Journal of Artificial Intelligence Research 74 (2022) 1011-1035 Submitted 07/2021; published 06/2022

Impact of Imputation Strategies on Fairness in Machine Learning

Simon Caton SIMON.CATON@UCD.IE
School of Computer Science, University College Dublin, Ireland

Saiteja Malisetty SMALISETTY@UNOMAHA.EDU
University of Nebraska at Omaha

Christian Haas CHRISTIAN.HAAS@WU.AC.AT
Department of Strategy and Innovation
Vienna University of Economics and Business (WU), Austria

Abstract

Research on Fairness and Bias Mitigation in Machine Learning often uses a set of reference datasets for the design and evaluation of novel approaches or definitions. While these datasets are well structured and useful for the comparison of various approaches, they do not reflect that datasets commonly used in real-world applications can have missing values. When such missing values are encountered, the use of imputation strategies is commonplace. However, as imputation strategies potentially alter the distribution of data they can also affect the performance, and potentially the fairness, of the resulting predictions, a topic not yet well understood in the fairness literature. In this article, we investigate the impact of different imputation strategies on classical performance and fairness in classification settings. We find that the selected imputation strategy, along with other factors including the type of classification algorithm, can significantly affect performance and fairness outcomes. The results of our experiments indicate that the choice of imputation strategy is an important factor when considering fairness in Machine Learning. We also provide some insights and guidance for researchers to help navigate imputation approaches for fairness.

1. Introduction

Over the last few years, many companies have come to rely on automated decision-making systems to classify people on the basis of their eligibility and/or suitability; often in the areas of credit assessments, job screening, insurance coverage, and marketing purposes (Zafar et al., 2017; Biega et al., 2018). In addition, they are also used in the public sector, including government service delivery and probation decisions in the criminal justice sector (Zafar et al., 2017; Hu & Chen, 2018; Noriega-Campero et al., 2020). Many of these systems rely on conventional statistical and Machine Learning techniques such as regression analysis (Logistic Regression), decision trees, and other classification algorithms.

Nonetheless, a significant new problem with Machine Learning systems is when and how they address potential bias(es) in the decision-making process. There have been many observations of systems unfairly treating specific demographics of individuals which often leads to different sociocultural biases and/or discrimination (for example on the basis of gender, race, age, etc.) (Corbett-Davies et al., 2017a; Edelman et al., 2017; Kleinberg et al., 2018). In response, many approaches to make Machine Learning models "fair(er)" have emerged which seek to find a balance between model performance and prediction fairness (see Caton and Haas (2020) for an overview of approaches to fairness in Machine Learning). Yet, much of the recent research into algorithmic

fairness considers approaches on how such bias and discrimination can be mitigated in a Machine Learning model and many of these works assume that the data is both clean and complete.

While many publications in Machine Learning fairness use the same set of selected datasets, real world datasets are often 'messier' and includes missing data. This missing information can be a problem when we try to train a model on real data or calculate its performance. In order to obtain a classification prediction for each observation (potentially with missing values), we need to deal with (or treat) this missing information. Two common strategies are deleting the entire observation with missing values, or imputing the missing values using various imputation techniques. Deleting the entire observation with missing values has less impact on the overall outcome (e.g., measured by performance metrics) if we have a large volume of data proportional to the number of observations with missing data. However, for smaller data sets, or when removing data is impractical (each observation needs to have a prediction), it is common to impute the missing data as we might lose valuable information while deleting sets of observations.

For imputation, a variety of different strategies exist that try to replace the missing data with estimates (Song & Shepperd, 2007). While imputation strategies can affect the performance of Machine Learning algorithms, their impact on the fairness of the resulting predictions is unclear. Due to the prevalence of missing data in real world applications, in this article we aim to analyze the impact of imputation strategies on various fairness and performance metrics, and in doing so address the following research question:

Research Question: Is there a significant effect of imputation strategy, classification algorithm, or both, on fairness and performance metrics in classification settings?

We consider the impact of missing values, and their corresponding imputation, on the classification outcome using three commonly used datasets in the fairness literature: Adult Income, COMPAS, and German Credit. As these three datasets are well structured and, in standard implementations, do not contain missing values, we randomly delete different percentages of values from the three datasets (i.e., data is missing completely at random) and impute the data using nine different imputation techniques to compare how these imputation techniques affect various fairness and performance metrics. We run structured experiments to iterate over a range of parameters, including different classification algorithms, imputation strategies, and percentage of missing values. Each scenario is independently repeated 100 times (totalling 208800 observations) to calculate distributions for the performance and fairness metrics corresponding to the classification outcomes. We analyze the resulting impact using statistical techniques and identify scenarios where imputation strategies and choice of Machine Learning model affect the fairness of the predictions. The results in this article aim to increase awareness of the impact of imputation strategies, and missing data as a whole, in fair Machine Learning pipelines, and help researchers to determine (potentially) promising imputation strategies that work best for given scenarios.

We discuss existing research in the field of Algorithmic Fairness in Section 2. Section 3 briefly describes our key implementation steps along with pseudo code of the experimental setup and gives an overview of fairness metrics and the imputation strategies that we use in our study. We discuss our empirical findings in Section 4, and discuss their implications in Section 5 with the goal of providing tangible guidance for the handling of missing data in fairness pipelines. Finally, we conclude our work in Section 6 providing an outlook on future work.

2. Related Work

In this section, we seek to situate this article in the literature and give relevant background on key topics. To this end, we start with a brief introduction to fairness in the context of Machine Learning and its associated challenges. We then discuss missing data and methods of handling, or treating, missing data both with and without fairness considerations.

2.1 Fairness in Machine Learning

The importance of fairness and related ethical principles is recognized as key to improving the trustworthiness of Machine Learning (and AI in general) (Commission, Directorate-General for Communications Networks, & Technology, 2019). However, Fairness in Machine Learning is a complex topic that has to try and balance many different contexts, multi-faceted sociocultural concepts (beyond concepts of race, gender, age), aspects of equality and diversity, and ethical principles such as whether a Machine Learning model is even appropriate. When specifically discussing fairness in Machine Learning, the literature typically focuses on either the technical aspects of bias and fairness in Machine Learning, or theorizes on the social, legal, and ethical aspects of discrimination (Goodman, 2016).

In this paper, we focus only on technical aspects of fairness, and thus refer the reader to other key survey papers for a more general introduction to fairness in Machine Learning as well as an overview of the current state of the art (Romei & Ruggieri, 2014; Mitchell et al., 2018; Hutchinson & Mitchell, 2019; Suresh & Guttag, 2019; Blodgett et al., 2020; Caton & Haas, 2020; Mehrabi et al., 2021). Similarly, for more ethical discussions around notions of fairness and related ethical principles, we refer the reader to the following studies: Skirpan and Gorelick (2017), Dignum (2021), Lepri et al. (2017), Binns (2018), Sokolovska and Kocarev (2018), Veale et al. (2018), Feldman et al. (2015). Specifically, as we highlighted in Caton and Haas (2020), there is a significant number of dilemmas that fairness in Machine Learning researchers still need to address. An important point to make is that this work does not (and cannot) emphasize whether the application of a Machine Learning model is appropriate or not, instead, our goal is to illustrate how current technical approaches to fairness and corresponding quantitative measures of fairness are affected by strategies to handle missing data.

In general, the basic premise of technical approaches to fairness in Machine Learning is that there exist some protected or sensitive attributes (age, sex, race etc., but essentially any feature of the data that involves or concerns people (Barocas et al., 2019)) which are either viewed as privileged (usually in receipt of a positive outcome such as being offered a loan) or unprivileged (usually in receipt of a negative outcome: not being offered a loan). These protected attributes, and the corresponding privileged and unprivileged observations, are then often used to define a variety of different fairness metrics and build up quantitative notions of how "fair" a Machine Learning model is. From the perspective of how Machine Learning is made "fair(er)" a variety of intervention-based approaches have been devised, which either: 1) treat the data prior to the training of a Machine Learning model: *pre-processing approaches*; 2) use quantitative measures of "fairness" within the Machine Learning model's objective function, thus altering how the model is trained: *in-processing approaches*; and 3) treat the output(s) of the Machine Learning model to "improve" some quantitative notion of fairness: *post-processing approaches*. In each of these different intervention types, a machine learning model is fit to the data, and used to produce a set of predictions. The fairness metric is applied to these predictions to derive how "fair" the Machine Learning model is. In this paper, we simplify this process somewhat to a "leaner" intervention: we treat the data (imputation) as a pre-processing step,

fit a Machine Learning model, and then compute a selection of commonly used fairness metrics. While it would be prudent to explore the effects of missing data on a large number of intervention approaches from the literature, we seek in this paper to highlight the impact(s) of missing data and imputation in general, thus motivating more expansive studies across the field.

