Received: 1 April 2020 | Revised: 22 January 2021 | Accepted: 22 February 2021

DOI: 10.1002/int.22415

RESEARCH ARTICLE

WILEY

Missing the missing values: The ugly duckling of fairness in machine learning

Martínez-Plumed Fernando[1,2] | **Ferri Cèsar**[2] | **Nieves David**[2] | **Hernández-Orallo José**[2,3]

[1]Joint Research Centre, European Commission, Seville, Spain

[2]Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València, València, Spain

[3]Leverhulme Centre for the Future of Intelligence, University of Cambridge, Cambridge, UK

Correspondence
Martínez-Plumed Fernando, Joint Research Centre, European Commission, Edificio Expo, Calle Inca Garcilaso, 3, 41092 Sevilla, Spain.
Email: Fernando.Martinez-Plumed@ec.europa.eu

Funding information
Ministerio de Economía, Industria y Competitividad, Gobierno de España (ES), Grant/Award Number: RTI2018-094403-B-C3; Generalitat Valenciana, Grant/Award Number: PROMETEO/2019/09; Future of Life Institute, Grant/Award Number: RFP2-15; European Commission, Grant/Award Number: DG JRC - HUMAINT project

Abstract

Nowadays, there is an increasing concern in machine learning about the causes underlying unfair decision making, that is, algorithmic decisions discriminating some groups over others, especially with groups that are defined over protected attributes, such as gender, race and nationality. Missing values are one frequent manifestation of all these latent causes: protected groups are more reluctant to give information that could be used against them, sensitive information for some groups can be erased by human operators, or data acquisition may simply be less complete and systematic for minority groups. However, most recent techniques, libraries and experimental results dealing with fairness in machine learning have simply ignored missing data. In this paper, we present the first comprehensive analysis of the relation between missing values and algorithmic fairness for machine learning: (1) we analyse the sources of missing data and bias, mapping the common causes, (2) we find that rows containing missing values are usually fairer than the rest, which should discourage the consideration of missing values as the uncomfortable ugly data that different techniques and libraries for handling algorithmic bias get rid of at the first occasion, (3) we study

the trade-off between performance and fairness when the rows with missing values are used (either because the technique deals with them directly or by imputation methods), and (4) we show that the sensitivity of six different machine-learning techniques to missing values is usually low, which reinforces the view that the rows with missing data contribute more to fairness through the other, nonmissing, attributes. We end the paper with a series of recommended procedures about what to do with missing data when aiming for fair decision making.

KEYWORDS
algorithmic bias, confirmation bias, data imputation, fairness, missing values, sample bias, survey bias

1 | INTRODUCTION

Because of the ubiquitous use of machine learning and artificial intelligence (AI) for decision making, there is an increasing urge in ensuring that these algorithmic decisions are fair, that is, they do not discriminate some groups over others, especially with groups that are defined over protected attributes, such as gender, race and nationality.[1–7] Despite all this growing research interest, fairness in decision making did not arise as a consequence of the use of machine-learning and other predictive models in data science and AI.[8,9] Fairness is an old and fundamental concept when dealing with data that should cover all data processing activities, from data gathering to data cleansing, through modelling and model deployment. It is not simply that data are biased and this can be amplified by algorithms, but rather that data processing can introduce more bias, from data collection procedures to model deployment.[10–17]

It is therefore no surprise that fairness strongly depends on both the quality of the data and the quality of the processing of these data.[18] One major issue for data quality is the presence of missing data, which may represent the absence of information but also some information that has been removed due to several possible reasons (inconsistency, privacy or other interventions).[19,20] Once missing data appears in the pipeline it becomes an *ugly duckling* for many subsequent processes, such as data visualisation and summarisation, feature selection and engineering, and model construction and deployment. It is quite common that missing values are removed or replaced as early as possible, so that they no longer become a nuisance for a bevy of theoretical methods and practical tools.

In this context, we ask the following questions: (1) Are missing data and fairness related? (2) Are those subsamples with missing data more or less unfair? (3) Is it the right procedure to delete or replace these values, as many theoretical models and machine-learning libraries do by default? In this paper we analyse all these questions through the first comprehensive analysis, to our knowledge, of the relation between missing values and fairness. We also give a series of recommendations and guidelines about how to proceed with missing values if we are—as we should be—concerned by unfair decisions.

Let us illustrate these questions with an example. The Adult Census Data[21] is one of the most frequently used data sets in the fairness literature, where race and sex are attributes that could be

used to define the protected groups. Adult has 48,842 records and a binary label indicating a salary of ≤$50K or >$50K. There are 14 attributes: eight categorical and six quantitative attributes. The prediction task is to determine whether a person makes over 50K a year based on their attributes. Not surprisingly, as we see in Figure 1, there is an important number of missing values, as it is usually the case for real-world data sets, and most especially those dealing with personal data. As we will see later, the *missingness distribution* in this data set is *not* missing completely at random (MCAR). This means that discarding or modifying the rows with missing values can bias the sample. But, more interestingly, these missing values appear in the `occupation`, `workclass` and `native.country` attributes, which seem to be strongly related to the protected attributes. As a result, the bias that is introduced by discarding or modifying the rows with missing values can have an important effect on fairness.

On the one hand, as missing values are so commonplace, there are many techniques for data cleansing, feature selection and model optimisation that have been designed to convert (or get rid of) missing data with the aim of improving some performance metrics. To our knowledge, however, fairness has never been considered in any of these techniques.

On the other hand, many theoretical methods for dealing with fairness—the so-called mitigation methods[22]—do not mention missing values at all. Conditional probabilities, mutual information and other distributional concepts get ugly when attributes can have a percentage of missing values. Much worse, when the theory is brought to practice and ultimately to implementation, we see that the most common libraries (AIF360 toolkit,[23] Aequitas,[24] Themis-ML[25] and Fairness-Comparison library[22]) simply remove the rows or columns containing the missing values, or assume the data sets have been preprocessed before being analysed (removing missing values), otherwise throwing an error to the user. As a result, the literature

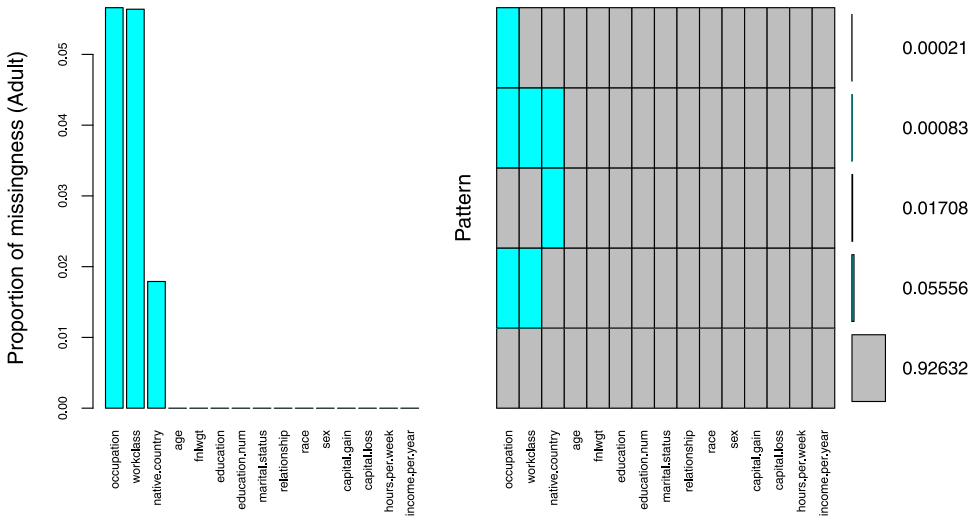

FIGURE 1 (Left) Histogram (%) of missing values per input variable in the adult data set. Missing values are concentrated in three of the categorical features: `workclass` (6%), `native.country` (2%) and `occupation` (6%). (Right) Aggregation plot of all existing combinations of missing and nonmissing values in the data set (missing values are expressed with blue cells). It is unlikely that missingness in the adult data set is MCAR because many missing values for `workclass` are also missing for `occupation` (about 6% of examples has missing values in both). MCAR, missing completely at random [Color figure can be viewed at wileyonlinelibrary.com]

using these techniques and libraries simply reports results about fairness as if the data sets did not contain the rows with the missing data. In other words, the data sets are simply mutilated.

Apart from the adult data set, in this paper we will also explore and analyse five other data sets: *Recidivism*, *Titanic*, *Autism*, *Credit Approval* and *Juvenile Offenders*. Given the intrinsic nature of the previous data sets they may have potential fairness issues and, as usual, contain missing values. The six data sets have important differences in terms of domain, protected attributes and ratio of missing values, but it is still manageable to understand them in some detail. In what follows, we will use all of them to give a comprehensive perspective to the general analysis, the conceptual questions and the results of different algorithms. These six data sets are the only ones that, to the best of our knowledge, are commonly scrutinised in the literature of fairness and *also* contain missing values. We also propose this collection of data sets (with two protected attributes each) as a benchmark to evaluate future studies on fairness *with* missing data, building on the insights of the foundational analysis performed in this paper.

The *Recidivism* data set contains variables used by the COMPAS algorithm[26] to assess potential adult recidivism risk in the United States, where the class indicates whether the inmate commits a crime in less than 2 years after being released from prison or not. Data for over 10,000 criminal defendants in Florida were gathered by ProPublica,[1] showing that black defendants were often predicted to be at a higher risk of recidivism than they actually were, compared with their white counterparts. In this data set, there are also three attributes with missing values: `days_b_screening_arrest`, `c_days_from_compas` and `c_charge_desc`. We also study the data for 891 of the real Titanic passengers, a case where bias is more explicit than that in the other cases. The class represents whether the passenger survived or not, where the conditional probability that a person survives given their `sex` and `passengerclass` is higher for females from higher classes. There are also three attributes with missing values: `age`, `fare` and `embarked`. Taking into account the strong semantic meaning of these attributes and possible association with protected groups by gender or race, it is very likely that the missing values of these attributes can have an effect on fairness, as we will study in the following sections. The *Autism* data refer to Autistic Spectrum Disorder (ASD) screening information for 704 adults. This data set includes attributes regarding the test takers' demographics (e.g., age, gender, ethnicity, etc.) as well as 10 screening test (binary) questions. The class represents early autism diagnosis, where the conditional probability that a person is diagnosed with this condition given their `Sex` and `Ethnicity` is higher for white-European females. In this data set, there are two attributes with missing values: `Age` and `Relation`. The *Credit Approval* data set collects a company's decisions to approve or deny credit card applications based on the applicant's information (e.g., prior default, years employed, credit score, income level, loan balances or number of individual's credit reports, etc.). It contains information about 690 applicants. Machine-learning models attempting to generalise this sort of data usually find patterns that may be controversial or even illegal. For instance, the loan repay model unveils that the applicant's ethnicity plays a significant role in the prediction of repayment because the training data set happened to have better repayment for white applicants compared with nonwhite applicants. In this data set, there are five attributes with missing values: `Age`, `Married`, `BankCustomer`, `EducationalLevel` and `ZipCode`. Finally, the *Juvenile Offenders* data set contains information about 4753 juvenile offenders who were incarcerated in the juvenile justice system of Catalonia (Spain). Their recidivism status (after their release) is assessed with SAVRY1. It has been found that, in general, machine-learning models trained with these data could end up discriminating by gender against men,

foreigners or people from specific national groups.[7] Compared with the numerous fairness-related studies published on COMPAS, the literature in juvenile criminal justice is limited,[28] and only a study analyses disparities between protected groups when using ML algorithms.[7] In this data set, there are six attributes with missing values: `Edat_fet_agrupat` (age), `Provincia` (province), `Comarca` (county), `Edat_fet` (age when offence) and `Fet` (offence). As we will see at the end of Section 2, the distribution of these missing values is not MCAR for any of these six data sets.

The main novel contribution in this paper is to clarify the relationship between missing data and fairness, and the best way forward in real situations. In this regard, we will combine a theoretical analysis of the causes of missing data and unfairness, with an experimental analysis of different kinds of classifiers using the six data sets above. We convert this general goal into more specific technical questions, such as whether the examples with missing values are fairer than the rest in terms of the Statistical Parity Difference (SPD), a common fairness metric,[23] or whether there is a relationship between protected attributes and the attributes with missing values. We will also formally characterise the space that derives from the trade-off between a metric of fairness and accuracy, and how different subsets of the data with or without missing values are represented in that space, also including the results replacing missing data with imputed values.

The rest of the paper is organised as follows. Section 2 overviews the reasons why missing values appear, what kinds of missingness there are and how missing values are handled (e.g., imputation). Next, in Section 3 we analyse the causes of fairness, along with some metrics, mitigation techniques and libraries. In Section 4, we put these two areas together and see why missingness and fairness are so closely entangled. We also analyse—for the running data sets and two scenarios each—whether the examples with missing values are fairer than the rest in terms of a common fairness metric, SPD. SPD sets the space of trade-offs with performance metrics (accuracy), representing a bounding octagon, which we derive theoretically. Once this is understood conceptually, in Section 5 we analyse a predictive model that can deal with missing values directly and see whether the bias is amplified or reduced. In Section 6 we study what happens when imputation methods (IMs) are introduced so that many other machine-learning techniques can be used. We analyse how sensitive different machine-learning algorithms are to changes in the attributes with missing values. We empirically evaluate how the results with imputed attributes compare with the models learnt by removing the missing values and with some reference classifiers (majority class and perfect classifier). In Section 7 we further analyse these results to make a series of recommendations about how to proceed when dealing with missing values if fairness is to be traded off against performance. Finally we close the paper with some final comments and takeaway messages.

2 | MISSING DATA

Missing data are a major issue in research and practice regarding real-world data sets. For instance, in the educational and psychological research domains, Peugh and Enders[29] found that, on average, 9.7% of the data was missing (with a maximum of 67%). In the same area, Rombach et al.[30] estimated that this percentage ranged from 1% to over 70%, with a median percentage of 25%. Another clear example of how pervasive missing data is can be found in the percentage of data sets (over 45%) that have missing values in the UCI repository,[31] one of the most popular sources of data sets for machine-learning and data science researchers.

Being such a frequent phenomenon, it is not surprising that 'missingness' has different *causes*. In the context of this paper, we need to analyse these causes if we want to properly understand the effect of missing data on fairness.

Three main patterns can be discerned in missing data[32]:

- *Partial completion (attrition)*. A partial completion (breakoff) is produced for a single record or sequence when, after collecting a few values of a record, at a certain point in time or place within a questionnaire or a data collection process, the remaining attributes or measurements are missing. That means that attributes that are at the end of a questionnaire, as well as users more prone to fatigue or problems, are more likely to be missing. Note that this case may have effect on full rows as well, as if only a few questions (attributes) are recorded, then the whole example (row) could be removed. This type of answering pattern usually occurs in longitudinal studies where a measurement is repeated after a certain period of time, or in telephone interviews and web surveys. This kind of missing data creates a dependency between attributes and their order, which may be used for imputation and other methods for treating missing values.
- *Missing by design*. This refers to the situation in which specific questions or attributes will not be posed to or captured for specific individuals. There are two main reasons for items to be missing by design. (1: *contingency attributes*) Certain questions may be 'non-applicable' (NA) to all individuals. In this case, the missingness mechanism is known and can be incorporated in the analysis. (2: *attribute sampling*) Specific design is used to administer different subsets of questions to different individuals (i.e., random assignment of questions to different groups of respondents). Note that in this case, all questions are applicable to all respondents, but for reasons of efficiency not all questions are posed to all respondents. Here, we also know the missingness mechanisms, but due to the random assignment of questions, this is the easiest case to treat statistically.
- *Item nonresponse*. No information is provided for some respondents on some variables. Some items are more likely to generate a nonresponse than others (e.g., private items, such as income). In general, surveys, questionnaires or interviews used to collect data suffer missing data in three main subcategories. (1: *not provided*) The information is simply not given for a question (e.g., an answer is not known, the value cannot be measured, a question is overlooked by accident, etc.); (2: *useless*) The information provided is unprofitable or useless (e.g., a given answer is impossible, unsuitable, unreadable, illegible, etc.); and/or (3: *lost*) usable information is lost due to a processing problem (e.g., error in data entry or data processing, equipment fails, data corruption, etc.). The former two mechanisms originate in the data collection phase, and the latter is the result of errors in the data processing phase.

