UNIVERSITÀ DEGLI STUDI DELL'AQUILA

DISIM
Dipartimento di Ingegneria e Scienze dell'Informazione e Matematica

BIAS E FAIRNESS NEL MACHINE LEARNING

Giordano d'Aloisio

Il problema del bias

- Nel 2018 una ricercatrice dell'IMT stava studiando i sistemi di riconoscimento facciale di Amazon
- Questi sistemi non erano in grado di riconoscere il suo volto
- All'inizio pensò fosse un errore del sistema
- Ma poi indossando una maschera bianca notò che il sistema era in grado di riconoscerla

Quindi il sistema non era in grado di riconoscere donne non bianche

Un altro esempio...

- Diversi giudici negli Stati Uniti hanno utilizzato per anni un sistema di intelligenza artificiale per decidere se liberare o meno un condannato
- Questo sistema prevedeva la possibilità che un condannato avrebbe ricommesso un crimine nei prossimi due anni
- Dopo uno studio attento del sistema è stato dimostrato che questo algoritmo date due persone con stesse caratteristiche, ma di etnia diversa, forniva una minore probabilità di recidiva alla persona bianca

Quindi il sistema favoriva sistematicamente le persone bianche solo in base alla loro etnia

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances — which belonged to a 6-year-old boy — a woman came running after them saying, "That's my kid's stuff." Borden and her friend immediately dropped the bike and scooter and walked away.

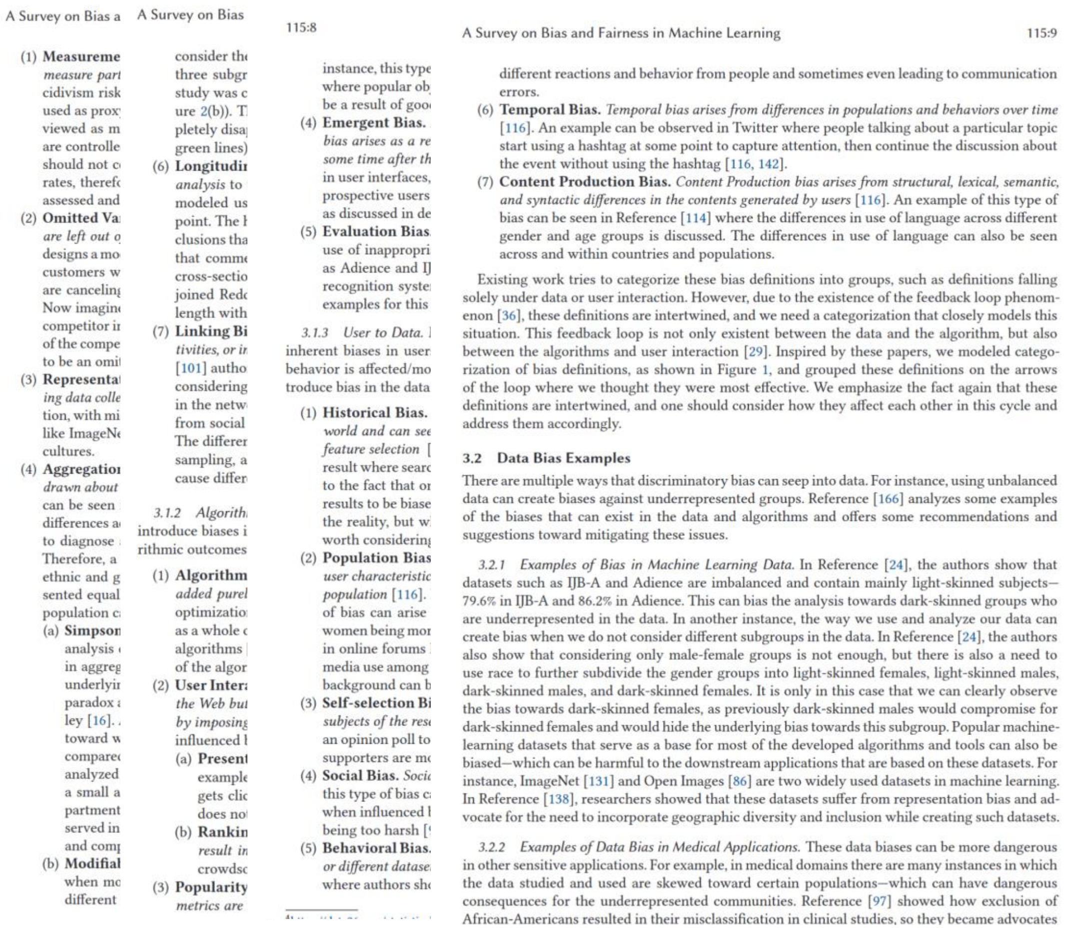
But it was too late — a neighbor who witnessed the heist had already called the police. Borden and her friend were arrested and charged with burglary and petty theft for the items, which were valued at a total of $80.

