Effects of Missing Data on Fairness: Analysing the Impact of Data Imputation in Automated Decision Making

João Miguel Peixoto Lamas
Student of the Master Degree in Artificial Intelligence at the
Faculty of Engineering & Faculty of Sciences of the University of Porto
Project for the Artificial Intelligence and Society Course
(Dated: January 5, 2026)

Real world datasets frequently contain missing values, which require the use of imputation methods that can silently affect the fairness of the decisions made by machine learning models. This study investigates the impact of data imputation on model performance and fairness and the trade-off between these metrics. Three distinct strategies are analysed: Median, K-Nearest Neighbours (KNN) and Multivariate Imputation by Chained Equations (MICE)[1], alongside a native handling baseline for XGBoost[2], across the Adult Census Income[3], COMPAS Recidivism[4] and German Credit[5] datasets. By implementing a rigorous pipeline and utilising a fairness-adjusted score ($\lambda = 1.5$) it is demonstrated that while imputation generally improves predictive accuracy, it also frequently "recovers" bias native to the data, which can make the model less fair compared to native handling. The analysis contrasts a standard single-seed evaluation with a robust 50-seed stability test to address the unreliability of isolated runs. Results indicate that while a simple method was superior in a specific single-seed run, the complex method (MICE) demonstrated greater stability and performance across multiple runs. The findings of this study highlight that imputation is not a neutral pre-processing step but a determinant and impactful modelling choice, which often requires multi-seed stability analysis to ensure accurate but fair deployment in sensitive contexts.

I. INTRODUCTION & MOTIVATION

The presence of Machine Learning models is increasing in high-stakes scenarios, such as credit scoring and hiring. The decision made by these models determine the access to life-changing opportunities, making fairness an important social-technical requirement. However, the quality of the decisions made by said models is heavily reliant on the quality of the data the model is trained with. The presence of missing data is a consistent challenge in real-world machine learning, which can happen due to numerous factors such as a user refusing to disclose some information or errors in data storage.

While modern tree-based models (e.g., XGBoost) can often handle missing data internally, classifiers such as Logistic Regression and Neural Networks cannot function with missing values. This results in practitioners relying on imputation, which is the process of replacing missing values with estimates, as a standard procedure. While this can be treated as a neutral pre-processing step, it is a determinant and impactful modelling choice. By imputing values taking into account the existing patterns in the data, imputation algorithms reconstruct these same patterns. This leads to a trade-off between performance and fairness. While accurate imputation manages to restore lost information, which can lead to potential improvements on the model's performance, if the data contains bias the imputation methods can propagate this bias to replace the missing values. In cases where missingness acted as a cover for sensitive attributes, accurately imputing that data might recover the bias, although inadvertently, which can lead to discriminatory decisions.

Furthermore, the evaluation of fairness in machine learning is notoriously unstable. As demonstrated in the study performed by Friedler et al.[6], fairness metrics are highly sensitive to fluctuations in dataset composition, which means that conclusions that were drawn from a single train-test split can be quite misleading. This study bridges the gap between missing data handling and algorithmic fairness by analysing the impact of three distinct imputation strategies (Median, KNN and MICE) alongside a native handling baseline for XGBoost, across the Adult Census Income, COMPAS and German Credit datasets. Beyond the standard performance metrics, this work highly values stability: by contrasting a detailed single-seed analysis with a robust 50-seed experiment, this report determines whether complex imputation methods offer a reliable advantage over simple heuristic across multiple setups.

II. BACKGROUND & RELATED WORK

The investigations on algorithmic fairness started by proposing multiple theoretical definitions, but underwent a major shift to where it now consists in rigorously evaluating methods that improve fairness across multiple scenarios, mainly on their stability and performance. The comparative study by Friedler et al.[6] was decisive in this change by introducing a standardised framework for benchmarking fairness algorithms. Their analysis highlighted the fact that "fairness interventions might be more brittle than previously thought", demonstrating that theoretical guarantees often fail when training data is slightly modified. This work also shows that fairness is not exclusively a property of the algorithm, but it is heavily dependent on the preprocessing done and the composition of the underlying data.

