as well compared to the other methods.

3.8. Metrics integration

Synthesizing the results from the different evaluation metrics to understand the impact of each technique on your ML models (Kazemi et al., 2024). These metrics include data completeness, model performance, stability, and computational efficiency. It is important to consider the trade-offs associated with various metrics to make an informed decision. A well-informed decision can be made that balances the trade-offs and aligns with the specific goals and requirements of the ML project. Let D represent the data completeness achieved by a specific technique. Let P represent the model performance obtained with the technique. Let S represent the stability of the results. Let C represent the computational efficiency required. To make an informed decision, we aim to optimize a composite objective function that incorporates these factors. This objective function can be defined as:

$$Objective = w_1 * D + w_2 * P + w_3 * S + w_4 * C$$

where, w_1, w_2, w_3, and w_4 are the weights assigned to each factor, representing their relative importance. These weights can be determined based on domain expertise, project requirements, or stakeholder preferences. A high objective means a higher composite score, which indicates the technique is more optimal or suitable based on the weighted evaluation of the different performance metrics (data completeness, model performance, stability, and computational efficiency). A low objective means a lower composite score, which indicates the technique is less optimal or suitable based on the weighted evaluation of the performance metrics.

Suppose that in an ML project to predict customer churn for a subscription service, a data scientist evaluates three techniques for handling missing data: Record deletion, mean imputation, and multiple imputations. To understand the overall impact of each method on the model, the scientist seeks to synthesize results across various evaluation metrics, including data completeness, model performance, stability, and computational efficiency.

For record deletion, the metrics indicate D=0.60 (60% data completeness), P=0.75 (75% for model accuracy), S=0.70 (70% for stability score), and C=0.50 (50% for computational efficiency score). Mean imputation shows better results with D=0.85, P=0.80, S=0.75, and C=0.80. Multiple imputation demonstrates the highest values, yielding D=0.90, P=0.85, S=0.80, and C=0.60. Next, the data scientist assigns weights to each factor based on their importance in the context of the project. For instance, the weights could be set as w_1=0.3 for data completeness, w_2=0.4 for model performance, w_3=0.2 for stability, and w_4=0.1 for computational efficiency. The objective function is then defined as:

$$Objective = w_1 * D + w_2 * P + w_3 * S + w_4 * C$$

Using this formula, the scientist calculates the composite scores for each technique. For record deletion, the objective score is 0.67. Mean imputation achieves a score of 0.805, while multiple imputation scores 0.83. The analysis shows that mean imputation has the highest objective score, closely followed by multiple imputation, while record deletion scores the lowest. This example illustrates how the data scientist integrates multiple evaluation metrics to make an informed decision about the best technique for handling missing data