While the three categories above originate from questionnaires involving people, they are general enough to include some other causes behind the presence of missing data in real applications (also known as incomplete data), including different kinds of failures: network, power, the own device or its sensors.[33] For instance, if a sensor capturing data stops working because it runs out of battery, we have a partial completion. If it is only installed in some specific models, devices or locations, we have missing by design. Finally, if it is affected by the weather, we have an item nonresponse. Internet of Things (IoT) applications analysing big data represent another example of a data source where data loss is very common due to, for instance, unreliable wireless link or hardware failure in the nodes (i.e., partial completion). Missing data may also have several effects in social media environments where actors are

linked together via multiple interaction contexts.[34] In this latter case, there is also a number of missing data situations, such as the noninclusion of actors, affiliations or some other incomplete registration data (i.e., item nonresponse), and censoring by vertex degree, as there is a practical limit on the number of neighbours of a vertex that can be explored (i.e., missing by design).

On many occasions, for simplicity or because the traceability to the data acquisition is lost, we can only characterise some statistical kinds of missing values. In general, a distinction is made between three types of missingness mechanisms[35,36]: (1) MCAR, where missing values are independent of both unobserved and observed parameters of interest and occur entirely at random (e.g., accidentally omitting an answer on a questionnaire). In this case, missing data are independent and simple statistical treatments may be used. (2) *Missing at random* (MAR), where missing values depend on observed data, but not on unobserved data (e.g., in a political opinion poll many people may refuse to answer based on demographics, then missingness depends on an observed variable [demographics], but not on the answer to the question itself). In this case, if the variable related to the missingness is available, the missingness can be handled adequately. MAR is believed to be more general and more realistic than MCAR (missing data IMs generally start from the MAR assumption). Finally, we have (3) *missing not at random* (MNAR), where missing values depend on unobserved data (e.g., a certain question on a questionnaire tends to be skipped deliberately by those participants with weaker opinions). MNAR is the most complex nonignorable case where simple solutions no longer suffice, and an explicit model for the missingness must be included in the analysis.

Note that all three mechanisms that generate missing data may be present at the same time for different attributes.[37] While it is possible to test the MCAR assumption (*t*-test), distinguishing between MAR and MNAR cannot be tested due to the fact that the answer lies within the absent data.[38,39] The literature also describes different techniques to handle missing data depending upon the missingness mechanism.[40] However, it is quite common that many practitioners apply a particular technique without analysing the missingness mechanism. Among the most common techniques, we can name the following:

- Row or listwise deletion (LD) implies that whole cases (rows) with missing data are discarded from the analysis. When dealing with MCAR data and the sample is sufficiently large, this technique has been shown to produce adequate parameter estimates. However, when the MCAR assumption is not met, LD will result in bias.[41,42]
- Column deletion (CD) simply removes the column. This is an extreme option as it totally removes the information of an attribute. Also, it can generate bias as well, as the values of other columns may be affected by the missing values in the removed column.
- Labelled category (LC). An attribute (usually quantitative) can be binned or discretised, adding a special category for missing values. However, the special label representing the missing value lacks any ordinal or quantitative value. Sometimes, the existence of a missing value can be flagged with a new Boolean attribute and the original attribute is removed.
- *IMs*. The missing value is replaced by a fictitious value. There are many methods to do this, such as replacing it by the mean (or the median) when the attribute is quantitative, and the mode when the attribute is qualitative, or estimating the value from the other attributes, even using predictive models.

LD is very common. Even if the MCAR assumption is not disproven, LD is claimed to be suboptimal because of the reduction in sample size.[29,40] Also, this technique for MCAR

does not use the available data efficiently.[43] Therefore, methodologists have strongly advised against the use of LD,[35,37] judging it to be "among the worst methods for practical applications"[44, p. 598]. IM, on the contrary, is increasingly more frequent as a preprocessing for many machine-learning techniques, as they cannot deal with missing values. However, many libraries do not explicitly state what IM they are using, and this may vary significantly depending on the machine-learning technique, library or programming language that is used. For example, random forest[45] handles missing values by imputation with average/mode or proximity-based measures whenever the implementation is based on *Classification and Regression Trees* (CARTs),[46] as it was originally proposed by its authors. However, in implementations where C4.5 decision tree learning[47] is used instead, the missing values are not replaced. The impurity score mechanisms take them into account (e.g., information gain is simply weighted by the proportion of missing values on a particular attribute[48]). This implicit (usually by-default) processing happens more in research than that in real practice, where the domain expert can evaluate several mechanisms and choose the best one, even replacing the by-default treatment done by a training algorithm. In research papers, many data sets are frequently used in experimental evaluations, and explicitly choosing the best choice for dealing with missing values would require a full understanding and analysis of each data set and technique. Using the same missingness mechanism, not to say the same IM, for a range of data set and machine-learning techniques is clearly suboptimal. There is no good-for-all missingness-dealing mechanism.

We can analyse the missingness mechanisms in our six running data sets (Adult, Recidivism, Titanic, Autism Screening, Credit Approval and Juvenile Offenders). We use Little's MCAR global test,[49] a multivariate extension of the *t*-test approach that simultaneously evaluates mean differences on every variable in the data set. Using the R package *BaylorEdPsych* 2, we reject the null hypothesis of MCAR with $p < 0.001$ for the six data sets. MCAR is discarded for the six of them, and hence LD is inappropriate.

Given that we do not have the complete traceability of the data sets, we can only hypothesise the causes for the missing values. For instance, in the Adult Census Data it may be due to item nonresponse, because of the distribution of missing values seen in Figure 1.

3 | FAIRNESS

Fairness in decision making has been recently brought to the front lines of research and the headlines of the media, but in principle, making fair decisions should be equal to making good decisions. The issue should apply both to human decisions and algorithmic decisions, but the progressive digitisation of the information used for decision making in almost every domain has facilitated the detection and assessment of systematic discrimination against some particular groups. The use of data processing pipelines is then a blessing and not a curse for fairness, as it makes it possible to detect (and treat) discrimination in a wider range of situations than when humans make decisions solely based on their intuition. As we will see, metrics for fairness have led to computerised techniques to improve fairness (mitigation techniques). However, it is still very important to know the *causes* of discrimination, to avoid oversimplifying the problem. Also, in this paper we want to map these causes with those of missing values seen in Section 2.

What are the causes that introduce unfairness in the decision process? We can identify common distortions originally arising in the data, but also those in the algorithms or the

humans involved in decision making that can perpetuate or even amplify some unfair behaviours. From the literature,[50–58] we can classify them into six main groups:

- *Sample or selection bias*. It occurs when the (sample of) data are not representative of the target population about which conclusions are to be drawn. This happens due to the sample being collected in such a way that some members of the intended population are less likely to be included than others. A classic example of a biased sample happens in politics, such as the famous 1936 opinion polling for the US presidential elections carried out by the American Literary Digest magazine, which overrepresented rich individuals and predicted the wrong outcome.[59] If some groups are known to be underrepresented and the degree of underrepresentation can be quantified, then sample weights can correct the bias.
- *Measurement bias (systematic errors)*. Systematic value distortion happens when the device used to observe or measure favours a particular result (e.g., a scale which is not properly calibrated might consistently understate weight producing unreliable results). This kind of bias is different from random or nonsystematic measurement errors whose effects average out over a set of measurements. On the contrary, systematic errors cannot be avoided simply by collecting more data, but by having multiple measuring devices (or observers of instruments) and specialists to compare the output of these devices.
- *Self-reporting bias (survey bias)*. This has to do with nonresponse, incomplete and inconsistent responses to surveys, questionnaires or interviews used to collect data. The main reason is the existence of questions concerning private or sensitive topics (e.g., drug use, sex, race, income, violence, etc.). Therefore, self-reporting data can be affected by two types of external bias: (1) social desirability or approval (e.g., when determining drug usage among a sample of individuals, the actual average value is usually underestimated); and (2) recall error (e.g., participants can erroneously provide responses to a self-report of dietary intake depending on her ability to recall past events).
- *Confirmation bias (observer bias)*. This bias places emphasis on one hypothesis because it involves favouring information that does not contradict the researcher's desire to find a statistically significant result (or its previously existing beliefs). This is a type of cognitive bias in which a decision is made according to the subject's preconceptions, beliefs or preferences, but can also emerge owing to overconfidence (with contradictory results/evidence being overlooked). Peter O. Gray[60] provides an example of how confirmation bias may affect a doctor's diagnosis: *"When the doctor has jumped to a particular hypothesis as to what disease a patient has may then ask questions and look for evidence that tends to confirm that diagnosis while overlooking evidence that would tend to disconfirm it."*
- *Prejudice bias (human bias)*. A different situation is when the training data that we have at hand already includes (human) biases containing implicit racial, gender or ideological prejudices. Unlike the previous categories, which mostly affect the predictive attributes (model inputs), this kind of bias is concerned with variables that are used as dependent variables (model outputs). Therefore, systems designed to reduce prediction error will naturally replicate any bias already present in the labelled data. We find examples of this in the reoffence risk-assessment tool COMPAS deployed in federal US criminal justice systems to inform bail and parole decisions that demonstrated biased against black people.[1] Another one is the Amazon's AI hiring and recruitment system that showed a clear bias against women,[61] having been trained from CVs submitted to the company over a 10-year period.
- *Algorithmic bias*. In this case the algorithm creates or amplifies the bias over the training data. For instance, different populations in the data may have different feature distributions

(also having different relationships to the class label). As a result, if we train a group-blind classifier to minimise overall error, as it cannot usually fit both populations optimally, it will fit the majority population. It may be plausible that the best classifier is one that always picks the majority class, or ignores the values for some minority group attributes leading, thus, to (potentially) higher distribution of errors in the minority population.

While the range of causes in general can be enumerated, the precise notion of fairness and what causes it in a particular case is much more cumbersome. It is nonetheless very helpful to become more precise with definitions or metrics of fairness, based on the notion of protected attribute and parity. Let us start with the definition of a decision problem, which is tantamount to a supervised task in machine learning. We will focus on classification problems, since it is the most common case in the fairness literature and the metrics are simpler.

Let us define a set of attributes X, where the subset $S \subset X$ denotes the protected attributes. A protected attribute is assumed to be categorical (e.g., race, gender and religion) and can partition a population into groups that should have parity in terms of the potential benefit obtained. For each protected attribute S_i, we have a set of values V_i (e.g., {`male`, `female`}). Groups can be created by setting the value of one or more protected attributes. We typically use the term 'privileged' group to highlight a group that has a systematic advantage in the context or domain of application (e.g., white males). Usually, fairness metrics are based on determining whether decisions are different between groups or just between the privileged groups and the rest. Now consider a label or class attribute Y, which can take values in C (e.g., {`guilty`, `non guilty`}). For the analysis of fairness we usually consider one of the classes as the "favourable" (or positive) outcome, denoted by c^+. Note that a favourable label value corresponds to an outcome that provides an advantage to the individual represented by the example. Similarly, c^- denotes the unfavourable class.[3] For instance, in this case, `nonguilty` is the favourable outcome. An unlabelled example x is a tuple choosing values in V_i for each $X_i \in X$, possibly including the extra value $\odot$, representing a missing value. A labelled example or instance $\langle x, y \rangle$ is formed from an unlabelled example with the class value y taken from C (e.g., $\langle${race = `black`, gender = `female`, income = $\odot$}, `guilty`$\rangle$). A decision problem is just defined as mapping x to $\hat{y}$, such that $\hat{y}$ is correct with respect to some ground truth y. Let us denote with M a mechanism or model (human or algorithmic) that tries to solve the decision problem. Data sets, samples and populations are defined over sets or multisets of examples, and examples are drawn or sampled from them, using the notation $\sim D$. Given a population or data set D we denote by $D_{X_i=a}$ the selection of instances such that $X_i = a$, with $a \in V_i$. This also applies to labelled data sets, so, $D_{y=c^+}$ is just the number of positive examples in D, usually shortened as $\mathrm{Pos}(D)$ (we do similarly for the negative examples $\mathrm{Neg}(D)$). By comparing the actual positives $y = c^+$ and negatives $y = c^-$ with predicted positives $\hat{y} = c^+$ and negatives $\hat{y} = c^-$ of a model, we can define TP, TN, FP and FN, as usual. Finally, given pairs $\langle x, y \rangle$ from a data set or distribution D, the probability of favourable outcome $p(y = c^+)$ when $x \sim D$ is simply denoted by $p^+(D)$.

Some fairness metrics choose an indicator that does not depend on the confusion matrix, but just on the overall probabilities. For instance, the percentage for the favourable class or the unfavourable class can be applied either to data sets or the predictions of a model. Other metrics are defined in terms of comparing predicted and true labels, for example, comparing the true positive rate (TPR) and the false positive rate (FPR), so they can only be applied to models when compared with the ground truth (or a test data set). In the end, several combinations of cells in the confusion matrix lead to dozens of fairness metrics.

A very common metric is the SPD, which—for the favourable class—can be defined for an attribute X_i with privileged value a as follows:

$$\mathrm{SPD}_i^+(D) = p^+(D_{X_i=a}) - p^+(D_{X_i\neq a}).$$

For instance, if the attribute is `Race` and the values are `caucasian`, `black` and `asian`, if we consider `caucasian` as the privileged group, SPD would be the probability of favourable outcomes for Caucasians minus the probability of favourable outcomes for non-Caucasians. A value of 0 implies both groups have equal benefit, a value less than 0 implies higher benefit for the unprivileged group, and a value greater than 0 implies higher benefit for the privileged group. Note that the sign of SPD will change if for a binary data set we swap the favourable and unfavourable class:

$$\mathrm{SPD}_i^-(D) = p^-(D_{X_i=a}) - p^-(D_{X_i\neq a}) = 1 - p^+(D_{X_i=a}) - (1 - p^+(D_{X_i\neq a})) = -\mathrm{SPD}_i^+(D).$$

The same happens if we swap the privileged groups. This is an interesting property, as the choice of the favourable class and especially the privileged group is sometimes arbitrary. For instance, in the adult data set, if a model is used to grant a subsidy, the favourable class is earning <$50K. The important thing is that a value closer to zero is fairer.

Other popular metrics that are applicable to both data sets and models are the Disparate Impact (DI),[62] which calculates a ratio instead of the difference as SPD does (it should be noted that the use of a ratio introduces some issues with extreme values in the metric). Other metrics that are only applicable to models are the Equal Opportunity Difference (EOD),[14,63] which is the difference in TPR between the groups, or the Average Odds Difference (OddsDif),[14] which also considers the FPR (see Reference [64] for a summary of fairness metrics definitions). There is usually some confusion about which fairness metric to look at, a question that is not very different from the choice of the 'right' performance metric in classification.[65,66] However, as we can see in their definitions, all of them are closely connected since they try to quantify differences between the unprivileged and privileged groups. For illustrative purposes, we have evaluated the strength of the relationships between all these metrics in a comprehensive simulated scenario. The absolute Spearman correlations between SPD, DI and OddsDif are all higher than 0.84, and EOD is always higher than 0.59.