Complementing these findings, recent work has examined specific preprocessing decisions such as the handling of missing data. By extending benchmarking to the data preparation pipeline, Caton et al.[7] investigated the effects of multiple imputation strategies on fairness. This work complements Friedler et al.'s by proving that the chosen imputation method can introduce bias before a fairness algorithm is even applied. These two studies show that both the learning algorithm and the data cleaning techniques must be considered in order to achieve algorithmic fairness.

III. METHODOLOGY

This study implements a robust evaluation pipeline to quantify the impact of different imputation methods on fairness and performance. It is designed to simulate realistic issues with data quality while preventing data leakage.

A. Datasets and Preprocessing

The analysis is conducted on three datasets that are widely recognised for containing historical biases against specific demographic groups: Adult Census Income, COMPAS Recidivism and German Credit. To ensure compatibility, the same preprocessing strategy was applied to all of them: countinous features were standardised with Z-Score normalisation and categorical features were one-hot encoded. The specific characteristics of each dataset are summarised in Table 1.

Dataset	Sample Size (N)	Sensitive Attribute	Target
Adult	45,222	Gender	Income (>50K)
COMPAS	5,278	Race	two_year_recid
German	1,000	Sex	Risk

TABLE I: Dataset Descriptions

B. Simulation of Missingness

To evaluate model robustness across different data quality scenarios, missing values were manually injected into the complete datasets. A baseline was established by training models on the original (unmodified) data to define an upper bound for performance. Three missingness mechanisms were simulated on both train and test sets:

- MCAR (Missing Completely At Random)[8]: Values were removed at random from a specific column at rates of 15% and 35% to simulate random data loss.
- MAR (Missing At Random): Missingness in a specific column was dependent on observed values of other columns (i.e., missingness in "education" dependent on "age"). The overall missing rate of the targeted column is 25%.
- Group-Based: Missingness in a specific column is dependent on the value observed in the sensitive column (unprivileged group vs privileged group). The base missing rate was defined as 25%, but entries that were part of the unprivileged group had their missing rate increased by half (1.5x), while entries that were part of the privileged group had their missing rate decreased by half (0.5x). This simulates a scenario where marginalised groups suffer from lower data quality.

C. Imputation Pipeline

A strict pipeline was designed and implemented to prevent data leakage, which could give false results. The dataset is first split into Training (80%) and Test (20%) sets using a stratified split to preserve the target class' representation across both sets. As illustrated in Figure 1, imputation models are fitted exclusively on the Training set and the learned parameters are then applied to transform both Train and Test sets.

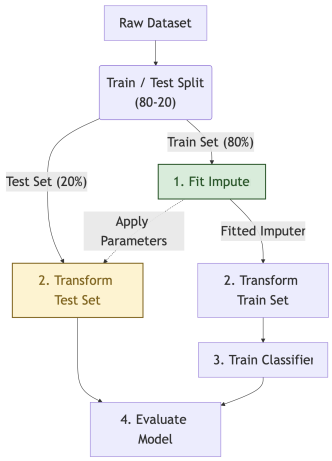

FIG. 1: Pipeline Diagram

D. Models and Imputation Methods

Three classifiers were evaluated, representing different complexity classes: Logistic Regression (linear), XGBoost (tree-based ensemble) and Multi-Layer Perceptron (MLP) (neural network). In order to isolate the effect of imputation, fixed default hyperparameters were used for all classifiers. For every missing data pattern, 4 imputation methods were considered:

- Median Imputation: Simple Imputation.

- KNN Imputation: Distance-weighted neighbour imputation (k=5).
- MICE: Iterative Multivariate Imputation
- Native Handling (XGBoost only): Making use of XGBoost native handling of missing values, therefore no imputation was performed.