Definiamo meglio il concetto di Bias e Fairness

- **BIAS:** sistematico favoritismo o discriminazione di individui da parte di un algoritmo sulla base di alcune loro caratteristiche (esempio il sesso o l’etnia)

- **FAIRNESS:** assenza di discriminazione o favoritismo da parte di un algoritmo

Il concetto di bias non è così semplice…



A Survey on Bias and Fairness in Machine Learning 115:9

different reactions and behavior from people and sometimes even leading to communication errors.

(6) **Temporal Bias.** *Temporal bias arises from differences in populations and behaviors over time* [116]. An example can be observed in Twitter where people talking about a particular topic start using a hashtag at some point to capture attention, then continue the discussion about the event without using the hashtag [116, 142].

(7) **Content Production Bias.** *Content Production bias arises from structural, lexical, semantic, and syntactic differences in the contents generated by users* [116]. An example of this type of bias can be seen in Reference [114] where the differences in use of language across different gender and age groups is discussed. The differences in use of language can also be seen across and within countries and populations.

Existing work tries to categorize these bias definitions into groups, such as definitions falling solely under data or user interaction. However, due to the existence of the feedback loop phenomenon [36], these definitions are intertwined, and we need a categorization that closely models this situation. This feedback loop is not only existent between the data and the algorithm, but also between the algorithms and user interaction [29]. Inspired by these papers, we modeled categorization of bias definitions, as shown in Figure 1, and grouped these definitions on the arrows of the loop where we thought they were most effective. We emphasize the fact again that these definitions are intertwined, and one should consider how they affect each other in this cycle and address them accordingly.

3.2 Data Bias Examples

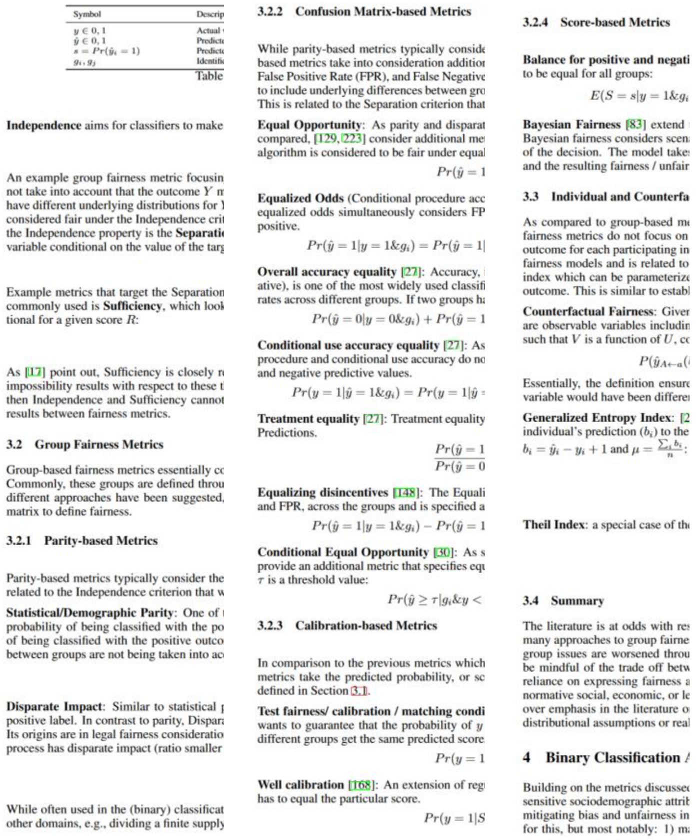
There are multiple ways that discriminatory bias can seep into data. For instance, using unbalanced data can create biases against underrepresented groups. Reference [166] analyzes some examples of the biases that can exist in the data and algorithms and offers some recommendations and suggestions toward mitigating these issues.

