

Detecting discriminatory risk through data annotation based on Bayesian inferences

Original

Detecting discriminatory risk through data annotation based on Bayesian inferences / Beretta, Elena; Vetro', Antonio; Lepri, Bruno; De Martin, Juan Carlos. - STAMPA. - (2021), pp. 794-804. (Intervento presentato al convegno ACM Conference on Fairness, Accountability, and Transparency 2021 (ACM FAccT 2021) tenutosi a Online nel March 3-10, 2021) [10.1145/3442188.3445940].

Availability:

This version is available at: 11583/2858625 since: 2021-05-08T13:11:01Z

Publisher:

Association for Computing Machinery

Published

DOI:10.1145/3442188.3445940

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

ACM postprint/Author's Accepted Manuscript

(Article begins on next page)

Detecting discriminatory risk through data annotation based on Bayesian inferences

Elena Beretta

elena.beretta@polito.it

Nexa Center for Internet & Society
Department of Control and Computer Engineering
Politecnico di Torino
Turin, Italy
Fondazione Bruno Kessler
Trento, Italy

Bruno Lepri

lepri@fbk.eu

Fondazione Bruno Kessler
Trento, Italy

Antonio Vetrò

antonio.vetro@polito.it

Nexa Center for Internet & Society
Department of Control and Computer Engineering
Politecnico di Torino
Turin, Italy

Juan Carlos De Martin

demartin@polito.it

Nexa Center for Internet & Society
Department of Control and Computer Engineering
Politecnico di Torino
Turin, Italy

ABSTRACT

Thanks to the increasing growth of computational power and data availability, the research in machine learning has advanced with tremendous rapidity. Nowadays, the majority of automatic decision making systems are based on data. However, it is well known that machine learning systems can present problematic results if they are built on partial or incomplete data. In fact, in recent years several studies have found a convergence of issues related to the ethics and transparency of these systems in the process of data collection and how they are recorded. Although the process of rigorous data collection and analysis is fundamental in the model design, this step is still largely overlooked by the machine learning community. For this reason, we propose a method of data annotation based on Bayesian statistical inference that aims to warn about the risk of discriminatory results of a given data set. In particular, our method aims to deepen knowledge and promote awareness about the sampling practices employed to create the training set, highlighting that the probability of success or failure conditioned to a minority membership is given by the structure of the data available. We empirically test our system on three datasets commonly accessed by the machine learning community and we investigate the risk of racial discrimination.

CCS CONCEPTS

• **Human-centered computing** → **Visualization application domains**.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference'17, July 2017, Washington, DC, USA

© 2021 Association for Computing Machinery.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...\$15.00

<https://doi.org/10.1145/nnnnnnnn.nnnnnnn>

KEYWORDS

human annotation, data ethics, race discrimination, sampling bias, data labeling, machine learning

ACM Reference Format:

Elena Beretta, Antonio Vetrò, Bruno Lepri, and Juan Carlos De Martin. 2021. Detecting discriminatory risk through data annotation based on Bayesian inferences. In *Proceedings of ACM Conference (Conference'17)*. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/nnnnnnnn.nnnnnnn>

1 INTRODUCTION

In the last decades machine learning systems are widely spreading in different academic domains, as well as many public and private sectors are increasing the exploitation of these systems. Their widespread and pervasiveness is mainly driven by the exponential growth of computational power and the extensive availability of large amounts of data [27]. Supervised machine learning models are also particularly widespread and now deeply rooted in different sectors due to their usage versatility. The predictive ability of supervised machine learning systems is deployed in disparate areas of application: credit reliability [27], justice system [2, 7], job recommendations [39], university selection process [25], cultural contents [38],[23] and purchases recommendations [35]. The key ingredient that supervised machine learning models have in common is the availability of a set of labeled data used to train the model in elaborating a response related to past events [17]. Since the known properties of the available set of data is used to create a classifier that makes predictions about new entities of the same type, the structure, properties and quality of the data are aspects that largely and directly influence the quality of the model and the results it produces [34], [1]. Although data-driven decision models have been shown to produce both economic and social benefits, many researchers have highlighted several problems and damages related to their use in different areas, especially if they are built on partial or incomplete data [20], [12]. As a matter of fact, in recent years several studies have found a convergence of issues related to

the ethics and transparency of these systems in the process of data collection and in the way they are recorded [30]. While the process of rigorous data collection and analysis is fundamental to the design of the model, this step is still largely overlooked by the machine learning community [6, 24]. As the practice of removing protected attributes from available data has been shown to potentially exacerbate further discrimination [44] - making bias even more difficult to detect - practices related to data collection, data transparency and data explainability become even more relevant and urgent. The aim of our work is to provide a data annotation system that serves as a diagnostic framework containing immediate information about the data appropriateness in order to more accurately assess the quality of the available data used in training models. We propose a data annotation method based on Bayesian statistical inference that aims to warn of the risk of discriminatory results of a given data set. In particular, our method aims to deepen the statistical knowledge related to the information contained in the available data, and to promote awareness of the sampling practices used to create the training set, highlighting that the probability of a discriminatory result is strongly influenced by the structure of the available data. We test our data annotation systems on three dataset widely spread in machine learning community: the COMPAS dataset [21], the Drug Consumption dataset [15], [16] and the Adult dataset [28].

1.1 Problem Statement

The majority of machine learning systems are based on historical data processing [32]. This is particularly true in supervised machine learning models. Several studies have shown evidence that many equity and discrimination issues are due to input data properties [5]. Most of today data sets used to train models are chosen through non-probabilistic methods, generating problems of data imbalance and representativeness [14, 32]. This means that different fractions of the population do not show the same opportunity to be represented within the sample - aka, training sets -, leading some groups of individuals to have a lower probability of being represented. Common observed effects of a bad sampling are underestimation and overestimation of some groups [4]. Undetected distortions in data may also easily represent a spurious statistical noise. This happens when the data structure induces dependence between two variables that are not linked by a real cause-effect relationship.