In conjunction with the performance metric, we have to select the performance metric that indicates how successful a trained model has been when scoring or predicting examples. Depending on the metric, performance can be measured using a (1) threshold and a qualitative understanding of error (e.g., accuracy, *F*-score, Kappa statistic, etc.); (2) probabilistic understanding of error (e.g., mean absolute error, LogLoss, Brier score, etc.); and on (3) how well the model ranks the examples (e.g., AUC). Our experience and empirical evidence, at least for the data sets we have selected for this study, show that we reach similar conclusions independently of the metrics of fairness and performance we choose for the analysis. As a result, in what follows, and for the sake of exposition and simplicity, we will use SPD as a representative fairness metric and accuracy as a metric of performance.

The introduction of formal metrics has fuelled the development of new techniques for discrimination mitigation. There are three main families: those who are applied to the data before learning the model, those that modify the learning algorithm, and those that modify or reframe the predictions (these are known as preprocessing, in-processing and postprocessing, respectively, in Reference [23]). Typically, maximising one fairness metric has a significant effect on the performance metrics, such as prediction error, so it is quite common to find trade-offs between fairness metrics and performance metrics.[67] A possible way of doing this is

through a Pareto optimisation, where different techniques can be compared to see whether they improve the Pareto front. We will use this approach in the rest of the paper.

Finally, both metrics and mitigation techniques are usually integrated into libraries. As of today, these are the most representative ones, in our opinion:

- *AIF360 toolkit*[23] is an open-source library to help detect and remove bias in machine-learning models. The AI Fairness 360 Python package includes a comprehensive set of metrics for data sets and models to test for biases, explanations for these metrics, and algorithms to mitigate bias in data sets and models.
- *Aequitas*[24] is an open-source bias audit toolkit for data scientists, machine-learning researchers and policymakers to audit machine-learning models for discrimination and bias, and to make informed and equitable decisions when developing and deploying predictive risk-assessment tools.
- *Themis-ML*[25] is a Python library that implements some fairness metrics (Mean difference and Normalised mean difference) and fairness-aware methods, such as Relabelling (preprocessing), Additive Counterfactually Fair Estimator (in-processing) and Reject Option Classification (postprocessing), providing a handy interface to use them. The library can be extended with new metrics, mitigation algorithms and test-bed data sets.
- *Fairness-Comparison library*[22] presented as a benchmark for the comparison of the different bias mitigation algorithms, this Python package makes available to the user the set of metrics and fairness-aware methods used for the study. It also allows its extension by adding new algorithms and data sets.

Some of these libraries come (or are illustrated) with data sets that are known to have or lead to fairness issues. In our case, we chose those data sets from the literature that has missing values originally. Titanic is uncommon in the literature of fairness, but its issues are also very relevant (although somewhat in the opposite direction as usual). This collection of data sets is proposed as a general benchmark to examine the relation between fairness and missingness. The details of these data sets were given in Section 1.

4 | MAPPING MISSINGNESS AND UNFAIRNESS

The new question we ask in this paper is the effect of missing values for fairness. As we will see, this is a complex relation for which we need to map the causes of missing values to the causes of unfair treatment, as illustrated in Figure 2. Looking at the left-hand side of the figure, we see that missing values might be the consequence of innumerable factors, from basic errors while processing and acquiring data to intentional action from human agents. When fairness is taken into consideration, one must realise that the missing data might not be evenly distributed between different groups, which in turn might lead to unwanted effects on the fairness of the data and models created, depending on how the missing data are handled. For example, on the side of errors, an important factor is that different groups might have different semantic constructs to answer the same query, leading to different interpretations and omissions,[68] for example, with more missing values being generated by a sensitive group than the other. On the other extreme, people might intentionally omit information as a natural coping mechanism when there is a belief that a truthful and complete answer might lead to a discriminatory and unfair decision.[69]

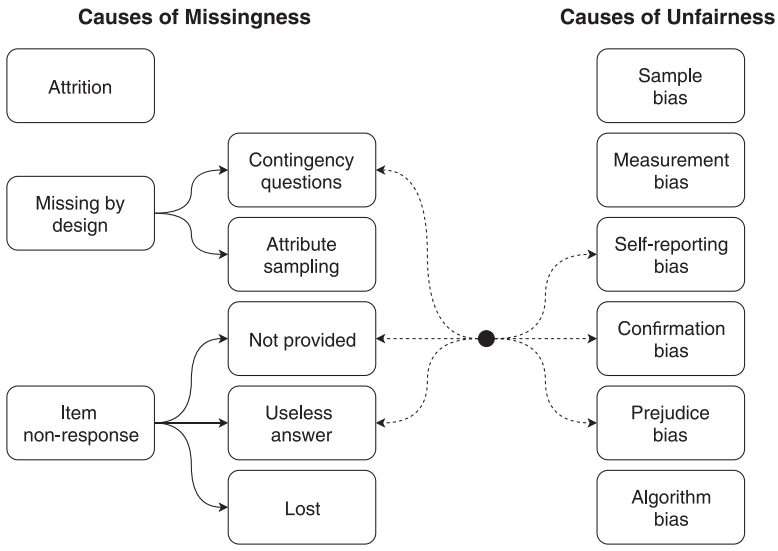

FIGURE 2 Mapping between causes of missingness and unfairness. Although many combinations are possible, dashed lines show those causes of missingness and unfairness that are most strongly related by having a common origin

If we go one by one from the causes of missingness to unfairness or vice versa, we see that many combinations are related. While *attrition*, *attribute sampling* (if random) and *lost information* may be less associated with those causes of unfairness, some others, such as *contingency questions*, *not provided* and *useless answers* may be strongly related to fairness. We can recognise common underlying origins for these three missingness situations and three causes of fairness: *self-reporting (survey) bias*, *confirmation (observer) bias* and *prejudice (human) bias* (see Figure 2). In general, we find common origins of missingness and fairness: the emergence of minorities or underrepresented groups that are reluctant to provide sensitive information, and discriminatory or unfavourable treatment to those individuals in the decision-making process on grounds, such as gender, disability or sexual orientation, or influenced by cultural norms. What we do not know is whether some of the conscious or unconscious actions done by the actors in the process may have a compensatory result. For instance, are women who do not declare their number of children treated in a more or less fair way than those who declare (or lie) on their number of children? In the end, filling a questionnaire or simply any other kind of behaviour when a person knows that she is being observed creates a bias. People would tend to conceal their real information or behaviour, to be classified in their desired group. In other words, some types of missing values can be used in an adversarial way by the person being modelled, trying to be classified into the favourable outcome. All this happens with university admissions, credit scoring, job applications, dating apps and so forth.

Bias, either intentional or unintentional can as well arise in some other use cases due to missingness, such as some of the following. For instance, discarding data due to a subpopulation having missing values more commonly (useless answers or not provided) may result in underrepresentation. Decision makers may also favour a privileged group or reinforce stereotypes influenced by cultural norms or one's own beliefs by, for instance, using underrepresentative data to train machine-learning models (e.g., gender and racial bias found in AI recognition technology[70,71]). They can also provide different subsets of questions to different

groups, thus hindering objectivity and leading to a selective and possibly misleading use of data to support decisions that have already been made. Furthermore, many real-world data sets in the health or criminal justice domain having missing values can also greatly influence explanations when machine-learning models remove or perform incorrect imputations leading to counterfactuals.[72] This is the case in health records where some tests are missing for most patients because of costs, risks or not applicable for them, but end up being imputed with unforeseen consequences when used for the rest of the analysis. Also, note that a biased population (e.g., hospitalised patients) does not provide the real distribution of the values to be imputed either.

We may even get more certainty about the relation between missingness and unfairness by analysing some real data. Let us look at our six running data sets, Adult, Recidivism, Titanic, Juvenile Offenders, Credit Approval and Autism Screening, each of them with two protected attributes. This allows us to perform 12 different evaluations.4 First, we do a simple correlation analysis, as shown in Figure 3. We observe that the protected group attributes are usually not very correlated with the class, which means that the bias must appear through other proxy attributes. Titanic is the only data set where the protected groups are really predictive about the class, which is what motivated us to include this data set. About the correlations of the occurrence of missing values, we found co-occurrences for Adult between `occupation` and `workclass` as one aggregates the other, Recidivism between `c_days_from_compass` and `c_charge_desc`, Juvenile Offenders between `Edat_fet_agrupat` and `Edat`, also because aggregation, and between `Provincia` and `Comarca`, similarly due to aggregation, and Credit Approval between all variables with missing values (except `age`). The influence and relation of the variables with missing values to the other variables are more diverse. What we see is that having a missing value or not shows small correlations with the protected attributes and the class (top right part of the correlation matrix), although this is affected by the proportion of missing values per attribute being relatively small. The purpose of this correlation matrix is actually to confirm that strong correlations are not found, and the relations, if they exist, are usually more subtle.

Now let us have a look at fairness. For each of these 12 cases (6 data sets $\times$ 2 privileged groups each), Table 1 shows the fairness metric (SPD) for different subsets of each data set and privileged group. Specifically, we focus on the following subsets of data: (a) the whole (original) data set ("all rows"), (b) the subset of examples that contain, at least, one missing value in any of their attributes ("with $\odot$") and (c) the subset of examples that do not contain any missing values ("w/o $\odot$"). If we look at the value of SPD for all rows for Adult, we see that this is positive. This means higher benefit for the privileged groups, whites and males. Nevertheless, it is important to note that the favourable class for Adult is earning more than \$50K. If a model is used to determine who is entitled to a subsidy, then having the label >\$50K would not be the favourable outcome. This suggests that we should not take the positive or negative sign as a good or bad bias, as this is application-dependent, but paying attention to the absolute magnitudes; the closer to 0 the better. A similar result is seen for Recidivism, where the positive values for SPD indicate that the privileged groups (Caucasian and female) are associated with no recidivism more often than the other groups. Then, for Titanic, the positive (and high) values for SPD indicate that the privileged groups (first-class passengers and females) fared very favourably (their survival rate was much higher than the complementary groups). Here, the bias is much stronger, as this followed an explicit protocol favouring females in ship evacuation at the time, and many other conditions favouring the first-class passengers very clearly. For the Autism data set, white-European females are more commonly diagnosed with Autism

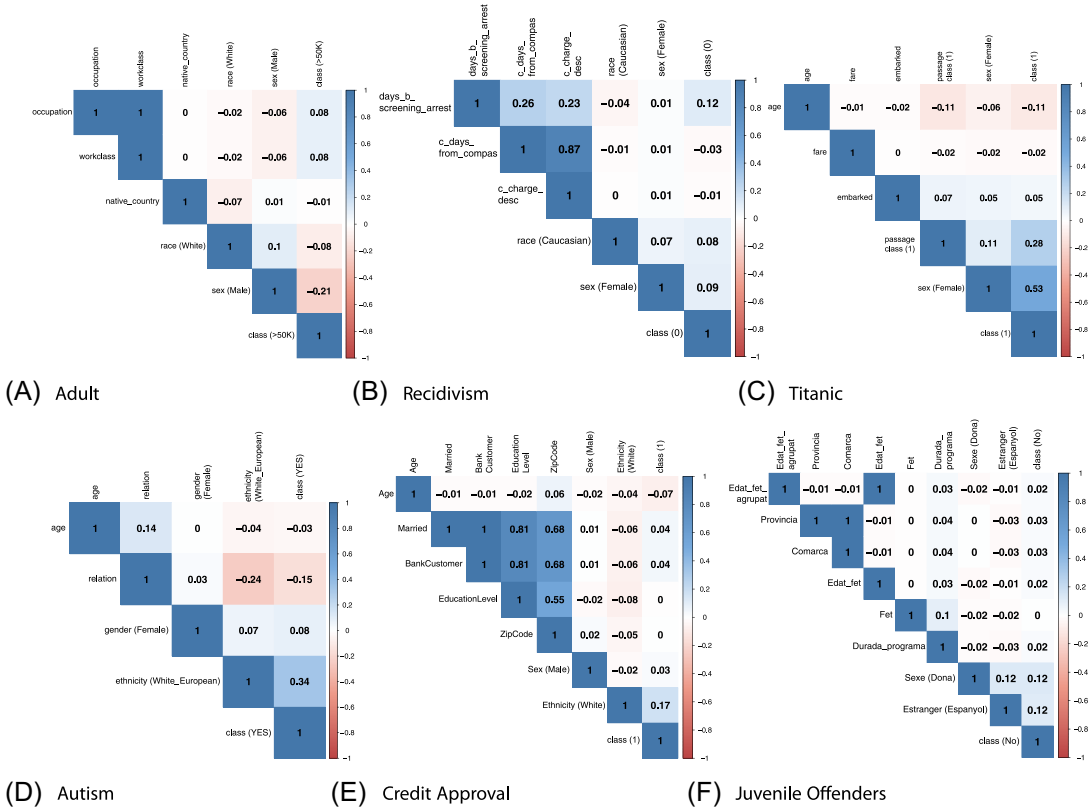

FIGURE 3 Correlation matrices for the Adult (A), Recidivism (B), Titanic (C), Autism (D), Credit Approval (E) and Juvenile Offenders (F) data sets. For each data set, we analyse Pearson's correlations between a discretised version of the attributes that may contain missing values (1 if it is missing and 0 otherwise), the protected attributes (1 if it is privileged and 0 otherwise) and the positive (favourable) class. For Juvenile Offenders, `Edat_fet` and `Edat_fet_agrupat` mean the age at offence time, nonaggregated and aggregated, respectively, `Provincia` and `Comarca` represent the province and county, respectively, `Fet` represents the offence, `Durada_programa` means the programme duration, `Sexe (Dona)` means gender (female), and `Estranger (Espanyol)` means nationality (Spanish) [Color figure can be viewed at wileyonlinelibrary.com]

compared with their counterparts, the bias also being much stronger for the ethnicity compared with the gender. For Credit Approval data, white males are the privileged groups granted with a higher proportion of credit approvals. Finally, for the Juvenile Offender data set, the positive values indicate that there is a bias against men and foreigners, they being associated with a major risk of reoffending.

The really striking result appears when we look at the metrics for the subsets of instances with missing values ("SDP (rows with ⊙)"). Ten out of the 12 cases analysed show that these subsets are *fairer* (their values closer to zero) than the subset of instances containing only clean rows (without missing values). And in 11 out of 12 cases the bias of the subset with missing values moves in the opposite direction from the bias shown in the whole data set (in Credit Approval with `Ethnicity` the compensation goes too far). The difference can be seen more clearly with the means in the last row. How do we interpret this finding? It is difficult to explain without delving into the particular characteristics of each data set. For instance, for Adult, this

TABLE 1 Data set description and fairness metrics (SPD) for different subsets of data

Data set	# Rows	# Rows with ⊙	# Cols. with ⊙	Protected attribute	Privileged values	Unprivlgd. values	c^+	Majority	SPD (all rows)	SPD (rows with ⊙)	SPD (rows w/o ⊙)
Adult	48,842	3620 (7.4%)	3	Race	White	≠ White	>$50K	≤$50K (76%)	0.1014	**0.0361**	0.1040
				Sex	Male	Female	>$50K	≤$50K (76%)	0.1945	**0.1117**	0.1989
Recidivism	7214	314 (4.4%)	3	Race	Caucasian	≠ Caucasian	0 (no crime)	0 (no crime) (55%)	0.0864	**0.0716**	0.0920
				Sex	Female	Male	0 (no crime)	0 (no crime) (55%)	0.1161	**0.0243**	0.1186
Titanic	1309	26 (20.3%)	3	Class	1	2–3	1 (survived)	0 (died) (62%)	0.3149	**0.2722**	0.3115
				Sex	Female	Male	1 (survived)	0 (died) (62%)	0.5365	**0.4727**	0.5458
Autism	704	95 (13.5%)	2	Ethnicity	White-European	Other	Yes (positive)	No (negative) (73.2%)	0.3101	**0.2062**	0.2967
				Gender	Female	Male	Yes (positive)	No (negative) (73.2%)	0.0713	**0.0572**	0.0782
Credit approval	690	27 (3.91%)	5	Ethnicity	White	No white	1 (approved)	0 (denied) (55.6%)	**0.1852**	−0.3478	0.1942
				Sex	Male	Female	1 (approved)	0 (denied) (55.6%)	0.0312	0.1786	**0.0245**
Juvenile offenders	4753	157 (3.3%)	6	Estranger	Spanish	≠ Spanish	No (no crime)	No (69.2%)	0.1134	**−0.0034**	0.1193
				Sexe	Female	Male	No (no crime)	No (69.2%)	0.1458	**0.1454**	0.1496
Mean	–	–	–	–	–	–	–	–	*0.18*	*0.10*	*0.19*

Note: We indicate the number of rows and columns with missing values (⊙), the protected attribute, its privileged and unprivileged values, the favourable class c^+ and the majority class (with the proportion for all rows). The last three columns show the SPD values for different subsets of data and privileged group: the complete (original) data set ("SPD (all rows)"), the subset of examples that contain, at least, one missing value in any of their attributes ("SPD (rows with ⊙)") and the subset of examples that are clean from missing values ("SPD (rows w/o ⊙)"). Bold figures represent the fairest subset for each combination of data and privileged group.

