Impact of missing values in AI fairness

Athanasios Tompras

Master Thesis

Supervisor Professor: Evaggelia Pitoura

Ioannina, February, 2023

ΤΜΗΜΑ ΜΗΧ. Η/Υ & ΠΛΗΡΟΦΟΡΙΚΗΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF IOANNINA

Acknowledgments

I would like to thank especially my supervisor, Professor Evaggelia Pitoura for her valuable help whenever I needed it and for her guidance during the writing of this thesis. Moreover, thanks for her human side and her support as it was a very difficult period for me.

Finally, I would like to thank my close people who supported me and my family without whom I would not have made it. I dedicate this work to my mom who is no longer with me but I know she will be proud of me from up there.

20-2-2023

Athanasios Tompras

Abstract

Nowadays, the use of machine learning is increasingly entering our lives. Machine learning algorithms are designed to learn from data, and the data they use to make their predictions may contain biases that can create unfairness and inequalities betweencertain groups of people, particularly groups that are protected variables such asethnicity and gender. In addition, the existence of missing values in the data for various reasons is a very common phenomenon, which creates even greater problems in terms of the fairness of the results of the algorithms. However, many algorithms as well as scientists ignore missing values. In this paper we analyze the reasons for missing values and study the relationship between fairness and missing values. We then handle missing values by different methods of replacing them and calculate the fairness of the original data sets as well as the developed models using metrics. Finally, we apply machine learning models to predict the missing values and see how this contributes to the fairness of the model by comparing them with the previous measurements.

Keywords: fairness, missing values, analysis, imputation methods, algorithms, machine learning

Table of contents

Chapter 1. Introduction

1.1 Field of study

An ever-increasing number of decisions in our daily lives are controlled by artificial intelligence and machine learning (ML) algorithms. Since they affect many areas of our lives, it is now very important to develop algorithms that are not only accurate but also objective and fair. Recent research has shown that decision making can discriminate between certain groups, especially groups that contain protected variables such as gender, gender and ethnicity. This algorithmic error in models is what fairness in machine learning seeks to correct and eliminate. Moreover, fairness depends on the quality of the data as well a s on the quality of its preprocessing. An important factor for data quality is the presence of missing values which may represent the absence of information or its removal for some serious reason. In this paper we consider the following issues:

1. We analyze the sources and reasons for missing values
2. The correlation of missing values with fairness
3. We study whether or not the existence of missing values results in more equitable outcomes
4. We apply different machine learning techniques to fill in missing values by examining whether they perform better in fairness

1.2 Organization of the analysis

Initially, the second chapter analyses the definitions discussed in the topic of the thesis as well as the theoretical background of the concepts needed to understand the measurements and their results.

Chapter three presents the libraries and data pre-processing techniques used to handle the data appropriately.

Chapter four applies the metrics to the data and then lists and comments on the research results for all the datasets and the various techniques and applications of machine learning models.

Finally, in chapter five we draw our conclusions based on our analyses and results and discuss future research.

CHAPTER 2. Theoretical Background

In this chapter we will present and analyze the theoretical framework on which our work is based as well as useful concepts for understanding the results.

2.1 Definition of missing values

Missing values are defined as data that have not been entered for a variable in its observation field. They are a major common problem in almost all areas of research and can have a major impact on the conclusions drawn from the data. According to the UCL repository one of the largest data sources for research and machine learning, missing values have been observed at 45% among the data provided. For this reason, several researches have focused on the handling of these values, the problems they cause as well as methods of handling them to reduce or avoid an error in their research. Such a phenomenon is likely to be caused by several causes which we discuss below to better understand the meaning and effect of missing values.

PARTIAL completion. Partial completion occurs when, after collecting many values for a record at a particular point in time through a survey or questionnaire, the remaining variables are missing. This means that variables towards the end of the questionnaire are more likely to be missing. This type of response is most common in long questionnaires as well as in telephone surveys.

LACK of planning. It refers to the case where certain questions or variables do not fit specific individuals. There are two main reasons for this lack of data.

1. Random variables. Specific questions do not apply (NA) to all individuals. In this case the missingness mechanism is known and can be included in the analysis.

2. Variable sampling. A specific design is used to manage different subsets of questions for different individuals. In this case all questions are addressed to all respondents but for efficiency reasons are not applied to all respondents. Here we know the missingness mechanism but because of the randomness of the questions, we handle them statistically.

NON-RESPONSE of OBJECTS. No information is provided for some respondents on some variables. Some items are likely to be non-responsive relative to others (e.g. Income) In general, surveys and interviews tend to collect data that contain missing values that fall into three main subcategories.

1. Not provided. The information is simply not given for the question.

2. Useless. The information provided is of no use or useless.

3. Lost. Useful information is lost because of a processing problem. The previous two problems occur in the data collection process and the last one in the processing process.

2.2 Categories and imputation methods of missing values

In many cases, for reasons of simplification or because it is not possible to trace the source of the data, we can only characterize some statistical types of missing values. The categorization is made between three types of missing value mechanisms.

1. MCAR (Missing Completely At Random), where the missing values are independent of both the unobserved and observed parameters of interest and occur completely at random (misprediction of a response to a questionnaire) . In this case, the missing values are independent and simple statistical methods can be used.

2. MAR (Missing At Random), missing values depend on observed data, but not on unobserved data (e.g. in a political opinion poll many people may refuse to respond based on demographics, then the missing data depends on the observed variable (demography), but not on the answer to the question itself. In this case, if the variable associated with the shortage is available, the shortage can be handled adequately. MAR is believed to be more general and more realistic than MCAR.

3. MNAR (Missing Not At Random), where the missing values depend on the observed data. MNAR is the most complex unidentifiable case where simple solutions cannot be applied and an explicit model for the missingness must be applied in the analysis.

The way in which missing prices are handled varies depending on the type of missing price mechanism. Below we explain the most common ways in which we can handle missing values and the appropriate one to use for our analysis.

- (LD) Line deletion: in this case the entire line containing even one missing value is deleted from the database. When we handle MCAR missing value problems because the sample is large this particular way of handling them gives satisfactory results.
- (CD) Delete column: This is an extreme case because it removes all the information of a variable from the data set. It can create a large error in case the deleted category is directly related to other categories.
- (LC) Special Category: a variable can be split by adding a new category only for missing values. However, this special category has missing quantitative values. Sometimes the missing value is replaced with a Boolean value and the original value is deleted.
- (IMs) Price replacement: There are many ways to do this such as replacing it with the mean if it is a quantitative variable and the average if it is qualitative or calculating its value from the other variables using prediction models.