Data Sampling. A key moment in the pipeline of a machine learning model is when the programmed algorithm is supplied with training data representing the entities on which the model itself trains its knowledge to make predictions. The quality of the data used in this phase is fundamental for the desired result, according to the principle of "garbage in - garbage out": even the most sophisticated models can present distorted results in the presence of low quality data [42]. One of the main causes of data distortion is the way the data is selected and provided to the algorithm displaying problems related to inaccuracy, lack of update or inadequate representativeness. However, while knowledge of bias typologies has proliferated over the years, less attention is paid to issues concerning data collection, notation and sampling [31]. In the spirit of fostering a broader awareness of data handling, we provide a reasoned list of issues that may arise during this phase:

- i *Data Selction*: the large proliferation of data sets availability on the same kind of problem to be analyzed, make hard the a priori choice of a given data set [17];
- ii *Inadequate sampling methods*: most models are trained with data sets that have been "found" and not subjected to probabilistic sampling methods, leading to limited or no data control [3];
- iii *Cost and Time Limit*: collecting large amounts of data that present proportional representations of each property with respect to a sensitive attribute is time consuming and often costly and labor-intensive [8];
- iv *Data set validation*: in the design of a machine learning model, more attention is paid to the mathematical basis of the classifier, restricting the data formation process to a black box [18, 19];
- v *Validation planning*: data validation, when applied, is often performed only after the model has been trained and used, making the feedback cycle inefficient and often ineffective [22];
- vi *Lack of statistical rigorousness*: the suitability of the data set varies depending on the task for which the data are prepared. For instance, models based on linear regression imply assumptions of normality on the measurement error [17, 43]. This specificity is often absent in the pipeline of machine learning models.

Miss-dependency. Two-dimensional or bivariate statistics is the study of the degree to which two distinct characters of the same statistical unit are connected. However, the connection only measures the degree of statistical dependency without inducing a cause-effect relationship or dependency between the variables. For instance, it can be shown that people with small feet make more spelling mistakes than people with large feet. However, this statistical dependency does not indicate that having small feet is the cause of spelling mistakes; the greater frequency of spelling mistakes may in fact be due to the younger age of people with small feet. In this case there could be a third variable, age, responsible for the cause-effect relationship. While in a human-centered model - where the human makes the decisions - this distinction is quite evident, in a machine learning model miss-dependency is not always deductible. This depends on two reasons: i) the machine does not recognize the meaning of the instance but looks at the properties of the variables; ii) the way in which the data are structured modifies the interpretation that the machine is having regarding the relation of statistical dependency. This means that, while in a human-centered model it is the human to verify that the relationships of statistical dependence detected in the available data are leading to a cause-effect relationship, in machine learning models the machine is not always able to recognize a spurious connection, erroneously assigning to two or more variables a cause relationship. In other words, the structure of the available data is responsible for the successful or failed relationships established with the protected attributes (ethnicity, gender, etc.) in the data. In addition, the rapid growth and spread of current machine learning systems is due in part to the ease of design of the models themselves, which thanks to modern software allows the construction of predictive models avoiding the

understanding and adoption of rigorous statistical analysis. The simplicity of design has therefore created a gap between predictive and analytical-explicative power, favoring misinterpretation between causality and statistical dependence. The distinction between statistical dependence and causal dependence in data is therefore a primary issue in machine learning models, especially to determine the causes of failure, potential biases encoded in the data and the reliability of application.

Based on the problems highlighted, our contribution aims to answer the following research questions:

RQ1 Is it possible to establish the probability of composition of the training data from the available data set?

RQ2 Do the available data known to the machine learning community present a discriminatory future risk based on their structure?

2 BACKGROUND

When machine learning model decisions are based on historical records, they tend to embed distortions that exist in reality and crystallize them. Prejudices and human bias therefore become part of the technology itself. This is particularly evident with regard to ethnic discrimination. Over the last years, the rise of machine learning models in various sectors is leading to a dramatic increase of discriminatory outcomes for ethnic minorities, across different fields of application. A striking and well known case is the COMPAS software, used in U.S. court to estimate the probability of defendants' recidivism, which has been shown to underestimate the risk of recidivism for white defendants and overestimate it for black defendants [21]. However, the COMPAS case is not an isolated phenomenon. In a 2017 experiment conducted on the Airbnb platform, applications from guests with typically African American names were found to be 16% less likely to be accepted than identical guests with typically white names [13]. Also in 2017, a geo-statistical analysis revealed that the design of the popular Pokémon GO game strengthens existing geographical prejudices, for example by benefiting urban areas and neighborhoods with smaller minority populations, economically disadvantaging ethnic minority areas [10]. Several studies have demonstrated the discriminatory potential of targeting advertising [41], [40], which is only recently receiving interventions to remove the prejudicial content of the model. For example, Facebook after years of scandals related to ads that exclude people based on race [2] has finally removed the racial targeting option for ads [29]. In a 2019 study, the commercial algorithm widely used in the U.S. health care system to guide health care decisions was found to discriminate against black patients [33]. The algorithm falsely assigned a healthier condition to black patients despite the risk of complications being the same for white patients, making black people less likely to receive more financial resources for extra care. Although facial recognition technologies are now used in several domains, they still present many discriminatory issues related to differences in margins of error - generally software has a 20% higher margin of recognition error for black women [36] -. As an example, we report what happened recently with Google Vision AI, a computer vision service for image labeling [26]. By providing the system with two images of people holding a body temperature thermometer, it labeled the image containing

the white person as an "electronic device", while in the image containing the black person the device held was labeled as a "gun". In a later experiment it was shown that it was sufficient to apply a pink mask on the black person's hand in order the software labeled the image as "tool". Racial bias encoded in machine learning systems is likely to spread silently and like wild fire in everyday technologies. The increasing and ubiquitous spread of such models also intended to make allocative decisions about people's lives makes the problem of prejudice and racial discrimination more urgent than ever. For this reason and for the historical moment we are experiencing, our work intends to focus on rational discrimination in data.

3 MOTIVATING EXAMPLE

Given a population composed of 60% Caucasians, 35% black people and 15% Asian people, the probability of positive outcome for the respective ethnic groups is 70% for Caucasians, 20% for Blacks and 60% for Asians. What is the probability of failure with respect to the protected attribute Ethnicity?

In this example the probabilities are given rather than the numerosity in order to simplify the following notation. To offer a better understanding of the Methodology this data will be used in Section 4. The data gives the probability of success, but the similar reasoning is also valid for cases where the probability of failure is known. The intent is to verify whether the probabilities of success or failure of a subgroup are influenced by group membership - and vice versa - and more specifically how these probabilities affect the composition of the training set.

4 METHODOLOGY

Our data annotation system is based on four modules:

- I Dependence:** assesses the degree of connection among the protected attribute - in our study, the ethnicity - and the target variable;
- II Diverseness:** provides the training diversification probability in respect to each level of the protected attribute and the target variable;
- III Inclusiveness:** provides the probability that two properties are simultaneously included in the training set;
- IV Training Likelihood:** provides the occurrence likelihood of the protected attribute levels given the target variable levels - and vice versa - before the training set is sampled.