Abbreviation: SPD, Statistical Parity Difference.

means that for those individuals with missing values in occupation, `workclass` or `native.country`, there is less difference in the probability of a favourable outcome (>50K) between privileged and unprivileged groups than for the other rows. Note that this observation is about the data, we are not yet talking about the bias that a machine-learning model can generate. Furthermore, there seems to be no general relationship between the degree of fairness and the ratio of missing values (at least for this set of data sets). In this regard, we have analysed different ratios of missing values per data set (undersampling) and multiple repetitions obtaining almost constant results for the fairness metric (see Appendix B for further details).

Algorithmic bias will be affected by the metric that is used for performance. If the metric is a proper scoring rule, the optimal value is attained with the perfect model, that is, a model that is 100% correct. But a perfect classifier will necessarily have exactly the same fairness metrics results as the test data set. This leads us to the following observation: the SPD of the test data set determines the space of possible classifiers, that is, a region that is bounded by the best trade-off between the performance metric and SPD. We can precisely characterise this region. For instance, if we choose accuracy as performance metric, the space of possible classifiers in terms of the balance of SPD and accuracy is bounded by an octagon. The proof and a graphical representation of this octagon can be found in Appendix A. In what follows, and especially in Section 6, the octagons for each particular problem (also in Appendix A) will be used to see how far the trade-offs that are achieved between fairness and accuracy are from the optimal result or from some trivial cases, such as the majority classifier.

The value of the SPD for each data set (which assumes a perfect model) determines the octagon. If we go back to Table 1, we can see that the SPD-accuracy space of the complete case given by SPD (all rows) can reach better values than the space configured by the data set without missing values in nine out of the 12 cases. This is even more extreme for the subset only including the missing values. A perfect classifier for these data sets could have almost perfect SPD for Adult with `Race`, Recidivism with `Sex` and Juvenile Offender with `Estranger`. On the contrary, the perfect classifier for the subset of examples without missing values would be unfair for all cases except for Credit approval with `Sex`. As a result we see two problems with choosing this mutilated data set, only considering rows without missing values. First, the perfect 'target' that this data set determines is wrong, because any model has to be evaluated with all the examples, not only a sample that is not chosen at random. Second, it will also lead to an unfair model. In the following sections we will explore the location of classifiers in this space and see how far we can approach the boundary of the possible space.

Independently of the explanation of the remarkable finding that the subset with missing values is usually fairer than the rest—for 83.3% of the cases in Table 1 with 11 out of 12 cases opposing the dominant bias—the main question is: if rows with missing values are usually fairer, why is everybody getting rid of them when learning predictive models? As we saw, some of the libraries seen in Section 3 apply either LD (e.g., AIF360 and Fairness-Comparison) or CD (Themis-ML), or they assume that the data set is clean of missing values (Aequitas). These libraries do this because many machine-learning algorithms (and all fairness mitigation methods) cannot handle missing values. Apparently, LD seems the easiest way to get rid of them. However, do we have better alternatives? As we saw, IMs could replace the missing values and keep these rows, which seem to have less biased information, as we have seen in the previous 12 cases. Before exploring the results with LD and a common IM, we need to explore how these missing values affect machine-learning models and the fairness metrics of the trained models, in comparison with the original fairness metrics. In other words, do rows with missing values contribute to bias amplification more or less than the other rows? That is what we explore in Section 5.

5 | REGAINING THE MISSING VALUES FOR FAIRNESS

Since missing values are ugly and uncomfortable, they are eliminated in one way or another before learning can take place. Actually, most machine-learning methods (or at least most of their off-the-shelf implementations) cannot handle missing values directly. There are a few exceptions, and analysing results with a method that does consider the missing values can shed light on the effect of missing values on fairness when learning predictive models. One such an exception is decision trees. During training, missing values are ignored when building the split at each node of the tree. During the application of the model, if a condition cannot be resolved because of missing values, the example can go through some of the children randomly or can go through all of them and aggregate the probabilities of each class. Another important reason why we choose decision trees for this first analysis is that they can become understandable (if of moderate size) and ultimately inspectable, which allows us to see what happens with the missing values and where unfairness is created. Also, the importance of each attribute can be derived easily from the tree.

In particular, we are using the classical CARTs[46] in the implementation provided by the *rpart* package.[73] This implementation treats missing values using *surrogate splitting*, which allows the use of the values of other input variables to perform a secondary split for observations with a missing value for the variable of the best (primary) split (see Section 5 in Reference [74] for further information). Because the six data sets we are using are not very imbalanced (at most 76% for Adult), and because accuracy is the most common metric in many studies of fairness, we will stick to this performance metric.

We separated 25% of the data for test from the original data set, disregarding the existence of missing values for the split. Consequently, this test data set has a mixture of rows with and without missing values approximately equal to the whole data set. Then, for training the decision tree, we used four different training sets. The "all rows" case used all the rows not used for test. The "with $\odot$" case used the subset of these that have missing values. The "without $\odot$" case used the subset of the "all rows" training set whose rows do not have missing values. Finally, for comparison, we made a small sample of the latter of the same size of the training set "with $\odot$." We used 100 repetitions of the training/test split, where the fairness metrics (and the accuracies) are calculated with the test set labelled with the decision tree, and then averaged for the 100 repetitions.

Table 2 shows the results of CART decision trees for the test set and the 12 cases we are considering. We first compare the fairness results of the model over the data with all rows against the original fairness results of the data set as we saw in Table 1. We see that for Adult and Autism the bias is reduced (represented by ∇), but it is worse for Recidivism, Titanic and Juvenile Offenders, for which bias is amplified (represented by Δ). For the Credit Approval data sets, the bias is augmented or reduced depending on the protected attribute. When we look at other subsets, again comparing them with the data fairness in Table 1, we see a similar picture for the subset without missing values, but almost the opposite situation with the subset with missing values. Again, the rows with missing values are having a very different behaviour. Still, if we analyse which row selection is best to get the least biased model, we see that learning from the rows with missing values is better for fairness, and they now reach the best SPD in 11 out of 12 cases. Finally, to really appreciate how much of the loss in accuracy was due to the characteristics of the rows rather than the number of examples, we have the "sample w/o $\odot$," the two rightmost columns in the table. From the results, we see that sample size seems to be the key factor. But not only accuracy is degraded, SPD becomes better than for the full sample, something that is easier to understand in the view of the octagons and the trend towards the perfect model, which is biased.

TABLE 2 Fairness metric (SPD) and accuracy (Acc) averaged for 100 repetitions using CART as predictive model

Data set	Protected attribute	Acc (all rows)	SPD (all rows)	Acc (with ⊙)	SPD (with ⊙)	Acc (w/o ⊙)	SPD (w/o ⊙)	Acc (sample w/o ⊙)	SPD (sample w/o ⊙)
Adult	Race	0.8504 ±0.003-2	▽0.0876 ± 0.0096	0.8264 ± 0.0075	△**0.0768 ± 0.0179***	0.8502 ± 0.0031	▽0.0881 ± 0.0102	0.8284 ± 0.006-3	0.0888± 0.0163
	Sex	0.8504 ± 0.0032	▽0.1868 ± 0.0072	0.8264 ± 0.0075	△**0.1643 ± 0.0220***	0.8502 ± 0.0031	▽0.1885 ± 0.0072	0.8284 ± 0.006-3	0.1959 ± 0.0182
Recidivism	Race	0.6227 ± 0.0089	△0.0982 ± 0.0289	0.5549 ± 0.0165	▽**0.0347 ± 0.0315***	0.6214 ± 0.0114	△0.1013 ± 0.0262	0.5793 ± 0.0193	0.0509 ± 0.0404
	Sex	0.6227 ± 0.0089	△0.1384 ± 0.0310	0.5549 ± 0.0165	▽**−0.0093 ± 0.0419***	0.6214 ± 0.0114	△0.1351 ± 0.0328	0.5793 ± 0.0193	0.0383 ± 0.0403
Titanic	Class	0.7819 ± 0.0210	△0.3507 ± 0.0701	0.7113 ± 0.0296	△**0.2875 ± 0.1213***	0.7724 ± 0.0227	△0.3641 ± 0.0766	0.7451 ± 0.028-2	0.2886 ± 0.1556
	Sex	0.7819 ± 0.0210	△0.6692 ± 0.0572	0.7113 ± 0.0296	▽**0.4418 ± 0.1308***	0.7724 ± 0.0227	△0.6471 ± 0.0670	0.7451 ± 0.028-2	0.6827 ± 0.1519
Autism	Ethnicity	0.8716 ± 0.0237	▽0.2906 ± 0.0681	0.7027 ± 0.0507	▽**0.1415 ± 0.2345***	0.8670 ± 0.0248	▽0.2785 ± 0.0722	0.8365 ± 0.0449	0.2214 ± 0.0915
	Gender	0.8716 ± 0.0237	▽0.0480 ± 0.0547	0.7027 ± 0.0507	▽**0.0024 ± 0.0358***	0.8670 ± 0.0248	▽0.0559 ± 0.0554	0.8365 ± 0.044-9	0.0362 ± 0.0621
Credit Approval	Ethnicity	0.8301 ± 0.0308	△0.1867 ± 0.0803	0.5766 ±0.05460	▽**0.0429 ± 0.1032***	0.8318 ± 0.0269	▽0.1919 ± 0.0777	0.6732 ± 0.148-5	0.1203 ± 0.1208
	Sex	0.8301 ± 0.0308	▽0.0245 ± 0.0716	0.5766 ±0.05460	▽0.0608 ± 0.0859	0.8318 ± 0.0269	▽**0.0187 ± 0.0971**	0.6732 ± 0.1485	0.0227 ± 0.0700

(Continues)

TABLE 2 (Continued)

Data set	Protected attribute	Acc (all rows)	SPD (all rows)	Acc (with ⊙)	SPD (with ⊙)	Acc (w/o ⊙)	SPD (w/o ⊙)	Acc (sample w/o ⊙)	SPD (sample w/o ⊙)
Juvenile Offenders	Estranger	0.6811 ± 0.0119	△0.1545 ± 0.0355	0.6211 ± 0.0431	△**0.0059 ± 0.0120***	0.6777 ± 0.0126	△0.1576 ± 0.0382	0.6268 ± 0.0376	0.1524 ± 0.1580
	Sex	0.6811 ± 0.0119	△0.1882 ± 0.0348	0.6211 ± 0.0431	▽**0.0392** ±	0.6777 ± 0.012-6	▽0.1834 ± 0.-0371	0.6268 ± 0.0376	0.0427 ± 0.0877
Mean	–	*0.77*	*0.20*	*0.67*	*0.11*	*0.77*	*0.20*	*0.71*	*0.16*

Note: We show results for different subsets of data: all rows, only the rows with missing values (⊙), only the rows without the missing values and a sample of the latter of the same size. The symbols Δ and ∇ represent whether the bias has been amplified or reduced, respectively, in comparison to the corresponding columns in Table 1. Bold figures represent the fairest result (closest to 0). The star symbol denotes statistical significance in a multiple pairwise-comparison between the means of the columns SPD (with ⊙), SPD (w/o ⊙) and SPD (sample w/o ⊙).

Abbreviations: CART, Classification and Regression Tree; SPD, Statistical Parity Difference.

Of course, this analysis disregards the performance of the model. We know that we can easily obtain a perfectly fair model by using trivial models, such as the one that always predicts the majority class (with a baseline accuracy). Figures 4 and 5 show a bidimensional representation with the performance metric (accuracy) on the y-axis and the fairness metric (SPD) on the x-axis, where we also plot this majority class model. By looking at the trade-off between accuracy and SPD, the picture gets more complex but still more interesting. While the model only learning from the rows with missing values shows poor accuracy (because the sample is small), it gets an intermediate fairness result usually midway between the majority model and the model with all data. This is again in accordance with the straight line between the majority model and the best model in the octagons. The result including all rows almost always has the highest accuracy, but it is only fairer than the subset without missing values for seven of the 12 cases, but these results are always very close. The big difference happens with the small samples. There are cases, such as Autism, where the model with only rows with missing values gets accuracies that are worse than majority, while in other cases this happens for all samples (Juvenile Offenders).[5]

To understand the effect of the missing values better, we analyse the effect of those attributes that have missing values. Table 3 uses the same configuration as Table 2, except that the columns with at least one missing value were removed for training (CD). The results are

FIGURE 4 Visualisation of the results from Table 2 also including the majority class model. The x-axis shows accuracy and the y-axis shows SPD. The dashed grey line shows 0 SPD (no bias and perfect fairness), and the grey area goes between −0.1 and 0.1, a reasonably fair zone that helps see the magnitudes better. SPD, Statistical Parity Difference [Color figure can be viewed at wileyonlinelibrary.com]

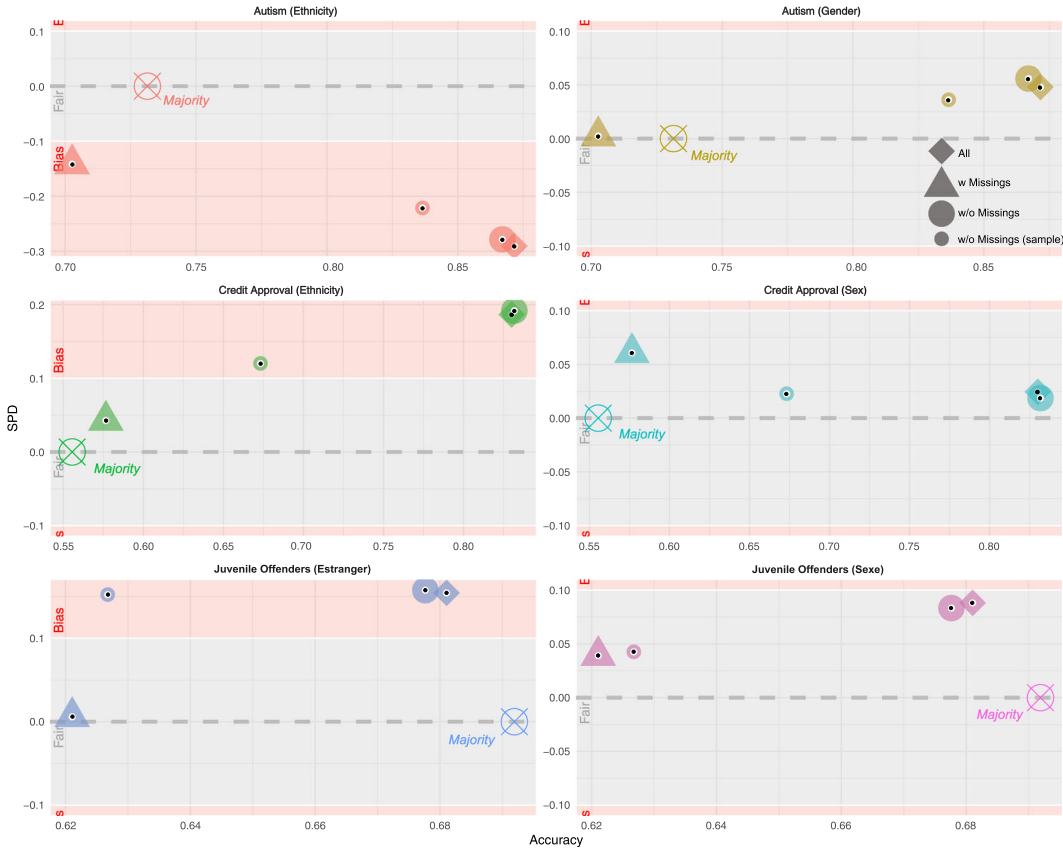

FIGURE 5 Visualisation of the results from Table 2 also including the majority class model. The x-axis shows accuracy and the y-axis shows SPD. The dashed grey line shows 0 SPD (no bias and perfect fairness), and the grey area goes between −0.1 and 0.1, a reasonably fair zone that helps see the magnitudes better. SPD, Statistical Parity Difference [Color figure can be viewed at wileyonlinelibrary.com]

different, especially for Titanic, where bias is importantly reduced for the passage `Class` and amplified for `Sex`, for all subsets and in comparison with the data and the model with all columns. Still, SPD for the subset of rows with missing values is the best, even more consistently than that in the case of training with all columns, now for the 12 cases. The surprise in this case comes when we look at accuracy, which is not strongly affected by removing the attributes (and in many cases the results are better).