The most common method is line deletion (LD). However, for problems of the MCAR class to which ours belong, the use of significantly reduces the sample size and not all data are used adequately. On the other hand, the value substitution method (IM) is now one of the most common and efficient data pre-processing techniques for many machine learning techniques in terms of handling missing values. However, many libraries do not specify the value substitution method they used and this varies depending on the machine learning method used as well as the programming language. Using the same IM methods for all datasets is not appropriate because there is no one mechanism good for all datasets. In many studies multiple value replacement and missing data mechanisms are used and in the end the most efficient one is chosen.

2.3 Definition of fairness

Fairness is a widely used term in the field of artificial intelligence (AI) and machine learning (ML). In order to make fair and unbiased decisions on research results we need to define the concept of fairness. However, before we analyze it we need to indicate the reasons for the existence of bias(bias) that lead us to the need to calculate fairness. We categorize them into six main categories.

SAMPLING or SELECTIVE bias. Occurs when the data sample is not representative of the target population according to the conclusions drawn. This occurs because the sample is collected in such a way that some members of the population are less likely to be included than others.

Measurement bias. Systematic distortion of values occurs when the device used to make the measurement favors a particular result. Systematic bias cannot be avoided by simply collecting more data but by using multiple measurement devices and the right experts to compare results.

Self-reported bias. It has to do with unanswered, incomplete and contradictory answers to surveys, questionnaires and interviews in order to collect data. The main reason for their existence is the presence of personal or sensitive questions.

Confirmation bias. This bias emphasizes a hypothesis because it contains favorable information that does not contradict the researcher's personal desire to find a significant statistical result. It is a type of cognitive error in which a decision is made according to prior beliefs, perceptions and biases and can lead to overlooking important data.

PRESET error. A different situation occurs when the training data we have collected contains human errors based on personal preferences, ideologies and racial discrimination. Unlike the other categories which mainly affect the predicted variables, this type of error is associated with variables used as dependent variables. For this reason, systems used to reduce prediction error will naturally reproduce any error present in the known data.

ALGORITHMIC bias. In this case the algorithm creates or reinforces bias in the training data. For example, different populations in the data may have different distributions of their data. As a result, if we train a blind group classifier to minimize error, since it cannot usually fully match all populations it will match the majority population.

Although the reasons can be categorized, the exact concept of fairness and what causes it in a particular case is more complex. Fairness refers to the various attempts to correct algorithmic error based on machine learning models. Decisions made after a machine learning process are considered unfair if they are based on variables that are considered sensitive are influenced by biases and discrimination. Such examples include variables related to gender, ethnicity, gender and others. Fairness seeks to treat all demographic groups equally. To identify unlawful or unlawful discrimination, the population is divided into two subsets based on one or more sensitive variables. One group has values for these variables that satisfy the treatment of their members, andthe other group which lacks these values. The first group is referred to as privileged and the second as non-privileged or protected depending on its content. Sensitive variables are also called protected variables because it is illegal or unlawful to discriminate according to their value. In the following we will discuss the categorization problems and their analysis in which fairness is a very common concept.

2.4 Definition analysis and symbolism

Before we move on to the analysis and presentation of the results, we need to talk about the symbolism and explanation of the terms we use in our research and results.

First, we deal with the classification problem which is part of supervised learning. Supervised learning is a subset of machine learning and aims to categorize data or predict outcomes using labeled datasets to train appropriate machine learning algorithms. To do this, a training set is used which contains inputs (also known as predictions or independent variables) and the corresponding outputs (also known as targets or dependent variables). The supervised learning algorithm analyzes this set and learns to map the inputs to the corresponding outputs. In the classification problem the goal is to predict a distinct category or label of a class for a given input. In other words, the input consists of a set of features and the output is a finite number of possible categories or classes. We continue by explaining the notations and definitions that will be used later in the paper.

We define a set of variables X, where the subset s < X indicates the protected variables. A protected variable is considered categorical (e.g., gender, sex, ethnicity) and can divide a population into groups that should have equal probability of being favored. For each protected variable Si, we have a set of values Vi (e.g. {man, woman}). Categories are created by assigning values to one or more protected variables. We usually use the term "privileged" category to emphasize that a category has an advantage in its content over others. (e.g. white males). We then consider a class Y variable, which takes values in C (e.g. guilty, not guilty). For fairness analysis we usually consider one of the classes as "favored" or positive outcome, denoting it by $c+$. Similarly, the "non-favored"class is denoted by $c-$. For example, in our case, innocent is the favored outcome. An uncategorized instance x takes values in Vi for each $Xi \in X$, possibly including an additional value. This value is the missing value and we will denote it by NaN. A categorized example is formed from an uncategorized example with the value of class y from C (e.g., {gender = colored, gender = female, income = NaN}, guilty). A decision problem is defined as the mapping of x to $y\hat{}$ (predicted variable) such that $y\hat{}$ *is* correct according to the ground truth.

In machine learning, we define ground truth as information that is known to be true after observation and measurement with respect to reality. Furthermore, given a dataset D we denote as DXi=a the selection of instances such that $Xi = a$, with $a \in Vi$. This also applies to categorized data, such that $Dy=c+$ is simply the number of positive results in D , also denoted as Pos(D) (similarly for negative ones as Neg(D)). The actual positive values are denoted as $y = c+$ and the negative ones as $y = c-$, while the corresponding predicted ones aredenoted as $y\hat{} = c+$ and $y\hat{} = c-$. Finally, given a set <x,y> of a dataset D, theprobability that the outcome is favorable is denoted as $p+(D)$.

2.5 Fairness and performance metrics

The main reason for the paper is the relationship between missing values and fairness and its illustration in real-life situations. For this reason we combine a theoretical analysis of the reasons for missing values and fairness with a practical analysis using different types of classifiers as well as metrics on our databases. To determine whether some examples with missing values are fairer than others and the relationship between protected variables and variables containing missing values, we use the Statistical Parity Difference (SPD) metric. This metric shows us the percentage independence between the protected variable and the outcome of the decision. Following the notations explained earlier, the SPD metric for the favored class is defined for a protected variable Xi with a favored value a as $SPDi+ (D) = p+ (DXi= a) - p+ (DXi \neq a)$.

More specifically, if the variable is gender and the possible values are male and female, if we consider women as the protected group, the SPD metric will be the probability of having favorable outcomes for women minus the probability of having favorable outcomes for men. The closer the value of the metric to 0, the more equitable the outcome between the two groups, while a value greater than 0 implies a higher benefit for the favored group and a value less than 0 for the non-favored group respectively. The sign of the metric can change depending on the favored class we choose as well as the protected variable as shown below: SPDi-$(D) = p- (DXi =a)$
$- p- (DXi \neq a) = 1 - p+ (DXi =a) - (1 - p+ (DXi \neq a)) = -SPDi+(D)$. We will focus on the absolute value of the metric and how close to 0 its value is and therefore how much closer to 0 the outcome is fairer.