4.1 Quantifying Dependence

Excluding some specific domains where the dependence of some protected attributes with the response variable is not considered problematic, but rather it is fundamental for the understanding of a certain problem (for example the gender attribute in the medical field in the detection of particular diseases [11]), in the broad field of machine learning systems the dependence between the protected attribute and the response variable has caused severe consequences [32, 34]. The dependence between the protected attribute and the response variable is therefore one of the major causes of discrimination and as such must be rigorously examined. The first step for a correct bias detection within the data is given by the dependency analysis between the different modalities of a protected attribute and the response variable. In statistics, the measurement of the

degree of dependence of two qualitative variables is called contingency; contingency measures the degree of connection of two categorical variables. To determine the degree of connection, the marginal frequencies and the combined frequencies of the bivariate table are used. Given two categorical variables x_i and y_j , the dependency or independence is established through the theoretical independence table $f'(x_i, y_j)$ once the table of the observed real data $f(x_i, y_j)$ is given. The contingency $C(x_i, y_j)$ is therefore given by the difference between the observed and theoretical frequencies:

$$C(x_i, y_j) = f(x_i, y_j) - f'(x_i, y_j) \quad (1)$$

If the table of the observed real data and the theoretical table of independence coincide - that is if for each cell the value is null - then the two variables are independent. Otherwise, it is necessary to measure the degree of connection between the variables. The degree of connection between two categorical variables is commonly measured by the Pearson connection index, obtained as the sum of the relative quadratic contingencies. The index assumes a value of zero in case of independence in distribution and increases as the degree of connection between variables increases:

$$\chi^2 = \sum_{i,j} \frac{C^2(x_i, y_j)}{n_{i,j}} = n \left(\sum_{i,j} \frac{n_{i,j}^2}{n_{i0}n_{0j}} \right) \quad (2)$$

In order to support Pearson's connection index, the contingency coefficient is adopted with the purpose of reducing the χ^2 in the range [0;1]:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \quad (3)$$

However, the effect size of the degree of connection between two categorical variables is not always easy to interpret, where by effect size we mean a quantitative measure of the magnitude of a phenomenon. To offer a better understanding of the relationship of dependency between two variables, several simplified methods of interpretation have been proposed, especially to guide social scientists in the interpretation of statistical test results. In the spirit of simplifying the interpretation of the dependency between the response variable and the protected categories for a data set user, we introduce the concept of the Effect Size Index w (ES w):

$$w = \sqrt{\sum_{i=1} \frac{(P_{1i} - P_{0i})^2}{P_{0i}}}, \quad (4)$$

where p_{0i} and p_{1i} are the value of the i th cells. Notice that unlike the contingency coefficient, the ES w is not derived from frequencies but from proportions. The relationship between the Pearson connection index, the contingency coefficient and the ES Index is given by the following formula:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{w^2}{w^2 + 1}} \quad (5)$$

Alternatively to the Formula 4 it is also possible to calculate the ES w from the contingency coefficient:

$$w = \sqrt{\frac{C^2}{1 - C^2}} \quad (6)$$

The size of the ES w between two variables is then evaluated through the use of Table 1, which relates the magnitude of the ES with a nominal label. The advantage of using the conventional

Table 1: Conventional definitions of Effect Size Index w magnitude

Magnitude	Value
SMALL	$w = 0.1$
MEDIUM	$w = 0.3$
LARGE	$w = 0.5$

conversion table for the user of the data set is that the magnitude of the dependency is displayed quickly and immediately without the need for more complex statistical tests.

4.2 Estimating Diverseness

Intuitively, the probability of an event represents how likely the event will occur. According to the classical definition the probability is given by the following ratio:

$$P = \frac{\text{number of favorable cases}}{\text{number of possible cases}} \quad (7)$$

We now apply this elementary theory to the problem of data collection in machine learning. When the data set is partitioned into training and test sets, a split with a more or less standard ratio (70/30 or 80/20) is generally performed, i.e. a sampling is performed on the available data. Let's consider the case in which the training data set is generated by random sampling on the original data set without considering further techniques (stratification or re-sampling) - for example in the case of a non expert user -. The probability an event occurs turns into the probability that the training set shows some existing properties contained in the original data set:

$$P = \frac{\text{number of favorable properties}}{\text{number of possible properties}} \quad (8)$$

In our data annotation this ratio is introduced to allow the dataset user to answer questions like: "If I perform a random sampling on the original dataset, what is the probability that the training set is mainly composed of positive examples? What is the probability of belonging to a certain group with respect to the target variable?"

Prior Probabilities. The a priori probability of a data property is the degree of belief of the property in the absence of other information, also known as the unconditional probability. The degree of belief is the probability of a property to be true in an uncertain environment. The probability is referred to the belief and not to the truth of the fact, as it is not possible for the user to know exactly the truth, that is if the original data are representative of the real world. Since the user does not have access to the complete information, several hypotheses on how the real data is structured have to be drawn, assigning to each of them a probability of being true. Formally:

$$\begin{aligned} P &= (Y = y) \\ P &= (A = a) \end{aligned} \quad (9)$$

We estimate the prior probabilities by using the data of the problem introduced in Section 3, where the target variable Y assumes value

1 in case of negative outcome, otherwise 0. In this specific case,

Table 2: Example of prior probabilities

Formula	Probability
$P(Y = 0)$	$P = 0.48$
$P(Y = 1)$	$P = 0.52$
$P(A = \text{white})$	$P = 0.6$
$P(A = \text{black})$	$P = 0.35$
$P(A = \text{Asian})$	$P = 0.15$

the prior probabilities indicate that the training set has probability 0.48 to be composed by individuals who display a positive outcome and 0.52 to be composed by individuals who display a negative outcome; finally, the probabilities that it is formed by individuals of white, black and Asian ethnicity are respectively 0.6, 0.35 and 0.15 (Table 2).

4.3 Estimating Inclusiveness

Posterior Probabilities. Given two events A and B, the probability $P(A|B)$ is said posterior probability because it allows to calculate the probability of A, knowing that B occurred. In our case the posterior probability means to compute the probability that $Y = y$, knowing that $A = a$ has occurred (and vice versa). In other words, the probability that the training set shows the property $Y = y$, knowing the property $A = a$ has occurred (and vice versa). We start by estimating the probability that two events occur simultaneously. From the definition of conditional probability:

$$\begin{aligned} P(A = a \cap Y = y) &= P(A = a)P(Y = y|A = a) \\ P(Y = y \cap A = a) &= P(Y = y)P(A = a|Y = y) \end{aligned} \quad (10)$$

Since from Compound Probability Theorem [37] $P(A = a \cap Y = y)$ is equal to $P(Y = y \cap A = a)$, i.e. the probability of both properties occurring is the same, either of the two formulas can be employed indistinctly.