Over the past 10 or so years, the definition of fairness in automated decision systems, the impact of biases leading to unfairness and how these affect both the fairness and classical performance of Machine Learning models, and different forms of interventions have been studied extensively. Initially, work in the area was concerned only with binary classification, but soon approaches to investigate biases in recommender systems, natural language processing systems, regression problems, and more recently even complex applications such as computer vision, speech processing and machine translation (among others) have also started to emerge. Rather than go into significant detail in these areas, the key facets of fairness, its definitions, and approaches, we instead focus our discussion of related work solely on missing data, and how missing data in the context of fairness in Machine Learning has been studied.

2.2 Missing Data and Imputation

Missing data is often a problem in many research studies across various disciplines (Myrtveit et al., 2001; Sinharay et al., 2001; Song & Shepperd, 2007; Hardy et al., 2009; Cheema, 2014). According to Soley-Bori (2013), there are multiple methods to handle missing data depending on the type of missing data. The authors suggest to use conventional techniques like list-wise deletion and imputation if the data is missing completely at random. In a study by Myrtveit et al. (2001), the authors compare the results of four techniques (list-wise deletion, mean imputation, pattern imputation & maximum likelihood) in handling missing data and determine that maximum likelihood works better when data is not missing completely at random. In a study by Donders et al. (2006), the authors show that the most commonly used techniques in handling missing data (mean imputation & missing indicator method) will always provide biased estimates. Many studies have used multiple imputation methods for missing data which yielded better results than single imputation techniques (Sinharay et al., 2001; Wayman, 2003). Previous results also show that choosing an imputation strategy will impact the results of classification error rate by up to 10 percent (Farhangfar et al., 2008), suggesting that not all imputation techniques are the same. In our study, we use multiple imputation strategies and compare their effect on both fairness and classification performance metrics.

2.3 Missing Data and Fairness

Specifically with regard to missing data and fairness, there is surprisingly little literature to discuss. Fernando et al. (2019, 2021) provide a comprehensive discussion on the considerations of missing data on fairness. They observe that the causes of unfairness and missing data are related, and that imputation strategies should be useful in the presence of missing data specifically for fairness, i.e., better than deletion. Yet, they note that there is still a lack of understanding with regard to how different imputation strategies affect fairness, and it is here that this article seeks to provide some guidance and insights. Another related study is Wang and Singh (2021), who look into the effects on fairness of missing values and selection bias for categorical data. They propose the uses of reweighting and different sampling methods (both pre-processing fairness interventions) to find a balance and fairness and performance (here: accuracy). We extend these works by looking at multiple imputation methods, applied as a pre-processing step for mixed data, i.e., datasets that have both

numeric (both discrete and continuous valued) as well as categorical features. We also try to bring to the fore how different Machine Learning models can affect the fairness vs. performance trade-off in different ways, and also seek to start some discussion on the notion of fairness preferences and how different treatment methodologies can impact canonical fairness preferences.

However, beyond explicit works that seek to consider fairness and missing data, there are some additional considerations. Imputation strategies often use other features in the data to derive values for missing data. Yet, caution is needed here specifically with respect to fairness because of the observed effects of proxy variables (sometimes referred to as quasi-identifiers), i.e., variables which are correlated with protected attributes. Not considering these correlations between proxy variables and protected attributes has been extensively shown to increase the risk of discrimination (Pedreshi et al., 2008; Dwork et al., 2012; Romei & Ruggieri, 2014; Lum & Johndrow, 2016; Zarsky, 2016; Veale & Binns, 2017; Calmon et al., 2017; du Pin Calmon et al., 2018; Lipton et al., 2018). We provide an overview of common proxy variables in Caton and Haas (2020). This has obvious implications for imputation strategies that use relationships within the data, i.e., are multivariate, and are used to repair or treat missing values. Research into the impact(s) of missing data on fairness is still in its early days, yet we highlight these challenges here as they are currently severely under-addressed in the literature, especially when considering missing data. Whilst, we do not explicitly address this challenge in this article, we highlight it as an area that will require significant research as the understanding of missing data within the fairness literature advances.

3. Methodology

To outline our approach and experimental setting, we first discuss different fairness metrics leveraged in this work (subsection 3.1) and follow this with a discussion of classical Machine Learning metrics (subsection 3.2) and base Machine Learning models we employ for classification tasks. These both provide the basis for discussing trade-offs between model fairness and performance in the presence of missing data. Next, we introduce the imputation strategies leveraged to treat missing data (subsection 3.3). Finally, we present the factorial experimental design used to explore the impacts of missing data and correspondingly imputation on model performance and fairness (subsection 3.4).

3.1 Fairness Metrics

Over the past years, building on fairness research in other domains, a variety of different fairness metrics have been suggested for Machine Learning. Fairness metrics are generally defined using protected attributes (e.g., race, gender & age) and privileged classes (e.g., white, male & adult) among those protected attributes that represent the groups (or individuals) that we want to compare. For example, these can be the groups that are potentially discriminated against, compared to other groups, by an automated decision making system. Typically, the literature discusses fairness from the perspective of individuals (everyone is treated the same) or groups (sociodemographically or socioculturally defined groups are treated equally). There is much debate over the "correct" manner(s) to define fairness for Machine Learning, and many notions of fairness are incompatible with each other (Chouldechova, 2017; Kleinberg et al., 2018). In this work, we use five commonly applied metrics to define the fairness of a Machine Learning model (below) allowing us to discuss the effect(s) of imputation from several perspectives.

- Statistical Parity (Demographic Parity) Difference: The difference between favourable outcomes received by the unprivileged group and privileged group. For a fair model/data this metric needs to be closer to zero (Kamishima et al., 2012; Zemel et al., 2013; Feldman et al., 2015; Corbett-Davies et al., 2017a).

- Equal Opportunity Difference: The difference between the true positive rate of unprivileged group and the true positive rate of privileged group. For a fair model/data this metric needs to be closer to zero (Hardt et al., 2016; Pleiss et al., 2017).

- Average Absolute Odds (Equalized Odds) Difference: The sum of the absolute differences between the true positive rate and the false positive rates of the unprivileged group and the true positive rate and the false positive rates of the privileged group. For a fair model/data this metric needs to be closer to zero.

- Disparate Impact: The ratio of favourable outcome for the unprivileged group to that of the privileged group. For a fair model/data this metric needs to be closer to one (Feldman et al., 2015).

- Theil Index: The inequality in benefit allocation for individuals. For a fair model/data this metric needs to be closer to zero (Speicher et al., 2018).

3.2 Classification Metrics and Models

For our study we use three standard classification algorithms: Logistic Regression, Random Forest, and Linear Support Vector Classifier to calculate predictions on our treated data. These algorithms have been chosen for both their simplicity (they require relatively little parameter tuning in comparison to other algorithms) and they are all fairly "standard" Machine Learning algorithms. They also have well established implementations in scikit-learn (Pedregosa et al., 2011) aiding the reproducibility of our results. These three classification methods also represent a broad range of approaches to classification. Thus, this enables the exploration of whether different Machine Learning algorithms respond differently to various imputation strategies in terms of fairness, classical performance, or both.

Traditionally, the predictive performance of classification algorithms is measured through a set of different performance metrics. These metrics are commonly used to evaluate the quality of the built Machine Learning model and are used alongside fairness metrics to study the effect of bias mitigation strategies, or other interventions, on the model outcome. In this article, we calculate the following standard Machine Learning classification metrics to evaluate the performance of the classifier in our evaluation:

- Accuracy: The ratio of number of correct predictions to the total number of predictions.

- Sensitivity: The ratio of how many predictions were correctly classified as positive to how many were actually positive.

- Specificity: The ratio of how many predictions were correctly classified as negative to how many were actually negative.

- F1 Score: The F1 score is a number between 0 and 1 and is the harmonic mean of precision and recall, where recall is same as sensitivity and precision determines how many actual positives are correctly classified among all the positive predictions.
- Area under Curve: AUC indicates how well the probabilities from the positive classes are separated from the negative classes.

We present both classical performance metrics alongside fairness metrics to capture the fairness vs. model performance trade-off. It has been observed by many researchers that improving either fairness or accuracy is often to detriment the other (Berk et al., 2018; Dwork et al., 2012; Corbett-Davies et al., 2017b; Hardt et al., 2016; Zliobaite, 2015; Calmon et al., 2017; Haas, 2019). Thus, in this paper, we present both perspectives and emphasize notions of fairness and accuracy over other possible effects of imputation as this best represents the fairness vs. performance trade-off.