Other popular metrics applied to the data and models are:

Disparate Impact(DI), which calculates a percentage instead of the difference like SPD. Metrics that only apply to data are Equal Opportunity Difference (EOD), which is the difference in True Positive Rates(TPR) between groups and Average Odds Difference which compares the difference in False Positive Rates (FPR). There is often confusion about which equity metric is the "best". However, as we can see in their definitions, they are all strongly linked to each other since they attempt to quantify differences between privileged and non-privileged categories.

In terms of performance measurement we need to choose the most appropriate one to show us how successfully a model is trained when predicting outcomes. Depending on the metric we can measure performance using a threshold (threshold) and a qualitative understanding of the error (e.g. Accuracy, F-score, Kappa statistic), 2.) probabilistic understanding of the error (mean absolute error, LogLoss, Brier score) and 3.) on how well the model ranks the predictions (AUC score).

There are several ways to achieve this goal but based on experience and data that we know at least for the data we have to analyze we will use the accuracy metric because it is less complicated than other methods and more understandable and leads to similar results as the other available metrics.

The accuracy metric is defined as the percentage of the number of correct predictions to the total number of predictions made.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

An important criterion for the accuracy result to be valid and reliable is that the number of ties between the classes under consideration should be equal. Precision takes values from 0 to 1 and the closer to unity the value is, the more well trained our model is and the predicted values are closer to the ground truth.

CHAPTER 3.

Data analysis and pre-processing

In this chapter we will analyze the data on which our research was based as well as the preprocessing methods we used to apply machine learning algorithms as well as performance and fairness metrics.

3.1 Data presentation

As mentioned above, our work is concerned with the analysis of databases and the existence of missing values in them, and how fairly the ways in which these values are handled by the algorithmic models we construct give us results. We start by using 3 different datasets from the UCI Machine Learning Repository that contain missing values in some of their variables.

The first database is called Adult Census Data and the race and sex variables define the protected groups for which white men are more favored over other groups. It consists of 48,842 records with 3620 missing values. In addition it has 14 variables, 8 of which are categorical and 6 of which are quantitative and the prediction objective is to determine if a man can earn over 50 thousand income per year based on the variables. The missing values are found in the variables occupation, workclass and native_country which seem to be strongly linked to the protected variables which means that the error resulting from modifying or deleting them will significantly affect the fairness.

The second dataset concerns 891 Titanic passengers (Titanic Dataset) with 12 variables and 866 missing data which are found in the variables: age, fare and embarked. The categorization class represents whether the passengers survived or not and the hypothetical probability that a passenger survived based on their gender and position on the ship is higher for women on the highest floor compared to the rest of the passengers on the ship. Again, variables containing missing values are closely related to the protected (Sex, Class) and as a result the error has a strong impact on fairness as wewill see below.

Finally, we looked at the Autism database which reports on 704 adults with Autism Spectrum Disorder. This database consists of variables related to the respondents (age, gender, ethnicity etc.) and 10 questions of the AQ-10 (Autism Spectrum Quotient) that they answered. The categorization class is related to early diagnosis of autism, where the hypothesized probability of being diagnosed with this condition according to gender and ethnicity is higher for white European women. In addition there are two variables with missing values Age and Relation. The distribution of the missing values in all three aforementioned datasets belongs to the not MCAR category and their handling is done by the analogous method that we will show below.

Table 3. Dataset description

Dataset	Rows	Rows with Nan	Cols with Nan	Protected attribute	Privileged values	Unprivileged values
Adult	48,842	3620	3	Race	White	non-White
				Sex	Male	Female
Titanic	1309	26	3	Class	1	2--3
				Sex	Female	Male
Autism	704	95	2	Ethnicity	White-European	Other
				Gender	Female	Male

3.2 Pre-processing methods

In this subsection we will present the libraries we used as well as the data preprocessing techniques that helped us to shape the data in a way that allowed us to better manipulate it and get more meaningful results.

3.2.1 Libraries

Initially, the tool we used to run our experiments was the Jupyter Notebook platform and the Python programming language. In order to be able to preprocess the data we needed to import some specific libraries through Python to implement the functions we needed. These are:

Sklearn: scikit-learn is a free machine learning library for the Python programming language. It offers a variety of categorization, recursion and clustering algorithms that have helped us create and implement machine learning models. It combines efficiently with Dataframes as well as with the Numpy and Matplotlib libraries that we will see below. It became our main tool for implementing algorithms as well as data processing techniques.

Aif-360: It is an open Python library that helps us to detect and remove bugs in machine learning models. The AI Fairness 360 package includes a broad collection of metrics for datasets and models that test errors and fairness between data classes. It was used to compute fairness metrics, convert to Binary datasets, and preprocess datasets.

Numpy: The Python Numpy library is used to perform any mathematical operation on the code. It helped us in displaying information and types of data as well as calculating various sums.

Pandas: The Pandas library helped us to load our data and process it as it offers a large number of functions to handle data sets with very good performance and productivity.

Matplotlib: Matplotlib is a drawing library for the Python programming language and its numerical math extension NumPy. It is used to create plots and various tables for our measurements.

Seaborn: Finally the Seaborn library uses the Matplotlib library to display diagrams as well as distributions. It was used as it gives more features than Matplotlib and works easily with DataFrames.

3.2.2 Data cleansing

Data cleaning includes techniques that remove duplicate values, redundant variables as well as manage missing values appropriately. In the paper, we removed variables that according to the Pearson test were not directly correlated with either the sensitive variables or the categories we wanted to predict, in order to make the analysis more accurate and have a more manageable dataset.

3.2.3 Data encoding

To be able to apply machine learning algorithms as well as fairness and performance metrics to our data, we need to convert categorical values into numerical values. For this reason we use the LabelEncoder() encoding method. This way we replace the categorical values with numbers and we c a n then process them and apply techniques and algorithms to them. In Figure 3.1 we see the alphanumeric values of the Adult dataset converted to the corresponding integer numeric values.

	education	marital.status	relationship	race	sex	income
0	HS-grad	Widowed	Not-in-family	White	Female	<=50K
1	HS-grad	Widowed	Not-in-family	White	Female	<=50K
2	Some-college	Widowed	Unmarried	Other	Female	<=50K
3	7th-8th	Divorced	Unmarried	White	Female	<=50K
4	Some-college	Separated	Own-child	White	Female	<=50K

	education	marital.status	relationship	race	sex	income
0	11	6	1	1	0	0
1	11	6	1	1	0	0
2	15	6	4	0	0	0
3	5	0	4	1	0	0
4	15	5	3	1	0	0

Figure 3.1 Converting alphanumeric values to integer values in the Adult Dataset

3.2.4 Data scaling

In machine learning algorithms if the values are closer together there are chances that the algorithm will be trained better and faster instead of the already existing values in our data which have large deviations between them which require more time to understand the data and the accuracy will be less. To shape our data to have small deviations between them we used the scaling technique.

