The Impact of Missing Values in Machine Learning

Michal Dawid Kowalski
Univeristy of Porto
Porto, Portugal
up202401554@up.pt

ABSTRACT

This study presents an empirical analysis of the impact of missing values in databases on machine learning models. The primary objective is to explore and examine several existing approaches to handling missing data, focusing on three main missing mechanisms: MCAR, MAR and MNAR. The experiment investigates how the choice of imputation strategy affects the model performance across various datasets. The findings obtained from this study form the basis of the following report.

1 INTRODUCTION

Dealing with missing data is a crucial part of working with real-world datasets. In the social sciences, it is actually inevitable that some respondents will refuse to participate or to answer certain questions [1]. It can be also related with other sources, like data errors. The past decades have seen significant advancements in the theory, methodologies, and software for addressing incomplete data challenges. Understanding the source and the impact of missing data is valid for selecting appropriate strategies and executing robust analysis.

There are 3 possible mechanisms which lead to the missing values introduction: MCAR, MAR and MNAR.

- **MCAR** (*Missing Completely At Random*): missing data is unrelated to other measured variables and unrelated to its values. It means that the probability of being missing is the same for all cases.
- **MAR** (*Missing At Random*): missing data is related to some other measured variables and not related to its values.
- **MNAR** (*Missing Not At Random*): missing data is not related to other measured variables, but related to its values. The probability of missing data varies due to factors that remain unknown to us.

Many well-known machine learning models are sensitive to missing data, but there are also models that can handle this issue without preprocessing in this case. However, since most models rely on a complete input data, the strategy for imputing missing values have to be applied. It is important that these imputed values properly reflect the real ones, which is why various tools have been developed. The most common approaches include statistical imputation, machine learning-based imputation, and case deletion, if there is enough data available.

The next points will present a detailed comparison of mentioned approached, evaluating their impact on classification performance, and exploring whether the missing mechanism influences the results, with a focus on both classification and predictive accuracy.

2 DATASETS

The study was based on three complete datasets from various domains of life, each well-suited for developing classification models. These datasets were sourced both from Kaggle and UCI, and were then used to apply different missing mechanisms in the next step of the analysis.

2.1 Autism Screening Adult Datset

This dataset focuses on autism screening for adults, including ten behavioral traits and personal characteristics, with the goal of identifying ASD syndrome [2]. Missing values were removed, as the remaining data was still sufficient for analysis. Irrelevant features were also eliminated, the data was encoded, and Min-Max scaling was applied. Moreover, the dataset was split into training and test sets with a 70/30 ratio.

2.2 Mushroom Dataset

This dataset contains cleaned and preprocessed information about mushrooms, including features like cap diameter, shape, and stem color, for the binary classification of edibility . The target class indicates whether the mushroom is edible or poisonous [3]. The dataset was split into training and test sets in an 80/20 ratio due to the large amount of data, and Min-Max scaling was applied.

2.3 Occupancy Detection Dataset

The dataset focuses on accurate occupancy detection of an office room using light, temperature, humidity, and CO2 measurements, and is used for the binary classification to predict room occupancy [4]. Relevant features were extracted from the datetime column, including hour and day of the week. Min-Max scaling was applied to the data. The dataset had already been split into training and test sets by the creators.

3 SIMULATING MISSING VALUES

In this step, the mentioned missing mechanisms were simulated, generating three variants of datasets for each of the datasets used. For this purpose *mdatagen* Python library was applied, which allows the controlled generation of missing data consistent with MCAR, MAR and MNAR.

The experiments focused mainly on univariate missing value generators, where only one feature contains missing data. The selected feature was the one with the highest correlation to the target. The missing rate was set to 50%, meaning that 50% of the values in the chosen feature were missing, in order to create a more 'extreme' environment for the analysis.

- **MCAR Function**- randomly removes values from a feature with the highest correlation to the target.

- **MAR Function** - generates missing values in the most correlated feature using the N/2 lowest values and N/2 highest values from an observed feature.
- **MNAR Function** - generates missing data by introducing missing values into the highest 80% of feature values.

Table 1: F1-Score Class 0/Class 1 for the RF Classificator

Dataset Type	ADS	Mushroom	Occupancy
Complete	0.95 / 0.87	0.99 / 0.99	0.97 / 0.94
MCAR	0.96 / 0.91	0.99 / 0.99	0.97 / 0.95
MAR	0.97 / 0.92	0.98 / 0.99	0.97 / 0.94
MNAR	0.96 / 0.91	0.98 / 0.98	0.95 / 0.92

4 CLASSIFYING WITH MISSING VALUES

For the purpose of evaluating how missing data impacts classification performance, the *Random Forest* classifier was used, as it can handle missing values by relying on other data to build splits. Metrics such as precision, recall, F1-score, and accuracy were calculated using the *classification_report* function to assess how the model performs with missing values compared to complete datasets.

The evaluation results indicate that the Random Forest classifier performs robustly in the presence of missing data under MCAR, MAR and MNAR mechanisms, with only smaill differences in performance compared to the complete dataset Table 1. In the case of the ADF dataset, the introduction of missing values surprisingly improved the generalization of the classification, resulting in better evaluation metrics. For other datasets, the metrics were only slightly worse, where the MNAR mechanism had the most significant impact on classification performance.