Table 3: Example of properties occurring simultaneously

Formula	Probability
$P(Y = 0 \cap A = \text{white})$	$P = 0.42$
$P(Y = 0 \cap A = \text{black})$	$P = 0.07$
$P(Y = 0 \cap A = \text{Asian})$	$P = 0.09$
$P(Y = 1 \cap A = \text{white})$	$P = 0.18$
$P(Y = 1 \cap A = \text{black})$	$P = 0.28$
$P(Y = 1 \cap A = \text{Asian})$	$P = 0.06$

4.4 Estimating Training Likelihood

From the definition of conditional probability, we derive the Bayes Theorem for the properties of the training set:

$$\begin{aligned} P(A = a|Y = y) &= \frac{P(A = a)P(Y = y|A = a)}{P(Y = y)} \\ P(Y = y|A = a) &= \frac{P(Y = y)P(A = a|Y = y)}{P(A = a)} \end{aligned} \quad (11)$$

In the case of binary classification and in the case of protected attributes we are in the presence of a certain event partition. This means that the events are disjointed from each other $Y_i \cap Y_j = \emptyset$ and $A_i \cap A_j = \emptyset$ if $i \neq j$ and that as a whole they are the only ones possible, i. e., if a certain property occurs, one and only one certainly appeared. In other words, it is not possible that the training set is composed of individuals who belong simultaneously to the black and white ethnic group, or who simultaneously show a positive and negative outcome. The union of the occurrence of the single properties is therefore the whole set of possible properties. For the properties outcome and ethnicity the generalization formula are respectively:

$$\begin{aligned} \Omega : \cup_{i=1}^N Y_i &= \Omega, \text{ hence } \sum_{i=1}^N P(Y_i) = P(\cup_{i=1}^N Y_i) \\ \Omega : \cup_{i=1}^N A_i &= \Omega, \text{ hence } \sum_{i=1}^N P(A_i) = P(\cup_{i=1}^N A_i) \end{aligned} \quad (12)$$

By applying Formulas 10 and 12 the Bayes Theorem can be generalized for each property of the training set:

$$\begin{aligned} P(Y = y|A) &= \frac{P(Y = y)P(A|Y = y)}{P(A)} = \frac{P(Y = y)P(A|Y = y)}{\sum_{i=1}^N P(A|Y_i)P(Y_i)} \\ P(A = a|Y) &= \frac{P(A = a)P(Y|A = a)}{P(Y)} = \frac{P(A = a)P(Y|A = a)}{\sum_{i=1}^N P(Y|A_i)P(A_i)} \end{aligned} \quad (13)$$

The first equation in Formula 13 derives the probability of the outcome property given the ethnic property, while the second equation derives the probability of the ethnic property given the outcome property. In other words, it derives the probability of composition of the training set based on the posterior probabilities of the outcome and ethnicity properties. Carried out a random sampling on the original data, the Formula answers the following questions:

- In the sampled training set what is the probability of belonging to an ethnic group with respect to the outcome variable?
- In the sampled training set what is the probability of obtaining a certain outcome with respect to the ethnic group?

Complementarily, the two equations can be interpreted as the probability of bias within the training set.

Table 4: Example of posterior probabilities

Formula	Probability
$P(Y = 0 A = \text{white})$	$P = 0.7$
$P(Y = 0 A = \text{black})$	$P = 0.2$
$P(Y = 0 A = \text{Asian})$	$P = 0.6$
$P(Y = 1 A = \text{white})$	$P = 0.3$
$P(Y = 1 A = \text{black})$	$P = 0.8$
$P(Y = 1 A = \text{Asian})$	$P = 0.4$
$P(A = \text{white} Y = 1)$	$P = 0.34$
$P(A = \text{white} Y = 0)$	$P = 0.87$
$P(A = \text{black} Y = 1)$	$P = 0.53$
$P(A = \text{black} Y = 0)$	$P = 0.15$
$P(A = \text{Asian} Y = 1)$	$P = 0.11$
$P(A = \text{Asian} Y = 0)$	$P = 0.18$

5 CASE STUDIES DATASETS

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions)¹ is a popular tool used by U.S. court to estimate the defendants' probability of recidivism. This dataset displays the probability of reoffending based on two year of further studies. The dataset has been shown to underestimate the risk of recidivism for white defendants and overestimate it for black defendants [21].

Drug Consumption [15, 16] contains information on the consumption of 18 drugs based on personality traits and socio-economic attribute. For simplicity of analysis we assumed the consumption of Cannabis as target variable but the annotation of the dataset can be made on each target drug.

Adult Dataset [28] The data set contains adult income annual census from the US Census Bureau. It is commonly employed in forecasting tasks in order to predict the factors leading to income below or above \$50,000.

Table 5: Summary of Datasets Prominent Properties

Property	COMPAS	Drug Consumption	Adult Dataset
Size	6172x9	1885x31	48842x15
Target variable	0 → no 1 → yes	0 → non user 1 → user	0 → > 50K 1 → ≤ 50K
Levels of ethnicity attribute	Asian Black Caucasian Hispanic NA ^c Other	Asian Black Black/Asian Caucasian White/Asian White/Black Other	AIE ^a API ^b Black Caucasian Other

^a American-Indian/Eskimo, ^b Asian-Pac-Islander, ^c Native American

6 RESULTS AND DISCUSSION

We performed the analyses that constitute our data annotation system for each of the datasets presented in Section 5. Sub-sections 6.1, 6.2, 6.3 and 6.4 report the analysis for each module - dependency, diverseness, inclusiveness, training likelihood, respectively - and contain an example graphic module. Figures 4 and 5 shows an illustrative example of the graphical visualization for the complete notation.

6.1 Dependence

This module aims to analyze the connection relationships between the protected attribute Ethnicity and the target variable that are established and depend on the available data. For instance, for the COMPAS dataset the module highlights the dependency relationships between recidivism and different ethnic minorities. Summary results for dependence module are shown in Table 6.

Table 6: Summary of Dependence Prominent Properties

	COMPAS	Drug Consumption	Adult Dataset
Contingency coefficient	0.1413	0.1558	0.0994
Effect size w variable	0.1427	0.1578	0.0999
Magnitude of Effect size w	SMALL	SMALL	VERY SMALL

None of the three datasets displays worrying dependency values among the protected attribute Ethnicity and the target variable, showing the magnitude of the Effect Size w as small or very small. However, the results of the COMPAS dataset - which is proven to contain bias - indicate that this module alone is not sufficient to show a latent bias risk. The degree of bias depends on the sample size and the value of the contingency coefficient of the target variable and the protected attribute [45]. Smaller samples lead to more bias and higher variance [46] and therefore the results of the dependency must be analyzed in relation to the amount of data available. In order to facilitate the interpretation of the connection

Dependence	
	VERY SMALL
	Range [0,1]
Contingency coefficient	< 0.1
Effect size w	0 < w < 0.1

Dependence	
	SMALL
	Range [0,1]
Contingency coefficient	< 0.3
Effect size w	0.1 ≤ w < 0.3

¹Retrieved from: <https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis>

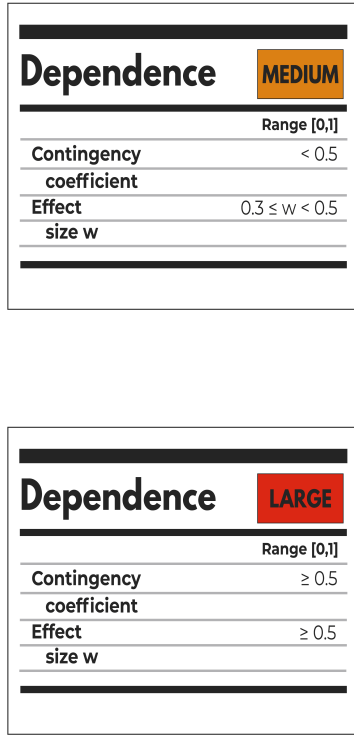


Figure 1: Example of Dependence graphic visualization

relations, we propose a graphic notation for dependence. Figure 1 shows the graphical representations of the dependency modules based on different connection magnitude.