More specifically, the MinMaxScaler() function normalizes the inputs of the variables by converting them to the value range [0,1], i.e. the minimum and maximum value of each variable to be 0 and 1 respectively following the following formula.

$$x_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Below is an example of the conversion of the variables in the [0,1] field and their graphical representation in the original and normalized dataset.

Table 3.2.1 Dataset before normalisation

	age	education.num	capital.gain	capital.loss	hours.per.week	workclass	education	marital.status	occupation	relationship	race	sex	native.country	income
0	90	9	0	4356	40	3	11	6	9	1	1	0	38	0
1	82	9	0	4356	18	3	11	6	3	1	1	0	38	0
2	66	10	0	4356	40	3	15	6	9	4	0	0	38	0
3	54	4	0	3900	40	3	5	0	6	4	1	0	38	0
4	41	10	0	3900	40	3	15	5	9	3	1	0	38	0

Table 3.2.2 Dataset after normalisation

	age	education.num	capital.gain	capital.loss	hours.per.week	education	marital.status	relationship	race	sex	income
0	1.000000	0.533333	0.0	1.000000	0.397959	0.733333	1.000000	0.2	1.0	0.0	0.0
1	0.890411	0.533333	0.0	1.000000	0.173469	0.733333	1.000000	0.2	1.0	0.0	0.0
2	0.671233	0.600000	0.0	1.000000	0.397959	1.000000	1.000000	0.8	0.0	0.0	0.0
3	0.506849	0.200000	0.0	0.895317	0.397959	0.333333	0.000000	0.8	1.0	0.0	0.0
4	0.328767	0.600000	0.0	0.895317	0.397959	1.000000	0.833333	0.6	1.0	0.0	0.0

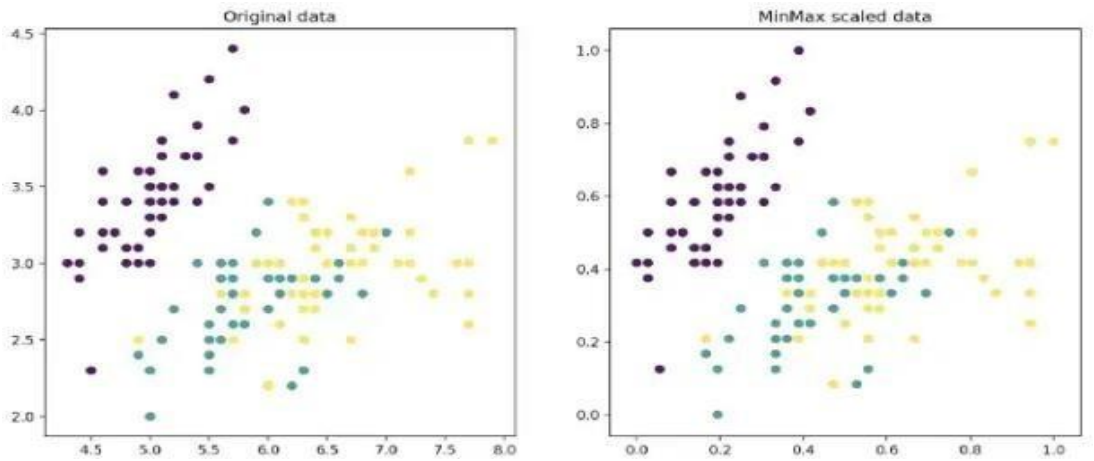


Figure 3.2 Illustration of initial and normalized variables

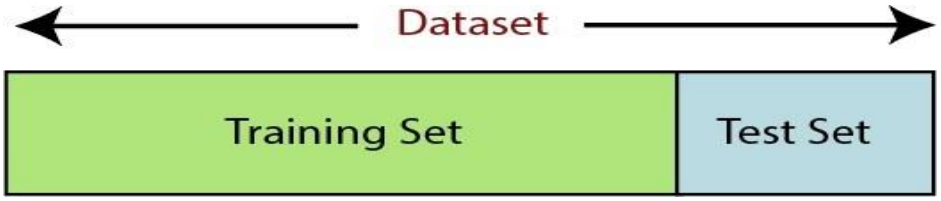
3.2.5 Handling missing values

Our next and most important step during data preprocessing is the handling of missing values. In our analysis all three datasets contain missing values in several categories. We first decided to divide their handling into three categories.

1. The first category contains all data except for rows for which there is at least one missing value in one of its categories. (rows w/o NaN)
2. The second category does not contain columns (categories) which contain even one missing value. (cols w/o NaN)
3. The third category contains all rows and columns of the data and missing values have been replaced by the corresponding average of the category values if the category is a numeric value and by the most frequent value if the category contains imputed cols.

3.2.6 Splitting the Dataset

Another important part of data preprocessing is the separation of the data into train and test set. It is a necessary step in order to improve the performance of our model and "train" it better. Dividing part of it as train test and the rest into test set gives us the ability to simulate the performance of our model on new or unobserved data.



Training Set : The subset of data that the machine learning model uses to train based on data that we know.

Test set : The subset of data we use to test our model by comparing the values it predicted with the values in the subset.

In our analysis using the sklearn library and the train test split() function we split our dataset into 75% for train set and the remaining 25% for test set.

3.2.7 Protected variables

Finally, as mentioned in chapter 2, there are some variables in our data that are considered protected and divide the population into subgroups and should be treated equally in terms of the benefit they can gain. We usually use the term "privileged group" to denote a group that has a systematic advantage over the others. Fairness metrics are based on determining whether decisions are different between protected groups and non-privileged groups. In our analysis in all 3 datasets we have privileged and unprivileged groups which we define by giving them a value of 1 for privileged groups and a value of 0 for unprivileged groups.

CHAPTER 4. Research and Results

This chapter presents the results of the research for each of the three datasets (Adult, Titanic, Autism). Personalised machine learning models have been applied to each of them and performance and fairness metrics have been calculated to help us draw conclusions from the research.

4.1 Calculation of SPD metric

In our first experiment and after we have prepared the data with the pre-processing techniques mentioned above, we will apply the fairness metric SPD.

for the different subsets that we have divided the datasets and privileged groups into. More specifically, we focus on the following three categories of data: a.) Imputed rows i.e. the whole data set with the missing values replaced by the corresponding average value(mean) of the category if it is a numeric value and by the most frequent value(mode) if the category contains categorical variables, b.) the subset of rows that do not contain any missing value (rows w/o NaN) and c.) the subset of columns that do not contain any missing value(cols w/o NaN).