3.3 Imputation Strategies

Imputation is a process of filling in the missing data with the some alternative values. There are many imputation strategies that we can choose based on the type of missing data (Soley-Bori, 2013). Here, we use nine of the most commonly used imputation strategies to impute missing data which are categorised in Table 1 below (and largely available in common Machine Learning libraries like scikit-learn).

Imputation Class	Strategy
Simple Imputer	Mean, Median, Most Frequent
kNN Imputer	Neighbours = 2[1]
Iterative Imputer	Bayesian Ridge
Single Imputer	Interpolate, Least Squares, Stochastic, Norm

Table 1: Imputation Strategies used in our Study

All imputation strategies, except for the most frequent imputation, were used to impute the missing data of numeric features. Whereas, for imputing missing categorical data, we have used only kNN & Most Frequent Imputation strategies. This means that we split our discussion on kNN (in key places) into when it operates on numeric data, and when it is used on categorical data; not doing so would add ambiguity to how it performs on different data types. Our goal is to observe whether different imputation strategies significantly affect the fairness, the classification performance, or both, of a Machine Learning model trained on imputed data.

The datasets that are commonly used in Machine Learning fairness research have no (or very few) missing data for all the variables. Hence, for our evaluation, we artificially introduce missing values first, and then compare various imputation strategies. To facilitate both categorical and numerical imputation strategies, we first split the dataset into two parts (numerical & categorical) based on

1. A valid question at this stage would be the motivation for $k = 2$, we recognize that higher values may perform better. Yet the results with kNN as an imputation strategy we feel are sufficiently adequate to not further explore higher values of k, as the general takeaways and discussions surrounding kNN as in imputation strategy would not likely be significantly different. We concede, however, that future work should empirically validate this expectation.

the data type of the variables. We then encode all the categorical variables including the protected variables to numerical values using a label encoder.

3.4 Experiment Setup

We define the protected attributes (Sex, Age, and Race) and privileged classes (Men, Adult, and White) for each of the datasets to calculate fairness metrics. Subsequently, we calculate the fairness metrics for the original dataset with no missing data using three classification algorithms (Logistic Regression, Random Forest, and Linear Support Vector Classifier) to create baseline observations of the performance and fairness for these datasets. We present results for protected attributes always as attribute 1 and attribute 2 with respect to different performance and fairness metrics. As "Sex" is shared between all datasets, attribute 1 always refers to sex, whereas attribute 2 refers to the remaining protected attributes for each dataset. We do this to try and simplify the presentation of results, otherwise the degree of results would be too fine grained across many different settings, datasets and performance / fairness metrics for structured discussion; at the cost of loss of generality in the discussion.

To conduct the main experimentation, we randomly delete values from each dataset at varying degrees of severity, increasing the number of affected features and apply the nine imputation strategies to perform data repair and fit the three Machine Learning models to the repaired data. Finally, we recompute the fairness and performance measures and compare these to the baseline observations. Algorithm 1 describes this approach more formally.

Variable	Values
Dataset with Protected Attributes	German (1: Sex, 2: Age), Adult (1: Sex, 2: Race), COMPAS (1: Sex, 2: Race)
Imputation Strategies	Mean, Median, Most Frequent, kNN, Iterative, Interpolate, Least Squares, Stochastic, Norm Imputation
Classification Algorithm	Logistic Regression, Random Forest, Linear Support Vector Classifier
Percent Deleted	1, 5, 10 Percent

Table 2: Overview of Experimental Parameters and their Values

To allow for a statistical analysis of the results, and to account for effects of random sampling, we independently repeat each scenario (parameterization of the methodology) 100 times. The set of parameters that are considered in the experiment are described in Table 2. The following analyses are based on these 100 iterations for each combination of dataset, classification algorithm, imputation strategy, and percentage of (artificially deleted) missing values. Note that previous publications have occasionally used different protected attributes for the three datasets, which is why we consider each of the two potential protected attributes separately. It should also be noted that until recently, few fairness interventions could handle multiple protected attributes concurrently, and it is more complicated to compute fairness metrics on the basis of multiple protected attributes.

We have two protected attributes for each dataset, sex & race for the COMPAS Dataset, sex & age for the German Dataset, and sex & race for the Adult Dataset. We have used 8 imputation strategies

Algorithm 1: Pseudocode of Evaluation Setup

```
Data: Datasets: ds ∈ {German, Adult, COMPAS}
Data: Imputation Strategies: 8 strategies for numerical imputation, 2 for categorical imputation
Data: Classification Algorithms: CA ∈ {LogisticRegression, LinearSVC, RandomForest}
Data: Repetition: i ∈ [1 : 100]
Data: Number of columns: j ∈ [1 : ncol(dataset)]
Data: Percentage of deleted values: p ∈ {0.01, 0.05, 0.1}
Result: Solution csv with Performance and Fairness metrics
begin
    Select dataset;
    for repetition i do
        for number of columns to consider j do
            Pick j columns from dataset at random;
            for percentage p do
                for column in columns to consider do
                    Randomly delete p percent of values in column;
                for imputation strategy imp do
                    Impute missing values based on strategy imp;
                    for algorithm algo do
                        Train classifier using algo;
                        Calculate performance and fairness metrics;
                        Add results to solution csv;
    Repeat for other datasets;
    return solution csv;
```

for numeric variables and two for imputing categorical variables, with kNN imputation applied to both data types. We have used mean & median strategies from simple imputer and interpolate, least squares, stochastic & norm strategies from single imputer for imputing numerical variables. Additionally, we have also used iterative imputer for numeric valuables and most frequent strategy for imputing categorical variables.

To conduct our analysis, we introduce three different percentages (1, 5, and 10%) of missing data per column. The columns as well as the number of columns where missing data is introduced are both randomly determined. Applying each imputation strategy, the three classification algorithms (Logistic Regression, Random Forest, and Linear Support Vector Classifier) are used to observe the effect(s) on fairness and classification metrics that introducing and subsequently correcting missing data had. To evaluate the results and provide a basis for discussion, we explore the metrics distributions, use a robust three way analysis of variance to determine if there is any influence of three categorical variables (Imputation Strategy, Classification Algorithm, and Percentage Missing Values) on both the fairness and classification metrics, and finally investigate ranking mechanisms for different imputation strategies.

For the implementation, we use Python as programming language and make use of several additional libraries. To facilitate the computation of various fairness metrics, we utilize the open source package AIF360 (Bellamy et al., 2019). The different classification algorithms use standard

scikit-learn implementations. All experiments and analyses were run on a 12 core workstation with 128 GB RAM to parallelise the independent repetitions.[2]

4. Statistical Results and Findings

To evaluate the effects of missing data and imputation strategies on model fairness and performance, we explore the experimental data aligned to the three main factors of the experimental design, i.e., i) percentage of missing data; ii) Machine Learning model, and iii) imputation strategy and their effect(s) on fairness and performance metrics. We note that our findings are on the basis of 208800 observations from 100 repetitions of 9 imputation strategies feeding 3 classification models for 3 datasets with 3 different extents of missing data for a range of affected columns. This results in 55,800 observations for the Adult dataset, 88,200 observations for the COMPAS dataset, and 64,800 observations for the German credit dataset. We also derive all performance measures on the original datasets, and thus can compare the results to what we would expect in the event of a "perfect" imputation, which although somewhat unrealistic, gives us an indication of how adversely affected the model is as a result of the imputation strategy applied, and by extension, the treatment of missing data in a more general sense.

To provide different views of the interplay between the main experimental factors, we structure our discussion into four parts. First, we look at the distribution of both fairness and classical performance metrics and comment on their practical implications in Section 4.1. Second, in Section 4.2, we discuss how the different Machine Learning models respond to remaining experimental factors (percentage of lost data, and imputation strategy) focusing our discussion on the performance (e.g., accuracy) vs. fairness trade-off. Thirdly, using a robust three factor analysis of variance, we look for empirical evidence that there are / are not relationships between the factors in Section 4.3, as this will shed some light on specific cases of interest and highlight interaction effects of note between the experimental factors. Finally, we use Friedman ranking (similar to Fernández-Delgado et al. (2014)) to construct a general ranked list of approaches in Section 4.4 to act as a means of discussion and future investigation and highlight how specific preferences with regard to the fairness vs. performance trade-off affect the methodological approach for imputation with fairness considerations. Upon the basis of these results, we discuss their general implications in Section 5.

We note that we do not concretely comment on how to make approaches "fair(er)". We abstain from this line of discussion as providing specific strategies on how to achieve fairness would be scenario-specific. As we see in in this section, there are explicit trade-offs evident in the treatment of missing data. Thus, how to concretely improve the fairness of a Machine Learning pipeline in the presence of missing data depends on the objectives and notion(s) of fairness under consideration.