6.2 Diverseness

This module aims to analyze the diverseness of the data available by estimating prior probabilities. They determine the probability that training set will display an a priori environment based on the original data available, i.e. they show the probability of training set composition stratified by each of target variable and protected attribute levels. For example, in our case study the module highlights the probability that training set will be equally composed by ethnic minorities and ethnic majorities. Summary results for diverseness module are shown in Table 7. In terms of target variable probabilities, the results show strong distortions for the Drug Consumption and Adult datasets with a high probability of positive examples - i.e. showing a negative outcome - while the probabilities of the COMPAS dataset are quite homogeneous. Regarding the probabilities of the protected attribute ethnicity, the distortions are even more pronounced than the target variable ones, revealing a very high probability of composition for the Caucasian ethnicity in the Drug Consumption and Adult datasets. In the case of the COMPAS dataset the probabilities are indeed distorted, although still not such as to predict at this point of the analysis more severe future distortions, which is why more in-depth analysis are required. Figure 4.2 shows the graphical representation of the diverseness module that simplifies the display of prior probabilities. In the example is given

Table 7: Summary of Diverseness Analysis Results

	COMPAS	Drug Consumption	Adult Dataset
0	0.545	0.329	0.239
1	0.455	0.671	0.761
Caucasian	0.341	0.912	0.855
Black	0.514	0.018	0.096
Asian	0.005	0.014	
Hispanic	0.082		
Native American	0.002		
Other	0.056	0.033	0.008
White/Black		0.011	
White/Asian		0.011	
Black/Asian		0.002	
Amer-Indian-Eskimo			0.010
Asian-Pac-Islander			0.031

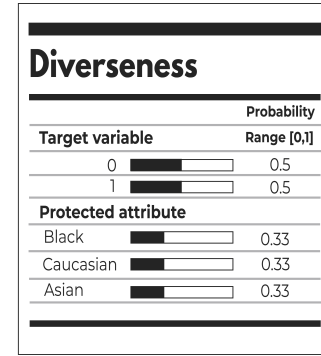


Figure 2: Example of Diverseness graphic visualization

the notation for a dataset where both the levels of the target variable and those of the protected attribute ethnicity are equiprobable.

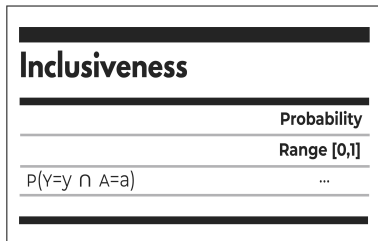
6.3 Inclusiveness

This module aims to analyze the inclusiveness of the data available by estimating the simultaneously probabilities. They determine the probability that training set will simultaneously display two by two the target variable and the protected attribute properties. For instance, in our case study the module highlights the probability that in training set the property Asian appears simultaneously with property success. Summary results for diverseness module are shown in Table 8. The results of this module show that the probability that two properties will occur simultaneously is related to the sample size. Evidence of this can be found in the results of the Drug Consumption and Adult datasets, where the highest probabilities of simultaneous events involve the Caucasian property. The COMPAS dataset shows quite homogeneous probabilities especially with regard to the Black property, while for the Caucasian property the highest probabilities are related to the simultaneous occurrence

Table 8: Summary of Inclusiveness Analysis Results

	COMPAS	Drug Consumption	Adult Dataset
0∩AIE ^a			0.0006
0∩Asian	0.0023	0.0019	
0∩API ^b			0.0041
0∩Black	0.1514	0.0023	0.0057
0∩Black/Asian		0.0000	
0∩Caucasian	0.1281	0.0555	0.1061
0∩Hispanic	0.0320		
0∩NA ^c	0.0006		
0∩Other	0.0219	0.0013	0.0005
0∩White/Asian		0.0004	
0∩White/Black		0.0006	
1∩AIE			0.0042
1∩Asian	0.0008	0.0007	
1∩API			0.0111
1∩Black	0.1661	0.0010	0.0412
1∩Black/Asian		0.0003	
1∩Caucasian	0.0822	0.1165	0.3115
1∩Hispanic	0.0189		
1∩NA	0.0005		
1∩Other	0.0124	0.0050	0.0036
1∩White/Asian		0.0016	
1∩White/Black		0.0014	

^a American-Indian/Eskimo, ^b Asian-Pac-Islander, ^c Native American

**Figure 3: Example of Inclusiveness graphic visualization**

with the Non-recidivist property. Since the simultaneous probabilities depend on the number of examples within the available data and the sample size, this result alone is not sufficient to establish a priori the certain presence of serious data distortions, although some evidence can already be seen. Figure 3 shows the graphical representation of the inclusiveness module that simplifies the display of simultaneously probabilities. In the example is given the notation for a dataset where all the properties of the target variable and those of the protected attribute ethnicity are equiprobable.

6.4 Training Likelihood

This module aims to analyze the training likelihood of the data available by estimating the posterior probabilities. They determine

the probability that in the training set the occurrence of the properties of the protected attribute is given by the properties of the target variable - and vice versa -. For example, in the COMPAS dataset they determine the probability that the occurrence of reoffending is given by the properties of the protected attribute ethnicity. Summary results for training likelihood module are shown in Table 9.