This suggests that including these rows, but treating the columns in a better way, could be beneficial. As we want to explore other machine-learning methods, and most do not deal with missing values—but we must keep those rows—Section 6 turns our analysis to imputation.

6 | TREATING MISSING VALUES FOR FAIRNESS: DELETE OR IMPUTE?

Even if most libraries simply delete the rows with missing values, some IMs are so simple that it is difficult to understand why this option is not given in these packages (or included in the literature of fairness research). Fortunately, we can apply a preprocessing stage to every data set

TABLE 3 Results with the same configuration as Table 2 but using the Column Deletion (CD) approach: columns with at least one missing value were removed for training (and hence not used by the tree during test either)

Data set	Protected attribute	Acc (all rows)	SPD (all rows)	Acc (with ⊙)	SPD (with ⊙)	Acc (w/o ⊙)	SPD (w/o ⊙)	Acc (sample w/o ⊙)	SPD (sample w/o ⊙)
Adult	Race	0.8467 ± 0.0029	∇0.0861 ± 0.0105	0.8282 ± 0.0056	$\triangle$**0.0818 ± 0.0204***	0.8463 ± 0.0030	∇0.0861 ± 0.0105	0.8260 ± 0.0055	0.0857 ± 0.0195
	Sex	0.8467 ± 0.0029	∇0.1854 ± 0.0072	0.8282 ± 0.0056	$\triangle$**0.1605 ± 0.0203***	0.8463 ± 0.0030	∇0.1864 ± 0.0067	0.8260 ± 0.0055	0.1943 ± 0.0182
Recidivism	Race	0.6394 ± 0.0104	$\triangle$0.1250 ± 0.0299	0.5333 ± 0.0218	∇**0.0048 ± 0.0552***	0.6374 ± 0.0103	$\triangle$0.1332 ± 0.0310	0.5921 ± 0.0222	0.0563 ± 0.0514
	Sex	0.6394 ± 0.0104	$\triangle$0.1393 ± 0.0374	0.5333 ± 0.0218	∇**−0.0105 ± 0.0614***	0.6374 ± 0.0103	$\triangle$0.1448 ± 0.0358	0.5921 ± 0.0222	0.0493 ± 0.0467
Titanic	Class	0.7862 ± 0.0179	∇0.1666 ± 0.0634	0.7478 ± 0.0266	∇**0.1471 ± 0.0740***	0.7827 ± 0.0180	∇0.1758 ± 0.0646	0.7665 ± 0.0268	0.2173 ± 0.1379
	Sex	0.7862 ± 0.0179	$\triangle$0.9000 ± 0.0432	0.7478 ± 0.0266	$\triangle$**0.7143 ± 0.1077***	0.7827 ± 0.0180	$\triangle$0.8991 ± 0.0496	0.7665 ± 0.0268	0.8769 ± 0.1284
Autism	Ethnicity	0.8720 ± 0.0234	∇0.2914 ± 0.0680	0.7027 ± 0.0507	∇**0.1415 ± 0.2345***	0.8674 ± 0.0248	∇0.2790 ± 0.0725	0.8349 ± 0.0493	0.2417 ± 0.1035
	Gender	0.8720 ± 0.0234	∇0.0480 ± 0.0559	0.7027 ± 0.0507	∇**0.0024 ± 0.0258***	0.8674 ± 0.0248	∇0.0564 ± 0.0563	0.8349 ± 0.0493	0.0398 ± 0.0676
Credit Approval	Ethnicity	0.8307 ± 0.0244	∇0.1847 ± 0.0838	0.5822 ± 0.0919	∇**0.0093 ± 0.1136***	0.8370 ± 0.0247	∇0.1927 ± 0.0725	0.7970 ± 0.0928	0.1836 ± 0.0925
	Sex	0.8307 ± 0.0244	$\triangle$0.0459 ± 0.0773	0.5822 ± 0.0919	∇**−0.0094 ± 0.0702***	0.8370 ± 0.0247	$\triangle$0.0437 ± 0.0692	0.7970 ± 0.0928	0.0259 ± 0.0755
Juvenile Offenders	Estranger	0.7049 ± 0.0128	$\triangle$0.1905 ± 0.0343	0.6668 ± 0.0234	$\triangle$**0.0568 ± 0.0718***	0.7027 ± 0.0129	$\triangle$0.1922 ± 0.0376	0.6510 ± 0.0391	0.1821 ± 0.1690

(Continues)

TABLE 3 (Continued)

Data set	Protected attribute	Acc (all rows)	SPD (all rows)	Acc (with ⊙)	SPD (with ⊙)	Acc (w/o ⊙)	SPD (w/o ⊙)	Acc (sample w/o ⊙)	SPD (sample w/o ⊙)
	Sexe	0.7049 ± 0.0128	$\triangledown$0.1113 ± 0.0351	0.6668 ± 0.0234	$\triangledown$**0.0395 ± 0.0589***	0.7027 ± 0.0129	$\triangledown$0.1163 ± 0.0366	0.6510 ± 0.0391	0.0564 ± 0.0732
Mean	–	*0.78*	*0.21*	*0.68*	*0.11*	*0.78*	*0.21*	*0.74*	*0.18*

Note: Removed columns: Adult (`workclass`, `occupation` and `native_country`), Recidivsm (`days_b_screening_arrest`, `c_days_from_compas` and `c_charge_desc`), Titanic (`age`, `fare` and `embarked`), Juvenile Offenders (`edat_fet_agrupat`, `provincia`, `comarca`, `edat_fet`, `fet` and `durada_programa`), Credit Approval (`age`, `married`, `bankcustomer`, `educationlevel` and `zipcode`) and Autism (`age` and `relation`).

with missing values where we can use any IM that we may have available externally. Nevertheless, IMs are not agnostic, and they can introduce bias as well, which can have an effect on fairness. As we saw in Section 2, there are many reasons for missing values and some IMs (e.g., those that impute using a predictive model) can amplify certain patterns, especially if the attributes with missing values are related to the class or the protected groups. In this first analysis we therefore prefer to use simple IMs, namely, the one that replaces the missing value by the mean if it is a quantitative attribute and the mode if it is a qualitative attribute.

Once the whole data set has gone through this IM, we can now extend the number of machine-learning methods that we can apply. We are using six machine-learning techniques from the *Caret* package[75]: Logistic regression (LR), Naive Bayes (NB), Neural Network (NN), Random Forest (RF), the RPart decision tree (DT) and a support vector machine (SV) using a linear kernel. The choice of these six methods was made to have a representative selection of machine-learning methods. They also represent very different ways of treating attributes,[76–78] so that we could better understand the impact of missing values (and to an extent of protected attributes) on the classifications.

As seen in Section 5, DTs use attributes very explicitly and it may be the case that an attribute is used (once or more) in a branch leading to the classification of an instance, and that can happen for all, some or no instances. LR, on the other hand, may derive a small coefficient for an attribute, and this sets its influence for all examples. A high coefficient (when all attributes are normalised) is assumed to have a high impact, but higher values for other attributes with lower coefficients may also affect a decision. For both DTs and LR, the relevance of one attribute can mask the relevance of another highly correlated attribute. This does not happen for NB, which assumes independence for the attributes. For NB, it is easy to understand the particular contribution for each decision, as this is just a product of the attribute and class estimates. The question is more convoluted for NNs (with at least a hidden layer), taking into account the nonlinearity. Apart from particular attribute importance metrics for NNs,[79] other especially designed techniques can unveil the relevance of an attribute for each decision.[80]

Instead of painstakingly analysing each model (rules, coefficients, weights, etc.) separately for each scenario, we will try to further understand the potentially different behaviour (and, thus, performance) of each of the six previous techniques when dealing with missing values. To do this in an effective way, we introduce a metric of *sensitivity* that could be applied to any machine-learning method (but comparable to some other technique-specific measures, such as the proportion of examples affected by a condition in a DT or the attribute importance in NNs). We want to anticipate how a model will behave during deployment for different contexts with missing values. In this way, sensitivity could be used as an extra variable to determine whether to include the rows with missing values.

Our sensitivity analysis has been performed as follows. Given a data set D split into train and test, and a trained model using the former, for each test example we modify the original value of a specific attribute (containing missing values), varying $X_i = a$ systematically, with value a following the original distribution of X_i, denoted as $a \sim X_i$. Next, we calculate the number of predicted positive classes for each example, denoted by $\text{Pos}_{X_i=a,a\sim X_i}$, and similarly for the negative classes. Therefore, *for each test example* and attribute X_i, the sensitivity is defined as

$$\text{sensitivity}_i = \frac{\min\left\{\text{Pos}_{X_i=a,a\sim X_i}, \text{Neg}_{X_i=a,a\sim X_i}\right\}}{\text{Pos}_{X_i=a,a\sim X_i} + \text{Neg}_{X_i=a,a\sim X_i}}.$$

This metric can be interpreted as follows: if we generate a minisample for one example by varying an attribute, this metric represents the percentage of examples in this minisample that

are predicted into the minority class by the model. Sensitivity is between 0.5 (maximum variance in the predictions) and 0 (the variations in the attribute do not affect the predictions).

Specifically, we calculate this metric as follows. From the original data sets D, we separated 25% of the data for testing. For each test example and attribute containing missing values, we vary the original value 100 times following the original categorical/numerical distribution. Then we aggregate (average) the sensitivity for all the test examples. These sensitivity results can be seen in Table 4. In general, the values are low, but before looking at them in more detail, we have to clarify that this effect depends, in the first place, on the variability of the attribute. Table 4 also shows the number of different values for each (categorical) attribute. For instance,

TABLE 4 Sensitivity analysis for the attributes possibly containing missing values in the Adult, Recidivism, Titanic, Credit Approval, Autism and Juvenile Offenders data sets for six different machine-learning models

			Technique					
Data set	**Attribute**	**# Values**	**LR**	**NB**	**NN**	**RF**	**DT**	**SV**
Adult	workclass	8	1.078	0.668	2.274	1.218	1.687	0.016
	occupation	14	4.593	2.216	6.479	6.012	8.330	0.115
	native.country	41	0.854	0.173	1.422	0.859	0.488	0.001
Recidivism	days_b_screening_arrest	Numeric	0.542	0.852	4.831	3.567	3.100	2.365
	c_days_from_compas	Numeric	0.057	0.621	2.989	5.032	3.391	0.069
	c_charge_desc	438	34.760	3.236	0	0.111	21.710	12.160
Titanic	age	Numeric	5.681	6.033	4.006	1.899	9.945	0
	fare	Numeric	0.119	7.012	5.938	3.064	4.027	0
	embarked	3	1.419	2.813	2.152	0.844	0	0
Autism	Age	Numeric	0	0.428	0	0.645	0	0
	Relation	5	0.765	0.725	0.651	0.08	0.571	0
Credit Approval	Age	Numeric	0	1.397	0	3.198	2.836	0
	Married	3	12.327	1.508	4.093	1.573	0	0.029
	BankCustomer	3	0.538	1.508	2.134	1.666	0	1.064
	EducationLevel	14	0.760	3.614	14.409	3.847	7.128	0.602
	ZipCode	170	15.233	3.005	0	0	8.538	2.578
Juvenile Offenders	provincia	4	10.406	1.910	8.309	1.172	0.239	6.342
	comarca	41	8.660	4.074	21.403	8.687	15.778	4.963
	edat_fet	Numeric	0	6.739	0	6.326	1.554	0
	fet	70	23.151	5.165	0	0.053	18.261	5.619
	durada_programa	Numeric	0	3.212	0	6.522	7.696	0

Note: Values represent the average sensitivity for the test data set multiplied by 100 (0 no effect and 50 maximum effect).

Abbreviations: DT, decision tree; LR, logistic regression; NB, naive Bayes; NN, neural Network; RF, random Forest; SV, support vector.

we see that attribute `c_charge_desc` has over 400 different values, and their corresponding sensitivity values can be very high. Now we can better understand the (magnitudes in the) values in the table. SV is the algorithm with the lowest sensitivity values overall. This can be explained by the use of the high-dimensional kernels combining many attributes, so a single attribute change does not move the example to the other side of the boundary. NB and RF also have low sensitivity for two of the three data sets. This robustness to one single attribute is already reported in the literature[45,81,82] and in line with previous studies about the effect of attribute noise.[83] Some other models have a high variability depending on the attribute (LR and DT), and this can be explained by the correlations between these attributes, as mentioned above (also shown in Figure 3). Finally, NN has high values overall, which is related to the well-known effect of outliers in some attributes, most specially if the networks are not too deep.[84]

This analysis leads to the conclusion that, in general, the low sensitivity values that are seen (they are lower than 10 in 115 out of 126) reinforce the view that the rows with missing data contribute more to fairness (and should be included in the analysis) than having a possible effect in the robustness of the machine-learning models.

Once we can determine whether performance may be affected by a wrong imputation of missing values, in Figures 6 and 7 we show the results of the models trained and evaluated on the imputed data sets for the six machine-learning models, the majority class model (Majority) and a perfect model (Perfect). We have introduced this perfect model as a reference to enrich our analysis, and direct comparison in the octagons. Note that this is tantamount to what we did for Table 1, but for the test only (which explains why the values are roughly the same as those in the table). These points are not achievable in practice (only theoretically), but help us understand the bias that is already in the data, and how much this is amplified. We can compare these plots with the octagons (the space of possible classifiers, following Proposition A1) for each of the 12 combinations that can be shown as separate plots6 in Appendix A. This also reminds us that the perfect model is not unbiased.

For each machine-learning model in Figures 6 and 7 we show the results when trained by removing the rows with missing values (Deletion) and when trained after imputation (Imputation). Note that all models are evaluated with a test set that has applied imputation, otherwise we would not be able to compare all the options with the same data.