4.1 Simple Distributional Observations

In order to observe the distributions of classification and fairness metrics with respect to both imputation strategy and classification algorithm, we have used box plots facetted by the fairness metric (in Figure 1) and performance metric (in Figure 2) and grouped them by the classification algorithm used. Note that "None" in in Figures 1 and 2 refers to the performance where no data was randomly deleted from the data set, and thus represents the expected performance were "perfect"

2. The code and datasets used to run our experiment can be found in the following GitHub Repository: `https://github.com/haas-christian/JAIR-Imputation`.

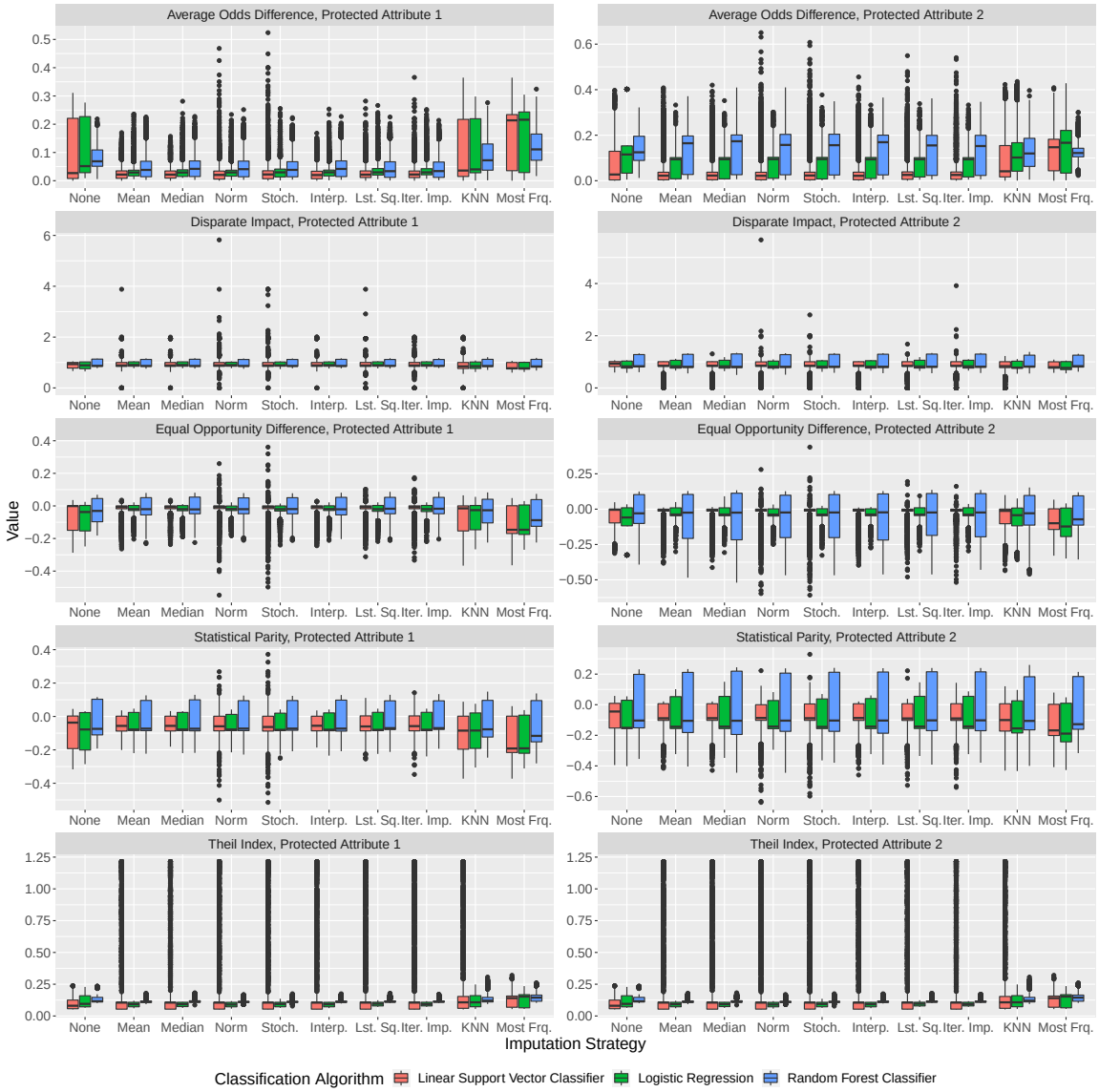

Figure 1: Fairness metrics w.r.t imputation and classification model. The fairness metrics are calculated separately depending on the considered sensitive attributed. Imputation strategies are shown on the x-axis. Different colors refer to the three considered classification algorithms.

imputation performed. Using the "None" scenario as baseline, we can see that although the variance is often higher than for some imputation strategies in some metrics, there is a general tendency for the variance to be lower in the majority (but not all) of cases, and rarely do we see a relatively high frequency of outliers.

In general, we can see that the presence alone of missing data leads to a change in distributional behaviour (i.e., more outliers, generally wider distribution, and sometimes also increases in variance)

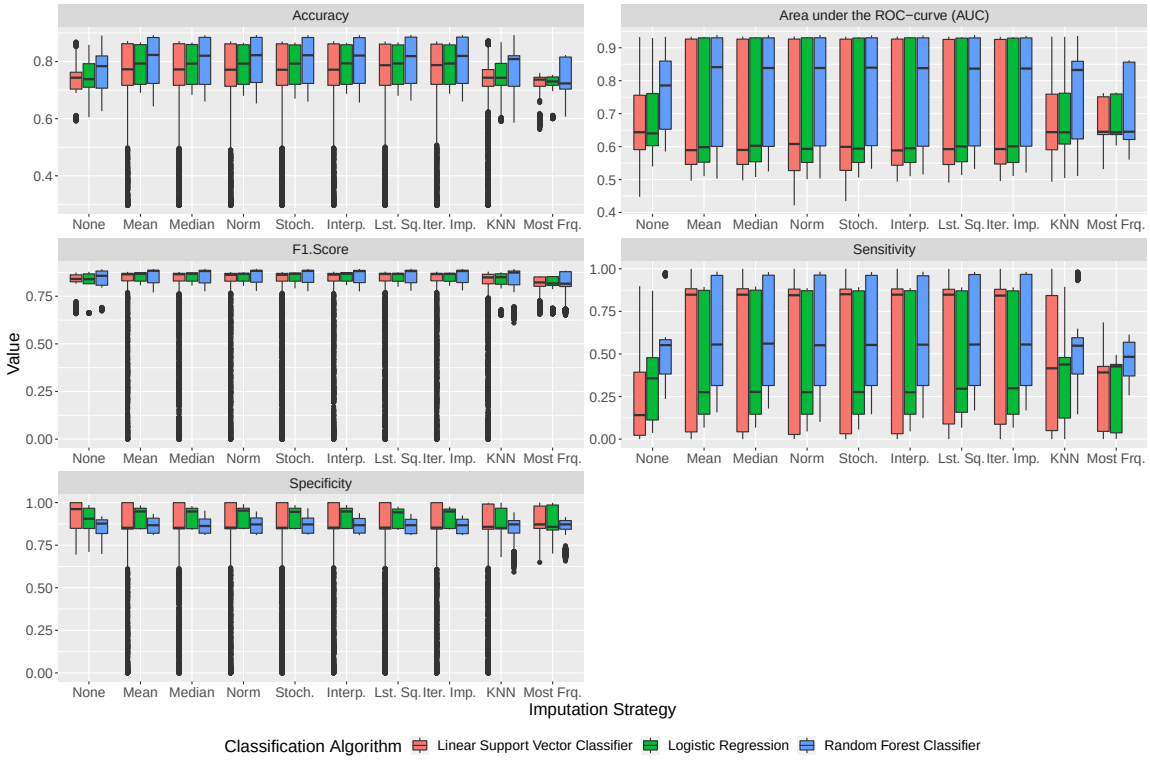

Figure 2: Performance metrics w.r.t imputation and classification model. Imputation strategies are shown on the x-axis. Different colors refer to the three considered classification algorithms.

across several performance and fairness metrics, i.e., models can become less fair sometimes considerably, and model performance (e.g., accuracy, AUC) is impacted as well. There is also a notable increase in outliers. Whilst neither of these observations is surprising, that different imputation strategies appear to affect specific performance and fairness metrics differently is. This would initially indicate that care is needed when handling missing values, and that significant testing is needed when treating missing data.

Considering the impact of imputation strategies, we can observe a clear difference in the distribution of both fairness and classification metrics while using the Most frequent and kNN imputation techniques. The other imputation strategies seem to result in more similar distribution of metrics, in particular fairness metrics as shown in Figure 1. This is not wholly surprising most frequent operates on categorical data, and kNN operates on both categorical and numeric data. Whereas all other strategies operate on numeric data only.