E. Model Evaluation and Validity

A model's performance was evaluated using F1 Score, while fairness is measured by the Demographic Parity Gap (DP) between the two groups. To rank the effectiveness of each method, a Model Score that favours fairness was considered:

$$\text{Model Score} = F1 - 1.5 \times \text{DP Gap}$$

Where λ=1.5 was chosen to penalise unfair models while still considering their performance.

To ensure the validity of the models, a strict filter was applied to all models to remove those that are trivial. The DP Gap can often be minimised (resulting in a model that is equally fair to both groups) in a trivial way by classifiers that predict the same class for all individuals, achieving maximum fairness but no utility value. To prevent such models from impacting the results, any model that contains a Selection Rate (SR) outside the interval [10%, 90%] was marked as trivial and removed.

F. Stability Measures

To handle the known fragile nature of fairness metrics, the experiment is divided into two distinct phases. Phase I conducts an in-depth analysis on a single random split (Seed 42) to investigate specific behavourial patterns of imputation methods. Phase II expands this into a stability test in 50 random seeds (controlled for reproducibility). This multi-seed validation quantifies the variance of the Model Score, distinguishing between methodological gains and variance introduced by the random train-test split.

IV. RESULTS AND DISCUSSION

The results analysis is structured to validate model reliability before investigating the specific impact of imputation on fairness and utility.

A. Impact of Validity Constraints

The filter was essential for correct evaluation, even before missingness was introduced. In the Clean Baseline (no missing data) for the German Credit dataset, the Multi-Layer Perceptron (MLP) was initially the "fairest" model, achieving a Demographic Parity Gap of ≈ 0.014. However, inspection revealed this result to be an artifact of trivial classification (Selection Rate $\approx 99\%$), where the model effectively accepted all applicants.

After applying the validity constraint ($10\% < SR < 90\%$), the MLP alongside other models were removed and XGBoost was identified as the model that actually performed the best. This confirms that fairness metrics without validity checks can prioritise trivial models.

After applying this same filter to all experiments, Table II presents the best valid strategies for the Group-Specific Missingness scenario.

Dataset	**Best Valid Strategy**	**F1**	**Gap**
Adult	XGBoost + Median	0.701	0.178
COMPAS	XGBoost + MICE	0.596	0.157
German	XGBoost + Median	0.826	0.093

TABLE II: Best Performing Valid Models (Post-Filter) for the Group-Specific Missingness scenario.

B. Fairness-Utility Trade-offs

Figure 2 illustrates the trade-off between utility (F1-Score) and fairness (DP Gap). The impact of missingness varies significantly by dataset structure:

- Dataset Sensitivity: The Adult dataset (left panel) exhibits a tight clustering of results, indicating that the fairness-utility relationship is relatively robust to the choice of imputation method, likely due to the large sample size. In contrast, the both the COMPAS (center panel) and German Credit dataset (right panel) display a wide vertical dispersion (DP Gap ranging from ≈ 0.16 to ≈ 0.33 for the COMPAS dataset, ≈ 0.03 to 0.26 for the German Credit dataset), suggesting that small changes in the imputation strategy can drastically alter the fairness outcome.
- Model Robustness: Across all datasets, the XGBoost classifier (marked by 'X') consistently maintains high F1-scores compared to Logistic Regression (circles) and MLP (squares), often being closer the preferred region of the graph than the other classifiers (bottom right corner, where F1 is Highest and DP Gap is lowest).