The first observation we can make is that in terms of accuracy, imputation is generally beneficial for all cases and almost all techniques. In terms of fairness, we see in Figures 6 and 7 that for Adult, Autism and Credit approval with `Sex` all methods *reduce* bias from the perfect model, which is quite remarkable. An opposite result happens for Recidivism, and Titanic with `Sex`, while Titanic with `Class` and Credit Approval with `Ethnicity` are more mixed depending on the technique. Juvenile Offenders present a particular behaviour, as the many models are worse than the majority class model.

When we look at the trade-off between fairness and performance comparing deletion and imputation, we see the expected trend of more performance for imputation implying less fairness, with the exception of Titanic with `class`. However, if we look more carefully method by method, we see that RF behaves quite differently: using imputation increases accuracy significantly but does not amplify (and sometimes reduces) bias. Actually, RF is the only method that draws an almost straight line between the Deletion point, the Imputation point and the Perfect point, except for Credit Approval with `Ethnicity` and the worse-than-majority behaviour for Juvenile Offenders.

Many other models are far from the segment joining the majority class model and the perfect model and hence very far from the boundary of ideal models that we show in Proposition A1 in Appendix A, but there are exceptions, such as Autism with `Gender` and Credit Approval `Sex`,

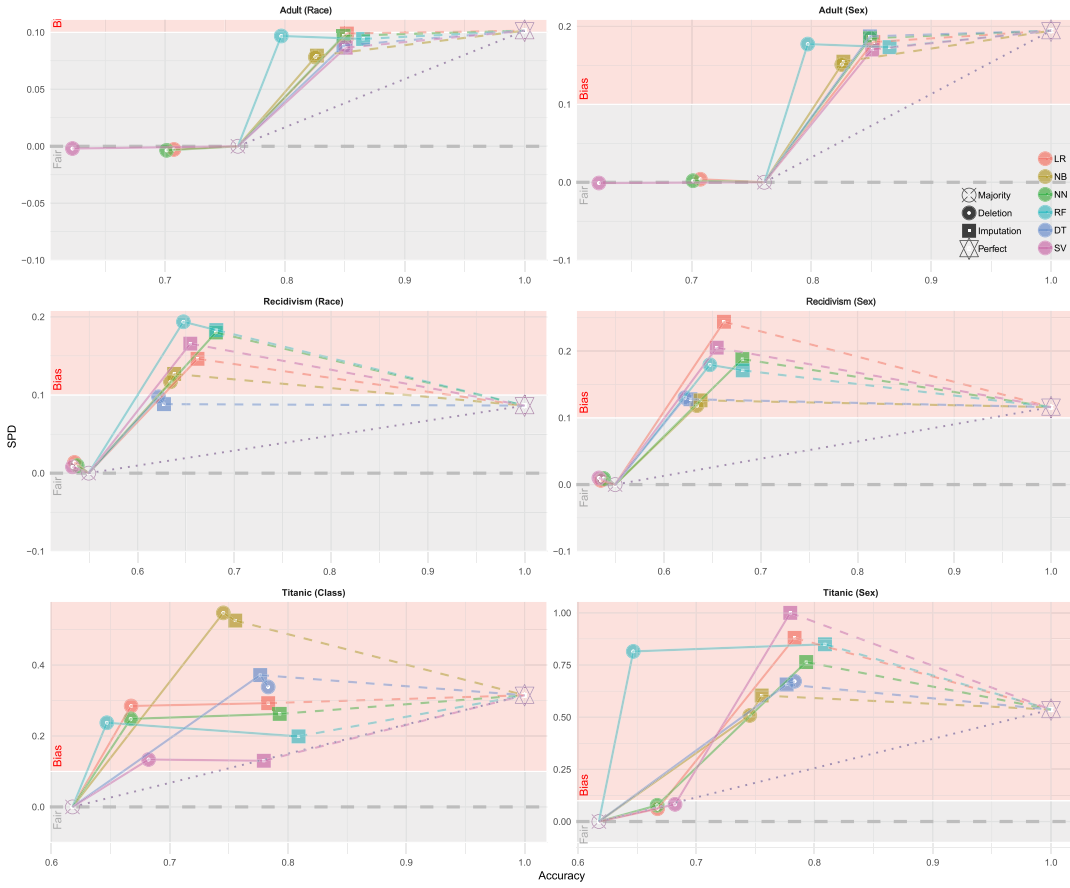

FIGURE 6 Performance (accuracy) and Bias (SPD) for six machine-learning models, using different subsets of data for training. Imputation: all data with the missing values being imputed. Deletion: only the data without the missing values. The test set is always with imputation. The plots also include two baselines: Majority (a majority class model) and Perfect (a perfect model, i.e., using the correct label from the test set). DT, decision tree; LR, logistic regression; NB, naive Bayes; NN, neural Network; RF, random Forest; SPD, StatisticalParityDierence; SV, support vector [Color figure can be viewed at wileyonlinelibrary.com]

and the particular case of model SV for Titanic with `class`. This would be the ideal behaviour, but it would depend more on the learning model than the imputation, as the separation already happens with the deletion results. DTs are usually in a good position dominating many other classifiers (closer to the straight line between majority and perfect). This is good news, as DTs have the advantage of being interpretable and easy to get their decisions conditioned to some attributes, and they could be manually modified to get to different positions of the accuracy-versus-bias space.

7 | DISCUSSION AND FUTURE WORK

While a predictive model can be trained on any particular subset of the data (e.g., excluding missing values), we need to evaluate any predictive model on *all* the examples (otherwise we would be cheating with the results and the model would ultimately need to delegate these

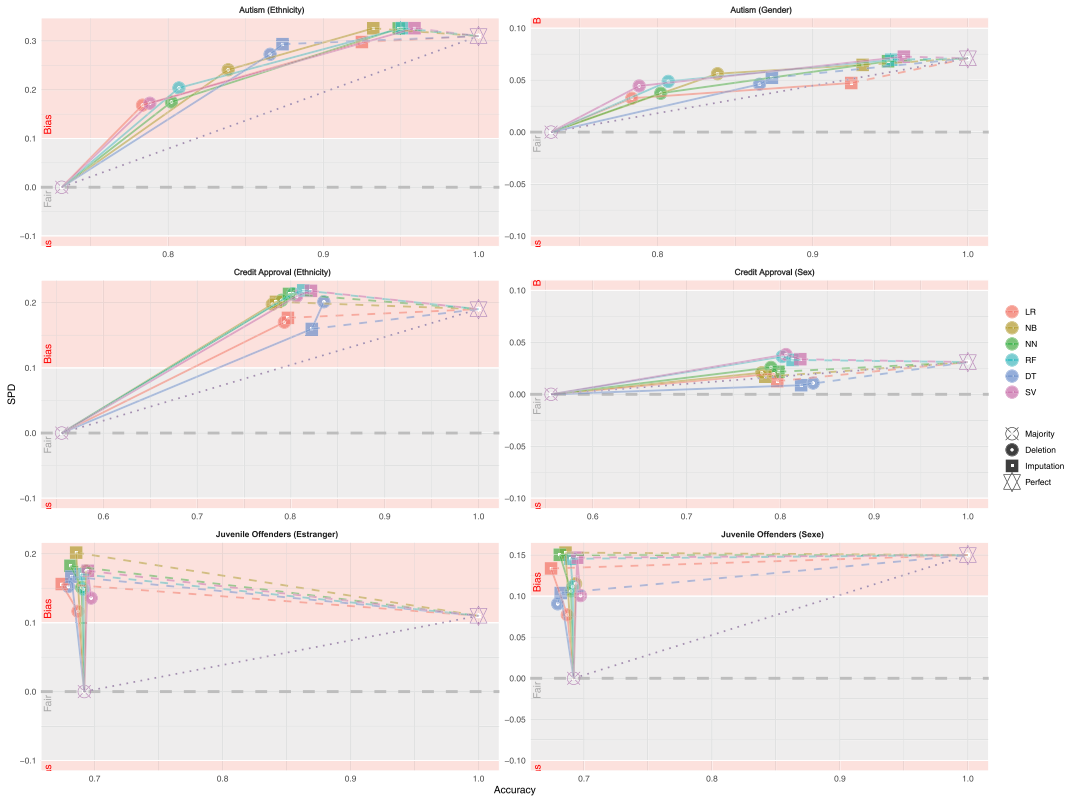

FIGURE 7 Performance (accuracy) and Bias (SPD) for six machine-learning models, using different subsets of data for training. Imputation: all data with the missing values being imputed. Deletion: only the data without the missing values). The test set is always with imputation. The plots also include two baselines: Majority (a majority class model) and Perfect (a perfect model, i.e., using the correct label from the test set). DT, decision tree; LR, logistic regression; NB, naive Bayes; NN, neural Network; RF, random Forest; SPD, StatisticalParityDierence; SV, support vector [Color figure can be viewed at wileyonlinelibrary.com]

examples to humans). Consistently, if we want to use techniques that do not handle missing values—during deployment—we need to delete the columns or use IMs. Still, we can use different techniques and learn the models without these rows, and see if they are better or worse than the models trained with the rows with imputation. This is exactly what we have done in the analysis in Section 6 and it is our recommendation as a general practice. Using these visualisations and performing the same procedures and analyses, we can locate the baseline classifiers and the space of possible classifiers, representing several techniques inside it. We can then have a clearer view of what to do for a particular problem, and which classifier we have to choose depending on the requirements of fairness (on the y-axis) or performance (on the x-axis).

At a theoretical level, new machine-learning techniques are being introduced where one can declare whether an attribute is protected. In the future, we should also investigate new methods to determine what to do with the missing values. We could also add causal models of how the attributes are related to (or controlled by) the primitive causes of missingness and unfairness, which would ensure better traceability than current methods. With the off-the-shelf methods we have available today, however, it is much more plausible that we need to analyse

the results for a particular problem (data set), favourable class and protected groups, and find the best compromise on the Pareto front between the chosen performance and fairness metrics, as we have illustrated with this paper.

Recently, some interesting papers have proposed new imputations methods for imputing missing values. Many of these new methods are based on machine-learning techniques that have gained popularity recently. Examples are Generative Adversarial Nets,[85] fuzzy-rough sets,[86] deep generative models[87] and autoencoders.[88] All these methods claim to improve the accuracy of the final machine-learning models. However, given the complexity of both phenomena (missing values and fairness) and their interaction, it is quite unlikely that a new (predictive) method of imputation works well for all techniques and all possible bias mitigation methods (and all the fairness and performance metrics). In fact, in our experiments, we detect the expected trend of more performance implying less fairness. Therefore, these methods could imply a significant degradation in the fairness of the models.

Another avenue of future analysis could be based on exploring the artificial generation of synthetic missing data, to complement the 12 scenarios with both fairness and missing values issues that we have analysed (and arranged as a publicly available benchmark). This could follow a univariate/multivatiate configuration, at different percentages (missing rates) and according to distinct missing mechanisms: MAR, MCAR and MNAR. However, to be meaningful, we would need to build a causal model and its relation to fairness (otherwise, conclusions derived from ill-defined configurations would be invalid). Similarly, there is a huge number of combinations of fairness metrics and performance metrics that could be explored. For instance, for very imbalanced data sets, it may be more interesting to at least compare the results with metrics, such as AUC, and understand the space in this case. Note that this study should be accompanied by classifiers that predict scores or probabilities, jointly with a proper analysis of the relation between calibration and both fairness and missing values.[89]

8 | CONCLUSIONS

We have presented the first comprehensive analysis of the relation between missing values and fairness. We investigated the causes of missing values and the causes of unfair decisions, and we saw many conceptual connections, and underlying causes for both. The surprising result was to find that, for the data sets studied in this paper, the examples containing missing values seem to be fairer than the rest. Unfortunately, missing values are usually ignored in the literature, even if we can easily transform them from ugly to graceful if handled appropriately. Indeed, we have seen that they incorporate information that could be useful to limit the amplification of bias, but they can have a negative effect on accuracy too if not dealt with conveniently. As a more precise list of lessons learnt, recovering the questions from the introduction, we give the following answers: (1) Missing data and fairness are closely related, and have common causes. From the empirical results in our experimental setting, we see that the correlations between the protected attributes and the attributes with missing values for the different data sets are usually small, so the effect may happen through other proxy variables. (2) Subsamples with missing data are less unfair, but using partial samples is usually counterproductive in terms of accuracy, as fairness and accuracy are usually presented as a trade-off. The octagons give us a perspective that while many plots get into the unfair areas as accuracy grows, this is usually worse than a straight line from the cases without imputation to those with imputation. (3) IMs can find a good trade-off between performance and fairness: imputing

missing values is an easy and effective way to deal with them, instead of discarding the rows with missing data. This is also reinforced by the sensitivity analysis performed for those attributes with this sort of values. What we found is that, apart from the well-known benefits of imputation for accuracy, IMs should be used and improved for the purpose of fairness.

Of course, as said above, more experimental analysis and new techniques could shed more light on the question. However, given the complexity of both phenomena (missing values and fairness) and their interaction, it is quite unlikely that a new (predictive) method of imputation works well for all techniques and all possible bias mitigation methods (and all the fairness and performance metrics). Instead, we have followed a reproducible methodology, based on representations and metrics, and it depends on each case to find the best choice, as we have recommended in Section 7. In the end, fairness is a delicate issue for which context-insensitive modelling pipelines may miss relevant information and dependencies for a particular problem. For the moment, we have learnt that ignoring the missing values is a rather ugly practice for fairness.

ACKNOWLEDGEMENTS

We are grateful to the anonymous reviewers for their suggestions, which have helped improve the paper. We thank the members of the DMIP group of the VRAIN institute for insightful discussions. This material is based upon work supported by the EU (FEDER), and the Spanish MINECO under grant RTI2018- 094403-B-C32, the Generalitat Valenciana PROMETEO/2019/098. F. Martínez-Plumed acknowledges funding of the HUMAINT project by DG CONNECT and DG JRC of the European Commission. J. Hernández-Orallo is also funded by an FLI grant RFP2-152.

AUTHOR CONTRIBUTIONS

F.M.P., C.F. and J.H.O. participated in the definition and refinement of the goals of this study and the hypotheses. F.M.P., C.F. and J.H.O. conceived the technical methodology. F.M.P., C.F. and D.N. processed the data and implemented the code. F.M.P., C.F. and J.H.O. discussed the results and contributed to the writing of the final manuscript.

ENDNOTES

* As opposed to COMPAS,[26] SAVRY is not a system based on machine-learning models, but an assessment protocol that guides the evaluating expert through the individual features that make up the overall risk assessment.[27]

† https://cran.r-project.org/web/packages/BaylorEdPsych/index.html

‡ In this study we only consider binary problems where one class is the "favourable" case (c^+) and the another is the "unfavourable" case (c^-). In the case of multiclass problems, favourable classes can be merged into a single label c^+ and the rest merged into c^-.

§ For the sake of reproducibility, all the code and data can be found in https://github.com/nandomp/missingFairness

¶ In this paper, we consider a limited version of the original data set. We only use 22 of the 100 original attributes. We exclude 78 attributes because they have more than 30% of missing values. The use of a limited number of attributes explains the poor results obtained by the machine-learning models.

We do not include the octagons in Figure 6 to keep the plots clean but also because the results in these plots are the average for several repetitions, while calculating means of octagons is not conceptually correct. Instead, we show the octagons for the whole data set in Appendix A. With the separation, it is also easier to have in mind that the bounding octagons can have small variations for each of the partitions.