Considering specific fairness metrics, for average odds difference using an imputation strategy (other than kNN and Most Frequent) generally leads to smaller values for average odds 1[3] as

3. Recall that 1 refers to the protected variable "sex" and 2 to the remaining protected variables (age and race) for each dataset). Thus, for example Theil Index 1 refers to the "sex" protected variable, and Theil Index 2 to either the race or age protected variable, depending the dataset in question.

compared to the base case: None, i.e., the distribution of average odds values is closer to 0 when these imputation strategies are applied. For equal opportunity, using an imputation strategy other than kNN or Most Frequent also leads to better distributions for the Linear Support Vector Classifier and Logistic Regression classifiers, and also a small shift towards better values (closer to 0) for the Random Forest classifier.

Overall, considering the first part of the research question posed (whether the imputation strategy affects the fairness and/or performance of a Machine Learning model), an initial inspection of the distribution of performance and fairness outcomes visualized in Figures 1 and 2 suggest that the choice of imputation strategy significantly affects both the performance and the fairness of the resulting predictions.

4.2 Implications for the Machine Learning Model(s) Used and the Training Process

Considering the impact of different classification algorithms, we see that the Linear Support Vector Classifier generates more outliers for both performance and fairness metrics, but has a smaller interquartile range, indicating that for some runs this particular classifier can be much worse. At the same time, the Random Forest classifier has performed best with respect to Accuracy, AUC & F1 Scores for all the imputation strategies which is expected based on the general findings of Fernández-Delgado et al. (2014). The distribution of fairness metrics look alike for all the imputation strategies except for kNN and Most frequent imputation when using a Logistic Regression or Random Forest classifiers. When using a Linear Support Vector Classifier, iterative and least squares imputer have metric distributions closer to the ideal value of the fairness metrics (1 for Disparate Impact 1, Disparate Impact 2, and 0 for the rest). Furthermore, for the numeric imputation strategies the median values of sensitivity metrics are higher when using a Linear Support Vector Classifier while it is the opposite for specificity. The most frequent imputation strategy has a lower dispersion of classification metrics for all the classification algorithms.

Overall, considering the second part of the research question posed (whether the classification algorithm affects the fairness and/or performance of a Machine Learning model), an initial inspection of the distribution of performance and fairness outcomes visualized in Figures 1 and 2 suggest each of the imputation strategies work differently for the three considered classification algorithms, indicating that there is no universal imputation strategy that leads to fairer outcomes and that the choice of Machine Learning model is important.

Putting this observation into the wider context of fairness in Machine Learning research, and the many different approaches to fairness interventions as we discuss in Caton and Haas (2020), there are obvious implications concerning how in-processing methods should handle missing data. In this article, our approach has been to treat missing data as part of a pre-processing step. Yet the results when viewed from the perspective of different "vanilla" Machine Learning models is highly suggestive of either handling the imputation as part of the model training, or creating a new "family" of fairness interventions specifically for the treatment of missing data. Yet, here there are several factors that would need to be considered, and would require significant additional research. Firstly, in-processing methods tend to incorporate quantitative notions of fairness into the classifier loss function (usually, but not always as a regularisation or penalty term) handling missing data at this point would significantly increase the complexity of the classifier optimisation function; potentially at the cost of losing or otherwise affecting convexity. Secondly, as this style of approach would fundamentally be some form of hybridisation of in-processing and transformation (the latter usually

a pre- or post-processing fairness intervention) would have implications for model explainability, as the process would make the model less interpretable (Lum & Johndrow, 2016; Lepri et al., 2018), which may be at odds with current data protection legislation (Caton & Haas, 2020). Thus, ultimately, a key conclusion from this part of the evaluation is that more research and innovation is needed to unravel exactly where in the fairness pipeline to treat missing data.

4.3 Effects of Imputation Strategy on Outcomes

To understand the effects of different factors on the performance and fairness metrics, we perform a robust three-factorial analysis of variance (ANOVA) design for trimmed means with interaction effects (Wilcox, 2017; Mair & Wilcox, 2020) using the imputation strategy as the first factor, the classification algorithm as the second factor, and the percentage of missing values as a third factor.[4] We ran a separate robust three-factor ANOVA for each performance and fairness metric (see Table 3) and summarized the results for the main effects (i.e., the rate at which a factor affects performance / fairness) and interaction effects (i.e., when one factor's effect on performance / fairness is dependent on another factor). Table 3 summarizes all specific aspects for ease of presentation (main effect or interaction effect) for each of the 15 robust ANOVAs (10 for different fairness metrics, 5 for performance metrics). This allows us to shed light on which factors of the study have leverage over performance or fairness to provide more context and implications for practitioners.

Considering the effect of the imputation strategy on the performance and fairness metrics, Table 3 shows that for all considered metrics, the main effect of the imputation strategy significantly impacts the average result for each metric. I.e., looking at the standard hypotheses for a three-factor (robust) ANOVA with interaction effects, we see that not all imputation strategies lead to the same average outcome for the performance and fairness metrics. Hence, the imputation strategy significantly affects the performance and fairness outcomes, and thus the choice of imputation strategy for maintaining performance and/or fairness is important.

Similar to the imputation strategy main effect, we see that the classification algorithm (in our experiment, three different algorithms) also significantly affects the average outcome for all considered performance and fairness metrics. Here, using the results illustrated in Figures 1 and 2, we would need to decide between an approach that on average performs well, but may have extreme outliers (here, the Linear Support Vector Classifier), or which on average has more consistent results (here, the Logistic Regression), or which on average has the better results, but may have some outliers (here, the Random Forest). This is not a decision that can be made in isolation and without consideration of the "cost" of specific trade-offs, which we discuss in Section 5.

For the third main effect, the percentage of missing values, we see that only for the Equal Opportunity 1, the Averaged Odds 1, the Theil metrics, and the F1 score outcomes this main effect is significant. While the p-values for the Equal Opportunity and Averaged Odds are small (0.006 and 0.001, respectively), we would not be able to rule out a type-I error for the corresponding values for the Theil metrics (0.012) and the F1 score (0.041), especially given the large number of observations. Hence, the percentage of missing values, at least in the range tested (between 1 and 10%), in general does not significantly impact the average performance and fairness results for most metrics and protected attributes. It is important to not immediately assume that 1%, 5% or 10% missing data is

4. The data did not follow the standard assumptions for a regular ANOVA, specifically normality and homoscedasticity. Hence, selecting the robust trimmed-means ANOVA still allows us to compare the average fairness and performance scores under heteroscedasticity.

Metric	Protected Attribute	Main Effects Strat	Alg	%	Interaction Effects Strat*Alg	Strat*Alg*%
Statistical Parity	1	0.0001***	0.0001***	0.558	0.001***	0.999
	2	0.0001***	0.0001***	0.263	0.001***	0.999
Equal Opportunity	1	0.0001***	0.0001***	0.006**	0.001***	0.999
	2	0.0001***	0.0001***	0.983	0.001***	0.999
Averaged Odds	1	0.0001***	0.0001***	0.001***	0.001***	0.088
	2	0.0001***	0.0001***	0.140	0.001***	0.953
Disparate Impact	1	0.0001***	0.0001***	0.228	0.001***	0.999
	2	0.0001***	0.0001***	0.377	0.001***	0.999
Theil	1	0.0001***	0.0001***	0.012*	0.001***	0.935
	2	0.0001***	0.0001***	0.012*	0.001***	0.935
Accuracy	1 & 2	0.0001***	0.0001***	0.084	0.001***	0.999
AUC	1 & 2	0.0001***	0.0001***	0.772	0.001***	0.999
F1 Score	1 & 2	0.0001***	0.0001***	0.041*	0.001***	0.999
Sensitivity	1 & 2	0.0001***	0.0001***	0.440	0.001***	0.999
Specificity	1 & 2	0.0001***	0.0001***	0.333	0.001***	0.998

Table 3: High-level summary of robust ANOVA results capturing the main effects and interaction effects for imputation strategy (Strat), classification algorithm (Alg), percent of missing data (%) on both fairness & classification metrics. Illustrates the p-value (where the min for the library used is 0.0001). For ease of readability, we have included asterisks for the corresponding alpha-levels of: * significant at the 0.05 level; ** significant at the 0.01 level; *** significant at the 0.001 level

not impactful enough to note a difference, as this is per feature. As such, as the percentage increases, the likelihood of multiple features per data instance (row) with missing data increases quickly.