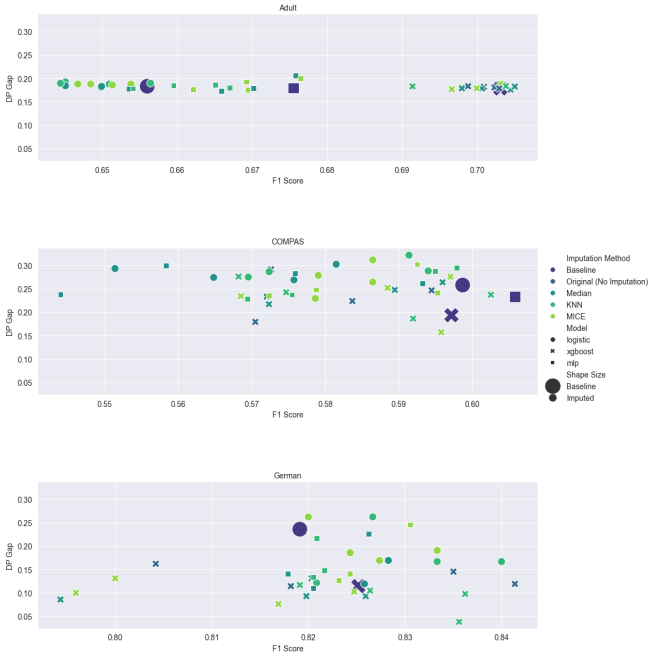

FIG. 2: Fairness vs. Performance Trade-off

C. Impact of Imputation on Fairness

The bar plot in Figure 3 quantifies the "Bias Reconstruction" effect. The red dashed line represents the baseline fairness gap of the original, complete data.

- Amplified Bias: In the COMPAS dataset (middle row), imputation strategies consistently result in a Demographic Parity Gap that exceeds the baseline (bars rising above the red line). This is particularly evident in the MCAR scenarios, all methods (even XGBoost's native handling) increase the DP Gap, confirming that "guessing" missing values can amplify existing societal biases in the data.

- Variable Impact: In contrast, the Adult dataset (top row) shows minimal deviation from the baseline, further supporting the finding that datasets with stronger signal-to-noise ratios are less susceptible to imputation-induced bias. As explained previously, the sample size of this dataset also likely played a role in stabilising results across the multiple scenarios.

- Small Sample Volatility: The German Credit dataset exhibits high susceptibility to variance due to its limited sample size. It was found that MCAR at a missing rate of 35% had lower DP Gap than the same missing mechanism at a lower rate (15%), which is counter intuitive. This irregular behaviour indicates that in small datasets, apparent "fairness" can be an artifact of statistical noise, where random data deletion accidentally disrupts discriminatory correlations, rather than a genuine algorithmic improvement.

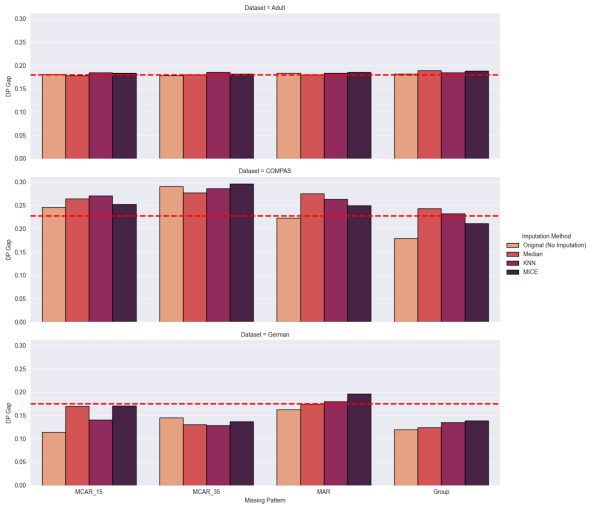

FIG. 3: Impact of Imputation on Fairness Metrics. The horizontal line marks the baseline performance (Clean Data)

D. Statistical Significance vs. Single-Split "Luck"

This study shows demonstrates that relying on single train-test splits is dangerous and can have misleading results. On the initial split (Seed 42), the simple Median baseline appeared superior to MICE. However, a T-Test across 50 seeds reveals that MICE is, in fact, statistically significantly better ($p < 0.05$) for this case.