ORCID

Martínez-Plumed Fernando http://orcid.org/0000-0003-2902-6477
Ferri Cèsar http://orcid.org/0000-0002-8975-1120
Nieves David https://orcid.org/0000-0003-4540-7237
Hernández-Orallo José http://orcid.org/0000-0001-9746-7632

REFERENCES

1. Angwin J, Larson J, Mattu S, Kirchner L. Machine bias: there's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*. 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
2. Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R. Fairness through awareness. In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, Cambridge, MA, USA, 8-10 January 2012. ACM; 2012:214-226.
3. Grgic-Hlaca N, Redmiles EM, Gummadi KP, Weller A. Human perceptions of fairness in algorithmic decision making: a case study of criminal risk prediction. In: *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, Lion, France, 23-27 April 2018. International World Wide Web Conferences Steering Committee; 2018:903-912.
4. Munoz C, Smith M, Patil D. *Big Data: a Report on Algorithmic Systems, Opportunity, and Civil Rights*. United States: White House Office; 2016.
5. Noble SU. *Algorithms of Oppression: How Search Engines Reinforce Racism*. United States: NYU Press; 2018.
6. Speicher T, Heidari H, Grgic-Hlaca N, et al. 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*, London, UK, August 2018. ACM; 2018:2239-2248.
7. Tolan S, Miron M, Gómez E, Castillo C. Why machine learning may lead to unfairness: evidence from risk assessment for juvenile justice in Catalonia. In: *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law, ICAIL 2019, Montreal, QC, Canada, June 17-21, 2019*. ACM; 2019:83-92.
8. Dignum V, Baldoni M, Baroglio C, et al. Ethics by design: necessity or curse? In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM; 2018:60-66.
9. Whittlestone J, Nyrup R, Alexandrova A, Dihal K, Cave S. *Ethical and Societal Implications of Algorithms, Data, and Artificial Intelligence: a Roadmap for Research*. London, UK: Nuffield Foundation; 2019.
10. Agarwal A, Beygelzimer A, Dudík M, Langford J, Wallach H. A reductions approach to fair classification. arXiv preprint arXiv:1803.02453, 2018.
11. Alvero A, Arthurs N, Antonio AL, et al. AI and holistic review: informing human reading in college admissions. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, New York, USA, 7-8 February 2020. 2020:200-206.
12. Berk R, Heidari H, Jabbari S, et al. A convex framework for fair regression. arXiv preprint arXiv:1706.02409, 2017.
13. Bolukbasi T, Chang K-W, Zou JY, Saligrama V, Kalai AT. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: *Advances in Neural Information Processing Systems*, Barcelona, Spain, 5-10 December 2016. 2016:4349-4357.
14. Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. In: *Advances in Neural Information Processing Systems*, Barcelona, Spain, 5-10 December 2016. 2016:3315-3323.
15. Kusner MJ, Loftus J, Russell C, Silva R. Counterfactual fairness. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, eds. *Advances in Neural Information Processing Systems*. Vol 30. United States: Curran Associates Inc.; 2017:4066-4076.
16. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635, 2019.
17. Zafar MB, Valera I, Rodriguez M, Gummadi K, Weller A. From parity to preference-based notions of fairness in classification. In: *Advances in Neural Information Processing Systems*, Long Beach, CA, USA. 2017:229-239.
18. Barocas S, Selbst AD. Big data's disparate impact. *Calif Law Rev*. 2016;104:671.

19. Lin B-R, Kifer D. Information measures in statistical privacy and data processing applications. *ACM Trans Knowl Discovery Data* (TKDD). 2015;9(4):28.
20. Romero-Tris C, Megías D. Protecting privacy in trajectories with a user-centric approach. *ACM Trans Knowl Discovery Data (TKDD)*. 2018;12(6):67.
21. Kohavi R. Data mining and visualization. In: *Sixth Annual Symposium on Frontiers of Engineering*. DC: National Academy Press; 2001:30-40.
22. Friedler SA, Scheidegger C, Venkatasubramanian S, Choudhary S, Hamilton EP, Roth D. A comparative study of fairness-enhancing interventions in machine learning. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Atlanta, GA, 29-31 January 2019. 2019:329-338.
23. Bellamy RK, Dey K, Hind M, et al. AI Fairness 360: an extensible toolkit for detecting, and mitigating algorithmic bias. *IBM J Res Dev*. 2019;63(4/5):1-15.
24. Saleiro P, Kuester B, Stevens A, et al. Aequitas: a bias and fairness audit toolkit. arXiv preprint arXiv:1811.05577, 2018.
25. Bantilan N. Themis-ML: a fairness-aware machine learning interface for end-to-end discrimination discovery and mitigation. *J Technol Hum Serv*. 2018;36(1):15-30.
26. Northpoint I. *Compas Risk and Need Assessment System*. Tech. rep. 2012.
27. Borum R, Lodewijks H, Bartel PA, Forth AE. Structured assessment of violence risk in youth (SAVRY). *Handbook of Violence Risk Assessment*. England, UK: Routledge; 2011:73-90.
28. Hilterman EL, Nicholls TL, van Nieuwenhuizen C. Predictive validity of risk assessments in juvenile offenders: comparing the SAVRY, PCL: YV, and YLS/CMI with unstructured clinical assessments. *Assessment*. 2014;21(3):324-339.
29. Peugh JL, Enders CK. Missing data in educational research: a review of reporting practices and suggestions for improvement. *Rev Educ Res*. 2004;74(4):525-556.
30. Rombach I, Rivero-Arias O, Gray AM, Jenkinson C, Burke O. The current practice of handling and reporting missing outcome data in eight widely used proms in RCT publications: a review of the current literature. *Qual Life Res*. 2016;25(7):1613-1623.
31. Dua D, Graff C. *UCI Machine Learning Repository*. Irvine, CA, USA: University of California, School of Information and Computer Sciences; 2017.
32. Lavrakas PJ. *Encyclopedia of Survey Research Methods*. United States: Sage Publications; 2008.
33. García S, Ramírez-Gallego S, Luengo J, Benítez JM, Herrera F. Big data preprocessing: methods and prospects. *Big Data Anal*. 2016;1(1):9.
34. Kossinets G. Effects of missing data in social networks. *Soc Netw*. 2006;28(3):247-268.
35. Little RJ, Rubin DB. *Statistical Analysis with Missing Data*. Vol 333. United States: John Wiley & Sons; 2014.
36. Rubin DB. Inference and missing data. *Biometrika*. 1976;63(3):581-592.
37. Enders CK. *Applied Missing Data Analysis*. United States: Guilford Press; 2010.
38. Enders CK. Missing not at random models for latent growth curve analyses. *Psychol Methods*. 2011;16(1):1.
39. Molenberghs G, Fitzmaurice G. Incomplete data: introduction and overview. In: *Longitudinal Data Analisis*. England, UK: Taylor and Francis; 2009:395-408.
40. Millsap RE, Maydeu-Olivares A. *The SAGE Handbook of Quantitative Methods in Psychology*. United States: Sage Publications; 2009.
41. Johnson DR, Young R. Toward best practices in analyzing datasets with missing data: comparisons and recommendations. *J Marriage Fam*. 2011;73(5):926-945.
42. Schafer JL, Graham JW. Missing data: our view of the state of the art. *Psychol Methods*. 2002;7(2):147.
43. Little TD, Schnabel KU. *Modeling Longitudinal and Multilevel Data: Practical Issues, Applied Approaches, and Specific Examples*. England, UK: Psychology Press; 2015.
44. Wilkinson L. Statistical methods in psychology journals: guidelines and explanations. *Am Psychol*. 1999; 54(8):594.
45. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32.
46. Breiman L. *Classification and Regression Trees*. England, UK: Routledge; 1984.
47. Quinlan JR. *C4. 5: Programs for Machine Learning*. Netherlands: Elsevier; 2014.
48. Witten IH, Frank E, Hall MA, Pal CJ. *Data Mining: Practical Machine Learning Tools and Techniques*. Netherlands: Morgan Kaufmann; 2016.

49. Little RJ. A test of missing completely at random for multivariate data with missing values. *J Am Stat Assoc*. 1988;83(404):1198-1202.
50. Chouldechova A, Roth A. The frontiers of fairness in machine learning. arXiv preprint arXiv:1810.08810, 2018.
51. Devine PG, Plant EA, Amodio DM, Harmon-Jones E, Vance SL. The regulation of explicit and implicit race bias: the role of motivations to respond without prejudice. *J Pers Soc Psychol*. 2002;82(5):835.
52. Donaldson SI, Grant-Vallone EJ. Understanding self-report bias in organizational behavior research. *J Bus Psychol*. 2002;17(2):245-260.
53. Heckman JJ. Sample selection bias as a specification error. *Econ: J Econ Soc*. 1979:153-161.
54. Kunz TP, Crone SF, Meissner J. The effect of data preprocessing on a retail price optimization system. *Decision Support Syst*. 2016;84:16-27.
55. Li H, Gupta A, Zhang J, Sarathy R. Examining the decision to use standalone personal health record systems as a trust-enabled fair social contract. *Decision Support Syst*. 2014;57:376-386.
56. Millsap RE, Everson HT. Methodology review: statistical approaches for assessing measurement bias. *Appl Psychol Meas*. 1993;17(4):297-334.
57. Nickerson RS. Confirmation bias: a ubiquitous phenomenon in many guises. *Rev Gen Psychol*. 1998;2(2): 175-220.
58. Wijnhoven F, Bloemen O. External validity of sentiment mining reports: Can current methods identify demographic biases, event biases, and manipulation of reviews? *Decision Support Syst*. 2014;59:262-273.
59. Squire P. Why the 1936 literary digest poll failed. *Public Opinion Q*. 1988;52(1):125-133.
60. Gray PO. *Psychology*. United States: Worth Publishers; 2002.
61. Dastin J. Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G, 2018.
62. Zafar MB, Valera I, Gomez Rodriguez M, Gummadi KP. Airness beyond disparate treatment & disparate impact: learning classification without disparate mistreatment. In: *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee; 2017: 1171-1180.
63. Kusner MJ, Loftus J, Russell C, Silva R. Counterfactual fairness. In: *Advances in Neural Information Processing Systems*, Long Beach, CA, USA. 2017:4066-4076.
64. Verma S, Rubin J. Fairness definitions explained. In: *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, Gothenburg, Sweden, 29 May 2018. IEEE; 2018:1-7.
65. Ferri C, Hernández-Orallo J, Modroiu R. An experimental comparison of performance measures for classification. *Pattern Recognition Lett*. 2009;30(1):27-38.
66. Hernández-Orallo J, Flach P, Ferri C. A unified view of performance metrics: translating threshold choice into expected classification loss. *J Mach Learn Res*. 2012;13:2813-2869.
67. Dutta S, Wei D, Yueksel H, Chen P-Y, Liu S, Varshney K. Is there a trade-off between fairness and accuracy? A perspective using mismatched hypothesis testing. In: *International Conference on Machine Learning*, 13-18 July 2020. PMLR; 2020:2803-2813.
68. Milfont TL, Fischer R. Testing measurement invariance across groups: applications in cross-cultural research. *Int J Psychol Res*. 2010;3(1):111-130.
69. Schwarz N, Groves RM, Schuman H. Survey methods. In: Gilbert DT, Fiske ST, Gardner L, eds. *Handbook of Social Psychology*. 4th ed. New York: McGraw-Hill; 1998:143-179. Chapter 4.
70. Buolamwini J, Gebru T. Gender shades: intersectional accuracy disparities in commercial gender classification. In: *Conference on Fairness, Accountability and Transparency*, New York University, New York, NY, 23, 24 February 2018. 2018:77-91.
71. Raji ID, Buolamwini J. Actionable auditing: investigating the impact of publicly naming biased performance results of commercial AI products. In: *AAAI/ACM Conference on AI Ethics and Society*, Honolulu, HI, USA, 27-28 January 2019. Vol 1. 2019.
72. Ahmad MA, Eckert C, Teredesai A. The challenge of imputation in explainable artificial intelligence models. arXiv preprint arXiv:1907.12669, 2019.
73. Therneau T, Atkinson B, Ripley B, Ripley MB. Package 'rpart'. cran.ma.ic.ac.uk/web/packages/rpart/rpart.pdf (accessed on 20 April 2016), 2015.

74. Therneau TM, Atkinson EJ. An introduction to recursive partitioning using the RPART routines (Technical Report). Vol. 61. United States: Mayo Foundation; 1997.
75. Kuhn M. Building predictive models in R using the caret package. *J Stat Software*. 2008;28(5):1-26.
76. Alpaydin E. *Introduction to Machine Learning*. United States: MIT Press; 2009.
77. Flach P. *Machine Learning: the Art and Science of Algorithms that Make Sense of Data*. England, UK: Cambridge University Press; 2012.
78. Hernández-Orallo J, Ferri C, Ramírez-Quintana MJ. *Introduction to Data Mining*. United States: Pearson Prentice Hall; 2004.
79. Setiono R, Liu H. Neural-network feature selector. *IEEE Trans Neural Netw*. 1997;8(3):654-662.
80. Ribeiro MT, Singh S, Guestrin C. Why should I trust you? Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, 13-17 August 2016. ACM; 2016:1135-1144.
81. Renooij S, Van Der Gaag LC. Evidence and scenario sensitivities in naive Bayesian classifiers. *Int J Approx Reason*. 2008;49(2):398-416.
82. Xu H, Caramanis C, Mannor S. Robustness and regularization of support vector machines. *J Mach Learn Res*. 2009;10:1485-1510.
83. Ferri C, Hernández-Orallo J, Martınez-Usó A, Ramırez-Quintana M. Identifying dominant models when the noise context is known. In: *First Workshop on Generalization and Reuse of Machine Learning Models Over Multiple Contexts*, Nancy, France, 19 September 2014. 2014.
84. Khamis A, Ismail Z, Haron K, Mohammed AT. The effects of outliers data on neural network performance. *J Appl Sci*. 2005;5(8):1394-1398.
85. Yoon J, Jordon J, van der Schaar M. GAIN: missing data imputation using generative adversarial nets. In: Dy J, Krause A, eds. *Proceedings of the 35th International Conference on Machine Learning (Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018), Proceedings of Machine Learning Research*. Vol 80. PMLR; 2018:5689-5698.
86. Amiri M, Jensen R. Missing data imputation using fuzzy-rough methods. *Neurocomputing*. 2016;205:152-164.
87. Camino RD, Hammerschmidt CA, State R. Improving missing data imputation with deep generative models. arXiv preprint arXiv:1902.10666, 2019.
88. Beaulieu-Jones BK, Moore JH. Missing data imputation in the electronic health record using deeply learned autoencoders. In: *Pacific Symposium on Biocomputing 2017*, Hawaii, United States, 3-7 January 2017. World Scientific; 2017:207-218.
89. Bella A, Ferri C, Hernández-Orallo J, Ramírez-Quintana MJ. Calibration of machine learning models. In: *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*. United States: IGI Global; 2010:128-146.

How to cite this article: Fernando M-P, Cèsar F, David N, José H-O. Missing the missing values: The ugly duckling of fairness in machine learning. *Int J Intell Syst*. 2021; 36:3217-3258. https://doi.org/10.1002/int.22415

APPENDIX A: THE SPD-ACCURACY SPACE

The following proposition shows the space for the possible classifiers in terms of the balance between SPD and accuracy.