Upon the basis that two of the three main effects were significant for all the performance and fairness metrics, we also consider the interaction effects in our robust three-factor ANOVA. We see that the interaction effects between the imputation strategy and classification algorithm are also consistently significant at the 0.001 level. This indicates that different imputation strategies affect the performance and fairness outcomes differently depending on the applied classification algorithm. Or in other words, that it is difficult to disentangle which combination(s) of imputation strategies and classification algorithms are "best". There is, however, no concrete evidence that there is a three-way interaction effect. This is in line with the previous main effects, which indicated that the percentage of missing values is not significantly affecting the performance and fairness results.

4.4 "Good" Combinations of Imputation Strategies and Classification Algorithms

In an attempt to provide some guidance to practitioners on combinations of imputation strategies and classification approaches, we present a Friedman ranking of the different permutations studied. We use Friedman ranking as an ordering mechanism, as this has been well instrumented in large scale

comparative studies (e.g., Fernández-Delgado et al. (2014)) to derive ranks for Machine Learning models across multiple datasets and performance metrics. The difference in this article is that we have competing metric categories within a fairness vs. performance trade-off. We also have a hierarchical structure through sub-categories: individual vs group fairness. The challenge here is that weighting metrics (or metric categories) differently arguably expresses different fairness preferences, which results in a different rank order. Our goal here is not to explore complex preference structures, but rather explore the impact that some simple canonical preference structures may have; we leave the derivation of meaningful preference structures for future investigation and note (Haas, 2019) as a first step in this direction.

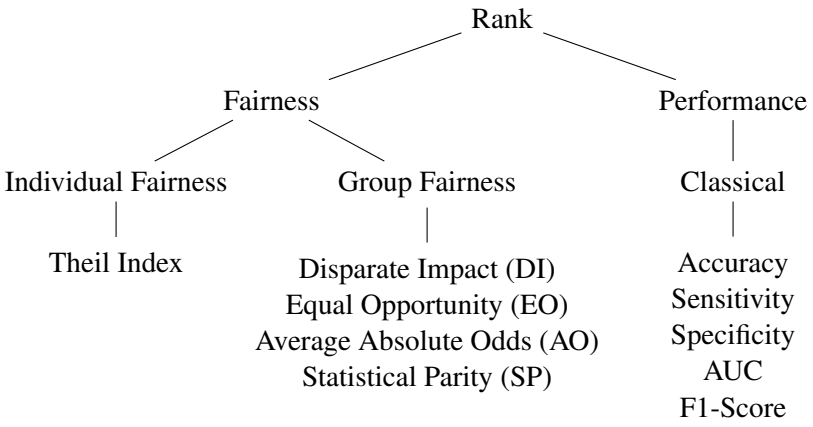

Figure 3: Contributing Metrics for the Rank Derivation of Imputation Strategies

A challenge that stems from the hierarchical nature of Figure 3 is the degree and ordering of aggregation in rank derivation, i.e., how does rank information propagate up the tree. For example, an imputation strategy could be "best" in terms of group fairness, but "worst" in terms of classical performance, and the way in which we combine this information significantly influences the rank order. Treated equally, this example would appear somewhere in the middle of the rank order. If instead, classical performance was used only for tie breaking, this example would appear near the top of the rank order. This is the fairness preference elicitation problem. It would make no sense to derive all permutations for this paper, as ultimately, how individual branches of the metric tree in Figure 3 are combined is context specific. Instead, we explore two canonical ranking structures, as simple linear combinations of the leaf nodes in Figure 3. Our goal in the formulation of these ranking mechanisms is to minimise the number of aggregation steps. We recognise, however, that there are many ways in which one could formulate these rankings to have similar intentions but different objectives. To derive rankings, we prioritise three aspects of interest: general fairness, specific fairness, and classification performance, i.e., the three main branches in Figure 3 and establish two potential trade offs: individual vs. group fairness, and fairness vs. performance. This gives five specific ranks that a permutation of imputation strategies with Machine Learning model could have: 1) general fairness (R_{fair}); 2) individual fairness ($Score_{Indiv}$); 3) group fairness ($Score_{Group}$); 4) classical performance (R_{perf}); and 5) and a linear combination of 1-4 (R_{EFP}).

To calculate a general fairness rank R_{fair}, all fairness measures are projected into $[0,1]$ (0 best, 1 worst) and averaged, with a rank order computed using all fairness metrics. This rank is then

averaged across the 3 datasets, producing a mean fairness rank. Due to the number of group fairness measures, caution is needed to ensure it does not crowd out individual fairness. Thus, we derive R_{fair} as follows:

$$\begin{aligned} R_{fair} &= 0.5 \times Score_{Group} + 0.5 \times Score_{Indiv} \\ Score_{Group} &= f(Score_{DI} + Score_{EO} + Score_{AO} + Score_{SP}) \\ Score_{Indiv} &= f(Score_{Th}) \end{aligned} \tag{1}$$

For classical performance, we apply a similar approach: again averaging the metrics within the $[0,1]$ scale and inverting them so that 0 is best and then averaging across the 3 datasets.

$$R_{perf} = f(Score_{accuracy} + Score_{sensitivity} + Score_{specificity} + Score_{AUC} + Score_{F1-Score}) \tag{2}$$

To capture the relationship between performance and general fairness, we compute an overall rank R_{EFP} as the average of the fairness rank R_{fair} and the performance rank R_{perf}:

$$R_{EFP} = 0.5 \times R_{fair} + 0.5 \times R_{perf} \tag{3}$$

Thus for R_{EFP} fairness corresponds to 50% of the rank (25% individual, and 25% group), and classical performance the other 50%. Again, we note that there is a diverse manner with which such rankings can be derived and that there is a large design space that potentially warrants investigation, but this is not our intention here. Table 4 illustrates the rank order for these ranking mechanisms and is ordered on the basis of R_{fair}. Ranks are rounded to 2 significant digits. To interpret the table, the minimum possible value is 1, this would mean that a combination is, on average, ranked first across all metrics. Similarly, the maximum value is 33, this would mean that a combination is, on average, ranked last. That no combination has an average rank of 33 indicates that there is no observable always "worst" combination. In fact, in none of the ranking procedures do we see an average rank above 20. The ranking process does not force tie-breaking, i.e., if multiple combinations have, on average, a similar rank distribution prior to aggregation, both will receive the same rank.

The most poignant takeaway from Table 4 is a clear trade off between performance (R_{perf}) and fairness (R_{fair}); the Random Forest (RF) tends to have on average better R_{perf} but worse R_{fair} and similarly the Linear Support Vector Classifier (SVC) has the opposite rank structure. Such an observation is not surprising, but key for practitioners and fairness researchers alike to consider. From Table 3, we identified that the degree of missing data is unlikely a cause of this. We should also note that even though the Linear Support Vector Classifier is on average performing well in terms of its R_{fair} score as we saw in Figure 1, it is prone to extreme outliers. We also see that "perfect" imputation, on average, outperforms the imputation strategies tested (as would be expected) but that some are not too far away in terms of their rank structure (e.g., interpolation + SVC and KNN_n + SVC, where KNN_n refers to kNN being applied on numeric variables).

Considering different types of data imputation, i.e., categorical vs. numerical, we see that on average imputing categorical data is less impactful on fairness metrics than imputing numerical data. Some caution is needed here in terms of how to interpret this. There are fewer categorical columns in the datasets selected, and this also explains the ranking of Most Frequent, which is the only imputation strategy to operate on only categorical data. kNN would also have received a slight increase in its rankings for this reason, as it operates on both numerical and categorical data; thus in Table 4 we differentiate between when it operates on numerical (KNN_n) and categorical (KNN_c)