The results of this module show that the posterior probabilities of target variable and protected attribute ethnicity are quite skewed in all dataset. In the case of the Adult dataset given as occurred event 1 or event 0, the probability of occurrence of the Caucasian ethnic group is respectively 0.908 and 0.839, - i.e. very high for both events - while the probabilities of all other ethnic groups conditioned to the target variable are all significantly lower; this means that the original data contain many examples of individuals belonging to the Caucasian ethnic group. In the case of Drug Consumption, a similar reasoning can be carried out for the ethnicity probabilities conditioned to the target variable; moreover, notice that given the property Black/Asian, the probability of occurrence of event 1, i. e. that the individual is a consumer, is 1 - while the probability of 0 is 0 - which means that in the available data there are no examples of individuals belonging to the ethnic group Black/Asian showing a positive outcome - i. e. negative examples -. Figures 4 and 5 shows the graphical visualization of our data annotation system for the COMPAS dataset. The analysis of the COMPAS dataset shows that if an individual is randomly sampled from the original data for the training set, the probability that this individual is black knowing that the re-offending property has occurred - i.e. knowing the outcome of the re-offending event - is 0.591, while the probability that the individual is white knowing that the re-offending property has occurred is 0.293. Instead, given as occurred the property Black the probability that the individual has not reoffended is 0.477, while the probability that the individual has reoffended is 0.523; given the property Caucasian, the probability that the individual has not reoffended is 0.609, while the probability that the individual has reoffended is 0.391, that is significantly lower. This means that in this dataset the reoffending is related to ethnicity, and that success or failure are determined by the membership to a specific ethnic group. The differences in probability between the properties highlight the risk of future bias, and in the case of the COMPAS dataset they anticipate the underestimation of recidivism for the Caucasian ethnic group and the overestimation of recidivism for the Black ethnic group proven in recent studies [21].

6.5 Final Remarks

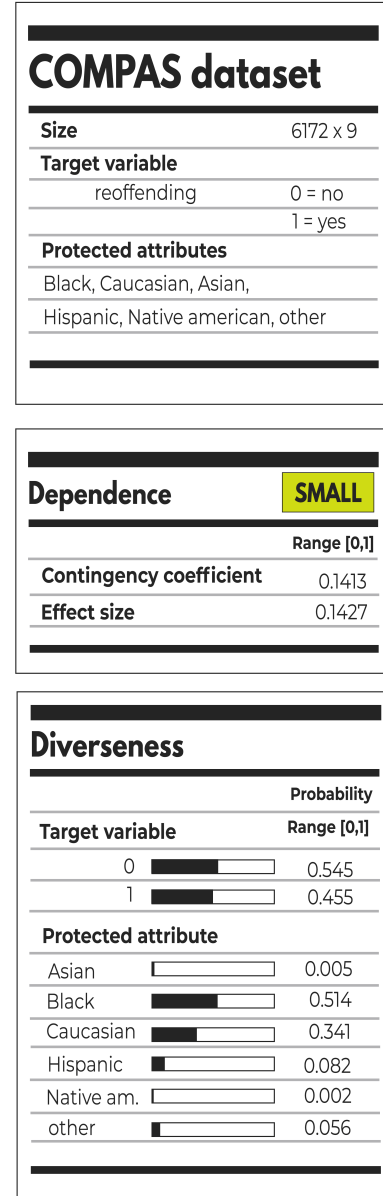
RQ1: in traditional sampling practices, instead of observing all the units of a population, only a subset of a population is detected, which must show certain probabilistic characteristics. In machine learning models the training set is sampled not from the real population but from the available data. While in classical sampling the empirical knowledge alone is effectively of a sample nature, in machine learning systems the available data are often of sample nature too, precisely due to the fact that it is not possible to make assumptions on the real population. Considering a random sampling from the

Table 9: Summary of Training Likelihood Analysis Results

	COMPAS	Drug Consumption	Adult Dataset
0 AIE ^a			0.117
0 Asian	0.742	0.731	
0 API ^b			0.269
0 Black	0.477	0.697	0.121
0 Black/Asian		0.000	
0 Caucasian	0.609	0.323	0.254
0 Hispanic	0.629		
0 NA ^c	0.545		
0 Other	0.638	0.206	0.123
0 White/Asian		0.200	
0 White/Black		0.300	
1 AIE			0.883
1 Asian	0.258	0.269	
1 API			0.731
1 Black	0.523	0.303	0.879
1 Black/Asian		1.000	
1 Caucasian	0.391	0.677	0.746
1 Hispanic	0.371		
1 NA	0.455		
1 Other	0.362	0.794	0.877
1 White/Asian		0.800	
1 White/Black		0.700	
AIE 0			0.005
AIE 1			0.011
Asian 0	0.007	0.031	
Asian 1	0.003	0.006	
API 0			0.035
API 1			0.030
Black 0	0.450	0.037	0.048
Black 1	0.591	0.008	0.111
Black/Asian 0		0.000	
Black/Asian 1		0.002	
Caucasian 0	0.381	0.895	0.908
Caucasian 1	0.293	0.921	0.839
Hispanic 0	0.095		
Hispanic 1	0.067		
NA 0	0.002		
NA 1	0.002		
Other 0	0.065	0.021	0.004
Other 1	0.044	0.040	0.010
White/Asian 0		0.006	
White/Asian 1		0.013	
White/Black 0		0.010	
White/Black 1		0.011	

^a American-Indian/Eskimo, ^b Asian-Pac-Islander, ^c Native American

available data, we have shown that the probability of composition of the training set can be predicted, highlighting that the structure of the data directly affects the probability of properties distribution;

**Figure 4: Data annotation visualization for COMPAS dataset**

RQ2: we analyzed three datasets frequently accessed by machine learning community. Of these, all three showed more or less pronounced distortions for the protected attribute Ethnicity. Although the COMPAS dataset is the sole one that has been shown to discriminate against black people, the Drug Consumption and Adult datasets reveal possible future bias in the detriment of ethnic minorities.

7 RELATED WORK

Although there are a number of papers that for ethical purposes deal with data annotation they are all very recent, indicating that

Inclusiveness	
	Probability
	Range [0,1]
$P(\text{Asian} \cap 0)$	0.0023
$P(\text{Asian} \cap 1)$	0.0008
$P(\text{Black} \cap 0)$	0.1514
$P(\text{Black} \cap 1)$	0.1661
$P(\text{Caucasian} \cap 0)$	0.1281
$P(\text{Caucasian} \cap 1)$	0.0822
$P(\text{Hispanic} \cap 0)$	0.0320
$P(\text{Hispanic} \cap 1)$	0.0189
$P(\text{Native american} \cap 0)$	0.0006
$P(\text{Native american} \cap 0)$	0.0005
$P(\text{other} \cap 0)$	0.0219
$P(\text{other} \cap 1)$	0.0124

Training likelihood	
	Probability
	Range [0,1]
$P(\text{Caucasian} 1)$	0.293
$P(\text{Caucasian} 0)$	0.381
$P(0 \text{Caucasian})$	0.609
$P(1 \text{Caucasian})$	0.391
$P(\text{Black} 1)$	0.591
$P(\text{Black} 0)$	0.450
$P(0 \text{Black})$	0.477
$P(1 \text{Black})$	0.523

Figure 5: Data annotation visualization for COMPAS dataset

this field of study is still partially explored and has only recently received considerable attention. Our contribution differs from the others because it induces a probabilistic reasoning on the causes of model discrimination based on sampling problems; our intention is to deepen the knowledge of data validation analysis, focusing on the meaning of probabilities. From a graphical point of view, our work has been inspired by the Data Nutrition Labels [22], a data labeling system mainly based on descriptive data statistics. A similar approach is addressed in [6], where an operational framework is proposed to identify the bias risks of automatic decision systems. In [17] the authors propose a data labeling system based on discursive data sheets. In [9] the authors propose a collaborative crowdsourcing system to improve the quality of the labels.