First, for the Adult Set with protected attributes Race and Sex and privileged values White and Male respectively, we know that the privileged class is to have an annual income of more than 50K while the majority (76%) earns less than 50K per year. Applying the SPD metric to the subset of substituted missing values(imputed NaN) we observe that it is positive with a value of 0.103 for privileged value Race and 0.196 for Sex. This implies a greater benefit for the privileged groups, Whites and males over the unprivileged groups non-White females. The positive sign however depends on the classes we consider biased and the favorable class. For this reason a positive or negative result in the SPD metric does not necessarily mean that bias is present and the importance is placed on how close this value is to 0. Based on the definition, the metric values range from 0-1 with a value of 0 meaning that our model is perfectly fair among the protected variables while a value of 1 means that there is a high bias among the sensitive categories . Furthermore, we observe that for the SPD metric in the data without the missing value categories the values are the same, which means that the missing values do not particularly affect the fairness of the dataset.

We observe similar results in the Autism dataset. Gender discrimination does indeed exist against men however it is quite less than that of ethnicity where the bias is quite large and White-European women are more likely to be diagnosed with Autism than men from the rest of the world. In the case of Titanic the values of the SPD metric are positive and quite high which means that the privileged groups i.e. people in the 1ης class and women were favoured to survive compared to people in lower classes and men. Here the bias is high which means some evacuation protocol was followed which favoured women and wealthier passengers clearly. Looking at Table 4.1 and comparing the average of the SPD metrics of all three categories we see a noticeable difference with fairer results for the imputed rows category compared to the other two categories. This observation has only to do with the data, we cannot yet talk about the error and bias that a machine learning model can create, we will see later. However from what we can see, replacing missing values contributes more to fairness than deleting them in any way. Next we will see how missing values affect machine learning models and the fairness metrics of the trained models relative to the original fairness metrics.

Table 4.1: Description of Dataset and calculation of initial SPD metrics

Dataset	Rows	Rows with Nan	Cols with Nan	Protected attribute	Privileged values	Unprivileged values	c+	Majority	SPD (imputed rows)	SPD (cols w/o Nan)	SPD (rows w/o Nan)
Adult	48,842	3620	3	Race	White	non-White	> $50K	<= $50K (76 %)	**0.103**	0.103	0.105
				Sex	Male	Female	> $50K	<= $50K (76 %)	**0.196**	0.196	0.201
Titanic	1309	26	3	Class	1	2--3	1 (Survived)	0 (died) (62 %)	**0.16**	0.20	0.15
				Sex	Female	Male	1 (Survived)	0 (died) (62 %)	**0.50**	0.55	0.51
Autism	704	95	2	Ethnicity	White-European	Other	Yes	No (73 %)	**0.24**	0.27	0.30
				Gender	Female	Male	Yes	No (73 %)	**0.07**	0.08	0.08

4.2 Application of classifier AND SPD calculation

Most algorithms in the Python programming language cannot handle missing values, so we have to delete or replace them somehow. We follow the pattern of the previous measurements in terms of separating categories for missing values.

For a metric of fairness we will continue to use the SPD. We split the dataset into 75% for training and 25% for testing and applied it to all three categories. To select the appropriate machine learning algorithm we created a function that selects among five algorithms [(LogisticRegression, SVC, KNeighborsClassifier, RandomForestClassifier, GaussianNB] using the Accuracy score performance metric as a benchmark and selects the one with the best score to train our data. To achieve more fair performance results, the function uses the Cross Validation() technique for 10 folds and at the end it finds the average accuracy score for each algorithm.

Based on this function, we choose the appropriate algorithm to train our data and apply the SPD metric to compare the new results with the original ones.

Table 4.2: SPD results for the best classifier based on Accuracy

Dataset	Protected attribute	Algorithm	Acc (imputed rows)	SPD (imputed rows)	Algorithm	Acc (cols w/o Nan)	SPD (cols w/o Nan)	Algorithm	Acc (rows w/o Nan)	SPD (rows w/o Nan)
Adult	Race	LR	0.88	0.09(↓)	KNN	0.87	0.1(↓)	KNN	0.82	0.13(↑)
	Sex	LR	0.88	0.16 (↓)	KNN	0.87	0.17(↓)	KNN	0.82	0.19(↓)
Titanic	Class	RF	0.83	0.41(↑)	RF	0.81	0.5(↑)	LR	0.78	0.43(↑)
	Sex	RF	0.83	0.6(↑)	RF	0.81	0.7(↑)	LR	0.78	0.8(↑)
Autism	Ethnicity	RF	0.97	0.20(↓)	RF	0.95	0.26(↓)	SVM	0.97	0.24(↓)
	Gender	RF	0.97	0.07(↓)	RF	0.95	0.13(↑)	SVM	0.97	0.3(↑)

Looking at Table 4.2 for the Adult Dataset we see that the LR and KNN algorithms give an accuracy of 0.86 and 0.87 respectively meaning that the predictions are very good compared to the actual data for the imputed rows and cols w/o NaN values.

The SPD metrics are even better than the first analysis and closer to zero for the imputed rows and cols w/o NaN cases, while for the rows w/o NaN category the metric for the race variable increases while for the sex variable it decreases significantly. The largest decrease and the best fairness value is given by the imputed rows category. In the Autism set we find an excellent performance rate which shows us that for this dataset the models of the RF and SVM algorithms have been trained almost perfectly. In the case of rows w/o NaN and cols w/o NaN the bias decreases or increases depending on the protected variable. But in the case of imputedrows we find once again that the results are more fair for the two protected variables. Finally, for the Titanic Dataset the results are different from the previous two datasets. In all three categories the bias is higher and the machine learning algorithms do not give us high accuracy rates which means that they also affect the error in our dataset. However in this case too if we had to choose a way to handle the missing values to get the lowest bias then this would again be the case to replace them. Concluding this analysis we come to the conclusion that in the cases where the model learns from the imputed lines the accuracy is very good and the model is more fair when I have little data whereas when I have more data the accuracy decreases while the bias increases. Regarding the cases of cols w/o NaN and rows w/o we observe that due to dataset reduction and deletion of useful information from the data the accuracy decreases and the models become more polarized(biased) except in some cases.

4.3 Replacing missing values with a prediction

Since so far our results have shown that replacing the missing values gives us fairer results in this last analysis we will try to predict the missing values with a machine learning algorithm and apply the SPD fairness metric to see if our model will be even fairer or not. In order to be able to predict the missing values we used the *HistGradientBoostingClassifier* algorithm, a recently built and much improved classification algorithm which is the only one that allows the use of missing values during training in the Python language. Its mode of operation is similar to the decision tree algorithm, however, its peculiarity which makes it extremely fast and accurate is that successive decision trees are trained which each time correct the error of the previous decision tree.

We place the missing values in a new category and replace it with the category we want to predict and create a test set with the rows containing the missing values while the train set contains the existing values.