Proposition A1. *Assuming that the majority class is the favourable class, the set of classifiers with respect to accuracy and SPD is bounded by the octagon composed of the following eight points:*

$$\begin{aligned}
&\text{Acc} = 1, \quad \text{SPD}_i^+ = p^+(D_{X_i=a}) - p^+(D_{X_i\neq a}),\\
&\text{Acc} = 1 - \frac{\text{Pos}(D_{X_i=a})}{|D|}, \quad \text{SPD}_i^+ = -\frac{\text{Pos}(D_{X_i\neq a})}{|D_{X_i\neq a}|},\\
&\text{Acc} = \frac{\text{Neg}(D_{X_i=a}) + \text{Pos}(D_{X_i\neq a})}{|D|}, \quad \text{SPD}_i^+ = -1,\\
&\text{Acc} = \frac{\text{Neg}(D_{X_i=a})}{|D|}, \quad \text{SPD}_i^+ = -1 + \frac{\text{Pos}(D_{X_i\neq a})}{|D_{X_i\neq a}|},\\
&\text{Acc} = 0, \quad \text{SPD}_i^+ = p^+(D_{X_i\neq a}) - p^+(D_{X_i=a}),\\
&\text{Acc} = \frac{\text{Pos}(D_{X_i=a})}{|D|}, \quad \text{SPD}_i^+ = \frac{\text{Pos}(D_{X_i\neq a})}{|D_{X_i\neq a}|},\\
&\text{Acc} = \frac{\text{Pos}(D_{X_i=a}) + \text{Neg}(D_{X_i\neq a})}{|D|}, \quad \text{SPD}_i^+ = 1,\\
&\text{Acc} = 1 - \frac{\text{Neg}(D_{X_i=a})}{|D|}, \quad \text{SPD}_i^+ = 1 - \frac{\text{Pos}(D_{X_i\neq a})}{|D_{X_i\neq a}|}.
\end{aligned}$$

Proof. As we have assumed that the majority class is the favourable class, we have the following point for the perfect model:

$$\begin{aligned}
&\text{Acc}[\text{Perfect}] = 1,\\
&\text{SPD}_i^+[\text{Perfect}] = p^+(D_{X_i=a}) - p^+(D_{X_i\neq a}).
\end{aligned}$$

We denote by $\hat{D}$ the data labelled with a classifier, so it will have accuracy and SPD as follows.

$$\begin{aligned}
\text{Acc}(\hat{D}) &= \frac{\text{TP}(\hat{D}) + \text{TN}(\hat{D})}{|D|} = \frac{\text{TP}(\hat{D}_{X_i=a}) + \text{TN}(\hat{D}_{X_i=a}) + \text{TP}(\hat{D}_{X_i\neq a}) + \text{TN}(\hat{D}_{X_i\neq a})}{|D|},\\
\text{SPD}_i^+(\hat{D}) &= p^+(\hat{D}_{X_i=a}) - p^+(\hat{D}_{X_i\neq a}) = \frac{\text{TP}(\hat{D}_{X_i=a}) + \text{FP}(\hat{D}_{X_i=a})}{|D_{X_i=a}|} - \frac{\text{TP}(\hat{D}_{X_i\neq a}) + \text{FP}(\hat{D}_{X_i\neq a})}{|D_{X_i\neq a}|}.
\end{aligned}$$

Let us start with the perfect classifier, and let us assume wlog that SPD_i^+ is positive.

SPD positive

We consider that the favourable class is the positive class, so we have four cases in which a prediction can change:

1. In $D_{X_i=a}$, a TN changes to a FP. This increments SPD_i^+ in $\frac{1}{|D_{X_i=a}|}$.
2. In $D_{X_i=a}$, a TP changes to an FN. This decrements SPD_i^+ in $\frac{1}{|D_{X_i=a}|}$.
3. In $D_{X_i\neq a}$, a TN changes to an FP. This decrements SPD_i^+ in $\frac{1}{|D_{X_i\neq a}|}$.
4. In $D_{X_i\neq a}$, a TP changes to an FN. This increments SPD_i^+ in $\frac{1}{|D_{X_i\neq a}|}$.

As SPD_i^+ is positive and we want to reduce it, then only the mistakes 2 and 3 are useful. If $|D_{X_i=a}| < |D_{X_i\neq a}|$, then it is more advantageous to do 2. Otherwise, it is more advantageous to do 3 first.

SPD positive and $|D_{X_i=a}| < |D_{X_i\neq a}|$

In this situation the strategy consists in moving one by one all the elements from TP to FN (exactly all the TP instances in $D_{X_i=a}$, which are simply the positives as we started with the perfect classifier), so that we move to a point that is located at

$$\text{Acc} = \text{Acc}[\text{Perfect}] - \frac{\text{Pos}(D_{X_i=a})}{|D|} = 1 - \frac{\text{Pos}(D_{X_i=a})}{|D|},$$
$$\text{SPD}_i^+ = \text{SPD}_i^+[\text{Perfect}] - \frac{\text{Pos}(D_{X_i=a})}{|D_{X_i=a}|} = -\frac{\text{Pos}(D_{X_i\neq a})}{|D_{X_i\neq a}|}.$$

As we did not have FP nor FN, because we started from the perfect classifier, and all the TPs for $X = a$ have been converted into FN, we only have TN and FN for $X = a$, which always predicts the negative class when $X = a$ and always correct when $X \neq a$.

Then, when case 2 is exhausted, we can start with case 3 and move one by one all the elements from TN to FP (exactly all the original TN instances in $D_{X_i\neq a}$, which are the negatives), and we get the point:

$$\begin{aligned}\text{Acc} &= \text{Acc}[\text{Perfect}] - \frac{\text{Pos}(D_{X_i=a})+\text{Neg}(D_{X_i\neq a})}{|D|} = 1 - \frac{\text{Pos}(D_{X_i=a})+\text{Neg}(D_{X_i\neq a})}{|D|}\\ &= \frac{\text{Neg}(D_{X_i=a})+\text{Pos}(D_{X_i\neq a})}{|D|},\\ \text{SPD}_i^+ &= \text{SPD}_i^+[\text{Perfect}] - \frac{\text{Pos}(D_{X_i=a})}{|D_{X_i=a}|} - \frac{\text{Neg}(D_{X_i\neq a})}{|D_{X_i\neq a}|}\\ &= -\frac{\text{Pos}(D_{X_i\neq a})}{|D_{X_i\neq a}|} - \frac{\text{Neg}(D_{X_i\neq a})}{|D_{X_i\neq a}|} = -1.\end{aligned}$$

Again, the last step is explained as we did not have FP nor FN, because we started from the perfect classifier, and all the TPs for $X = a$ have been converted into FN, so we only have TN and FN for $X = a$, which means it always predicts the negative class when $X = a$. As all the TNs for $X \neq a$ have been converted into FP, we only have TP and FP for $X \neq a$, which means it always predicts the positive class when $X \neq a$. This is the most negative bias, -1.

Note that as the value of SPD is negative already after all the 2s, we know that perfect fairness for SPD has been achieved before exhausting the 2s. However, the points on the segments connecting these points and the original perfect classifier do not need to contain the majority classifier, which will usually be left out of these two segments.

And now if we want to explore the other two cases, we are at $\text{SPD} = -1$ and we can only increase. If $|D_{X_i=a}| < |D_{X_i\neq a}|$ then now we want to increase SPD as slowly as possible to cover the whole space, so it is now more advantageous to do 4 as the steps are smaller.

As in the previous cases, now we change TP to FN, so it is the positives in $D_{X_i\neq a}$ that we lose, so incrementally over the previous point this would lead to

$$\begin{aligned}\text{Acc} &= \frac{\text{Neg}(D_{X_i=a})+\text{Pos}(D_{X_i\neq a})-\text{Pos}(D_{X_i\neq a})}{|D|}\\ &= \frac{\text{Neg}(D_{X_i=a})}{|D|},\\ \text{SPD}_i^+ &= -1 + \frac{\text{Pos}(D_{X_i\neq a})}{|D_{X_i\neq a}|}.\end{aligned}$$

The other four points of the octagon are mirrored with the above (symmetric with the limits of Acc and SPD, respectively).

SPD positive and $|D_{X_i=a}| > |D_{X_i\neq a}|$

The situation when $|D_{X_i=a}| > |D_{X_i\neq a}|$ follows similarly but does 2 before 3 and then gets $D_{X_i\neg a}$ and $D_{X_i=a}$ swapped. Similarly for the rest of the octagon. Basically, we are exploring a convex octagon, so making a wrong choice does not give us the whole space (there would be concave parts).

SPD negative. Finally, the scenario when SPD of the perfect classifier is negative is analogous. □

An example of the octagonal SPD-Accuracy space for a test data set is included in Figure A1. The data set presents the following metrics: $\mathrm{SPD}_i^+ = 0.3$, $p^+(D_{X_i=a}) = 0.8$, $p^+(D_{X_i\neq a}) = 0.5$ and $|D_{X_i=a}|/|D| = 0.3$. Figures A2 and A3 show the corresponding SPD-Accuracy spaces for the six data sets and protected groups included in Table 1.

APPENDIX B: THE RELATIONSHIP BETWEEN FARINESS AND THE PROPORTION OF MISSING VALUES

Are the fairness metrics affected by different ratios of missing values in a particular data set? Following this general question, here we empirically analyse whether there is a relationship between the SPD values and the ratio of missing values. For illustrative purposes, we focus on the Titanic data set as it is the one with the highest percentage of rows with missing values (20%) in our experimental setting (see Table 1). We have varied this percentage from 20% to 2% subsampling the sets of rows containing missing values to see whether there is any effect wrt to the fairness metric between subsamples. We perform 100 repetitions for each variation

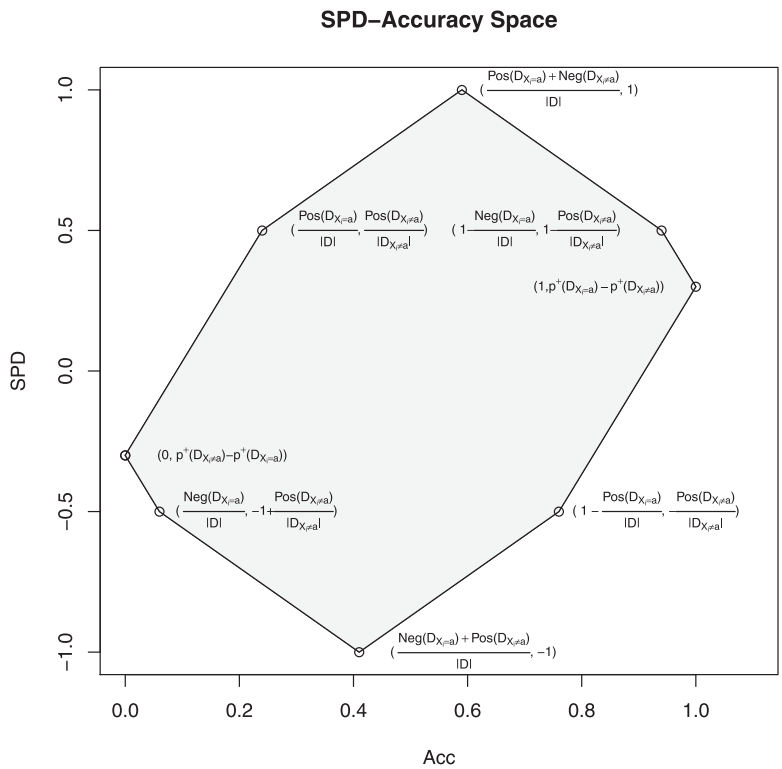

FIGURE B1 Variation of SPD values for different subsamplings of data with or without missing values for the Titanic data set. Proportion of missing values varies from 2% to 20%. SPD values are shown considering only the data ("SPD (w Missing)" and "SPD (w/o Missing)") and using CART as a predictive model ("SPD (w Missings - Rpart)" and "SPD (w/o Missings - Rpart)"). SPD, Statistical Parity Difference

FIGURE A1 SPD-accuracy space for a data set with $\mathrm{SPD}_i^+ = 0.3$, $p^+(D_{X_i=a}) = 0.8$, $p^+(D_{X_i\neq a}) = 0.5$ and $|D_{X_i=a}|/|D| = 0.3$. SPD, Statistical Parity Difference

FIGURE A2 SPD-accuracy spaces for the data sets Adult, Recidivism and Titanic of Table 1. We also show the majority class classifier with a cross, and connect it to the perfect classifier (represented by a star). SPD, Statistical Parity Difference

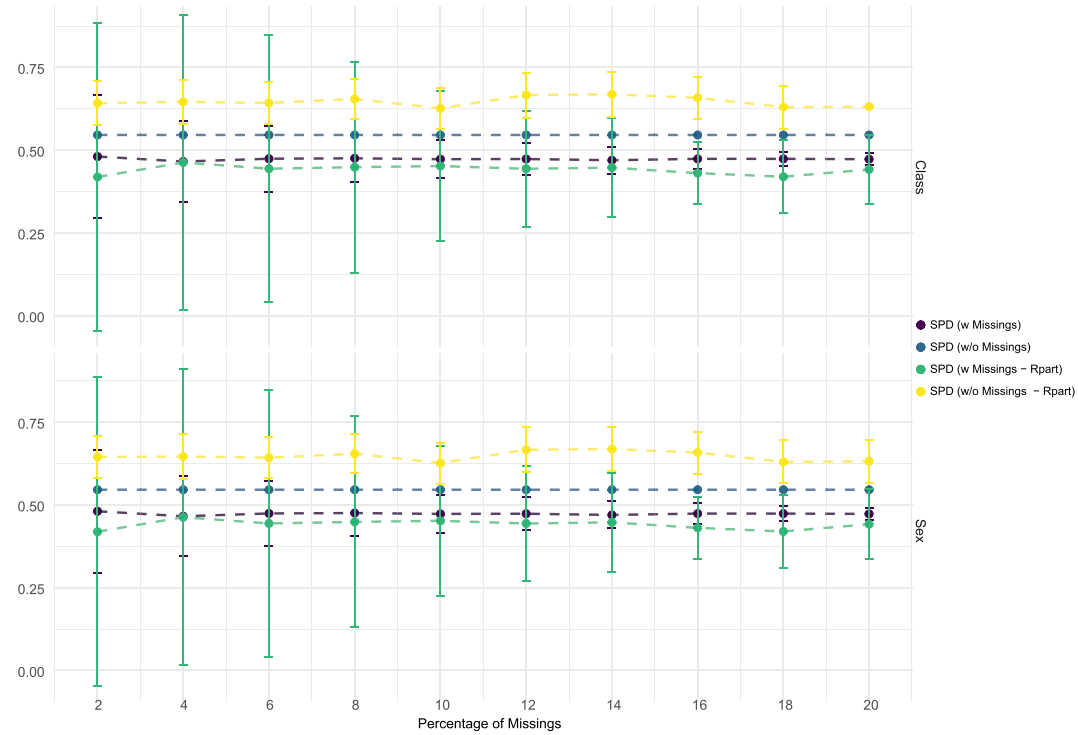

FIGURE A3 SPD-accuracy spaces for the data sets Autism, Credit Approval and Juvenile offenders of Table 1. We also show the majority class classifier with a cross, and connect it to the perfect classifier (represented by a star). SPD, Statistical Parity Difference [Color figure can be viewed at wileyonlinelibrary.com]

(obtaining different subsets of rows) and then we average the results. In Figure B1 we show the average SPD values (and error bars) for different subsets of data and privileged group ("Class or Sex"). We can see:

- The subset of examples that do not have missing values ("SPD (w/o Missing)"), represented in blue.
- The subset of examples that contain, at least, one missing value in any of their attributes ("SPD (w Missing)"), represented in purple.
- The subset of examples that do not have missing values using CART as a predictive model ("SPD (w Missings - Rpart)," represented in yellow.
- The subset of examples that contain, at least, one missing value in any of their attributes using CART as a predictive model ("SPD (w Missings - Rpart)"), represented in green.

The first item ("SPD (w/o Missing)") is just included for reference, and should be constant, as it is not affected by the subsampling. If we look at the second one ("SPD (w Missing)"), we can see that the fairness metric remains almost constant for the percentages of missing values (obviously, only the standard deviation decreases as the set of rows containing missing values increases). Similarly, for the latter two we cannot find significant differences for the fairness metric as they remain almost constant for the different subsets of rows containing

missing values. Note that similar results can be obtained with the rest of data sets in our paper, but with even smaller variations due to the lower number of instances with missing values.

In brief, this illustrative example shows us that there seems to be no general relationship between the degree of fairness and the ratio of missing values (at least for the set of data sets in our paper).