Table 2: Evaluation metrics for missing data handling techniques

Metric	Description	Evaluation method	Interpretation
Data completeness	Measures missing value reduction	Calculating the percentage of missed values before and after applying the technique.	Effective: $\text{Missing}_{\text{after}}$ < $\text{Missing}_{\text{before}}$ No impact: $\text{Missing}_{\text{after}}$ =$\text{Missing}_{\text{before}}$ Ineffective: $\text{Missing}_{\text{after}}$ > $\text{Missing}_{\text{before}}$
Model performance	Impact on model performance	Calculating $\text{P}_{\text{Original}}$, $\text{P}_{\text{Deleted}}$ and $\text{P}_{\text{imputed}}$.	No difference: $Impact_{deleted}$ =0 and $Impact_{Imputed}$ = 0 Positive difference: $Impact_{deleted}$ > 0, Improved the model performance: If $Impact_{Imputed}$ > 0 and $Impact_{Imputed}$ > 0 Worse: $Impact_{Imputed}$ < 0
Stability of results	Consistency and variability	Computing variance, confidence interval width	Higher stability, consistent and reliable results: $If\ Tech_{var}$ < $Mean_{var}$. Lower stability, more variability, and potential inconsistency. $If\ Tech_{var}$ > $Mean_{var}$ Similar stability as the average, consistent, and reliable performance: $If\ Tech_{var}$ = $Mean_{var}$ Lower variance and smaller CI: Higher stability
Bias and variance	Impact on model bias and variance	Bias difference, variance difference	High bias: if $Bias_{impact}$ > Threshold Low bias: if $Bias_{impact}$ > Threshold Minor impact, variability preserved: if $Variance_{impact}$<Threshold Variance impacted, predictability altered: if $Variance_{impact}$<Threshold
Robustness to new data	Generalization performance	Generalization performance (Deleted) - Generalization performance (Imputed)	G_{impact} = $GPerf_{imputed}$ - $GPerf_{deleted}$ Higher generalization: If G_{impact} is High Limited generalization: If G_{impact} is Low
Computational efficiency	Resource usage and time required	Time and resource requirements	$Tech_{effeciency}$ = $Tech_{time}$ / $Time_{ref}$ + $Tech_{resources}$ / $Resources_{ref}$ High efficiency: Low $Tech_{effeciency}$ Low efficiency: High $Tech_{effeciency}$
Domain-specific metric	Domain-specific metrics guide decision	Evaluating data quality, accuracy, completeness, and reliability within a specific problem domain	$SCORE_{T_i}$= W_{dq} * DQ_{T_i} + W_{acc} * ACC_{T_i} + W_{dc} * DC_{T_i} + W_{rel} * REL_{T_i} Highest score, best fit Lowest score, least fit
Metrics integration	Integrated assessment of technique performance	Data completeness, model performance, result stability, and computational efficiency factors are balanced using a weighted objective function to optimize the chosen approach	$Objective = w_1 * D + w_2 * P + w_3 * S + w_4 * C$ Optimal technique with highest weighted score. Suboptimal technique with lower weighted score

Table 3 presents a comprehensive analysis of the advantages and disadvantages associated with the various methods employed to evaluate the effectiveness of techniques for addressing and managing missing data. The metrics highlighted in Table 3 aim to strike a balance between simplicity, performance, robustness, and domain-specific considerations. On the strength side, the metrics prioritize the reduction of missing values, the evaluation of model accuracy, the assessment of consistency, the measurement of bias and variance, and the consideration of computational efficiency and domain-specific factors. However, these approaches are not without their weaknesses. They may oversimplify the patterns of missingness, exhibit limited generalizability, be computationally expensive, necessitate domain expertise, and present challenges when dealing with large datasets.

The integration of multiple metrics is widely regarded as a strength, as it offers a more comprehensive understanding of the techniques. However, this integration can also be computationally complex, particularly when dealing with large datasets.

Table 3: Critique of missing data handling evaluation metrics

Metric	Strengths	Weaknesses
Data completeness	The standardized approach focuses on missing value reduction	Oversimplifies missingness patterns, may not capture nuances
Model performance	Evaluate the impact on model accuracy, guide decision-making	Limited generalizability, thresholds may need adjustments
Stability of results	Assesses consistency of model performance across techniques	Computationally expensive for large datasets
Bias and variance	Measures impact on model bias and variance	Requires domain expertise to set thresholds
Robustness to new data	Evaluates model performance on unseen data	Relies on validation techniques (cross-validation)
Computational efficiency	Considers resource usage and execution time	Can penalize more complex techniques
Domain-specific metrics	Tailored to specific problem domains, considers unique characteristics	Requires deep understanding of the domain for weighting
Metrics integration	Combines results for a comprehensive understanding	Computationally complex for large datasets

4. Conclusion

Accurate ML predictions rely on high-quality datasets that are relevant, sufficient, representative, balanced, and diverse. However, values can be necessary to maintain these qualities. Effective handling and imputation techniques are crucial for preserving dataset integrity. This analysis compares three approaches to address missing values: Zero imputation, deletion, and various imputation methods. While deletion is a quick option, it can lead to information loss and bias. Imputation maintains the dataset but may introduce uncertainty. The choice of method depends on the dataset's characteristics and the problem at hand. A structured eight-step process is outlined for evaluating missing value-handling techniques. This process involves defining metrics, selecting methods, and validating results. A two-step decision-making process is also proposed: First, assessing techniques using diverse metrics, and second, scoring each method based on project goals. The article provides a comprehensive overview of tools for handling missing values, emphasizing the need to balance data preservation, bias avoidance, and analytical accuracy. It is a valuable resource for navigating the complexities of missing value handling in ML. Future research will focus on empirically validating the proposed metrics, exploring the integration of deep learning techniques, applying transfer learning to address data limitations, developing methods to quantify uncertainty in predictions, creating metrics tailored to specific domains, improving the computational efficiency of missing data handling, and investigating how these techniques can be integrated with other data quality measures.