Strat	Alg	R_{fair}	$Score_{Indiv}$	$Score_{Group}$	R_{perf}	R_{EFP}
No Missing	SVC	1.40	1.67	1.33	2.33	1.87
No Missing	RF	2.27	2.67	2.17	1.40	1.83
No Missing	LR	2.33	1.67	2.50	2.27	2.30
Most Frq.	SVC	2.53	1.67	2.75	3.13	2.83
KNN_c	RF	3.07	4.33	2.75	3.13	3.10
Most Frq.	RF	3.27	4.00	3.08	3.13	3.20
KNN_c	SVC	3.33	3.00	3.42	3.73	3.53
Most Frq.	LR	4.10	3.33	4.29	3.67	3.88
KNN_c	LR	4.70	4.67	4.71	4.20	4.45
Norm	SVC	5.87	10.00	4.83	17.07	11.47
Interpl.	SVC	5.97	9.67	5.04	16.73	11.35
KNN_n	SVC	6.73	10.67	5.75	16.27	11.50
Stoch.	SVC	7.53	11.00	6.67	16.67	12.10
Median	SVC	7.83	10.00	7.29	15.07	11.45
Mean	SVC	8.00	10.00	7.50	15.47	11.73
It. Imp.	SVC	8.63	11.67	7.88	16.80	12.72
Least Sq.	SVC	9.23	12.67	8.38	16.80	13.02
Norm	LR	11.03	5.00	12.54	13.13	12.08
Mean	LR	12.00	9.33	12.67	10.80	11.40
Interpl.	LR	12.27	6.67	13.67	11.60	11.93
Median	LR	12.80	10.33	13.42	11.00	11.90
KNN_n	LR	13.23	7.33	14.71	11.80	12.52
Stoch.	LR	13.93	9.00	15.17	13.47	13.70
It. Imp.	LR	14.57	11.67	15.29	13.67	14.12
Least Sq.	LR	15.27	12.33	16.00	13.40	14.33
Stoch.	RF	15.87	16.00	15.83	7.80	11.83
Norm	RF	16.13	15.67	16.25	8.47	12.30
It. Imp.	RF	16.13	19.67	15.25	8.80	12.47
Least Sq.	RF	16.40	20.00	15.50	9.53	12.97
KNN_n	RF	16.53	17.67	16.25	8.93	12.73
Mean	RF	17.57	18.67	17.29	9.00	13.28
Interpl.	RF	17.70	16.00	18.12	9.00	13.35
Median	RF	18.77	19.00	18.71	8.73	13.75

Table 4: Friedman Ranking for Imputation Strategies (Strat) with different classification algorithms (Alg) considering different fairness vs. performance trade-offs. We distinguish between kNN imputing categorical features (KNN_c) and numeric features (KNN_n).

data. However, as protected attributes are often categorical (race, ethnicity, etc.), the observation that categorical imputation tends to be "better" (depending on the specific fairness preference structure) is somewhat reassuring.

Considering different perspectives on fairness, i.e., individual (via the Theil Index) vs. group (via the remaining fairness metrics), we see a clear divide in the rankings dependent on the promotion of

specific notions of fairness. There are few combinations of imputation strategies and classification algorithms that can maintain high(er) ranks for both individual and group fairness. Again, this is to be expected on the basis of many scholars having discussed at length the trade-offs between individual and group fairness notions (e.g., Chouldechova (2017), Kleinberg et al. (2018)).

4.5 Summary

To summarize the results, we see that for the types of classification algorithms, imputation strategies, and percentage of missing values in the data, the classification algorithm and imputation strategy main effects significantly affect the average outcome for all performance and fairness metrics, and there is also a highly significant interaction effect between the classification algorithm and the imputation strategy. On the other hand, we can see that the percent of missing data generally has no effect on both the fairness and classification metrics, neither when considering the corresponding main effect, nor the interaction effect with the classification algorithm and imputation strategy. We also see a trade-off between performance and fairness, approaches that have "better" performance are less fair, and vice versa. We also see challenges with respect to average performance / fairness vs. variance in performance / fairness, i.e., approaches that are on average "better" can have higher variance and be more prone to outliers. However, we also note that more analysis is needed here: we experimented with only three classification models, and those models exhibit structurally different behaviours.

5. Discussion

Deciding on an appropriate imputation strategy to handle missing data within a Machine Learning pipeline that considers fairness is not a trivial undertaking. Primarily, we see that the trade-off between performance and fairness – well known in the fairness literature – is also strikingly apparent (see Table 4) when considering missing data. Whilst we only scratch the surface of possible interplays between Machine Learning models, imputation strategies and fairness interventions, it is clear that much work is needed to address the question of what is(are) the "right" approach(es) for imputing missing data in a fairness pipeline. Whilst initial work by Fernando et al. (2019, 2021) suggests that doing "something", i.e., imputing missing data, is better than doing nothing, i.e., deleting rows with missing data, our results indicate that doing "something" is not straightforward. In fact, it suggests that a key part of the notion of exploratory fairness analysis (first suggested by Veale and Binns (2017), Corbett-Davies and Goel (2018)) will be to deliberate the handling of missing data, and explore the potential effects of available options. It also suggests that researchers in the fairness community need to more explicitly consider how missing data may impact their approaches and interventions for fairness in Machine Learning.

When considering the impacts for unprivileged groups, in all cases we see that the fairness of the models studied reduces when we attempt to treat missing data via imputation. The implications of this for unprivileged groups within the data is that they will be treated (even) less fairly than when data is complete, i.e., the tendency for them to receive a negative classification outcome increases sometimes significantly. This is not to say do not impute missing data, but rather that the methods we have explored here can reduce the overall fairness when compared to "perfect" imputation, and that sometimes this difference is quite considerable. Whilst we have investigated this only in a within-group setting, i.e., one protected attribute in isolation (e.g., people of colour within a race variable), there is no reason to expect this not also be the case for between-groups, i.e., combinations

of unprivileged protected attributes (e.g., women of colour), and also for this effect (reduced fairness) to be even more apparent.

We also observed that some combinations of imputation strategies with Machine Learning models more significantly affect fairness than others. Whilst Table 4 may suggest this is pretty clear cut it would be short-sighted to not explore more fairness interventions, of which there are many! We summarised these in Caton and Haas (2020), and would suggest an expansive review of their (as well as new and emerging approaches) with regard to how they respond to both missing data and data which has been treated, e.g., via imputation. We also note that in this article, we have data which is missing completely at random (we deleted it at random from one or more features), and that in practical settings, data may be missing for more structured reasons (e.g., subgroups of the data not answering a question), and thus our findings are (hopefully) a worst case scenario. Yet, there is a strong argument to increase the "diversity" of data explored in fairness studies.

We note that a new "family" of fairness interventions in Machine Learning is likely needed. Our observations (albeit only with a few different models) are indicative that the type of Machine Learning model used has sometimes quite dramatic effects on both fairness and classical performance metrics in the presence of missing data. Our intuition would be that a natural starting point for this new "family" of fairness interventions would be in the in-processing subset of interventions. In this subset of approaches, researchers could try to balance trade-offs between imputation effects (or approaches) and other notions of fairness or classical performance within the loss or more generally model optimisation function.

As a final part of the discussion, is a note of some of the wider Machine Learning pipeline considerations. In this paper, we have assumed that the only pre-processing necessary is data imputation to treat missing values. However, a Machine Learning process needs to carefully consider and explore the data to ensure it is fit for analysis (multicollinearity, heteroscedasticity, normalisation, basic feature encoding, dimensionality reduction, centering etc.). One of the motivations for this study was an observation that many fairness metrics or interventions often make assumptions (often implicitly) concerning how "ready for analysis" or "perfect" data is (or needs to be). Similarly, there may be different versions of the data (where a key example here is the COMPAS dataset; there exist different variants of the dataset, some with more, some with less preprocessing already performed). As such, whilst we have highlighted that how missing data is handled is important for considerations of fairness, there are still many other aspects of the "normal" data (pre)processing pipeline where the impact(s) on fairness are not well understood, and both fairness researchers and Machine Learning practitioners should not lose sight of how other methodological decisions may affect fairness. Especially as some studies (e.g., Barenstein (2019)) have illustrated how even some simple considerations (here: correctly implementing a temporal cutoff) can affect fairness in pronounced ways. Our expectation is that using different "versions" of a dataset (i.e., through different forms of pre-processing) will not significantly change the main takeaways of this study, but rather change the magnitude of the effect, i.e., how much fairness or performance is affected, rather than whether they are affected by imputation strategies or not.

6. Conclusion and Future Work

Missing data is ubiquitous in real-world datasets, yet much of the previous work on fairness in Machine Learning uses well structured and (mostly) complete reference datasets for their evaluation. Hence, to determine the potential impact of imputation strategies aimed at replacing missing data

with substitute values, this study analyzes if the choice of imputation strategy, along with other characteristics of the data and the applied Machine Learning algorithm, has an impact on various performance and fairness metrics for binary classification. To simulate missing values in datasets while using the reference datasets commonly used in fairness research, we artificially, and randomly, introduce missing values. We run a structured experiment, iterating over different characteristics of the dataset (e.g., percentage of missing values, classification algorithm, and imputation strategy), to systematically analyze the impact of different factors on the performance and fairness metrics.

In total, we consider three different percentages of missing data, 9 imputation strategies, and 3 classification algorithms (Logistic Regression, Random Forest, Linear Support Vector Classifier) on 3 reference datasets (German Credit, Adult Income, and COMPAS), and tested each combination of factors to observe potential effects on the performance and fairness metrics. We find that the choice of imputation strategy and classification algorithm selected has an impact on both the fairness and classification metrics. Certain imputation strategies lead to more consistent outcomes, when averaged over a number of independent repetitions, and the classification algorithm also determines the potential range of values for the performance and fairness metrics. In particular, we see that the kNN and the most frequent imputation strategies lead to a wider distribution of performance and fairness metrics as compared to the other imputation strategies, indicating that other imputation strategies lead to a more "stable" outcome. The findings of this article will allow practitioners to factor the imputation strategy into their design of a Machine Learning fairness pipeline, in particular when missing values are encountered in real-world datasets. We hope that fairness researchers will take on board the findings of this study and include in their presentation of (new) fairness interventions an appropriately scaled study on missing data.