Table 4.3 : test set for the category *'native country'* which contains missing values

age	workclass	education.num	capital.gain	capital.loss	hours.per.week	education	marital.status	occupation	relationship	race	sex	income	native.country
41	3.0	10	0	3004	60	15	4	2.0	4	1	1	1	NaN
22	3.0	12	0	2824	40	7	4	5.0	1	0	1	1	NaN
60	4.0	11	0	2415	70	8	2	6.0	0	1	1	1	NaN
39	5.0	15	0	2415	50	14	2	9.0	0	1	1	1	NaN
43	0.0	15	0	2415	55	14	2	9.0	0	0	1	1	NaN

After, I train my model on the train set and apply it to the test set in which I predict the missing values. I do the same work for all categories containing missing values until I replace them all.

To check how fair the values predicted by the model are, I compare their distributions with the already observed variables to see if there is any overlap. Looking at the distributions of the Adult Dataset we see that the distributions of the predicted categories(in red) take values close to the already known ones(in blue) and there is a large overlap of values, which means that our new dataset is fair and not biased.

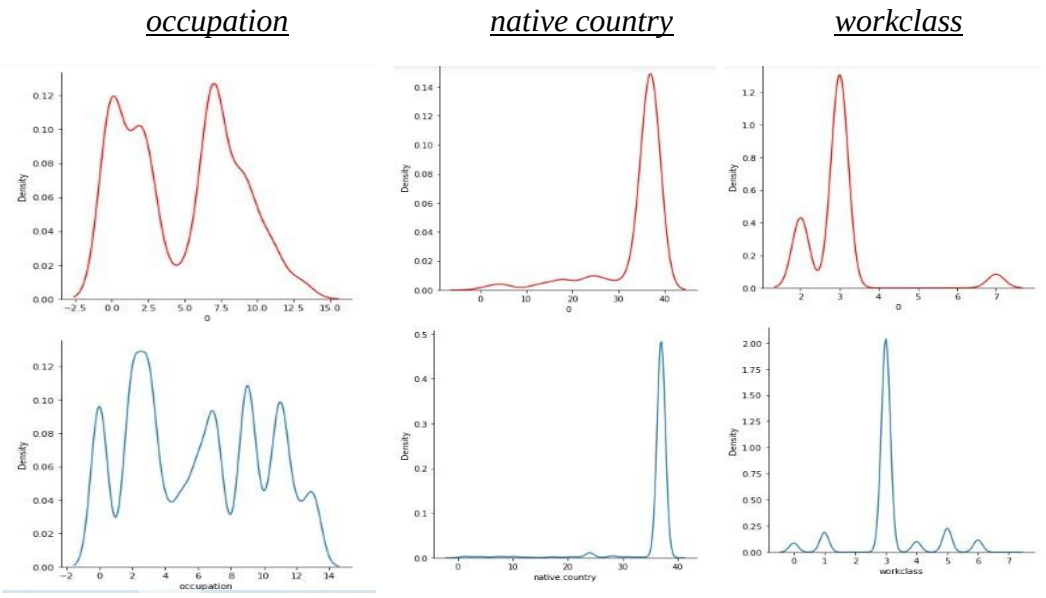


Figure 4.3 Distributions of categories with missing values for the Adult Dataset

Regarding the Autism Dataset we have two missing values that we replace. The observed values of the relation variable range from 1.5 to 2.5 in a few cases and from 3 to 4.5 in most of the cases. The predicted values range at similar levels in the intervals (1.5-2.5), (2.5-3.5) and largely in (3.5-4.5) very close to the ground truth of the values for these variables. Similarly the values of the second variable with missing values ('Age') show a large overlap as well as the predicted values follow a similar distribution to the observed values in the (-10.50) range as shown in the figure below.

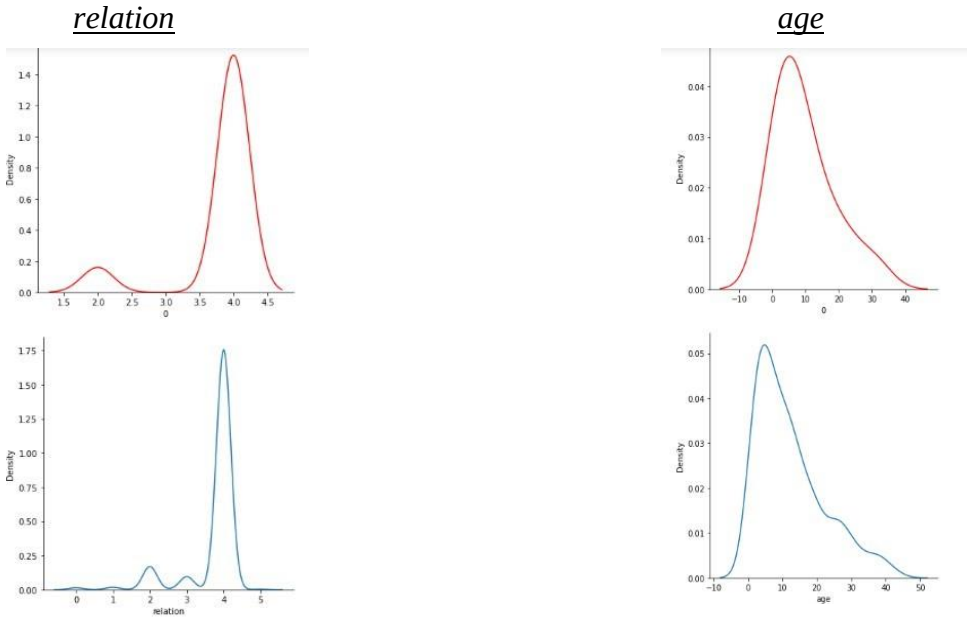


Figure 4.3.1 : Distributions of categories with missing values for the Autism Dataset

In the Titanic dataset we have three variables with missing values (Cabin, Age, Embarked).

In the case of the variable Embarked we have only 2 variables and it is difficult to train the model and as a result there is a bias in the model as shown in the distributions which are not identical but have a small overlap in the field (-0.5, 2.5). As for the variable 'Cabin' the number of missing values is large and helps the model to train correctly and generate fair results. The distribution of predicted values shows overlap in (0-150) with a large fraction of values in (50,150) which i s also the case for the observed variables except that in -(50,50) the predicted values either have low overlap or are not even certain. Finally, the 'Age' variable seems to show almost complete overlap with the observed variables which further strengthens the fairness of the model.