Since ethically data annotation represent a quite new field of study, there are several works that provide different types of labels. We believe that at present the focus should not be on achieving a unified data annotation system in the short term, but rather on the fact that the fair machine learning community is working together to focus attention on the data collection problem. Especially because awareness of data issues is often not rooted outside of this community. It is important that this field and this work inspire greater awareness of the possible causes of discrimination due to the fundamental ingredient that all users and designers of machine learning systems (from the most to the least experienced) use, data.

8 CONCLUSIONS

The purpose of the current study was to detect the potential race discriminatory risk for future machine learning system by providing a data annotation system based on Bayesian Inference. Our notation serves as a diagnostic framework to immediately visualize data appropriateness and potential bias occurring when sampling the training set from an available dataset. The investigation of the probabilities of the training set sampling has shown that it is possible to establish a risk of future bias by observing prior and posterior probabilities of the ethnicity and target variable properties. The empirical findings in this study provide a new perspective on data annotation practices by showing that Bayesian inferences may reveal the risk of bias in three different widespread dataset. Furthermore, this study has raised important questions about the awareness of most widely data sampling practices in machine learning community. The findings of this investigation complement those of earlier studies. Our data annotation system is limited to the binary case and to the analysis of categorical variables for classification tasks. This would be a fruitful area for further work. Our intent is to expand the work in the following directions: i) extend the notation to multiple protected attributes - the probabilities of the training set will then be given by the vectors of the protected attribute combinations - ; ii) extend the notation to the non-binary case - for prediction tasks involving regression analysis for example - ; iii) extend the probabilistic notation to non-labeled data.

REFERENCES

- [1] Aws Albarghouthi and Samuel Vinitzky. 2019. Fairness-Aware Programming. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (Atlanta, GA, USA) (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 211–219. <https://doi.org/10.1145/3287560.3287588>
- [2] Julia Angwin and Terry Jr. Parris. 2016. *Facebook Lets Advertisers Exclude Users by Race*. ProPublica. Retrieved September 12, 2020 from <https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race>
- [3] Abolfazl Asudeh, Assessing Jin, Remedying Coverage for a Given Dataset, and Hosagrahar Visvesvaraya Jagadish. 2019. Assessing and Remedying Coverage for a Given Dataset. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, New Jersey, US, 554–565.
- [4] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2018. Fairness and Machine Learning. <http://www.fairmlbook.org>.
- [5] Ruha Benjamin. 2019. Assessing risk, automating racism. *Science* 366, 6464 (2019), 421–422. <https://doi.org/10.1126/science.aaz3873> arXiv:<https://science.sciencemag.org/content/366/6464/421.full.pdf>
- [6] Elena Beretta, Antonio Santangelo, Bruno Lepri, Antonio Vetró, and Juan Carlos De Martin. 2019. The Invisible Power of Fairness. How Machine Learning Shapes Democracy. In *Advances in Artificial Intelligence, Proceedings of 32nd Canadian Conference on Artificial Intelligence, Canadian AI 2019 (Kingston, ON, Canada)*, Marie-Jean Meurs and Frank Rudzicz (Eds.), Vol. 11489. Springer, Cham, Germany, 238–250. https://doi.org/10.1007/978-3-303-03055-9_19
- [7] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. Fairness in Criminal Justice Risk Assessments: The State of the Art.