While the awareness that the imputation strategy can affect various performance and fairness metrics in the application of fairness interventions is an important first step, we intend to extend this line of research in several directions. First, adding additional datasets (besides the three reference datasets), algorithms, and imputation strategies is a straightforward extension that can help to further generalize the results in this article. Second, we want to also consider the relative effects, or trade-offs, that imputation strategies have on the performance and fairness metrics. For example, it would be interesting to analyse if certain imputation strategies that improve one set of metrics either improve or deteriorate other metrics at the same time (consider individual fairness metrics vs. group fairness metrics as a stark example here). Thirdly, in the general context of fairness in Machine Learning, various bias mitigation strategies have been suggested over time. Hence, we plan to augment our analysis by including a variety of pre-, in-, and post-processing strategies into our experimental setup to determine if imputation strategies also affect the performance of different bias mitigation strategies. While the majority of fairness interventions focus on binary classification problems, there is an ever increasing number of studies and approaches focusing both on regression and unsupervised Machine Learning (usually clustering) scenarios. Thus, further study on the impact(s) of missing data are needed here. We have considered only individual protected variables in this article. We made this choice on the basis that few works in the fairness in Machine Learning literature can concurrently handle multiple protected variables. However, moving forward, multinomial considerations of protected variables should also be considered. Finally, we note that there is a need to unravel where in the Machine Learning process it is most appropriate to handle missing data, and that this may change based up the type of fairness intervention used, and notion(s) of fairness being considered.

References

Barenstein, M. (2019). Propublica's compas data revisited. *arXiv preprint arXiv:1906.04711*.

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning*. fairmlbook.org.

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., & Mojsilović, A. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, *63*(4/5), 4–1.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 0049124118782533.

Biega, A. J., Gummadi, K. P., & Weikum, G. (2018). Equity of attention: Amortizing individual fairness in rankings. In *The 41st international ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 405–414.

Binns, R. (2018). Fairness in Machine Learning: Lessons from Political Philosophy. In *Conference on Fairness, Accountability and Transparency*, pp. 149–159.

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized Pre-Processing for Discrimination Prevention. In *Advances in Neural Information Processing Systems*, pp. 3992–4001.

Caton, S., & Haas, C. (2020). Fairness in Machine Learning: A Survey. *arXiv preprint arXiv:2010.04053*.

Cheema, J. R. (2014). A review of missing data handling methods in education research. *Review of Educational Research*, *84*(4), 487–508. Publisher: Sage Publications Sage CA: Los Angeles, CA.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, *5*(2), 153–163.

Commission, E., Directorate-General for Communications Networks, C., & Technology (2019). *Ethics guidelines for trustworthy AI*. Publications Office.

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017a). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 797–806.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017b). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 797–806, New York, New York, USA. ACM, ACM Press.

Dignum, V. (2021). The Myth of Complete AI-Fairness. In *International Conference on Artificial Intelligence in Medicine*, pp. 3–8. Springer.

Donders, A. R. T., Van Der Heijden, G. J., Stijnen, T., & Moons, K. G. (2006). A gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, *59*(10), 1087–1091. Publisher: Elsevier.

du Pin Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2018). Data pre-processing for discrimination prevention: Information-theoretic optimization and analysis. *IEEE Journal of Selected Topics in Signal Processing*, *12*(5), 1106–1119.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226. ACM.

Edelman, B., Luca, M., & Svirsky, D. (2017). Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment. *American Economic Journal: Applied Economics*, *9*(2), 1–22.

Farhangfar, A., Kurgan, L., & Dy, J. (2008). Impact of imputation of missing values on classification error for discrete data. *Pattern Recognition*, *41*(12), 3692–3705. Publisher: Elsevier.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems?. *The Journal of Machine Learning Research*, *15*(1), 3133–3181.

Fernando, M.-P., Cèsar, F., David, N., & José, H.-O. (2019). Fairness and missing values. *arXiv preprint arXiv:1905.12728*.

Fernando, M.-P., Cèsar, F., David, N., & José, H.-O. (2021). Missing the missing values: The ugly duckling of fairness in machine learning. *International Journal of Intelligent Systems*.

Goodman, B. W. (2016). Economic models of (algorithmic) discrimination. In *29th Conference on Neural Information Processing Systems*, Vol. 6.

Haas, C. (2019). The price of fairness-a framework to explore trade-offs in algorithmic fairness. In *40th International Conference on Information Systems, ICIS 2019*. Association for Information Systems.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pp. 3315–3323.

Hardy, S. E., Allore, H., & Studenski, S. A. (2009). Missing data: a special challenge in aging research. *Journal of the American Geriatrics Society*, *57*(4), 722–729. Publisher: Wiley Online Library.

Hu, L., & Chen, Y. (2018). A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference*, pp. 1389–1398.

Hutchinson, B., & Mitchell, M. (2019). 50 Years of Test (Un)Fairness: Lessons for Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 49–58. ACM.

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 35–50. Springer.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018). Algorithmic fairness. In *AEA Papers and Proceedings*, Vol. 108, pp. 22–27.

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology*, *31*(4), 611–627.

Lepri, B., Staiano, J., Sangokoya, D., Letouzé, E., & Oliver, N. (2017). The tyranny of data? the bright and dark sides of data-driven decision-making for social good. In *Transparent Data Mining for Big and Small Data*, pp. 3–24. Springer.

Lipton, Z., McAuley, J., & Chouldechova, A. (2018). Does mitigating ml's impact disparity require treatment disparity?. In *Advances in Neural Information Processing Systems 31*, pp. 8125–8135.

Lum, K., & Johndrow, J. (2016). A statistical framework for fair predictive algorithms. *arXiv:1610.08077*.

Mair, P., & Wilcox, R. (2020). Robust statistical methods in r using the wrs2 package. *Behavior Research Methods*, *52*(2), 464–488.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1–35.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2018). Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867*.

Myrtveit, I., Stensrud, E., & Olsson, U. H. (2001). Analyzing data sets with missing data: An empirical evaluation of imputation methods and likelihood-based methods. *IEEE Transactions on Software Engineering*, *27*(11), 999–1013.

Noriega-Campero, A., Garcia-Bulle, B., Cantu, L. F., Bakker, M. A., Tejerina, L., & Pentland, A. (2020). Algorithmic targeting of social policies: fairness, accuracy, and distributed governance. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pp. 241–251.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Pedreshi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 560–568. ACM.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. In *Advances in Neural Information Processing Systems*, pp. 5680–5689.

Romei, A., & Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, *29*(5), 582–638.

Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data.. *Psychological Methods*, *6*(4), 317–329.

Skirpan, M., & Gorelick, M. (2017). The Authority of "Fair" in Machine Learning. *arXiv:1706.09976*.

Sokolovska, A., & Kocarev, L. (2018). Integrating technical and legal concepts of privacy. *IEEE Access*, *6*, 26543–26557.

Soley-Bori, M. (2013). Dealing with missing data: Key assumptions and methods for applied analysis. Tech. rep., Boston University.

Song, Q., & Shepperd, M. (2007). Missing data imputation techniques. *International Journal of Business Intelligence and Data Mining*, *2*(3), 261–291. Publisher: Inderscience Publishers.

Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., & Zafar, M. B. (2018). A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2239–2248.

Suresh, H., & Guttag, J. V. (2019). A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*.

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, *4*(2).

Veale, M., Van Kleek, M., & Binns, R. (2018). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–14.

Wang, Y., & Singh, L. (2021). Analyzing the impact of missing values and selection bias on fairness. *International Journal of Data Science and Analytics*, 1–19.

Wayman, J. C. (2003). Multiple imputation for missing data: What is it and how can I use it. In *Annual Meeting of the American Educational Research Association, Chicago, IL*, Vol. 2, p. 16.

Wilcox, R. R. (2017). *Introduction to robust estimation and hypothesis testing*. Elsevier.

Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2017). Fairness Constraints: Mechanisms for Fair Classification. *arXiv:1507.05259 [cs, stat]*. arXiv: 1507.05259.

Zarsky, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values*, *41*(1), 118–132.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. In *International Conference on Machine Learning*, pp. 325–333.

Zliobaite, I. (2015). On the relation between accuracy and fairness in binary classification. *arXiv preprint arXiv:1505.05723*.