To understand this discrepancy, we analyzed the imputation fidelity. MICE achieved an RMSE of 4.387 compared to Median's 5.174 (a 15.2% fidelity gain). This confirms that MICE was objectively better at reconstructing the data, but the specific random partition of Seed 42 masked this advantage. This highlights that "fairness" results on single splits can be deceptively influenced by random data partitioning, validating the need for multi-seed rigor.

V. CONCLUSION AND FUTURE WORK

This study evaluated the impact of missing data handling on algorithmic fairness. Our results demonstrate that while complex imputation strategies like MICE can recover lost utility (RMSE improvement of 15.2%), they often do so at the cost of fairness and stability. Specifically, we found that:

1. **Imputation requires validity checks:** Without strict triviality filters, fairness metrics can be easily gamed by degenerate models that reject all candidates.

2. **Complexity introduces instability:** Multivariate imputation exhibited higher variance across random splits compared to simple heuristics, confirming that complex reconstruction is sensitive to data partitioning.

3. **Native handling is robust:** For tree-based models, allowing the algorithm to handle missingness internally (XGBoost) consistently provided the most stable Fairness-Utility trade-off.

A. Future Directions

Future work should extend this analysis beyond standard demographic parity. Potential avenues include:

- **Fairness-Aware Imputation:** Standard imputers optimize solely for data fidelity (RMSE). Future research could explore "fair imputers" that incorporate a regularization term into the reconstruction loss, explicitly penalizing the correlation between imputed values and sensitive attributes.

- **Causal Modeling of Missingness:** While this study assumed a Group-Specific (MNAR) mechanism, future work should apply Causal Structural Models (SCM) to distinguish between missingness caused by systemic bias versus random attrition, allowing for more targeted data correction strategies.

- **Intersectional Analysis:** Systemic bias is rarely unidimensional. Validating these stability protocols on intersectional subgroups (e.g., Race × Gender) is crucial to ensure that "robust" methods do not hide bias within smaller, multiply-marginalized cohorts.

VI. ACKNOWLEDGMENTS

This document was refined with the assistance of Large Language Models, which helped check grammar, correct typos and enhance clarity. The overall content and ideas remain solely the responsibility of the author

REPRODUCIBILITY AND SUPPLEMENTAL MATERIAL

To ensure full reproducibility of the experiments and statistical tests reported in this study, we have released the complete code, detailed experimental logs, and high-resolution versions of all figures.

The repository includes:

- **Source Code:** Python notebooks for the missingness generation (MCAR/MNAR), imputation (MICE, KNN), and model training pipelines.
- **Detailed Results:** Full CSV logs containing precision, recall, and F1-scores for all 50 random seeds across all datasets.

- **Supplementary Figures:** Extended visualizations of the MCAR severity analysis and the prefilter triviality rankings.

Repository URL: https://github.com/jmplamas/Effects_Missing_Data_Fairness

[1] S. Van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in r," *Journal of Statistical Software*, vol. 45, no. 3, pp. 1–67, 2011.

[2] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[3] W. Liu, "Adult income dataset," Kaggle, 2016. [Online]. Available: https://www.kaggle.com/datasets/wenruliu/adult-income-dataset

[4] ProPublica, "Compas analysis repository," https://github.com/propublica/compas-analysis, 2016.

[5] H. Kabure, "German credit data with risk," Kaggle, 2018. [Online]. Available: https://www.kaggle.com/datasets/kabure/german-credit-data-with-risk/data

[6] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, "A comparative study of fairness-enhancing interventions in machine learning," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 2019, pp. 329–338.

[7] S. Caton, S. Malisetty, and C. Haas, "Impact of imputation strategies on fairness in machine learning," *Journal of Artificial Intelligence Research*, vol. 74, pp. 1011–1035, 2022.

[8] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*, 3rd ed. John Wiley & Sons, 2019.