3.2.1 Examples of Bias in Machine Learning Data. In Reference [24], the authors show that datasets such as IJB-A and Adience are imbalanced and contain mainly light-skinned subjects—79.6% in IJB-A and 86.2% in Adience. This can bias the analysis towards dark-skinned groups who are underrepresented in the data. In another instance, the way we use and analyze our data can create bias when we do not consider different subgroups in the data. In Reference [24], the authors also show that considering only male-female groups is not enough, but there is also a need to use race to further subdivide the gender groups into light-skinned females, light-skinned males, dark-skinned males, and dark-skinned females. It is only in this case that we can clearly observe the bias towards dark-skinned females, as previously dark-skinned males would compromise for dark-skinned females and would hide the underlying bias towards this subgroup. Popular machine-learning datasets that serve as a base for most of the developed algorithms and tools can also be biased—which can be harmful to the downstream applications that are based on these datasets. For instance, ImageNet [131] and Open Images [86] are two widely used datasets in machine learning. In Reference [138], researchers showed that these datasets suffer from representation bias and advocate for the need to incorporate geographic diversity and inclusion while creating such datasets.

3.2.2 Examples of Data Bias in Medical Applications. These data biases can be more dangerous in other sensitive applications. For example, in medical domains there are many instances in which the data studied and used are skewed toward certain populations—which can have dangerous consequences for the underrepresented communities. Reference [97] showed how exclusion of African-Americans resulted in their misclassification in clinical studies, so they became advocates

- Più di 23 definizioni di bias esistono oggi in letteratura

Il bias e la fairness possono essere misurati…



3.2.2 Confusion Matrix-based Metrics

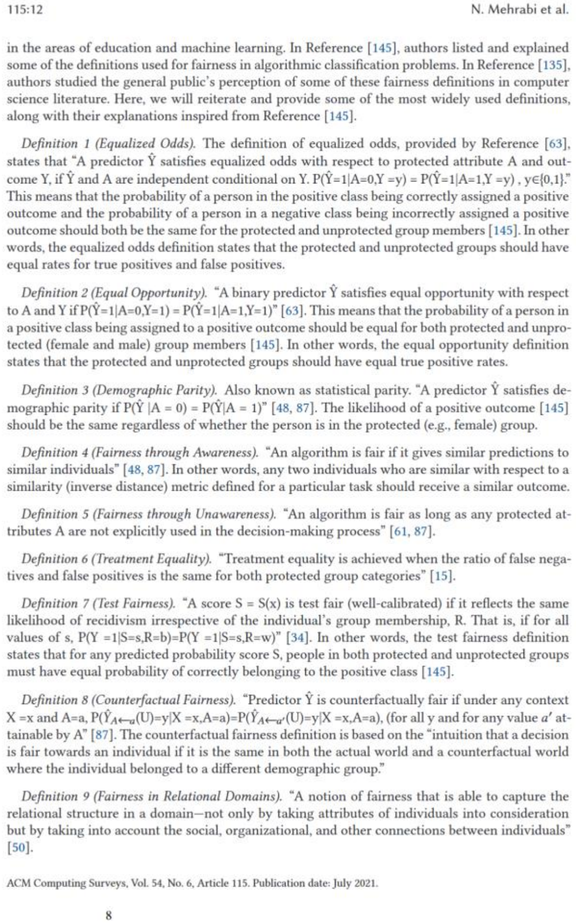
3.2.4 Score-based Metrics

3.2 Group Fairness Metrics

3.2.1 Parity-based Metrics

3.2.3 Calibration-based Metrics

3.4 Summary



115:12 N. Mehrabi et al.

in the areas of education and machine learning. In Reference [145], authors listed and explained some of the definitions used for fairness in algorithmic classification problems. In Reference [135], authors studied the general public's perception of some of these fairness definitions in computer science literature. Here, we will reiterate and provide some of the most widely used definitions, along with their explanations inspired from Reference [145].

Definition 1 (Equalized Odds). The definition of equalized odds, provided by Reference [63], states that "A predictor $\hat{Y}$ satisfies equalized odds with respect to protected attribute A and outcome Y, if $\hat{Y}$ and A are independent conditional on Y. $P(\hat{Y}=1|A=0,Y=y) = P(\hat{Y}=1|A=1,Y=y)$, $y\in\{0,1\}$." This means that the probability of a person in the positive class being correctly assigned a positive outcome and the probability of a person in a negative class being incorrectly assigned a positive outcome should both be the same for the protected and unprotected group members [145]. In other words, the equalized odds definition states that the protected and unprotected groups should have equal rates for true positives and false positives.

Definition 2 (Equal Opportunity). "A binary predictor $\hat{Y}$ satisfies equal opportunity with respect to A and Y if $P(\hat{Y}=1|A=0,Y=1) = P(\hat{Y}=1|A=1,Y=1)$" [63]. This means that the probability of a person in a positive class being assigned to a positive outcome should be equal for both protected and unprotected (female and male) group members [145]. In other words, the equal opportunity definition states that the protected and unprotected groups should have equal true positive rates.