Compliance with ethical standards

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Abidin NZ, Ismail AR, and Emran NA (2018). Performance analysis of machine learning algorithms for missing value imputation. International Journal of Advanced Computer Science and Applications, 9(6): 442-447. https://doi.org/10.14569/IJACSA.2018.090660

Agarwal S (2023). An intelligent machine learning approach for fraud detection in medical claim insurance: A comprehensive study. Scholars Journal of Engineering and Technology, 11(9): 191-200. https://doi.org/10.36347/sjet.2023.v11i09.003

Albahri AS, Zaidan AA, AlSattar HA, Hamid RA, Albahri OS, Qahtan S, and Alamoodi AH (2023). Towards physician's experience: Development of machine learning model for the diagnosis of autism spectrum disorders based on complex T-spherical fuzzy-weighted zero-inconsistency method. Computational Intelligence, 39(2): 225-257. https://doi.org/10.1111/coin.12562

Alzubaidi L, Bai J, Al-Sabaawi A, Santamaría J, Albahri AS, Al-dabbagh BSN, Fadhel MA, Manoufali M, Zhang J, Al-Timemy AH, and Duan Y et al. (2023). A survey on deep learning tools dealing with data scarcity: Definitions, challenges, solutions, tips, and applications. Journal of Big Data, 10: 46. https://doi.org/10.1186/s40537-023-00727-2

Alzyadat W, Shaheen A, Ala'a Al-Shaikh AA, and Al-Khasawneh Z (2024). A proposed model for enhancing e-bank transactions: An experimental comparative study. Indonesian Journal of Electrical Engineering and Computer Science, 34(2): 1268-1279. https://doi.org/10.11591/ijeecs.v34.i2.pp1268-1279

Aubaidan BH, Kadir RA, Ljab MT, and Taha BA (2024). Intelligent imputation of missing data using bidirectional neighbor graph modeling for diabetic risk prediction. Journal of Theoretical and Applied Information Technology, 102(8): 3508-3522.

Başakın EE, Ekmekcioğlu Ö, and Özger M (2023). Providing a comprehensive understanding of missing data imputation processes in evapotranspiration-related research: A systematic literature review. Hydrological Sciences Journal, 68(14): 2089-2104. https://doi.org/10.1080/02626667.2023.2249456

Blázquez-García A, Wickstrøm K, Yu S, Mikalsen KØ, Boubekki A, Conde A, Mori U, Jenssen R, and Lozano JA (2023). Selective imputation for multivariate time series datasets with missing values. IEEE Transactions on Knowledge and Data Engineering, 35(9): 9490-9501. https://doi.org/10.1109/TKDE.2023.3240858

Buttia C, Llanaj E, Raeisi-Dehkordi H, Kastrati L, Amiri M, Meçani R, Taneri PE, Ochoa SAG, Raguindin PF, Wehrli F, and Khatami F et al. (2023). Prognostic models in COVID-19 infection that predict severity: A systematic review. European Journal of Epidemiology, 38(4): 355-372. https://doi.org/10.1007/s10654-023-00973-x **PMid:36840867 PMCid:PMC9958330**

Chen AY and McCoy J (2024). Missing values handling for machine learning portfolios. Journal of Financial Economics, 155: 103815. https://doi.org/10.1016/j.jfineco.2024.103815

Emmanuel T, Maupong T, Mpoeleng D, Semong T, Mphago B, and Tabona O (2021). A survey on missing data in machine learning. Journal of Big Data, 8: 140. https://doi.org/10.1186/s40537-021-00516-9 **PMid:34722113 PMCid:PMC8549433**