5 HANDILING MISSING VALUES WITH IMPUTATIONS

In the next step, SVM classifier with rbf kernel was evaluated using different imputation strategies: Statistical Imputation, KNN imputation and Multiple Imputation (MICE). The classification performance (same metrics as with RF) was compared across these approaches to assess how the choice of imputation method impacts the model's ability to handle missing data.

Table 2: F1-Score Class 0/Class 1 Comparison for Statistical Imputation Methods - Mushroom Dataset

Imputation Method	MCAR	MAR	MNAR
Mean	0.85/0.87	0.83/0.86	0.78/0.84
Median	0.84/0.87	0.80/0.85	0.78/0.84

5.1 Statistical Imputation

Statistical imputation replaces missing values with statistical measures, in this case the mean and median of the respective feature, providing a simple approach to handle missing data.

Similar trends were observed in all datasets. Table 2, using the Mushroom Dataset as an example, shows that imputation methods performed consistently across different types of missing data, but they were less effective than Random Forest, suggesting the introduction of erroneous information. Additionally, the difference between the mean and median strategies was minimal (the mean performed slightly better), with the worst results seen for the MNAR mechanism, while results for MCAR and MAR were almost identical.

Table 3: F1-Score Class 0/Class 1 KNN Imputation

Dataset	MCAR	MAR	MNAR
ADS	0.95/0.93	0.96/0.94	0.94/0.90
Mushroom	0.86/0.88	0.83/0.86	0.77/0.83
Occupancy	0.98/0.97	0.96/0.92	0.88/0.72

5.2 KNN Imputation

KNN Imputation fills missing data by using the values from the nearest neighbors in the feature space. In that case, KNN imputation with n_neighbors=5 was applied, and again it performed best for MAR and MCAR, while showing poorer results for MNAR, as shown in Table 3. Additionally, comparing the classifier evaluation metrics, the results are not significantly different from those obtained using Simple Imputer, in relation to the complexity of the datasets.

Table 4: F1-Score Class 0/Class 1 MICE Imputation

Dataset	MCAR	MAR	MNAR
ADS	0.97/0.94	0.97/0.93	0.96/0.90
Mushroom	0.86/0.87	0.84/0.86	0.78/0.84
Occupancy	0.97/0.95	0.97/0.94	0.88/0.72

5.3 MICE Imputation

MICE imputation, where a series of regressions are modeled for each feature with missing data, in theory should only be applied with MAR. However, it was applied to the other missing data mechanisms as well, to investigate this approach on various scenarios. The process was repeated multiple times (in this case 100 times) to enhance imputation stability. Once again, similar trends are observed as with previous strategies, but the results from the MICE imputation are the best among all methods Table 4.

6 PERFORMANCE COMPARISON

In the last step, predictive accuracy is assessed using the *Mean Absolute Error* (MAE), focusing on the model's ability to correctly predict outcomes. After evaluating classification performance, the model's predictive capability is checked to ensure consistency in the presence of missing data. This step ensures that the model is not only classifying correctly but also making accurate predictions, despite the challenges introduced by missing values.

The patterns observed in the predictive accuracy (MAE) for different imputation methods are consistent across the other datasets as well. Table 5 presents the results for the Occupancy dataset.

Table 5: Predictive Accuracy (MAE) for Different Imputation Methods - Occupancy Dataset

Missing Mechanism	Mean Imputation	Median Imputation	KNN Imputation	MICE Imputation
MCAR	0.0077	0.0056	0.0004	0.0036
MAR	0.0090	0.0091	0.0036	0.0097
MNAR	0.0110	0.0110	0.0110	0.0110

In this analysis, KNN imputation outperforms other methods across MCAR and MAR mechanisms. It means that KNN Imputation can recover missing values in the most accurate way. Mean and Median imputation methods exhibit similar performance. However, for MNAR, all methods yield the same and the highest MAE, suggesting there are some other methods to predict accurately this type of missing data.

7 CONCLUSIONS

- The Random Forest Classifier performed well in the above cases, providing satisfactory performance handling missing values.
- The various missing mechanisms affected performance in different ways, but with the most negative impact observed for MNAR.
- Among the various imputation strategies, KNN imputation performed the best, particularly in cases of MCAR and MAR, proving to be the most robust method.
- Mean and Median imputation provided similar results.
- Current methods struggled in the MNAR scenario. Future research may explore more advanced techniques or hybrid approaches to further improvement.

REFERENCES

[1] Stef Van Buuren. Flexible imputation of missing data, 2018.
[2] Fadi Thabtah. Autism screening adult. https://archive.ics.uci.edu/dataset/426/autism+screening+adult, 2017. UCI Machine Learning Repository, Donated on: 2017-12-23, Accessed: 2025-01-10.
[3] Prisha Sawhney. Mushroom dataset. https://www.kaggle.com/datasets/prishasawhney/mushroom-dataset, 2023. Accessed: 2025-01-10.
[4] Luis Candanedo. Occupancy detection. https://archive.ics.uci.edu/dataset/357/occupancy+detection, 2016. UCI Machine Learning Repository, Donated on: 2016-02-28, Accessed: 2025-01-10.