Definition 3 (Demographic Parity). Also known as statistical parity. "A predictor $\hat{Y}$ satisfies demographic parity if $P(\hat{Y}|A=0) = P(\hat{Y}|A=1)$" [48, 87]. The likelihood of a positive outcome [145] should be the same regardless of whether the person is in the protected (e.g., female) group.

Definition 4 (Fairness through Awareness). "An algorithm is fair if it gives similar predictions to similar individuals" [48, 87]. In other words, any two individuals who are similar with respect to a similarity (inverse distance) metric defined for a particular task should receive a similar outcome.

Definition 5 (Fairness through Unawareness). "An algorithm is fair as long as any protected attributes A are not explicitly used in the decision-making process" [61, 87].

Definition 6 (Treatment Equality). "Treatment equality is achieved when the ratio of false negatives and false positives is the same for both protected group categories" [15].

Definition 7 (Test Fairness). "A score $S = S(x)$ is test fair (well-calibrated) if it reflects the same likelihood of recidivism irrespective of the individual's group membership, R. That is, if for all values of s, $P(Y=1|S=s,R=b)=P(Y=1|S=s,R=w)$" [34]. In other words, the test fairness definition states that for any predicted probability score S, people in both protected and unprotected groups must have equal probability of correctly belonging to the positive class [145].

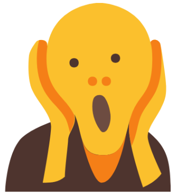
Definition 8 (Counterfactual Fairness). "Predictor $\hat{Y}$ is counterfactually fair if under any context $X=x$ and $A=a$, $P(\hat{Y}_{A\leftarrow a}(U)=y|X=x,A=a)=P(\hat{Y}_{A\leftarrow a'}(U)=y|X=x,A=a)$, (for all y and for any value a' attainable by A" [87]. The counterfactual fairness definition is based on the "intuition that a decision is fair towards an individual if it is the same in both the actual world and a counterfactual world where the individual belonged to a different demographic group."

Definition 9 (Fairness in Relational Domains). "A notion of fairness that is able to capture the relational structure in a domain—not only by taking attributes of individuals into consideration but by taking into account the social, organizational, and other connections between individuals" [50].

ACM Computing Surveys, Vol. 54, No. 6, Article 115. Publication date: July 2021.

8

- Più di 15 metriche di fairness diverse esistono oggi in letteratura



Il bias può essere mitigato...

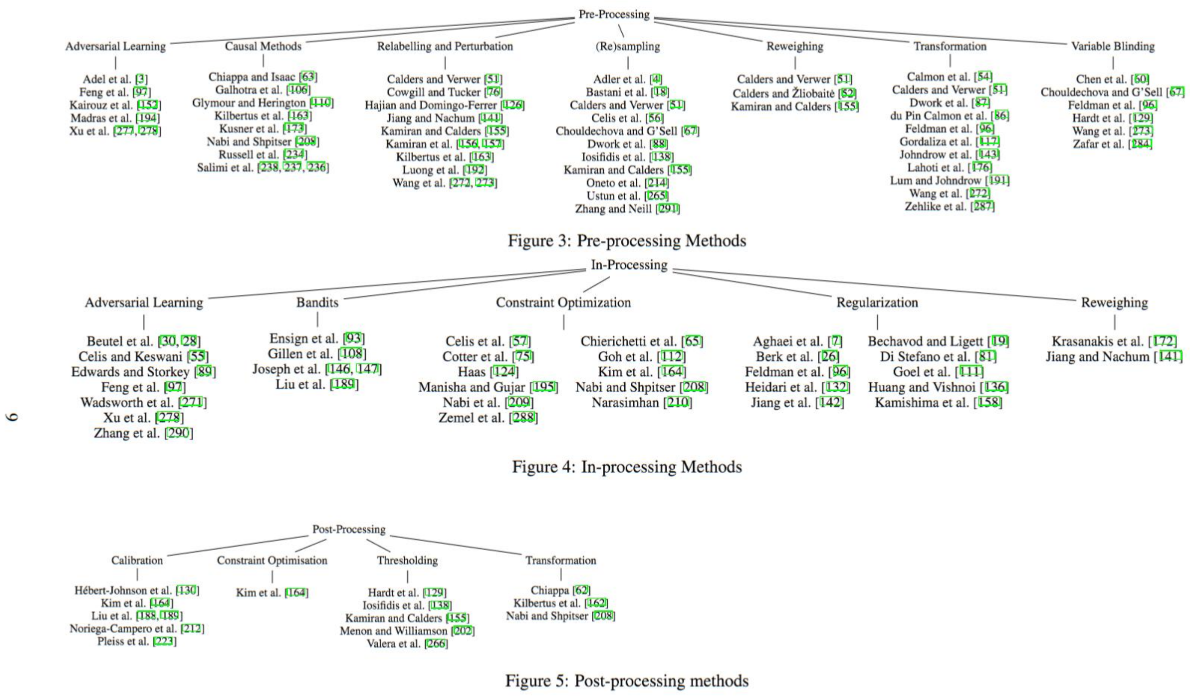
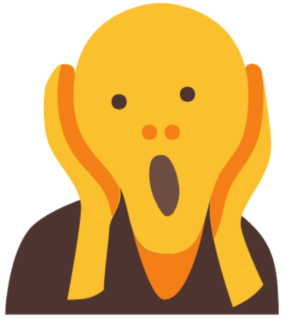