Gorriz JM, Segovia F, Ramirez J, Ortiz A, and Suckling J (2024). Is K-fold cross validation the best model selection method for Machine Learning? Arxiv Preprint Arxiv:2401.16407. https://doi.org/10.48550/arXiv.2401.16407

Huang B, Zhu Y, Usman M, and Chen H (2024). Semi-supervised learning with missing values imputation. Knowledge-Based Systems, 284: 111171. https://doi.org/10.1016/j.knosys.2023.111171

Jaradat Y, Masoud M, Manasrah A, Alia M, and Jannoud I (2024). Review of data imputation techniques in time series data: Comparative analysis. The Eurasia Proceedings of Science, Technology, Engineering and Mathematics, 27: 122-129. https://doi.org/10.55549/epstem.1518433

Kadhim MA and Radhi AM (2023). Heart disease classification using optimized Machine learning algorithms. Iraqi Journal for Computer Science and Mathematics, 4(2): 31-42. https://doi.org/10.52866/ijcsm.2023.02.02.004

Kazemi A, Rasouli-Saravani A, Gharib M, Albuquerque T, Eslami S, and Schüffler PJ (2024). A systematic review of machine learning-based tumor-infiltrating lymphocytes analysis in colorectal cancer: Overview of techniques, performance metrics, and clinical outcomes. Computers in Biology and Medicine, 173: 108306. https://doi.org/10.1016/j.compbiomed.2024.108306 **PMid:38554659**

Koukaras P, Mustapha A, Mystakidis A, and Tjortjis C (2024). Optimizing building short-term load forecasting: A comparative analysis of machine learning models. Energies, 17(6): 1450. https://doi.org/10.3390/en17061450

Li C, Ren X, and Zhao G (2023). Machine-learning-based imputation method for filling missing values in ground meteorological observation data. Algorithms, 16(9): 422. https://doi.org/10.3390/a16090422

Li J, Guo S, Ma R, He J, Zhang X, Rui D, Ding Y, Li Y, Jian L, Cheng J, and Guo H (2024). Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets. BMC Medical Research Methodology, 24: 41. https://doi.org/10.1186/s12874-024-02173-x **PMid:38365610 PMCid:PMC10870437**

Liguori A, Markovic R, Ferrando M, Frisch J, Causone F, and van Treeck C (2023). Augmenting energy time-series for data-efficient imputation of missing values. Applied Energy, 334: 120701. https://doi.org/10.1016/j.apenergy.2023.120701

Liu X, Hasan MR, Ahmed KA, and Hossain MZ (2023). Machine learning to analyse omic-data for COVID-19 diagnosis and prognosis. BMC Bioinformatics, 24: 7. https://doi.org/10.1186/s12859-022-05127-6 **PMid:36609221 PMCid:PMC9817417**

Liu Y, Li B, Yang S, and Li Z (2024). Handling missing values and imbalanced classes in machine learning to predict consumer preference: Demonstrations and comparisons to prominent methods. Expert Systems with Applications, 237: 121694. https://doi.org/10.1016/j.eswa.2023.121694

Mitra R, McGough SF, Chakraborti T, Holmes C, Copping R, Hagenbuch N, Biedermann S, Noonan J, Lehmann B, Shenvi A, and Doan XV et al. (2023). Learning from data with structured missingness. Nature Machine Intelligence, 5(1): 13-23. https://doi.org/10.1038/s42256-022-00596-z

Mundargi Z, Khedkar S, Kumbhar S, Mohod K, and Meshram Y (2024). Revolutionizing cerebral stroke prediction: Mastery unveiled through stratified k-fold and k-fold cross validation techniques for imbalanced datasets. Grenze International Journal of Engineering and Technology, 10: 2407-2413.