- Sociological Methods & Research* 50, 1 (2018), 3–44. <https://doi.org/10.1177/0049124118782533>
- [8] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186. <https://doi.org/10.1126/science.aal4230> arXiv:<https://science.sciencemag.org/content/356/6334/183.full.pdf>
- [9] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) (CHI '17). Association for Computing Machinery, New York, NY, USA, 2334–2346. <https://doi.org/10.1145/3025453.3026044>
- [10] Ashley Colley, Jacob Thebault-Spieker, Allen Yilun Lin, Donald Degraen, Benjamin Fischman, Jonna Häkkinen, Kate Kuehl, Valentina Nisi, Nuno Jardim Nunes, Nina Wenig, Dirk Wenig, Brent Hecht, and Johannes Schöning. 2017. The Geography of Pokémon GO: Beneficial and Problematic Effects on Places and Movement. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) (CHI '17). Association for Computing Machinery, New York, NY, USA, 1179–1192. <https://doi.org/10.1145/3025453.3025495>
- [11] D. Dahiawade, G. Patle, and E. Meshram. 2019. Designing Disease Prediction Model Using Machine Learning Approach. , 1211–1215 pages.
- [12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (Cambridge, Massachusetts) (ITCS '12). Association for Computing Machinery, New York, NY, USA, 214–226. <https://doi.org/10.1145/2090236.2090255>
- [13] Benjamin Edelman, Michael Luca, and Dan Svirsky. 2017. Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment. *American Economic Journal: Applied Economics* 9, 2 (April 2017), 1–22. <https://doi.org/10.1257/app.20160213>
- [14] Virginia Eubanks. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, Inc., USA.
- [15] Elaine Fehrman, Vincent Egan, and Evgeny M. Mirkes. 2015. UCI Machine Learning Repository. <http://archive.ics.uci.edu/ml>
- [16] Elaine Fehrman, Awaz K. Muhammad, Evgeny M. Mirkes, Vincent Egan, and Alexander N. Gurban. 2017. The Five Factor Model of Personality and Evaluation of Drug Consumption Risk. , 231–242 pages. https://doi.org/10.1007/978-3-319-55723-6_18
- [17] Timnit Gebru, Jamie Morgenstern, WJennifer Vecchione, Brianna Vaughan, Hanna Wallach, Hal III Daumé, and Kate Crawford. 2018. Datasheets for Datasets. arXiv:arXiv:1803.09010
- [18] R. Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage in, Garbage out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From?. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) (FAT* '20). Association for Computing Machinery, New York, NY, USA, 325–336. <https://doi.org/10.1145/3351095.3372862>
- [19] Yolanda Gil, Cédric H. David, Ibrahim Demir, Bakinam T. Essawy, Robinson W. Fulweiler, Jonathan L. Goodall, Leif Karlstrom, Huikyo Lee, Heath J. Mills, Ji-Hyun Oh, Suzanne A. Pierce, Allen Pope, Mimi W. Tzeng, Sandra R. Villamizar, and Xuan Yu. 2016. Toward the Geoscience Paper of the Future: Best practices for documenting and sharing research from data to software to provenance. *Earth and Space Science* 3, 10 (2016), 388–415. <https://doi.org/10.1002/2015EA000136> arXiv:<https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2015EA000136>
- [20] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Barcelona, Spain) (NIPS'16). Curran Associates Inc., Red Hook, NY, USA, 3323–3331.
- [21] Jeff Harry Thornburg Larson, Surya Mattu, and Lauren Kirchner. 2016. *Machine Bias*. ProPublica. Retrieved September 2, 2020 from <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- [22] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. *CoRR* abs/1805.03677 (2018), 21 pages. arXiv:1805.03677 <http://arxiv.org/abs/1805.03677>
- [23] K. Indira and M. K. Kavithadevi. 2019. Efficient Machine Learning Model for Movie Recommender Systems Using Multi-Cloud Environment. *obile Networks and Applications* 24, 6 (2019), 1872–1882. <https://doi.org/10.1007/s11036-019-01387-4>
- [24] Eun Seo Jo and Timnit Gebru. 2020. Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) (FAT* '20). Association for Computing Machinery, New York, NY, USA, 306–316. <https://doi.org/10.1145/3351095.3372829>
- [25] Sumitkumar Kanoje, Debajyoti Mukhopadhyay, and Sheetal Girase. 2016. User Profiling for University Recommender System Using Automatic Information Retrieval. In *Procedia Computer Science*, Vol. 78. Elsevier, Netherlands, 5–12. <https://doi.org/10.1016/j.procs.2016.02.002> 1st International Conference on Information Security & Privacy 2015.
- [26] Nicolas Kayser-Bril. 2020. *Google apologizes after its Vision AI produced racist results*. AlgorithmWatch. Retrieved August 17, 2020 from <https://algorithmwatch.org/en/story/google-vision-racism/>
- [27] Jon Kleinberg. 2018. Inherent Trade-Offs in Algorithmic Fairness. In *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems* (Irvine, CA, USA) (SIGMETRICS '18). ACM Press, New York, NY, USA, 40–40. <https://doi.org/10.1145/3219617.3219634>
- [28] Ronny Kohavi and Barry Becker. 1996. UCI Machine Learning Repository. <http://archive.ics.uci.edu/ml>
- [29] Joe Kukura. 2020. *Facebook (Finally) Removes Racial Ad Targeting*. SFist. Retrieved September 12, 2020 from <https://sfist.com/2020/08/31/facebook-finally-removes-racial-ad-targeting/>
- [30] Vidushi Marda and Shivangi Narayan. 2020. Data in New Delhi's Predictive Policing System. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) (FAT* '20). Association for Computing Machinery, New York, NY, USA, 317–324. <https://doi.org/10.1145/3351095.3372865>
- [31] Daniel McDuff, Roger Cheng, and Ashish Kapoor. 2018. Identifying Bias in AI using Simulation. arXiv:arXiv:1810.00471
- [32] Safiya Umoja Noble. 2018. *Algorithms of oppression: How search engines reinforce racism*. NYU Press, New York, NY, USA.
- [33] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453. <https://doi.org/10.1126/science.aax2342> arXiv:<https://science.sciencemag.org/content/366/6464/447.full.pdf>
- [34] Cathy O'Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, New York.
- [35] Oladapo Oyebode and Rita Orji. 2020. A hybrid recommender system for product sales in a banking environment. , 11 pages. <https://doi.org/10.1007/s42786-019-00014-w>
- [36] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. 2020. Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA) (AI/ES '20). Association for Computing Machinery, New York, NY, USA, 145–151. <https://doi.org/10.1145/3375627.3375820>
- [37] Sheldon M Ross. 1996. *Stochastic processes*. Wiley, New Jersey, US. <https://books.google.de/books?id=ImUPAQAAIAAJ>
- [38] Markus Schedl, Hamed Zamani, Ching-Wei Chen, Yashar Deldjoo, and Mehdi Elahi. 2018. Current challenges and visions in music recommender systems research. *International Journal of Multimedia Information Retrieval* 7, 2 (2018), 95–116. <https://doi.org/10.1007/s13735-018-0154-2>
- [39] Z. Siting, H. Wenxing, Z. Ning, and Y. Fan. 2012. Job recommender systems: A survey. In *2012 7th International Conference on Computer Science Education (ICCSE)*. IEEE Xplore Digital Library, New York, 920–924.
- [40] Lin Song. 2020. Two-Sided Price Discrimination by Media Platforms. *Marketing Science* 39, 2 (2020), 317–338. <https://doi.org/10.1287/mksc.2019.1211> arXiv:<https://doi.org/10.1287/mksc.2019.1211>
- [41] Toll Speicher, Muhammad Ali, Giridhari Venkatadri, Filipe Nunes Ribeiro, George Arvanitakis, Fabrício Benevenuto, Krishna P. Gummadi, Patrick Loiseau, and Alan Mislove. 2018. Potential for Discrimination in Online Targeted Advertising. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, New York, NY, USA, 5–19. <http://proceedings.mlr.press/v81/speicher18a.html>
- [42] Tatiana Tommasi, Patricia Novi, Barbara Caputo, and Tinne Tuytelaars. 2017. *A Deeper Look at Dataset Bias*. Csukka G. (eds) Domain Adaptation in Computer Vision Applications. Advances in Computer Vision and Pattern Recognition, Springer, Cham, Swiss. 37–55 pages. https://doi.org/10.1007/978-3-319-58347-1_2
- [43] Yan Wang and Xuelei Sherry Ni. 2017. Predicting Class-Imbalanced Business Risk Using Resampling, Regularization, and Model Emsembling Algorithms. *International Journal of Managing Information Technology (IJMIT)* 11, 1 (2017), 15 pages. <https://ssrn.com/abstract=3366806>
- [44] Betsy A. Williams, Catherine F. Brooks, and Yotam Shmargad. 2018. How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications. *Journal of Information Policy* 8 (2018), 78–115. <https://www.jstor.org/stable/10.5325/jinfopoli.8.2018.0078>
- [45] Yao Zhou, M. Isabel Vales, Aoxue Wang, and Zhiwu Zhang. 2017. Systematic bias of correlation coefficient may explain negative accuracy of genomic prediction. *Briefings Bioinform* 18, 5 (2017), 744–753. <https://doi.org/10.1093/bib/bbw064>
- [46] Donald W. Zimmerman, Bruno D. Zumbo, and Richard H. Williams. 2017. Bias in estimation and hypothesis testing of correlation. *Psicología* 24, 1 (2017), 133–158.