Embarked *Cabin* *Age*

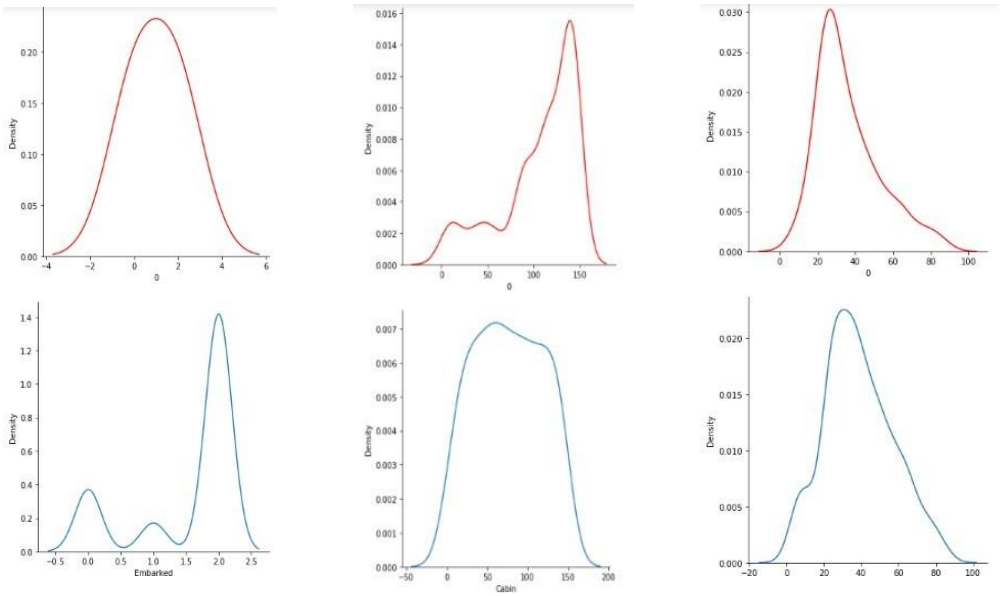


Figure 4.3.2 : Distributions of categories with missing values for the Titanic Dataset

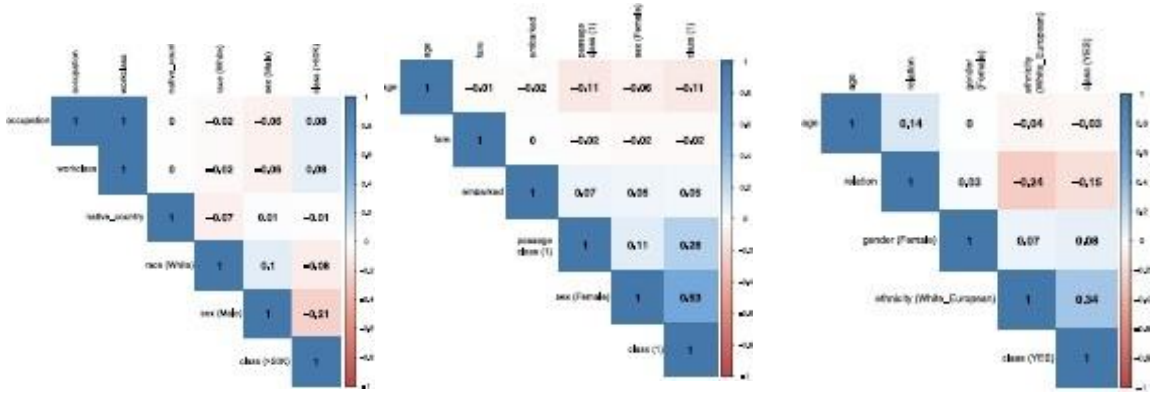
4.3.1 Implementation of SPD metrics in the new Dataset

At this point we will apply the SPD metric to the new dataset we created earlier and compare it to the metric for imputed rows which has so far proven to be the fairest technique for dealing with missing values. We will then also apply a trained machine learning model to see how it behaves towards fairness and performance.

Table 4.3.2: SPD metric for the classifier with the best accuracy score

Dataset	Algorithm	Protected attribute	SPD before training	Accuracy	SPD after training
Adult	HGB	Race	0.103	0.88	0.09 (↓)
	HGB	Sex	0.18	0.88	0.15 (↓)
Titanic	HGB	Class	0.16	0.87	0.38(↑)
	HGB	Sex	0.50	0.87	0.55(↑)
Autism	NB	Gender	0.07	0.95	0.12(↑)
	NB	Ethnicity	0.27	0.95	0.12(↓)

Looking at Table 4.3.2 we see that compared to the first analysis where we applied the SPD metric to our original datasets the values of the metric either remain the same (Adult, Autism) or have very small deviations (Titanic). We conclude that the categories with missing values are not directly related to the protected variables to affect the fairness of the results. We confirm this through the Pearson test in which we show the correlations between variables and find that the variables with missing values have very little correlation with the protected variables in the Adult and Autism datasets and a little more in Titanic.



1. Adult 2. Titanic 3. Autism

Figure 4.3.3 correlation matrices for the 3 datasets

Finally, we tried to train a machine learning model to see how it behaves on the new data and compare the fairness and performance with the values of the previous corresponding analysis with imputed rows.

We worked in a similar way to the previous analysis in terms of selecting the appropriate algorithm based on its Accuracy score and using the 10-fold cross validation technique. This time we expanded the number of algorithms and added the HistGradientBoostingClassifier algorithm which we used to predict the missing values.

Going back to Table 4.3.2 and observing the results we see that in two of the three cases the HistGradientBoostingClassifier algorithm performed better than all the others while in the case of the Autism dataset NaiveBayes was chosen. Let us examine what happens with fairness though. The Adult bundle set seems to have become even fairer with respect to replacing the missing values with mean and mode values since the SPD metric was reduced to a small percentage however it is sufficient to characterize our model as the most fair compared to the previous ones. In the case of individuals with Autism we observe an increase in bias with respect to gender and a decrease with respect to ethnicity which means that our model becomes more fair with respect to the racial discrimination of individuals who may develop Autism. Finally, in the Titanic data the fairness metric follows the opposite trend and increases in both protected variables with a satisfactory 87% accuracy rate. However, again in compared to imputed rows are more fair results. We conclude that missing value prediction gives even fairer results than any other missing value replacement technique tested in this analysis.

CHAPTER 5.

Conclusions AND future research

5.1 Conclusions

We presented an analysis of the relationship between missing values and fairness. We analyzed the reasons for the existence of incomplete prices and the reasons for unjust decisions and the ways we can manage these prices. The result was to find that replacing missing values with some technique is much better in terms of fairness of results than eliminating them from the dataset. Unfortunately in many studies they are often ignored and deleted and thus harm the sample and falsify the results however if handled correctly they can provide very useful information. In more detail we concluded the following:

1. While missing values and fairness are directly related, in our datasets the relationships between protected variables and missing value variables are not strong so the effect is driven by other proxy variables.

2. Subsets that delete the rows or columns of missing values may have satisfactory accuracy rates when applied to machine learning models however a large sample of information is lost and the model becomes more unfair to the protected variables.

3. Gap-filling techniques can achieve a balance of accuracy and fairness and can significantly contribute to obtaining fairer results than other techniques.