Munshi RM (2024). Novel ensemble learning approach with SVM-imputed ADASYN features for enhanced cervical cancer prediction. PLOS ONE, 19(1): e0296107. https://doi.org/10.1371/journal.pone.0296107 **PMid:38198475 PMCid:PMC10781159**

Nezami N, Haghighat P, Gándara D, and Anahideh H (2024). Assessing disparities in predictive modeling outcomes for college student success: The impact of imputation techniques on model performance and fairness. Education Sciences, 14(2): 136. https://doi.org/10.3390/educsci14020136

Pagano TP, Loureiro RB, Lisboa FV, Peixoto RM, Guimarães GA, Cruz GO, Araujo MM, Santos LL, Cruz MA, Oliveira EL, and Winkler I et al. (2023). Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. Big Data and Cognitive Computing, 7(1): 15. https://doi.org/10.3390/bdcc7010015

Palanivinayagam A and Damaševičius R (2023). Effective handling of missing values in datasets for classification using machine learning methods. Information, 14(2): 92. https://doi.org/10.3390/info14020092

Parhi SK and Patro SK (2023). Prediction of compressive strength of geopolymer concrete using a hybrid ensemble of grey wolf optimized machine learning estimators. Journal of Building Engineering, 71: 106521. https://doi.org/10.1016/j.jobe.2023.106521

Park K, Ergan S, and Feng C (2024). Quality assessment of residential layout designs generated by relational generative adversarial networks (GANs). Automation in Construction, 158: 105243. https://doi.org/10.1016/j.autcon.2023.105243

Santos KC, Miani RS, and de Oliveira Silva F (2024). Evaluating the impact of data preprocessing techniques on the performance of intrusion detection systems. Journal of Network and Systems Management, 32: 36. https://doi.org/10.1007/s10922-024-09813-z

Sedaghat A, Arbabkhah H, Jafari Kang M, and Hamidi M (2024). Deep learning applications in vessel dead reckoning to deal with missing automatic identification system data. Journal of Marine Science and Engineering, 12(1): 152. https://doi.org/10.3390/jmse12010152

Shamji MH, Ollert M, Adcock IM, Bennett O, Favaro A, Sarama R, Riggioni C, Annesi-Maesano I, Custovic A, Fontanella S, and Traidl-Hoffmann C et al. (2023). EAACI guidelines on environmental science in allergic diseases and asthma–leveraging artificial intelligence and machine learning to develop a causality model in exposomics. Allergy, 78(7): 1742-1757. https://doi.org/10.1111/all.15667 **PMid:36740916**

Sharma B, Sharma L, Lal C, and Roy S (2023). Anomaly based network intrusion detection for IoT attacks using deep learning technique. Computers and Electrical Engineering, 107: 108626. https://doi.org/10.1016/j.compeleceng.2023.108626

Sierra-Porta D (2024). Assessing the impact of missing data on water quality index estimation: A machine learning approach. Discover Water, 4: 11. https://doi.org/10.1007/s43832-024-00068-y

Sun Y, Li J, Xu Y, Zhang T, and Wang X (2023). Deep learning versus conventional methods for missing data imputation: A review and comparative study. Expert Systems with Applications, 227: 120201. https://doi.org/10.1016/j.eswa.2023.120201

Tahyudin I, Solikhatin SA, Tikaningsih A, Lestari P, Nambo H, Winarto E, and Hassa N (2024). Forecasting hospital length of stay for stroke patients: A machine learning approach. International Journal of Advances in Soft Computing and Its Applications, 16(1): 99-117.

Veetil IK, Chowdary DE, Chowdary PN, Sowmya V, and Gopalakrishnan EA (2024). An analysis of data leakage and generalizability in MRI based classification of Parkinson's disease using explainable 2D Convolutional Neural Networks. Digital Signal Processing, 147: 104407. https://doi.org/10.1016/j.dsp.2024.104407

Zhou Y, Shi J, Stein R, Liu X, Baldassano RN, Forrest CB, and Huang J (2023). Missing data matter: An empirical evaluation of the impacts of missing EHR data in comparative effectiveness research. Journal of the American Medical Informatics Association, 30(7): 1246-1256. https://doi.org/10.1093/jamia/ocad066 **PMid:37337922 PMCid:PMC10280351**