Figure 3: Pre-processing Methods

Figure 4: In-processing Methods

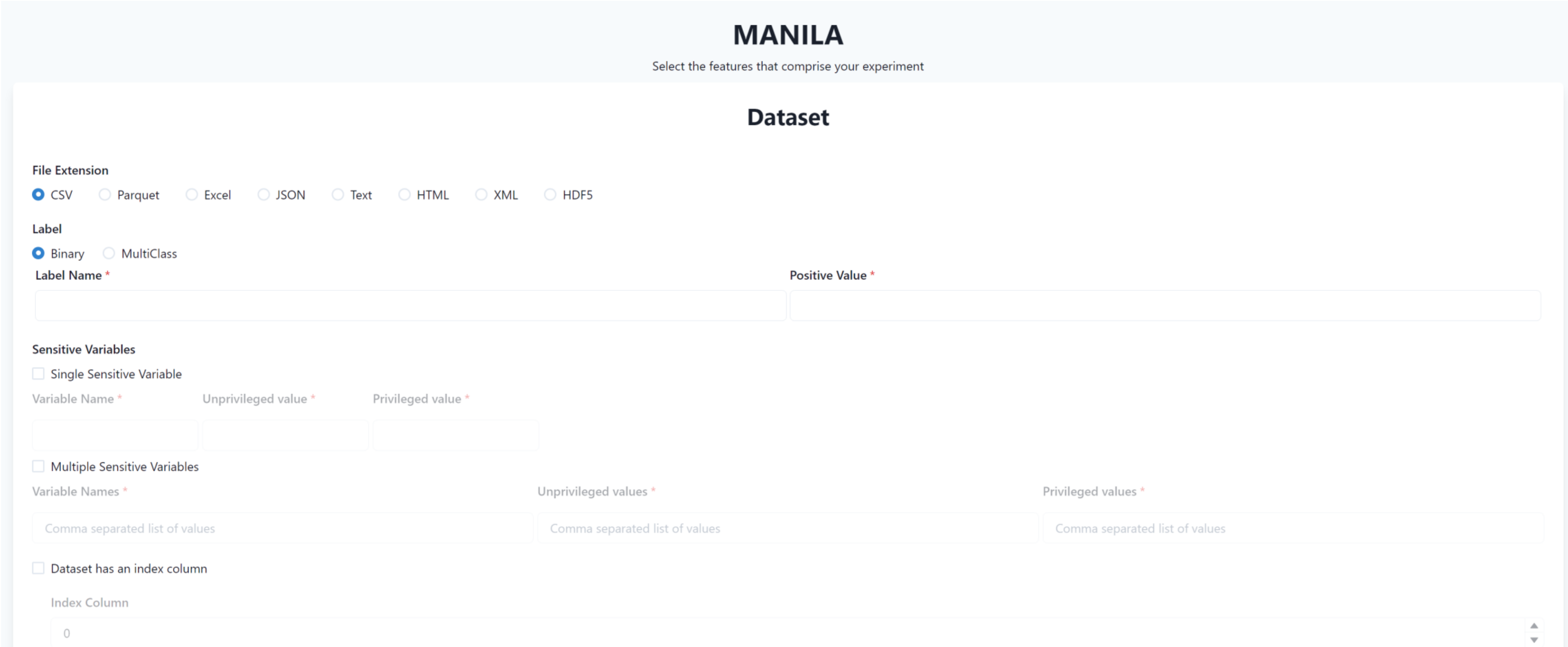
Figure 5: Post-processing methods



Come possiamo migliorare questa situazione?

Guidando l'utente nella selezione di definizioni, metriche e metodi appropriati in base al contesto

MANILA in SoBigData RI



DOMANDE?🤔