4. The technique of predicting missing values through an algorithmic categorization model formulates even fairer results when trained with appropriate machine learning models. Finally, we found that in addition to the already known advantages of missing value imputation techniques in accuracy, they must now be used and developed for the sake of fairness.

5.2 Future research

Based on our analysis and the conclusion that if the missing values are replaced in some way they produce better and fairer results, other ways of replacing values could be tested by applying new algorithms and techniques that would under certain circumstances improve fairness and accuracy even more.

CHAPTER 6.

Bibliography

1. Bellamy, Rachel K. E., et al. *AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias*. arXiv, 3 Oct. 2018. *arXiv.org*, https://doi.org/10.48550/arXiv.1810.01943.
2. Besse, Philippe, et al. *A Survey of Bias in Machine Learning through the Prism of Statistical Parity for the Adult Data Set*. arXiv, 6 Apr. 2020. *arXiv.org*, https://doi.org/10.48550/arXiv.2003.14263.
3. 'Fairness and Missing Values' *DeepAI*, 29 May 2019, https://deepai.org/publication/fairness-and-missing-values.
4. Fouad, Khaled M., et al. 'Advanced Methods for Missing Values Imputation Based on Similarity Learning'. *PeerJ Computer Science*, vol. 7, July 2021, p. e619. *PubMed Central*, https://doi.org/10.7717/peerj-cs.619.
5. Hort, Max, and Federica Sarro. 'Privileged and Unprivileged Groups: An Empirical Study on the Impact of the Age Attribute on Fairness' *Proceedings of the 2nd International Workshop on Equitable Data and Technology*, Association for Computing Machinery, 2022, pp. 17-24. *ACM Digital Library*, https://doi.org/10.1145/3524491.3527308.
6. *Introduction to Data Mining.* https://www.users.cse.umn.edu/~kumar001/dmbook/index.php. Accessed 3 Feb. 2023.
7. Jäger, Sebastian, et al. 'A Benchmark for Data Imputation Methods' *Frontiers in Big Data*, vol. 4, 2021. *frontiers*, https://www.frontiersin.org/articles/10.3389/fdata.2021.693674.
8. Kang, Hyun. 'The Prevention and Handling of the Missing Data'." *Korean*

Journal of Anesthesiology, vol. 64, no. 5, May 2013, pp. 402-06.

Central, https://doi.org/10.4097/kjae.2013.64.5.402.

9. Kumar, Ajitesh. 'Fairness Metrics - ML Model Sensitivity for Bias Detection'. *Data Analytics*, 31 Oct. 2018, https://vitalflux.com/fairness-metrics-ml-model-sensitivity-bias-detection/.
10. Kumar, Satyam. '7 Ways to Handle Missing Values in Machine Learning'. *Medium*, 28 Sept. 2021, https://towardsdatascience.com/7-ways-to-handle-missing-values-in-machine-learning-1a6326adf79e.
11. 'ML Algorithm That Natively Supports Missing Values'. *medium*, 18 Jan. 2022, https://towardsdatascience.com/ml-algorithm-that-natively-supports-missing-values-40b42559c1ec.
12. Leurent, Baptiste, et al. 'Sensitivity Analysis for Not-at-Random Missing Data in Trial-Based Cost-Effectiveness Analysis: A Tutorial'. *Pharmacoeconomics*, vol. 36, no. 8, 2018, pp. 889-901. *PubMed Central*, https://doi.org/10.1007/s40273-018-0650-5.
13. Mattu, Julia Angwin, Jeff Larson,Lauren Kirchner,Surya. 'Machine Bias'. *ProPublica*, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. accessed 3 Feb. 2023.
14. *Mining of Massive Datasets*. http://mmds.org/. Accessed 3 Feb. 2023.
15. Oneto, Luca, and Silvia Chiappa. *fairness in machine learning*. 2020, pp. 155-96. *arXiv.org*, http://arxiv.org/abs/2012.15816.
16. Pesach, Dana, and Erez Shmueli.'A Review on Fairness in Machine Learning'. *ACM Computing Surveys*, vol. 55, no. 3, Feb. 2022, p. 51:1-51:44. *March 2023*, https://doi.org/10.1145/3494672.
17. 'Improving Fairness of Artificial Intelligence Algorithms in Privileged-Group

Selection Bias Data Settings'.*Expert Systems with Applications: An International Journal*, vol. 185, no. C, Dec. 2021. *Dec 2021*, https://doi.org/10.1016/j.eswa.2021.115667.

18. Stewart, Matthew. 'Programming Fairness in Algorithms'. *Medium*, 29 July 2020, https://towardsdatascience.com/programming-fairness-in-algorithms-4943a13dd9f8.
19. Su, Thom. 'Just How Robust Are Tree-Based Classifiers In Handling Missing Data As Is?' *Medium*, 5 Oct. 2019, https://towardsdatascience.com/just-how-robust-are-tree-based-classifiers-in-handling-missing-data-as-is- 67d62adaf40c.
20. Trevethan, Robert. 'Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice' *Frontiers in Public Health*, vol. 5, 2017. *Frontiers*, https://www.frontiersin.org/articles/10.3389/fpubh.2017.00307.
21. Wang, Yanchen, and Lisa Singh. 'Analyzing the Impact of Missing Values and Selection Bias on Fairness'. *International Journal of Data Science and Analytics*, vol. 12, no. 2, Aug. 2021, pp. 101-19. *Springer Link*, https://doi.org/10.1007/s41060-021-00259-z.
22. *When and How Do Fairness-Accuracy Trade-Offs Occur?* https://wearepal.ai/blog/when-and-how-do-fairness-accuracy-trade-offs-occur. Accessed 3 Feb. 2023.
23. Yu, Zhe, et al. *Fairer Machine Learning Software on Multiple Sensitive Attributes With Data Preprocessing*. arXiv, 21 June 2022. *arXiv.org*, https://doi.org/10.48550/arXiv.2107.08310.
24. Zhang, Yiliang, and Qi Long. *Assessing Fairness in the Presence of Missing Data*. arXiv, 7 Dec. 2021. *arXiv.org*, https://doi.org/10.48550/arXiv.2112.0489

25. Lee, J. *"Missing Data Imputation in Clinical Trials Using Recurrent Neural Network Facilitated by Clustering and Oversampling." ResearchGate, 22 June 2021*,https://www.researchgate.net/publication/359146814_Missing_data_imputation_in_clinical_trials_using_recurrent_neural_network_facilitated_by_clustering_and_oversampling

26. Allison, Paul D. *"Handling Missing Data by Maximum Likelihood." Statistical* Horizons, *Haverford, PA, USA, n.d.*, https://statisticalhorizons.com/wp-content/uploads/2022/01/MissingDataByML.pdf.